# My presentation

Rui Dong

2022-10-14

## Contents



# 1 How I can contribute to the SAPPHIRE project

## 1.1 Update the `glottodist` with metrics other than Gower's distance

We can implement other kinds of metrics like listed in [1], Take Eskin's distance as an example:

$$S_k(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{n_k^2}{n_k^2+2} & \text{otherwise} \end{cases}$$

## 1.2 Add a function to compute the Hausdorff distance

def Hausdorff

## 1.3 Topological data analysis (TDA)

sdfs [2]

## 1.4 Spectral graph theory methods

```
library(glottospace)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(TDAstats)
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
# load the dataset wals
wals <- glottoget("wals")
```

```
#select the data wrt Aouth America
wals_sam <- subset(wals, continent == "South America")
```

```
# Define a function select.features(dataset, feature_names, a) to get all features of dataset that the
count.na <- function (dataset, feature.name) {
  num.features <- length(st_drop_geometry(dataset))
  return(sum(is.na(dataset[[feature.name]])) / nrow(dataset))

}

select.features <- function(dataset, feature_names, a) {
  counts <- c()
  for (x in feature_names) {
    counts <- append(counts, count.na(wals_sam, x))
  }
  result <- t(as.matrix(counts[counts<0.5]))
  colnames(result) <- feature_names[which(counts<0.5)]

  return(result)
}
```

```
feature_names <- colnames(wals_sam)[-c(1, 194:208)]
select.features(wals_sam, feature_names, 0.5)
```

```
##         81A   82A    83A       86A   129A
## [1,] 0.4875 0.425 0.4125 0.4833333 0.4875
```

```
data <- select(wals_sam, 'glottocode', '81A', '82A', '83A', '86A', '129A')
data.obj <- st_drop_geometry(data)

data.obj$`81A` <- as.factor(data.obj$`81A`)
data.obj$`82A` <- as.factor(data.obj$`82A`)
data.obj$`83A` <- as.factor(data.obj$`83A`)
data.obj$`86A` <- as.factor(data.obj$`86A`)
data.obj$`129A` <- as.factor(data.obj$`129A`)
```

```
gtd.unique <- match(unique(data.obj$glottocode), data.obj$glottocode)
data.obj.unique <- data.obj[gtd.unique, ]
str(data.obj.unique)
```

```
## 'data.frame':    235 obs. of  6 variables:
##  $ glottocode: chr  "abip1241" "acha1250" "ache1246" "achu1248" ...
##  $ 81A       : Factor w/ 7 levels "No dominant order",..: 5 5 NA NA NA NA NA NA NA 1 ...
##  $ 82A       : Factor w/ 3 levels "No dominant order",..: 2 2 NA NA NA NA NA NA NA 1 ...
##  $ 83A       : Factor w/ 3 levels "No dominant order",..: 3 3 NA 2 NA NA NA NA NA 2 ...
##  $ 86A       : Factor w/ 3 levels "Genitive-Noun",..: 2 1 NA 1 1 NA NA NA NA 1 ...
##  $ 129A      : Factor w/ 2 levels "Different","Identical": 1 1 NA 1 NA 1 NA NA NA 1 ...
```

```
structure <- glottocreate_structuretable(varnames = c("81A", "82A", "83A", "86A", "129A"))
structure$type <- rep("factor", 5)
```

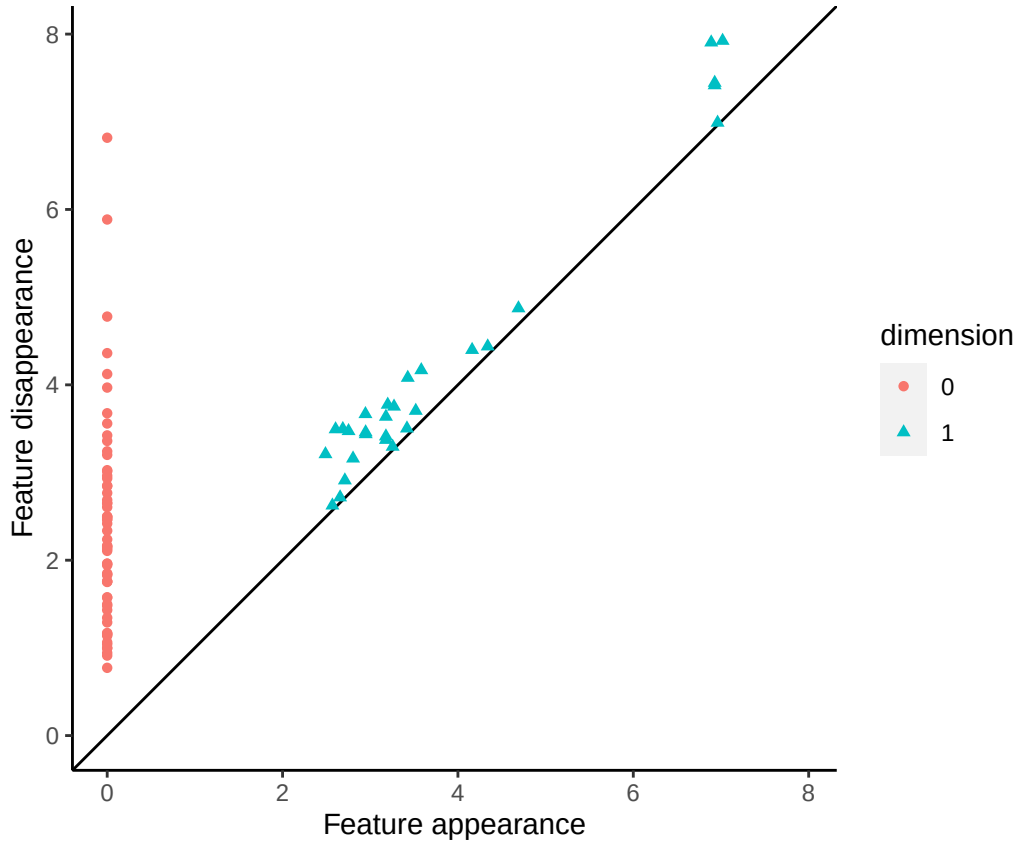
```
my.glottodata <- glottocreate_addtable(data.obj.unique, structure, name="structure")
```

```
my.glottodist <- glottodist(my.glottodata)
```

