ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# Monsoon Prediction Using Data Mining

**Suhas Suresh**
R.L.Jalappa Institute of Technology, Bangalore, India.

## *Manuscript Info*

***Manuscript History:***

Received: 10 October 2013
Final Accepted: 22 October 2013
Published Online: November 2013

***Key words:***
Data mining, Satellite images, Weather forecasting, Rainfall forecasting, K-NN Algorithm, K-mean Algorithm, Multiple Linear Regressions.

## *Abstract*

Monsoon is an integral part of the Indian economy, as it determines the economic and social domain of the nation as a whole. India, being an agricultural country, it is important to predict the average rainfall for that year. Heavy rains or prolonged dry period-both are unacceptable. In this paper, we throw some light on the different strategies to predict monsoon using satellite images.



## I. Introduction

Prediction of monsoon is important for the development of any country, both in social and in economic domain. For a country like India, the economy of the entire nation is dependent on the onset of the monsoon. It is known that 75% of the agriculture in India is dependent on monsoon, the prediction is also essential for other fields like aviation, shipping, trading, drought management, power generation and consumption, reservoir storage management, etc. Hence, rainfall prediction is a much-awaited act by all and their slightest fluctuation is of great concern.

There are two ways of monsoons that occur in India. The first one is the South-western Monsoon and the second being the North-eastern Monsoon. The south western monsoon is the rainfall that origates due to the depression the Bay of Bengal Ocean. This causes the winds to blow from the region of lower pressure to the region of higher pressure, thus blowing the air of rain through the regions of South-western parts of India- namely the states of Kerala, Karnataka, parts of Tamil Nadu and Maharashtra. The Northeastern monsoon, on the other hand is the inverse of this. It originates in the region of West Bengal and traverse through the states of Andhra Pradesh, Tamil Nadu, and few parts of Kerala.

There are a wide range of weather forecast methods that are employed at national and international levels. But out of them, the two methods widely used are: Empirical method and Dynamical method. Empirical method is a historic approach of data of rainfall, on statistics and other oceanic variables[2] .
The most widely used empirical approaches are fuzzy logic, artificial neural logic, regression, etc. In dynamic approach, the predictions are generated using the models that are developed based on the evolutions of global climatic conditions over the initial/original climatic conditions[2].

A section of this paper deals with the satellite images used for the analysis of the monsoon while the other section deals with the methods used in analysis and prediction of monsoon using different algorithms and principles.

## II. SATELLITE IMAGES

Satellite images are the images taken from the sensors of the satellite. It should be notes that, these images are not photographs taken from the satellite. These images are the images which are visible outside the visible region like the infrared region, near infrared region and far infrared region unlike those images of the photographs which are taken only in the visible region. The IR spectrum is used to differentiate between lands, sea, aloud or thin cloud. It is used to capture picture all time (day or night). Other important advantages of using IR spectrum is that it is used in calculating the cloud height, temperature on top of the cloud, sea surface temperature, water vapor wind, etc[1]. While the visible spectrum helps in determining the peripheral parameters like overall cloud coverage, smoke, pollution density, thin clouds, fog and smog in some cases [1].

Another noteworthy point is that these images are taken from different satellites namely- Polar satellites and Geostationary satellites. Polar satellites are used to produce good resolution images while Geostationary images are used to produce low resolution images. Geostationary satellites are stationary satellites which are stationed at a high altitude from the earth and they regularly monitor the movement of the earth.

The onset of southwestern monsoon is considered when there is a substantial amount of rainfall in the state of Kerala. Ideally, this would happen on the 1st of May or June but the dates, however keeps fluctuating. In spite of this, the weather conditions in the southern states remain to be same and this is the key factor for measuring through satellite images[1]. There are a few factors which heavily determine the monsoon. They are cloud height, cloud top temperature, cloud density and water vapor wind.

