AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Kamau Gladys Muthoni	Assignment Number: 1
Course: Research Methods in Climate Science	Date: January 18, 2020

1 Task 1: Cluster Analysis

2 Introduction

The data choosen for this study was extracted from Climate Dataset (CRU) over Turkana City in Kenya with coordinates 3.3122° N, 35.5658° E. The study is based on the month of April which was recorded with the highest amount of precipitation over the span of 51 years from 1960 to 2010. The aim of the study is to identify the characteristics of three Cluster Analysis algorithms. CA groups climate variables into groups of similar type. The importance of this study is to learn how different methods of hierarchical cluster analysis work.

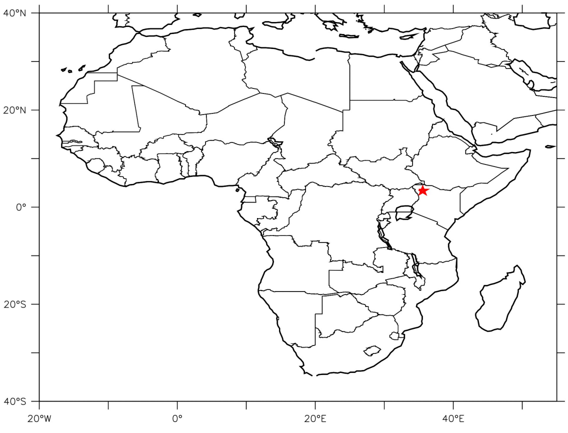


Figure 1

There's no clear history of analysing climate data of Turkana using CA. This study therefore groups the various climate variables over Turkana city.

The extracted data had 51 rows and 10 columns with the following key variables:

- CLDD - Cloud Cover Days
- PRED - Precipitation Days
- WETD - Wet Days
- VAPD - Vapour Pressure
- TMND - Minimum Temperature
- TMPD - Mean Temperature
- TMXD - Maximum Temperature
- PETD - Potential Evapotranspiration
- DTRD - Diurnal Temperature Range

3 Methodology

After extracting the needed data using microsoft excel it was subjected to CA analysis methods. This study focused on 3 hierarchical CA methods; single linkage, average linkage and ward algorithm. Vertical cluster tress was the main tool used to visualise the data. R was used as the programming software and language to up with the cluster trees. Clustering was done by grouping the data matrix to explore the grouping of climate variable and years (1960 -2010) over the city.

4 Results and Interpretation

The data was analysed using Hierarchical cluster method which comprises of single linkage, average linkage and ward algorithm. The study explores the grouping of climate variables and years.

4.1 Climate Viariables

i) Single Linkage

Starts with two closest data points and link. Next two closest data points and link until all data points were linked as shown in figure 2.

Cutting the cluster tree at height 9 creates two groups:

- CLDD, PRED, WETD - common feature is wetness.
- DTRD, PETD, VAPD, TMND, TMPD, TMXD - common feature is temparature.

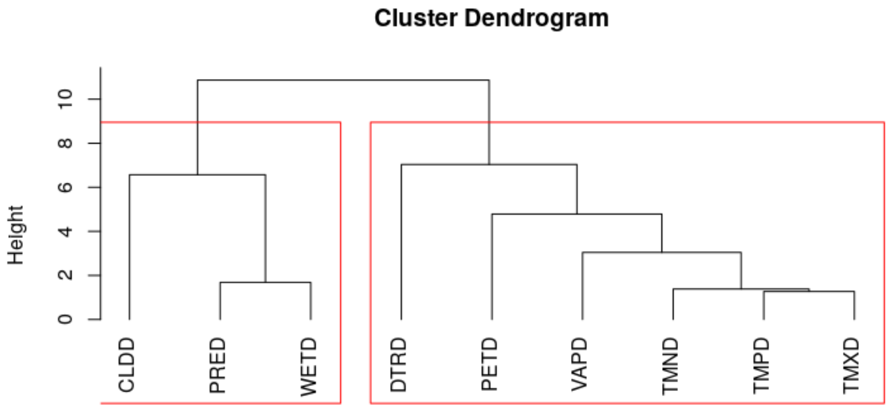


Figure 2: Single linkage

PRED and WETD are more similar to each other compared to CLDD indicated by the short linkage distance. TMPD, TMXD and TMPD are also more similar to each other than VAPD, PETD and DTRD.

ii) Average Linkage

Similar to single linkage but based on the average linkage distance. Grouped the data as shown in figure 3.

