Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 6, pp. 1407–1419
DOI : 10.47974/JSMS-1151

Rainfall and outlier rain prediction with ARIMA and ANN models

N. Shobha *
Department of Computer Science and Design
Dayananda Sagar College of Engineering
Bengaluru
Karnataka
India

T. Asha †
Department of Information Science and Engineering
Bangalore Institute of Technology
Bengaluru
Karnataka
India

K. Seemanthini §
Department of Machine Learning
B.M.S College of Engineering
Bengaluru
Karnataka
India

V. Jagadishwari ‡
Department of Computer Science and Engineering
C.M.R Institute of Technology
Bengaluru
Karnataka
India

**E-mail:* shobha-csd@dayanandasagar.edu (Corresponding Author)
† E-mail: ashat@bit-bangalore.edu.in
§ E-mail: seemanthinik.mel@bmsce.ac.in
‡ E-mail: jiswariv@yahoo.com

Abstract

The precipitation level, a vital agro meteorological factor, holds immense significance in the decision-making process for the promotion of sustainable agriculture, preserving natural resources and improving quality of life. Rainfall prediction is necessary to explore crop environment relationship, water availability, soil erosion, floods and drought disasters. By leveraging Artificial Neural Networks (ANNs) and Autoregressive Integrated Moving Average (ARIMA) techniques, the proposed method utilizes ten input parameters and day-to-day meteorological observations to accurately forecast rainfall events at the Bengaluru station from 2013 to 2017. ANN method is also used to find an outlier during non-monsoon season. The suggested ARIMA model c(2,0,2) forecast daily rainfall 3 days in advance and c(1,0,0) anticipate monthly rainfall 5 months in prior. The model evaluation results are tabulated separately with MSE, RMSE, MAE and R2 values.

Subject Classification: *68P45.*

Keywords: *Artificial neural network, ARIMA, Meteorological data, Rain fall.*

1. Introduction

Rainfall is one of the important factors in agro met observatories. The efficient utilization of weather or climate data will help to conserve natural resources and promotes economic benefit to the farmer. Rainfall becomes major challenge and accurate rainfall prediction is an important goal with critical consequences for agriculture, maintaining water for major cities and minimizing flood disasters. The rainfall prediction is stochastic process where rainfall observation y(tn+1) can be predicted at some future time tn+1 with its interaction with other observations such as temperature, humidity, cloud amount, wind direction and surface pressure denoted as { y1 y2…yn(t1) , y1 y2…yn(t2), …, y1 y2…yn(tn)} in a time sequence t1,t2, …,tn. Literature survey shows that many researchers worked on ANN models [1] [2] [3][4][5][6] and ARIMA model [7] for rainfall prediction. This work focusses on the techniques of applying ANN and ARIMA on five years' meteorological data to predict rainfall and to determine an outlier during nonmonsoon period. In the first step ANN analyze the input parameters and give statistical score near to predicted variable rainfall. In the second step day to day rainfall measure and monthly mean rainfall value is given as input to ARIMA to forecast daily rainfall observations three days in advance and monthly rainfall five months in prior. In the third step input data is partitioned into non monsoon months (40 data set), ANN experimented on each data set to predict an outlier. The model evaluation results are meticulously documented, presenting the values of root mean square (RMSE), mean absolute error (MAE), and coefficient of determination (R2) in a comprehensive tabular format.

The fundamental goal of machine learning algorithm is to discover non trivial relationship in training data and interpret the result in test data. Cluster wise linear regression method was proposed for predicting monthly rainfall in Victoria by (Adil et. al., 2016) [1]. The data over the period 1889-2014 was collected from eight weather stations of Australia with five input variables. By using cluster wise linear regression method, the data was partitioned into k clusters and regression function was applied within clusters to find trends in the data. To assess the models performance, various evaluation metrics such as accuracy, root mean squared error, mean absolute error, mean absolute scaled error, and coefficient of efficiency were employed, comparing the results with established methodologies including CR-EM, Multiple Linear Regression, SVMreg, and Artificial Neural Networks.

