Indian Journal of Agricultural Sciences **89** (3): 540–4, March 2019/Article
https://doi.org/10.56093/ijas.v89i3.87602

Forecasting crop yield through weather indices through LASSO

K N SINGH[1], K K SINGH[2], SUDHEER KUMAR[3], SANJEEV PANWAR[4] and BISHAL GURUNG[4]

ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110 012

Received:15 May 2018; Accepted: 20 November 2018

ABSTRACT

Reliable forecast of crop production before the harvest is important for advance planning, formulation and implementation policies dealing with food procurement, its distribution, pricing structure, import and export decisions, and storage and marketing of the agricultural commodities. Weather plays a very important role in crop growth and development. Therefore, model based on weather variables can provide reliable forecast. Weather variables used can be employed for crop production forecast by making appropriate models. In this study, a statistical model is used for crop yield forecast at different growth stages of wheat crop. This model uses maximum and minimum temperature, rainfall, morning and evening relative humidity during crop growing period. The forecast model was developed using generated weather indices as regressors in model. In order to select significant weather variables affecting the yield of crop least absolute shrinkage and selection operator (LASSO) as well as stepwise regression methodology is applied. The result of lasso gives a better result as compared to stepwise regression. The R^2 of lasso and stepwise regression are 0.84 and 0.85, respectively. The mean square error (MSE) and root mean square error (RMSE) of Lasso regression were better than stepwise regression, which leads to improvement of crop yield forecasting. It can be inferred that for the data under consideration, lasso works better than stepwise regression for variable selection.

Key words*:* Correlation coefficient, Forecast model, Lasso, Stepwise regression, Weather variables

Agriculture plays a vital role in India's economy and it has been considered as the backbone of the country's economy. It is a country with almost 70% of its population directly or indirectly engaged in agricultural practices for meeting up their daily livelihood. The agriculture sector provides employment to 58.4% of the country's workforce. It is the single largest private sector occupation and share of agriculture and allied sectors (including agriculture, livestock, forestry and fishery) was 15.35%. In such a country where agriculture is considered important sector of economy and livelihood for population, reliable and routine forecasting of crop yield plays vital role for advance planning, formulation and implementation of a number of policies dealing with food procurement, its distribution, pricing structure, import and export decisions, and for exercising several administrative measures related to storage and marketing of the agricultural commodities. Accurate crop yield forecasting can also become important in a changing climate characterized by increasingly variable weather (Salinger 2005).

There are many techniques that have been developed to forecast crop production. However, there are two main approaches for predicting crop yield: simulation models and multiple regression models. The simulation models designed to forecast crop yield use details about crop biology and requires extensive information such as soil type, plant parameters, and weather data related to the crop development stage, which are often not readily available. In order to overcome this difficulty, statistical models have been developed based on the weather parameters. Multiple regression approach used to predict the crop production is not only easier to use, but is likely to be more accurate than the simulation model approach as the parameters estimated using Ordinary Least square are best linear unbiased estimate (BLUE). Tannura *et al*. (2008) and other studies have proven that multiple regression models have high explanatory power and can represent relationships between weather conditions and crop yield.

The effect of weather on crop growth varies with different stages of crop growth. It has been found that the influence of weather on the yield of a crop depends on the magnitude of weather variables as well as on the manner in which the weather gets distributed over the different growth stages of the crop because different growth stages of the crop growth have different sensitivities towards weather parameters; few of them are very sensitive to weather fluctuations whereas others are less sensitive. Hence, for precise forecasting we need to divide the entire crop growth phase into very fine intervals. As a result there

[1]Head (e mail: kn.singh@icar.gov.in). [2,3]India Meteorological Department, New Delhi. [4]ICAR-Krishi Bhawan, New Delhi.

is an increase in the number of variables in the model and consequently more parameters have to be evaluated from the data. Hence, in such situations a very long series data is a prerequisite for precise estimation of the parameters, which may be practically very difficult to obtain. Hence, the solution to this situation lies in the fact that one has to seek a model which is based on less number of parameters that could be easily evaluated and at the same time it should also take into account the pattern or the manner in which the weather is distributed over the entire crop growth phase.