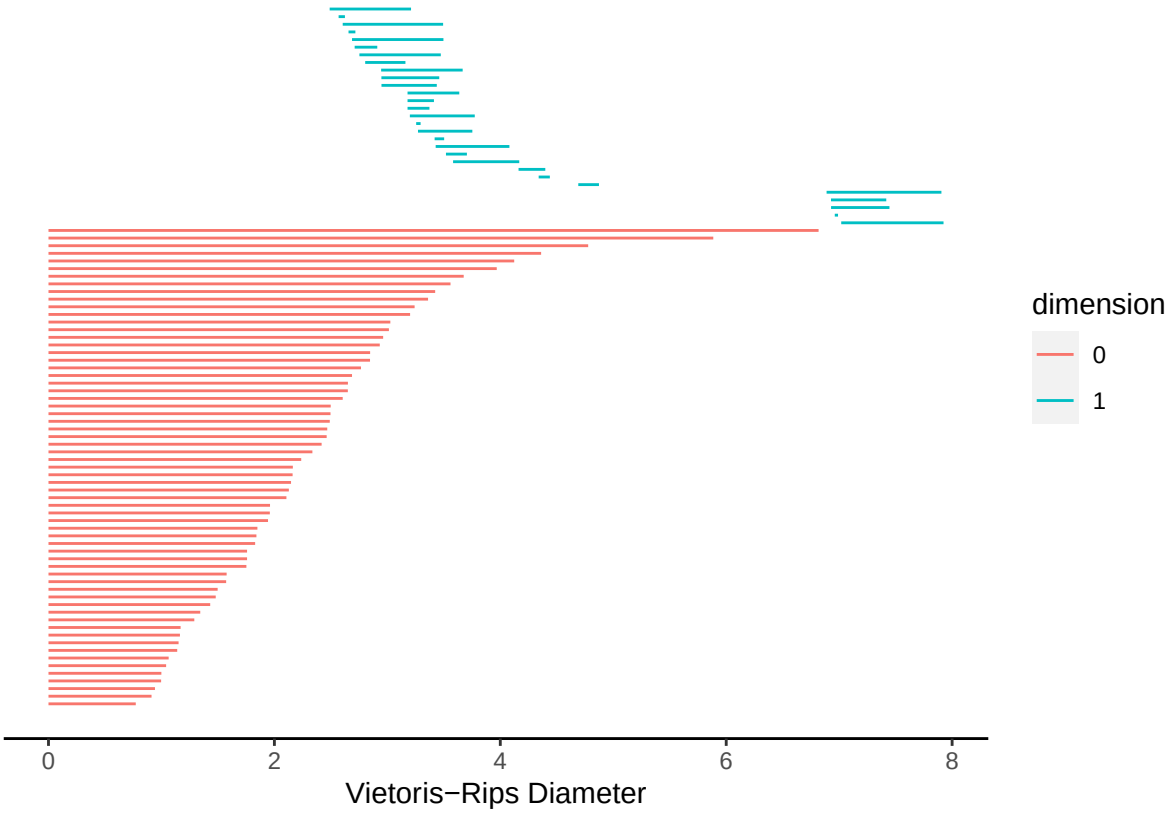
```
## Missing values recoded to NA
## All variables have two or more levels (excluding NA)
```

```
my.glottodist[is.na(my.glottodist)] <- 0
```

