Precipitation and Mosquito-Borne Diseases in Africa

Hariharan Sreenivas (sreenh@rpi.edu), Jagrati Sharma (sharmj@rpi.edu), Ridhi Gulati (gulatr@rpi.edu)

[1]Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States.

Abstract

Mosquitoes are classified as one of the deadliest animals in the world. There are different types of mosquitoes, which can cause various diseases like zika, dengue, chikungunya, yellow fever, and many more. According to the World Health Organization more than 50 percent of the world's population is presently at risk from mosquito-borne diseases. Africa is the world's second largest and second most populated continent. It covers 6% of the Earth's total surface area and 20% of its land area. The African region has a high risk of mosquito-borne diseases. Mosquitoes are cold-blooded animals and prefer temperatures over 80-degree Fahrenheit. Mosquitoes birth are mostly dependent on the water and temperature in the surrounding for egg to hatch into larvae. Africa has less than 1000 millimeters of rainfall and it decreases with the distance from the equator. Generally the rainfall is most abundant in eastern and central part of Africa.

This poster will mainly focus on data visualization and finding a correlation between vector-borne diseases like dengue, Eastern Equine Encephalitis, Filariasis, Malaria, Rift Valley Fever, Tularemia, West Nile Virus, Yellow fever, and Zika virus and the rainfall over Africa from June 2019 to November 2019. The poster visualizes rainfall and the location of the outbreak of the diseases. Then, we use the KNN and Hierarchical Clustering algorithms to find major clusters of disease outbreaks based on the location of the disease.

Dataset

The dataset is collection from two different sources. The rainfall data is collected from TAMSAT in **.netCDF** format, a satellite-based rainfall estimation for Africa. The health data is collected from HealthMap in **.CSV** format, a freely available website that delivers real-time intelligence on a broad range of emerging infectious diseases for a diverse audience including libraries, local health departments, governments, and international travelers. Both the datasets are between the months June to November 2019.

Exploratory Data Analysis

The analysis is done on various diseases such as dengue, Eastern Equine Encephalitis, Filariasis, Malaria, Rift Valley Fever, Tularemia, West Nile Virus, Yellow fever, and Zika virus. The malaria has the maximum occurrence and Eastern Equine Encephalitis has the least.

