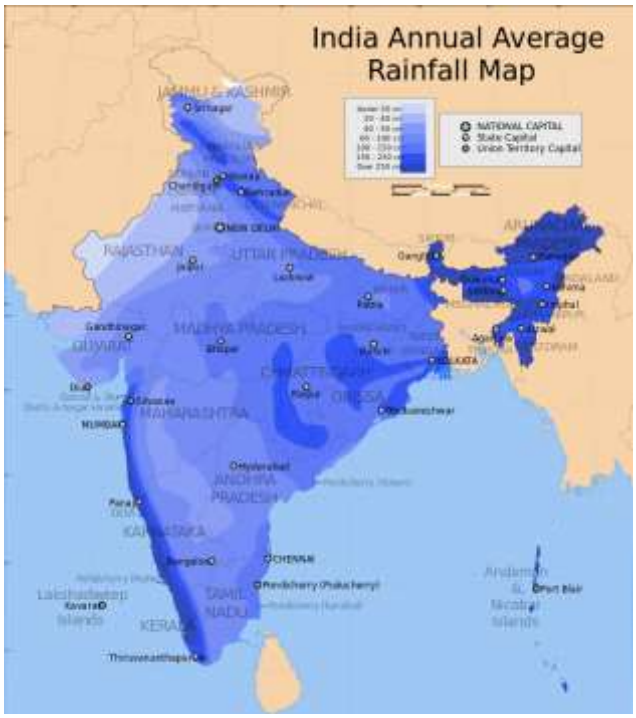
TIME SERIES ANALYSIS: ANALYSIS OF RAINFALL DATA IN INDIA (PART 1)

-Aastha Sumra, Kartikeya Sinha, Navya Garg, Nishika Taneja, Riya Khandelwal

ABSTRACT

Preliminary research of the rainfall phenomenon in India shows that it is dependent on various climatic control factors, which can be broadly divided as- latitudes and altitudes, pressure and wind system, distance from the sea (continentality), ocean currents and relief features. These are heavily interdependent and interconnected with each other, hence providing a seasonal alteration of the wind systems. This brings provides a rhythmic cycle of seasons. Some factors, such as El-Nino and La-Nina, affect the Indian rainfall on a cyclical basis. However, their effect is comparatively less pronounced than the seasonal variations. Moreover, factors like global warming, climate crisis, human intervention in various natural phenomenon effect the rainfall patterns in the long run.



Due to variations in the geographical conditions of the country, the pattern of rainfall varies significantly as one goes from one region to another. We consider for our study the states of **Kerala and Punjab.** Kerala receives maximum rainfall during *south-west monsoon* (months June-September), along with some rainfall during *north-east monsoon* (months October-December), while Punjab receives rainfall during the *monsoon season* (July-September) only.

Due to the interdependency of the climatic factors, we have assumed a multiplicative model for our study.

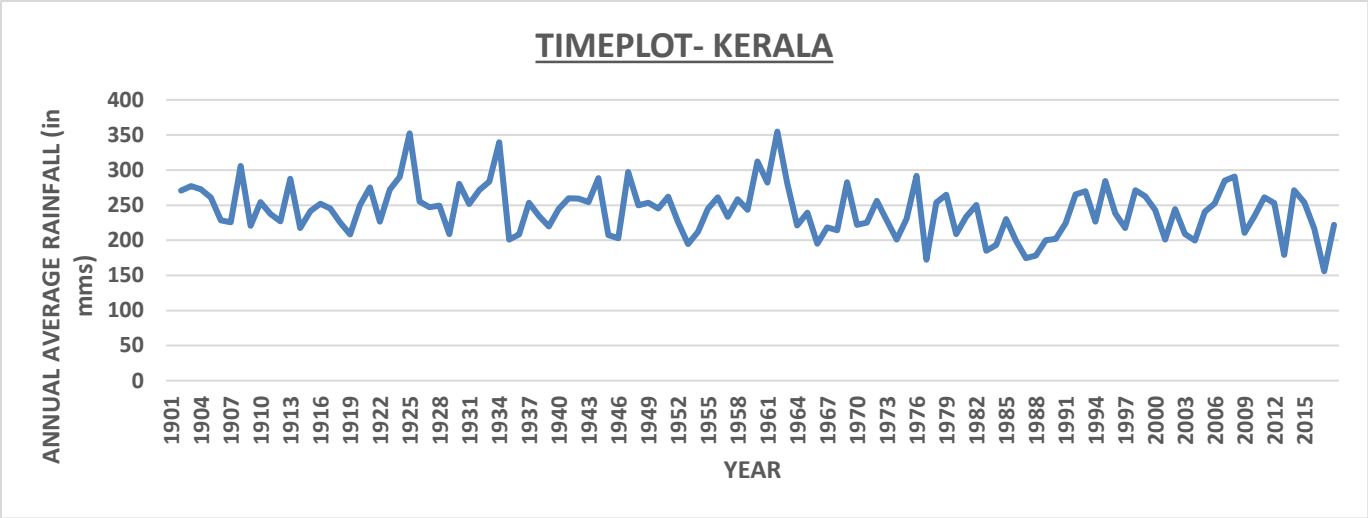
Our study verifies the above stated research. The state of Kerala shows a linear decreasing trend, with seasonal indices indicating that the rainfall occurs in the state mainly in 2 periods- maximum during June-September, and during October-November.

