Иерархический кластерный анализ

Аббакумов
Вадим Леонардович

Происхождение термина

- Кластер – калька слова «cluster»,
- «сгусток», «гроздь (винограда)», «скопление (звезд)» и т.п.

Ранее использовались другие термины

- распознавание образов без учителя,
- стратификация,
- таксономия,
- автоматическая классификация.

Задача

- Кластерный анализ разбивает набор объектов на группы
- Попутно определяется число групп

Определение

- Группы, на которые разбита выборка, называются кластерами.

Еще раз:

- при иерархическом кластерном анализе заранее неизвестно число кластеров (групп, на которые разбивается набор объектов).

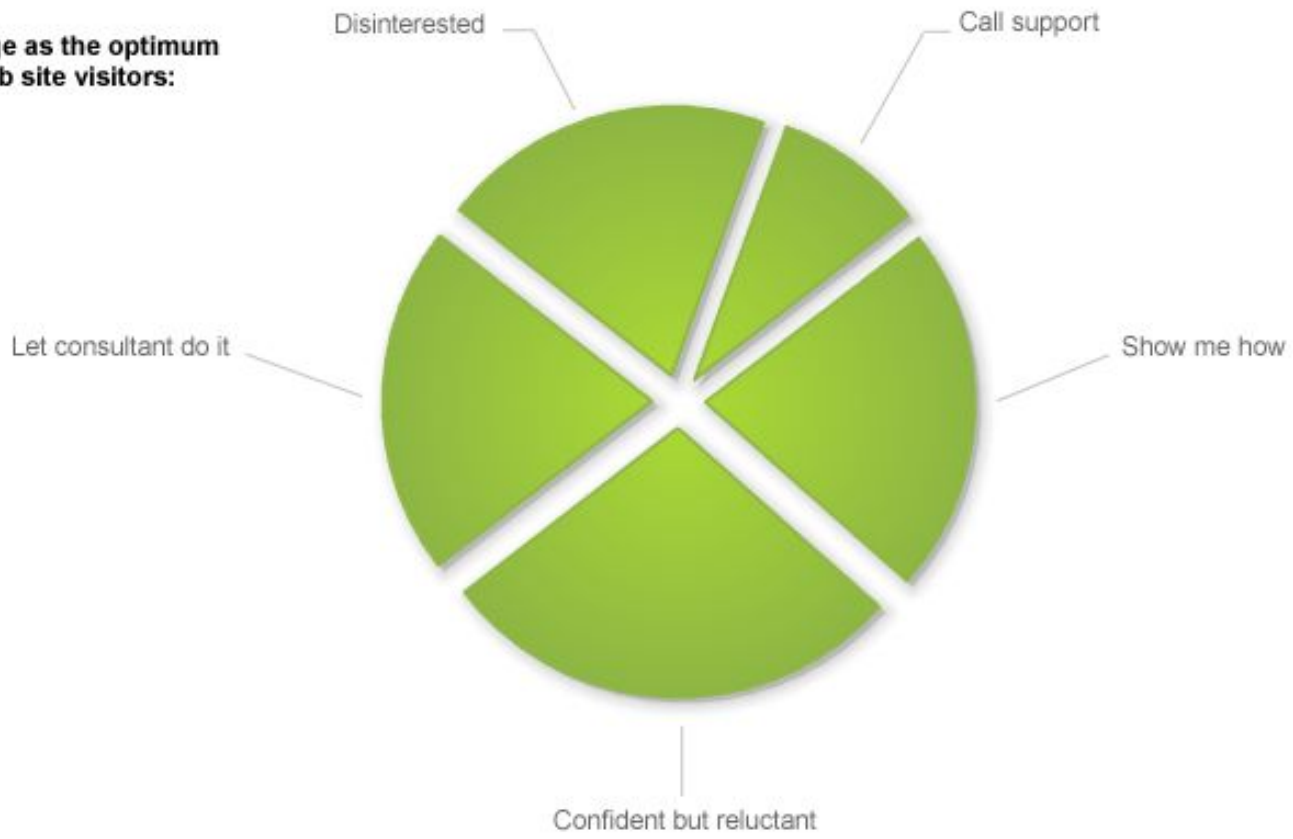
Другие методы кластеризации

- Метод к-средних
- Самоорганизующиеся карты Кохонена (SOM)
- Смесь (нормальных) распределений
- …

