J. Cotton Res. Dev. **29** (2) 339-343 (July, 2015)

Principal component technique for pre harvest estimation of cotton yield based on plant biometrical characters

U. VERMA*, D.R. ANEJA AND B.K. HOODA
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar-125004
**E-mail : vermas21@hotmail.com*

ABSTRACT : An attempt has been made to estimate the yields of cotton hybrids using the principal components of the plant biometrical characters spread over five six successive stages within the growth period of cotton crop. The results indicate the possibility of yield prediction of cotton hybrids RCH 134BG I, RCH 134BG II and Bioseeds 6488BG II, one month ahead of the harvest time. The estimated yields of these hybrids during *kharif*, 2011-2012 were 30.96 q/ha, 31.51q/ha and 31.53q/ha against the observed yields 30.24 q/ha, 29.42q/ha and 32.72q/ha, respectively.

Key words : Eigen value, eigen vector, estimated yield, multicollinearity, principal component score

Reliable, accurate and timely information on types of crop grown and their acreages, crop yield and crop growth conditions are vital components for planning efficient management of natural resources. This involves formulating and implementing appropriate prices of agricultural commodities and import/export of these commodities from time to time. Various organizations in India and abroad are engaged in developing methodology for pre harvest forecasting of crop yield using various approaches. With the advent of remote sensing technology during 1970s, its great potential in the field of agriculture have opened new vistas of improving the agricultural statistics system all over the world.

In Haryana State, Hisar and Sirsa are the two major cotton producing districts, accounting for 80 per cent of the acreage and 86 per cent of the cotton production in the state. Cotton is the dominant crop grown in these districts during *kharif* season and occupies almost 40 per cent of the geographical area. Cotton is mostly grown under irrigation due to the prevailing arid conditions. Being a long duration crop, it is generally sown before the onset of the monsoon (May-June) and harvested during the early part of winter (November-December).

Prominent among the methods of forecasting are the statistical models that utilize data on crop biometrical characters, weather parameters and remotely sensed crop reflectance observations etc., utilized either separately or in an integrated approach. In Haryana state, the first attempt to estimate cotton acreage and condition assessment was made during *kharif*, 1990-1991 season in Hisar and Sirsa districts using IRS-1A LISS-1 digital data and the stratified random sampling approach (Sharma *et al.*, 1992). Subsequently, the efforts were made to improve the accuracy and timeliness of this process by modifying the stratification and sampling procedure by Yadav *et al.*, (1994). Larson *et al.*, (2002) have studied cotton defoliation and harvest timing effects on yields and quality. Rai *et al.*, (2003) have conducted a study on pre-harvest cotton yield based on plant biometrical characters. Viator *et al.*, (2005) have worked to observe the effect of climatic factors on cotton boll formulation. A study on the relationship between leaf area index and IRS LISS-III spectral vegetation indices of cotton in Hisar district was conducted by Kalubarme *et al.*, (2006).

MATERIALS AND METHODS

The primary data on plant biometrical characters (2011-2012) collected from Cotton

Research Area, Department of Genetics and Plant Breeding, CCS HAU, Hisar have been used to develop suitable models for predicting the yield of cotton hybrids RCH 134BG I, RCH 134BG II and Bioseeds 6488BG II. The Hisar district, a part of the Indo Gangetic alluvial plain is situated between $28^0 53'45''$ to $29^0 49'15''$N latitudes and $75^0 13'15''$ to $76^0 18'15''$E longitudes. It occupies an area of 3788 sq km and experiences a sub tropical climate. The climate is influenced by westerly winds in summer months raising temperature as high as 48^0C, whereas in winter north-westerly cold winds provide low temperature touching even 0^0C. The average rainfall in the district is 334.4 mm. About 85 per cent of annual rainfall is received during the short south western monsoon period.

Six fields; two each of hybrids RCH 134 BG I, RCH 134BG II and Bioseeds 6488BG II were selected and a number of plants ranging from 13 to 20 plants from each hybrid at both the fields were selected at random for recording observations on biometrical characters. These plants were tagged and the recordings were made at regular interval of a fortnight. The row to row spacing was 67.6cm and plant to plant spacing was 60cm. The data were recorded from the last week of August to first week of November. Following data on biometrical characters of each selected plant for all the three hybrids were recorded during the crop growth period:

RCH 134BG I and II (Last week of August to 1st week of November)

Stage I	
(Last week of August)	X_1 Height (cm)
	X_2 Girth (cm)
	X_3 Unopened bolls
Stage II	
(2nd week of September)	X_4 Height (cm)
	X_5 Girth (cm)
	X_6 Unopened bolls
Stage III	
(Last week of September)	X_7 Height
	X_8 Girth
	X_9 Unopened bolls
	X_{10} Opened bolls
	X_{11} Total bolls
Stage IV	
(1st week of October)	X_{12} Unopened bolls
	X_{13} Opened bolls
	X_{14} Total bolls
	X_{15} Yield of 1st pick (g)
Stage V	
(3rd week of October)	X_{16} Unopened bolls
	X_{17} Opened bolls
	X_{18} Total bolls
	X_{19} Yield of 2nd pick (g)
	X_{20} Yield of (1st +2nd) picks (g)
Stage VI	
(1st week of November)	X_{21} Unopened bolls
	X_{22} Opened bolls
	X_{23} Total bolls
	X_{24} Yield of 3rd pick (g)
	Y Total yield (g)