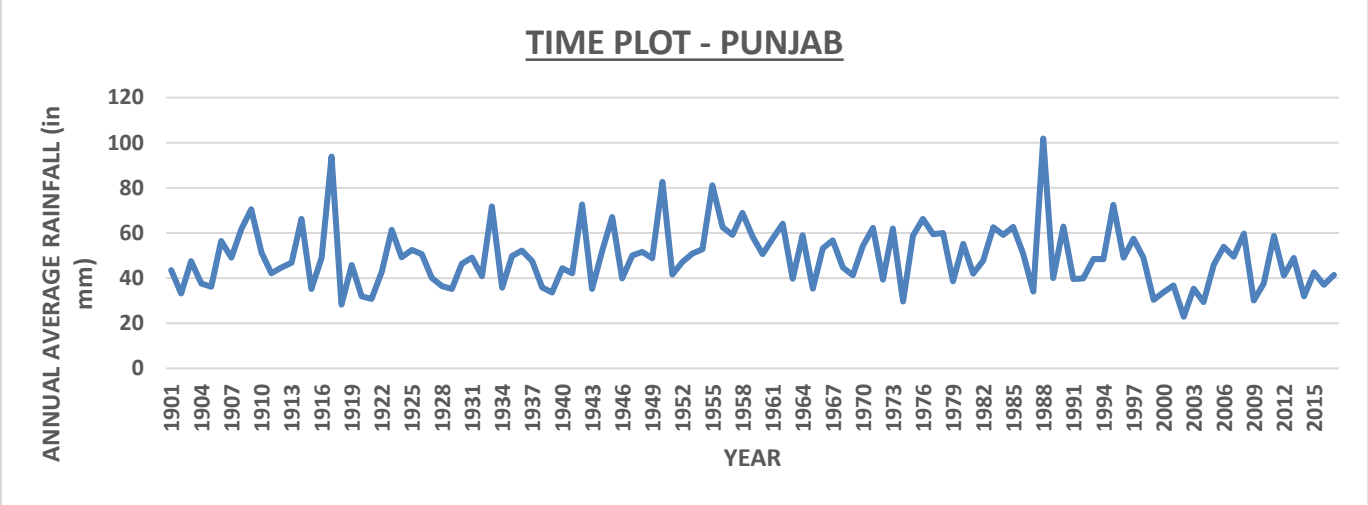
The state of Punjab shows a quadratic trend which has been decreasing in the recent years. The seasonal indices, confirm a concentrated rainfall period from June to September, with a dry spell in winter months specific to the country.

SOURCE OF DATA AND DATA DESCRIPTION

Source of Data - https://data.gov.in/resource/sub-divisional-monthly-rainfall-1901-2017
Data Description– The data covers the monthly rainfall (in millimeters, mms) for 36 regional subdivisions of India, from the year 1901 to 2017 (117 years).

