International Journal of Agricultural Sciences
Volume 14 | Issue 1 | January, 2018 | 165-172 ■ e ISSN-0976-5670
☞DOI:10.15740/HAS/IJAS/14.1/165-172
Visit us : *www.researchjournal.co.in*

RESEARCH PAPER

A study on cross validation for model selection and estimation

Fehim Jeelani Wani*, S.E.H. Rizvi[1], Manish Kumar Sharma[1] **and** M. Iqbal Jeelani Bhat[1]
Division of Agicultural Statistics (SKUAST-K), SHALIMAR (J&K) INDIA
(Email : faheemwani149@gmail.com)

Abstract : In the present study, *k*-fold cross validation method was examined for performance evaluation of different regression models. A multistage sampling technique was adopted for the selection of samples in which districts, villages within districts and fodder trees in the selected village formed the first stage, second stage and third stage units, respectively. A total number of 10 trees were randomly selected from each village so as to constitute a predetermined total sample size of 60 trees. Primary data on height, bole height, diameter at breast height (dbh), no. of primary branches, secondary branches, average no. of leaves per secondary branch, age, canopy diameter and green fodder yield (dependent variable) for each selected tree were collected through visiting farmers field in the selected area and by adopting standard forest mensuration procedures. Regression analysis was used to study the relationship between fodder yield (dependent variable) and other parameters. Different regression models were tried and on the basis of adj. R^2, the best five models were selected. Goodness of fit of the selected models was tested by applying chi-square test. The chi-square test results came out to be insignificant indicating thereby that the models under study were qualified for goodness of fit and could be used for further study. The models were validated for its adequacy through different criteria, namely, adj. R^2, bias, variance, root mean square error and coefficient of dispersion. On the basis of set criteria, the models were ranked. After applying the Wilcoxon signed rank test on fitting data set, one can arrive at the final ranks by considering ranks of both fitting (R_f) and validating (R_v) data sets. Finally, on the basis of all the criteria adopted in the present investigation, the regression model obtained as $\hat{Y} = 8.480 + 0.000004L^2S$ ranked first, where $\hat{Y}$ = estimated fodder yield, L = avg. no. of leaves per secondary branch (S) and hence, recommended for fodder yield prediction of *Grewia optiva* for the present study area.

Key Words : Cross validation, Regression analysis, Goodness of fit, *Grewia optiva*

View Point Article : Wani, Fehim Jeelani, Rizvi, S.E.H., Sharma, Manish Kumar and Bhat, M. Iqbal Jeelani (2018). A study on cross validation for model selection and estimation. *Internat. J. agric. Sci.*, **14** (1) : 165-172, **DOI:10.15740/HAS/IJAS/14.1/165-172.**

Article History : Received : 16.03.2017; **Revised :** 26.11.2017; **Accepted :** 09.12.2017

INTRODUCTION

Regression analysis has received wide use in data analysis and the development of empirical models. Once a model which gives an adequate fit to the data has been found, the next step in the process is to use the model for prediction, or to learn about the mechanism, generated

*** Author for correspondence:**
[1]Division of Statistics and Computer Science, Faculty of Basic Science (SKUAST-J), Main Campus, CHATHA (JAMMU) INDIA

by the data. But before the model is to be used, its validity should be checked. A valid comparison of real data and model output in the validation stage requires an understanding of the nature of the problem plus the availability of statistical procedures which had been designed to fit the conditions of the problem. Ideally this can be done by using data not used earlier in either model formulation or calibrating. The importance of validation to those who construct and use models is well recognized (Caswell, 1976; Gentil and Blake, 1981; Reynolds *et al.*, 1981; Mayer and Butler, 1993; Oreskes *et al.*, 1994; Rykiel, 1996; Loehle, 1997; Vanclay and Skovsgaard, 1997 and Robinson and Ek, 2000). To develop a prediction model, different candidate models are often compared and the 'best' model is selected based on some model selection criteria. Once the 'best' prediction model is determined, its predictive performance on new data needs to be assessed. It is well known that model fitting statistics may not be a good indication on how well a model will predict. It is easy to over-fit the data by including too many covariates to subsequently inflate model fitting statistics. The best way to measure the predictive ability of a model is to test it on an independent dataset not used in parameter estimation. But an independent dataset is often not available or difficult and expensive to collect (Snee, 1977). Wani *et al.* (2015) fitted different regression models for fodder yield estimation of *Grewia optiva* in Jammu Shiwaliks.

