Use of Statistical Models in Predicting Groundnut Yield in Relation to Weather Parameters

Abhinaya D[1*], Patil S G[1], Dheebakaran Ga[2], Djanaguiraman M[3], Arockia Stephen Raj[1]

[1*]Department of Physical Sciences & Information Technology, Agricultural Engineering College & Research Institute, Tamil Nadu Agricultural University, Coimbatore-641 003.
[2]Agro Climate Research Centre, Tamil Nadu Agricultural University, Coimbatore.
[3]Department of Crop Physiology, Agricultural College & Research Institute, Tamil Nadu Agricultural University, Coimbatore.

ABSTRACT

In Tamil Nadu, groundnut is an essential and major oilseed crop, mainly grown under rainfed conditions. The changes in weather parameters might affect the productivity of groundnut. Hence, crop yield forecasting based on weather parameters is essential for proper planning, decision-making, and buffer stocking policy formulation. As for the data with multicollinearity, penalized regression models i.e.Ridge, Least Absolute Selection and Shrinkage Operator (LASSO) and Elastic Net (ENet), are better alternatives to classical linear regression. The data on weather parameters such as maximum temperature(Tmax), minimum temperature (Tmin), morning relative humidity (RH I), evening relative humidity (RH II), and rainfall were collected for 29 years from 1991-2019. The weather indices approach was used in this study. The collected data were partitioned into training, and testing datasets and the hyperparameters of penalized regression models were tuned using cross-validation. The performance of the models was evaluated using an adjusted coefficient of determination (R^2adj), Root Mean Squared Error (RMSE), normalized RMSE (nRMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) as the goodness-of-fit criteria. The results revealed that all the Penalized regression models provide a better fit to data. The SMLR and ENet were found to predict with better accuracy. Hence, these methods can be used for groundnut yield forecasting during Kharif season for the Coimbatore district of Tamil Nadu.

Keywords: *Stepwise Multiple Linear Regression; Ridge regression; LASSO; Elastic Net; Groundnut yield prediction, Weather indices.*

INTRODUCTION

Groundnut is one of the most essential oilseed crops of India. In Tamil Nadu, it is annually grown in about 0.35 million hectares with about 0.98 million tons production. In India, Tamil Nadu ranks third in production contributing 9.74% of the total production of groundnut crop in the country, with an average yield of 2840 kg ha^{-1} (Directorate of Economics and Statistics, 2019-20). In Tamil Nadu, two-thirds of the groundnut cultivation area is under rainfed condition and the remaining one-third is under irrigated condition.

The overall growth of the Indian economy relies on the performance of agriculture, which depends , upon the weather conditions every year. A timely and reliable forecast of crop yield is of great importance for monsoon-dependent countries like India, where the economy is mainly based on agricultural production (Vinaya *et al*., 2017).

Crop yield is influenced by technological change and weather variability. Technological factors increase yield smoothly through time and therefore, years or some other parameters of time can be used to study the overall effect of technology on yield (Agarwal *et al.,* 1980). Generally, there are two approaches for crop yield forecasting: crop simulation and empirical statistical models (Bocca and Rodrigues,2016). Though crop simulation models are precise, they are input data-intensive and the lack of sufficient data sets makes their application limited to smaller scales rather than regional scales.

Hence, empirical statistical models with simple regression techniques have been largely used as an alternative to process-based simulation models (Lobell and Burke,2010; Shi *et al* ,2013). Calibrated and tested statistical models lead to successful crop yield forecasting based on weather

parameters. Since groundnut is mainly cultivated under rainfed conditions, weather conditions highly affect its productivity. Most of the previous studies have used Multiple Linear Regressions (MLRs) to develop a statistical crop yield prediction model (Rai *et al*, 2013; Dhekle *et al*,2014; Kumar *et al*,2014). However, MLR results in over-fitting when (a) the number of samples is less than the number of predictors, (b) multicollinearity exists i.e. when the independent variables are correlated (Verma *et al*., 2016). One of the consequences of multicollinearity is the large standard errors of regression coefficients, making the inaccurate inference based on the fitted model (Yakubu, 2010; Dormann *et al*., 2013).

