Clustering (2020 Jun)

Hwang Seong-Yun

2022 9 15

SOM cluster

reference1 : https://data-make.tistory.com/91 (https://data-make.tistory.com/91)

reference2 : https://www.statmethods.net/advstats/cluster.html
(https://www.statmethods.net/advstats/cluster.html)

```
water <- read.csv("C:/Users/HSY/Desktop/영산강 수질악화 관련 데이터 정리_결과 포함(220915)/월별 평균 자료/2020년 6월.csv",
sep=",", header=T)
water_name <- water[,1]
water <- water[,-1]
rownames(water) <- water_name
```

Distance matrix

```
water_scale <- scale(water)
d <- dist(water_scale, method="euclidean")
as.matrix(d)
```

```
##               우치      광주1     방류수   광주천2      광주2      광주3    황룡강5
## 우치     0.000000 1.637389 4.004057 2.222388 2.7915515 2.763285 2.476397
## 광주1    1.637389 0.000000 3.939917 2.492541 2.9647931 3.473414 1.499912
## 방류수   4.004057 3.939917 0.000000 3.287200 4.7926310 4.485092 3.341339
## 광주천2  2.222388 2.492541 3.287200 0.000000 3.0009386 3.010273 2.383733
## 광주2    2.791552 2.964793 4.792631 3.000939 0.0000000 1.231712 3.659956
## 광주3    2.763285 3.473414 4.485092 3.010273 1.2317121 0.000000 4.056628
## 황룡강5  2.476397 1.499912 3.341339 2.383733 3.6599560 4.056628 0.000000
## 광산     2.932647 3.120806 4.937223 3.294799 0.3859292 1.252986 3.805520
##                광산
## 우치     2.9326466
## 광주1    3.1208061
## 방류수   4.9372229
## 광주천2  3.2947990
## 광주2    0.3859292
## 광주3    1.2529858
## 황룡강5  3.8055197
## 광산     0.0000000
```

Decide number of clusters

find the optimal number of clusters using Total within-cluster sum of squares

```
tot_withinss <- c()
for (i in 1:7){
  set.seed(1004) # for reproducibility
  kmeans_cluster <- kmeans(water_scale, centers = i, iter.max = 1000)
  tot_withinss[i] <- kmeans_cluster$tot.withinss}
plot(c(1:7), tot_withinss, type="b",
     main="Optimal number of clusters",
     xlab="Number of clusters",
     ylab="Total within-cluster sum of squares")
```

Optimal number of clusters
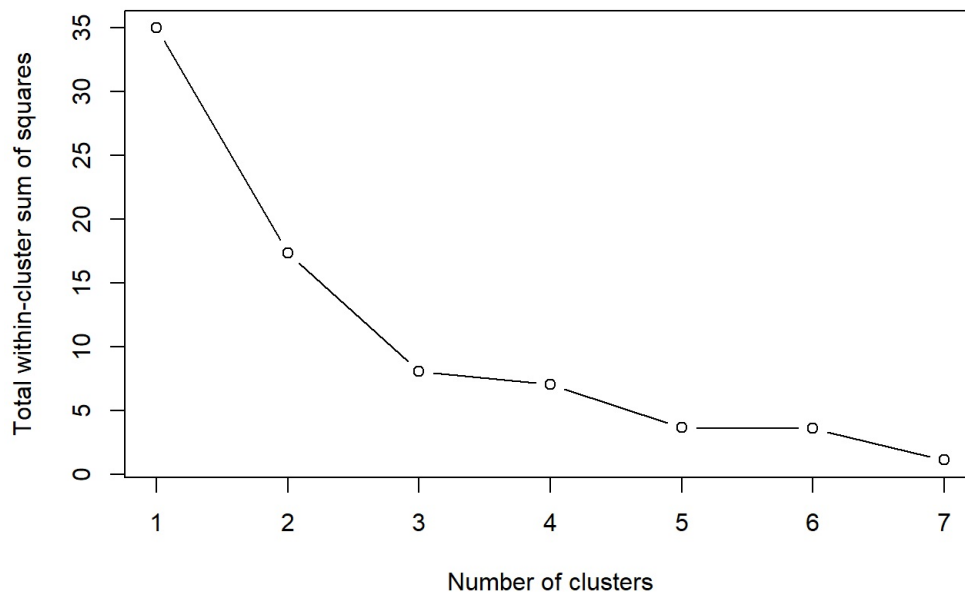

```
fit <- hclust(d, method="ward.D")
plot(fit)
rect.hclust(fit, k=3)
```

