https://doi.org/10.56093/ijas.v90i6.1047786

Indian Journal of Agricultural Sciences **90** (6): 1140–4, June 2020/Article

Qualitative analysis of random forests for evaporation prediction in Indian Regions

RAKHEE*, ARCHANA SINGH, MAMTA MITTAL and AMRENDER KUMAR

Amity School of Engineering and Technology, Noida, Uttar Pradesh 201 303, India

Received: 21 May 2019; Accepted: 17 September 2019

ABSTRACT

The performance of logistic regression, discriminant analysis, and random forest has been compared for the prediction of evaporation of different regions of India during 2019 at ICAR-IARI, New Delhi . The present experiment was performed at Raipur (Chhattisgarh), Karnal (Haryana), Pattambi (Kerala) and Anantpur (Andhra Pradesh). Evaporation and other weather parameters are collected from the year 1985-2012, 1973-2005, 1991-2005 and 1958-2010 respectively. The performance of the techniques is compared using classification, misclassification, and sensitivity of the model along with the Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC) value. The combinations of variables as independent variables are used in two sets. In the first set, maximum & minimum temperature, relative humidity morning & evening, wind speed, rainfall, and bright sunshine hours are used. In the second set mean temperature, mean relative humidity, bright sunshine hours, and wind speed is used to see the effect on evaporation. It is found that more accuracy is obtained using the second set as predictors. The model validation accuracy is checked via running developed model on out of sample data, i.e. testing data (last three years). The study demonstrates that the random forest approach predict evaporation in a much better way than logistic regression, discriminant analysis. The random forest model can provide timely information for the decision-makers to make crucial decisions impacting due to evaporation conditions in India.

Key words: Discriminant analysis, Logistic regression, Random forest, Sensitivity

Weather is an essential factor that influences the crop yield and pest infestation in the field; underestimating the impacts of weather parameters on agricultural production could lead to severe complacency about the potential changes happening across the country. Among different weather variables, evaporation is being influenced by other parameters and stands to be of utmost importance to monitor the crop water requirement. In this view, there is a need to develop methodology that can predict the evaporation for the management of resources. Many notable classification and prediction techniques (Agrawal *et al*. 2007 and Gang *et al.* 2006) have been adapted by researchers to classify the data and predict future outcomes. Much attention has been received on determining pest dynamics by prediction models (Kumar *et al.* 2012 and Kumar *et al*. 2013). Some preliminary work is carried out focusing on logistic regression and discriminant analysis for classification purpose (Zibaei and Bakhshoodeh 2008 and Bhowmik 2009). Ordinal logistic method was employed to identify environmental factors associated with mating flight (Kim *et al*. 2009). Another Forecast model was developed for wheat yield in the Kanpur district using a discriminant function analysis technique that provided a reliable yield forecast about two months before harvest (Agrawal *et al*. 2012). Random forest and logistic regressions are applied to vast dataset for the prediction analysis and to provide insight to the farmers to plant which crop to their farm field (Sahu *et al.* 2017 and Manuel *et al*. 2019). Researchers have applied other data mining techniques such as support vector machines (Baydaroglu *et al*. 2014), artificial neural networks (Benzaghta *et al*. 2012) to predict evaporation and other weather parameters. They have estimated the daily evaporation from humid environments. Prediction of evaporation using M5 model tree algorithm technique is also explained (Deswal *et al.* 2008). This study aims to perform the prediction of evaporation for different regions of India. Qualitative classification is illustrated using logistic regression, discriminant analysis, and random forest. Timely prediction of evaporation may provide a time frame for the farmers and researchers to take care of the future needs of crop fields.

