Environment and Ecology 39 (3) : 520—524, July—September 2021
ISSN 0970-0420

Development of Forecast Model using Principal Component Regression Approach

Sunil Kumar, Manoj Kumar, Ashok Dhillon, Hemant Poonia

Received 6 April 2021, Accepted 18 May 2021, Published on 7 July 2021

ABSTRACT

In this paper an attempt has been made to compare the Principal Component Regression and Multiple Regression analysis approach by developing the pre-harvest forecast model for chickpea seed yield and its attributing traits based on biometrical character. The data of one year has been used in this study. The data have been divided into two component testing and training data set. The training data set consists of 90% of the observation and remaining 10% data have been used for validation purpose. The coefficient of determination (R_2) for Principal Component Regression and Stepwise Multiple Regression model (SMR) model were 92% and 34% respectively with mean square error 11.60 and 29.60. Therefore, model developed on principal component regression performed better than the stepwise multiple regression approach for the forecasting purpose.

Keywords Biometrical Character, Experimental data, Multiple regression analysis, Principal component regression, R-software.

Sunil Kumar
Assistant Scientist, Pulses Section, Department of Plant Breeding, CCS HAU, Hisar 125004, Haryana, India
Manoj Kumar*
Assistant Professor, Department of Mathematics & Statistics, CCS HAU, Hisar 125004, Haryana, India
Ashok Dhillon
DES, KrishiVigyan Kendra, Mahendragargh, CCSHAU, Hisar 125004, Haryana, India
Hemant Poonia
Assistant Professor, Department of Mathematics & Statistics, CCS HAU, Hisar 125004, Haryana, India
Email: m25424553@gmail.com
*Corresponding author: Dr Manoj Kumar, Assistant Professor (Statistics), Department of Mathematics & Statistics, CCS HAU, Hisar125004, Haryana, India

INTRODUCTION

Agriculture is backbone of Indian economy, contributing about 40 % towards the Gross National Product (GDP) and provides livelihood to about 70 % of the population. Chickpea (*Cicer arietinum* L.) is the only cultivated species within the genus *Cicer*. The crop is self-pollinated diploid (2n= 2x= 16) seeds contain 20-30 % protein, approximately 40 % carbohydrates and only 3-6 % oil (Gill *et al.* 1996) and moreover, they are a good source of calcium, magnesium, potassium, phosphorus, iron, zinc and manganese (Ibrikci *et al*. 2003). Commercially the species is grouped into two distinct types of chickpea *Desi* and *Kabuli* types: *Desi* (also known as microsperma) whereas *Kabuli* (also known as macrosperma). The Desi types are found in central Asia and in the Indian subcontinent while the *Kabuli* types are mostly found in the Mediterranean region. *Kabuli* types are usually taller and have large beige or cream seed color and "ram's head" seeds with white flowers. *Desi* types are generally shorter, possessing small leaflets, pods, seeds and

predominantly pink-colored flowers (Moreno and Cubero 1978). India is largest producer and consumer of chickpea in the world. In India, the area under chickpea was 9.67 million hectares with a production of 10.09 million tones and productivity of 1043 kg/ha during *rabi* 2018-19. In Haryana, area under chickpea was 32 thous and hectares with a total production of 36 thousand tones and productivity of 1125 kg/ha during 2018-19 (Anonymous 2019). Madhya Pradesh, Uttar Pradesh, Maharashtra, Rajasthan, Gujarat, Andhra Pradesh, Karnataka and Bihar are the major chickpea growing states in the country. As we know that reliable forecast of crop yield is required by the Government for making policy decision with regard to procurement, distribution, import-export and buffer-stocking. All the agro-based industries, traders and agriculturists need the forecast model for proper planning of their operations. A number of research papers have been published by the various authors viz., Annu *et al.* (2015, 2017), Azfar *et al*. (2015), Basso *et al.* (2012), Esfandiary *et al.* (2009), Gill *et al.* (1996), Lobell *et al.* (2010), Manoj *et al*. (2019),Verma *et al.* (2012),Yadav *et al.* (2014), to develop statistical models for forecasting crop yield based on biometrical characters using experimental and survey data in different region of the country. Keeping in view, importance of yield for policy decision makers and other government agencies, the pre-harvest forecast model for chickpea seed yield based on biometrical character haven been developed by applying the techniques of principal component regression and multiple regression approach. For this purpose, the experimental data have been taken from Pulses Section, Genetic and Plant Breeding, CCS Haryana Agricultural University, Hisar.

