American Journal of Applied Sciences 10 (10): 1172-1180, 2013
ISSN: 1546-9239

doi:10.3844/ajassp.2013.1172.1180 Published Online 10 (10) 2013 (http://www.thescipub.com/ajas.toc)

Resemblance of Rain Fall in Bangladesh with Correlation Dimension and Neural Network Learning

[1]Abu Nasir Mohammad Enamul Kabir, [1]Hussain Muhammad Imran Hasan, [2]Mohd Abdur Rashid, [2]Azralmukmin Azmi, [1]Md. Zakir Hossain and [1]Md. Shahjahan

[1]Department of Electrical and Electronic Engineering,
Faculty of Electrical and Electronic Engineering,
Khulna University of Engineering and Technology, Khulna-9203, Bangladesh
[2]School of Electrical Systems Engineering,
University Malaysia Perlis, Pauh Putra Campus, 02000 Arau, Perlis, Malaysia

Received 2013-06-12, Revised 2013-08-26; Accepted 2013-08-28

ABSTRACT

Rain fall and Temperature are undoubtedly two important factors that balance water in the environment. Adequate study of the rain behavior helps to forecast it. The time series obtained from different stations of the country throughout the several years are collected and analyzed. The dynamics of rain fall time series is analyzed with Correlation Dimension (CD) to characterize the several zones of Bangladesh. In addition a Neural Network (NN) predictor model was designed to realize complexity of rain fall. We found the interesting similarity between CD and NN predictor. The findings are useful in explaining why several zones show behavioral regularity and change.

Keywords: Rain Fall, Time Series Analysis, Complexity, Neural Network, Learning and Prediction

1. INTRODUCTION

Adequate knowledge of the rainfall is needed for, among other things, (a) optimal design of water storage and drainage networks; (b) management of extreme events, such as floods and droughts; and (c) determination of the rate of pollution. Rainfall influences the social behaviour of human. For example, rainfall shocks around the time of birth have been shown to causally influence later health (Thai and Myrskyla, 2012). In addition rain fed crops substantially affects the classification of radar data acquisition (Larranaga *et al*., 2013). In Bangladesh, it is well-known that Cox's bazar and sylhet are very important zones for national and international travellers. However, these two zones are very much rain affected zone. Erratic behavior of rain fall often prevents travellers to travel. This study explains the reason why this two zones exhibit such behavior using chaos and neural network analyses. Bangladesh is tropical country. A large amount of rain fall is seen at different zones in Bangladesh. It is famous in the world as a most flooding country. There are a number of famous place including Cox's bazar beach and Sylhet hill tracks. Appropriate study of rain fall dynamics may help national and international people to determine their travel plan. Temperature influences evaporation and the precipitation phase (rain, snow) and therefore plays an important role in the water balance. Precipitation is characterised by both high spatial and temporal variability. Long-term forecasting can be done only in a stochastic way, due to the highly nonlinear relationships governing the rainfall dynamics. This renders long term deterministic forecast impossible. It is important to determine the attributes of rain behavior at different zones in this country (Kannan *et al*., 2010). These attributes therefore will be useful in charactering different zones of interest.

Corresponding Author: Mohd Abdur Rashid, School of Electrical Systems Engineering, University Malaysia Perlis, Pauh Putra Campus, 02000 Arau, Perlis, Malaysia

