Development of machine learning model for studying the seasonality of aphids in wheat based cropping systems of terai zone of Darjeeling, West Bengal

Biwash Gurung[1], Suprakash Pal[2], Md. Wasim Reza[3], Bishal Gurung[4 *], and Achal Lama[5]

[1]School of Agriculture, ITM University, Gwalior-474001, Madhya Pradesh

[2]Directorate of Research (RRS-TZ), Uttar Banga Krishi Viswavidyalaya, Pundibari, Cooch Behar - 736 165, India.

[3]Regional Research Sub-station (Terai Zone) Kharibari, Uttar Banga Krishi Viswavidyalaya, Pundibari, Cooch Behar-736 165, India.

[4]Deaprtment of Statistics, North Eastern Hill University, Shillong-793022

[5]ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

*Corresponding author email: bishalgurung@nehu.ac.in

The primary goal of this research work is to study the effect of weather variables on aphid population and development of weather-based forewarning model by using a powerful machine learning technique called the Random Forest technique. The developed forewarning model could be employed to make proper management strategy by the farming community for controlling aphid infestation.

Keywords: Wheat based cropping systems, Aphid infestation, Forewarning model, Weather parameters, Machine learning, Random Forest

Wheat *Triticum aestivum* (Linnaeus) is a crop belonging to the family Gramineae and is commercially very important crop grown worldwide in wide range of climatic conditions. It is the second most important food crop and contributes about 35% of the total food grain production, thus a major contributor to the agrarian economy of the country (Nagarajan, 2000). Wheat is also a staple food predominantly in the North and North-Western part of India. It is the most important cereal crop in temperate areas of the world and a staple food for more than 35 per cent of the world population covering at least 43 countries and occupies 23 per cent of global cultivated area (Khakwani *et al.*, 2012). According to FAOSTAT (2014), Wheat production has increased from 235 million tons in 1961 to an estimated 733 million tons in 2015. One of the major constraints limiting yield of agricultural products is the insect pests' attacks. Wheat production is regularly suffering from threats actuated by diseases and pest with annual capital loss due to insect pests amounting to around Rs. 413.68 billion. Insect pest are one of these biotic factors and are known to attack wheat worldwide (Hatcheet *et al.*, 1987). In India several insect pests infesting wheat crop have been reported from planting till harvest time (Pal, 1996). Among the various biotic stresses reported on this crop, aphid is one of the most important and destructive pests in West Bengal and many parts of India. Gurung et al. (2018) studied the effect of weather parameters on population dynamics of coccinellids on different crop ecosystems and established the relationship between populations of predatory

coccinellids with abiotic factors. Gurung et al. (2023) have also developed forewarning models based on weather variables for tomato leaf curl infestation by employing Beta regression methodology.

Previously, the problem of insect-pest was not severe in wheat but with fluctuating climate, monocropping and promotion of new crop production technologies like conservation agriculture technologies, minor and random pests are now becoming consistent and major pests of wheat which requires regular monitoring. Keeping this in mind, a statistical model employing machine learning technique is developed for modelling the seasonality of aphids in wheat (*Triticum aestivum* L.) based cropping systems of Terai zone of Darjeeling, West Bengal.

In this manuscript, one of the machine learning techniques has been applied to select important weather variables that is related to aphid population. So, it has resulted in development of prediction models wherein Random forest methodology is employed to model the variability in the pest infestation data. The developed aphid forewarning model could be employed to make proper management strategy by the farming community.

Material and Methods

Meteorological data and study location

Field experiments were conducted at experimental plots in research station farm at the Regional Research Sub-station (Terai Zone) Kharibari, Uttar Banga Krishi Viswavidyalaya, Darjeeling, West Bengal during *Rabi* season of 2018-19 and 2019-20. Geographically speaking, the farm is located at an elevation of 113 m above mean sea level and at 26.55° N latitude and 88.19° E longitude. The area comes under the sub-Himalayan Terai agro climatic zone and the administrative jurisdiction of Darjeeling of West Bengal. Terai zone comprises of entire portion of Jalpaiguri and Cooch Behar districts, Islampur sub-division of North Dinajpur district and Siliguri sub-division of Darjeeling district. The experiment was conducted to detect, categorize and to document the pests and natural enemies associated with wheat as it helps in adopting preventive measures and timely implementation of management tactics.