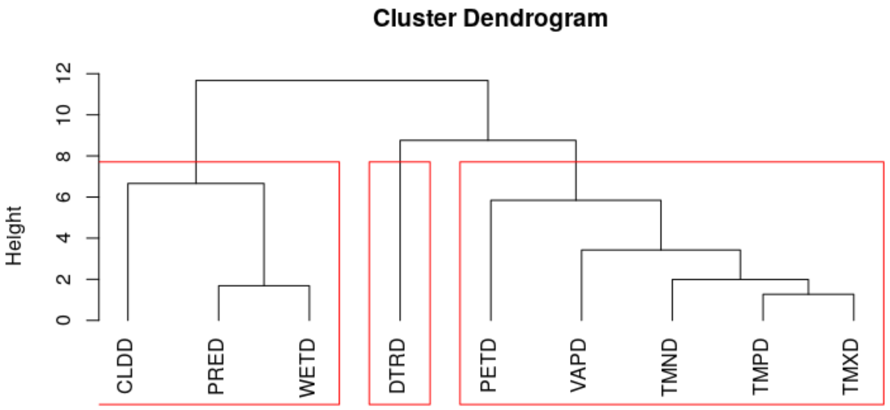


Figure 3: Average linkage

Cutting the cluster tree at the height of almost 8 created 3 groups:

- CLDD, PRED, WETD - common feature is wetness.
- DTRD
- PETD, VAPD, TMND, TMPD, TMXD - common featrue is temperature.

DTRD is an outlier. PRED and WETD are more similar to each other compared to CLDD which joins the cluster later. TMPD and TMXD are also more similar to each other more compared to PETD, VAPD and TMND.

iii) Ward Algorithm

Based on variance analysis. It seeks to minimize within group variance and maximize between group variance. Grouped the data as shown in figure 4.

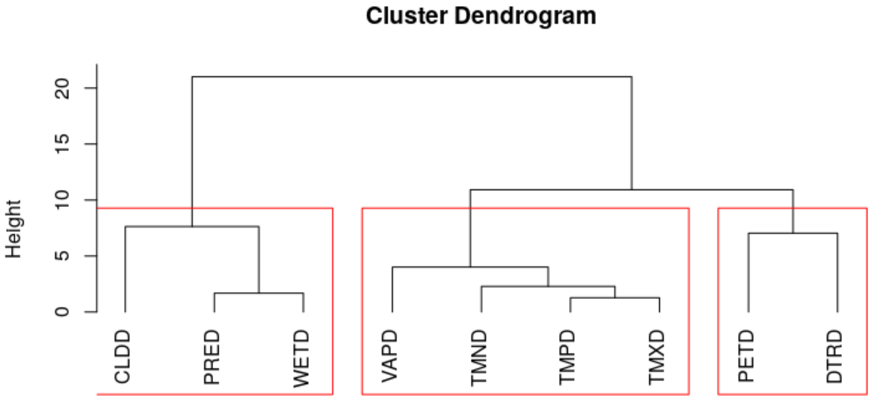


Figure 4: Ward algorithm

Cutting the cluster tree at the height of almost 10 created 3 groups:

- CLDD, PRED, WETD - common feature is wetness.
- PETD, DTRD - potential evapotranspiration causes diurnal tempareture change.
- VAPD, TMND, TMPD, TMXD - common feature is temperature.

DTRD and PETD are of similar type. PRED and WETD are more similar to each other compared to CLDD which joins the cluster later. TMPD and TMXD are also more similar to each other more compared to VAPD and TMND.

4.2 Years (1960 -2010)

i) Single Linkage

Grouped the data points of years as shown in figure 5. Cutting the cluster tree at the

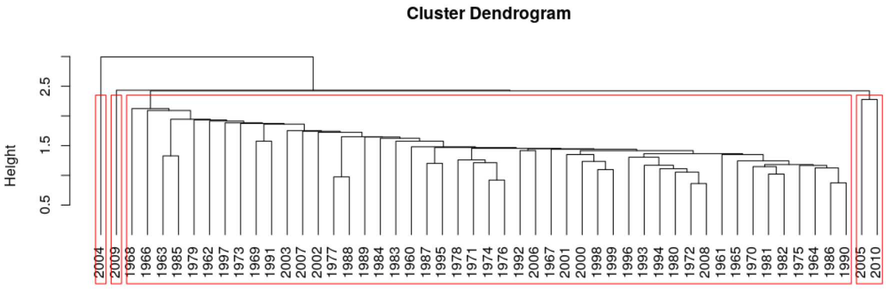


Figure 5: Single linkage

height of almost 2.5 groups the data points into 2 clusters with 2004 and 2009 as outliers. The largest group captures most of the data points indicating that all the years in this group share a common feature. It is also evident from the cluster tree that 1977 and 1988, 1974 and 1976, 1972 and 2008, 1986 and 1990 are more similar to each other since they are closest to each other.

ii) Average Linkage

Data points of years were grouped as shown in figure 6.

Cutting the cluster tree at the height of 3.5 groups the data points into 6 clusters with 2004 as an outliers. The 2 largest group captures most of the data points indicating that all the years in these groups share a common feature respectively.

