



MLP-ANN and SMN-ANN based models for sediment yield prediction in upper Ganga basin (India)

Saurabh Singh¹ and Akhilesh Kumar²

¹Research Scholar, ²Professor

^{1,2}Department of Soil and Water Conservation Engineering,

Govind Ballabh Pant University of Agriculture & Technology, Pantnagar, 263145, Uttarakhand, INDIA.

Abstract: The estimation of sediment yield was investigated by using multi-layer perceptron (MLP) and single multiplicative neuron (SMN) based artificial neuron network models. The feed-forward back propagation network is adopted with the fastest MLP training algorithm that is Levenberg-Marquardt algorithm and learning function gradient descent for model development. Different combinations of transfer function and number of neurons in the hidden layer were adopted. It was observed that for MLP-ANN a model structure with 3 neurons with 16 epochs and tansig as an activation function while for SMN-ANN with single neuron and logsig as an activation function with learning rate ($\alpha=0.5$) and 90 epochs resulted in best performance. The performance of models was evaluated by using statistical indices such as R, RMSE, IOA, ISE, CE and VE. The values of these performance indices for MLP were obtained as 0.961 & 0.981; 0.841 & 1.744 tons/sec; 97.9% & 98.6%; 0.024 & 0.08; 92.4% & 95.1% and -6.596% & 1.565% where as for SMN these values were obtained as 0.934 & 0.983; 1.099 & 1.459 tons/sec; 96.2% & 99.1%; 0.032 & 0.073; 86.9% & 96.6% and -2.179% & -2.097% respectively during the training and cross validation.

Keywords: Multi-layer perceptron; single multiplicative neuron; sigmoidal function.

I. Introduction

The Artificial Neural Network (ANN) approach comprises linear and non-linear concepts in model building, and can be operated with dynamic or memory less input-output system. The ASCE Task Committee recommended ANNs for their application in mapping from one multivariate space to another without providing the physics of the process. It has the following major advantages as: i. An ANN model does not require a prior knowledge of the system and, therefore, can be applied to solve the problems not clearly defined. ii. The model has more tolerance to noise and incomplete data, and thus, requires less data for model development and the results are the out-come of the collective behaviour of data, and thereby, the effect of outlier is minimized. iii. In ANN, the gradient descent search optimization embedded with back propagation algorithm is capable of exploring diverse areas. Agarawal *et al.* carried out the study of sediment yield using back propagation ANN model at Vamsadhara river basin, India. Jha and Jain investigated the use of ANNs in rainfall-runoff modelling in Kentucky River basin, USA. Raghuwanshi *et al.* investigated the performance of ANN in predicting runoff and sediment yield at Upper Siwane River, India. Hydrologic applications of ANN include the modelling of daily rainfall-runoff-sediment yield process, snow-rainfall process, assessment of stream's ecological and hydrological responses to climate change, rainfall-runoff forecasting, ground water quality prediction and ground water remediation. ASCE (2000A;B), Tokar (2000), Rajurkar (2004) among others provided a good overview of the ANN application to rainfall-runoff simulation and prediction.

Recently, single multiplicative neuron (SMN) based ANN model has been applied as an effective alternative to the general MLP type ANN. Yadav *et al.* (2007) applied SMN-ANN approach to develop a business model which involves only one neuron in the hidden layer and provides less complex network architecture and eliminates the hit and trial process for selection of appropriate number of neurons in the hidden layer as in case of multilayer perceptron based ANN models.. The same approach has been used by Singh (2014) to model rainfall-runoff process in a hilly watershed in India. In this study, therefore the SMN-ANN approach was tried to model runoff-sediment yield process from upper Ganga basin and to compare it with MLP-ANN approach.