Weather parameters

The available meteorological data on weather variables viz. rainfall in mm, maximum (RH max), minimum relative humidity (RH min), maximum temperature (T max) and minimum temperature (T min) and their differences (T max – T min) were collected from All

India Coordinated Research Project (AICRP) on Agro-Meteorology, Uttar Banga Krishi Viswavidyalaya, Darjeeling, West Bengal.

Model Description: Random Forest Regression

There are many machine learning (ML) techniques in the literature. One such important ML techniques is the Random Forest (RF), which is an ensemble of decision trees. RF is made up of various trees that are assembled in an unambiguous "random" manner. Each tree is built on a distinct sample of rows, and each node is fragmented up into various set of features. Prediction is obtained from each tree and the predictions so obtained is averaged to produce a single result. In random forest variable selection, which is also called feature selection in machine learning jargon, is carried out by estimating the variable importance of each variable or feature.

The usual method of calculating the variable importance is the average decrease in impurity, which is also called as Gini Importance. Using RF algorithm it is possible to calculate how much each variable or feature decreases the impurity. The less a feature decreases impurity, the less important it is and vice-versa. The final variable importance in random forests is determined by averaging the decrease in impurity from each variable or feature across all the trees.

Accuracy-based importance: For each tree, there is an out-of-bag (OBB) sample of data that was not employed during the development process. The OBB sample is used to decide the importance of a particular feature by checking the OBB sample's prediction accuracy. The shuffling of values of the variable in the out-of-bag-sample is done at random, while all other variables remain unchanged. Then, the decrease in accuracy on the shuffled data will be noted. These values of this measure helps us in calculating the reduction in accuracy when a particular variable is eliminated and, conversely, how the accuracy increases by including a variable. Accuracy is calculated by:

$$Accuracy = \frac{\text{number of correctly identified members of a class}}{\text{Total no. of times the model predicted that class}}$$

Gini-based importance: The Gini impurity criterion is employed in selecting the variable which is to be split at each node when assembling a tree. Every time a variable is selected to split a node, the sum of the Gini reduction over all trees of the forest is accumulated for that particular variable.

The functional form of Gini impurity is given by:

$$GI = 1 - \sum_{i=1}^{n} (p)^2$$

RF are usually employed for regression and discrimination tasks. In discrimination, the goal is to predict the group label of each sample in the dataset. In regression, the goal is to predict the dependent variable (e.g., the yield or infestation of pest/diseases) based on the independent variables of the data. Random forests are widely used because they are easy to train, can be applied for high-dimensional data, and are extremely accurate. They also have the capacity to handle missing values and can be employed for imbalanced datasets. After training the data using random forest algorithm, it can be employed to make predictions. For prediction process, the random forest uses the predictions of each decision trees and combines them by averaging.

Results and Discussion

The feature selection of available meteorological data on weather variables viz. rainfall in mm, maximum (RH max), minimum relative humidity (RH min), maximum temperature (T max) and minimum temperature (T min) and their differences (T max – T min) were carried out using the powerful random forest technique and important variables were selected.

The plot of random forest for 250 number of trees. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.

