Academic Journal of Applied Mathematical Sciences
ISSN(e): 2415-2188, ISSN(p): 2415-5225
Vol. 4, Issue. 9, pp: 98-106, 2018
URL: http://arpgweb.com/?ic=journal&journal=17&info=aims

Academic Research Publishing Group

Original Research **Open Access**

Boosted Regression Tree for Modeling Evaporation Piche Using Other Climatic Factors Over Ilorin

Ezekiel I. D.*
Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria

Alabi N. O.
Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria

Abstract

Evaporation is one of the climatic/meteorological factors influenced by causes of climate change. Interest in the topic of climate change has been growing over the last three decades. The threat it poses cannot be overemphasized particularly in developing economies largely due to the connection it has with national development issues. We present a regression tree grown using recursive binary splitting, cost complexity pruning and boosting to study the relationship between evaporation piche and other climatic factors such as relative humidity, solar radiation, sunshine hours, wind speed, temperature and rainfall over the city of Ilorin in Nigeria. These factors are generally seen to change with rising climatic changes in a place. Analysis of the fitted tree reveals that relative humidity, temperature and rainfall are the most important meteorological factors affecting the level of evaporation piche in the city of Ilorin.

Keywords: Climate change; Meteorological factor; Recursive binary splitting; Cost complexity pruning; Boosting.

1. Introduction

Climate change has been an undesirable environment threat of the 21st century which the world is currently experiencing and seeking measures to adapt and mitigate its impact. Globally the effect of climate change affect the economic development of any country. Global warming is closely associated with climate change especially as co-traveler in the interplay of the equilibrium between the natural and man-made components of the Green House Gases (GHG_S) that have been eminently adjusted globally as the culprit for the warming of the earth's atmosphere and oceans. Global warming is caused by increase in the emission of GHG_S through the burning of fossil fuels (oils, natural gas and coal), burning of wood and solid wastes landfills, combustion of solid wastes and fossil fuels in industrial and agricultural activities, bush burning and deforestation. These human factors have been proven to be responsible for the ongoing climate change.

Nigeria, like many other countries is exposed to climate induced dangers of desertification, erosion, flooding and other ecological problems which have impacts on the welfare of millions of people. Persistent droughts and flooding, off season rain and dry spells have sent growing seasons out of orbit, on a country dependent on a rain fed agriculture. Over 80% of Nigeria's population depends on rain fed agriculture and fishing as their primary occupation and means of livelihood, climate change has affects crop production in a number of ways, growth rates of maize, guinea corn, millet and rice are reduced by rises in temperature, warming trends also make the storage of roots crops and vegetables difficult for those without access to refrigerators. During the drought of the 1970's and 1980's, close to one million livestock were lost, affecting meat and dairy supply throughout the country. Animal production is affected due to the increase in disease and pest (including PPR, foot rot, mange etc.) under the influence of climate change impacts that cut investment profits in livestock production system. Climate change also affects human directly or indirectly in many ways changes in temperature, precipitation, rising sea levels, increasing frequencies have great implication on human health in the area of injury, illness, morbidity and mortality. Rising sea level is anticipated as a result of climate change, hence flooding may occur which is likely to increase the vulnerability of the poor to common diseases like cholera, typhoid, malaria and pneumonia. Energy services are necessary inputs for every nation's development and growth. And also the fuel driving the engine of growth and sustainability development is a nation's access to reliable and adequate energy. No economy can sufficiently thrive without adequate access to reliable and adequate energy. The supply of energy entails the generation, transmission and distribution of energy, notably electricity. Nigeria has an abundant supply of energy resources and it's endowed with thermal, hydro solar, oil resources and yet still described as an energy poor country. Nigeria like most developing countries is highly vulnerable to the impact of climate change because its economy is mainly dependent on income generated from the agriculture, manufacturing, food processing and export of fossil fuels and energy intensive products. Building Nigeria's Response to Climate Change (BNRCC) report of 2011 stressed that likelihood of influence of climate change will be pronounce in the hydro-power generation sector. This is particularly the case since it is generally susceptible to the geographical pattern of rainfall as well as changes in temperature. The report also stated that a reduced flow in river and higher temperature reduces the capability of thermal electric generation as higher temperature also reduces transmission capacity. Evaporation is one of the major processes in the hydrological

*Corresponding Author

cycle. It is perhaps the most difficult component because of complex interactions of the component of the land-plant-atmosphere system [1]. Evaporation depends on climatic factors such as temperature, wind speed, solar radiation amongst other factors. However the connection between evaporation and these other major factors is a complex process.

The ancient city of Ilorin, the capital of Kwara state, is located between latitude 8^0 24' and 8^0 36' north of the equator and between longitude 4^0 10' and 4^0 36' east of the Greenwich meridian. Ilorin is a transition zone between the deciduous wood land of the south and dry savanna of north. The city has an approximate area of 150sq.km and population of about 847,582. Ilorin has a humid tropical climate which is characterized by wet and dry seasons. The temperature in the city is uniformly high throughout the year and open air insulation can be very uncomfortable during the dry season. The mean monthly temperatures are usually very high varying between 25.1^0c in august and 30.3^0c in March. The area's daytime range of temperature is also high. Rainfall in Ilorin is produced by tropical continental air mass and it exhibits great variability both temporarily and spatially with mean annual rainfall of about 1200mm; relative humidity in the city during the wet season is between 75 to 80% while in the dry season it is about 65%. Vegetation in the city comprises of tall grasses which are interspersed with scattered trees. Examples of trees in the area include lophira lanceolata, terminaalia glauscens and andropogon gayanus are few examples of grass species [2]. Due to the urbanization process of deforestation activities is on the increase within and around Ilorin with its consequence effect on runoff generation.

