Знакомство с многомерными данными

Анализ и визуализация многомерных данных с использованием R

Вадим Хайтов, Марина Варфоломеева

Вы сможете

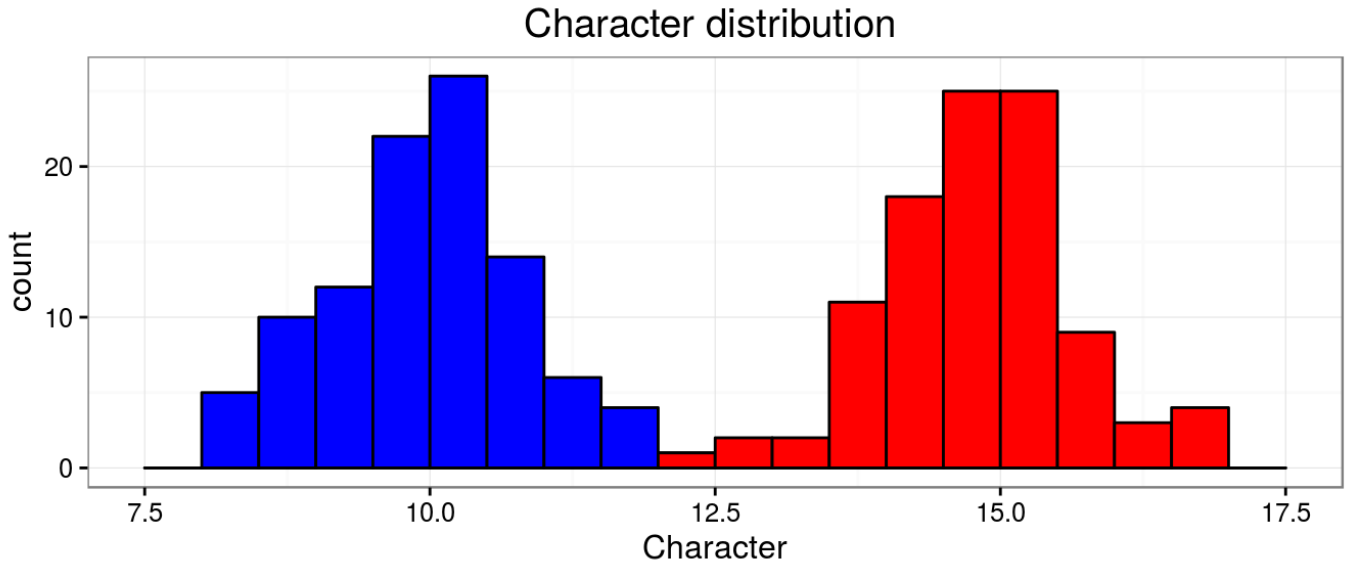
- Объяснить почему для некоторых задач больше подходят многомерные даные
- Объяснить суть понятия "многомерое пространство признаков"
- Представить многомерые данные в виде матриц описания значений признаков для объектов
- Оценить сходство/различие между объектами с помощью специальных коэффициентов
- Описать *взаиморасположение* объектов многомерном пространстве признаков с помощью матриц
- Визуализировать взаиморасположение объектов с помощью простейших методов

Часть 1. Общая характеристика многомерных методов

Почему нужны многомерные методы?

Пусть у нас меется две группы объектов, у которых мы изучили некий признак. Мы хотим тестировать гипотезу о том, что эти две группы различаются.

Вспомним логику тестирования гипотез.



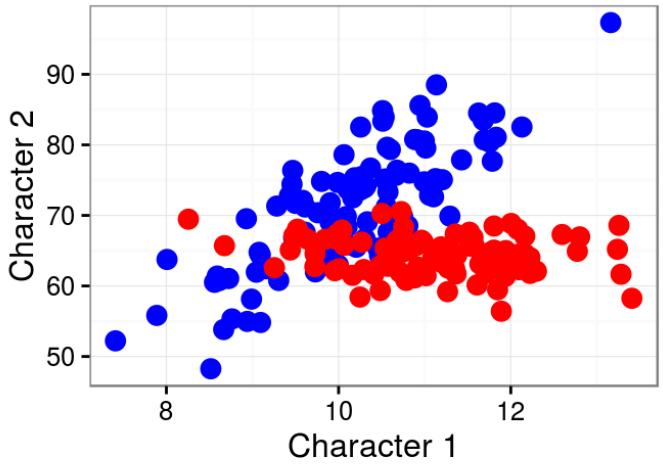
Почему нужны многомерные методы?

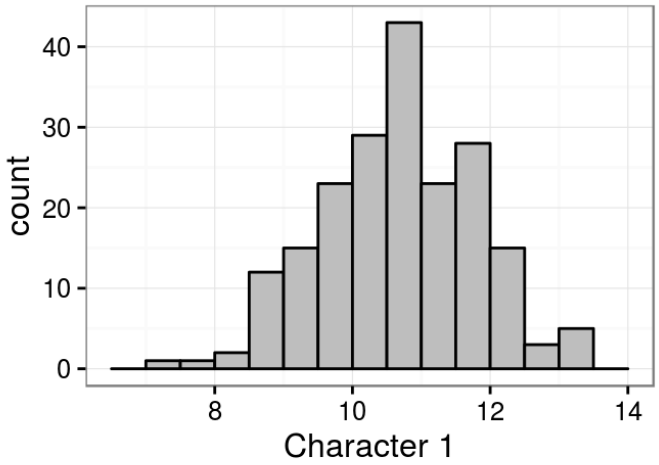
Теперь представим, что наш объект, по своей природе, не может быть описан только по одному признаку

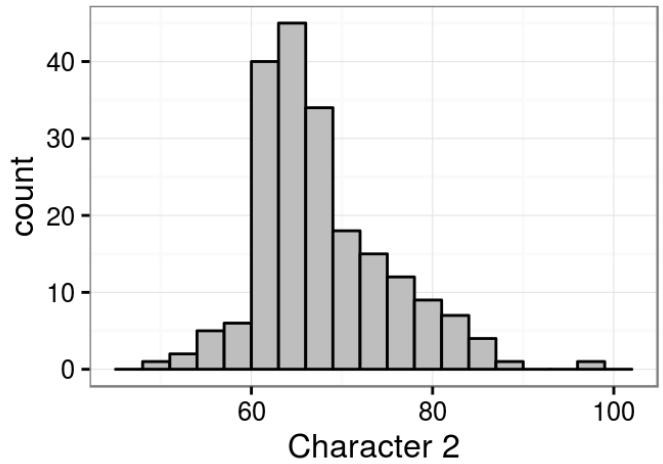
- Сообщества (признаки - виды)
- Форма тела (признаки - размеры тех или иных частей)
- Социальная активность животного (признаки - проявление того или иного паттерна)
- Общественное мнение (признаки - ответы на разные вопросы анкет)
- Транскриптом (признаки - транскрипты)

Почему нужны многомерные методы?