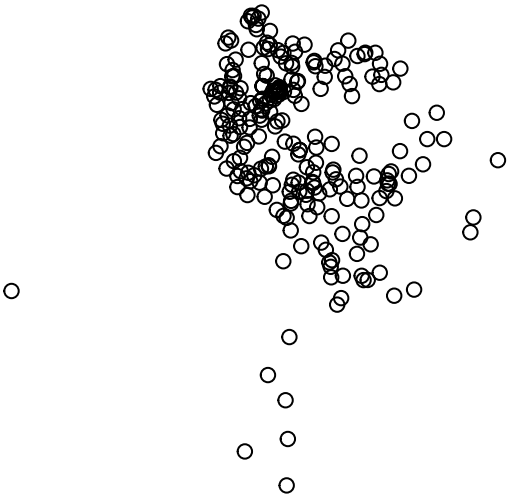
```
my.glottodata.phom <- calculate_homology(my.glottodist)
plot_persist(my.glottodata.phom)
```




```
plot_barcode(my.glottodata.phom)
```




```
plot(st_geometry(wals_sam))
```



```
my.glottodata_1 <- glottoclean(my.glottodata)
```

```
## Missing values recoded to NA
```

```
structure <- my.glottodata[["structure"]]
my.glottodata_1 <- glottosimplify(my.glottodata_1)
my.glottodata_1 <- tibble::column_to_rownames(my.glottodata_1, "glottocode")
```

```
# Hausdorff distance
glotto.hausdorff.dist <- function(glottodata1, glottodata2) {
```

```
}
```

```
overlap.dist <- function (x, y){
  sum(x != y) / length(x)
}

overlap.x.Y.dist <- function(x, Y) {
  result <- c()
  for (i in 1:nrow(Y)) {
    result <- append(result, overlap.dist(x, Y[i,]))
  }
  return(min(result))
}

overlap.supX.Y.dist <- function(X, Y) {
  result <- c()
  for (i in 1:nrow(X)) {
    result <- append(result, overlap.x.Y.dist(X[i, ], Y))
  }
  return(max(result))
}

overlap.hausdorff.dist <- function(X, Y) {

  result <- max(overlap.supX.Y.dist(X, Y), overlap.supX.Y.dist(Y,X))
  return(result)
}
```

```
wals_asia <- subset(wals, continent=="Asia")
data_asia <- select(wals_asia, 'glottocode', '81A', '82A', '83A', '86A', '129A')
data_asia <- st_drop_geometry(data_asia)

data_asia$`81A` <- as.factor(data_asia$`81A`)
data_asia$`82A` <- as.factor(data_asia$`82A`)
data_asia$`83A` <- as.factor(data_asia$`83A`)
data_asia$`86A` <- as.factor(data_asia$`86A`)
data_asia$`129A` <- as.factor(data_asia$`129A`)

str(data_asia)
```

```
## 'data.frame':    599 obs. of  6 variables:
##  $ glottocode: chr  "nort3139" "abkh1244" "abun1252" "abui1241" ...
##  $ 81A       : Factor w/ 6 levels "No dominant order",..: NA 3 4 3 1 NA 3 6 6 NA ...
##  $ 82A       : Factor w/ 3 levels "No dominant order",..: NA 2 2 2 1 2 2 3 3 NA ...
##  $ 83A       : Factor w/ 3 levels "No dominant order",..: NA 2 3 2 1 2 2 3 3 NA ...
##  $ 86A       : Factor w/ 3 levels "Genitive-Noun",..: NA 1 1 1 3 1 1 3 3 NA ...
##  $ 129A      : Factor w/ 2 levels "Different","Identical": NA 1 NA NA NA NA NA NA NA NA ...
```

```
data_asia.unique <- match(unique(data_asia$glottocode), data_asia$glottocode)
data_asia.unique <- data_asia[data_asia.unique, ]
str(data_asia.unique)
```

```
## 'data.frame':    561 obs. of  6 variables:
##  $ glottocode: chr  "nort3139" "abkh1244" "abun1252" "abui1241" ...
##  $ 81A       : Factor w/ 6 levels "No dominant order",..: NA 3 4 3 1 NA 3 6 6 NA ...
```

```
##  $ 82A       : Factor w/ 3 levels "No dominant order",..: NA 2 2 2 1 2 2 3 3 NA ...
##  $ 83A       : Factor w/ 3 levels "No dominant order",..: NA 2 3 2 1 2 2 3 3 NA ...
##  $ 86A       : Factor w/ 3 levels "Genitive-Noun",..: NA 1 1 1 3 1 1 3 3 NA ...
##  $ 129A      : Factor w/ 2 levels "Different","Identical": NA 1 NA NA NA NA NA NA NA NA ...
```

```
glottodata.asia <- glottocreate_addtable(data_asia.unique, structure, name="structure")
```

```
glottodata.asia_1 <- glottoclean(glottodata.asia)
```

```
## Missing values recoded to NA
```

```
structure <- glottodata.asia_1[["structure"]]
glottodata.asia_1 <- glottosimplify(glottodata.asia_1)
glottodata.asia_1 <- tibble::column_to_rownames(glottodata.asia_1, "glottocode")
```

```
my.glottodata_1 <- data.frame(lapply(my.glottodata_1, as.character), stringsAsFactors = FALSE)
my.glottodata_1[is.na(my.glottodata_1)] <- "unknown"

glottodata.asia_1 <- data.frame(lapply(glottodata.asia_1, as.character), stringsAsFactors = FALSE)
glottodata.asia_1[is.na(glottodata.asia_1)] <- "unknown"
```