A proposal was put forward by K.S. Kasiviswanathan et. al., [2] to determine the prediction interval for artificial neural network rainfall runoff models. The input to the ANN model was the rainfall received hourly and runoff from kolar basin. In the first step, genetic algorithm was used to set weights and biases vector to train the ANN model . In the second step, ANN parameters were optimized to create model with minimum value of mean and maximum estimated prediction. Finally, model validation and calibration was done using mean, standard deviation, correlation coefficient, Nash-efficiency and RMSE.

Priya Narayanan et. al. (2013) [7] introduced a methodology that utilizes a univariate ARIMA model to conduct trend analysis and examine the pre-monsoon rainfall behavior. The study involved analyzing monthly rainfall data for March, April, and May from six stations spanning a 60-year period (1949-2009). Patterns within the time series data were investigated using the pre-whitened Mann Kendall test, and auto-correlation elimination was performed at a significance level of 10% (α=10%).

2. Research Method

This section the methodology adopted in the work is described along with the explanation of the data and artificial neural network and ARIMA algorithms employed in the analysis.

2.1 *The meteorological dataset*

The dataset considered includes meteorological data of five years (2013 to 2017) collected from IMD Bengaluru [17] [18]. Input parameters include temperature wet 700, relative humidity 700, cloud 700, wind 700 and surface pressure 700 recorded at Indian Standard Time (IST)

7.20 and temperature dry 1400, humidity 1400, cloud 1400, wind 1400, surface pressure 1400 recorded at IST 14.20, and the target variable rainfall recorded at IST 8.30.

2.2 *Artificial Neural Network*

Artificial Neural Network (ANN) is a significant engineering concept within the field of Artificial Intelligence, utilized to extract patterns from historical data. ANN can be described as a large collection of interconnected units. The interconnection follows some structure of patterns for communication between the units.

ANN model has been used for rainfall forecasting for the interval of five years (2013 to 2017) daily data. Input variables include temperature, humidity, cloud amount, wind and surface pressure. There are recurrent patterns and relationship existing between input variables and target variable Rainfall. The model is developed using R language. Steps to build ANN model is as follows.

1. Initially data is normalized by using min-max method and scaled in the interval [0, 1].
2. Sample function is used to split 70% (1260) of data into training and remaining 30% (540) is used as test set. Supervised learning algorithm was applied by providing the network with specified number of inputs and target output rainfall.
3. Fitting the model by calling the function neural net using the configuration 10-10-2-1 where input layer, first hidden layer and the second hidden layer has 10,10 and 2 neurons respectively. A single neuron represents the output layer. During training, weights of neurons are adjusted to minimize error and various configuration was tested in order to identify the optimal model, various techniques and methodologies are employed.
4. Correlation between the predicted and actual values is calculated by applying compute function with trained model and values of predicted variables of the test set as arguments. Finlay coefficient of determination (R^2) is calculated by using the formula

$$R = \frac{TSS - RSS}{TSS} \tag{1}$$

The value of R2, which is approximately 0.9999 and approaching 1, signifies a superior model as it indicates that the total sum of squares

(TSS) associated with the outcome variable is well explained, while the residual sum of squares (RSS) is minimal.

5. To find an outlier in the data set, five years meteorological data is partitioned into non monsoon months. The months between January to May, and October to December a total of 8 months were considered as non- monsoon according to Indian monsoon. ANN was applied on month wise data to find instances of rainfall, where occurrence of rainfall during non-monsoon days was treated as an outlier.70% of data was taken as training data remaining data was considered as test data.

ANN algorithm was experimented on each data set and model was evaluated using mean squared error (MSE) and R^2 value.

2.3 *ARIMA*

ARIMA (Auto Regressive Moving Average) model has been used for rainfall prediction. The input data comprises of five years (2013 to 2017) daily data with two parameters date and rainfall. Every day rainfall data and mean value of monthly rainfall data fed as input to ARIMA model to predict future values by using past observations. The ARIMA (p, d, q) model can be described as having p autoregressive terms, where p represents the number of such terms

$$y_t = c + \varnothing_1 y_{t-1} + \varnothing_2 y_{t-2} + e_t \tag{2}$$

φ1 & φ2 are parameters for the model, the degree of differencing terms is represented by d, q component represents the error of the model denoted by

$$y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t \tag{3}$$

Including autoregressive, moving average and differencing results in ARIMA model is as follows:

$$y_t = c + \varnothing_1 y_{d\ t-1} + \varnothing_p y_{d\,t-p} + \ldots + \varnothing_1 e_{t-1} + \varnothing_q e_{t-q} + e_t \tag{4}$$

The data has to be fed into ARIMA in stationary form. Hence it is required to reduce constant variance and skewness present in the data. Procedure to develop ARIMA model is as follows.