Предположим, что объекты характеризуются только двумя признаками







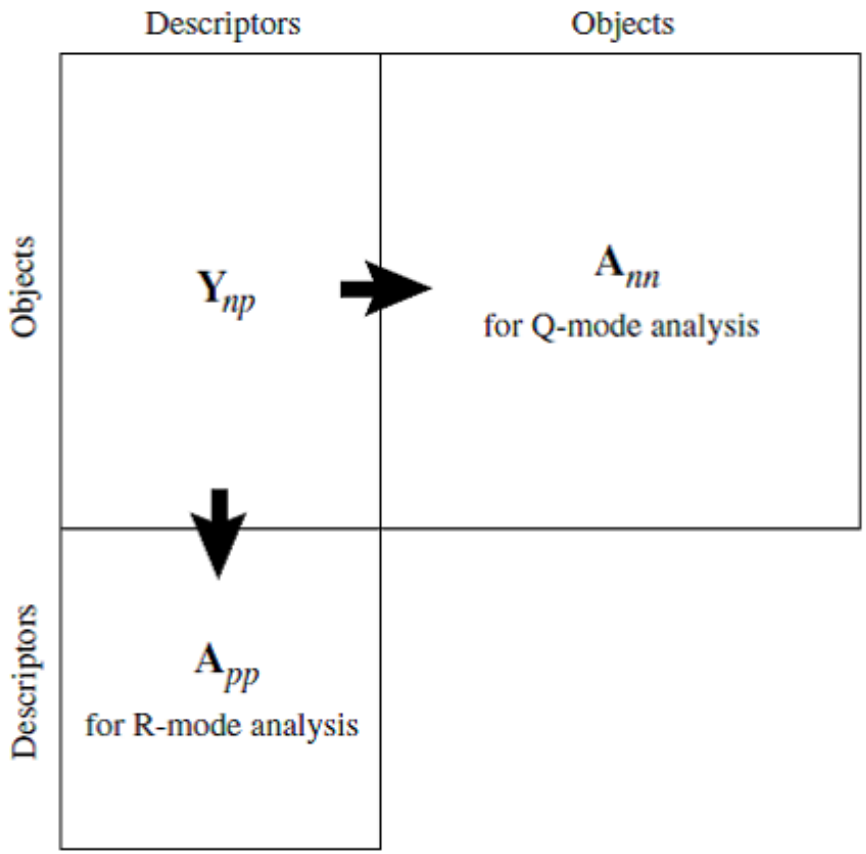
Какие задачи решаются методами могомерной статистики?

- Выявление взаимоотношений (сходства-различия) между объектами (или признаками):
  - Классификация (Кластерный анализ)
  - Ординация. В том числе картирование пространственно выраженных объектов (nMDS, PCA, RDA).
- Тестирование гипотез о различиях между группами объектов (ANOSIM, PERMANOVA).
- Выявление связи между группами признаков (тест Мантела, BIOENV, CCA).

Признаки и объекты

Данные представляются в виде таблицы (матрицы), где строками являются обекты (Objects), а столбцами признаки (Descriptors).

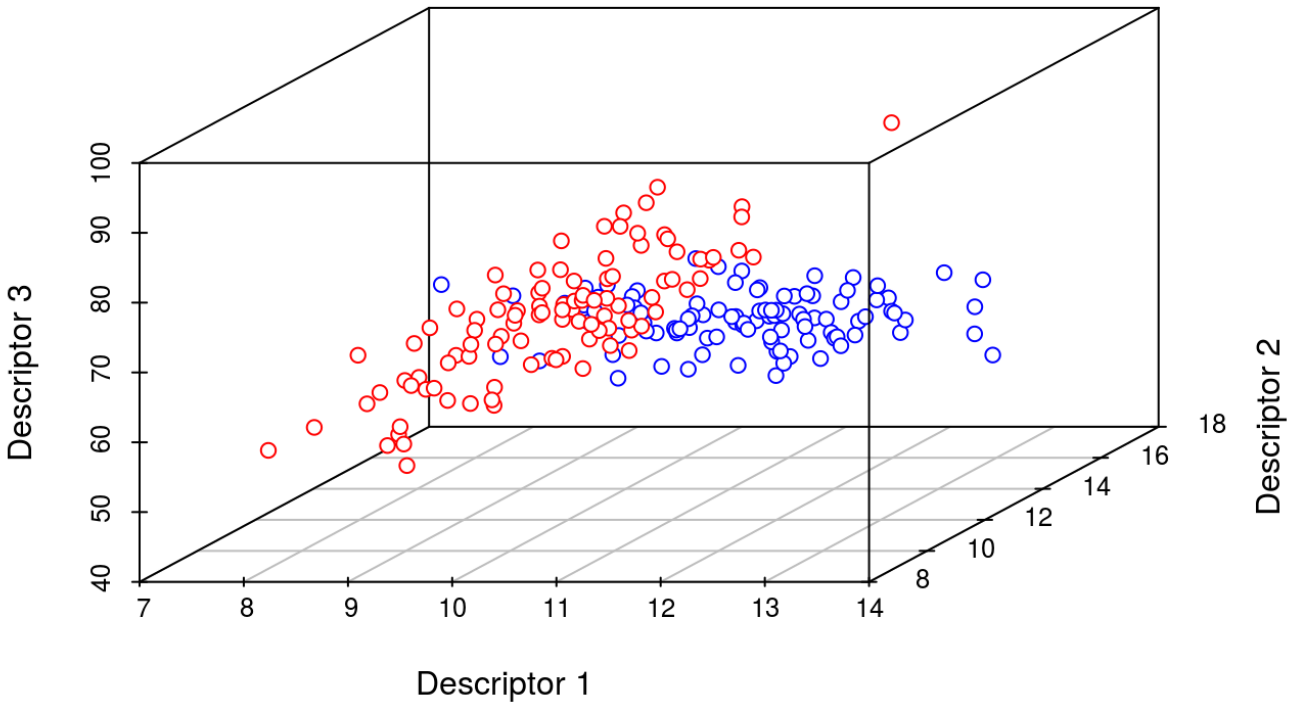
R и Q анализы



- R-анализ: Выясняем взаимоотношения между признаками
- Q-анализ: Выясняем взаимоотношения между объектами

Геометрическая интерпретация Q-анализа

- Признаки - оси
- Объекты - точки

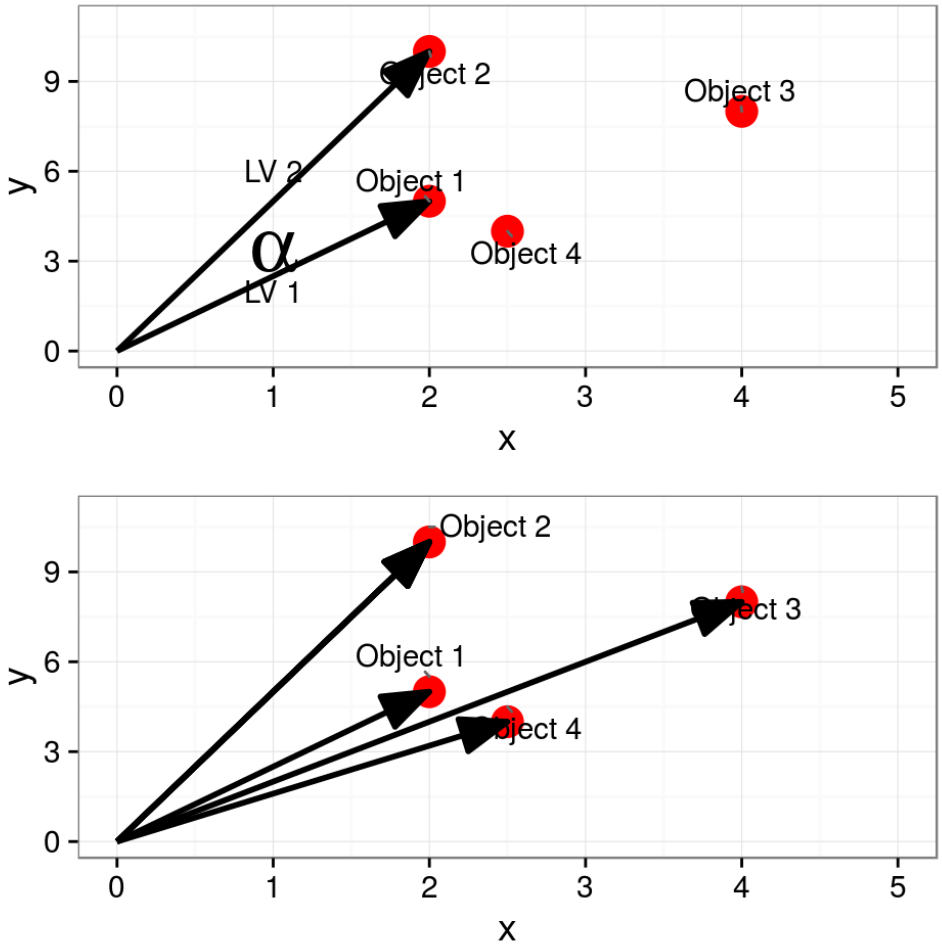


Описание расположения объектов в многомерном пространстве признаков

В большинстве случаев нас интересуют не абсолютные значения координат (признаков), а *взаиморасположение* точек в многомерном пространстве.