It is interesting to see rain behavior as a function of chaos. There are two properties in the rain-nonlinearity and repetitive periods. Therefore, there is a chaotic behavior. Though a chaotic system and a stochastic system may exhibit very similar, apparently random behaviour as exemplified in their time series, there is a crucial difference between them. The chaotic system is governed by deterministic dynamics and therefore, it can be modeled, its underlying mechanisms can be divined and pre-dictions can be made. However, in reality this is often very difficult to do as such systems contain a good deal of noise, which may be due to measurement errors. Nevertheless, merely identifying a system as chaotic presents a step forward as approximate predictions can then be made. Following the development of Chaos science, many complex natural systems have been identified or at least suspected to be chaotic, such as rainfall (Ghorbani *et al*., 2010). We here use correlation integral and correlation dimension in order to deterministic chaotic dynamics of rain fall. On the other hand, a standard neural network predictor is used in order to find how difficult the time sequence for the network. Neural network predictor is a model to predict the future value of rainfall using previous rainfall sequence. We used this as well to determine the complexity of rain fall. The daily rainfall time series of Bangladesh of varying record lengths, obtained from eight stations situated around it, are analyzed using chaos and neural network in order to investigate their complexity. In this study, we attempt to explain why different zones behave similarly and differently using available time series with correlation dimension and neural net-work.

2. MATERIALS AND METHODS

In order to find the regularity of time course behavior, one needs to analyze the underlying time series. We use two different tools. Firstly, the correlations dimension. This will indicate whether the time series is very complex or not. In this regard autocorrelation function is used for finding an optimal delay between two lagged time series. Secondly, neural network is used to find the complexity of the time series in terms of their output. We first briefly outline theoretical explanation of them.

2.1. Autocorrelation Function

Autocorrelation is the correlation of a data set with itself (Shumway and Stoffer, 2010). The autocorrelation can be used for the following two purposes (i) To detect non-randomness in data and (ii) To identify an appropriate time series model if the data are not random. Given measurements, Y1, Y2, ...,YN at time X1, X2, ..., XN, the lag k autocorrelation function is defined as Equation 1:

$$\tau_k = \frac{\sum_{i=1}^{N-k}\left(Y_i - \bar{Y}\right)\left(Y_{i+k} - \bar{Y}\right)}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2} \quad (1)$$

Although the time variable, X, is not used in the formula for autocorrelation, the assumption is that the observations are equi-spaced. Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times Xi and Xi+k. When the autocorrelation is used to identify an appropriate time series model, the autocorrelations are usually plotted for many lags. Therefore, we have plotted for many time lags and chosen the delay for first minimum value of autocorrelation.

2.2. Correlation Dimension

The Correlation Dimension (CD) (Acharya *et al*. 2009; Xingyuan *et al*., 2013) is a measure of the dimensionality of the space occupied by a set of random points, often referred to as a type of fractal dimension. This is computed with correlation integral described below. The deterministic chaos can be distinguished in a time series by various factors-the power spectrum should be continuous, the largest Lyapunov exponent is positive, the autocorrelation function is converged to zero at the infinite time, points in Poincare map be limited within a certain finite space and so on. Among them, the presence of the fractal dimension (a non-integer value) has been considered as the strong evidence for the presence of the deterministic chaos. The correlation integral is a metric invariant, which characterizes the metric structure of the attractor by quantifying the density of points in the phase space. It achieves this through a normalized count of pair of points lying within a radius r. formally, correlation integral C(r) is defined as Equation 2:

$$C(r) = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Theta\left(r - \left\|x_i - x_j\right\|\right) \quad (2)$$

where, Θ is the Heaviside function. Note that, xi in this case refers to a point in the phase space i.e., it corresponds to i-th row vector of X. In our experiments, we computed C(r) for a fixed value of r and used it as a feature vector.

Correlation dimension measures the change in the density of phase space with respect to the neighborhood radius r. In this study, the correlation integral C(r) of one-dimensional time series is calculated according to sphere counting method can be found in (Xingyuan *et al*., 2013). In order to estimate the correlation dimension m, we estimate the slope of the linear relation between log(C(r)) and log(r). The first step is to choose the scaling region; we do this by looking at plots of the log(C(r)) against log(r).

The scaling region is chosen to be the values of r for which the linear relationship between log(C(r)) and log(r) appears to hold. We estimate the slope of the li-near relationship using the least squares estimator. The presence (or absence) of chaos can be identified by plotting the correlation exponent values against the corresponding embedding dimension values. If the value of the correlation exponent is finite, low and non-integer, then the system is generally considered to exhibit low-dimensional chaos. The saturation value of the correlation exponent is defined as the correlation dimension of the attractor. The nearest integer above the saturation value is generally considered to provide the minimum number of phase-space or variables necessary to model the dynamics of the attractor.