One way to address the problem is through cross validation (CV). CV is a data resampling method by partitioning a dataset into two: a training dataset and a testing dataset. The training dataset is used to fit a model, and the testing dataset is used to evaluate the predictive performance of the fitted model through prediction errors. The idea of CV originated in the 1930s (Larson, 1931) and was further developed by Mosteller and Turkey (1968) and others (Stone, 1974; Gelfand *et al*.,1992 and Shao, 1993).

MATERIAL AND METHODS

Study area:

The present study was conducted in Jammu region of Jammu and Kashmir State covering Shiwalik belt. Samba, Kathua and Udhampur districts were purposely selected. From each district two villages were randomly selected in order to select fodder trees from these villages as the ultimate unit for study purpose. The Jammu and Kashmir state is situated between 32^0 17´–36^0 58´ North latitude and 37^0 26´–80^0 30´ East longitude. The state is the northern most part of the India which is girdled by Tibet to the east, China and Afghanistan to the north, to its west is Pakistan, to its south lies the states of Punjab and Himachal Pradesh and have been divided into three agro climatic divisions *viz.*, outer plain and outer hills, middle mountains and Kashmir valley and inner Himalayas (Ladakh). The outer plains and outer hills include Kathua and Jammu districts, extends up to Shivalik hills in the north. The three districts which were purposely selected for the present study falls under sub-tropical zone as per NARP classification has hot and dry climate in summer and cold climate in winter. The sub-tropical region receives regular monsoons whereas the northern part prone to hailstorms experiences excessive rains. The mean temperature of the Samba, Reasi and Jammu district of the Jammu region varies between the minimum of 3^0–4^0C from December to January to the maximum of 43^0–47^0C from May to June. The rainy season of the area normally starts from the end of June or in the first fortnight of July.

Description of the tree species:

Grewia optiva is one of the most important fodder trees of north-western and central Himalayas. The tremendous popularity of *Grewia optiva* has added its importance in social and agroforestry programme. It is a moderate sized tree, with a spreading crown, reaching height upto 12 m with clear bole of 3-4 m and girth 80 cm, when fully grown. Its bark yields a fibre which is used for basket making, its fruit is edible. Its leaves provide very nutritious fodder; the leaves and edible green twigs are palatable, nutritious and easily digestible. The leaves are rated as good fodder (Laurie, 1945). The green leaves constitute about 70 per cent of the total green weight of branches (Chandra and Sharma, 1977). The leaf fodder, when fed with straw or other inferior dry roughage can profitably substitute concentrates. Nutritious young leaves are converted into a protein-rich meal after in the sun. The calorific value of the tree is 4920 k cal kg^{-1}, which makes it a very good fuelwood tree and alternate source of energy (Joshi and Dhiman, 1992). The species is raised by planting seedlings or stumps usually at 8m spacing in single rows on field bund terrace rivers. It is very heavily lopped for fodder which is the main use of the species (Sehgal and Chauhan, 1989). This system is practised in Jammu and Kashmir, Himachal Pradesh and Uttar Pradesh (Garhwal and Kumaon regions) Himalaya at the elevation of about 550 to 2,300m .

Sampling structure :

In order to have a true representative sample from entire study area, a multistage sampling technique was adopted for the selection of sampling units in which districts were the first stage, villages within each district formed second stage units and *Grewia optiva* trees in the selected villages were considered as the ultimate units. Samba, Reasi and Udhampur districts were purposely selected. From each district two villages were selected randomly. For the selection of the ultimate unit (fodder tree), a simple random sampling without replacement (SRSWOR) was adopted. The procedure of equal allocation was made in each village. For the selection of trees, in each village the total number of trees were counted and with the help of random number table, a total number of 10 trees were selected from each village. A predetermined sample size of 60 trees was selected from three districts following the same procedure in each village. After selecting the trees the different parameters were recorded.

The primary data regarding the different parameters (height, bole height, diameter at breast height (dbh), no. of primary branches, secondary branches, average no. of leaves per secondary branch, age, canopy diameter and green fodder yield) of *Grewia optiva* was collected for each randomly selected units (fodder tree). The collection of information regarding various aspects of *Grewia optiva* was done by visiting the fields of farmers and contacting the farmers directly during the study period.