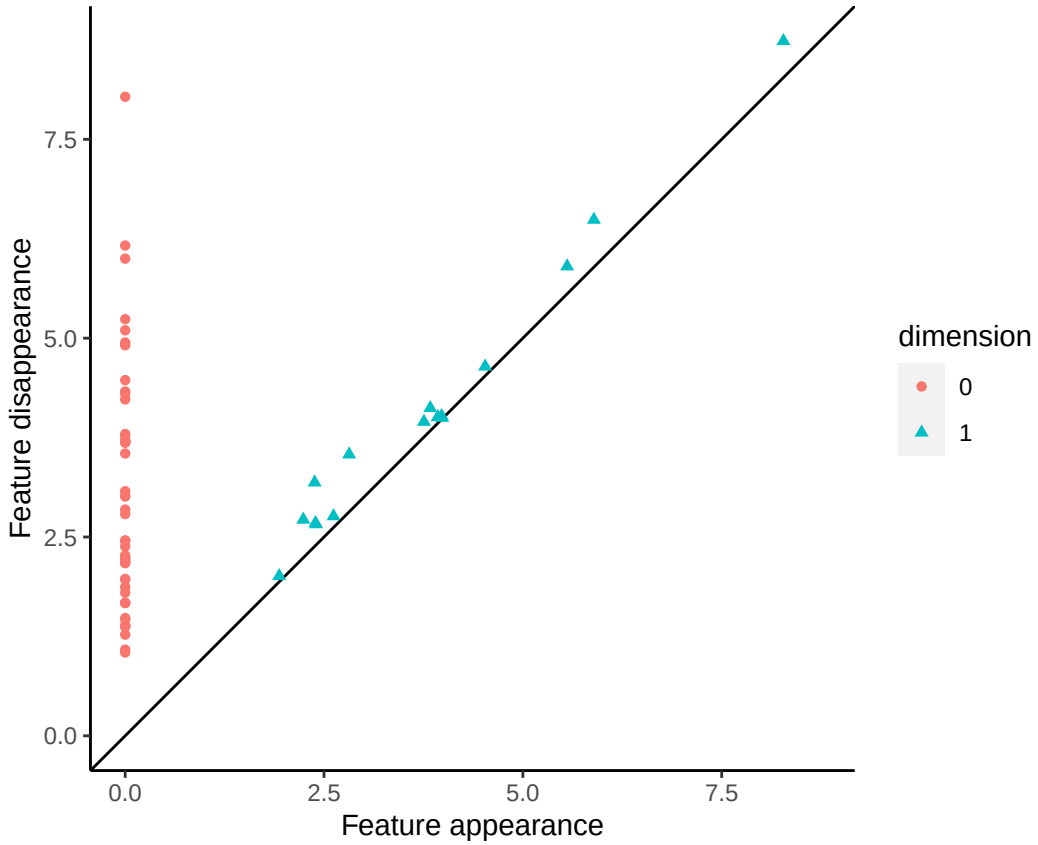
```
# overlap.hausdorff.dist(glottodata.asia_1, my.glottodata_1)
```

```
asia.dist <- glottodist(glottodata.asia)
```

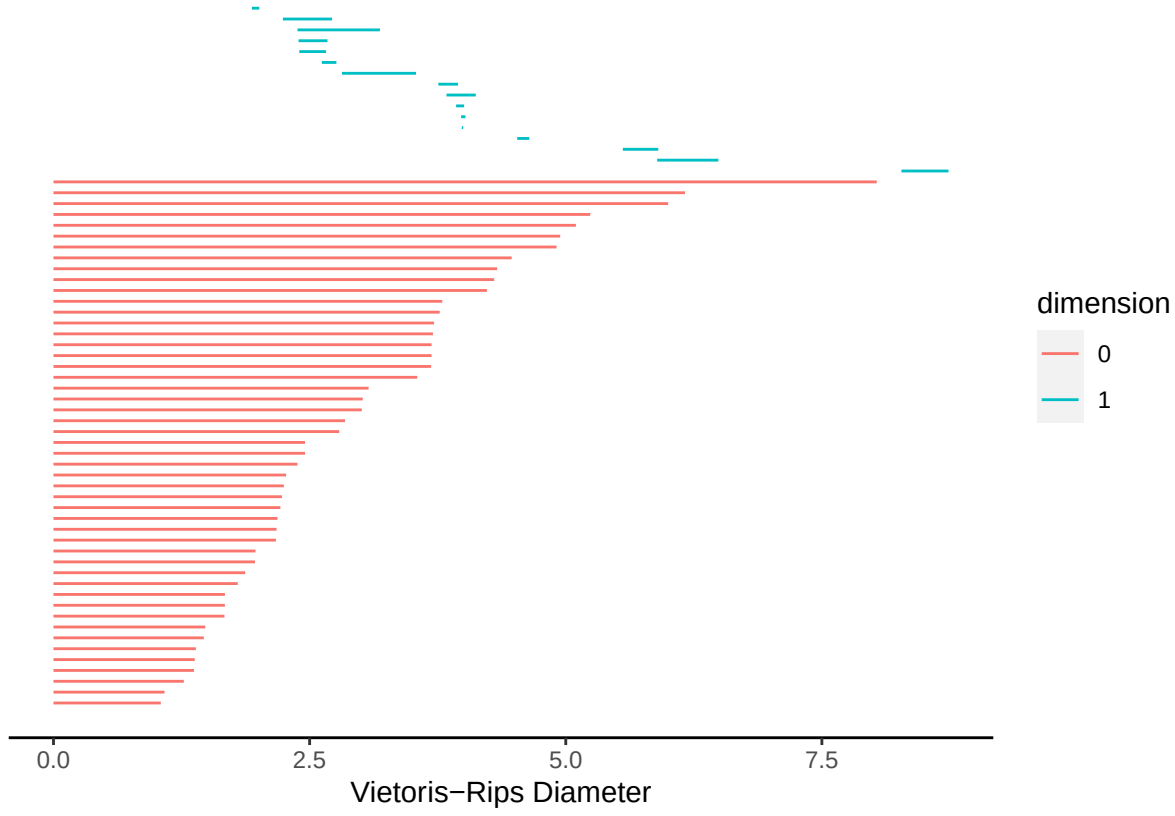
```
## Missing values recoded to NA
```

```
## All variables have two or more levels (excluding NA)
```

