RESEARCH ARTICLE Legume Research- An International Journal, Volume 45 Issue 7: 822-827 (July 2022)

Trait Based Modelling Approach for Selection of Elite Germplasm Accessions in Soybean [*Glycine max* (L). Merrill]

K. Shruthi[1], R. Siddaraju[1], K. Naveena[2], T.M. Ramanappa[1], C. Gireesh[3], K. Vishwanath[1], K.S. Nagaraju[4]

10.18805/LR-4567

ABSTRACT

Background: Identification of suitable factors that influence significantly to the response is crucial for the traits based breeding program to make a better decision about improvement in productivity. Multiple linear regression (MLR) is the benchmark method commonly using to identify suitable factors for crop improvement. It doesn't work always due to stringent assumption (Multicollinearity, Linearity) behind the MLR model. Here we tried to develop an efficient model for the selection of major traits that contribute to seed yield in soybean by comparing different models.

Methods: Field experiment was conducted using 98 soybean core population through augmented design. 18 morphometric traits obtain from soybean core population were considered under the study as regressors. Multiple linear regression (MLR), principle component regression (PCR), regression tree and Random Forest models were compared to select traits based on prediction accuracy.

Result: All the models identified the number of pods per plant (NPP) has the most influencing variable to the soybean yield. However random forest has a much higher prediction power (RMSE=4.59, MAPE=0.18) compared to other models under study. The results of random forest revealed that the number of pods per plant, number of branches per plant and other associated characters like plant height at harvest as highly influencing traits for seed yield in soybean. Finally, tried to identify genotypesthat possess superiority about most influencing morphological characters on seed yield using cluster analysis.

Key words: Multiple linear regression, Principle component analysis, Random forest, Regression tree, Seed yield.

INTRODUCTION

Soybean [*Glycine max* (L.) Merril] is the world's most important seed legume and contributes~25 % of the global edible oil and about two-third of the world's protein concentrate for livestock feeding (Singh and Hymowitz, 1999). It has earned epithets like "Cow of the field" or "Gold from soil", "poor man's food" and "wonder crop". It is globally grown over an area of 125.64 mhawith a production of 358.65 MT and productivity of 2.85 metric ton per ha during 2018-19. India ranks 4th in terms of global soybean area sown (10.40 m ha) and 5th (10.93 mt) in terms of soybean production after USA, Brazil, Argentina and China. India has less productivity (0.96 metric ton per ha) compare to average world productivity (Anonymous 2020).

Morphological characters plays critical role in the selection of desirable parents in plant breeding program. Additionally, yield and yield contributingcharacters are very helpful through which overall performance of genotypes could be determined (Hasan *et al*., 2015). Seed yield is an important parameter influenced by several other characters, where few of them only significantly contributein yield formation. Hence, characterization of genotypes based on these major characters will improve the accuracy for selection of parents. Therefore, identifying traits which are closely related and have significant contribution to yield becomes highly essential. Germplasm is the ultimate source of genetic variations in soybean improvement program. Globally, there are 1,70,000 accessions of soybean germplasm available (Husain and Shrivastav, 2011) and in India approximately 3443 accessions of soybean germplasm maintained at National Active Germplasm Sites (Gireesh *et al*., 2015). Genetic assessment of germplasm diversity is imperative to identify the promising accessions for trait of interest that can be utilized for genetic improvement of soybean.

[1]University of Agricultural Sciences, National Seed Project, Bengaluru-560 065, Karnataka, India.
[2]Centre for Water Resources Development and Management, Calicut-673 001, Kerala, India.
[3]ICAR-Indian Institute of Rice Research, Hyderabad-500 030, Telangana, India.
[4]Gandhi Krishi Vigyan Kendra, University of Agricultural Sciences, Bengaluru-560 065, Karnataka, India.

Corresponding Author: K. Shruthi, University of Agricultural Sciences, National Seed Project, Bengaluru-560 065, Karnataka, India. Email: shruthikns3@gmail.com

How to cite this article: Shruthi, K., Siddaraju, R., Naveena, K., Ramanappa, T.M., Gireesh, C., Vishwanath, K. and Nagaraju, K.S. (2022). Trait Based Modelling Approach for Selection of Elite Germplasm Accessions in Soybean [*Glycine max* (L). Merrill]. Legume Research. 45(7): 822-827. DOI: 10.18805/LR-4567.