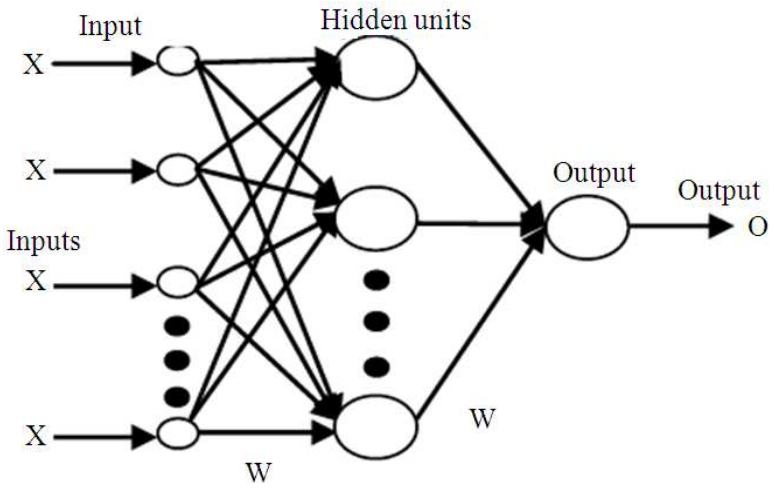

Fig. 1. Model of a fully connected feed forward neural network

1. Firstly a feed-forward network is created with n_{in} inputs, n_{hidden} hidden units, and n_{out} output units
2. Initialize weights to some small random values
3. Until the termination condition is met, Do

For each (input $\underline{x}$, target output $\underline{t}$), in training_examples, Do

Propagate the input forward through the network

a) Input the instance $\underline{x}$ to the network and compute the output o_u of every unit u in the network.

Errors are propagated to backward through the network

b) Error term ∂_q, for each output unit q is

$$\partial_q = o_q(1 - o_q)(t_q - o_q) \dots\dots\dots\dots\dots\dots (3)$$

c) Error term ∂_h, for each hidden unit h is

$$\partial_h = o_h(1 - o_h) \sum_{q \in outputs} w_{qh}\,\partial_q \dots\dots\dots (4)$$

d) For each weight, Do

$$w_{ji} = w_{ji} + \Delta w_{ji} \dots\dots\dots (5)$$

Where, $\Delta w_{ji} = \eta \partial_j x_{ji}$

Fig. 2. Weight updates process of SBP

The value of the embed-ding dimension at which the saturation of the correlation exponent occurs generally provides an upper bound on the number of variables sufficient to model the dynamics (Liang *et al*., 2013). Conversely, if the correlation exponent increases without bound with increase in the embedding dimension, then the system under investigation is generally considered as stochastic.

2.3. Artificial Neural Network

An artificial neural network is a mathematical model inspired by biological neural networks (Jaiyen *et al*., 2010) as shown in **Fig. 1**. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases, it is an adaptive system changing its structure during a learning phase. The researchers often apply it in many application domains like classification, pattern recognition, signal processing, weather prediction. The main advantages of using such a tool are the universal approximation and generalization ability. It can tolerate noise within a margin. Therefore, one can easily apply it to time series prediction. The outcome of the neural network will be the mean square error. One can consider if this error is smaller, then prediction task of that zone is easy because network easily predict that zone otherwise difficult. In the figure, we designate in-puts as X_1 to X_n and output as 'O' and weights as W. The weight from unit i to unit j is denoted W_{ji} and the weight from hidden unit j to output unit k is denoted W_{kj}.