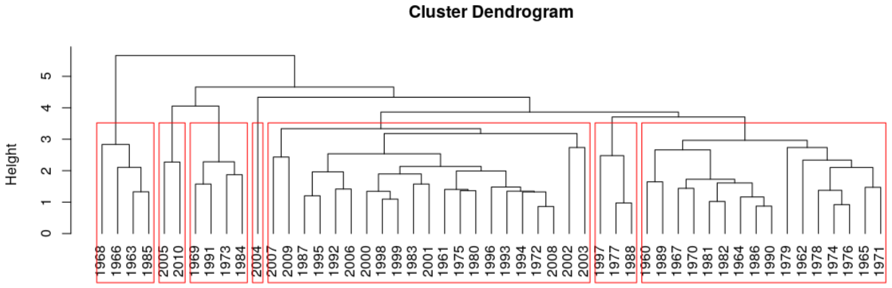


Figure 6: Average linkage

iii) Ward Algorithm

Grouped the data points of years as shown in figure 7. Cutting the cluster tree at the height of 7 groups the data points into 6 clusters.

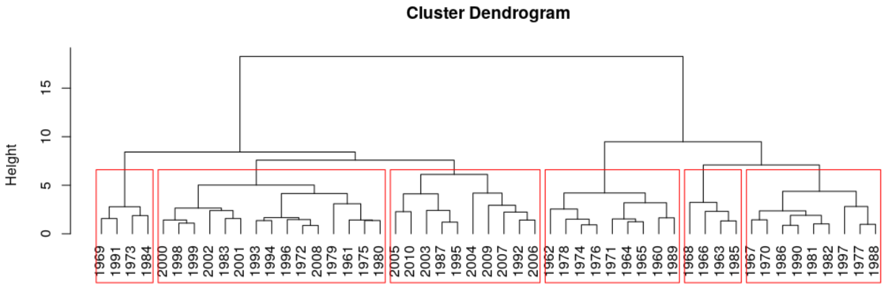


Figure 7: Ward algorithm

5 Conclusion

From the three hierarchical cluster analysis methods used in the 2 cases of variables its clear that single linkage and average linkages identify the outliers but the groups are not well defined. Ward algorithm clearly defines the groups but outliers are not well identified. It is also noted that cluster trees for climate variables and year variables are different.The years grouped together indicate that the climate feature was common to those years.

6 Task 2: Principal Component Analysis (PCA)

7 Introduction

PCA is a procedure that focuses on dimesionality reduction. It extracts fewer and independent underlying dimensions around which the data variance is organised. It also identifies the main processes that explain the largest percentage variance of the dataset. The aim of the task is to study the characteristics of PCA and compare them with that of CA. The main importance of this tast is to learn how PCA works. PCA has been used on various studies like climate variability and change on vulnerability and adaptation among Turkana pastoralists, establishing vegetable cover change over time and assessing the impact of climate change on food security of communities in Turkana. This study will try to get a deeper insight of Turkana climate data and explain the main processes that the city has been experiencing over a span of 51years.

8 Methodology

The study used two methods of dimentionality reduction; rotated and non-rotated. These two methods had different outputs but rotated was more efficient in reducing the number of dimensions. It allows changing of the factor analysis to identify new patterns of the factor structure while unrotated PCA tries to illustrate the maximum variance value with minimal number of factors. Rotated PCA helps to extract meaningful adat that accurately represents the original dataset. The study looked at component loadings, variation and component scores to select the most important principal factors and better understand the dataset. The results were compared to ward algorithm to better understand PCA. R software was used for programming.

9 Results and Interpretation

i) Variation (Rotated and Unrotated PCA)
A total of 9 principle components were obtained as shown in table 1. The standard deviation measures variability across each principle component. Proportion of variance is the total variance percentage explained by each principle component in the original data set. PC1 explains 50% of the total variance in the original dataset. PC2 explains 23% of the total variance. In cumulative proportion PC1,PC2 and PC4 explains 85% of the total variance. As a result PC1, PC2 and PC4 were selected to explain the dataset thus dimentionality reduction.
Both unrotated and rotated differ in principle components that best describe the data set. Rotated method was prefered to unrotated as it reduces the dimensions to PC1,PC2 and PC4 which explain a variance of 44%, 23% and 18% respectively.

	Unrotated			Rotated		
	Std dev	Proportion of Variance	Cumulative Proportion	Std dev	Proportion of Variance	Cumulative Proportion
PC1	2.276	0.576	0.576	3.947	0.453	0.453
PC2	1.419	0.224	0.800	2.082	0.261	0.714
PC3	1.048	0.122	0.922	1.622	0.213	0.927
PC4	0.722	0.058	0.979	1.152	0.053	0.980
PC5	0.386	0.017	0.996	0.156	0.012	0.992
PC6	0.163	0.005	0.997	0.027	0.005	0.997
PC7	0.093	0.001	1	0.014	0.003	1.000
PC8	0.025	0	1	0.001	0	1.000
PC8	0.025	0	1	0.001	0	1.000
PC9	2.635	0	1	0	0	1.00

Table 1: PCA Summary

Selection of the most important principle components can also be selected using the scree plot show in 8 with a default setting that the varinace should be atleast 1.

