ABSTRACT

Due to the climate of India, it is prone to many kinds of natural disaster at different times of the year. For this the Government of India has allocated two types of funds -State Disaster Response Fund (SDRF) and National Disaster Response Fund (SDRF). In this project we are mostly interested in SDRF. For SDRF, the contribution is made by both the Central Government and State Governments for all states. The SDRF shall be used only for meeting the expenditure for providing immediate relief to the victims of cyclone, drought, earthquake, fire, flood, tsunami, hailstorm, landslide, avalanche, cloud burst, pest attack and frost & cold wave. Initially we do an exploratory analysis to find the insight of the data,association and pattern.

Next, we try to see any particular factors like population, area and coastline length of the states are affecting the allocation of SDRF. So we only focus on the 9 states with coastline areas (West Bengal, Madhya Pradesh, Maharashtra, Tamil Nadu , Kerala, Karnataka, Goa, Gujrat and Odisha). We also check the assumptions of our model and if any assumption is violated(i.e if the data has heteroscedasticity or multicollinearity or autocorrelation) we will take measures to remove any model inadequacies.

Lastly we would like to do a time series analysis on the data mainly to forecast the 2024 SDRF for the 28 states. Our entire analysis will be focused within the period 2011-2023 and will check how close the forecasted SDRF values are for year 2024(the 2024 data has released recently). We also want to do a bivariate time series to incorporate any possible correlation between states and central to get better results.

CONTENTS

1.INTRODUCTION:

India is a large country with a geographical area of 3.28 million sq. kms. Situated between the latitudes 8 4' N and 37 6' N and Longitudes 68 7' E and 97 25', India has a tropical and subtropical Climate. The country is bounded in the north by the Himalayan mountain ranges . India, on account of its geographical position, climate and geological setting, has had from time immemorial, a fair share of natural disasters. There is hardly a year when some part of the country or other does not face the spectre of drought, due to the failure of monsoons in vulnerable areas. One or two cyclones strike the peninsular region of the country every year. Similarly, floods are a regular feature of the Eastern India where Himalayan rivers inundate large parts of its catchment areas uprooting people, disrupting livelihood and damaging infrastructure. The fragility of the Himalayan mountain ranges are a continuing source of concern for their high vulnerability to earthquakes, landslides and avalanches. The recent earthquakes in Maharashtra and Madhya Pradesh have demonstrated that the areas considered comparatively safe till now, are really not so. Some of the most common natural disasters in India are -flood ,drought ,avalanches, cyclone, earthquake and landslides.

The Government allots annual funding to each state in response to the natural disaster and for providing immediate relief to the victims.

<u>State Disaster Response Fund (SDRF) and National Disaster Response Fund (NDRF)</u>

The Government of India supplements the effort of the State Government by providing assistance for relief of immediate nature through two ways (i) State Disaster Response Fund (SDRF) and (ii) National Disaster Response Fund (NDRF) as per established procedure.

The allocation of funds under SDRF and NDRF is based on the recommendations of the successive Financial Commissions. For SDRF, the contribution is made by the Central Government and State Governments.

In this project we mainly focus on the State Disaster Response Fund (SDRF) considering both central share and state share for analysis.

2.ABOUT THE DATA

The following data consists of 13(+1) years of allocated State Disaster Response Fund (SDRF) (including both central and states share) for 29 states from year 2011 to 2023.The 2024 SDRF data is also collected after the national budget is sanctioned.

Disaster (s) covered under SDRF: Cyclone, drought, earthquake, fire, flood, tsunami, hailstorm, landslide, avalanche, cloudburst, pest attack, frost and cold waves.

The State Disaster Response Fund (SDRF), constituted under Section 48 (1) (a) of the Disaster Management Act, 2005, is the primary fund available with State Governments for responses to notified disasters. The Central government also contributes a significant amount.

3.SOURCE OF THE DATA

The State Disaster Response Fund data is collected from the following website:

https://ndmindia.mha.gov.in/response-fund

The population and density of states data is collected from

https://en.wikipedia.org/wiki/List_of_states_in_India_by_past_population

the coastline length data is collected from:

https://testbook.com/ias-preparation/coastal-states-of-india

4.OBJECTIVE

The main objective of this project is to

1. Forecast of State Disaster Response Fund (SDRF), using time series analysis

2. To see if there is any influence of area, population and coastline length on the SDRF allocation on certain states.

5.METHODOLOGY

5.1 Visualisation

- The data collected on State Disaster Response Fund (SDRF) for both state and central share for 29 states from the year 2011-2023 has been organised into two tables for time series analysis:

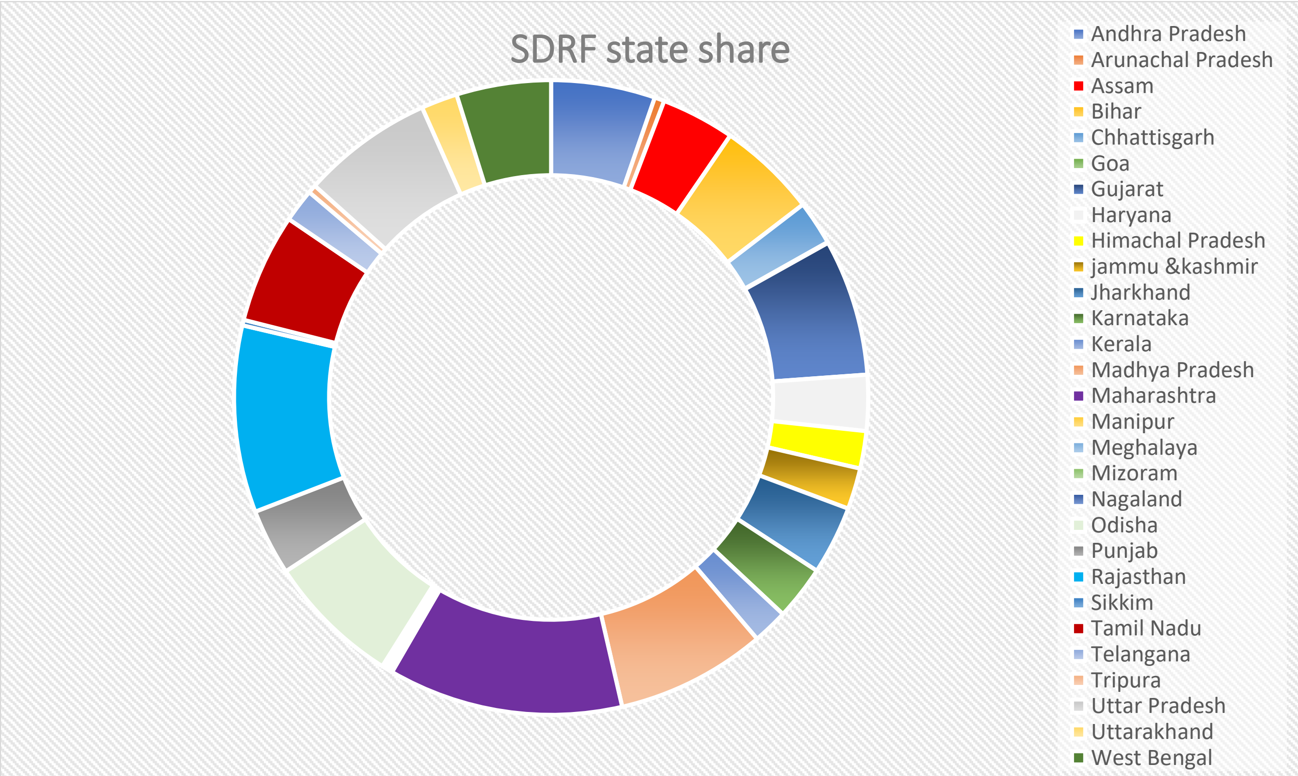
1. For state share
2. For Central share

Firstly, we visualize our data

So we take the means SDRF for each state over 13 years and draw the doughnut diagrams accordingly:

5.1.1 Doughnut chart for SDRF state share

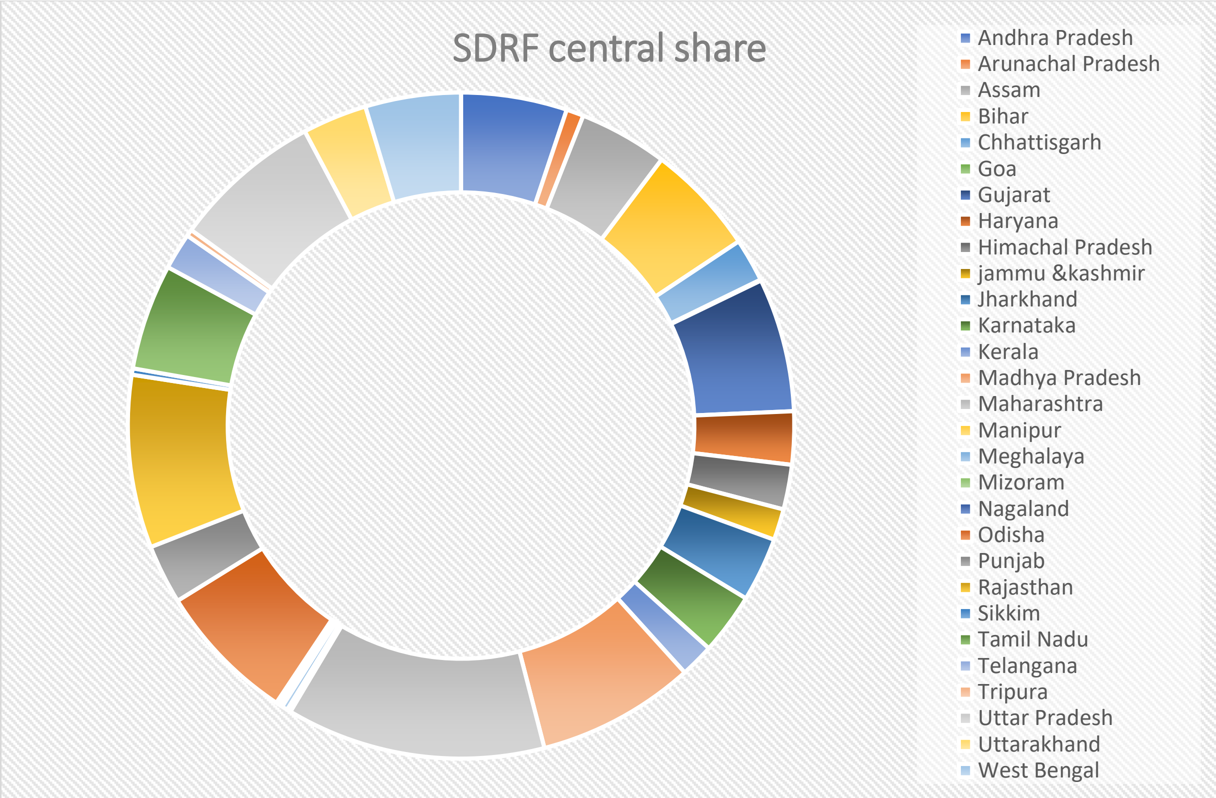
Fig5.1.1: Chart showing State Disaster Response Fund (SDRF) state share



Comment : The doughnut chart shows with the most fund gets allocated to Odisha ,Maharashtra,Gujrat ,Madhyapradesh, Haryana, Rajasthan and Tamil Nadu.

5.1.2 Doughnut chart for SDRF central share

Fig5.1.2: Chart showing State Disaster Response Fund (SDRF) central share



Comment : The doughnut chart shows with the most fund gets allocated to Odisha ,Maharashtra,Gujrat ,Madhyapradesh, Haryana, Rajasthan and Tamil Nadu.

<u>Information regarding Natural disaster and the states associated with it.</u>

Table 5.1.1Breakdown of natural disasters in India per type of event and nature of losses

	Natural disasters	Material loss	Human loss
Floods	52%	63%	32%
Hurricanes	30%	19%	32%
Landslides	10%	-	2%
Earthquakes	5%	10%	33%
Droughts	3%	5%	1%
Total	**100%**	**100%**	**100%**

Table 5.1.2 List of major natural disasters that have occurred over the last 20 years in India

Date	Place	Nature of the event	Economic losses (in billion USD)	Insured losses (in billion USD)	Number of fatalities	Number of affected persons
May 2020	West Bengal	Hurricane Amphan	13.5	ND	103	500 000 homeless
August 2018	Kerala	Floods	3.52	0.37	504	223 139 homeless
November 2015	Chennai (Tamil Nadu)	Floods	2.37	0.98	289	-
April 2015	Himalaya	Storm	-	-	78	20 000 injured
October 2014	Andhra Pradesh	Storm	7.56	0.68	68	43 injured
September 2014	Jammu and Kashmir	Floods	6.45	0.26	665	-
June 2013	Uttarranchal	Floods	1.21	0.55	5 748	4 473 injured 271 931 homeless
September 2009	Andhra Pradesh	Floods	5.63	0.06	300	2 000 000 homeless
August 2006	Gujarat	Floods	4.3	0.52	350	4 000 000 homeless
July 2005	Maharashtra	Floods	4.36	0.93	1 150	15 000 homeless
January 2001	Gujarat	Earthquake	6.13	0.14	19 737	166 850 injured 1 790 000 homeless

Source: https://www.atlas-mag.net/en/article/natural-disasters-risk-in-india

Note: The most occurring disasters are noted in states -Maharasthra,Gujrat,Andra Pradesh ,West Bengal and some more mentioned in the above table. Hence they are allocated with more disaster funds. Interesting enough most of this states mentioned above contain prominent coastline areas, so we may want to check if the length of coastline has any effect or not. Also most of the state are moderately large in population and area. We will also take that into account.

5.2 Explorative Study on SDRF data

First we cluster the data and see if there is any interesting feature of fund allocation from state or central for any state:

5.2.1 Clustering using K-medoid

The K-Medoids algorithm is similar to the K-Means Algorithm, the only difference being that the clusters are centered on their medians, not their means. The algorithm divides the dataset into k (where k is specified from beforehand) clusters or groups, in such a manner that that variation within each cluster is minimized. This ensures that the observations within a cluster are as similar to each other as possible, and the clusters themselves are such that the properties of one cluster are distinct from the other clusters. In our context, this would ensure that each cluster would contain countries with distinct socio-economic conditions, and the countries within a cluster would be as similar to each other as possible. Here, similarity is measured in terms of Euclidean distance.

Let, C_1, C_2, C_3,...,C_k denote the sets containing the indices of the observations belonging to the k clusters. Then we have:

- $C_1 \cup C_2 \cup \ldots \cup C_k = \{1,\ldots,n\}$ this means that each observation belongs to at least one cluster.
- $C_1 \cap C_2 \cap \ldots \cap C_k = \emptyset$ this means that no observation can belong to more than one cluster simultaneously.

The Algorithm

- Step 1: A number from {1, 2,..., k} is randomly assigned to each of the observations. These are the initial cluster assignments; the observations start out as belonging to the cluster they have been assigned.
- Step 2: The p-feature median is computed for each cluster.
- Step 3: Then each observation is assigned to the cluster whose median it is closest to. Here, closeness is defined in terms of Euclidean distance.
- Step 4: Steps 2 and 3 are repeated till the cluster assignments stop changing, that is, convergence is achieved.

- For a cluster C_k the within cluster sum of squares $W(C_k)$ is a measure of how much the observations inside the cluster Ck differ from each other. So, our objective is to minimize:
- $minC_1,...,C_k \sum W\ (C_k)$
- We achieve this using the concept of Euclidean distance. The within-cluster variation for the k - th cluster is the sum of all of the pairwise squared Euclidean distances between the observations in that cluster, divided by the total number of observations in the cluster. The within cluster sum of squares for the cluster Ck can alternatively also be defined as:
- $W\ (C_k)\ =\ (1/\ |\ C_k\ |) \sum \sum\ (xij\ -\ xi\ 'j)^2$
- Where: | C_k | C_k denotes the number of observations in the k-th cluster.
- Thus, within cluster sum of squares for each cluster is written in terms of the squared Euclidean distance.
- The problem now becomes to minimize:
- $minC_1, \ldots, C_k \sum_{k'=1}^{k} \frac{1}{|C'k|} \sum_{i,i' \in C'k} \quad \sum_{j=1}^{p} (xij - xi'j)$
- This is achieved with the help of the algorithm. This algorithm is run on our dataset. However, before we can finalize the number of clusters and further study them, we need to determine the optimum number of clusters. In order to do so, we use the Elbow Method.
- We run the algorithm on the dataset for a range of k. Next, we plot the total within sum of squares, summed over all the clusters, versus k.

