Wheat Crop Yield Estimation using Geomatics Tools in Saharanpur District

Ankush Kumar Gupta¹, Pramod Soni²

ABSTRACT
Background: For a primarily agriculture-based country like India, reliable, accurate and timely information on types of crops grown, their acreages and yield are of vital importance. In this study, attempts have been made to predict wheat crop yield in Saharanpur district of Uttar Pradesh for the year 2016-17 using various spectral indices (Normalized Difference Vegetation Index, Soil Adjusted Vegetation Index, Ratio Vegetation Index and Transformed Normalized Difference Vegetation Index) and agrometeorological variables.

Methods: The spectral indices were derived using agrometeorological data (Growing Degree Days, Temperature Difference and Helio Thermal Units) and the ground truth information was collected using the Global Positioning System in the field. Observed acreage data for the wheat crop were collected from the Saharanpur district headquarter.

Result: The NDVI showed the highest correlation with wheat crop yield. The spectral yield model using only the NDVI showed an R² value of 0.81 with an RMSE of 88.02 kg/ha. However, by incorporating the NDVI with temperature, the model shows much better performance with an R² value of 0.95 and an RMSE of 49.86 kg/ha. The rainfall showed lesser correlation with the wheat crop yield as compared to the temperature, probably due to ground water dominated region.

Key words: Crop estimation, Spectral index, Remote sensing, Regression.

INTRODUCTION
The agricultural crop production is usually estimated as a product of area under the crop and the average yield per unit area of the crop. There are several methods for crop yield estimation. The traditional method is based on Crop Cutting Experiments (CCE), which involve sample surveys for the crop estimation. The crop production estimates are obtained by taking the product of crop acreage and the corresponding crop yield. The method is laborious, intensive, expensive and imprecise (Sukhatme and Panse, 1951).

Remote-sensing data acquired by satellite have a wide scope for agricultural applications owing to their synoptic and repetitive coverage. Since the early 1970s, Remote Sensing approach has shown great potential in agricultural domains all around the world for improving the agricultural statistics. In the past, spectral data acquired via satellite have been extensively utilized for crop yield modeling in various parts of the world. Dubey et al. 1994 found that yield obtained from crop-cutting sites was found to be linearly related to Ratio Vegetation Index (RVI) derived from Landsat MSS data of corresponding sites in Ludhiana and Patiala districts of Punjab. Prasad et al., 2006 considered NDVI, soil moisture, surface temperature and rainfall data of Iowa state, US, for 19 years for crop yield assessment and prediction using piecewise linear regression method with break-point. The predicted values were found to be very close to the observed values (R² = 0.78 for corn and R² = 0.86 for soybean). Rojas, 2007 reported the development of an operational Spectro-agrometeorological yield model for maize using a spectral index, the NDVI derived from Spot-vegetation, meteorological data obtained from the European Centre for Medium-Range Weather Forecast (ECMWF) model and crop-water status indicators estimated by the Crop-Specific Water Balance model (CSWB) for the six large maize-growing provinces in Kenya. When the Jack-knife resampling technique was applied the forecast capability of the model was improved with (R² = 0.81 and RMSE = 0.359 t/ha). Wheat yield prediction using different agrometeorological indices, spectral index NDVI and trend predicted yield (TPY) were developed in Hoshiarpur and Rupnagar districts of Punjab by Bazgeer et al. 2006. It was found that Agromet-Spectral-Trend-Yield model could explain 96% (SE = 87 kg/ha) and 91% (SE = 146 kg/ha) of wheat yield variations for Hoshiarpur and Rupnagar districts, respectively.

Verma et al., 2011 predicted wheat yield in various districts of Haryana state for the period 2001-02 to 2007-08 with a linear regression model that showed an R² value of 0.81 and standard error of 192.73. Iqbal et al. 2012...
constructed a statistical model of wheat yield in the Punjab province, Pakistan by using multi-regression analysis. Bazgeer et al. 2007 used different meteorological variables with agrometeorological indices for wheat yield prediction in Hamedan district during 2003-04 and 2004-05. Verma et al., 2016 developed zonal weather models (R² of 0.81) for district-level mustard yield estimation on an agro-climatic zone basis in Haryana State based on five weather variables, i.e., maximum temperature, minimum temperature, rainfall, relative humidity and sunshine hours.

