

ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

27/11/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Co to znaczy „eksploracja danych”

Klastrowanie (grupowanie) hierarchiczne

Klastrowanie (grupowanie) k-means

Graficzna reprezentacja danych

3

- Graficznie pokazujemy najistotniejsze charakterystyki danych
 - Tabela ze statystycznym podsumowaniem
 - Histogramy
 - Box-ploty
 - Scatter-ploty
- Mogą być wykonane bardzo szybko/łatwo, powinny pozwolić na pierwsze wrażenie „co jest w danych”
- Staramy się robić stosunkowo dużo plotów, tabel sumarycznych w różny sposób grupując zmienne.

Przykład

4

- Danie: zanieczyszczenie drobinkami materiałów powietrza na terenie USA.
 - <http://www.epa.gov/air/ecosystem.html>
- Dla poziomu zanieczyszczeń ustalona jest norma $12\mu\text{g}/\text{m}^3$
- Zestawienie dzienne jest dostępne ze strony
 - <http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqdata.htm>
- Pytanie: czy są stany w których ta norma jest przekraczana?

Dane:

5

Tu są dane za okres 2008-2010

```
pollution <- read.csv("data/avgpm25.csv", colClasses = c("numeric", "character",  
  "factor", "numeric", "numeric"))  
head(pollution)
```

```
##      pm25  fips region longitude latitude  
## 1  9.771 01003  east    -87.75    30.59  
## 2  9.994 01027  east    -85.84    33.27  
## 3 10.689 01033  east    -87.73    34.73  
## 4 11.337 01049  east    -85.80    34.46  
## 5 12.120 01055  east    -86.03    34.02  
## 6 10.828 01069  east    -85.35    31.19
```

Czy w jakiś stanach przekroczona jest ta norma?

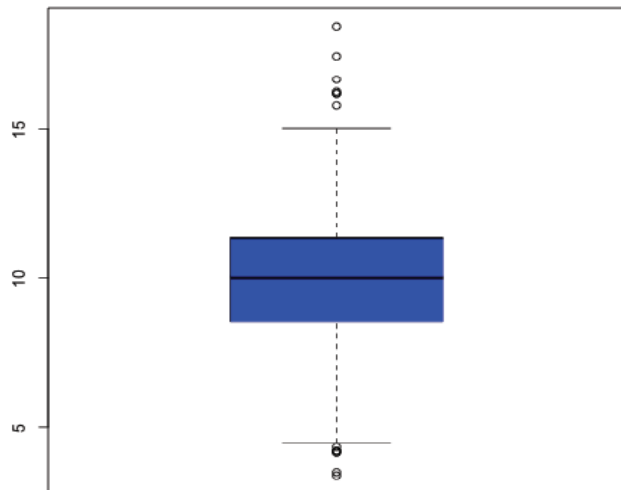
Tabela summaryczna, box plot, histogram

6

```
summary(pollution$pm25)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      3.38   8.55   10.00   9.84  11.40   18.40
```

```
boxplot(pollution$pm25, col = "blue")
```



```
hist(pollution$pm25, col = "green")
```

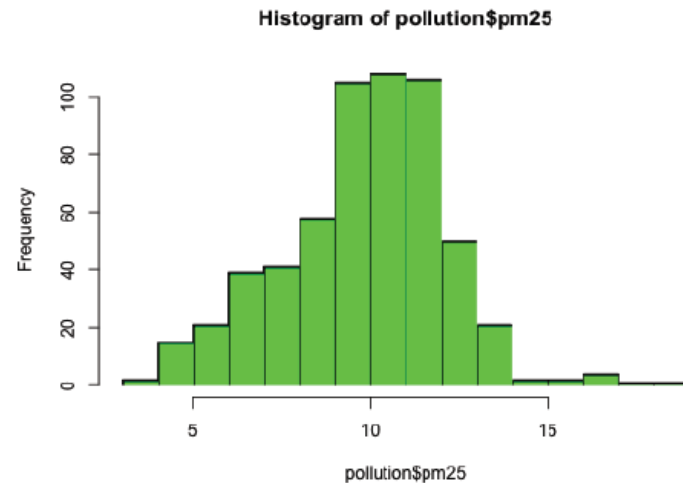


Tabela summaryczna, box plot, histogram

7

```
hist(pollution$pm25, col = "green")  
rug(pollution$pm25)
```

```
hist(pollution$pm25, col = "green", breaks = 100)  
rug(pollution$pm25)
```

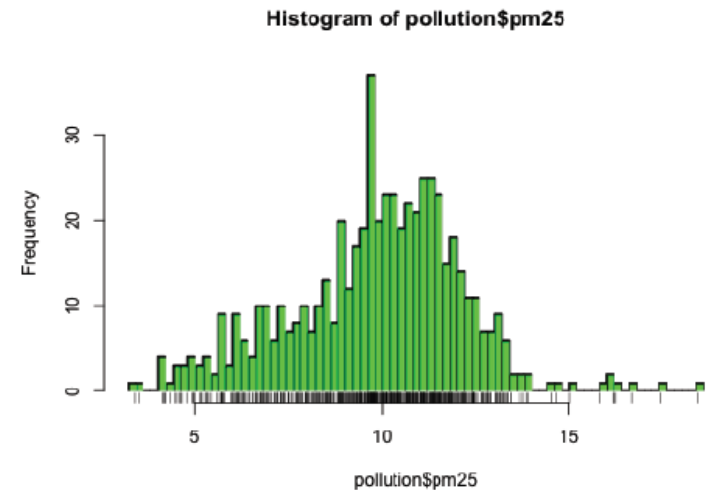
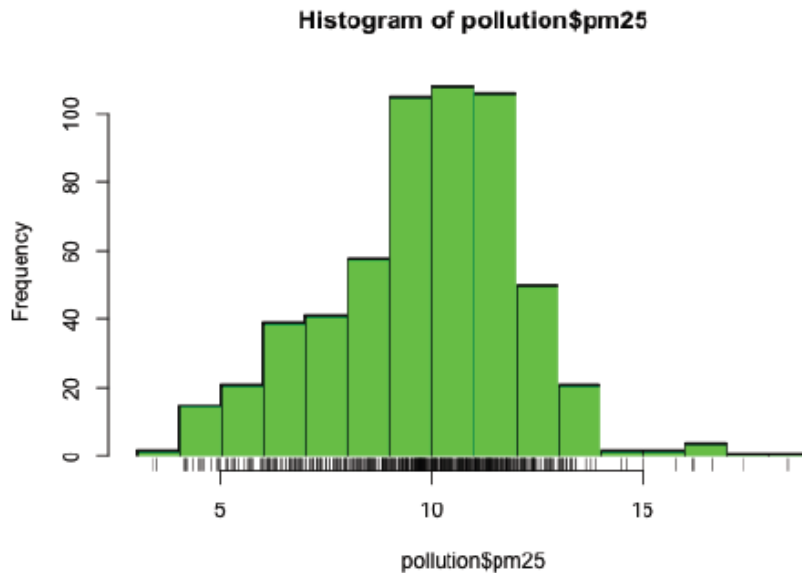
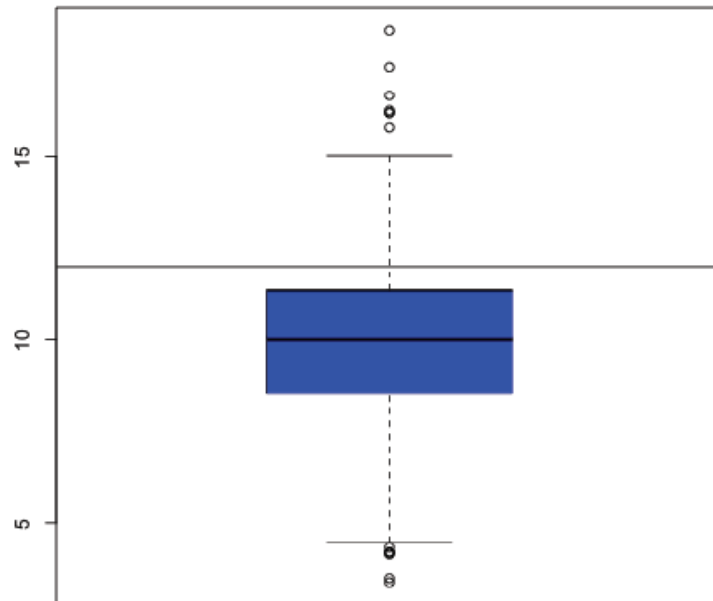


