UNRAVELING CHENNAI'S RAINFALL PATTERNS USING CLUSTERING AND FORECASTING TECHNIQUES

M.Sc. Project

Submitted by

Joel Jossy

P222605

Guided by

Dr. M.Manoprabha

Department of Statistics and Applied Mathematics
Central University of Tamil Nadu
Thiruvarur - 610005, INDIA

Month & Year of Submission: May, 2024

CENTRAL UNIVERSITY OF TAMIL NADU
DEPARTMENT OF STATISTICS AND APPLIED MATHEMATICS

CERTIFICATE

This is to certify that the project entitled **"UNRAVELING CHENNAI'S RAINFALL PATTERNS USING CLUSTERING AND FORECASTING TECHNIQUES"** is the bonafide project work carried out by **Joel Jossy**, Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvarur 610 005 during the academic year 2023-2024, in partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics and Applied Mathematics by the Central University of Tamil Nadu.

No part of this report has been submitted elsewhere for the award of any other degree.

(Dr.M.Manoprabha)
Signature of Guide

Date:

Place: Thiruvarur

DECLARATION

I, Joel Jossy of II year M.Sc Statistics and Applied Mathematics, Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvarur, hereby declare that the project work entitled **"UNRAVELING CHENNAI'S RAINFALL PATTERNS USING CLUSTERING AND FORECASTING TECHNIQUES"** submitted to the Department of Statistics and Applied Mathematics, Central University of Tamil Nadu during the academic year 2023-24 under the guidance of **Dr.M.Manoprabha**, Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvarur, is a bonafide work done by me. This project work is submitted in partial fulfilment of the requirements for the award of the degree of **"Master of Science"** in Statistics and Applied Mathematics. I further declare that the results of this work have not been submitted for any other degree.

Name of Student: Joel Jossy

Signature of student:

Reg. No :P222605

Date:

Place: Thiruvarur

ACKNOWLEDGEMENT

I'd want to convey my heartfelt thanks to **Dr.M.Manoprabha**, whose advice, support, and encouragement have been vital throughout my academic career. My way of thinking has been greatly influenced by Sir's experience, wisdom, and commitment, which have also improved my knowledge and abilities. Sir has taught me so much, and his enthusiasm has inspired me. I am very grateful to Sir for his helpful criticism, patience, and eagerness to go the extra mile in order to assist me reach my academic objectives. I am incredibly appreciative of Sir's guidance, and I will always treasure the knowledge and experiences I have acquired while working under his guidance.

I would also like to express my gratitude towards **Dr Deepak M Sakate**, Head of the Department and all other Faculty Members of the Department for giving me this opportunity to do a project. Without their support and suggestions, this project would not have been completed.

Abstract

This paper delves into the analysis of monthly rainfall data for Chennai, employing a multifaceted approach encompassing clustering techniques and forecasting methodologies. Initially, the monthly rainfall data is meticulously visualized to comprehend its underlying patterns and variability. Subsequently, employing the K-means clustering technique, the data is segmented into distinct clusters, facilitating a comprehensive examination of each cluster's properties, including percentage composition of each cluster to the dataset as a whole , measures of central tendency and measures of variability.

Moreover, the distribution of data points within each cluster is studied, shedding light on the underlying structure of the rainfall patterns. The dataset is then partitioned into two subsets, enabling the application of various forecasting methods to predict rainfall for the time period of the second set using the data from the first set. Forecasting techniques such as SARIMA, STL decomposition, and seasonal naïve forecasting are deployed, with performance evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE).

Notably, STL decomposition emerges as the optimal forecasting method, exhibiting superior performance owing to its ability to capture underlying trends, seasonal patterns, and irregular components. Leveraging this finding, STL decomposition is utilized to forecast rainfall for the entire dataset. The forecasted values are integrated with the original data, and K-means clustering is reapplied to ascertain whether the distribution of data in the new clusters aligns with the previous clustering results.

Remarkably, the analysis reveals a remarkable similarity in the distribution of data across the new clusters, indicating the robustness of the clustering methodology and the stability of rainfall patterns. This not only contributes to a deeper understanding of rainfall variability in Chennai but also underscores the efficacy of clustering and forecasting techniques in analyzing and predicting complex environmental phenomena.

Keywords: K-means clustering, SARIMA, STL decomposition, Seasonal naïve forecasting, Performance metrics

.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Extreme rainfall events have become increasingly prevalent in the wake of global climate change, and urban flooding disasters are increasingly a routine as opposed to an uncommon one. In climate research and management of resources, rainfall pattern recognition and forecasting are critical, particularly in regions like Chennai which are vulnerable to unpredictable weather. Chennai, which is located on India's southeast coast, has a tropical wet and dry climate. The monsoon season brings with it a great deal of rain, which is essential for agriculture, maintaining water resources, and urban development. However, effective resource allocation and disaster preparedness are severely hindered by the variability of the rainfall in the area in question. Chennai, the capital city of Tamil Nadu, India, has experienced its share of these fluctuations, witnessing both periods of drought and devastating floods in recent memory. Understanding the underlying patterns in rainfall and accurately forecasting future extreme climate events in Chennai is paramount for effective risk mitigation and adaptation strategies. Uncertainty in flood forecasting arises from various aspects such as rainfall, model structure, model parameters, etc. Therefore, improving the accuracy of flood forecasting from these aspects has become an essential part of flood prevention and mitigation research.[8]

Understanding historical rainfall patterns and developing accurate forecasting models are essential for mitigating the impact of extreme weather events and ensuring sustainable development. The advent of advanced data analytics and machine learning techniques offers unprecedented opportunities to enhance the accuracy and reliability of rainfall forecasting models. Constructing a forecast system for accurate rainfall is a challenging issue for researchers. The common question here is how to analyse the past and predict the future. One such solution is Time Series Modelling, involves functioning on time-based data to derive concealed insights for informed decision making.[4]

This master's thesis aims to explore and enhance pattern recognition methodologies for future rainfall forecasting in Chennai. By leveraging monthly rainfall data for little over a century

using clustering and forecasting methods to identify underlying patterns and trends. The thesis will focus on developing robust forecasting models capable of accurately predicting rainfall patterns on a monthly basis, with a particular emphasis on long-term forecasting. By analysing historical monthly rainfall data, spanning several decades, this research seeks to identify trends, cycles, and anomalies that characterize Chennai's rainfall regime. Furthermore, by using time series methods, this study endeavours to develop robust predictive models capable of forecasting future rainfall patterns and extreme climate events with enhanced accuracy.

Figure 1.1: Geographical area of the study

This thesis seeks to bridge the gap between theoretical research and practical applications in the field of rainfall forecasting, with a specific focus on Chennai. By leveraging advanced data analytics and temporal analysis, the thesis aims to enhance our understanding of rainfall patterns and improve the accuracy of future forecasts, thereby contributing to the resilience and sustainability of Chennai's infrastructure and economy. The insights gained from this research can inform policymakers, urban planners, and agricultural stakeholders in making informed decisions, thereby enhancing resilience to climate variability and facilitating sustainable development in the region. Understanding the spatiotemporal dynamics of rainfall patterns enables proactive decision-making, such as optimizing water resource management strategies, designing resilient infrastructure, and formulating effective disaster preparedness plans tailored to Chennai's unique climatic challenges.

By employing clustering algorithms, we seek to identify distinct rainfall regimes within the city, which can aid in localized planning and risk mitigation strategies. In the context of rainfall analysis for Chennai, these algorithms offer a powerful tool to delineate distinct rainfall regimes within the city. By grouping together similar patterns of rainfall behaviour over time, clustering enables the identification of localized zones or regions with homogeneous rainfall characteristics.

This integrated approach provides a comprehensive understanding of long-term rainfall trends and patterns, offering valuable insights for stakeholders and policymakers. Through this analysis, we aim to contribute to the field of meteorology and climate science by leveraging data-driven techniques to enhance our understanding of rainfall variability and improve predictive capabilities for better decision-making in various sectors. Overall, these studies highlight the importance of accurate rainfall data and the need for innovative approaches in flood forecasting to account for changing environmental conditions.

1.1. Objectives of the study

- Pattern classification for rainfall data using K-means clustering techniques.
- To study the properties of each cluster by using their measures of central tendency and variability.
- Obtain the best forecasting model for highly seasonal data
- Study changes in properties of cluster with the elapse of time

CHAPTER-2

METHODOLOGY

This chapter discusses the clustering techniques and time series forecasting models used in this thesis. It also concerns the selection of the best forecasting model and the metrics of MAE, RMSE and MASE used for such a selection in great mathematical detail. The clustering technique used is K-means clustering algorithm and the time series forecasting methods are ARIMA, STL decomposition and seasonal naïve forecasting.

2.1.K-means clustering algorithm

This is a popular unsupervised machine learning technique used for partitioning a dataset into k distinct, non-overlapping clusters.

A data set $\boldsymbol{X}$= $\{\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_n}\}$ is a set of n data points in a d-dimensional space. To divide **X** into **k** clusters the following algorithm is used

Step 1: Randomly initialize k cluster centroids $\mathbf{C}= \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots\ldots, \mathbf{C_n}\}$

Step 2: For each data point $\mathbf{x_i}$,calculate the distance to each centeroid C_j using the formula,

$$\mathbf{d}(\boldsymbol{x_i}, \boldsymbol{C_j}) = \sqrt{\sum_{\boldsymbol{l=1}}^{\boldsymbol{d}} (\boldsymbol{x_{il}} - \boldsymbol{c_{jl}})^{\mathbf{2}}} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.1})$$

Select the clusters $\mathbf{S_k}$ such that,

$$\mathbf{S_k} = \{\mathbf{d}(\mathbf{x_i}, \mathbf{C_k}) \leq \mathbf{d}(\mathbf{x_i}, \mathbf{C_l}) \forall \mathbf{l} \neq \mathbf{k}\} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.2})$$

Step 3: For each cluster S_j,for some j, update the centroid as the mean of the data points in that cluster.

$$\boldsymbol{C_j} = \frac{\mathbf{1}}{\boldsymbol{N_j}} \sum_{\boldsymbol{x_i} \in \boldsymbol{S_j}} \boldsymbol{x_i} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.3})$$

, where N_j is the number of elements in cluster S_j.