Standard Back Propagation (SBP) is a supervised learn-ing method that is a generalization of the delta rule. It is most useful algorithm for feed-forward networks. It re-quires a dataset of the desired output for many inputs, making up the training set. The weights of the network are updated according to the input and output training data. A predictor model usually uses several points in the time series as input and several with time version as out-puts. For each (x, t), in training examples, propagate the input forward through the network (Kurokawa *et al*., 2011; Russell and Norvig, 2010). **Figure 2** shows how to update weights of a two layer network in a standard BP algorithm.

2.4. Data Collection and their Characteristics

We describe area and data taken and the analysis of the autocorrelation function and correlation dimension for different zones of Bangladesh. We tried to cover several greater zones throughout Bangladesh so that a complete picture may come out.

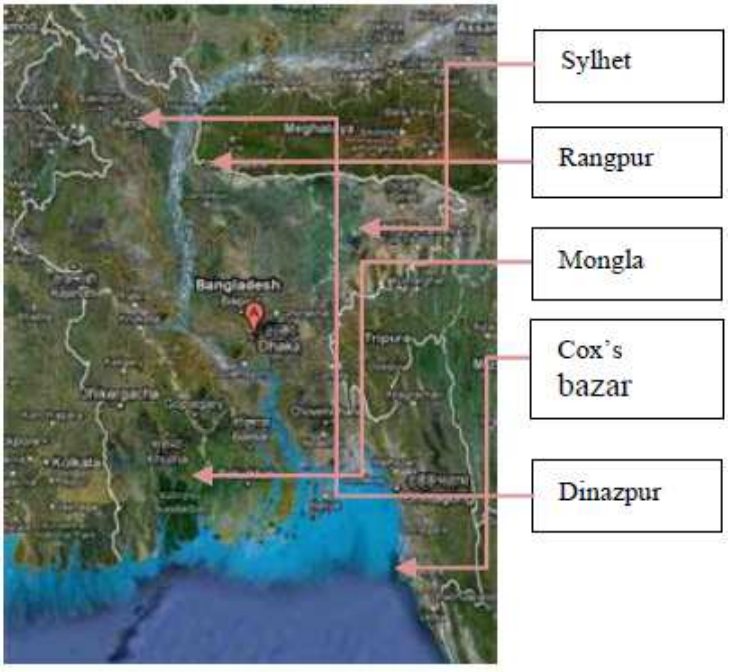

Fig. 3. The NASA image of Bangladesh (Nov 9, 2011)

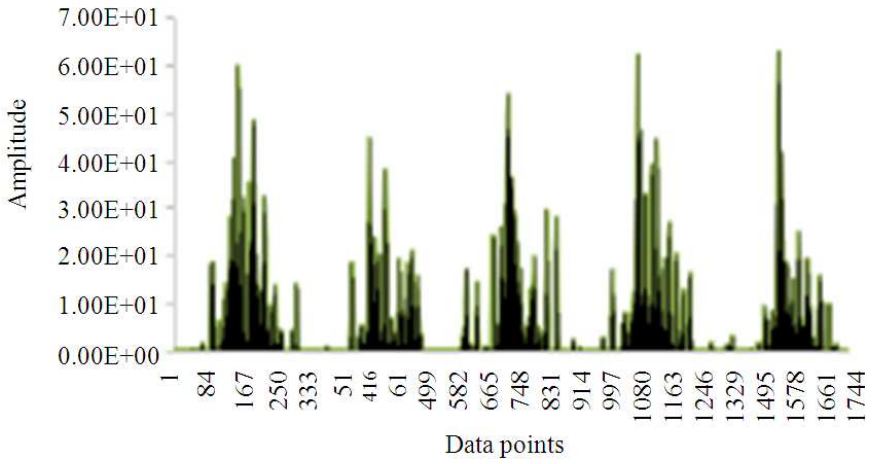

Fig. 4. Daily rain falls in mm for Cox's Bazar zone

Table 1. Rainfall characteristics of different zones in Bangladesh

Zone	Mean	Stand. Dev.	Min	Max
Cox's Bazar	10.92	19.36	0	176.0
Dhaka	5.98	10.48	0	120.0
Dinazpur	5.84	11.44	0	118.6
Khulna	5.35	9.37	0	86.0
Mongla	5.71	9.34	0	69.0
Rangpur	6.78	13.78	0	216.4
Sylhet	10.85	16.36	0	140.8

