Clustering (2020 Mar)

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 3월.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      광산
## 우치     0.000000 1.871498 3.135013 2.753959 4.126206 3.900584 1.060684 3.847092
## 광주1    1.871498 0.000000 3.806471 1.982194 3.078613 3.092292 1.934801 3.120481
## 방류수   3.135013 3.806471 0.000000 3.245658 3.993056 3.203741 3.850040 3.503029
## 광주천2  2.753959 1.982194 3.245658 0.000000 3.071559 2.412160 3.376742 2.923344
## 광주2    4.126206 3.078613 3.993056 3.071559 0.000000 1.555009 4.115113 2.168028
## 광주3    3.900584 3.092292 3.203741 2.412160 1.555009 0.000000 4.260142 1.194037
## 황룡강5  1.060684 1.934801 3.850040 3.376742 4.115113 4.260142 0.000000 4.161144
## 광산     3.847092 3.120481 3.503029 2.923344 2.168028 1.194037 4.161144 0.000000
```

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
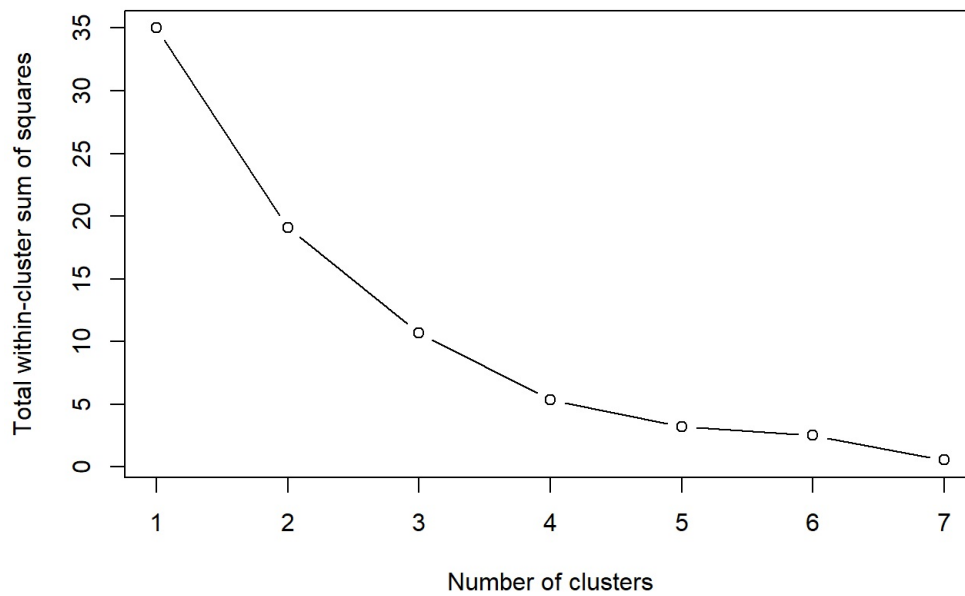

```
fit <- hclust(d, method="ward.D")
plot(fit)
rect.hclust(fit, k=3)
```

Cluster Dendrogram

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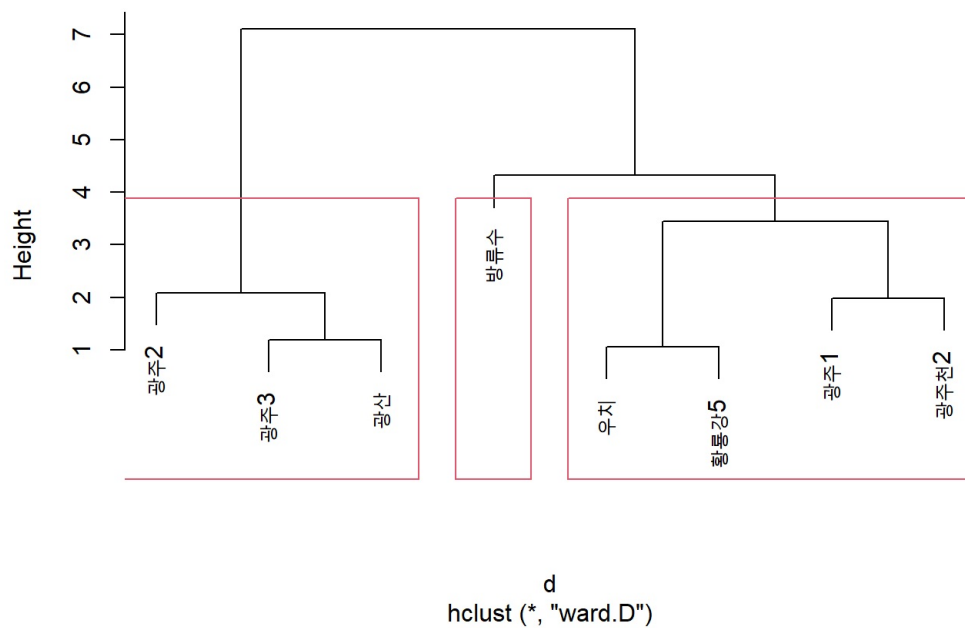
SOM cluster

```
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'
```

```
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")
```
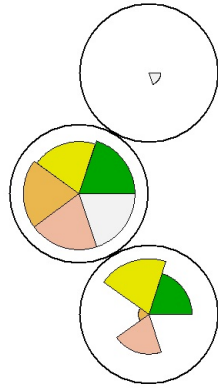
Visualization

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

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

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

feature distribution

Legend: BOD, COD, T_N, T_P, Q

```
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