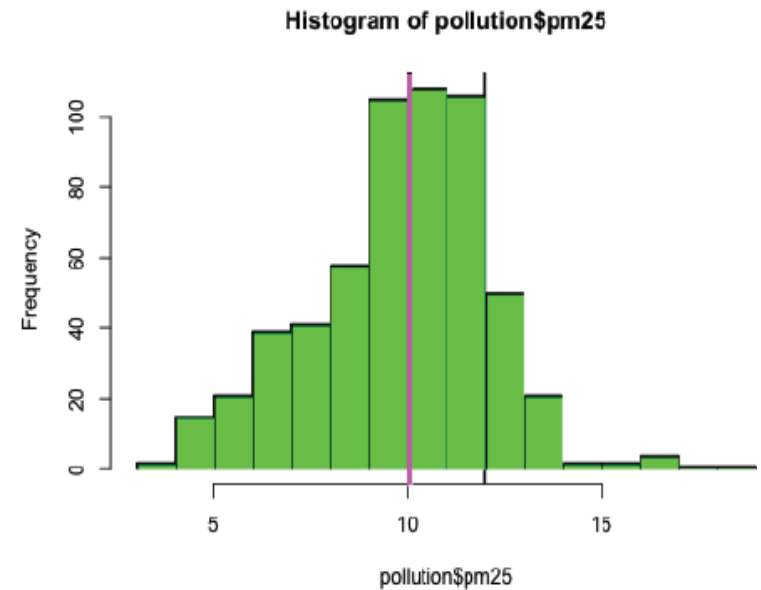
Tabela summaryczna, box plot, histogram

8

```
boxplot(pollution$pm25, col = "blue")  
abline(h = 12)
```



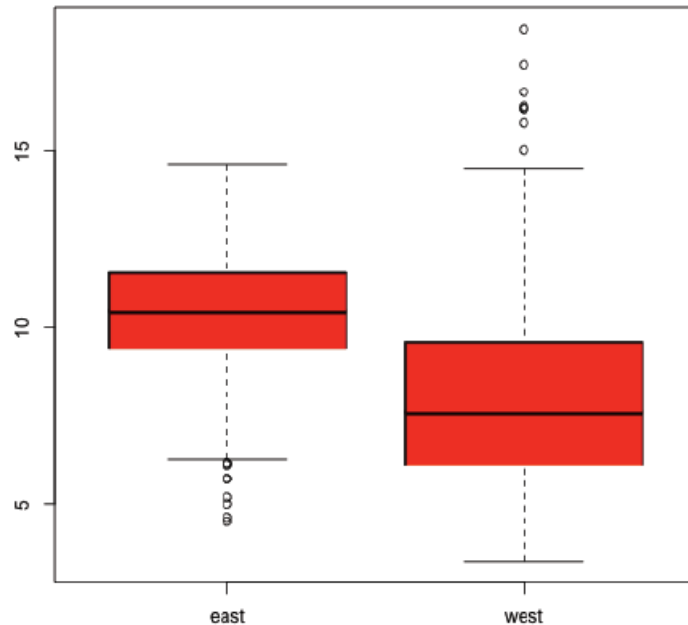
```
hist(pollution$pm25, col = "green")  
abline(v = 12, lwd = 2)  
abline(v = median(pollution$pm25), col = "magenta", lwd = 4)
```