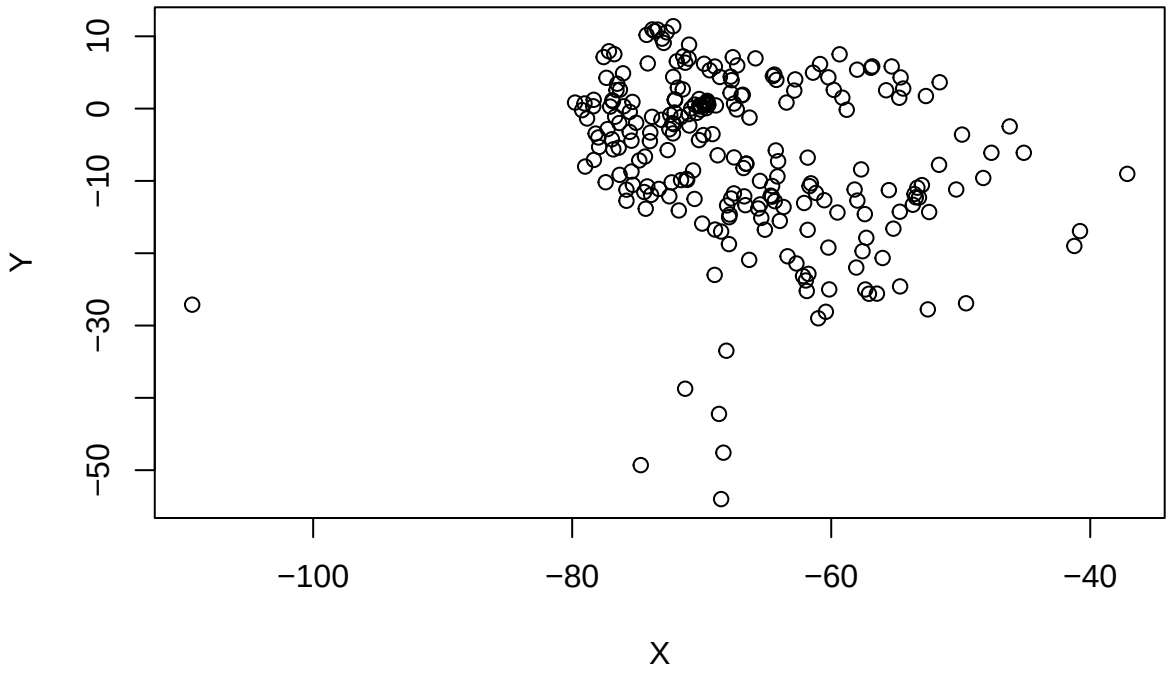
```
asia.dist[is.na(asia.dist)] <- 0
asia_phm <- calculate_homology(asia.dist)
plot_persist(asia_phm)
```



