Journal of Agrometeorology 10(1) : 33- 38 (June 2008)

Modeling of evaporation using M5 model tree algorithm

S. DESWAL
Civil Engineering Department, National Institute of Technology, Kurukshetra-136119, Haryana
Email: deswal.leo@gmail.com

ABSTRACT

This paper investigates the prediction of pan evaporation using M5 Model Tree technique, evaluated for its applicability for predicting evaporation from meteorological data. Different combinations of input data were considered and the resulting values of evaporation were analysed and compared with those of existing techniques. The results suggest that the M5 Model could be successfully employed in estimating the evaporation from the available meteorological data set, within a scatter of ±15%, using the combination of air temperature, wind speed, sunshine hours and relative humidity) using M5 Model Tree algorithm. This study suggests the usefulness of M5 Model Tree technique with all the meteorological parameters considered together in predicting the pan evaporation from reservoirs.

Keywords: pan evaporation, M5 model tree.

Evaporation is a necessary components in any water balance assessment for different water resources planning, design, operation and management studies including hydrology, agronomy, forestry and land resources, irrigation management, river flow forecasting, investigation of lake ecosystem and modeling, and many more. Among the various components of hydrological cycle, evaporation is probably the most difficult to estimate due to complex interactions between the components of land-plant-atmosphere system (Singh and Xu, 1997). This is particularly true for lakes/reservoirs for which detailed daily pan evaporation measurements are difficult for long time periods. Thus, it becomes necessary to develop approaches to estimate the evaporation rates from other available meteorological parameters, which are comparatively easier to measure (Terz and Keskn, 2005).

Evaporation has been estimated from meteorological parameters through empirically developed methodologies or statistical and stochastic approaches in addition to mass-balance based formulations by many researchers (Bruin, 1978; Anderson and Jobson, 1982; Stewart and Rouse, 1976; Abtew, 2001). Recently, Murthy and Gawande (2006) proposed simple linear relationships between evaporation and individual meteorological parameters for predicting evaporation from reservoirs by using linear regression approach. Murthy and Gawande (2006) did not consider the combined effect of all the meteorological parameters on evaporation, which seems to be the major limitation. Many other studies too suffer from similar limitation (Singh *et al*., 1981; Senapati *et al.*, 1985; Chandra *et al.*, 1988; Bhakar and Singh, 2004; Kadhane and Tatewar, 2006).

In view of the limitation of above studies, this paper presents the results of a M5 Model Tree based approach in predicting Pan Evaporation using different combinations of input data. Further, its performance was compared with Linear Regression approach as suggested by Murthy and Gawande (2006).

M5 model tree algorithm

This machine-learning technique uses the following idea: to create a model by using the input parameters, split the parameter space into areas (subspaces) and build in each of them a linear regression model. In fact the resulting model can be seen as a modular model, or a committee machine, with the linear models being specialized on the particular subsets of the input space. The M5 Model Tree approach (Quinlan, 1992; Witten and Frank, 1999), based on the principle of information theory, makes it possible to split the multi-dimensional parameter space and to generate the models automatically according to the overall quality criterion. It also allows for varying the number of models.

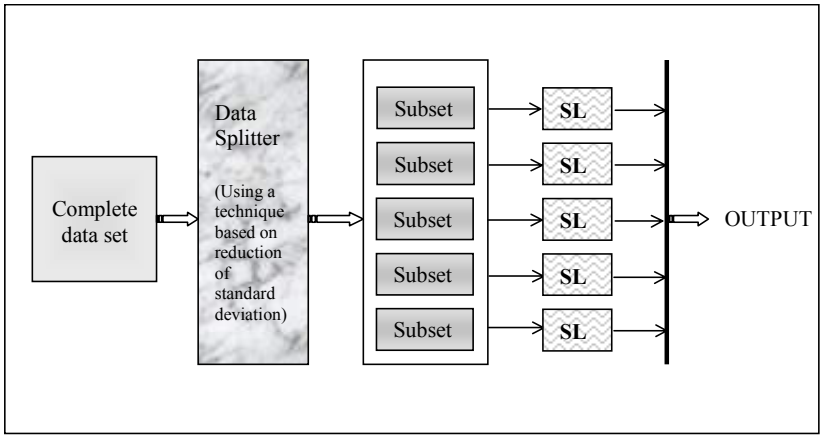

Fig. 1: The induction of M5 model tree as modular model, where SL indicates simple linear regression.

The splitting in model tree follows the idea of a decision tree, but instead of the class labels it has linear regression functions at the leaves, which can predict continuous numerical attributes. Model trees generalize the concepts of regression trees, which have constant values at their leaves (Witten and Frank, 1999). So, they are analogous to piece-wise linear functions and hence nonlinear. Computational requirements for model trees grow rapidly with dimensionality. The major advantage of model trees over regression trees is that they are much smaller than regression trees, the decision strength is clear, and regression functions do not normally involve many variables.