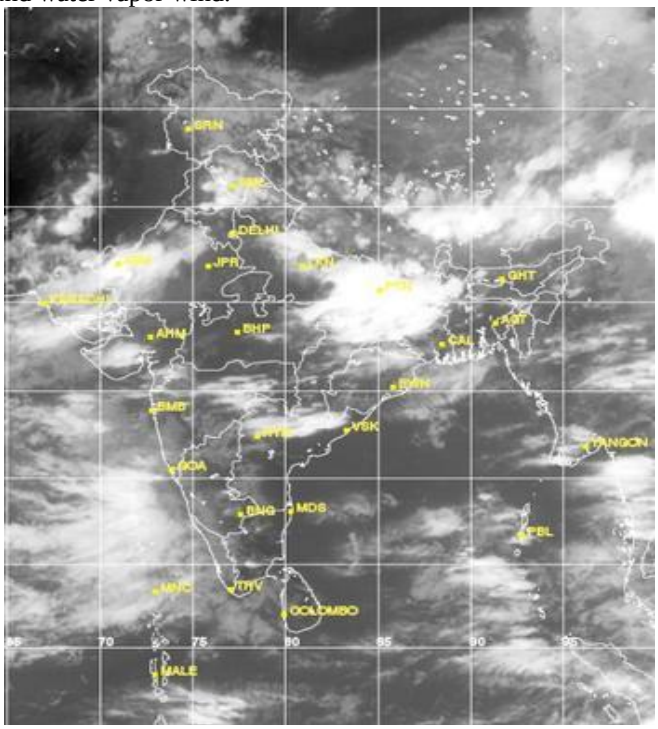


**Fig 1: Satellite image of south-western monsoon in the month of May/June.**

The above figure shows the satellite image of south-western monsoon. It can be observed that few parts of the image has thick cloud coverage in few parts of Bay of Bengal Ocean and in few parts of Bihar and West Bengal states of India. Not to forget, these are the only few topics which help us predict rainfall.

## III. MONSOON PREDICTION ALGORITHM.

The entire algorithm consists of three stages: the first stage consists of data set (image data) obtained from a span of 4 years. The second stage is a cluster stage where the data set is segregated into multiple clusters and the thirds stage is a comparison stage where the data set (images) is compared with the centroids of each clusters created in the second stage[1]. Then using K-NN algorithm, monsoon onset is predicted. The KNN algorithm scheme is used to predict weather, climatic variables, etc using spatial and temporal dependencies at multiple stations.

The following algorithm is one which is derived from a paper presented by Dinu John, K.K. Sindhu and B.B Meshram[1]. The first step if the algorithm is slightly a time consuming process while the other stages are more accurate and faster.

Begin

1. Set size of K-Array=0,
   MaxDistK-Array=999999,
   MinDistK-Array=999999,
2. Extract all the features from every region of the Water Vapor, Infra-red and Visible images.

3. Repeat for each record of cluster

   a. Calculate the Euclidean distance.
   b. If (Distance < MaxDistK-Array)
      a.1. While size of K-Array less than 7
         a.1.1. Insert record into K-array
         a.1.2. Increment the size of K-Array by 1.
   Else
      Replace record with MaxDist in K-Array with new record.
      a.2. Update MaxDistanceK-Array value
      a.3. If the distance is less than MinDistK-Array value, update MinDistK-Array value.

4. Calculate the remaining onset days left using the K-Array value.

   End.

For our work, we have used two algorithms namely, K-mean clustering algorithm and K-NN algorithm. Clustering algorithm is a technique to group the objects such that the objects in the cluster share more properties than with objects outside the cluster. This means that all like objects are grouped together and this would help us in classifying and summarizing it. The following image shows the first stage of obtaining the data set.

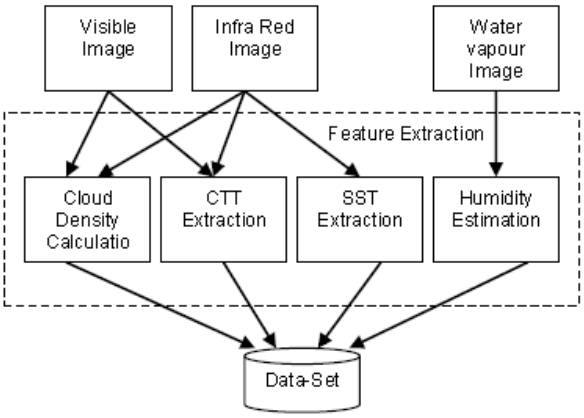


***Fig 2: Block diagram for obtaining the data set values.***

In the above figure, all the different types of images are grouped together to extract several parameters like cloud density, CTT, SST and humidity of different regions that we've considered. All these data are then collected wholly to form a data set.

