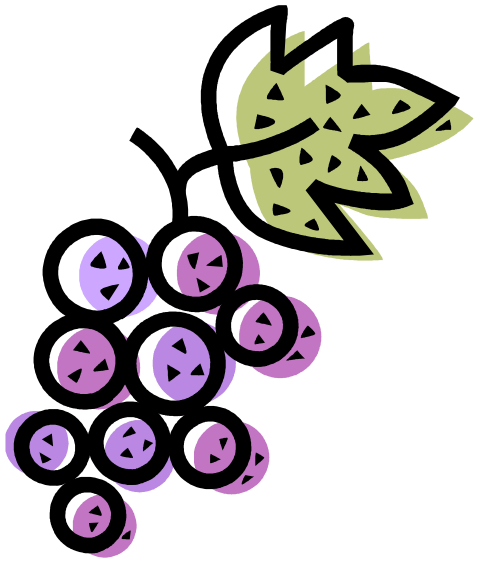
クラスター分析

専門演習B

1176638c 吉田 智洋

14/6/12

Cluster analysis



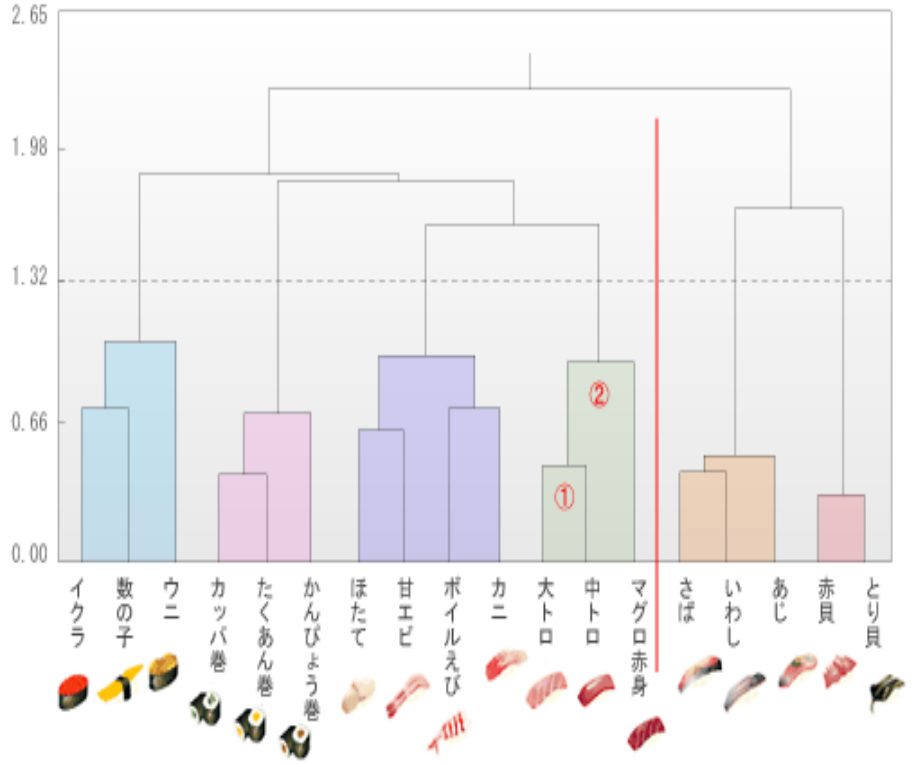
- Cluster
  1.(ブドウ、サクランボなどの）房
  2.(同種のもの、人の）群れ、集合