В маркетинге: Сегментирование рынка

Persona Segments

5 “Persona” segments emerge as the optimum solution representative of web site visitors:



Пример:
Определение групп потребителей

- – По данным о покупателях (результаты опроса, поведение на сайте) выявить и описать/понять рыночные сегменты.
- – Прежде, чем фирма определится, какие сегменты рынка создают для нее наибольшие возможности, надо решить, какие сегменты уже существуют.

- Страховая компания интересуется группами, на которые разделяются потенциальные клиенты.
- Результаты классификации используются, чтобы для разных групп определять оптимальные цены на услуги, оптимальные тарифы

Пример:
Определение групп потребителей

– Для разбиения потребителей на группы можно выбирать разные наборы характеристики объектов, например возраст, образование, место жительства, тип личности, и так далее.

Несложно разделить покупателей на сегменты по одной (или по каждой) характеристике.

Кластерный анализ может помочь выявить уже сложившееся разбиение потребителей на «*группы со схожими потребностями в отношении конкретного товара или услуги, достаточными ресурсами, а также готовностью и возможностью покупать*» учитывая **все** выбранные показатели одновременно.

Пример: товарные группы для рекомендательной системы

На рынке присутствует большой выбор товаров схожего назначения под разными торговыми марками. Надо разбить товары на группы.

Иногда такое разбиение известно и получается без применения статистической техники. Например, компьютеры бывают «для дома», «для офиса», «серверы» и «специализированные».

Кластерный анализ применяется, если нет классификации, признанной всеми.

Важно! Результат будет зависеть от выбора набора показателей.

Пример

- Определение целевой аудитории баннерной рекламной компании в интернете.
- 100000 сайтов
- Каждый из них указывает на интересы куки, на текущее настроение куки…
- Надо отождествить схожие сайты

Другие задачи классификации

- Machine Learning
- Классификация с учителем
- Распознавание образов

Отличие

- Заранее известно, к какому классу принадлежит каждое из наблюдений.
- Технологически - среди переменных присутствует так называемая группирующая переменная.

Что тогда классифицировать?

- Надо придумать правило.
- Для классификации новых наблюдений.

Другие задачи классификации

- Классификация с обучающей выборкой
  - наивный байесовский классификатор
  - дискриминантный анализ
  - деревья классификации
  - К-го ближайшего соседа
  - Нейронная сеть прямого распространения
  - SVM
  - Случайный лес
  - Gradient boosting machine