Results

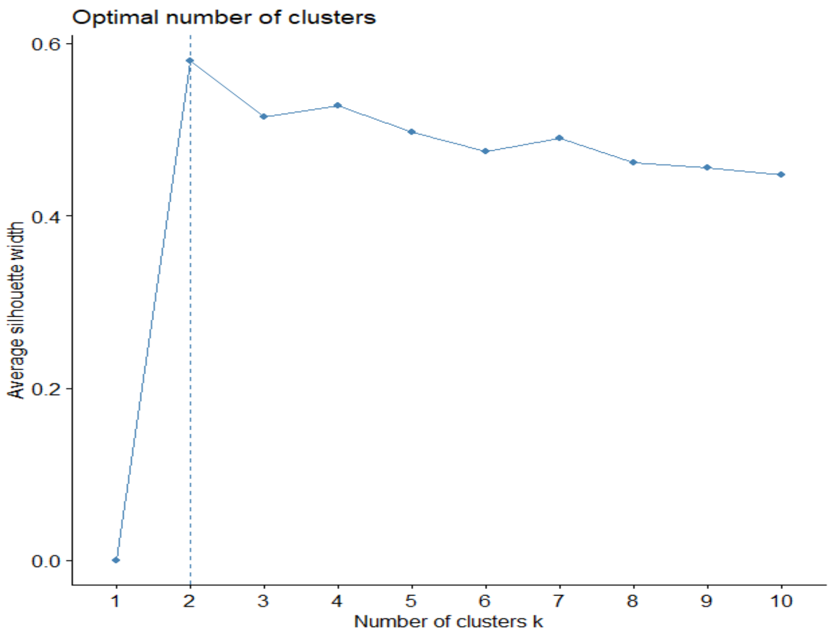
Now we cluster the data on the basis of the budgets allotted by SRDF over the course of 13 years. Here we take the 13 years of SRDF fund for each state as the data points

For SRDF state share

Elbow method:

We run the algorithm on the dataset for a range of k. Next, we plot the total within sum of squares, summed over all the clusters, versus k.

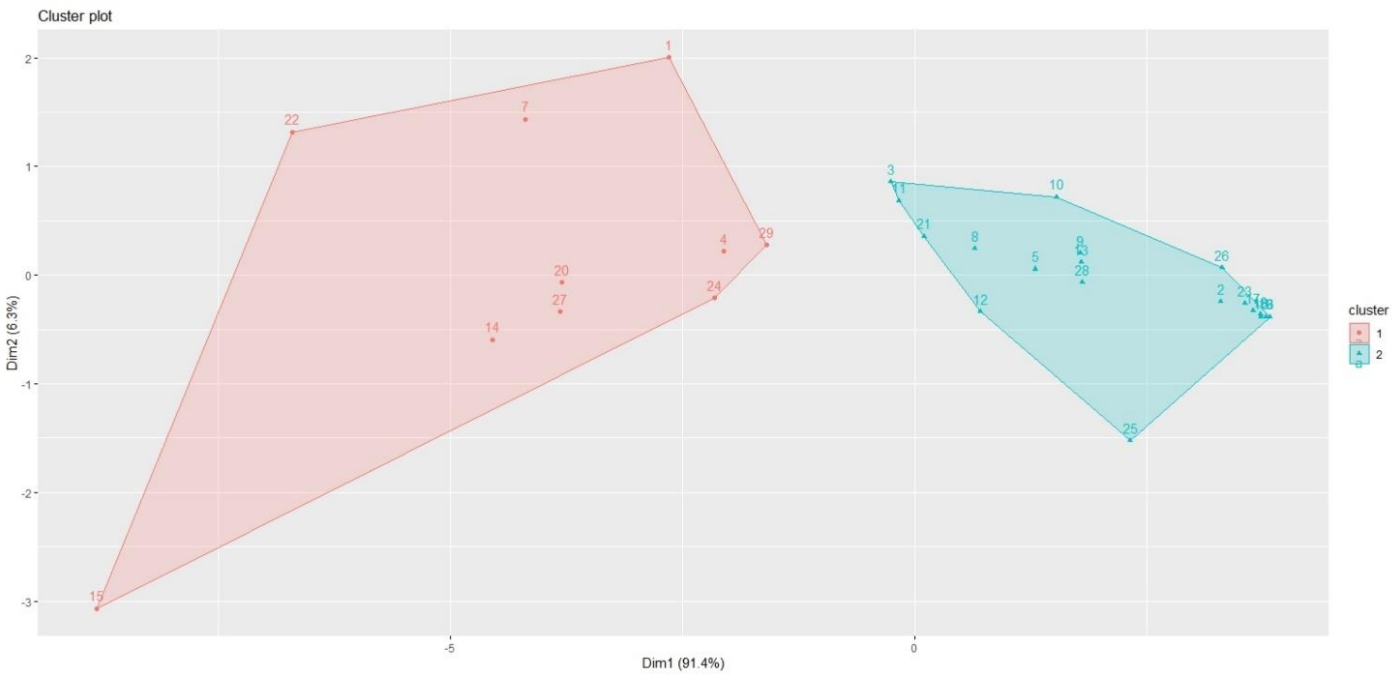
Fig 5.2.1



Clearly the optimum no of clusters we need is 2

So the clusters are as follows:

Fig -5.2.2



Now we study the characteristics of both the clusters:

Table 5.2.1 table showing cluster means

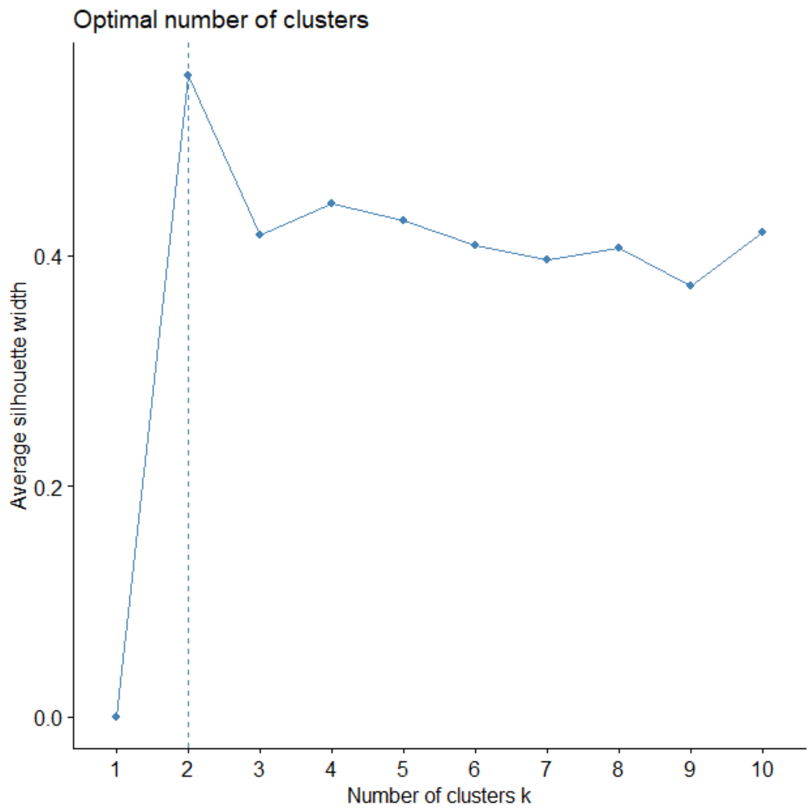
Year	Cluster 1 means	Cluster 2 means
2011	415.677	101.07526
2012	436.471	115.70895
2013	458.293	111.43632
2014	481.207	117.00842
2015	480.144	122.86000
2016	769.400	178.26316
2017	807.900	187.15789
2018	848.200	196.42105
2019	890.800	206.15789
2020	935.100	215.10526
2021	531.700	78.00000
2022	425.360	62.40000
2023	446.560	65.47368

Studying the cluster means it is evident the clusters are well formed with obviously states with higher SDRF goes to cluster 1 and the states with lower SDRF goes to cluster 2.