Bioseeds 6488BG II (Last week of August to 1st week of November)

Stage I	
(Last week of August)	X_1 Height (cm)
	X_2 Girth (cm)
	X_3 Unopened bolls
	X_4 Opened bolls
Stage II	
(2nd week of September)	X_5 Height (cm)
	X_6 Girth (cm)
	X_7 Unopened bolls
	X_8 Opened bolls
	X_9 Total bolls
Stage III	
(Last week of September)	X_{10} Height
	X_{11} Girth
	X_{12} Unopened bolls
	X_{13} Opened bolls
	X_{14} Total bolls
Stage IV	
(1st week of October)	X_{15} Unopened bolls
	X_{16} Opened bolls
	X_{17} Total bolls
	X_{18} Yield of 1st pick (g)

Stage V
(3rd week of October)
X_{19} Unopened bolls
X_{20} Opened bolls
X_{21} Total bolls
X_{22} Yield of 2nd pick(g)
X_{23} Yield of (1st +2nd) picks (g)

Stage VI
(1st week of November)
X_{24} Unopened bolls
X_{25} Opened bolls
X_{26} Total bolls
X_{27} Yield of 3rd pick (g)
Y Total yield (g)

RESULTS AND DISCUSSION

The use and interpretation of a multiple regression model often depends explicitly or implicitly on the assumption that the explanatory varia-bles are not strongly interrelated. In most regression applica-tions, the explanatory variables are not orthogonal. Usually the lack of orthogonality is not serious enough to affect the analy-sis. However in some situations, the explanatory variables are so strongly interrelated that the regression results are ambiguous. Typically, it is impossible to estimate the unique effects of individual variables in the regression equation. The estimated values of the coefficients are very sensitive to slight changes in the data and to the addition or deletion of variables in the equation. The regression coefficients have large sampling errors, which affect both inference, and forecasting that is based on the regression model. The condition of severe non orthogonality is also referred to as the problem of multicollinearity. To overcome the problem of multicollinearity observed among plant biometrical characters (Verma *et al*., 2013), the crop yield models are developed within the framework of principal component analysis (PCA).

Principal component method was used for the extraction of factors which consists of finding the eigen values and eigen vectors. Principal components P_i (i=1,2,....) were obtained as P = kX , where P and X are the column vectors of transformed and the original variables, respectively and k is the matrix with rows as the characteristic vectors of the correlation matrix R. The variance of P_i is the i^{th} characteristic root $ë_i$ of the correlation matrix R; $ë_s$ were obtained by solving the equation $|R- ëI| =0$. For each ë, the corresponding characteristic vector k was obtained by solving $|R-ëI|$ k=0

Under this study, first 4 (for RCH 134BG 1 and II) and 5 (for Bioseeds 6488BG II) eigen values (Table1) of correlation matrix of explanatory variables (plant biometrical variables used for PC analysis *i.e.* X_1 to X_{15} for RCH 134BG 1, II and X_1 to X_{18} for Bioseeds 6488BG II) suggested four/five factor solution. It is clear that the remaining components accounted for a smaller amount of total variation (eigen values beyond 9^{th} PC are not shown being very small in magnitude). Hence, those components were not considered to be of much practical significance. For the hybrids RCH 134BG I and II, first four PCs (out of 15 PCs obtained on the basis of X_1 to X_{15} plant biometrical variables) were retained explaining 93 per cent of variation in the data set. In case of hybrid Bioseeds 6488BG II, first five PCs (out of 18 PCs obtained on the basis of X_1 to X_{18} plant biometrical variables) explained 92 per cent of variation in the data set. Eigen vectors being the weights

Table 1. Eigen values and variance (%) explained by different principal components

Compo-nents	Eigen value (% variance explained)		
	RCH 134 BG I	RCH 134 BG II	Bioseeds 6488 BG II
1	7.92 (52.77)	9.37 (62.44)	5.67 (35.46)
2	3.53 (23.51)	2.36 (15.75)	3.87 (24.18)
3	1.36 (9.04)	1.47 (9.83)	2.68 (16.72)
4	1.08 (7.17)	0.69 (4.58)	1.34 (8.37)
5	0.42 (2.85)	0.47 (3.11)	1.08 (6.76)
6	0.33 (2.21)	0.33 (2.19)	0.54 (3.43)
7	0.12 (0.77)	0.12 (0.80)	0.21 (1.28)
8	0.08 (0.58)	0.08 (0.59)	0.18 (1.12)
9	0.06 (0.44)	0.04 (0.29)	0.15 (0.91)

