Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 25 (3) : 432-439 (September - 2023)
https://doi.org/10.54386/jam.v25i3.2185
https://journal.agrimetassociation.org/index.php/jam

Research paper

Comparative evaluation of penalized regression models with multiple linear regression for predicting rapeseed-mustard yield: Weather-indices based approach

AJITH S[1*], MANOJ KANTI DEBNATH[1], DEB SANKAR GUPTA[1], PRADIP BASAK[1], SUBHENDU BANDYOPADHYAY[2], SHYAMAL KHEROAR[3] and RAGINI HR[1]

[1]*Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya, West Bengal, India.*
[2]*Department of Agronomy, Uttar Banga Krishi Viswavidyalaya, West Bengal, India.*
[3]*All India Network Project on Jute and Allied Fibres, Uttar Banga Krishi Viswavidyalaya, West Bengal, India*
*Corresponding author e-mail id: ajithagristat@gmail.com

ABSTRACT

Rapeseed-mustard (*Brassica spp*.) is one of the important edible oilseeds crops in India. The same level of weather condition impacts the growth and establishment of rapeseed-mustard plant differently in different stages of crop which lead to large intra-seasonal yield variations. Hence it is essential to give weightage to weekly weather conditions while fitting predictive model. In this present study, path-coefficient based weighted index was proposed along with existing unweighted and correlation based weighted index. The performance of penalized regression models *viz.* Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENET) were compared with Multiple Linear Regression (MLR) for predicting rapeseed-mustard yield using weather-indices. The results revealed that the path-coefficient based weighting of weather parameters to the yield were stable than correlation based weighted-indices. Path-coefficient based weighted indices of maximum temperature, minimum temperature and windspeed were important variables in projection of yield. The performance of MLR was poor during validation of model due to overfitting issue. The performance of penalized models was stable in both calibration and validation of the model. The LASSO and ENET models that accompanied with coefficient shrinkage and variable selection were found to be the best fitted models for predicting Rapeseed-Mustard yield.

Keywords: Rapeseed-Mustard, Weather Indices, Multiple Linear Regression, Ridge Regression, LASSO, Elastic Net.

Rapeseed-mustard (*Brassica spp*.) is one of the important edible oilseeds crop in India. The Rapeseed-mustard crop is grown in 6.86 million hectares of land area in India that produce 9.12 million tonnes of grains with an average productivity of 1331 kg ha^{-1}. The crop is grown in diverse agroclimatic conditions ranging from north-western to north-eastern hills. Rapeseed-mustard is grown as rabi crop in sub-tropical regions of West Bengal. AICRP on Rapeseed & Mustard has reported that there was a negative growth in the yield of rapeseed-mustard crop in country level as well as in major growing states in the last ten years. But West Bengal has registered 12.00% growth in the cultivation area with very high growth in productivity (32.70%) that resulted in 41.10% increase in the production.

The same level of weather condition impacts the growth and establishment of the plant differently in different stages of crop. Heat stress during flowering to maturity stage of the rapeseed-mustard crop, accelerates the growth and development of the plant that leads to significant reduction in the yield due to premature end of reproductive stage (Chugh and Sharma, 2022). The movement of photosynthates to the sinks is adversely affected by heat stress during post anthesis to seed filling stages of the crop that lower the seed weight and seed yield (Kumar *et al.,* 2017). For potential productivity of mustard crop, the maximum temperature ranges between 20.0-27.3°C and minimum temperature ranging 4.6-11.8°C required during flowering stage and it is 11.3-20.4°C and 3.8-7.2°C during pod formation stage respectively (Kaur and Gill, 2017). Hence it is necessary to give weightage to weekly weather

Article info - DOI: *https://doi.org/10.54386/jam.v25i3.2185*
Received: 06 April 2023; Accepted: 30 June 2023; Published online : 31 August 2023

conditions while fitting the predictive model to forecast the crop yield. In order to give weightage to the respective week's weather conditions, correlation coefficient based weighted indices can be used (Jain *et al.,* 1980; Gupta *et al.,* 2018).

