Principle Component analysis for Forecasting of pre-harvest Rapeseed and Mustard yield based on meteorological parameters

A B S T R A C T

The pre-harvest forecast model was developed using time series data on Rapeseed and Mustard yields as well as weekly data on six weather variables from the 40th SMW of one year to the 8th SMW of the next year, which covers the years 1990–1991 to 2014–2015. Multiple regression and principal component analysis are two statistical techniques that have been reported for creating pre-harvest forecast models. The most accurate model found by applying step-by-step regression analysis to weekly weather data was based on adj R^2, RMSE, and%SE. One and a half months prior to harvest, these models can be utilised to obtain a trustworthy forecast of the yield of Rapeseed and Mustard.

Key word: Meteorological parameters, Crop yield, multiple regression Principle Component function analysis, Forecast model

Introduction

The remarkable narrative of Indian agriculture is well-known around the world for its multi-functional success in producing employment, a means of subsistence, and ecological, nutritional, and food security. In 2014–15, agriculture and related industries contributed 11% to the GDP (Gross Domestic Product) [Economic survey (2015–16)].It remains India's largest economic sector and contributes significantly to the country's overall socio-economic growth, despite a steady fall in its share of the GDP.

Rapeseed and mustard are significant crops that offer high-quality food and nutritional security for eradicating malnutrition among the underprivileged. They are crucial in preserving natural resources, which are necessary for sustainable agriculture, and boosting soil fertility. Oil is a preferred crop of small land owners due to its many applications and contribution to sustainable agriculture. Among the different grain legumes farmed, legumes are known to provide food proteins in the developing world and are typically grown on risk-prone marginal areas with modest inputs.With the aforementioned point and debate, it is clear that the weather variable, specifically the

meteorological parameters, are crucial to two stages of the cultivation of rapeseed and mustard for their final production/productivity.

An approach to reduce dimensions using multiple variables is principal components analysis. Let's say there are twelve correlated variables. Our 12 measurements could be condensed to a few principle components using principal component analysis. In this case, getting the component scores—variables that are added to the data set—and/or examining the data's dimensionality may be of the utmost interest. We might state that two dimensions in the component space account for 68% of the variation, for instance, if two components are extracted and those two components account for 68% of the overall variance. Principal components analysis is not typically employed to find underlying latent variables, in contrast to factor analysis.The loadings onto the components are therefore not understood as factors as they would be in a factor analysis. As in this example, principal components analysis, like factor analysis, can be carried out on raw data as well as on a matrix of correlation or covariance. When using raw data, the technique will produce the original covariance or correlation matrix, depending on what the user specifies. The variables are standardised when the correlation matrix is employed, and the total variance will match the number of variables used in the analysis (because each standardised variable has a variance equal to 1). The variables will keep their original metric if the covariance matrix is applied. Yet it's important to pick variables whose variances and scales are comparable. In contrast to factor analysis, which examines the common variance, a principal component analysis' original matrix examines the overall variance. Principal component analysis also makes the assumption that there were no measurement errors in the original measurements.

Materials and Methods:

Yield data

From the Bulletins of the Directorate of Agricultural Statistics and Crop Insurance, Government of Uttar Pradesh, we have collected time series data on yield of Rapeseed & Mustard for Sultanpur district for 25 years (1990-91 to 2014-15).

Development of Statistical forecast models:-

Let PC_1, PC_2, ..., PC_k be first k (k< p) principal components explaining variability up to 90 percent. Using these k principal components as regressor variables and crop yield (y) as dependent, the following pre-harvest forecast models are proposed.

Model-P_1:

Six unweighted weather indices representing six different weather variables have been employed in this technique as p=6 variables in the main component analysis. Let's say that, according to loading, the first k principle components from principal component analysis are the most important ones because they account for more than 75% of the total variance. As a result, the yield was used as the response in the multiple linear regression model together with these first k principal components and the trend variable (T). The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

where Y is the crop yield, $\beta_i^{'s}(i = 0,1,2,\ldots.k)$ and δ are model parameter, PC_1, PC_2,.....PC_k are principal components, T is the trend variable and eis error term assumed to follow normal distribution with mean 0 and variance σ^2.