The objective of the present study is to prepare and evaluate a suitable statistical model for crop yield prediction using remote sensing data for Saharanpur district of Uttar Pradesh.

MATERIALS AND METHODS

Study area

The present study was conducted over the Saharanpur district of Uttar-Pradesh, India (Fig 1). The district is situated between 29°34’45” and 30°21’30” N latitude and 77°09’00” and 78°14’45” E longitude having an area of 3,689 km². According to 2011 census, with a total population of 34,66,382, Saharanpur district is at 24th position in Uttar Pradesh state. Urban population in the district is 30.8% as compared to the state average of 22.3%. The net sown area in the district is about 2.742 lakh ha out of which 2.55 lakh ha is irrigated. The major source of economy for the district is agriculture with sugarcane as the main cash crop. The other main crops raised in the district include wheat, maize, rice, potato and tomato. Wheat, gram, barley, potatoes and winter vegetables are harvested during the *Rabi* season, whereas paddy, maize, cotton and pulses are harvested during the *Kharif* season. Saharanpur district comes under tropical atmosphere due to the proximity of the Himalayan region.

Data used

Different meteorological parameters, satellite data and ground truth data were collected from various sources. The collected information helped in the identification of the best period for the remote sensing data acquisition and development of wheat crop prediction. Digital boundary map of Saharanpur district was prepared using the Geographic Information System. Remote sensing data from the Landsat-7 (2006-2013) and Landsat-8 (2013-2016) satellites were collected for the last ten years (2006-2016) during the flowering stage (maximum vegetation growth) of wheat crop (mainly the month of March) from United States Geological Survey (USGS). Total 11 images were uses for the present study.

Ground truth information (40 locations, 28th March 2017) were also collected rigorously during the flowering stage of the wheat crop (Fig 2). In wheat dominant areas, fieldwork was carried out and ground control points were collected using the Global Positioning System (GPS). A total of 40 location were collected for ground truthing on 28th march, 2017. Crop information and knowledge collected during the fieldwork facilitated mapping of wheat crop. The wheat crop

---

Wheat Crop Yield Estimation using Geomatics Tools in Saharanpur District
prevailing areas were located and marked in the satellite data. These points were then used to decide the Vegetation Index (VI) ranges for the wheat and were used as an input in the classification.

The various agrometeorological data, like rainfall, maximum temperature, minimum temperature and sunshine hours, were collected from various government agencies and websites. Table 1 lists the various meteorological and wheat crop yield data for the district.

### Methodology

Initially, the district boundary mask of Saharanpur was generated using the topographic map with a scale 1:250,000 in Q-GIS. False Colour Composite (FCC) using spectral data from band-2 (Green), band-3 (Red) and Band-4 (near infrared) was generated in ERDAS (Earth Resources Data Analysis System) IMAGINE software. Further, the ground truth information obtained from the GPS was used to identify the wheat crop plots. The GPS was taken to the plots and location of the plots in terms of longitude and latitudes were recorded. These locations were then identified on the FCC’s. The coordinates of each plot in terms of scan-line and column number were recorded to identify these plots on the normalized difference vegetation indices (Table 2) and area was calculated as given in Table 3.

Land use land cover (LULC) of the district were classified into six categories: (1) Wheat crop (2) Other Crop, (3) Urban, (4) Forest (5) Water and (6) Sand. Supervised classification was carried out by preparing the separability curve. The bands 3, 5 and 6 showed the maximum separability; hence, they were used for supervised classification using maximum likelihood classification method. The accuracy assessment was carried out with a threshold of 85%. The total area for each class was calculated by counting the number of pixels and multiplying it with the spatial resolution of the image (30m x 30m). The relative deviation of the wheat crop was estimated by comparing it with the actual data obtained from the district headquarters.

