Prediction of Climate Variable using Multiple Linear Regression

E. Sreehari[1]
Assistant Professor
School of computing science and engineering
Galgotias University
Anantapur, India
sreehari.0088@gmail.com

Dr.Satyajee Srivastava[2]
Associate Professor
School of computing science and engineering
Galgotias University
Uttar Pradesh, India
drsatyajee@gmail.com

***Abstract*—The change in global temperatures, recent past three years natural disasters, rising sea levels, reducing Polar Regions can be causing the problem of understanding and predicting these climate phenomena. Prediction is a prime importance and they can be run and simulated as computer simulations to predict climate variables temperature, precipitation, rainfall and etc. The state of Kerala saw the worst flood of the century in august 2018, beginning on 15 August 2018. Severe floods affected the south Indian state of Kerala, due to unusually high rainfall during the monsoon season. It was the worst flooding in Kerala in nearly a century. Over 483 people died, and 15 are missing. At least a million people were evacuated. The agricultural country called India in which 60% of the people depending upon the agriculture. Rain fall prediction is the most important task for predicting early prediction of rainfall May helps to peasant's as well as for the people because most of the people in India can be depends upon the agriculture. This paper explains about multiple linear regression technique for the rainfall estimation or prediction. It can helps to farmers for taking appropriate decisions on crop yielding. As usually at the same time there may be a scope to analyze the occurrence of floods or droughts. The multiple linear regression analysis methodology applied on the dataset collected over six years of Nellore district from Andhra Pradesh state. The experiment and our multiple linear regression methodology exploit the appropriate results for the rain fall than simple linear regression methodology.**

Keywords: Regression, simple linear regression, prediction, correlation coefficient.

I. INTRODUCTION

Regression can be one of the techniques used for climate prediction and many other areas. The climate can be changes strongly day by day. The natural disaster occurred the year 2015 Nov-Dec south India floods occurred and affected in the regions of Andhra Pradesh and Tamil Nadu. In which it may results in the loss of property approximately 100,000 cores of money, reducing natural resources and more than 500 people were killed. In 2017 more than 1200 people died in flood related which had been reported by respective state governments.

The state of Kerala saw the worst flood of the century in august 2018, started on 15 August 2018. Severe floods affected the south Indian state of Kerala, due to unusually high rainfall during the monsoon season. It was the worst flooding in Kerala in nearly a century. Over 483 people died, and 15 are missing. At least a million people were evacuated. Kerala losts the highest loss of lives due to floods and Uttar Pradesh saw 204 deaths, in West Bengal 195, Karnataka 161 and Assam 46. In Kerala, 54 lakh people were affected and14.52 lakh people were living in relief camps. In Assam, 11.46 lakh people were affected and 2.45 lakh were in relief camps. An estimate by the National Disaster Management Authority (till 2005) put the loss of lives at an average 1,600 every year due to floods. The damage caused to crops, houses and public utilities was in excess of Rs 4,745 crore annually with 12% of the country's geographical area being flooding prone.

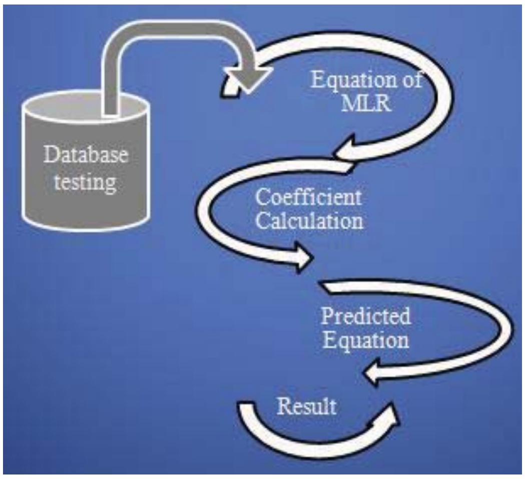

Fig. 1: Regression working process

Empirical approach as well as Dynamical approach can be two ways used for rainfall prediction. With the help of the historical we will accomplish Empirical methodology. The physical models can help us to form equations with that we will implement prediction values. One of the concepts called numerical rainfall forecasting method [6, 26].