Cross validation :

There are different types of CV methods that adopt different data resampling techniques. The simplest method is the holdout method, in which a dataset is divided into a training set and a testing set. *k*-fold CV, the basic form of CV, is one way to improve over the holdout method. A dataset is randomly partitioned into k mutually exclusive equal (or nearly equal) subsets, and the holdout method is repeated *k* times. Each time, one of the k subsets is used as a testing set and the remaining *k-1* subsets are put together to form a training set. Then the average prediction error on the testing data across all *k* trials is computed. For the k-fold method, data resampling is done without replacement and it matters less how the data get divided. Every data point gets to be in the testing sets exactly once, and gets to be in the training sets *k-1* times. For the k-fold method, typical choice of *k* value is 5 or 10 (Hastie *et al.,* 2009). For larger datasets, smaller k values are sufficient. For smaller datasets, however, larger k values are needed. In the present study, 10-fold cross validation technique was adopted.

For model validation, the estimates of Apparent error, True error and Excess error of model are critical. Apparent error (also called resubstitution error) is computed by applying the fitted equation to the data used in calibration of the model and will normally give an optimistic view of the quality of a model. True error is estimated by fitting the model to independent data (computed by applying the fitted equation to the data not used for calibration of the model). Apparent error underestimates the true error ("it is downwardly biased"). The difference between true error and apparent error is known as Excess error. The relationship can be formulated as :

True error = Apparent error + Excess error

The predictive ability of the different models were to be assessed on the basis of following evaluation criteria by using second data set (known as validating data set).

Average residual or prediction bias (B), $\mathbf{B = \frac{\sum r_i}{n}}$

where, r_i represents the difference between the observed and predicted fodder yield for ith tree in the validating data set. The variance of B is obtained by using formula:

$$\mathbf{Var\,(B) = \frac{\sum_{i=1}^{n}(r_i - B)^2}{n-1}}$$

The root mean square error (RMSE) provides a composite measure (combining bias and precision) of the overall accuracy of prediction. The smaller these values the better the prediction.

$$\mathbf{RMSE = \sqrt{\left[B^2 + Var(B)\right]}}$$

This co-efficient of dispersion (CD) based on standard deviation, which measures the proportion variation in bias provides a composite measure of overall accuracy of prediction. The smaller the value, the better the prediction. Moreover, it is unitless too. The CD is obtained by using following formula :

$$\mathbf{CD = \frac{\sqrt{Var(B)}}{B}}$$

All the procedures discussed so far belong to parametric tools. In addition some non-parametric technique could also be used so as to arrive at final decision criterion for selection of developed model. In this regard, Wilcoxon's signed rank test (Wilcoxon, 1945) was used to test bias produced by each equation. This non-parametric test assumes that there is information in the magnitudes of the differences between paired observations and rank them from smallest to largest by absolute value. Add all the ranks associated with positive differences and then negative differences. Finally, the p-value associated with this statistic is found from an appropriate table. A rank was assigned to each equation based on each evaluation criteria (Cao *et al*., 1980). The smaller the rank value the better the performance of the model. These ranks of all criterion are then summed up to arrive at the final fit rank for each equation, which is the indicative of model's performance with respect to all the criteria considered.

RESULTS AND DISCUSSION

The results obtained from the present investigation as well as relevant discussion have been summarized under following heads :

Model construction:

For the development of the model, the fodder yield was the dependent variable and yield attributing characters *viz.*, height, bole height, dbh, primary branch, secondary branch, average number of leaves per secondary branch, age and canopy diameter were the independent variables. The values of all the simple correlation co-efficient along with p-values between fodder yield and yield attributing parameters have been given in Table 1. The highest value of correlation co-efficient was recorded between fodder yield and number of leaves per secondary branch (0.903) followed by number of secondary branch per tree (0.725) and bole height (0.639). The different forms of models were tried with individual independent variables and their feasible combinations were also considered. In total, 30 models were tried for the present investigation. Then the models with the parameter co-efficients obtained at fitting level were applied on the validation data set. Table 2 showed all the fitted linear models. Out of 30 models, the best five models were selected on the basis of highest R^2. From Table 3 we can depict the performance of model at serial number 4 as best followed by serial number 3 and 1, respectively. Table 4 gives the other related characteristics like e*, e**, collinearity statistics and error index of the fiited models. The collinearity statistics given in Table 4 shows there is no multicollinearity present in the models. The value of tolerance for the first model 1 is maximum followed by model 4, whereas the VIF is maximum for model 2 and minimum for model 1. The 1st and 2nd column of Table 4 gives the lower and upper bounds of error index. On the basis of the criterion of adj. R^2 the model at serial number 4 as best followed by model at serial number 3. Whereas on the basis of the criterion of error index the model at serial number 5 as best followed by model at serial number 4.