MATERIALS AND METHODS

In this study, evaporation, along with other weather variables from different locations of India, is considered for the analysis of three classification techniques. Weekly meteorological data are collected for different locations,

*Corresponding author e-mail: rakheesharma234@gmail.com

viz. Raipur (21.25° N, 81.62° E) during 1985-2012; Karnal (29.68° N, 76.99° E) during year 1973-2005; Pattambi (10.80° N, 76.19° E) during year 1971-2005 and Anantpur (14.6819° N, 77.6006° E) during 1985-2010. Weather variables in two independent sets are examined: First set consists of maximum and minimum temperature, relative humidity in morning and evening, wind speed, rainfall and bright sunshine hours; second set consist mean temperature, mean relative humidity, bright sunshine hours and wind speed.

Logistic regression: Logistic regression is widely used when it comes to classification problems. It provides the probability of observation to be part of a particular class or not. The probabilities are in the range of 0 to 1. A probability close to 1 means it is likely to be part of that category. To generate the probability for dependable variable (evaporation) following equation is used.

$$p\left(Y=\frac{1}{X}\right)=\pi(\mathrm{X})=\frac{e^{\beta' x}}{1+e^{\beta' x'}}$$

where $\beta' X = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$ and n is the number of independent variables. The formula also shows that $\pi(X)$ changes as an S-shaped function of independent variables. The probability distribution of dependent variable and the log of likelihood function is also explained (Arno De *et al.* 2018).

Discriminant analysis: Discriminant analysis may be a variable technique involved with separating distinct sets of objects (or sets of observations) and allocating new objects (or representations) to the antecedently outlined groups. As a grouping procedure, it's typically used to research ascertained variations once causative relationships aren't well understood. In building the discriminant analysis-based prediction model, a large number of training samples are needed to understand the relationship between dependent variables and independent variables. In this paper, weekly weather data (1-52 weeks) were used to create training samples. Then, discriminant functions were built to establish the relationship between dependent and independent variables, respectively. The basic form of the function is presented as:

$$D = \beta_1(X_1) + \beta_2(X_2) + \beta_a(X_a) + ... + C$$

Here, D is the discriminant function score, β is the function coefficient, and C is the intercept, whereas X is the value of the independent variable.

Random forests: Random forests (Breiman 2001) are learning algorithm which can be used for developing classification as well as regression models. It can predict categorical variables and also involves predicting a continuous response variable, respectively. The models created through random regression and classification of forests fit a set of data into the utmost decision tree models. The data are divided recursively into more homogeneous units for each tree; these units are called nodes, which enhance the response variable's predictability. The split points on the nodes are depended on the value of the predictor variable of the data. While growing trees, random forest contributes randomness to the model. It searches for the highest function among a random subset of characteristics instead of looking for the most important function while dividing a node. This results in a wide diversity that generally results in a better model. In this paper, the random forest classification model is used for the two-class response variable of evaporation for four different regions across India. These models were developed using the "random forest" package written in R statistical software (http://www.R-project.org version 3.6). The random forest was trained using different weather variables affecting evaporation. There are three parameters for random forest which are essential while training the dataset:

- 'ntree'- number of trees in the forest
- 'mtry'- The number of distinct descriptors tested for each division and
- node size- The minimum size of the node below which leaves are not subdivided further.

The correct value of training parameters was selected between the ranges node size= from 1 to 50 unit, ntree= from 1 to 5000 unit mtry= from 1 to 126 respectively (Palmer *et al.* 2007). Hence, for this dataset what we have observed here in this paper, the optimum value is reached at, ntree=20, mtry=2, and node size=14, respectively. The sample used for tree growth in the random forest is selected, leaving the last three years from each location. The subsequent years are not used for tree growth are termed as out of bag samples and considered for validation purposes. Since we use mtry variables at each node, the complexity to build one tree would be O (mtry*nlog(n)). For building a random forest with 'ntree' number of trees, the complexity will be O (ntree*mtry*nlog(n)).