As mentioned earlier, we have used K-NN algorithm in this approach and the K-nearest neighbor value is supposedly taken to be 7.

The following figure (fig: 3) shows the block diagram of first of the algorithm. In stage-1, various images are extracted from the satellite and are processed and extracted.

The resultant data is stored on the shared memory and is then processed to obtain the required dataset.

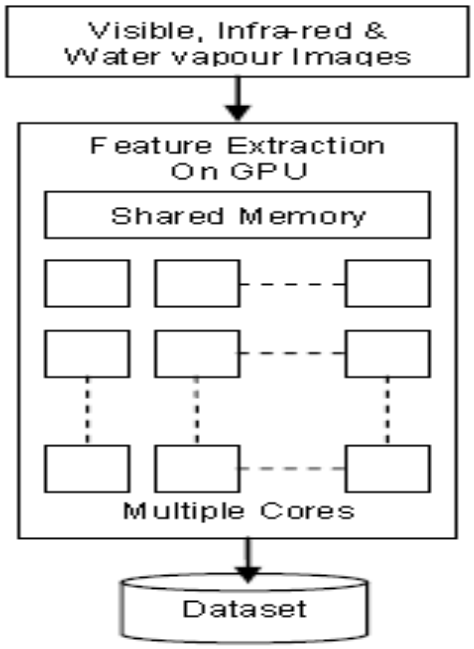


***Fig 3: Block diagram of stage one of the algorithm***.

Now comes the task of finding the Euclidean distance. For this, we use the following formula.

$$\text{Euclidean distance} = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

, where 'n' is the number of regions that we selected.

In the following stage, i.e. the clustering stage, we group all the like objects into a class so as to obtain a class of relative objects. The below figure (fig 4) shows the flowchart of the clustering stage (stage-2) of our algorithm.

When IR, Visible and water vapor images of a day are presented for onset prediction, the parameters for 'n'

regions are extracted. Then the parameter's Euclidean distance is calculated using the centroids of each cluster. Using K-NN algorithm, each record in the closest cluster is compared. The considered $k_n$ value gives us an idea of the "Onset Days" left. Hence, monsoon can be predicted [1].

The K-NN model does not produce the exact value from the historic record. Therefore, it can be concluded that this approach is fairly generic and can be changed from places to places with changing data set.

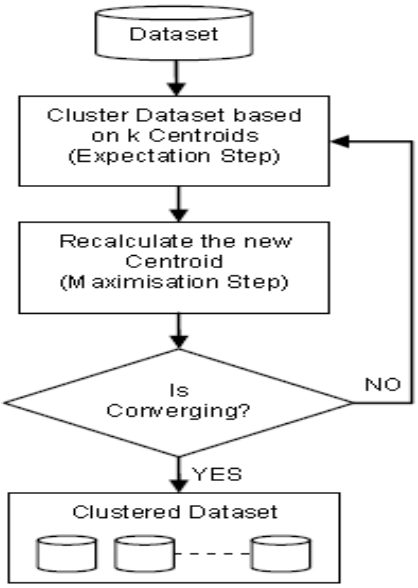


**Fig 4: Stage-2 (Clustering) flowchart.**

## IV. K-MEANS ALGORITHM.

K-means algorithm is a well-known algorithm to cluster the objects. It is popular because of its minimal time complexity as compared to other clustering algorithms. The other main advantage of using K-mean algorithm is that it is very efficient for very large dataset and the clustering quality is very good.

A. Disadvantages of using K-Means Algorithm:

- Difficulty in comparing the quality of clusters produced, which in-turn affects the resultant K.
- A fixed number of clusters make it difficult to predict the exact value of K.
- Fails to work with non-globular clusters.
- Clustering the initial partitions results in different final cluster, this again makes the prediction difficult.