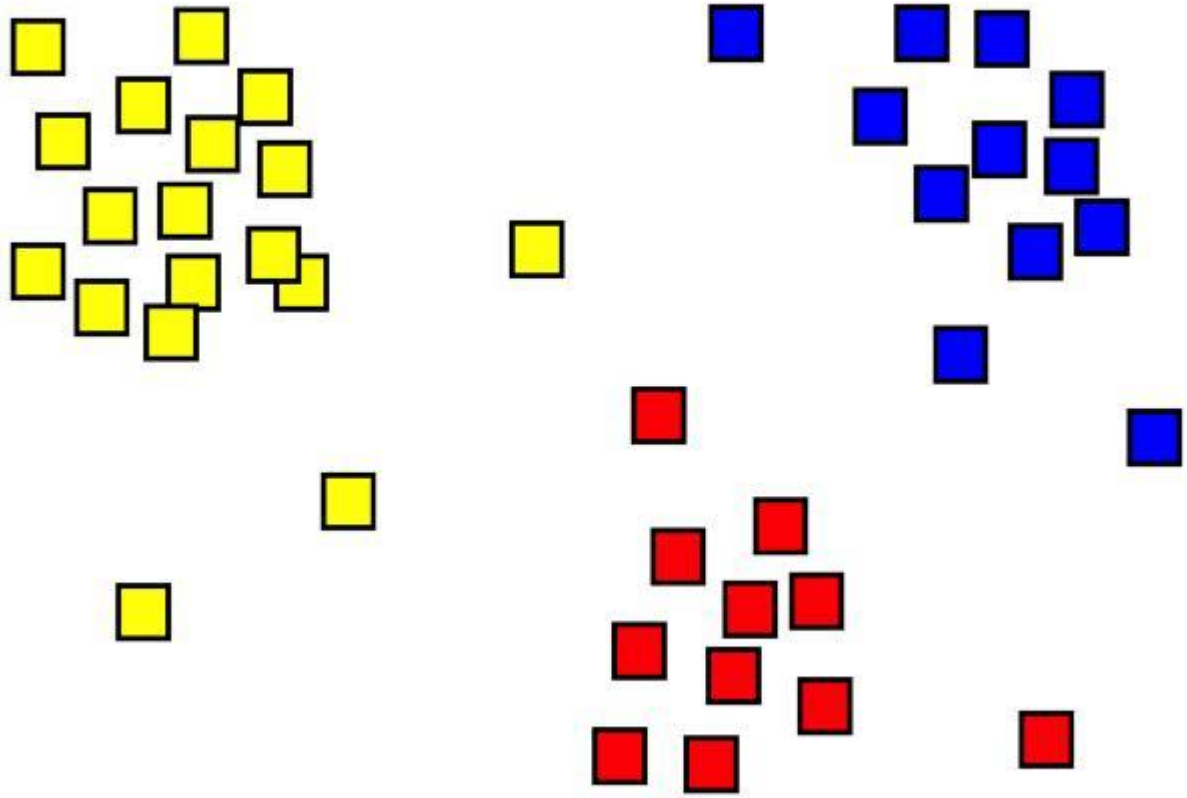
- Вернемся к кластерному анализу

Идея метода

- Сведем задачу к геометрической

Сведем задачу к геометрической

- Каждый объект – точка.
- Похожие объекты расположены «близко» друг к другу
- Различающиеся объекты расположены «далеко»
- Скопления точек – кластер.



Расстояние между объектами

- Евклидово расстояние
- Квадрат Евклидова расстояния
- Блок (Манхеттен, сити-блок)
- и так далее…

Расстояние Евклида

- *Две точки*

  (x1, x2, x3)

  (y1, y2, y3)

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Квадрат евклидова расстояния не является расстоянием...

Расстояние Block

(Manhatten, таксиста).

Расстояние Block

(Manhattan, таксиста, Минковского при p=1).

