"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2015, 96-105 | Science Signpost Publishing

# Linear Mixed Modeling for Mustard Yield Prediction in Haryana State (India)

U. Verma
*Department of Mathematics and Statistics, CCS Haryana Agricultural University Hisar, India.*
H. P. Piepho, K. Hartung, J. O. Ogutu
*Bioinformatics Unit, University of Hohenheim, Stuttgart, Germany.*
A. Goyal
*University College, Kurukshetra University, Kurukshetra (Haryana), India.*

## Abstract

Crop forecasting is a formidable challenge. 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. This study deals in developing a methodology for pre-harvest crop yield prediction of major mustard growing districts in Haryana (India). Zonal yield models using agro-meteorological parameters were generated using multiple linear regression and mixed model procedures. The common weather-based approach to yield forecast is linear regression with constant coefficients over time. This may be restrictive and of limited prediction power since it does not account for the year-to-year dependence in the yield variable. A mixed model procedure provided a flexible way to fit a multi-level model for crop yield prediction. The linear mixed effects models with random time/weather effects at district, zone and state level were fitted for crop yield estimation. The percent deviation(s) of district-level yield forecasts from the real time yield(s) data show a preference for using linear mixed models. The purpose of this paper is also to show the usefulness of the mixed model framework for pre-harvest crop yield forecasting.

*Keywords*: Multiple linear regression, linear mixed model, weather variables, pre-harvest crop forecast and percent relative deviation.

**Corresponding author:**
U. Verma, Department of Mathematics and Statistics, CCS Haryana Agricultural University Hisar, India. E-mail: vermas21@hotmail.com.

# 1. Introduction

India is one of the largest rapeseed-mustard growing countries in the world, occupying the first position in area and the third position in production after the EU27 and China, and contributing around 11% of the world's total production. India's contributions to the world acreage and production are 28.3 and 19.8 per cent, respectively. Rapeseed is a major crop in India, grown on nearly 13% of the cropped land. Rapeseed and mustard crops are adapted to tropical as well as temperate environments and require relatively cool temperatures for optimal growth. It is basically a winter crop and is grown in the rabi season from September-October to February-March. The crop grows well in areas receiving 25 to 40 cm of rainfall and this is provided by the monsoon rains during the sowing season of the crop in India. Brassica (rapeseed-mustard) is the second most important edible oilseed crop in India after groundnut and accounts for nearly 30% of the total oilseeds produced in the country. The major rapeseed-mustard growing states of India comprise Haryana, M.P., Rajasthan and U.P. and collectively represent 81 percent of the national acreage and contribute 82.9 per cent to the total rapeseed-mustard production.

The Haryana state, located in northern India, has a total geographical area of 4.42 m ha. Like most of the other states in India, the principal occupation in the villages of Haryana is agriculture. About 70% of the total population of Haryana are dependent upon agriculture to earn their livelihood and this has made the state self-sufficient in food grains production. There are two major types of crops namely Rabi and Kharif cultivated in the villages of Haryana, depending on the two cultivation seasons. The Haryana state is one of the top contributors of food grains to the Indian market. In fact, Haryana together with Punjab are called the `Grain Bowl` of India.

Climate change is a major concern today and researchers are engaged in understanding its impact on growth and yield of crops. Changes in seasonal temperatures affect crop yield, mainly through their effect on phenological and developmental processes. Winter crops are especially vulnerable to high temperatures during the reproductive stages and their differential responses to rising temperatures can have important consequences for crop yield. In many previous studies, yield forecasting models have incorporated a series of weather predictors (Kandiannan *et al.* 2002, Verma *et al.* 2003, Dadhwal *et al.* 2003). Models developed by Mehta *et al.* (2000), Agarwal *et al.* (2001) and Ramasubramanian *et al.* (2004) were successfully used for forecasting yields of various crops at district as well as agro climatic zone level in different states of India. As well; Bazgeer *et al.*, 2007, Andarzian, 2008, Esfandiary *et al.*, 2009, Xingjie *et al.*, 2010, Adrian, 2012 and Verma *et al.*, (2014) have developed and used agromet indices in the context

of crop yield prediction.