Table 2. Selected cotton yield models based on PC scores of plant biometrical characters

RCH134BG I Model variable Model 1	Coefficients	RCH134BG II Model variable Model 2	Coefficients Model 3	6488 BG II Model variable	Coefficients
Constant	**$c_1$135.79**	**Constant**	**c1 158.06**	**Constant**	**$c_1$152.22**
PC_1	**$b_1$65.50**	**PC_1**	**b_1 89.64**	**PC_1**	**b_1 38.04**
PC_3	**$b_2$16.22**	**PC_2**	**b_2 41.07**	**PC_2**	**$b_2$39.59**
PC_4	**$b_3$19.31**	**PC_3**	**$b_3$23.65**	**PC_3**	**$b_3$11.01**
				PC4	b_4-15.67
R^2= 0.923 adj.R^2 =0.916	**SE =21.26**	**R^2= 0.908 adj.R^2 =0.898**	**SE= 33.91**	**R^2= 0.856 adj.R^2 =0.837**	**SE= 25.35**

RCH 134BG 1

$$Yield_{est} \text{ (model 1)} = \{c_1 + (b_1 \times PC_1) + (b_2 \times PC_3) + (b_3 \times PC_4)\}$$

RCH 134BG II

$$Yield_{est} \text{ (model 2)} = \{c_1 + (b_1 \times PC_1) + (b_2 \times PC_2) + (b_3 \times PC_3)\}$$

Bioseeds 6488BG II

$$Yield_{est} \text{ (model 3)} = \{c_1 + (b_1 \times PC_1) + (b_2 \times PC_2) + (b_3 \times PC_3) + (b_4 \times PC_4)\}$$

where $Yield_{est}$ - Model predicted yield

PC_i - i^{th} principal component score (i = 1,2,3,4,5)

SE - Standard error of yield estimate

R^2 - Coefficient of determination

were used to compute PC scores. Plant biometrical data starting from last week of August to 1st week of October *i.e.* one month before harvest were utilized for the model building. So for quantitative forecasting, regression models *via* step-wise regression (Draper and Smith, 2003) were fitted, considering PC scores as regressors and total yield (Y) as dependent variable (Table 2). Further, the developed models were used to obtain the yield estimates of the cotton hybrids under consideration.

CONCLUSION

The yield of cotton hybrids RCH 134BG I, RCH 134BG II and Bioseeds 6488BG II can be predicted in the first week of October using the principal components of plant biometrical characters. The estimated yields of these hybrids during *kharif*, 2011-2012 were 30.96 q/ha, 31.51q/ha and 31.53q/ha against the observed yields 30.24q/ha, 29.42q/ha and 32.72q/ha, respectively.

ACKNOWLEDGEMENT

The primary data collected on plant biometrical characters from Cotton Research Area, Department of Genetics and Plant Breeding, CCS HAU, Hisar are gratefully acknowledged.

REFERENCES

Draper, N.R. and Smith, H. 2003. *Applied Regression Analysis*. 3rd edition, John Wiley and Sons. New York.

Kalubarme, M.H., Hooda, R.S., Yadav, M. and Saroha, G.P. 2006. Relationship between leaf area index and IRS LISS-III spectral vegetation indices of cotton. *Scientific Note,* EOAM/ SAC/CAPE-II/SN/ 98.

Larson, J.A., Gwathmey, C.O. and Hayes, R.M. 2002. Cotton defoliation and harvest timing effects on yields, quality and net revenues. *J. Cotton Sci.* **6** : 13-27.

Rai, L., Aneja, D.R., Saxena, K.K. and Grover, D.K. 2003. Pre harvest cotton yield based on plant biometrical characters. Proceedings of workshop on *"Remote Sensing and GIS for Rural Development with special reference to Haryana"* held at HARSAC, CCS HAU, Hisar Sep., 29-30: pp. 203-08.

Sharma, S.A., Ajai, Hooda, R.S., Mothikumar, K.E., Yadav, M. and Manchanda, M.L.1992. Cotton acreage and condition assessment for Hisar and Sirsa districts of Haryana (1990-91). *Scientific Note*, RSAM/SAC/CACA/SN.

Verma, U., Aneja, D.R. and Rai, L. 2013. Forecasting the yield of *Bt* cotton using biometrical characters in Hisar district of Haryana (India). *Envir. Eco.* **31** : 527-31.

Viator, R.P., Nuti, R.C., Keith, L., Edmisten and Wells, R. 2005. Predicting cotton boll maturation period using degree days and other climatic factors. *Agron. J.* **97** : 494-99.

Yadav, M., Hooda, R.S., Mothikumar, K.E., Ruhal, D.S., Khera, A.P., Singh,C.P., Hooda, I.S., Verma, U., Dutta, S. and Kalubarme, M.H. 1994. Cotton acreage in Hisar and Sirsa districts of Haryana using remote sensing techniques. *Tech. Report,* HARSAC/TR/ 03.

Recieved for publication : August 13, 2013
Accepted for publication : April 6, 2014