II. Material and methods

The Ganga River basin, largest of the basins of India with an area of 8,61,452 Sq.km in India, draining into the 11 states of the country, Uttarakhand, Uttar Pradesh, Haryana, Himachal Pradesh, Delhi, Bihar, Jharkhand, Rajasthan, Madhya Pradesh, Chhattisgarh and West Bengal. The upper Ganga River basin with an area of 2, 48,000 Km² has been considered for this study. The Ganga basin lies between east longitudes 73°2' to 89°5' and north latitudes 21°6' to 31°21' having maximum length and width of approx. 1,543 km and 1024 km. The

average water resource potential of the basin has been assessed as 525020 Million Cubic Meters (MCM). Climate is not uniform over the Ganga basin; it varies from alpine, temperate, sub-tropical and tropical.

III. Collection of data

In the development of these models runoff and sediment yield data for 31 years were used. The ten-daily discharge and sediment load data were obtained for the period of 1980-2011 at Kanpur station (Station Code: GG000G7) from Middle Ganga Division-II, Central Water Commission, Lucknow.

IV. Inputs to models

The ANN models used in this study are multi layer feed forward back propagation networks with one input layer, one hidden and one output layer. The SMN-ANN model used single neuron with different parameters. The different models were developed by using discharge and sediment load data for Ganga river basin to forecast sediment yield from stream flow. The transfer function used in the hidden layer is non linear hyperbolic tan sigmoid whose output ranges between -1 and +1. Accordingly, in this study, the training data was rescaled and the inputs and targets patterns were standardized so that the all values fall in the range of -1 and 1. The expression used for standardization of the raw data series is given below;

$$x_n = 2 \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \right) - 1 \quad \dots(1)$$

The obtained output is then de-standardized to get the predicted target values in their original unit by using the following relationship.

$$x = 0.5 (x_n + 1) (x_{\max} - x_{\min}) + x_{\min} \quad \dots(2)$$

where, x_n is standardized input, x is actual data, $x_{\min}$ is minimum value of data vector and $x_{\max}$ is maximum value of the data vector.

V. Lag selection criteria

The runoff-sediment yield in a watershed being a dynamic process is affected by the watershed memory. In order to model the dynamic runoff-sediment process, determination of the appropriate lag of runoff and sediment yield data affecting the present day sediment yield from the watershed becomes very important. Therefore, the selection of most relevant lag of input data for building the model is very crucial to ensure model quality in terms of prediction, higher accuracy and efficiency.

Some researchers have suggested the use of cross correlation analysis technique to decide the number of inputs. Kisi (2007) used a correlation analysis for understanding the effects of preceding flows, whereas Sudheer *et al.* (2002) suggested a statistical procedure, based on cross-correlation, autocorrelation and partial autocorrelation properties of the series for identifying the appropriate input vector of the model. In the present study, the input data lag has been determined using cross-correlation, autocorrelation and partial autocorrelation values at 95% confidence level. Based on the above analysis a lag of 5days was found relevant to the study area. Therefore, the general form of the dynamic representation of runoff- sediment yield for the study area could be expressed as,

$$Y_t = A_0 + \sum_{i=1}^{i=6} A_i Q_{t+1-i} + \sum_{i=1}^{i=5} B_i Y_{t-i} \quad \dots(3)$$

where Q is the discharge in m^3/s , Y is the sediment load in tons/s, A_0 , A_i , and B_i are coefficients on i^{th} day and subscripts t is the present day.

VI. Model development

On the basis of autocorrelation, partial autocorrelation and cross-correlation analysis, a lag of 5days was found to be applicable in this study area. Accordingly, total number of variables in the model becomes 11. To reduce the number of variables of the model, and to make the model structure more compact and better representative of the runoff and sediment yield process, the input variables found significant at the 5% level of significance were retained, and the remaining variables not significant at this level were dropped, by applying the technique of stepwise regression. At first, the least significant variable was removed and the process was repeated until all the remaining variables were found to be significant at the 5% level. Based on the stepwise regression only six variables namely, Q_t , Q_{t-1} , Q_{t-2} , Q_{t-3} , Y_{t-1} , and Y_{t-4} were found to be affecting sediment yield on any particular day in the study area and accordingly were considered in the development of models. The model structure with relevant variables can thus be expressed as,

