COMPARISON OF LEARNING ALGORITHMS WITH THE ARIMA MODEL FOR THE FORECASTING OF ANNUAL RAINFALL IN TAMILNADU

NIRMALA .M[1], SUNDARAM .S.M[2]
[1]Research Scholar, Department of Mathematics, Sathyabama University, Chennai, INDIA
[2]Department of Mathematics, Sathyabama University, Chennai, INDIA
Email: [1]monishram5002@gmail.com

Abstract

Rainfall is one of the most complex and difficult elements of the hydrology cycle to understand and to model due to the tremendous range of variation over a wide range of scales both in space and time. The complexity of the atmospheric processes that generate rainfall makes quantitative forecasting of rainfall an extremely difficult task. Thus, accurate rainfall forecasting is one of the greatest challenges in operational hydrology, despite many advances in weather forecasting in recent decades. In this article, a data set containing the monthly rainfall of Tamilnadu for a period of 136 years (1871 - 2006) was analysed and the prediction of annual rainfall of Tamilnadu was made through Box-Jenkins ARIMA model, traditional statistical time series forecasting model and some of the learning algorithms of artificial neural networks. The results are compared using the error measure MAPE. It is found that the Radial Basis Function learning algorithm gives better prediction when compared to the other models discussed in this paper.

Keywords : ARIMA, Learning Algorithm, MAPE.

I. INTRODUCTION

Weather forecasting is one of the most imperative and demanding operational responsibilities carried out by meteorological services all over the world. Several stochastic models have been attempted to forecast the occurrence of rainfall, to investigate its seasonal variability, to forecast monthly or yearly rainfall of a region. This paper evaluates the soft computing techniques and conventional statistical approach

based on Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) model to predict the annual rainfall in Tamilnadu.

II. LITERATURE SURVEY

The association between the southwest and northeast monsoon rainfall over Tamilnadu have been examined for 100 years period from 1877-1976 through a correlation analysis by O.N. Dhar and P.R. Rakhecha [13]. The average rainfall series of Tamilnadu for the northeast monsoon months of October to December and the season as a whole were analysed for trends, periodicities and variability using standard statistical methods by O. N. Dhar, P. R. Rakhecha, and A. K. Kulkarni [15]. The local and teleconnective association between Northeast Monsoon Rainfall (NEMR) over Tamilnadu and global Surface Temperature Anomalies (STA) were examined using the monthly grid-ded STA data for the period 1901-2004 by Balachandran S., Asokan R. and Sridharan S [3]. The trends, periodicities and variability in the seasonal and annual rainfall series of Tamilnadu were analysed by O. N. Dhar, P. R. Rakhecha, and A. K. Kulkarni [14]. The5 prediction of annual rainfall over Tamilnadu have been made by using AutoRegressive Integrated Moving Average method by M. Nirmala and S.M. Sundaram [12].

III. STUDY AREA

Tamilnadu stretches between 8o 5' N and 13o 35' N latitudes and between 78o 18' E and 80o 20' E longitudes. Tamilnadu is in the southeastern portion of the Deccan which extends from the Vindhya mountains in the north to Kanyakumari in the south. It is bounded on the north by the states, Andhra Pradesh and Karnataka, on the east by the Bay of Bengal, on the south by the Indian Ocean and

Fig. 1. Map of Tamilnadu

In India, the state Tamilnadu receives rainfall in both southwest and northeast monsoons. Agriculture is more dependent on the northeast monsoon. Important crop seasons such as Samba, Thaladi and Navarai depend on the northeast monsoon. Hence, the rainfall during October to December plays an important role in deciding the fate of the agricultural economy of the state. Another important agro – climatic zone is the Cauvery river delta zone, which depends on the southwest monsoon. The period October to December is referred to as Northeast Monsoon season over peninsular India. Northeast Monsoon season is the major period of rainfall activity over south peninsula, particularly in the eastern half comprising of the meteorological Tamilnadu. This is the main rainy season accounting for about 48% of the annual rainfall. Coastal districts of the State get nearly 60% of the annual rainfall and the interior districts get about 40-50% of the annual rainfall. Tamilnadu should normally receive 979 mm of rainfall every year. Approximately 33% is from the southwest monsoon and 48 % is from the northeast monsoon.