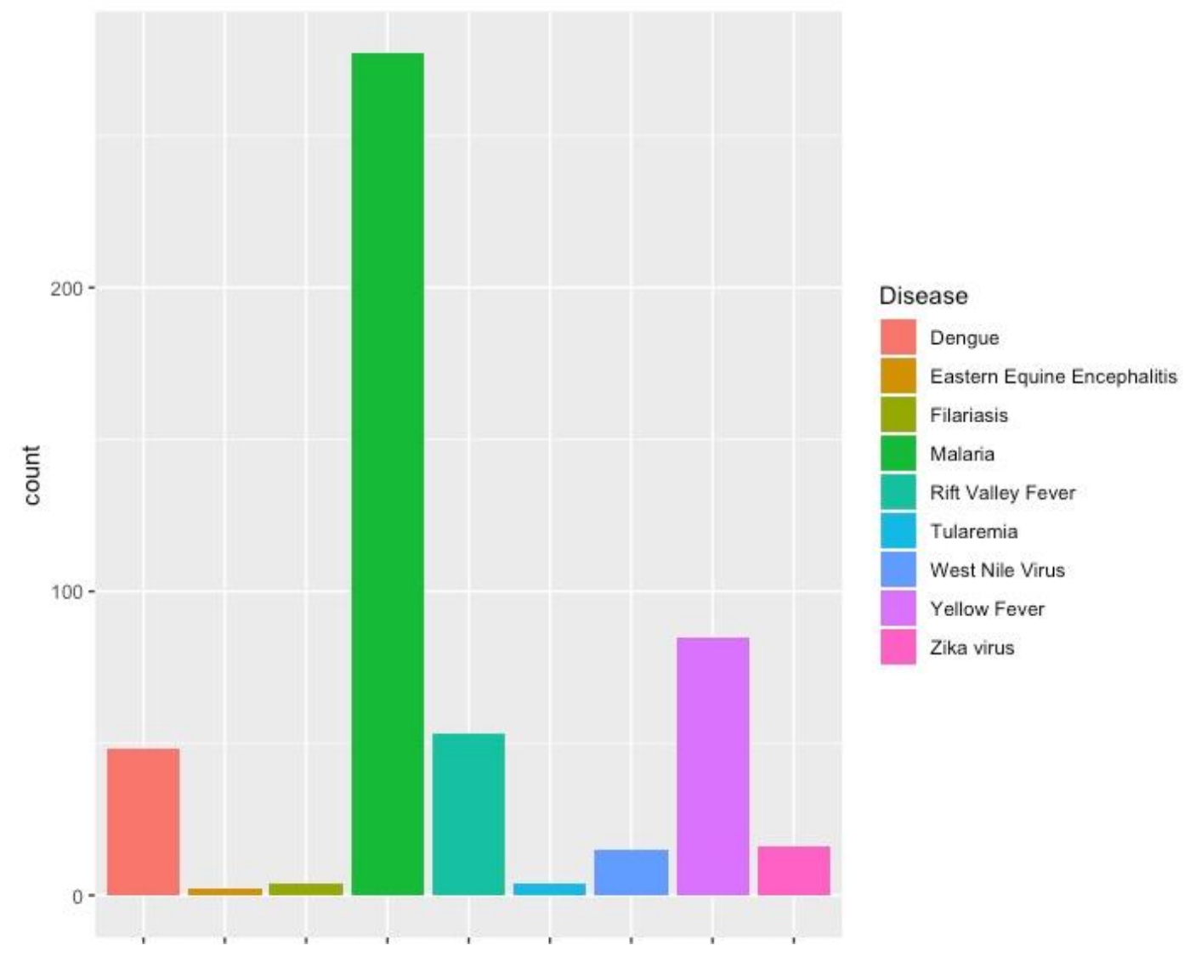
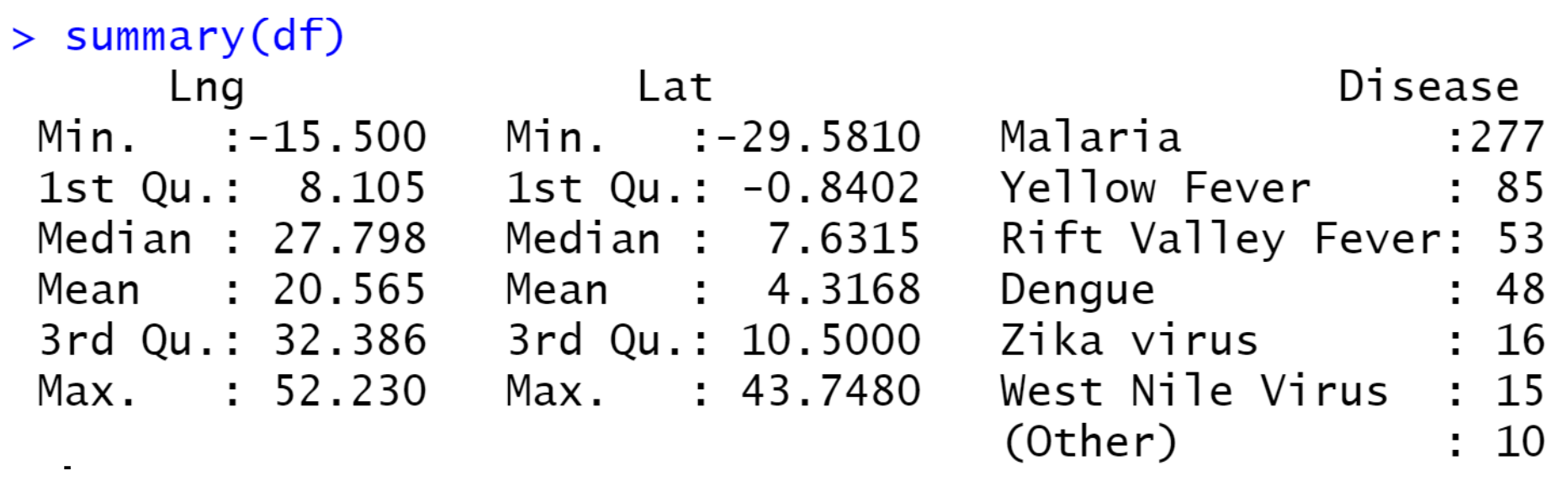


Fig 1- Bar plot of the disease count

Sponsors:



```
> summary(df)
      Lng              Lat                       Disease   
 Min.   :-15.500   Min.   :-29.5810   Malaria          :277  
 1st Qu.:  8.105   1st Qu.: -0.8402   Yellow Fever     : 85  
 Median : 27.798   Median :  7.6315   Rift Valley Fever: 53  
 Mean   : 20.565   Mean   :  4.3168   Dengue           : 48  
 3rd Qu.: 32.386   3rd Qu.: 10.5000   Zika virus       : 16  
 Max.   : 52.230   Max.   : 43.7480   West Nile Virus  : 15  
                                      (Other)          : 10  
```

Fig-2 Summary of the Disease dataset

Correlation between Rainfall and Disease

The visualization of rainfall and vector-borne diseases data like dengue, Eastern Equine Encephalitis, Filariasis, Malaria, Rift Valley Fever, Tularemia, West Nile Virus, Yellow fever, and Zika virus. in Africa between June 2019 and November 2019.