Regression, artificial neural network, fuzzy logic, Support vector machines [18] can be empirical based methods and they are supervised learning methods with that we can able to analyze and define new values. A simple climate forecasting [7] can be done by regression techniques. These regression techniques are discussed

978-1-5386-6947-1/18/$31.00 ©2018 IEEE

further. The use of data mining in the field of hydrology can be increasing in the recent years. The work and concept which have been performed and explained in the mentioned references [14-16]. Keskin explained about integrated evaporation model [17], using DM process for three lakes in Turkey. Artificial Intelligent methods used in the estimation of rainfall can be defined in the following references [19-25].

II. RELATED WORK

The overall work of the of the paper can be organized as follows. In Section 2, we given overview of the related concept work. In Section 3, we defined our working strategy and solution for the prediction using regression. In Section 4, we implemented the model and other several baselines for regression. Result can be discussed in section 5. In chapter 6 we presented the conclusion along with future work. Acknowledgement and references defined at the end.

III. MULTIPLE LINEAR REGRESSION

In the concept of statistical solving simple linear regression is a empirical approach and it can solves the tasks by considering the historical data set of the climate values or parameters. It can be only consists of single dependent variable and independent variable. In the simple linear regression model can be exists only two variables.

The representation of multiple linear regression will be like

$$Y=a_0+a_1x_1+ a_2x_{2+}\ a_3x_{3+}\ a_4x_{4+\text{-------}}\ +a_nx_n$$

In which Y=dependent variable
$x_1,x_{2,}x_{3,}x_4$=independent variables
$a_0,a_1,a_{2,}a_3,a_4$=regression coefficients

In multiple linear regression we can represent in the form of mathematical equation by calculating slope and regression. coefficents. The strength and direction of the association between the two variables can be estimated by using the regression coefficient formula. Similarly there are various correlation coefficient formulas can also be available in the mathematical and statistical evolution processing.

The mathematical for r is given as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

The coefficient determination measures how well data can be represented in the regression line. It can defines strength and direction of relationship between the dependent and independent variables.

Error Calculation

The error can be calculated after calculating the predicted values and the difference can be calculated using the actual and predicted values. The formula for calculating error is

$$RMSE = \sqrt{\frac{(y_i - \bar{y}_i)^2}{n}}$$

RMSE can be known as root mean square error or root mean square deviation. It can be used measure of difference between sample and population values.

IV. RAIN FALL PREDICTION USING MLR

The following architecture will helps you regarding how to accomplish the task of rainfall prediction in a sequential manner.

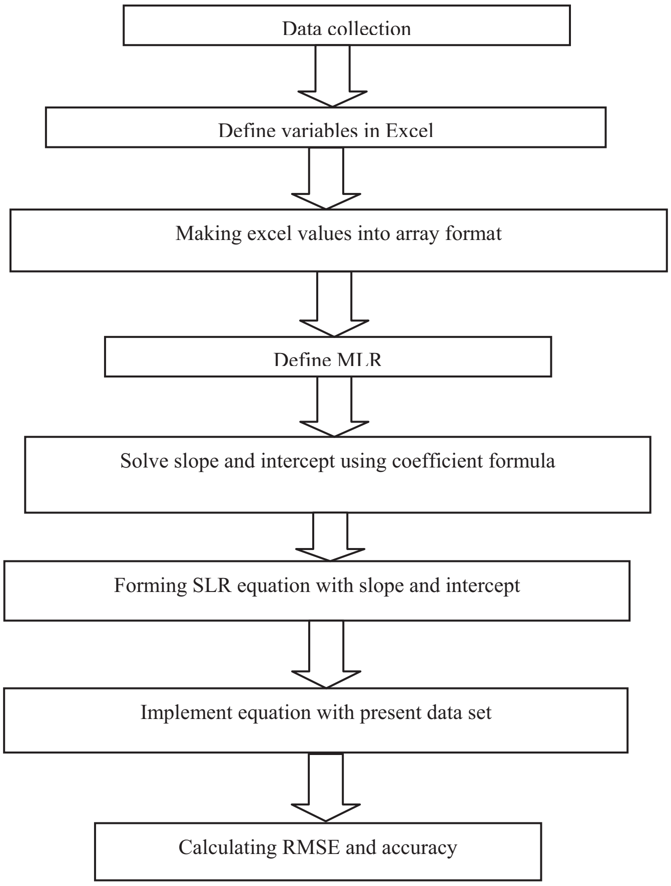
Data collection
Define variables in Excel
Making excel values into array format
Define MLR
Solve slope and intercept using coefficient formula
Forming SLR equation with slope and intercept
Implement equation with present data set
Calculating RMSE and accuracy