MATERIALS AND METHODS

The experiment was conducted at pulses section CCS HAU, Hisar. The average temperature of the areas is 21.65°C. The major crops grown widely are wheat, chickpea, barley, oat, Indian mustard, under rain fed and irrigation but cereal mono cropping is the predominant grown in the study area. The soil type of the area is light loamy soil with a pH ranging from 6.5 to 8.0. The field was loose tilt and well drained. A smooth seedbed was prepared to avoid packing of the cloddy surface due to winter rains and to facilitate soil aeration and for easy seedling emergence (Mehta *et al.* 2000).The experiment was laid out in augmented design without replications. Each genotype had 2 rows in a plot of 4.0 m row length with a row to row and plant to plant spacing of 30 cm x 10 cm respectively.

The secondary data have been taken from the technical report of the genetic and plan breeding section for the year 2018-19(Fig. 1). The following character was taken namely Plant Height (PH), No of branches (NB), no of pods/plant (NP), Seed Index (SI), Days to 1st flower (DF), Days to maturity (DM), Biological Yield (BY) (kg/ha), Seed yield (SY) (kg/ha) and Total Biomass Yield (TBY). Principal Component, Principal Regression and Multiple Regression Analysis was performed to forecast the chickpea seed yield based on the biometrical character using the following R-code.

```
data1<-read.csv (file.choose(),header=TRUE)
attach(data1)
numeric=c('X1','X2','X3','X4','X5','X6','X7')
target=c('Y')
set.seed(42)
train = sample(nrow(data1), 0.9*nrow(data1))
test = setdiff(seq_len(nrow(data1)),train)
model<-step(lm(Y~.,data= data1[train,c(target,numeric)]))
summary(model)
Estimate=predict(model,type='response',newdata=-
data1[test,c(target,numeric)])
Observed=subset(data1[test,c(numeric,target)],select=target)
Format(cor(Estimate,Observed$target)^2,digits=4)
Require(FactoMineR)
Data_for_PCA&amp;amp;amp;amp;lt;-data1[,numeric]
pca1 = PCA(data1)
PCA_data=as.data.frame(cbind(data1[train,target],pca1$ind$coord[train,]))
Step_PCA_Reg =step(lm(V1~.,data = PCA_data))
summary(Step_PCA_Reg)
PCA_Estimate=predict(Step_PCA_Reg,type='response',newdata=cbind(data1[test,c(target,numeric)],pca1$ind$coord[test,]))
summary(PCA_Estimate)
accuracy(PCA_Estimate)
format(cor(PCA_Estimate, Observed$House_Price)^2, digits=4)
```

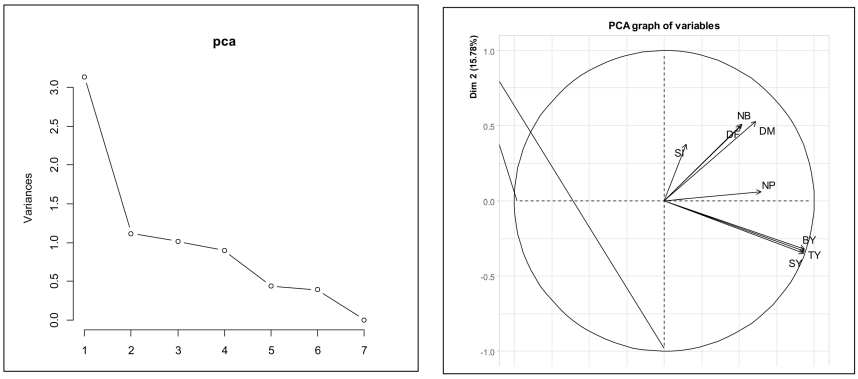

Fig. 1. PCA graph variables.