To overcome these drawbacks, feature selection and penalized regression methods such as Stepwise Multiple Linear Regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net (ENET) and ridge regression techniques can be used (Das *et al*.,2017). In this context, the main objective of our study is to develop and select a statistical groundnut yield forecasting model for the Coimbatore district of Tamil Nadu with the predictive performance and efficiency of the developed models.

MATERIAL AND METHODS

Data Collection

Time series data of groundnut yield (*Arachis hypogea L*) for Coimbatore district of Tamil Nadu for 29 years (1991 to 2019) has been collected from the Season and Crop Report, Department of Economics and Statistics. Daily weather data were collected from Agro Climate Research Centre, TNAU. The data on five weather variables namely maximum temperature (Tmax,^{o}C), minimum temperature (Tmin,^{o}C), morning and evening relative humidity (RH I & RH II (%)) and rainfall (mm) for a total of 18 weeks of crop cultivation which includes 14th to 31st standard meteorological week (SMW) has been used in the study as the sowing of the groundnut in Coimbatore district is usually carried out during the month of April-May (Chithiraipattam). Daily data of Tmax, Tmin, RH I and RH II had been converted into its weekly average, whereas the weekly sum of rainfall has been considered. Out of the 29-year data, 24 years were used for calibration, while the remaining 5 years were used for validation.

Detrending of Yield Time Series Data

The fluctuations in yield data over the years due to the technology differences, climatic variability, etc., leads to a nonlinear and non-stationary trend which has to be removed. The correlation between detrended yield and weather parameters is used to calculate weight for model development(Wu *et al*., 2007). In the present investigation, a simple linear regression model has been applied to detrend the yield of groundnut.

$$Y_t = \beta_0 + \beta_1 * t \qquad (1)$$

Where, t is the time period, Y_t is the crop yield at time t, and β_0 and β_1 are the coefficients. This model's residuals (detrended yield) were used for indices calculation (Trnka *et al*.,2009).

Weather Indices Approach

For each weather variable, two indices are developed, one as the total values of weather parameters over different weeks and the other one as weighed total; the weights are the correlation coefficient between detrended yield and weather variable in respective weeks.

Unweighed weather indices: $\quad Z_{ij} = \sum_{w=1}^{m} X_{iw}$

Weighed weather indices: $\quad Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw}$

Where,

X_{iw} = value of ith weather variable in w-th week

r_{iw}^{j} = correlation coefficient of detrended yield with i-th weather variable

m = week of forecast

For j=0, we have unweighed indices and for j=1, weighed indices. Totally 11 weather variables were generated as per the procedure mentioned above is presented in Table.1.

YIELD FORECAST MODELS

Stepwise Multiple Linear Regression

Multiple Linear Regression (MLR) is the most straightforward approach for the development of statistical models. However, its application for the dataset with more significant explanatory variables and is not always successful (Balabin *et al*., 2011). A stepwise regression procedure was adopted to select the best regression variables among many independent variables (Singh *et al*, 2014). A fundamental problem with stepwise regression is that some real explanatory variables that have causal effects on dependant variables may happen to be statistically insignificant, while nuisance variables may be coincidentally significant (Smith *et al*., 2018). Hence, we opt for alternative methods such as penalized regression methods.

Penalized Regression

Penalized regression is a better alternative for the linear regression model (or the ordinary least squares method). The penalized regression adds a constraint (penalty) in the equation. The consequence of imposing this penalty is to reduce the coefficient values towards zero. This allows the less contributive variables to have a coefficient close

to zero or equal to zero. The logic behind penalized regression is to reduce the impact of multicollinearity since all independent variables in the study are related.

Ridge Regression

Ridge regression shrinks the regression coefficients so that variables with a minor contribution to the outcome have their coefficients close to zero. The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called L_2-norm, which is the sum of the squared coefficients (Zou and Hastie, 2005).

$$L_2 = \sum (\hat{Y}_i - Y_i)^2 + \lambda \sum \beta^2 \qquad (2)$$

Where y is the independent variable, β is the corresponding coefficient and λ is the L_2 norm penalty. A large value of λ means a more significant amount of shrinkage. Ridge regression keeps all the predictors in the model without making any variable selection.