Submitted: 09-12-2020 **Accepted:** 14-04-2021 **Online:** 26-05-2021

Statistical modeling is one of the way to identify significantly associated charcters to yield. Classical variable selection method like Multiple regression approach (Ghanbari *et al*., 2018; Vu *et al*., 2019) is the benchmark statistical technique commonly used for analysing the relationship between the triats. But sometimes it misleads the researchers due to its stringent assumptions like Multicollinearity, Linearity *etc*. If yield attributing triats possess a multicollinearity problem then multiple regression analys is overestimates the relationship between the yield and its associating variables (Johnston *et al.,* 2018). If there is non linear relationship between yield and its explanatory charcters then MLR predict yield with higher bias, hence, it is necessary for developing the model which works well under the above problems and explains the actual relationship between yield and its associated variables to take a better decision for selection of parents .

Several studies have explained the factor identification for improving the yield using statistical models (Roberts *et al.,* 2017; Shi *et al*., 2013; Michel *et al*., 2013). Eledum 2016 observed that multicollinearity problem of MLR [Variance inflation factor (VIF)=58.21] can be solved by principle component regression analysis (VIF=1.078) and it is also superior in performance about model accuracy compare to MLR. Jeong *et al.* (2016) found superiority of random forest modelsover MLR models for predicting the crop yields, where the root mean square errors (RMSE) ranged between 6 and 14% of the average observed yield in all test cases whereas RMSE ranged from 14% to 49% for MLR models. This paper focused on the analysis of the soybean morphological data for finding optimal parameters to maximize the yield and precise prediction of yield using different Statistical and machine learning models.

MATERIALS AND METHODS

The material for the study comprised a core set of 98 germplasm accessions which included indigenous and exotic germplasm accessions of soybean along with five high yielding varieties as a check (DSB-21, MAUS-2, KB-79, JS-335, KBS-23) procured from All India Coordinated Research Project (AICRP) on Soybean, UAS, GKVK, Bengaluru.

The 98 accessions and five checks were sown in Augmented design (Federer, 1956) in four blocks during Kharif 2015 and 2016. Each block consisted of 25 germplasm accessions and five checks (replicated twice). Each entry was sown in a single row of 2.5 meters length with a row spacing of 0.45 m and 0.2 m between plants within a row. A basal dose of 25:50:25 Kg NPK $ha^{-1}$was applied to the experimental plot. Recommended crop management practices are followed during the crop growth period to raise a healthy crop.

Observations on different quantitative characters like shoot length (SL), root length (RL), hypocotyl length (HL), epicotyl length (EL), plant height at 30 days (PH@30), plant height at 40 days (PH@40), plant height at harvest (PH@HVT) were recorded using measuring scale and also days to flowering (DF), days to maturity (DM), pod length (POD_L), number of branches per plant (NBP), number of pods per plant (NPP), seed size (SS) and 100 seed weight (TW), shoot length (Shoot_L), seed weight (SW), seed length (SL), seed thickness (ST) and seed yield (SY) were recorded on five randomly selected plants from each germplasm accession and check variety following DUS and UPOVA descriptors (Anonymous, 2009). The number and per cent accessions belonging to each class were counted and computed, respectively. To identify major factors that contribute to seed yield and for prediction of seed yield, we used different statistical tools like multiple linear regression (MLR), principle component regression (PCR), regression tree and random forest technique. Pearson correlation and variance inflation factor (VIF) approaches are used to decide multicollinearity in independent variables. If VIF value of any independent variable is more than 10 indicates multicollinearity (Olivoto *et al.,* 2017).

The popular prediction evaluation methods like coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute percentage error (MAPE) used to evaluates the accuracy of prediction models (Naveena *et al*., 2017) as given in the Framework of the proposed system is portrayed in Fig 1. To check the prediction accuracy of the above models the data was divided into 2 sets *viz*. training and testing. 80 per cent observations were used for training the model and 20 per cent observation for testing of models. Different packages under R studio were applied to analyse above mentioned models.