Data_for_PCA = data1[numeric]
pca=PCA(Data_for_PCA)
pca$eig

*Correlation_Matrix=as.data.frame(round(cor(Data_for_PCA,pcaindcoord)^2*100,0))*
Correlation_Matrix[with(Correlation_Matrix, order(-Correlation_Matrix[,1])),]
Only for principal component and scree plot
standardisedconcentrations<- as.data.frame(scale(ta1[2:8]))
pca<- prcomp(standardisedconcentrations)
pca$sdev
sum((pca$sdev)^2)
screeplot(pca, type="lines")

Evaluation of the model

The adequacy of the model was find out using the relative mean square error and percentage forecast error.

$$\text{Relative mean square error (RMSE)} = \sqrt{\frac{(\sum_{i=1}^{N} (\text{Actual value-Forecast value})^2}{N}}$$

RESULTS AND DISCUSSION

The descriptive statistics of the data are presented in the Table 1. Principal component regression and stepwise multiple regression have been used to forecast the plant yield based on biometrical character. Table 2 show that first four principal components explain

Table 1. Descriptive statistics for various parameter in chickpea. PH= Plant Height, NB = No of branches, NP= no of pods/plant, SI= Seed Index, DF= Days to 1st flower, DM= Days to maturity, BY= Biological Yield (kg/ha), SY= Seed yield (kg/ha), TBY = Total Biomass Yield.

	PH	NB	NP	SI	DF	DM	BY	SY	TBY
Mean	73.31	4.22	77.35	17.78	96.71	148.90	1197.86	930.02	2127.88
Standard Error	1.91	0.15	6.41	0.71	1.29	1.45	68.72	54.84	123.43
Range	66.00	5.00	235.00	19.10	40.00	42.00	1910.48	1480.30	3390.78
Minimum	42.00	2.00	15.00	9.80	66.00	126.00	190.35	120.62	310.98
Maximum	108.00	7.00	250.00	28.90	106.00	168.00	2100.83	1600.92	3701.76

Table2. Principal component analysis for the seed yield and its contributing traits in chickpea during 2018-2019.

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Eigen Value	3.95	1.26	1.01	0.927	0.443	0.397	0.003
Proportion of Variance	49.3	15.7	12.6	11.5	5.54	4.97	0.004
Cumulative Proportion	49.38	65.16	77.84	89.43	94.98	99.95	100.00

Table 3. Comparison of step-wise multiple regression (SMR) and Principal component regression (PCR) during 2018-2019 in chickpea. ***Regression coefficients are significant at 0.001 level,**Regression coefficients are significant at 0.01 level, *Regression coefficients are significant at 0.05 level.

	Step-wise multiple regression (SMR)		
	Coefficients (SE)	t-value	p-value
Intercept	41.85*** (7.250)	5.77	0.0243*
SI	0.9116* (0.348)	2.617	
BY	0.012 (0.003)	3.499**	
Multiple R-squared = 0.3485		Adjusted R-squared: 0.3167	
MSE of the model =			
	Principal component regression (PCR)		
	Coefficients (SE)	t-value	p-value
Intercept	72.89 (0.599)	121.63***	
PC_1	3.82 (0.296)	12.87**	
PC_2	5.718 (0.503)	11.359**	
PC_3	-4.33 (0.598)	-7.254**	
PC_4	-1.379 (0.584)	-2.360**	
PC_5	-6.42 (0.827)	-8.366***	
Multiple R-squared: 0.92	Adjusted R-squared: 0.91		
MSE of the model =			

about 89.94% of the total variability and these first four components have been used as repressor for forecasting the plant yield and it is also clear from the "scree" plot of the principal component that there is change in slope in the scree plot occurs at component 4. The data have been divided into two component testing and training data set. The training data set consists of 90% of the observation and reaming 10% data have been used for validation purpose.

