A multiple linear regression model for precipitation forecasting over Cuttack district, Odisha, India

Conference Paper · April 2017

DOI: 10.1109/I2CT.2017.8226150

3 authors, including:

S. Swain
Indian Institute of Technology Roorkee
43 PUBLICATIONS **1,406** CITATIONS

SEE PROFILE

Saswata Nandi
University of California, Merced
25 PUBLICATIONS **330** CITATIONS

SEE PROFILE

A Multiple Linear Regression Model for Precipitation Forecasting over Cuttack District, Odisha, India

S. Swain, P. Patel
Inter-Disciplinary Programme in Climate Studies
IIT Bombay
Mumbai, India
sabyasachiswain16@gmail.com

S. Nandi
Department of Civil Engineering
IIT Bombay
Mumbai, India
ce.saswata@gmail.com

Abstract— **Estimation of precipitation is necessary for optimum utilization of water resources and their appropriate management. The economy of India being heavily dependent on agriculture becomes vulnerable due to lack of adequate irrigation facilities. In this paper, a multiple linear regression model has been developed to reckon annual precipitation over Cuttack district, Odisha, India. The model forecasts precipitation for a year considering annual precipitation data of its three preceding years. The model testing was performed over a century-long dataset of annual precipitation i.e. for 1904-2002. Assuming the intercept or constant of the multiple linear regression model as zero, the equation developed thereby displayed a superb result. The model predictions showed an excellent association with the observed data i.e. the coefficient of determination (R^2) and adjusted R^2 value was obtained to be 0.974 and 0.963 respectively. This reconciliation justifies the application of the developed model over the study area to forecast rainfall, thereby aiding in proper planning and management.**

Keywords—multiple linear regression, precipitation, forecast, Cuttack district

I. Introduction

Precipitation is very important parameter for all types of water resources project development, planning and management [1]. In India, about two-third of the population have agriculture as their primary occupation and agriculture mainly depends on rainfall in most of the parts [2]. Due to lack of adequate irrigation facilities, any changes or variability in the annual precipitation significantly affect the agriculture and food security of the country [3], [4]. So, it is always advisable to be ready for any kind of inconvenience due to precipitation anomaly. In order to achieve that, the estimation of annual precipitation is quite necessary [5], [6].

Several methods are employed to quantify the precipitation beforehand i.e. rainfall forecasting. All these methods are categorized into two approaches i.e. Dynamical approach and Empirical approach. The dynamical approach predicts rainfall through physics-based models. Basically, numerical rainfall prediction methods are applied in Dynamical approaches. The Empirical approach deals with analysis of past rainfall records and its relationship with other hydro-meteorological parameters e.g. regression, fuzzy logic, artificial neural network (ANN) etc. Among them, multiple linear regression is widely applied because, it easily depicts the relationship of multiple influential variables to form a statistically significant rainfall [7], [8]. It also helps in determining the dominance of a variable over the others.

It is generally believed that the changes in rainfall pattern should be studied over a long period i.e. at least on a scale of 30 years. So, it can be expected that, the behavior of rainfall over an area on a particular year will resemble well with its few previous years rainfall. The hypothesis of this paper is that, the rainfall over a year can be predicted well by considering the rainfall data of its three previous years, using multiple linear regression.

II. Study Area

The area selected for this study is Cuttack District of the state Odisha, India. Area of Cuttack district is 3932 square kilometers (1518 square miles). It is second most populated district of Odisha. The geographical location of the district is 20.517^0 N latitude and 85.726^0 E longitude. The location of Cuttack district is shown in the Fig. 1.

From the point of view of climatology, the average annual precipitation over the district is about 1440millimetres, most of which occurs during south-west monsoon period (June-September). The temperature seems moderate for the area throughout the year except for the summer season (March-June), where the average maximum temperature is 41 ^{0}C. The average minimum temperature over the district is 10^0C.