$$Y_t = A_0 + B_1 Y_{t-1} + B_4 Y_{t-4} + A_1 Q_t + A_2 Q_{t-1} + A_3 Q_{t-2} + A_4 Q_{t-3} \quad \dots(4)$$

i. Development of MLP-ANN model

An ANN is a highly interconnected network of many simple processing units called neurons, which are analogous to the biological neurons in the human brain. Neurons having similar characteristics in an ANN are arranged in groups called layers. The neurons in one layer are connected to those in the adjacent layers, but not

to those in the same layer. The strength of connection between the two neurons in adjacent layers is represented by what is known as a 'connection strength' or 'weight'. An ANN normally consists of three layers, an input layer, a hidden layer and an output layer. In a feed-forward network, the weighted connections feed activations only in the forward direction from an input layer to the output layer. The three-layer network also has connections from the input to all three layers. In this study, Tan sigmoid (tansig) transfer function was selected for forward back propagation networks to reach the optimized status.

In the present study the back propagation algorithm is used in multi layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. The back propagation algorithm uses supervised learning, which means that provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) was calculated. The idea of the back propagation algorithm was to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal was to adjust them to minimize the error. The model was developed with six variables as an input parameter for the sediment yield prediction. The selection of an optimum number of hidden layer and number of neurons in each hidden layer is very important in MLP-ANN modeling and it will result in better efficiency and comparatively less training of the network. Hecht-Nielsen (1987) reported that any continuous function can be represented by a neural network that has only one hidden layer with exactly $2n+1$ nodes, where n is the number of input nodes for a specific activation function. Accordingly, in this study the number of neurons has been considered in the range varying from 2 to $2n+1$ which satisfy the criteria for MLP.

As discussed above, in this study the number of neurons in the hidden layer has been considered in the range varying from 2 to 13 which satisfy all the criteria as the number of inputs are six and one output. In this way in total 12 model architectures were attempted for the selection of the best model architecture for MLP. The varying number of neurons in the hidden layer were trained using 1000 epochs on ten daily runoff-sediment series for the data of 30 years (1980-2010) whereas the ten daily data series of the year 2010-2011 was used for validation. Model performance was assessed based on performance indices both in training and validation to select best performing architecture. The MLP-ANN model having network architecture 6-3-1 as shown in (Fig1) yielded lowest RMSE and VE values for training and validation and finally selected for the sediment yield prediction.

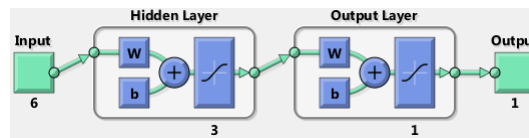


Fig1. MLP-ANN network architecture

ii. Development of SMN-ANN model

Single multiplicative neuron based artificial neural network (SMN-ANN) model was developed using log sigmoid transfer function as proposed by Yadav *et al.* (2007). In this model the weights and the biases are applied for each connection from the input to the Single multiplicative neuron. The schematic diagram of a generalized single multiplicative neuron is shown in Fig2.

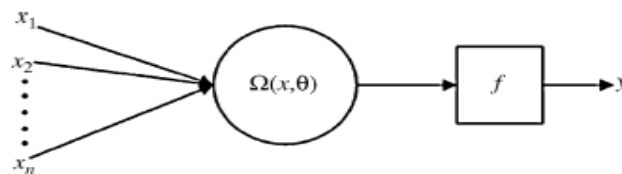


Fig2. A generalized single multiplicative neuron

Here, the operator Ω is a multiplicative operation with the weights w_i and biases b_i being the parameter Θ of the operator.