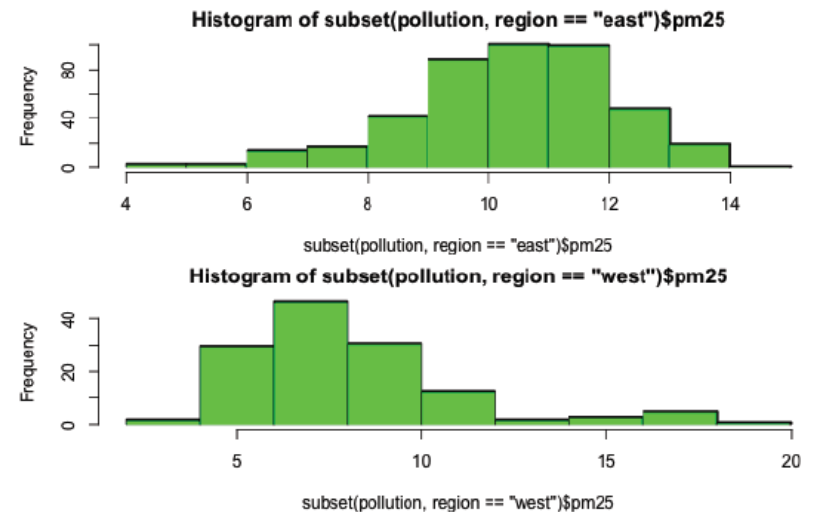
Multi-box plot, multi-histograms

9

```
boxplot(pm25 ~ region, data = pollution, col = "red")
```



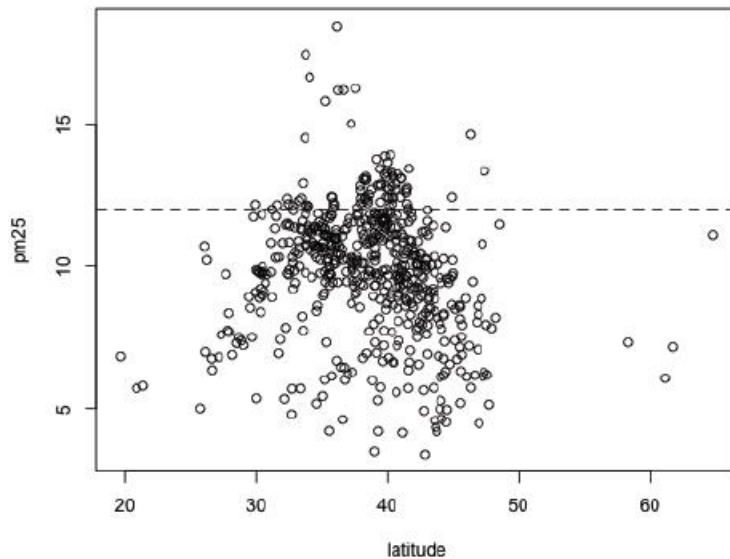
```
par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))  
hist(subset(pollution, region == "east")$pm25, col = "green")  
hist(subset(pollution, region == "west")$pm25, col = "green")
```



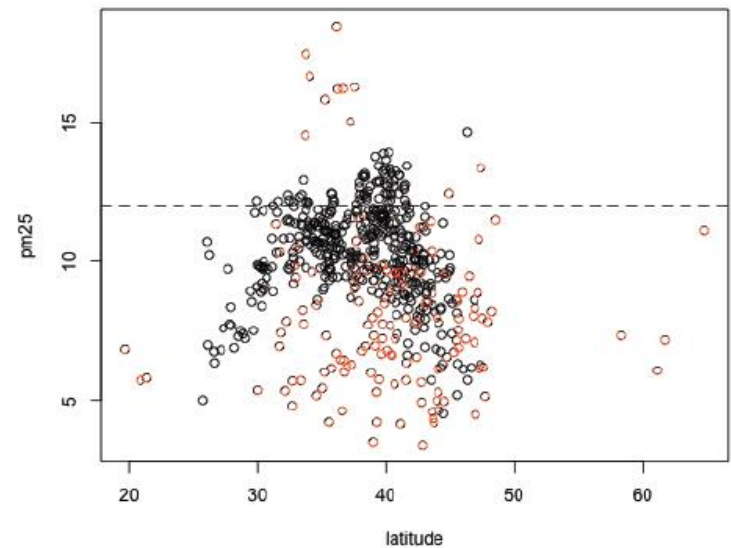
Scatter-plot

10

```
with(pollution, plot(latitude, pm25))  
abline(h = 12, lwd = 2, lty = 2)
```



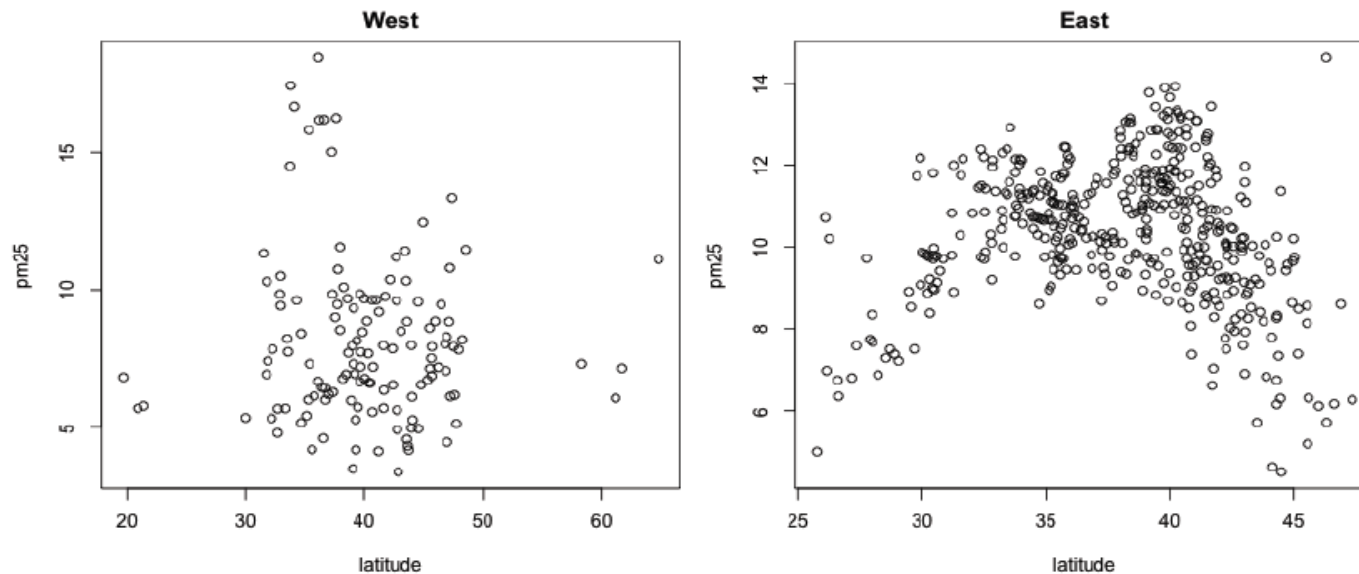
```
with(pollution, plot(latitude, pm25, col = region))  
abline(h = 12, lwd = 2, lty = 2)
```



Multiple scatter plots

11

```
par(mfrow = c(1, 2), mar = c(5, 4, 2, 1))  
with(subset(pollution, region == "west"), plot(latitude, pm25, main = "West"))  
with(subset(pollution, region == "east"), plot(latitude, pm25, main = "East"))
```