There are 64 districts in Bangladesh. There are 32 stations in entire Bangladesh for recording daily rain falls. We did not take the rain falls for all the places rather we take seven different yet important zones. Daily rain fall was takenfor zones Cox's Bazar, Dhaka, Dinazpur, Khulna, Mongla, Rangpur and Sylhetas marked in **Fig. 3**. A daily rain fall time series is shown for Cox's Bazar district in **Fig. 4**.

The characteristics of data at different zones are summarized in **Table 1**. The mean is the average over daily rain fall over the period 1998-2007.

3. RESULTS AND DISCUSSION

3.1. Experimental Results

A typical example of autocorrelation function, correlation integral and dimension are described in **Fig. 5**. **Table 2** describes the relative CDs of different zones of Bangladesh. This measurement was taken with a uniform experimental setup. The highest value of CD is for Cox's Bazar 1.29, sec highest Sylhet 1.17. In reality, these two zones are famous for different instantaneous rainfall. As a result, they have the highest rainfall. The minimum CD is found for Dinazpur-a northern part of Bangladesh. This means there is less dynamic neighborhood change among the surrogate data.

Now we validate afore mentioned findings with a neural network training to determine the tuff and rough zones here. Basically we prepare a neural network prediction model to comment how difficult the time series to be learnt. This determines the easy and difficult zones in terms of rain fall in Bangladesh. 'Easy' means regular characteristics of rain fall. 'Difficult' means the rain fall changes instantaneously and irregularly. We can determine this using neural network. A neural network or a learning system obviously takes time to a set of complex input patterns as human takes more time to remember a complex sequence than an easy sequence of patterns. A time sequence x(t-1), x(t-2), x(t-3) and x(t-4) are given in the neural network input and trained for the target of x(t). A training graph is shown in **Fig. 6** exhibiting error versus iteration for ten years data.

Three situations were taken into consideration using a data set consisting of 5 years, 10 years and 20 years. A more stable and complete data will be the one which involves larger number of years, because entire dynamics of the rain fall came into consideration during training.