The two main goals of linear regression model are prediction accuracy, and complexity of the model should be as less as possible. Complexity of model depends on set of predictor variable and determining this subset is called variable selection problem. The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Many techniques have been developed for selection of predictor variables like forward selection, backward selection, stepwise regression, ridge regression, etc. Stepwise regression is a modification of the forward selection, after each step in which variable are added, all predictor variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model. Ridge Regression is a technique used when the data suffers from multicollinearity. In multicollinearity, even though the least squares estimate (OLS) are unbiased; their variances are large which deviates the observed value far from the true value.

Ridge regression reduces the standard errors. But it does not set value of any coefficients to zero and thus ridge regression cannot produce a parsimonious model. To overcome this limitation lasso can perform well which is able to perform variable selection by shrinking the estimated value of some of the regression coefficients which are less important exactly equals to zero. A vastly popular and successful approach in statistical modeling is to use regularization penalties in model fitting (Hoerl and Kennard 1970). By jointly minimizing the empirical error and penalty, one seeks a model that not only fits well and is also "simple" to avoid large variation which occurs in estimating complex models. Lasso (Tibshirani 1996) is a successful idea that falls into this category. In order to improve the accuracy of forecast of crop production lasso can better perform for selection of predictor variables which can reduce the number of predictors in a regression model and identify important predictors more effectively.

MATERIALS AND METHODS

Crop yield and weather data are required to forecast the crop yield using the statistical yield forecast model. This study was done for wheat crop for Delhi region. The daily data on weather parameters such as maximum and minimum temperature, morning and evening relative humidity, amount of rainfall for 23 years period have been collected from weather station located at IARI New Delhi. Wheat yield data (1984-2015) were also collected from IARI, New Delhi.

The wheat crop yield forecast was done before harvest using the following statistical model:

$$Y = A_0 + \sum_{i=1}^{p}\sum_{j=0}^{1} a_{ij}Z_{ij} + \sum_{i\neq i'=1}^{p}\sum_{j=0}^{1} a_{ii'j}Z_{ii'j} + cT + e \quad (1)$$

where,

$$Z_{ij} = \sum_{w=1}^{m} r^{J}_{iw}X_{iw} \quad \text{and} \quad Z_{ijw} = \sum_{w=1}^{m} r^{J}_{ii'w}X_{iw}X_{i'w}$$

where, Y is the wheat yield (kg/ha), r_{iw} is correlation coefficient of yield with i-th weather variable in w-th period, rii'w is correlation coefficient (adjusted for trend effect) of yield with product of i-th and i'-th weather variables in w-th period, p is number of weather variables used, m is period of forecast, e is error term and is normally distributed and has a mean zero.

This model was successfully used for forecasting yields of various crops at district level as well as agro-climatic zone level (Agrawal *et al* 2001, Mehta *et al*.2000). The approach has been successfully used for forecasting yields of rice, wheat and sugarcane for Uttar Pradesh. (Agrawal *et al*. 2001). Using same above approach, wheat crop yield indices are estimated, stepwise and lasso regression were used to get coefficient. Further, this coefficient estimated by two regression approach was used to developed crop yield model and to make comparison between them so as to find which one gives best result.

Stepwise regression is a modification of the forward selection so that after each step in which in this regression a variable was added, all predictor variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non significant variable is found, it is removed from the model. Stepwise regression requires two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop.

Lasso is a regularization technique which reduces the number of predictors in a regression model and identifies important predictors. Lasso is a shrinkage estimator with potentially lower predictive errors than ordinary least squares and it also includes a penalty term that constrains the size of the estimated coefficients. Therefore, it resembles ridge regression. It generates coefficient estimates that are biased to be small. Nevertheless, a lasso estimator can have smaller mean squared error than an ordinary least-squares estimator when you apply it to new data. Unlike ridge regression, as the penalty term increases, lasso sets more coefficients to zero. This means that the lasso estimator is a smaller model, with fewer predictors. As such, lasso is an alternative to stepwise regression and other model selection and dimensionality reduction techniques.

The lasso technique solves this regularization problem. For a given value of λ, a non-negative parameter, lasso

Table 1 Coefficient of stepwise regression.