Several mixed models have also been developed and used to forecast crop yield. Hall and Clutter (2004) have proposed the use of multivariate multilevel nonlinear mixed effect models for timber yield predictions. Hebel *et al.* (1993) have applied shrinkage estimators to the prediction of French winter wheat yield. A study on crop yield modelling under a mixed modelling framework has been conducted by Lenny *et al.* (2006).

## 2. Study Region and Statistical Methodology

The Haryana state comprised of 21 districts (geographical area: 4.42 m ha) is situated between 74° 25' to 77° 38' E longitude and 27° 40' to 30° 55' N latitude. A time-series of state Department of Agriculture (DOA) mustard yield data spanning 1966-67 to 2006-07 and weather data from 1980-81 to 2006-07 (Source: esaharyana.gov.in/StateStatisticalAbstract/ and different meteorological observatories in Haryana) were collected for the purpose. The major mustard growing districts; Rohtak, Mahendergarh, Rewari, Faridabad, Gurgaon, Bhiwani, Sirsa and Fatehabad were grouped into different zones based on their physiographic/soil or agroclimatic conditions.

Multiple linear regression and linear mixed model procedures incorporating different alternative variance-covariance structures were used for the development of zonal weather-yield models. The predictive abilities of the models were compared by fitting the zonal models using weather variables namely average maximum temperature, average minimum temperature and accumulated rainfall calculated over different fortnights, as covariates. Since, the weather variables affect the crop differently during different phases of its growth period. Thus, to integrate the weather variables over different growth phases, the crop growth period (September to February) was divided into 12 fortnights and daily weather data summarized on a fortnightly basis were prepared for the model building and model testing period(s).

### 2.1 Modeling Procedures

#### 2.1.1 Multiple Linear Regression

The multiple linear regression model was used to relate crop yield(s) to the average maximum temperature, average minimum temperature calculated for 10 fortnights covering the period October to February, and accumulated rainfall for 12 fortnights over the period September to February.

In this method, a dependent (response) variable is regressed with a set of independent (explanatory) variables which may or may not be inter-related among themselves. The standard linear regression model

considered may be written in the form $\mathbf{Y=Xb+\varepsilon}$; where $\mathbf{Y}$ is an (n×l) vector of observations (DOA yields), $\mathbf{X}$ is an (n×p) matrix of known form (weather variables & trend yield), $\mathbf{b}$ is a (p×l) vector of parameters, $\boldsymbol{\varepsilon}$ is an (n×l) vector of errors with the assumptions $E(\boldsymbol{\varepsilon})=\mathbf{0}$ and $V(\boldsymbol{\varepsilon})= \mathbf{I}\sigma^2$; so the elements of $\boldsymbol{\varepsilon}$ are uncorrelated. The normal equations $(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{Y}$ are fitted by least squares technique (here $\mathbf{Y}$, $\mathbf{X}$ & $\mathbf{b}$ are same as above and $(\mathbf{X}'\mathbf{X})$ is the dispersion matrix) providing the solution $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Data for the last one month of the crop season were excluded, as the idea behind the study is to predict yield(s) about one month before the actual harvest. The best subsets of weather variables were selected using a stepwise regression method (Draper and Smith, 1981) in which all variables are 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 $R^2$ and lowest standard error (SE) of yield at a given step. The predictive performance(s) of the zonal yield equations were compared on the basis of adj-$R^2$ and percent deviations of yield estimates from the real-time yields.

### 2.1.2 Linear Mixed Modeling

The approach we adopted under the linear mixed modeling procedure differs from other studies (Hall & Clutter 2004, Hebel et al. 1993, Lenny et al. 2006 etc.) in that the hierarchical structure of the data at the geographical level is exploited. For mixed modeling, the hierarchical data structure of yield is represented as.

$$y_{ijt} = s_t + z_{it} + d_{ijt}$$

where, $y_{ijt}$ = yield in the $j$-th district within $i$-th zone in the $t$-th year

$s_t$ = general state effect in the $t$-th year

$z_{it}$ = effect of the $i$-th zone within state in the $t$-th year

$d_{ijt}$ = effect of the $j$-th district within $i$-th zone in the $t$-th year

For each of the three effects (state: $s_t$, zone: $z_{it}$, district: $d_{ijt}$), we have set up a time-series model with three components: Regression + Time Trend + White noise. *Regression* is a fixed part comprising regression on time as well as on the meteorological covariates. *Time Trend is* comprised of a random part for serial correlation with AR(1) type covariance structure and regression splines (Ruppert et al. 2003). *White noise* is an additional independently distributed random error term. To exemplify the modeling

framework, we considered the simple form illustrated below.