RESULTS AND DISCUSSION

An attempt was made to develop the model for identify suitable morpho metric variables which project seed yield of soybean germplasm accessions which is having higher genetic variability (Shruthi *et al.,* 2021). The results from the Multiple linear regression (MLR) indicates the VIF values of the most of the variables are more than 10 (Table 1) and high correlation between independent variables (Fig 2) indicting multicollinearity problem in the data set. So this problem effecting the results of MLR and leads to wrong interpretation. Even 85.2 per cent of variation of seed yield explained by selected cause variables (R^2 = 0.852), only few variables (number of pod per plant (0.70**) and days to maturity (-0.28**) are significantly contributing to changes in the seed yield (Table 1). To overcome this problem of multicollinearity observed among biometric data and for the identification of major factors of influence, the principal component regression is used (Goyal and Verma 2018). The eigenvalues corresponding to each principal component represents the variance connected with the particular principal component. The first four eigenvalues had eigen value more than 1 and explains a total of 80.28% variability present in the data. So, the first four eigenvalues are selected to build principle component regression model. The rotated component factor loadings are presented in Table 1. The factor loadings represent the weights assigned to each of

the variables in the linear combination corresponding to each eigen value.

The linear combination of these factor loadings with the corresponding variables gives the corresponding principal components. To assess the degree of relationship between principal components and seed yield, we tried principal component regression by considering the principal components as independent variables and seed yield as the dependent variable. Here first (5.78**) and fourth principal component regression coefficients (6.30**) are significantly contributing to seed yield, so variables which having factor loadings more than 0.7 under first and fourth principal component considered as important variables for seed yield improvement. So as per Table 1 principle component regression showed quantitative variables like shoot length (0.87), root length (0.87), hypocotyl length (0.81), epicotyl length (0.74), plant height at 30 days (0.84), plant height at 40 days (0.87), plant height at harvest (0.83), number of pods per plant (0.70) and shoot length(0.87) are significantly contributing to the seed yield. Even though PCR works better under the multicollinearity situation but prediction accuracy of this model is less (R^2=0.582) so further we tried regression tree and random forest models which works well under multicollinearity situation with high prediction accuracy.

The regression tree ranks the variables based on its contribution to predicting the seed yield using the classification and regression tree (CRT) method, the part

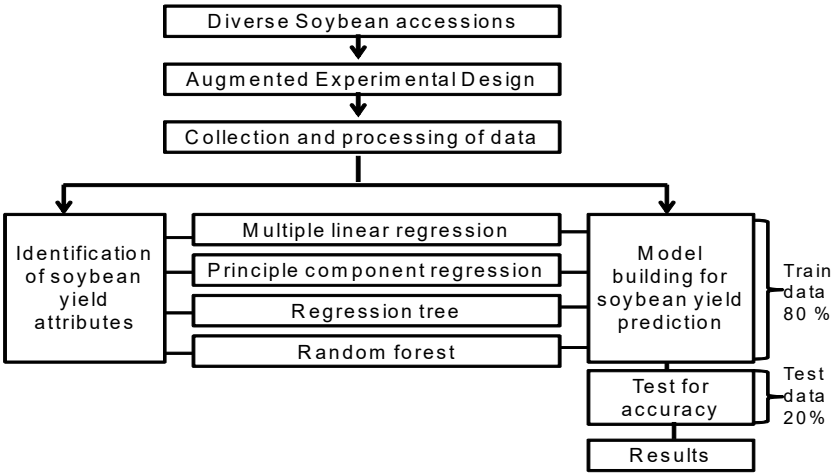

Fig 1: Frame work for proposed system.

Table 1: Multiple liner regression and principal component regression forselection of morphological traits.