For evaluation of the performance of the techniques, Sensitivity is calculated for the correct classification of evaporation. It represents the percentage of events of dependent variables successfully predicted as events. The statistics are provided with Confusion matrix, which gives information about the actual and predicted classifications done by a classification system. The Performance of the system is commonly evaluated using the data in the matrix. The confusion matrix is represented as True negatives (the sample are correctly classified as negative); false negatives (samples are incorrectly classified as negatives); false positives (sample are incorrectly classified as positives); true positives (samples are correctly classified as positive). The analysis has been done by using SAS (Statistical Software System) version 9.3 software for performing logistic and deterministic approaches and R Studio version 3.6 for random forest technique.

RESULTS AND DISCUSSION

The purpose of this paper is to determine the analysis between three well-known techniques for the prediction of evaporation covering four different regions of India. For the same purpose, other weather variables, viz. mean temperature, mean relative humidity, wind speed, and

bright sunshine hours are kept as independent variables. The performance of the model is evaluated by dividing all data training set and validation set. The data of the validation set, are not included while developing models for prediction, it is used only for model evaluation. Subsequent years for locations include Raipur (2010-2012), Karnal (2003-2005), Pattambi (2002, 2004-2005), Anantpur (2008-2010) respectively. When the techniques were attempted to classify evaporation, two independent combinations of variables were considered, these sets are set-I (Maximum and Minimum Temperature, Maximum and Minimum Relative Humidity, Bright Sunshine Hours, Rainfall, Wind speed) and set-II (Mean Temperature, Mean Relative Humidity, Bright Sunshine Hours and Wind speed). The classification accuracy with both sets was compared and is illustrated (Table 1), which provides the classification and misclassification percentage of random forest using both variables set for training and validation data. It is evident from the table that using set-II, viz. Mean Temperature, Mean Relative Humidity, Bright Sunshine Hours, and Wind speed as predictor gives more accuracy to classify the data. It can be observed that for training set as well as validation only set II is providing high classification percentages for all the considered locations. Moreover, we have also attempted the same experiment with logistic and discriminant techniques and got high accuracy with set-II. Thus, for all the three methods we have considered the same independent weather variable (Mean Temperature, Mean Relative Humidity, Bright Sunshine Hours, and Wind speed) for the model development. In this study, logistic regression, discriminant analysis, and random forest techniques were utilized for the classification of data into two groups and are used for qualitative forecasting.

The first technique, i.e. logistic regression, describes the effect of predictor variables on a categorical variable. It is used for prediction of the probability of occurrence of an event by fitting data to a logit function. Discriminant function analysis is used to determine which variables discriminant between two or more naturally occurring groups. This technique makes use of the information provided by the X variables to achieve the most precise possible separation between two groups (in our case, the two groups are high and low evaporation). Random Forests have been used widely to predict agricultural data and have been incorporated into many climate and crop yield forecasting. Random forest can be used extensively for classification purposes; it can predict a categorical response variable. These techniques are outlined to develop models and to see the impact of other weather variables for classifying evaporation.

Model validation: To evaluate the performance of the models, all data are divided into training set and validation set. The data of validation set (last three years weekly weather data) are not included while developing the models for prediction, it is used for model evaluation. Subsequent years included for Raipur (2010-2012), Karnal (2003-2005), Pattambi (2002, 2004-2005) and Anantpur (2008-2010) respectively. The developed models for all the three techniques are validated by running the model over out of sample data. The percentage of classified and misclassified data for validation set was obtained (Table 2). This clearly suggested that the random forest model has greater accuracy. The qualitative comparison also shows that, the number of misclassification is minimum for random forest approach. Thus, in forthcoming term the reliable prediction of evaporation can be obtained using random forest over these other two techniques. It can also be seen from the table that misclassification in Anantpur region is slightly higher as compared to other considered location. It may happen due to uneven or very low rainfall received

Table 1 Training of model using Random Forest Classification using different set of weather variables