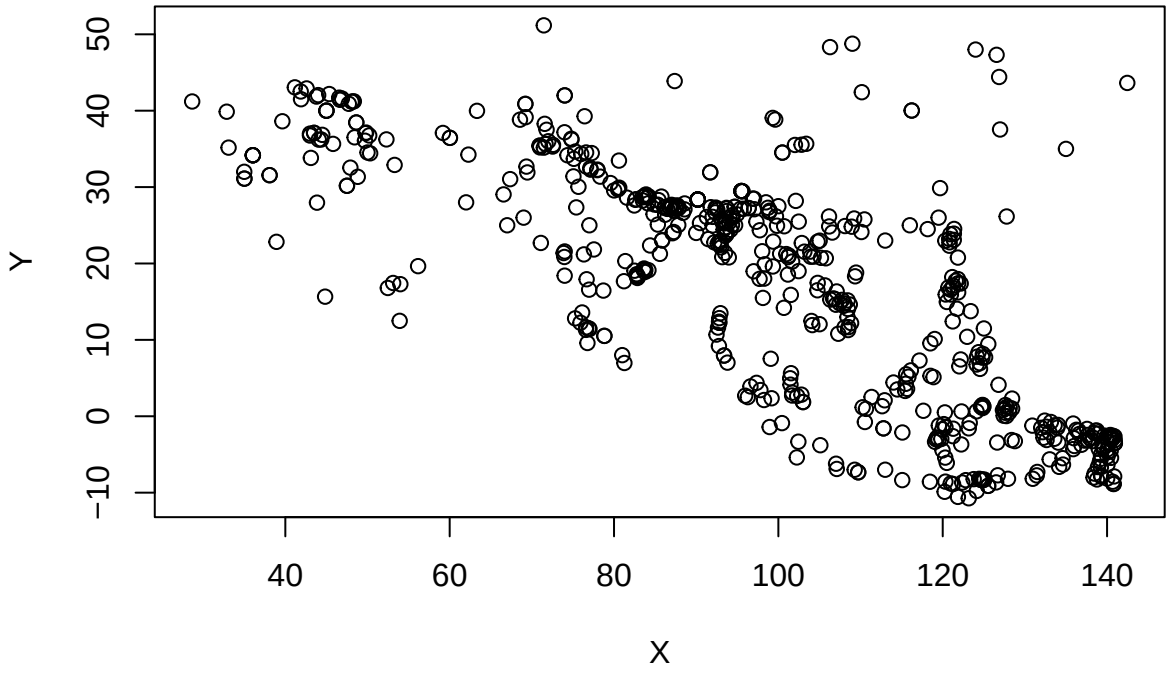

```
plot_barcode(asia_phm)
```



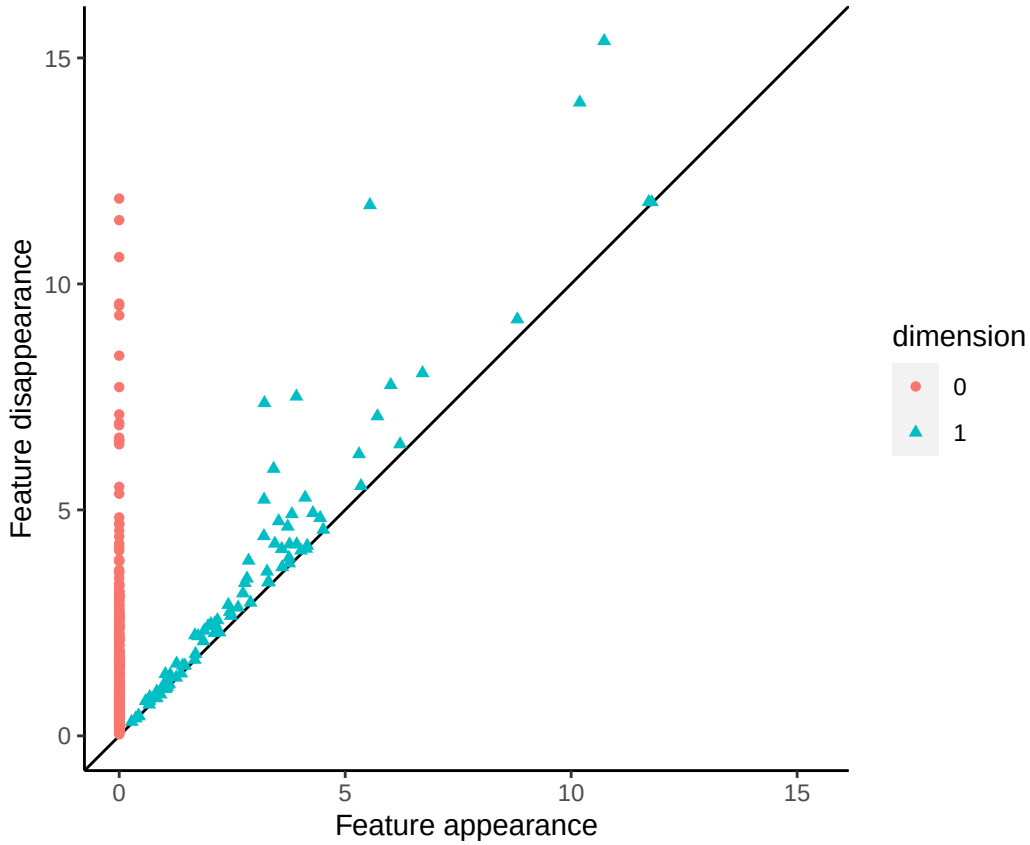
```
plot(st_coordinates(wals_sam))
```



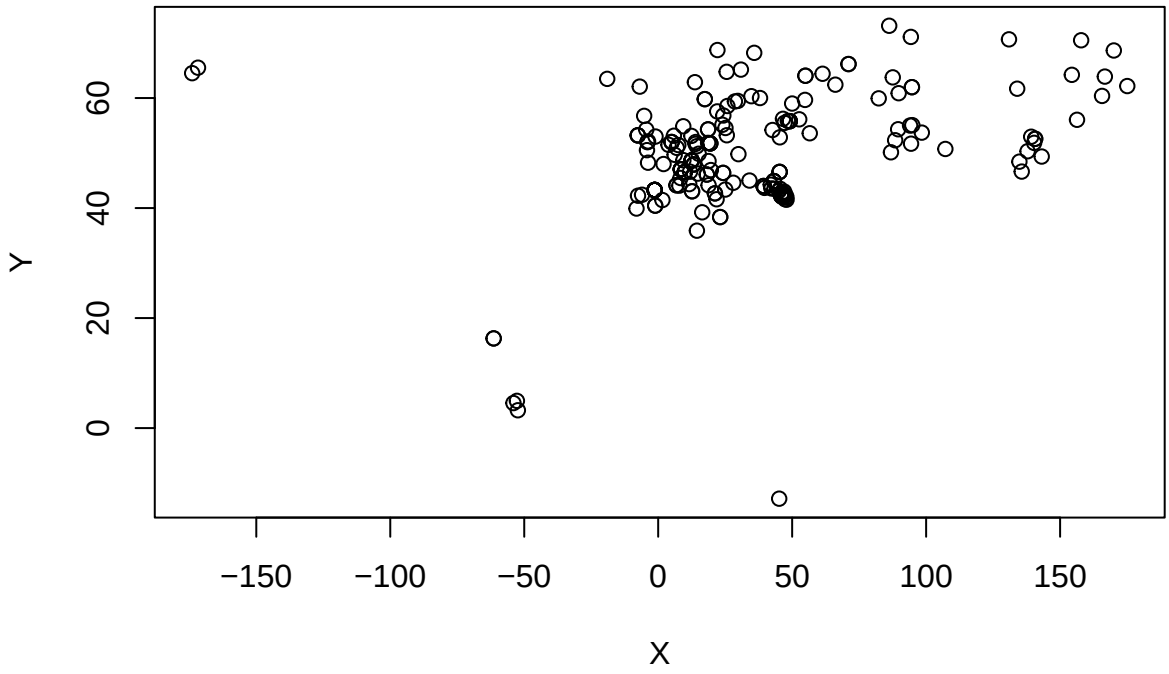

```
plot(st_coordinates(wals_asia))
```