$$X = (x_1, x_2, \ldots, x_k)$$

$$Y = (y_1, y_2, \ldots, y_k)$$

$$d_{XY} = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_k - y_k|$$

Расстояние Хэмминга

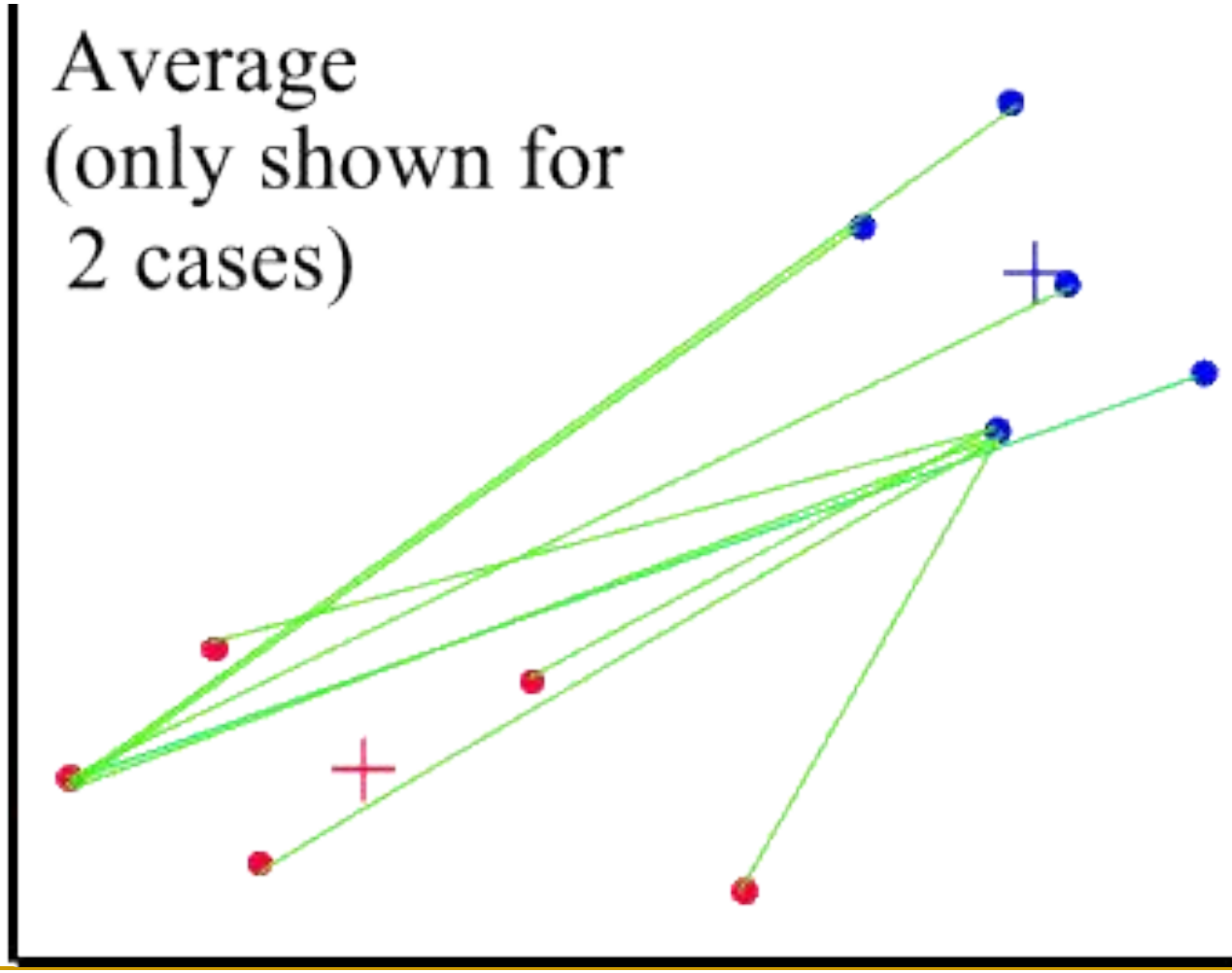
- число позиций, в которых соответствующие символы двух слов одинаковой длины различны
-
- D(1011101, 1001001) =
- D(2173896, 2233796) =
- D(toned, roses)

- Вопрос:
- Когда выбирать евклидово расстояние, а когда растояние Манхэттен?