Finally, the linear regression techniques were applied to formulate a suitable model. In the first step, the correlation analysis was performed to find out the preliminary relationship between spectral indices/agrometeorological and wheat crop yield. Next, the step-wise linear regression was performed to find the best set of variables that can effectively predict the wheat crop yield. In each step of regression, the variable having the non-significant coefficients (at a 95% Confidence Interval) were removed from further steps of regression. The general equation of regression is given below as:

\[
\text{Yield} = b_1 + b_2 \times T_{\text{max}} + b_3 \times T_{\text{min}} + b_4 \times \text{Rain} + b_5 \times SH + b_6 \times GDD + b_7 \times TD + b_8 \times HTU + b_9 \times \text{NDVI}
\]

Here, \(b_1\) to \(b_9\) are the regression coefficients and, \(T_{\text{min}}\) = Minimum Temperature, \(T_{\text{max}}\) = Maximum Temperature, \(\text{Rain}\) = Rainfall, \(SH\) = Sunshine Hours, \(GDD\) = Growing degree days, \(TD\) = Temperature difference, \(HTU\) = Helio-thermal units, \(PTU\) = Photo thermal units.

### Results and Discussion

The various agrometeorological variables and spectral indices from the year 2006 to 2016 are shown in Table 4. The ground truth information about wheat crop yield was not available for the year 2016-17.

### Table 2: Definitions of Spectral Indices.

| Indices                  | Description                                                                 | Formula                                                                 |
|--------------------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------|
| Growing Degree Days (GDD)| Explains the relationship between crop growth and temperature              | GDD = \(\sum b_i \left[ \frac{T_{\text{max}} + T_{\text{min}}}{2} \right] - T_b \) |
| Temperature difference (TD) | The difference in daily maximum and minimum temperature                      | TD = \(\sum b_i \left( T_{\text{max}} + T_{\text{min}} \right) \)         |
| Photo Thermal Units (PTU) | Product of GDD and sunshine hours potentially influenced by region geographical width (N) | PTU = \(\sum \frac{b_i}{a_i} \left( \text{GDD} \times N \right) \)         |
| Helio Thermal Units (HTU) | Product of GDD and actual sunshine hours (n)                                 | HTU = \(\sum \frac{b_i}{a_i} \left( \text{GDD} \times n \right) \)         |
| Normalized Difference Vegetation Index (NDVI) | A vegetation index to delineate vegetation from soil | NDVI = \( \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}} \) |
| Soil-adjusted Vegetation Index (SAVI) | This index is very helpful to minimize the effect of soil background.        | SAVI = \( \frac{(1+L)(\text{NIR}-\text{RED})}{\text{NIR}+\text{RED}+L} \) |
| Ratio Vegetation Index (RVI) | Separate green vegetation from soil background using landsat MSS imagery | RVI = \( \frac{\text{NIR}}{\text{RED}} \) |
| Transformed Normalized Difference Vegetation Index (TNDVI) | It has higher coefficient of assurance for the similar variable and this is the distinction amongst TNDVI and NDVI | TNDVI = \( \sqrt{\text{NDVI}} \) |
Land use and land cover classification

Fig 3 shows the classified image of Saharanpur district. We can see that the wheat crop covers most of the district with some urban area in the central part and forests in the Northern part of the district.

Accuracy assessment was also performed after supervised classification. The classified image is accepted only when accuracy is above 85%. The Accuracy assessment report is shown in Table 5. In the table, the accuracy assessment is based on the comparison of two maps; one based on the analysis of remote sensing data known as classified map and second based on information derived from actual ground also known as the reference map. Reference total is true class on reference map and classified total is total class on classified map. Based on this, the accuracy of the wheat crop was found to be 85.71% with the overall average accuracy of 81.22%.