**A hierarchical mixed model for crop yield estimation**

| | State-level | Zone-level | District-level |
|---|---|---|---|
| | $s_t = \alpha + \beta t + e_t + h_t$ | $z_{it} = \gamma_i + \delta_i t + f_{it} + k_{it}$ | $d_{ijt} = \eta_{ij} + \xi_{ij} t + g_{ijt} + m_{ijt}$ |
| Regression: | $\alpha + \beta t$ | $\gamma_i + \delta_i t$ | $\eta_{ij} + \xi_{ij} t$ |
| Time Trend: | $e_t \sim AR(1)$ | $f_{it} \sim AR(1)$ | $g_{ijt} \sim AR(1)$ |
| White noise: | $h_t \sim N(0, \sigma_h^2)$ | $k_{it} \sim N(0, \sigma_k^2)$ | $m_{ijt} \sim N(0, \sigma_t^2)$ |

So, the linear mixed effects models with random time/weather effects at district, zone and state level with AR(1) type covariance structure were fitted for crop yield estimation. The SAS Proc Mixed /Proc Glimmix statement for fitting the linear mixed models are given in the end.

## 3. Results and Discussion

The objective of the agromet yield modeling was to assess the predictive accuracies of the contending models for estimating district-level crop yields and how the accuracies are influenced by grouping the districts into zones. Hence, the crop growth period was split into 12 fortnights and the fortnightly weather variables were used as covariates to select the suitable zonal models to estimate the pre-harvest crop yields shown in Table-1.

**Zonal trend- agromet yield relationships based on multiple linear regression analysis**

**Zone-1 (Rohtak)**

$Yield_{est}$ = -1105.84+79.43 $TMX_3$+54.89 $TMN_1$-20.06 $ARF_7$-82.82 $TMX_4$+.951Tr

$R^2$ = 0.85, Adj. $R^2$ = 0.82 & SE = 148.8

**Zone-2 (Bhiwani, Sirsa, Fatehabad)**

$Yield_{est}$ = -950.23-10.94 $ARF_3$-51.32 $TMN_{10}$+35.74 $TMX_9$+ 29.77 $TMX_7$ -21.44 $TMX_6$+1.55 Tr

$R^2$ = 0.82, Adj. $R^2$ = 0.80 & SE = 150.2

**Zone-3 (Faridabad, Gurgaon, Mahendergarh, Rewari)**

$Yield_{est}$ = 2327.83-78.17 $TMX_7$+49.27 $TMN_5$+163.47 $TMN_9$-16.79 $ARF_4$-51.65 $TMX_2$-103.92 $TMX_8$-9.55 $ARF_9$+ 93.52 $TMX_4$

$R^2$ = 0.75 , Adj. $R^2$ = 0.72 & SE = 168.0

where $Yield_{est}$ - Model predicted yield (q/ha)

Tr - Linear time-trend based yield

TMX - Av. maximum temperature

TMN - Av. minimum temperature

ARF - Accumulated rainfall (1,2,3,…,10/12 are different fortnights covering the crop growth period)

SE - Standard error of the yield estimate

**Table 1. District-specific estimated mustard yields (Est. yield) based on zonal models and their associated percentage deviations (RD (%) = 100 × (Est. yield -observed yield)/ observed yield)**

i) Weather variables and trend yield were used as regressors

| Districts/Years | Fatehabad | | Faridabad | | Gurgaon | | Mahendergarh | | Rewari | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) |
| 2004-05 | 13.93 | 21.83 | 11.08 | -6.01 | 11.08 | -4.64 | 11.08 | -7.27 | 11.08 | -20.79 |
| 2005-06 | 16.61 | 17.02 | 6.52 | -57.50 | 6.52 | -47.68 | 6.52 | -44.19 | 6.52 | -57.24 |
| 2006-07 | 16.45 | 12.56 | 14.52 | -2.75 | 14.52 | 14.24 | 14.52 | 2.76 | 14.52 | -0.42 |
| **Av. abs. deviation** | | **17.13** | | **22.08** | | **22.18** | | **18.07** | | **26.15** |