Hence, K-means algorithm is seldom preferred and less implemented.

## V. REGRESSION.

Regression is a statistical empirical method, that uses the relation between two or more relations obtained from an observational dataset so that the outcome from the relation can be predicted from the others.

There are two types of regressions namely, Simple Linear Regression and Multiple Linear model.

Simple linear regression is of the form:

$Y= w_0 +w_1x$

where, 'x' is the predictor value.

Multiple Linear Regression (MLP): is a type of regression technique which has more than 2 predictor variables.

In this paper, we use MLP to predict the summer monsoon for the current year using the summer monsoon dataset obtained from the previous year. After computation, the MLP we get is,

Multiple linear regression is of the form:

$Y= ax_1+bx_2+cx_3$
where a, b and c are regression coefficients.

And,
$x_1$= rainfall in the month of September in the previous year
$X_{2=}$ rainfall in the month of October in the previous year.
$X_3$= rainfall in the month of November in the previous year.
Y= average rainfall of the present year.

It can be noted that, this method does not employ the prediction of rainfall only for the current year. Rather, this method can be employed to predict rainfall for the coming years using the previous year's rainfall data.

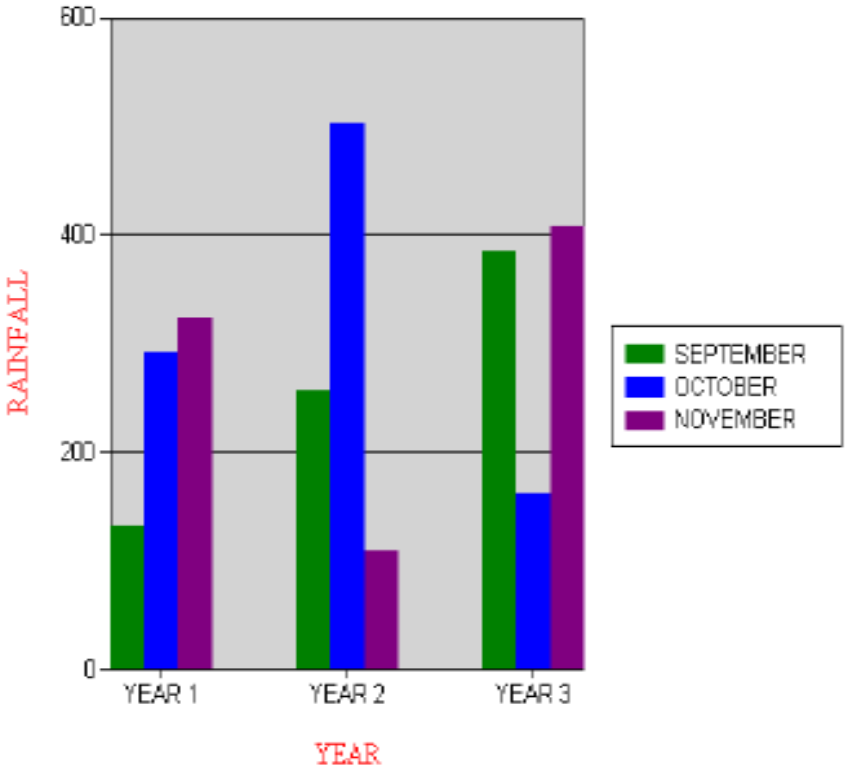


**Fig 5. Prediction of rainfall for the next 3 years.**

The above figure shows the implementation of the relevant data set to predict rainfall for the months September through November for the next three years. From the figure, it can be notes that there is neither continuous nor consistent amount of rainfall.

A. Disadvantages of Multiple Linear Regression:

The main and probably only disadvantage of MLR is that this becomes very complex if simple statistics is not used.