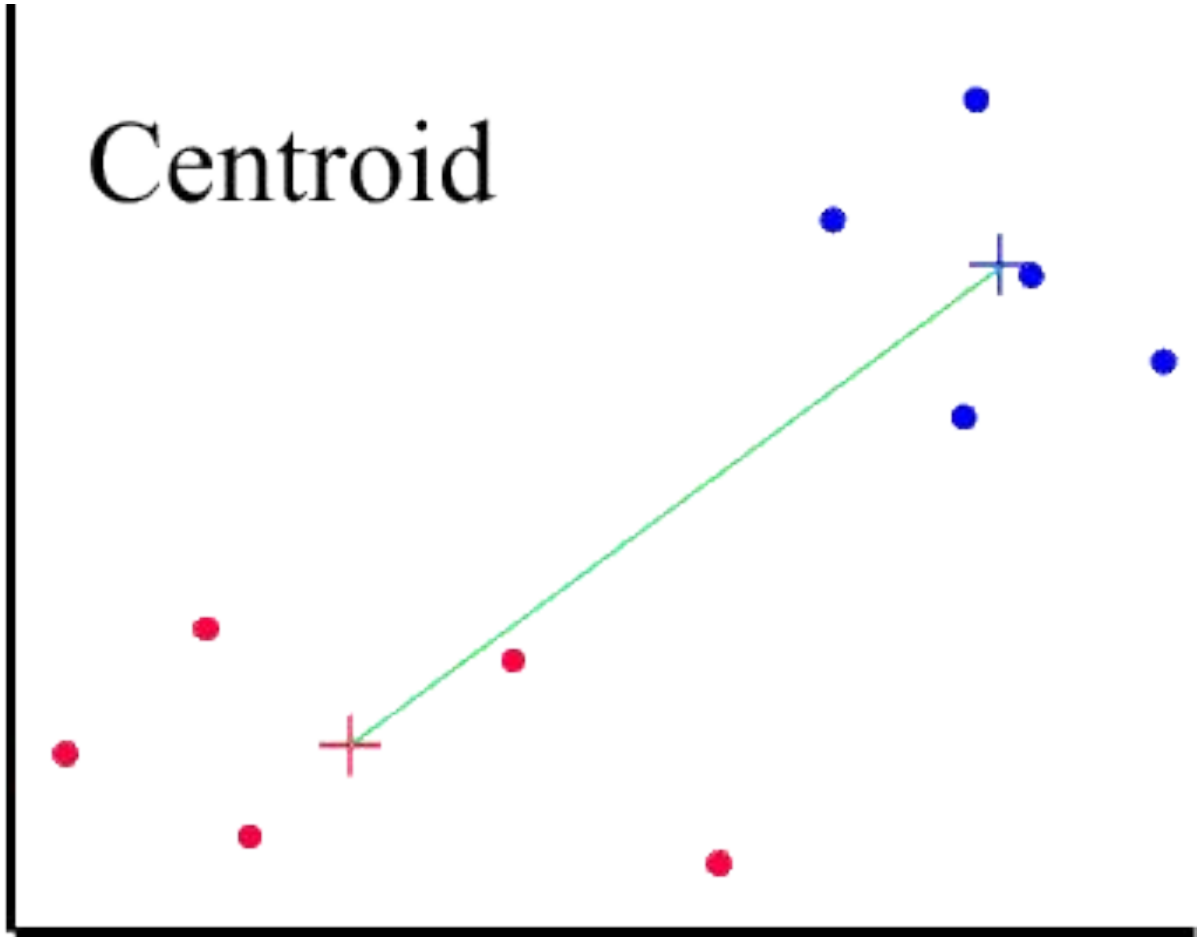
Расстояние между кластерами

- Среднее невзвешенное расстояние (Average linkage clustering).
- Центроидный метод (Centroid Method).
- Метод дальнего соседа, максимального расстояния (Complete linkage clustering).
- Метод ближайшего соседа (Single linkage clustering).
- Метод Варда (Ward's method).