The algorithm known as the M5 algorithm is used for inducing a model tree (Quinlan, 1992), which works as follows (Fig.1). Suppose that a collection T of training examples is available. Each example is characterized by the values of a fixed set of (input) attributes and has an associated target (output) value. The aim is to construct a model that relates a target value of the training cases to the values of their input attributes. The quality of the model will generally be measured by the accuracy with which it predicts the target values of the unseen cases.

Tree-based models are constructed by a divide-and-conquer method. The set T is either associated with a leaf, or some test is chosen that splits T into subsets corresponding to the test outcomes and the same process is applied recursively to the subsets. The splitting criterion for the M5 model tree algorithm is based on treating the standard deviation of the class values that reach a node as a measure of the error at that node, and calculating the expected reduction in this error as a result of testing each attribute at that node.

The formula to compute the standard deviation reduction *(SD*R) is:

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (1)$$

where T represents a set of examples that reaches the node; T_i represents the subset of examples that have the i^{th} outcome of the potential set; and *sd* represents the standard deviation. After examining all possible splits, M5 chooses the one that maximizes the expected error reduction. Splitting in M5 ceases when the class values of all the instances that reach a node vary just slightly, or only a few instances remain. The relentless division often produces over-elaborate structures that must be pruned back, for instance by replacing a subtree with a leaf. In the final stage, a smoothing process is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree, particularly for some models constructed from a smaller number of training examples. In smoothing, the adjacent linear equations are updated in such a way that the predicted outputs for the neighbouring input vectors corresponding to the different equations are becoming close in value. Details of this process can be found in Quinlan (1992) and are given by Witten and Frank (1999).

Table 1: Average weekly meteorogical data of Manasgaon (1990-2004).

Month	Weeks	Pan evaporation , E (mm per day)	Air temperature, T (^{O}C)	Wind speed , U_2 (m sec^{-1} at 2 m height)	Sunshine hours, S_h (hrs day^{-1})	Relative humidity, R_h (%)
Jan	1	3.4	19.64	3.3	8.8	64.3
	2	3.3	20.29	2.9	8.3	61.3
	3	3.4	21.47	3.2	8.5	60.1
	4	3.0	20.86	3.8	6.9	60.4
Feb	1	3.4	21.99	3.5	9.0	59.3
	2	3.9	23.59	3.8	8.9	55.7
	3	4.3	24.52	4.4	9.1	54.2
	4	4.7	25.62	4.1	8.9	48.4
Mar	1	6.2	26.31	4.7	10.1	45.4
	2	5.6	26.89	4.9	8.6	45.1
	3	6.5	29.18	5.0	9.3	40.7
	4	5.9	30.04	5.4	8.2	40.2
Apr	1	8.1	31.52	5.8	10	42.5
	2	6.8	32.44	6.7	8.3	40.4
	3	8.8	33.36	6.4	9.8	38.5
	4	9.8	34.70	8.0	9.3	36.4
May	1	12.4	35.46	8.5	11.2	37.6
	2	11.9	34.85	9.5	10.2	44.4
	3	12.2	34.41	11.1	9.3	49.3
	4	11.2	34.68	10.3	8.3	46.6
June	1	15.3	34.05	10.7	9.4	53.3
	2	9.8	31.86	9.1	5.0	65.2
	3	7.5	30.07	10.4	4.1	74.0
	4	5.7	29.68	10.8	4.1	73.9
July	1	7.5	29.66	8.8	5.7	73.0
	2	5.7	29.01	9.3	3.9	77.6
	3	4.9	27.52	8.2	2.6	82.8
	4	3.3	27.37	8.1	3.4	83.9
Aug	1	3.9	27.10	6.7	3.2	87.6
	2	3.1	26.70	6.8	2.5	86.9
	3	2.9	26.56	6.6	2.4	86.2
	4	3.0	26.88	6.1	3.7	84.8
Sep	1	3.1	26.85	6.1	4.0	84.7
	2	3.3	27.43	5.4	5.6	79.3
	3	4.1	27.86	4.3	5.8	78.7
	4	4.4	28.58	2.9	6.3	78.9
Oct.	1	4.9	27.83	2.9	8.0	77.5
	2	4.3	27.43	2.8	7.3	75.6
	3	4.0	25.65	2.2	8.8	69.2
	4	3.5	24.62	2.5	7.6	65.8
Nov	1	4.5	24.78	2.5	10.1	66.0
	2	3.5	24.47	2.5	7.4	69.3
	3	3.7	23.03	2.1	8.9	60.5
	4	3.4	22.47	2.2	8.7	58.5
Dec	1	3.7	20.78	2.4	9.6	61.4
	2	3.3	20.99	2.4	8.3	61.2
	3	3.3	20.21	2.5	8.1	61.0
	4	2.8	20.38	3.2	6.5	60.2