Model validation:

To assess the predictive ability of different functions, it is pertinent to use an independent data set (validating data set) for model validation. In the present study, dataset was randomly partitioned into 10 mutually exclusive equal (or nearly equal) subsets. Each time, one of the subsets is used as a testing set and the remaining subsets are put together to form a training set. The validation process is necessary so that the model can be used with some confidence. The equations (with the parameters obtained from the fitting data set) were applied to the validating

Table 1 : Simple correlation co-efficients between fodder yield and other related parameters

Parameters	Correlation co-efficient (r)	p- value
Height	0.635	< 0.01
Bole height	0.639	<0.01
DBH	0.271	<0.05
No. of primary branches	0.460	<0.01
No. of secondary branches	0.725	<0.01
No. of leaves per secondary branch	0.903	<0.01
Age	0.671	<0.01
Canopy diameter	0.575	<0.01

p-value <0.01 shows highly significant p-value <0.05 shows significant data set.

For model validation, the estimates of Apparent error, True error and excess error of model are critical. Less the value of excess error, the better the predictive ability of the equation. From Table 5 which gives the values of errors of the equations, it can be concluded that Equations 1, 2, 4 and 5 has given reasonable values of excess error, whereas Equation 3 has given the highest value of excess error, which means that the predictive ability of Equation 3 is poorer than other equations in case of excess error criterion. Whereas the predictive ability of Equation 1 may be judged as best than other equations. Table 6 compares the validation statistics for the five equations used over the validating data set. It can be observed from table that Equation 5 has the lowest value of bias (0.53) whereas Equation 2 has highest value of bias (0.80). In case of variance, Equation 2 has highest value (1.53) whereas Equation 5 has lowest value (0.21). The combined effect of bias and variance is expressed as RMSE. With regard to RMSE, Equation 5 has the least