1. Differencing the series will help to remove constant variance and skewness present in the data and make it much more stationary. Differencing term can be written as

$$diff(1) = y_{dt} = y_t - y_{t-1} \tag{5}$$

$$diff(2) = y_{d2t} = y_{dt} - y_{dt-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \tag{6}$$

2. The next step is to fit AR (auto regressive) term by applying ACF (auto correlation function).
3. Fit MA (moving average) term by applying PACF (partial auto correlation function) on time series data. ACF and PACF provides q and p values respectively. When PACF applied on difference 1 term of daily rainfall data, nine values were going out of bound. The term C (9, 1, 1) is given as input to ARIMA and results were tabulated. Different orders were experimented to select best model for forecast. Finally the order C (2, 0, 2) was chosen as best fit by calling auto. arima function on daily rainfall data.
4. Finally, model is tested by Ljung-Box chi square statistics test to check the presence of auto correlation between the residuals. P value was around 0.2298, hence the constructed ARIMA model was utilized to forecast rainfall for the upcoming three periods.
5. In the same way the order term for monthly mean value of rainfall series was selected by repeating the above procedure. The order C (1, 0, 0) was chosen as best fit for forecasting using auto.arima function. The Ljung-Box test gives the P value around 0.0626, the proposed model was able to forecast rainfall for next five months.

3. Results and Discussions

The empirical analysis was done using R, which composes of different packages to develop Artificial Neural Network and ARIMA model, nnet function is used for ANN implementation.

In this section, the outcomes of rainfall data analysis using Artificial Neural Networks (ANN) and Autoregressive Integrated Moving Average (ARIMA) are presented. The architecture of the ANN model, denoted as 10-10-2-1, is illustrated in Figure 1, where the first and last numbers (10 and 1) represent the number of neurons in the input and output layers, respectively. The numbers between them indicate the number of neurons in the hidden layers. In this particular architecture, there are two hidden

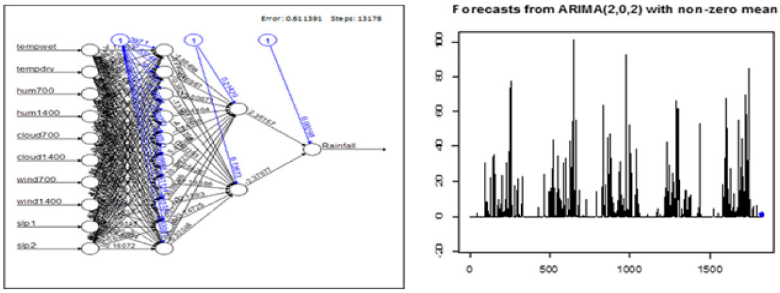

Figure 1

Rainfall Prediction using ANN 10-10-2-1 and ARIMA C (2,0,2) for daily rainfall data

layers consisting of 10 and 2 neurons, respectively. Learning rate and epoch have been adjusted to 0.611391 and 13178 respectively. As shown in the Table 1, the first ANN architecture 10-10-2-1 has produced mean square error (MSE) of 0.00281009 for training data, MSE of 0.05841202 for test data and coefficient of determination was 0.9999923462. The second ANN architecture 10-12-4-1 was used epoch of 28740 with error rate of 0.486538, produced MSE of 0.00118531 for training data, MSE of 0.02133956 for test data and the value of coefficient of determination was 0.9999939092. The third ANN architecture 10-10-4-1 was used epoch of 58601 with error rate of 0.583014, produced MSE of 0.00197107 for training data, MSE of 0.02409290 for test data and the value of coefficient of determination was 0.9999927014. Among all three models based on the epochs, error rate and MSE, the first model of ANN 10-10-2-1 was selected as best model. Training results of rainfall data shows good prediction accuracy.

ARIMA modeling has been done for rainfall prediction by considering five years daily and monthly rainfall data of Bengaluru station. To build ARIMA model 5 years daily rainfall data is considered, nearly 1800 events.

