A Comparative Study on Decision Tree Methods in Predicting Daily Reference Crop Evapotranspiration

N.Manikumari
Department of Civil Engineering
Faculty of Engineering & Technology, Annamalai University

Abstract- *Evapotranspiration plays a significant role in the water and energy balance on earth's surface and it is of specific significance to agricultural and irrigation methods. Reference Evapotranspiration (ETo) rates for irrigation system is estimated using the FAO56 Penman-Monteith (P-M) method using various climatic parameters such as temperature, solar radiation, wind speed, and relative humidity. In this study, decision tree (DT) models are applied in prediction of ETo. Various kinds of decision trees such as REP Tree, Random Tree, Random Forest and M5P are employed in ETo prediction. The various decision tree models employed in this study are analyzed using the WEKA tool. The prediction performance of the DT methods is presented in terms of correlation coefficient. The results show that M5P Tree and random forest are the best in terms of correlation coefficient.*

Keywords- Decision tree; irrigation; evapotranspiration.

I. INTRODUCTION

Reference Evapotranspiration (ETo) is defined as the rate at which readily available soil water is vaporized from specified vegetated surfaces. In order to analyze the evaporative demand, independent of crop type, the need to measure ETo arises. The ETo values calculated irrespective of locations and seasons are analogous as they refer to the evapotranspiration from the unchanged reference surface. The factors affecting ETo includes various climatic parameters [1,2]. Many empirical methods have been investigated to estimate evapotranspiration from various climatic variables. P-M equation is one such widely accepted method to estimate evaporation from solil, water and grass [4,6].

In the past decade, machine learning methods have got wide focus due to its importance of prediction models. The comparison of prediction models is a complex and open problem. The performance of the prediction models are evaluated using various metrics such as accuracy, speed, cost, reliability etc. In order to measure such performance metrics various machine learning tools are available. The model predicted values are then compared with the estimated values. The selection of the optimum prediction algorithm for a particular dataset is an extensive problem. In this logic it is essential to make a number of methodological choices. Among the various methodological choices, this study focuses on the decision tree algorithms for prediction of ETo [3,5,7].

The decision tree method is a dominant statistical prediction tool that has numerous prospective real time applications. Decision trees can be adapted for a diverse range of applications. The DT methodology is also easy for the programmers to implement and interpret the results of DT models. DT methods can also be used to diagnose mechanical failures and troubleshooting. DT also find its application in various business level applications. Higher level officials use decision tree methods to analyze the company's decision-making process.

II. BACKGROUND

Decision Trees are considered to be one of the most accepted method for prediction problems. Many researchers from different research areas have employed decision tree models for prediction. A brief survey of existing methods in prediction of ETo using decision tree classifiers is discussed in this section. Rahimikhoob, A. (2014) performed an analysis on the performance of Artificial Neural Network (ANN) method and M5 model tree for prediction of ETo. He employed feed forward neural network approach at various meteorological sites. The input variables chosen were the maximum air temperature and minimum air temperature, air humidity and extraterrestrial radiation. The performance of the ANN and M5 model in prediction of ETo was assesed using P-M method as a reference model. The results showed that ANN estimated ETo superior than the M5 model tree. Root mean square error and Correlation coefficient for ANN model are 5.6 % and 0.98 respectively. RMSE and R^2 for M5 model are 5.6 % and 0.98 respectively..

Prediction of ETo by modelling is vital in reservoir management, planning regional water resources and evaluation of drinking water supplies. Kisi, O. (2016) analysed the effect of various regression methods such as Least Square Support Vector Regression, Multivariate Adaptive Regression Splines and M5 Model Tree in modeling of ETo. The results showed that the M5Tree models performed better than other

models with respect to various performance metrics used. Gholami, A. (2016) studied the use of machine learning methods in diverse hydraulic sciences. He investigated two ANN models such as Multi Layer Perceptron (MLP) and Radial Basis Function (RBF) with decision trees (DT). He designed two hybrid models, namely decision tree based multilayer perceptron (DT-MLP) and decision tree based radial basis function (DT-RBF). The performance of the DT-MLP and DT-RBF is better when compared with MLP and RBF models.