Lasso Regression (Least Absolute Shrinkage And Selection Operator)

It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called L_1-norm, which is the sum of the absolute coefficients. In the case of lasso regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero (Tibshirani, 1996). One obvious advantage of lasso regression over ridge regression is that it produces more straightforward and more interpretable models that incorporate only a reduced set of predictors.

$$L_1 = \sum (\hat{Y}_i - Y_i)^2 + \lambda \sum |\beta| \qquad (3)$$

Where y is the independent variable, β is the corresponding coefficient and λ is the L_1 norm penalty.

Elastic Net Regression

Elastic Net combines characteristics of both lasso and ridge, i.e., penalized with both the L_1 and L_2 norm (Hoerl and Kennard, 1970). The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (like in LASSO). Hence it reduces the impact of different features while not eliminating all of the features (Cho et al., 2009).

$$L = \sum (\hat{Y}_i - Y_i)^2 + \lambda_2 \sum \beta^2 + \lambda_1 \sum |\beta| \qquad (4)$$

Where y is the independent variable, β is the corresponding coefficient and λ is the penalty.

These methods have two parameters, namely lambda and alpha, which need to be optimized. The optimal lambda values were selected by minimizing the average mean square error in leave-one-out cross-validation (Piaskowski et al., 2016). The overall strength of the penalty is controlled by tuning parameter λ (Hastie and Qian,2014). The other tuning parameter alpha was set at 0 for Ridge, 1 for LASSO and 0.5 for ELNET. The data were analyzed using 'glmnet' R-package (Friedman et al., 2009).

Model Performance

The performance of the developed statistical models is tested using, adjusted R^2, root mean square error (RMSE), normalised RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE) were calculated using the following formula:

$$RMSE = [1/n \sum_{(i=1)}^{n} (Y_i - \hat{Y}_i)^2]^{1/2} \qquad (5)$$

$$nRMSE = [1/n \sum_{(i=1)}^{n} (Y_i - \hat{Y}_i)^2]^{1/2} \times 100/(mean(y_i)) \qquad (6)$$

$$MAE = 1/n \sum_{(i=1)}^{n} |Y_i - \hat{Y}_i| \qquad (7)$$

$$MAPE = 100/n \sum_{(i=1)}^{n} |(Y_i - \hat{Y}_i) Y_i| \qquad (8)$$

y_i = actual value

$\hat{Y}_i$ = Model output

R^2_{adj} towards 1 and RMSE towards 0 indicates better performance of the developed models. Also lesser the MAE and MAPE values, the better fit the model is. According to nRMSE, the model performance is judged as excellent, good, fair and poor when the values are in the range of <10%, 10–20%, 20–30% and >30%, respectively (Jamieson et al., 1991).

RESULTS AND DISCUSSION

Summary Statistics Of Yield Data

The summary statistics of groundnut yield data (1991-2019) of the Coimbatore district of Tamil Nadu is presented in Table 2. The maximum yield was 2877 kgha^{-1}, whereas the minimum yield was 1519 kg ha^{-1}. The coefficient of variation of yield was found to be 17.84%. A normal Q-Q plot was constructed for testing the normality of yield data and it was affirming the normality, thus satisfying the basic assumptions of parametric models (Fig.1). Figure.2. Shows the Pearson's coefficient of correlation between all variables. Significant positive correlations (correlation coefficient greater than 0.5) were found between yield and Z11, Z30, Z31,

Table1. Weather indices used in the development of multivariate yield forecasting regression models

Parameter	Unweighed Indices	Weighed Indices
Tmax	Z10	Z11
Tmin	Z20	Z21
Rain	Z30	Z31
RH I	Z40	Z41
RH II	Z50	Z51

Z51 (P<0.05). The yield was found to be strongly correlated with those variables (p<0.01) and the correlation coefficients ranged between 0.50 and 0.59 (Iqbal et al., 2019).

