# Лабораторна робота №3
# Студента 2 курсу магістратури
# Групи "статистика"
# Варіант №4

Горбунов Даніел Денисович

20 жовтня 2022 р.

## Вступ.

У даній роботі використано оптимальні кластеризації на даних з першої лабораторної роботи для порівняння з тими кандидатами, які отримані за допомогою методів ієрархічної кластеризації.

## Хід роботи.

### Ієрархічна кластеризація за даними з варіанту.

Першочергово треба розібратися з тим, що за дані записані у файлі.

```
> # Зчитуємо дані
> data <- read.table("./mult4.txt", header=T)
> # Стандартизуємо дані
> data.std <- scale(data)
```

Тепер можна переходити до основної частини роботи. Спочатку підрахуємо матриці відстаней.

```
> # Обчислення L2-відстаней
> dist.l2 <- dist(data.std, method = "euclidean")
> # Обчислення L1-відстаней
> dist.l1 <- dist(data.std, method = "manhattan")
> # Обчислення minimax-відстаней
> dist.mm <- dist(data.std, method = "maximum")
```

Далі вже розглядаємо ієрархічні кластеризації в залежності від метрики та методів зв'язку.

**Ієрархічна кластеризація: метод одного зв'язку.**

```
> # Кластеризація за l2-метрикою
> data.std.link.single.l2 <- hclust(
+   d = dist.l2, method = "single"
+ )
> # Кластеризація за l1-метрикою
> data.std.link.single.l1 <- hclust(
+   d = dist.l1, method = "single"
+ )
> # Кластеризація за мінімаксною метрикою
> data.std.link.single.mm <- hclust(
+   d = dist.mm, method = "single"
+ )
```

Тепер покажемо дендрограми кластеризації:

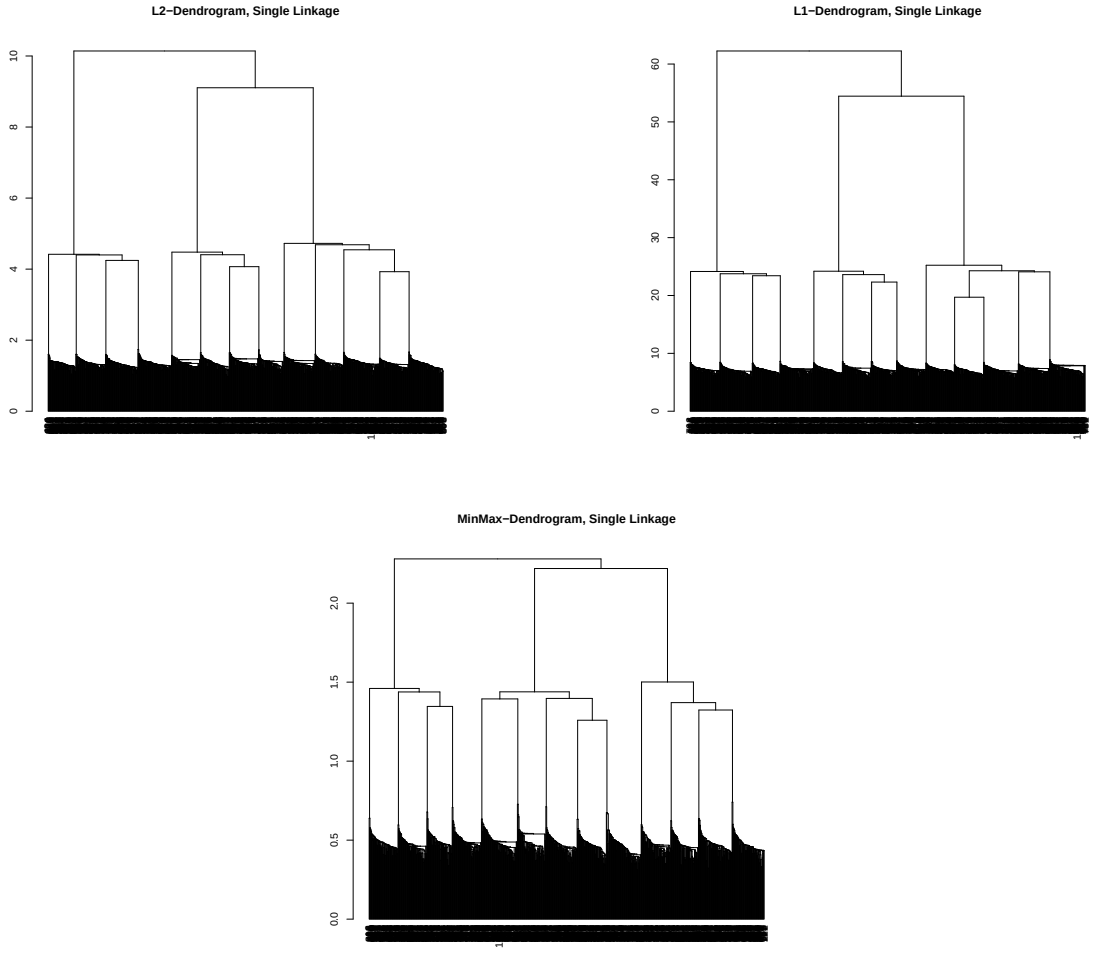



Рис. 1: Дендрограми кластеризацій в залежності від метрики. Метод одного зв'язку.