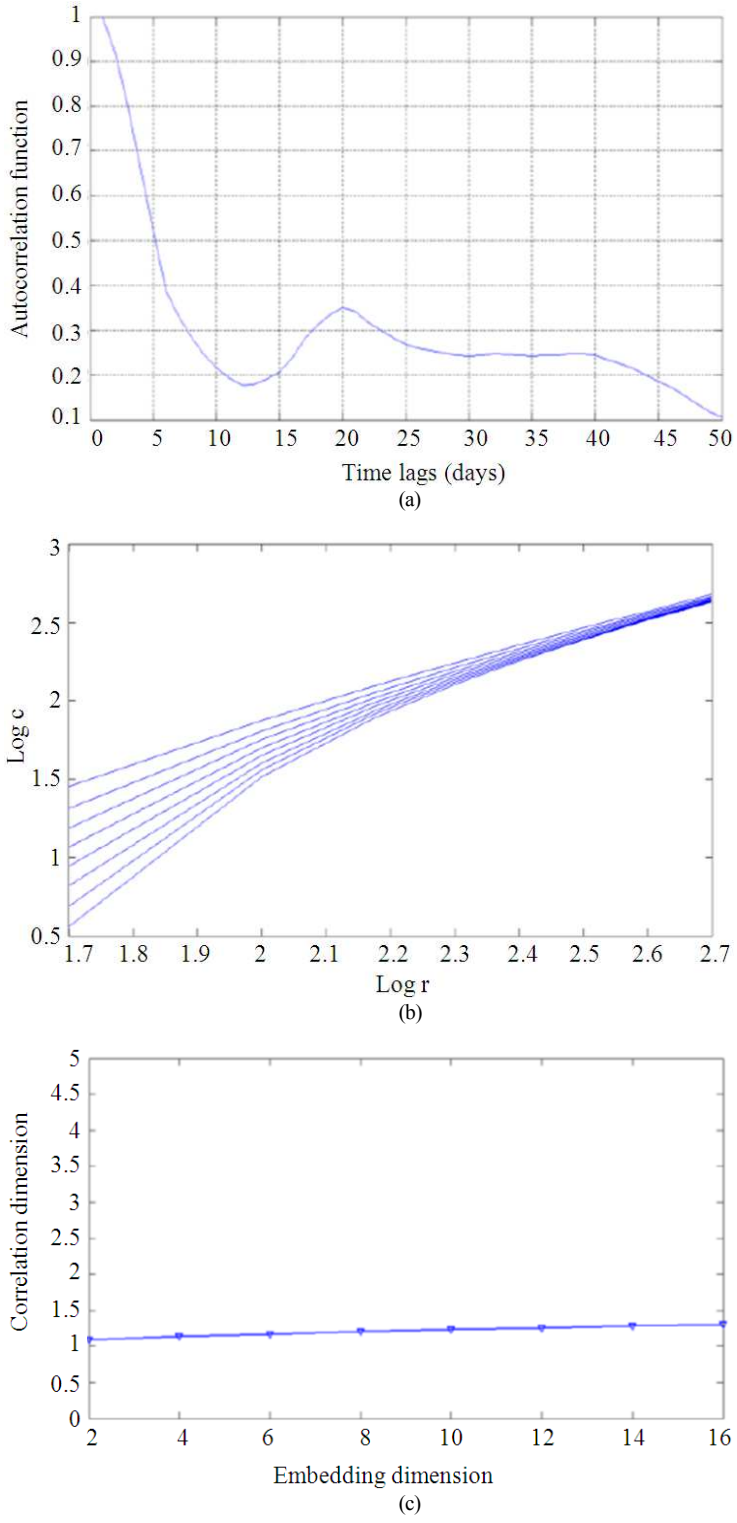

Fig. 5. (a) Autocorrelation function (b) Correlation integral (c) Correlation dimension of Cox's bazar rain fall

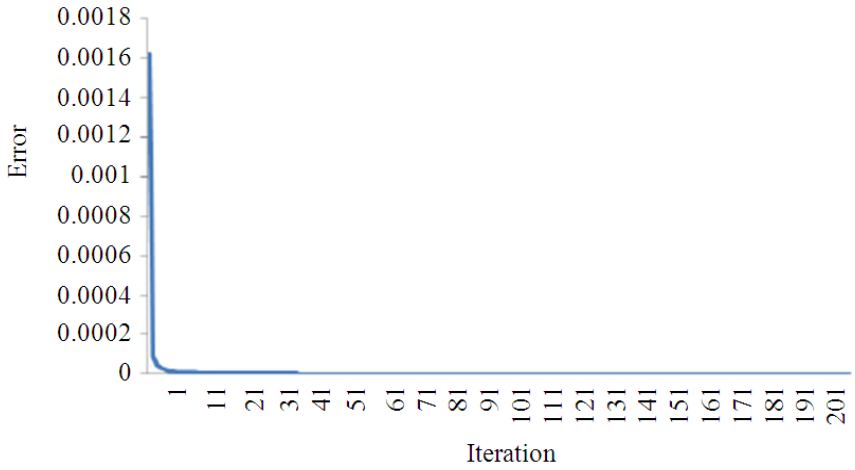

Fig. 6. Neural network error Vs iteration

Table 2. Comparative study of CD of different zones

Zone	Time lag	CD
Cox's Bazar	12	1.29
Dhaka	08	1.15
Dinazpur	08	1.07
Khulna	10	1.11
Mongla	12	1.10
Rangpur	10	1.13
Sylhet	8	1.23