IV. MATERIALS AND METHODS

A dataset containing a total of 136 years (1871 - 2006) monthly rainfall totals of Tamilnadu was obtained from Indian Institute of Tropical Meteorology (IITM), Pune, India. The annual rainfall totals of 136 years were calculated from the dataset. The data set was splitted into a training data (1871 - 1980) consisting of 110 years and a test data (1981 - 2006) with 26 years.

A. Box-Jenkins ARIMA Model

A powerful model for describing stationary and non - stationary time series is Autoregressive Integrated Moving Average process (ARIMA) of order (p, d, q). A simple regression model can be represented as

$$Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + \ldots\ldots + b_p Y_{t-p} + e_t \quad [1]$$

where Y_t is the forecast variable, $Y_{t-1}, \ldots\ldots, Y_{t-p}$ are explanatory variables and e_t is the error term. The name Auto Regression is used to denote the above equation due to the time-lagged values of the explanatory variable. The moving average model uses the past errors as the explanatory variables. A simple moving average model is represented as follows:

$$Y_t = b_0 + b_1 e_{t-1} + b_2 e_{t-2} + \ldots\ldots\ldots + b_p e_{t-p} + e_t \quad [2]$$

The autoregressive model and the moving average model are efficiently coupled to form a general and useful class of time series model called Autoregressive Moving Average Model (ARMA). This class of models can be extended to non-stationary time series by differencing the data series. These models are called Box-Jenkins Autoregressive Integrated Moving Average Models [7].

B. Artificial Neural Network

An Artificial Neural Network is a computational model inspired by biological neural networks both structurally and functionally. It consists of a group of interconnected computation units called neurons. A neuron takes inputs and transforms them into certain form of output based on its specific activation function. Different types of Artificial Neural Networks take variant forms of activation functions and inter-neuron connections to model the underlying complex mathematical relationships between the inputs and the outputs of different types of systems. Numerous Artificial Neural Network models have been developed for time series prediction. Among them, the two major types of Artificial Neural Networks are Multi-Layer Perceptron (MLP) networks and Radial Basis Function (RBF) networks. Multi-Layer Perceptron and Radial Basis Function are feed forward networks. These layered networks have been applied to various problems such as pattern recognition, prediction, and function approximation. Multi-Layer Perceptron networks and Radial Basis Function networks conduct off-line iterative learning over a set of stored training data. Because the learning process has full control of the training data set, many methods like batch learning, multiple random initialization, etc. can be used to optimize and speedup the learning towards the minimum of the objective function, which is some form of error evaluation. Artificial Neural Networks do not have any stationarity constraint on the time series to be learned and predicted.

A multilayer perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that. It can distinguish data that is not linearly separable, or separable by a hyperplane. The Multilayer feedforward network consist of a set of sensory units that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input signal propagates through the network in a forward direction, on a layer-by-layer basis. Mathematically, a multilayer perceptron network is a function consisting of composition of weighted sum of the functions corresponding to the neurons

The Backpropagation training algorithm was first described by Rumelhart and McClelland in 1986. It was the first practical method for training neural networks. Because the error information is propagated backward through the network, this type of training method is called backward propagation. The idea of Backpropagation was to minimize some output error function by easily computable updates of weights. In this algorithm, at first step, the input and desired output are identified. Then an arbitrary weight vector w_0 is initialized. Then the feed forward neural network is iteratively adopted according to the recursion formula,

$$w_{k+1} = w_k + \eta\, d_k \qquad [3]$$

where w_k, denotes the weight matrix at epoch k. The positive constant is called the learning rate. The direction vector, d_k is negative of the gradient of the output error function,

$$d_k = -\nabla E(w_k). \qquad [4]$$

Batch learning and On-line learning are the two standard learning methods in Backpropagation algorithm. In batch learning, all the pairs consisting of input and target patterns in the training set are treated as a batch and after training the training set, the network is updated. But whereas in on-line learning, the weights of the network are updated immediately after the presentation of each pair of input and target patterns.

The purpose of neural network training is to produce appropriate output patterns for corresponding input patterns. It is achieved by an iterative learning process that updates the neural network weights based on the neural network response to a set of training input patterns. Learning can be categorized as supervised, reinforcement, unsupervised or hybrid. Supervised learning occurs when the correct output pattern is known and used during training. In reinforcement learning the correct output is not known but a measure of the correctness of a network output response can be computed. Unsupervised learning does not require a correct output to be available during training. Different learning rules form the basis of different training algorithms and their applicability is dependent on the neural network architecture and the learning category being used.