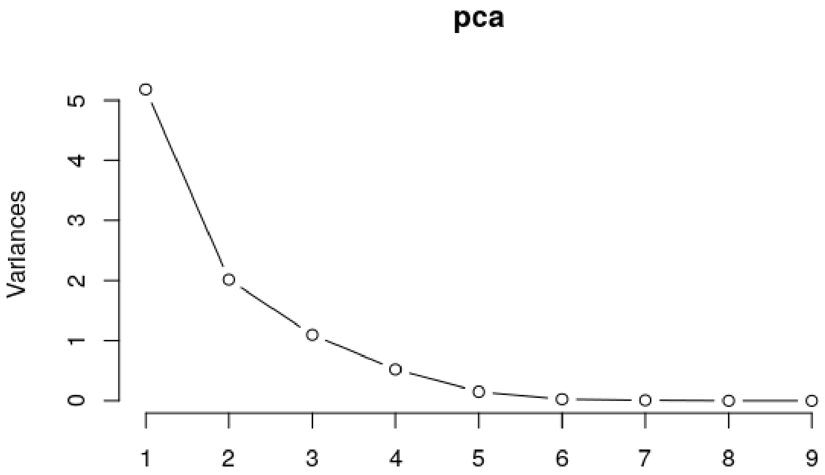


Figure 8

ii) Component Loadings (Rotated)

These are the correlation coefficients of the linear combination of the original variables from which principle components are constructed. Rotated PCA determines the loadings cutoff for the importnat variables that contribute highest to the factors as shown in red in table 2.

TMPD, TMND, TMXD and VAPD all increase in PC1. Due to the high temperatures in the dataset, this process is drought. In PC2 both PRED and WETD decreases. In drought seaons there's no rain and the land is dry. In PC3 as CLDD decreases,DTRD increases. When there's no cloud during the night the temparatures tend to decrease and increase during the day. This results to increase in dirunal range temparatures.

Variables	**PC1**	**PC2**	**PC3**
PRED	-0.162010	-0.953727	-0.171107
PETD	0.546160	0.396223	0.678831
CLDD	-0.141978	-0.508158	-0.714389
DTRD	0.011618	0.029256	0.893948
TMPD	0.965685	0.142101	0.194406
TMND	0.983804	0.137589	0.028010
TMXD	0.917354	0.138291	0.349367
VAPD	0.935681	0.154074	-0.057063
WETD	-0.167719	-0.952899	-0.155570

Table 2: Principle component loadings

iii) Component Scores (Rotated)
Figure 9 figuratively describes percentage variance of each principle component on the original data and the component scores. The score shows how each process varies in different years. PC1 most active in 2010, most inactive in 1968 but was dormant in 1975. PC2 was most active in 1966, most inactive in 1969 but was dormant in 1960,1961 and 2002. PC4 most active in 2005, most inactive in 1963 with no dormance in any year.