Существует два основных способа описания.

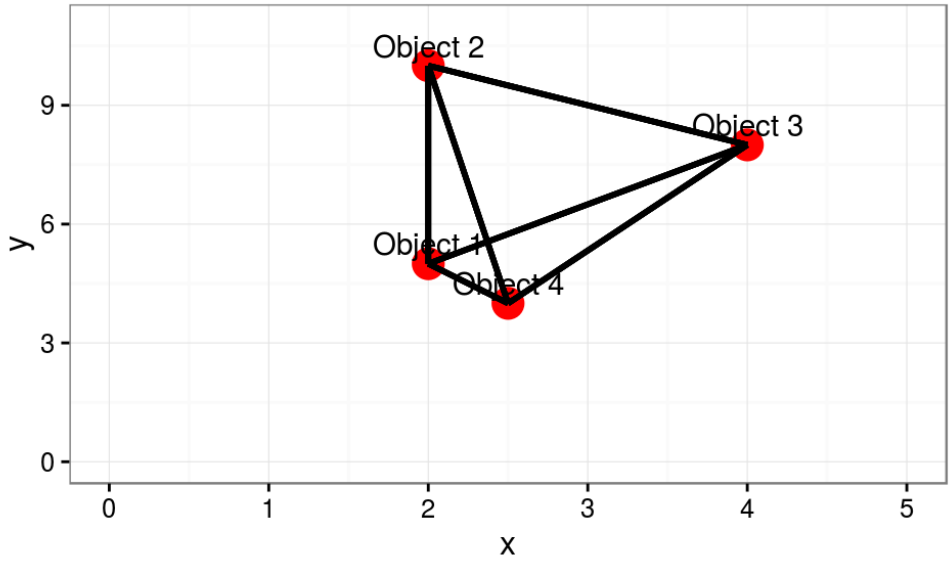
Способ 1. Геометрическое описание (линейная алгебра)



Для описания взаморасположения точек необходимо иметь два набора данных

Матрицу углов между векторами (косинусов углов)

Способ 2. Через вычисление матрицы попарных расстояний (Similarity/Dissimilarity matrix)



В анализ вовлекается матрица попраных расстояний (сходств) между объектами. Эта матрица однозначно описывает взаиморасположение между объектами.

Этот способ представления взаиморасположения лежит в основе *Иерархического кластерного анализа, MDS, теста Мантела, ANOSIM, PERMANOVA, процедуры BIOENV*

Три способа изображения матрицы расстояний

```
##     1   2   3   4
## 1 0.0 5.0 3.6 1.1
## 2 5.0 0.0 2.8 6.0
## 3 3.6 2.8 0.0 4.3
## 4 1.1 6.0 4.3 0.0
```

```
##     1   2   3   4
## 1 0.0
## 2 5.0 0.0
## 3 3.6 2.8 0.0
## 4 1.1 6.0 4.3 0.0
```

```
##     1   2   3
## 2 5.0
## 3 3.6 2.8
## 4 1.1 6.0 4.3
```

Количество чисел в треугольной матрице сходства/различия:

$$N = \frac{n^2 - n}{2}$$

Матрица расстояний в развернутом виде (Unfolded similarity/dissimilarity matrix)

```
##     1   2   3
## 2 5.0
## 3 3.6 2.8
## 4 1.1 6.0 4.3
```

```
## [1] 5.0 3.6 1.1 2.8 6.0 4.3
```

Часть 2. Меры сходства и различия между объектами (Resemblance coefficients)

Кто что ест? Потребление белка разного происхождения в Европе

9 признаков 25 объектов:

- Country: Страна
- RdMeat: Красное мясо
- WhMeat: Белое мясо
- Eggs: Яйца
- Milk: Молоко
- Fish: Рыба
- Cereal: Зерновые культуры
- Starch: Крахмал-содержащая пища
- Nuts: Бобовые, орехи и масличные культуры
- Fr&Veg: Фрукты и овощи

Читаем данные

```
protein <- read.table("data/Protein.txt", header = TRUE, sep = "\t")
head(protein)
```

```
##           Country RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts
## 1         Albania    10.1       1.4  0.5  8.9  0.2    42.3    0.6  5.5
## 2         Austria     8.9      14.0  4.3 19.9  2.1    28.0    3.6  1.3
## 3         Belgium    13.5       9.3  4.1 17.5  4.5    26.6    5.7  2.1
## 4        Bulgaria     7.8       6.0  1.6  8.3  1.2    56.7    1.1  3.7
## 5 Czechoslovakia     9.7      11.4  2.8 12.5  2.0    34.3    5.0  1.1
## 6         Denmark    10.6      10.8  3.7 25.0  9.9    21.9    4.8  0.7
##   Fr.Veg
## 1    1.7
## 2    4.3
## 3    4.0
## 4    4.2
## 5    4.0
## 6    2.4
```

Подготовка данных для анализа

Переменные могут быть измерены в разных шкалах

- численность, биомасса и проективное покрытие организмов
- температура воды, соленость и концентрация биогенов
- Размеры частей тела, количество частей тела, площадь частей тела

Для таких случаев необходима стандартизация величин.
Этот прием наиболее важен для геометрического способа представления многомерных данных.

$$x_{stand} = \frac{x_i - \bar{x}}{\sigma_x}$$

- **Вопрос:** Какими свойствами обладают стандартизованные величины?
- Среднее зачние равно нулю.
- Среднеквадратичное отклонение равно единице.

Подготовка данных для анализа

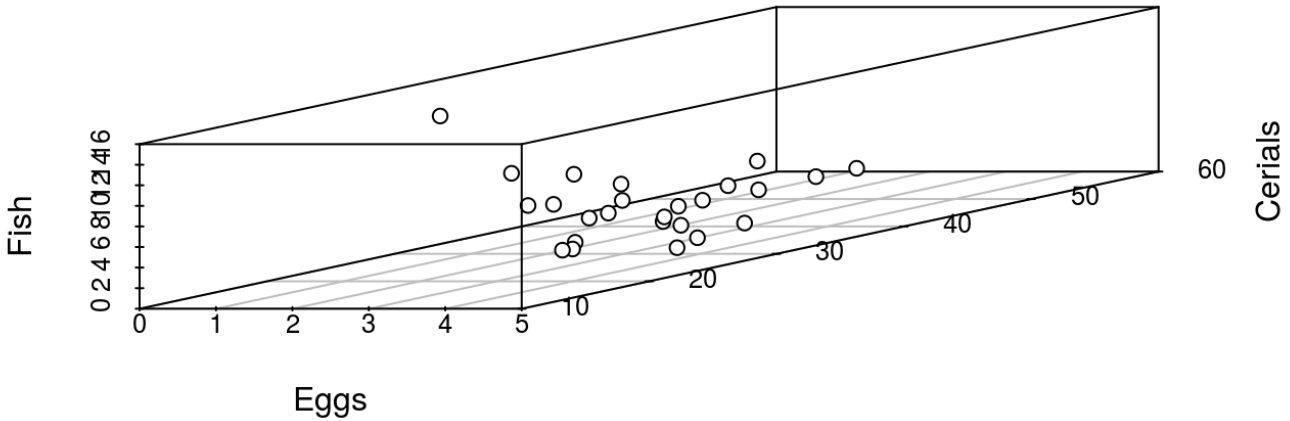
В ряде случаев (особенно в экологических исследованиях) необходимо перевести абсолютные значения в относительные (доли от суммы или от максимума)

$$x_{rel} = \frac{x_i}{\sum x_i} \times 100\%$$

$$x_{rel} = \frac{x_i}{max(x_i)} \times 100\%$$

Подготовка данных для анализа

Часто возникает ситуация, когда один признак (или несколько прзнаков) имеет существенно более высокие абсолютные значения, чем все остальные.



