

# Clustering (2022 May)

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/2022년 5월.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)
```

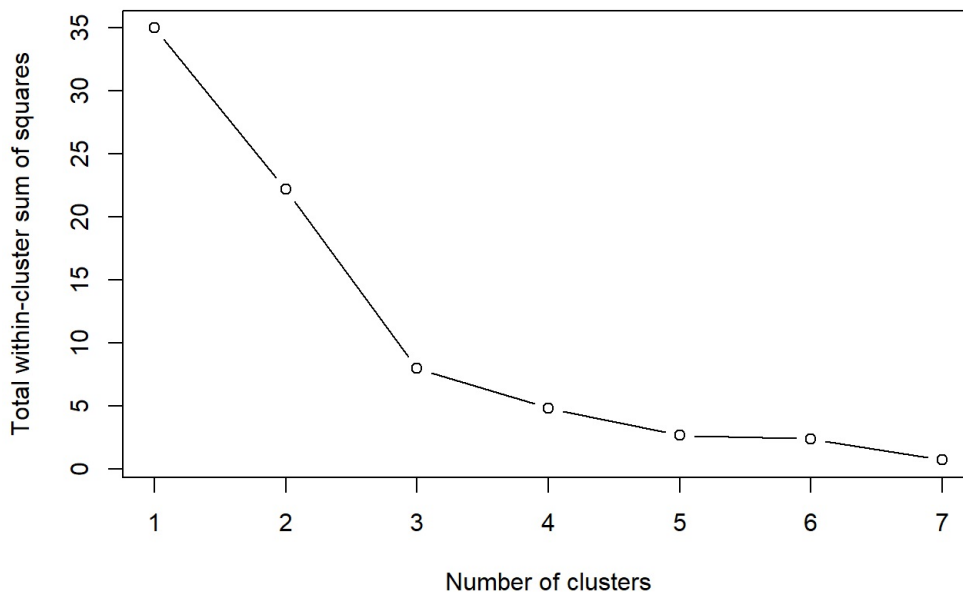
```
##           위치   광주1   방류수   광주천2   광주2   광주3   황룡강5
## 위치      0.000000  1.846787  3.282259  1.808728  2.9255256  3.2200204  1.578270
## 광주1      1.846787  0.000000  4.103433  2.306000  2.7752820  2.7734673  1.524568
## 방류수      3.282259  4.103433  0.000000  4.164367  3.6373379  3.8119557  4.204896
## 광주천2     1.808728  2.306000  4.164367  0.000000  2.5368592  2.7446806  1.200826
## 광주2      2.925526  2.775282  3.637338  2.536859  0.0000000  0.7275358  2.345647
## 광주3      3.220020  2.773467  3.811956  2.744681  0.7275358  0.0000000  2.627415
## 황룡강5     1.578270  1.524568  4.204896  1.200826  2.3456469  2.6274154  0.000000
## 광산       4.758643  3.899447  5.853874  3.870797  2.3410779  2.0615216  3.645594
##           광산
## 위치       4.758643
## 광주1      3.899447
## 방류수      5.853874
## 광주천2     3.870797
## 광주2      2.341078
## 광주3      2.061522
## 황룡강5     3.645594
## 광산       0.000000
```

## Decide number of clusters

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

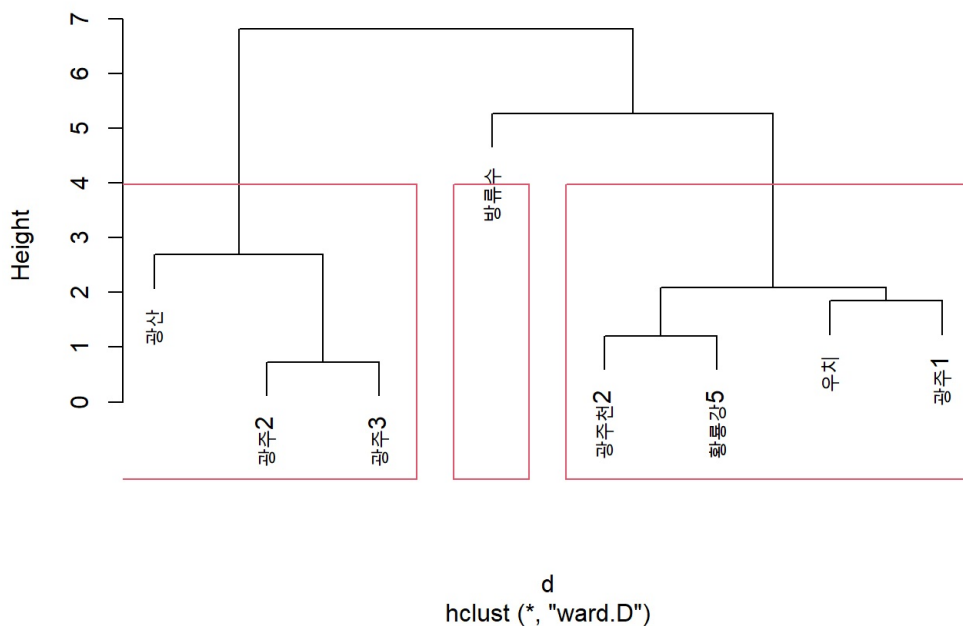
```
tot_withinss <- c()
for (i in 1:7){
  set.seed(1004) # for reproducibility
  kmeans_cluster <- kmeans(water_scale, centers = i, iter.max = 1000)
  tot_withinss[i] <- kmeans_cluster$tot.withinss}
plot(c(1:7), tot_withinss, type="b",
     main="Optimal number of clusters",
     xlab="Number of clusters",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

## Optimal number of clusters



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

## Cluster Dendrogram



## 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")
```

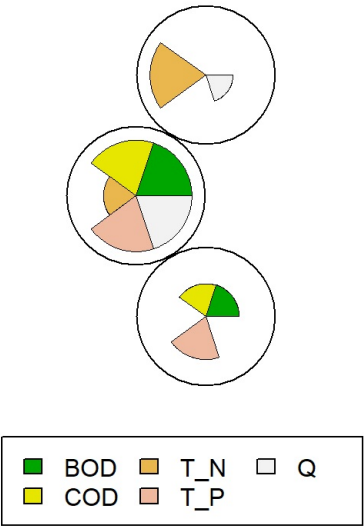
## Visualization

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

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

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

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