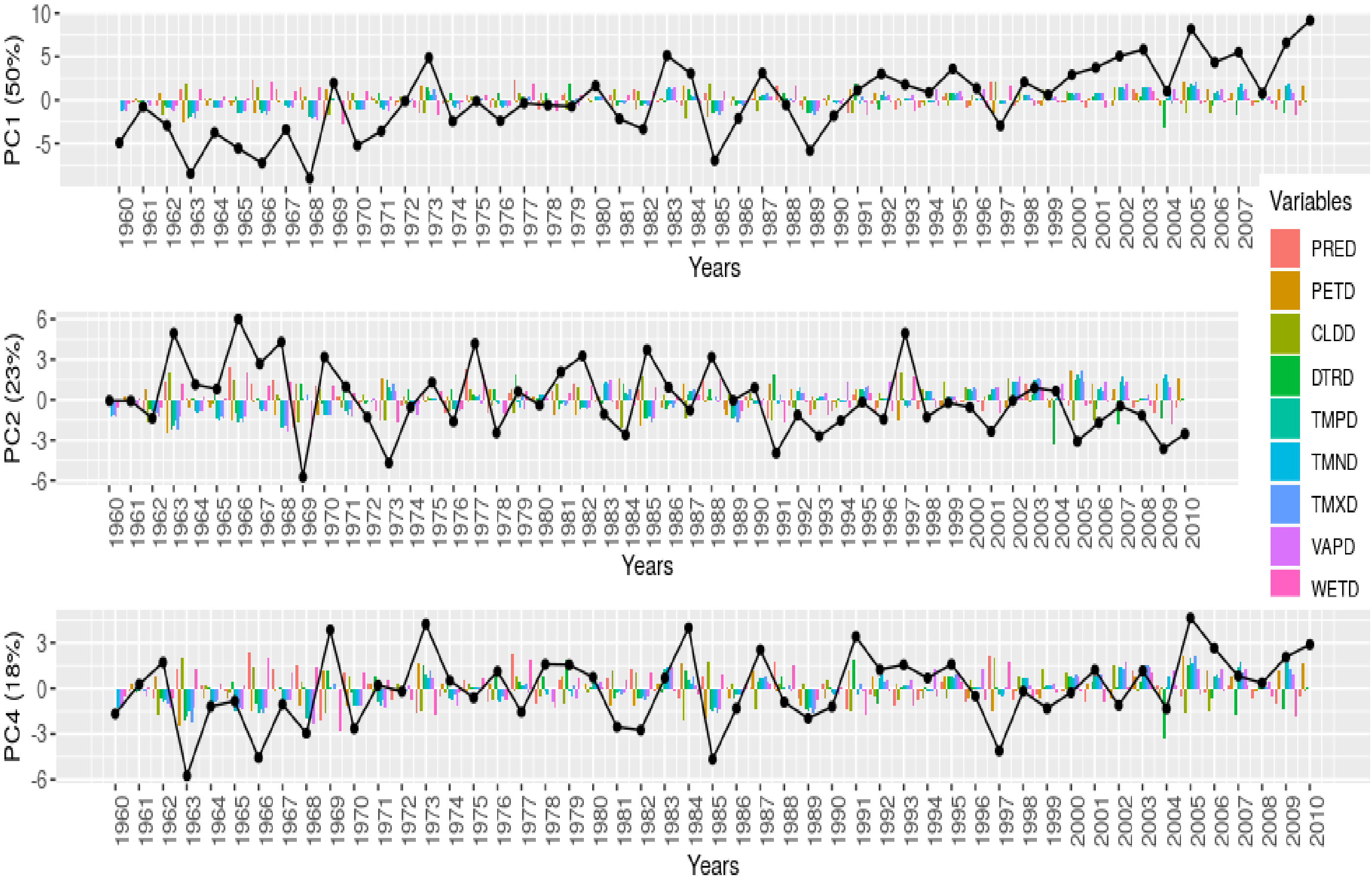


Figure 9: Component scores

10 Conclusion

PCA analysis identifies the processes that cause the climate variables to vary. PCA gives a deeper understanding of the dataset as it aslo reduces dimentionality which can be helpful for better climate predictions and data analysis. Both loadings and scores play an important role in analysing data using PCA. Ward algorithm defines the groups well but you can barely deduce how they are grouped and the processes involved.

11 Task 3: Time-Series Analysis

12 Introducttion

Time series analysis is a sequence of data(continuous) that follow non-random orders. The data must be equally spaced time intervals. To deploy time series analysis the dataset must satisfy these two conditions. Time series analysis comprises of various methods that analysizes time series data for the purpose of extracting meaningful climate statistics and other featurs of the data. The aim of this task was to explore the time series patterns in precipitation (PRED) and temperature (TMPD) data and the relationships between the two variables. This is important for the purpose of learning and analysing the patterns of temperature and precicpitation in Turkana city over a span of 51years from 1960-2010. As a result it helps to study the past and predict the future.

13 Methodology

There are various methods of time series analysis which can be classified into frequency domain (Fourier and Wavelet analysis) and time domain (autocorrelation and cross-correlation). This task focused on time domain and time series to identify if there was trend in the data. Dataset was re-extracted again using excel so as to include the month variable since the data was indexed monthly from January 1960 to December 2010. Using statistica TMPD and PRED values were subjected to time series to identify the patterns. At some point it was required to detrend the temperature variable for better analysis of the data. Detrending of the data was done using microsoft Excel.

14 Results and Interpretation

14.1 Plot of PRED and TMPD time series

It was evident from figure 10 that temperature variable had deterministic and global trend; change in mean over time. Line of best fit showed the temperature values were increasing with 0.0025^0C every month from 1960 to 2010. Precipitation variable presented a very small trend as shown in figure 11 which was neglected. This implied the slope of the best line of fit was insignificant but we can see the pattern of highest amount of rainfall recorded at 120 was recieved in 1960 and this has reduced over the years.