В такой ситуации необходима *трансформация* , которая "уравнивает" силу влияния признаков.

По силе эффекта трансформирующие функции распределяются так:

Отсутствие трансформации ⇒ Квадратный корень ⇒ Корень четвертой степени ⇒ Логарифм $log(x_i + 1)$ ⇒ Присутствие отсутствие (1, 0)

Задание:

На основе датасета `protein` создайте матрицу, содержащую относительные величины (доли протеина каждого типа в общей количестве потребляемых протеинов)

Hint: Воспользуйтесь функцией `apply()`

Решение

```
total <- apply(protein[, -1], MARGIN = 1,FUN = sum)
protein_stand <- protein[, -1] / total
protein_stand$Country <- protein$Country

head(protein_stand)
```

```
##   RedMeat WhiteMeat    Eggs   Milk    Fish Cereals  Starch   Nuts
## 1  0.1419    0.0197 0.00702 0.1250 0.00281   0.594 0.00843 0.0772
## 2  0.1030    0.1620 0.04977 0.2303 0.02431   0.324 0.04167 0.0150
## 3  0.1546    0.1065 0.04696 0.2005 0.05155   0.305 0.06529 0.0241
## 4  0.0861    0.0662 0.01766 0.0916 0.01325   0.626 0.01214 0.0408
## 5  0.1171    0.1377 0.03382 0.1510 0.02415   0.414 0.06039 0.0133
## 6  0.1180    0.1203 0.04120 0.2784 0.11024   0.244 0.05345 0.0078
##   Fr.Veg        Country
## 1 0.0239        Albania
## 2 0.0498        Austria
## 3 0.0458        Belgium
## 4 0.0464       Bulgaria
## 5 0.0483 Czechoslovakia
## 6 0.0267        Denmark
```

Наиболее частые коэффициенты сходства/различия и расстояния

Знакомимся с пакетом `{vegan}` (Oksanen et al., 2015)

```
##        1     2     3     4     5     6     7     8     9    10    11
## 2  23.18
## 3  21.65  7.87
## 4  15.69 32.30 32.79
## 5  15.15 10.31 10.61 24.01
## 6  30.16 11.96 11.12 40.33 19.42
## 7  22.87 10.74  8.93 33.61 10.61 15.18
## 8  30.99 17.42 17.60 40.34 24.02 12.25 23.83
## 9  23.17 11.01  6.01 33.26 13.44 12.72 13.86 18.18
## 10 12.14 19.53 18.25 19.32 15.03 24.47 22.11 24.11 18.25
## 11 13.16 16.97 18.78 18.40  9.18 26.73 17.52 29.98 21.26 14.93
## 12 27.90 10.04  9.15 38.36 17.58  8.94 16.18 11.57 10.15 22.98 25.02
## 13 10.62 14.69 13.57 21.01  8.71 21.60 15.62 23.76 15.16  7.98 10.70
## 14 28.30  6.76  9.68 38.53 16.35  8.36 13.27 14.68 12.35 24.06 23.21
## 15 26.81 13.68 10.80 38.17 18.73  6.69 14.99 11.69 13.15 21.42 25.67
## 16 17.64  9.94 12.20 24.49  8.26 17.81 14.87 19.36 14.01 12.36 11.99
## 17 23.11 22.93 19.20 33.29 19.06 23.93 15.16 31.18 21.85 22.16 22.03
## 18 10.32 25.26 25.88  8.33 17.31 33.29 26.73 33.40 27.12 13.34 11.64
## 19 17.15 17.44 13.92 28.89 13.07 21.19 11.78 26.57 17.22 16.27 16.17
## 20 29.99 13.03 11.63 41.48 20.43  4.80 15.58 11.91 14.03 25.16 27.66
## 21 24.93  7.58  7.53 35.51 14.97  9.65 14.65 13.05  8.18 20.04 22.14
## 22 24.31 12.92  6.83 36.42 16.50 11.74 14.78 15.95  6.99 20.32 24.17
## 23 11.03 19.04 18.42 16.68 12.64 25.33 21.37 24.69 19.43  8.18 12.49
```

Сходства и различия

- Сходство (S) достигает максимума, когда объекты обладают идентичными признаками, различие (D), наоборот - достигает минимума.
- Обычно (но не всегда) коэффиценты сходства распределены от 0 до 1.
- Тогда $D = 1 - S$, или $D = \sqrt{1 - S}$, или $D = \sqrt{1 - S^2}$.
- Показатели расстояния можно нормировать. $D_{norm} = \frac{D}{D_{max}}$, или $D_{norm} = \frac{D - D_{min}}{D_{max} - D_{min}}$.

Проблема двойных нулей (Double zeros probem)

```
##         D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
## Object1  0  0  0  0  0  0  0  2  2   1
## Object2  0  0  0  0  0  4  5  0  0   1
```

О чем говорит то, что признаки D1, D2, D3, D4, D5 не были отмечены у двух объектов?

- Вариант 1. Это ничего не означает.
- Вариант 2. Отсутствие признака - дополнительное сходство между объектами.

От ответа на этот вопрос зависит какой коэффициент выбрать.

Два типа коэффициентов

Меры сходства/различия не учитывающие двойные нули. Эти коэффициенты не изменяются если в данные будут добавлены двойные нули (например, при увеличеннии количества описанных объектов).

Примеры:

- Евклидово расстояние
- расстояние по Манхеттену.

Меры сходства/различия учитывающие двойные нули. Эти коэффициенты изменяются при появлении двойных нулей. Сходство возрастает за счет того, что отсутсвие признака считается тоже сходством.

Пример:

- Коэффициенты корреляции.

Меры расстояния, или метрики

Свойства:

- Если $a = b$, то $D(a, b) = 0$
- Симметричность $D(a, b) = D(b, a)$
- Справедливо неравество треугольника $D(a, b) + D(b, c) \geq D(a, c)$

Наиболее популярные меры расстояния

Важно: метрики неадекватно оценивают степень различия при большом количестве нулей.Очень чувствительны к выбросам.

Нестандартизованные

- Евклидово расстояние:

$$D = \sqrt{\sum (x_{i,j} - x_{i,k})^2}$$

- Расстояние по Манхеттену (Manhattan metric, taxicab metric, city-block metric):

$$D = \sum | x_{i,j} - x_{i,k} |$$

Стандартизованные

Удобнее так как признаки могут быть измерны в разных шкалах

- Расстояние по Канберре (Canberra metric):

Неметрические коэффициенты сходства/различия (Similarity/Dissimilarity)

- Корреляция Браве-Пирсона:

$$R = \frac{cov(X, Y)}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

Коэффициент Браве-Пирсона варьирует от -1 до 1.

Обычно используется в R анализе

- Коэффициент Брея-Куртиса (Bray-Curtis dissimilarity):

$$D = \frac{\sum | x_{i,j} - x_{i,k} |}{\sum x_{i,j} + \sum x_{i,k}}$$

Это самый распространенный в экологии коэффициент.

Коэффициенты для бинарных данных

В основе лежит четырехпольная таблица

	+	−
+	a	b
−	c	d

- a - сходство объектов по наличию признака
- b - различие
- c - различие
- d - сходство по отсутствию признака

Коэффициенты для бинарных данных

- Доля несовпадений:

$$D = \frac{b + c}{a + b + c}$$

- Коэффициетт Жаккара:

$$S = \frac{a}{a + b + c}$$

- Коэффициент Сёренсена:

$$S = \frac{2a}{2a + b + c}$$

NB! Коэффициент Сёренсена - это коэффициент Брея-Куртиса, вычисленный для значений, оцененных как 1 или 0.

