Time series analysis and Data Classification using Cluster Analysis (CA) and Principal Component Analysis (PCA) on Precipitation in Africa

Stephen Kiilu
African Institute for Mathematical Sciences (AIMS) Rwanda
Research Methods for Climate

Contents

1 Introduction

When we talk about data classification the first thing which comes to our mind are the two popular methods in statistical data classification, namely Cluster Analysis (CA) and Principal Component Analysis (PCA). What is Cluster Analysis (CA)? What is Principal Component Analysis (PCA)? To answer these two questions we need to understand the concept of statistical machine learning. Most statistical learning problems fall into two categories namely; supervised and unsupervised learning. In supervised statistical learning, the response variable is provided for analysis and our main task is to exploit the relationship between the response and predictors and also to make predictions e.g normal regression (simple and multiple regression) and logistic regression. Unsupervised machine learning the response is not provided for our analysis i.e the response is not measured and our main task is to explore the relationship between the variables. Cluster Analysis (CA) is an unsupervised statistical learning method that is used to determine whether observations fall into two relatively distinct groups. PCA is a tool for unsupervised machine learning that is popular for reducing data dimension while retaining key information from a large set of variables. Then what is time series? Why are we interested in time series analysis? Time series can be described as a collection of random variables that are indexed in the order they we collected in time. The main objective of time series is to develop a statistical model that will help us describe this kind of data that fluctuates randomly over time.

2 Aim of the study

The aim of this study is to explore data classification methods and time series analysis. We are interested in studying how CA and PCA algorithms work and how the two methods compare using precipitation data for ten cities in Africa. We are also interested in investigating the time-series pattern of the first two principal factors from PCA.

3 Study question

- What are the characteristics of the three (CA) algorithms?
- How is the grouping using the three CA algorithms?
- What are the characteristics of (PCA)?
- How do CA and PCA compare?
- Do the Principal factors show any trend?
- Do the Principal factors show any pattern?

4 Data

The data used in the analysis has 11 columns representing the 10 cities in Africa and the years our key climate variable i.e precipitation was measured. The years run from 1960 to 2010 and our main variables in the study are:

- Year

. Yaounde

. Niamey

. Nalohou

. Njombe Penja

. Ouagadougou

. Rubavu

. Saint Louis

. Suyani

. Tamale

. Abuja

For the time analysis, we are going to use the score of the two principal factors from PCA.

5 Methodology

5.1 Cluster Analysis

CA is a statistical learning tool that is employed to ascertain if observations fall into relatively distinct groups. CA fall into two basic categories.
Category one, clusters are predefined by the user before the analysis procedure. In the second category, the clusters are determined by the user after the analysis procedure.
In this study, we focus on three algorithms;

- **Single linkage** start by linking the two closest data points, continue to the next two closest data points and continue with the same procedure until you create a dendrogram as a function of the distance between the data points.
- **Average linkage** is an improved procedure of single linkage which uses the mean group points when clustering, and the proceeds in the same as single linkage.
- **Wards algorithm** is based on the analysis of variance. Data points are joined to a new group based on if their inclusion in the group improves the variance of the group or not compared to the inclusion in some other group.

5.2 Principal Component Analysis (PCA)

PCA is a statistical learning tool for deriving low dimensional dataset from a large set of variables while maintaining the main information. CPA falls into categories based on the distribution of variance in the principal factors.

- Rotated PCA - variance is distributed in the first two principal factors.
- Unrotated PCA - emphasis on variance is put in the first principal factor.

Practically, rotated PCA is preferred because it has more physical interpretation than unrotated PCA.

5.3 Time series Analysis

Time series refers to a sequence of measurements that follow a non-random pattern measured at equally spaced time series. Our main objective in time series analysis is to identify patterns in the data and make predictions. We are going to focus on:

- **Identifying patterns** - we focus on identifying systematic and random noise from our time series data.
- **Trend analysis** - we remove trend from the time series through smoothing to only study the signal in the data.
- **Analysis of seasonality** - we detrend our time series data because we are only not interested in the extremes but only in studying signal in the data.

6 Results and Discussion

6.1 Cluster Analysis (CA)

In this section, we discuss our findings from our CA using the three algorithms and compare their sensitivity. We note that classification is subjective and grouping is determined by the interest of the one doing data analysis and interpretation. One big appealing characteristic of this hierarchical clustering is that one dendrogram can be used to obtain any clusters or groups. In the three CA algorithms, we choose an eye-appealing number of clusters based on the height of fusion and the cluster we desire to have.