$$\Omega(x, \Theta) = \prod_{i=1}^n (w_i x_i + b_i) \quad \dots(5)$$

This approach involves only one neuron in the hidden layer and thereby provides less complex network architecture and eliminates the hit and trial process for selection of appropriate number of neurons in the hidden layer as in case of MLP based ANN models. In the present study selected learning rates of 0.1, 0.3, 0.5 and 0.7 were tried to study the effect of learning rate on the performance of the SMN-ANN model. The different combinations of learning rate and epochs were taken for the model development. The same criterion was used for the evaluation of best performing architecture as used in MLP-ANN model. The network architecture with single neuron, learning rate of 0.5 and 90 epochs yielded the lowest RMSE and VE values for training and validation and finally selected for the sediment yield prediction.

VII. Performance evaluation indices

A number of statistical criteria have been suggested by researchers to evaluate the performance of runoff-sediment models. To assess the accuracy of runoff-sediment models, more than one criterion should be used. The model performance was evaluated using following statistical parameters such as Correlation coefficient (R), Root mean square error (RMSE), Integral square error (ISE), Index of agreement (IOA), Coefficient of efficiency (CE) and Volumetric error (VE). The details of each criterion are described as:

1. Root mean square error (RMSE)

The root mean square error (RMSE) also called the root mean square deviation (RMSD), is often used to assess the prediction accuracy of a model. The root mean square error between observed and predicted values is determined by the following relationship.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{n}} \quad \dots(6)$$

2. Integral Square Error (ISE)

The integral square error, another measure of goodness of fit between the observed and predicted runoff is in fact proportion to the ratio of the root mean square error to the sum of observed runoff flow ordinate. The integral square error (ISE) is calculated by the following relationship as proposed by (Diskin *et al.* 1978).

$$ISE = \frac{\sqrt{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}}{\sum_{i=1}^n Q_i} \times 100 \quad \dots(7)$$

3. Index of agreement (IOA)

The Index of Agreement (d) developed by Willmott (1981) as a dimensionless standardized measure of the degree of model prediction error and varies between 0 and 1. A value of 1 indicates a perfect match, and 0 indicates no agreement at all (Willmott, 1981). It was not designed to be a measure of correlation but of the degree to which a model's predictions are error free. The index of agreement can detect additive and proportional differences in the observed and simulated means and variances; however, it is overly sensitive to extreme values due to the squared differences (Legates and McCabe, 1999).

Legates and McCabe (1999) highlighted a number of deficiencies with relative measure such as CE and R². They note that R² is particularly sensitive to outliers and insensitive to additive and proportional differences between modeled and observed data. Index of Agreement (d) is determined by using the following relationship.

$$d = 1 - \frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^n \left(\left| \hat{Q}_i - \bar{Q} \right| + \left| Q_i - \bar{Q} \right| \right)^2} \quad \dots(8)$$

4. Coefficient of efficiency (CE)

The Coefficient of Efficiency (CE) developed by Nash and Sutcliffe (1970) is an improvement over R² statistic which is used very commonly in hydrology. It gives the proportions of variance of the observation accounted for by the model. It is expressed mathematically as,

$$CE = 1 - \frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^n (Q_i - \bar{Q})^2} \quad \dots(9)$$

Where, Q_i is the observed value of daily runoff, $\hat{Q}_i$ is the corresponding predicted value of daily runoff.

5. Volumetric error (VE)

This index is also known as absolute prediction error (Kachroo *et al.* 1992) and it is used to compare the performance of the model, when the same data is used for their development. It is estimated as follows.

$$VE = \left[\frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)}{\sum_{i=1}^n (Q_i)} \right] \times 100\% \quad (10)$$

VIII. Results and discussion

The analysis has been carried out to develop sediment yield model namely, MLP-ANN and SMN-ANN using MATLAB (R2013a) for the upper Ganga basin. The MLP-ANN model was developed based on learning function gradient descent with different iterations and transfer functions tan sigmoid. A three layer back propagation network model trained by Levenberg-Marquardt optimization algorithm is chosen for this study. In the development of SMN-ANN model, based on the single neuron in the hidden layer with a learning rate of 0.5 with different iterations and log sigmoid as a transfer function was considered.