Parameters	MLR		PCR	
	Regression coefficients	VIF	PC1	PC2
(Constant)	25.53^{NS}	-	-	-
Seed length	-10.57*	36.15	0.46	-0.09
Seed weight	-10.52^{NS}	41	0.30	-0.16
Seed thickness	-17.75*	55.94	0.34	-0.12
Seed size	37.3^{NS}	293.74	0.42	-0.15
Test weight	0.48^{NS}	2.98	0.42	0.01
Plant height @ 30 days	0.38^{NS}	25.94	0.84	0.10
Plant height @ 40 days	-0.22^{NS}	37.15	0.87	0.14
Plant height @ harvest	0.01^{NS}	13.53	0.83	0.19
Days to flowering	0.21^{NS}	1.45	0.11	0.07
Number of branches per plant	0.24^{NS}	2.59	0.61	0.51
Number of pods per plant	0.70*	2.94	0.57	0.70
Days to maturity	-0.28*	1.48	-0.01	0.19
Shoot length	-0.16^{NS}	30.61	0.87	-0.25
Root length	-0.73^{NS}	20.17	0.87	-0.26
Hypocotyl length	-0.06^{NS}	22.95	0.81	-0.32
Epicotyl length	1.95^{NS}	9.05	0.74	-0.42
Pod length	-0.68^{NS}	1.49	0.26	0.50
F-value	13.94			
R^2	0.736			

algorithm of R software used to build the model. Fig 3 represents the results of regression tree modeling about the importance of morphological character on seed yield. Which defined higher the importance of variable when it possesses higher importance score. The order of performance of the variables was as follows number of pods per plant (9425.73) > number of branches per plant (2153.73) > plant height at harvest (1823.25) *etc.* as given in Fig 3. While, seed size, seed thickness, days to maturity having importance scores near to zero so they never apperars as primary or a surrogate splitters and regression tree model eliminate this varibles from tree. Number of pods per plant, number of branches per plant, plant height at harvest, plant height at 30 and plant height at 40 days will be considered as important traits besed on the high impotance score as given in Fig 3. Overall prediction accuracy of this model is (R^2=0.766) much better than Principle component regression (R^2=0.582) as given in Table 2 hence, further we are trying random forest model.

Random forest predict the seed yield using the random forest algorithm of R software. Tune grid function used to identify optimal number of variables available for splitting at each tree node (mtry), Number of trees to grow (ntree) and the minimum number of observations in a terminal node (max nodes) of the model. Among all possible combinations optimal parametrs, mtry=10 (R^2=0.74, RMSE=6.15, MSE=4.95), ntree=130 (R^2=0.79, RMSE=5.15, MSE=4.90) and MAX nodes=8 (R^2=0.71, RMSE=6.12, MSE=4.94) having high level of accuracy of prediction. Overall prediction accuracy of this model (R^2=0.925) is much better than all other models as given in Table 2. Fig 3 also explain the rankings of the relative importance of each morphological charcter on seed yield. Higher the value of purity indicates the higher the importance of variable. Here number of pods per plant possess most importance with higher rank (4765.41). The importance of variables according to purity values obtained by random forest is number of pods per plant (4765.41) > number of branches per plant (1265.36)

Table 2: Evaluation of models for prediction of soybean seed yield.

Dataset	Methods	RMSE	MAPE	R^2
Overall set (100%)	Multiple linear regression	5.74	0.21	0.74
	Principle component regression	7.22	0.26	0.58
	Regression tree	5.40	0.18	0.77
	Random forest	3.10	0.10	0.92
Training set (80%)	Multiple linear regression	5.24	0.20	0.76
	Principle component regression	7.20	0.27	0.67
	Regression tree	5.10	0.16	0.77
	Random forest	3.07	0.11	0.92
Testing set (20%)	Multiple linear regression	9.15	0.27	0.47
	Principle component regression	7.31	0.23	0.66
	Regression tree	9.10	0.24	0.49
	Random forest	4.59	0.18	0.84