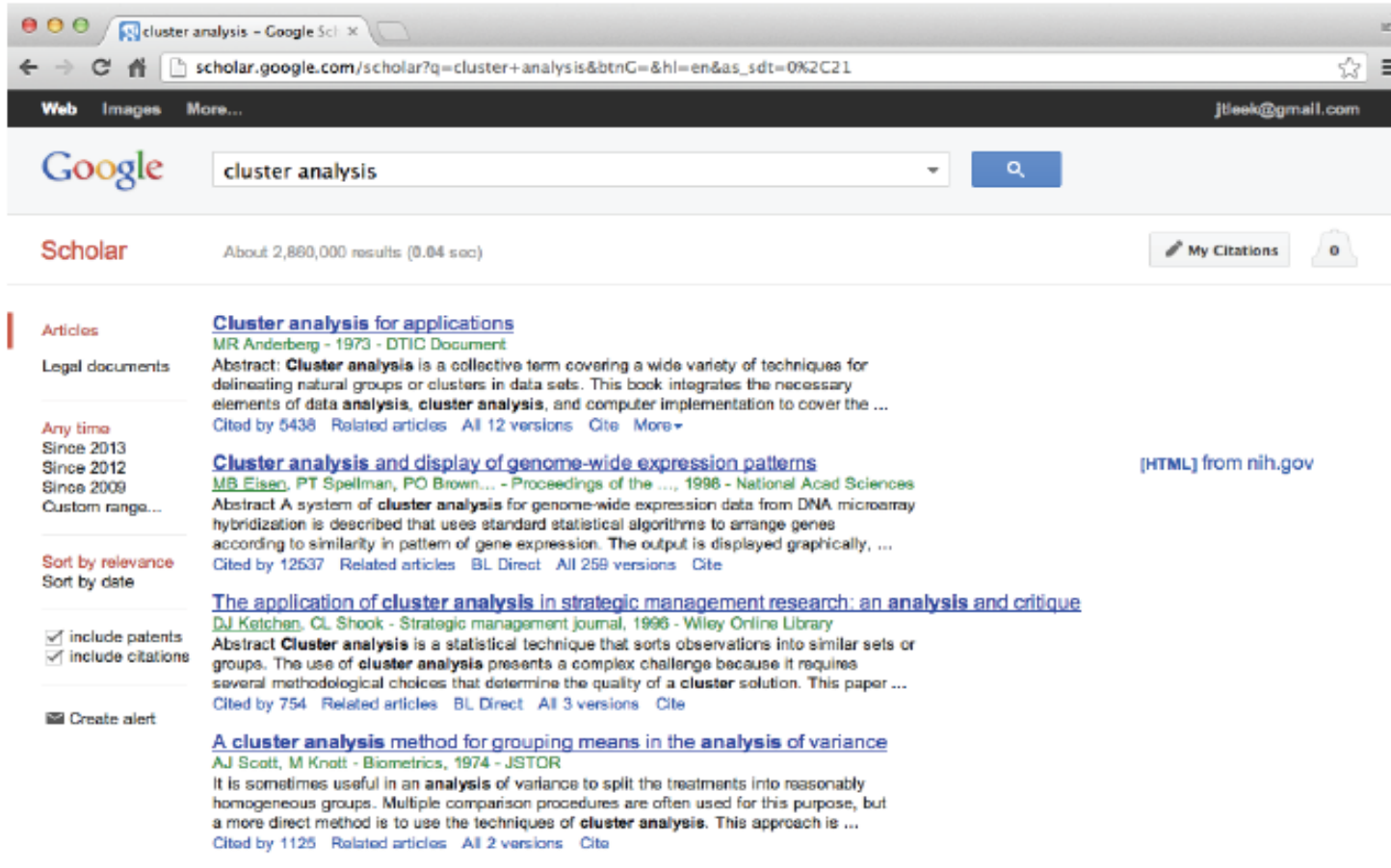
Co się dzieje w danych które są w wielu wymiarach

12

- Klastrowanie organizuje dane które są „blisko” w pewne grupy czyli klastry.
 - Co to znaczy że dane są blisko
 - Co to znaczy że grupujemy?
 - Jak pokazać graficznie grupowanie?
 - Jak interpretować grupowanie?

Klastrowanie jest bardzo ważną techniką

13



The screenshot shows a Google Scholar search interface. The search query is "cluster analysis", and the results page displays several articles. The first article is "Cluster analysis for applications" by MR Anderberg (1973). The second is "Cluster analysis and display of genome-wide expression patterns" by MR Eisen et al. (1998). The third is "The application of cluster analysis in strategic management research: an analysis and critique" by DJ Ketchen and CL Shook (1996). The fourth is "A cluster analysis method for grouping means in the analysis of variance" by AJ Scott and M Knott (1974). The interface includes a search bar, navigation tabs (Web, Images, More...), and a sidebar with filters for articles, legal documents, and sorting options.

cluster analysis - Google Sci x

scholar.google.com/scholar?q=cluster+analysis&btnG=&hl=en&as_sdt=0%2C21

Web Images More... jtlesk@gmail.com

Google cluster analysis

Scholar About 2,880,000 results (0.04 sec) My Citations 0

Articles

Legal documents

Any time

Since 2013

Since 2012

Since 2009

Custom range...

Sort by relevance

Sort by date

include patents

include citations

Create alert

Cluster analysis for applications
MR Anderberg - 1973 - DTIC Document
Abstract: **Cluster analysis** is a collective term covering a wide variety of techniques for delineating natural groups or clusters in data sets. This book integrates the necessary elements of data **analysis**, **cluster analysis**, and computer implementation to cover the ...
Cited by 5438 Related articles All 12 versions Cite More

Cluster analysis and display of genome-wide expression patterns
MR Eisen, PT Spellman, PO Brown... - Proceedings of the ..., 1998 - National Acad Sciences [HTML] from nih.gov
Abstract A system of **cluster analysis** for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, ...
Cited by 12537 Related articles BL Direct All 259 versions Cite

The application of cluster analysis in strategic management research: an analysis and critique
DJ Ketchen, CL Shook - Strategic management journal, 1996 - Wiley Online Library
Abstract **Cluster analysis** is a statistical technique that sorts observations into similar sets or groups. The use of **cluster analysis** presents a complex challenge because it requires several methodological choices that determine the quality of a **cluster** solution. This paper ...
Cited by 754 Related articles BL Direct All 3 versions Cite

A cluster analysis method for grouping means in the analysis of variance
AJ Scott, M Knott - Biometrics, 1974 - JSTOR
It is sometimes useful in an **analysis** of variance to split the treatments into reasonably homogeneous groups. Multiple comparison procedures are often used for this purpose, but a more direct method is to use the techniques of **cluster analysis**. This approach is ...
Cited by 1125 Related articles All 2 versions Cite

Hierarchiczne klastrowanie

14

- Aglomeracyjne podejście (bottom-up)
 - Znajdź dwa najbliższe położone punkty
 - Połącz je ze sobą w „super-punkt”
 - Znajdź kolejne najbliższe punkty (traktując już połączone jako jeden super-punkt)
- Parametry algorytmu
 - definicję odległości punktów
 - definicję „łączenia” punktów
- Wynik
 - Prezentujemy jako pewne „drzewo” (dendogram)

Jak definiujemy odległość

15

- Odległość:
 - Ciągła: euklidesowa metryka
 - Ciągła: stopień podobieństwa lub korelacji
 - Dyskretna: „Manhattan”
- Wybieramy taką definicję która stosuje się do naszych danych

Jak definiujemy odległość

16

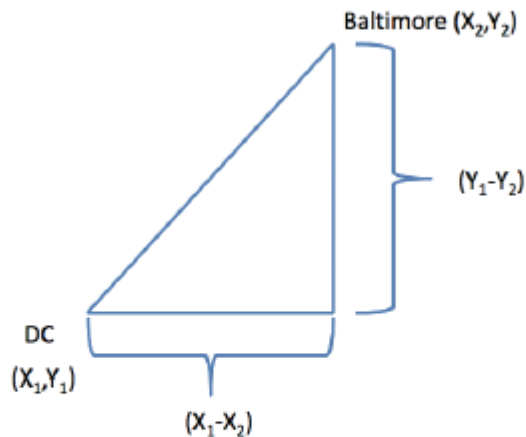
- Odległość:
 - Ciągła: euklidesowa metryka
 - Ciągła: stopień podobieństwa lub korelacji
 - Dyskretna: „Manhattan”
- Wybieramy taką definicję która stosuje się do naszych danych