Gradient Descent Algorithm

The original procedure used in the gradient descent algorithm is to adjust the weights towards convergence using the gradient. It uses a gradient search technique to minimize a cost function equal to the sum square difference between desired and estimated net outputs. Derivatives of error (called delta vectors) are calculated for the network's output layer, and then backpropagated through the network until delta vectors are available for each hidden layer of the network.

Scaled Conjugate Gradient Algorithm

A search is made along conjugate directions rather in steepest descent directions for faster convergence in Conjugate Gradient Algorithm. Then a line search is performed to determine the optimal distance to move along the current search direction. In Scaled Conjugate Gradient Algorithm, the time consuming line search at each iteration is avoided by combining the model trust region approach in conjugate Gradient Algorithm.

Radial Basis Function

Radial Basis Function network is also a feedforward network but has only one hidden layer. It uses radial combination functions in the hidden layer, based on the squared Euclidean distance between the input vector and the hidden vector. Radial Basis Function is a kernel function that is symmetric with respect to the origin. Hence its variable is the norm distance from the origin. In Neural Networks, Radial basis functions have been applied as a replacement for the sigmoidal hidden layer transfer characteristic in Multi-Layer Perceptron. The Gaussian Radial Basis function uses exponential or softmax activation function in the hidden layer. Radial Basis Function networks have the advantage that they do not suffer from local minima as Multi-Layer Perceptron do, since the only parameter that are adjusted in the learning process are the linear mapping from hidden layer to output layer. It has any number of inputs and any number of outputs with any activation function. It has connections between the input and the hidden layer and connections between the hidden and the output layer.

V. RESULTS AND DISCUSSIONS

A. Box-Jenkins ARIMA Model:

A good starting point for time series analysis is a graphical plot of the data.

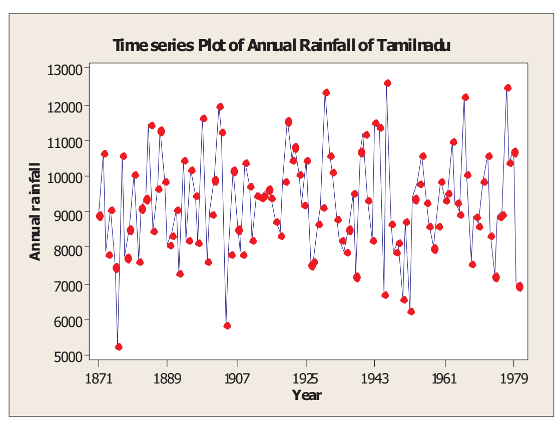

Fig. 2. Time series plot of Annual rainfall of Tamilnadu

To check the normality of the rainfall series, Anderson – Darling test is made. Anderson-Darling test compares the empirical cumulative distribution function of the data with the distribution expected if the data is normal. If this observed difference is sufficiently large, the test will reject the null hypothesis of population normality. The test statistic value is 0.155 and p > 0.01. This shows that the annual rainfall series of Tamilnadu is normal.

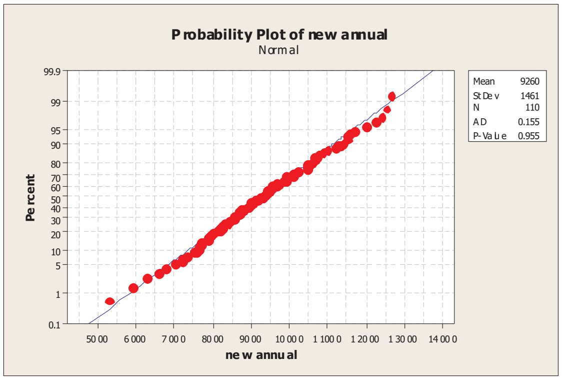

Fig. 3. Normality plot of Annual Rainfall in Tamilnadu

The autocorrelation coefficient is a valuable tool for investigating properties of an empirical time series. Autocorrelations can be used to determine whether there is any pattern like AR, MA, ARMA and ARIMA in the time series. The absence of such patterns shows that the time series is random.

