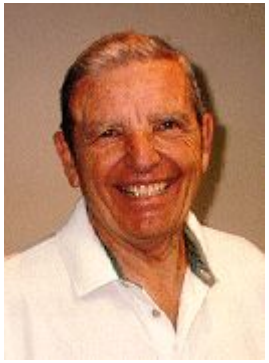
Метод Варда

Джо Вард



Доктор Д. Вард работал в таких направлениях, как

- Педагогическая психология
- Статистика
- И другие.

Он был консультантом ВВС, армии и флота США по применению статистических методов для подбора и оценки персонала, поступающего на службу.

Последние годы жизни он посветил волонтерской работе в начальной школе, которая названа в честь него - *Dr. Joe Ward Elementary School*

Метод Варда

- Метод Варда – это альтернативный подход для проведения кластерного анализа. В основном, вместо использования метрик и мер связей данный метод большее рассматривает проблему с точки зрения дисперсионного анализа. Он подходит скорее для анализа количественных переменных, а не для бинарных переменных.

Метод Варда

- Метод Варда – это альтернативный подход для проведения кластерного анализа. В основном, вместо использования метрик и мер связей данный метод большее рассматривает проблему с точки зрения дисперсионного анализа. Метод Вада подходит скорее для анализа количественных переменных, а не для бинарных переменных.

Метод Варда

- Основываясь на том, что кластеры многомерных наблюдений должны иметь примерно эллиптическую форму, считается, что данные из каждого кластера будут реализованы в многомерное распределение. То есть, если построить p-мерную точечную диаграмму, кластеры будут похожи на эллипс.

Метод Варда

Пусть

- X_{ijk} – означает значение k- переменной в j – наблюдении, принадлежащему i – кластеру.
- При этом для реализации данного метода мы должны определить следующее:

Метод Варда

- **Ошибка суммы квадратов:**

$$\mathrm{ESS} = \sum_i \sum_j \sum_k \left| X_{ijk} - \overline{x}_{i \cdot k} \right|^2,$$

Здесь суммируется все переменные во всех подчастях каждого кластера и сравнивается отдельное наблюдение для каждой переменной со средней этой переменной из кластера. Если ESS имеет малые значения, то данные близки к средним по кластеру, подразумевая, что мы уже имеем кластер, как единицу анализа.

Метод Варда

- **Общая сумма квадратов:**

$$\mathrm{TSS} = \sum_i \sum_j \sum_k \left| X_{ijk} - \overline{x}_{..k} \right|^2$$

В данном случае сравнивается отдельные наблюдения в каждой переменной с общей средней по переменной.