- ϕ-корреляция Пирсона

$$\phi = \frac{ad - bc}{(a + b)(c + d)(a + c)(b + d)}$$

Используется в R-анализе

Коэффициент Говера

Обобщенный коэффициент, который применяется для случаев, когда одни признаки объекта описаны, как количественные величины, а другие - как бинарные данные (или даже качественные даные).

$$D = \frac{1}{p} \frac{\sum W_i \frac{|x_{i,j} - x_{i,k}|}{\max x_{i,j} - \min x_{i,k}}}{\sum W_i}$$

W_i = 0 Если отсутсвует информация о $x_{i,j}$ и/или $x_{i,k}$ отсутствует

W_i = 1 Если присутсвует информация как о $x_{i,j}$ так и о $x_{i,k}$

Отчего зависит выбор коэффициента?

1. От природы материала (признаки могут иметь количественную, бинарую и качественную оценку).
2. От характера тестируемой гипотезы (какой аспект природы сходства-различия хочет выразить автор).
3. В экологии: от того, насколько мы хотим учитывать вклад редких и малочисленных видов.
4. От взглядов исследователя на природу сходства/различия между объектами.
5. От типа анализа (R или Q)

На практике для Q-анализа (в экологии и смежных дисцилинах)

Для количественных данных:

- Если мало нулей (данные, описывающие параметры среды), то Евклидово расстояние
- Если много нулей (данные, описывающие обилия организмов) коэффициент Брея-Куртиса или Говера

Для бинарных данных - Коэффициенты Сёренсена или Жаккара

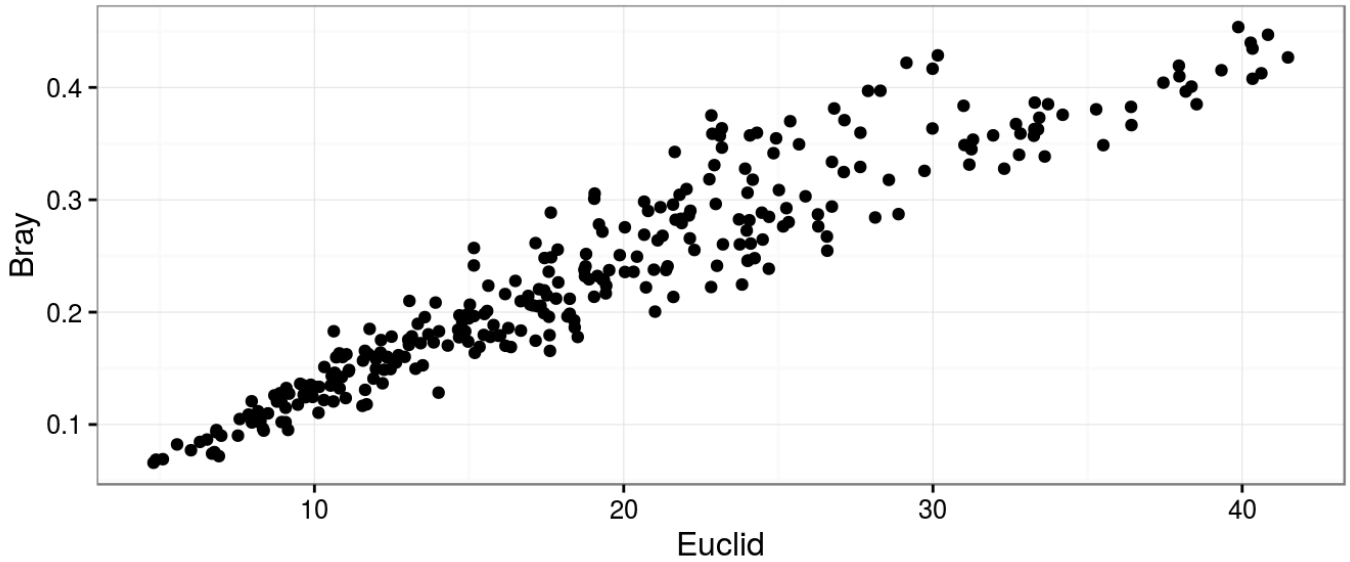
Задание:

Отразите графически как взаимосвязаны коэффициент Брея-Куртиса и Евклидово расстояние

Решение

```
BCD <- as.vector(vegdist(protein[,-1], method = "bray"))
ED <- as.vector(vegdist(protein[,-1], method = "euclidean"))

distances <- data.frame(Bray = BCD, Euclid = ED)
ggplot(distances, aes(x = Euclid, y = Bray)) + geom_point() + theme_bw()
```



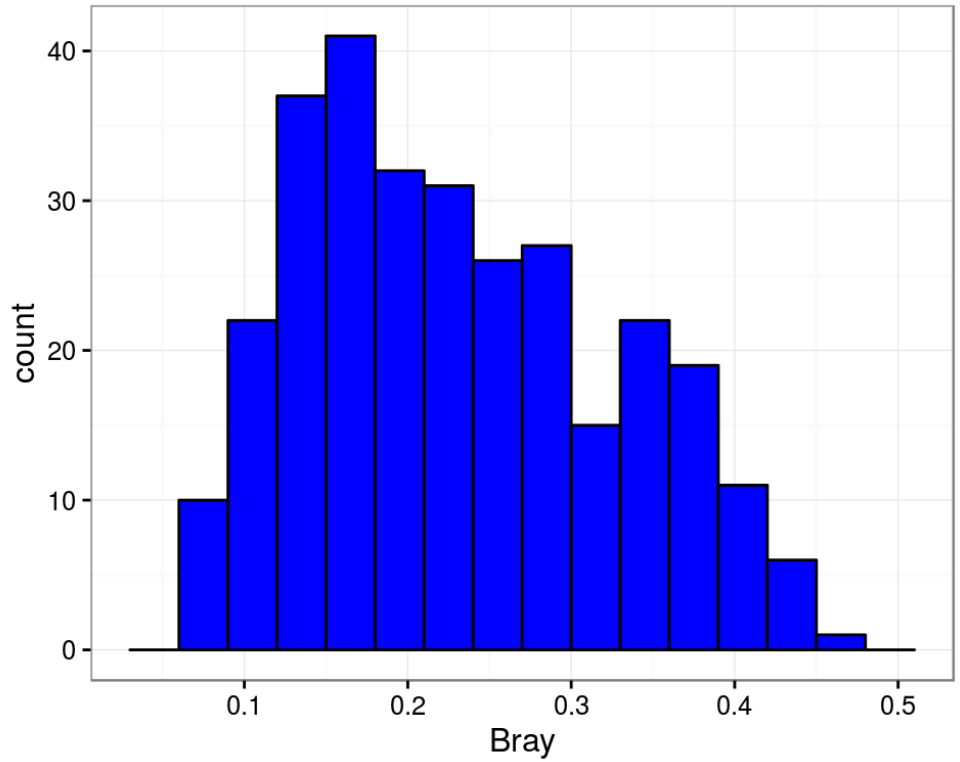
Часть 3. Простейшие методы анализа матриц сходства/различия

Задание:

Постройте гистограмму, отражающую частотное распределение коэффициентов Брея-Куртиса, расчитанных для матрицы `protein`