Xie [3], suggested that using partial correlation coefficients to describe the linear relationship between two variables would be more reasonable and more reliable for practical application. This is particular so since partial correlation coefficient reflect marginal degree of dependency between evaporation and meteorological factors by deducting all other affecting variables. Shen, *et al.* [4] used a non-parametric test such as Mann-Kendall test to identify the difference in trend of E601 pan evaporation over China from 1960 to 2006. The results show that during the last 50years, temperature has presented a significantly increasing trend while E601 pan evaporation shows a decreasing trend for most climate zones of China before the late 1990_s. The result reveals that daily temperature range; sunshine duration and average wind speed correlated well within the E601 pan evaporation, and significantly decreasing trends of these influencing factors are main reason explaining the decreasing rates of E601 pan evaporation. Kay and Davies [5] investigated the differences in the estimation of monthly potential evaporation in Britain when using outputs from either GCM_s or RCM_s. They utilized the Penman-Monteith model and a simple temperature based model in the estimation of evaporation for the current climate. The results were compared with a dataset derived from a modified Penman-Monteith formulation using meteorological factors. According to their results, RCM'_s outputs are able to generate monthly evaporation rates that show much closer agreement with evaporation rates derived from observed data as compared to GCM'_s outputs.

Roderick, *et al.* [6] performed a study on global pan evaporation trends and found that there is an overall declining trend over a 30 to 50 year period. More specifically, they found most analyzed sites to range from -1 to -4mm y_r^{-1} after a 30 years period. Roderick, *et al.* [6] showed that there has been an overall decrease of $4.8W_m^{-2}$ over the past 30 years. Shih [7] employs nine meteorological variables to estimate pan evaporation using multiple regression with the ordinary least square analysis or the ridge regression analysis. The Thornthwaite method for estimation of evaporation and ET is basically dependent on the average monthly temperature, number of days in the month, and number of hours between sunrise and sunset. Because of simple data requirements, the method is relatively attractive for use in the Northwestern Ontario. Data requirements in both the Hargreaves and Christiansen-Mehta methods are identical. For example, the mean monthly temperature, the mean monthly relative humidity at noon, the wind velocity and the percentage of possible sunshine. The Hargreaves, the Christiansen-Mehta, and the Morton methods were selected for further investigation in the estimation of evaporation and ET in Northwestern Ontario. The result shows that the Hargreaves method has been found to be consistent under -estimate of evaporation. In most cases, the estimated values for the months of June and September are low, while for the months of July and August are comparable to pan evaporation. It is apparent from this that the percent error curves exhibit a high degree of curvature. Consequently, there exists wide-ranging error in obtaining estimates of evaporation. In this method, evaporation is a direct function of temperature. One reason for consistent under-estimates of evaporation given by Hargreaves is the values of monthly temperature. In Northwestern Ontario, the mean monthly temperature varies widely from month to month e.g. mean monthly temperature reaches as low as -30^0c in January and high as 25^0c in July. The Morton method is capable of estimating not only the potential evaporation but also the areal evaporation. The areal estimates of evaporation by this method are relatively low. It is noted that the areal evaporation by definition is the evaporation encompassing the effects of lakes, vegetation, and bare soils within an area. For example, the lowest under-estimate of evaporation occurred at the Rawson lake climatological station. The Morton method of estimating potential evaporation gives very reasonable results compared to the pan evaporation. It is apparent that the Morton method provides the least error at the Atikokan and pickle lake climatological stations. The Hargreaves method is the predictor of evaporation at the Lansdowne house and Rawson lake climatological stations. At the pickle lake climatological station, the Morton method and the Hargreaves method provide evaporation estimates within percent error of 8.2% and -8.5% respectively. Kisi [8] investigated the abilities of three different Artificial Neural Network (ANN) techniques and it was found that the MLP and Radial Basis Neural Network (RBNN) computing techniques could be employed successfully to model the evaporation process using the available climatic data.

Xu, *et al.* [9] based on various machine learning statistical methods such as NNARX ANN, and Marciano Method concluded out of four meteorological analyzed in the Yangtze basin of china, air temperature and relative humidity showed an increasing trend while wind speed and net radiation showed a declining trend. Peterson, *et al.*

[10] based on pan evaporation data (1945-1990) analysis from the eastern and western United States, Europe, Middle Asian and Siberian regions of the former Soviet Union, a significant decline of pan evaporation was reported. The paper assert for the western United States during the past 45 years, the largest change reported was 97mm increase in a warm season. They also concluded that a decrease in potential evaporation (ET_p) is an important feature of recent climate change. The decrease in pan evaporation was attributed to a decrease in the diurnal temperature range and an increase in low cloud cover. Chang, *et al.* [11] proposed a Self-Organizing Map Neural Network (SOMN) to assess the variability of daily evaporation on meteorological variables. The results demonstrated that the topological variables and the networks could well estimate the daily evaporation.