Table 1 presents the calculated values of root mean square error (RMSE), mean absolute error (MAE), and p-value for various orders, assessing the accuracy of the model fitting. The first value in the order c(2,0,2) represent pacf, second and third value represent difference and acf terms. The order c(2,0,2) holds 8.6352 RMSE, 3.8743 MAE and 0.2298 p-value. The second order c(6,1,1) and third order c(2,2,2) has value of RMSE, MAE, p-value was around 8.6662, 3.8216, 0.3707 and 8.6942, 3.8346, 0.2428 respectively. By considering the results of all three orders, the first order c(2,0,2) holds smaller values of RMSE and p-value with non zero mean of 2.6501, which indicates the better fitting model. Based on the

Table 1

ANN and ARIMA models for daily rainfall data.

Model	ANN Architecture	MSE training	MSE test	R^2	ARIMA model	RMSE	MAE	Check for Randomness p-value
1	10-10-2-1	0.0028190	0.0584120	0.999992346	C(2,0,2)	8.6352	3.8743	0.2298
2	10-12-4-1	0.0011853	0.0213395	0.999993909	C(6,1,1)	8.6662	3.8216	0.3707
3	10-10-4-1	0.0019710	0.0240929	0.999992701	C(2,2,2)	8.6942	3.8346	0.2428

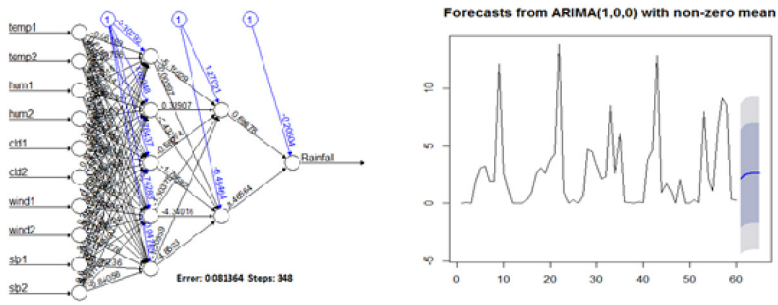

Figure 2

Rainfall Prediction using ANN 10-5-2-1 and ARIMA C (1,0,0) for monthly rainfall data

results the graph was plotted for future forecasting is shown in Figure 1, which describes rainfall variations for 1800 records and blue dot shows rainfall prediction for next three periods. The rainfall prediction showed zero for next three days because the next time interval would come in the month of January.

The ANN architecture 10-5-2-1 is as shown in the Figure 2, the ANN architecture 10-5-2-1 has produced mean square error (MSE) of 0.00250009 for training data, error rate 0.081364 and 348 epochs. ARIMA C(1,0,0) is as shown in the Figure 3, in ARIMA model blue curved line predicts the amount of rainfall for next five months Jan, Feb, March, April and May of 2018 of Bengaluru station.

Daily and month wise rainfall forecasting results obtained from ANN and ARIMA model is compared with the actual rainfall data and shown in the Figure 3. The number of days is plotted on X axis while Y axis represents day-to-day rainfall observations. Blue line represents actual rainfall, brown line represents predicted value of ANN and green line represents predicted value of ARIMA. Forecasting accuracy of ANN model is closer

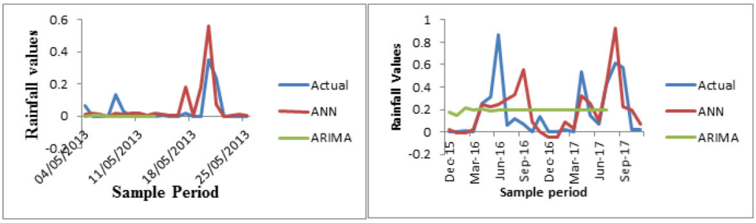

Figure 3

Comparison of ANN and ARIMA model with actual day to day rainfall data (left) and monthly rainfall data (right).

to actual rainfall data when compared to ARIMA. The prediction values of ARIMA are not close to actual data and become constant after some point. ANN achieved better performance over ARIMA model. Processing the data using ANN provides more accurate results. Figure 3 illustrates the comparison of ANN and ARIMA model with actual and monthly rainfall data of March, May, Nov and Dec of 2013 to 2017.