Model-P_2:

In this model, principal component analysis was utilised to weight six weather variables into six weighted weather indices. According to loading, the first k principal components are the most significant ones, explaining more than roughly 75% of the total variance, according to the principal component analysis. As a result, the yield was used as the regressand in the multiple linear regression model, along with the trend variable (T) and these first k principal components. The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

where the notations stand as usual as described in model-1.

Model-P_3:

All 42 weather indicators (21 weighted and 21 unweighted) were included in this approach as 42 variables (p=42) in principal component analysis. These variables included 6 weighted weather indices, 15 weighted interaction indices, and 6 unweighted weather indices. If the first k (kp) principal components, which are the most important ones according to loading, have explained more than roughly 75% of the total variance, then these first k principal components along with the trend variable (T) were used as the regressors and the yield as the response in a multiple linear regression model. The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

where the notations stand as usual as described in model-1.

Model-P_4:

In this model, the p=12 variables for the main component analysis were 6 weighted and 6 unweighted weather indices of 6 weather variables. Let's say that, according to loading, the first k principal components from principal component analysis are the most important ones, explaining more than roughly 75% of the overall variance. As a result, the yield was used as the response in the multiple linear regression model together with these first k principal components and the trend variable (T). The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

where the notations stand as usual as described in model-1.

Model-P_5:

Principal component analysis was used in this model to combine six unweighted and fifteen unweighted interactions between weather indices from six different weather variables. Let's say that, according to loading, the first k principal components from principal component analysis are the most important ones, explaining more than roughly 75% of the overall variance. As a result, the yield was used as the response in the multiple linear regression model together with these first k principal components and the trend variable (T). The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

where the notations stand as usual as described in model-1.

Model-P_6:

Principal component analysis was used in this model to combine six weighted weather indices and fifteen weighted interactions of six weather variables. Let's say that, according to loading, the first k principal components from principal component analysis are the most important ones, explaining more than roughly 75% of the overall variance. As a result, the yield was used as the response in the multiple linear regression model together with these first k principal components and the trend variable (T). The type of model taken into account is as follows:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots\ldots + \beta_k PC_k + \delta T + e$$

Measures for validation and comparison of models:

Percent Deviation:

This calculates the forecast's percentage departure from the yield data's actual value. The following formula can be used to get the percent forecast deviation:

$$\textit{Percentage deviation} = \frac{(\text{actual yield} - \text{forecasted yield})}{(\text{actual yield})} \times 100$$

R^2 (Coefficient of Determination):

It is in general used for checking the adequacy of the model. R^2 is given by the following formula;

$$R^2 = 1 - \frac{ss_{res}}{ss_t}$$

Where ss_{res} and ss_t are the residual sum of square and the total sum of square respectively.

Percent Standard Error of the Forecast (CV):

Let $\hat{y}_f$ be forecast value of crop yield and X_0 be the vector of selected values for regressor variables for the yield is forecasted.

The variance of $\hat{y}_f$ as given in (Draper and Smith, 1998) is obtained as

$$V(\hat{y}_f) = \hat{\sigma} X_0^{'} (X^{'} X)^{-1} X_0$$

Where $X^{'}X$ is the dispersion matrix of the sum of square and cross products of regressor variables used for the fitting the model and $\hat{\sigma}^2$ is the estimated residual variance.

The percent standard error (PSE) of forecast yield $\hat{y}_f$ is given by

$$PSE = \frac{\sqrt{V(\hat{Y_f})}}{Fore\cos tyield} \times 100$$

In fact, the PSE is the coefficient of variation (C.V.) of forecast yield.

Root Mean Square Error (RMSE):

It is also a measure of comparing two models. The formula of RMSE is given below

$$RMSE = \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} (O_i - E_i)^2 \right\} \right]^{\frac{1}{2}}$$

where O_i and the E_i are the observed and forecasted value of the crop yield respectively and n is the number of years for which forecasting has been done.