Table.2. OLS estimates of regression coefficients in MLR and VIF

Variable	Coefficients	VIF
Intercept	5249.95 ± 5597.78	0
Z10	-14.37 ± 11.0972	12.01887
Z11	28.88 ± 38.93	7.859344
Z20	6.62 ± 7.73	3.753992
Z21	-21.24 ± 41.89	6.089400
Z30	-23.48 ± 16.18	474.9300
Z31	57.18 ± 36.39	507.5759
Z40	-0.39 ± 0.91	6.801643
Z41	4.58 ± 5.49	5.729535
Z50	-1.01 ± 0.86	4.468881
Z51	13.30 ± 3.41	1.651192
T	9.71 ± 8.84	3.438897
Adj R^2	0.8215	
RMSE	184.09	

Groundnut Yield Forecasting Models

In Multiple Linear Regression (MLR), regression coefficients along with their standard errors and VIF values are shown in Table.3. All the predictors included in the model explained 82.15% of the variation. The VIF values of more than 5 observed for most of the variables may be considered a cause of concern, whereas a value of more than 10 indicates severe multicollinearity (Sheater, 2009; Kutner et al., 2004). As the data needed further examination, we opted for alternative approaches to fix this problem of multicollinearity.

Figure.1. Normal Q-Q plot for groundnut yield of Coimbatore district.
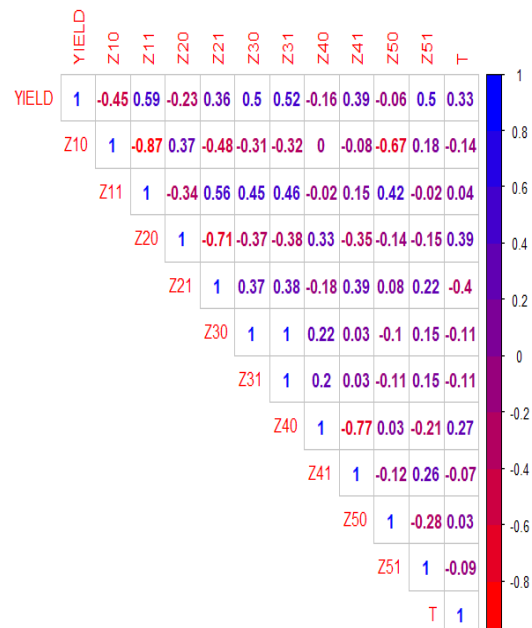
This study utilized Ridge, LASSO, ENet regression and SMLR as alternative methods to MLR due to the presence of multicollinearity. In the regularization techniques such as ridge, LASSO and ENet cross-validation is done for selecting the optimal lambda (λ_{min}) values (Fig.3). The results of applying these methods for groundnut yield prediction are shown in Table.4.

Table.3. Descriptive statistics of variables

	Tmax (^{0}C)	Tmin (^{0}C)	RH-I (%)	RH-II (%)	Rainfall (mm year^{-1})	Yield (kg ha^{-1})
Mean	33.13	23.69	86.09	41.24	1194	2045
Maximum	38.10	26.20	92.43	59.71	1303	2877
Minimum	25.83	19.53	78.86	21.86	1097	1519
Standard deviation	1.44	1.34	2.88	7.59	46.15	365
CV (%)	4.34	5.65	3.35	18.40	3.86	17.84

Stepwise Multiple Linear Regression (SMLR)

The developed SMLR model explains about 83.76% variation in the yield due to weather parameters. The most critical parameters identified using SMLR were maximum temperature followed by morning relative humidity, minimum temperature and rainfall. This model is considered excellent according to nRMSE value (Jamieson et al., 1991)

Ridge Regression

The ridge regression contains all the predictors, whereas other methods consisted of a reduced number of predictors thus reducing the model complexity. Ridge regression explains 87.85% of variation in yield due to all the predictors. The most influential parameter was found to be maximum temperature followed by rainfall. The developed ridge model is considered good in accordance with nRMSE value.

LASSO (Least Absolute Shrinkage and Selection Operator)

In LASSO, feature selection is made along with regularization of parameters, thus preventing the model from overfitting. The developed model explains about 87.46% of the variation in yield. The most influential parameter was found to be maximum temperature followed by rainfall. The developed model is considered excellent in agreement with nRMSE value.