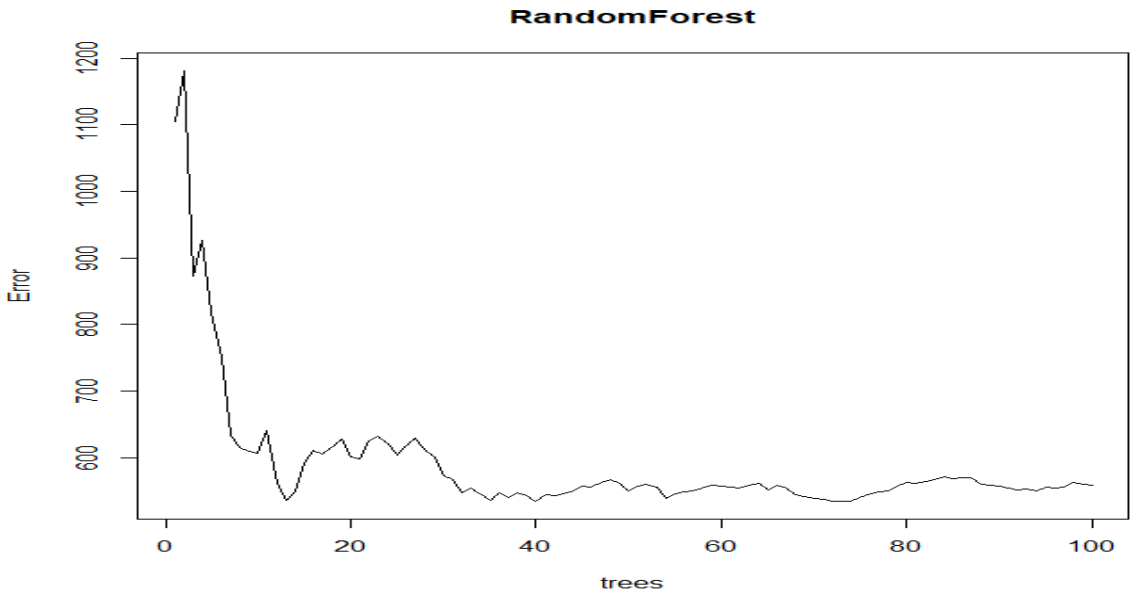


Fig 1. Plot of random forest

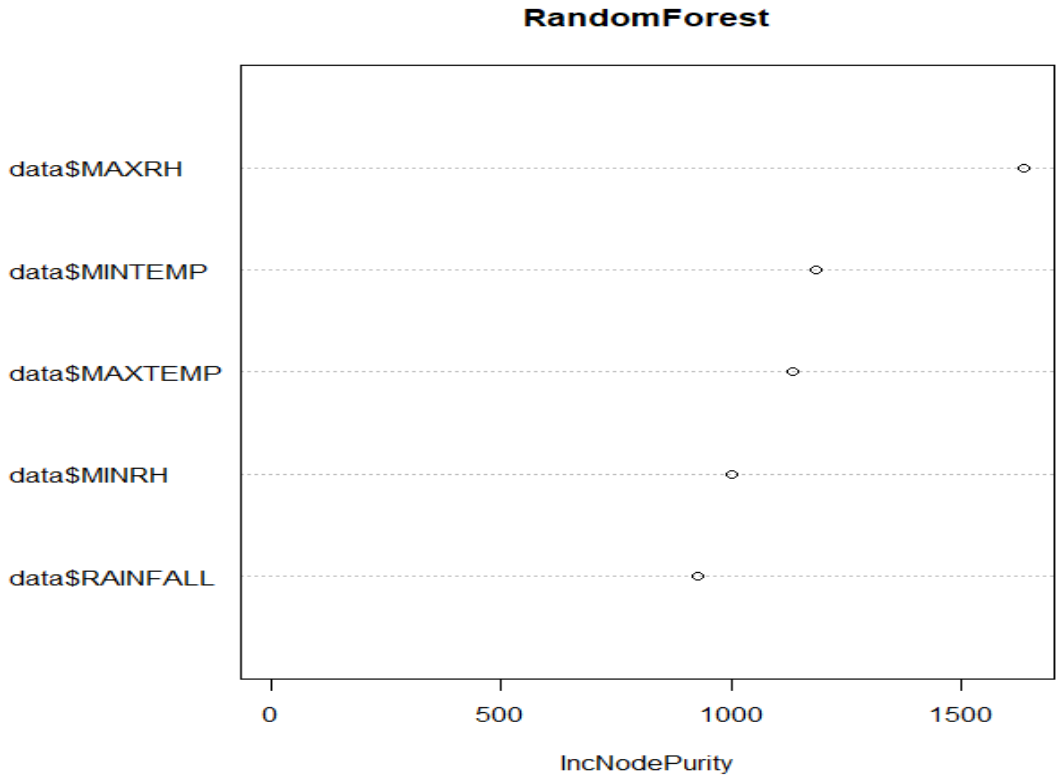


Fig 2. Plot of feature selection using increase in purity importance criterion

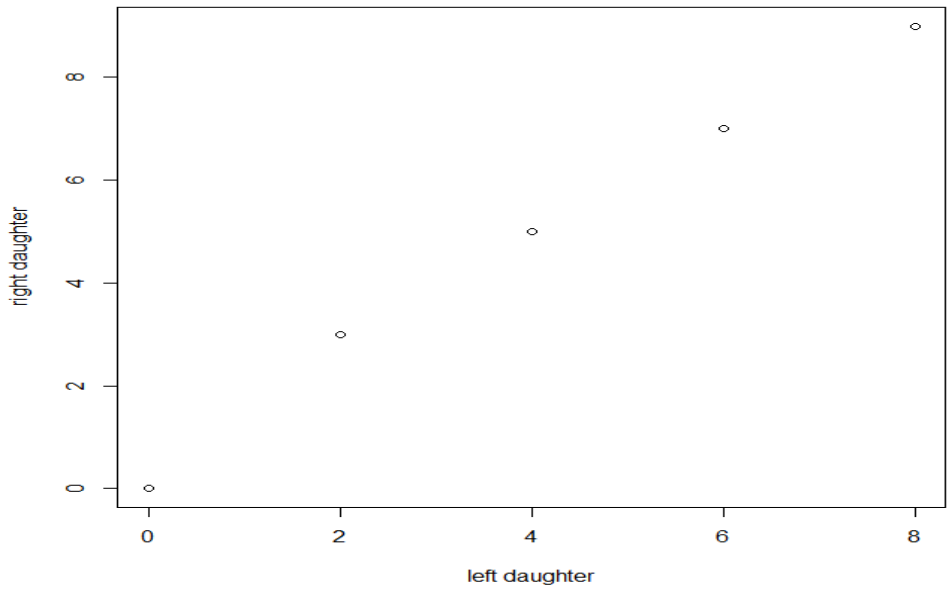


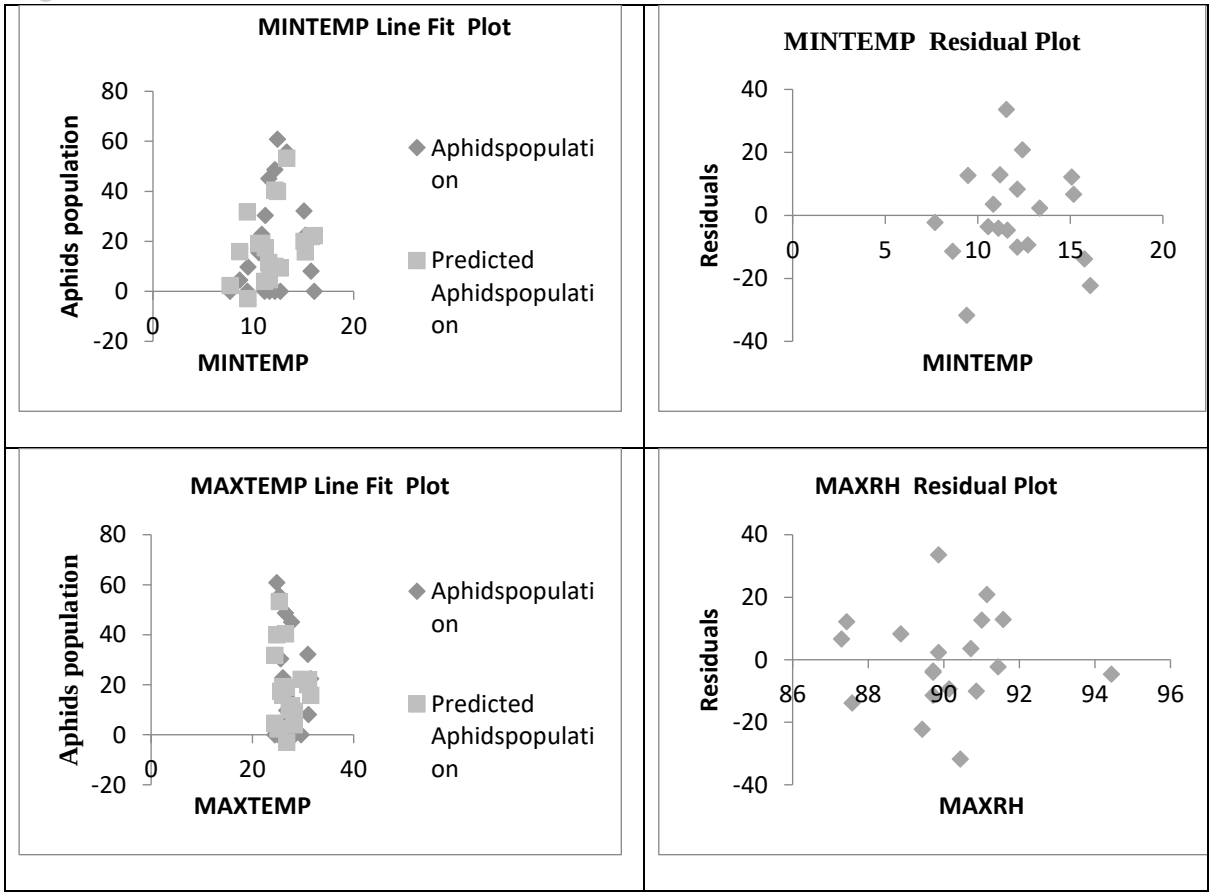
Fig 3. Tree Diagram

From the analyzed data, we find out that Maximum RH, Minimum temperature and Maximum temperature are affecting significantly the aphid population. We have considered both the criteria, %Increase in MSE and Increase in Node Purity, for making the decision on feature selection. After feature selection, the variable selected are employed for building a forecasting model for seasonality of aphids. The fitted model after feature selection by Random forest technique is given below.

Table 1. Fitted model by employing feature selection through Random Forest

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	*Lower 95%*	*Upper 95%*
Intercept	1220.02	425.65	2.86	0.011	312.75	2127.28
MAXRH	-10.48	3.96	-2.64	0.018	-18.93	-2.03
MAXTEMP	-12.39	3.64	-3.40	0.003	-20.14	-4.63
MINTEMP	6.65	2.53	2.62	0.018	1.25	12.04

From the above fitted model we conclude that feature selection has provided statistically significant variables as all three variables selected using random forests are seen to have significant effect on the aphid infestation.