Kim, *et al.* [12] was able to show that ANN such as Multi-Layer Perceptron Neural Networks (MLP), Generalized Regression Neural Networks (GRNN) and Support Vector Machine Neural Networks (SVM-NNM) performed better than the empirical Linacre model and Multiple Linear Regression (MLR) model in estimating evaporation piche in temperate and arid climatic zones. In Japan, pan evaporation is experiencing a downward trend in 13 sites using a 34 years data on evaporation [13]. In India, Goyal, *et al.* [14] studied the abilities of soft computing models such as ANN, Least Squares Support Vector Machine (LSSVM), Fuzzy Logic (FG), Adaptive Neuro-Fuzzy Inference System (ANFIS) techniques, Hargreaves and Samani method (HGS), as well as the Stephens-Stewart (SS) method and concluded that the soft computing models outperformed the HGS and SS methods in terms of model accuracy in estimating daily evaporation. The study also showed that the LSSVM, and FG models produced the highest accuracies. Herch and Burn [15] applied the Meyer's formula to analyze the total evaporation and pan evaporation over Canada using 30, 40 and 50year data. Produced varying trends of gross evaporation and pan evaporation with generally June, July, October and annual evaporation producing significant decreasing trends. Kisi [16] investigated the accuracy of LSSVM, Multivariate Adaptive Regression Splines (MARS) and M5 model tree (M5 tree) in modeling evaporation using local input and output data while the MARS model performed better than the LSSVM model in the case of without local input and outputs. Hess [17] analysis of evaporation measurements (1964-1998) in the central coastal plains of Israel showed a small but statistically significant increase in pan evaporation from screened class pans.

2. Methods, Empirical Analyses and results

The meteorological data over Ilorin is a sample with size n = 372 observations on evaporation piche (**ev**), solar radiation (sr), wind speed (ws), sunshine hours (sh), rainfall (rf), relative humidity and change in temperature (te). This data was extracted from the database of Nigeria Meteorological Agency (NIMET). We divided the monthly data training and test datasets. The former was used to train the decision tree while the latter to test the performance of the fitted tree model. This method derive motivation from a tree analogy in which the leaves are referred to as the *terminal nodes* and the points along which the regressor space is split are called the *internal nodes*. The internal and terminal nodes are connected by sections called the *branches*. We draw our inspiration from the procedure outlined by Hastie, *et al.* [18] which involves dividing the regressor space $X = [x_1, x_2, x_3,\ldots\ldots,x_6]$ where x_1 = sr, x_2 = sh, x_3 = ws, x_4 = rf, x_5 = te, x_6 = rel into j distinct and disjoint box-shaped regions $R_1, R_2, \ldots..R_j$ in which the response averages in each box are as different as possible. The splitting procedures describing the regions are related to each through a binary tree done recursively. The prediction is then done by determining the region R_j in which an observation falls into and using the mean of the response values of the training observations in that region R_j as the predicted value for that observation. The response y = **ev** was modeled as a constant c_j in each region as

$$f(x) = \sum_{j=1}^{J} c_m I(x \in R_j) \qquad 1$$

The regions are determined such that the residual sum of squares, RSS, given in equation 1.1 is minimized

$$RSS = \sum_{j=1}^{J} \sum_{i \in Rj} (y_i - \bar{y}_{Rj})^2 \qquad 1.1$$

where $\bar{y}_{Rj}$ is the mean response (evaporation piche, **ev**) for training observations in the j^{th} region. Our goal is not to generate every J partition of the regression space which is costly and numerically infeasible. Rather we employed an efficient procedure called the *recursive binary splitting* (RBS) in growing our decision tree as discussed below.

2.1. Recursive Binary Splitting (RBS) On the Fitted Evaporation Piche Tree Model

We adopted the *recursive binary splitting* (RBS, hereon) which start with a single region at the top of the tree and gradually perform splitting in an optimal fashion at each step of the tree construction. The RBS algorithm select automatically a regressor $\boldsymbol{x}_j$ in $\boldsymbol{X}$ and a cutoff u such that the regressor space is split into two regions

$R_1 = \{X \mid x_j < u\}$ and $R_2 = \{X \mid x_j \geq u\}$

With the ultimate aim of reducing the RSS in equation 1.1. This implies that we attempt to generate the value of j and u, which minimize equation 1.2

$$\sum_{i:x_i \in R_1(j,u)} (y_i - \bar{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,u)} (y_i - \bar{y}_{R_2})^2 \qquad 1.2$$

where $\bar{y}_{R_1}$ is the mean response (ev) for training observations in the $R_1(j, u)$ and $\bar{y}_{R_2}$ is the mean response for training observations in the $R_2(j, u)$ regions. This process was repeated in the subsequent steps minimizing RSS in each step. This resulted in the tree in Figure 1 with the value of j and u that minimizes the RSS in equation 1.1 are 6 and 77.5 per cent respectively. That is relative humidity (rel) is the regressor at the top of the tree used for the initial split such that

$$R_1 = \{X \mid rel < 77.5\} \text{ and } R_2 = \{X \mid rel \geq 77.5\}$$

The value of RSS = 225, MSE = 1.257 and the number of terminal nodes = 7. The splitting was terminated as soon as we have not more than 5 observations in each region Table 1. The unpruned tree used up about 88 per cent of the training observations.