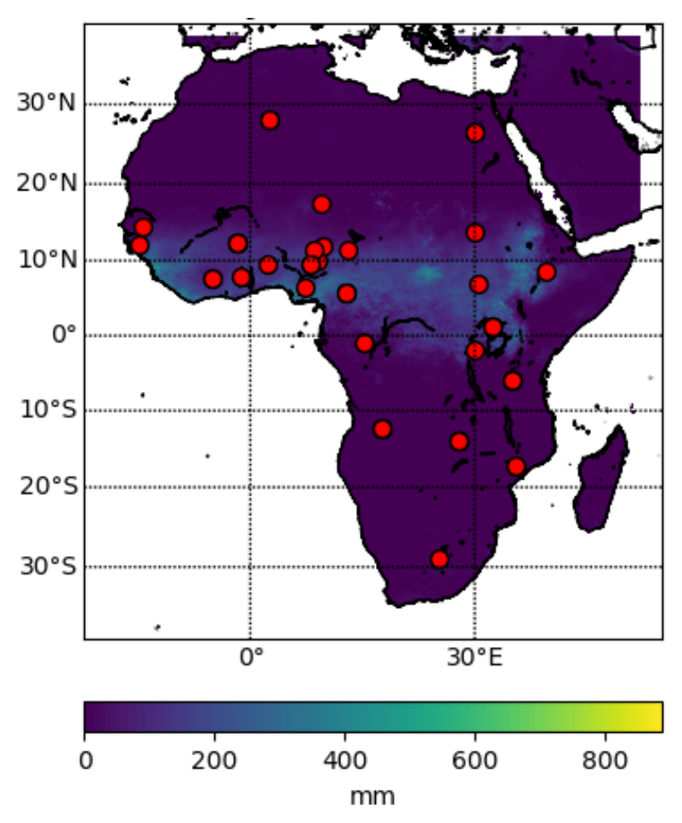


Fig 3: June 2019

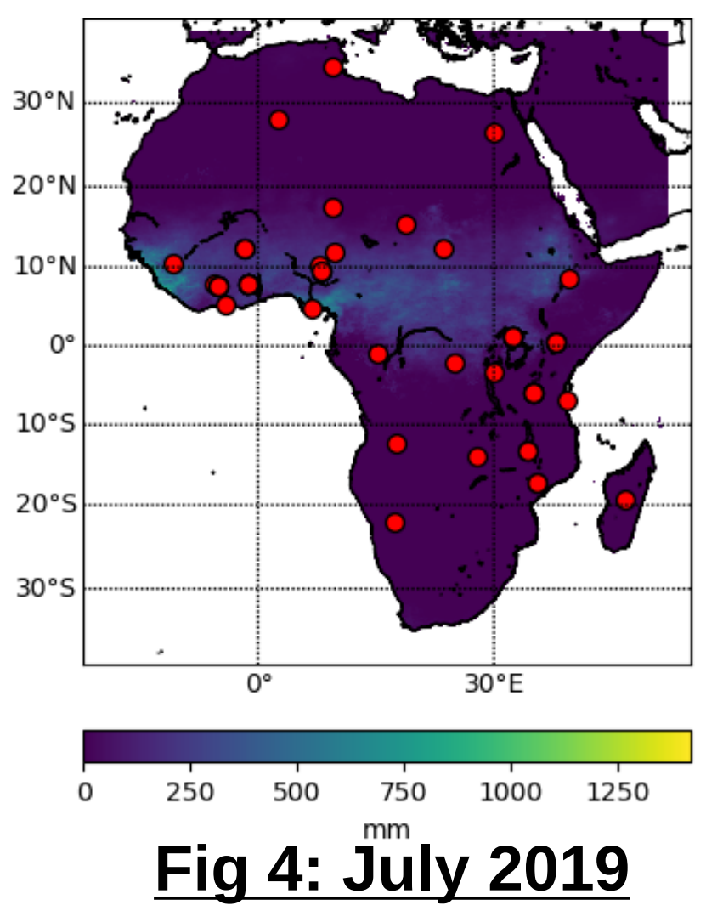
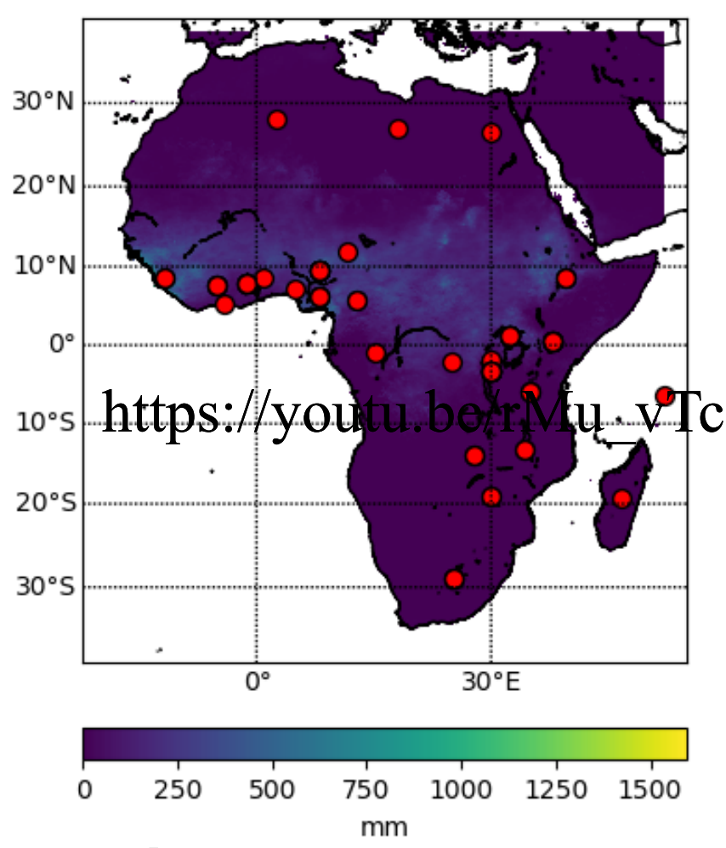