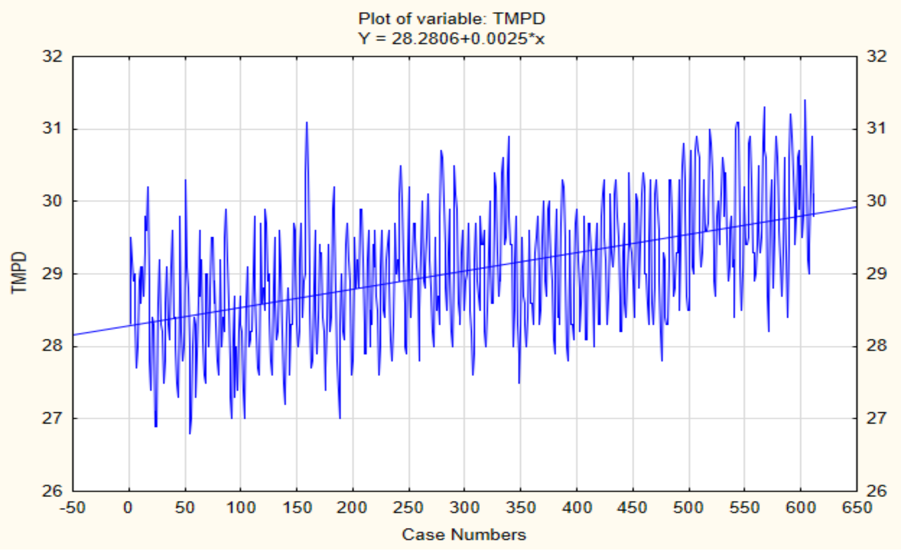


Figure 10

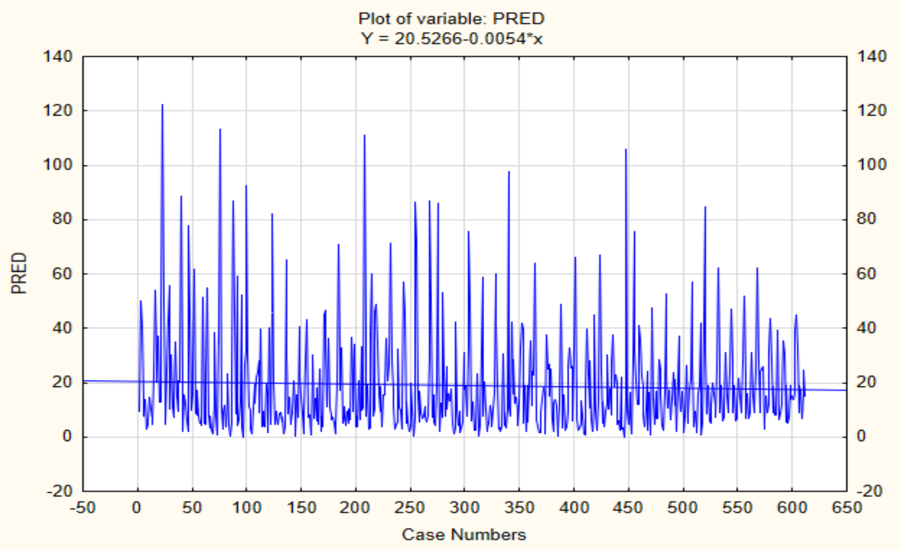


Figure 11

14.2 Detrended TMPD

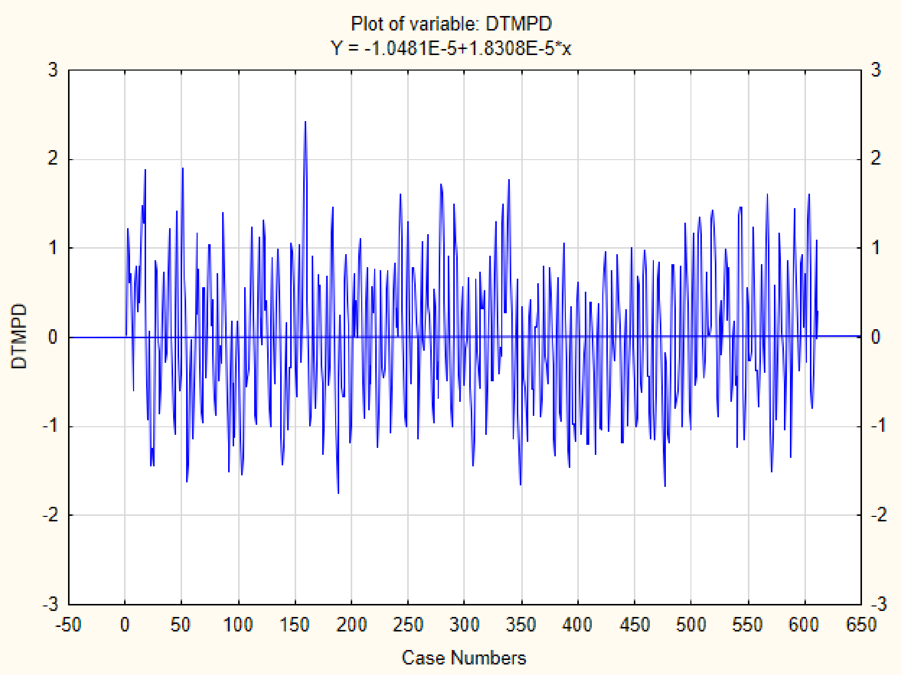


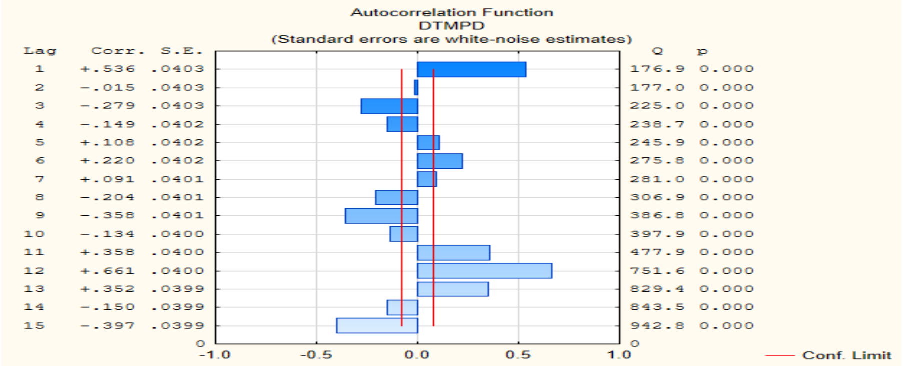
Figure 12: Detrended TMPD

There are many methods of detrending data but linear regression was used in this case. Detrending was done to remove distortion in TMPD variable to allow better analysis of its variability. Plot of the detrended data gave a clearer visual of the cyclic patterns that is increase and decrease of temperature with time compared to trended data. It was clear to identify extreme maximum temperature of ......

14.3 Correlogram for PRED and DTMPD

Correlogram is synonym for autocorrelation. It is the correlation between a series data and lagged version of itself. It is used to find repeating patterns. As shown in figure 13 DTMPD autocorrelation shows a significant sinusidal cycle. With a lag of 1 month, it