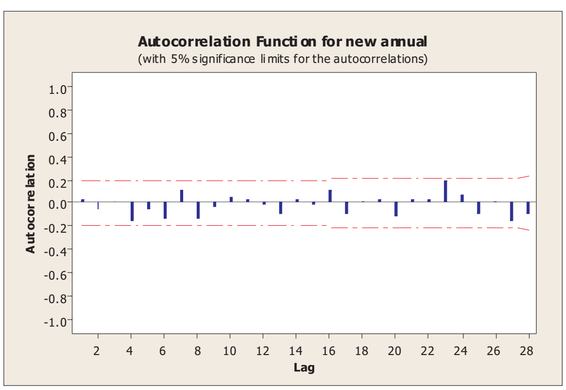

Fig. 4. The autocorrelation function of the annual rainfall in Tamilnadu

Seasonality is defined as a pattern that repeats itself over fixed intervals of times. For a stationary data, seasonality can be found by identifying those autocorrelation coefficients of more than two or three time lags that are significantly different from zero. The autocorrelation function of the annual rainfall of Tamilnadu for first five lags is given table1 and the graph is shown in fig.4. Since the time series considered here has the autocorrelation coefficients which are near to zero, the Tamilnadu annual rainfall series has no seasonality.

Table 1. Autocorrelation function coefficient for first five lags

Lag	ACF	T
1	0.031102	0.33
2	-0.055561	-0.58
3	0.011549	0.12
4	-0.154584	-1.61
5	-0.040724	-0.42

Box – Jenkins ARIMA Model identification can be made through the autocorrelation function and partial autocorrelation function plots. The autocorrelation plot shows how values of the time series are correlated with the past values of the series and the partial autocorrelation measures the degree of association between yt and yt+k when the effect of other time lags 1,2,3… k – 1 are somehow removed. The final part of model identification is the check for white noise. This is an approximate statistical test of the hypothesis that no autocorrelation of the time series up to the given lag are significantly different from zero.

For the annual rainfall in Tamilnadu, since the autocorrelation function dies out rapidly, that is, it reaches zero within one or two lag periods, the time series is stationary. Hence there is no need of differencing. Therefore d = 0. The partial autocorrelation function decays exponentially and the corresponding autocorrelation function coefficients are non-zero for first 4 lags, then it suggests that p = 0 and q = 4. For the partial autocorrelation function plot and the corresponding coefficients, refer figure 5 and table 2 and for autocorrelation function plots and the corresponding coefficients refer figure 4 and table 1. Hence the suitable model for predicting the annual rainfall in Tamilnadu is ARIMA (0,0,4).

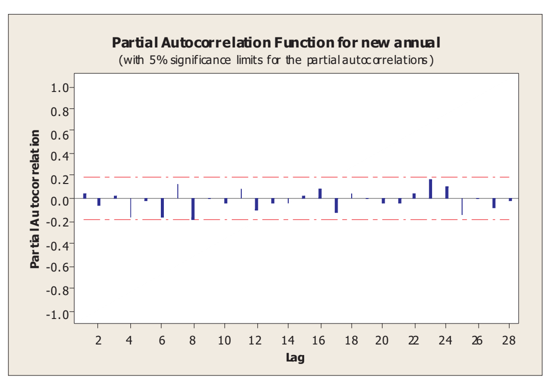

Fig. 5. The partial autocorrelation function of the annual rainfall in Tamilnadu

Table 2. Partial autocorrelation function coefficient for first five lags

Lag	PACF	T
1	0.031102	0.33
2	-0.056583	-0.59
3	0.015198	0.16
4	-0.159318	-1.67
5	-0.029139	-0.31

The following table 3 shows that the chi-square values at different lag periods and the corresponding p values are greater than 0.05. This is to check for white noise. Since p values are greater than 0.05 the hypothesis that that no autocorrelation of the time series up to the given lag are significantly different from zero is true. Using the Schwartz Bayesian Criteria, the suitable ARIMA model was chosen. Hence the model, ARIMA (0,0,4) is a suitable model. The stationary R − squared value obtained while fitting this model is 0.077 and the corresponding R − squared value is 0.077. The Modified Box-Pierce (Ljung-Box) Chi-Square statistic values are given in the following table 3.