The statistical models are often used to study the complex association between agricultural systems and climatic parameters. The linear regression is a standard statistical model to study cause and effect relationship between a dependent variable and one or more independent variables. The penalized regression models such as, Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENET) are widely applied as an alternative to traditional multiple linear regression to predict the crop yield (Setiya *et al.,* 2022). With this view, the present study has been formulated to find a best statistical model for predicting rapeseed-mustard yield in four Northern districts of West Bengal using unweighted and weighted weather indices.

MATERIALS AND METHODS

The yield data of rapeseed-mustard crop of four selected districts of West Bengal namely, Cooch Behar, Jalpaiguri, Malda and Uttar Dinajpur from 1997-98 to 2020-21 were collected from Directorate of Economics and Statistics, Ministry of Agriculture & Farmers Welfare, Govt. of India. Weekly weather data of these districts were collected from Regional Meteorological Centre, Kolkata. The weather parameters considered for the study were maximum temperature (°C), minimum temperature (°C), relative humidity (%), rainfall (mm) and windspeed (m/s). Weekly weather data for the period from 47th Standard Meteorological Week (SMW) of a year to 11th Standard Meteorological Week (SMW) of next year in which Rapeseed-mustard crop is grown were used to develop yearly weather indices.

Developing of different weather indices

The weekly weather data were converted to yearly weather indices. Three types of weather indices were developed. The unweighted indices are the simple average of weekly weather variables of the weeks in which the particular crop was grown. Two types of weighted indices were calculated as given below,

Correlation coefficient based weighted weather indices: The correlation based weighted indices is the weighted average of weather variables where the weight is correlation coefficient between weather variables in respective week and crop yield. The correlation coefficient based weighted indices are calculated as follows,

$$\mathbf{C_{ij}} = \frac{\sum_{k=1}^{m} \mathbf{r_{jk}.X_{ijk}}}{\sum_{k=1}^{m} \mathbf{r_{jk}}} \tag{1}$$

Where, C_{ij} is weighted weather index for jth weather parameter in ith year with correlation coefficient as weight, r_{jk} is the correlation coefficient between detrended crop yield and jth weather parameter at kth week, X_{ijk} is the observation of jth weather parameter in kth week of ith year and m is the number of weeks in which the crop is grown.

As the detrended yield represents only the actual effect of weather factors, the yield was adjusted for trend effect while calculating correlation coefficient (Agrawal *et al.,* 1986). The detrended yield remove the effect of technological development over the years such as varietal improvement, advancement in agronomic practices *etc*. (Huzsvai *et al.,* 2022).

Path coefficient based weighted weather indices: The correlation between a response variable and an independent variable is the sum of direct and indirect effect of that particular independent variable on the response (Wright, 1921). The path coefficient analysis the correlation coefficient into direct and indirect effects. The path coefficient accounts for only the direct effect of an independent variable on the ultimate response variable (Alwin and Hauser, 1975). Hence the path coefficient based weighted indices were proposed as weighted average of weather variables where weight is the path coefficient.

$$\mathbf{P_{ij}} = \frac{\sum_{k=1}^{m} \mathbf{p_{jk}.X_{ijk}}}{\sum_{k=1}^{m} \mathbf{p_{jk}}} \tag{2}$$

Where P_{ij} is weighted weather index for jth weather parameter in ith year with path coefficient being the weight and p_{jk} is the path coefficient between detrended crop yield and jth weather parameter at kth week.

In total, fifteen weather indices that consist of five unweighted, five correlation coefficient based and five path coefficient based indices were calculated from five weather variables. The description of weighted and unweighted indices pertaining to each weather variable is given in the Table 1.

Multiple linear regression (MLR)

Linear regression model is being used to study the functional relationship between a response variable and one or more than one explanatory variables. The function relationship can be represented as,

$$\mathbf{Y = A + \sum \boldsymbol{\beta}_i . X_i + \boldsymbol{\varepsilon}} \tag{3}$$

Where, Y is the response variable, is ith explanatory variable, A is intercept, βi is regression coefficient corresponds to ith variable and ε is error term or unexplained part of the model. The regression coefficients were estimated using Ordinary Least Square (OLS) estimation procedure.