TIMEPLOTS FOR THE DATA USED IN THE STUDY





ASSUMPTIONS FOR OUR STUDY

1) The factors discussed above heavily affect the seasonal variations in the Indian Average Rainfall. Hence, the seasonal component in the time series is highly pronounced.
2) The factors such as global warming, climate crisis, increasing human intervention in nature, etc. affect the rainfall phenomenon. Hence, a trend is present in the data.
3) The cyclic component of the data is not pronounced. (On the basis of the attached time plots).
4) A multiplicative model is a better fit for our data, i.e., $assumed\ model, Y_t = T_t * S_t * R_t$, where

$Y_t - observation\ at\ time\ t,$
$T_t - Trend\ value\ at\ time\ t,$
$S_t - seasonal\ component\ at\ time\ t,$
$R_t - random\ component\ influencing\ the\ observation\ at\ time\ t$

DECOMPOSITION OF TIME SERIES

A. VARIATE DIFFERENCE METHOD

1. KERALA

n	117	116	115
k	0	1	2
μ_2'	60253.21	1978.724	5576.735
σ_k^2	60253.21	989.3622	929.4559
R_k	10.51949	1.33611	

Since the absolute R_k value for k=1 is less than 1.96, therefore, a **polynomial of degree 0** will be a suitable curve for our data. However, it is inferred from the time plot that a polynomial of higher degree will be more suitable to represent trend in our data. The variance of the random component will be either 989.3622 or 929.4559.

2. PUNJAB

n	117	116	115
k	0	1	2
μ_2'	2611.583	369.825	1181.733791
σ_k^2	2611.583	184.913	196.9556318
R_k	9.937835	-1.43712	

Since the absolute R_k value for k=1 is less than 1.96, therefore, a **polynomial of degree 0** will be a suitable curve for our data. However, it is inferred from the time plot that a polynomial of higher degree will be more suitable to represent trend in our data. The variance of the random component will be either 184.913 or 196.956.

We further fitted mathematical models on our data to det more precise results.

B. TREND

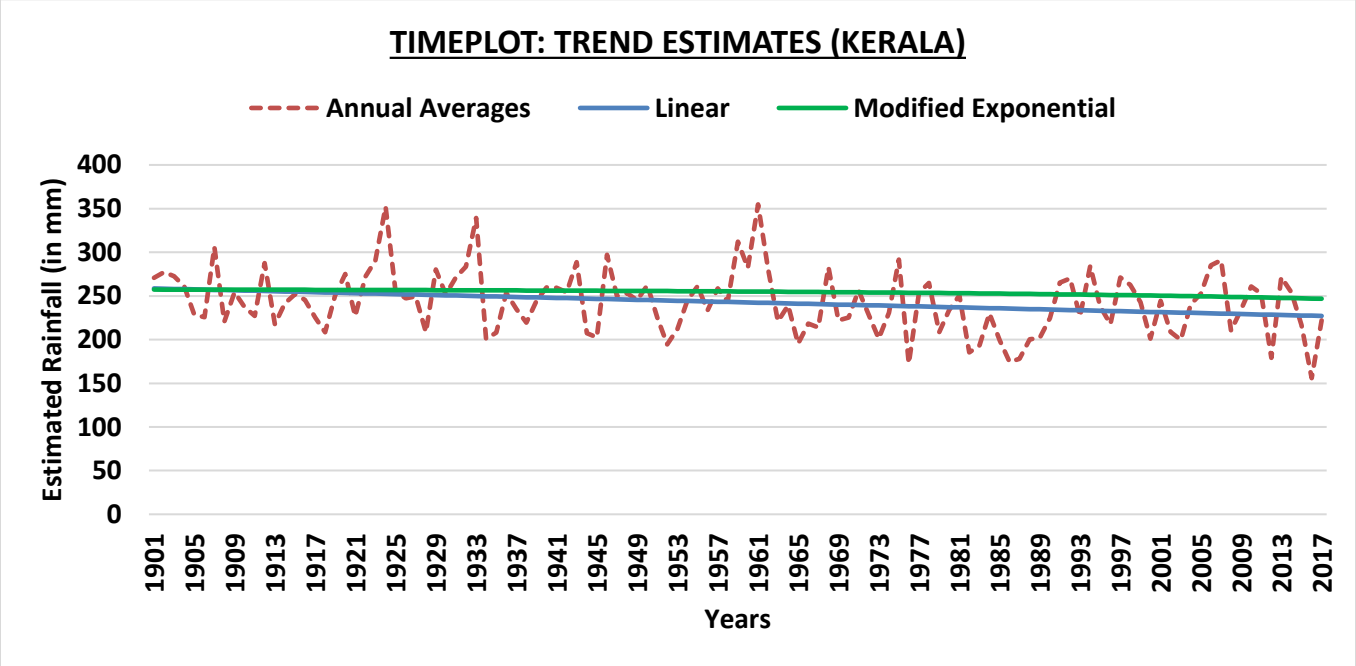
1. KERALA

The time plot shows the presence of a **linear trend**. 3 curves were mathematically-

$$Linear\ Trend\ Line: T_t = 242.8547 - 0.2718292 * t$$

$$Quadratic\ Trend\ Line: T_t = 242.80446 - 0.2718292 * t + 4.405E - 05 * t^2$$

$$Modified\ Exponential\ Curve: T_t = 258.15541 - 3.0549405 * 1.022879t^2$$



(Quadratic trend line has not been plotted as it was coinciding with the linear trend line.)