(i) **Single linkage method**
The figure 1 below shows single linkage clustering, where we choose to have four groups and cut the dendrogram horizontally at height 19. At this height, it results in two distinct clusters and two outliers. The cities Saint Louis, Njombe Penja, Nalohou, Tamale, Niamey, and Ouagadougou are placed in the same cluster. The cities Sunyani and Abuja are in another distinct cluster, while Rubavu and Yaounde are clustered as outliers. That is;

. Saint Louis, Njombe Penja, Nalohou, Tamale, Niamey, and Ouagadougou
. Suyani and Abuja
. Rubavu
. Yaounde

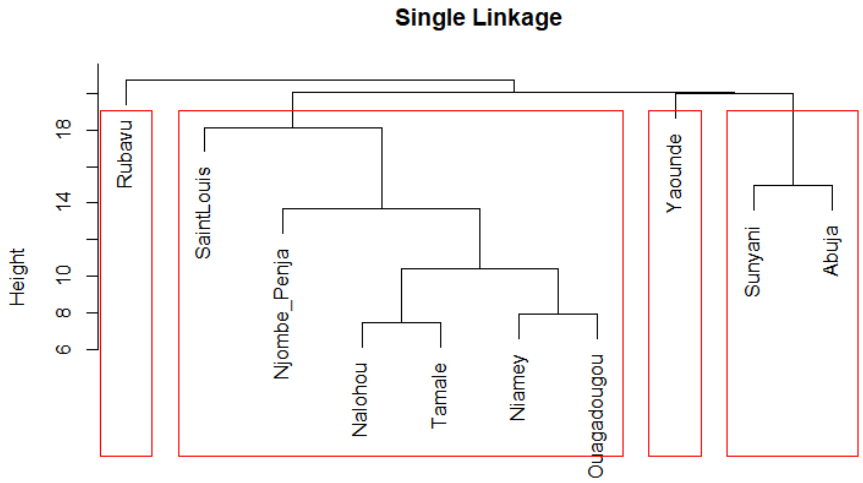

Figure 1: single linkage

- Saint Louis, Njombe Penja, Nalohou, Tamale, Niamey, and Ouagadougou - are similar in terms of precipitation
- Suyani and Abuja - show similar observation in precipitation
- Rubavun - are very different from the above two groups and cannot be grouped into either cluster
- Yaounde - is also an outlier city in terms of precipitation.

In cluster analysis observations that fuse at the very bottom of the cluster, a tree is very similar, while the observations which fuse at the very top of the cluster tree are very different. We can say Nalohou and Tamale are very similar, as compared to Saint Louis and Njombe Penja in terms of precipitation.

(ii) **Average linkage**
This algorithm clusters data points just like a single linkage but it uses group mean data points distance to cluster two data points. We make a horizontal cut across the height of 20 and obtain clusters as shown in the figure 2 below.

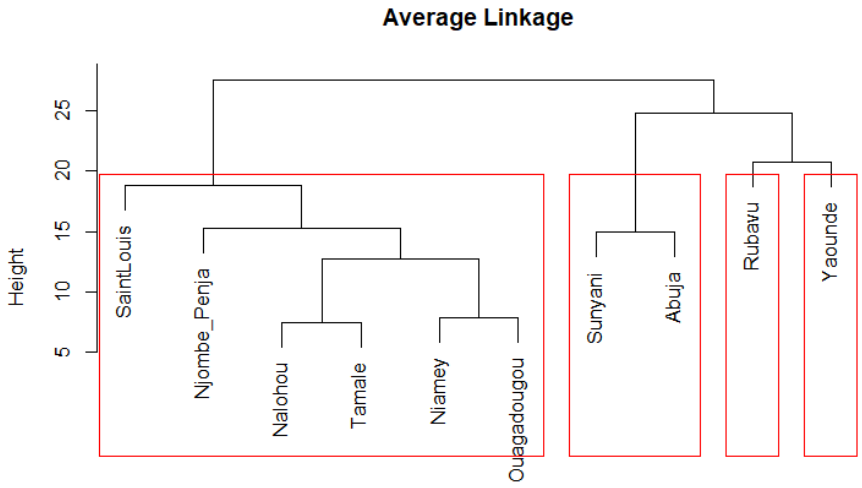

Figure 2: average linkage

Just like single linkage the cities are classified together in similar way. That is;

- Saint Louis, Njombe Penja, Nalohou, Tamale, Niamey, and Ouagadougou
- Suyani and Abuja
- Rubavu
- Yaounde

Where Yaounde and Rubavu are outliers.

