

Clustering (2020 Jan)

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 1월.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)
```

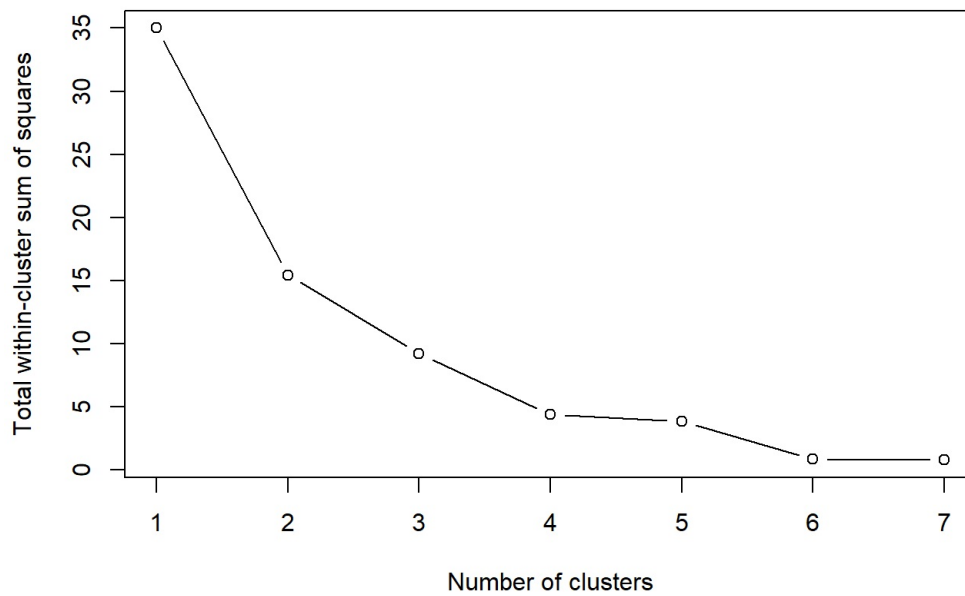
```
##           우치      광주1   방류수   광주천2   광주2   광주3   황룡강5
## 우치      0.000000  0.4079300  2.918294  3.591481  4.189637  4.649102  1.0839641
## 광주1     0.407930  0.0000000  2.982857  3.461234  3.985967  4.598498  0.7308101
## 방류수     2.918294  2.9828565  0.000000  3.366605  3.316307  3.155743  3.4158196
## 광주천2    3.591481  3.4612343  3.366605  0.000000  2.870120  2.863500  3.3496269
## 광주2     4.189637  3.9859674  3.316307  2.870120  0.000000  2.039866  3.8393598
## 광주3     4.649102  4.5984982  3.155743  2.863500  2.039866  0.000000  4.6971792
## 황룡강5    1.083964  0.7308101  3.415820  3.349627  3.839360  4.697179  0.0000000
## 광산       3.078029  2.9131946  2.729264  2.581090  1.270547  2.351402  2.7989795
##           광산
## 우치       3.078029
## 광주1      2.913195
## 방류수      2.729264
## 광주천2     2.581090
## 광주2       1.270547
## 광주3       2.351402
## 황룡강5     2.798980
## 광산        0.000000
```

Decide number of clusters

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

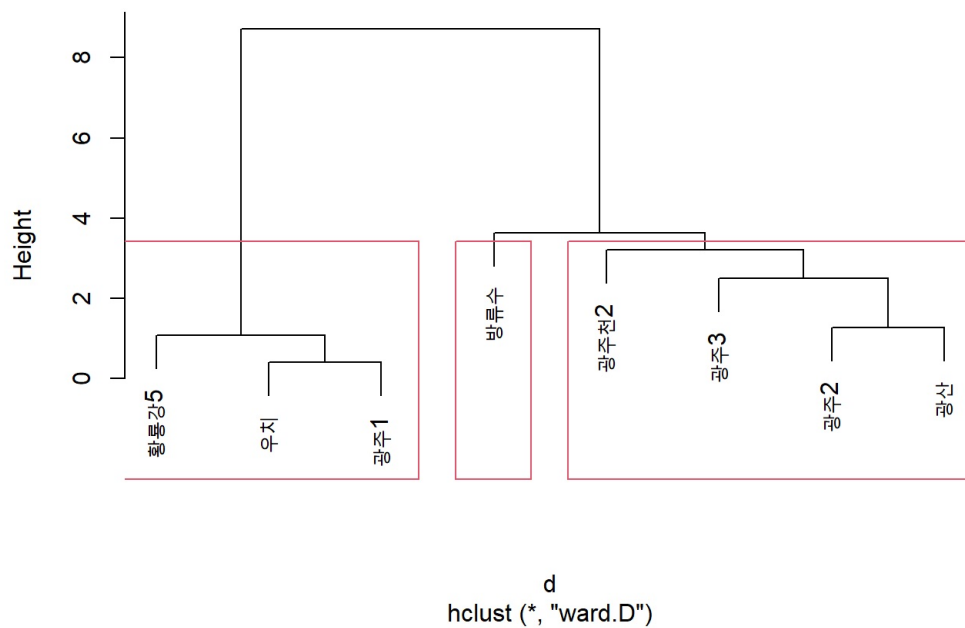
```
tot_withinss <- c()
for (i in 1:7){
  set.seed(1004) # for reproducibility
  kmeans_cluster <- kmeans(water_scale, centers = i, iter.max = 1000)
  tot_withinss[i] <- kmeans_cluster$tot.withinss}
plot(c(1:7), tot_withinss, type="b",
     main="Optimal number of clusters",
     xlab="Number of clusters",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

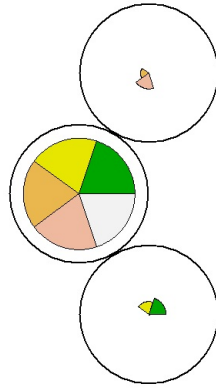
Visualization

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

```
##  
## 1 2 3  
## 4 1 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

3

황룡강5 광주1우치

2

방류수

1

광주천2 광주2 광산 광주3