Penalized regression models

The regression model fitted with OLS estimator often causes overfitting issue. An overfitted model describes the data very well. But its performance for the new data is generally poor (Zhang, 2014). The penalized regression models impose penalty for each independent variables added to the model by including a penalty term along with Mean Square Error (MSE) function while estimating the parameters. Due to the added penalty, the parameter values are shrunken. Thus, overall slope of the fitted model tends to decrease, thereby overfitting of the model can be avoided. Hence the performance of three penalized regression models *viz.* ridge regression (RR), least absolute shrinkage and selection operator (LASSO) and elastic net (ENET) were compared with multiple

Table 1: Description of weighted and unweighted indices

S. No.	Indices	Description
1	TMAX	Unweighted indices of maximum temperature
2	TMIN	Unweighted indices of minimum temperature
3	RH	Unweighted indices of relative humidity
4	Windspeed	Unweighted indices of windspeed
5	RF	Unweighted indices of rainfall
6	CC_TMAX	Correlation coefficient based weighted indices of maximum temperature
7	CC_TMIN	Correlation coefficient based weighted indices of minimum temperature
8	CC_RH	Correlation coefficient based weighted indices of relative humidity
9	CC_Windspeed	Correlation coefficient based weighted indices of windspeed
10	CC_RF	Correlation coefficient based weighted indices of rainfall
11	PC_TMAX	Path coefficient based weighted indices of maximum temperature
12	PC_TMIN	Path coefficient based weighted indices of minimum temperature
13	PC_RH	Path coefficient based weighted indices of relative humidity
14	PC_Windspeed	Path coefficient based weighted indices of windspeed
15	PC_RF	Path coefficient based weighted indices of rainfall

linear regression (MLR).

Ridge regression (RR): Given a linear regression with predictors X_{ij} and response Y_i, the Ridge regression solves the L_2 penalized regression problem to estimate β={βi} by minimizing

$$\sum_{i=1}^{n}\left(y_i - \sum x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left(\beta_j\right)^2 \tag{4}$$

Where, λ is turning parameter which ranges between 0 and ∞. The λ=0 gives the same result as OLS method. As the λ increases, the amount of the penalty added to the model increases. Due to which the coefficient values shrunken toward zero but it never reaches zero.

Least absolute shrinkage and selection operator (LASSO): LASSO solves the L_1 penalized regression problem to estimate β={βi} by minimizing

$$\sum_{i=1}^{n}\left(y_i - \sum x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \tag{5}$$

Because of imposition of L_1 penalty in LASSO, coefficients correspond to some of the variables are shrunken to zero, thereby the variable selection can be achieved. Hence, the LASSO does variable selection and shrinkage, whereas ridge regression only shrinks the coefficients.

Elastic net (ENET): Elastic net (ENET) is a combination of Ridge and LASSO regression. ENET imposes both L_1 and L_2 penalty together. The β={βi} is estimated by minimizing

$$\sum_{i=1}^{N}\left(y_i - \sum x_{ij}\beta_j\right)^2 + \lambda\left(\alpha \sum_{j=1}^{p}\left|\beta_j\right| + (1-\alpha)\sum_{j=1}^{p}\left(\beta_j\right)^2\right) \tag{6}$$

Two penalty parameters Alpha (α) and Lambda (λ) have to be optimized in ENET. The alpha value ranges between 0 and 1. As like LASSO, the ENET also does both variable selection and shrinkage. ENET is advantages if the independent variables are highly correlated.

The parameters Lambda (λ) and Alpha (α) of the penalized regression models were optimized using 10-fold cross validation procedure.

Model evaluation criteria

The performance of the models was examined using the following model evaluation criteria.

Where, Yi and Ŷi are the observed and predicted yield correspond to i^th year and Ȳi is average yield.

Coefficient of Determination: $R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

Mean Absolute Error: $MAE = \frac{\sum_{i=1}^{n}|(Y_i - \hat{Y}_i)|}{n}$

Root Mean Square Error: $RMSE = \sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / n}$

normalized Root Mean Square Error: $nRMSE = \frac{RMSE}{\bar{Y}}$

Eighty percent of the data were randomly chosen for calibration of the model and remaining twenty percent of the data were utilized for validation of the fitted model. The random splitting was done in order to ensure the representation of both recent as well as past year data in both training and testing datasets. The data pertaining to the years 1997-98, 2001-02, 2002-03, 2011-12 and 2017-18 were in testing dataset and remaining data were in training dataset. The model that performs better in both calibration and validation stages were considered as the best fitted model. The best fitted model for each district were selected. The rapeseed-mustard yield was forecasted for the year 2020-21 for each district using the respective best fitted model. The forecasted yield was compared with actual yield.