(iii) **Ward's algorithm**
In this algorithm, we cut the dendrogram at height 22 and obtain four different groups as shown in the figure 3 below;

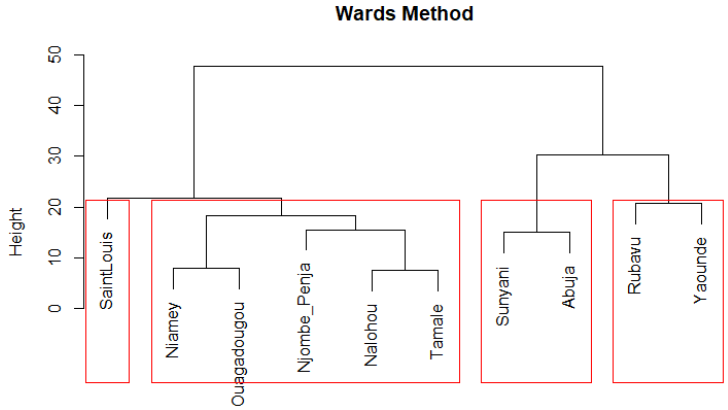

Figure 3: ward's algorithm

- Njombe Penja, Nalohou, Tamale, Niamey, and Ouagadougou - are similar and are placed in the same cluster
- Suyani and Abuja - are also clustered together
- Rubavu and Yaounde are also put in the same cluster.
- Saint Louis is an outlier after the classification, which means it very different from the other cities in precipitation, and it can't be placed in any of the above clusters.

(iv) **Map showing different cities after CA**
The following figure 4 show the location of different cities after the classification using Ward's algorithm. The cities with the same key mean that they are clustered together.

Figure 4: Map showing location of the 10 cities
source: Google map

Single linkage and average linkage have done a similar job in clustering the cities and have identified Rubavu and Yaounde as outliers. Ward's algorithm has clustered the cities differently by grouping them in three distinct clusters with Saint Louis as the only outlier. This is because single linkage and average linkage tend to identify outliers while ward's algorithm outliers are not well identified.

6.2 Principal Component Analysis (PCA)

The focus of this task is to study how PCA works and the sensitivity of the two methods in PCA, namely; rotated and unrotated PCA. Before we proceed, we need to summarize our data for to understand it better and at same time draw some crucial descriptive statistics.

Summary of the data

Nalohou	Niamey	Njombe_Penja	Ouagadougou
Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : 0.00
1st Qu.: 3.75	1st Qu.: 0.00	1st Qu.: 59.0	1st Qu.: 0.00
Median : 82.40	Median : 9.25	Median : 206.5	Median : 18.35
Mean : 99.74	Mean : 49.04	Mean : 211.7	Mean : 63.22
3rd Qu.:171.03	3rd Qu.: 84.25	3rd Qu.: 321.0	3rd Qu.:118.28
Max. :391.10	Max. :327.80	Max. :1028.0	Max. :362.00

Rubavu	SaintLouis	Sunyani	Tamale
Min. : 5.80	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 80.28	1st Qu.: 0.00	1st Qu.: 37.62	1st Qu.: 5.70
Median :121.70	Median : 0.20	Median :102.45	Median : 82.55
Mean :118.64	Mean : 22.76	Mean :102.08	Mean : 91.09
3rd Qu.:153.80	3rd Qu.: 23.88	3rd Qu.:153.53	3rd Qu.:151.62
Max. :281.60	Max. :205.70	Max. :302.60	Max. :382.00

Yaounde	Abuja
Min. : 0.00	Min. : 0.00
1st Qu.: 67.03	1st Qu.: 32.60
Median :155.25	Median : 71.55
Mean :155.96	Mean : 85.29
3rd Qu.:220.15	3rd Qu.:123.08
Max. :513.10	Max. :463.60

Eligibility of PCA

PCA mainly relies on covariance matrix and correlation, and it important before we do the PCA procedure that we determine if there is a correlation between the variables. Figure 5 below shows the correlation plot of the 10 cities in our data set.