Elastic Net Regression

In ENet method, which is a combination of both ridge and LASSO, the maximum temperature was found to be the most critical parameter followed by rainfall. The developed model explains about 87.48%

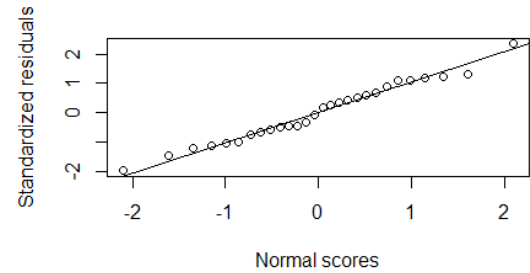of the variation in yield. The nRMSE value depicted that the model performance was excellent.

Figure.2. Pearson correlation coefficient between groundnut yield and weather indices of Coimbatore district.

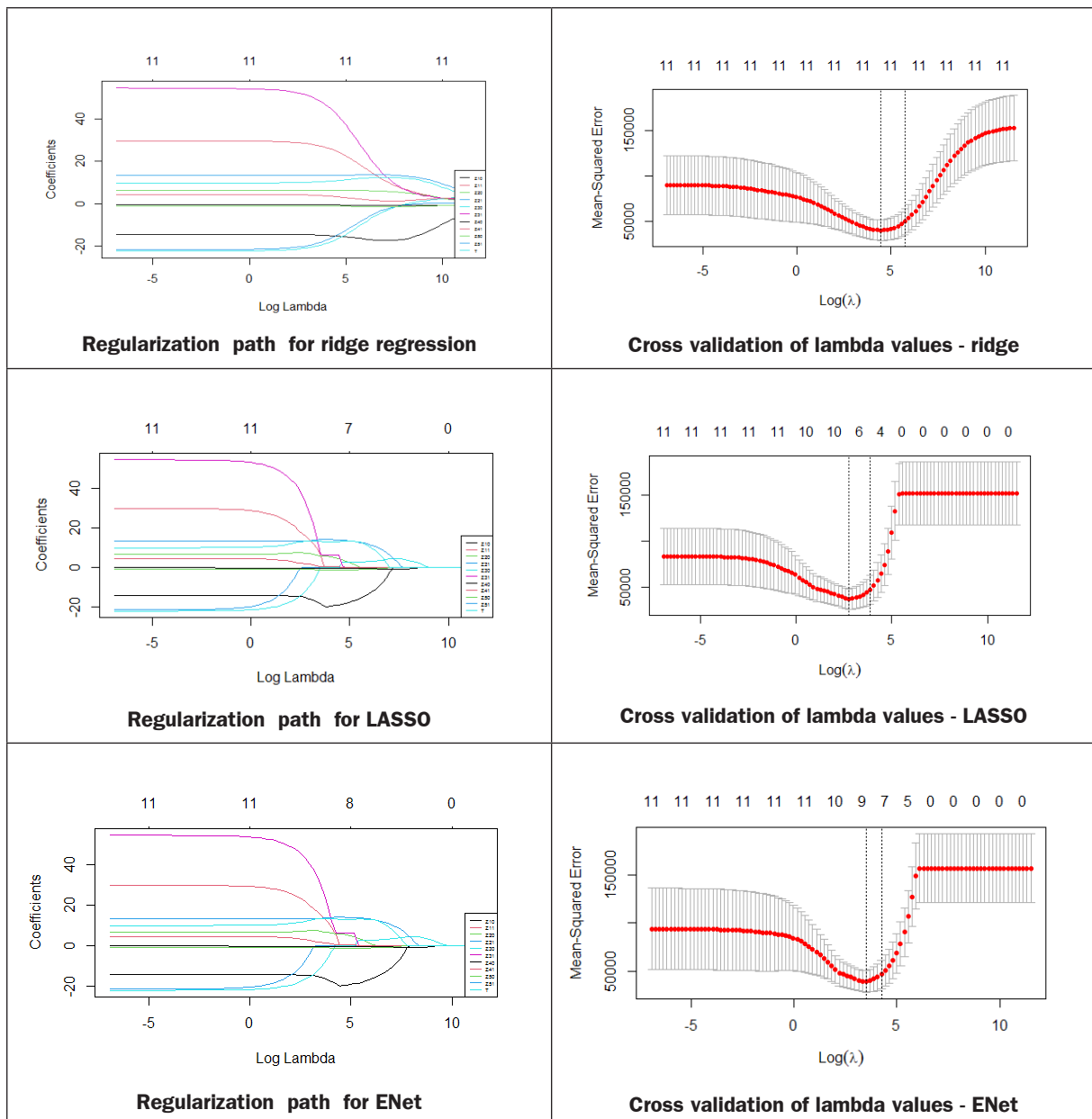For comparing the performance of SMLR and penalized regression techniques, we used goodness-