Acreage estimation and relative deviation

For each Vegetation Index (VI), using the respective range for the wheat crop, the thematic map is prepared with two layers, as shown in Fig 4. Wheat pixels were collected using the attribute table and the wheat acreage was calculated by multiplying the total number of pixels with the spatial resolution (30x30) of the input Image. The estimated acreage using various VIs are shown in Table 5. The NDVI shows the least relative deviation of 5.97, whereas TNDVI shows the maximum deviation of 9.79.

Correlation analysis

Initially, correlation coefficients between wheat yield and Spectral indices/agrometeorological variables were
obtained, as shown in Fig 5. From the figure, we can see that the crop yield is strongly correlated to NDVI and minimum temperature, having correlation coefficients of 0.89 and -0.71, respectively. The yield has a feeble dependency on sunshine hours and temperature difference. Although rainfall does play an important role in the crop yield production, it is not very effective in the Saharanpur district as crop yield shows a small positive correlation of 0.54 with rainfall, probably due to availability of other water sources in the district.

Stepwise linear regression

Step-1 of regression

In the first step, all eight variables were included and multivariate linear regression was performed. The general equation of the regression is given below:

\[ \text{Yield} = b_1 + b_2 \times T_{\text{min}} + b_3 \times T_{\text{max}} + b_4 \times \text{Rain} + b_5 \times \text{SH} + b_6 \times \text{GDD} + b_7 \times \text{TD} + b_8 \times \text{HTU} + b_9 \times \text{NDVI} \]

The best-fitted equation of regression is shown below.

\[ \text{Yield} = \left\{ 401.4 - 50.32 \times T_{\text{min}} + 12.99 \times T_{\text{max}} + 7.69 \times \text{Rain} + 10.45 \times \text{SH} \right\} + 23.96 \times \text{GDD} + 21.93 \times \text{TD} + 1.09 \times \text{HTU} + 2348.7 \times \text{NDVI} \]

The plot of observed and fitted wheat crop yield is shown in Fig 6 and the coefficients along with their uncertainty are given in Table 6. The t-stat is the coefficient divided by its standard deviation in the regression and the p-value

\[ \text{Table 6: The relative deviation of vegetation indices.} \]

| VI    | Predicted acreage (hectare) | Actual acreage (hectare) | Relative deviation (%) |
|-------|-----------------------------|---------------------------|------------------------|
| NDVI  | 124813.79                   | 117849.13                 | 5.97                   |
| SAVI  | 125649.9                    | 117849.13                 | 8.84                   |
| TNDVI | 126986.87                   | 117849.13                 | 9.79                   |
| RVI   | 115314.36                   | 117849.13                 | 7.63                   |
represents the significance of the coefficient. The t-stat value of more than 1 and p-value of less is 0.05 is considered a good significance of the coefficient and corresponding variable.

With all the variables included in the regression, the model shows an RMSE of about 28.6 kg/ha and an $R^2$ value of 0.993. However, from the Table 7, we can see that the coefficients $b_5$ and $b_6$ (coefficients of SH and GDD) have very small t-stat (less than 0.1) and stand out of other variables, therefore, in the next step of regression, they were removed from the analysis.

### Step-2 of regression

In the second step, six variables were included after discarding SH and GDD and regression was performed. The general equation for the regression is:

$$
Yield = b_1 + b_2 \times T_{\text{min}} + b_3 \times T_{\text{max}} + b_4 \times \text{Rain} + b_5 \times TD + b_6 \times HTU + b_7 \times NDVI
$$

After performing the linear regression, the best-fitted equation is given below.

$$
Yield = \left\{ 379.13 \times T_{\text{min}} + 18.21 \times T_{\text{max}} + 7.69 \times \text{Rain} + 26.49 \times TD + 1.71 \times HTU + 2348.7 \times NDVI \right\}
$$