Решение

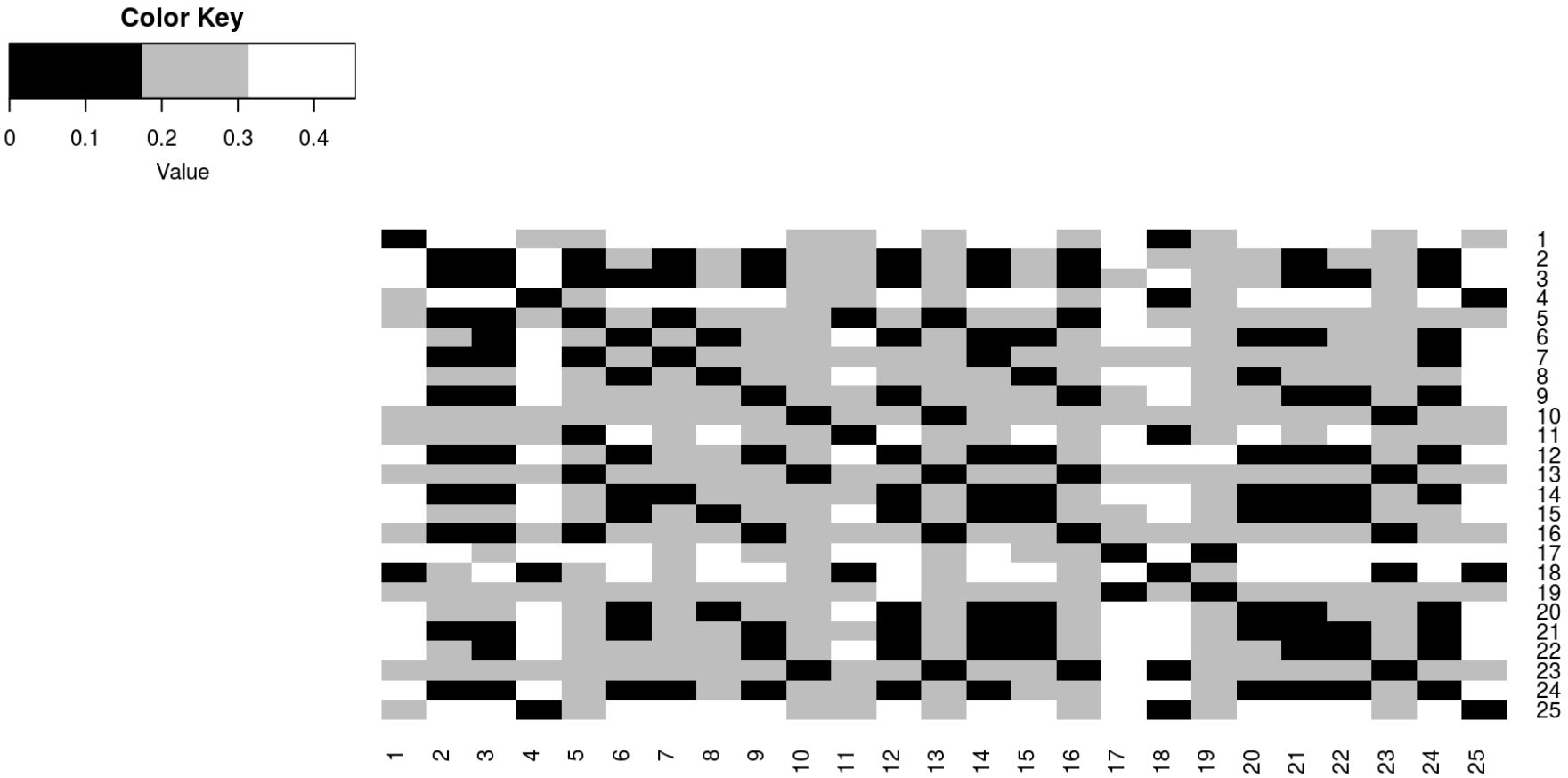
```
ggplot(distances, aes(x = Bray)) +
  geom_histogram(binwidth = 0.03,
                 fill = "blue",
                 color = "black") +
  theme_bw()
```



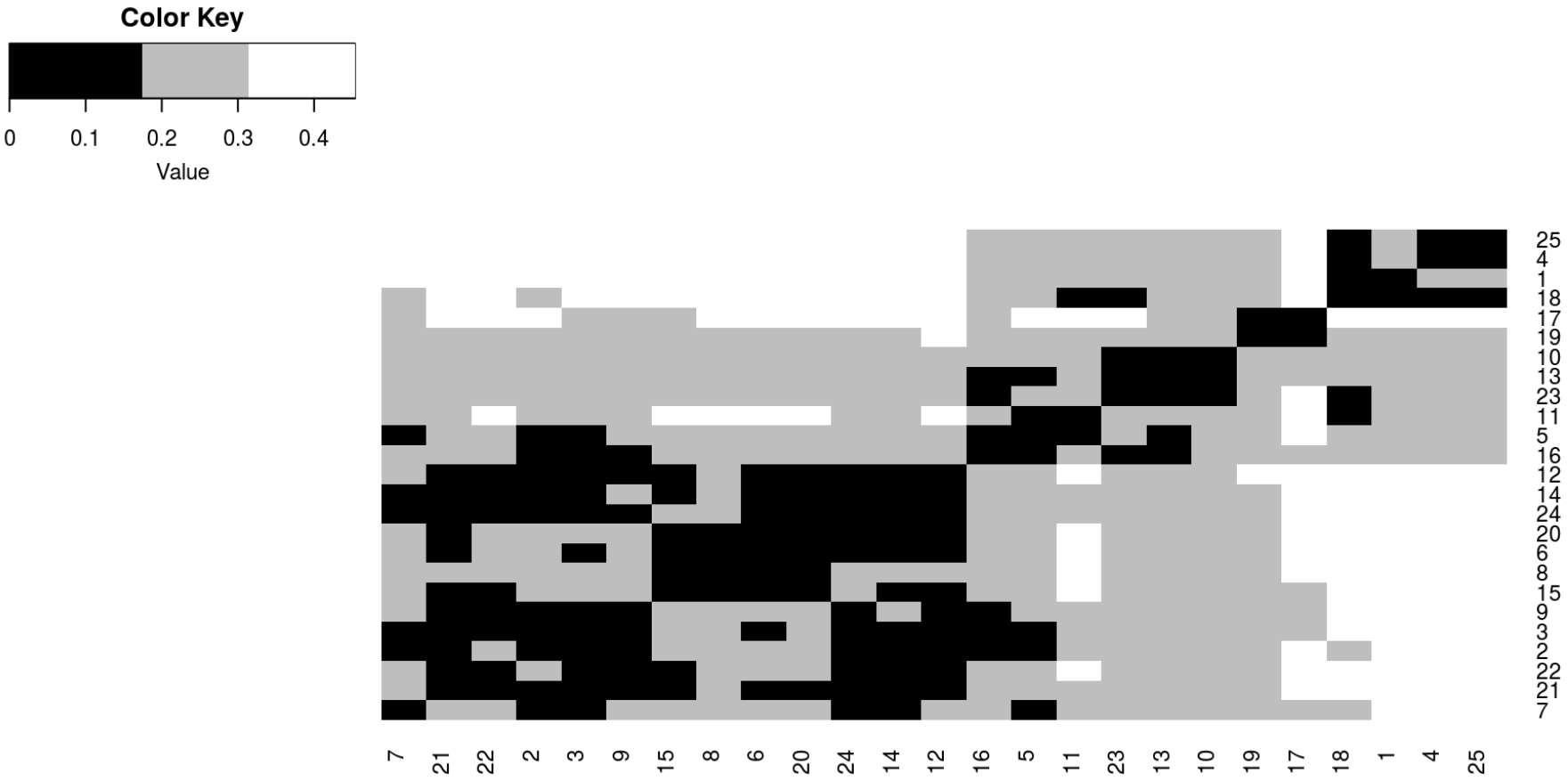
Что вы можете сказать о характере взаиморасположения объектов, глядя на это распределение?

- Частотное распределение коэффициентов указывает на наличие двух совокупностей расстояний: большие (межгрупповые) и маленькие (внутригрупповые).