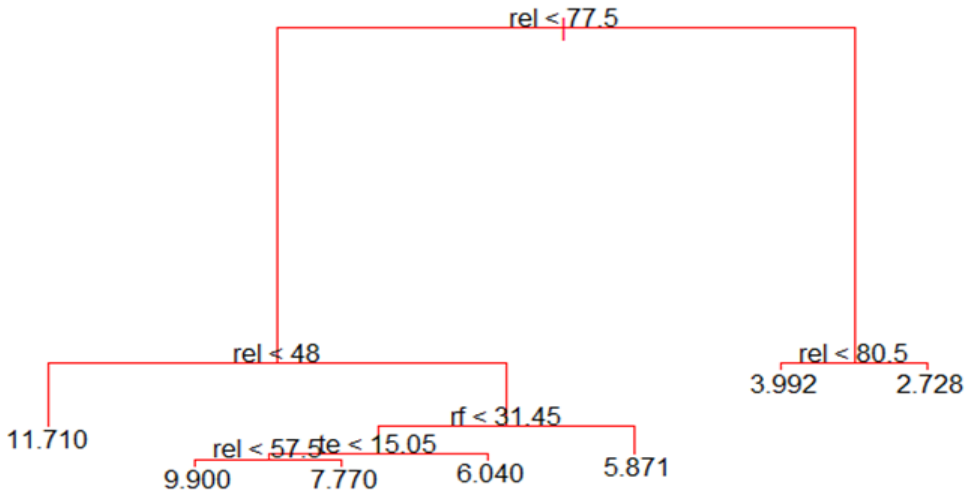
Table-1. Analysis of Recursive Binary Splitting on Evaporation Piche Tree Model

Node	Split	Number of Observations	Residual Sum of Squares	Predicted Evaporation Piche
1	Root	186	1502	5.209
2	rel < 77.5	94	482.900	7.462
4	rel < 48	9	16.050	11.710*
5	rel > 48	85	287.100	7.012
10	rf < 31.45	50	161.900	7.810
20	te < 15.05	45	124.800	8.007
40	rel <57.5	5	10.640	9.900*
41	rel > 57.5	40	94.000	7.770*
21	te > 15.05	5	19.710	6.040*
11	rel > 31.45	35	47.810	5.871*
3	rel > 77.5	92	54.620	2.907
6	rel < 80.5	13	4.169	3.992*
7	rel > 80.5	79	32.600	2.728*

*Terminal node (leave) ***Source***: *Personal Computations using* ***R*** *language*

2.1.1. Recursive Binary Splitting of Regression Tree on Evaporation Piche Using Meteorological Factors over Ilorin

Figure-1. *Unpruned decision tree with 7 leaves and 6 internal nodes. This tree model shows that relative humidity (rel) is the most important regressor in the evaporation piche tree model since it minimizes the RSS in equation 1.1. Splitting was done using the recusrsive binary algorithm. At each step, a regressor is selected for binary splitting. The the entire binary tree splitting process resulted in relative humidity rel, rainfall rf and temperature te as internal and terminal nodes. At a given internal node the left-hand branch is represented by $x_j < q_k$ resulting from the split and $x_j > q_k$ indicates the right hand branch. At the top of the tree, the split resulted in two branches in which the left hand branch corresponds to rel < 77.5 per cent and the right hand branch corresponds to rel ≥ 77.5 per cent. Splitting ensures simplicity and ease of interpretability of the regression tree model*



Source: *Personal computation using R language*

This procedure of RBS resulting in Figure 1 produced good predictions on the training dataset but eventually overfitting the evaporation piche data. One problem is that this might lead to a very poor performance of the decision tree model on the test dataset. We were able to achieve a better test performance by generating a smaller tree that contain fewer splits using *cost complexity pruning*. One benefit of this pruning method is that we were able to fit an evaporation piche tree model with lower variance but slightly higher bias in a manner that utilizes only few number of subtrees unlike the cross-validation and the validation set pruning. The cost complexity pruning involves a positive tuning parameter α such that for every value of this quantity, there exists a subtree $T \subset T_0$ for which equation 1.3 is as small as possible.

$$C_\alpha|T| = \sum_{m=1}^{|T|} n_m H_m(T) + \alpha|T|$$

where,

$$H_m(T) = \frac{1}{n_m} \sum_{x_i=R_m} (y_i - \hat{c}_m)^2 \qquad 1.3$$

$$\hat{c}_m = \frac{1}{n_m} \sum_{x_i=R_m} y_i$$

$$n_m = \#\{x_i \in R_m\}$$

The quantity $|T|$ is the number of leaves on the decision tree T, R_m is the regions relating to the m^{th} leave and $\hat{c}_m$ is the mean of the training dataset's evaporation piche (ev) corresponding to the region R_m. Increasing the value of α from zero in equation 1.3 prunes the branches and controls the tradeoff between the complexity of the regression subtree and its goodness of fit on the training dataset. We employed the 10-fold cross validation to determine the values of the positive tuning parameter and a most complex tree. We computed a tuning parameter $\alpha = 18.785$ with an RSS value of 403.652 using cross validation. Table 2 shows the result of the cost complexity pruning for various values of α at each split. We set the value of the best number of terminal nodes to 4 which is calculated using the cross validation. This pruned tree was generated from a large tree on the training dataset containing 186 observations and varying the nonnegative tuning parameter α in equation 1.3.