| Districts/Years | Rohtak | | Bhiwani | | Sirsa | |
|---|---|---|---|---|---|---|
| | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) |
| 2004-05 | 14.11 | 12.34 | 12.86 | 18.73 | 13.63 | 18.12 |
| 2005-06 | 13.09 | 48.78 | 15.26 | 36.36 | 16.10 | 32.63 |
| 2006-07 | 13.81 | 2.27 | 14.82 | 15.13 | 15.73 | 19.35 |
| **Av. abs. deviation** | | **21.12** | | **23.40** | | **23.36** |

ii) Linear mixed model for yield with weather and penalized spline smoothing of time trend

| Districts/Years | Fatehabad | | Faridabad | | Gurgaon | | Mahendergarh | | Rewari | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) |
| 2004-05 | 13.81 | 17.26 | 14.56 | 19.05 | 12.42 | 6.43 | 13.22 | 9.61 | 14.77 | 5.28 |
| 2005-06 | 14.01 | -1.25 | 14.69 | -4.47 | 12.51 | 0.31 | 13.31 | 12.19 | 14.97 | -1.91 |
| 2006-07 | 14.21 | -2.78 | 14.82 | -0.73 | 12.60 | -0.89 | 13.40 | -5.41 | 15.18 | 3.94 |
| **Av. abs. deviation** | | **7.09** | | **8.08** | | **2.54** | | **9.07** | | **3.71** |

| Districts/Years | Rohtak | | Bhiwani | | Sirsa | |
|---|---|---|---|---|---|---|
| | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) | Est. Yield (q/ha) | RD (%) |
| 2004-05 | 11.99 | -4.42 | 11.59 | 7.14 | 13.84 | 16.64 |
| 2005-06 | 8.17 | -7.00 | 10.02 | -10.45 | 14.04 | 13.52 |
| 2006-07 | 13.81 | 2.38 | 12.17 | -5.40 | 14.23 | 7.39 |
| **Av. abs. deviation** | | **4.60** | | **7.66** | | **12.51** |

The predictive accuracies of the zonal regression models and linear mixed models expressed in terms of the per cent deviations of the estimated yields from the observed yields, differed markedly for all the three zones. Though, all the weather variables were statistically significant as predictors of crop yield, however, the percent relative deviations of the estimated yields from the observed yields were too wide for practical purposes for the sample period itself in case of regression models. The yields estimated had rather wide percent deviations from the observed yields, sometimes too wide than considered acceptable for reliable yield prediction purposes. We have attempted to improve the predictive accuracy of the zonal yield models by using linear mixed modeling and the linear mixed models substantially improved the predictive accuracy and produced what we consider to be satisfactory district-level yield(s) estimation. Hence, based on this empirical study, we recommend the use of linear mixed models for pre-harvest yield forecasting of mustard crop to enhance the predictive accuracy of the zonal models. This work has demonstrated the utility of understanding and quantifying the relationships between mustard yield and weather variables. The relationships can be employed in studies that explore the impact of climate change on probable future crop yields at regional scales. Moreover, it has also shown the usefulness of the mixed model framework for pre-harvest crop yield forecasting.

## Acknowledgement

The weather data received from Haryana Space Applications Centre, Hisar and Department of Agro-meteorology, CCS HAU, Hisar, India are gratefully acknowledged.

## References

[1]. D. Adrian, A model-based approach to forecasting corn and soybean yields. USDA, National Agricultural Statistics Service, Research & Development Division (2012).

[2]. R. Agarwal, R.C. Jain and S.C. Mehta, Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis. Indian J. of Agril. Sci. 71 (7) (2001), 487-490.

[3]. B. Andarzian, A.M. Bakhshandeh, M. Bannayan, Y. Emam, G. Fathi and S. Alami, Wheat Pot : a simple model for spring wheat Yp using monthly weather data. Biosyst. Eng. 99(2008), 487–495.

[4]. S. Bazgeer, Gh. Kamali, and A. Mortazavi, Wheat yield prediction through agrometeorological indices for Hamedan, Iran. Biaban 12 (2007), 33-38.