Step 4: Repeat step 2 and 3 until convergence. Convergence is reached when the cluster centres stop changing significantly or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

The algorithm converges to a local minimum of the within-cluster variance. Since K-means can be sensitive to the initial random centres, it's common to run the algorithm multiple times with different initializations and choose the clustering with the lowest within-cluster variance.

The main optimization objective of K-means can be expressed as:

$$\mathbf{W(C_k)} = \sum_{\mathbf{x_i \in S_j}} \mathbf{\left(x_i - C_j\right)^2} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.4})$$

The idea that this intra cluster variation should be minimised is the main objective of K-means clustering technique. The total within cluster variation is defined as

Total within cluster variation

$$= \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{W(C_k)} = \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\mathbf{x_i \in S_j}} \mathbf{\left(x_i - C_j\right)^2} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.5})$$

The total within-cluster sum of square measures the compactness (i.e goodness) of the clustering and we want it to be as small as possible.

2.2. Measures of variability

The following measures of standard deviation, variance, Inter Quantile Range (IQR), kurtosis and skewness are used to find the variability of each cluster S_j,with center C_j and N_j is the number of elements in that given cluster

2.2.1. Standard Deviation

For a given cluster S_j,with center C_j the standard deviation shows the spread of the data points in the cluster within it around the centroid. Higher values of standard deviation indicate spread out points while the lower values show that the data points are close to the centroid. The standard deviation is given by,

$$\sigma_j = \sqrt{\frac{1}{N_j} \sum_{x_i \in S_j} \|x_i - C_j\|^2} \quad \ldots\ldots\ldots\ldots\ldots (Eqn\ 2.6)$$

Where,

$\| \|$ is the Euclidian distance and N_j is the number of elements in the cluster.

2.2.2. Variance

Variance measures the average of squared deviations from mean. It gives the measure of the overall data points within the cluster. The variance is given by,

$$\sigma_j^2 = \frac{1}{N_j} \sum_{x_i \in S_j} \|x_i - C_j\|^2 \ldots\ldots\ldots\ldots\ldots (Eqn\ 2.7)$$

Where,

$\| \|$ is the Euclidian distance and N_j is the number of elements in the cluster

2.2.3. Interquartile range (IQR)

IQR represents the range between the first quartile (25th percentile) and the third quartile (75th percentile) of the data. It gives a measure of the spread of the middle 50% of the data, which can be more robust to outliers compared to the standard deviation. A larger IQR indicates a wider spread of the central data values, while a smaller IQR indicates a tighter clustering of the central data values. The IQR for cluster S_j,with center C_j is gives by.

IQR_j =first quartile of the cluster-third quartile of the cluster. $\ldots\ldots\ldots\ldots\ldots (Eqn\ 2.8)$

2.2.4. Skewness

Skewness measures the asymmetry of the distribution of the data within the given cluster, say S_j. It indicates whether the data points are concentrated more on one side of the mean compared to the other. There are several methods to calculate skewness, the approach used here is Pearson's skewness coefficient which is given by,

$$SKEW_j = \frac{3(mean_j - median_j)}{Standard\ Deviation_j} \ldots\ldots\ldots\ldots\ldots (Eqn\ 2.9)$$

Skewness can be positive, negative, or zero. Positive skewness indicates that the data is skewed to the right, with a tail extending towards the higher values. Negative skewness indicates that the data is skewed to the left, with a tail extending towards the lower values. Zero skewness indicates a symmetric distribution. Skewness provides insights into the shape of the distribution of data within the cluster, helping to understand the overall pattern and behaviours of the data points.

2.2.5. Kurtosis

Kurtosis measures the peakdness or flatdness of the distribution of data points within the cluster S_j.It can be calculated using various formulas, with one common approach being the Pearson's kurtosis coefficient.

$$\boldsymbol{KURT_j = \frac{\sum_{i=1}^{N_j}(x_i - mean_j)^4}{N_{j-1} * (Standard\ Deviation)^4}} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.10})$$

Kurtosis values can be positive or negative. Positive kurtosis (leptokurtic) indicates that the distribution has a sharper peak and heavier tails compared to a normal distribution. Negative kurtosis (platykurtic) indicates that the distribution is flatter and has lighter tails compared to a normal distribution. A kurtosis value of zero (mesokurtic) indicates that the distribution has similar peakedness and tail behaviour as a normal distribution. Kurtosis provides insights into the shape of the distribution of data within the cluster, helping to understand the degree of outliers and the tails' behaviour.

2.3. SARIMA

Seasonal Auto-Regressive Integrated Moving Average (SARIMA) is an extension of ARIMA model that incorporates seasonality in addition to non-seasonal components. ARIMA model is represented as

ARIMA(p,d,q)(P,D,Q)[m]

Where the first bracket represents the non-seasonal components and the second bracket represents the seasonal components and ‘m’ is the number of observations per year or the period of the model.

p= Number of Auto-regressive terms of the non-seasonal component

d=Number of the differencing of raw observations to allow the time series to become stationary

q=Number of Moving Average terms of the non-seasonal component

P=Number of seasonal AR terms. This component captures the relationship between the current value of the series and its past values, specifically at seasonal lags.

D=Number of seasonal differences. Similar to the non-seasonal differencing, this component accounts for the differencing required to remove seasonality from the series.

Q=Number of Seasonal Moving Average terms. This component models the dependency between the current value and the residual errors of the previous predictions at seasonal lags.

For example, the ARIMA (0,0,0) (2,0,0) [12] can be expressed mathematically as,

$$\mathbf{Y_t = c + \Phi_1 Y_{t-12} + \Phi_2 Y_{t-24} + \in} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.11})$$

Where,Y_t is the time series data at time t

Intercept is represented by c.

The auto regressive parameters at lags 12 and 24 are represented by Φ_1 and Φ_2 respectively.

2.4.STL decomposition

Seasonal-Trend decomposition using LOESS (Locally weighted regression and scatterplot smoothing) is a method used to decompose the time series into 3 components seasonal, trend and remainder components.

$$\boldsymbol{Y_t = S_t + T_t + R_t} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.12})$$

Here the trend component T_t is calculated using the formula,

$$\mathbf{T_t = \boldsymbol{LOESS}(Y_t)} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.13})$$

LOESS stands for locally weighted regression and scatterplot smoothing.

Now,

$$\boldsymbol{Y_t - T_t = S_t + R_t} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.13})$$

Use moving average method to the RHS of the above equation ie, $S_t + R_t$ to obtain the seasonal component of the time series

$$\boldsymbol{S_t = Moving\ Average(S_t + R_t)} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.14})$$

Use this seasonal component to obtain the remainder component.

$$\boldsymbol{R_t = S_t + R_t - S_t} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.15})$$

After having obtained trend, seasonal and remainder component forecast for each component using appropriate forecasting method. Then combine this forecast to obtain the required forecast,

$$\boldsymbol{\widehat{Y}_t = \widehat{T}_t + \widehat{S}_t + \widehat{R}_t} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.16})$$

2.5. Seasonal Naïve Forecasting

The idea behind seasonal naïve forecasting is to use the observation from the previous season as forecast for the corresponding season in the future. Mathematically this can be given as,

$$\boldsymbol{\widehat{Y}_{t+h} = Y_{t+h-k(m\lfloor \frac{h}{m} \rfloor)}} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.17})$$

Here k is the number of seasons ago and m is length of seasonal cycle (here since the data is monthly m=12).$\hat{Y}_{t+h}$ is the forecast for time period t+h.

2.6. Performance of forecasting models

In this paper, performance of the forecasting models is assessed using 3 metrics MAE, RMSE, MASE. Lower values of this metrics indicate better performance of the model.

2.6.1. Mean Absolute Error (MAE)

MAE measures the mean of the errors between actual value and forecasted value. It can be mathematically expressed as,

$$\boldsymbol{M.A.E} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\ \widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \boldsymbol{Y}_{\boldsymbol{i}}| \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.18})$$

Here, n is the number of observations, $\hat{Y}_i$ is the forecasted value and Y_i is the actual value.

2.6.2. Root Mean Squared Error (RMSE)

RMSE is similar to MAE but it penalises large errors more heavily because of the squaring of errors. It is expressed mathematically as,

$$\boldsymbol{R.M.S.E} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \boldsymbol{Y}_{\boldsymbol{i}})^{\mathbf{2}}} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.19})$$

Here, n is the number of observations, $\hat{Y}_i$ is the forecasted value and Y_i is the actual value.

2.6.3: Mean Absolute Scaled Error (MASE)

MASE is a normalized version of MAE that compares the performance of a forecasting model to that of a naive model. It's particularly useful when dealing with data that has seasonality. It is expressed mathematically as,

$$\boldsymbol{M.A.S.E} = \frac{\boldsymbol{M.A.E}}{\frac{\mathbf{1}}{\boldsymbol{n-m}}\sum_{\boldsymbol{i=m+1}}^{\boldsymbol{n}} |\boldsymbol{Y_i - Y_{i-m}}|} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.20})$$

Here, n is the number of observations, m is the seasonal period,Y_i is the actual value at i, Y_{i-m} is the actual value at(i-m) and M. A. E is mean absolute error.

2.7. Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) is a measure of the similarity between two data clustering. It accounts for chance agreement between the clusters. ARI returns a value between -1 and 1, where 1 indicates perfect agreement between the clustering, 0 indicates the clustering is no better than random, and negative values indicate disagreement worse than random.

Here's the formula for the Adjusted Rand Index:

$$\boldsymbol{ARI} = \frac{\boldsymbol{Adjusted\ Agreement - Expected\ Agreement}}{\boldsymbol{Maximum\ Possible\ Agreement - Expected\ Agreement}} \quad \ldots\ldots\ldots\ldots\ldots (\boldsymbol{Eqn\ 2.21})$$

Where:

Adjusted Agreement: The proportion of pairs of elements in the data that are in the same cluster in both the true and predicted clustering, adjusted for chance.

Expected Agreement: The expected proportion of pairs of elements that would be in the same cluster by chance.