	MSE	RMSE	R^2	MAE
Linear	1211.220873	34.8026	0.980241	27.30566
Quadratic	1211.218819	34.80257	0.980241	27.30442
Modified Exponential Curve	1384.324683	37.20651	0.977418	29.28828

As values of MSE and MAE are least and R^2 maximum for the Quadratic Curve, hence, it can be considered as the best fit for the data. However, the coefficient of the square term is almost negligible in the quadratic equation, and the MSE, MAE and R^2 are approximately the same as that of the Linear Trend Line. Hence, **linear trend** is the best fit for the data.

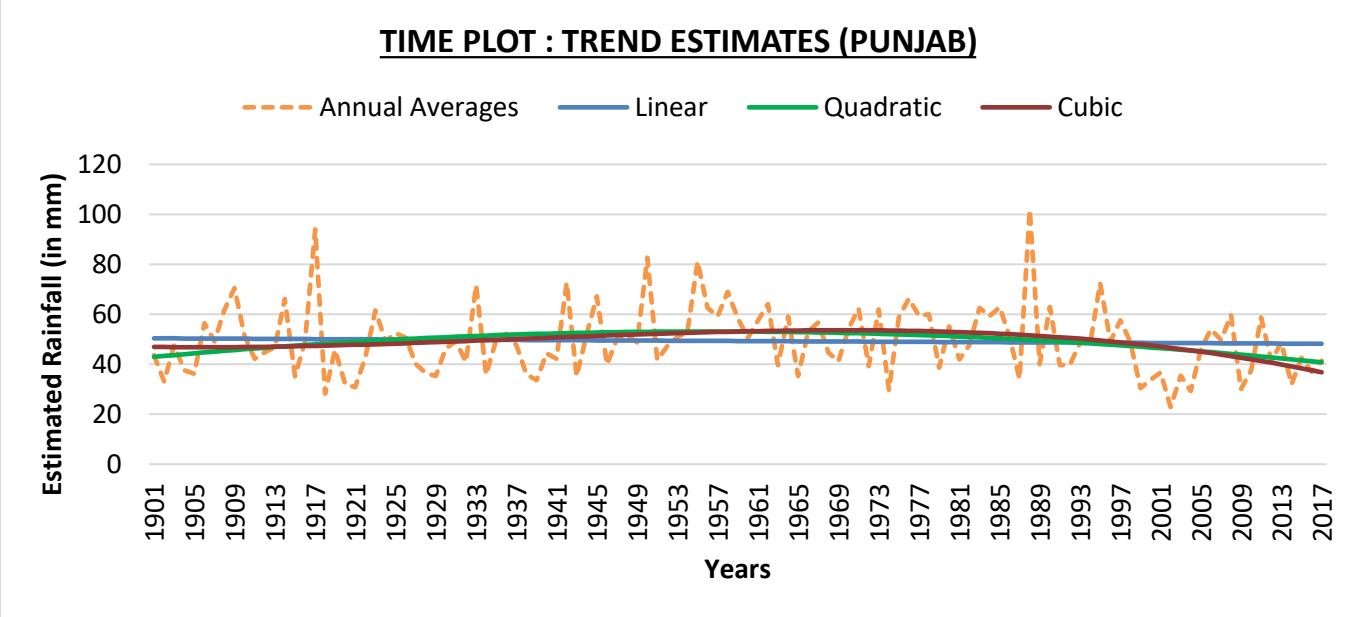
2. **PUNJAB**

The time plot of Punjab shows the presence of trend. So, we fit linear, quadratic and cubic models to the given data to check the best fit for the given data. The estimated trend lines are as follows:

LINEAR TREND LINE: $T_t=49.2868 - 0.01869*t$

QUADRATIC TREND LINE: $T_t=53.0981 - 0.01869*t - 0.0033*t^2$

CUBIC TREND LINE: $T_t=53.0981 - 0.08903*t - 0.0033*t^2 - 0.00005*t^3$



	MSE	RMSE	R^2	MAE
Linear	181.9932884	13.49049	0.930313	10.406
Quadratic	170.3751785	13.05278	0.934762	9.886168
Cubic	167.8553619	12.9559	0.935727	9.720769