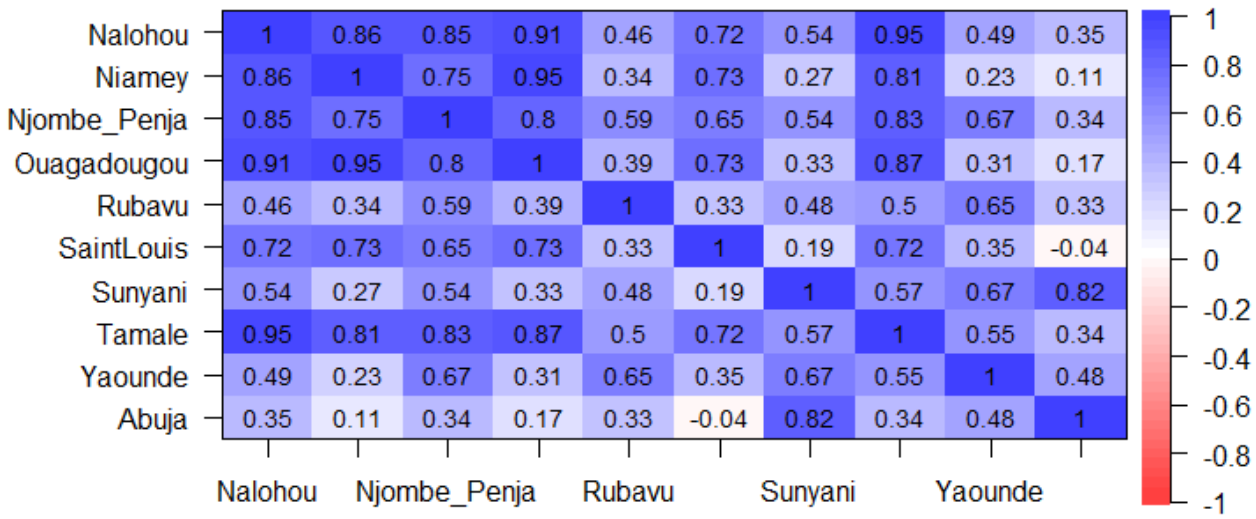

Figure 5: corplot

We see most of the variables in our cities dataset have some reasonable correlation and therefore CPA can be a good approach in dimensionality reduction.

(i) **Selection of most important Principal components**
We use a screen plot to select the most important Principal Components (PC), as shown in the figure 6 below.

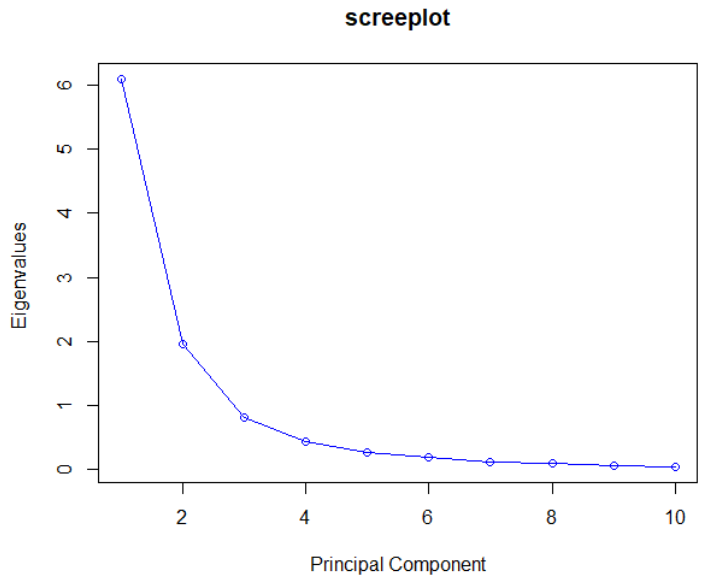

Figure 6: screen plot

This is a plot of the Eigenvalues against Principal Components. The Eigenvalues show the amount of variance explained by each Principal Components, while Principal Components themselves are just linear combinations of our initial variable in the data set. By default, we use Principal components above Eigenvalues 1 (they explain most of the variance in the data set). From our screen plot, we select the first two Principal Components, i.e PC1 and PC2.

(ii) **Percentage of explained variance by the Principal components**
Figure 7 below shows a graph of the cumulative percentage of variance explained by the Principal Components. The Principal Components are constructed in a way that the first principal component accounts for the largest possible variance possible, the second Principal Component accounts for the next variance possible, and so on until we have calculated the amount of variance explained by all Principal Components equal to the number of variables. From the above PVE graph the PC1 accounts for 61% of the variance in the data set, which means it accounts for at least 61% of the information in the data set. PC1 and PC2 account for about 81% of variance in the data set, and PC1, PC2, and PC3 accounting for about 90% of all variance in the data set, which implies that other Principal Components only account for less than 10% of the information in the data set. This means that we can decide to maintain the first 3 Principal Components i.e PC1, PC2, and PC3 and have at least 90% of all important information retained and drop other Principal Components which are almost insignificant (they account for very little information). In PCA, this is the main idea and first step towards dimensionality reduction.