Maximum Possible Agreement: The maximum possible value for Adjusted Agreement given the number of clusters and the distribution of elements.

CHAPTER-3

RESULTS AND DISCUSSION

Data Analysis and Interpretations

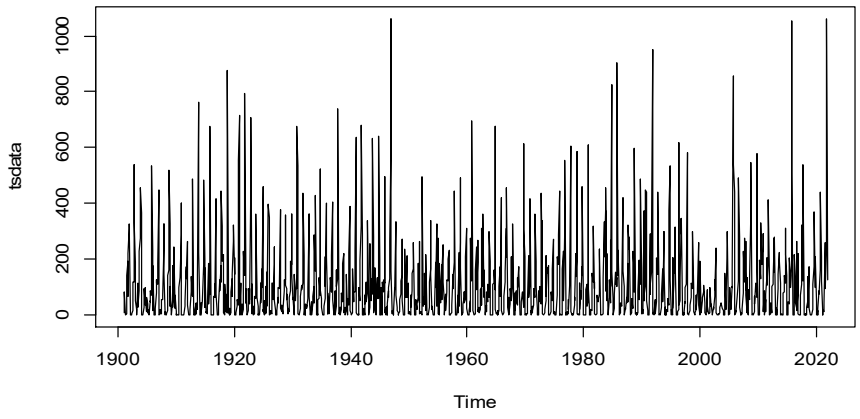
3.1. Data Source

Monthly rainfall data for Chennai city for the years 1901 to 2021 have been obtained from https://data.opencity.in. This extensive dataset offers a valuable resource for understanding the historical precipitation patterns in one of India's major urban centers. Over the course of more than a century, these records encapsulate the fluctuations and trends in rainfall that have impacted Chennai's environment.

3.2. Visualization of the data

The data from January 1901 to December 2021 is plotted in the figue 3.1.This figure depicts the fluctuation in rainfall for the city. This visualization provides valuable insights into the temporal variability of rainfall patterns, highlighting periods of abundance and scarcity.

Figure 3.1:Graph of Rainfall



3.3.K-means Cluster

The rainfall data is subjected to a K-means clustering algorithm, resulting in the classification of the dataset into seven distinct clusters. These clusters represent groups of data points with similar patterns or characteristics in terms of rainfall variability. Figure 3.2 visually presents these clusters, providing a graphical representation of how the data points are grouped together based on their respective rainfall patterns. These 7 clusters are then used to carry out investigation into the rainfall data in a very comprehensive way.

Figure 3.2:K-Means Clustering of Rainfall Data

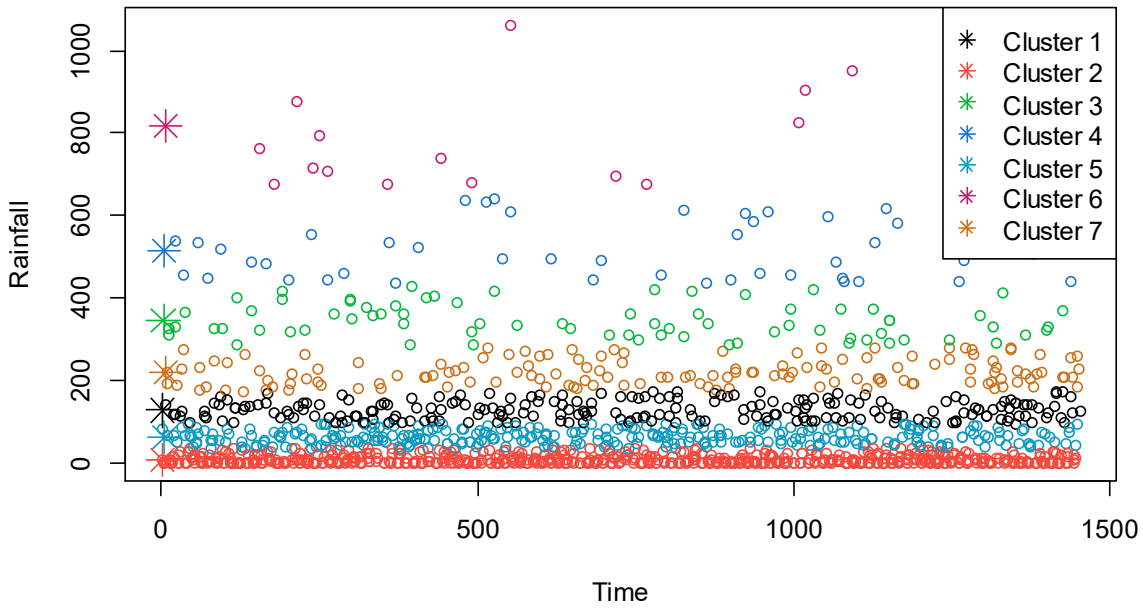


Figure 3.2 gives the graph of the clustering done using K-means clustering algorithm. This graph clearly shows that cluster 2 is associated with low rainfall and cluster 6 is associated high levels of rainfall. The graph of each cluster is given by Figures 3.3-3.9. These graphs are used to understand the temporal trends in each cluster and through it the entire data