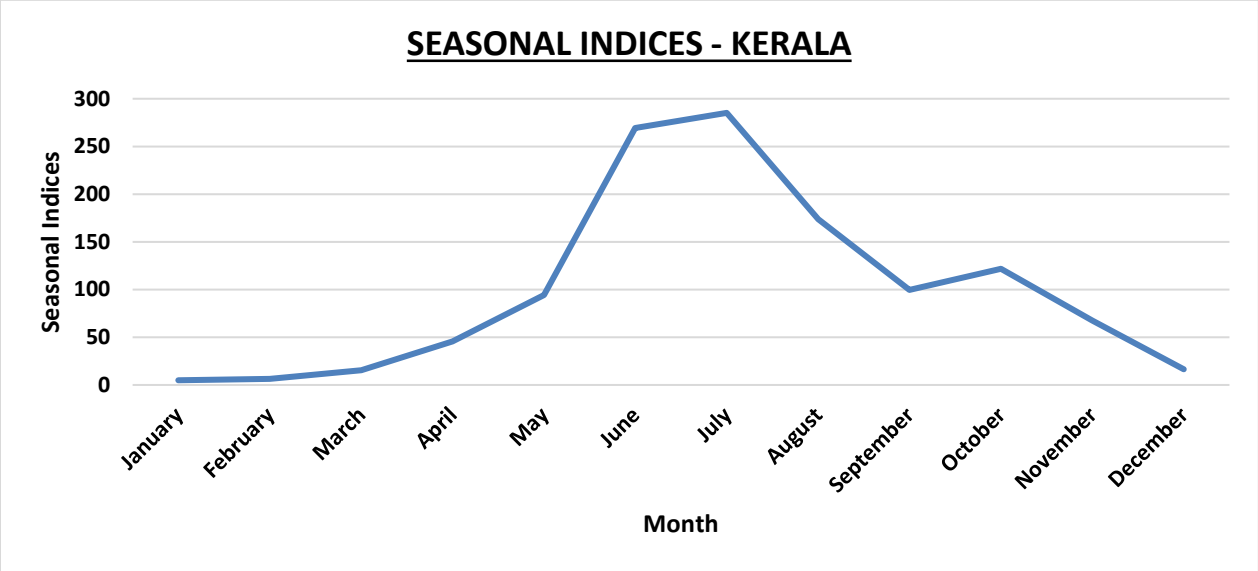
As values of MSE and MAE are least and R^2 maximum for the Cubic Curve, hence, it can be considered to be the best fit for the data. However, with a small coefficient for the cubic term in our cubic model, it suggests a very small effect. Therefore, in this case, the cubic term might not contribute significantly to the model's explanatory power.

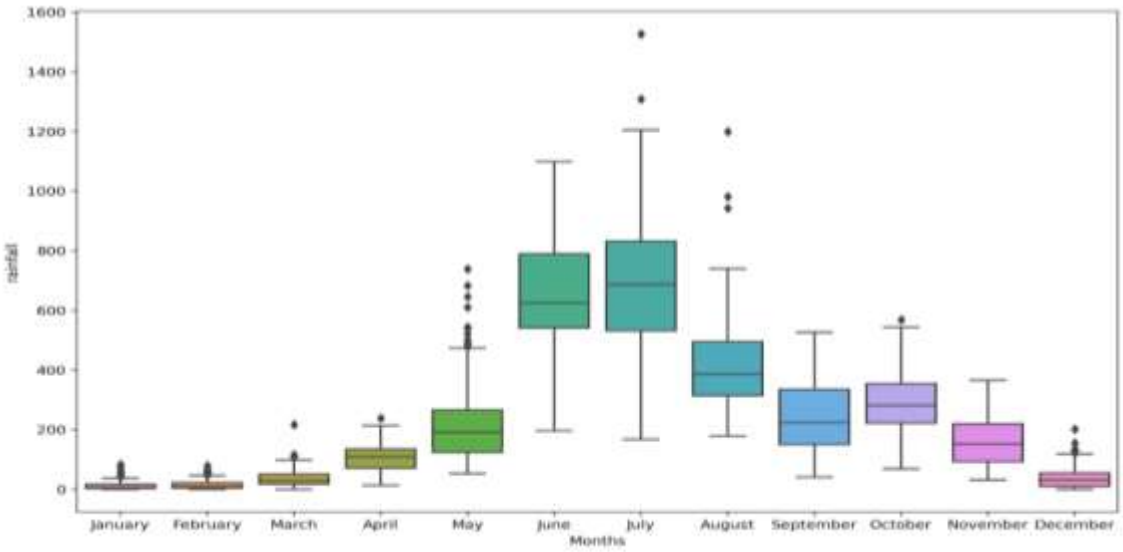
So, we might consider favoring the quadratic model, as the cubic term's impact appears negligible and the MSE, MAE and R2 are approximately the same as that of the Quadratic Trend Line.
Hence, **quadratic trend** is the best fit for the data.