Среднее невзвешенное расстояние



Центроидный метод



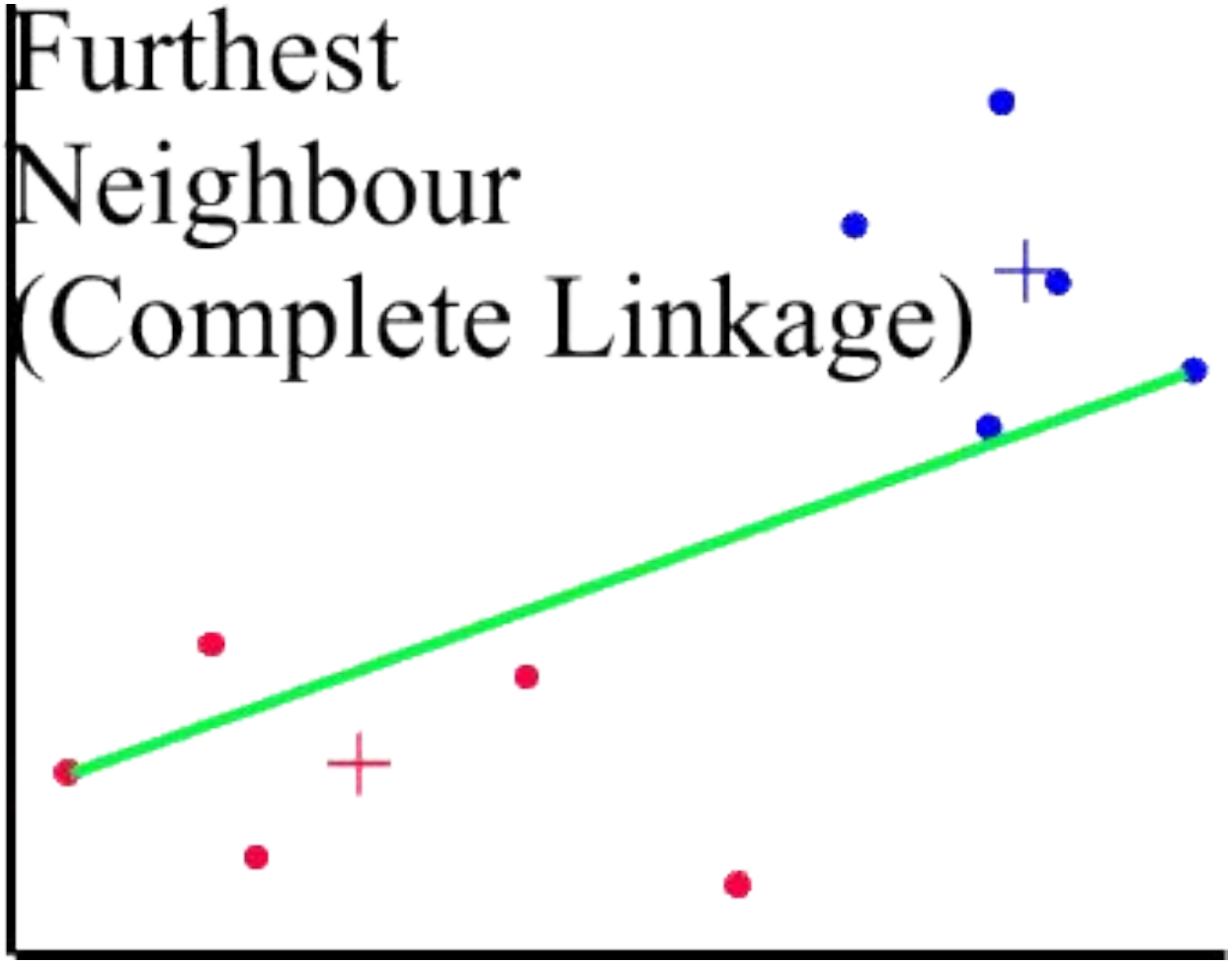
Центроидный метод

- Вычислительная простота
- Объем кластера не влияет.

