Clustering (2020 Feb)

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 2월.csv",
sep=",", header=T)
water_name <- water[,1]
water <- water[,-1]
rownames(water) <- water_name
```

Distance matrix

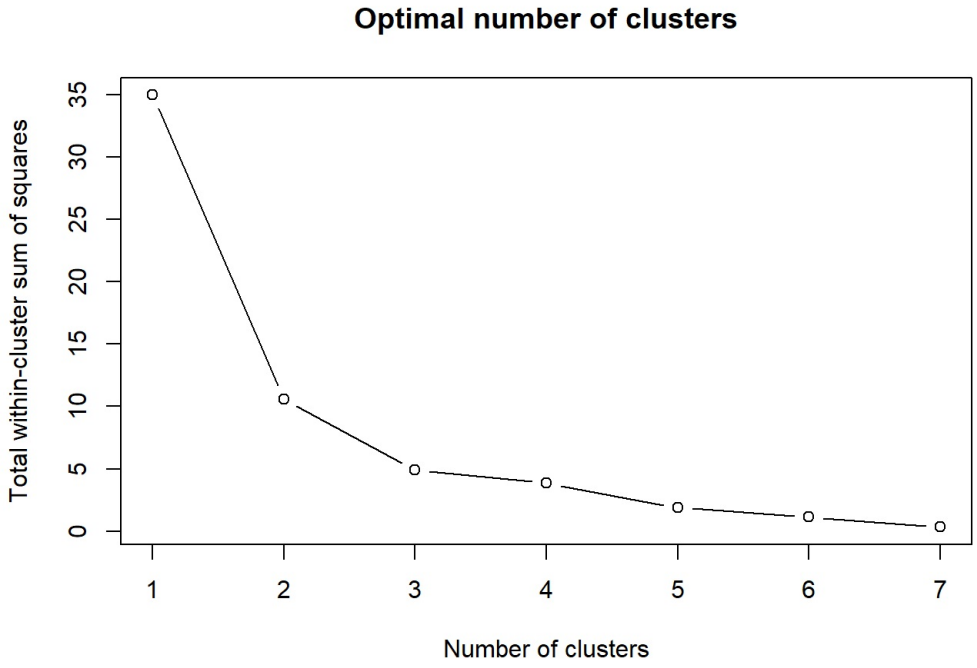
```
water_scale <- scale(water)
d <- dist(water_scale, method="euclidean")
as.matrix(d)
```

```
##              우치    광주1   방류수  광주천2    광주2    광주3   황룡강5
## 우치    0.0000000 1.447162 2.763250 2.118762 4.813944 4.119954 0.8115935
## 광주1   1.4471617 0.000000 2.897150 1.255349 3.843078 3.308855 0.9917850
## 방류수  2.7632502 2.897150 0.000000 2.548941 3.833343 2.960333 2.9744465
## 광주천2 2.1187624 1.255349 2.548941 0.000000 3.714253 3.275791 1.5754790
## 광주2   4.8139442 3.843078 3.833343 3.714253 0.000000 1.237957 4.6214303
## 광주3   4.1199536 3.308855 2.960333 3.275791 1.237957 0.000000 4.0753445
## 황룡강5 0.8115935 0.991785 2.974447 1.575479 4.621430 4.075344 0.0000000
## 광산    4.4625503 3.450359 3.828059 3.687806 1.559395 1.256418 4.3485815
##             광산
## 우치    4.462550
## 광주1   3.450359
## 방류수  3.828059
## 광주천2 3.687806
## 광주2   1.559395
## 광주3   1.256418
## 황룡강5 4.348582
## 광산    0.000000
```

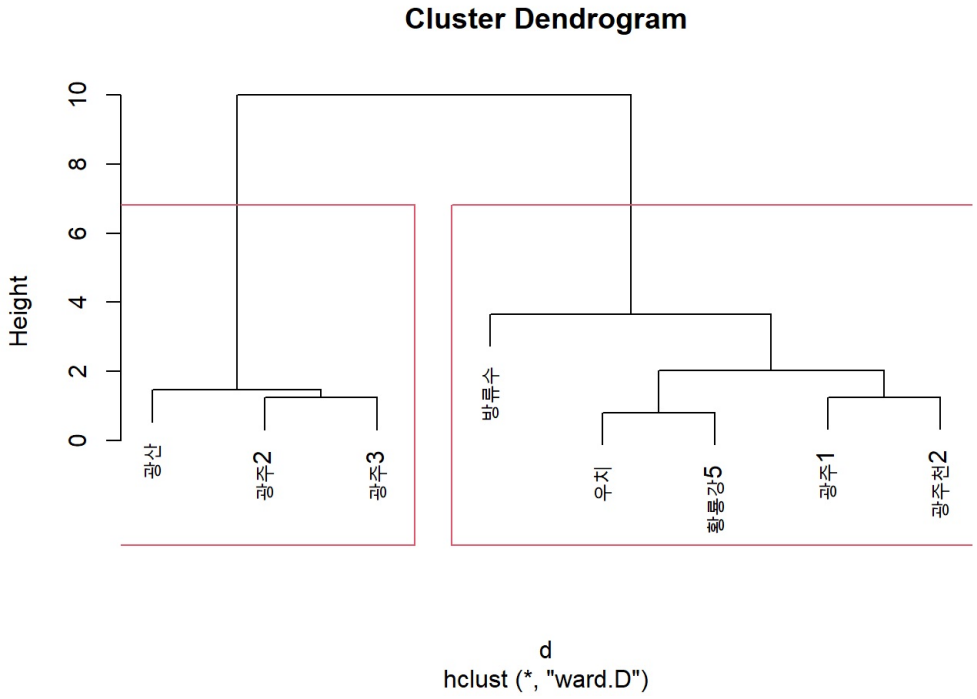
Decide number of clusters

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")
```




```
fit <- hclust(d, method="ward.D")
plot(fit)
rect.hclust(fit, k=2)
```



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

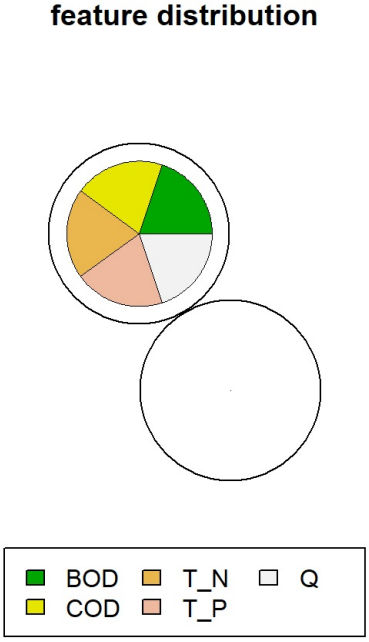
```
som_grid <- somgrid(xdim=1, ydim=2, topo="hexagonal")
som_model1 <- som(water_scale_matrix, grid=som_grid)
som_model2 <- trainSOM(x.data=water_scale, dimension=c(1,2),
                       nb.save=10, maxit=2000, scaling="none",
                       radius.type="letremy")
```

Visualization

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

```
##
## 1 2
## 5 3
```

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




```
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