Goyal, M. K. & Ojha, C. S. P. (2013) analyzed the effect of multiple decision tree methods such as single conjunctive rule learner, decision stump, M5P model tree, decision table and REP Tree. In his modelling approach, various parameters such as air temperature, zonal wind, meridional wind and geo-potential height were used as predictor variables. The investigation is carried out using data set for the period 1948–2000 collected from the National Centers for Environmental Prediction (NCEP). The results obtained on several DT methods was evaluated using various performance metrics. The M5P model tree algorithm was found to be better in performance than other DT methods.

Kisi, O. et al. (2015) investigated the effect of Artificial Neural Networks and M5 model tree in modeling ETo. The analysis was carried out using data collected from six different stations operated by California Irrigation Management Information System. The papremeters used in thie analysis were daily climatic data, average temperature, solar radiation, wind speed, and relative humidity The ANN and M5 Tree models were found to be better in predicting ETo of six stations compared to empirical methods.

III. DECISION TREES

Decision tree (DT) methodology is a frequently employed data mining approach for prediction of a target variable. The DT method predicts by forming a tree like structure with branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The DT method is a non-parametric approach. DT methods can effectively handle huge complex datasets. A root node, is the decision node. The decision node decides the subdivision of all data samples into mutually exclusive subgroups. Internal nodes represents the potential choices available at that node in the tree. The internalt node is connected to its parent node and its child nodes. Leaf nodes in the tree structure represent the final result of a combination of decisions. Branches represent chance outcomes from root nodes and internal nodes.

a. Reduced Error Pruning Tree (REPTree)

Reduced Error Pruning (REP Tree) is a procedure that reduces the inappropriate branches from the decision tree (DT). It is a post pruning technique in which the DT structure is first constructed. After constructing, each leaf is detached one by one and then accuracy is measured at every step using the bottom up approach. After removing ther leaves of the particular branch, the accuracy is remeasured. If there is no degradation in the accuracy, the branch is removed permanently. REP Tree uses the regression tree logic and creates multiple trees in different iterations. After that, it selects the best one from all generated trees. REP Tree is a rapid decision tree learner which constructss a regression tree using information gain (IG) as the splitting principle.

b. Random Tree (RT)

Random Tree algorithm learns by using multiple individual learners. Random tree makes use of a random subset of data for building a decision tree. Random trees is a collection of decision tree methods called as forest. In this prediciton, the random tree method response is the average of the responses over all the decision trees in the forest. Random Trees are fundamentally the grouping of existing decision tree algorithms in machine learning. The training data is sampled to form subsets with replacement for each single decision tree employed. Then, when building a tree, the optimum probable split for each node for random subset of data samples is considered at every node.

c. Random Forest

Random Forest is also a collection of simple tree predictors. Each predictor tree is capable of producing a response when presented with a set of predictor values. A Random Forest consists of an random number of simple trees, which are used to determine the final outcome [3]. The predictor set is randomly selected from the same distribution and for all trees. Given the above, the mean-square error for a Random Forest is given by:

$$mean\ error = (observed - tree\ response)^2$$

d. M5P Tree

M5P tree is used for building trees of regression models. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. M5P is a decision tree induction algorithm is used to build a tree. A splitting condition is used that minimizes the intra subset variation in the class values in top down manner in each branch [5]. The splitting procedure in M5P stops if the class values of all instances that reach a node vary very slightly, or

only a few instances remain. Then the tree is pruned in bottom up manner from leaf node. A smoothing method is applied that combines the leaf model prediction with each node along the path to the root.

IV. MODELING

In this study, daily data on maximum temperature, minimum temperature, maximum humidity, minimum humidity, actual bright sunshine hours and wind speed observed at Indian Meteorological Observatory, Annamalainagar was collected for computing Daily ETo. Dataset was developed for the period 1997 -2007. The data set consists of measured parameter values such a Maximum Relative Humidity (RHMX),Minimum Relative Humidity (RHMN) ,Wind Speed (WS), Maximum Temperature (TMAX), Minimum Temperature (TMIN), Sunshine Hours (SSH) and Reference Evapotranspiration (ETo in mm) for all 365 days in a year (366 days in case of leap year).The detailed description of the dataset used is given in Table 1.