### Table 7: Regression coefficients statistics in Step-1.

| Coefficient | Estimate  | SE       | t-Stat    | p-Value |
|-------------|-----------|----------|-----------|---------|
| $b_1$       | 401.39698 | 75.97500 | 5.28328   | 0.01323 |
| $b_2$       | -50.32348 | 106.23607| -0.47369  | 0.66806 |
| $b_3$       | 12.99614  | 105.24146| 0.12349   | 0.90953 |
| $b_4$       | 7.69645   | 3.40553  | 2.25998   | 0.10893 |
| $b_5$       | 10.45805  | 151.03080| 0.06924   | 0.94915 |
| $b_6$       | 23.96104  | 274.27769| 0.08736   | 0.93589 |
| $b_7$       | 21.93600  | 4.57156  | 4.79836   | 0.01722 |
| $b_8$       | 1.09442   | 8.81967  | 0.12409   | 0.90909 |
| $b_9$       | 2348.67312| 189.39856| 12.40069  | 0.00113 |

![Fig 6: Step-1 of stepwise regression.](image)

![Fig 7: Step-2 of stepwise regression.](image)
The plot of observed and fitted crop yield is shown in Fig 7 and the coefficients with their uncertainty statistics are shown in Table 8.

By excluding the SH and GDD from regression, the model shows an improved RMSE of about 24.85 and \( R^2 \) value of 0.993. However, from Table 8, we can see that the coefficients \( b_1, b_4 \) and \( b_6 \) (Constant, Rain and HTU) have small t-stat and p-value more than 0.05. Thus, they are not significant at a 95% Confidence Level; therefore, in the next step of regression, these variables were removed from the analysis.

**Table 8: Regression coefficients statistics in Step-2.**

| Coefficient | Estimate | SE    | t-Stat | p-Value |
|-------------|----------|-------|--------|---------|
| \( b_1 \)   | 379.1349 | 288.0667 | 1.3161 | 0.2585 |
| \( b_2 \)   | -35.8990 | 7.5810 | -4.7354 | 0.0091 |
| \( b_3 \)   | 18.2112 | 6.1801 | 2.9468 | 0.0421 |
| \( b_4 \)   | 7.6998 | 2.9513 | 2.6089 | 0.0986 |
| \( b_5 \)   | 26.4949 | 3.9438 | 6.7181 | 0.0026 |
| \( b_6 \)   | 1.7018 | 0.7937 | 2.1442 | 0.0986 |
| \( b_7 \)   | 2348.4196 | 164.1243 | 14.3088 | 0.0001 |

**Step-3 of regression**

In the third step, only four variables were included after discarding the constant term, Rain and HTU and regression was performed. The general form of the regression equation is:

\[
\text{Yield} = b_1 \times T_{\text{min}} + b_2 \times T_{\text{max}} + b_3 \times TD + b_4 \times NDVI
\]

The best-fitted equation is given below.

\[
\text{Yield} = 37.66 \times T_{\text{min}} + 25.51 \times T_{\text{max}} +34.21 \times TD + 2728.5 \times NDVI
\]

The plot of observed and fitted crop yield is shown in Fig 8 and the coefficients with their uncertainty statistics are shown in Table 9.

Even by incorporating only four variables, the model shows an RMSE of about 49 and an \( R^2 \) value of 0.95. Hence it explains the importance of NDVI and temperature in the estimation of wheat crop yield for Saharanpur district. The p-value of NDVI is the smallest and hence has the most significant coefficient in the regression. To further check the sole dependence of NDVI on wheat crop yield, another

**Table 9: Regression coefficients statistics in Step-3.**

| Coefficient | Estimate | SE    | t-Stat | p-Value |
|-------------|----------|-------|--------|---------|
| \( b_1 \)   | -37.667351 | 8.553209 | -4.403885 | 0.003142 |
| \( b_2 \)   | 25.516836 | 2.825259 | 9.031679 | 0.000042 |
| \( b_3 \)   | 34.217667 | 6.024787 | 5.679481 | 0.000751 |
| \( b_4 \)   | 2728.501353 | 245.578004 | 11.110528 | 0.000011 |

**Fig 8: Step-3 of Stepwise Regression.**

**Fig 9: The Spectral Yield Model.**
Table 10: Spectral Yield Model statistics.

| Coefficient | Estimate | SE  | t-Stat | p-Value |
|-------------|----------|-----|--------|---------|
| b₁          | 2009.41  | 153.97 | 13.051 | 0.0000012 |
| b₂          | 2399.56  | 413.52 | 5.8028 | 0.000403  |

The spectral yield model was evaluated by incorporating just the NDVI index.