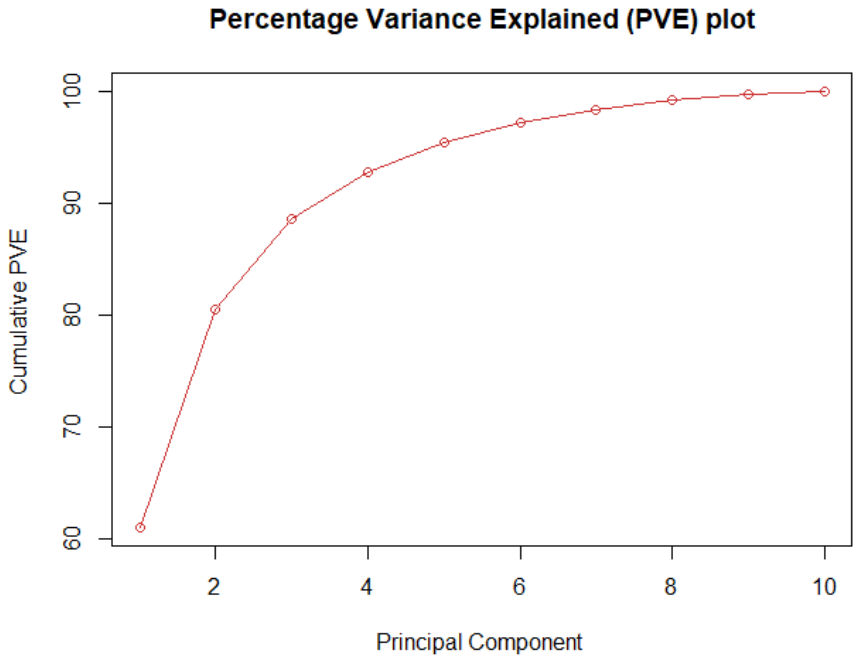

Figure 7: Percentage of explained variance by each PC

(iii) **Explained variation (Unrotated PCA)**
The table 1 below summarizes the loadings and the amount of variation explained by

each Principal Component under Unrotated PCA.

	PC1	PC2	PC3
Loadings	6.10	1.95	0.81
Proportion Var	0.61	0.2	0.08
Cumulative Var	0.61	0.81	0.89
Proportion Explained	0.69	0.22	0.09
Cumulative Proportion	0.69	0.91	1.00

Table 1: Unrotated PCA

We study a total of 3 Principal Components. From results in the above table; PC1 explains 61% of the total variation, PC2 about 20% of the total variation and PC3 explains about 8% of the total variation. We see unrotated PCA puts more emphasis on the variance explained by the first Principal Component.

(iv) **Explained variation (Rotated PCA)**
The table 2 below is a summary of the loadings and the amount of variation explained by each Principal Component under Rotated PCA.

	PC1	PC2	PC3
Loadings	4.74	2.10	2.03
Proportion Var	0.47	0.21	0.20
Cumulative Var	0.47	0.68	0.89
Proportion Explained	0.53	0.24	0.23
Cumulative Proportion	0.53	0.77	1

Table 2: Rotated PCA

We investigate a total of 3 Principal Components. From results in the above table; PC1 explains 47% of the total variation, PC2 about 21% of the total variation and PC3 explains about 20% of the total variation. We see in rotated PCA distributes total variance explained among the 3 Principal Components.

(v) **Component Loadings (Rotated PCA)**
We focus on Rotated PCA because it has a more practical physical interpretation of results than Unrotated.

	RC1	RC2	RC3
Nalohou	0.90	0.34	0.19
Niamey	0.96	0.05	0.13
Njonge-Peja	0.74	0.25	0.49
Ouagadougou	0.96	0.10	0.11
Rubavu	0.24	0.15	0.81
SaintLouis	0.82	-0.16	0.11
Sunyani	0.22	0.85	0.36
Tamale	0.86	0.30	0.18
Yaonde	0.20	0.40	0.81
Abuja	0.04	0.95	0.15

Table 3: Rotated PCA

From the table 3 above, the loadings in colour red refers to the cities grouped together using PCA method. The cities Nalohou, Niamey, Njonge - Peja, Ouagadougou, Saint Louis, and Tamale are grouped, Suyani and Abuja are classified together and also Rubavu and Yaounde are in same group.

(vi) **Map showing different cities after PCA**
The figure 8 below shows how the cities are grouped using PCA. The cities with the same key mean that they are grouped.

Figure 8: Map showing location of the 10 cities
Source:Google map

(vii) **Scores plot (Rotated PCA)**
We compare the scores plot for the Rotated Component 1 (RC1) starting from the year 1960 to 2010. When we talk of scores we are concerned about time series for a particular period. We ask the question, which years were our climate variable i.e precipitation most active? Which years were in negative mode? The figure 9 below shows the time series for RC1 from the year 1960 to 2010.