- 物事を整頓する際
  機能、形状など似ているものを同じところに集めて整頓する

2種類のクラスター分析

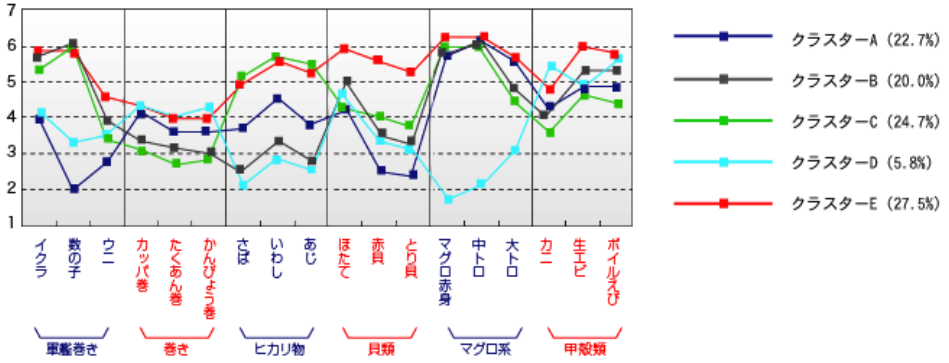
階層クラスター分析

- 樹形図（デントログラム）



非階層クラスター分析

- K平均法
- 大量のデータ分析に向く



階層的クラスター分析

（１）データから距離（あるいは類似度）を求める
（２）分析の方法選択
※最近隣法、最遠隣法、群平均法、重心法、メディアン法、ウォード法
（３）選択された方法のコーフェン行列を求める
（４）コーフェン行列に基づいて樹形図を作成
（５）結果について考察を行う

実際にやってみる

（1）データから距離（あるいは類似度）を求める

R Console

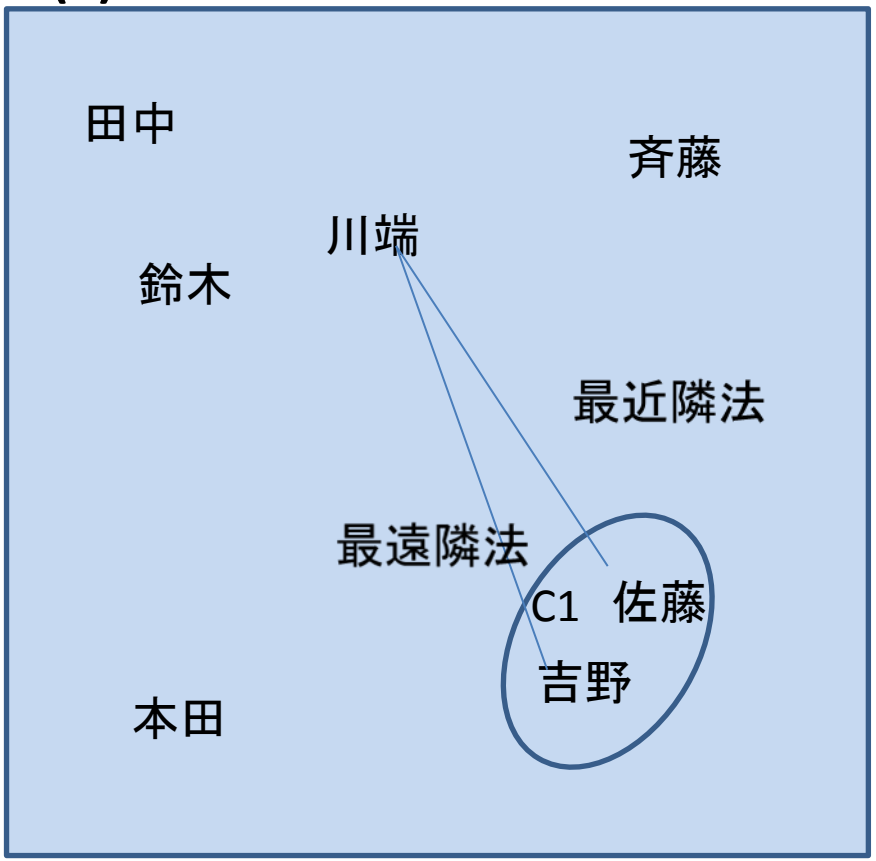
```
> TMP <- read.csv('D://seiseki.csv')
> DAT <- TMP[,2:ncol(TMP)]
> colnames(DAT)
[1] "Math" "Sci"  "Lang" "Eng"  "Soc"
> rownames(DAT)
[1] "1" "2" "3" "4" "5" "6" "7"
> rownames(DAT) <- TMP[,1]
> rownames(DAT)
[1] "Tanaka"   "Sato"     "Suzuki"
[7] "Saito"
> DAT
         Math Sci Lang Eng Soc
Tanaka     89  90   67  46  50
Sato       57  70   80  85  90
Suzuki     80  90   35  40  50
Honda      40  60   50  45  55
Kawabata   78  85   45  55  60
Yoshino    55  65   80  75  85
Saito      90  85   88  92  95
> seiseki.d <- dist(DAT)
> round(seiseki.d)
         Tanaka Sato Suzuki Honda Kawabata Yoshino
Sato         69
Suzuki       34   81
Honda        60   64     53
Kawabata     28   61     21    47
Yoshino      63   12     76    54       56
Saito        68   38     88    92       68      46
```