Fig 4: July 2019



https://youtu.be/rMu_vTcL__g

Fig 5: August 2019

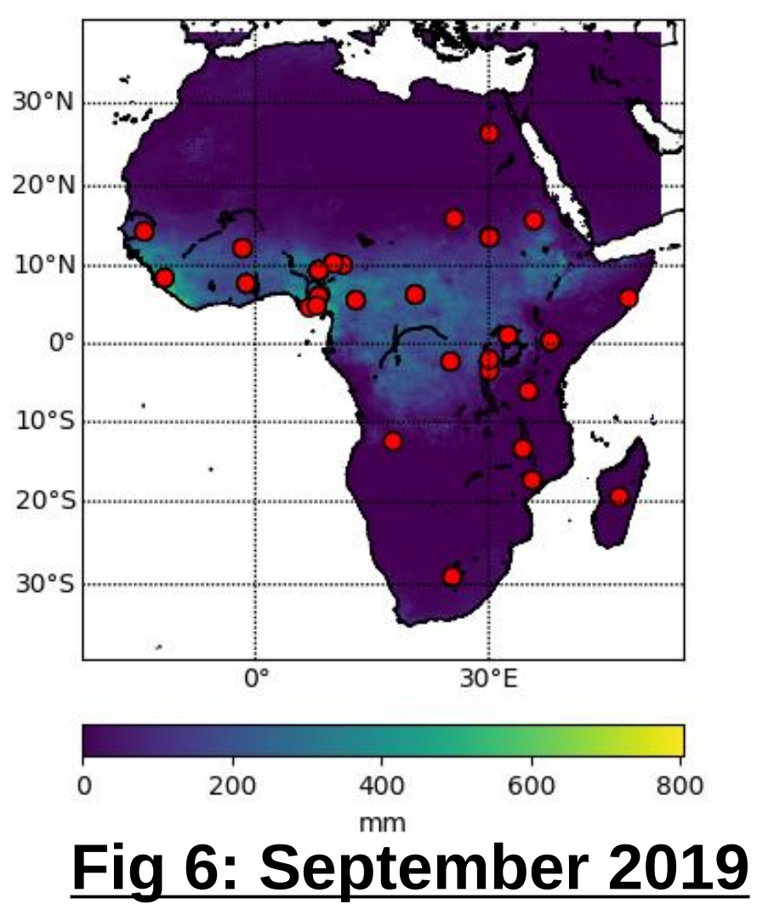


Fig 6: September 2019

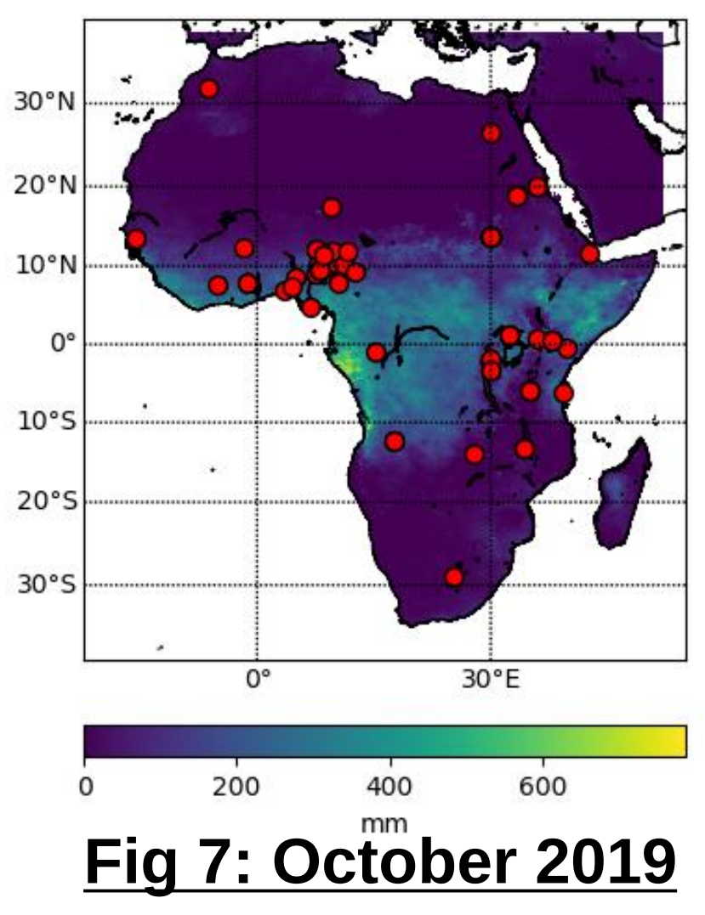


Fig 7: October 2019

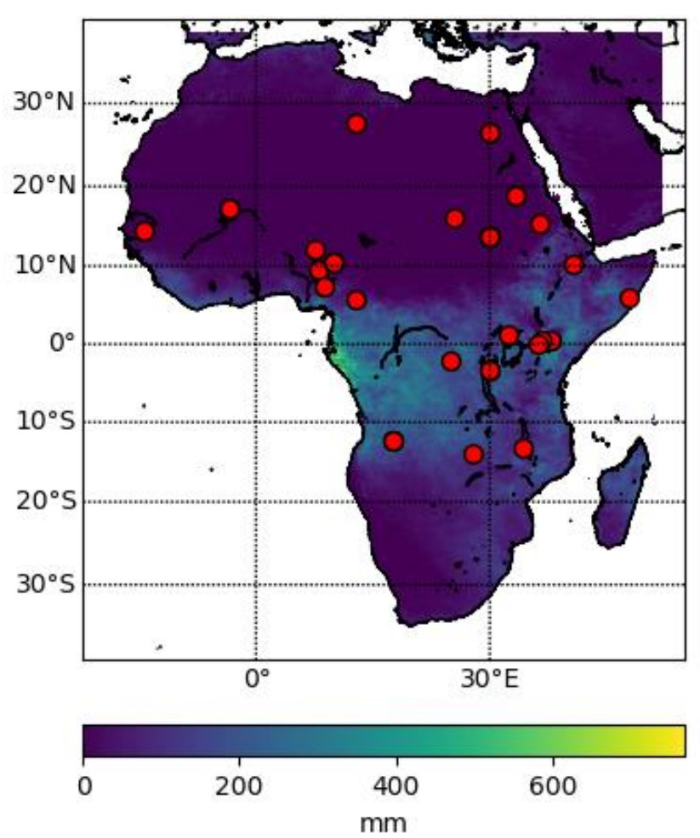


Fig 8: November 2019

(Fig 3-8: Mapping rainfall and disease data)

K-Nearest Neighbor Model

For KNN, we first created a matrix of coordinates. Using the *knearneigh* function in *spdep* package, a matrix of Great Circle distances is created. This matrix is converted into a list of *nb* (spatial neighbours) vectors. The code for KNN is:

```
library(spdep)
#Matrix of coordinates
df.coords <- cbind(df$Lng,df$Lat)

#Creating spatial neighbours
df.5nn <- knearneigh(df.coords, k=5, longlat = TRUE)
df.5nn.nb <- knn2nb(df.5nn)
plot(df.5nn.nb,df.coords)
```