Location	Using Set-I*		Using Set-II*		Using Set-I*		Using Set-II*	
	Training Set				Validation Set			
	Classified	Misclassified	Classified	Misclassified	Classified	Misclassified	Classified	Misclassified
Raipur	92.47	7.53	95.23	4.77	94.23	5.77	95.69	4.31
Karnal	93.42	6.58	94.89	5.11	95.32	4.68	96.39	3.61
Pattambi	91.88	8.12	93.52	6.48	91.54	8.46	92.67	7.33
Anantpur	85.23	14.77	85.79	14.21	88.84	11.16	89.35	10.65

* Set-I: Maximum & Minimum Temperature, Maximum & Minimum Relative Humidity, Bright Sunshine Hours, Rainfall, Wind speed. Set-II: Mean Temperature, Mean Relative Humidity, Bright Sunshine Hours and Wind speed.

Table 2 Percentage of classified and misclassified of training data using three techniques

Location	Year of validation	Number of data points	Discriminant analysis		Logistic regression		Random forest	
			Classified	Misclassified	Classified	Misclassified	Classified	Misclassified
Raipur	2010-12	1275	94.98	5.02	92.21	7.79	95.23	4.77
Karnal	2003-05	1402	90.15	9.85	90.01	9.99	94.89	5.11
Pattambi	2002, 2004-05	1441	89.79	10.21	89.38	10.62	93.52	6.48
Anantpur	2008-10	1196	78.67	21.33	73.98	26.02	85.79	14.21

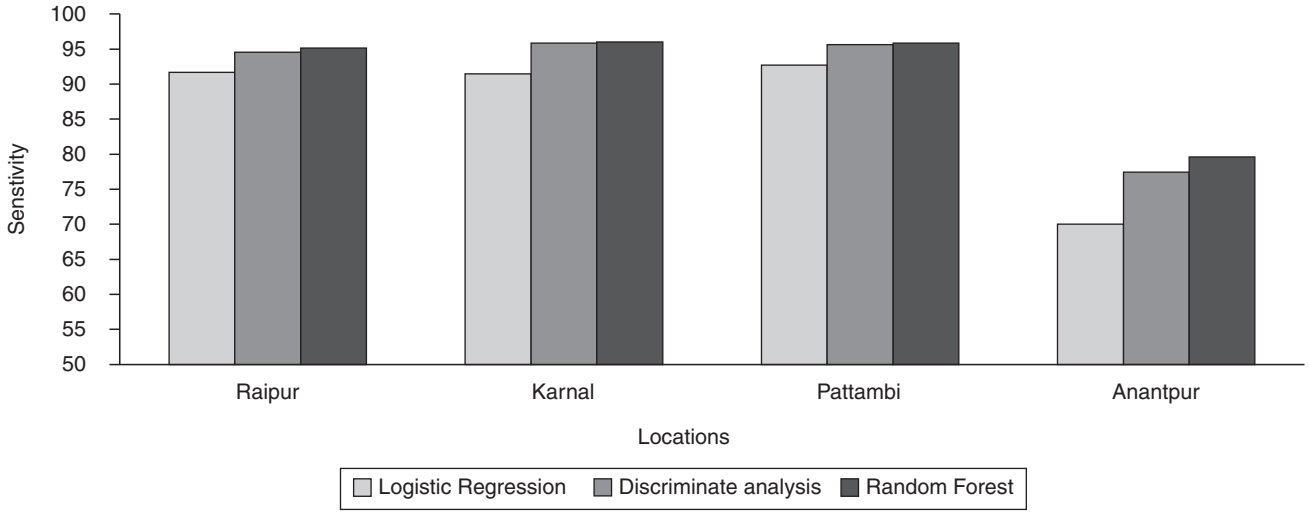

Fig 1 Sensitivity of the developed models constructed for all four location using three different techniques.

by this region (Naveen *et. al*. 1991). Also during the crop growing season the evaporation valued varied between 8.2mm/day (July) to 5.2 mm/day (October).