距離が近いもの同士組み合わせていく
コーフェン距離最小 ”12” C1{吉野、佐藤}
次に小さいのは “21” {川端、鈴木}
{川端、鈴木}？{C1、川端}？{C1、鈴木}？
次は{川端、田中}
{C2、田中}？{C1、田中}？、、、、、、、、、、

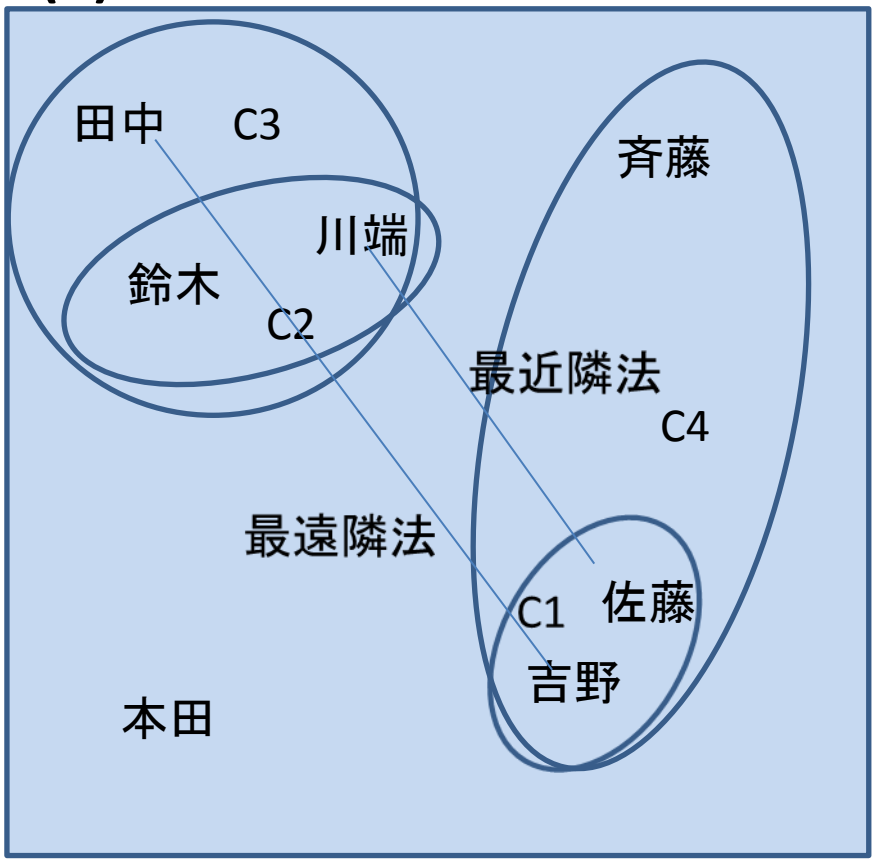
コーフェン距離
値が小さいほど距離が近い

コーフェン距離　参考

(a)クラスターと個体（川端とC1）



(b)クラスターの間（C1とC3）



（2）分析の方法(最近隣法、最遠隣法等）選択

R Console

```
> (sei.hc <- hclust(seiseki.d))

Call:
hclust(d = seiseki.d)

Cluster method   : complete
Distance         : euclidean
Number of objects: 7

> summary(sei.hc)
            Length Class  Mode
merge       12     -none- numeric
height       6     -none- numeric
order        7     -none- numeric
labels       7     -none- character
method       1     -none- character
call         2     -none- call
dist.method  1     -none- character
> sei.hc$merge
     [,1] [,2]
[1,]   -2   -6
[2,]   -3   -5
[3,]   -1    2
[4,]   -7    1
[5,]   -4    3
[6,]    4    5
> sei.hc$height
[1] 12.40967 21.30728 33.77869 45.58509 60.13319 91.53142
> sei.hc$order
[1] 7 2 6 4 1 3 5
```