Метод Варда

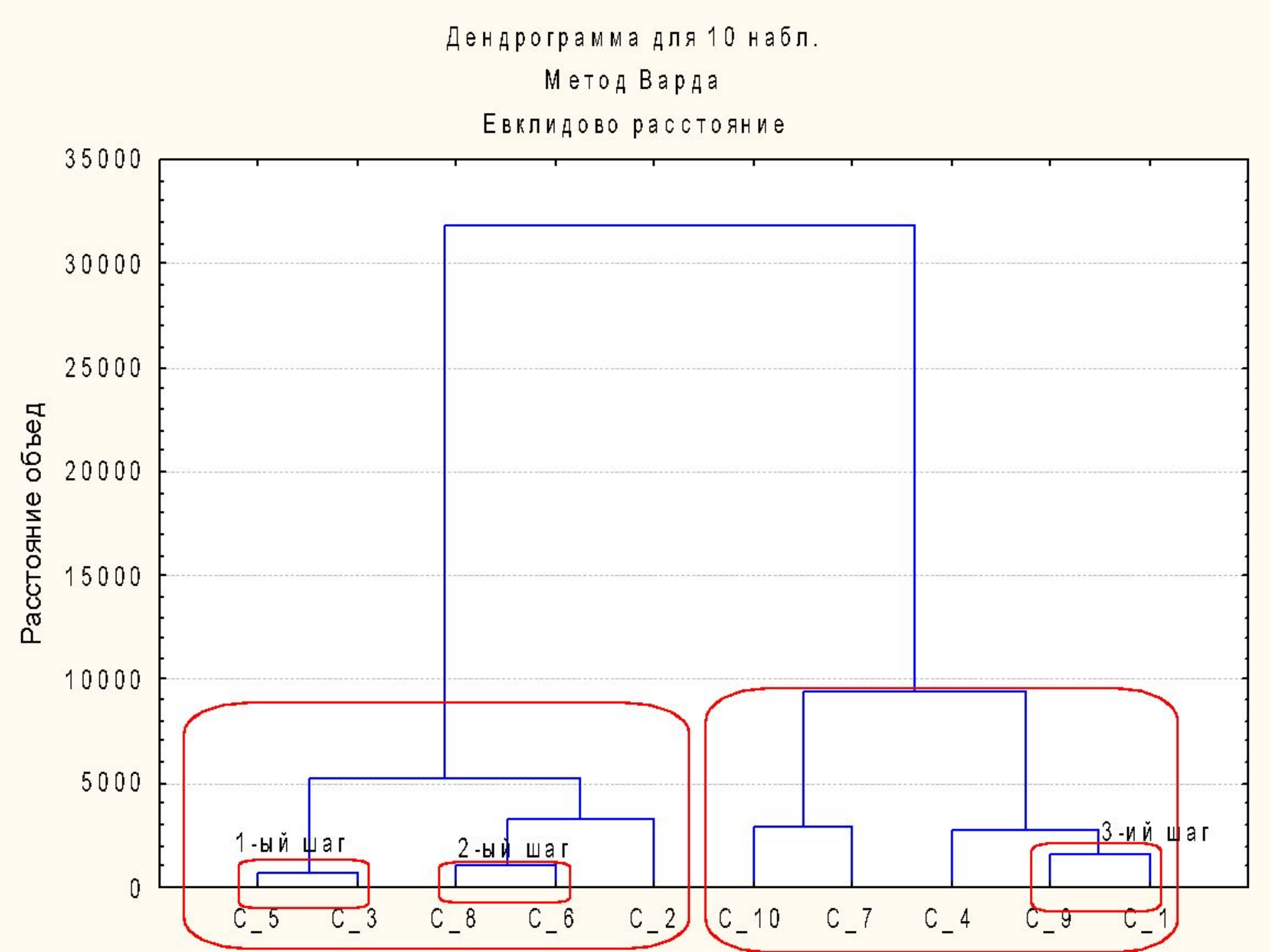
- **R-квадрат:**

$$r^2 = \frac{\text{TSS - ESS}}{\text{TSS}}$$

Значение интерпретируется, как доля вариации, объясняемая специфической кластеризацией наблюдений.

Метод Варда

- Использование метода Варда начинается с образования n кластеров, куда входит по одному наблюдению. На первом шаге формируется n-1 кластер, где в одном из кластеров объединяется два наблюдения. Вычисляется ошибка сумм квадратов и r- квадрат. На следующем этапе образуется n-2 кластера, при этом в двух из кластерах может оказаться по два наблюдения, а во всех остальных по одному, или в одном кластере 3 наблюдения, а во всех остальных по одному. Таким образом на каждом шаге кластеры или наблюдения комбинируются таким образом, чтобы свести к минимуму ошибки суммы квадратов и максимизировать значение r – квадрат. Реализация алгоритма завершается, когда образуется один большой кластер, куда входят все наблюдения.


Дендрограмма для 10 набл.
Метод Варда
Евклидово расстояние
Расстояние объед
35000
30000
25000
20000
15000
10000
5000
0
1-ый шаг
2-ый шаг
3-ий шаг
C_5
C_3
C_8
C_6
C_2
C_10
C_7
C_4
C_9
C_1