Figure 3. 3 : Cluster 1

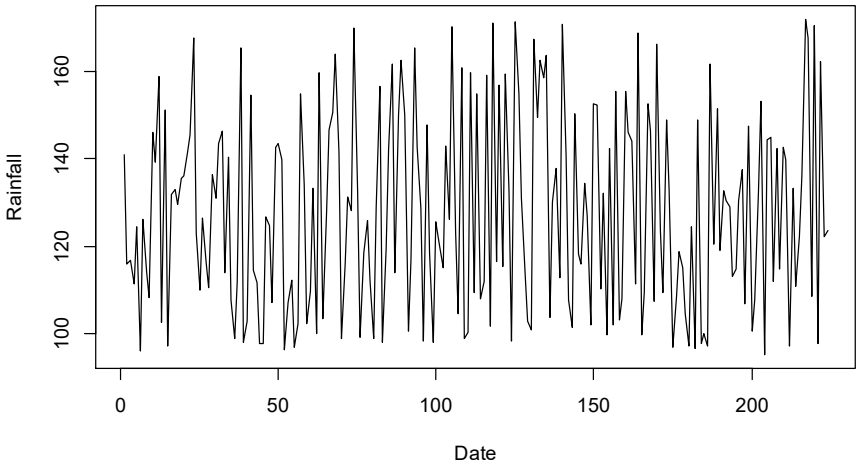


Figure 3. 4 : Cluster 2

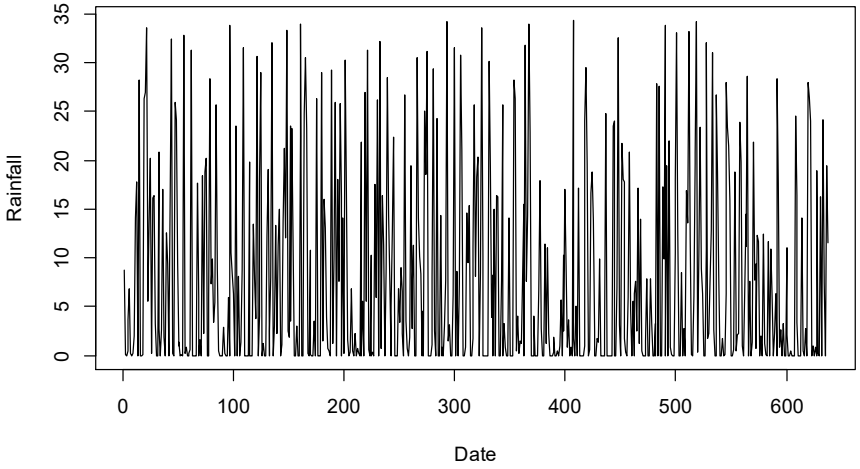


Figure 3. 5 : Cluster 3

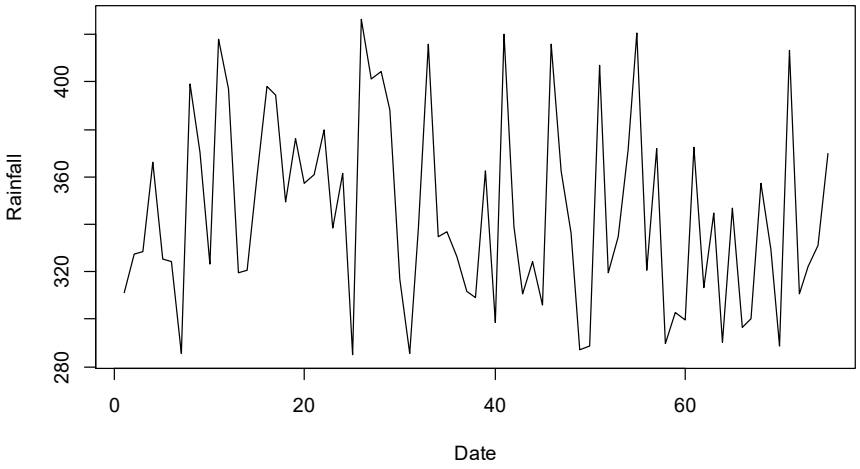


Figure 3. 6 : Cluster 4

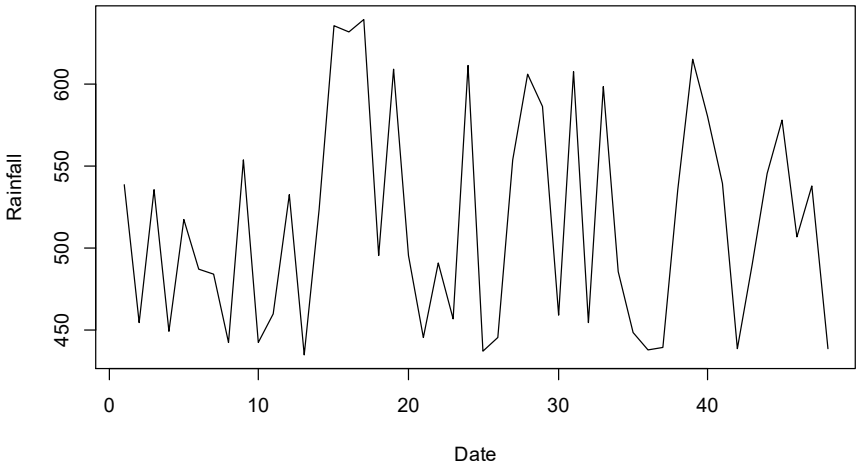


Figure 3. 7 : Cluster 5

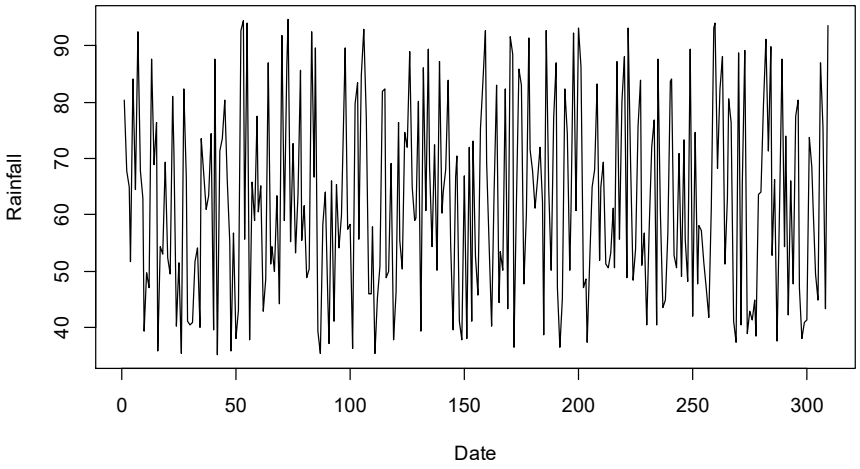


Figure 3. 8 : Cluster 6

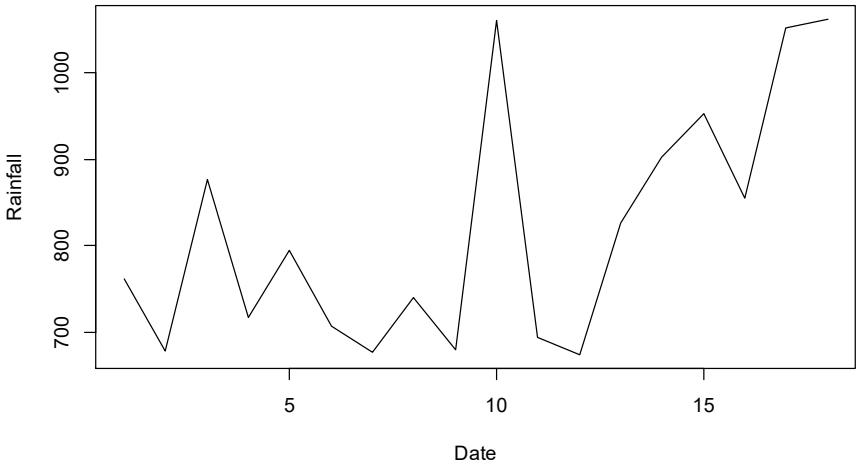
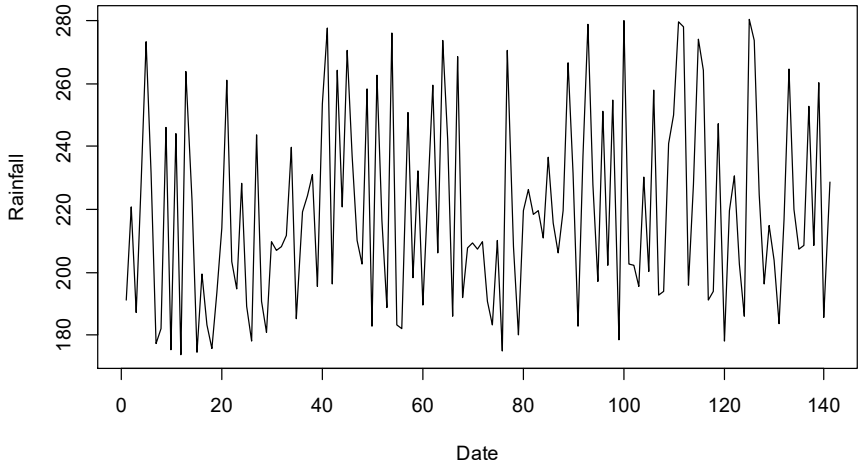


Figure 3. 9 : Cluster 7



In table 3.1, C1, C2…..., C6 corresponds to cluster 1, cluster 2…..., cluster 6. The inference drawn from Figure 3.2 is true since heavy rainfall is associated with C6 with a mean rainfall of 817.22 while C2 is associated with low rainfall with a mean rainfall of 7.64.

We can comment about the distribution of each cluster using the table. For example, Cluster 1 represents a moderate rainfall level, with values between 95.22 and 171.82. The rainfall distribution in this cluster is approximately symmetric, as indicated by the skewness and kurtosis values close to zero. The value of mode in this cluster is 95.22. The mean rainfall in this cluster is approximately 127.61, close to the median value of 126.04, indicating a relatively balanced distribution. Given a standard deviation of 22.21, the data points are relatively close to the mean, there is only a moderate level of variability in the rainfall.

Table 3.1. Characteristics of each cluster

Descriptive Statistics	C1	C2	C3	C4	C5	C6	C7
Data count	224	637	75	48	309	18	141
Data percent	15.43	43.87	5.17	3.31	21.28	1.24	9.71
Min	95.22	0	285.19	435.03	35.21	674.19	173.93
Max	171.82	34.31	426.35	639.64	94.69	1061.64	280.46
Mean	127.61	7.64	344.68	515.36	62.63	817.22	219.55
Median	126.04	2.21	336.48	501.13	61.12	778.08	214.22
Mode	95.22	0	285.19	435.03	35.21	674.19	173.93
Standard Deviation	22.21	9.97	40.53	65.63	17.05	139.35	30.67
Varience	493.25	99.42	1642.57	4307.47	290.71	19417.4	940.6
IQR	37.28	13.53	59.16	106.79	27.38	198.20	48.21
Skewness	0.3	1.18	0.40	0.39	0.22	0.66	0.45
Kurtosis	1.9	3.14	2.08	1.84	1.90	2.08	2.08

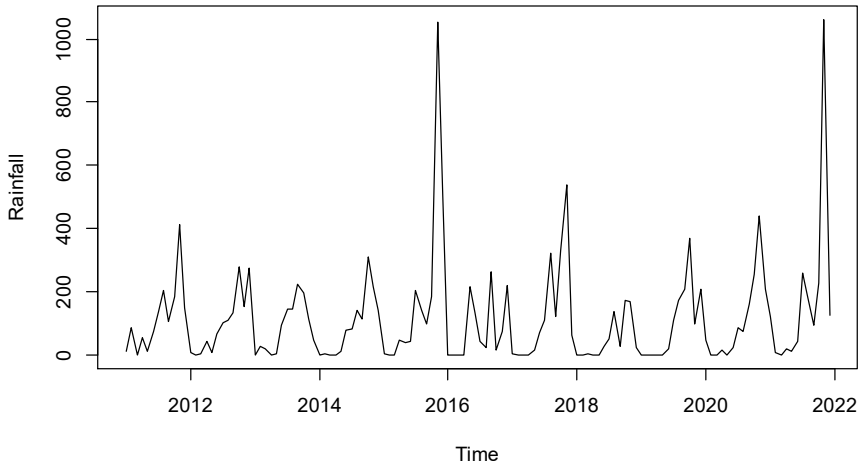
3.4. Forecasting the rainfall

ARIMA, STL Decomposition and Seasonal Naïve Forecasting methods are used to predict the rainfall. The best predictive model has been obtained by splitting the data in to two sets. The one set of data will contain the rainfall data from January 1901 to December 2010 while the second set will contain the rest of data from January 2011 to December 2021.We used the first set of data to forecast the rainfall during Jan 2011 – Dec 2021 using the three methods and find the actual vs forecasted value plot as well as the residual to find the metrics of Mean Absolute Error(MAE),Root Mean Squared Error(RMSE) and Mean Absolute Scaled Error(MASE).The method with least values of this metrics is the best method for this case.

3.4.1. Splitting of the data

The graph of the first partition is given by Figure 3.10 and that of the second partition is given by Figure 3.11. It is clear from figure 3.10 and 3.11 that both partitions have similar levels of variability. This implies that both partitions have requisite level of similarity to the original data for helping in the task in this paper of finding the better forecast model SARIMA,STL decomposition and seasonal naïve forecasting .

Figure 3.11:Graph of 2nd Partition



3.4.2. Forecasting for the first partition

SARIMA, STL decomposition and Seasonal naïve forecasting methods are used in the first partition to forecast for the time period of the second partition and use the difference between actual vs forecasted value to find the residuals of the model.