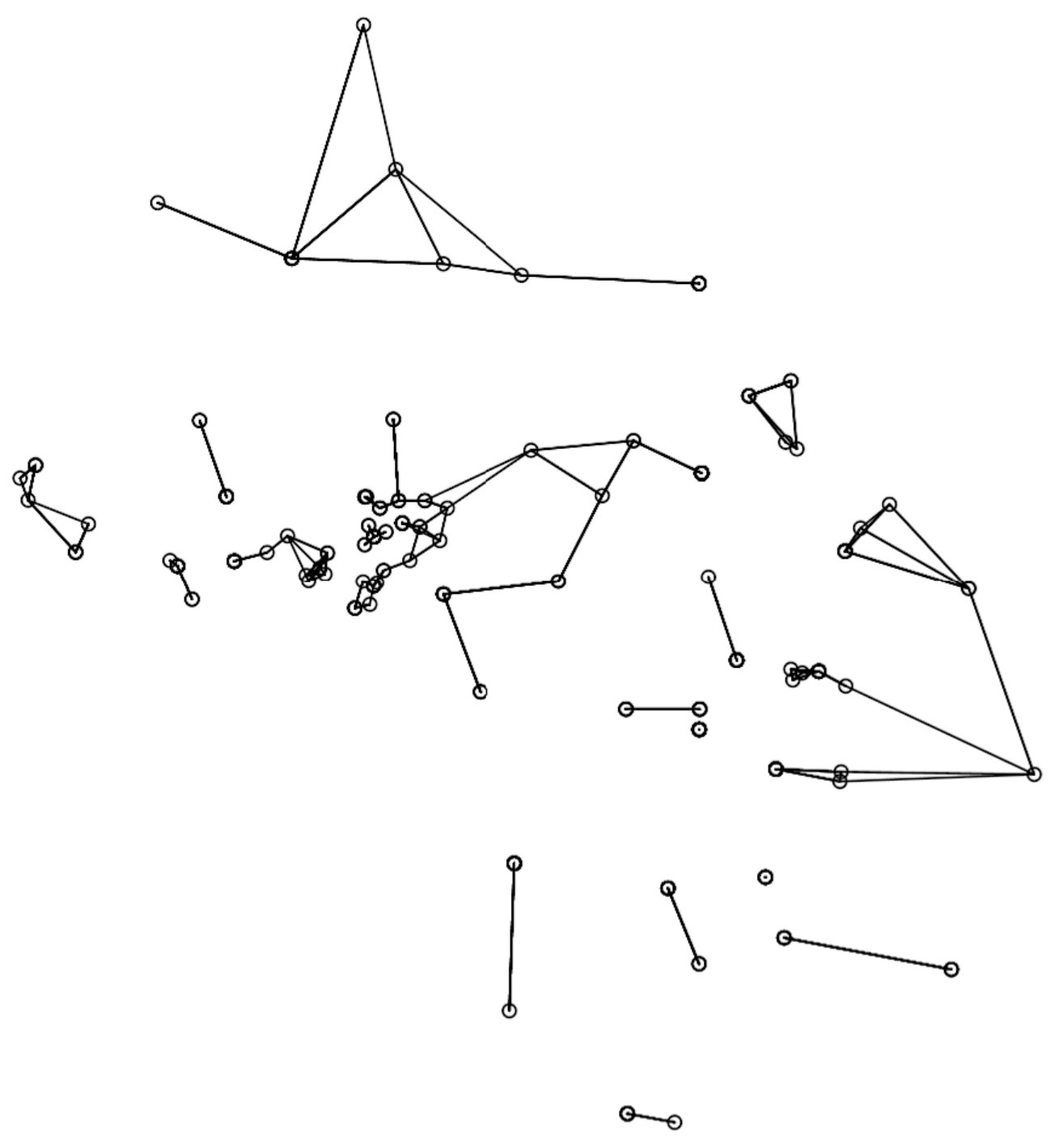