Fig. 2: Architecture to implement MLR

The Algorithm for multiple Linear Regression's

Step1: Function (Min.tmp, Max.tmp, CC, VP, Y)
Step2: Collect all the data in to excel.
Step3: Read the values from the excel sheet f1=new file (D:\\sample.xls).
Step4: CC= [j] = (a. get contents ()).
Step5: VP = [j] = (b. get contents ())…similarly.
Step6: To calculate the correlation coefficient value.
Step7: Forming regression equation.
Step8: obtaining constant value from equation and data set
Step9: By using this equation with considered data set we can obtained the predicted value which is closest to dependent variable.
Step10: we have calculate the RMSE

Step11: we can calculate accuracy by using confusion matrices.

Initially Consideration of Requirements for MLR

- Data set which is in the form of climate variables with numerical values.
- Data base maintenance by Excel software.
- JDK software.
- The IDE software called Net Beans.
- The package of java Jxl.jar.
- Weka software.

Steps regarding the implementation of a project

- Defining the data set variables in excel.
- Installing the NetBeans IDE along with JDK if JDK is not installed in your system.
- Creation of project in NetBeans framework.
- Importing the jxl.jar package into the project library folder.
- Declaring the variables and source file path in the by creating a file.
- Implementing correlation coefficient calculation using programming.
- Define regression equation.
- Calculating slope value from using simple regression equation and coefficient value.
- Final forming the predicted equation with slope and coefficient values.
- Implementing the equation with the climate parameters and comparing the actual values with the predicted values.
- Calculate the RMSE using past data set and obtained values.
- Install the WEKA software.
- Obtaining confusion matrix using WEKA.
- Calculating specificity and sensitivity.
- Finally implementing the accuracy formula using specificity and sensitivity.

Considered data set is:

TABLE I: DATA SET MONTHLY AND YEAR WISE

year/month	min.tmp	max.tmp	cc	wdf	precipitation	vapor pressure
1997/january	18.15	29.25	23.337	0.946	3.471	19.28
1998/january	20.13	31.24	18.378	0.1362	0.243	21.38
1999/january	18.28	29.37	21.862	0.4267	0.524	19.24
2000/january	19.31	30.4	18.759	0.0014	0.001	20.488
2001/january	19.34	30.44	31.592	0.9734	3.245	20.444
2002/january	19.27	30.36	32.868	1.3846	11.512	20.487

V. RESULTS

In this section we present the experimental results for the regression methodologies simple regression and multiple regression, we have introduced briefly in the previous sections regarding regression concepts. The concept of regression can be implemented by calculating coefficient, slope and the considered climate data set either day, monthly or annual wise. As well as the performance of predicting the future values can be calculated by using multiple linear regression algorithms. By using Net beans and weka framework we implemented our project. We can form the regression equation by calculating regression coefficient will be like in this form

Y=605.04+57.73(min.temp)-55.80(max.temp)-0.15179(clodcover) +3.0560(wdf)-0.1745(precipitation)

By substituting the considered data set in the regression equation, we can obtain and carried out the experimental values as shown in table.

The final graph representing after implementing my proposed system called multiple linear regression and existing system multiple linear regression.

The graph we can obtained like in this way

TABLE II. RESULTS AFTER EVALUATION

ACTUAL VALUES	SLR VALUES	MLR VALUES
19.28	20.22082	19.43
21.38	20.21034	21.53
19.24	20.21125	19.39
20.488	20.20955	18.41
20.444	20.22009	20.59
20.487	20.24695	20.44

In figure 2 we compare the both the regression methodologies simple linear regression and multiple linear regressions by considering the same data set. Here we consider the same climate variables for predicting the future values. The existing methodology called multiple linear regressions can require and uses more climate parameters. In the other way simple linear regression can be easy to implement and analyze but it may use one independent and dependent variable