C. SEASONAL COMPONENT

1. KERALA

We have used the ratio-to-moving average method for estimating the seasonal component of Kerala's rainfall data. The seasonal indices have been plotted below from which we can observe that rainfall occurs mainly in **2 periods**, first being from **June-September** out of which June and July receive the highest rainfall (also known as Southwest monsoon) and the second being **October-November** (also known as Northeast monsoon or Retreating monsoon). We can observe a similar pattern in rainfall from the boxplot as well.

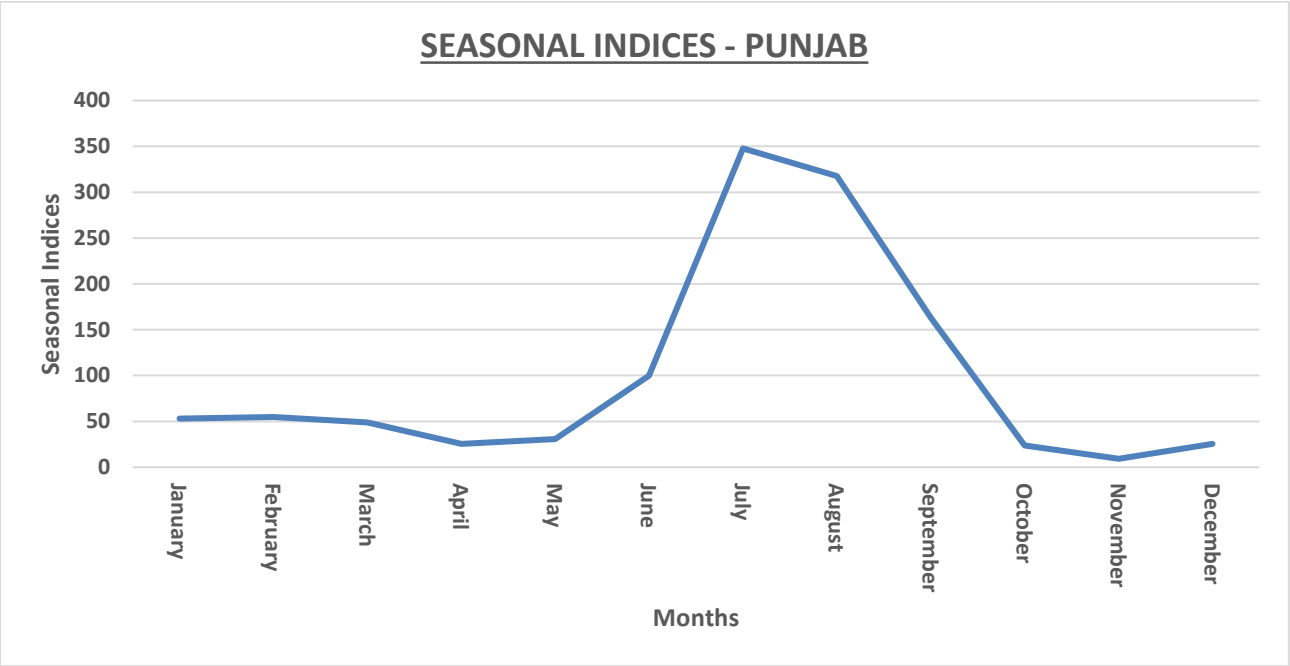




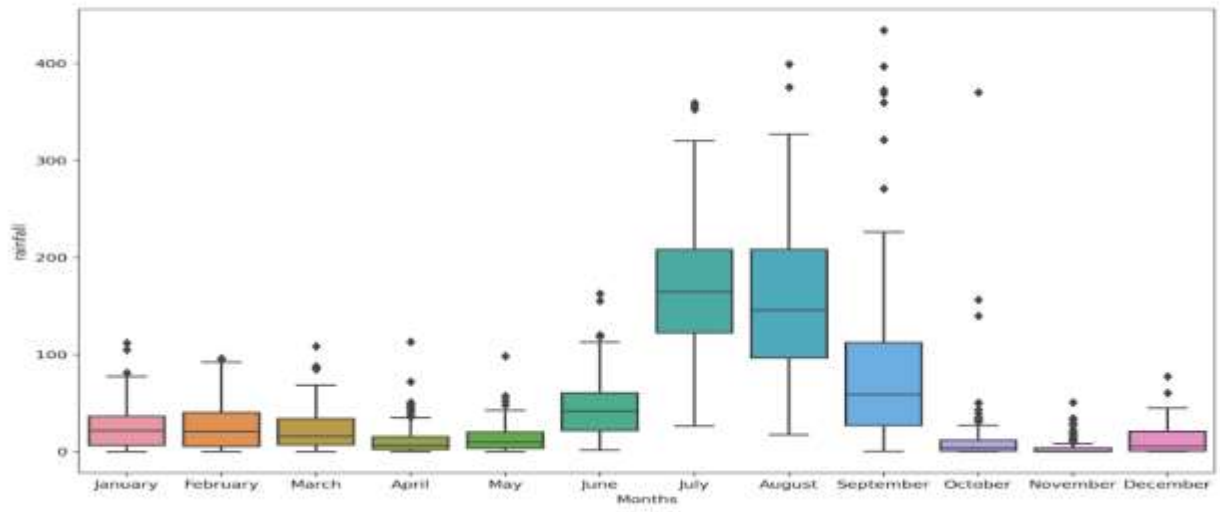
From research, the southwest monsoon is the principal rainy season when the State receives about 70% of its annual rainfall. The rainfall amount in the State decreases towards the south with the decrease of height of Western Ghats. The southernmost district of Thiruvananthapuram where Western Ghats are nearest to the sea coast and its average height is also least in the State receives minimum amount of rainfall.

2. **PUNJAB**

Applying the ratio-to-moving-average method of extent 12, to Punjab's highly seasonal rainfall data, the moving averages consisted of the trend component mainly, assuming minimal cyclic influence. The resulting seasonal indices, showcased below in the chart, confirm a **concentrated rainfall period from June to September**. Notably, the time plot illustrates peak rainfall in July and August.



This streamlined approach underlines the data's heavy seasonality, offering a clearer perspective on the distribution of rainfall. Pinpointing peak months holds practical implications for regional planning, emphasizing the need for targeted strategies during heightened precipitation. From our research, in Punjab, the Kharif season usually starts with the onset of the monsoon, which is around July. Farmers aim to take advantage of the ample rainfall during this season to ensure proper growth and development of these crops. It can be verified from the computations and charts above that July serves as an ideal month for such crops.