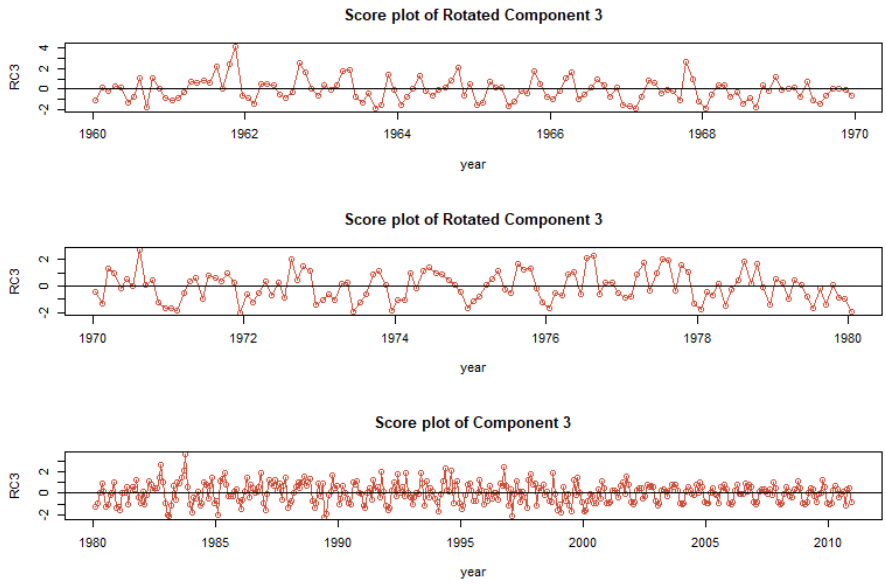

Figure 9: Time series

Generally, we can see a seasonal pattern in the time series, for example, in the years 1962, 1967, 1971, 1975, and 1995 are in positive mode, meaning we're more active in terms of precipitation. The years, 1970, 1977, 1984, and 2005, for instance, are in negative mode.

6.3 Time series Analysis

In this subsection, we are going to plot time series and check if there is a trend and investigate for patterns using autocorrelation and cross-correlation.

(i) **Time series analysis for PC1 and PC2**
The following figure 10 shows a time series plot for Principal Component one (PC1) and Principal Component two (PC2) with a linear line of best fit. The upper panel of figure 10 shows the time series for Principal Component 1 and the lower panel of figure 10 shows the time series for PC2. The fitted line for PC1 and PC2 is almost stationary, i.e it is not continuously increasing over time, which means there is a small or no trend at all.

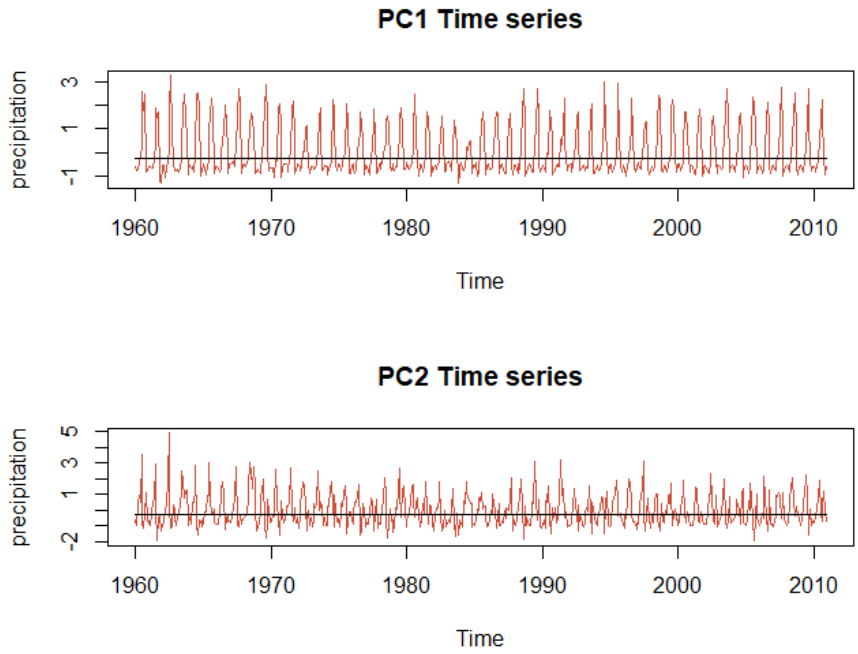

Figure 10: Time series plot for PC1 and PC2

(ii) **Time analysis for detrended PC1 and PC2**
Figure 11 below compares time series for PC1 and detrended PC1. Ideally, when we are dealing with time series, it is very necessary to have time-series data that is stationary so that averaging of time series values over some time is a sensible thing to do. In time series, we are very interested in studying the dependencies between values, and it is very hard to measure this dependency when the dependencies are changing over every time point. One of the ways to make time-series data stationary is by detrending. Figure 11 below compares time series for PC1 original data (upper panel) and detrended PC1. The detrended data has no trend, meaning the mean value is stationary over some time. We can use the detrended PC1 time series data to carry out statistical time series analysis, e.g autocorrelation, and cross-correlation.