Complete(最遠隣法)
距離は、ユークリッド

結果オブジェクトに格納された
データのリスト

クラスター形成の履歴
マイナス付が個体の番号、付い
てないのがクラスターの番号

クラスターを形成する樹木図の枝
の長さが記録されている

樹木図の左から個体の番号
が記録されている

（4）コーフェン行列に基づいて樹形図を作成

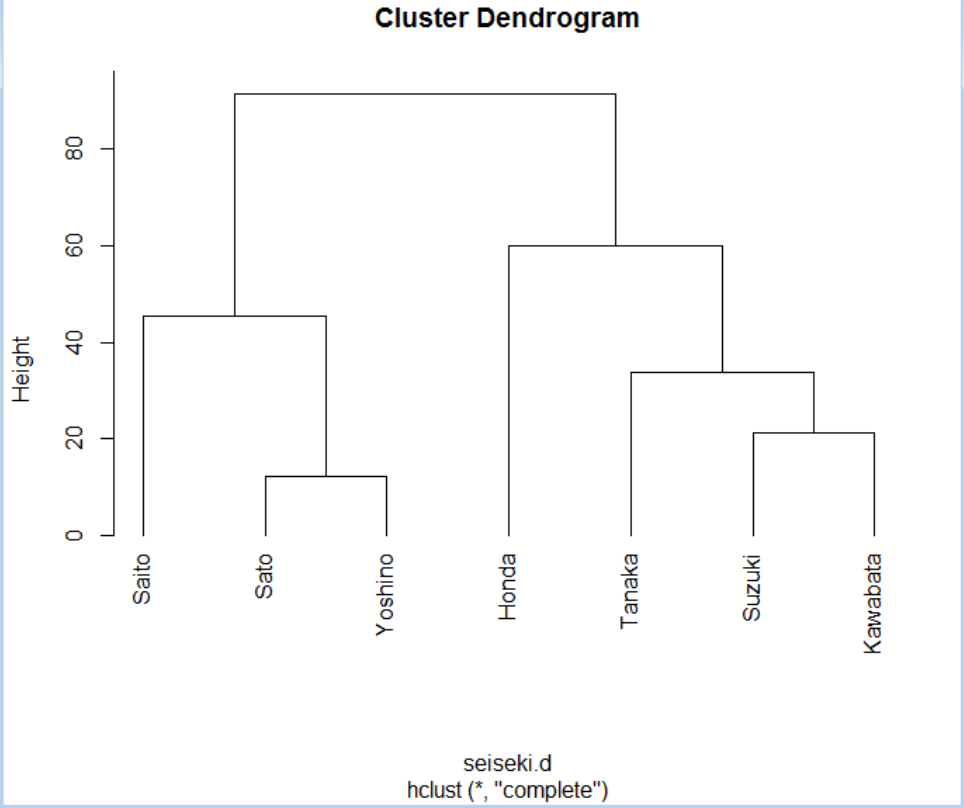
RGui (32-bit)

