STUDY OF RAINFALL PATTERN IN BANGALORE

Research work submitted to CHRIST (Deemed to be University)

Bachelor of Science
in
COMPUTER SCIENCE, MATHEMATICS, STATISTICS

by
ADARSH JAYAKUMAR
K HARI
PRAKHAR SRIVASTAVA

Under the Supervision of
Dr. Hemlata Joshi
Assistant Professor

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

DEPARTMENT OF STATISTICS
CHRIST (Deemed to be University)
BANGALORE

September 2019

DECLARATION

We declare that the research work entitled "STUDY OF RAINFALL PATTERN IN BANGALORE" is a record of original research work undertaken by us under the supervision of Dr. Hemlata Joshi, Assistant Professor, Department of Statistics, CHRIST (Deemed to be University), Bangalore and has not formed the basis for the award of any degree, diploma, associateship, fellowship etc.

Place: Bangalore

Date:

Signature of candidates

Adarsh Jayakumar(1740201)

K. Hari(1740212)

Prakhar Srivastava(1740214)

CERTIFICATE

This is to certify that the research work submitted by Adarsh Jayakumar , K Hari, Prakhar Srivastava, (1740201, 1740212, 1740214), entitled “STUDY OF RAINFALL PATTERN IN BANGALORE” is a record of original research work carried out during the academic year 2018 – 2019 under my supervision.

Place: Bangalore Signature of the Supervisor

Date:

Study of Rainfall Pattern in Bangalore

****

Prakhar Srivastava
Department of statistics
Christ (Deemed to be university)
Bengaluru, India
prakhar.srivastava@science.christuniversity.in

Adarsh Jayakumar
Department of statistics
Christ (Deemed to be university)
Bengaluru, India
adarsh.jayakumar@science.christuniversity.in

K.Hari
Department of statistics
Christ (Deemed to be university)
Bengaluru, India
k.hari@science.christuniversity.in

ABSTRACT

Temporal change in rainfall erosivity varies due to the rainfall characteristic (amount, intensity, frequency, duration), which affects the conservation of soil and water. Examining the spatiotemporal dynamics of meteorological variables in the context of changing climate, particularly in countries where rainfed agriculture is predominant, is vital to assess climate-induced changes and suggest feasible adaptation strategies. This study illustrates the variation of rainfall erosivity due to changing rainfall in the past and the future. Variation in these parameters due to climate change might have an impact on the future water resource of the study area, which is mainly an agricultural based region, and will help in proper planning and management. Examining the spatiotemporal dynamics of meteorological variables in the context of changing climate, particularly in countries where rainfed agriculture is predominant, is vital to assess climate-induced changes and suggest feasible adaptation strategies. Time series modeling and forecasting has fundamental importance to various practical domains. Time series analysis can be used in a multitude of business, weather forecasting and describing a quantity into the future and explaining its historical patterns. Here are just a few examples of possible use cases, explaining seasonal patterns in sales, predicting the amount of seasonal changes in weather across an area, estimating the effect of a newly launched product on number of sold units, detecting unusual events and estimating the magnitude of their effect

The paper implores data collected from 2009 – 2018 of the rainfall patterns in Bangalore to analyze an array of data to find the best prediction of rainfall in further years along with.

Keywords—Rainfall, Future Prediction, Forecasting, ARIMA model, Time Series Analysis.

I. INTRODUCTION

The Indian economy is based on agriculture and its products, and crop yield is heavily dependent on the summer monsoon (June-September) rainfall. Therefore, any decrease or increase in annual rainfall will always have a severe impact on the agricultural sector in India. About 65% of the total cultivated land in India is under the influence of rainfed agriculture system . Therefore, prior knowledge of the monsoon behavior (during which the maximum rainfall occurs in a concentrated period) will help the Indian farmers and the Government to take advantage of the monsoon season. This knowledge can be very useful in reducing the damage of crops during less rainfall periods in monsoon season. Therefore, forecasting the monsoon temporally is a major scientific issue in the field of monsoon meteorology. The ensemble of statistics and mathematics has increased the accuracy of forecasting of ISMR up to some extent. But due to the non-linear nature of ISMR, its forecasting accuracy is still below the satisfactory level. In 2002, IMD failed to predict the deficit of rainfall during ISMR, which led to considerable concern in the meteorological community. In

2004, drought was again observed in the country with a deficit of more than 13% rainfall which could not be predicted by any statistical or dynamic model.