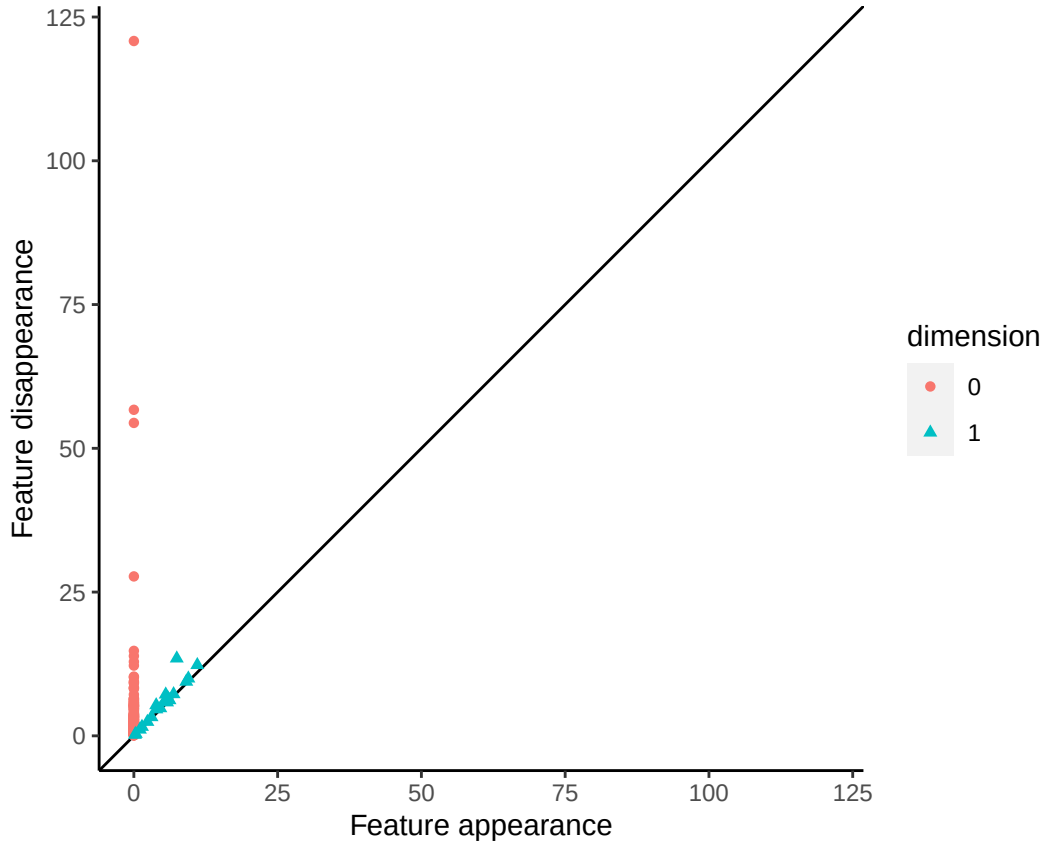

```
plot_persist(calculate_homology(st_coordinates(wals_asia)))
```



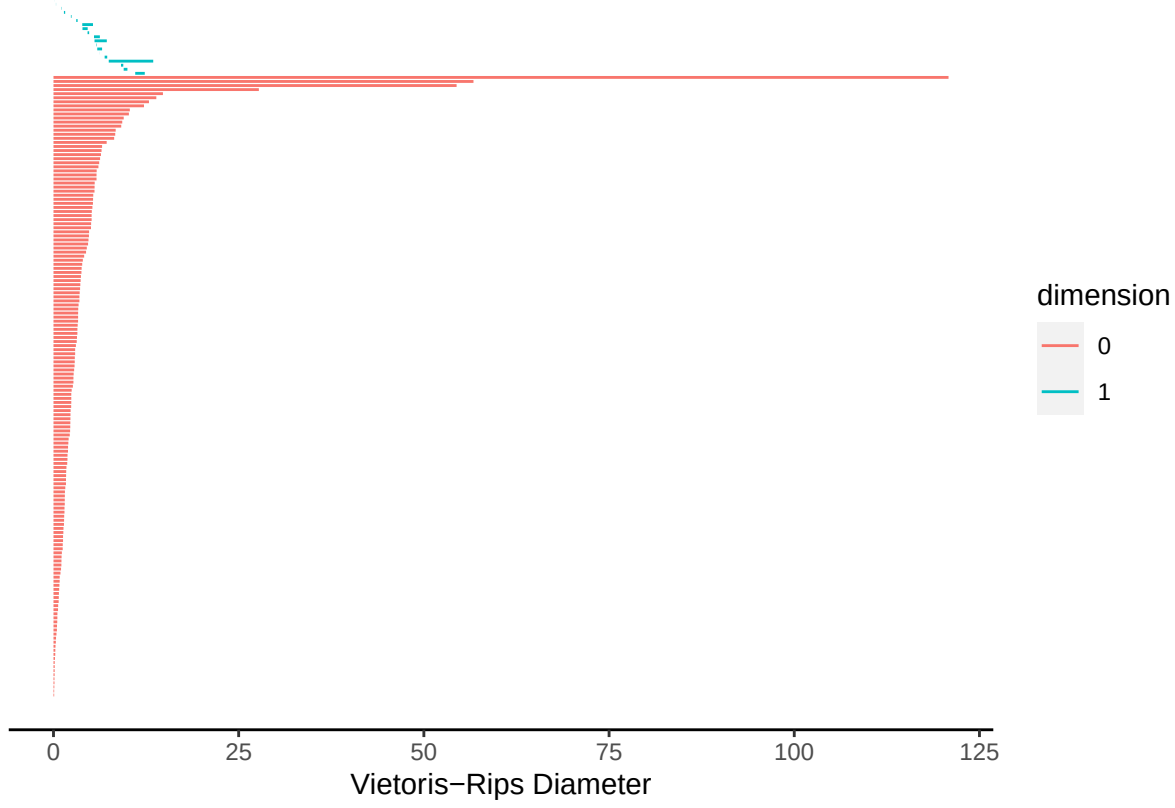

```
wals_europe <- subset(wals, continent=="Europe")
plot(st_coordinates(wals_europe))
```




```
europe_phm <- calculate_homology(st_coordinates(wals_europe))
plot_persist(europe_phm)
```