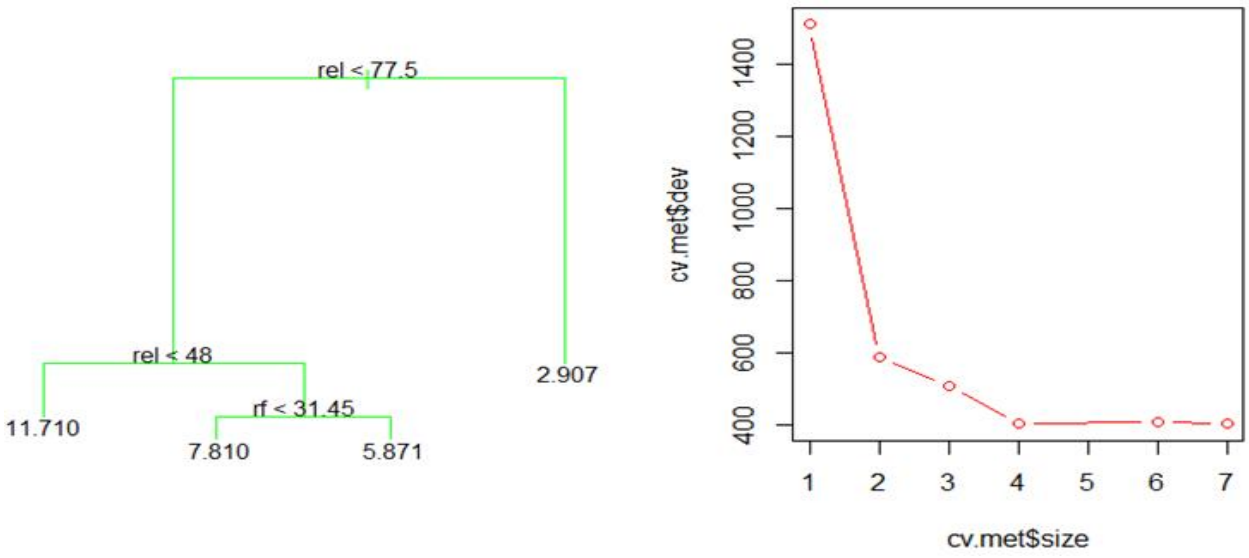
Table-2. *Cost Complexity Pruning on Regression Tree Model*

Node	Split	Number of Observations	Residual Sum of Squares	Predicted Evaporation Piche
1	root	186	1502	5.209
2	rel < 77.5	94	482.900	7.462
4	rel < 48	9	16.050	11.710*
5	rel > 48	85	287.100	7.012
10	rf < 31.45	50	161.900	7.810*
11	rf > 31.45	35	47.81	5.871*
3	rel > 77.5	92	54.620	2.907*

Terminal node (leave)* ***Source**: Personal Computation using* ***R*** *language*

2.1.2. Cost Complexity Pruning of Decision Tree on Evaporation Piche Using Other Meteorological Factors

Figure-2. *Analysis of cost complexity pruning on the fitted decision tree of Figure 2. Left panel: Prunned tree with 4 terminal nodes, 3 internal nodes. At a given internal the left-hand branch is represented by $X_j < q_k$ resulting from the split and $X_j > q_k$ indicates the right hand branch. At the top of the tree, the split resulted in two branches in which the left-hand branch corresponds to rel < 77.5 per cent and the right-hand branch corresponds to rel ≥ 77.5 per cent. Right panel: The result of 10-fold cross validation showing the cross-validation error (cv.met$dev) as a function of the terminal nodes (cv.met$size). It indicates that the CV error dipped at terminal node value of 4* with RSS value = 403.652

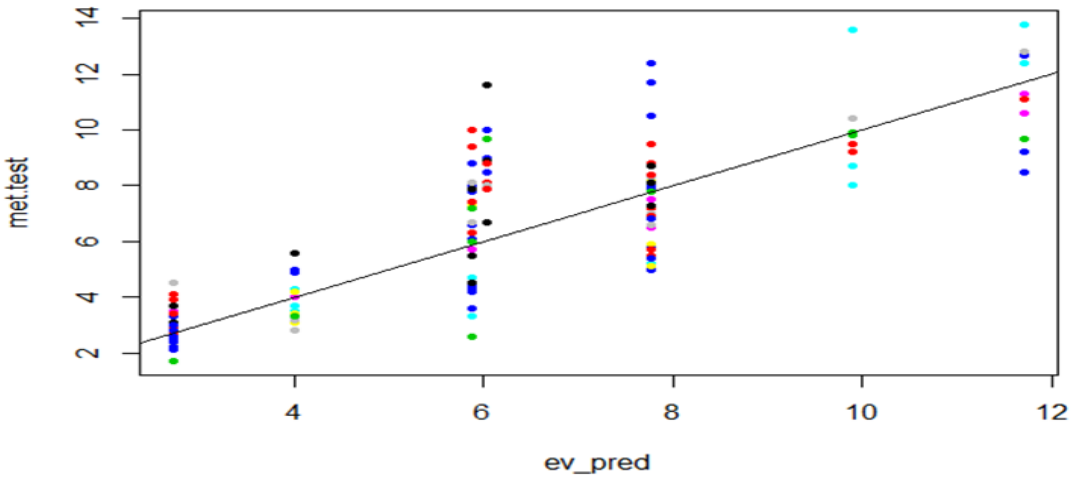


Source*: Personal Computation using* ***R*** *language*

The pruned tree in Figure 2 reveals that out of the six regressors, relative humidity rel is the most important meteorological factor for predicting evaporation piche over Ilorin city. Using 94 training observations, given that rel

is greater or equal to 77.5 per cent, the mean of evaporation piche is 2.907 ml which represent the first terminal node. The fitted regression was further split by rel given that the relative humidity at the top of the tree is less than 77.5 per cent. At this step, if rel is less than 48 per cent, a total of 9 training observations produced a mean response of 11.710 ml. Otherwise another split was carried out which resulted in two terminal nodes. These two terminal nodes are 7.81 ml if rainfall (rf) is less than 31.45 mm and 5.871 ml otherwise. A total of 50 and 35 training observations were involved in calculating these two mean responses of the evaporation piche. Using the result of this pruned tree, we predicted the evaporation piche for the 186 test observations to determine its performance. The mean square error (MSE) of the pruned model using test dataset is 2.196.