We already know that heavily populated cities like Chennai are already under serious water crisis and nothing but rainfall can help them to get through that water crisis. Due to urbanization we see a lot of abnormalities in the trends of rainfall. It has decreased over a period of time. This could have been caused due to multiple reasons such as forest cover, atmospheric pressure, temperature and so on. We have noticed serious urbanization especially in past decade and the cities are still urbanizing at a very alarming rate. Therefore, the past decade is considered for collecting data of precipitation and various other factors. We believe it would help us analyze and the significant factors and predict the future rainfall patterns on the basis of the current environment.

A common assumption in many time series techniques is that the data are stationary.

A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

We have chosen the Bangalore rainfall in particular since of a combined variability in it's trend, which is different from the rest of India.

Our general aim was to study the dataset(obtained from various government sites)for accuracy, which gives the rainfall statistics of all the 10 years of Bangalore rainfall, from the years 2009 to 2018 to infer.

II. DATASET

This dataset contains the rain statistics for 10 years, from 2009 to 2018 with each row representing a different precipitation amount for each individual month of each individual year.

A. *Methodology*

We used for collecting data in a tabular format using excel. For this particular paper a literature study of 9 papers were done. EBSCO, Google Scholar, ScienceOpen, Academia.edu were used to get access to published researches.

Bibliographies of several papers were also examined to get other relevant studies. Certain websites that did not include published scientific articles were also used as a point of reference. Our main focus was on Rainfall in Bangalore.

The sample size was calculated from various government sites but most had data about precipitation(rain+snowfall+thunderstorm+hailstorm) , thus the data was filtered after collection to just Rainfall for our study to be significant.

Statistical tools used for the analysis of data are ARIMA, SARIIMA, R (with RStudio), Microsoft Excel was used to visualize, clean, pre-process and analyze data.

A. Stationarity: For practical purposes, stationarity can usually be determined from a run sequence plot.

Transformations to Achieve Stationarity

If the time series is not stationary, we can often transform it to stationarity with one of the following techniques.

1. We can difference the data. That is, given the series Zt, we create the new series

   $Yi=Zi-Zi-1$.

   The differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is usually sufficient.
2. If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove long term trend, a simple fit, such as a straight line, is typically used.
3. For non-constant variance, taking the logarithm or square root of the series may stabilize the variance. For negative data, you can add a suitable constant to make all the data positive before applying the

transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

The above techniques are intended to generate series with constant location and scale. Although seasonality also violates stationarity, this is usually explicitly incorporated into the time series model.

B. ARIMA MODEL

ARIMA stands for auto-regressive integrated moving average. ARIMA models are the most general class of models for
time series forecasting which can be stationarized by transformations such as differencing and logging. It was
introduced by Box and Jenkins (1970) which includes autoregressive as well as moving average parameters, and
explicitly includes differencing in the formulation of the model. It's a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

1. A pattern of growth/decline in the data is accounted for (hence the "auto-regressive" part)
2. The rate of change of the growth/decline in the data is accounted for (hence the "integrated" part)
3. Noise between consecutive time points is accounted for (hence the "moving average" part)

$$Y_t = c + \phi_1 y_{dt-1} + \phi_p y_{dt-p} + \ldots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t$$

where e is the error term and c is the constant

ARIMA models are typically expressed like "ARIMA(p,d,q)", with the three terms p, d, and q defined as follows:

- p means the number of preceding ("lagged") Y values that have to be added/subtracted to Y in the model, so as to make better predictions based on local periods of growth/decline in our data. This captures the "autoregressive" nature of ARIMA.

- d represents the number of times that the data have to be "differenced" to produce a stationary signal (i.e., a signal that has a constant mean over time). This captures the "integrated" nature of ARIMA. If d=0, this means that our data does not tend to go up/down in the long term (i.e., the model is already "stationary"). In this case, then technically you are performing just ARMA, not AR-I-MA. If p is 1, then it means that the data is going up/down linearly. If p is 2, then it means that the data is going up/down exponentially.

- q represents the number of preceding/lagged values for the error term that are added/subtracted to Y. This captures the "moving average" part of ARIMA.

C. Time Series Forecasting:

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends.

It is important because there are so many prediction problems that involve a time component. These problems are neglected because it is this time component that makes time series problems more difficult to handle.

Time series analysis involves developing models that best capture or describe an observed time series in order to understand the underlying causes. This field of study seeks the "why" behind a time series dataset.