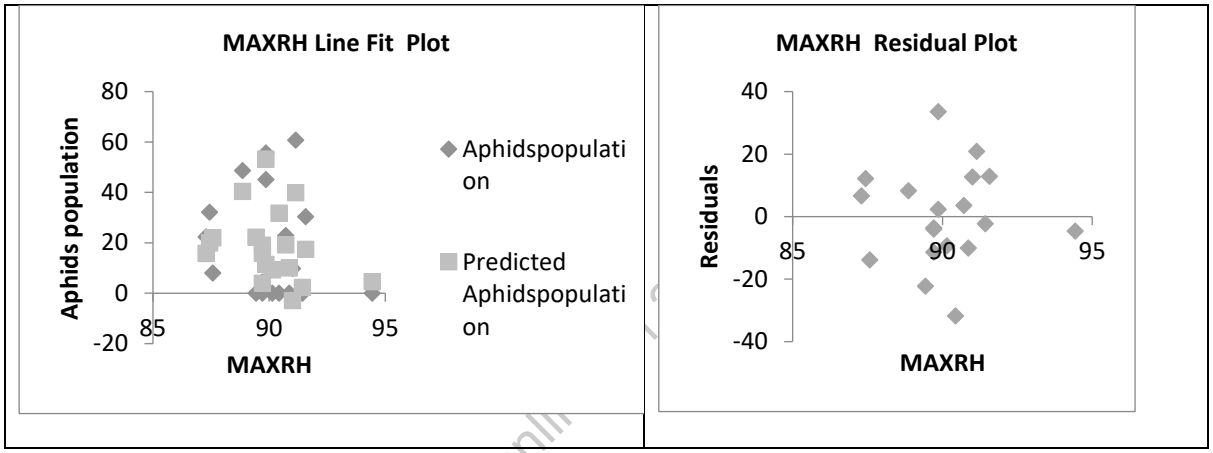
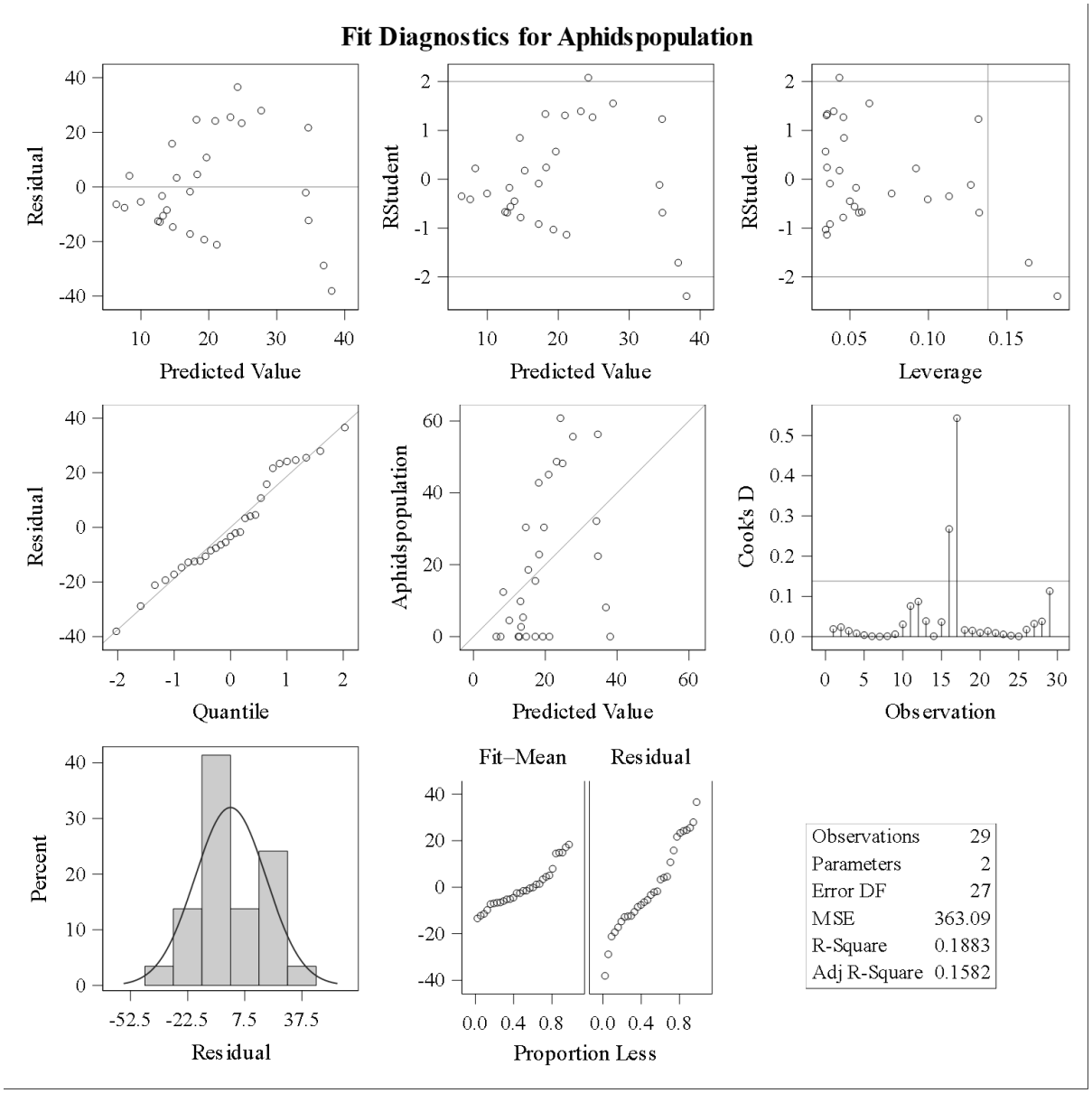


Fig. 2. Line fit plot and Residual Plot of independent weather variables

Further, we have also made comparison with the usual regression methodology with variable selected through stepwise regression. The comparison is made with respect to modeling as well as forecasting performance of the two competing models, viz. RF and stepwise regression. Through stepwise regression, we create a regression model from the available regressor variables by selecting and removing the regressor variables based on the p values of the variables in a stepwise manner till there is no variable left to be selected or removed.

Table 2. Fitted model by employing feature selection through stepwise regression

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	*Lower 95%*	*Upper 95%*
Intercept	-22.71	17.33	-1.31	0.20	-58.28	12.84
MINTEMP	3.78	1.511	2.50	0.01	0.68	6.88



For comparison purpose we have calculated the modelling and forecasting performance of the fitted model. For forecasting purpose we have taken the last five points of the datasets.

Table 3: Comparison of goodness-of-fit performances.

Criterion	Aphid population	
	Stepwise Regression	Random Forest Regression
MAE	15.30	14.83
MSE	336.32	333.17

Table 3: Forecast performance for hold-out data.