Figure.3. Regularization paths and cross validation of lambda values for ridge, LASSO and ENet

of-fit measures i.e., R^2_{adj}, RMSE, MAE, MAPE. The adjusted coefficient of determination (R^2_{adj}) was significant for all the models included in the study.

When considering the R^2_{adj}, ridge regression was found to have a better fit compared to other models. The RMSE value of 114.40 was found the least for

Table.4 . Goodness of fit measures obtained for Groundnut yield prediction of Coimbatore district using different statistical models.

MODEL	Adj R^2	RMSEC	RMSEV	MAE	MAPE	nRMSEV	F statistic
MLR	0.8215	117.78	184.09	167.17	0.08	9.64	10.62
SMLR	0.8376 (2)	114.34	178.01 (1)	158.99 (3)	0.08 (1)	9.32	14.18
Ridge	0.8785 (1)	136.51	210.08 (4)	166.93 (4)	1.84 (4)	11.00	5.67
LASSO	0.8746 (4)	136.01	188.09 (3)	158.54 (2)	1.76 (3)	9.85	10.39
ELNET	0.8748 (3)	138.63	181.99 (2)	149.61 (1)	1.65 (2)	9.53	7.77

**Values in parenthesis refers to the rank of the measures.

SMLR followed by MLR and other models. ENet method was found to have a minimum MAE of 149.62 followed by LASSO and SMLR. In the view of MAPE, SMLR is found to have the least value of 0.082 followed by MLR and other models. However, the best prediction is not necessarily provided by a model that fits the data well. Hence, we evaluated the predictive performance of the models using a validation data set and found that the RMSE value of 178.06 was found to be the least for SMLR, followed by ENet and other models. In accordance with the nRMSE value, all models except ridge were considered excellent (<10%), whereas ridge was considered good (10-20%).

CONCLUSION

In the present investigation, five different multivariate models were compared to predict the groundnut yield of Coimbatore district during Kharif season. The classical Multiple Linear Regression (MLR) model shows the presence of multicollinearity. So, the performance of other models such as SMLR, ridge, LASSO and ENet were ranked based on the different goodness-of-fit measures (Table.4). SMLR and ENet provide better data fitting, revealing that these models can be used for groundnut yield forecasting for the studied region.

ACKNOWLEDGMENT

ETHICS STATEMENT

No specific permits were required for the described field studies because no human or animal subjects were involved in this research.

CONSENT FOR PUBLICATION

All the authors agreed to publish the content.

COMPETING INTERESTS

There were no conflict of interest in the publication of this content

REFERENCES

Agrawal, R., Jain, R. C., Jha, M. P., and Singh, D. 1980. Forecasting of wheat yield using climatic variables. *Ind. J. Agric. Sci.*, **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.

Balabin, R. M., Lomakina, E. I., and Safieva, R. Z. 2011. Neural network (ANN) approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. *Fuel.*, **90(5):** 2007-2015.

Bocca, F. F., and Rodrigues, L. H. A. 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and electronics in agriculture.*, **128:** 67-76.

Cho, S., Kim, H., Oh, S., Kim, K., and Park, T. 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. In *BMC proceedings* .3(**7**) 1-6.

BioMed Central.,Das, B., Sahoo, R. N., Pargal, S., Krishna, G., Verma, R., Chinnusamy, V., ... and Gupta, V. K. 2017. Comparison of different uni-and multi-variate techniques for monitoring leaf water status as an indicator of water-deficit stress in wheat through spectroscopy. *Biosystems Engineering.*, **160:** 69-83.