D. ADF test

An augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity
This is a unit root test for stationarity.

E. PACF Test

In time series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags

B. Literature Review

Components of Time Series

Level. The baseline value for the series if it were a straight line.

Trend. The optional and often linear increasing or decreasing behavior of the series over time.

Seasonality. The optional repeating patterns or cycles of behavior over time.

Noise. The optional variability in the observations that cannot be explained by the model.

Concerns of forecasting:

How much data do you have available and are you able to gather it all together? More data is often more helpful, offering greater opportunity for exploratory data analysis, model testing and tuning, and model fidelity.

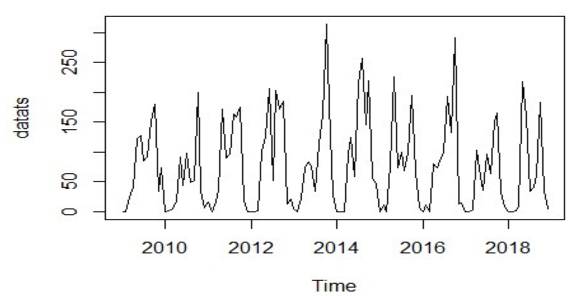
What is the time horizon of predictions that is required? Short, medium or long term? Shorter time horizons are often easier to predict with higher confidence.

Can forecasts be updated frequently over time or must they be made once and remain static? Updating forecasts as new information becomes available often results in more accurate predictions.

At what temporal frequency are forecasts required? Often forecasts can be made at a lower or higher frequencies, allowing you to harness down-sampling, and up-sampling of data, which in turn can offer benefits while modeling

III. ANALYSIS

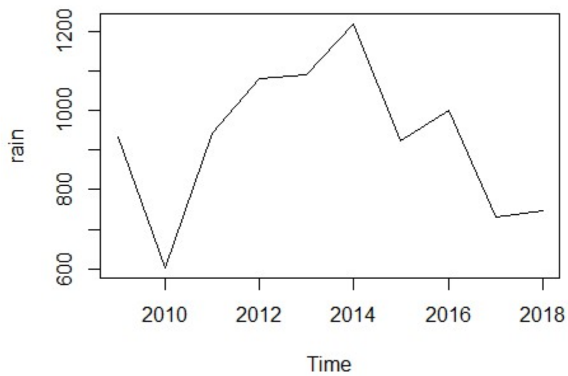
DATA VISUALISATION



This graph shows us the proper distribution of rain according to months for the previous 10 years for which the data was collected.

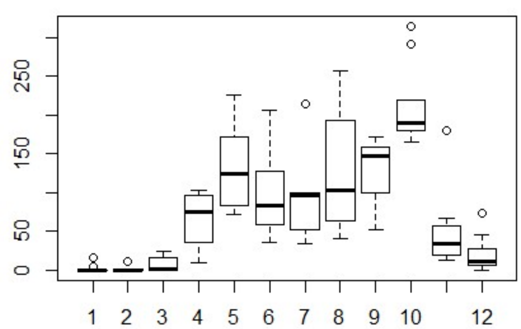
You can see from the plot that there is roughly constant level (the mean stays constant at about 150 mm). The random fluctuations in the time series seem to be roughly constant in size over time, so it is probably appropriate to describe the data using an additive model. Thus, we can make forecasts using simple exponential smoothing.

Thus from the plotting we can assume that 2014 experienced the maximum amount of rainfall, so



By this analysis we can now conclude that in the last 10 years, 2014 was the year of highest rainfall with most of it being experienced in October.

SUMMARY OF THE DATA



This graph shows us that from the months January to March and December the mean rainfall is similar i.e. the amount of rainfall experienced by Bangalore in those months is almost the same.

Then for the month of April a pretty huge variability is observed than March, such that even the minimum rainfall of April is almost equal to the maximum rainfall of March.

Further in the proceeding months the mean of May is showing high difference than the mean of June, July and August which are almost equal.