Table1. Description of dataset

Properties	Value
Number of instances	3651
Number of attributes	7(6 + 1)
Period of analysis	1997-2007
Training instances	2556
Test instances	1095
Dependent variable	ETo
Independent variables	RHMAX,RHMIN,WS,SSH,TMAX,TMIN

All data mining computations were done with the free data mining software WEKA. The methods included variations of decision trees like model trees, REP Tree, random tree, random forest and M5P tree. The total size of the tree obtained by REP Tree method is 79. Fig. 1 shows the sub tree of the tree model generated by REP Tree method. The Sub tree shown depicts that the attribute SSH is the deciding (root) node with attribute value used for comparison is 15.6. If the attribute value is greater than 15.6. Then the SSH value is checked if it is greater than 22.25. The node number of SSH is 59. This proceeds in top down approach till a ETo value is predicted.

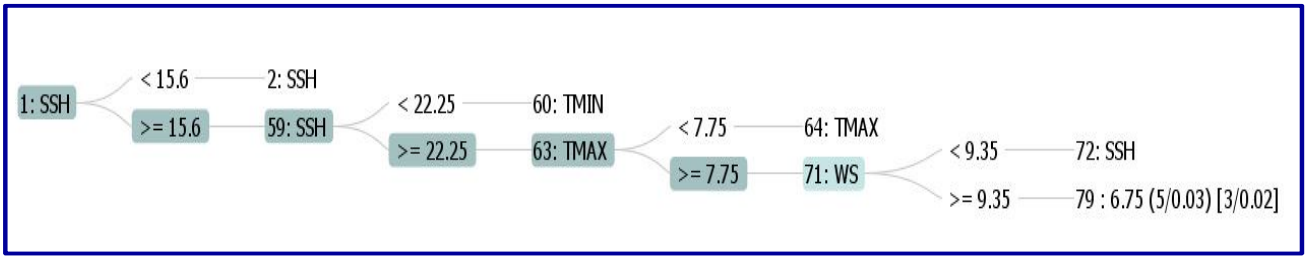

Fig. 1. Subtree of REP Tree Model

A total of 20 subtrees are generated in M5P model. Fig. 2 shows the subtree model obtained for ETo prediction by M5P tree method. The M5P subtree in Fig 2 is constructed only by using three attributes ie. TMAX, TMIN and SSH values. The subtree shown in Fig 2 is formed by 3 linear models (LM1, LM2 and LM3). In the LM nodes, the first number in parentheses is the number of instances in that branch. The second number is the root mean squared error of the predictions of the model at that node. If the value of TMIN is less than 2.65, then we have to follow LM1 to predict the ET0 , else we have to follow the other linear models.

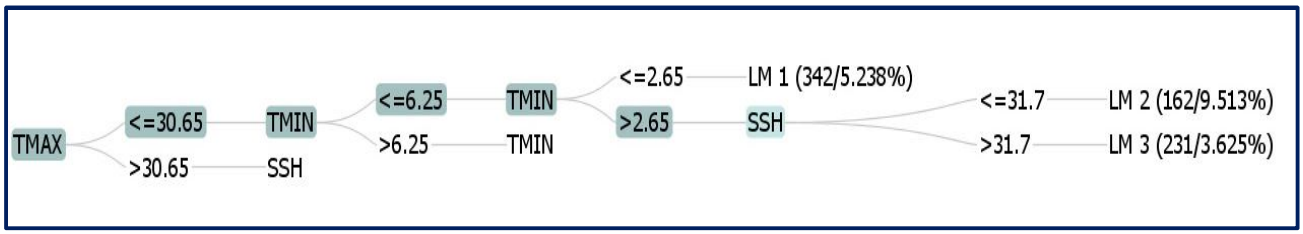

Fig. 2. Subtree of M5P Tree Model

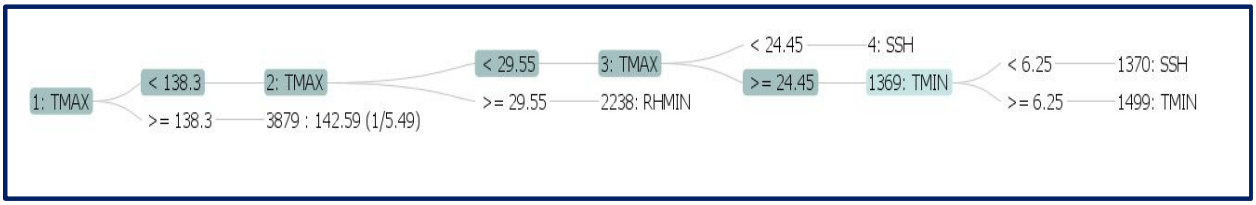

Fig. 3. Subtree of Random Tree Model

Fig. 3 shows the subtree generated by random tree method in Weka. The root node is TMAX with branch value of 138.3. The child node in the subtree are TMAX, TMIN, SSH and RHMX. In random forest implementation 100 trees are generated for prediction.