Jak definiujemy odległość

17

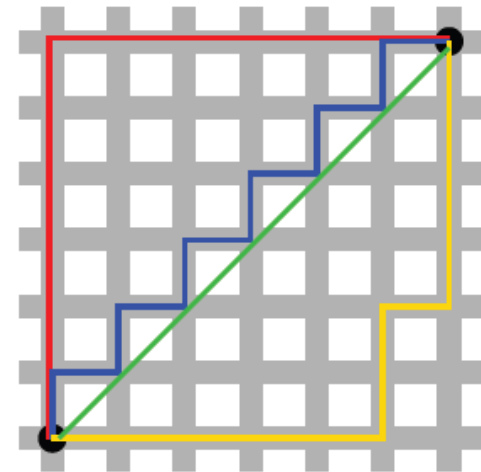
- Euklidesowa metryka vs Manhattan metryka

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$



Naturalnie rozszerzalne do wielu wymiarów

$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \dots + (Z_1 - Z_2)^2}$$



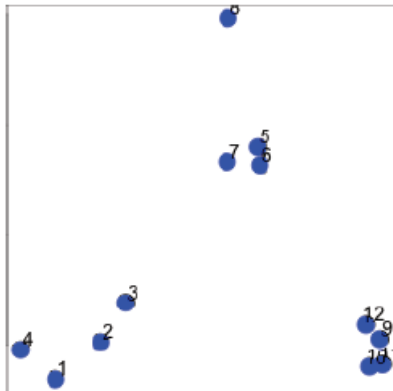
Musimy chodzić po ulicach

$$|A_1 - A_2| + |B_1 - B_2| + \dots + |Z_1 - Z_2|$$

Hierarchiczne klastrowanie

18

```
set.seed(1234)
par(mar = c(0, 0, 0, 0))
x <- rnorm(12, mean = rep(1:3, each = 4), sd = 0.2)
y <- rnorm(12, mean = rep(c(1, 2, 1), each = 4), sd = 0.2)
plot(x, y, col = "blue", pch = 19, cex = 2)
text(x + 0.05, y + 0.05, labels = as.character(1:12))
```



Symulujemy dane: 12 punktów, 3 klastry

Hierarchine klastrowanie: **dist**

19

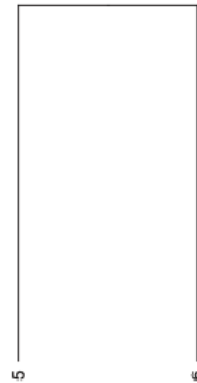
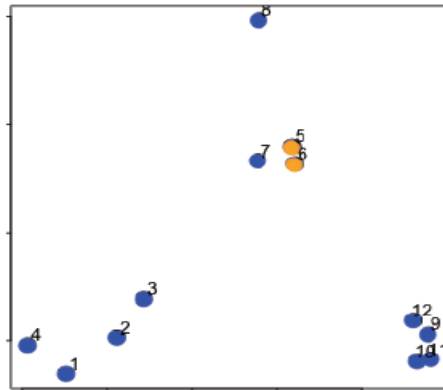
```
dataFrame <- data.frame(x = x, y = y)
dist(dataFrame)
```

Obliczamy odległość pomiędzy
wszystkimi parami punktów
Funkcja `dist()` ma jako parametr opcje
wyboru metryki

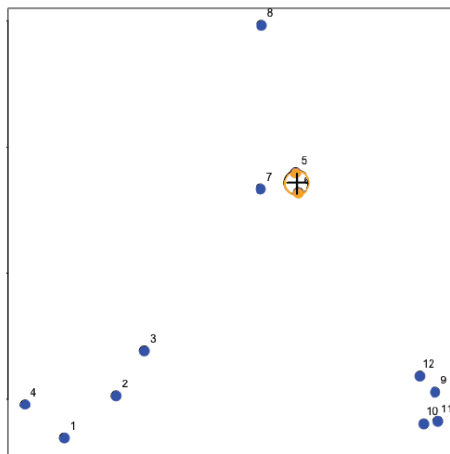
```
##           1           2           3           4           5           6           7           8           9
## 2  0.34121
## 3  0.57494  0.24103
## 4  0.26382  0.52579  0.71862
## 5  1.69425  1.35818  1.11953  1.80667
## 6  1.65813  1.31960  1.08339  1.78081  0.08150
## 7  1.49823  1.16621  0.92569  1.60132  0.21110  0.21667
## 8  1.99149  1.69093  1.45649  2.02849  0.61704  0.69792  0.65063
## 9  2.13630  1.83168  1.67836  2.35676  1.18350  1.11500  1.28583  1.76461
## 10 2.06420  1.76999  1.63110  2.29239  1.23848  1.16550  1.32063  1.83518  0.14090
## 11 2.14702  1.85183  1.71074  2.37462  1.28154  1.21077  1.37370  1.86999  0.11624
## 12 2.05664  1.74663  1.58659  2.27232  1.07701  1.00777  1.17740  1.66224  0.10849
##           10           11
```

Hierarchiczne klastrowanie

20



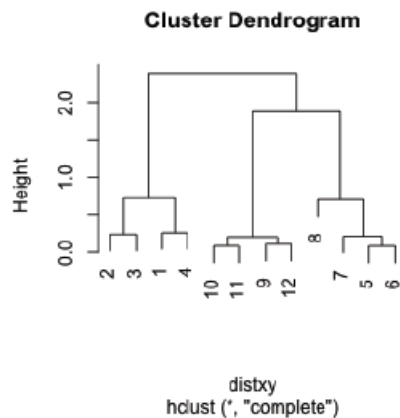
Stratujemy z punktów które są od siebie najbliższe, formułujemy z nich pojedynczy super-point



Hierarchiczne klastrowanie

21

```
dataFrame <- data.frame(x = x, y = y)
distxy <- dist(dataFrame)
hClustering <- hclust(distxy)
plot(hClustering)
```



Ile mamy klastrów ?
To zależy gdzie utniemy to drzewo.
To jest już zadanie związane z interpretacją analizy