Table 2: Summary of different weather indices

Indices	Cooch Behar		Jalpaiguri		Malda		Uttar Dinajpur	
	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
TMAX (°c)	26.51	0.94	26.21	0.94	26.86	1.18	27.17	1.32
TMIN (°c)	12.46	1.11	13.30	0.68	15.80	0.80	14.66	1.59
RH (%)	75.45	3.24	76.22	3.01	71.15	4.19	80.10	6.43
Windspeed (m/s)	1.82	0.08	1.94	0.13	2.55	0.39	2.33	0.44
RF (mm)	45.05	48.58	32.29	35.03	67.04	70.82	15.26	16.36
CC_TMAX	22.63	2.79	25.36	1.74	22.36	2.60	26.25	1.82
CC_TMIN	11.85	1.31	12.74	2.50	13.72	1.58	15.56	1.88
CC_RH	37.43	19.63	87.67	6.28	217.66	282.76	83.05	5.88
CC_Windspeed	1.51	0.24	1.84	0.18	2.58	0.51	2.36	0.50
CC_RF	10.57	60.41	14.12	35.24	-1.44	29.05	1.32	4.42
PC_TMAX	25.49	1.58	26.34	1.10	23.64	2.25	27.60	1.31
PC_TMIN	11.94	1.33	13.74	1.07	14.00	1.34	12.40	1.64
PC_RH	69.26	5.66	7.24	33.20	73.97	10.60	80.29	6.03
PC_Windspeed	1.73	0.14	2.14	0.45	2.66	0.75	2.44	0.51
PC_RF	1.60	9.28	2.34	4.60	3.11	10.45	-0.41	13.53

RESULTS AND DISCUSSION

Developing weather indices

The weather parameters from 47th SMW of a year to 11th SMW of next year were used develop the weather indices. The summary of unweighted and weighted indices was given in the Table 2. The average maximum temperature during the crop growing period was around 26°C-27°C in the four districts. The minimum temperature was low in Cooch Behar (12.46°C) followed by Jalpaiguri (13.30°C) and Malda (14.66°C). The highest relative humidity was observed in Uttar Dinajpur (80.10%) followed by Jalpaiguri (76.22%) and Cooch Behar (75.45%) and it was comparatively low level of 71.15% in Malda. The average windspeed was greater than 2 m/s in Malda and Uttar Dinajpur and it was less than 2 m/s in Cooch Behar and Jalpaiguri districts. As the crop is grown in winter season, average cumulative rainfall was less than 70 mm during the crop growing period. It can be observed that standard deviation of all the weather parameters were low in all districts except rainfall which indicates the occurrence of rainfall was not stable during the crop period.

The average of both correlation and path coefficient based weighted indices were lower than the unweighted indices over all weather parameters in all four districts which indicates weighted indices lowering the indices value according to the relationship between crop yield and weekly weather condition. It can be observed that standard deviation of path-coefficient indices was less than correlation coefficient based weighted indices which indicates path coefficient-based indices were stable than correlation-based indices. The unstable unweighted indices of rainfall were also stabilized by path-coefficient based weighted-indices.

Optimizing penalty parameters for penalized regression models

The penalty parameters namely Alpha (α) and Lambda (λ) have to be optimized while fitting the penalized regression modes. The alpha value for Ridge and LASSO regression was fixed were 0 and 1 respectively. Hence only lambda (λ) value has to be optimized for Ridge and LASSO models. For ENET model, both alpha (α) and lambda (λ) have to be optimized. The penalty parameters were optimized using 10-fold cross validation. The optimum values of penalty parameters for penalized regression models were given for each district in the Table 3.