Table 3. Chi-square statistic values for different lag periods

Lag	12	24	36	48
Chi-Square	6.3	18.1	23	30.3
DF	7	19	31	43
P-Value	0.509	0.518	0.848	0.928

The Mean Absolute Percentage Error (MAPE) measure is given by

$$MAPE = 100 \times \frac{\sum_{t=1}^{N} \frac{|E_t|}{Y_t}}{N} \qquad [5]$$

where Y_t and E_t represent desired outputs and corresponding errors at t = 1,2,...,N respectively. For the model ARIMA (0,0,4), the error measure, Mean Absolute Percentage Error is 11.7706 for the training data and 14.8981 for the testing data which shows that the fit is a good fit.

B. Artificial Neural Network Training

The objective of the training is to find the set of weights between the neurons that determines the global minimum of the error function. Training involves the following steps: Deciding the number of iterations, selection of learning rate and the momentum values. In this paper, three predictors for the year Y are used to predict the annual rainfall in Tamilnadu in the year (Y + 1). The three predictors are the Tamilnadu Rainfall amount in the Northeast Monsoon, Tamilnadu Rainfall amount in the Southwest Monsoon and the annual Rainfall in Tamilnadu. For developing an Artificial Neural Network model, the whole dataset is divided into training set consisting of 110 years (1871 - 1980) and testing set consisting of 26 years (1981 - 2006).

First, using Gradient Descent Algorithm, a three layered architecture was constructed with three units in the input layer, three units in the hidden layer and one unit in the output layer. Activation functions are mathematical formulae that determine the output of the processing unit. Each input takes its net input and applies an activation function on it. Transfer functions such as sigmoid are commonly used since they are nonlinear and continuously differentiable which are required for network learning. In this algorithm, the sigmoidal function is chosen in the hidden and the output layers as the activation function. The sigmoid function is given by

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

This transfer function produces an output between 0 and 1. Thus the output is a smooth function of the inputs. Every cycle in which each one of the training patterns is presented once to the neural network is called an epoch. The data was trained upto 1000 epochs. After the training process, the Neural Network was tested over the test set. The mean absolute percentage error obtained in both training and testing sets are given in table 5. The architecture of the Artificial Neural Network trained using Gradient Descent Algorithm for the annual rainfall in Tamilnadu is shown in fig 6.

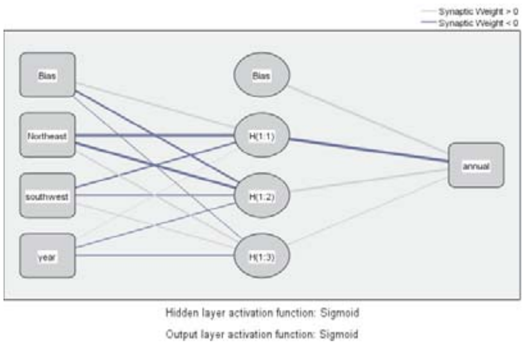

Fig. 6. Architecture of an ANN model using Gradient Descent Algorithm for Tamilnadu annual rainfall series

Table 4. Table showing the parameters of the various learning algorithms of ANN

Parameters	SCG	GDA	RBF
Number of Layers	3	3	3
Number of Input Nodes	3	3	3
Number of Output Nodes	1	1	1
Number of Hidden Nodes	3	3	6
Activation function in the Hidden Layer	Sigmoid	Sigmoid	Softmax
Activation function in the Output Layer	Sigmoid	Sigmoid	Identity

Secondly, the Artificial Neural Network was trained through the scaled conjugate gradient algorithm with a three layered architecture. The number of units in the input and the hidden layers are three whereas in the output layer it is one. The activation function in the hidden and the output layers was sigmoidal. The data was trained upto 1000 epochs. The training was tested using the test data set. The results of the error measure, MAPE obtained in the training and testing are tabulated in table 5.

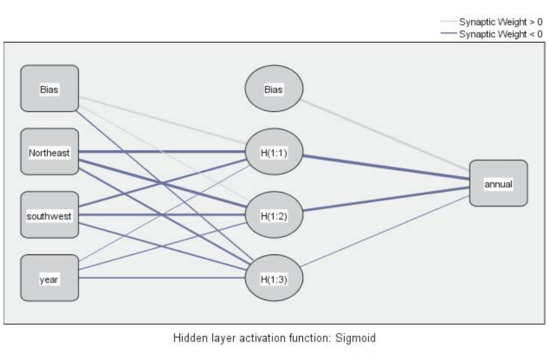

Fig. 7. Architecture of an ANN model using Scaled Conjugate Gradient Algorithm for Tamilnadu annual rainfall series.