Ładniejsza graficzna reprezentacja

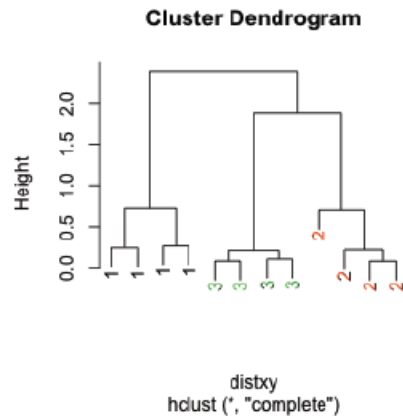
22

```
myplclust <- function(hclust, lab = hclust$labels, lab.col = rep(1, length(hclust$labels)),
  hang = 0.1, ...) {
  ## modification of plclust for plotting hclust objects *in colour*! Copyright
  ## Eva KF Chan 2009 Arguments: hclust: hclust object lab: a character vector
  ## of labels of the leaves of the tree lab.col: colour for the labels;
  ## NA=default device foreground colour hang: as in hclust & plclust Side
  ## effect: A display of hierarchical cluster with coloured leaf labels.
  y <- rep(hclust$height, 2)
  x <- as.numeric(hclust$merge)
  y <- y[which(x < 0)]
  x <- x[which(x < 0)]
  x <- abs(x)
  y <- y[order(x)]
  x <- x[order(x)]
  plot(hclust, labels = FALSE, hang = hang, ...)
  text(x = x, y = y[hclust$order] - (max(hclust$height) * hang), labels = lab[hclust$order],
    col = lab.col[hclust$order], srt = 90, adj = c(1, 0.5), xpd = NA, ...)
}
```

Hierarchiczne klastrowanie

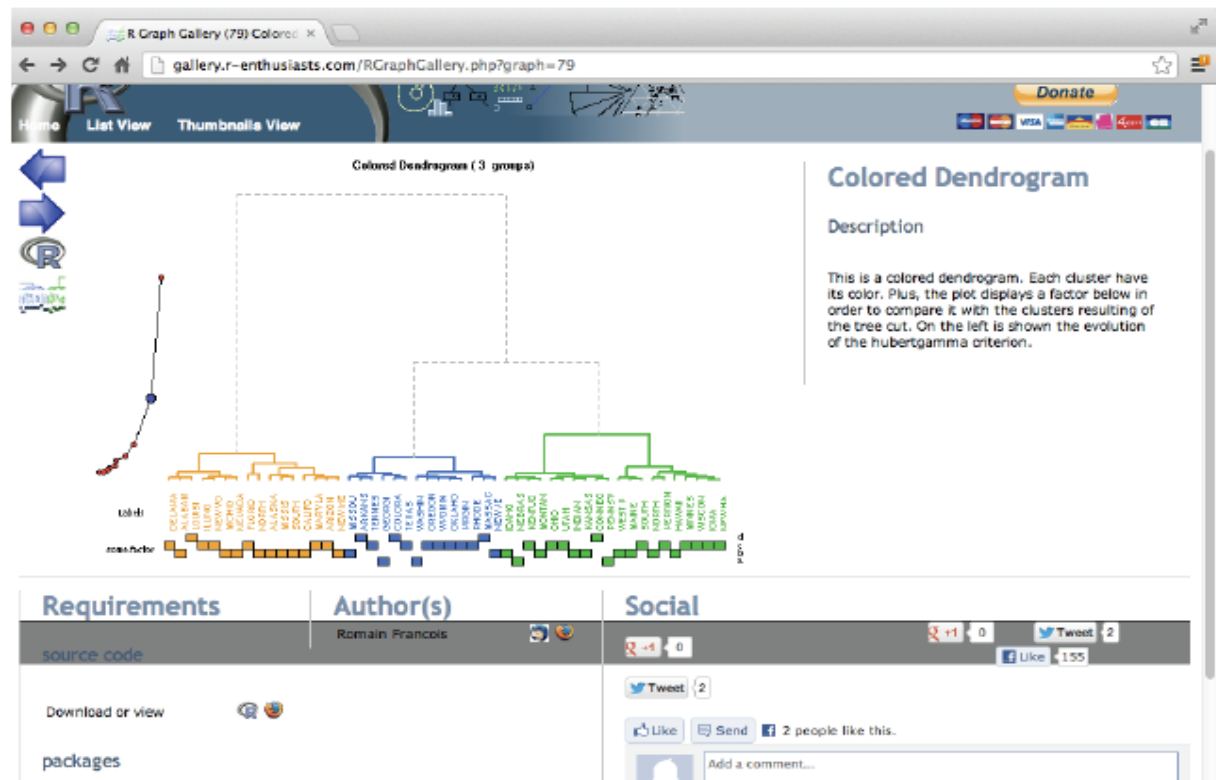
23

```
dataFrame <- data.frame(x = x, y = y)
distxy <- dist(dataFrame)
hClustering <- hclust(distxy)
myplclust(hClustering, lab = rep(1:3, each = 4), lab.col = rep(1:3, each = 4))
```



Jeszcze ładniejsza graficzna reprezentacja

24



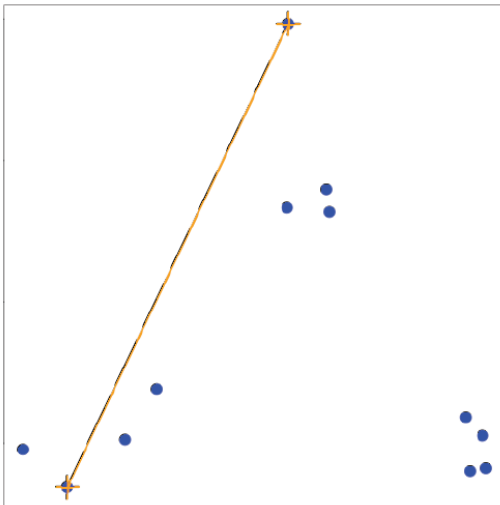
<http://gallery.r-enthusiasts.com/RGraphGallery.php?graph=79>

Łączenie punktów

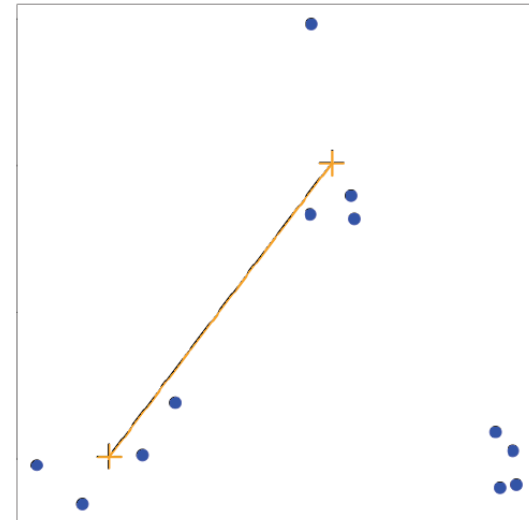
25

Jaka jest odległość pomiędzy klastrami?