Table 3: Optimum penalty parameters for penalized regression models for each district

District	Ridge	LASSO	Elastic Net	
	Lambda	Lambda	Alpha	Lambda
Cooch Behar	299	25	0.3	66
Jalpaiguri	70	12	0.3	20
Malda	62	12	0.8	14
Uttar Dinajpur	130	27	0.6	38

The optimum lambda values of Ridge regression were more than LASSO model. Thus, Ridge regression exerts more penalty to the model than LASSO. The better combination of alpha and lambda values of ENET model was optimized. The alpha values for Cooch Behar and Jalpaiguri districts were less than 0.5. Hence the ENET models for these districts were of Ridge type. The ENET models for Malda and Uttar Dinajpur were of LASSO type as their alpha values were near one. Further, the lambda values of ENET models were ranged between Ridge and LASSO models. Thus, ENET imposed moderate penalty than Ridge and LASSO models.

Model fitting

The MLR, Ridge, LASSO and ENET models were fitted for each district by taking Rapeseed-Mustard yield as response or dependent variable and the fifteen weather indices of the respective district as independent variables. The year number also included as a time trend variable. The penalized regression models were fitted using their respective optimum penalty parameter values. The estimate of coefficients of the fitted models for each district were given in Table 4. In comparison with MLR, regression coefficients of each weather indices were shrunken in Ridge model as it imposes penalty to the model. But the coefficients values were

Table 4: The coefficients of the fitted models for each district

Indices	Cooch Behar				Jalpaiguri				Malda				Uttar Dinajpur			
	MLR	Ridge	LAS-SO	ENET	MLR	Ridge	LASSO	ENET	MLR	Ridge	LASSO	ENET	MLR	Ridge	LASSO	ENET
(Intercept)	2710.20	22.00	-216.97	-287.15	1634.40	486.37	438.11	409.69	1712.10	1004.34	868.48	896.9	12944.84	-1846.93	-1000.99	-1187.82
Year Number	19.31	0.50	-	-	20.38	6.53	10.28	9.23	24.24	11.66	15.57	17.21	134.99	6.84	16.53	15.12
TMAX	22.79	3.12	-	-	-213.91	13.29	-	10.56	19.96	-3.19	-	-	291.31	14.5	-	-
TMIN	-217.13	-2.20	-	-	-58.14	-33.21	-10.63	-29.99	-31.14	-1.98	-	-	-65.99	5.38	-	-
RH	-71.42	-0.97	-	-	-56.74	2.68	-	-	-0.03	1.54	-	-	-71.18	3.61	1.41	5.17
Windspeed	-415.22	-21.58	-	-	-505.39	25.88	-	-	-151.76	16.81	-	-	927.95	94.01	-	-
RF	-2.31	-0.01	-	-	0.59	-0.39	-	-	2.69	-0.39	-	-	-23.56	-0.82	-	-
CC_TMAX	75.41	5.00	8.56	6.46	30.58	5.57	-	3.63	-19.00	3.4	-	-	-118.28	15.88	18.57	23.71
CC_TMIN	706.91	-5.10	-	-	-27.17	-4.65	-0.11	-5.55	3.88	-1.68	-	-	-56.49	13.76	2.16	3.91
CC_RH	-11.03	1.03	4.42	1.93	30.37	-1.56	-	-0.15	0.10	0.12	0.14	0.14	-13.56	6.66	9.63	10.27
CC_Windspeed	1811.70	61.14	137.69	99.79	589.54	-40.8	-	-0.98	-329.36	-48.35	-47.12	-47.87	606.34	-19.71	-	-
CC_RF	-10.08	0.14	-	-	0.81	-0.56	-0.73	-0.69	12.54	1.61	2.2	2.21	-153.46	-2.63	-	-
PC_TMAX	-6.54	9.03	13.3	12.94	245.19	10.78	26.54	21.06	29.04	5.68	-	-	-259.54	-2.83	-	-
PC_TMIN	-529.94	-5.50	-	-0.04	-19.43	-14.19	-33.26	-16.51	-80.51	-27	-10.83	-12.45	-241.14	21.48	13.49	19.57
PC_RH	35.20	1.22	-	-	5.28	0.56	0.52	0.62	5.87	3.73	3.73	3.81	4.21	2.87	-	-
PC_Windspeed	-1376.56	79.88	-	74.34	-23.32	-26.64	-19.24	-23.38	215.18	-14.71	-	-	-1300.38	17.15	-	-
PC_RF	41.96	0.71	-	-	8.68	-0.99	-	-	-24.95	1.51	-	-	16.33	1.12	-	-
Sum of absolute values of coefficients	**8063.71**	**219.13**	**380.94**	**482.65**	**3469.92**	**674.65**	**539.42**	**532.04**	**2662.35**	**1147.7**	**948.07**	**980.59**	**17229.55**	**2076.18**	**1062.78**	**1265.57**