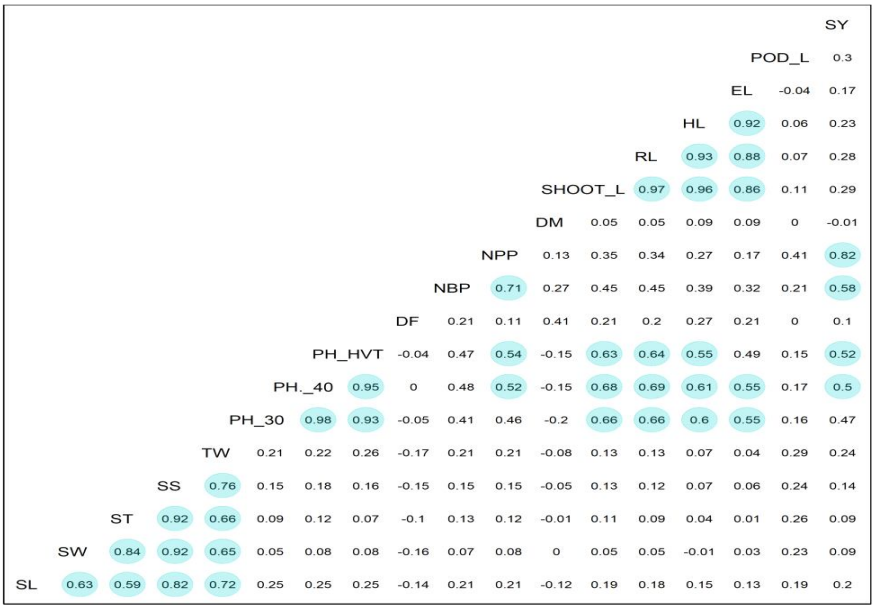

Fig 2: Correlation matrix between biometric variables.

> plant height at harvest (1113.83) *etc.* as given in Fig 3. Hence, number of pods per plant, plant height at harvest, number of branches per plant, plant height at 30 days and plant height at 40 days will be considered as important variables as like in regression tree model because of significantly high purity values and this parameters have positive significant relation with seed yield as given in Fig 1.

To check the capability of each model to predict seed yield the data was divided into 2 sets *viz.* training and testing data. 80% *i.e.* 83 genotypes observations were used for training and 20% *i.e.* 20 genoypes observations were used for testing models. The models were trained saperatly to build model and the best model was selected on the basis of its prediction accuracy in the testing period. The comparative results for the best model between multiple linear regression, Principle component regression, Regression tree and Random forest models are given in Table 2. As assessed by prediction accuracy measures like RMSE, MAPE and R^2 statistic indicates the superiority of the random forest for prediction of soybean seed yield for germplasm accessions. It indicates number of pods per plant, number of branches per plant, plant height at harvest, plant height at 30 days and plant height at 40 days will be considered as most influencing morphological characters on seed yield.