Finally, the Artificial Neural Network was trained with the Radial Basis Function algorithm. The three layered architecture of Radial Basis Function consists of three input units, six hidden units and one output unit. After the training, it was tested with the test data set. The mean absolute percentage error obtained during training and testing is tabulated in table 5.

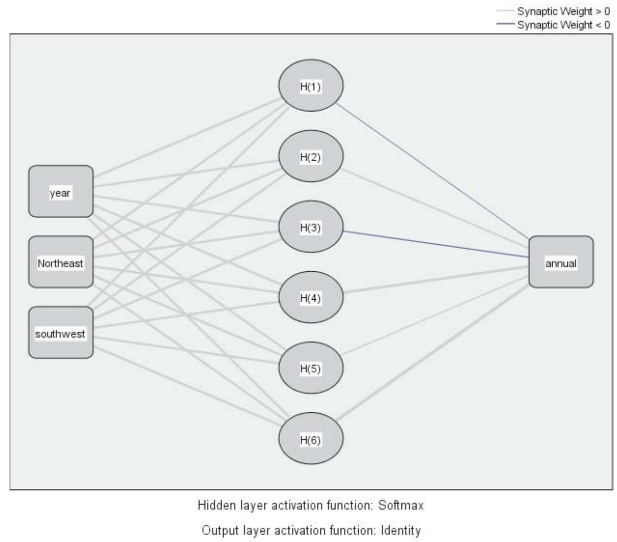

Fig. 8. Architecture of an ANN model using Radial Basis Function for Tamilnadu annual rainfall series

Error-correction learning rules in supervised learning uses the difference between the correct and actual output patterns to adjust connection weights with the aim of reducing this error. The following table shows the Mean Absolute Percentage Error measure obtained during the training of the network through the various learning algorithms and the corresponding testing error measures.

Table 5. Table showing the values of the error measure, MAPE of the models

Models of ANN	Training error (MAPE)	Testing error (MAPE)
GDA	5.1448	6.6061
SCG	5.352	6.5063
RBF	0.4526	0.3926
ARIMA	11.7706	14.8981

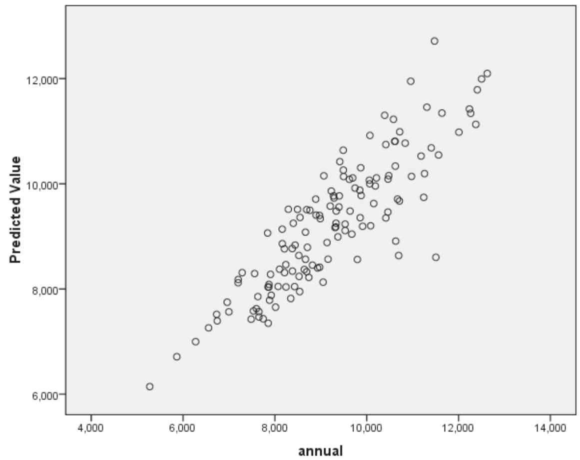

Fig. 9. Graph showing the predicted values of annual rainfall in Tamilnadu using Radial Basis Function.

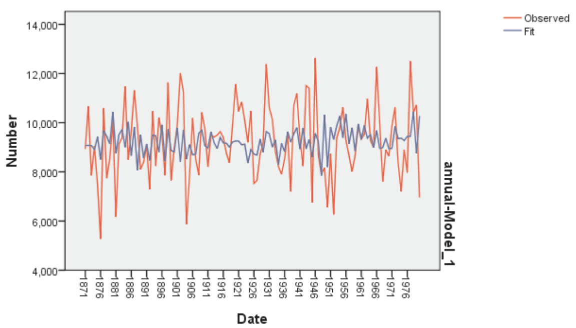

Fig. 10. Graph showing the observed and the fitted values of annual rainfall in Tamilnadu using ARIMA(0,0,4).

VI. CONCLUSION

Traditional statistical time series forecasting methods, including moving average, exponential smoothing, and Auto-Regressive Moving Average (ARMA), all assume stationarity of the time series. But Artificial Neural Networks do not have any stationarity constraint on the time series to be learned and predicted. It has highly flexible nonlinear regressive structure to fit the target pattern space. For some real world problems, Artificial Neural Network will never replace the existing conventional techniques but because of the fast growing applications it can be an alternative to those existing techniques. The Radial Basis Function algorithm gives better forecasting as the MAPE value is the least of the four techniques discussed, for the analysis of the annual rainfall in Tamilnadu. In contrast, ARIMA model which is supposed to be a powerful tool had the largest MAPE value despite satisfying so many constraints.