The scatter plot shown in Figures 4 – 7 reflect the strong relationship between actual and predicted values. The predicted value is plotted on the x-axis of a graph and the actual value is plotted in the y-axis of the plot. If the actual and predicted values are equal, then the point should fall on the regression line. The graph shows that almost all the decision tree models can predict ETo values accurately. Among the decision tree models employed, the M5P and random forest seems to be the methods that better approximates the diagonal line (i.e., smaller prediction error). Though these models are to be considered better than random tree and REPtree, random tree model still deviate in a minor marginal value .

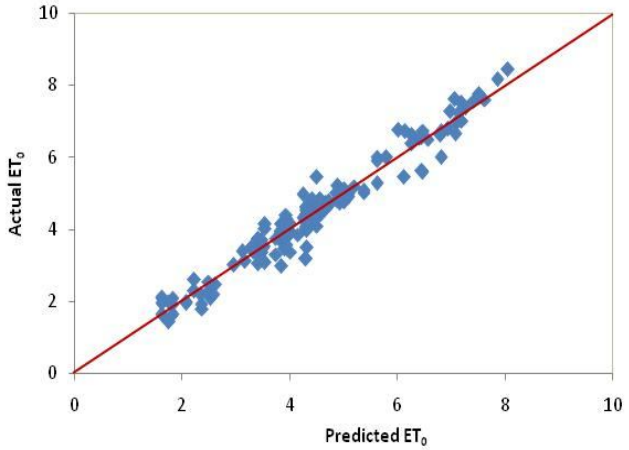

Fig.4. Comparison Plot for REP Tree model

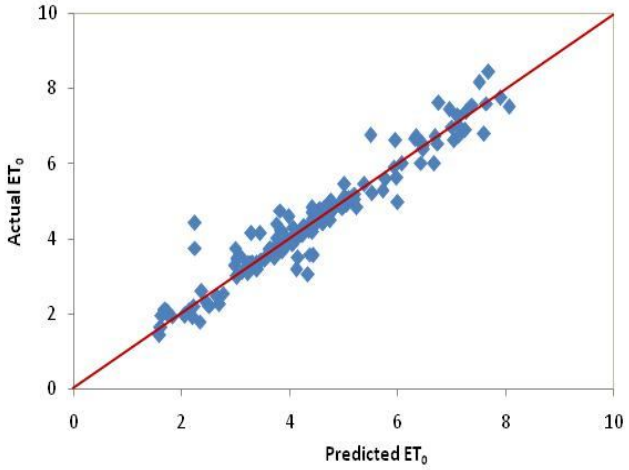

Fig.5. Comparison Plot for Random tree model

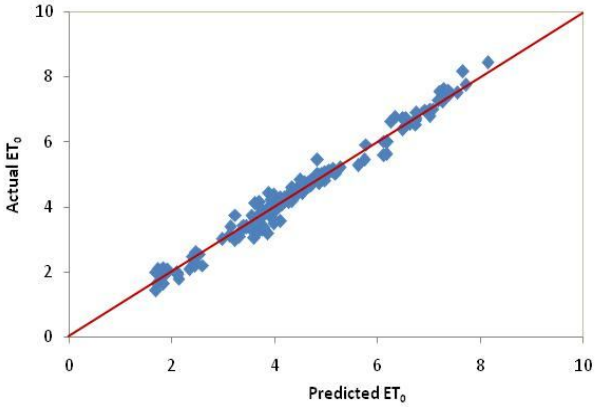

Fig.6. Comparison Plot for Random forest model

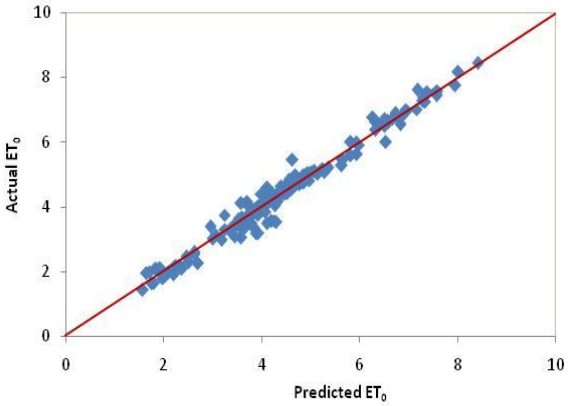

Fig.7. Comparison Plot for M5P model

Figures 4 and 5 show the plot of REP Tree and random tree models respectively. The plot obtained for these models are much scatterred than M5P plot and random forest plot as shown in Figs 6 and 7 respectively. Here, the solid line represents the condition of perfect agreement and the points plotted away from the solid line represents discrepancies of. It is seen that in Fig. 6 and 7 the very minimal scatter in ETo is evident for M5P and random forest models.

Fig. 8 shows the regression coefficient values obtained based on the computed and predicted values of the decision tree models employed. These regression correlation values ranged between 0.74 to 0.99. It is observed that M5P and random forest has highest regression coefficient of 0.99. Random tree ranks next in the hierarchy with a minimum deviation of 0.96. The RepTree performed the least.

V. CONCLUSION