Results for Sultanpur district

On the basis of weekly data on weather variables such as Minimum Temperature, Maximum Temperature, Relative Humidity 07 hrs, Relative Humidity 14 hrs, Wind-Velocity, and Sunshine hours using principal component, statistical models for pre harvest forecast of the Rapeseed & Mustard yield in Sultanpur district of Eastern Uttar Pradesh have been developed. Three models have been created using the best methods.In Sultanpur district, rapeseed and mustard sowing typically begins in the first week of October. The first week of October marks the pre-sowing of the 40^{th} SMW of crop, hence weekly data on meteorological variables have been taken into account since then. Pre-harvest forecasting of the rapeseed and mustard yield has been suggested to be done at the stage of milking/dough, or roughly one and a half months before the harvest. As a result, the pre-harvest forecast week has been designated as the eighth SMW of the following year. In order to create the statistical models, data on the meteorological variables from the 40^{th} SMW of the previous year to the 8^{th} SMW of the following year were used for all 21 weeks.

Comparison of the model

The forecast yields for the years 2012–2013, 2013–2014, and 2014–2015 have been calculated based on these two forecast models, and the results are shown in Table 1. For the best three models, the values of R^2 adj, percent deviation of forecast from actual yield, RMSE, and%SE (CV) have also been computed and are also shown in Table 2. The results of Table 1 have been represented graphically in Fig. 1 to highlight the models' propensity towards forecasting.

Table.1 Comparison between actual and forecasted yield of different years of Sultanpur District

Model	Year	Actual yield	Predicted yield	Percent deviation)	PSE	R^2	Adj R^2	RMSE
I	2012-13	7.32	9.32	27.36	4.19	84.7	82.2	1.47
	2013-14	4.99	6.37	27.68	9.67			
	2014-15	4.45	5.24	17.75	15.88			
II	2012-13	7.32	10.08	37.82	3.82	96.5	94.3	2.10
	2013-14	4.99	6.71	34.49	5.38			
	2014-15	4.45	6.07	36.54	8.73			
III	2012-13	7.32	9.73	32.96	5.63	88.2	84.5	1.89
	2013-14	4.99	6.73	35.06	8.13			
	2014-15	4.45	5.83	31.13	13.01			

Figure :-1 Graphical Representation of data.

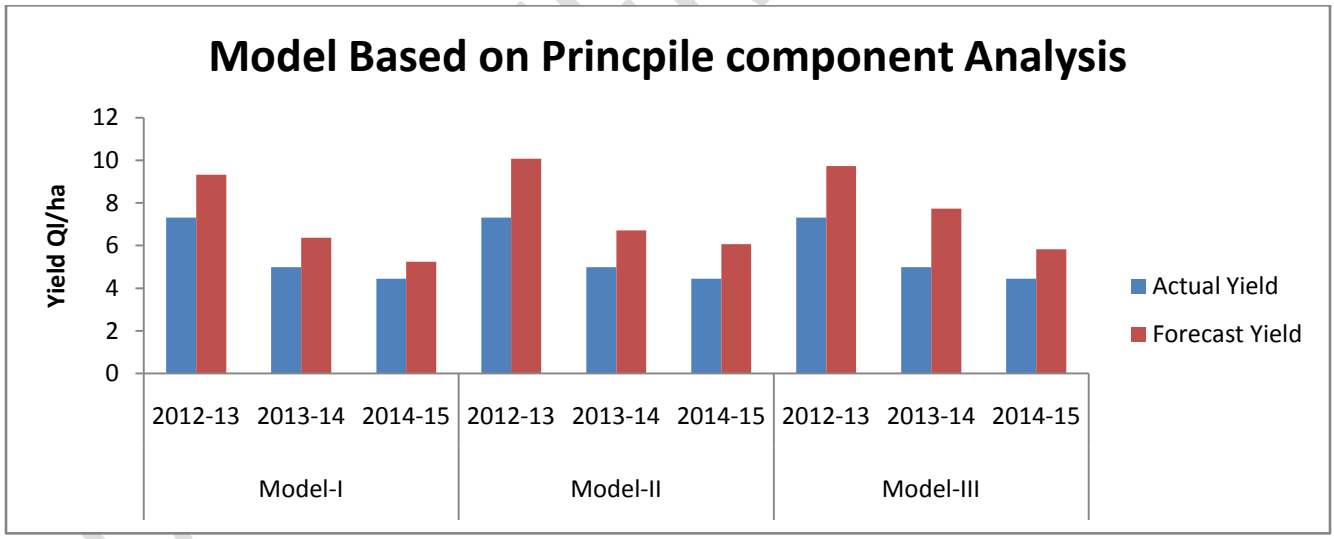

Table.2:-Best model from the application of Principle Component analysis of weekly weather