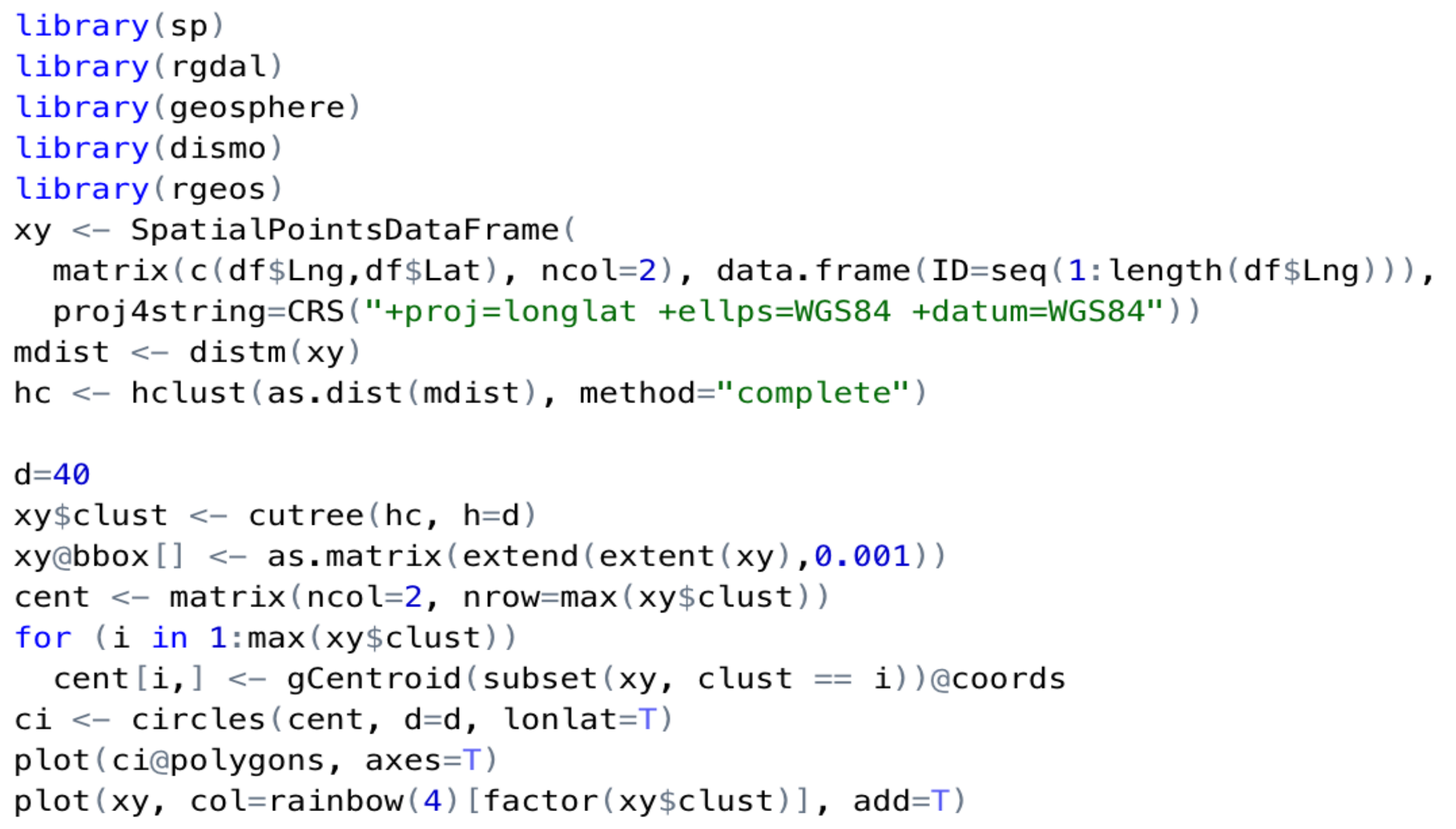