Source: Murthy and Gawande (2006)

MATERIALS AND METHODS

The data used in the present study (Table 1) are taken from Murthy and Gawande (2006). The mean weekly pan evaporation per day were collected from Anand Sagar Reservoir in Shegaon for one year only ; while other meteorological data for a period of fifteen years (from 1990 to 2004) was obtained from a full

Table 2: Relationship of pan evaporation with different input combinations of meteorological parameters

S. No	Modeling tool	Input parameters	Prediction equation	Correl. coeff.	RMSE
*1	Linear Regression	T	$E = 0.5706\ T - 9.9064$	0.840	1.64
		U_2	$E = 0.7775\ U_2 + 1.261$	0.716	2.108
		S_h	$E = 0.455\ S_h + 2.2303$	0.361	2.817
		R_h	$E = -0.1082\ R_h + 12.345$	0.546	2.531
2	M5 Model Tree	T	$E = 0.266\ T - 2.619$ for $T \leq 28.8$ $E = 0.811\ T - 17.336$ for $T > 28.8$	0.906	1.303
		U_2	$E = 0.778\ U_2 + 1.261$	0.666	2.258
		S_h	$E = 0.131\ S_h + 3.870$ for $S_h \leq 9.2$ $E = 0.263\ S_h + 5.110$ for $S_h > 9.2$	0.369	2.823
		R_h	$E = -0.056\ R_h + 10.710$ for $R_h \leq 53.8$ $E = -0.033\ R_h + 6.673$ for $R_h > 53.8$	0.522	2.587
		$T + U_2$	$E = 0.302\ T - 0.103\ U_2 - 3.082$ for $T \leq 28.8$ $E = 0.768\ T + 0.19\ U_2 - 17.559$ for $T > 28.8$	0.874	1.473
		$T + S_h$	$E = 0.340\ T + 0.310\ S_h - 6.658$ for $T \leq 28.8$ $E = 0.807\ T + 0.202\ S_h - 18.711$ for $T > 28.8$	0.922	1.77
		$T + U_2 + S_h$	$E = 0.229\ T + 0.671\ U_2 + 0.679\ S_h - 9.326$	0.946	0.982
		$T + U_2 + S_h + R_h$	$E = 0.23\ T + 0.72\ U_2 + 0.85\ S_h + 0.03\ R_h - 12.83$	0.951	0.941

* Results of Murthy and Gawande (2006)

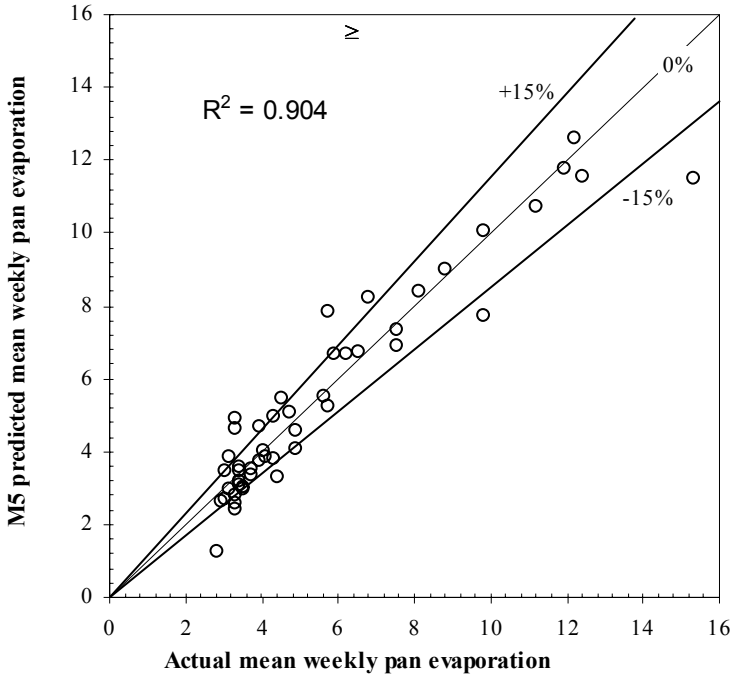

Fig. 2: Scatter plot between actual and estimated pan evaporation (mm day^{-1})

climatic station at Manasgaon, about 9 Km from Shegaon, lying under water resources division, Amravati Hydrology Project (Government of Maharashtra). Class *A* Pan Evaporameter was used for all measurements. The maximum and minimum values are averaged so as to get weekly data, which were used

for the analysis. The same data is used for estimating pan evaporation by using M5 Model Tree technique in the present analysis.