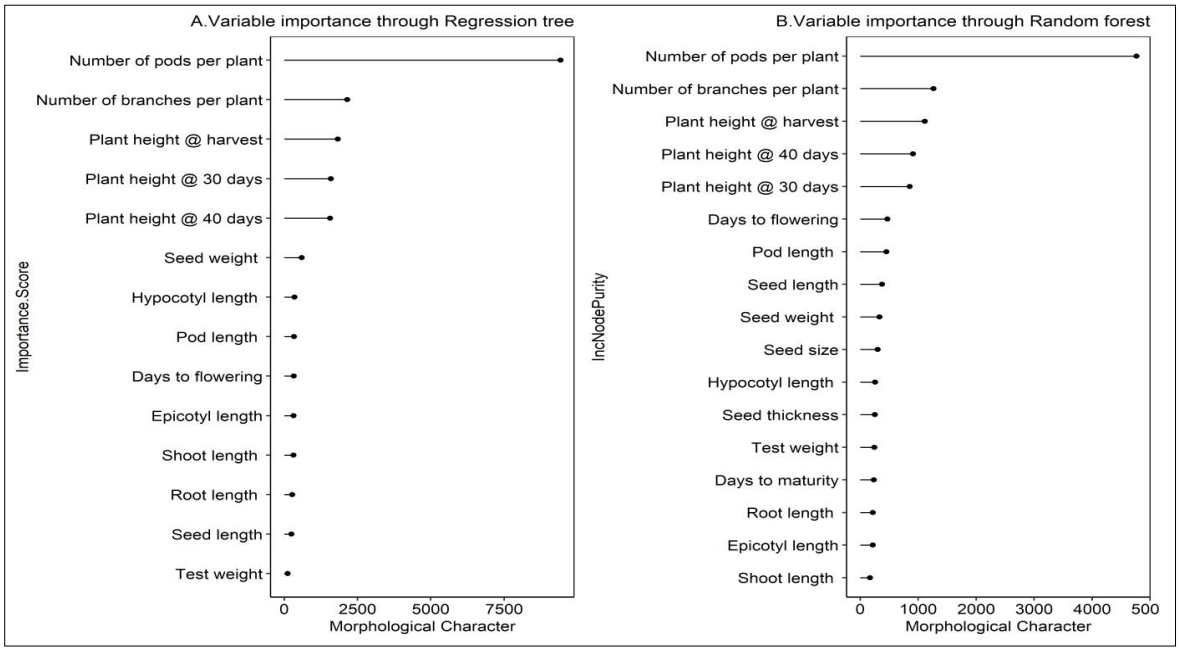

Fig 3: Importance of morphological characters vs seed yield through regression tree (A) and random forest (B).

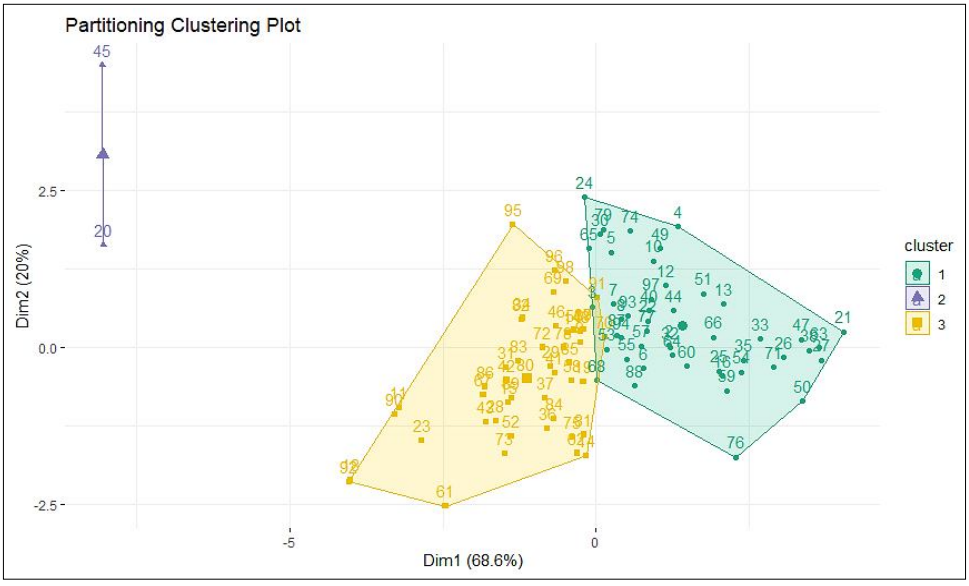

Fig 4: Classifiction of 98 soybean accessions based on selected factors.

Finally, we tried to identify genotypes that possess superiority about most influencing morphological characters on seed yield using cluster analysis. Fig 4 displays the k mean clusters analysis results based on the major morphological characters identified from the best model (random forest) across all the genotypes using f*viz*_cluster function in r. Here, genotypes were made into three final clusters having 49, 2 and 47 genotypes respectively. The seed yield (gram/plant) mean values of each group (Cluster 1: 18.999±0.658, Cluster 2: 45.940± 3.920, Cluster 3: 34.170±3.653) are varying significantly and second group gentotypes showing superirioty in seed yield. CAT-586 and JS-SH-1310 genotypes of second group has superiority of seed yield (45.94 gm/plant), Number of pod per plant (57.00), Plant height at harvest (85.50 cm), Plant height at 30 days (52.10 cm), Plant height at 40 days (66.15 cm) and Root length(12.89 cm) compare to other two groups.

