Advances in Research

17(4): 1-8, 2018; Article no.AIR.42670
ISSN: 2348-0394, NLM ID: 101666096

Principal Component Technique for Pre-harvest Crop Yield Estimation Based on Weather Input

Megha Goyal[1*], Salinder[1], Suman[1] and Urmil Verma[1]

[1]Department of Mathematics, Statistics and Physics, CCS Haryana Agricultural University, Hisar-125004, India.

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AIR/2018/42670
Editor(s):
(1) Dr. Magdalena Valsikova, Professor, Horticulture and Landscape Engineering, Slovak University of Agriculture, Nitra, Slovakia.
(2) Dr. Slawomir Borek, Department of Plant Physiology, Adam Mickiewicz University, ul. Umultowska, Poland.
(3) Dr. Francisco Marquez-Linares, Professor of Chemistry, Nanomaterials Research Group, School of Science and Technology, University of Turabo, USA.
Reviewers:
(1) Abhijit M. Zende, Dr. Daulatrao Aher college of Engineering, India.
(2) Ritambhara Singh, International Agribusiness Management Institute, Anand Agricultural University, India.
(3) Yahaya, Tayo Iyanda, Federal University of Technology, Nigeria.
(4) Timothy Denen Akpenpuun, University of Ilorin, Nigeria.
Complete Peer review History: http://www.sciencedomain.org/review-history/27406

Short Research Article

Received 11 June 2018
Accepted 17 November 2018
Published 26 November 2018

ABSTRACT

Forecasting of crop production is one of the most important applications of statistics in agriculture. Such predictions before harvest are needed by the national and state governments for various policy decisions relating to storage, distribution, pricing, marketing, import- export, etc. Therefore, a methodology for the estimation of wheat yield, ahead of harvest time, is developed specifically for wheat growing districts in Haryana (India). The Haryana state, having a total geographical area of 44212 sq. km, was divided into four zones for pre-harvest crop yield forecasts. An attempt has been made in this paper to estimate the yield of the wheat crop using principal components of the weather parameters spread over the crop growth period. Principal component analysis has been used for the purpose of developing zonal yield forecast models because of multicollinearity present among weather variable. The results indicate the possibility of district-level wheat yield prediction, 4-5 weeks ahead of the harvest time, in Haryana. Zonal weather models had the desired predictive accuracy and provided considerable improvement in the district-level wheat yield estimates. The estimated yield(s) from the selected models indicated good agreement with State Department of

**Corresponding author: E-mail: meggoel@yahoo.com;*

Agriculture (DOA) wheat yields by showing 2-10 percent average absolute deviations in most of the districts except for the Rohtak district observing 12.81 percent average absolute deviation from the real-time data.

Keywords: Linear time trend; Eigen value; eigen vector; weather variables; multicollinearity; principal component score.

1. INTRODUCTION

Crop yield models are abstract presentation of the interaction of the crop with its environment and can range from simple correlation of yield with a finite number of variables to the complex statistical models with predictive end. Forecasts can be formed in many different ways. The method chosen depends on the purpose and importance of forecasts as well as the cost of alternative forecasting methods. 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.

Crop yield is affected by technological change and weather variability. It can be assumed that the technological factors will increase crop yield smoothly through time and therefore, a year or some other parameter of time can be used to study the overall effect of technology on yield. Weather variables affect the crop differently during different stages of development. This increases the number of variables in the model and thus, a technique based on a relatively smaller number of manageable parameters and at the same time, taking care of entire weather distribution may solve the problem. Principal Component Analysis (PCA) was carried out for pre-harvest wheat yield estimation on agro-climatic zone basis in Haryana.

Some similar studies concerning this work viz., weather models developed by Mehta et al. [1], Agarwal et al. [2] and Ramasubramanian et al. [3] were successfully used for forecasting yields of various crops at district as well as agro-climatic zone level in different states of India. Hoogenboom [4], Kandiannan et al. [5], Bazgeer et al. [6], Esfandiary et al. [7], Lobell and Burke [8], Basso et al. [9] etc. have used a series of weather predictors for crop yield forecasting. Verma et al. [10] & [11] and Goyal and Verma [12] have used agromet/spectral indices for the development of crop yield models of different crops in Haryana (India). Azfar et al. [13] used principal component analysis for rapeseed and mustard yield forecast models for Faizabad district of U.P. (India).