Table 5: Comparison of fitted models for each district

District	Model	During Calibration				During Validation		
		R^2	MAE	RMSE	nRMSE	MAE	RMSE	nRMSE
Cooch Behar	MLR	0.98	14.98	21.85	0.04	232.02	326.05	0.60
	Ridge	0.68	60.42	79.37	0.15	91.87	101.91	0.19
	LASSO	0.75	54.80	70.86	0.13	83.11	96.35	0.18
	ENET	0.73	55.87	73.23	0.14	85.49	95.52	0.18
Jalpaiguri	MLR	0.99	5.26	6.59	0.01	290.94	360.77	0.56
	Ridge	0.91	29.16	39.31	0.06	76.38	86.51	0.13
	LASSO	0.92	26.10	36.70	0.06	60.63	67.19	0.10
	ENET	0.94	23.80	32.88	0.05	67.57	75.56	0.12
Malda	MLR	0.99	4.17	5.51	0.01	168.19	191.71	0.18
	Ridge	0.92	44.01	54.54	0.05	39.72	49.75	0.05
	LASSO	0.94	39.68	49.79	0.05	45.12	50.96	0.05
	ENET	0.94	39.78	49.59	0.05	46.49	51.77	0.05
Uttar Dinajpur	MLR	0.99	19.20	22.64	0.03	855.36	981.30	1.20
	Ridge	0.83	61.28	88.20	0.11	128.77	178.37	0.22
	LASSO	0.82	67.10	91.87	0.12	116.28	134.74	0.16
	ENET	0.82	65.43	90.78	0.11	109.50	136.95	0.17

Table 6: Categorization of the models based on RMSE

Model	Stage	Cooch Behar	Jalpaiguri	Malda	Uttar Dinajpur
	Calibration	Excellent	Excellent	Excellent	Excellent
MLR	Validation	Poor	Poor	Poor	Poor
	Overall	Poor	Poor	Poor	Poor
	Calibration	Good	Good	Good	Fair
Ridge	Validation	Poor	Poor	Good	Poor
	Overall	Poor	Poor	Good	Poor
	Calibration	Good	Good	Good	Fair
LASSO	Validation	Fair	Fair	Good	Fair
	Overall	Good	Good	Good	Fair
	Calibration	Good	Good	Good	Fair
ENET	Validation	Fair	Fair	Good	Fair
	Overall	Good	Good	Good	Fair

Table 7: Forecasted yield for the year 2020-21

District	Actual (kg ha^{-1})	Forecasted (kg ha^{-1})	Deviation (kg ha^{-1})	Percent Deviation (%)
Cooch Behar	870	859	11	1.28
Jalpaiguri	900	886	14	1.58
Malda	1390	1409	-19	1.35
Uttar Dinajpur	1290	1265	25	1.98

never shrunken to zero. But in LASSO and ENET models some of the coefficients were shrunken to zero, thereby variable selection was accomplished. Only important variables were retained in the optimum LASSO and ENET models.

Cooch Behar: At the optimum lambda level of ridge model, all the sixteen values were retained with non-zero coefficients as given in the Table 4. Thus, ridge model does not do variable reduction. But in LASSO and ENET models, only four and six indices were retained at optimum lambda value. The CC_TMAX, CC_RH, CC_Windspeed and PC_TMAX were retained in the optimal LASSO model. Along with the above indices, PC_TMIN and PC_Windspeed were also retained in ENET model. The sum of absolute value of the coefficients were very low in penalized models than MLR. Due to the imposition of higher penalty, the sum value was comparative lower in Ridge regression.

Jalpaiguri: Eight and twelve indices were retained at the optimum penalty values of LASSO and ENET models as given in the Table 4. All the path coefficient based weighted indices except PC_RF and CC_TMIN, TMIN and time trend variable were retained in the optimal LASSO model. All the weighted indices except PC_RF and TMAX, TMIN and time trend variable were retained in the final ENET model. The sum of absolute coefficient value was comparative lower in ENET followed by LASSO and Ridge regression.