The developed models for predicting sediment yield were evaluated for their performance using statistical indices during training and cross validation period. The performance of developed models during training and cross validation is presented in Table I. From Table I, it was observed that the values of R for the multi-layer perceptron based ANN model and single multiplicative neuron based ANN model during the training period were 0.961 and 0.934 respectively whereas in cross validation period are 0.981 and 0.983 respectively. Kachroo *et al.* (1992) reported that a model can be considered satisfactory if R value exceeds 90 per cent and considered fairly good for R in the range of 80 per cent to 90 percent. This indicates that model's performance is very good in predicting the sediment yield during training and cross validation.

Table I. Values of performance indices of developed MLP-ANN and SMN-ANN models of upper Ganga basin for sediment yield prediction

	MLP-ANN		SMN-ANN	
	Training Period (1980-2010)	Validation Period (2010-2011)	Training Period (1980-2010)	Validation Period (2010-2011)
R	0.961	0.981	0.934	0.983
RMSE	0.841 tons/sec	1.744 tons/sec	1.099 tons/sec	1.459 tons/sec
IOA	0.979	0.986	0.962	0.991
ISE	0.024	0.08	0.032	0.073
CE	0.924	0.951	0.869	0.966
VE	-6.596%	1.565%	-2.179%	-2.097%

The values of other indices such as RMSE for both the models were 0.841 and 1.744 tons/sec respectively, under training phase whereas, 1.099 and 1.459 tons/sec respectively, under cross validation phase. The minimum values of RMSE indicate the model performance is good and higher values reduce the performance. The RMSE values are in acceptable range. It was also observed that the other performance indices such as IOA, ISE, CE and VE were 97.9% and 96.2%, 0.024 and 0.032, 92.4% and 86.9%, -6.596% and -2.179% respectively, under training period whereas, for cross validation period were 98.6% and 99.1%, 0.08 and 0.073, 95.1% and 96.6%, 1.565% and -2.097% respectively. It can also be seen from the Fig 3 that comparison of observed and estimated sediment yield for both the models during the validation period performing very well and are in acceptable range.

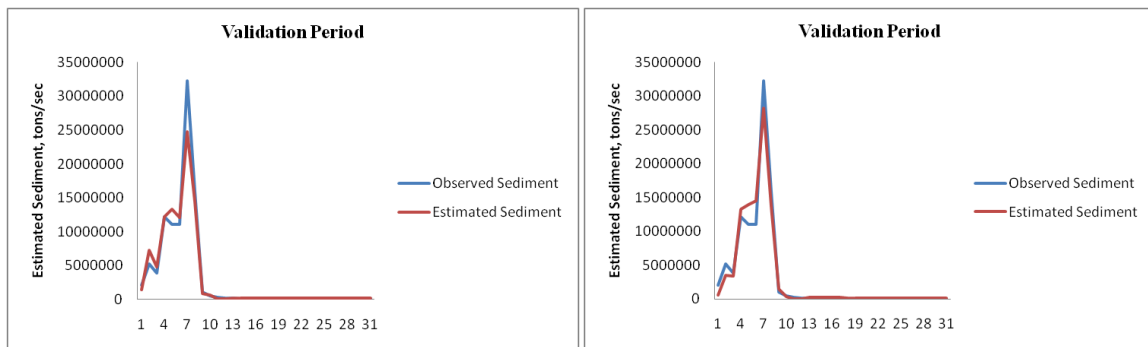


Fig. 3. Comparison of Observed and Estimated Sediment Yield for MLP-ANN and SMN-ANN models during Validation Period