3.4.2.1. SARIMA

SARIMA (0,0,0) (2,0,0) [12] with non-zero mean is used to forecast for the first partition for the time period of the second partition. The graph of the forecast is plotted in figure 3.12

Figure 3.12:Graph of forecast using ARIMA

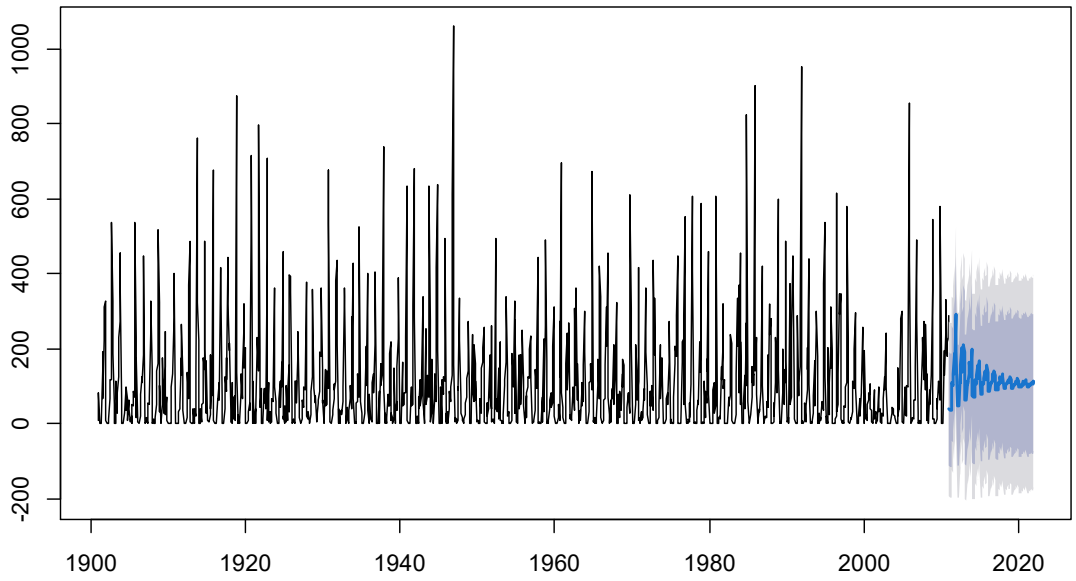


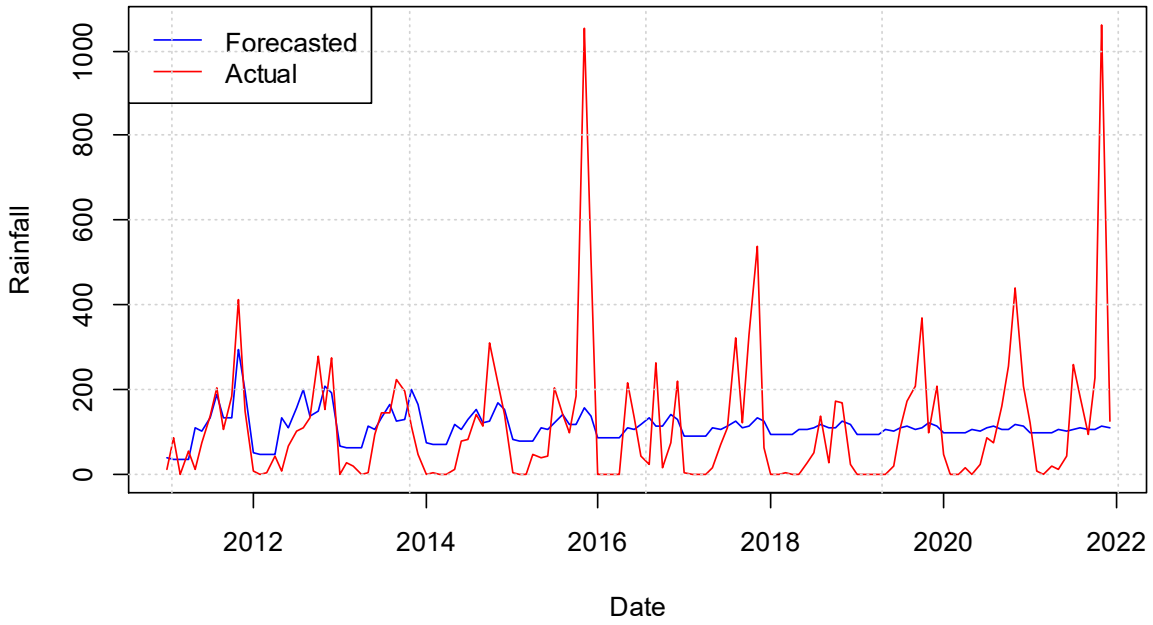
Figure 3.12 gives the plot of the forecast values obtained using ARIMA (0,0,0) (2,0,0) [12] in the shaded region. This forecast value is then compared with the values of the data at the same time point in the second partition and the difference between these actual values and forecasted values is taken as the residual at that time point, as shown in table 3.2.

Table 3.2: Sample of Forecast, Actual values & residuals using SARIMA

	Date	Forecasted	Actual	Residuals
1	2011-01-01	38.94756	11.6260862	-27.3214729
2	2011-02-01	35.01608	87.6357594	52.6196808
3	2011-03-01	35.57392	0.0000000	-35.5739248
4	2011-04-01	34.97325	54.2870996	19.3138509
5	2011-05-01	107.55448	10.9515980	-96.6028822
6	2011-06-01	101.44639	73.9905854	-27.4558005
7	2011-07-01	134.04769	137.4666191	3.4189313
8	2011-08-01	186.33735	202.7637950	16.426445
…….				
……..				
……..				
126	2021-06-01	103.03186	43.4090546	-59.6228016
127	2021-07-01	106.23923	260.4401859	154.2009536
128	2021-08-01	109.69055	185.4993285	75.8087834
129	2021-09-01	105.08696	93.6575608	-11.4293949
130	2021-10-01	105.69043	228.6255028	122.9350758
131	2021-11-01	112.39101	1061.6422220	949.2512091
132	2021-12-01	109.40223	123.6062256	14.2040003

The actual and forecasted values are plotted in Figure 3.13 against time to get a more general sense of the difference between them. This visualization allows for a comprehensive examination of how well the forecasted values align with the actual observed data over the entire time period of the 2nd partition.

Figure 3.13:Forecasted vs Actual Rainfall

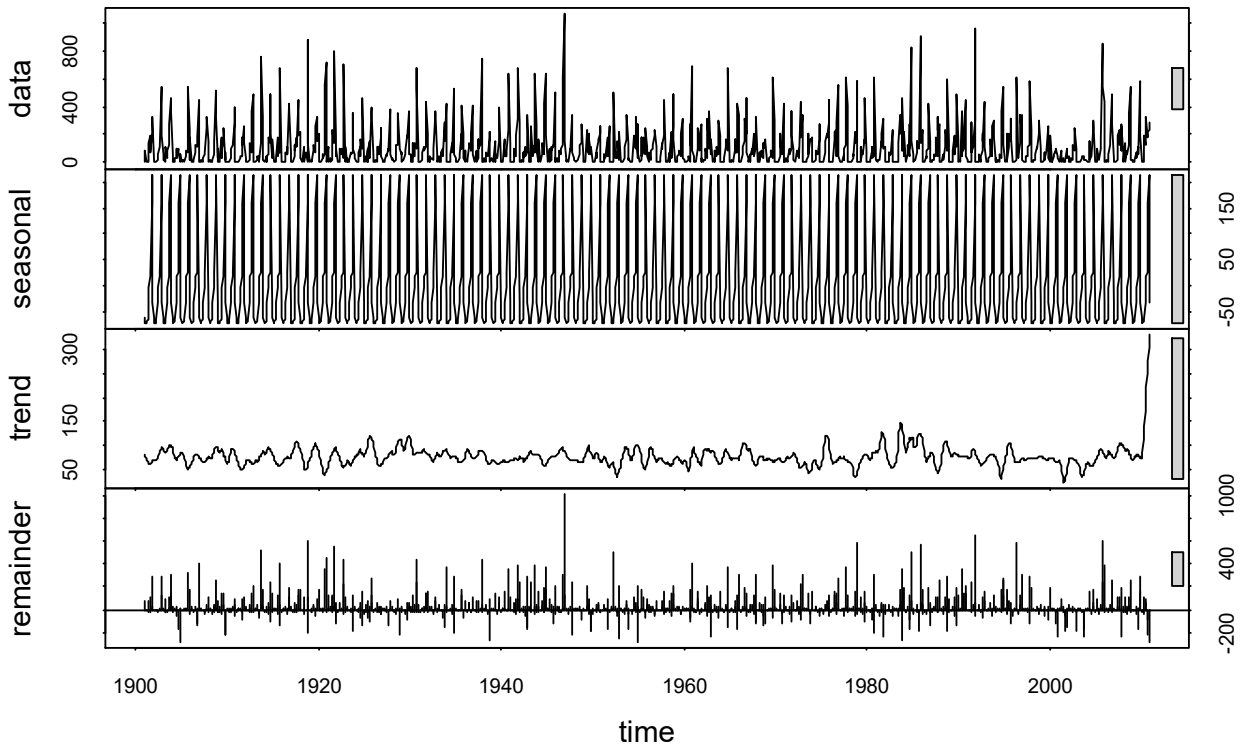


From figure 3.13, it is observed that while the forecast aligns with the actual data in the first few years it gradually has very big fluctuations from the actual values that is very large values of residuals.

3.4.2.2.STL Decomposition

The STL decomposition of the first partition is given by Figure 3.14.

Fig 3.14 : STL Decomposition of Rainfall



The forecast obtained using STL decomposition is plotted in figure 3.15.