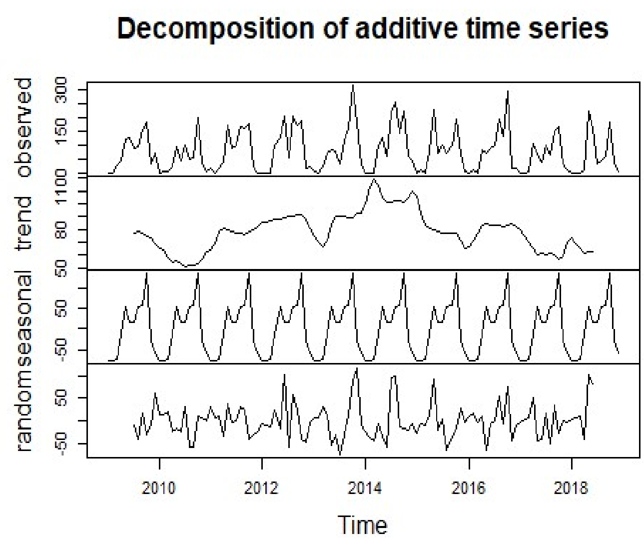
The maximum values of June and July are almost the same.

May having high similar mean with September but have a higher amount of maximum rainfall.

The highest rainfall of the whole year is experienced in October, and also the maximum amount of rainfall in the whole year is also experienced in October as the Mean of October is having a properly noticeable difference as compared to other months.

But after October as we move onto November there is a huge drop in mean Rainfall of both the months, with November experiencing almost the same amount of rainfall as the first 3 months. And in only few rare cases the rainfall of November is able to reach the lower limit of October.

Decomposition of the Data



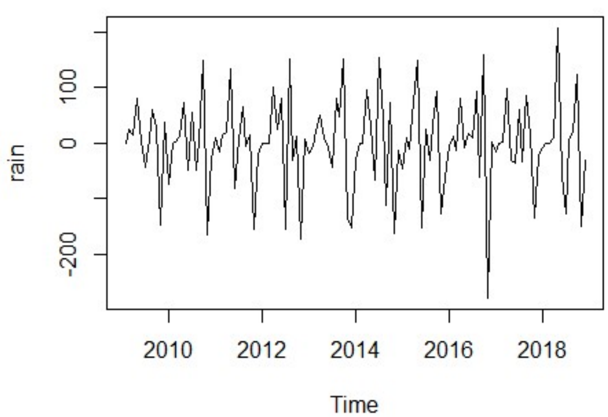
- The graph shows a number of component series from the observed time series, where each one of these has a certain behavior. It is decomposed into the following:

- the original time series (1st),
- the estimated trend component (2nd)
- the estimated seasonal component (3rd)
- The estimated irregular component (4th).

We see that the estimated trend component shows a small decrease from about 80 in late 2009 to about 50 in early 2011, followed by similar fluctuations till 2014 and then gradually decreasing trend till 2018. The sum of the trend, seasonal, and random components is equal to the observed series.

Prediction

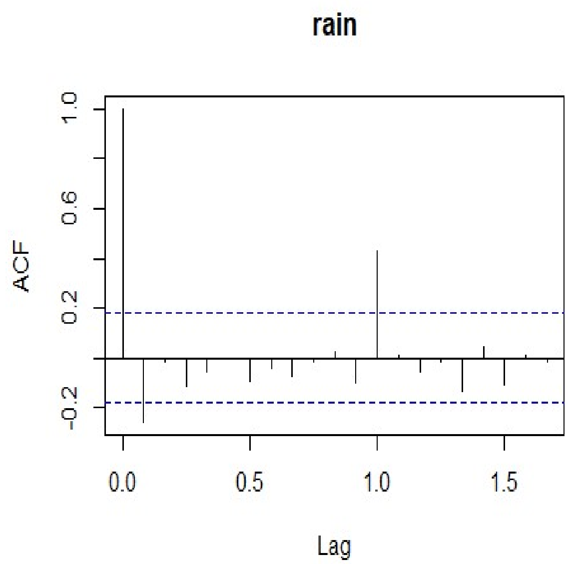
The forecast package provides functions for the automatic selection of exponential and ARIMA models. The **ets()** function supports both additive and multiplicative models. The auto.arima() function can handle both seasonal and nonseasonal ARIMA models. Models are chosen to maximize one of several fit criteria.



This graph is the plot of the values from the diff() function of the data This function takes two arguments of note. The first is the lag, which is the number of periods, and the second is differences which is the order of the difference (e.g. how many times diff()is called) which in this graph lag value = 4 and Difference is called 120 times.