Table 3. Weights between inputs to hidden layer for Cox's bazaar data

	B	I1	I2	I3	I4	I5
H1	0.3129	0.2335	-0.4652	0.6141	0.4888	-0.7492
H2	0.7524	-0.2446	-0.6566	0.2040	-0.1805	-0.5571
H3	0.5805	0.0242	0.5507	-0.3692	0.7088	0.2955
H4	0.6328	0.8186	-0.6216	0.2489	-0.3923	0.7064
H5	-0.7414	0.7534	-0.1794	0.0438	-0.2671	-0.2736
H6	-0.2771	0.2457	-0.6702	-0.0116	0.5571	-0.7890
H7	0.8074	0.3935	-0.7080	0.8560	0.7587	-0.8688
H8	-0.4446	-0.6488	0.9032	-0.3756	-0.7789	-0.0765
H9	0.4225	0.1938	-0.5099	0.4914	0.8135	-0.7908

The training of neural network was initially started with the same set of weights for each experiment for 200 iterations. The final error reached by the network was assessed as indicator of how difficult the training sequence for the network. The same initial weights are used to train the neural network.

Table 3 shows weight between hidden units H, bias unit B and input units I. From **Table 3** it is shown that there have total eight hidden units with a bias unit and five input units. For example weight between hidden unit 2 (H2) and Input unit 4 (I4) is -0.180541. **Table 4** also shows weight between output unit O and hidden unit H with a bias unit B which is output bias.

We found a similarity between Cox's bazar and Sylhet, whereas other zones exhibit different errors starting with same initial conditions and network structure. However, optimization of network parameters is not our current interest. These two zones have errors of 0.000001208020160 and 0.000001302082068 which are larger than other zones for 20 years data as observed in **Table 5**. This means these two zones are difficult for the network. In case of 5 years data, the scenario becomes identical except Rangpur and Khulna zones having larger error than that of Cox's bazar and Sylhet. This is because all variations possibly did not include in the training data set. Immaturity in the data set may dominant.

Table 4. Weight between outputs to hidden layer

	B	H1	H2	H3	H4
O	-2.5346	-1.5970	-1.1776	-0.6678	-0.3153
	H5	H6	H7	H8	H9
O	0.1709	-0.4317	-1.3872	-0.2552	-1.2467

Table 5. The mean square error of the neural network output

No of Years	05	10	20
Cox-bazar	0.0020561	0.000001673278076	0.000001208020160
Sylhet	0.002777	0.000001712843447	0.000001302082068
Dhaka	0.001825	0.000001319979452	0.000000832485065
Khulna	0.002358	0.000001648679731	0.000001104426431
Rangpur	0.002971	0.000001133464411	0.000000630250753
Dinajpur	0.001890	0.000001492534616	0.000000937526408

However, when we took ten and 20 years data the results are very reliable since now the entire input Data contains full information.

3.2. Comparison of Results

A linear dependencies between MSE and C.D are observed when artificial data created by adding noise to base time series of stock prices. However, the observations are more spreaded in case of meteorological data (Acharya *et al*., 2009). It is worth noticing that the linear dependence holds regardless of the size of MSE. A similar characteristic is observed when the MSE is below 0.004 and for MSE reaching 0.05. Since our data set belongs to rain fall, the similar characteristic was found. The C.D linearly depends on MSE.

3.3. Implications

The main interest of this research is to assess the outcome of correlation dimension computed from rainfall with neural network predictions. The resemblance between these two findings was observed. The complexity of time series was judged with the help of correlation dimension and the prediction for the same time series was done with the neural network. This finding can also be applied for any time series such as daily transactions in the bank, stock values, electrical load prediction. Visual Evoked Potentials (VEPs) may be analyzed by examination of the morphology of their components, such as Negative (N) and Positive (P) peaks using correlation dimension.