For SDRF central Share

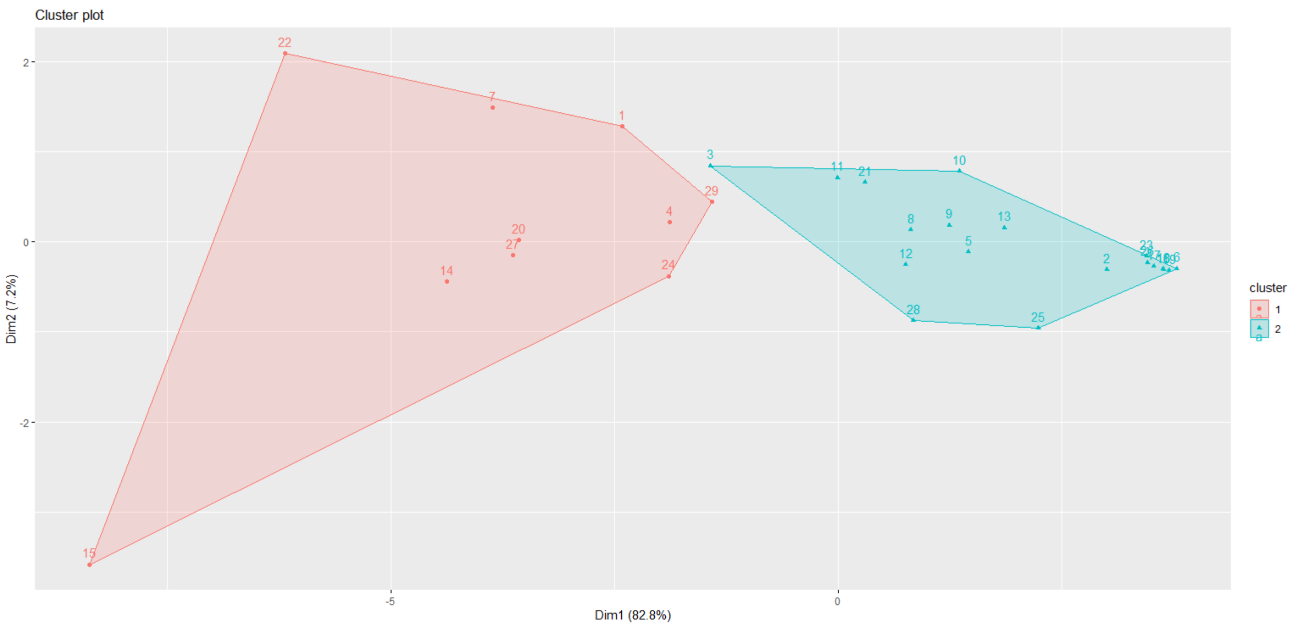
Elbow method:

Fig 5.2.3



Clearly the optimum no of clusters we need is 2

Fig 5.2.4



Now we study the characteristics of both the clusters:

Table 5.2.2 table showing cluster means

Year	Cluster 1 means	Cluster 2 means
2011	310.3720	64.95553
2012	311.3160	61.38316
2013	331.7130	100.06000
2014	407.2555	103.22158
2015	334.2690	112.25474
2016	590.4280	150.09000
2017	547.5380	152.60895
2018	696.4440	127.28211
2019	629.4210	177.97474
2020	739.8970	186.24053
2021	1595.0000	328.10526
2022	1276.0000	262.48421
2023	1340.0000	275.53684

Studying the cluster means it is evident the clusters are well formed with obviously states with higher SDRF goes to cluster 1 and the states with lower SDRF goes to cluster 2.

The clusters are given as follows:

Table 5.2.3 Table showing with state has been placed in which cluster

State	cluster(stateshare)	cluster(centralshare)
Andhra Pradesh	1	1
Bihar	1	1
Gujarat	1	1
Madhya Pradesh	1	1
Maharashtra	1	1
Odisha	1	1
Rajasthan	1	1
Tamil Nadu	1	1
Uttar Pradesh	1	1
West Bengal	1	1
Arunachal Pradesh	2	2
Assam	2	2
Chhattisgarh	2	2
Goa	2	2
Haryana	2	2
Himachal Pradesh	2	2
jammu &kashmir	2	2
Jharkhand	2	2
Karnataka	2	2
Kerala	2	2
Manipur	2	2
Meghalaya	2	2
Mizoram	2	2
Nagaland	2	2
Punjab	2	2
Sikkim	2	2
Telangana	2	2
Tripura	2	2
Uttarakhand	2	2

Comment: From the table we can see that the state that is in cluster 1 for SDRF state share is also in cluster 1 for SDRF Central share, eg. Kerala is in cluster 2(the cluster with low SDRF fund allocation) for both central and state share. Also fig 5.2.3 and fig 5.2.4 are very similar in nature and structure though the clusters are not same. The data for state and central may have some kind of association or correlation.

To see the relationship between the SDRF state share and Central share we do a rank correlation test.

5.2.2 Rank correlation Test

The Spearman's rank coefficient
The Spearman's rank coefficient of correlation or Spearman correlation coefficient is a nonparametric measure of rank correlation (statistical dependence of ranking between two variables).
It measures the strength and direction of the association between two ranked variables.

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2 - 1)}$$

Here,

n= number of data points of the two variables
di= difference in ranks of the "ith" element
The Spearman Coefficient, ρ, can take a value between +1 to -1 where,

- A ρ value of +1 means a perfect association of rank
- A ρ value of 0 means no association of ranks
- A ρ value of -1 means a perfect negative association between ranks.

Results

Hence for cluster 1(cluster with high disaster-prone states)

The rank correlation between state and central share of SDRF

rho

0.9151515

For cluster 2(cluster with low disaster-prone states)

The rank correlation between state and central share of SDRF

0.9245614

Overall rank correlation between the SDRF State Share and Central Share.

Rho= 0.9753695

Comment: Here we can see the SDRF state share and central share is highly associated among each other. This also indicate that the state that gets higher SDRF share from state also gets higher share from central .This means that there is no unfairness in SDRF allocation to any state from central. . Each state has uniform allocation of funds with higher disaster-prone state getting comparatively more fund than the low disaster -prone states.

5.3 Fitting of linear Regression

Now we perform multiple regression on both the clusters of SDRF state and also SDRF central share.

5.3.1 Multiple Regression

A linear model is usually fit to a data set where the response (y) and the regressor ($xi's$) variables have a clear linear relationship.
Then our linear model is of the form
$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta p-1xp-1+e$
Where the $\beta i's$ are the parameters and e is the error.

However, this model is valid only under the following assumptions:

- ➢ There is a linear relationship between the response and each of the regressor variables.
- ➢ The mean of the errors must be equal to 0.
- ➢ The errors are independently distributed.
- ➢ The errors follow a normal distribution with mean 0 and constant variance.
- ➢ The errors must be uncorrelated
- ➢ The covariates must be independent of each other
- ➢ Covariates is independent of errors.

Variables in the data

The variables that are used in the study are:

1. y = SDRF state share mean/SDRF central share mean
2. x_1 =population of state
3. x_2= coastal length(km)
4. x_3= Area(km^2)

Here *y* is the response variable or the dependent variable and x_1, x_2, x_3 are the regressor or the independent variables.

5.3.1.1 Check if the model fits the data

Firstly, we try to fit linear multiple regression to the given data. Once a linear model is fitted to the data, next we check if the model fits the data well. This is done by calculating the F-

Statistic and comparing it with the tabulated F-Statistic. If the calculated F-Statistic is greater than the tabulated F-Statistic, we reject the null hypothesis $H0$: $\beta1=\cdots=\beta p=0$, with the alternative hypothesis $H1$: at least one βi, i=1(1)p is non-zero which implies that our linear model fits the data well.

5.3.1.2 Check for possible model inadequacies

When we have a data in hand, and we have already fitted a linear model to it, it is not always necessary that our fitted model will be adequate for the data.
So we will check if the assumptions of linear model are violated or not.

5.3.1.3 Check for Multicollinearity

Multicollinearity

- Multicollinearity is a statistical concept where several independent variables in a model are correlated.
- Multicollinearity among independent variables will result in less reliable statistical inferences.

Detection of Multicollinearity:
The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} = \frac{1}{\mathrm{Tolerance}}$$

Where $\mathbf{R_i^2}$ represents the unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones.
Vif>5 indicates presence of multicollinearity.