**Ієрархічна кластеризація: метод повного зв'язку.**

```
> # Кластеризація за l2-метрикою
> data.std.link.complete.l2 <- hclust(
+   d = dist.l2, method = "complete"
+ )
> # Кластеризація за l1-метрикою
> data.std.link.complete.l1 <- hclust(
+   d = dist.l1, method = "complete"
+ )
> # Кластеризація за мінімаксною метрикою
> data.std.link.complete.mm <- hclust(
+   d = dist.mm, method = "complete"
+ )
```

Тепер покажемо дендрограми кластеризації:

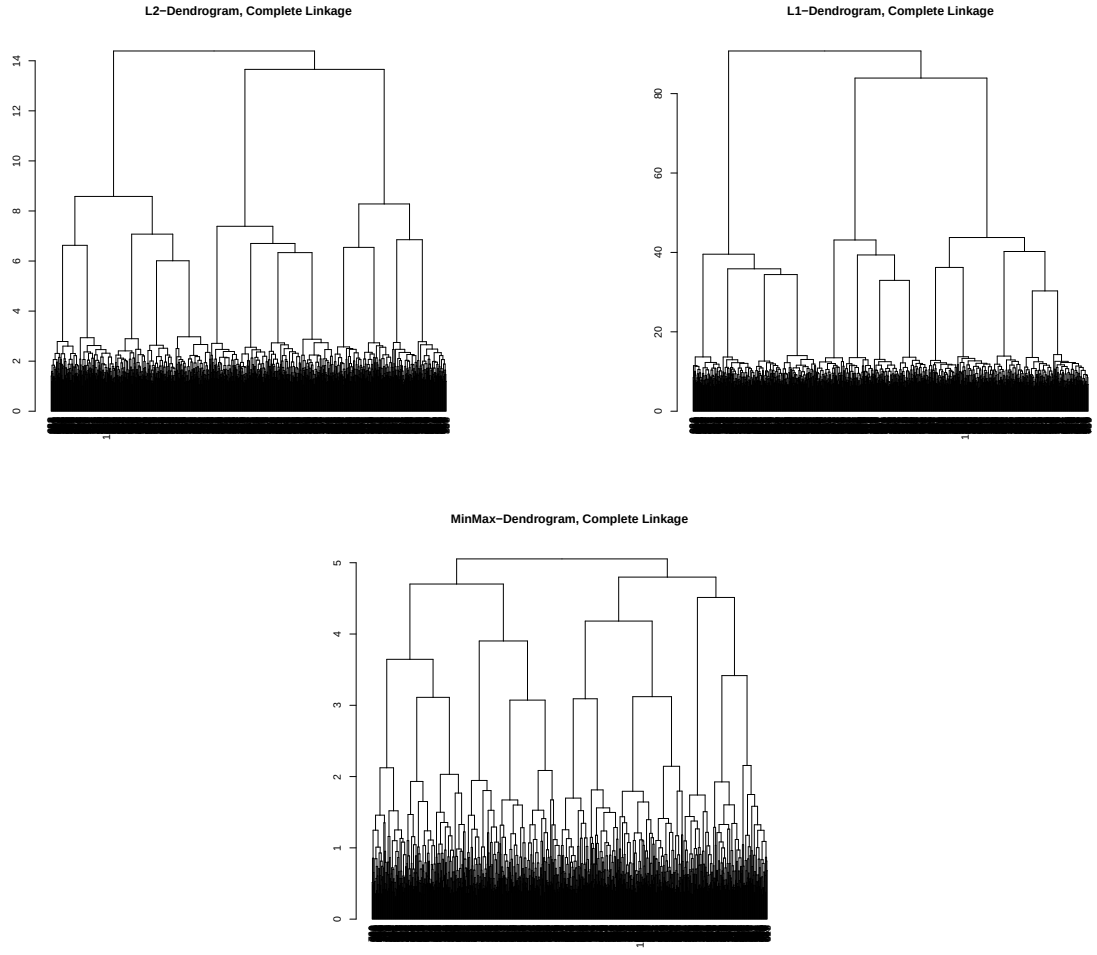



Рис. 2: Дендрограми кластеризацій в залежності від метрики. Метод одного зв'язку.

**Ієрархічна кластеризація: метод середнього зв'язку.**

```
> # Кластеризація за l2-метрикою
> data.std.link.average.l2 <- hclust(
+   d = dist.l2, method = "average"
+ )
> # Кластеризація за l1-метрикою
> data.std.link.average.l1 <- hclust(
+   d = dist.l1, method = "average"
+ )
> # Кластеризація за мінімаксною метрикою
> data.std.link.average.mm <- hclust(
+   d = dist.mm, method = "average"
+ )
```

Тепер покажемо дендрограми кластеризації:

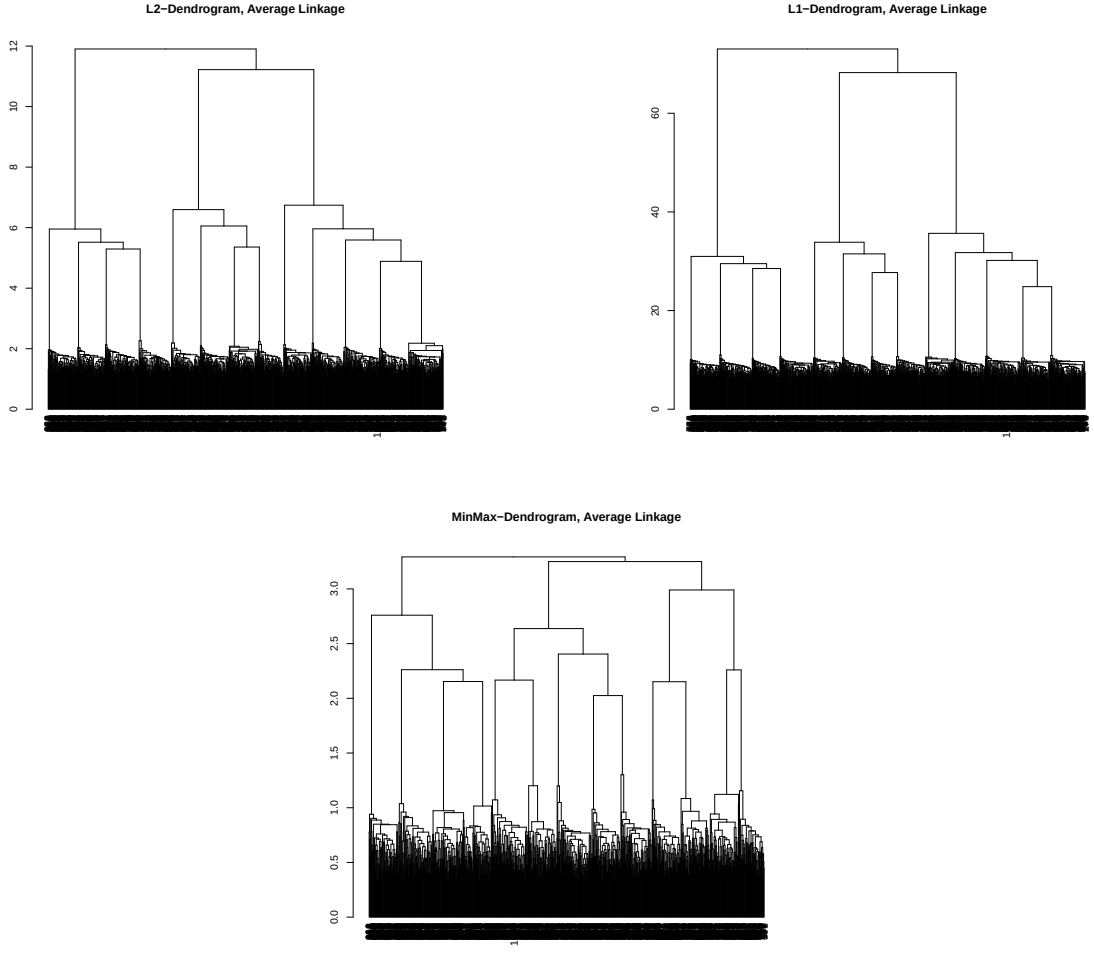



Рис. 3: Дендрограми кластеризацій в залежності від метрики. Метод середнього зв'язку.

**Підрахунок кофенетичних відстаней для методу одного зв'язку:**

```
> # L2-відстань
> cor(cophenetic(data.std.link.single.l2), dist.l2)

[1] 0.9780738

> # L1-відстань
> cor(cophenetic(data.std.link.single.l1), dist.l1)

[1] 0.9796792

> # MinMax відстань
> cor(cophenetic(data.std.link.single.mm), dist.mm)

[1] 0.8406677
```

**Підрахунок кофенетичних відстаней для методу повного зв'язку:**

```
> # L2-відстань
> cor(cophenetic(data.std.link.complete.l2), dist.l2)

[1] 0.973988

> # L1-відстань
> cor(cophenetic(data.std.link.complete.l1), dist.l1)

[1] 0.9826179

> # MinMax відстань
> cor(cophenetic(data.std.link.complete.mm), dist.mm)

[1] 0.8284624
```

**Підрахунок кофенетичних відстаней для методу середнього зв'язку:**

```
> # L2-відстань
> cor(cophenetic(data.std.link.average.l2), dist.l2)

[1] 0.9831204

> # L1-відстань
> cor(cophenetic(data.std.link.average.l1), dist.l1)

[1] 0.9837843

> # MinMax відстань
> cor(cophenetic(data.std.link.average.mm), dist.mm)

[1] 0.8834431
```

Кофенетичні кореляції, незалежно від методу зв'язку, виходять досить високими для кожної з метрик. Варто відмітити, що якість більш-менш така сама, якщо використовувати або евклідову відстань, або відстань сіті-блок. На дендрограмах кластеризації можна побачити, що майже всюди утворюються по три великі кластери (знову ж таки, це узгоджується з нашими дослідженнями у попередніх роботах). "Майже всюди" – для довільного методу зв'язку, окрім повного зв'язку (там виділилися чотири кластери). Тому наше припущення про те, що варто розбивати дані на три частини, справді має місце. Зокрема можна побчити виділення $\tilde{1}3$ груп на самому дні дендрограми, тому таке теж можна дослідити.

Далі зробимо обрізання за вказаною кількістю кластерів, тобто за трьома на довільній з метрик з найбільшим кофенетичним коефіцієнтом (наприклад евклідова):

```
> single.cut.3 <- cut(as.dendrogram(data.std.link.single.l2), h = 3)
> complete.cut.3 <- cut(as.dendrogram(data.std.link.complete.l2), h = 4)
> average.cut.3 <- cut(as.dendrogram(data.std.link.average.l2), h = 4)
```

Перевіримо, чи має місце наше припущення про розбиття на три підгрупи хоча б на одному із запропонованих методів зв'язування.