Sensitivity analysis, ROC and AUC: Sensitivity analysis is performed to access the fitness of the developed model and furthermore to appreciate the strength of outcome drawn from such models. The results of sensitivity, scrutinize the relationship between the inputs and outputs of the model. The rationalization for sensitivity analysis is that if tested on the dataset from which it was obtained, a model will always perform better. The sensitivity analysis of the models evaluated using all three techniques. It can be observed (Fig 1) that the random forest model shows the highest sensitivity in all locations.

ROC curve is the ratio between sensitivity and 1-specificity at different threshold values. The ROC curve (Fig 2) in random forest for Karnal region is represented, for other regions also similar results were obtained. Sensitivity also known as true positive rate can be derived as True Positive/True Positive + False Negative whereas specificity is true negative rate derived as True Negative/True Negative + False Positive. These values can be calculated from a confusion matrix. The graphical plot obtained explains the ability of a classifier. For the performance measurement, AUC is obtained, which provides the degree of separability, i.e. how much the model is capable of distinguishing between two classes. Higher AUC represents better predictability of the model. It was observed from the experiment that random forest AUC, i.e. 0.95, is higher than logistic and discriminant analysis, i.e. 0.77 and 0.92 respectively.

A comparison of logistic regression and discriminant analysis with random forest shows that the later technique provides more accuracy while predicting evaporation treating other weather variables as predictors for all

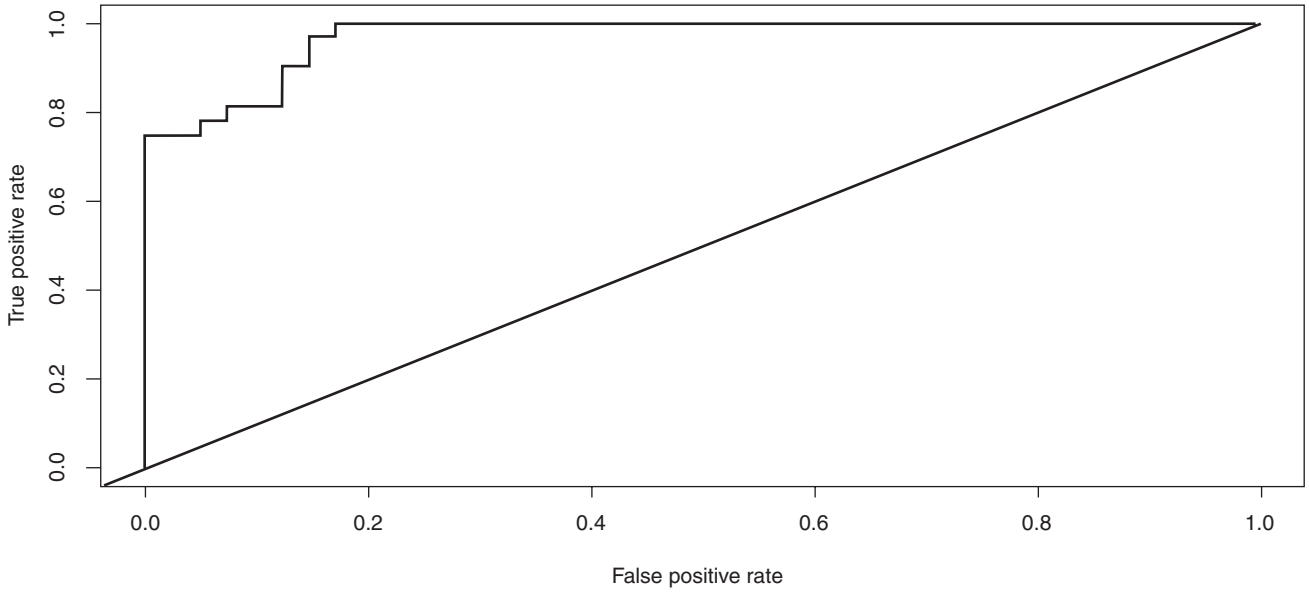

Fig 2 ROC curve for Karnal location using Random Forests approach.

considered locations. In this study, the validation of models is also performed for three consecutive years, including Raipur (2010-2012), Karnal (2003-2005), Pattambi (2002, 2004-2005), Anantpur (2008-2010) respectively. The validation of the model developed using the variable set (mean temperature, mean relative humidity, bright sunshine hours, and wind speed) as independent variables have the highest classification rate. Random forest performance dominates over logistic and discriminate while validating the data in terms of accuracy and sensitivity of the models. Sensitivity analysis has also been performed to show the model performance and appropriateness of the developed model.