Augmented Dickey-Fuller Test

Dickey-Fuller value = -6.5321
Lag order(df) = 4
p-value=0.01

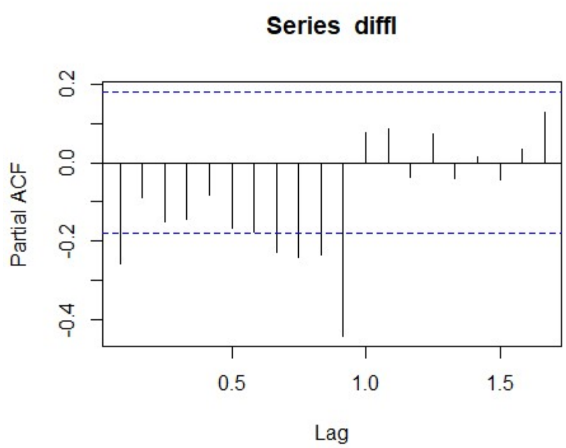


The autocorrelation function acf() gives the autocorrelation at all possible lags. The autocorrelation at lag 0 is included by default which always takes the value 1 as it represents the correlation between the data and themselves. As we can infer from the graph above, the autocorrelation continues to decrease as the lag increases, confirming that there is no linear association between observations separated by larger lags. From this graph the p value obtained is 1

The autocorrelation function acf() gives the autocorrelation at all possible lags. The autocorrelation at lag 0 is included by default which always takes the value 1 as it represents the correlation between the data and themselves. As we can infer from the graph above, the autocorrelation continues to decrease as the lag increases, confirming that there is no linear association between observations separated by larger lags.

Partial Autocorrelation Function(PACF)

Partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. Pacf() at lag k is autocorrelation function which describes the correlation between all data points that are exactly k steps apart after accounting for their correlation with the data between those k steps. It helps to identify the number of autoregression (AR) coefficients(p-value) in an ARIMA model.



Pacf of the data will give us the q value which in this case is 0.

Interpretation

```
Coefficients:
         ar1      ma1    sar1     sma1  intercept
      0.3143  -0.0728  0.9998  -0.9745    77.2111
s.e.  0.2776   0.2866  0.0009   0.0559    21.7558

sigma^2 estimated as 1887:  log likelihood = -638.37,  aic = 1288.74
```

This is the data obtained for seasonal arima model with very less standard error(s.e.)

Then for Arima Fit,

```
test=ts(datats, frequency=12, start=c(2009,1), end=c(2018,12))
fit=arima(test,  c(0,1, 1), seasonal=list(order=c(0,1,1), period=12))
fit
```

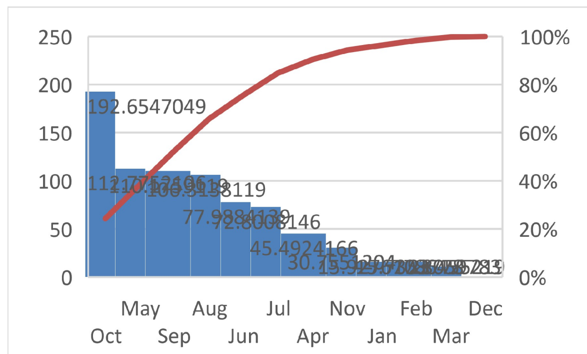
This gave the fitted value for the ARIMA model which resulted to

```
Coefficients:
          ma1     sma1
      -0.9103  -1.0000
s.e.   0.0481   0.1115

sigma^2 estimated as 1950:  log likelihood = -572.1,  aic = 1150.21
```

And in the fitted values the standard error was still almost negligible.

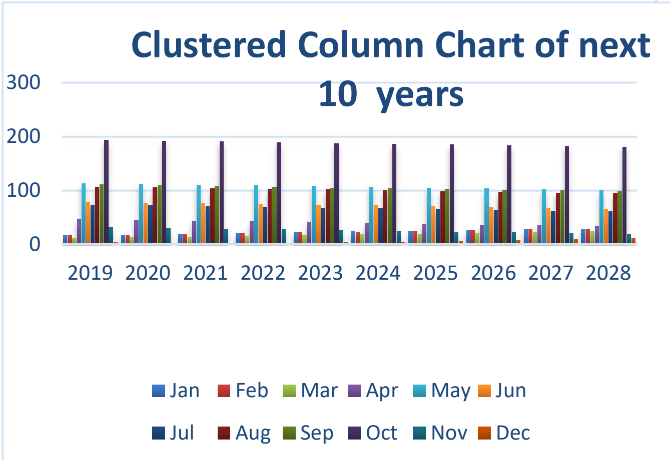
Later the future predictions for the next 10 years is made using the Predict() function.