4. Results of ANN on Non-monsoon Months to Predict an Outlier

Results of ANN on non- monsoon months to predict outlier was discussed. ANN was tested on each month data to find rainfall during non-monsoon season, rainfall during non-monsoon will be considered as an outlier. The ANN architecture includes ten inputs of weather parameters and one predicted variable rainfall. Few rainfall instances are found in Apr, May, Oct and Nov month because Mar to May was considered as pre monsoon and Oct to Nov months were recognized as post monsoon season. ANN algorithm accurately predicts outlier during these periods. Medium or average rainfall was expected during post monsoon and pre monsoon season, if amount of rainfall was excess that cause damage to crops and crops suffer from pest related diseases. The farmer has to invest money on pesticides which in turn reduces the yield and vegetables such as tomato and onion would spoil because of excess rainfall.

Accuracy, Precision and Recall Values were calculated for Actual Value and Predicted Value of test data and it is shown in the Table 2. Error can be calculated as actual output minus predicted output, R2 is measured based on error value at prediction. MSE and accuracy values was calculated based on the results. It has been observed that prediction accuracy of non-monsoon months was nearly 88% to 100% means no rainfall observed. In some cases, the months having rainfall during premonsoon and post monsoon having accuracy 55% to 66%, since these periods are having less

Table 2

Validation measures of ANN in outlier prediction.

2013							
April		**May**		**Oct**		**Nov**	
R^2	0.99	**R^2**	0.99	**R^2**	0.99	**R^2**	0.99
SE	0.0370	**MSE**	0.7927	MSE	0.0916	**MSE**	0.10
Accuracy	66%	**Accuracy**	88%	**Accuracy**	66%	**Accuracy**	66%
2014				**2015**			
May		**Oct**		**Apr**		**May**	
R^2	0.99	**R^2**	0.99	**R^2**	0.99	**R^2**	0.99
MSE	0.3076	**MSE**	0.2369	**MSE**	0.005	**MSE**	0.0776
Accuracy	55%	**Accuracy**	44%	**Accuracy**	66%	**Accuracy**	55%
2015		**2016**					
Nov		**Mar**		**Apr**		**May**	
R^2	0.99	**R^2**	1	**R^2**	1	**R^2**	0.99
MSE	0.0320	**MSE**	0.1108	MSE	0.1229	**MSE**	0.0444
Accuracy	44%	**Accuracy**	100%	**Accuracy**	88%	**Accuracy**	55%
2017							
Apr		**May**		**Oct**		**Nov**	
R^2	1	**R^2**	0.99	**R^2**	0.99	**R^2**	1
MSE	0.1101	**MSE**	0.0315	**MSE**	0.4327	**MSE**	0.1516
Accuracy	88%	**Accuracy**	55%	**Accuracy**	55%	**Accuracy**	55%

rainfall, the rainfall amount was about 0.1 mm to 2mm, the model was recognizing this as 0 instead of 1. R2 value are in the range 0.99 to 1 and MSE values of all observations are least which signifies that model was able to predict outliers in the data set and prediction rate is also accurate.

Conclusion

The proposed model constitutes of ANN and ARIMA having the ability to accept more number of time series observations and related climate inputs to find required result. Ten input variables, 1800 observations are considered and tried to find different combinations of ANN architecture. Models are built with training datasets and evaluated using test data sets. ANN model 10-10-2-1 is considered as best based on 0.00281009 MSE for training data, 0.05841202 MSE for test data and coefficient of determination is equal to 0.9999. To build ARIMA model two input variables, 1800 records for daily rainfall series and 60 instances for monthly rainfall series are examined. The model c(2,0,2) for daily rainfall series produces 8.6352 RMSE, 3.8743 MAE, 0.2298 p-value and c(1,0,0) for monthly rainfall series generates 3.2485 RMSE, 2.3869 MAE and 0.0626 p-value, these models are selected as best forecasting model for rainfall based on lowest RMSE and non zero mean value. The model c(2,0,2) predicts rainfall forecasting of 1st,2nd and 3rd day in January 2018 and c(1,0,0) predicts rainfall event on Jan, Feb, March, April and May months in 2018. The ANN model also predicts an outlier by examining non-monsoon periods in the dataset and model validation was done using accuracy, MSE and R2. In case the rainfall is more during non monsoon days that causes harm to cultivated crops and also diseases can affect crops which in turn reduces yield. ANN model achieved better performance over ARIMA model. Processing the data using ANN provides more accurate results.