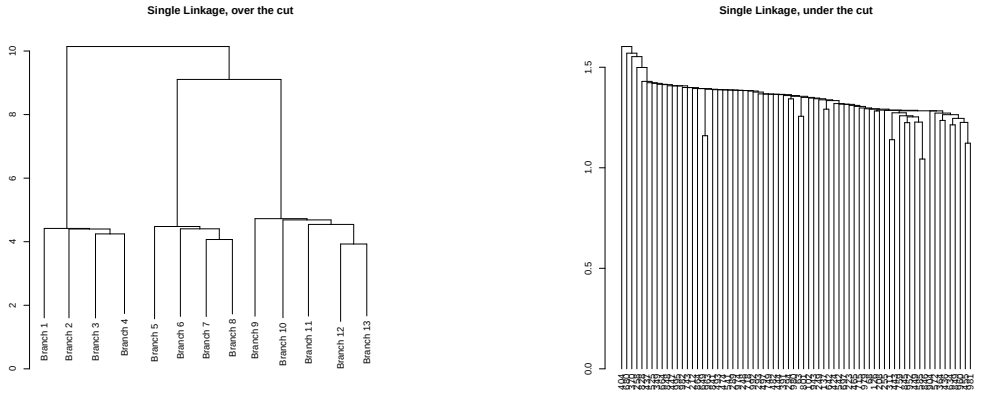



Рис. 4: Дерево кластеризації над та під обраним зрізом. Метод одного зв'язку.

Обрізання на вказаному рівні показує доцільність вибору трьох кластерів на основі методу одного зв'язку. Можна побачити, що дендрограма, утворена нижнім зрізом основного дерева, розміщує усі об'єкти відносно недалеко та майже рівномірно (в сенсі кофенетичних відстаней між ними). Тобто виокремити конкретні кластери з цієї дендрограми доцільно лише один.

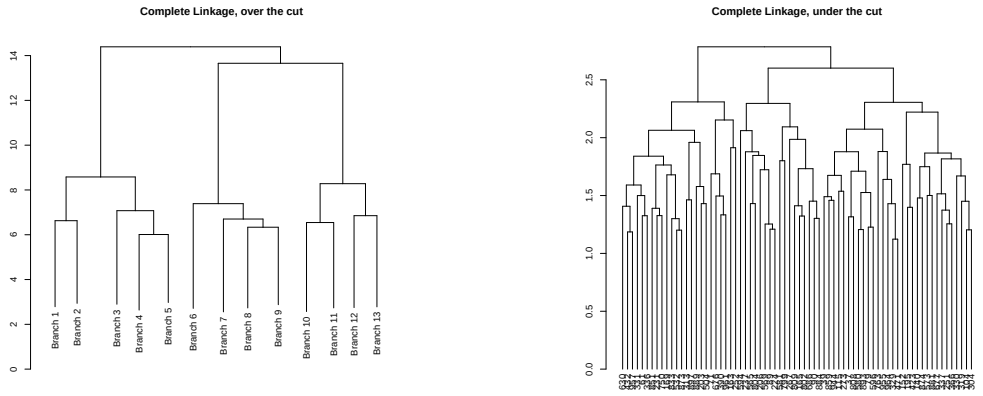



Рис. 5: Дерево кластеризації над та під обраним зрізом. Метод повного зв'язку.

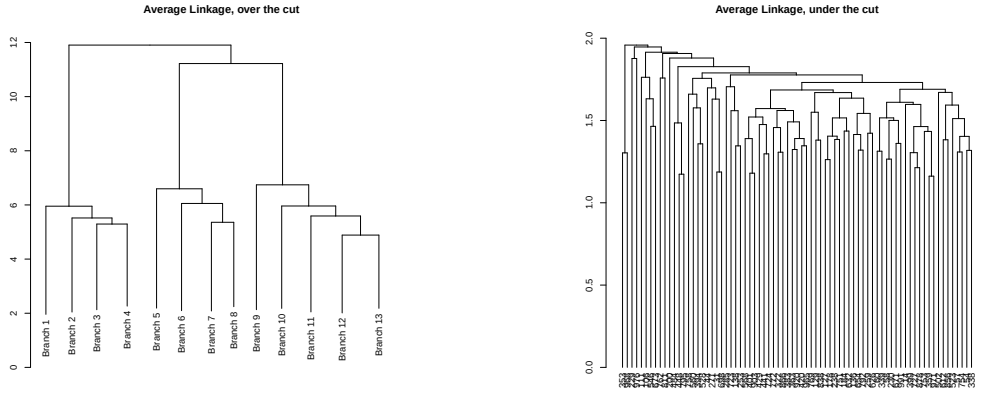



Рис. 6: Дерево кластеризації над та під обраним зрізом. Метод середнього зв'язку.

Аналогічна ситуація до попередньої не спостерігається при використанні інших методів: на нижніх дендрограмах утворюються підгрупи з трьох або чотирьох частин. Перевіримо, чи співпадає ієрархічне розбиття методом одного зв'язку на три групи з тим, що вдалося отримати у першій роботі методом центроїдів.

```
> # Отримуємо кластеризацію на основі методу центроїдів
> kmeans.cluster <- kmeans(data.std, centers = 3, nstart = 50)$cluster
> # Отримуємо кластеризацію на основі зрізу в ієрархічній кластеризації
> hierarchical.cluster <- cutree(data.std.link.single.l2, k = 3)
> # Підрахунок міри відповідності
> rand.index(kmeans.cluster, hierarchical.cluster)

[1] 1

> # Таблиця спряженості на основі обраних кластеризацій
> table(kmeans.cluster, hierarchical.cluster)

             hierarchical.cluster
kmeans.cluster   1   2   3
             1 284   0   0
             2   0   0 312
             3   0 404   0
```