shows that temperature data is correlated with the past data and it varies with time. It also shows a repetitive annual pattern. The red line indicates the approximate 95% confidence interval.

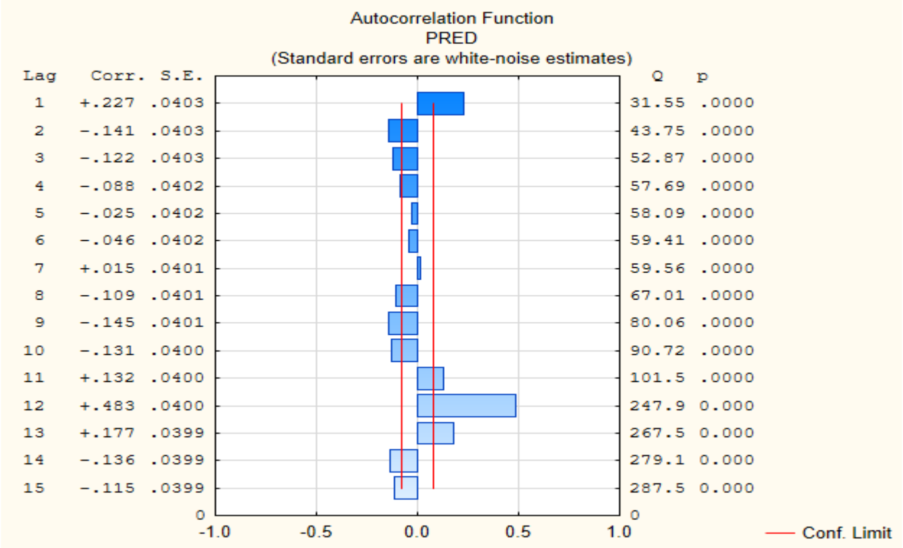
Figure 14 shows a weak sinosidal autocorrelation of precipitation. At lag 1 correlation is at approximately 0.25. This shows that the correlation is less significant and precipitation values are almost random in the area.



Autocorrelation Function
DTMPD
(Standard errors are white-noise estimates)

Lag	Corr.	S.E.	Q	p
1	+.536	.0403	176.9	0.000
2	-.015	.0403	177.0	0.000
3	-.279	.0403	225.0	0.000
4	-.149	.0402	238.7	0.000
5	+.108	.0402	245.9	0.000
6	+.220	.0402	275.8	0.000
7	+.091	.0401	281.0	0.000
8	-.204	.0401	306.9	0.000
9	-.358	.0401	386.8	0.000
10	-.134	.0400	397.9	0.000
11	+.358	.0400	477.9	0.000
12	+.661	.0400	751.6	0.000
13	+.352	.0399	829.4	0.000
14	-.150	.0399	843.5	0.000
15	-.397	.0399	942.8	0.000