Fig 3.15-Forecast using STL Decomposition

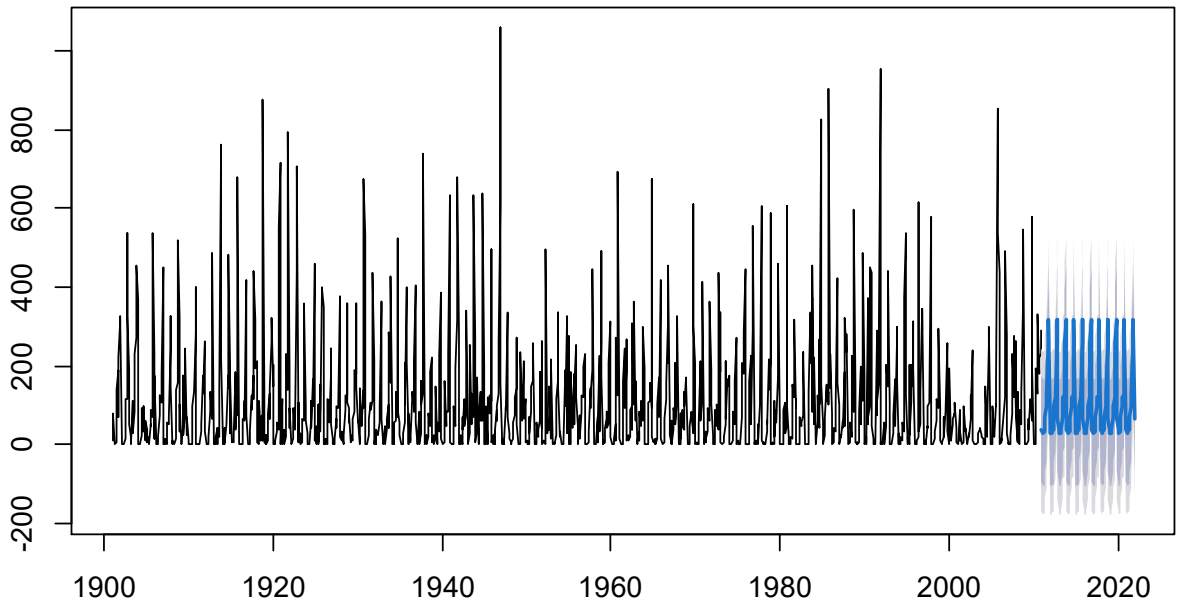


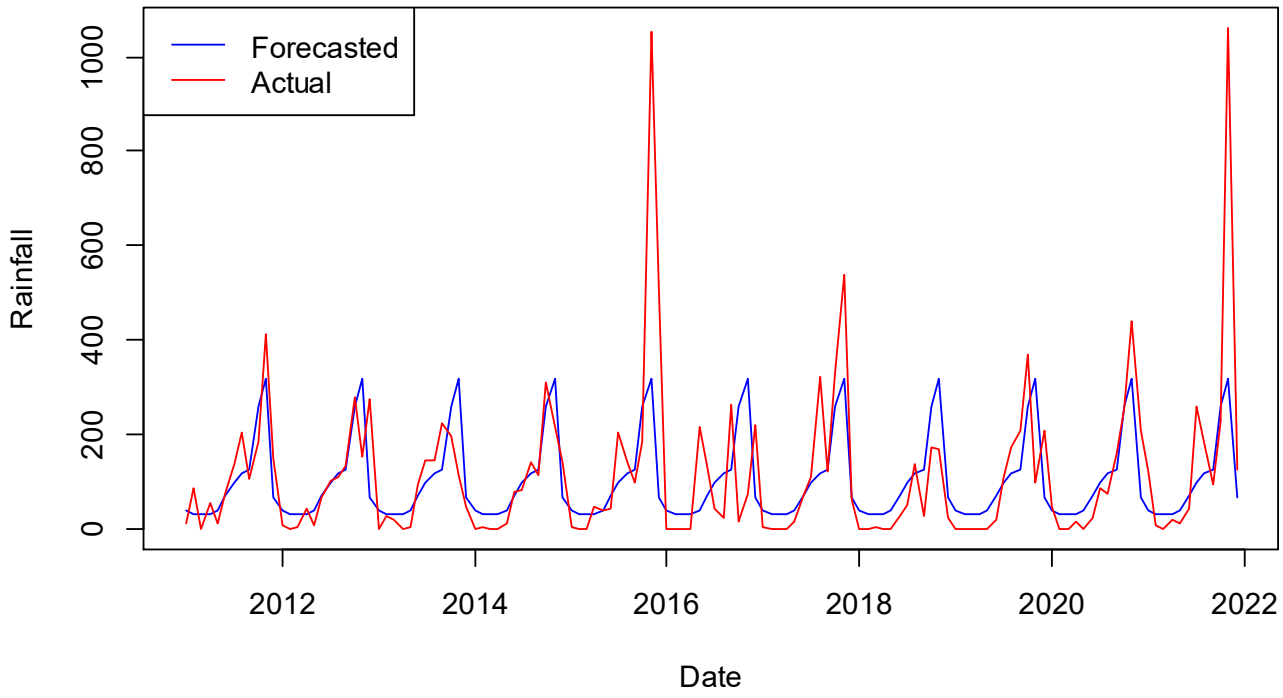
Figure 3.15 gives the plot of the forecast values obtained using STL decomposition in the shaded region. This forecast value is then compared with the values of the data at the same time point in the second partition and the difference between these actual values and forecasted values is taken as the residual at that time point, as shown in table 3.3

Table 3.3: Sample of Forecast, Actual values & residuals using STL decomposition

	Date	Forecasted	Actual	Residuals
1	2011-01-01	39.42936	11.6260862	-27.80327647
2	2011-02-01	30.00971	87.6357594	57.62604538
3	2011-03-01	29.53496	0.0000000	-29.53495902
4	2011-04-01	32.10036	54.2870996	22.18673950
5	2011-05-01	37.22194	10.9515980	-26.27034112
6	2011-06-01	68.57386	73.9905854	5.41673025
7	2011-07-01	95.78349	137.4666191	41.68312607
8	2011-08-01	119.10549	202.7637950	83.65830593
…….				
……..				
……..				
……				
126	2021-06-01	68.57386	43.4090546	-25.16480056
127	2021-07-01	95.78349	260.4401859	164.65669287
128	2021-08-01	119.10549	185.4993285	66.39383943
129	2021-09-01	124.78231	93.6575608	-31.12474885
130	2021-10-01	259.57759	228.6255028	-30.95208840
131	2021-11-01	316.96123	1061.6422220	744.68099606
132	2021-12-01	66.79381	123.6062256	56.81241627

The actual and forecasted values are plotted in Figure 3.16 against time to get a more general sense of the difference between them. This visualization allows for a comprehensive examination of how well the forecasted values align with the actual observed data over the entire time period of the 2nd partition.

Figure 3.16:Forecasted vs Actual Rainfall using STL



It is observed from figure 3.16 and figure 3.13, that the forecasted value aligns with the actual value more in the case of figure 3.16

3.4.2.3. Seasonal Naive method

The forecasted value obtained using Seasonal naïve method are plotted in figure 3.17.

Figure 3.17:Forecast using Seasonal Naive

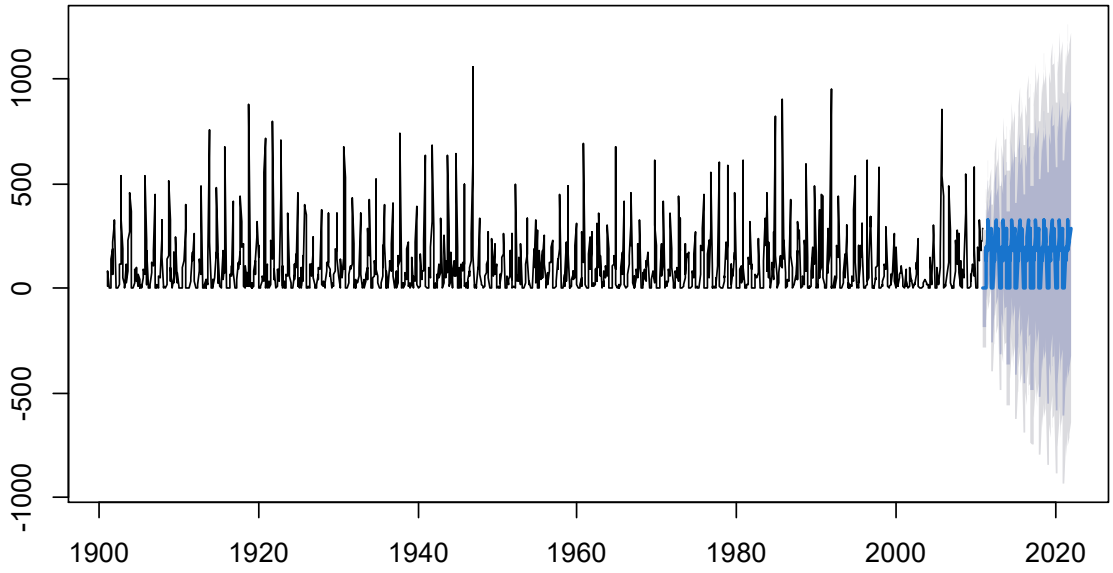


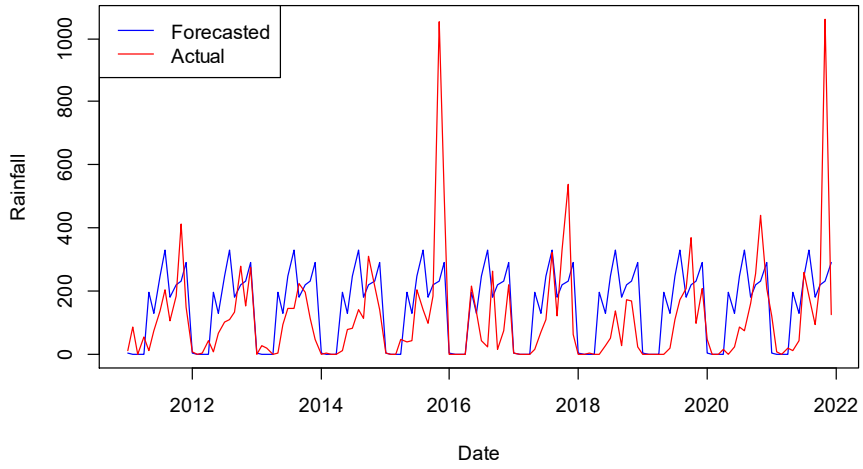
Figure 3.17 gives the plot of the forecast values obtained using STL decomposition in the shaded region. This forecast value is then compared with the values of the data at the same time point in the second partition and the difference between these actual values and forecasted values is taken as the residual at that time point, as shown in table 3.4

Table 3.4: Sample of Forecast, Actual values & residuals obtained using seasonal naïve forecasting

	Date	Forecast	Actual	Residuals
1	2011-01-01	1.2308154	11.6260862	10.39527086
2	2011-02-01	0.1207892	87.6357594	87.51497013
3	2011-03-01	0.0000000	0.0000000	0.00000000
4	2011-04-01	0.0000000	54.2870996	54.28709960
5	2011-05-01	193.9845879	10.9515980	-183.03298985
6	2011-06-01	130.4748111	73.9905854	-56.48422571
7	2011-07-01	247.1086310	137.4666191	-109.64201190
8	2011-08-01	329.7319173	202.7637950	-126.96812230
…….				
…….				
126	2021-06-01	130.4748111	43.4090546	-87.06575652
127	2021-07-01	247.1086310	260.4401859	13.33155490
128	2021-08-01	329.7319173	185.4993285	-144.23258880
129	2021-09-01	178.2481117	93.6575608	-84.59055087
130	2021-10-01	219.6521431	228.6255028	8.97335970
131	2021-11-01	230.7219979	1061.6422220	830.92022410
132	2021-12-01	288.8349686	123.6062256	-165.22874300

The actual and forecasted values are plotted in Figure 3.18 against time to get a more general sense of the difference between them. This visualization allows for a comprehensive examination of how well the forecasted values align with the actual observed data over the entire time period of the 2nd partition

Figure 3.18:Forecasted vs Actual Rainfall



It is observed that even though the forecast aligns the majority of the time to actual value there are large Residuals in the data.