III. Data and Methodology

The annual rainfall data for 1901-2002 for Cuttack district was collected from India Water Portal. The statistical regression technique is applied over this data in order to develop a model to predict the annual rainfall values. Regression is a statistical technique that uses the relation between two or more variables on observational database so as to predict an outcome from other variables. There are many kinds of regression analysis out of which linear regression is very much applied as it is very simple to use [9], [10]. The quantitative variables are assumed to be linearly related to each other. There are basically two types of linear regression i.e. simple linear regression and multiple linear regression [11]. The multiple linear regression that is used in this study can be represented as shown in (1).

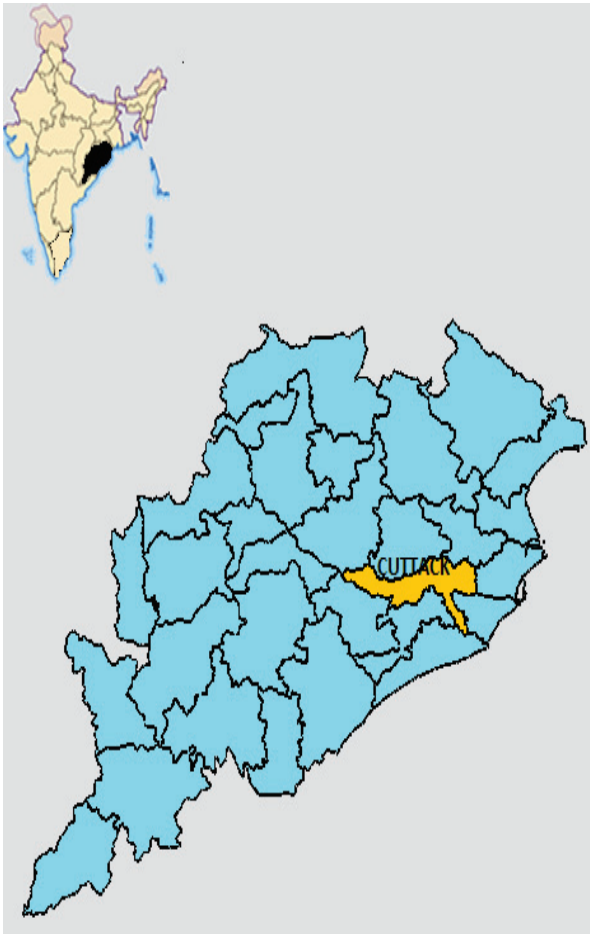


Fig. 1. Location of Cuttack District in Odisha Map

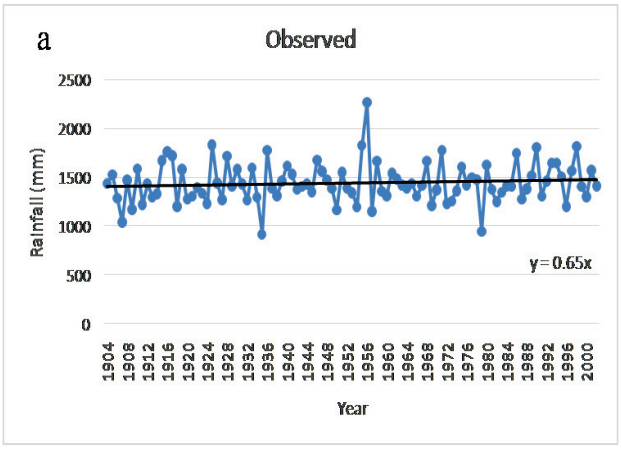
$$P = aX_{-1} + bX_{-2} + cX_{-3} + k \quad (1)$$

Equation (1) represents the general multiple linear regression model, where, P is the annual rainfall to be obtained. The variables X_{-1}, X_{-2} and X_{-3} are the annual rainfall values of three previous years. In a typical multiple linear regression equation, a variable has to be estimated, which is regarded as Predictand. The variables from which the Predictand will be estimated are regarded as Predictors. The coefficients of Predictors and the intercept will govern the value of Predictand.By looking at (1), it can be clearly noted that, P is the Predictand, whereas, the variables X_{-1}, X_{-2} and $X_{-3}$are the Predictors. The coefficients a, b and c determines the dominance of relationship of the predictor variables with annual rainfall to be obtained. The term k is intercept of the regression equation, which is a constant. The intercept is used to adjust the shift of the mean predicted value with that of mean observed value [12]-[15]. In a good multiple linear regression model, the coefficients of Predictors and the intercepts are controlled in such a way that the Predictand will possess acceptable association with that of observed values. As the fluctuations in the rainfall over a year will not be extremely high, a mere high value of constant may cause almost negligible dependence of precipitation in previous three years. Therefore, here we have assumed the intercept to be zero. Since the study was about the rainfall estimation using its previous three year values, the testing and validation was applied on rainfall from 1904-2002. These 99-year values were then compared to that of the observed data.