## VI. MULTIPLE LINEAR REGRESSION VERSUS ARTIFICIAL NEURAL NETWORK.

From the above study, it can be noted that the Multiple Linear Regression is,

$$\hat{y} = a_0 + \sum_{i=1}^{n} a_i x_i.$$

**Fig 6. MLR general form.**

While the general method for predicting the rainfall in Artificial Neural Networks(ANN) involves three stages. The first stage is the input stage, the second stage is the hidden stage while the third stage is the output stage. With the complexity associated with this algorithm, it is certain that ANN is the best approach to predict the rainfall, although the MLR is no less. Since, our paper is not concerned with ANN method, we just bring it to the notice and not deal with it in detail.

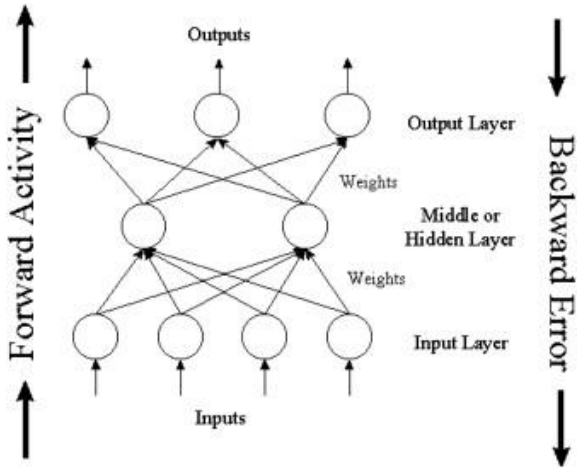


**Fig 7. Typical ANN Architecture.**

Given above is a typical ANN architecture, using which we can deduce a system, which is capable of accepting a dataset and using which, a monsoon forecasting model is constructed.

The main and possibly the only advantage of this model is that it can be used to find a possible error in the model, which is not possible in the other models stated above.

## VII. IMPLEMENTATION.

For the stated methods to be implemented, we make use of a certain chosen areas of our sub-continent. Once we mark the required region to be considered, we collect the previous year's statistical data and feed them to the algorithms that are mentioned above. Here, it is important to note that we have considered only two regions. As we mainly concentrate on the south western monsoon, we keep in mind the two main regions and mark them in red to show the relevance in acquiring the data for our implementation.

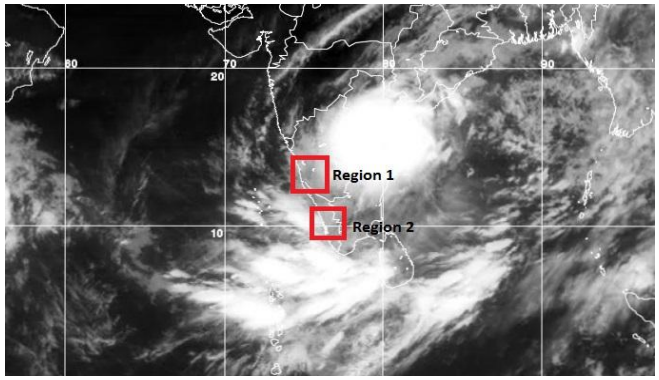


**Fig 8. The 2 regions from where the features would be extracted.**

The table given below shows the data set acquired for our study. This table mainly consists of the dates under which we consider the data, the Sea Surface Temperature (SST), Cloud Top Temperature (CCT), cloud density and Humidity of the 2 regions. Since the south western monsoon starts in month of April, we consider 4 days in the same month for the calculation of the onset date.

| Regions → | Region 1 | | | | Region 2 | | | |
|---|---|---|---|---|---|---|---|---|
| **Dates** | **SST** | **CCT** | **Cloud density** | **Humidity** | **SST** | **CCT** | **Cloud density** | **Humidity** |
| 25/04/2006 | 53 | 0 | 0 | 25 | 50 | 0 | 0 | 25 |
| 26/04/2006 | 90 | 0 | 0 | 28 | 58 | 0 | 0 | 27 |
| 27/04/2006 | 90 | 0 | 0 | 30 | 55 | 0 | 0 | 32 |
| 28/04/2006 | 94 | 0 | 0 | 28 | 56 | 0 | 0 | 28 |

| **Onset Days Left** |
|---|
| 31 |
| 29 |
| 30 |
| 27 |

**Table 1: Sample data set for 2 regions.**

Thus, it is found that 0 is given in places of cloud density and humidity. This shows that there are no clouds present. Thus, by employing the algorithms, we find the offset days left for the approaching monsoon.