M5 MODEL

The parameters comprising the data set includes pan evaporation (E) as the output parameter and up to four input parameters representing air temperature (T), wind speed (U_2), sunshine hours (S_h) and relative humidity (R_h). A total of 48 data values were used in the present study for model building and validation. The M5 Model Tree algorithm is used to calculate correlation coefficient and root mean square error (RMSE) by using cross-validation to generate the model on different combinations of the input data set in predicting the weekly mean pan evaporation per day. Cross validation was used to test the models due to the availability of small number of data sets. The cross-validation is a method of estimating the accuracy of a classification or regression model in which the input data set is divided into several parts (a number defined by the user), with each part in turn used to test a model fitted to the remaining parts. For this study, a ten-fold cross-validation was used.

Two parameters namely correlation coefficient and root mean square error (RMSE) values were used for the performance evaluation of the models and comparison of the results for prediction of evaporation. The results of the M5 model tree technique are provided in Table 2 in terms of the correlation coefficient and RMSE. For comparison, the results of Murthy and Gawande (2006) using linear regression are included in Table 2.

As far as the significance of individual parameters is concerned, the study revealed that the highest correlation and least value of RMSE were obtained for evaporation with air temperature, followed by wind speed and relative humidity (Table 2). By using M5 Model tree algorithm (Table 2), the results start improving when the combined effect of parameters is considered for predicting evaporation, and yields maximum correlation (0.951) with minimum of RMSE (0.941).

The relationship as obtained from the application of M5 Model Tree algorithm is:

$$E = 0.23T + 0.72U_2 + 0.85S_h + 0.03R_h - 12.83 \quad (2)$$

A plot between actual and predicted values (Fig. 2) by Eq. 2 shows closer with 1:1 Line and most of the values lie within 15% deviation. The results suggest better performance by M5 model tree approach in comparison to linear regression approach.

CONCLUSIONS

There are many available direct and indirect methods used to estimate evaporation. The results suggest the usefulness of M5 Model Tree based technique in accurate estimation of the evaporation as an alternative to the simple linear regression approach. Most of the predicted values with M5 Model Tree algorithm are lying near the 45° line and the scatter range is within ±15% line.

REFERENCES

Abtew, W. (2001). Evaporation Estimation for Lake Okeechobee in South Florida. *Irrigation Drainage Eng*., 127:140-147.

Anderson, M. E. and. Jobson, H. E (1982). Comparison of Techniques for Estimating Annual Lake Evaporation using Climatological Data. *Water Resources Res.*, 18:630-636.

Bhakar, S. R. and Singh, R. V. (2004). Effect of Climatic Parameters on Evaporation and Evapotranspiration. *J. Indian Water Res. Soc*., 24(1): 12-18.

Bruin, H. A. R. (1978). A Simple Model for Shallow Lake Evaporation. *Applied Meterol*., 17:1132-1134.

Chandra, A., Shrikhande, V. J. and Kulshreshtha, R. (1988). Relationship of Pan Evaporation with Meteorological Parameters. *J. Indian Water Resour. Soc*., 8:41-44.

Kadhane, R. L. and Tatewar, S. P. (2006). Development of Relationship of Climatic Parameters with Evaporation and Evapotranspiration. *J. Pradushan Nirmulan.*, 3:29-33.

Murthy, S. and Gawande, S. (2006). Effect of Metrological Parameters on Evaporation in Small Reservoirs 'Anand Sagar' Shegaon- A Case Study.

J. Prudushan Nirmulan.,3:52-56.

Quinlan, J. R. (1992). Learning with Continuous Classes. Proc. on Artificial Intelligence, World Scientific: Singapore, pp.343-348.

Senapati, P. C., Mishra, N. and Lal, R. (1985). Relationship between Pan Evaporation and Meteorological Parameters. *J. Indian Water Resour. Soc*., 5:27-32.

Singh, R.V., Chauhan, H. S. and Ali, A. B. M. (1981). Pan Evaporation as Related to Meteorological Parameters. *J. Agri. Engg*., 18:48-53.

Singh, V. P. and Xu, C. Y. (1997). Evaluation and Generalization of 13 Mass-Transfer Equations for Determining Free Water Evaporation. Hydrological Processes, 11:311-323.

Stewart, R. B. and Rouse, W. R. (1976). A Simple Method for Determining the Evaporation from Shallow Lakes and Ponds. Water Resour. Res., 12:623-627.

Terz, O. and Keskn, M. E. (2005). Modeling of Daily Pan Evaporation. *Appled Sc*., 5:368-372.

Witten, I. H. and Frank, E. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kan Finann, San Francisco.

Received : May 2007; Accepted : December 2007