Model	Unstandardized coefficients			Standardized Coefficients		
		B	Standard error	Beta	t	Significance
3	(Constant)	2166.712	141.1756		15.34763	1.41E-13
	VAR00001	48.87302	2.584799	0.923325	18.90786	1.67E-15
	VAR00022	1.690226	0.38532	0.228406	4.386557	0.000215
	VAR00014	3.053721	0.861431	0.184309	3.544939	0.001728
a	Dependent Variable: VAR00002					

solves the problem by:

$$\min_{\beta_0 \beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (Y_i - \beta_0 - X_i^T \beta^2) + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right)$$

where, N is the number of observations, y_i is the response at observation i, x_i is data, a vector of p values at observation i, λ is a nonnegative regularization parameter corresponding to one value of Lambda, The parameters β_0 and β are scalar and p-vector respectively, As λ increases, the number of nonzero components of β decreases.

RESULTS AND DISCUSSION

The crop yield was modelled 30 days in advance before harvesting of crop using statistical tool. The regression equation was developed using weather parameters and predictor variables were selected using SPSS software. The result of SPSS is presented in Table 1.

Crop yield forecast model

The yield forecast equation has been developed using the significant generated weather variables with significant at 1% level. The final yield forecast model using weather variables has been presented below.

$$Y = 2166.71 + 48.87X_1 + 3.05X_{14} + 1.69X_{22} \quad (2)$$
$$(64.23) \quad (2.23) \quad (0.78) \quad (0.31)$$

Crop yield forecast

Developed crop yield forecast model were used to predict the crop yield for twenty seven years crop yield. Results were presented in Table 2.

Lasso regression

Same data was performed for selection of variable using lasso regression method. This work was accomplished using R software using the glmnet package. The results of lasso performance for regression coefficient are presented in Table 3.

Crop yield forecast model

The yield forecast equation has been developed using the significant generated weather variables based on equation 1. The final yield forecast function using important weather variables has been presented below.

$$Y = 1202.3 + 47.1X_1 + 8.23X_3 + 4.54X_9 \quad (3)$$
$$(32.32) \quad (2.12) \quad (1.21) \quad (0.52)$$

Table 2 Comparison of actual yield and predicted yield by stepwise regression.

Year	Actual yield	Predicted yield using stepwise regression	Residual
1984	2563	2699	-136
1985	2712	2729	-17
1986	2873	2986	-113
1987	2903	2796	107
1988	2875	2950	-75
1989	3120	2978	142
1990	3180	3146	34
1991	3580	3455	125
1992	3376	3472	-96
1993	3188	3178	10
1994	3331	3222	109
1995	3330	3332	-2
1996	3365	3294	71
2000	4008	4108	-100
2001	3522	3559	-37
2002	3510	3578	-68
2003	4005	3888	117
2004	3516	3516	0
2005	3944	4013	-69
2008	4069	3816	254
2009	4002	3986	16
2010	3748	3921	-173
2011	3850	3951	-101
2012	4220	4299	-79
2013	3939	4009	-70
2014	4134	4311	-177
2015	4811	4484	327

Crop yield forecast

Using the lasso regression technique, yield forecast model was developed as a given equation 3. The performance of the crop yield forecast equation has been tested by comparing the simulated values with the observed values for period from (1984-2015) which are presented in Table 4.

Statistical and graphical comparison

Graphical and statistical indicators were used in order

Table 3 Coefficient of LASSO regression

Variable	Coefficient
Intercept	1202.3
X_1	47.1
X_3	8.23
X_9	4.54

Table 4 Comparison of actual yield and predicted yield by Lasso regression

Year	Actual yield	Predicted yield by using LASSO	Residual (LASSO)
1984	2563	2767	-204
1985	2712	2762	-50
1986	2873	2989	-116
1987	2903	2855	48
1988	2875	2952	-77
1989	3120	3020	100
1990	3180	3148	32
1991	3580	3464	116
1992	3376	3437	-61
1993	3188	3154	34
1994	3331	3282	49
1995	3330	3354	-24
1996	3365	3285	80
2000	4008	4028	-20
2001	3522	3491	31
2002	3510	3568	-58
2003	4005	3895	110
2004	3516	3567	-51
2005	3944	3973	-29
2008	4069	3825	245
2009	4002	3970	33
2010	3748	3913	-165
2011	3850	3920	-70
2012	4220	4224	-4
2013	3939	3991	-52
2014	4134	4341	-207
2015	4811	2767	313