Fig 9: KNN Model

This KNN creates clusters using Spatial neighbours. The model uses great circle distances for clustering.

Hierarchical Cluster Analysis

For Hierarchical clustering, we first created a spatial data frame using "*SpatialPointDataFrame"* function. It is then converted to a distance matrix using '*distm'* function and this matrix is used in clustering. The code for Hierarchical Cluster Analysis is:



```
library(sp)
library(rgdal)
library(geosphere)
library(dismo)
library(rgeos)
xy <- SpatialPointsDataFrame(
  matrix(c(df$Lng,df$Lat), ncol=2), data.frame(ID=seq(1:length(df$Lng))),
  proj4string=CRS("+proj=longlat +ellps=WGS84 +datum=WGS84"))
mdist <- distm(xy)
hc <- hclust(as.dist(mdist), method="complete")

d=40
xy$clust <- cutree(hc, h=d)
xy@bbox[] <- as.matrix(extend(extent(xy),0.001))
cent <- matrix(ncol=2, nrow=max(xy$clust))
for (i in 1:max(xy$clust))
  cent[i,] <- gCentroid(subset(xy, clust == i))@coords
ci <- circles(cent, d=d, lonlat=T)
plot(ci@polygons, axes=T)
plot(xy, col=rainbow(4)[factor(xy$clust)], add=T)
```