2. MATERIALS AND METHODS

2.1 Crop Status and Data Description

Wheat is one of the most important cereal crops in India as it forms a major constituent of the staple diet of a large part of the population. India is the second largest producer among wheat growing countries of the World (Source: www.mapsofindia.com/indiaagriculture). Haryana occupies the third position in wheat production among the various states in India (www.agricoop.nic.in/statistics). Haryana is self-sufficient in food grains production and also one of the top contributors of food grains to the central pool. Wheat occupies the foremost position followed by rice, not only regarding acreage and production but also in the versatility in adopting different soils and climatic conditions.

The Haryana state comprised of 22 districts and is situated between 74^{o} 25' to 77^{o} 38' E longitude and 27^{o} 40' to 30^{o} 55' N latitude. The total geographical area of the state is 44212 sq. km. Location of the study area is provided in Fig. 1.

The DOA wheat yield data published by Bureau of Economics and Statistics, Haryana were compiled for the period 1980-1981 to 2013-2014 for Ambala, Kurukshetra, Rohtak, Karnal, Jind, Sonepat, Gurgaon, Faridabad, Mahendergarh, Hisar, Sirsa and Bhiwani districts, 1989-1990 to 2013-2014 for Yamunanagar, Panipat, Kaithal and Rewari, 1995-1996 to 2013-2014 for Panchkula, 1997-1998 to 2013-2014 for Jhajjar and Fatehabad and 2006-2007 to 2013-2014 of Mewat district(s) and the same were used to carry out linear time-trend analysis and then computing the district-level trend based yield TY = a + bt, where TY = Trend yield, a = Intercept, b = Slope and t = Year. The weather data for the last 34 years (1980-1981 to 2013-2014) were collected from India Meteorological Department (IMD) and different meteorological observatories

of Haryana. Since the climatic data from adequate number of stations were not available; the districts having equable climatic conditions were grouped into four agro-climatic zones based on their physiography/soils and agro-climatic conditions in Haryana viz., zone-1: Ambala, Panchkula, Yamuna Nagar, Kurukshetra, zone-2: Karnal, Kaithal, Jind, Panipat, Sonipat, Rohtak, zone-3: Mahendergarh, Rewari, Jhajjar, Gurgoan, Faridabad, Mewat and zone-4: Sirsa, Fatehabad, Hisar, Bhiwani.

2.2 Computation of Weather Parameters

Weather data starting from 1st fortnight of November to 1 month before harvest were utilised for the model building (crop growth period: 1st November to 15th April). The various fortnightly weather parameters were computed as follows:

Average Maximum Temperature (TMX) = $\frac{\sum_{i=1}^{15} \text{TMXi}}{15}$

Average Minimum Temperature (TMN) = $\frac{\sum_{j=1}^{15} \text{TMNj}}{15}$

Accumulated Rainfall (ARF) = $\sum_{k=1}^{15} ARF_K$

Where

TMX_i = i^{th} day maximum temperature
TMN_j = j^{th} day minimum temperature
ARF_K = k^{th} day rainfall
i,j,k = daily meteorological data

2.3 Principal Component Analysis

The use and interpretation of a multiple regression model often depends explicitly or implicitly on the assumption that the explanatory variables are not strongly interrelated. In most regression applications, the explanatory variables are not orthogonal. Usually the lack of orthogonality is not severe enough to affect the analysis. However in some situations, the explanatory variables are so strongly interrelated that the regression results are ambiguous. Under such situation, it is impossible