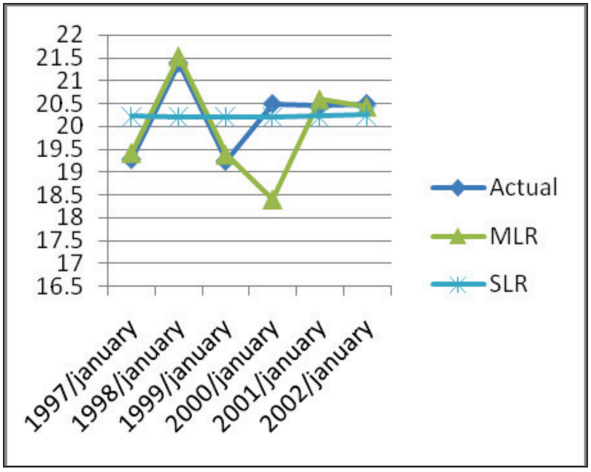

Fig. 3: Graph showing results

VI. CONCLUSION

The natural incidents may not possible to stop and cannot estimate in a efficient and accurate manner. In general by using the concept of future estimation concept or events or values there may be a scope to minimize lot of problems. In this project we have implemented the simple regression methodology, multiple regression and we predicted the values, the multiple regression error rate also less when comparing with simple linear regression. Finally concluding that multiple linear regressions can be more better than simple linear regression. By considering vapor pressure value as a dependent variable with other values as independent we successfully implemented the simple linear regression method multiple linear regression. This is the concept of prediction but not in a accurate manner because we know that climate factors changes due to different reasons and impacts on it.

VII. FUTURE WORK

As a future study the logistic regression we can also able to implement by considering the same data set and to accomplish the logistic regression the dependent variable must in be in binary format. Similarly we can obtain results that are called predicted values and the base line of multiple regression concepts can be considered for doing some mathematical calculations. So in future we can also implement the concept of ARIMA model by using different correlation coefficient formulas with the same methodology. By using different correlation coefficient formulas the predicted equation also changes with different values. However, by implementing this concept, we can achieve the prediction in a better and appropriate manner.

REFERENCES

[1] Divya Chauahan and Jawhar Thakur "Data Mining Techniques for Weather Prediction" published in 2015.

[2] Ozlem terzi, Hindwani Monthly Rainfall Estimation by Data-Mining Process Publishing Corporation in the year 2012.

[3] Nikhil Sethi and Dr. Kanwal Garg"Exploiting Data Mining Technique for Rainfall Prediction" in 2014.

[4] Imran Ahmed, Sruthi Menon and Nikitha"Rainfall Prediction using Multiple Regression Technique" in the year 2014.

[5] Pinky Saikia Dutta and Hitesh Tahbilder Prediction of Rainfall by Datamining Technique in Assam in the year 2013.

[6] Z.Ismail et. al, Forecasting Gold Pieces Using Multiple Linear Regression Method in American Journal of Applied Sciences in the year 2009.

[7] Paras, et.al, "A simple Weather Forecasting Model Using mathematical Regression" in Indian Research Journal of extension Education Special Issue (Vol. 1) in the year 2012.

[8] C. Damle and A. Yalcin, Flood prediction using time series data mining, Journal of Hydrology, vol. 333, no.2-4, pp. 305-316, 2007.

[9] K. W. Chau and N. Muttil, Data mining and multivariate statistical analysis for ecological system in coastal waters by, Journal of Hydroinformatics, vol. 9, no. 4, pp. 305-317,2007.

[10] E. P. Roz, Water quality modeling and rainfall estimation: a data driven approach [M.S.thesis], University of Iowa, Iowa city, Iowa, USA, 2011.

[11] M. E. Keskin and O. Terzi Datamining process for integrated evoparation model in the year 2009.

[12] Y. Radhika and M. Shashi , "Atmospheric Temparature prediction using support vector machines" in the year 2009.

[13] T. B. Trafalis.M.B. Richman, A.White, and B. Santosa Data mining techniques for improved WSR-88D rain fall estimation in 2002.

[14] J. E. Ball and A. Sharma, An Application of artificial neural networks for rainfall forecasting in the year 2001.

[15] M. Zhang, A. Scofield and J. Fulcher Rainfall estimation using artificial neural network group in 1997.

[16] T. Shoji and H. Kitaura Statistical and geostatistical analysis of rainfall in central Japan in 2006.

[17] M. C. V. Ram´ırez, H. F. C. Velho, and N. J. Ferreira, "Artificial neural network technique for rainfall forecasting applied to the S˜ao Paulo region in 2005.

[18] R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records in 2005.

[19] Y.-M. Chiang, F. J. Chang, B. J. D. Jou, and P. F. Lin, "Dynamic ANN for precipitation estimation and forecasting from radar observations 2007

[20] H.Hasani, ," A New Approach to Polynomial Regression and Its Application to Physical growth of Human Height.

[21] E. Sreehari, J. Velmurugan, Dr. M. Venkatesan "A Survey Paper on Climate Changes Prediction Using Data mining, 2016.