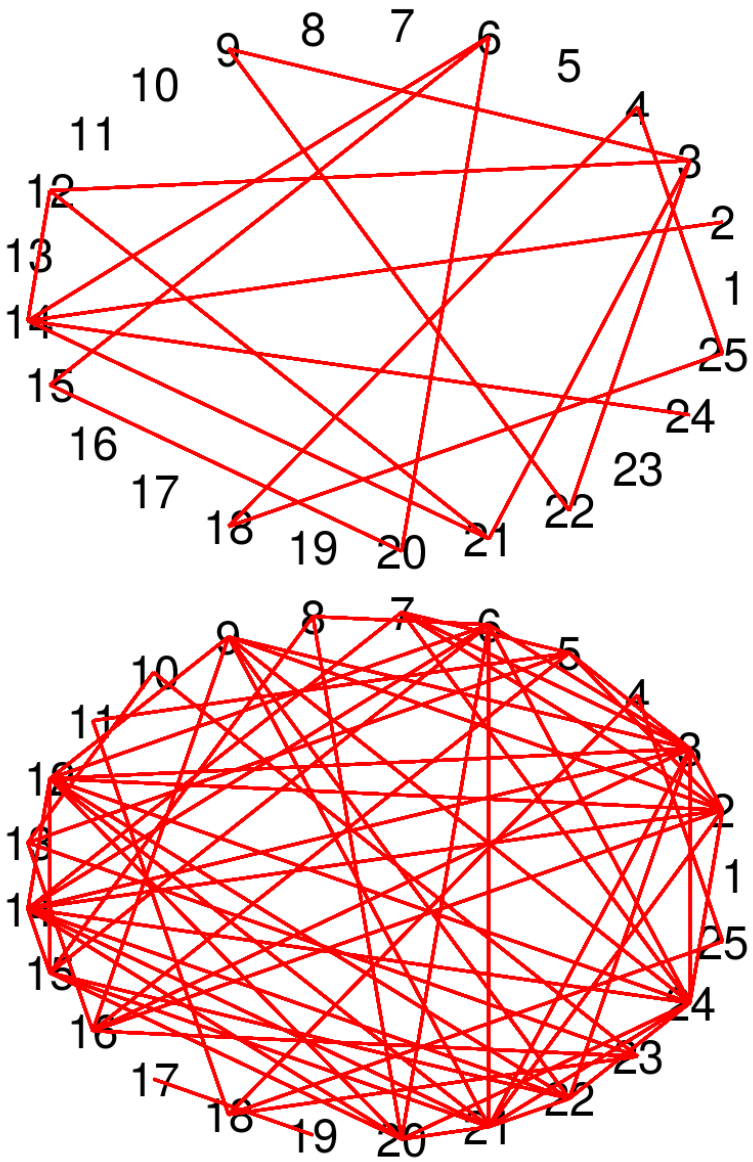
Диагонализация матрицы расстояний (Метод Кульчинского, Kulczynski, 1928)



Диагонализация матрицы расстояний (Метод Кульчинского, Kulczynski, 1928)