Cluster Dendrogram

d
hclust (*, "ward.D")
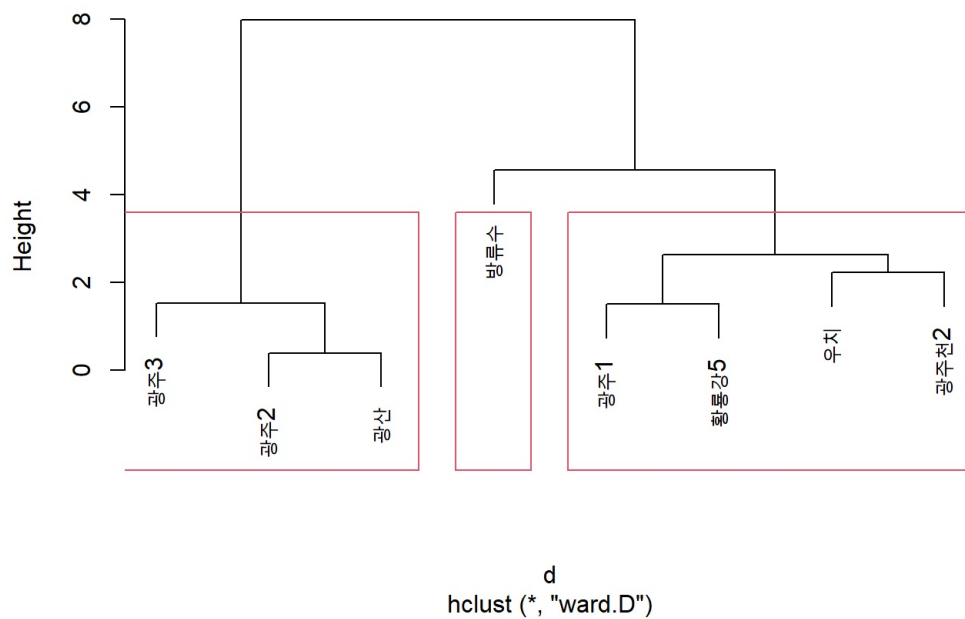
SOM cluster

```
library(SOMbrero)
```

```
## Warning: 패키지 'SOMbrero'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: igraph
```

```
## Warning: 패키지 'igraph'는 R 버전 4.1.2에서 작성되었습니다
```

```
##
## 다음의 패키지를 부착합니다: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
## 필요한 패키지를 로딩중입니다: markdown
```

```
##
```

```
## **********************************************************
```

```
##
```

```
##         This is 'SOMbrero' package, v 1.4.1
```

```
##
```

```
## Citation details with citation('SOMbrero')
```

```
##
```

```
## Further information with help(SOMbrero)...
```

```
##
```

```
## Use sombreroGUI() to start the Graphical Interface.
```

```
##
```

```
## **********************************************************
```

```
library(kohonen)
```

```
## Warning: 패키지 'kohonen'는 R 버전 4.1.3에서 작성되었습니다
```

Normalization of data

```
water_scale <- data.frame(scale(water))
water_scale_matrix <- as.matrix(water_scale)
```

Training the SOM model

```
som_grid <- somgrid(xdim=1, ydim=3, topo="hexagonal")
som_model1 <- som(water_scale_matrix, grid=som_grid)
som_model2 <- trainSOM(x.data=water_scale, dimension=c(1,3),
                       nb.save=10, maxit=2000, scaling="none",
                       radius.type="letremy")
```
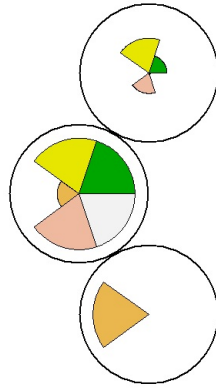
Visualization

```
table(som_model2$clustering)
```

```
##
## 1 2 3
## 3 2 3
```

```
plot(som_model1, main="feature distribution")
```

feature distribution

■ BOD	■ T_N	□ Q
■ COD	■ T_P	

```
plot(som_model2, what="obs", type="names", print.title=T, scale=c(1,1))
```

```
## Warning in plot.somRes(som_model2, what = "obs", type = "names", print.title =
## T, : 'print.title' will be deprecated, please use 'show.names' instead
```

Observations overview

repartition of row.names values

3
방류수 광주천2 황룡강5

2
광주1우치

1
광주2광산 광주3