Figure-3. *The scatter plot on the residuals of the optimally-pruned regression tree using the test dataset. The straight line through the dotted points is the trend line*



***Source**: Personal Computation using **R** language*

2.2. Boosting the Fitted Evaporation Piche Decision Tree Model

Decision trees are known to present some setbacks in terms of predictive accuracy (higher variance) causing instability and lack of robustness to changes in data. Furthermore, the pruned tree using the cost complexity pruning performed poorly on the test dataset as revealed by the MSE values. Therefore, we employed boosting to improve the accuracy of the predictions generated by the regression tree model. Boosting is a bootstrap aggregation procedure used in reducing the variance of a statistical learning method such as our fitted regression tree. Averaging a set of observations reduces the variance at the expense of higher bias. Here the bootstrapping involves taking single training dataset, we sequentially grow G = 40,000 different bootstrapped regression trees. Subsequent trees are grown utilizing information from the residuals ($\boldsymbol{r}_i$) of the previously grown trees. That is these trees were fitted using the residuals of previously grown trees rather than the original response values $\boldsymbol{y}_i$. Using G = 40,000 subtrees, we computed.

$$\hat{f}^{1}(x), \hat{f}^{2}(x)..... \hat{f}^{40,000}(x)$$

based on the following algorithm.

2.2.1. Algorithm on Boosting For Evaporation Piche Decision Tree Model

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training dataset
2. Fit g = 1, 2, 3,……,40,000, repeat:

a. Fit a tree $\hat{f}^{g}$ with $\boldsymbol{v}$ splits to the training data (x, r)

b. Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \delta\hat{f}^{g}(x)$$

c. Update the residuals,

$$r_i \leftarrow r_i \leftarrow \delta\hat{f}^{g}(x_i)$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{g=1}^{40,000} \delta\hat{f}^{g}(x)$$

Because these trees are grown deep and unpruned, they possess higher variance but low bias. The boosting algorithm terminated with G = 4,074 trees aggregated. In order to improve the performance of our boosted model by minimizing the variance, we set the shrinkage parameter δ = 0.0159. The MSE of the boosting procedure using 15-fold cross-validation on test data set is 1.3953 which is about 64 per cent improvement over the cost complexity

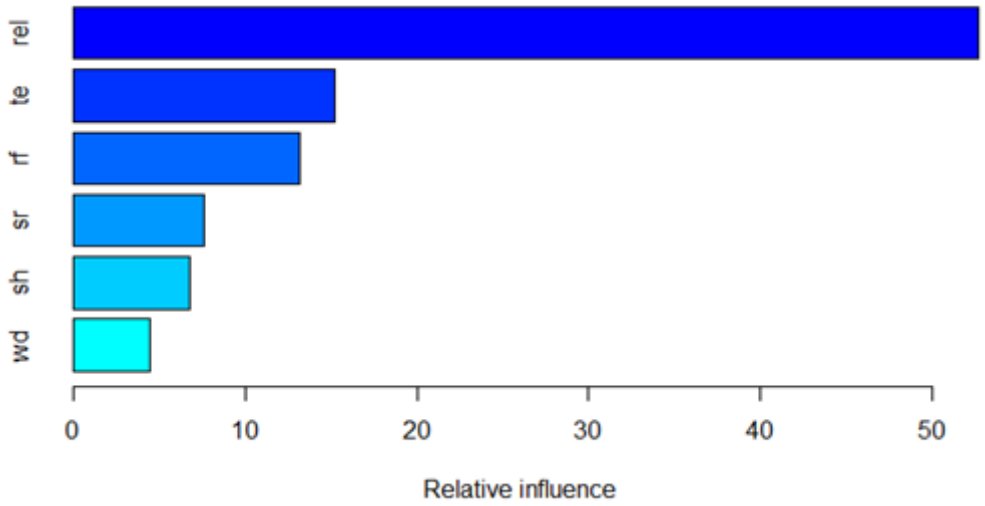
pruned model. We were able to generate the relative influence of the regressors in the evaporation piche tree model using the boosting procedure Table 3 and Figure 4.