3.4. Limitations

The correlation dimension has limitations to calculate in an analytically closed form. Therefore, there are many sophisticated methods used to estimate the correlation dimension. The value of C.D. depends on several factors: level of data stationarity, the length of a stationary segment, external noise, sampling rate and embedding dimension. For example, if we choose "stationary" region of time series, the length of this region must be about 10^(D+2) points. Therefore to determine D = 10 we need about 10^6 points of sufficiently stationary observable. It is observed that computation of correlation dimension involves the region selection of log-log plot. The bad selection of the linear region may result deviated slope estimate.

3.5. Future Scope

Features of the correlation dimension can be easily and properly utilized for a classification task. Generally a classification model ignores such kind of information as a feature. The classification accuracy may improve for such information inclusion in feature space. This is essential for many time series such as biomedical gene data bases. All information regarding the problem domain is not often appeared in feature space. This approach can be extended for such application especially for gene microarray data.

4. CONCLUSION

This study presents an interesting scenario-the correlation dimension of rain fall correlates the neural network prediction output when the input space sufficiently contains full information of the rainfall. The correlation dimension and mean square error of neural network exhibit identical outcomes. Therefore, these two zones- cox's bazar and sylhet are difficult zones in terms of their rain fall behavior. The real scenario also resembles with this findings.

5. REFERENCES

Acharya, U. R., K.C. Chua, T.C. Lim, Dorithy and J.S. Suri, 2009. Automatic identification of epileptic EEG signals using nonlinear parameters. J. Mech. Med. Biol., 9: 539-553. DOI: 10.1142/S0219519409003152

Ghorbani, M.A., R. Khatibi, B. Sivakumar and L. Cobb, 2010. Study of discontinuities in hydrological data using catastrophe theory. Hydrol. Sci. J., 55: 1137-1151. DOI: 10.1080/02626667.2010.513477

Jaiyen, S., C. Lursinsap and S. Phimoltares, 2010. A very fast neural learning for classification using only new incoming datum. IEEE Trans. Neural Netw., 21: 381-392. DOI: 10.1109/TNN.2009.2037148

Kannan, M., S. Prabhakaran and P. Ramachandran, 2010. Rainfall forecasting using data mining technique. Int. J. Eng. Technol., 2: 397-401. Kurokawa, F., K. Ueno, H. Maruta and H. Osuga, 2011. A new control method for dc-dc converter by neural network predictor with repetitive training. Proceedings of the 10th International Conference on Machine Learning and Applications, Dec. 18-21, IEEE Xplore Press, Honolulu, HI, pp: 292-297. DOI: 10.1109/ICMLA.2011.17

Larranaga, A., J. Alvarez-Mozos, L. Albizua and J. Peters, 2013. Backscattering behavior of rain-fed crops along the growing season. IEEE Geosci. Remote Sensing Lett., 10: 386-390. DOI: 10.1109/LGRS.2012.2205660

Liang, J.W. and S.L. Chen and C.M. Yen, 2013. Identification and verification of chaotic dynamics in a missile system from experimental time series. Int. J. Syst. Sci., 44: 700-713. DOI: 10.1080/00207721.2011.618639

Russell, S.J. and P. Norvig, 2010. Artificial Intelligence: A Modern Approach. 1st Edn., Prentice Hall, Upper Saddle River, ISBN-10: 0136042597, pp: 1132.

Shumway, R.H. and D.S. Stoffer, 2010. Time Series Analysis and Its Applications: With R Examples. 1st Edn., Springer Publisher, New York, ISBN-10: 1441978658, pp: 607.

Thai, T.Q. and M. Myrskyla, 2012. Rainfall shocks, parental behavior and breastfeeding: Evidence from rural Vietnam. Max Planck Institute for Demographic Research.

Xingyuan, W., Z. Liu and M. Wang, 2013. The correlation fractal dimension of complex networks. Int. J. Modern Physics C, 24: 9-9. DOI: 10.1142/s0129183113500332