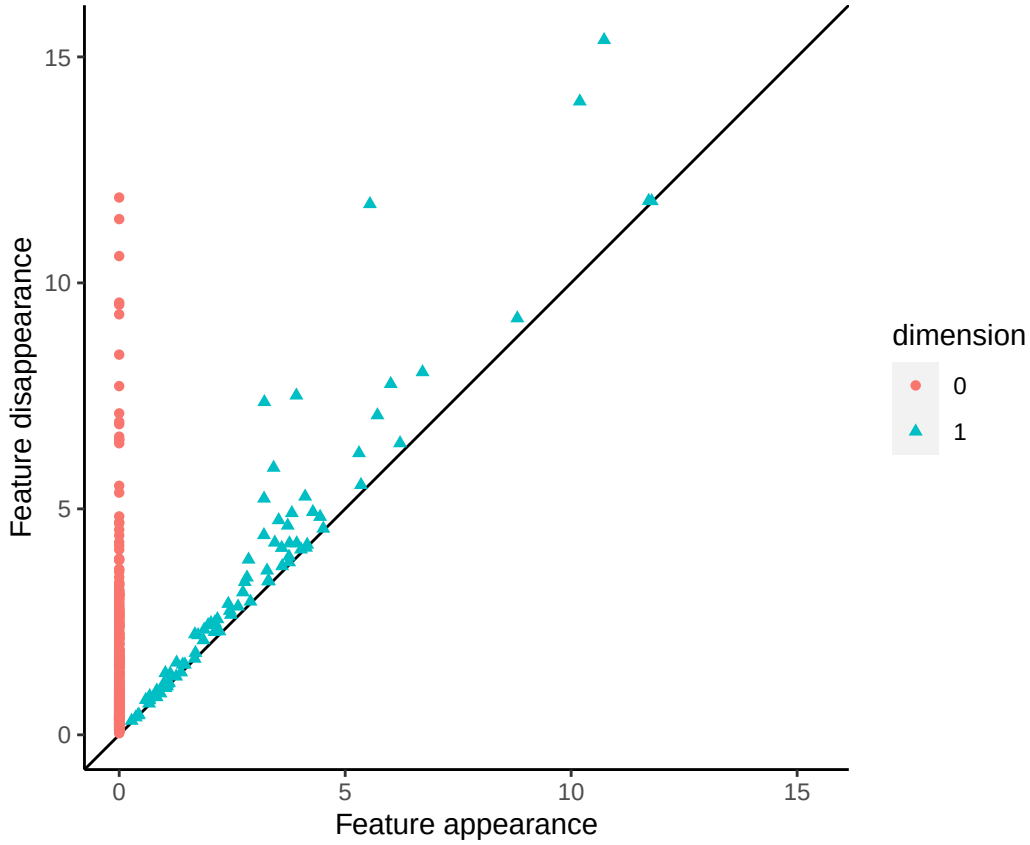

```
plot_barcode(europe_phm)
```



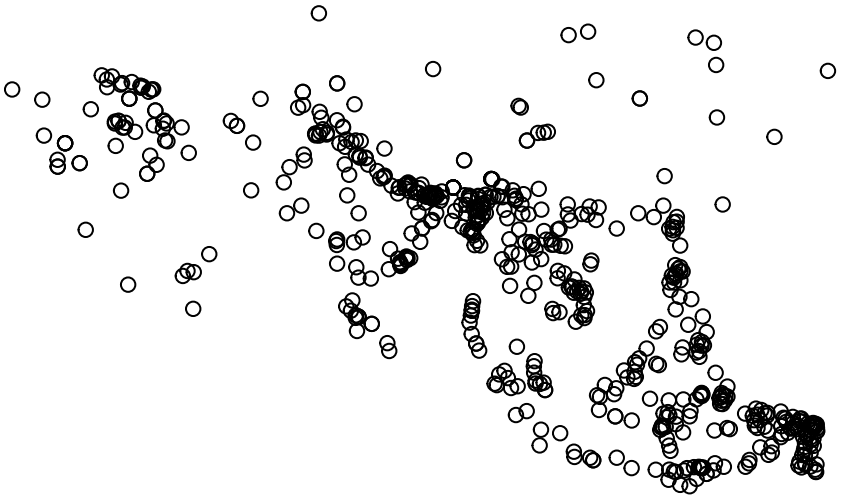
```
wals_asia_geo <- st_geometry(wals_asia)
wals_asia_geo
```

```
## Geometry set for 599 features
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 28.6632 ymin: -10.7326 xmax: 142.4617 ymax: 51.17
## Geodetic CRS:  WGS 84
## First 5 geometries:
## POINT (36.0468 34.1709)
## POINT (41.15911 43.05622)
## POINT (132.416 -0.57073)
## POINT (124.588 -8.31058)
## POINT (96.6032 3.90757)
```

```
plot_persist(calculate_homology(st_coordinates(wals_asia_geo)))
```




```
plot(st_geometry(wals_asia))
```



## References


[1] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM*, 2008.

[2] Nieves Atienza, Rocio Gonzalez-Díaz, and Manuel Soriano-Trigueros. On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognition*, 107:107509, nov 2020.