„Najbardziej odległe punkty”



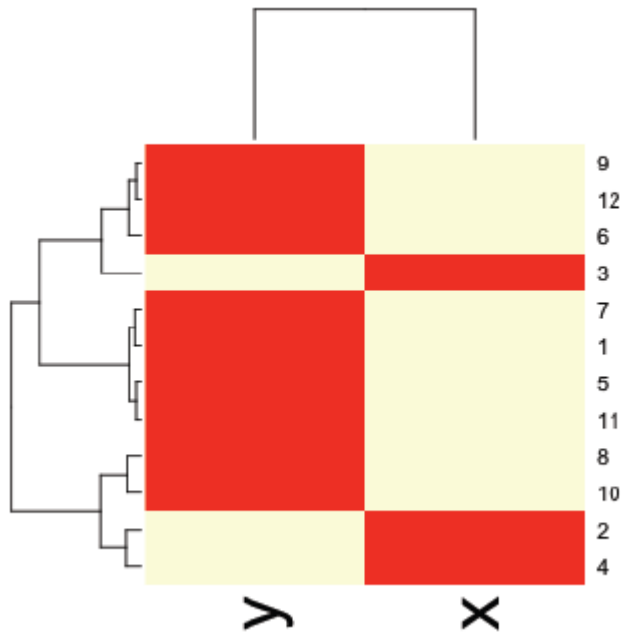
Środek ciężkości punktów



heatmap()

26

```
dataFrame <- data.frame(x = x, y = y)
set.seed(143)
dataMatrix <- as.matrix(dataFrame)[sample(1:12), ]
heatmap(dataMatrix)
```



heatmap() to funkcja dla ilustrowania macierzy odległości.

Ta funkcja wykonuje hierarchiczne klastrowanie na wierszach i kolumnach macierzy odległości

Hierarchiczne klastrowanie

27

- Daje pewny obraz związku pomiędzy danymi, organizuje je w hierarchiczny sposób.
 - Jako opcje algorytmu wybieramy metrykę oraz metodę łączenia w „super-punkty”
- Wynik algorytmu klastrowania ulegnie zmianie jeżeli:
 - Usuniemy kilka punktów
 - Dane mają różne brakujące punkty
 - Zmienimy metrykę
 - Zmienimy strategię łączenia
 - Zmienimy skalę w jednym z wymiarów
- To klastrowanie jest deterministyczne (powtarzalne dla tego samego zestawu danych, nie ma losowości)
- Powinno być używane przede wszystkim dla „eksploracji” danych.

K-means klastrowanie

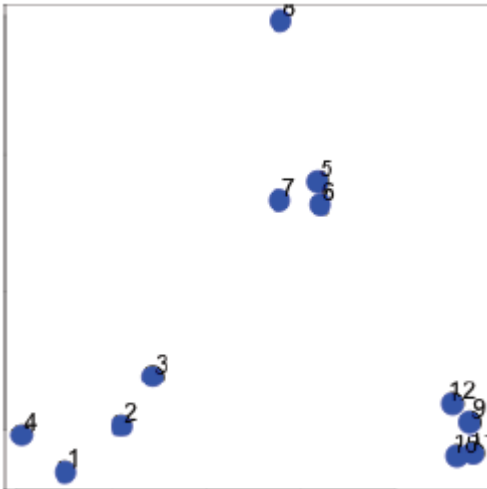
28

- Sposób na podzielenie danych na grupy
 - Zdecyduj ile chcesz mieć klastrów
 - Zdefiniuj pozycję „centrum” dla każdego klastra
 - Przypisz punkty pomiarowe do klastrów
 - Przelicz na podstawie przypisanych danych pomiarowych pozycje „centrum” dla każdego klastra
- Parametry wejściowe: metryka, ilość klastrów, punkt startowy dla centrum klastra
- Wynik: przeliczona pozycja centrum, asocjacja punktów pomiarowych do danych

Przykład:

29

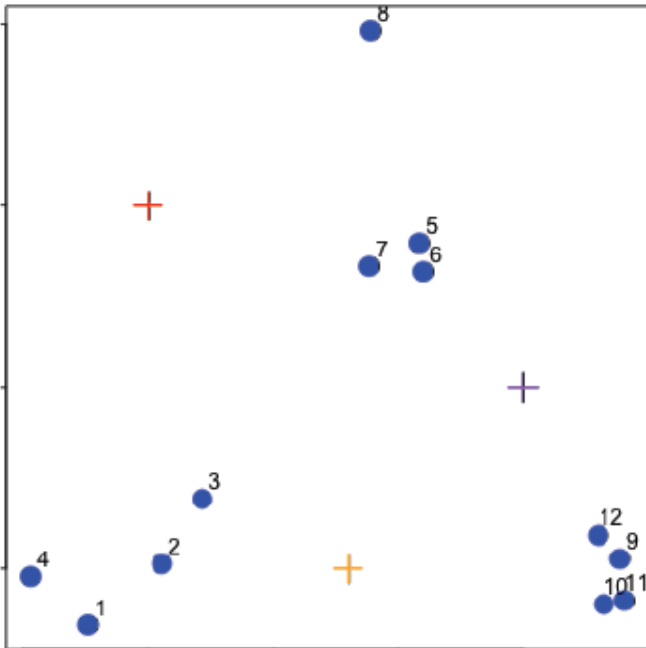
```
set.seed(1234)
par(mar = c(0, 0, 0, 0))
x <- rnorm(12, mean = rep(1:3, each = 4), sd = 0.2)
y <- rnorm(12, mean = rep(c(1, 2, 1), each = 4), sd = 0.2)
plot(x, y, col = "blue", pch = 19, cex = 2)
text(x + 0.05, y + 0.05, labels = as.character(1:12))
```



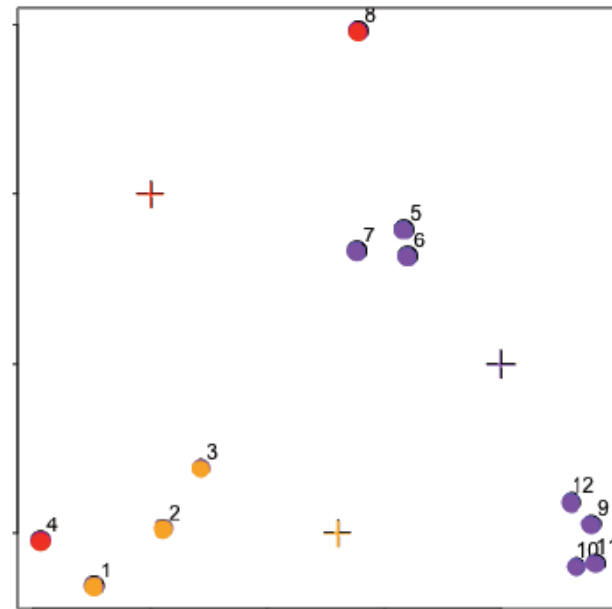
Przykład

30

Wybieramy startowe centra



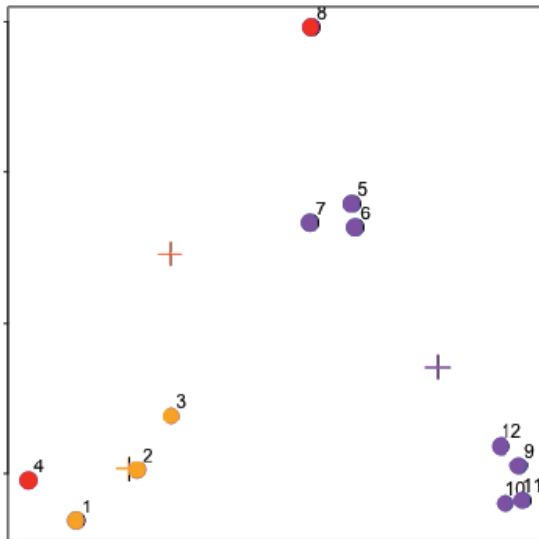
Przypisujemy punkty do klastrów początkowe grupowanie