The fitted models along with parameter estimation and R^2 valueare shown in Table 3. It is clear from the table that the value of coefficient of determination (R^2) for the principal component regression (PCR) showed 92% as compare to stepwise multiple regression 34%. It can also be observed that first principal component (PC1) and third component showed highly positive significant effect on the yield. However, the first and fourth have no significant effects on the seed yield in chickpea. The forecast yields for both the models are shown in the Table 4 it is observed that root mean square error is less for principal component egression model as compared to the multiple regression model. From the analysis it can be concluded that the application of principal component regression techniques has provided a suitable forecast model harvest forecast of seed yield in chickpea if the proper measurements on biometrical characters under consideration are available. using biometrical character. Therefore, the proposed model can be used to obtain reliable pre harvest forecast of seed yield in chickpea if the proper measurements on biometrical characters under consideration are available.

Table 4. Actual and forecast of seed yield of chickpea based on the above fitted model.

Actual yield(q/ha)	Forecast yield(q/ha)		RMSE	
	SMR	PCR	SMR	PCR
70	70.42	66.44	29.60	11.60
88	77.00	80.01		
52	68.15	53.02		
83	60.89	80.77		
75	77.32	67.78		

REFERENCES

Annu, Sisodia BVS, Kumar Sunil (2015) Pre-harvest forecast models for wheat yield based on biometrical characters. *Econ Affairs* 60(1):89—93.DOI: 10.5958/0976-4666.2015.00012.1

Annu, Sisodia BVS, Rai VN (2017) An Application of principal component analysis for pre-harvest forecast model for wheat crop based on biometrical character. *Int Res J Agric Econ Stat* 8(1): 83—87. DOI: 10.15740/HAS/IRES/8.1/83-87

Anonymous (2019) Project Coordinator's Report, AICRP on chickpea. ICAR- IIPR, Kanpur, pp19—21.

Azfar Mohd, Sisodia BVS, Rai VN, Monika Devi (2015) Pre-harvest of rapeseed and mustard yield based on weather variables-An application of principal component analysis of weather variables. Mausam 66 (4) : 761—766.

Basso B, Fiorentino C, Cammarano D, Cafiero G, Dardanelli J (2012) Analysis of rainfall distribution on spatial and temporal patterns of wheat yield in Mediterranean environment. *Europ J Agron* 41:52—65. DOI:1016/.a.2012.03.007

Esfandiary F, Aghaie G, Mehr AD (2009) Wheat yield prediction through agro meteorological indices for Ardebil district. World Acad Sci: *Engg Technol* 49 : 32—35. Doi=10.1.1.912.2241

Gill J, Nadal S, Luna D, Moreno MT, Haro A de (1996) Variability of some physio-chemical characters in *Desi* and *Kabuli* Chickpea types. *J Sci Food Agric* 71: 179—184. https://doi.org/10.1002/(SICI)1097-0010(199606)71:2<179::AID-JSFA566>3.0.CO;2-7

Ibrikci H, Knewtson S, Grusak MA (2003) Chickpea leaves as a vegetable green for humans: Evaluation of minerals composition. *J Sci Food Agric* 83: 945—950. https://doi.org/10.1002/jsfa.1427

Kumar Manoj, Battan KR, Sheoran OP (2019) Pre-harvest forecast model for rice yield using principal component regression based on biometrical character with R-software. *Int J Agricult Stat Sci* 15 (1) : 323—326.

Lobell DB, Burke M (2010) On the use of statistical models to predict crop yield responses to climate change. *Agric For Meteorol* 150:1443—1452. https://doi.org/10.1016/j.agrformet.2010.07.008

Mehta SC, Agarwal R, Singh VPN (2000) Strategies for com posite forecast. *J Ind Soc Agric Statis*t 53(3) : 262—272. http://isas.org.in/jsp/volume/vol53/issue3/S.C.Mehta.pdf

Verma U, Dabas DS, Hooda RS, Kalubarme MH, Yadav M, Sharma MP (2011) Remote sensing based wheat acerage and spectraltrend-agrometeorological yield forecasting: Factor analysis approach. Soc Stat, Computers Ap p l9 (1&2):1—13. https://www.ssca.org.in/media/1Urmil_Verma.pdf