Table-3. Analysis of regressor relative influence on evaporation piche over Ilorin

Regressor	Relative Influence (%)
Relative Humidity (rel)	52.704
Temperature (te)	15.221
Rainfall (rf)	13.140
Solar radiation (sr)	7.657
Sunshine hours (sh)	6.806
Windspeed (wd)	4.474

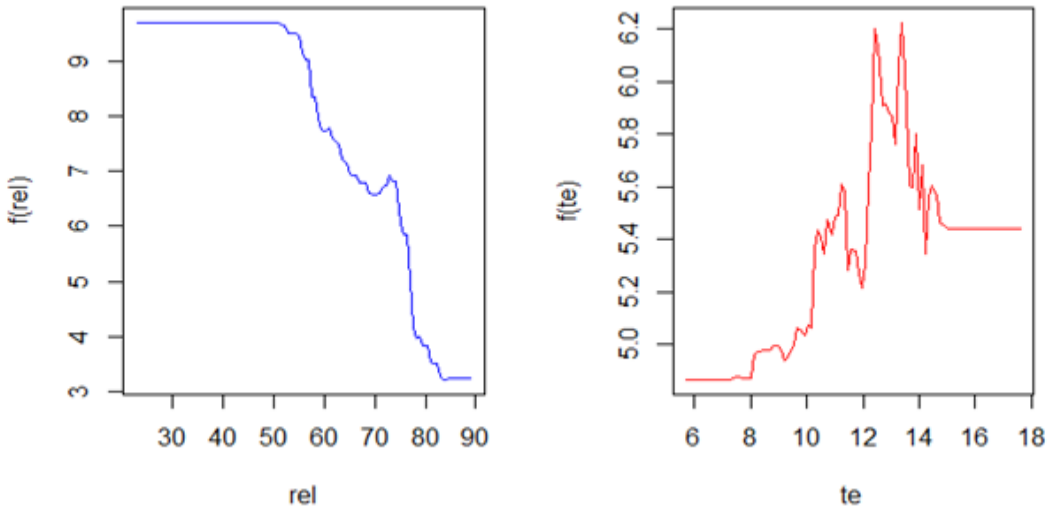
Source: Personal Computation using **R** language

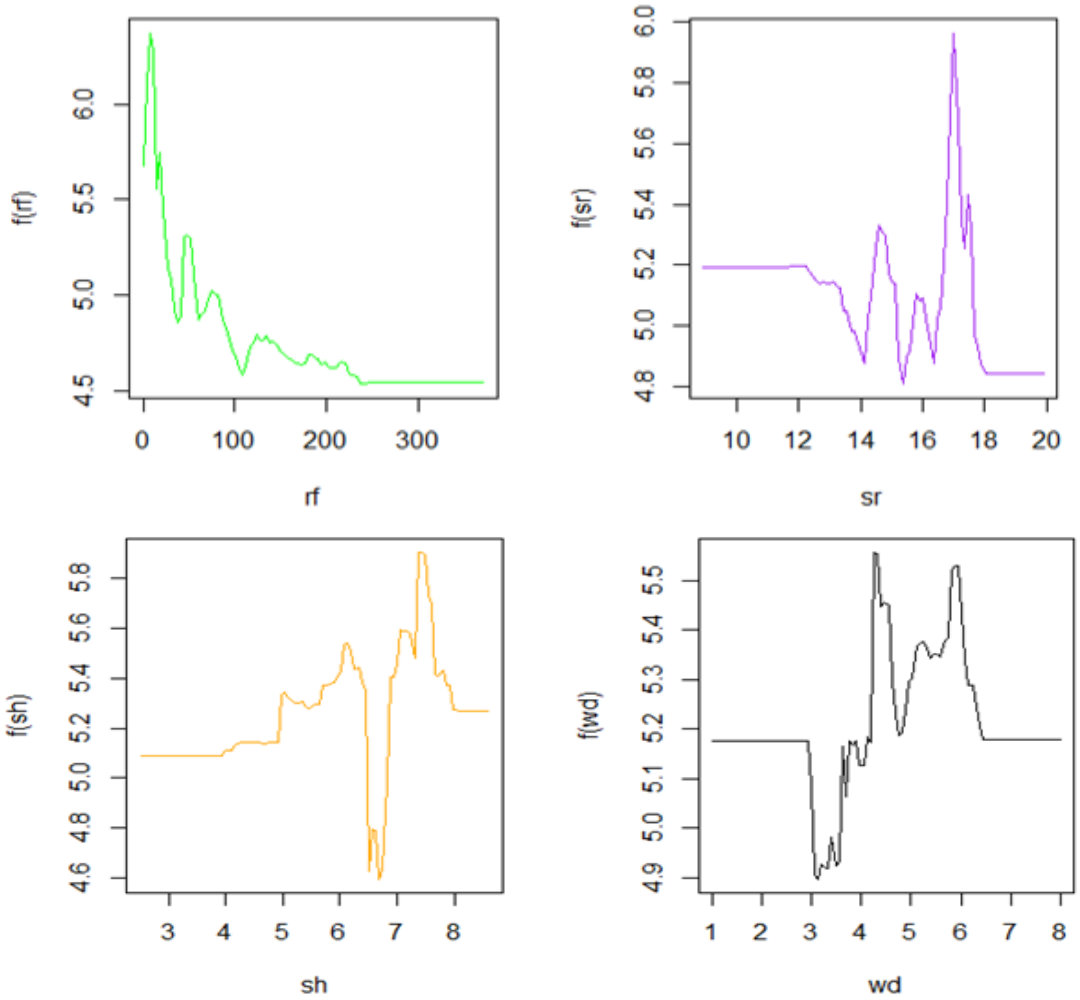
Figure-4. *Relative Influence of each meteorological factor is displayed in this plot. The plot reveals that relative humidity is largely the most important factor (accounting for more than 50 per cent of the total variability) affecting evaporation piche in Ilorin. This is followed by temperature (about 15.3 per cent), then rainfall (about 13.2 per cent). Wind speed has the least impact (less than 7 per cent of the total variability) on evaporation piche in the city*