## VIII. CONCLUSION.

Rainfall time series may be unfounded. The topic of monsoon-rainfall data series is highly complex; the role that multiple linear regressions might play in this topic is one for future research—it appears, from the evidence here, not to be useful as a predictive model.

Firstly, an introduction of satellite images is given. Following that, a two-stage Monsoon Prediction Algorithm is described. Later, the Multiple Regression Technique is shown. Along with this an introduction of Artificial Neural Network is also given, thereby drawing comparisons between them.

From all the above models that are given, it can be noted that MLR is the best algorithm to predict the rainfall. However, it should be noted that this model, like any other model, has few flaws. So, ANN can be used instead. As mentioned earlier, the main advantage of this model is that it has error finding capability, which is not present in other models. And, another main advantage is that- the ANN algorithm can be used to predict the monsoon in spite of uncertain dataset and all other chaos. While this feature is not found in any other method, the ANN method surely seems the future.

## IX. FUTURE STUDY.

As depicted above, the ANN method is an excellent and efficient way to predict monsoon. While it is complex, it is also effective and decisive in many ways over other techniques. Given its advantage of finding the error and accurate prediction in spite of uncertain and sometimes unavailable dataset, this method surely here to stay.

With further deep understanding and research of the topic, this technique can be used to provide the accurate and trust-worthy prediction of monsoon showers for the country.

## X. REFERENCES.

[1]. Dinu John,K. K. Sindhu,B. B. Meshram, "*Two Stage Data Mining Technique for Fast Monsoon Onset Prediction*",
International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, 2012.

[2]. M.Kannan, S.Prabhakaran, P.Ramachandran, "*Rainfall Forecasting Using Data Mining Technique*", International Journal of Engineering and Technology Vol.2 (6), 2010

[3]. G. D'souza, E.C. Barrett, C.H. Power (1990): "*Satellite rainfall estimation techniques using visible and infrared imagery"*, Remote Sensing Reviews, 4:2, 379-414

[4]. J. K. Mishra, O. P. Sharma, " *Cloud top temperature based precipitation intensity estimation using INSAT-1D data*", International Journal of Remote Sensing 2001, 22:6, 969-985

[5]. Tao Chen, Milcio Talagi, ―"*Rainfall prediction of geostationary meteorological satellite images using artificial neural network"*, International Geoscience and Remote Sensing Symposium 1993

[6]. Indian Meteorological Department, http://www.imd.gov.in

[7]. Klaush Juliseh, "*Data mining for Intrusion Detection – A critical review*", Applications of Data mining in computer security, Daniel Barbara, Sushil jajodia, Published by Springer.

[8]. Guhathakurta, P (2005) "*Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network*", Current Science, 90, 773-779

[9]. P.T. Nastos , K.P. Moustris, I.K. Larissi , A.G. Paliatsos, " Rain Intensity Forecasting Using Artificial Nural Networks in in Athens, Greece", published in Elsevier, Volume 119, January 2013.

[10]. Ashis K. Mitra, E. N. Rajagopal, G. R. Iyengar, D. K. Mahapatra, I. M. Momin, A. Gera, K. Sharma, J. P. George, R Ashrit, M. Dasgupta, S. Mohandas, V. S. Prasad, Swati Basu, A. Arribas, S. F. Milton, G. M. Martin, D. Barker and M. Martin, "Prediction Of Monsoon Using Seamless Coupled Modelling System", Current Sciences, Volume . 104, No. 10, 25 May 2013.

[11]. Charney, J. G. and Shukla, J., Predictability of monsoons. In *Monsoon Dynamics* (eds Lighthill, J. and Perace, R. P.), Cambridge
University Press, Cambridge, 1981, pp. 99–109.

[12]. Ghelli, A., Cloke, H. and Kulkarni, A., Monsoons: prediction, variability and impact. *Meteorol. Appl*. (*Spec. Issue*), 2012, **19**