ファイル　履歴　サイズ変更　ウインドウ

R Console

```
merge        12     -none- numeric
height        6     -none- numeric
order         7     -none- numeric
labels        7     -none- character
method        1     -none- character
call          2     -none- call
dist.method   1     -none- character
> sei.hc$merge
     [,1] [,2]
[1,]   -2   -6
[2,]   -3   -5
[3,]   -1    2
[4,]   -7    1
[5,]   -4    3
[6,]    4    5
> sei.hc$height
[1] 12.40967 21.30728 33.77869 45.58509 60.13319 91.53142
> sei.hc$order
[1] 7 2 6 4 1 3 5
>
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(hclust(seiseki.d,"centroid"),hang=-1)
> plot(hclust(dist(seiseki,"canberra"),"ward"),hang=-1)
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
 以下にエラー as.matrix(x) :  オブジェクト 'seiseki' がありません
> plot(hclust(seiseki.d,"canberra"),"ward"),hang=-1)
 エラー:  予想外の ',' です  in "plot(hclust(seiseki.d,"canberra"),"ward"),"
> cutree(sei.hc,k=2)
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       1        2        1        1        1        2        2
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc,hang=-1)
>
```

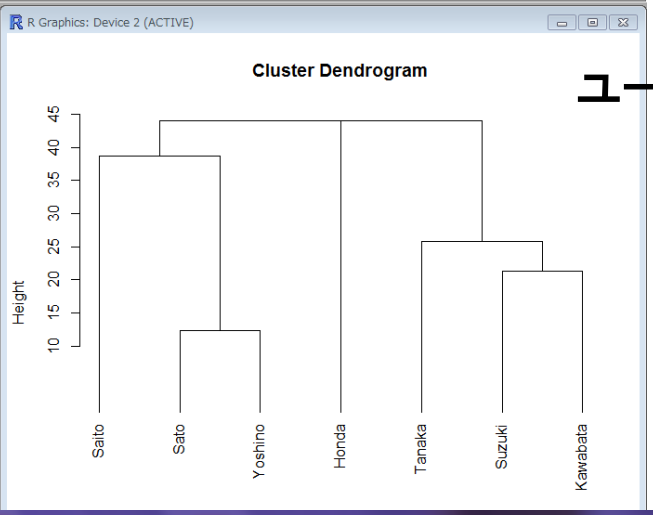
R Graphics: Device 2 (ACTIVE)



14:43
2014/05/27

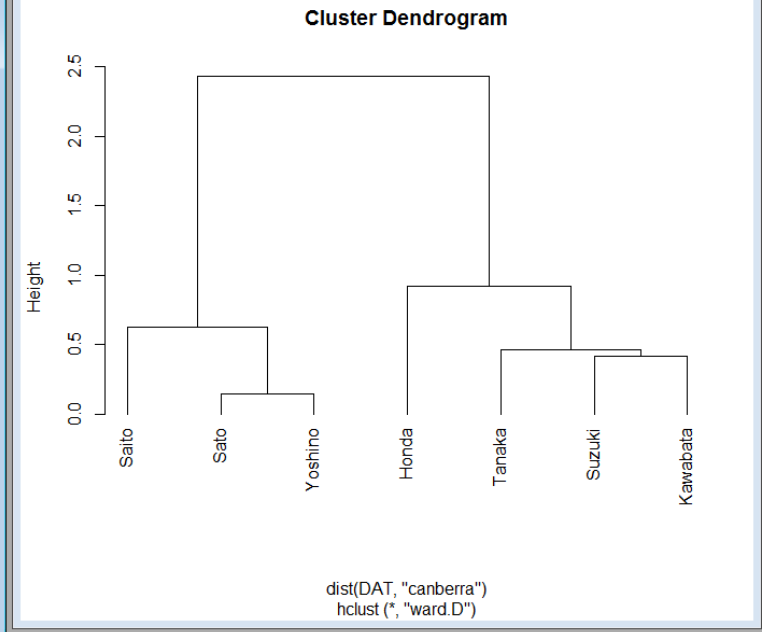
ユークリッド距離、重心法

```
height       6     -none- numeric
order        7     -none- numeric
labels       7     -none- character
method       1     -none- character
call         2     -none- call
dist.method  1     -none- character
> sei.hc$merge
     [,1] [,2]
[1,]   -2   -6
[2,]   -3   -5
[3,]   -1    2
[4,]   -7    1
[5,]   -4    3
[6,]    4    5
> sei.hc$height
[1] 12.40967 21.30728 33.77869 45.58509 60.13319 91.53142
> sei.hc$order
[1] 7 2 6 4 1 3 5
>
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(hclust(seiseki.d,"centroid"),hang=-1)
> plot(hclust(dist(seiseki,"canberra"),"ward"),hang=-1)
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
 以下にエラー as.matrix(x) :  オブジェクト 'seiseki' がありません
> plot(hclust(seiseki.d,"canberra"),"ward"),hang=-1)
 エラー:  予想外の ',' です  in "plot(hclust(seiseki.d,"canberra"),"ward"),"
> cutree(sei.hc,k=2)
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       1        2        1        1        1        2        2
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc,hang=-1)
> plot(hclust(seiseki.d,"centroid"),hang=-1)
>
```