Malda: At the optimum lambda level of LASSO and ENET models, only six indices were retained as given in the Table 4. The CC_RH, CC_Windspeed, CC_RF, PC_TMIN, PC_RH and time trend variable were retained in the optimal LASSO and ENET models. The sum of absolute coefficient value was comparative lower in LASSO followed by ENET and Ridge regression.

Uttar Dinajpur: Only six indices were retained in LASSO and ENET models at the optimum level of lambda. The RH, CC_TMAX, CC_TMIN, CC_RH, PC_TMIN and time trend variable were retained in the final LASSO and ENET models as given in the Table 4. The sum of absolute coefficient value was comparative lower in LASSO followed by ENET and Ridge regression.

Variables importance in projection (VIP)

The variables importance in projection of penalized regression models were calculated for each district. The weighted indices of windspeed were the most important variable in most of the models for all the districts. PC_TMAX was found have importance in all the districts. PC_TMIN was the important variable for prediction in all the districts except Cooch Behar. Time trend variable was found to be the important variable in all the districts except Cooch Behar. In overall, the weighted indices were more important than unweighted indices.

Goodness of fit and model validation

The performance of fitted models were compared in terms of goodness of fit of the models. Further their performance for new data were also validated. The goodness of fit criteria and validation of fitted models for each district were given in the Table 5. The MLR was the best fitted for all the districts with R^2 values close to one and low MAE, RMSE and nRMSE values during calibration stage. But its performance was poor during validation of the model with very huge error values of MAE, RMSE and nRMSE. It clearly indicates the chance of overfitting of the model.

The LASSO model was found to be the best fitted model for Cooch Behar and Malda districts with higher R^2 of 0.75 and 0.94 respectively as well as comparatively lowest MAE, RMSE and nRMSE values of 54.80, 70.86 and 0.13 for Cooch Behar and 39.68, 49.79 and 0.05 for Malda respectively. Its performance was also stable during validation of the model with new data with low error values. Similar results were obtained by Kumar *et al.,* 2019 while selecting the best model for predicting the yield of wheat crop. Elastic Net (ENET) model was the best fitted model for Jalpaiguri and Uttar Dinajpur districts with higher R^2 of 0.94 and 0.82 respectively with lowest MAE, RMSE and nRMSE values in both calibration and validation as given in the Table 5. Similar results were obtained by Das *et al.,* 2020.

The fitted models were ranked based on their performance with respect to RMSE values during calibration, validation as well as overall performance and the same has been given in the Table 6. The performance of MLR was excellent during calibration stage for all the districts. But its performance was poor during validation. For both Cooch Behar and Jalpaiguri districts, there was a good performance by LASSO and ENET models and the same were performed fairly during validation. Hence, both the models were performed well on overall. For Malda district, the performance of all the three penalized models were good in both the stages. There was a fair performance by the LASSO and ENT models with respect to RMSE for Uttar Dinajpur district.

It can be observed that the LASSO and ENET models that were fitted with very few indices performed better than MLR and Ridge Regression models using all the indices. The results were in agreement with the results obtained by Aravind *et al.,* 2022. The penalized regression models provide stable performance in both calibration and validation stages than MLR (Sridhara *et al.,* 2023).

Forecasting performance of fitted models

The rapeseed-mustard yield was forecasted for the year 2020-21 by the respective best fitted model using unweighted and weighted weather indices of the year. The forecasted yield was compared with actual yield. The results of actual and forecasted yield was given in the Table 7. The forecasted yield was close with the actual yield in all the districts. The yield was overestimated in Malda by a meagre quantity and it was underestimated in the remaining three districts by a small quantity. The percent deviation of less than 2% indicates that the models were performing better for forecasting the rapeseed-mustard yield.

CONCLUSIONS

In the present study, an attempt has been made in order to select the best statistical model to predict the yield of Rapeseed-Mustard crop using weather indices. Path coefficient based weighted index was proposed along with the existing correlation based weighted indices. The path coefficient-based weighting of weather parameters to the yield were stable than the correlation based weighted indices. The proposed path coefficient based weighted indices were also having greater importance while fitting the predictive model. The MLR model had highest R^2. But its performance was poor during validation of the model due to overfitting issue. The performance of the penalized regression models was stable in both calibration and validation stages and these models were performing better for forecasting also with less than 2% deviation with actual yield. The LASSO and ENET models that accompanied with coefficient shrinkage and variable selection were found to be the best fitted models for predicting the Rapeseed-Mustard yield.