This graph is used to show the trends of rainfall in the next year i.e. 2019 by plotting the predicted values obtained using predict function. The values obtained show the estimate amount of rainfall that each month will experience in the future years.

According to the plot, in Bangalore almost 78% of rainfall contribution will be experienced during the month of October having a maximum value of 192.6547 mm, followed by May with a rainfall amount of 112.7752 mm and the rest months are following them in a descending order.

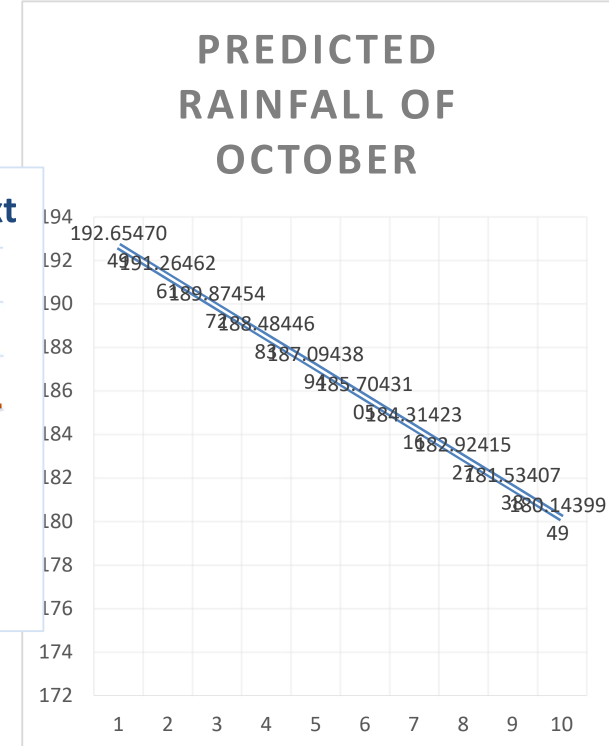
As we interpreted earlier through Boxplot about the rainfall pattern which showed that October experiences the maximum rain for the last 10 years with a high marginal difference between it's mean and other months mean, is continuing in the year 2019. And also the second highest rainfall receiving month May is still at its position but the months next to it don't have a high difference from May, thus it can be assumed that in further years May might loose it’s second place to either September or August.



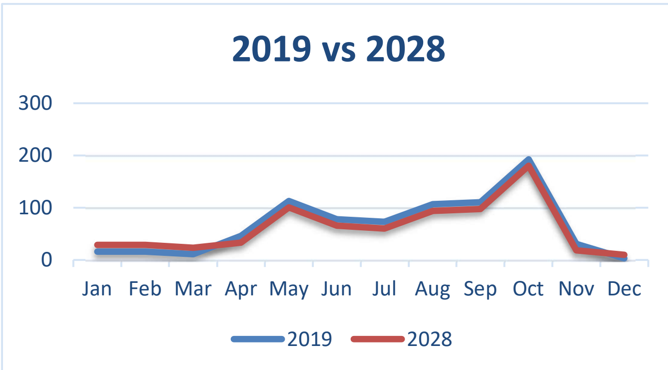
This graph is showing the trends of rainfall in the next 10 years i.e. 2019-2028, clustered with the data of each month for the next 10 years. From the graph we can see that the month of October receives the highest rainfall. we can easily study the trend followed by the graph. but it is gradually decreasing as the years go through. Same is the case with every month. Not a very big difference is observed between two consecutive values of the same month in different years

This graph is the rainfall values of October in the next 10 years i.e. from 2019 – 2028. Now this data is obtained by plotting the predicted values obtained from the predict function.

Now in this graph we can clearly observe a falling trend in the amount of rainfall experienced in October, which is from



192.6547049 in 2019 the amount of rainfall reduced to 180.1439949 in 2028.This means that in just 10 years there is a 12mm decrease in the rain amount experienced which further might change obviously, but assuming that it will reduce will be invalid because Temperature is a factor that affects the amount rainfall, i.e. more the temperature more the evaporation thus more the rainfall, and the temperature in increasing at an alarming rate, so predicting the amount of rainfall with previous year’s data and not including temperature in the analysis might be considered incomplete.Thus, the rainfall pattern can now be considered to decrease but also it is just estimate data so proper prediction can’t be made.