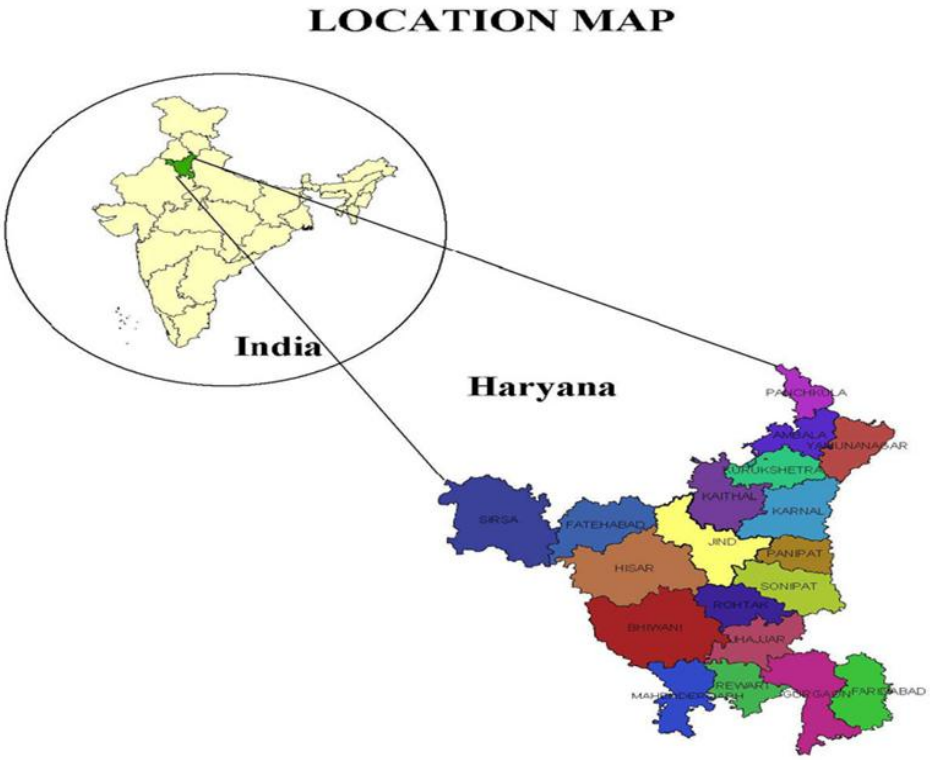

Fig. 1. Location map of the study area

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. To overcome the problem of multicollinearity observed among weather data, the crop yield models are developed within the framework of PCA.

PC method was used for the extraction of factors which consists of finding the eigen values and eigen vectors. Principal components (i.e. P_i, i=1,2,….) were obtained as P = kX , where P and X are the column vectors of transformed and the original variables 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 λ_i of the correlation matrix R; λ_s were obtained by solving the equation $|R- \lambda I| =0$. For each λ, the corresponding characteristic vector k was obtained by solving $|R-\lambda I| k=0$.

3. RESULTS AND DISCUSSION

PC method consists of finding the eigen roots and eigen vectors of the correlation matrix of explanatory variables. The most frequently used convention is to retain the components whose eigen values are greater than one. Weather data starting from first fortnight of November to first fortnight of March *i.e.* one month before crop harvest were utilised for the model development. In this study, first ten eigen values of the correlation matrix of explanatory variables (weather parameters) suggested ten factor(s) solution (Table 1). However, the remaining components accounted for a smaller amount of total variation. Hence, those components were not considered to be of much practical significance. Eigen vectors being the weights were used to compute PC scores.

Multiple linear regression models *via* step-wise regression (Draper and Smith [14]) were fitted by considering PC scores as regressors and wheat yield as the dependent variable. The best subsets of weather variables were selected using stepwise regression method in which all variables were first included in the model and eliminated one at a time with decisions at any particular step conditioned by the result of the previous step. The best supported weather variables were retained in the model if they had the highest adjusted adj. R^2 and lowest standard error (SE) of estimate at a given step. The selected zonal trend-agromet-yield models are as follows:

Zone 1: $Yield_{est}$ = -2.93 +1.08 Tr + 0.77 PC_3 -0.81 PC_4 + 0.85 PC_6

R^2= 0.880 adj.R^2 =0.875 SE =2.64

Zone 2: $Yield_{est}$ = 0.31 +0.99 Tr -0.73 PC_1 + 0.81 PC_2 -0.51 PC_7

R^2= 0.911 adj.R^2 =0.909 SE =2.14

Zone 3: $Yield_{est}$ = 1.31 +0.99 Tr + 0.84 PC_3 -0.50 PC_6

R^2= 0.860 adj.R^2 =0.857 SE =2.55

Zone 4: $Yield_{est}$ = -1.78+1.05 Tr -1.10 PC_6 + 1.02 PC_{10}

R^2= 0.827 adj.R^2 =0.822 SE =2.93

where Yieldest - Model predicted yield
T_r - Linear time trend based yield
PC_i - i^{th} principal component score (i = 1,2,…,10)
SE - Standard error of yield estimate
R^2 - Coefficient of determination

There is a good agreement between the forecast yields and State DOA yield estimates i.e. the real time data. The model predicted yields along with observed yields and per cent relative deviations are given in Table 2.