Раніше зазначили про те, що виділяються в районі 13 кластерів внизу кожної дендрограми. Зробимо зріз на рівні цих кластерів та подивимося, чи є співпадіння з кластеризацією на основі методу центроїдів:

```
> # Отримуємо кластеризацію на основі методу центроїдів
> kmeans.cluster <- kmeans(data.std, centers = 13, nstart = 50)$cluster
> # Отримуємо кластеризацію на основі зрізу в ієрархічній кластеризації
> hierarchical.cluster <- cutree(data.std.link.single.l2, k = 13)
> # Підрахунок міри відповідності
> rand.index(kmeans.cluster, hierarchical.cluster)

[1] 0.9843083

> # Таблиця спряженості на основі обраних кластеризацій
> table(kmeans.cluster, hierarchical.cluster)

              hierarchical.cluster
kmeans.cluster  1  2  3  4  5  6  7  8  9 10 11 12 13
            1  34  0  0  0  0  0  0  0  0  0  0  0  0
            2   0  0  0  0  0  0  0  0  0  0  0 72  0
            3  39  0  0  0  0  0  0  0  0  0  0  0  0
            4   0 79  0  0  0  0  0  0  0  0  0  0  0
            5   0  0 88  0  0  0 74  0  0  0  0  0  0
            6   0  0  0  0  0  0  0 85  0  0  0  0  0
            7   0  0  0 91  0  0  0  0  0  0  0  0  0
            8   0  0  0  0  0  0  0  0  0  0 64  0  0
            9   0  0  0  0  0  0  0  0  0  0  0  0 70
            10  0  0  0  0  0 82  0  0  0  0  0  0  0
            11  0  0  0  0  0  0  0  0 73  0  0  0  0
            12  0  0  0  0 74  0  0  0  0  0  0  0  0
            13  0  0  0  0  0  0  0  0  0 75  0  0  0
```

Обрана ієрархічна кластеризація вийшла майже такою, якою б могла вийти на основі методі центроїдів. Є нев'язка для п'ятої групи за другим методом: точки звідти розкинули по двом іншим групам.

Зобразимо розмітку на просторовій діаграмі розсіювання, спроектувавши попередньо на головні компоненти.

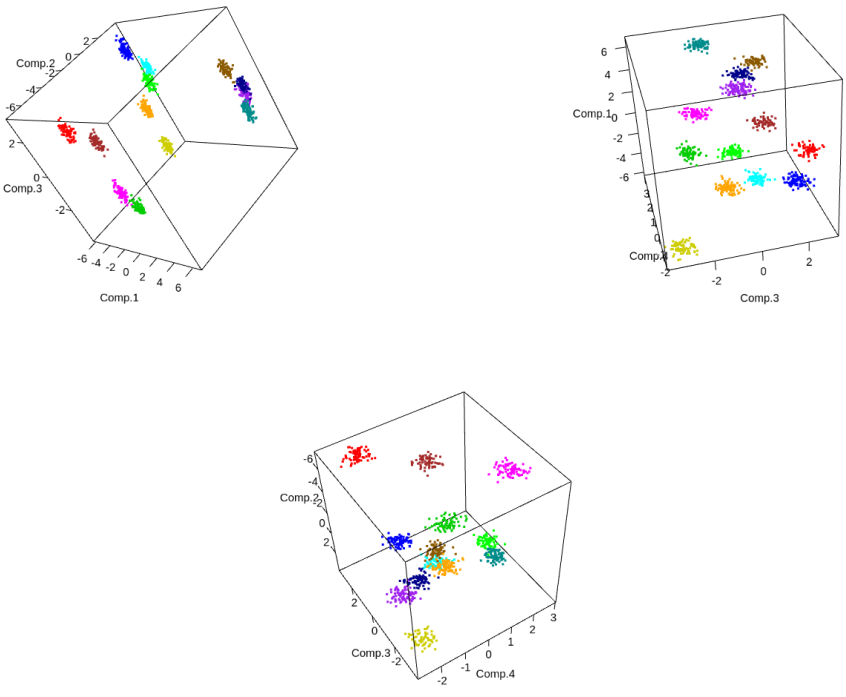



Рис. 7: Просторова діаграма розсіювання на перші чотири головні компоненти.

## Ієрархічна кластеризація на напоях.

```
> # Зчитуємо дані про напої
> data.drinks <- read.csv("./drinks_100ml.csv", header=T)
> data.drinks <- data.drinks[,c("Вуглеводи", "Ціна")]
> colnames(data.drinks) <- c("Carbohydrates", "Price")
```

Аби було доречно використовувати метричні методи, треба звести до єдиної шкали (стандартизуємо дані):

```
> # Стандартизуємо дані
> data.drinks.std <- scale(data.drinks)
```

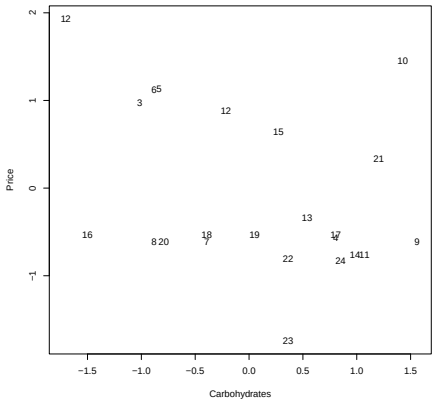
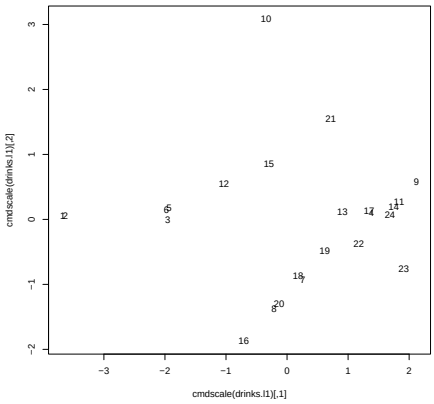



Рис. 8: Діаграма розсіювання даних після стандартизації.