This graph is a final comparison between the years 2019 and 2028 with respect to rainfall showing the change in amount of rainfall experienced in Bangalore in these 2 years having a 10 years gap between them. This will help us to observe all the trends of rainfall and how they changed over a period of time and during which time period was the change most..

RESULT TABLE

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2019	15.92768	15.73538	10.80598	45.49242	112.7752	77.98841	72.80081	106.3138	110.1759	192.6547	30.75512	2.645522
2020	17.31776	17.12546	12.19606	44.10234	111.3851	76.59834	71.41074	104.9237	108.7858	191.2646	29.36504	1.255443
2021	18.70784	18.51554	13.58614	42.71226	109.9951	75.20826	70.02066	103.5337	107.3958	189.8745	27.97496	0.134636
2022	20.09791	19.90561	14.97621	41.32218	108.605	73.81818	68.63058	102.1436	106.0057	188.4845	26.58488	1.524715
2023	21.48799	21.29569	16.36629	39.9321	107.2149	72.4281	67.2405	100.7535	104.6156	187.0944	25.1948	2.914794
2024	22.87807	22.68577	17.75637	38.54202	105.8248	71.03802	65.85042	99.36342	103.2255	185.7043	23.80473	4.304873
2025	24.26815	24.07585	19.14645	37.15194	104.4347	69.64794	64.46034	97.97334	101.8354	184.3142	22.41465	5.694952
2026	25.65823	25.46593	20.53653	35.76186	103.0447	68.25786	63.07026	96.58326	100.4454	182.9242	21.02457	7.08503
2027	27.04831	26.85601	21.92661	34.37179	101.6546	66.86778	61.68018	95.19318	99.05528	181.5341	19.63449	8.475109
2028	28.43839	28.24609	23.31669	32.98171	100.2645	65.4777	60.2901	93.8031	97.6652	180.144	18.24441	9.865188

CONCLUSION:

Analysis of long term rainfall data from Bangalore showed a clear trend in it’s occurrence. Also a proper estimate of rainfall was obtained for the next 10 years using Time Series Analysis. Mainstreaming of climate change and variability into policies and promulgation of adaptation strategies via education, research, and extension is critical. Further studies are necessary to advance our understanding on rainfall and variability to validate the close nexus between climate change and alterations in agro-ecosystems, and examine and strengthen community adaptation strategies in Bangalore.

There was a need for development of a more homogeneous (spatially and temporally) rainfall time series for the Bangalore region using the latest data. The newly constructed rainfall series focuses upon Bangalore and it represents almost all the exist-ing administrative districts. The present study brings out some of the interesting and also significant changes in the rainfall pattern of the country. The alternating sequence of multi-decadal periods of 20 years having frequent droughts and flood years are observed. Most of the studies on rainfall trend/analysis were only limited to seasonal monsoon rainfall only. We have also studied contribution of each of the major rain producing month’s (i.e. June, July, August and September) in annual rainfall and examine whether there is any significant change in their contribution. The iJune rainfall is getting importance as its contribution to annual rainfall is increasing. But contribution of August rainfall is increasing in all these areas. Significant increasing trend is also observed in the annual rainfall.

References

1. Aftab, S., Ahmad, M., Hameed, N.,

Salman, M., Ali, I. and Nawaz, Z. (2018). Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications*, 9(5).

2. Haviluddin and Alfred, R. (2014). Forecasting Network Activities Using ARIMA Method. *Journal of Advances in Computer Networks*, 2(3), pp.173-177.
3. Choden and Unhapipat, S. (2018). ARIMA model to forecast international tourist visit in Bumthang, Bhutan. *Journal of Physics: Conference Series*, 1039, p.012023.
4. Mahanti, N. (2005). *Predicting meteorological events*. New Delhi: Narosa Pub.
5. Mondal, A., Khare, D. and Kundu, S. (2016). Change in rainfall erosivity in the past and future due to climate change in the central part of India. *International Soil and Water Conservation Research*, 4(3), pp.186-194.
6. Kashid, S. and Maity, R. (2012). Prediction of monthly rainfall on homogeneous monsoon regions of India based on large scale circulation patterns using Genetic Programming. *Journal of Hydrology*, 454-455, pp.26-41.
7. Asfaw, A., Simane, B., Hassen, A. and Bantider, A. (2018). Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. *Weather and Climate Extremes*, 19, pp.29-41.