3.4.3. Comparing the Forecasting models

Table 3.5: Best method of forecasting

	ARIMA	STL	Seasonal naive
MAE	90.73846	67.99901	95.83648
RMSE	150.7801	125.0251	154.3388
MASE	0.8951691	0.6708358	0.945463

It is observed from table 3.5 that because of the low values of MAE, RMSE and MASE for STL decomposition that STL decomposition is the best method for forecasting in this case.

3.4.4. Forecasting for the entire data

STL decomposition is used to forecast for the entire data due to reasons given in section 3.4.3. The STL decomposition graph for the data is given in figure 3.19.

Figure 3.19:STL Decomposition of Rainfall

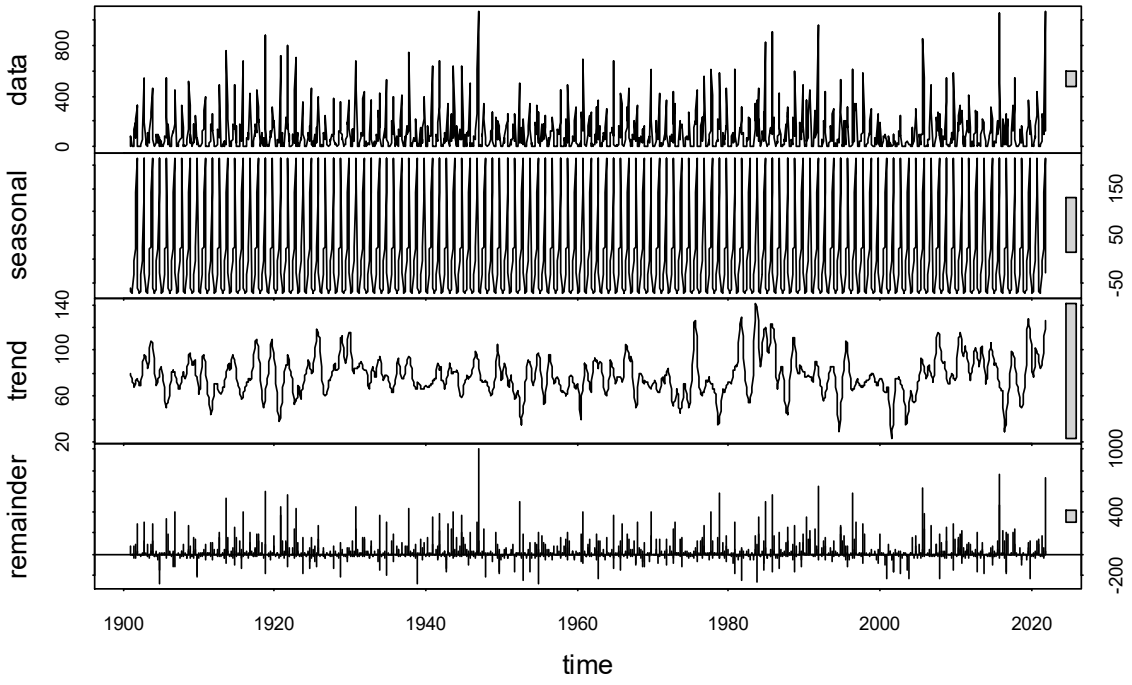


Table 3.6 gives the final forecast for the entire data using STL decomposition.

Table 3.6: Forecast for entire data

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2022	40.08102	-97.39145	177.5535	-170.16498	250.3270
Feb 2022	31.55911	-105.91336	169.0316	-178.68689	241.8051
Mar 2022	30.54613	-106.92635	168.0186	-179.69988	240.7921
Apr 2022	34.45305	-103.01942	171.9255	-175.79295	244.6991
May 2022	37.72072	-99.75175	175.1932	-172.52528	247.9667
Jun 2022	70.79082	-66.68165	208.2633	-139.45519	281.0368
Jul 2022	99.38107	-38.09141	236.8535	-110.86494	309.6271
Aug 2022	124.42615	-13.04633	261.8986	-85.81986	334.6722
Sep 2022	127.33421	-10.13827	264.8067	-82.91180	337.5802
Oct 2022	244.59069	107.11821	382.0632	34.34468	454.8367
Nov 2022	317.75316	180.28068	455.2256	107.50714	527.9992
Dec 2022	72.15624	-65.31624	209.6287	-138.08977	282.4023
Jan 2023	40.08102	-97.39146	177.5535	-170.16499	250.3270
Feb 2023	31.55911	-105.91337	169.0316	-178.68690	241.8051
Mar 2023	30.54613	-106.92636	168.0186	-179.69989	240.7921
Apr 2023	34.45305	-103.01943	171.9255	-175.79296	244.6991
May 2023	37.72072	-99.75176	175.1932	-172.52530	247.9667
Jun 2023	70.79082	-66.68166	208.2633	-139.45520	281.0368
Jul 2023	99.38107	-38.09142	236.8536	-110.86495	309.6271
Aug 2023	124.42615	-13.04634	261.8986	-85.81987	334.6722
Sep 2023	127.33421	-10.13828	264.8067	-82.91182	337.5802
Oct 2023	244.59069	107.11820	382.0632	34.34467	454.8367
Nov 2023	317.75316	180.28067	455.2256	107.50713	527.9992
Dec 2023	72.15624	-65.31625	209.6287	-138.08978	282.4023
Jan 2024	40.08102	-97.39147	177.5535	-170.16500	250.3271
Feb 2024	31.55911	-105.91338	169.0316	-178.68692	241.8051
Mar 2024	30.54613	-106.92636	168.0186	-179.69990	240.7922
Apr 2024	34.45305	-103.01944	171.9255	-175.79298	244.6991
May 2024	37.72072	-99.75177	175.1932	-172.52531	247.9668
Jun 2024	70.79082	-66.68167	208.2633	-139.45521	281.0369
Jul 2024	99.38107	-38.09142	236.8536	-110.86497	309.6271

Aug 2024	**124.42615**	**-13.04635**	**261.8986**	**-85.81989 334.6722**
Sep 2024	**127.33421**	**-10.13829**	**264.8067**	**-82.91183 337.5802**
Oct 2024	**244.59069**	**107.11820**	**382.0632**	**34.34465 454.8367**
Nov 2024	**317.75316**	**180.28066**	**455.2257**	**107.50712 527.9992**
Dec 2024	**72.15624**	**-65.31625**	**209.6287**	**-138.08980 282.4023**
Jan 2025	**40.08102**	**-97.39147**	**177.5535**	**-170.16502 250.3271**
Feb 2025	**31.55911**	**-105.91339**	**169.0316**	**-178.68693 241.8052**
Mar 2025	**30.54613**	**-106.92637**	**168.0186**	**-179.69992 240.7922**
Apr 2025	**34.45305**	**-103.01944**	**171.9256**	**-175.79299 244.6991**
May 2025	**37.72072**	**-99.75178**	**175.1932**	**-172.52532 247.9668**
Jun 2025	**70.79082**	**-66.68168**	**208.2633**	**-139.45522 281.0369**
Jul 2025	**99.38107**	**-38.09143**	**236.8536**	**-110.86498 309.6271**
Aug 2025	**124.42615**	**-13.04635**	**261.8986**	**-85.81990 334.6722**
Sep 2025	**127.33421**	**-10.13829**	**264.8067**	**-82.91184 337.5803**
Oct 2025	**244.59069**	**107.11819**	**382.0632**	**34.34464 454.8367**
Nov 2025	**317.75316**	**180.28065**	**455.2257**	**107.50711 527.9992**
Dec 2025	**72.15624**	**-65.31626**	**209.6287**	**-138.08981 282.4023**
Jan 2026	**40.08102**	**-97.39148**	**177.5535**	**-170.16503 250.3271**
Feb 2026	**31.55911**	**-105.91339**	**169.0316**	**-178.68694 241.8052**
Mar 2026	**30.54613**	**-106.92638**	**168.0186**	**-179.69993 240.7922**
Apr 2026	**34.45305**	**-103.01945**	**171.9256**	**-175.79300 244.6991**
May 2026	**37.72072**	**-99.75178**	**175.1932**	**-172.52533 247.9668**
Jun 2026	**70.79082**	**-66.68169**	**208.2633**	**-139.45524 281.0369**
Jul 2026	**99.38107**	**-38.09144**	**236.8536**	**-110.86499 309.6271**
Aug 2026	**124.42615**	**-13.04636**	**261.8987**	**-85.81991 334.6722**
Sep 2026	**127.33421**	**-10.13830**	**264.8067**	**-82.91185 337.5803**
Oct 2026	**244.59069**	**107.11818**	**382.0632**	**34.34463 454.8368**
Nov 2026	**317.75316**	**180.28065**	**455.2257**	**107.50709 527.9992**
Dec 2026	**72.15624**	**-65.31627**	**209.6288**	**-138.08982 282.4023**