Table 2: Models and other related characteristics

Sr. No.	Models	R^2	Adj. R^2	¡ ^{2}
1.	$\hat{Y} = 8.871 + 0.000003L^2S$	0.822 (0.000)	0.818 (0.000)	2.13
2.	$\hat{Y} = 3.522 + 0.256S + 0.00007L^2$	0.825 (0.000)	0.817 (0.000)	2.06
3.	$\hat{Y} = 2.869 + 2.301B + 0.00007L^2$	0.847 (0.000)	0.840 (0.000)	1.89
4.	$\hat{Y} = 0.476 + 0.474A + 0.00007L^2$	0.855 (0.000)	0.849 (0.000)	1.81
5.	$\hat{Y} = -9.669 + 0.480A + 0.237\sqrt{LS}$	0.824 (0.000)	0.817 (0.000)	2.02
6.	$\hat{Y} = 8.992 + 0.000003L^2S$	0.862 (0.000)	0.859 (0.000)	1.69
7.	$\hat{Y} = 3.567 + 0.250S + 0.00007L^2$	0.866 (0.000)	0.860 (0.000)	1.62
8.	$\hat{Y} = 3.117 + 2.068B + 0.00007L^2$	0.884 (0.000)	0.879 (0.000)	1.48
9.	$\hat{Y} = 1.154 + 0.402A + 0.00007L^2$	0.889 (0.000)	0.885 (0.000)	1.45
10.	$\hat{Y} = -8.809 + 0.351A + 0.255\sqrt{LS}$	0.860 (0.000)	0.854 (0.000)	1.73
11.	$\hat{Y} = 8.480 + 0.000004L^2S$	0.880 (0.000)	0.878 (0.000)	1.53
12.	$\hat{Y} = 1.936 + 0.343S + 0.00007L^2$	0.875 (0.000)	0.870 (0.000)	1.56
13.	$\hat{Y} = 2.741 + 2.235B + 0.00007L^2$	0.880 (0.000)	0.875 (0.000)	1.51
14.	$\hat{Y} = 0.526 + 0.443A + 0.00007L^2$	0.884 (0.000)	0.879 (0.000)	1.49
15.	$\hat{Y} = -10.14 + 0.411A + 0.259\sqrt{LS}$	0.871 (0.000)	0.866 (0.000)	1.59
16.	$\hat{Y} = 9.470 + 0.000003L^2S$	0.836 (0.000)	0.833 (0.000)	2.02
17.	$\hat{Y} = 5.039 + 0.201S + 0.00006L^2$	0.847 (0.000)	0.841 (0.000)	1.91
18.	$Y = 4.756 + 1.662B + 0.00006L^2$	0.861 (0.000)	0.855 (0.000)	1.64
19.	$\hat{Y} = 3.829 + 0.266A + 0.00007L^2$	0.849 (0.000)	0.843 (0.000)	1.87
20.	$\hat{Y} = -5.905 + 0.229A + 0.241\sqrt{LS}$	0.805 (0.000)	0.797 (0.000)	2.23
21.	$\hat{Y} = 8.990 + 0.000003L^2S$	0.810 (0.000)	0.806 (0.000)	2.17
22.	$\hat{Y} = 2.609 + 0.300S + 0.00007L^2$	0.812 (0.000)	0.804 (0.000)	2.25
23.	$\hat{Y} = 2.100 + 2.703B + 0.00006L^2$	0.842 (0.000)	0.835 (0.000)	1.95
24.	$\hat{Y} = 0.136 + 0.526A + 0.00007L^2$	0.846 (0.000)	0.840 (0.000)	1.92
25.	$\hat{Y} = -8.980 + 0.521A + 0.222\sqrt{LS}$	0.831 (0.000)	0.824 (0.000)	2.07
26.	$\hat{Y} = 8.790 + 0.000003L^2S$	0.861 (0.000)	0.858 (0.000)	1.71
27.	$\hat{Y} = 2.448 + 0.331S + 0.00007L^2$	0.856 (0.000)	0.850 (0.000)	1.77
28.	$\hat{Y} = 1.534 + 2.963B + 0.00006L^2$	0.883 (0.000)	0.878 (0.000)	1.49
29.	$\hat{Y} = 0.018 + 0.502A + 0.00007L^2$	0.880 (0.000)	0.875 (0.000)	1.52
30.	$\hat{Y} = -9.683 + 0.452A + 0.244\sqrt{LS}$	0.862 (0.000)	0.856 (0.000)	1.70

Figures given in parentheses indicate p-value p-value<0.01 indicates highly significant

$\hat{Y}$ = Estimated fodder yield per tree, L = Avg. number of leaves per secondary branch, S = Number of secondary branches, B = Bole height, a = age of the tree value (0.63) whereas Equation 2 has the highest value (1.47). Co-efficient of dispersion has also been calculated to evaluate the proportion variation in the mean, standard deviation being considered as the total variation in the

Table 3: Best five selected models

Sr. No.	Models	$\mathbf{R}^2$	Adj. $\mathbf{R}^2$	χ^2
1.	$\hat{Y} = 8.480 + 0.000004L^2S$	0.880 (0.000)	0.878 (0.000)	1.53
2.	$\hat{Y} = 1.936 + 0.343S + 0.00007L^2$	0.875 (0.000)	0.870 (0.000)	1.56
3.	$\hat{Y} = 3.117 + 2.068B + 0.00007L^2$	0.884 (0.000)	0.879 (0.000)	1.48
4.	$\hat{Y} = 1.154 + 0.402A + 0.00007L^2$	0.889 (0.000)	0.885 (0.000)	1.45
5.	$\hat{Y} = -10.14 + 0.411A + 0.259\sqrt{LS}$	0.871(0.000)	0.866(0.000)	1.59

Figures given in parentheses indicate p-value p-value < 0.01 indicates highly significant

$\hat{Y}$ = Estimated fodder yield per tree, L = Avg. number of leaves per secondary branch, S = Number of secondary branches, B = Bole height, a = Age of the tree