**Spectral yield model**

In the spectral yield model, a linear relationship between crop yield and NDVI was assumed.

\[ \text{Yield} = b_1 + b_2 \times \text{NDVI} \]

The best-fitted regression equation is given below:

\[ \text{Yield} = 2009 + 2399.56 \times \text{NDVI} \]

The plot of observed and fitted crop yield is shown in Fig 9 and the coefficients with their uncertainty statistics are shown in Table 10.

We found an RMSE of about 88 and \( R^2 \) of 0.808 by just incorporating NDVI. Hence although NDVI is the most important index for crop yield estimation, the performance of the model was improved by including temperature and its indices.

**SUMMARY AND CONCLUSION**

The major conclusions of the present study are:

I. The acreage obtained from the NDVI showed the lowest Relative Deviation of 5.97% from actual acreage compared to other indices.

II. The stepwise regression revealed that Rainfall has a nonsignificant relationship in wheat crop yield prediction due to the dominance of groundwater sources availability.

III. The spectral yield model with only NDVI shows an \( R^2 \) value of 0.808 and RMSE of 88. However, by incorporating the minimum and maximum temperature along with temperature difference index, the \( R^2 \) is improved to 0.95 and RMSE to 49.

**ACKNOWLEDGEMENT**

The authors would like to thank Saharanpur district headquarter for providing various agrometeorological data.

We would also like to thank USGS for providing land use land cover data.

**REFERENCES**

Bazgeer, S., Mahey, R.K., Sharma, P.K., Sood, A. and Sidhu, S.S. (2006). Pre-harvest wheat yield prediction using agromet-spectral-trend-yield models for Hoshiarpur and Rupnagar districts of Punjab. Journal of the Indian Society of Remote Sensing. 34(3): 269-277.

Bazgeer, S., Kamali, G. and Mortazavi, A. (2007). Wheat yield prediction through agrometeorological indices for Hamedan, Iran. Desert. 12(1): 33-38.

Dubey, R.P., Ajwani, N., Kalubarme, M.H., Sridhar, V.N., Navalgund, R.R., Mahey, R.K. and Na Rang, R.S. (1994). Pre-harvest wheat yield and production estimation for the Punjab, India. International Journal of Remote Sensing. 15(10): 2137-2144

Iqbal, M.J., Ali, Z.U. and Ali, S.S. (2012). Agroclimatic modelling for estimation of wheat production in the Punjab Province, Pakistan. Proceedings of the Pakistan Academy of Sciences. 49(4): 241-249.

Prasad, A.K., Chai, L., Singh, R.P. and Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. International Journal of Applied Earth Observation and Geoinformation. 8(1): 26-33

Rojas, O. (2007). Operational maize yield model development and validation based on remote sensing and agro meteorological data in Kenya. International Journal of Remote Sensing. 28(17): 3775-3793

Sukhatme, P.V. and Panse, V.G. (1951). Crop surveys in India-II. Journal of the Indian Society of Agricultural Statistics. 3(2): 95-168.

Verma, U., Dabas, D.S., Hooda, R.S., Kalubarme, M.H., Yadav, M., Grewal, M.S. and Prawasi, R. (2011). Remote sensing based wheat acreage and spectral-trend-agrometeorological Yield Forecasting: Factor Analysis Approach. Statistics and Applications. 9(1): 1-13.

Verma, U., Piepho, H.P., Goyal, A., Ogutu, J.O. and Kalubarme, M.H. (2016). Role of climatic variables and crop condition term for mustard yield prediction in Haryana. Int. J. Agric. Stat. Sci. 12: 45-51.