Dhekale, B. S., Sawant, P., and Upadhye, T. 2014. Weather based pre-harvest forecasting of rice at Kolhapur (Maharashtra). *Trends Biosci.*, **7:** 39-41.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ...and Lautenbach, S. 2013.Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography.*, **36(1):** 27-46.

Eyduran, E., A. Waheed, M.M. Tariq, F. Iqbal, and S. Ahmed 2013. Prediction of live body weight from morphological characteristics of commercial goat

in Pakistan using factor and principal component scores in multiple linear regression. *J. Anim. Plan. Sci.*, **23(6):** 1532-1540.

Friedman, J., Hastie, T., and Tibshirani, R. 2009. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version.*, **1(4):** 1-24.

Hastie, T., and Qian, J.2014. Glmnet vignette.,**9:** 1-30.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics.*, **12(1):** 55-67.

Iqbal, F., Ali, M., Huma, Z. E., and Raziq, A.2019. Predicting live body weight of harnai sheep through penalized regression models. *J. Anim. Plant Sci.*, **29(6):** 1541-1548.

Jamieson, P. D., Porter, J. R., and Wilson, D. R. 1991. A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. *Field crops research.*, **27(4):** 337-350.

Kumar, H. V., Shivamurthy, M., and Lunagaria, M. M. 2017. Impact of rainfall variability and trend on rice yield in coastal Karnataka. *Journal of Agrometeorolog.y*, **19(3):** 286-287.

Kumar, N., Pisal, R. R., Shukla, S. P., and Pandey, K. K. 2014. Crop yield forecasting of paddy, sugarcane and wheat through linear regression technique for south Gujarat. *Mausam.*, **65(3):** 361-364.

Kutner, M., Nachtsheim, C., and Neter, J. 2004. Applied Linear Regression Models. Richard D. Irwin. *Inc.: Homewood, Illinois*.

Lobell, D. B., and Burke, M. B. 2010. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and forest meteorology.*, **150(11):** 1443-1452.

Piaskowski, J. L., Brown, D., and Campbell, K. G. 2016. Near infrared calibration of soluble stem carbohydrates for predicting drought tolerance in spring wheat. *Agronomy Journal.*, **108(1):** 285-293.

Rai, Y. K., Ale, B. B., and Alam, J. 2011. Impact assessment of climate change on paddy yield: a case study of Nepal Agriculture Research Council (NARC), Tarahara, Nepal. *Journal of the Institute of Engineering.*, **8(3):** 147-167.

Sangun, L., Cankaya, S., Kayaalp, G. T., and Akar, M. 2009. Use of factor analysis scores in multiple regression model for estimation of body weight from some body measurements in Lizardfish. *Journal of Animal and Veterinary Advances.*, **8(1):** 47-50.

Sheather, S. (2009). *A modern approach to regression with R*. Springer Science and Business Media.

Shi, W., Tao, F., and Zhang, Z. 2013. A review on statistical models for identifying climate contributions to crop yields. *Journal of Geographical Sciences.*, **23(3):** 567-576.

Singh, R. S., Patel, C., Yadav, M. K., and Singh, K. K. 2014. Yield forecasting of rice and wheat crops for eastern Uttar Pradesh. *Journal of Agrometeorology.*, **16(2):** 199.

Smith, G. 2018. Step away from stepwise. *Journal of Big Data.*, **5(1):** 1-12.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological.)*, **58(1):** 267-288.

Trnka, M., Hlavinka, P., Semerádová, D., Dubrovsky, M., Zalud, Z., and Mozny, M. 2007. Agricultural drought and spring barley yields in the Czech Republic. *Plant Soil and Environment.*, **53(7):** 306.

Verma, U., Piepho, H. P., Goyal, A., Ogutu, J. O., and Kalubarme, M. H.2016. Role of climatic variables and crop condition term for mustard yield prediction in Haryana. *Int. J .Agric. Stat. Sci.*, **12:** 45-51.

Wu, Z., Huang, N. E., Long, S. R., and Peng, C. K. 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences.*, **104(38):** 14889-14894.

Yakubu, A. 2010. Fixing multicollinearity instability in the prediction of body weight from morphometric traits of White Fulani cows. *Journal of Central European Agriculture*.

Zou, H., and Hastie, T.2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology).*, **67(2):** 301-320.