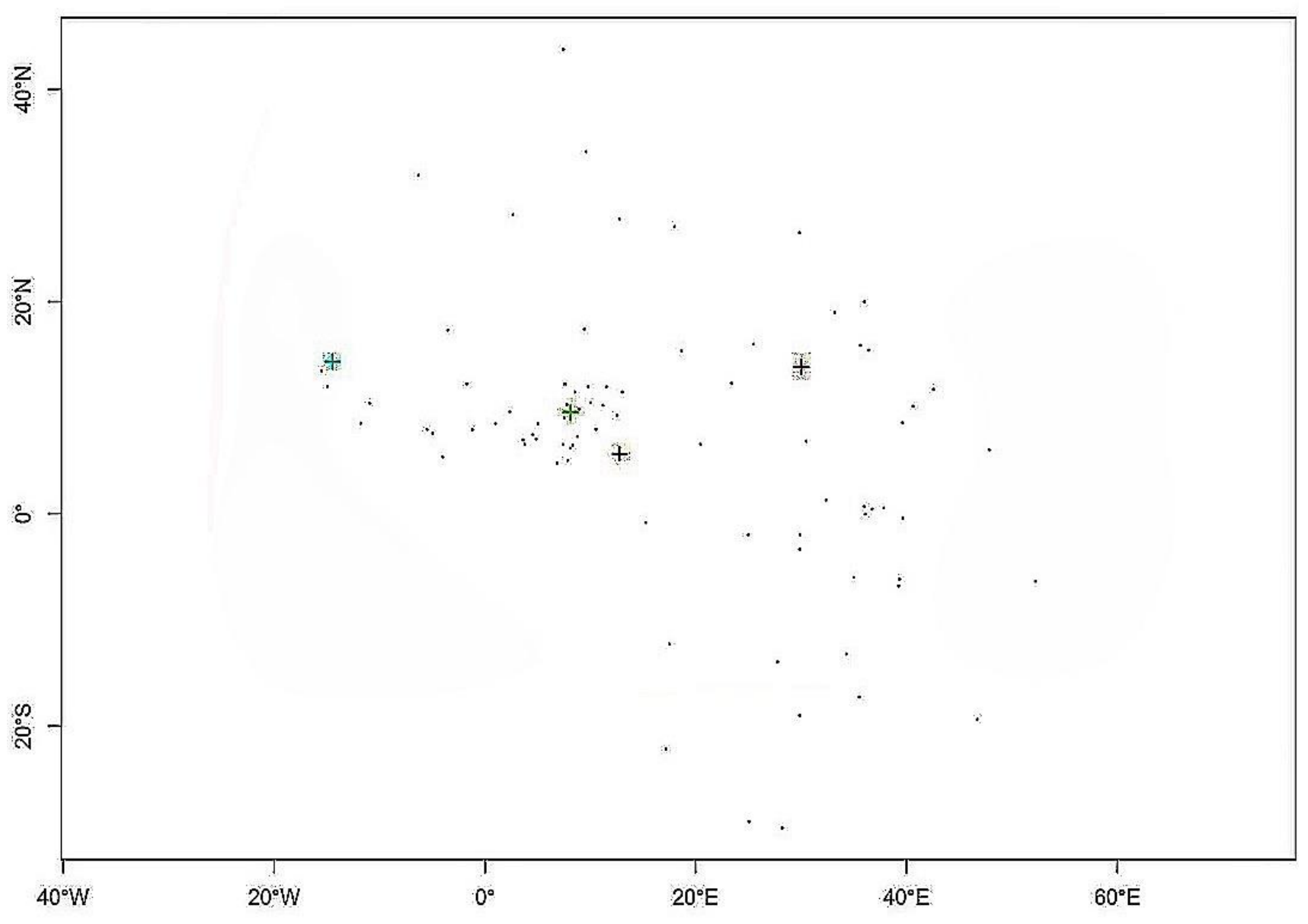


Fig 10: Hierarchical clustering model based on distance rule.

The centroids represent the major outbreak points of the diseases.

Conclusion

From the visualizations above, it can be seen that there isn't a clear correlation between rainfall and the outbreak of mosquito-related diseases. This is because several other factors affect the outbreak of mosquito-related diseases that include surface temperature, air temperature, precipitation, soil moisture, vegetation, and evapotranspiration. For building an accurate model all these factors must be taken into consideration.

Glossary:

- Python – A programming language, capable of processing data/statistical analysis
- R – A program to process data and perform statistical analysis
- Pandas – A useful data manipulation package in python
- Df, dataframe – Data manipulation structure in R & python pandas
- KNN- A non-parametric method used for classification and regression
- Hclust-Type of agglomerative clustering is to repeatedly combine the two nearest clusters into a larger cluster

Resources:

- Rainfall dataset: https://www.tamsat.org.uk/data/archive
- Mosquito Dataset: https://www.healthmap.org/en/
- Hierarchical Clustering: https://gis.stackexchange.com/questions/17638/clustering-spatial-data-in-r
- Panoply: https://www.giss.nasa.gov/tools/panoply/
- Python Script: https://www.youtube.com/watch?v=XiZbrii49pI
- KNN for Spatial data: https://www.youtube.com/watch?v=MtkuQxxQj5s

SCAN THIS CODE FOR VIDEO!