Clustering (2020 Apr)

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 4월.csv",
sep=",", header=T)
water_name <- water[,1]
water <- water[,-1]
rownames(water) <- water_name
```

Distance matrix

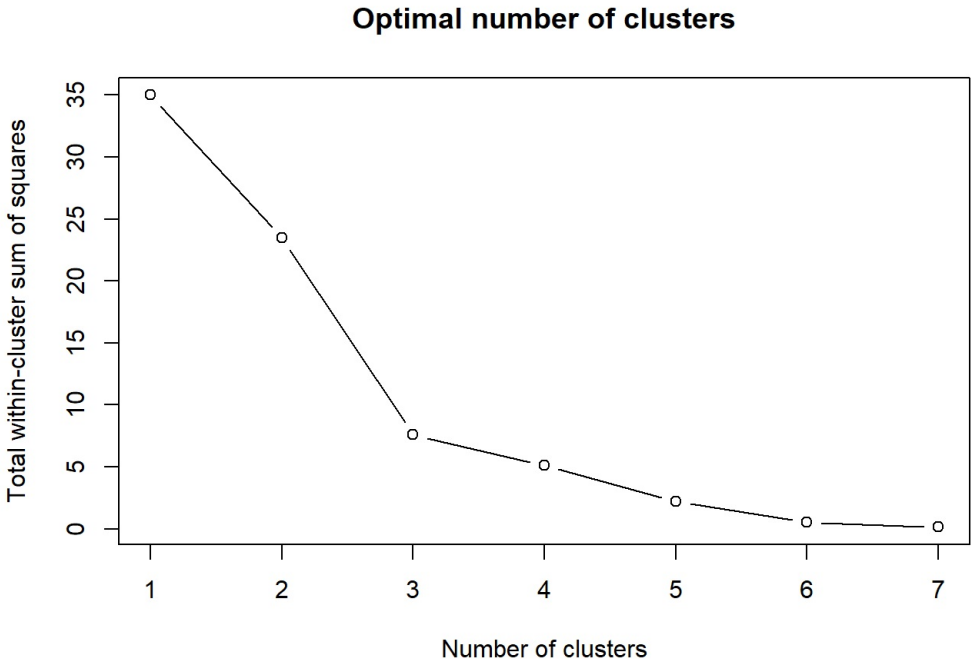
```
water_scale <- scale(water)
d <- dist(water_scale, method="euclidean")
as.matrix(d)
```

```
##              우치     광주1   방류수  광주천2    광주2    광주3   황룡강5
## 우치    0.0000000 0.7471910 3.878018 1.868324 3.679517 2.669511 0.5404575
## 광주1   0.7471910 0.0000000 4.287481 1.887646 3.298897 2.518380 0.8400785
## 방류수  3.8780182 4.2874812 0.000000 3.317576 4.676966 3.222272 4.1895448
## 광주천2 1.8683244 1.8876461 3.317576 0.000000 3.238438 2.133493 2.2636393
## 광주2   3.6795173 3.2988968 4.676966 3.238438 0.000000 1.849849 3.8821752
## 광주3   2.6695113 2.5183798 3.222272 2.133493 1.849849 0.000000 3.0276271
## 황룡강5 0.5404575 0.8400785 4.189545 2.263639 3.882175 3.027627 0.0000000
## 광산    4.0217971 3.7197736 4.631859 3.401056 2.382719 1.834410 4.4005017
##             광산
## 우치    4.021797
## 광주1   3.719774
## 방류수  4.631859
## 광주천2 3.401056
## 광주2   2.382719
## 광주3   1.834410
## 황룡강5 4.400502
## 광산    0.000000
```

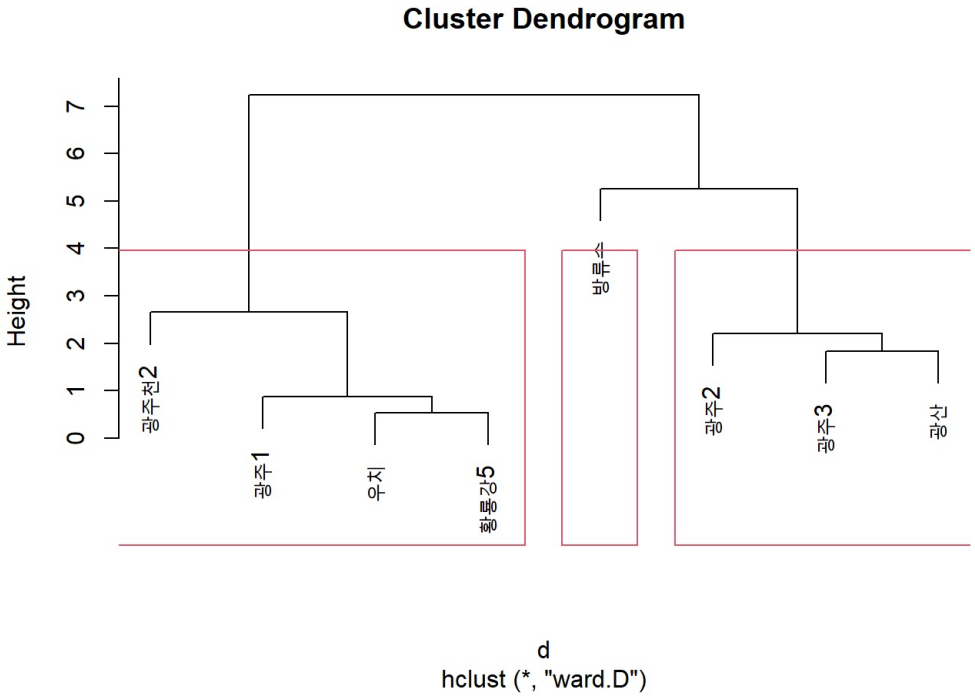
Decide number of clusters

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=3)
```



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

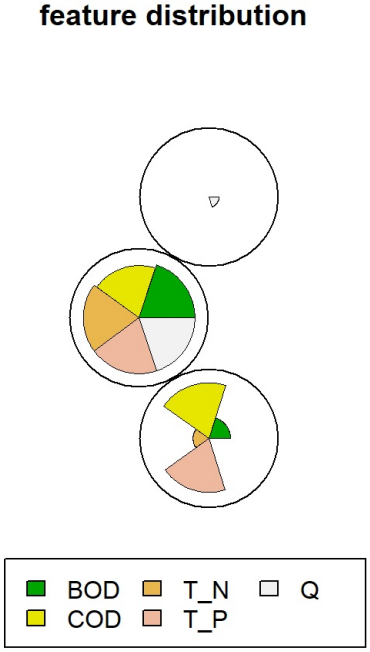
```
som_grid <- somgrid(xdim=1, ydim=3, topo="hexagonal")
som_model1 <- som(water_scale_matrix, grid=som_grid)
som_model2 <- trainSOM(x.data=water_scale, dimension=c(1,3),
                       nb.save=10, maxit=2000, scaling="none",
                       radius.type="letremy")
```

Visualization

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

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

```
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
```