Removal of multicollinearity

There are several ways to remove multicollinearity (PCA, Ridge Regression, Lasso etc)
Here we will use Lasso:

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model

Letting X be the covariate matrix, so that $X_{ij} = (x_i)_j$ and x_i^T is the i^{th} row of X, the expression can be written more compactly as

$$\min_{\beta_0,\beta} \left\{ \|y - \beta_0 - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t,$$

where $\|u\|_p = \left(\sum_{i=1}^{N} |u_i|^p \right)^{1/p}$ is the standard ℓ^p norm.

5.3.1.4 Heteroscadasticity

Heteroskedasticity refers to a situation where the variance of the residuals is unequal over a range of measured values.

Detection: Gold-feld Quant Test

Steps to Perform Goldfeld-Quandt Test:

Step 1: Arrange the observations in ascending order of X_i. If there are more than one explanatory variables(X) then you choose the one regarding which you have a concern that with this variable the error variance is positively related and arrange in ascending order according to this variable. In other words, you can choose any one of them to arrange.

Step 2: Omit 'c' central observations and divide the remaining (n-c) observations into two groups containing (n-c)/2 observations each. The first (n-c)/2 observations belong to the first group(the smaller variance group) and the remaining (n-c)/2 observations belong to the second group(the larger variance group).

Step 3: Fit a separate regression model for the first group and obtain RSS_1. Also, fit a separate regression model on the second group and obtain RSS_2.

This RSS each have **(n-c)/2 – k** or **(n-c-2k)/2 degrees of freedom**, where k is the number of parameters to be estimated.

For a model with only one explanatory variable(X) the value of k =2 and increases with an increase in the number of explanatory variables.

Step 4: Compute the Test Statistic F

Step 5: Find out the critical value

Use the F Table to find out the critical value for the given level of significance(alpha). In this test, the values of df_1 and df_2 are the same(df1=df2).

Step 6: Compare $F_{critical}$ and $F_{calculated}$ and state the result.

5.3.1.5*Autocorrelation*

Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

Detection: Durbin Watson Test

Durbin Watson Test: A test developed by statisticians professor James Durbin and Geoffrey Stuart Watson is used to detect autocorrelation in residuals from the Regression analysis. It is popularly known as Durbin-Watson d statistic, which is defined as

$$d = \frac{\sum_{t=2}^{t=n} (u_t - u_{t-1})^2}{\sum_{t=1}^{t=n} u_t^2}$$

Results

Now we check if the coastline ,area and population has has any effect on the SDRF

Note: we have total 9 states with coastal area ,hence we have 9 data points.

Since SDRF state share and central share is highly correlated we will just check for the state share.

Hence the SDRF state share model is given by

```
Call:
lm(formula = statemean ~ est.population + area.in.sq.km. + coastline.km.,
    data = c_d)

Residuals:
        1         2         3         4         5         6         7         8 
 -75.5965  -24.7579  -42.1251 -314.2992    0.5579  137.7355  119.1441   51.2337 
        9 
 148.1074 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.412e+00  1.371e+02   0.032    0.976
est.population -2.283e-07  4.372e-06  -0.052    0.960
area.in.sq.km.  2.819e-03  1.691e-03   1.667    0.156
coastline.km.   8.553e-02  2.067e-01   0.414    0.696

Residual standard error: 181.4 on 5 degrees of freedom
Multiple R-squared:  0.7714,	Adjusted R-squared:  0.6342 
F-statistic: 5.623 on 3 and 5 DF,  p-value: 0.04655
```

Comment: The results shows none of the covariates have any influence on the response ,which is contradictory to the fact that in table 5.1.2 we found that most notable natural disaster have occurred in states with prominent coastal areas. Also the R-squared is 0.77 and adjusted R-squared is 0.63. The linear model is not giving best fit but not worst fit as well. So we check if the model violates any assumption of linear model.

Observations

- No heteroscadatiscity(using Gold_feld Quandt test)
- No autocorrelation (using Durbin Watson test)
- Presence of multicollinearity (using variance inflation factor)
- population area.in.sq.km. coastline.km.
- 6.636687 5.741128 1.455836

  We see for both population and area the vif >5 ,so we suspect multicollinearity and proceed to remove it by Lasso

Lasso

Best lamda(tunning parameter)= 21.99194

The estimates of lasso

```
                         s0
(Intercept)     57.574201105
est.population   .
area.in.sq.km.   0.002560779
coastline.km.    0.037420223
```

Comment : Hence both the area and length of coastline has a significance influence on the statemean . This implies states with coastal areas (with larger area)are more prone to disaster and should be allotted with higher SDRF for immediate relief of the casualties.. Also this applies for central share.

5.4 Time series Analysis & Forecasting

Time Series Analysis is a way of studying the characteristics of the response variable concerning time as the independent variable.

To perform the time series analysis, we have to follow the following steps:

- Collecting the data and cleaning it
- Preparing Visualization with respect to time vs key feature
- Observing the stationarity of the series
- Developing charts to understand its nature.
- Model building – AR, MA, ARMA and ARIMA
- Extracting insights from prediction
- Forecasting.

5.4.1 Forecasting using univariate time series

Step 1: Organizing the data:

We set up 29 univariate time series model(each for one state) for SDRF state share over the span of 13 years starting from 2011 to 2023 and forecast the SDRF state share 2024 for 29 states from this 29 models.

Similarly for SDRF central share we take 29 univariate time series model(each for one state) over the span of 13 years starting from 2011 to 2023 and forecast for 2024.

Step 2: Initially we try to visualize the SDRF central and state share data over the course of 13 years for some states.