References

[1] Adil M Bagirov, et. al., "Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach," Elsevier, Atmospheric Research (2017), doi:10.1016/j.atmosres.2017.01.003.

[2] K.S.Kasiviswanathan, et. al., "Constructing prediction interval for artificial neural network rainfall runoff models based on ensemble simulations," *Elsevier, Journal of Hydrology*, 499, 275-288 (2013).

[3] Narasimha Prasad, et. al., "An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree," 2013 4th International Conference on Intelligent Systems, *Modelling and Simulation*, DOI 10.1109/ISMS.2013.27.

[4] Wassamon Phusakulkajorn, et. al., "Wavelet-Transform Based Artificial Neural Network for Daily Rainfall Prediction in Southern Thailand,"ISCIT 2009.

[5] Mislana, et. al.,"Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia,"International Conference on Computer Science and Computational Intelligence (ICCSCI 2015), Elsevier, Procedia Computer Science, 59, 142–151 (2015).

[6] John ABBOT and Jennifer MAROHASY,"Application of Artificial Neural Networks to Rainfall Forecasting in Queensland, Australia,"Advances in Atmospheric Sciences, Vol. 29, No. 4, 717–730 (2012), doi: 10.1007/s00376-012-1259-9.

[7] Priya Narayanan, et. al.,"Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India,"Comptes Rendus Geoscience,C.R. Geoscience 345, 22-27 (2013).

[8] Shifa Chen and Xuan Zha,"Effects of the ENSO on rainfall erosivity in the Fujian Province of southeast China,"Elsevier, Science of the Total Environment (2017).

[9] Anil K.Jain, Jianchang Mao,"Artificial Neural Networks: A Tutorial,"IEEE-1996.

[10] Matheus Soares Geraldi and Enedir Ghisi, "Influence of the length of rainfall time series on rainwater harvesting systems: A case study in Berlin,"Elsevier, Resources, Conservation and Recycling, 125, 169-180 (2017).

[11] G. Evin, et. al.,"A regional model for extreme rainfall based on weather patterns subsampling," *Elsevier, Journal of Hydrology,* 541, 1185-1198 (2016).

[12] Jiansheng wu,et. al.,"Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm,"Elsevier, Neurocomputing, 148, 136-142 (2015).

[13] A. Troncoso, et. al.,"Imbalanced classification techniques for monsoon forecasting based on a new climatic time series," Elsevier, Environmental Modelling & Software, 1-9 (2017).

[14] Philip Hall,"Climate Divergence – Fact or Fiction: Synoptic characterisation as a methodology for short-to-medium term climate analysis and forecasting," 2nd International Conference on Signal Processing and Integrated Networks (SPIN) (2015).

[15] H. Vathsala,et. al.,"Closed Item-Set Mining for Prediction of Indian Summer Monsoon Rainfall A Data Mining Model with Land and Ocean Variables as Predictors,"Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science, 54, 271–280 (2015).

[16] Jamileh Farajzadeh, et. al.,"Modeling of monthly rainfall and runoff of Urmia lake basin using feed-forward neural network and time

series analysis model, "Elsevier, Water Resources and Industry, 7-8, 38–48 (2014).

[17] http://www.imd.gov.in

[18] http://www.uasbangalore.edu.in/index.php/research/agromoterology.

[19] Kalpana Polisetty, Afrifa Yamoah Ebenezer, "An empirical study on rainfall patterns of monsoon season in the north-west India using time series models", *Journal of Statistics and Management Systems*, PP 559-572 (2020).

[20] Mahboob Alam, Mohd Amjad, "Weather forecasting using parallel and distributed analytics approaches on big data clouds", *Journal of Statistics and Management Systems*, PP 791-799 (2019).

[21] Sanku Dey, Enavetur Raheem, Saikat Mukherjee, Hon Keung Tony, "Two parameter exponentiated Gumbel distribution: properties and estimation with food data example", *Journal of Statistics and Management Systems*, PP 197-233 (2017).

Received February, 2023