Figure 13: Autocorrelation DTMPD



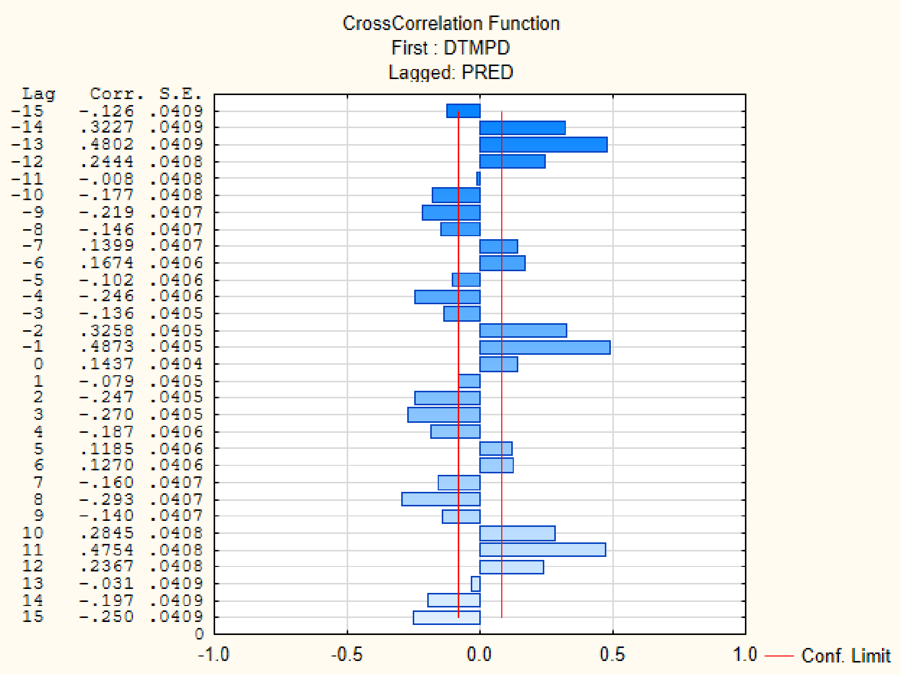
Autocorrelation Function
PRED
(Standard errors are white-noise estimates)

Lag	Corr.	S.E.	Q	p
1	+.227	.0403	31.55	.0000
2	-.141	.0403	43.75	.0000
3	-.122	.0403	52.87	.0000
4	-.088	.0402	57.69	.0000
5	-.025	.0402	58.09	.0000
6	-.046	.0402	59.41	.0000
7	+.015	.0401	59.56	.0000
8	-.109	.0401	67.01	.0000
9	-.145	.0401	80.06	.0000
10	-.131	.0400	90.72	.0000
11	+.132	.0400	101.5	.0000
12	+.483	.0400	247.9	0.000
13	+.177	.0399	267.5	0.000
14	-.136	.0399	279.1	0.000
15	-.115	.0399	287.5	0.000

Figure 14: Autocorrelation PRED

14.4 Cross-correlation of PRED and DTMPD

Cross-correlation describes the degree of correlation between two different time series. It is useful in determining whether changes in one time series has effect on the other time series. Figure 15 shows temperature is signficantly correlated to precicpitation. At lag -13 and 11 the correlation is 0.5. This implies that increase in temperature increases the level of evaporation which can later condense to form clouds which can lead rainfall.



Lag	Corr.	S.E.
-15	-.126	.0409
-14	.3227	.0409
-13	.4802	.0409
-12	.2444	.0408
-11	-.008	.0408
-10	-.177	.0408
-9	-.219	.0407
-8	-.146	.0407
-7	.1399	.0407
-6	.1674	.0406
-5	-.102	.0406
-4	-.246	.0406
-3	-.136	.0405
-2	.3258	.0405
-1	.4873	.0405
0	.1437	.0404
1	-.079	.0405
2	-.247	.0405
3	-.270	.0405
4	-.187	.0406
5	.1185	.0406
6	.1270	.0406
7	-.160	.0407
8	-.293	.0407
9	-.140	.0407
10	.2845	.0408
11	.4754	.0408
12	.2367	.0408
13	-.031	.0409
14	-.197	.0409
15	-.250	.0409

Figure 15

15 Conclusion

Autocorrelation, cross-correlation and time series played a vital role in identifying the underlying patterns of how temperature and precipitation varied with time. To get a deeper and better insight of the data more analysis needs to be done. Both autocorrelation an dcross-correlation are based on correlation.

16 Task 4: Spectral and Wavelet Analysis

17 Introduction

Spectral analysis is also referred to as fourier analysis. It is useful in identifying seasonal fluctuatins of different lengths in the dataset. From the cyclic pattern identified in fourier analysis, wavelet analysis explains the behaviour and change of the cycles and amplitude with time. The aim of this task is to explore and compare spectral analysis and wavelet analysis using the precipitation data. The study was important for learning and analysing how the cycles of precicpation varied within the span of 51years from 1960-2010 in Turakana City. This task will also be an extension of the analysis done in task 3.

18 Methodology

19 Results and Interpretation

20 Conclusion