LIMITATIONS

1. The boxplots show presence of outliers in the data. Further analysis is required to treat them.
2. Cyclic component has been assumed to be negligible, when in actuality it can be influencing our data to a significant extent.
3. Variate difference method indicates the degree of the polynomial to be 0 in both the case, however, we have used polynomials of higher orders for fitting our data.
4. Due to usage of ratio to moving average method for computation of the seasonal indices, we did not obtain any seasonal index for the first six and the last six observations of the data.
5. A mixed model can be a better fit to our data rather than a purely multiplicative or a purely additive model.
6. Presence of sampling error- Rain gauges are used to measure rain and they are usually placed at places where eddies of air will not interfere with the normal fall of the raindrops are used for measuring the levels of rainfall. Rain gauge gives relatively accurate point measurement of rainfall but observations are not available over most remote land areas and over oceanic areas. Land rain gauge observations gives sampling error if the network is not adequately dense.

SOURCES

https://kerenvis.nic.in/Database/CLIMATE_829.aspx
https://www.agrimetassociation.org/journal/fullpage/fullpage-202001291249227994.pdf
https://www.mospi.gov.in/sites/default/files/Statistical_year_book_india_chapters/Rainfall.pdf
https://ncert.nic.in/textbook/pdf/iess104.pdf
https://www.mapsofindia.com/punjab/geography-and-history/climate.html
https://india.mongabay.com/2022/04/explainer-what-factors-affect-the-indian-summer-monsoon/