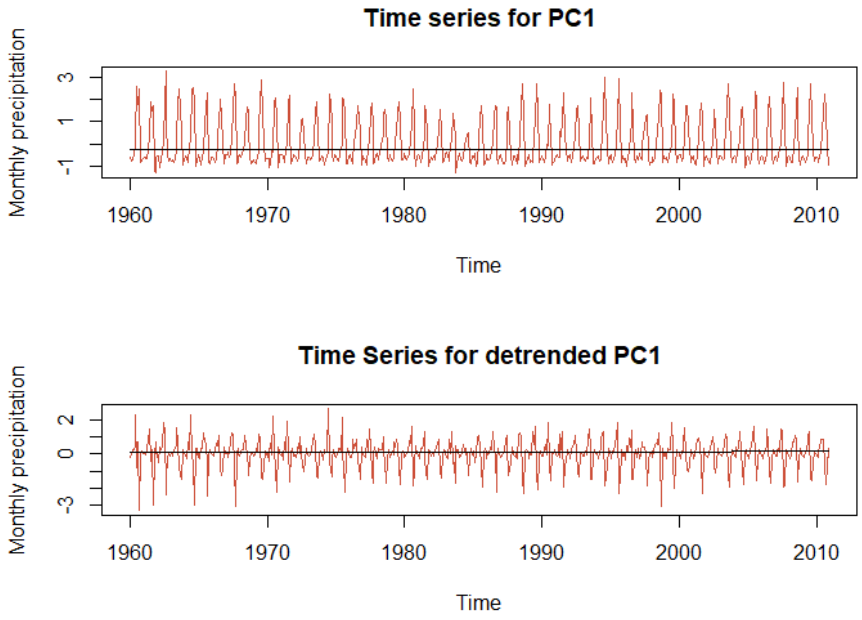

Figure 11: Time series plot for PC1 and detrended PC1

We now compare time series for PC2 and detrended PC2 as shown in figure 12 below. The time series is similar to one in figure 11. The time series plot on the lower panel in figure 12 shows a detrended data with a stationary fitted line, which implies there is no trend in the time series.

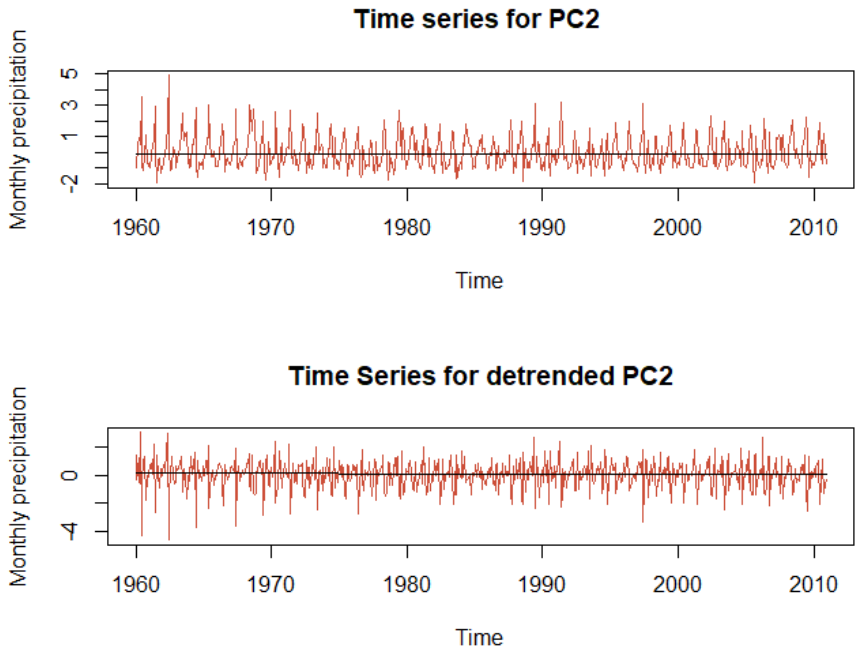

Figure 12: Time series plot for PC2 and detrended PC2

(iii) **Autocorrelation analysis for PC1 and PC2**

The figure 13 below shows a correlogram for Principal component 1 and Principal Component 2. Autocorrelation refers to correlating two-time series among themselves. The upper panel of figure 13 shows a correlogram for detrended PC1, with a positive correlation close to 0.8. The lower panel of figure 13 shows a correlogram for detrended PC2, with a positive correlation close to 0.5. Both PC1 and PC2 in figure 13 show sinusoidal autocorrelation meaning there is a signal in the time series which keeps repeating from time to time. For both PC1 and PC2, there is a pattern repeating after lag 12 i.e at the highest correlation, which implies a signal repeating after every 12 months, i.e annual cycle.