IV. Results and Discussion

The multiple linear regression analysis produced excellent results. The coefficients obtained for the first, second and third previous years are 0.492, 0.277 and 0.226 respectively. The results obtained are quite convincing as the dominance of the predictor variables are decreasing with that increasing the temporal gap. The equation or multiple linear regression model developed is presented in (2).

$$P = 0.492X_{-1} + 0.277X_{-2} + 0.226X_{-3} \quad (2)$$



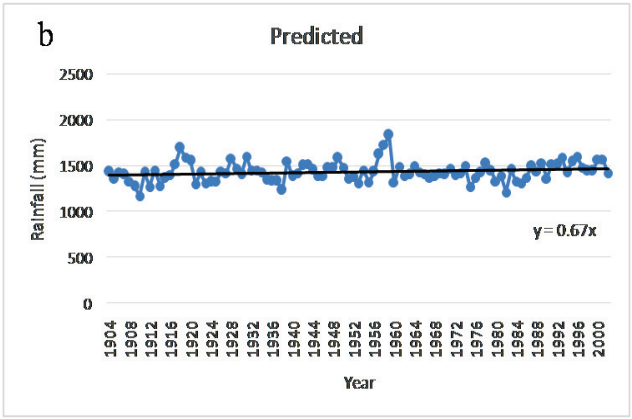


Fig. 2. Comparison of (a) Observed Annual Rainfall and (b) Predicted Annual Rainfall, for 1904-2002

From the validation of this model on annual rainfall data for 1904-2002, the regressed value has shown a very good match with that of observed values. The correlation coefficient (r) and coefficient of determination (R^2) are the efficacy measure used for validation of this model. The expressions for R^2 and adjusted R^2 are presented in (3) and (4) respectively.

$$R^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N \, \sigma_x \sigma_y} \right)^2 \quad (3)$$

where, R^2 is the coefficient of determination, x_i and y_i are ith elements, $\bar{x}$ and $\bar{y}$ are mean, σ_x and σ_y are standard deviation of the x and y value of observation respectively, N is the total sample size.

$$Adjusted\ R^2 = 1 - \left(\frac{(1-R^2)(N-1)}{N-p-1}\right) \quad (4)$$

where, R^2 is sample coefficient of determination and p is the number of predictors.

Generally, in the multiple linear regression models, the predictions may get biased due to the fact that, more number of predictors causes over fitting of the model. It is obvious that if the number of predictors increases, the accuracy of the model also increases. So each time a new predictor will be introduced, it may contribute to a better agreement of model with observed values, even if due to chance alone. Secondly, more predictors with higher ordered polynomials lead to creation of random noises, which also produces high values of R^2, which can be misleading. To combat such problems, the Adjusted R^2 is considered, which gets adjusted in accordance with the number of predictor variables. The value of Adjusted R^2 increases when a new predictor improves the model more than would be expected by chance and decreases if a predictor improves the model less than what is expected by chance [14]. The interesting thing to note is that the value of Adjusted R^2 can even be negative, but generally it does not occur so. Its value will be always lesser than R^2. For a good model, the difference between R^2 and Adjusted R^2 is small.