***Source**: Personal Computation using **R** Language*

Figure-5. *Partial Dependence plots showing the marginal effects of the meteorological factors on evaporation piche over Ilorin. These plots are the marginal effects of the corresponding meteorological factor integrating out the other factors in the fitted decision tree model. These plots show that evaporation piche declines steadily with increase in relative humidity. Similarly, as rainfall rises in the city, evaporation piche declines. However, rising temperature resulted in increased evaporation piche*





***Source**: Personal Computation using **R** Language*

3. Conclusion

Decision tree for regression has been used to study the relationship between evaporation piche and six other meteorological factors influencing it over Ilorin. The analysis of the tree was done using the recursive binary splitting (RBS), cost complexity pruning, and boosting. Recursive Binary Splitting produced a tree with seven leaves or terminal nodes. Though this single large tree produced a low MSE for the model on the training dataset, its performance on the test dataset was weak. Hence, we applied the cost complexity pruning (CCP) to trim down the number of terminal nodes (leaves) and improve the predictive capacity alongside the interpretability of our model. This model resulted in lower model MSE for the test dataset and 4 terminal nodes. Relative humidity minimizes the residual sum of squares (RSS). Therefore, this meteorological factor was used as the initial splitting variable at the top of the tree. The Cost Complexity Pruning procedure was based on varied values of a tuning parameter which was used to control the tradeoff between the complexity of the model and it overfitting model. We employed boosting to further improve the predictive capacity of the postulated decision tree model. Boosting is a procedure that involve growing large number of separate trees sequentially using the residuals from previously grown tree as the response in the new tree. The results of boosting revealed improved performance of the regression tree model on the test data set in terms of lower MSE. Partial dependence analysis of the regressors in terms of marginal effects indicates that evaporation piche rises with rising temperature and sunshine hours. But declines with rising relative humidity and rainfall in Ilorin. This study shows that relative humidity, temperature and rainfall are the most important meteorological factors affecting the level of evaporation piche in the city of Ilorin.

References

[1] Singh, V. P. and Xu, C. Y., 1997. "Evaluation and generalization of 13 mass transfer equations for determining free water evaporation." *Hydrological Processes*, vol. 11, pp. 311-323.

[2] Iroye, K. A., 2013. "Water position and its implications on basin management in urbanized catchment in a tropical city of ilorin, Nigeria." *Ethiopian Journal of Environmental Studies and Management*, vol. 6, pp. 358-364.

[3] Xie, P., 2009. *Spatial-temporal variability and simulation of evapotranspiration in East River Basin*. China: Sun Yat-Sen University.

[4] Shen, Y. J., Liu, C. M., and Liu, M., 2009. "Change in pan evaporation over the past 50years in the arid region of China." *Hydrol. Processes*, vol. 24, pp. 225-231.

[5] Kay, A. L. and Davies, H. N., 2008. "Calculating potential evaporation climate model data. A source of uncertainty for hydrological climate change impacts." *J. Hydrol.*, vol. 358, pp. 221-239.

[6] Roderick, M. L., Hobbins, M. T., and Farquhar, G. D., 2009. "Pan evaporation trends and the terrestrial water balance." *Geography Compass*, vol. 3, pp. 746-760.

[7] Shih, S. F., 1984. "Data requirements for evaporation estimation." *ASCE*, vol. 110, p. 263.

[8] Kisi, O., 2009a. "Daily pan evaporation modeling using multi-layer perceptrons and radial basis neural networks." *Hydrol. Process*, vol. 23, pp. 213-223.

[9] Xu, C. Y., Gong, L., Tong, J., and Chen, D., 2006. "Decreasing reference evapotranspiration in a warming climate. A case of Changjiang (Yangtze) River catchment during 1970-2000." *Advances in Atmospheric Sciences*, vol. 23, pp. 513-520.

[10] Peterson, T. C., Golubev, V. S., and Groisman, P. Y., 1995. "Evaporation losing its strength." *Nature*, vol. 377, pp. 687-688.

[11] Chang, F. G., Chang, L. C., Kao, H. S., and Wu, G. R., 2010. "Assessing the effort of meteorological variables for evaporation estimation by self-organizing map neural network." *J. Hydrol.*, vol. 384, pp. 118-129.

[12] Kim, S., Shiri, J., and Kisi, O., 2012. "Pan evaporation modeling using neural computing approach for different climatic zones." *Water Resour. Manage*, vol. 26, pp. 3231-3249.

[13] Jun, A. and Hideyuk, K., 2004. "Pan Evaporation trends in Japan and its relevance to the variability of the hydrologic cycle." *Tenki*, vol. 51, pp. 667-678.

[14] Goyal, M. K., Bharti, B., Quilty, J., Adamowski, J., and Pandey, A., 2014. "Modeling of daily pan evaporation in Sub-tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS." *Expert Syst. Appl.*, vol. 41, pp. 5267-5276.

[15] Herch, N. M. and Burn, D. H., 2005. *Analysis of trends in evaporation- phase 1*. ON Canada: University of Waterloo.

[16] Kisi, O., 2015. "Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree." *J. Hydrol.*, vol. 528, pp. 312-320.

[17] Hess, T. M., 1998. "Trends in reference evapotranspiration in the North East Arid Zone in Nigeria." *Journal of Arid Environments*, vol. 38, pp. 99-115.

[18] Hastie, T., Tibshirani, R., and Friedman, J., 2008. *Elements of statistical learning, data mining, inference and prediction*. Second Edition ed. Stanford California: Springer.