[5]. V. K Dadhwal, V. K., Sehgal, R. P. Singh, and D.R. Rajak, Wheat yield modeling using satellite remote sensing with weather data; recent Indian experience. Mausam 54 (2003), 253-262.

[6]. N. Draper and H. Smith, Applied Regression Analysis. 2nd edition. New York: Wiley (1981).

[7]. F. Esfandiary, G. Aghaie, and A. D. Mehr, Wheat yield prediction through agro meteorological indices for Ardebil district. World Academy of Science, Engineering and Technology 49(2009), 32-35.

[8]. D. B. Hall and M. Clutter, Multivariate multilevel nonlinear mixed effects models for timber yield predictions. Biometrics 60 (2004), 16-24.

[9]. P. Hebel, R. Faivre, B. Goffinet, and D. Wallach, Shrinkage estimators applied to prediction of French winter wheat yield. Biometrics 49 (1993), 281-293.

[10]. K. Kandiannan, K.K. Chandaragiri, N. Sankaran, T.N. Balasubramanian and C. Kailasam, Crop-weather model for turmeric yield forecasting for Coimbatore District, Tamil Nadu, India. Agricultural and Forest Meteorology 112 (2002), 133- 137.

[11]. W. Lenny L. Denis and ML Pierre, The early explanatory power of a satellite based measure in crop yield modelling. Les Cahier du CREF 06-15 (2006), 1-25.

[12]. S.C. Mehta, R. Agarwal and V.P.N. Singh, Strategies for composite forecast. J. Ind. Soc. Agril. Statist. 53 (3) (2000), 262-272.

[13]. V. Ramasubramanian and R.C. Jain, Use of growth indices in Markov Chain models for crop yield forecasting. Biom. J. 41 (1) (1999), 99-109.

[14]. D. Ruppert, M.P. Wand and R.J. Carroll, Semiparametric regression. Cambridge University Press, Cambridge (2003).

[15]. U. Verma, D.S. Ruhal, R.S. Hooda, M. Yadav, A.P. Khera, C.P. Singh, M.H. Kalubarme and I.S. Hooda, Wheat yield modelling using remote sensing and agrometeorological data in Haryana State. J. Ind. Soc. Agric. Statist. 56, 2 (2003), 190-98.

[16]. U. Verma, H.P. Piepho, M. Goyal and A. Goyal, Impact of climatic variables on sugarcane yield prediction in Haryana (India). Advances and Applications in Statistics 39, 1 (2014), 25-35.

[17]. Ji. Xingjie, Yu Yongqiang and Wen Zhang, The Harvest Index Model of Winter Wheat in China based on meteorological data. Scientia Agricultura Sinica 43, 20 (2010), 4158-68.

**SAS code:**

```
Proc glimmix data=mustard Noclprint Method=RSPL;
class district zone;
Model Yield=Zone*district Zone*district*time /noint dist=normal link=identity ddfm=kr
solution;
random time/sub=intercept     type=pspline knotmethod=equal(20);
random time/sub=zone          type=pspline knotmethod=equal(20);
```

```
random time/sub=zone*district  type=pspline knotmethod=equal(20);
output out=mustard_yld_pred Pred(ilink)=mu LCL(ilink)=lower UCL(Ilink)=Upper;
nloptions tech=NEWRAP Maxiter=1000 maxfunc=1000;
run;
```

```
Proc glimmix data=mustard Noclprint Method=RSPL itdetails;
class district zone;
Model Yield=Zone*district Zone*district*time / dist=normal link=identity ddfm=kr solution
htype=1;
random time/sub=intercept        type=pspline knotmethod=equal(10);
random time/sub=zone             type=pspline knotmethod=equal(10);
random time/sub=zone*district    type=pspline knotmethod=equal(10) knotinfo ;
random RF1-RF12/ sub=intercept   type=ar(1) s;
random Tmx1-Tmx10/  sub=intercept   type=ar(1) s;
random Tmn1-Tmn10/sub=intercept   type=ar(1) solution;
output out=mustard_yld_pred Pred(ilink)=mu LCL(ilink)=lower UCL(Ilink)=Upper;
nloptions tech=NEWRAP Maxiter=10000 maxfunc=10000;
run;
```