Fig 5.4.1 Andhra Pradesh

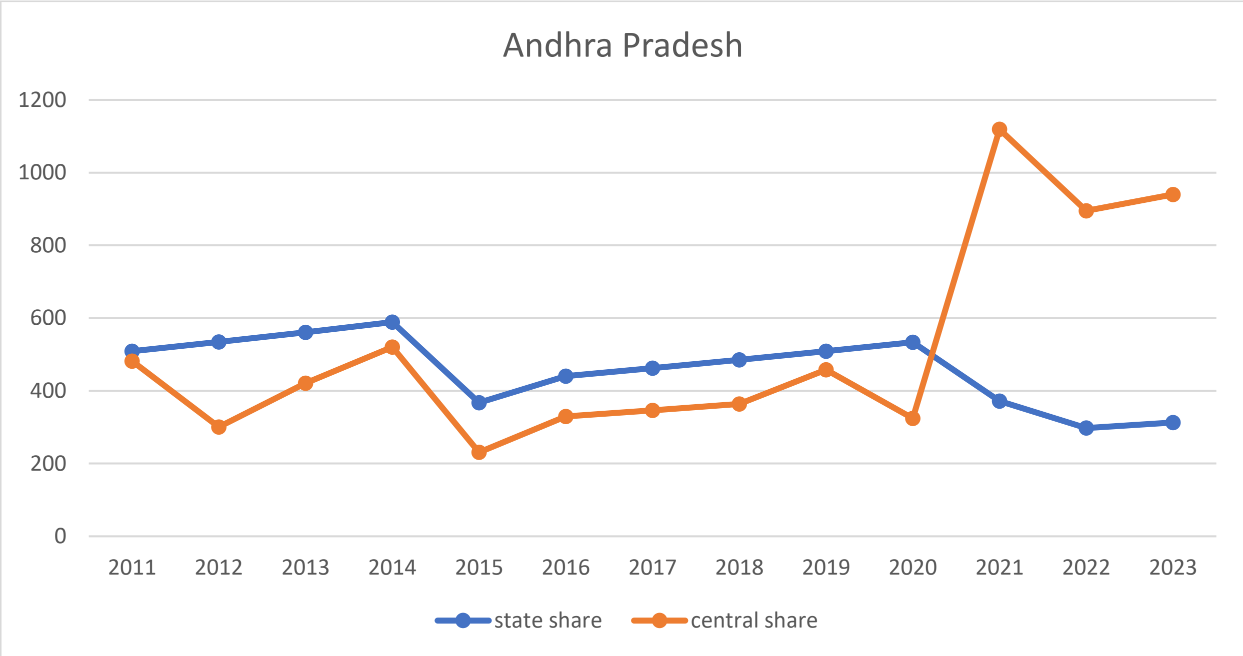


Fig. 5.4.2Bihar

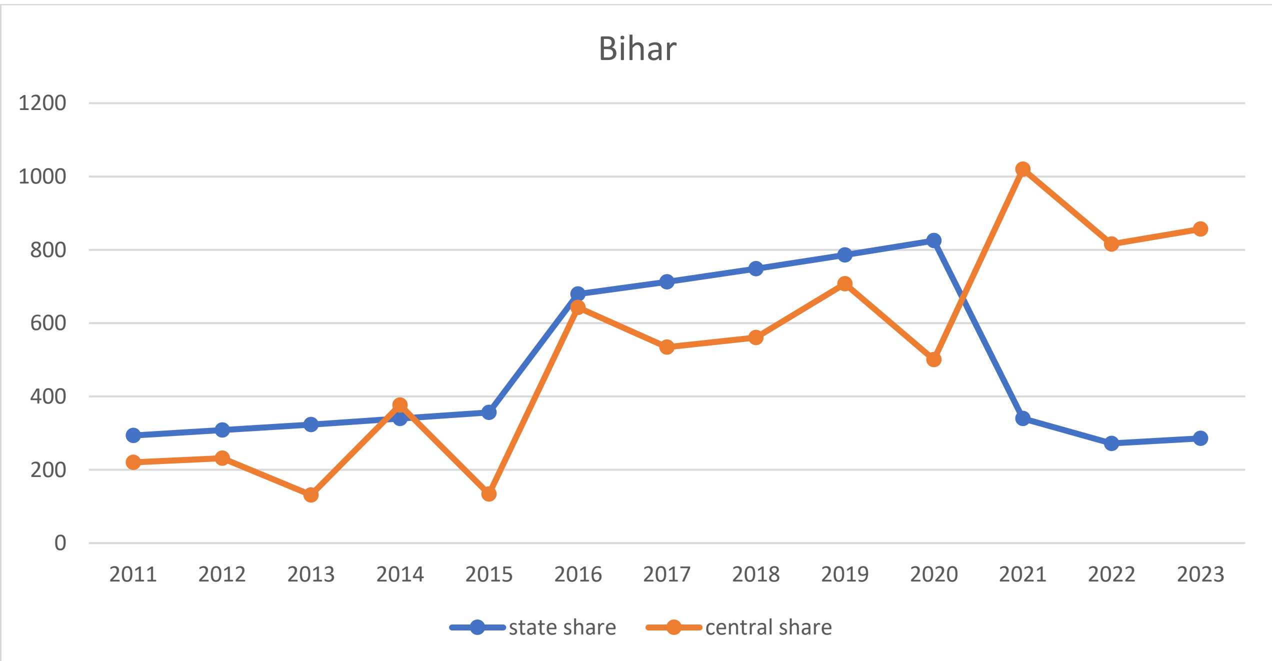


Fig 5.4.3 Madhyapradesh

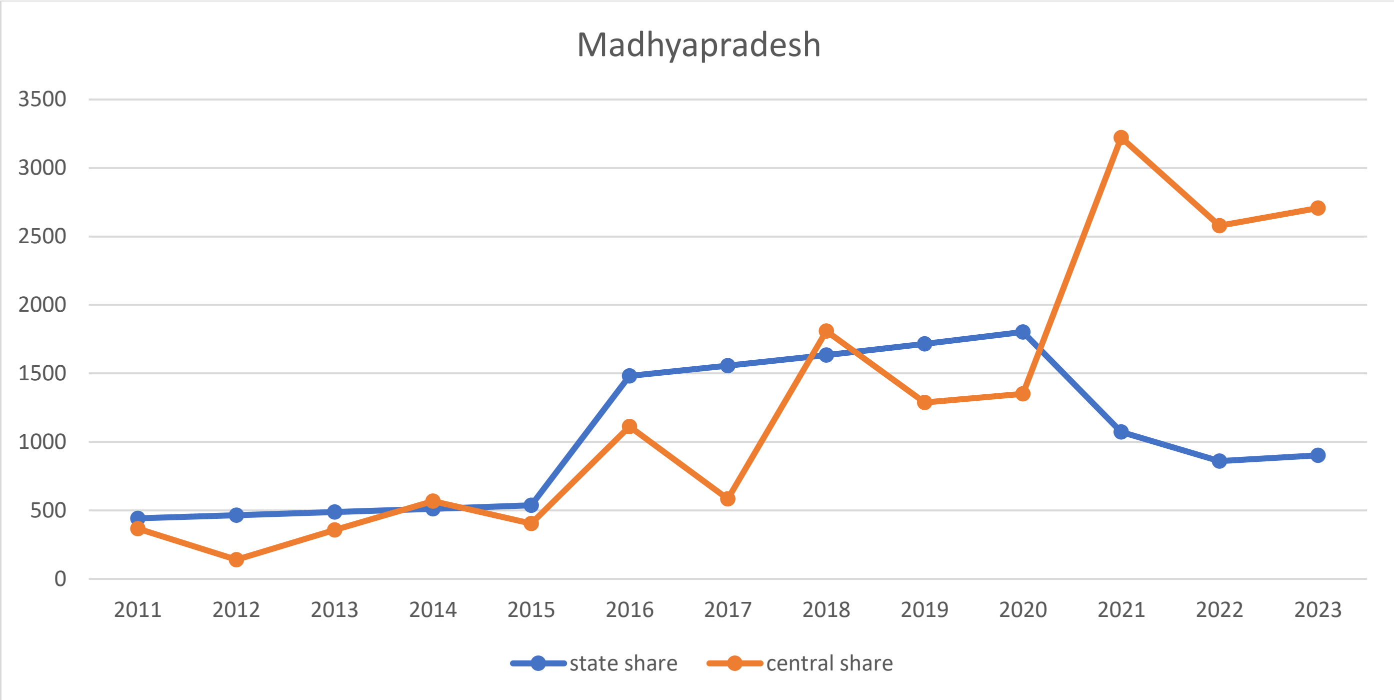
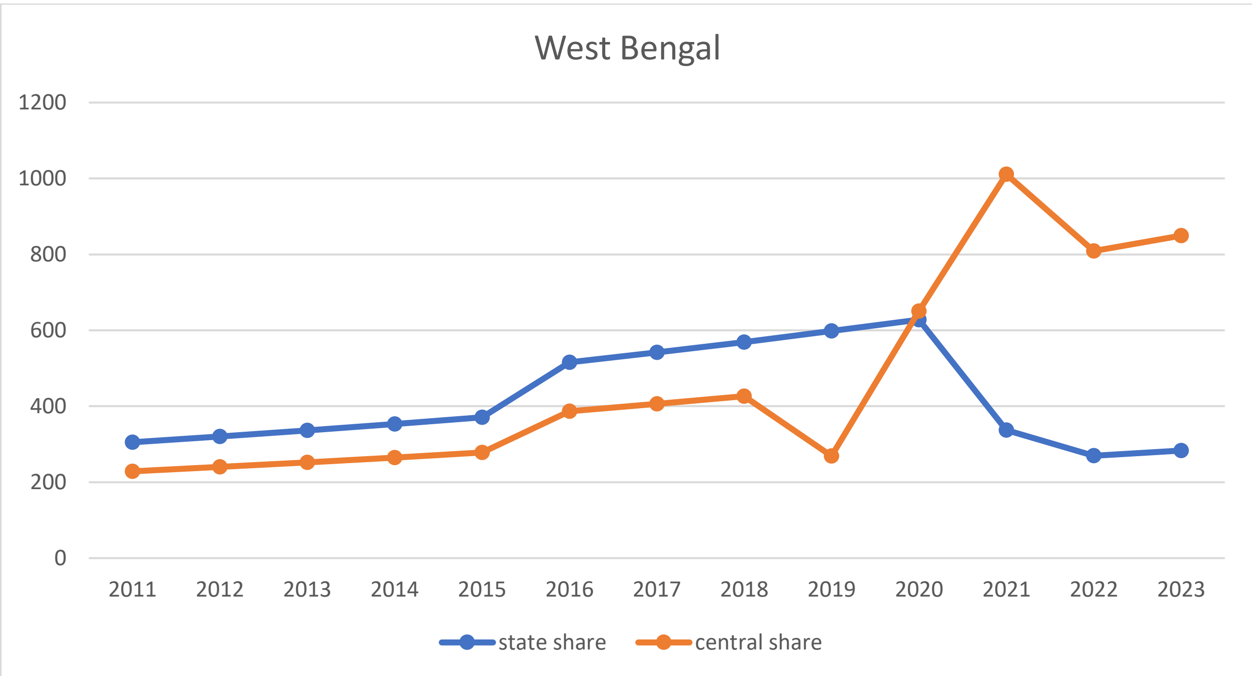


Fig5.4.4 West Bengal



Comment: Not much can be said except we can see trend for certain states and there is a sudden rise in central share around the year 2019-2020. This indicates the central allocated huge funding during the corona pandemic . Corona pandemic can be assumed as the irregular component in this analysis. No seasonality present. Overall it can be assumed none the time series model for any state is stationary.

step 3: checking for stationarity of the data

To check stationarity
Dickey Fuller Test

Dickey Fuller test is a statistical test that is used to check for stationarity in time series. This is a type of unit root test, through which we find if the time series is having any unit root.