Table 5 Statistical indicators of stepwise and lasso regression

Parameter	Stepwise	Lasso
R^2	0.84	0.85
MSE	14836.14	13385.20
RMSE	121.80	115.70

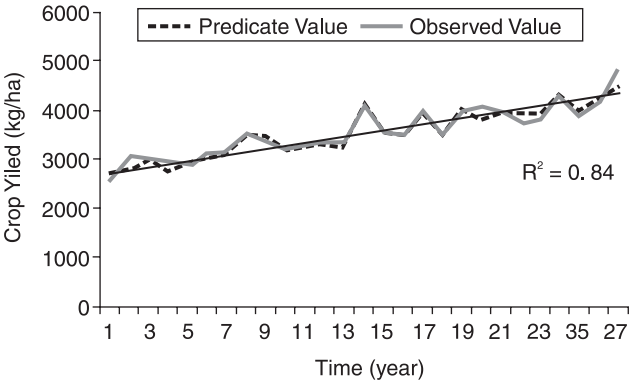

Fig 1 Correlation Coefficient between actual and predicted yield using stepwise regression.

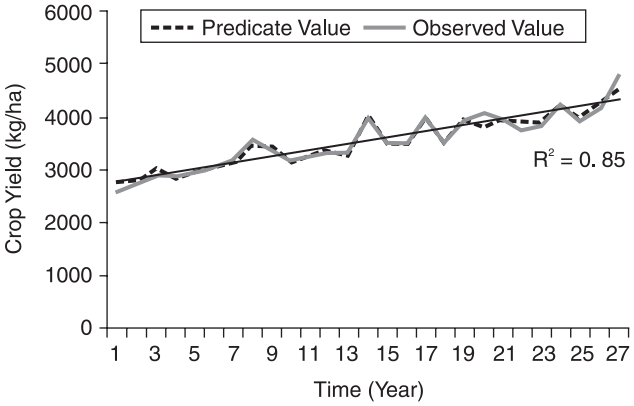

Fig 2 Correlation Coefficient between actual and predicted yield using Lasso regression.

to make comparison of better result between stepwise and Lasso regression. R^2, mean square error and Root mean square error can show how its predicated value is closer to the observed value. Correlation coefficient between actual and predicted yield using stepwise regression and lasso regression is shown in Fig 1 and Fig 2, respectively. R^2 for stepwise and Lasso regression are 0.84 and 0.85, respectively. Statistical indicators for both regressions methods are shown in Table 5. MSE and RMSE of lasso regression are lower than stepwise regression, which represent better result of lasso.

This study was done to find an improved technique of variable selection for crop yield forecast. We have employed LASSO as well as stepwise regression for selection of variables. Using appropriate measures of goodness of fit, we can infer that model where LASSO was employed to forecast yield is able to give better results compared to model where stepwise regression is employed. Further, graphically also it can be seen that model using LASSO gives a better result compared to stepwise regression. Thus, LASSO is more efficient in selecting the variables.

REFERENCES

Agrawal R, Jain R C and Jha M P. 1983. Joint effects of weather variables on rice yield. *Mausam* **34**(2): 189–94.

Agrawal R, Jai R C, Jha M P and Singh D. 1980. Forecasting of rice yield using climatic variables. *Indian Journal of Agricultural*

Sciences **50**(9): 680–4.

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

Hoerl A E and Kennard R W. 1970. Ridge regression: Biased estimation for no orthogonal problems. *Techno metrics* **12**(55)–67.

Mehta S C, Agrawal R and Singh V P N. 2000. Strategies for composite forecast. *Journal of Indian Society of Agricultural Statistics* **53**(3).

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*. Series B **58**(1): 267–88.

Salinger M J. 2005. Climate variability and change: past, present and future—an overview. *Climate Change* **70**(1–2): 9–29.

Tannura M A, Irwin S H and Good D L. 2008. A Review of the Literature on Regression Models of Weather, Technology, and Corn and Soybean Yields in the U.S. Marketing and Outlook Research Report 2008-02, Department of Agricultural and Consumer Economics, University of Illinois at Champaign-Urbana.