REFERENCES

[1] Abraham A, Philip NS and Joseph KB, (2001), Will we have wet summer? Soft Computing models for long – term rainfall Forecasting, Proceedings of 15th European Simulation Conference ESM2001, Prague.

[2] Anderson J.A. (2003) An Introduction to Neural Networks, Prentice Hall.

[3] Balachandran S., Asokan R. and Sridharan S (2006). Global surface temperature in relation to northeast monsoon rainfall over Tamilnadu, Journal of Earth System Science, vol. 115, no.3, pp. 349-362.

[4] Damodar N. Gujarati and Sangeetha (2007). Basic Econometrics, 4th edition,the Tata Mcgraw – Hill Publishing Company Limited.

[5] Fallah G-A , Mousavi M. Habiby M, (2008), Using Artificial neural network (ANN) for seasonal rainfall forecasting based on tele-connection patterns, Geophysical Research Abstracts, Vol. 10, EGU 2008 -A- 11664.

[6] Gadgil, S., Rajeevan, M. & Nanjundiah, R., (2005) Monsoon prediction – Why yet another failure? Current Science 88: 1389-1499.

[7] George E. P. Box, Gwilym M. Jenkins and Gregory C. Reinsel (1994). Time Series Analysis – Forecasting and Control, 3rd Edition, Pearson Education, Inc.

[8] Hung N Q, Babel M S, Thailand, Weesakul S, and Tripathi N K, An artificial neural network model for rainfall forecasting in Bangkok, Hydrology and Earth System Sciences discussions, vol. 5, pp.183–218, 2008.

[9] Khaing Win Mar and Thinn Thu Naing (2008), Optimum Neural Network architecture for Precipitation Prediction of M y a n m a r , Proceedings of World Academy of Science, Engineering and Technology, vol.36, December.

[10] Meaza Demissie (2003), Characterizing trends and Identification of rainfall predicting model: An Application to Melksa observatory Nazreth rainfall series, Proceedings of Biometric society, pp. 132 – 134.

[11]. Moller A F(1993), A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning, Neural Networks, Vol.6, pp.525 – 533.

[12]. Nirmala M and Sundaram S.M.(2009), Rainfall Predicting Model: An Application to Tamilnadu Rainfall Series, Proceedings of National Conference on Effect of Climatic Change and Sustainable Resource Management, SRM University and India Meteorological Department, Chennai.

[13]. O. N. Dhar, P. R. Rakhecha, A. K. Kulkarni (1982). Fluctuations in northeast monsoon rainfall of Tamilnadu, International Journal of Climatology Volume 2, Issue 4, Pages339 – 345.

[14]. O. N. Dhar, P. R. Rakhecha, A. K. Kulkarni (1982). Trends and fluctuations of seasonal and annual rainfall of Tamilnadu, Proceedings of Indian Academy of Sciences (Earth planetary Sciences), vol. 91, No.2, pp. 97 –104.

[15]. O. N. Dhar and P. R. Rakhecha (1983). Foreshadowing Northeast Monsoon Rainfall over Tamilnadu, Monthly Weather Review, vol.3, issue 1,pp. 109 - 112

[16]. Philip N S and Joseph K B, On the Predictability of Rainfall in Kerala: An application of ABF Neural Network, Proceedings of Workshop on Intelligent Systems Design and Applications, International Conference on Computational Sciences, ICCS 2001, San Francisco, May 2001.

[17]. Spyros Makridakis, Steven C. Wheelwright, Victor E. McGee. Forecasting – Methods and Applications , 2nd edition

[18] Sunyoung Lee, Sungzoon Cho and Patrick M. Wong (1998) Rainfall Prediction Using Artificial Neural Networks, Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 233 – 242.

[19]. Yegnanarayana, B (2000) Artificial Neural Networks, Prentice-Hall of India Pvt Ltd, New Delhi, India.

Nirmala .M, faculty Department of Mathematics - Sathyabama University. she has presented 3 papers in international conferences. Her research interests are Time Series analysis, Multi variate analysis and Statistical Modeling.