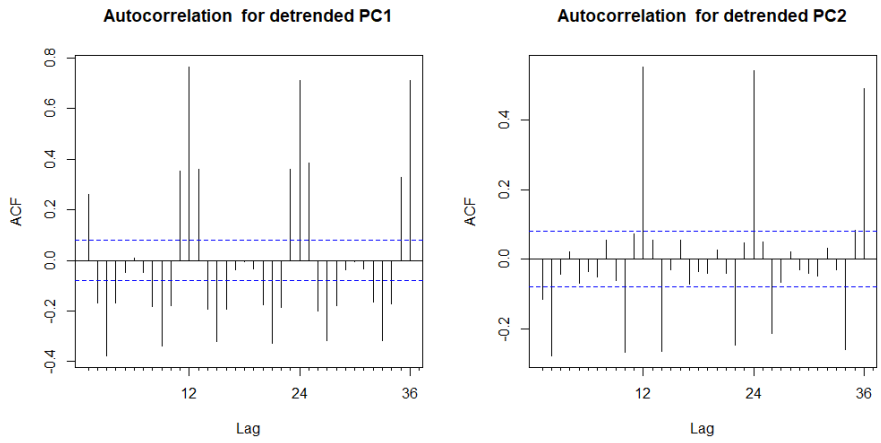

Figure 13: correlogram for PC1 and PC2

(iv) **Cross-correlation analysis of PC1 and PC2**

The figure 14 below shows a plot of cross correlation between PC1 and PC2.

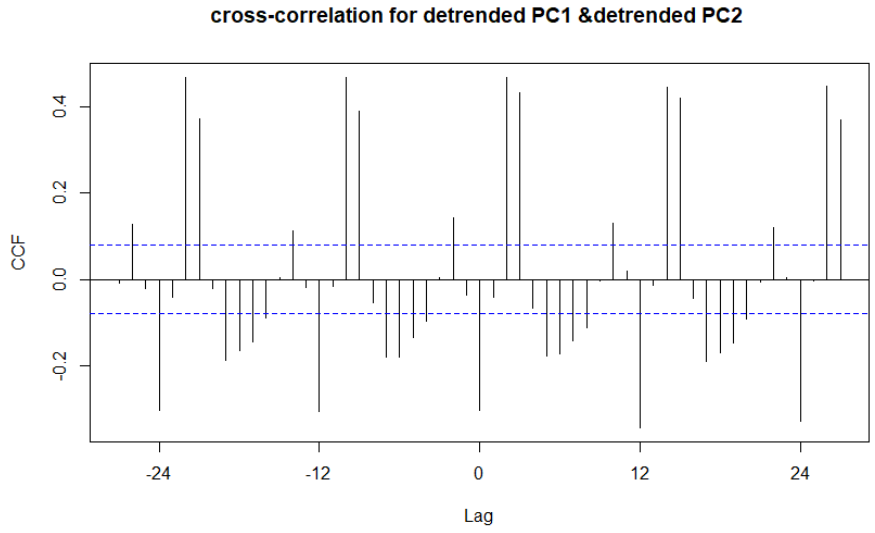

Figure 14: cross-correlation of PC1 and PC2

Cross-correlation is the correlation between two-time series. The figure 14 shows autocorrelation between PC1 and PC2. The PC1 and PC2 have a correlation of 0.5, which means that they have a positive linear relationship, i.e as PC1 is increasing also PC2 is increasing. The cross-correlation also shows a pattern that is occurring after every 12 months i.e annual cycle. The cross-correlation bars are also extending beyond the threshold dotted line, meaning the correlations between PC1 and PC2 are statistically significant, further confirming a pattern in the time series.

7 Conclusion

Comparing my results from PCA and CA, PCA is more preferred to CA because it gives more understanding of the dataset and same time quantifies the classification. PCA also reduces data dimensionality while retaining on most important information from the entire data set. Ward's algorithm is also preferred because grouping is clearly defined. PCA method reduced the dimension of the data set to 3 Principal Components while retaining the most important information, the first 3 Principal Components i.e PC1, PC2, and PC3 accounted for close to 89% of variance in the data set. Wards algorithm in the same way came up with 3 groups of cities and one outlier. This essentially means the two classification methods CA and PCA did the same job. To sum up, PCA is most preferred because it does not only reduce the dimensionality of the data set but also quantifies the amount of variance (information) explained by each PCA. Whereas data classification methods like PCA and CA are tools for data generalization, time series tends to study for patterns in non-random data obtained from equal time intervals and use the patterns to do predictions. Both PCA, CA, and time series are powerful tools in identifying patterns in data sets but time series analysis performs better when it comes to making a prediction based on the observed patterns.

References

[1] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani Introduction to Statistical Learning with Applications in R 2013.

[2] Babatunde Abiodun *Lecture notes,* 2021.

[3] Robert H. Shumway David S. Stoffer Time Series Analysis and Its Applications With R, Fourth Edition 1999, 2012, 2016, 2017.