Unit root is a feature of time series that indicates if there is any stochastic trend in the time series that drives it away from its mean value. Presence of unit root makes a time series non-stationary and as a result it leads to difficulties in deriving statistical inferences from the time series and future predictions.

Dickey Fuller test assumes a AR(1) type time series model and it is represented mathematically as,

$$y_t = \mu + \varphi_1 y_{t-1} + \varepsilon_t$$

After we substract yt-1 from both the side, we get:

$$\Delta y_t = \mu + \delta y_{t-1} + \varepsilon_t$$

where,
μ: Constant
φ: Co-efficient
yt-1: Value in the time series at lag of 1
Et: Error component

The test statistic formula is:

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

Augmented Dickey Fuller(ADF) Test

Augmented Dickey Fuller(ADF) test is an extension of Dickey Fuller test for more complex models than AR(1). The primary difference between the two tests is that the ADF is utilized for a larger sized set of time series models which can be more complicated.

Augmented Dickey Fuller test assumes a AR(p) type time series model and it is represented mathematically as,

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i \, y_{t-1} + \varepsilon_t$$

After we substract yt-1 from both the side, we get:

$$\Delta y_t = \mu + \delta y_{t-1} + \sum_{i=1}^{p} \beta_i \, \Delta y_{t-1} + \epsilon_t$$

ADF is the same equation as the DF with the only difference being the addition of differencing terms representing a larger time series.

The test statistic formula is:

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Assumptions

The test is conducted under following assumptions:

1. **Null Hypothesis (H0)**: There exists a unit root in the time series and it is non-stationary. ***Unit root = 1 or δ = 0***

2. **Alternate Hypothesis (H1)**: There exists no unit root in the time series and it is stationary. ***Unit root < 1 or δ < 0***

Condition to reject H0 and accept H1

If the ***test statistic is less than the critical value*** or if the ***p-value is less than a pre-specified significance level (e.g., 0.05)***, then the null hypothesis is rejected and the time series is considered ***stationary***.

If the test statistic is greater than the critical value, the null hypothesis cannot be rejected, and the time series is considered ***non-stationary***.

The critical value is found from the Dickey Fuller table (similar to t-table that we use for t-test, we have a table with critical values for Dickey Fuller test)

Results

We will use dickey fuller test for some state and print the p-values to find whether the time series is stationary or not.

Table 5.4.1 table showing the adf test p values

state	State share	Central share
Andhra Pradesh	0.4495	0.99
Bihar	0.9422	0.99
Gujarat	0.9635	0.9606
Madhya Pradesh	0.9434	0.99
Maharashtra	0.9316	0.9822
Odisha	0.9505	0.99
Tamil Nadu	0.9482	0.5261
Uttar Pradesh	0.9469	0.99
Uttarakhand	0.959	0.9786
West Bengal	0.9572	0.9856

Comment: We reject the null if p-value is <0.05. Clearly for all states the p-value>0.05 and none of the time series processes are stationary.

Step4: We fit the best models for 29 states using auto.arima() and hence forecast:

Note: We have recently collected the SDRF data of 2024. So we also check the errors and compare the forecasting with observed data.

Table 5.4.2. table showing the forecasted values of SDRF state Share 2024 using Univariate time series

state	SDRF predicted	SDRF observed	error
Andhra Pradesh	312.8	328	15.2
Arunachal Pradesh	28.50794	24.8	-3.70794
Assam	64.54316	76	11.45684
Bihar	396.8	416	19.2
Chhattisgarh	139.3082	127	-12.3082
Goa	3.512316	3.2	-0.31232
Gujarat	429.1213	388.8	-40.3213
Haryana	166.114	144	-22.114
Himachal Pradesh	34.33364	40	5.66636
Jharkhand	197.4236	166.4	-31.0236
Karnataka	221.6	232	10.4
Kerala	105.8558	92	-13.8558
Madhya Pradesh	484.3477	535.2	50.8523
Maharashtra	902.4	947.2	44.8
Manipur	3.510959	4	0.489041
Meghalaya	5.879444	6.4	0.520556
Mizoram	3.631959	4.8	1.168041
Nagaland	3.726901	4	0.273099
Odisha	473.649	471.2	-2.449
Punjab	129.8937	145.6	15.7063
Rajasthan	386.0188	435.2	49.1812
Sikkim	4.396841	4.8	0.403159
Tamil Nadu	263.8184	300	36.1816
Telangana	125.6	132	6.4
Tripura	37.03846	6.4	-30.6385
Uttar Pradesh	541.6	568	26.4
Uttarakhand	81.94894	92	10.05106
West Bengal	313.0195	297.6	-15.4195

Note: The observations for Jammu & Kashmir is absent in 2024 SDRF data ,so we have omitted it while model fitting & forecasting.

Table 5.4.3. table showing the forecasted values of SDRF central Share 2024 using Univariate time series

state	SDRF predicted	SDRF observed	error
Andhra Pradesh	940	987.2	47.2
Arunachal Pradesh	210.4	220.8	10.4
Assam	628.1271	680.8	52.6729
Bihar	1189.6	1248.8	59.2
Chhattisgarh	483.8088	380.8	-103.0088
Goa	9.6	9.6	0
Gujarat	1112	1168	56
Haryana	402.4662	433.6	31.1338
Himachal Pradesh	342.4	360.8	18.4
Jharkhand	476.8	500.8	24
Karnataka	664	697.6	33.6
Kerala	257.1296	277.6	20.4704
Madhya Pradesh	1528.8	1605.6	76.8
Maharashtra	2706.4	2841.6	135.2
Manipur	35.2	37.6	2.4
Meghalaya	54.4	58.4	4
Mizoram	39.2	41.6	2.4
Nagaland	34.4	36.8	2.4
Odisha	1348	1415.2	67.2
Punjab	416	436.8	20.8
Rajasthan	1244.8	1307.2	62.4
Sikkim	42.4	44.8	2.4
Tamil Nadu	831.6612	900	68.3388
Telangana	377.6	396	18.4
Tripura	55.47763	60.8	5.32237
Uttar Pradesh	1624	1705.6	81.6
Uttarakhand	787.2	826.4	39.2
West Bengal	849.6	892	42.4

Note: The observations for Jammu & Kashmir is absent in 2024 SDRF data ,so we have omitted it while model fitting & forecasting.

Comment: We can see overall errors are less and the forecasted values are close to the observed for most of the cases except some like Chhattisgarh,Odisha,Maharashtra etc. So we go bivariate time series to get more better forecasting since bivariate time series will take account of the correlation between the two data sets (SDRF state share and SDRF central share) ,if exists.

5.4.2 Bivariate time series using VAR(vector autoregressive) model

The vector autoregressive (VAR) model is a workhouse multivariate time series model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.