ETo can be adequately estimated by decision tree models from values of meteorological variables of routine use. Decision trees are an efficient methodology to estimate daily ETo using a number of meteorological parameters. The better estimates of daily ETo shows that decision tree based

predictions are strongly correlated to the data set used. Among the decision tree models investigated M5P and random forest performed better. In future the combination of decision tree models may be investigated to further improve the performance.

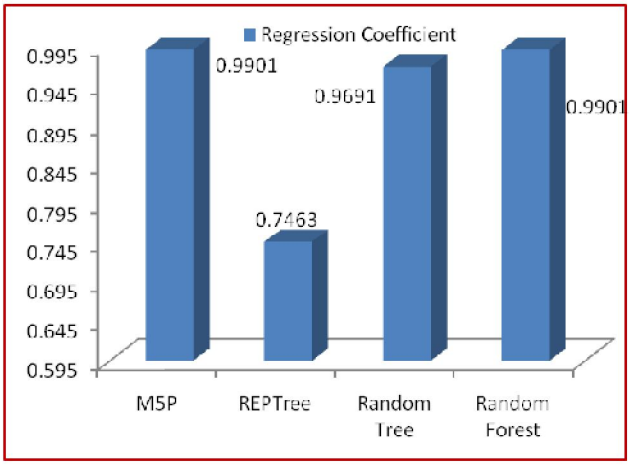

Fig. 8. Regression coefficient of decision tree models

REFERENCES

[1] Gholami, A., Bonakdari, H., Zaji, A. H., Michelson, D. G., & Akhtari, A. A. (2016). Improving the performance of multi-layer perceptron and radial basis function models with a decision tree model to predict flow variables in a sharp 90° bend. Applied Soft Computing, 48, 563-583.

[2] Goyal, M. K., & Ojha, C. S. P. (2013), Evaluation of rule and decision tree induction algorithms for generating climate change scenarios for temperature and pan evaporation on a Lake Basin. Journal of Hydrologic Engineering,19(4), 828-835.

[3] Kisi, O. (2016), Modeling reference evapotranspiration using three different heuristic regression approaches. Agricultural Water Management, 169, 162-172.

[4] Kisi, O., & Kilic, Y. (2015), An investigation on generalization ability of artificial neural networks and M5 model tree in modeling reference evapotranspiration. Theoretical and Applied Climatology, 1-13.

[5] Manikumari N. & Vinodhini G., Regression Models for Predicting Reference Evapotranspiration, International Journal of Engineering Trends and Technology, 38(3),134-139 ,2016.

[6] Pradhan, T., Walia, V., Kapoor, R., & Saran, S. (2014, September), Optimizing land use classification using decision tree approaches, In Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on(pp. 1-5). IEEE.

[7] Rahimikhoob, A. (2014), Comparison between M5 model tree and neural networks for estimating reference evapotranspiration in an arid environment.Water resources management, 28(3), 657-669.