Table 4 : Other related characteristics of best selected models

Models	e*	e**	Collinearity statistics		
			Tolerance	VIF	Error index
1	0.062	0.044	1	1	10.09
2	0.065	0.039	0.543	1.841	9.96
3	0.059	0.041	0.630	1.587	9.58
4	0.057	0.036	0.713	1.403	9.35
5	0.052	0.034	0.611	1.636	8.91

Table 5: Values of errors of equations

Equations	Apparent error	True error	Excess error
1	0.116	0.176	0.060
2	0.510	0.763	0.253
3	0.423	0.858	0.435
4	0.544	0.876	0.332
5	0.592	0.815	0.223

Table 6 : Validation statistics for equations with independent data set

Equations	Bias	Var (B)	RMSE	CD	Σ Rank	Final rank
1	0.78 (4)	1.45 (4)	1.41 (4)	1.47 (2)	14	4
2	0.80 (5)	1.53 (5)	1.47 (5)	1.56 (5)	20	5
3	0.71 (2)	1.29 (2)	1.37 (2)	1.51 (4)	10	2
4	0.74 (3)	1.38 (3)	1.34 (3)	1.49 (3)	12	3
5	0.53 (1)	0.21 (1)	0.63 (1)	0.69 (1)	4	1

Table 7 : Wilcoxon's signed rank test and combined result of the criterion ranks

Equations	Z	Asymptotic significance	R_f	R_v	Σ rank	Final rank
1	-1.284	0.128	1	4	5	1
2	-1.029	0.239	2	5	7	3
3	-0.163	0.864	4	2	6	2
4	-0.148	0.782	3	3	6	2
5	-0.113	0.895	5	1	6	2

R_f – Rank of fitting set R_v – Rank of validating set mean and for this Equation 5 has the least value (0.69) whereas Equation 3 has the highest value (1.51). For each criterion, the ranks has been assigned and these ranks are then summed upto give final ranks. After considering all the ranks, the Equation 5 was ranked first followed by Equations 3, 4, 1 and 2 at last.

After the validation statistics with independent data set, a non-parametric test (Wilcoxon's signed rank test) was used to test bias produced by each equation. This non-parametric test assumes that there is information in the magnitudes of the differences between paired observations. This test begins by transforming each value into its absolute value, which is accomplished simply by removing all the positive and negative signs. The absolute differences are then ranked from lowest to highest, with tied ranks included where appropriate. Add all the ranks associated with positive differences and then negative differences. Finally, the Z-value can be calculated. The asymptotic significance of Wilcoxon's signed rank test for validating data set for all equations showed that the null hypothesis of test, *i.e*. the difference between sum of the positive and negative rank is zero is accepted. It can be depicted from Table 7 that Equation 1 has the lowest asymptotic significance whereas the Equation 5 has the highest asymptotic significance. The equation with lowest asymptotic significance has been ranked first and equation with highest significance has been ranked last. Table 7 shows that Equation 1 has been ranked first and Equation 5 has been ranked last. By considering ranks of both fitting (R_f) and validating (R_v) data sets, one can arrive at the final ranks. Table 7 shows that overall rank of Equation 1 is lowest, which means that Equation 1 would perform better for predicting fodder yield estimation of *Grewia optiva*. Therefore, it can be concluded that :

$$\hat{Y} = 8.480 + 0.00000\,4\,L^2\,S$$

where $\hat{y}$ is the estimated fodder yield, L = no. of leaves per secondary branch, S = no. of secondary branch, has been found best among all the five best considered models obtained through the applications of parametric as well as non-parametric (Wilcoxon test) tests and therefore, recommended for fodder yield estimation of *Grewia optiva* on the basis of present investigation.

Conclusion:

The data collected were subjected to various statistical analyses for studying the technique of cross validation for model selection and estimation. Regression analysis was used to study the relationship between fodder yield (dependent variable) and other parameters (independent variables) like height, bole height, dbh, average number of leaves per secondary branch, number of primary branch, secondary branch, canopy diameter and age of the tree. All the yield attributing characters were found to be positively and significantly correlated with fodder yield (dependent variable). For model development, the data recorded on green fodder yield and yield attributing characters for all the 60 randomly selected trees were utilized. In total, 30 models were tried on the fitting data set. Out of total models tried, the best five models were selected on the basis of adj. R^2 value. Goodness of fit of the selected models was also tested by applying chi-square test. The critical error, error index and collinearity statistics were calculated for each model. The collinearity statistics include tolerance and Variance Inflation Factor (VIF), and it was found that multicollinearity was not present in the seclected models. The equations obtained from the fitting data set were applied on an independent data set for validation purpose. For model validation, the estimates of Apparent error, True error and Excess error are critical. The model adequacy was ranked on the basis of different criteria, namely, adj. R^2, bias, root mean square error (RMSE) and co-efficient of dispersion (CD). Finally, the non-parametric test (Wilcoxon signed rank test) was also used to test bias produced by each equation. By considering ranks of both fitting (R_f) and validating (R_v) data sets, one can arrive at the final ranks. The overall rank of Equation 1 was lowest indicating thereby that this equation would perform better for predicting green fodder yield per tree of *Grewia optiva* specie. Therefore, it can be concluded that Equation 1 should be preferred over other equations considered for present study.

REFERENCES

Cao, Q.V., Burkhart, H.E. and Max, T. A. (1980). Evaluation of two methods for cubicvolume prediction of loblolly pine to any merchantable limit. *Forest Sci.,* **26** (1) : 71-80.

Caswell, H. (1976). The validation problem. In: Patten, B. (Ed.), *Systems analysis and simulation in ecology*, vol. 4. Academic Press, New York, pp. 313–325.

Chandra, J.P. and Sharma, R.K. (1977). Note on nursery technique of beul (*Grewia oppositifolia*). *Indian Forester*, **103** (10) : 684-685.

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods. Technical Report No. 462, Department of Statistics, Stanford University, Stanford, California, 38 pp.

Gentil, S. and Blake, G. (1981). Validation of complex ecosystem models. *Ecol. Modelling,* **14** : 21–38.

Hastie, T., Tibshirani, R. and Friedman J. (2009). *The elements of statistical learning: data mining, inference and prediction 2009.* 2nd Ed. Springer Series in Statistics,745.

Joshi, N.K. and Dhiman, R.C. (1992). Lopping yield studies of *Grewia optiva* Drummond. *Van Vigyan,* **30**(2) : 80–85.

Larson, S. (1931). The shrinkage of the co-efficient of multiple correlation. *J. Edu. Psychol.,* **22** : 45–55.

Laurie, M.V. (1945). *Fodder trees in india*. pp. 17-82. FRI Dehradun.

Loehle, C. (1997). A hypothesis testing framework for evaluating ecosystem model performance. *Ecol. Modelling* **97** : 153–165.

Mayer, D.G. and Butler, D.G. (1993). Statistical validation. *Ecol. Modelling,* **68** : 21–32.

Mosteller, F. and Turkey, J.W. (1968). Data analysis, including statistics. In: *Handbook of social psychology*. Addison-Wesley, pp. 601–720.

Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earthsciences. *Science,* **263** : 641–646.

Reynolds, Jr. M.R., Burkhart, H.E. and Daniels, R.F. (1981). Procedures for statistical validation of stochastic simulation models. *Forest Sci*., **27** (2) : 349–364.

Robinson, A.P. and Ek, A.R. (2000). The consequences of hierarchy for modelling in forest ecosystems. *Can. J. Forest Res.*, **30** (12) : 1837–1846.

Rykiel, E.J. (1996). Testing ecological models - the meaning of validation. *Ecol. Modelling,* **90** (3) : 229–244.

Sehgal, R. N. and Chauhan, V. (1989). *Grewia optiva* an ideal agroforestry tree of western Himalaya. Farm Forestry News 5. Winrock International, USA.

Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Stat. Assoc*., **88** : 486–494.

Snee, R.D. (1977). Validation of regression models: methods and examples. *Technometrics,* **19** : 415–428.

Stone, M. (1974). Cross-validatory choice and the assessment of statistical predictions. *J. Roy. Stat. Soc.* Ser B., **36**:111–133.

Vanclay, J.K. and Skovsgaard, J.P. (1997). Evaluating forest growth models. *Ecol. Modelling,* **98** (1) : 1–12.

Wani, F. J.., Rizvi, S. E. H. and Sharma, M. K. (2015). Statistical Modelling for fodder yield estimation of *Grewia optiva* in Jammu Shiwaliks. *Internat. J. Agric. & Statistical Sci.,* **11**(1) : 139-142.

Wilcoxon, Frank (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,* **1** (6) : 80-83.