VAR models are characterized by their *order*, which refers to the number of earlier time periods the model will use. Continuing the above example, a 5th-order VAR would model each year's wheat price as a linear combination of the last five years of wheat prices. A *lag* is the value of a variable in a previous time period. So in general a *p*th-order VAR refers to a VAR model which includes lags for the last *p* time periods. A *p*th-order VAR is denoted "VAR(*p*)" and sometimes called "a VAR with *p* lags". A *p*th-order VAR model is written as

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t,$$

The variables of the form y_{t-i} indicate that variable's value *i* time periods earlier and are called the "i*th* lag" of y_t. The variable *c* is a *k*-vector of constants serving as the intercept of the model. A_i is a time-invariant (*k* × *k*)-matrix and e_t is a *k*-vector of error terms. The error terms must satisfy three conditions:

1. $\mathrm{E}(e_t) = 0$. Every error term has a mean of zero.
2. $\mathrm{E}(e_t e_t') = \Omega$ The contemporaneous covariance matrix of error terms is a *k* × *k* positive-semidefinite matrix denoted Ω.
3. $\mathrm{E}(e_t e_{t-k}') = 0$ for any non-zero *k*. There is no correlation across time. In particular, there is no serial correlation in individual error terms.[1]

The process of choosing the maximum lag *p* in the VAR model requires special attention because inference is dependent on correctness of the selected lag order.[2][3]

Table 5.4.4 The forecasted SDRF state share using VAR for 2024 is given as follows:

state	SDRF state share observed	SDRF state share forecasted	error
Andhra Pradesh	328	327.1744	0.8256
Arunachal Pradesh	24.8	27.44237	-2.64237
Assam	76	89.3588	-13.3588
Bihar	416	433.2009	-17.2009
Chhattisgarh	127	140.5525	-13.5525
Goa	3.2	3.991417	-0.79142
Gujarat	388.8	418.0767	-29.2767
Haryana	144	123.8707	20.1293
Himachal Pradesh	40	36.3116	3.6884
Jharkhand	166.4	207.0528	-40.6528
Karnataka	232	243.6302	-11.6302
Kerala	92	93.13062	-1.13062
Madhya Pradesh	535.2	512.7026	22.4974
Maharashtra	947.2	978.216	-31.016
Manipur	4	3.525181	0.474819
Meghalaya	6.4	3.10103	3.29897
Mizoram	4.8	6.435243	-1.63524
Nagaland	4	3.869455	0.130545
Odisha	471.2	431.9485	39.2515
Punjab	145.6	114.4376	31.1624
Rajasthan	435.2	469.0038	-33.8038
Sikkim	4.8	4.425299	0.374701
Tamil Nadu	300	279.5414	20.4586
Telangana	132	101.6309	30.3691
Tripura	6.4	11.19773	-4.79773
Uttar Pradesh	568	536.2882	31.7118
Uttarakhand	92	102.4633	-10.4633
West Bengal	297.6	281.3147	16.2853

Table 4.5.5 The forecasted SDRF central share using VAR for 2024 is given as follows:

state	SDRF central share observed	SDRF central share forecasted	error
Andhra Pradesh	987.2	918.0885	69.1115
Arunachal Pradesh	220.8	212.334	8.466
Assam	680.8	702.6123	-21.8123
Bihar	1248.8	1329.5483	-80.7483
Chhattisgarh	380.8	319.3916	61.4084
Goa	9.6	9.996663	-0.39666
Gujarat	1168	1221.416	-53.416
Haryana	433.6	412.0805	21.5195
Himachal Pradesh	360.8	369.9809	-9.1809
Jharkhand	500.8	467.2192	33.5808
Karnataka	697.6	628.5106	69.0894
Kerala	277.6	207.0521	70.5479
Madhya Pradesh	1605.6	1576.173	29.427
Maharashtra	2841.6	2894.173	-52.573
Manipur	37.6	29.64423	7.95577
Meghalaya	58.4	54.77315	3.62685
Mizoram	41.6	41.71454	-0.11454
Nagaland	36.8	29.27142	7.52858
Odisha	1415.2	1351.2	64
Punjab	436.8	370.8643	65.9357
Rajasthan	1307.2	1264.482	42.718
Sikkim	44.8	39.90876	4.89124
Tamil Nadu	900	822.9605	77.0395
Telangana	396	422.3875	-26.3875
Tripura	60.8	58.45076	2.34924
Uttar Pradesh	1705.6	1644.355	61.245
Uttarakhand	826.4	808.2952	18.1048
West Bengal	892	878.9571	13.0429

Comment: Forecasted values by VAR gives much closer results to 2024 SDRF thatn univariate time series forecasting with significance decrease in errors

A chart for comparison of errors:

Table 5.4.6 table showing SDRF state share errors

state	SDRF state share errors predicted by univariate time series	SDRF state share errors predicted by VAR
Andhra Pradesh	15.2	0.8256
Arunachal Pradesh	-3.70794	-2.64237
Assam	11.45684	-9.3588
Bihar	19.2	-17.2009
Chhattisgarh	-12.3082	-13.5525
Goa	-0.31232	-0.19142
Gujarat	-40.3213	-29.2767
Haryana	-22.114	20.1293
Himachal Pradesh	5.66636	3.6884
Jharkhand	-31.0236	-40.6528
Karnataka	10.4	-6.6302
Kerala	-13.8558	-1.13062
Madhya Pradesh	50.8523	22.4974
Maharashtra	44.8	-31.016
Manipur	0.489041	0.474819
Meghalaya	0.520556	3.29897
Mizoram	1.168041	-1.63524
Nagaland	0.273099	0.130545
Odisha	-2.449	9.2515
Punjab	15.7063	-8.8376
Rajasthan	49.1812	-33.8038
Sikkim	0.403159	0.374701
Tamil Nadu	36.1816	20.4586
Telangana	6.4	30.3691
Tripura	-30.6385	-4.79773
Uttar Pradesh	26.4	31.7118
Uttarakhand	10.05106	-10.4633
West Bengal	-15.4195	16.2853

Comment: The combined MSE for SDRF state share using Univariate Time series is 624.1676 and the combined MSE for SDRF state share using VAR is 238.3536. There is a significance decrease in MSE. The predicted model using VAR is giving better forecasted values for 2024 data.

Table 5.4.7 table showing SDRF state share errors

state	SDRF central share errors predicted by univariate time series	SDRF central share errors predicted by VAR
Andhra Pradesh	15.2	0.8256
Arunachal Pradesh	-3.70794	-2.64237
Assam	11.45684	-9.3588
Bihar	19.2	-17.2009
Chhattisgarh	-12.3082	-13.5525
Goa	-0.31232	-0.19142
Gujarat	-40.3213	-29.2767
Haryana	-22.114	20.1293
Himachal Pradesh	5.66636	3.6884
Jharkhand	-31.0236	-40.6528
Karnataka	10.4	-6.6302
Kerala	-13.8558	-1.13062
Madhya Pradesh	50.8523	22.4974
Maharashtra	44.8	-31.016
Manipur	0.489041	0.474819
Meghalaya	0.520556	3.29897
Mizoram	1.168041	-1.63524
Nagaland	0.273099	0.130545
Odisha	-2.449	9.2515
Punjab	15.7063	-8.8376
Rajasthan	49.1812	-33.8038
Sikkim	0.403159	0.374701
Tamil Nadu	36.1816	20.4586
Telangana	6.4	30.3691
Tripura	-30.6385	-4.79773
Uttar Pradesh	26.4	31.7118
Uttarakhand	10.05106	-10.4633
West Bengal	-15.4195	16.2853

Comment: The combined MSE for SDRFcentral share using Univariate Time series is 27740.77and the combined MSE for SDRF central share using VAR is 8468.891
. There is a significance decrease in MSE. The predicted model using VAR is giving better forecasted values for 2024 data.

Conclusion

From the project we found that

- From the clustering we see the SDRF central share and state share have high positive association. In other words ,the central and states allocates high SDRF for higher disaster-prone states and lower SDRF for less disaster-prone state. The allocation is fair. We confirm this by using Rank correlation which comes out to be positive indicating high association between the two shares.
- Using regression it can be said that the length of coastline and area of each state also have significance influence on the allocation of SDRF. Geometrically the coastline plays a significant role in the natural disaster as the coastline area is more prone to cyclone , hurricane, flood etc. Statistically it is also evident that the states with coastal region are allocated with more SDRF.
- We also forecasted the SDRF (both state and central share) for year 2024. The VAR model gives better values than the forecasted values by univariate time series model by reducing the MSE significantly due to the consideration of correlation between the 2 sets of data(state share and central share).
- Another interesting find is that central allocated huge SDRF share during the corona pandemic in 2019-2020.

References:

1. T. Hastie, R. Tibshirani & J. Friedman : The Elements of Statistical Learning
2. R.A. Johnson & D.W. Wichern : Applied Multivariate Statistical Analysis
3. C. Chatfield : The Analysis of Time Series – An Introduction
4. G.E.P. Box ,G.M. Jenkins & G.C.Reinsel : Time Series Analysis – Forecasting & Control
5. P.J. Brockwell & R.A. Davis : Introduction to Time Series Analysis and Forecasting
6. A.Pankratz : Forecasting with Univariate Box-Jenkins Model
7. G. Janacek and L. Swift : Time Series –Forecasting, Simulation, Applications