キャンベラ距離
ウォード法

```
dist.method  1     -none- character
> sei.hc$merge
     [,1] [,2]
[1,]   -2   -6
[2,]   -3   -5
[3,]   -1    2
[4,]   -7    1
[5,]   -4    3
[6,]    4    5
> sei.hc$height
[1] 12.40967 21.30728 33.77869 45.58509 60.13319 91.53142
> sei.hc$order
[1] 7 2 6 4 1 3 5
>
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(hclust(seiseki.d,"centroid"),hang=-1)
> plot(hclust(dist(seiseki,"canberra"),"ward"),hang=-1)
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
 以下にエラー as.matrix(x) :  オブジェクト 'seiseki' がありません
> plot(hclust(seiseki.d,"canberra"),"ward"),hang=-1)
 エラー:  予想外の ',' です  in "plot(hclust(seiseki.d,"canberra"),"ward"),"
> cutree(sei.hc,k=2)
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       1        2        1        1        1        2        2
> plot(sei.hc)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc)#plot(sei.hc,hang=-1)
> plot(sei.hc,hang=-1)
> plot(hclust(seiseki.d,"centroid"),hang=-1)
> plot(hclust(dist(seiseki,"canberra"),"ward"),hang=-1)
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
 以下にエラー as.matrix(x) :  オブジェクト 'seiseki' がありません
> plot(hclust(dist(DAT,"canberra"),"ward"),hang=-1)
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
>
```



コーフェン行列

```
> cutree(sei.hc,k=2)
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       1        2        1        1        1        2        2
> cophenetic(sei.hc)
           Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino
Sato     91.53142
Suzuki   33.77869 91.53142
Honda    60.13319 91.53142 60.13319
Kawabata 33.77869 91.53142 21.30728 60.13319
Yoshino  91.53142 12.40967 91.53142 91.53142 91.53142
Saito    91.53142 45.58509 91.53142 91.53142 91.53142 45.58509
> cor(seiseki.d, cophenetic(sei.hc))
[1] 0.8944869
```

ユークリッド距離と最遠隣法による
コーフェン相関係数

階層的クラスター分析は、データの構造が複雑になると少数個の個体を入れ替えるだけで、結果が大きく変わることも珍しくない。どの結果を信じるべきかに関しては、階層的クラスター分析の結果だけではなく、他のデータ解析方法も用いて探索的にさまざまな角度でデータを眺めて、総合的に判断することが必要である。

非階層的クラスター分析

- 大量のデータ解析
- K平均法
  - k個のクラスター中心(seeds)の初期値を与える
  - すべてのデータをk個のクラスター中心との距離を求め、最も近いクラスターに分類する
  - 形成されたクラスターの中心を求める
  - クラスターの中心が変化しない時点まで繰り返す

非階層的クラスター分析

```
> (seiseki.km <- kmeans(DAT,2))
K-means clustering with 2 clusters of sizes 3, 4

Cluster means:
      Math      Sci     Lang  Eng   Soc
1 67.33333 73.33333 82.66667 84.0 90.00
2 71.75000 81.25000 49.25000 46.5 53.75

Clustering vector:
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       2        1        2        2        2        1        1

Within cluster sum of squares by cluster:
[1] 1228 2754
 (between_SS / total_SS =  62.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> seiseki.km$cluster
  Tanaka     Sato   Suzuki    Honda Kawabata  Yoshino    Saito
       2        1        2        2        2        1        1
`
```