Обчислимо відстані.

```
> drinks.l2 <- dist(data.drinks.std, method = "euclidean")
> drinks.l1 <- dist(data.drinks.std, method = "manhattan")
> drinks.mm <- dist(data.drinks.std, method = "maximum")
```

Цікаво подивится на те, як зміниться форма даних після шкалування.



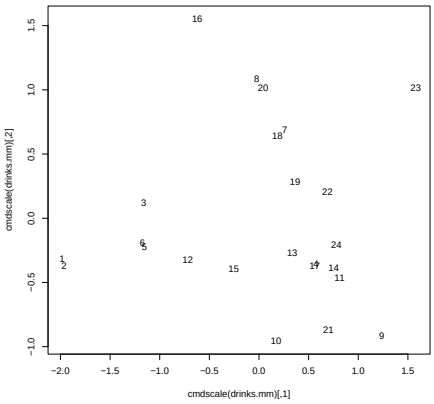




Рис. 9: Діаграма розсіювання даних після шкалування (відносно відстані сіті-блок та мінімаксної).

**Ієрархічна кластеризація: метод одного зв'язку.**

```
> # Кластеризація за l2-метрикою
> drinks.std.link.single.l2 <- hclust(
+   d = drinks.l2, method = "single"
+ )
> # Кластеризація за l1-метрикою
> drinks.std.link.single.l1 <- hclust(
+   d = drinks.l1, method = "single"
+ )
> # Кластеризація за мінімаксною метрикою
> drinks.std.link.single.mm <- hclust(
+   d = drinks.mm, method = "single"
+ )
```

Тепер покажемо дендрограми кластеризації:

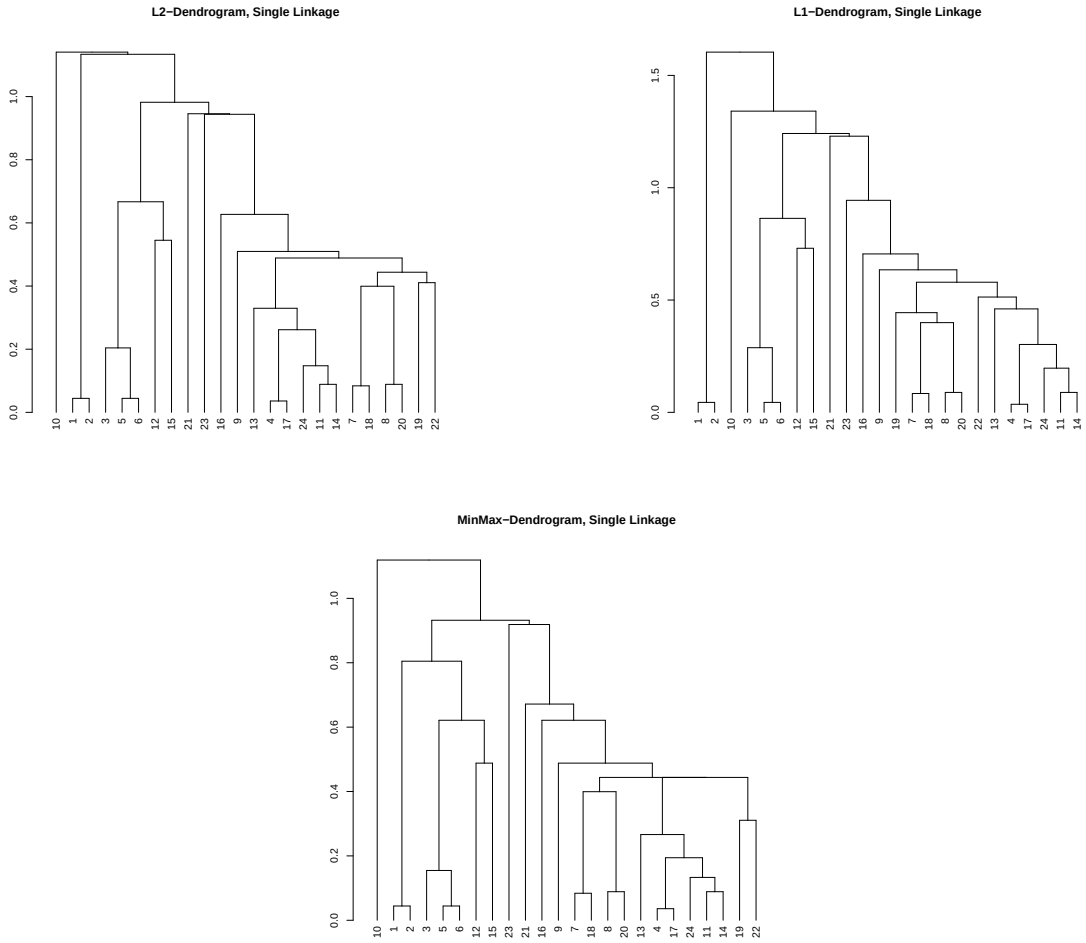



Рис. 10: Дендрограми кластеризацій в залежності від метрики. Метод одного зв'язку.

Метод одного зв'язку не впорався з загальних відокремленням даних по частинам, судячи з рисунків. Ні, звісно, те що потрібно було зробити – представити ієрархію, то добре, але чіткого розбиття на ній неможливо побачити, лише розглядаючи конкретні частини в ній (тобто коли копаємося вглиб гілок, то побачимо локальні розбиття).