IX. Conclusions

The present study was conducted to predict the sediment yield from the Ganga River. The feed forward back propagation model was developed by using the discharge and sediment load as an input parameter of lag of 5 days. The best combination of different neurons, learning rate and transfer function were tried using the MATLAB software (R2013a) and analyzed on the basis of different statistical indices. The performance of models was evaluated by using statistical indices such as R, RMSE, IOA, ISE, CE and VE. The values of these performance indices for MLP were obtained as 0.961 & 0.981; 0.841 & 1.744 tons/sec; 97.9% & 98.6%; 0.024 & 0.08; 92.4% & 95.1% and -6.596% & 1.565% where as for SMN these values were obtained as 0.934 & 0.983; 1.099 & 1.459 tons/sec; 96.2% & 99.1%; 0.032 & 0.073; 86.9% & 96.6% and -2.179% & -2.097% respectively during the training and cross validation. The values of these performance indices were well within acceptable range and the model estimated values of sediment yield compare well with their respective values of observed sediment yield. Hence, it is concluded that both these models are satisfactorily applicable in the study area. However, it was observed that SMN-ANN model which is providing slight edge on MLP-ANN model also

being based on single neuron is easier to develop as it eliminates the process of appropriate neuron selection in the model.

References

- [1]. Agarwal, A., Singh, R. D., Mishra, S. K., and Bhunya, P. K. 2005. ANN-based sediment yield models for amsadhara river basin (India). *Water SA*, 31(1): p-95.
- [2]. ASCE. 2000A. Task committee on Application of Neural Networks in hydrology. Artificial neural networks in Hydrology I: Preliminary concepts. *Journal of Hydrologic Engineering*, 5(2): 115-123.
- [3]. ASCE. 2000B. Task committee on Application of Neural Networks in Hydrology. Artificial neural networks in Hydrology II: Hydrology application. *Journal of Hydrologic Engineering*, 5(2): 124-137.
- [4]. Gharde, K. D., Kothari, M., Mittal, H. K., Singh, P. K., and Dahiphale, P. A. 2015. Sediment Yield Modelling of Kal River in Maharashtra Using Artificial Neural Network Model. *Research Journal of Recent Sciences*, 4: 120-130.
- [5]. Hecht-Nielsen, R., 1987. Kolmogorov's mapping neural network existence theorem. In *IEEE First Annual International Conference on Neural Networks*. 3: 11-13
- [6]. Kisi, O. 2007. Streamflow forecasting using different artificial neural network algorithms. *Journal of Hydrologic Engineering*, 12(5): 532-539.
- [7]. Kumar, A., and Das, G. 2000. Dynamic Model of Daily Rainfall, runoff and sediment yield for a Himalayan Watershed. *Journal of agricultural engineering research*, 75(2): 189-193.
- [8]. Raghuwanshi, N. S., Singh, R., and Reddy, L.S. 2006. Runoff and Sediment Yield Modeling Using Artificial Neural Networks: Upper Siwane River, India. *Journal of Hydrologic Engineering*, 11(1): 71-79.
- [9]. Rajurkar, M.P., Kothari, U.C., and Chaube, U.C. 2004. Modeling of the daily rainfall-runoff relationship with artificial neural network. *Journal of Hydrology*, 285(1): 96-113.
- [10]. Singh, P.V., Kumar, A., Rawat, J.S., and Kumar, D. 2013. Artificial neural networks based daily rainfall-runoff model for an agricultural hilly watershed. *International Journal of Engineering and management sciences*, 4(2): 108-112.
- [11]. Sudheer, K. P., Gosain, A. K., and Ramasastri, K. S. 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological processes*, 16(6): 1325-1330.
- [12]. Tokar, A.S. and Markus, M. 2000. Precipitation-runoff modeling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering*, 5(2): 156-161.
- [13]. Yadav, R. N., Kalra, P. K., and John, J. 2007. Time series prediction with single multiplicative neuron model. *Applied soft computing*, 7(4): 1157-1163.