CONCLUSION

Accurate identification of influencing traits to the response is crucial for plant breeding. Advanced models like regression tree, random forestwerefoundtooutperform for variable selection compares to basic techniques like multiple linear regression, principal component regression with high prediction accuracy. The study reveals that the random forest model found as the best modeling approach in the assessment of most contributing morphometric factors for seed yield in soybean germplasm accession. Traits like the number of pods per plant, plant height at harvest, number of branches per, plant height at 30 and plant height at 40 days were noticed as the most influencing factors for seed yield enhancement. Hence considering the nature and magnitude of character association it can be inferred that improvement of seed yield is possible through simultaneous manifestation these above find traits.

REFERENCES

Anonymous (2009). Guidelines for the conduct of test for distinctiveness, uniformity and stability (DUS) on soybean [*Glycine max* (L.) Merrill]. Plant Variety Journal of India. 3(10): 289-98.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters. 13(11): 114003.

Eledum, H. (2016). A comparison study of ridge regression and principle component regression with application. International Journal of Research. 3B(8): 283.

Ghanbari, S., Nooshkam, A., Fakheri, B.A. and Nafiseh M. (2018). Assessment of yield and yield component of soybean genotypes (*Glycine max* L.) in north of Khuzestan. J. Crop Sci. Biotechnol. 21: 435-441.

Gireesh, C., Husain, S.M., Shivakumar, M., Satpute, G.K., Kumawat, G., Arya, M., Agarwal, D.K. and Bhatia, V.S. (2015). Integrating principal component score strategy with power core method for development of core collection in Indian soybean germplasm. Plant Genetic Resources: Characterization and Utilization. 11: 1-9.

Goyal, M. and Verma, U. (2018). Principal component technique for pre-harvest crop yield estimation based on weather input. Advances in Research. pp.1-8.

Hasan, M.M., Yusop, M.R., Ismail, M.R., Mahmood, M., Rahim, H.A. and Latif, M.A. (2015). Performance of yield and yield contributing characteristics of BC2F3 population with addition of blast resistant gene. Ciência e Agrotecnologia. 39(5): 463-476.

Husain, S.M. and Shrivastav, R.N. (2011). Personal communication, Directorate of Soybean Research (ICAR). pp. 1-13.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E. and Kim, S.H. (2016). Random forests for global and regional crop yield predictions. PLoS One. 11(6): 1-9.

Johnston, R., Jones, K. and Manley, D. (2015). Confounding and collinearity in regression analysis: A cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Qual Quant. 52(4): 1957-1976.

Michel, L. and Makowski, D. (2013). Comparison of statistical models for analyzing wheat yield time series. Plos One. 8(10): e78615.

Naveena, K., Singh, S., Rathod S. and Singh, A. (2017). Hybrid time series modelling for forecasting the price of washed coffee (Arabica plantation coffee) in India. Intl. J. of Agri. Sci. 9(10): 4004-4007.

Olivoto, T., de Souza, V.Q., Nardino, M., Carvalho, I.R., Ferrari, M., de Pelegrin, A.J. andSchmidt, D. (2017). Multicollinearity in path analysis: A simple method to reduce its effects. Agronomy Journal. 109(1): 131-142.

Pearl, J. (2000). Causality: Models, Reasoning and Inference. Cambridge University Press, New York.

Roberts, M.J., Noah, O Braun, N.O., Sinclair, T.R., Lobell, B.D. and Wolfram, S. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. Environ. Res. Letters. 12(9): 095010.

Shruthi, K., Siddaraju, R., Naveena, K., Ramanappa, T.M. and Vishwanath, K. (2021). Assessment of variability based on morphometric characteristics in the core set of soybean germplasm accessions. Legume Research. 44(4): 375-381. DOI: 10.18805/LR-4286.

Singh, R.J. and Hymowitz, T. (1999). Soybean genetic resources and crop improvement. Genome. 42: 605-616.

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.

Vu, T.T. H., Le, T.T.C., Vu, D.H., Nguyen, T.T. and Ngoc, T. (2019). Correlations and path coefficients for yield related traits in soybean progenies. Asian Journal of Crop Science. 11(2): 32-39.