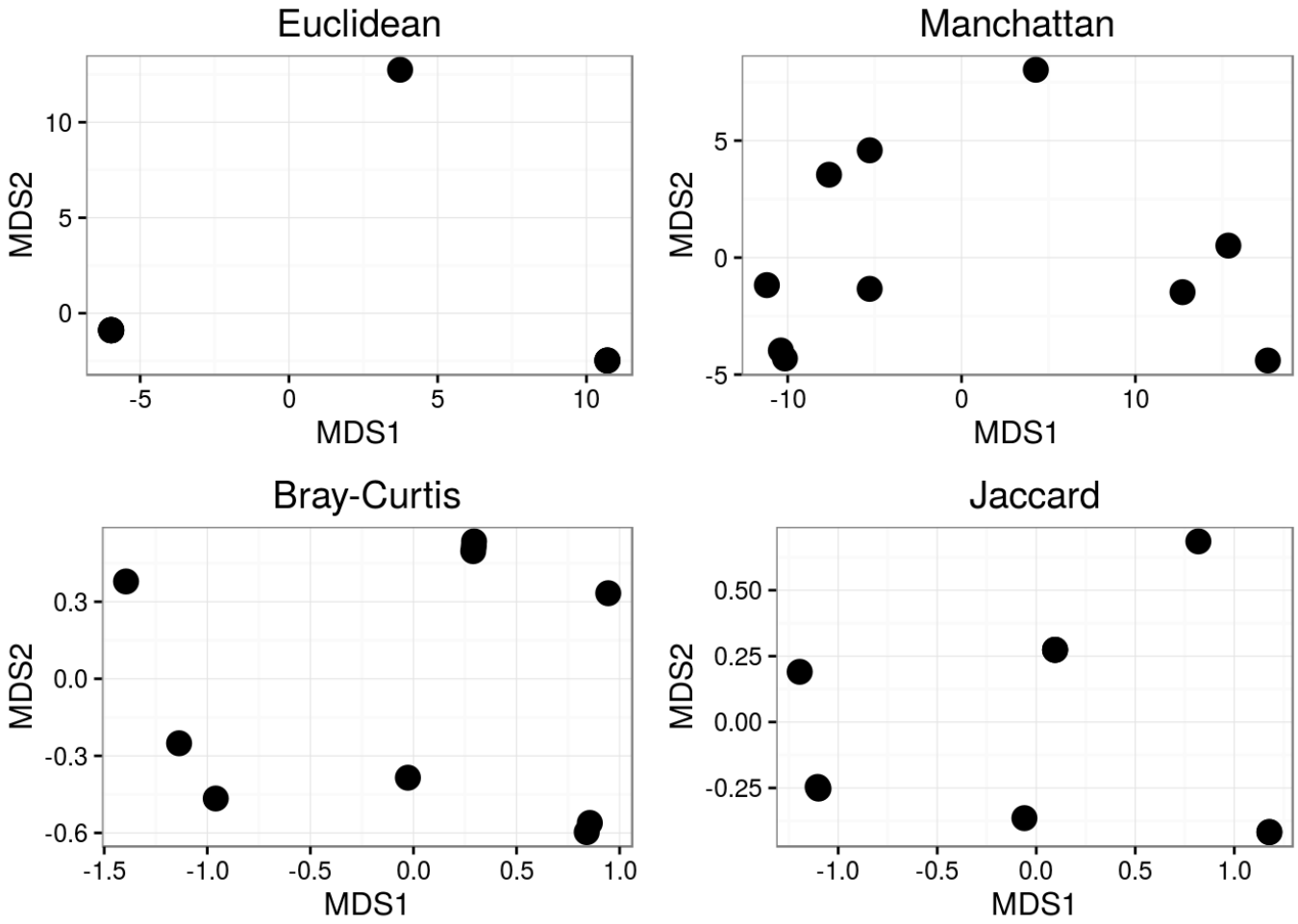


Анализ плеяд

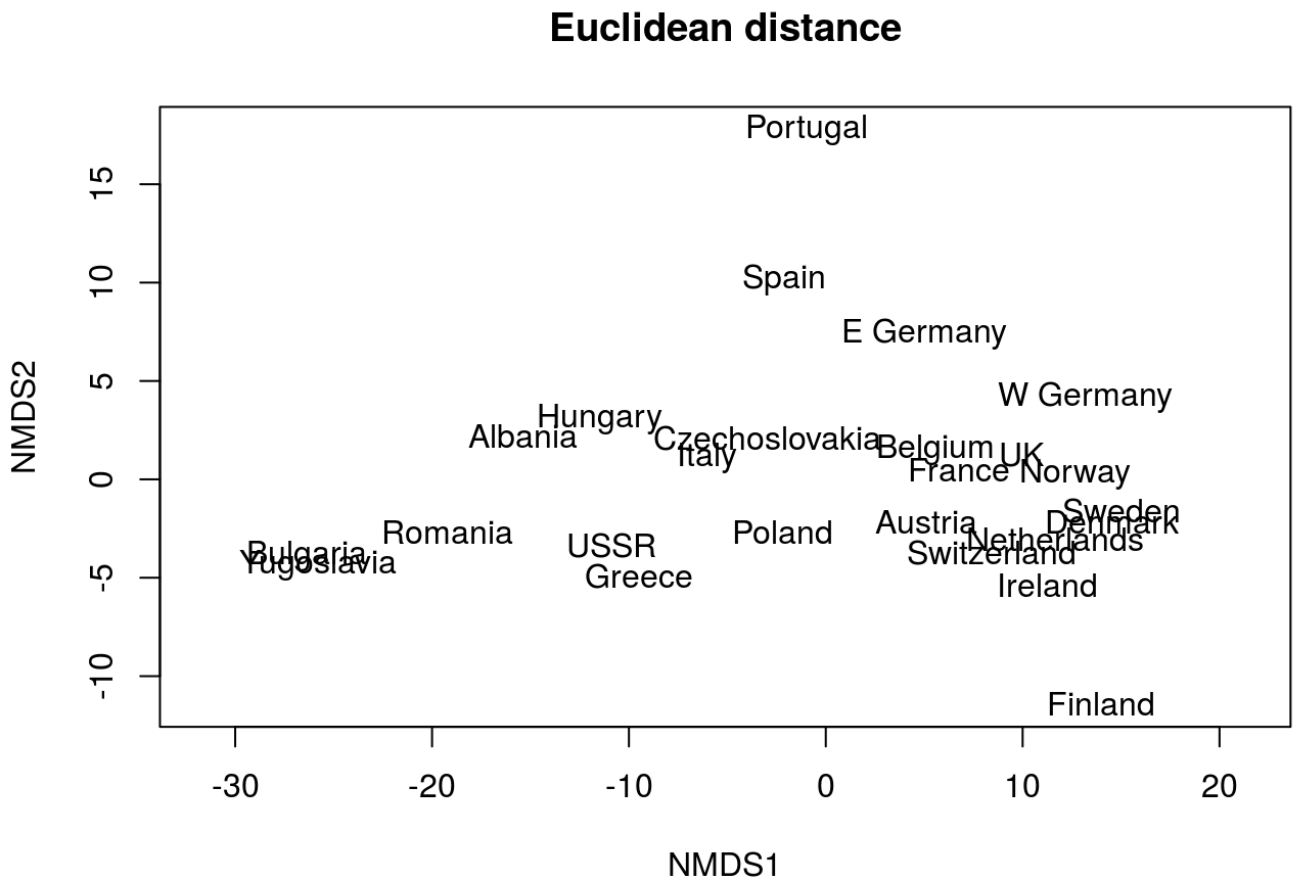


Часть 4. Ординация в пространстве со сниженной размерностью (первое знакомство)

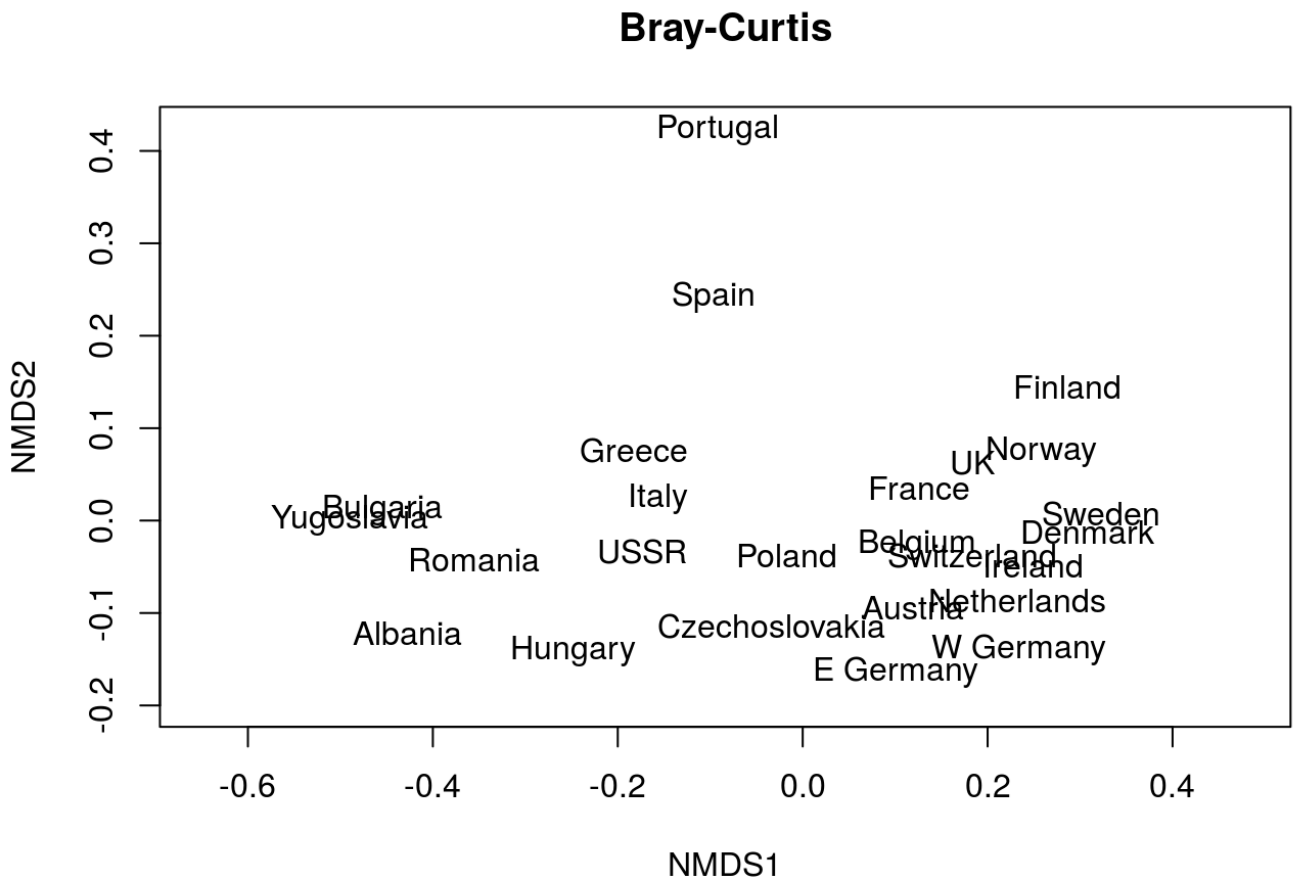
Разные коэффициенты могут давать разные результаты



Евклидово расстояние



Коэффициент Брея-Куртиса



Summary

- Существует два подхода к анализу многоменых данных:
  - 1. -подход, основанный на линейной алгебре,
  - 1. -подход, основанный на исследовании матриц сходства/различия между объектами.
- Выбор коэффициентов сходств/различия - непростая задача, решение которой зависит от структуры материала и поставленных задач. Разные коэффициенты потенциально могут давать разные результаты.
- Получить важную информацию о взаиморасположении объектов можно с помощью некоторых простейших методов.

Что почитать

- Legendre P., Legendre L. (1998) Numerical ecology. Second english edition. Elsevier, Amsterdam. (Фундаментальный труд, описывающий большинство методов. Дается исчерпывающее обсуждение разнообразных коэффициентов сходства/различия)
- Zuur, A. F., Ieno, E. N., Smith, G. M. Analysing Ecological Data. Springer 2007 (Практичеки все то же самое, что в L&L, но написанное простым языком)
- Clarke, K. R., Gorley R. N. (2006) PRIMER v6: User Manual/Tutorial. PRIMER-E, Plymouth. (Очень доходчиво написанное руководство, дающее общее представление о "механике" работы многомерных методов)
- Миркин Б.М., Розенберг Г.С. Фитоценология. Принципы и методы. М., 1978. (Руководство написано в докомпьютерную эпоху, но простейшие методы изложены очень хорошо)
- Василевич В.И. Статистические методы в геоботанике. - Л.: Наука, 1969.
- Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977