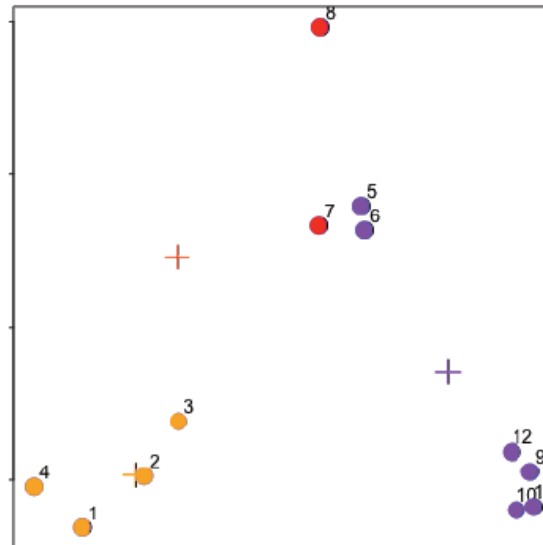
Przykład

31

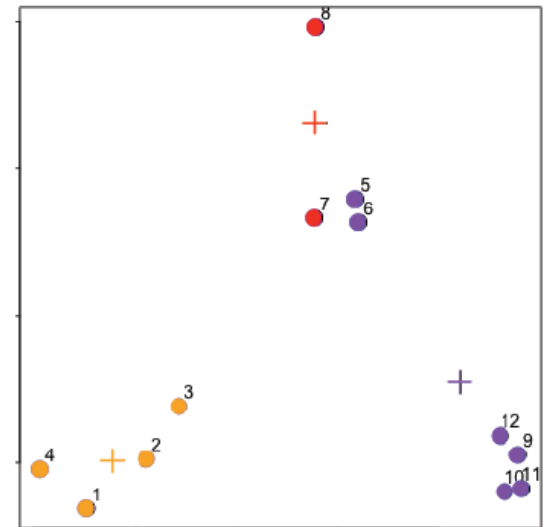
Przeliczamy pozycje centrów



Modyfikujemy przyporządkowanie



Przeliczamy raz jeszcze pozycje centrów



kmeans()

32

```
dataFrame <- data.frame(x, y)
kmeansObj <- kmeans(dataFrame, centers = 3)
names(kmeansObj)
```

```
## [1] "cluster"      "centers"      "totss"       "withinss"
## [5] "tot.withinss" "betweenss"   "size"        "iter"
## [9] "ifault"
```

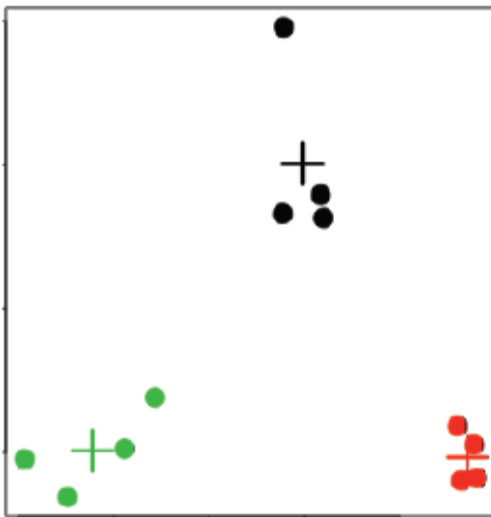
```
kmeansObj$cluster
```

```
## [1] 3 3 3 3 1 1 1 1 2 2 2 2
```


kmeans()

33

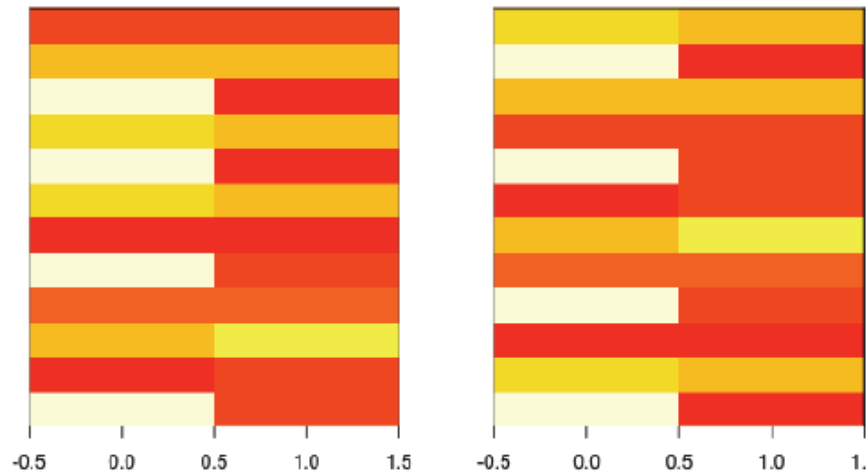
```
par(mar = rep(0.2, 4))  
plot(x, y, col = kmeansObj$cluster, pch = 19, cex = 2)  
points(kmeansObj$centers, col = 1:3, pch = 3, cex = 3, lwd = 3)
```



heatmaps()

34

```
set.seed(1234)
dataMatrix <- as.matrix(dataFrame)[sample(1:12), ]
kmeansObj2 <- kmeans(dataMatrix, centers = 3)
par(mfrow = c(1, 2), mar = c(2, 4, 0.1, 0.1))
image(t(dataMatrix)[, nrow(dataMatrix):1], yaxt = "n")
image(t(dataMatrix)[, order(kmeansObj2$cluster)], yaxt = "n")
```



Reorganizujemy wiersze macierzy danych

K-means: podsumowanie

35

- K-means klastrowanie wymaga ustalenie liczby klastrów jako parameter wejściowy
 - ▣ Ocena wzrokowa scatter-plotów, intuicja
 - ▣ Techniki: cross-validation, teoria informacji, etc.
 - ▣ Inne techniki: patrz dodatkowe materiały
- Wynik algorytmu K-means:
 - ▣ Zależy od ustalonej liczby klastrów
 - ▣ Zależy od liczby iteracji