The graph of the forecast is given in Figure 3.20

Figure 3.20:Forecast for the entire data

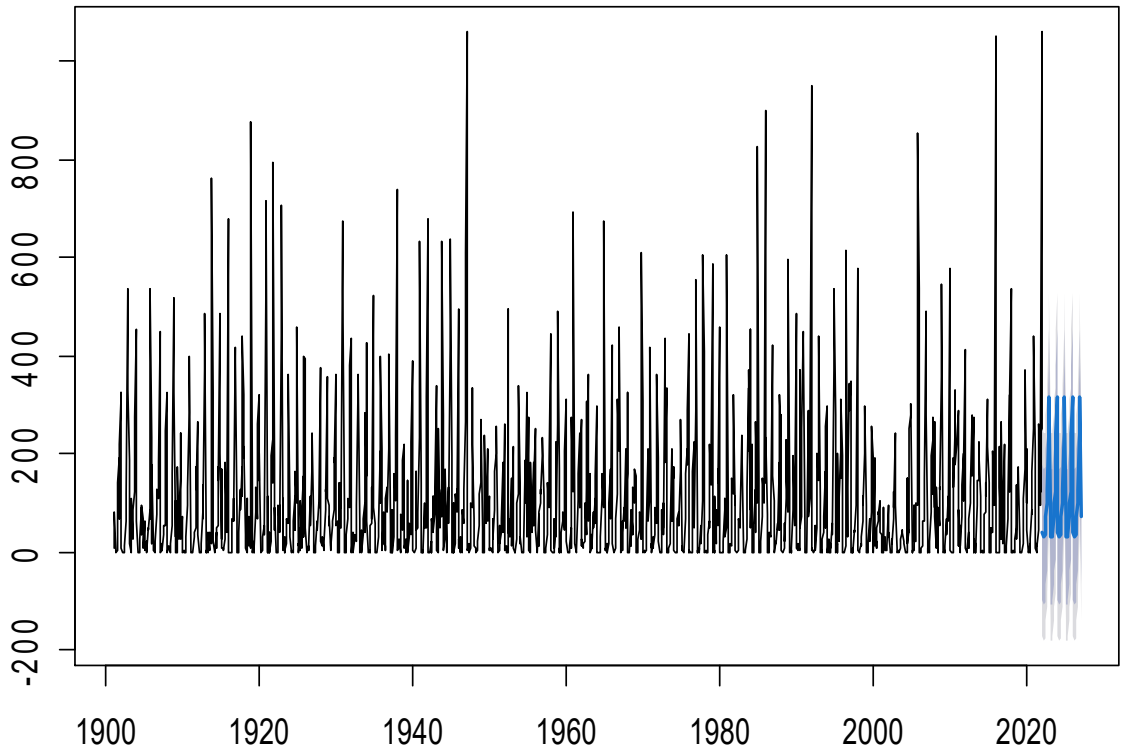


Figure 3.20 gives the plot of the forecast for the entire data using STL decomposition for the time period of January 2022 to December 2026(ie,5 years)

3.5.Final Pattern

The final forecasted value is combined with the original data to form a combined new data and K-means clustering is done in this combined data to form 7 clusters and investigation is done on this new 7 clusters and the original 7 clusters.

Clustering in the combined data is given by Figure 3.21.

Figure3.21:Combined Clustering of Original and Forecasted Data

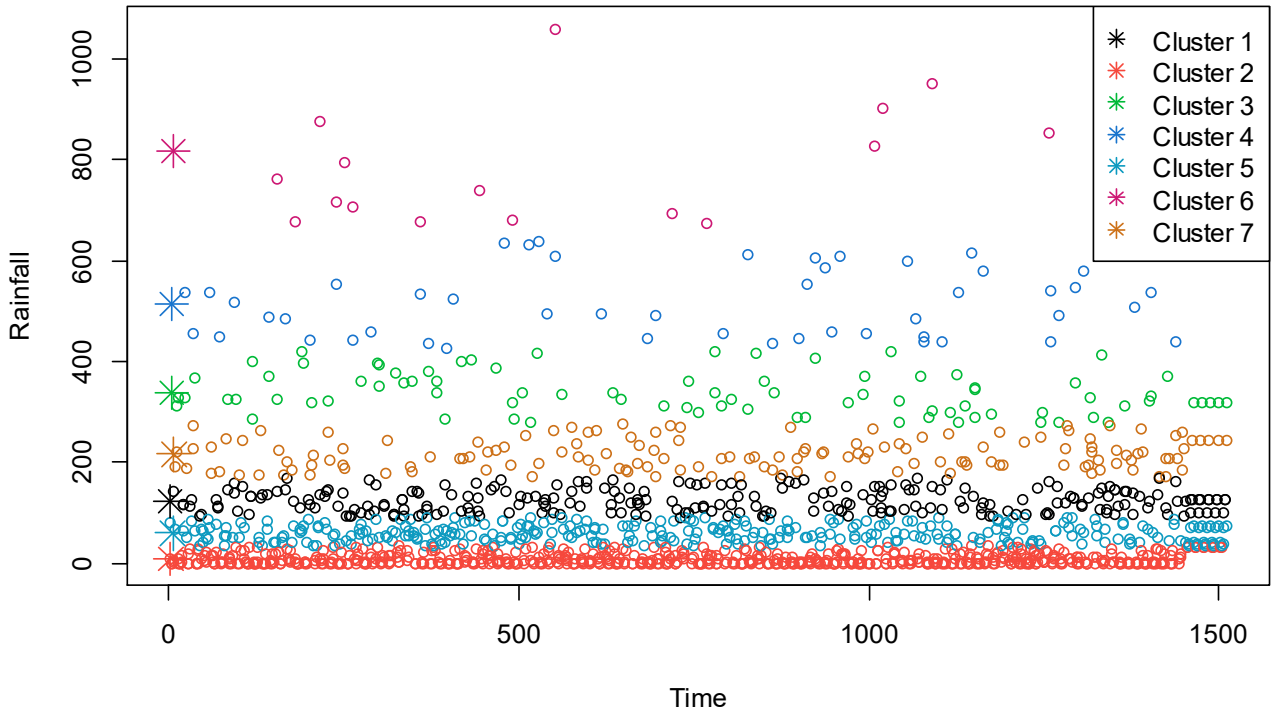


Table 3.7 gives the properties of the new clusters.

Table 3.7: Characteristics of each cluster

	C1	C2	C3	C4	C5	C6	C7
Data count	249	638	85	49	326	18	147
Data percent	16.47	42.2	5.62	3.24	21.56	1.19	9.72
Min	91.58	0	277.81	426.35	33.58	674.19	169.95
Max	168.70	33.25	420.79	639.64	91.28	1061.64	276.03
Mean	123.40	7.64	337.51	513.55	59.36	817.22	215.64
Median	122.19	2.21	327.47	495.7	58.28	778.08	210.13
Mode	99.38	0	317.75	426.35	34.45	674.19	244.59
Standard Deviation	21.71	9.91	40.80	66.17	16.47	139.35	29.6
Variance	471.4	98.1	1655.05	4379.43	271.15	19417.4	875.98
IQR	38.48	13.61	55.74	105.28	26.94	198.20	47.15
Skewness	0.3	1.18	0.47	0.41	0.19	0.66	0.35
Kurtosis	1.95	3.01	2.22	1.85	1.90	2.08	2.07

From visual inspection of tables 3.1 and 3.7 it is apparent that the new and old clusters have similar properties. The 2 clusters have an Adjusted Rand Index (ARI) of 0. 950.This indicates a high level of agreement between the old and new clusters

CHAPTER-4

CONCLUSION

In this thesis, K-means clustering algorithm was used to divide monthly rainfall in Chennai from the year 1901 to 2021 into 7 clusters. This clusters were then studied extensively using their contribution to the whole dataset, their measures of central tendency and variability in the distribution of the data points among each cluster. These properties were then used to comment about the distribution of the datapoints in the cluster.

After the splitting of the original data into 2 partition, extensive study was done to find the beast method of forecasting among SARIMA, STL decomposition and seasonal naïve forecasting. In this case because of the low values of MAE, RMSE and MASE it was inferred that STL decomposition is the best method for forecasting for highly seasonal data.

This information was then used to do forecasting for the entire data using the best method of STL decomposition. The forecast and the graph of the forecast are provided. This forecast value was then added to the original data. In this new combined data, clustering was done much like in the earlier case and the properties of the clusters were observed. It is noted that the properties of the clusters especially measures of variability have only minimal divergence. This suggests that the distribution of the data points within the cluster have only minimal changes after incorporating the forecasted values. This consistency indicates that the forecasting method, based on the STL decomposition, effectively captured the underlying patterns in the data.

Bibliography

[1].Bentivoglio, R., Isufi, E., Jonkman, S. N., & Taormina, R. (2022). Deep learning methods for flood mapping: a review of existing applications and future research directions. Hydrology and earth system sciences, 26(16), 4345-4378.

[2].Bora, S., & Hazarika, A. (2023, April). Rainfall time series forecasting using ARIMA model. In 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) (pp. 1-5). IEEE.

[3].Chan, F. K. S., Yang, L. E., Scheffran, J., Mitchell, G., Adekola, O., Griffiths, J., ... & McDonald, A. (2021). Urban flood risks and emerging challenges in a Chinese delta: The case of the Pearl River Delta. Environmental Science & Policy, 122, 101-115.

[4].Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. Disaster Adv, 11(11), 27-33.

[5].Precipitation Time Series Analysis and Forecasting for Italian Regions Ebrahim Ghader pour, Hanieh Dadkhah , Hamed Dabiri, Francesca Bozzano, Gabriele Scarascia Mugnozza and Paolo Mazzanti

[6].Stedinger, J. R. (2012). Statistical Methods for Assessing Flood Risk and the Climate Change Challenge. Revista de Ingeniería, (36), 48-53.

[7].Stedinger, J. R. (2012). Statistical Methods for Assessing Flood Risk and the Climate Change Challenge. Revista de Ingeniería, (36), 48-53.

[8].Tang, Y., Sun, Y., Han, Z., Wu, Q., Tan, B., & Hu, C. (2023). Flood forecasting based on machine learning pattern recognition and dynamic migration of parameters. *Journal of Hydrology: Regional Studies*, *47*, 101406.

[9].Wang, X., Zhang, T., Wang, Y., Huang, X., Gong, H., & Chen, B. (2023). Temporal and Spatial Distribution Characteristics of Flood Disasters with Different Intensities in Arid-Semiarid Region in Northern Xinjiang, China. In E3S Web of Conferences (Vol. 394, p. 01009). EDP Sciences.

[10].Zhao, J., Wang, J., Abbas, Z., Yang, Y., & Zhao, Y. (2023). Ensemble learning analysis of influencing factors on the distribution of urban flood risk points: a case study of Guangzhou, China. Frontiers in Earth Science, 11, 1042088.