The model generated values show an extremely satisfactory agreement to actual data, thereby obtaining a correlation coefficient = 0.987 and consequently, a coefficient of determination = 0.974. The adjusted coefficient of determination is also 0.963, which is enough to justify the predictability of the model (as difference between R^2 and Adjusted R^2 is very small). A comparative analysis of the variation of actual data and the model generated values for the 100 year duration are presented in Fig. 2. An interesting thing to notice is that, the regressed values show lesser variance compared to observed data. But the mean values are very well captured. Also, looking at the equations provided on the chart in Figure 1, it can be noticed that, setting the intercept as zero, the observed values obtain a relation of y= 0.65x, whereas regressed values obtain a relation y= 0.67x with respect to time (y is the annual rainfall and x is the year). Hence, the regression model produce quite accurate results.

V. Conclusion

It is very important to estimate rainfall properly for an improved water resources planning, development and management. A multiple linear regression model was developed to estimate the annual rainfall over Cuttack District, Odisha, India, using the annual rainfall values of three previous years. The model is able to produce very good result and delivered an excellent matching with the actual data thereby obtaining a very high coefficient of determination (R^2) equal to 0.974 and an adjusted R^2 of 0.963. Such a high R^2 value is enough to justify the capability of the model to estimate annual rainfall over the area, that may aid for further hydro-meteorological investigations in future.

Acknowledgment

We express our deep sense of gratitude to all the organizations and individuals who have directly or indirectly helped to complete this work. We acknowledge India Water Portal for the rainfall data. We thank our parents, teachers, colleagues and friends for their extensive support and motivation. We also thank Almighty for his blessings.

References

[1] S. Swain, M. Verma, and M. K. Verma, "Statistical trend analysis of monthly rainfall for Raipur District, Chhattisgarh", Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March, 2015, pp. 87-89.

[2] M. Kannan, S. Prabhakaran, and P. Ramachandran, "Rainfall forecasting using data mining technique", International Journal of Engineering and Technology, vol. 2, 2010, pp. 397-401.

[3] S. Swain, "Impact of climate variability over Mahanadi river basin", International Journal of Engineering Research and Technology, vol. 3, 2014, pp. 938-943.

[4] M. K. Verma, M. K. Verma, and S. Swain, "Statistical Analysis of Precipitation over Seonath River Basin, Chhattisgarh, India", International Journal of Applied Engineering Research, vo. 11, 2016, pp. 2417-2423.

[5] C. Gagiu, E. M. Pica, X. Querol, and C. S. Botezan, "Analysis of predictors related to soil contamination in recreational areas of Romania", Environmental Science and Pollution Research, vol. 22, 2015, pp.18885-18893.

[6] I. Mamani, "Modeling of Thermal Properties of Persian Walnut Kernel as a Function of Moisture Content and Temperature Using Response Surface Methodology", Journal of Food Processing and Preservation, vol. 39, 2015, pp. 2762-2772.

[7] D. F. Andrews, "A robust method for multiple linear regression", Technometrics, vol. 16, 1974, pp. 523-531.

[8] S. S. Amiri, M. Mottahedi, and S. Asadi, "Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the US", Energy and Buildings, vol. 109, 2015, pp. 209-216.

[9] K. J. Preacher, P. J. Curran, and D. J. Bauer, "Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis", Journal of educational and behavioral statistics, vol. 31, 2006, pp. 437-448.

[10] B. T. Nolan, M. N. Fienen, and D. L. Lorenz, "A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA", Journal of Hydrology, vol. 531, 2015, pp. 902-911.

[11] M. R. Piña-Monarrez and J. F. Ortiz-Yañez, "Weibull and lognormal Taguchi analysis using multiple linear regression", Reliability Engineering & System Safety, vol. 144, 2015, pp. 244-253.

[12] M. Krzywinski and N. Altman, "Points of Significance: Multiple linear regression", Nature methods, vol. 12, 2015, pp. 1103-1104.

[13] M. Novotná, O. Mikeš, and K. Komprdová, "Development and comparison of regression models for the uptake of metals into various field crops", Environmental Pollution, vol. 207, 2015, pp. 357-364.

[14] S. Yu, W. Kang, S. Ko, and J. Paik, "Single image super-resolution using locally adaptive multiple linear regression", JOSA A, vol. 32, 2015, pp. 2264-2275.

[15] M. Krzywinski and N. Altman, "Points of Significance: Simple linear regression", Nature methods, vol. 12, 2015, pp. 999-1000.