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

Components	Eigen value (% variance explained)			
	Zone 1	Zone 2	Zone 3	Zone 4
1	4.91(18.17)	5.10(18.90)	4.48(16.59)	4.52(16.75)
2	3.48(12.90)	3.74(13.85)	3.41(12.62)	3.45(12.79)
3	2.67(09.89)	3.02(11.19)	2.73(10.12)	2.77(10.25)
4	2.33(08.62)	2.81(10.39)	2.33(08.65)	2.59(09.61)
5	2.24(08.28)	2.07(07.67)	2.11(07.81)	2.09(07.74)
6	1.94(07.20)	1.89(07.02)	2.04(07.56)	1.79(06.63)
7	1.73(06.40)	1.48(05.48)	1.71(06.34)	1.62(05.99)
8	1.36(05.03)	1.12(04.15)	1.28(04.74)	1.34(04.96)
9	1.24(04.61)	1.06(03.92)	1.17(04.34)	1.10(04.08)
10	1.06(03.94)	0.90(03.32)	0.96(03.56)	1.07(03.96)

Table 2. District-specific wheat yield estimates alongwith percent deviations from DOA yield(s) using fitted models

Districts/ Years	Ambala			Kurukshetra			Yamunanagar		
	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)
2010-11	36.08	41.12	-13.96	44.72	49.45	-10.58	45.57	41.21	9.57
2011-12	48.61	43.58	10.34	54.38	51.93	4.51	53.64	43.56	18.80
2012-13	42.06	42.30	-0.58	46.57	50.66	-8.77	43.39	42.15	2.85
2013-14	46.31	43.27	6.56	48.78	51.64	-5.85	45.31	43.00	5.09
Average absolute deviation	7.86			7.43			9.08		

Districts/ Years	Rohtak			Karnal			Jind		
	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)
2010-11	45.52	43.28	4.92	44.47	49.36	-10.99	45.45	48.26	-6.19
2011-12	50.2	43.08	14.18	56.7	49.24	13.15	52.35	48.23	7.88
2012-13	37.38	44.38	-18.72	46.7	50.62	-8.39	42.7	49.69	-16.36
2013-14	39.24	44.52	-13.44	49.12	50.84	-3.51	45.49	49.99	-9.89
Av. abs. Dev	12.81			9.01			10.08		

Districts/ Years	Sonipat			Panipat			Kaithal		
	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)	Obs. yield (q/ha)	Fitted yield (q/ha)	(RD%)
2010-11	46.43	47.99	-3.35	40.88	43.21	-5.70	47.2	47.36	-0.34
2011-12	55.21	47.94	13.17	39.17	43.48	-11.00	54.51	47.06	13.66
2012-13	45.2	49.40	-9.29	42.47	45.24	-6.53	46.84	48.26	-3.04
2013-14	46.5	49.69	-6.86	42.83	45.85	-7.06	48.15	48.31	-0.33
Av. abs. Dev	8.17			7.57			4.34		

Districts/ Years	Gurgaon			Faridabad			Mahendergarh		
	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)
2010-11	47.12	42.48	9.86	45.45	44.44	2.23	46.39	44.53	4.00
2011-12	49.62	43.82	11.69	48.37	45.67	5.58	46.11	45.68	0.93
2012-13	45.54	44.07	3.23	45.39	45.82	-0.94	47.77	45.74	4.25
2013-14	44.87	44.59	0.62	46.63	46.23	0.86	50.23	46.06	8.29
Av. abs. Dev	6.35			2.40			4.37		

Districts/ Years	Rewari			Jhajjar			Hisar		
	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)
2010-11	48.58	43.38	10.70	45.08	40.44	10.30	46.22	48.02	-3.90
2011-12	50.02	44.33	11.38	48.58	41.33	14.92	50.98	49.92	2.07
2012-13	49.28	44.19	10.33	39.71	41.14	-3.59	42.73	48.48	-13.45
2013-14	49.37	44.31	10.25	45.13	41.20	8.70	44.77	49.17	-9.83
Av. abs. Dev	10.66			9.38			7.31		

Districts/ Years	Sirsa			Bhiwani			Fatehabad		
	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)	Obs. Yield (q/ha)	Fitted Yield (q/ha)	(RD%)
2010-11	51.3	47.67	7.08	44.65	41.79	6.41	50.81	48.71	4.13
2011-12	53.57	49.56	7.49	43.06	43.67	-1.41	54.72	50.54	7.64
2012-13	48.42	48.10	0.67	40.55	42.20	-4.06	46.81	49.02	-4.72
2013-14	53.47	48.77	8.79	42.28	42.85	-1.34	53.18	49.63	6.67
Av. abs. Dev	6.01			3.31			5.79		

Percent Relative Deviation (RD%) = 100 × [(observed (obs.) yield – fitted yield) / observed yield]

The analysis was carried out to see the impact of weather parameters for pre-harvest wheat yield forecasting on agro-climatic zone basis in Haryana state. The developed zonal models are based on time-series data of weather parameters from 1980-81 to 2009-10 and trend based yield as well, however, the data from 2010-11 to 2013-14 were used for validation of the models. Year/time variable was included to take care of variation between districts within zone as the weather data were not available for all the districts, though the zonal model utilised the same weather data in the adjoining districts under the zone. Data for the last one month of wheat crop season were excluded from the analysis, as the idea behind the study was to predict yield(s) about one month in advance of the actual harvest.