**Ієрархічна кластеризація: метод повного зв'язку.**

```
> # Кластеризація за l2-метрикою
> drinks.std.link.complete.l2 <- hclust(
+   d = drinks.l2, method = "complete"
+ )
> # Кластеризація за l1-метрикою
> drinks.std.link.complete.l1 <- hclust(
+   d = drinks.l1, method = "complete"
+ )
> # Кластеризація за мінімаксною метрикою
> drinks.std.link.complete.mm <- hclust(
+   d = drinks.mm, method = "complete"
+ )
```

Тепер покажемо дендрограми кластеризації:

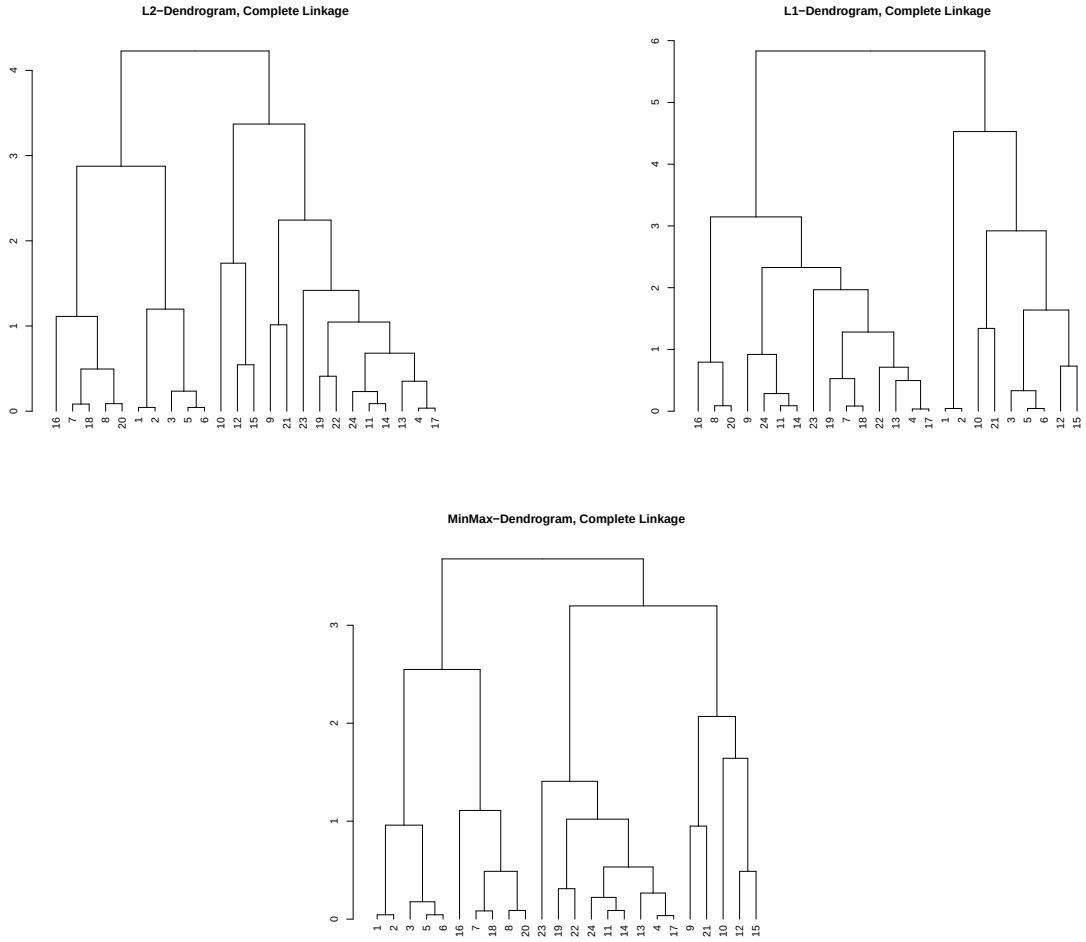



Рис. 11: Дендрограми кластеризацій в залежності від метрики. Метод одного зв'язку.

Добре видно, що відокремлюються по два великі кластери в цілому. А кожен з них розбивається на одну чи дві частини.

**Ієрархічна кластеризація: метод середнього зв'язку.**

```
> # Кластеризація за l2-метрикою
> drinks.std.link.average.l2 <- hclust(
+   d = drinks.l2, method = "average"
+ )
> # Кластеризація за l1-метрикою
> drinks.std.link.average.l1 <- hclust(
+   d = drinks.l1, method = "average"
+ )
> # Кластеризація за мінімаксною метрикою
> drinks.std.link.average.mm <- hclust(
+   d = drinks.mm, method = "average"
+ )
```

Тепер покажемо дендрограми кластеризації:

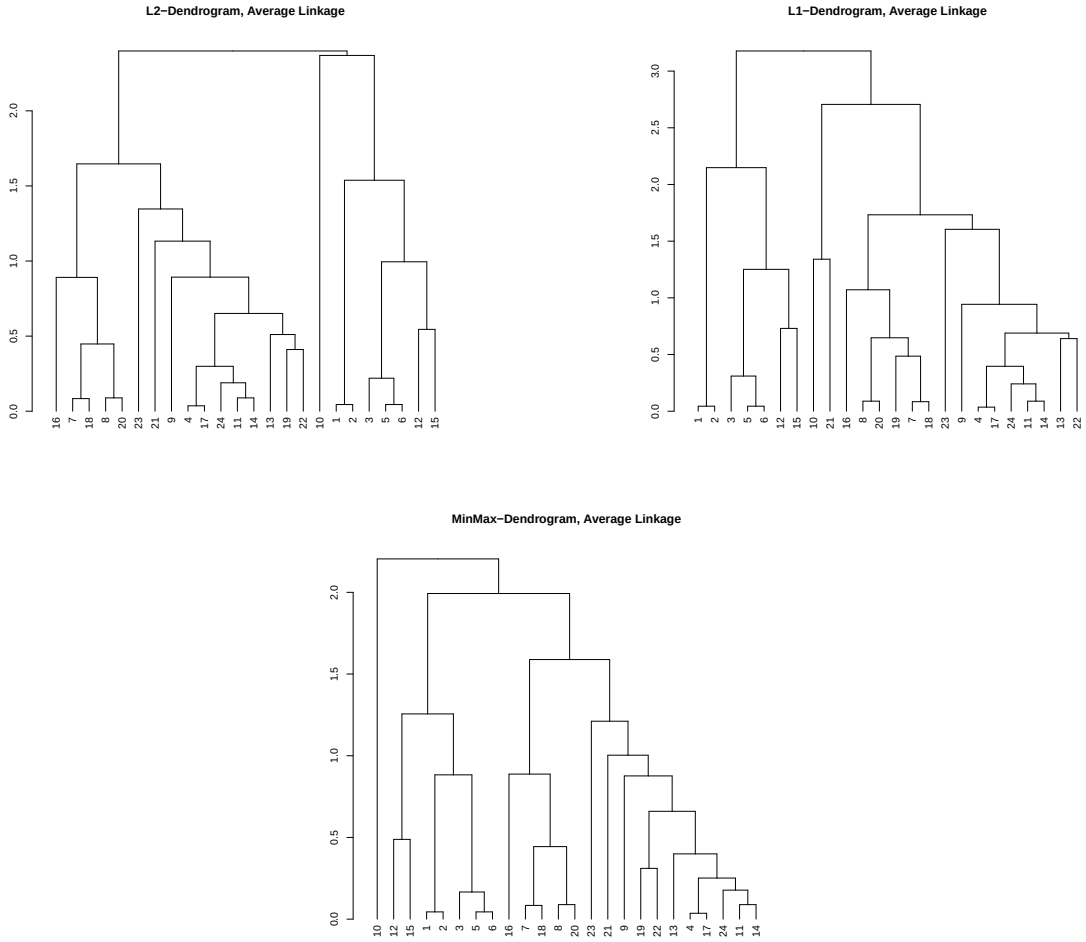



Рис. 12: Дендрограми кластеризацій в залежності від метрики. Метод середнього зв'язку.

Картина виходить схожою як в попередньому випадку (в характері кластеризації), хоча цікаво що за мінімаксною метрикою, середній зв'язок виокремлює одну з точок в окремий кластер.

**Підрахунок кофенетичних відстаней для методу одного зв'язку:**

```
> # L2-відстань
> cor(cophenetic(drinks.std.link.single.l2), drinks.l2)

[1] 0.7199813

> # L1-відстань
> cor(cophenetic(drinks.std.link.single.l1), drinks.l1)

[1] 0.7725963

> # MinMax відстань
> cor(cophenetic(drinks.std.link.single.mm), drinks.mm)

[1] 0.6687227
```

**Підрахунок кофенетичних відстаней для методу повного зв'язку:**

```
> # L2-відстань
> cor(cophenetic(drinks.std.link.complete.l2), drinks.l2)

[1] 0.6701033

> # L1-відстань
> cor(cophenetic(drinks.std.link.complete.l1), drinks.l1)

[1] 0.7060126

> # MinMax відстань
> cor(cophenetic(drinks.std.link.complete.mm), drinks.mm)

[1] 0.6466011
```

**Підрахунок кофенетичних відстаней для методу середнього зв'язку:**

```
> # L2-відстань
> cor(cophenetic(drinks.std.link.average.l2), drinks.l2)

[1] 0.7673505

> # L1-відстань
> cor(cophenetic(drinks.std.link.average.l1), drinks.l1)

[1] 0.7410921

> # MinMax відстань
> cor(cophenetic(drinks.std.link.average.mm), drinks.mm)

[1] 0.7726483
```

Наприлад, візьмемо кластеризацію на основі методу середнього зв'язку та з евклідовою метрикою. Підрахуємо індекс відповідності з кластеризацією за методом центроїдів:

```
> # Кластеризації за ''хорошими'' підходами
> cut.l2 <- cutree(drinks.std.link.average.l2, k = 2)
> cut.mm <- cutree(drinks.std.link.average.mm, k = 3)
> # Кластеризація на оснвоі методу центроїдів
> set.seed(182)
> kmeans.res <- kmeans(data.drinks, centers = 2)$cluster
> # Індекс Ренда для двох кластерів
> c(rand.index(kmeans.res, cut.l2), rand.index(kmeans.res, cut.mm))

[1] 0.9166667 0.8913043

> # Таблиця відповідностей
> table(kmeans.res, cut.l2)

          cut.l2
kmeans.res  1  2
         1  8  1
         2  0 15

> # Таблиця відповідностей
> table(kmeans.res, cut.mm)

          cut.mm
kmeans.res  1  2  3
         1  7  1  1
         2  0 15  0
```

Побудовані кластеризації дещо схожі на ті, що були отримані у першій роботі.

# Висновки.

Ієрархічна кластеризація добре застосовна на використаних даних з першої роботи. Отриманим розбиттям на основі ієрархій властиві ті сам закономірності, що були виявлені нами у першій роботі.