Criterion	Aphid population	
	Stepwise Regression	Random Forest Regression
MAPE	25.24	25.10
MSPE	778.15	757.08
RMAPE	147.23	139.26

Further, a study to compare the goodness-of-fit performance of fitted models was carried out using Mean square error (MSE), and Mean absolute error (MAE) criteria and the results are reported in Table 2. The forecasting performance was also calculated using Mean square prediction error (MSPE), Mean absolute prediction error (MAPE) and the results are reported in Table 3. A look at the two tables indicate the superiority of Random Forest Regression over Stepwise Regression for modelling as well as forecasting for the data set under consideration.

Conclusion

The current investigation was conducted for feature selection in aphid population through random forest machine learning technique. As such, it has led to development of statistical model by means of random forest for prediction of aphid population which is an important and destructive pests in various parts of India. A comparison was also carried out to check the improvement of RF over the usual variable selection method from modelling as well as forecasting point of view. Thus, the model developed, can be employed by various plant protection agencies for improvement of management strategies against aphids in Terai zone of Darjeeling, West Bengal. As future work, possibility of employing XGboost, Support vector machine (SVM), Gaussian Process Regression (GPR), balanced repeated replication (BRR) may be explored.

References

Balogun, A. O., Basri, S., Abdulkadir, S. J., & Hashim, A. S. (2019). Performance analysis of feature selection methods in software defect prediction: a search method approach. *Applied Sciences*, **9**(13), 2764.

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.

Breiman, L. (2001). Random forests. Machine learning, **45**, 5-32. DOI:10.1023/A:1010933404324

Gopal, P. M. and Bhargavi, R. (2019). Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches. *Applied Engineering in Agriculture*, **35**, 9-14.

Gurung, B., Dutta, S., Singh, K.N., Lama, A., Vennila, S., and Gurung, B. (2022). Development of weather-based forewarning model for tomato leaf curl infestation. *Journal of Agrometeorology*. **24**(4): 424-426.

Gurung, B., Ponnusamy, N. and Pal, S. (2018). Effect of weather parameters on population dynamics of coccinellids on different crop ecosystems. *Journal of Agrometeorology*, **20**(3): 254-255.

Hatchett, A. H., Stacks, K. J. and Webster, J. A. (1987). Insect and mite pests of wheat. In: E.G. Heyne (ed.) wheat and wheat important. Madison, Wisconsin, USA, pp 625.

Khakwani, A. A., Dennett, M. D., Muni, M. and Abid, M. (2012). Growth and yield response of wheat varieties to water stress at booting and anthesis stages of development. *Pakistan J. Biotech.* **44**, 879-886.

Nagarajan, S. (2000). Wheat production in India a success story and future strategies. *Indian Farming*, **9**, 915.

Oreski, D., Oreskib, S. and Klicek, B. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, **52**, 109-119.

Pal. B. P. (1996). Wheat. Indian Council of Agriculture Research. New Delhi. Pp 244-246.

Whitmire, C. D., Vance, J. M., Rasheed, H. K., Missaoui, A., Rasheed, K. M. and Maier, F. W. (2021). Using Machine Learning and Feature Selection for Alfalfa Yield Prediction. *AI*, **2**, 71-88.

Annexure

SAS Code for stepwise regression

```
data regression;
input RAINFALL    MAXRH MINRH MAXTEMP     MINTEMP     Aphidspopulation;
cards;
;
.
.
.

ods rtf;


proc reg;

model Aphidspopulation = RAINFALL   MAXRH MINRH MAXTEMP
      MINTEMP/selection=STEPWISE;

run;

ods rtf close;
```

R Code for random forest variable selection

```
install.package("randomForest")

library("randomForest")

data=read.csv(file.choose())

RandomForest=randomForest(data$Aphidspopulation~data$RAINFALL+data$MAXRH+dat
        a$MINRH+data$MAXTEMP+data$MINTEMP, data, ntree=1000,
        keep.forest=TRUE, importance=TRUE)

plot(RandomForest)

importance(RandomForest)

varImpPlot(RandomForest, sort=TRUE, n.var=min(30,
        nrow(RandomForest$importance)),type=2, class=NULL, scale=TRUE,
        main=deparse(substitute(RandomForest)))

getTree(RandomForest,k=10,labelVar=FALSE)
```