Model- I	**Y** = 5.878– 1.787 PC_1 + 0.378PC_2+ 0.143 T	**84.7**	**82.2**

Model-II	**Y** = 6.124-1.408 PC_1 - 0.866 PC_2- 0.063 PC_3+ 0.088 PC_4+0.741 PC_5- 0.350 PC_6 +0.075PC_7 + 0.130**T**	**96.5**	**94.3**
Model-III	Y = 5.896 -1.736 PC_1 -0.408 PC_2 - 0.177 PC_3+ 0.021 PC_4 + 0.146 T	**88.2**	**84.5**

Based on the overall results of Table 2, it can be concluded that Model-II, followed by Model-I and Model III, is the most suitable model to forecast rapeseed and mustard yield in Sultanpur district of Eastern Uttar Pradesh. It is clear from the results of Table 1 that coefficient of determination (R^2) has been found to be 84.7%, 96.5%, and 88.2% for the Model, respectively. Hence, the Model-I can be used to accurately predict the yield of rapeseed and mustard around 1.5 months before harvest.

Summary and conclusion of the study are as follows:

Both models use the stepwise regression method and principal component analysis for the construction of their forecasting models. Table 2 lists the top model discovered through the use of principal component analysis on weekly weather data. For the Sultanpur district's forecast of rapeseed and mustard yield, model II is more trustworthy.

References:-

Agrawal,Ranjana, Jain, R.C. and Jha, M.P. (1986). Models for studying rice crop-weather relationship. *Mausam*, 37(1), 67-70.

Fisher, R. A. (1924). The influences of rainfall on the yield of wheat at Rothamsted. Philosophical Transaction of Royal Society of London, Series B, Vol. 213,pp. 89-142.

Hendricks, W.A. and Scholl, G.C. (1943). Technique in measuring joint relationship: The joint effects of temperature and precipitation on crop yield. N. Carolina Agric. Exp. Stat.

Jain, R.C., Agrawal, Ranjana and Jha, M.P. (1980). Effect of climatic variables on rice yield and its forecast. *Mausam*, 31(4), 591-96.

Jain, R.C., Sridharan, H. and Agrawal, Ranjana (1984). Principal component technique for forecasting of sorghum yield. *Indian Journal of Agril. Sci.*Vol.LI, 1: 61-72.

Jain, R.C., Jha, M.P. and Agrawal, Ranjana, (1985). Use of growth indices in yield forecast. Biometrical Journal, Vol.27(4), pp. 435-439.

Pandey, K.K., Rai, V.N., Sisodia, B.V.S., Bharti, A.K., Gairola, K.C. (2013). Pre - Harvest Forecast Models Based On Weather Variable And Weather Indices For Eastern U.P.*Adv. Biores*., Vol. 4 (2): 118- 122.

Sisodia, B. V. S., Yadav, R. R., Kumar, S. and Sharma, M. K. (2014). Forecasting of Pre-harvest crop yield using discriminant function analysis of meteorological parameter. *Journal ofAgrometeorology*. Vol.16(1), pp.121- 125.

Yadav, R.R., Sisodia, B.V.S., Kumar, S. (2014). Application of Principal Component Analysis in Developing Statistical Models to Forecast Crop Yield Using Weather Variables. *Mausam*, Accepted

AzfarMohd., Sisodia, B. V. S, RaiV. N. ,DeviMonoka, (2015).Pre-harvest forecast models for rapeseed & mustard yield using principal component analysis of weather variables *Journal ofAgrometeorology* Vol. 66 No. 4.