4. CONCLUSION

The developed zonal models were used to obtain yield forecasts at district level in Haryana. Trend yield is an important parameter appearing in all the models, indicating that most of the variability in yield is explained by *Tr*, which is an indication or technological advancement, improvement in fertiliser/insecticide / pesticide/weedicide used and increased use of high yielding varieties. The other important aspect was to see the usefulness of the zonal models for district-level pre-harvest yield prediction. The predictive performance(s) of the fitted models were observed in terms of the percent deviations of wheat yield forecasts in relation to real time wheat yield(s). The estimated yield(s) from the selected models indicated good agreement with DOA wheat yield estimates by showing 2-10 percent average absolute deviations for most of the districts except for the Rohtak district observing 12.81 percent average absolute deviation from the real-time yield data. Moreover, the fitted models may be used to provide reliable yield forecasts of wheat crop about one month in advance of the crop harvest while the state DOA yield estimates are obtained quite late after the actual harvest of the crop. Thus, the zonal models developed may be considered usable for district-level operational wheat yield forecasting in Haryana. Although this empirical analysis produced a model with adequate accuracy for pre-harvest forecasting purposes but it would also be worthwhile exploring if other summaries of the weather variables using alternative time windows besides the fortnight summaries we used, may improve the models' performance.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Mehta SC, Agarwal R, Singh VPN. Strategies for composite forecast. J. Ind. Soc. Agril. Statist. 2000;53(3):262-272.
2. Agarwal R, Jain RC, Mehta SC. Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis. Indian J. of Agril. Sci. 2001;71(7): 487-490.
3. Ramasubramanian V, Agrawal R, Bhar LM. Forecasting sugarcane yield using multiple Markov chains. (IASRI, New Delhi publication); 2004.
4. Hoogenboom G. Contribution of agrometeorology to the simulation of crop production and its applications. Agricultural and Forest Meteorology. 2000;103:137-157.
5. Kandiannan K, Chandaragiri KK, Sankaran N, Balasubramanian TN, Kailasam C. Crop-weather model for turmeric yield forecasting for Coimbatore District, Tamil Nadu, India. Agricultural and Forest Meteorology. 2002;112:133-137.
6. Bazgeer S, Kamali Gh, Mortazavi A. Wheat yield prediction through agrometeorological indices for Hamedan, Iran. Biaban. 2007;12:33-38.
7. Esfandiary F, Aghaie G, Mehr AD. Wheat yield prediction through agro meteorological indices for Ardebil district. World Academy of Science: Engineering and Technology. 2009;49:32-35.
8. Lobell DB, Burke M. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology. 2010;150:1443–1452.
9. Basso B, Fiorentino C, Cammarano D, Cafiero G, Dardanelli J. Analysis of rainfall distribution on spatial and temporal patterns of wheat yield in Mediterranean environment. European Journal of Agronomy. 2012;41:52–65.
10. Verma U, Dabas DS, Hooda RS, Kalubarme MH, Yadav M, Sharma MP. Remote sensing based wheat acerage and spectraltrend-agrometeoro-logical yield forecasting: Factor analysis approach. Society of Statistics, Computers and Applications. 2011;9(1&2):1-13.

11. Verma U, Piepho HP, Goyal A, Ogutu JO,d kalubarme MH. Role of climatic variables and crop condition term for mustard yield prediction in Haryana (India). International J. of Agricultural and Statistical Sciences. 2016;12(1):45-51.
12. Goyal M, Verma U. Development of weather-spectral models for pre-harvest wheat yield prediction on agro-climatic zone basis in Haryana, International J. of Agricultural and Statistical Sciences. 2015;11(1):73-79.
13. Azfar M, Sisodia BVS, Rai VN, Devi M. Pre-harvest forecast models for rapeseed and mustard yield using principal component analysis of weather variables. Mausam. 2015;66(4):761-766.
14. Draper NR, Smith H. Applied regression analysis. 3rd edition, John Wiley & Sons. New York; 2003.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/27406