ACKNOWLEDGEMENT

Authors are thankful to Directorate of Economics and Statistics, Ministry of Agriculture & Farmers Welfare, Govt. of India for providing yield data and Regional Meteorological Centre, Kolkata for providing meteorological data.

Funding: No external fund received for this study.

Data availability statement: Data shall be available from the corresponding author on request.

Conflict of Interests: The authors declare that there is no conflict of interest related to this article.

Authors contributions: **Ajith S:** Designed and Conceptualized the study, Statistical Analysis and original draft preparation; **M.K. Debnath:** Conceptualization, Supervision and Critical Review; **D.S. Gupta & P. Basak:** Supervision and Critical Review; **S. Bandyopadhyay:** Meteorological data supply; **S. Kheroar & HR Ragini:** Critical Review and Shaping of final manuscript.

Disclaimer: The contents, opinions, and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

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

Alwin, D.F., and Hauser, R.M. (1975). The decomposition of effects in path analysis. *Amer. Soc. Rev.*, 37-47.

Aravind, K.S., Vashisth, A., Krishanan, P., and Das, B. (2022). Wheat yield prediction based on weather parameters using multiple linear, neural network and penalised regression models. *J. Agrometeorol.,* 24(1):18-25.https://doi.org/10.54386/jam.v24i1.1002

Chugh, P., and Sharma, P. (2022) Terminal heat stress in Indian mustard (Brassica juncea L.): Variation in dry matter accumulation, stem reserve mobilization, carbohydrates translocation and their correlation with seed yield. *Indian J. Exp. Bio.*, 60:423-431.

Das, B., Nair, B., Arunachalam, V., Reddy, K. V., (2020). Comparative evaluation of linear and nonlinear weather-based models for coconut yield prediction in the west coast of India. *Int. J. Biometeorol.*, 64:1111-1123.

Gupta, S., Singh, A., Kumar, A., Shahi, U. P., Sinha, N. K., and Roy, S. (2018). Yield forecasting of wheat and mustard for western Uttar Pradesh using statistical model. *J. Agrometeorol.,* 20(1):66-68. https://doi.org/10.54386/jam.v20i1.508

Huzsvai, L., Zsembeli, J., Kovács, E., and Juhász, C. (2022). Response of winter wheat (*Triticum aestivum* L.) yield to the increasing weather fluctuations in a continental region of four-season climate. *Agronomy*, 12(2):314.

Jain, R. C., Agrawal, R., and Jha, M. P. (1980). Effect of climatic variables on rice yield and its forecast. *Mausam*, 31(4):591-596.

Kaur, B., and Gill, K.K. (2017). Development of Weather based Weekly Thumb Rules for Potential Productivity of Mustard Crop in Punjab. *Vayu Mandal*, 43(1):72-81.

Kumar, Y., Singh, R., Kumar, A., and Dhaka, A. K. (2017). Effect of growth and yield parameters on Indian-mustard genotypes under varying environmental conditions in western Haryana. *J. Apl. Nat. Sci.*, 9(4):2093-2100.

Kumar, S., Attri, S. D., and Singh, K. K. (2019). Comparison of LASSO and stepwise regression technique for wheat yield prediction. *J. Agrometeorol.,* 21(2):188-192. https://doi.org/10.54386/jam.v21i2.231

Setiya, P., Satpathi, A., Nain, A. S., and Das, B. (2022). Comparison of weather-based wheat yield forecasting models for different districts of Uttarakhand using statistical and machine learning techniques. *J. Agrometeorol.,* 24(3):255-261. https://doi.org/10.54386/jam.v24i3.1571

Sridhara, S., Manoj, K.N., Gopakkali, P. (2023). Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India *Int. J. Biometeorol.*, 67(1): 165-180.

Wright, S. (1921). Correlation and causation. *J. Agric. Res.,* 20:557-585.

Zhang, Z. (2014). Too much covariates in a multivariable model may cause the problem of overfitting. *J. Thorac. Dis.*, 6(9): E196-E197.