ACKNOWLEDGMENTS

The work would not have been feasible without the data. Authors are grateful to Agricultural Knowledge Management Unit, ICAR-IARI, New Delhi for providing us data from different Indian places.

REFERENCES

Agrawal R and Mehta S C. 2007. Weather based forecasting of crop yields, pests and diseases - IASRI models. *Journal of the Indian Society of Agricultural Statistics* **61**(2): 255–63.

Agrawal R, Chandrahas and Aditya K. 2012. Use of discriminant function analysis for forecasting crop yield. *Mausam* **36**(3): 455–58.

Arno D C, Kristof C and Bock K W. 2018. A new Hybrid Classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research* **269**(2): 760–72.

Breiman L. 2001. Random Forests. *Machine Learning* **45**(2): 5–32.

Bhowmik A. 2009. 'A study on logistic regression modeling for classification in agriculture'. MSc thesis, ICAR-Indian Agricultural Research Institute, New Delhi.

Baydaroglu O and Kocak K. 2014. SVR-based prediction of evaporation combined with chaotic approach. *Journal of Hydrology* **508**(16): 356–63.

Benzaghta M A, Mohammed T A, Ghazali A H, Mohd A and Mohd S. 2012. Prediction of evaporation in tropical climate using artificial neural network and climate based models. *Scientific Research and Essays* **7**(36): 3133–48.

Deswal S. 2008. Modeling of evaporation using M5 model tree algorithm. *Journal of Agrometeorology* **10**(1): 33–38.

Gang C, Haiguang W and Zhanhong M. 2006. Forecasting wheat stripe rust by discrimination analysis. *Plant Protection* **32**(4): 24–27.

Kumar V, Kumar A and Chattopadhyay C. 2012. Design and implementation of web-based aphid (*Lipaphis erysimi*) forecast system for oilseed Brassicas. *Indian Journal of Agricultural Sciences* **82**(7): 608–14.

Kumar A, Agrawal R and Chattopadhyay C. 2013. Weather based forecast models for diseases in mustard crop. *Mausam* **64**(4): 663–70.

Kim H, Li J and Wang S. 2009. *Ordinal logistic regression modelling to predict mating flights through meteorological cues.* Texas A&M University, College Station, Texas.

Naveen P and Seetharaman N. 1991. 'An anlaysis of Anantpur climate, drought research seminar forum'. MSc thesis, The International Crops Research Institute for the Semi-Arid Tropics, Telangana.

Manuel D P, Angel C O, Jose A S A and Callejon A S. 2019. Logistic regression to evaluate the marketability of pepper cultivars. *Agronomy* **9**(3): 125–43.

Palmer D S, Boyle N M, Glen R C and Mitchell J B O. 2007. Random Forest models predict aqueous solubility. *Journal of Chemical Information and Modeling* **47**(1): 150–58.

R Development Core Team. R: A language and environment for statistical computing, URL http://www.R-project.org (accessed Sept 06, 2019).

Sahu S, Chawla M and Khare N. 2017. An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. *(In) Proceeding of International Conference Computing, Communication and Automation (ICCCA),* Greater Noida, India, May 5-6, pp. 53–57.

Zibaei M and Bakhshoodeh M. 2008. Investigating determinants of sprinkler irrigation technology discontinuance in Iran: Comparison of logistic regression and discriminant analysis. *Journal of Agricultural and Environmental Sciences* **22**(5): 46–55.