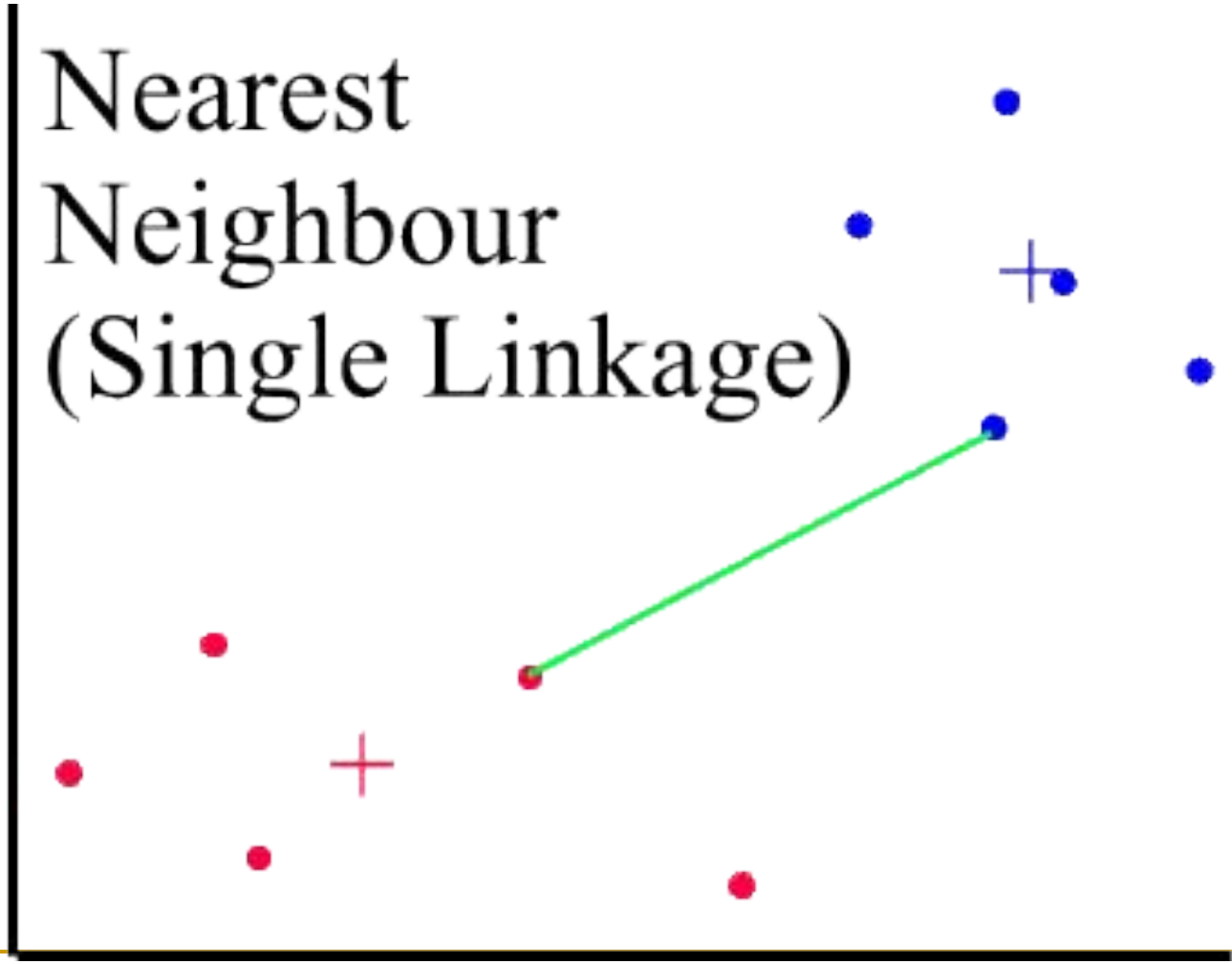
- Дендрограмма может иметь самопересечения

- Выходит из употребления

Метод дальнего соседа



Метод ближнего соседа



Растояние Sørensen–Dice

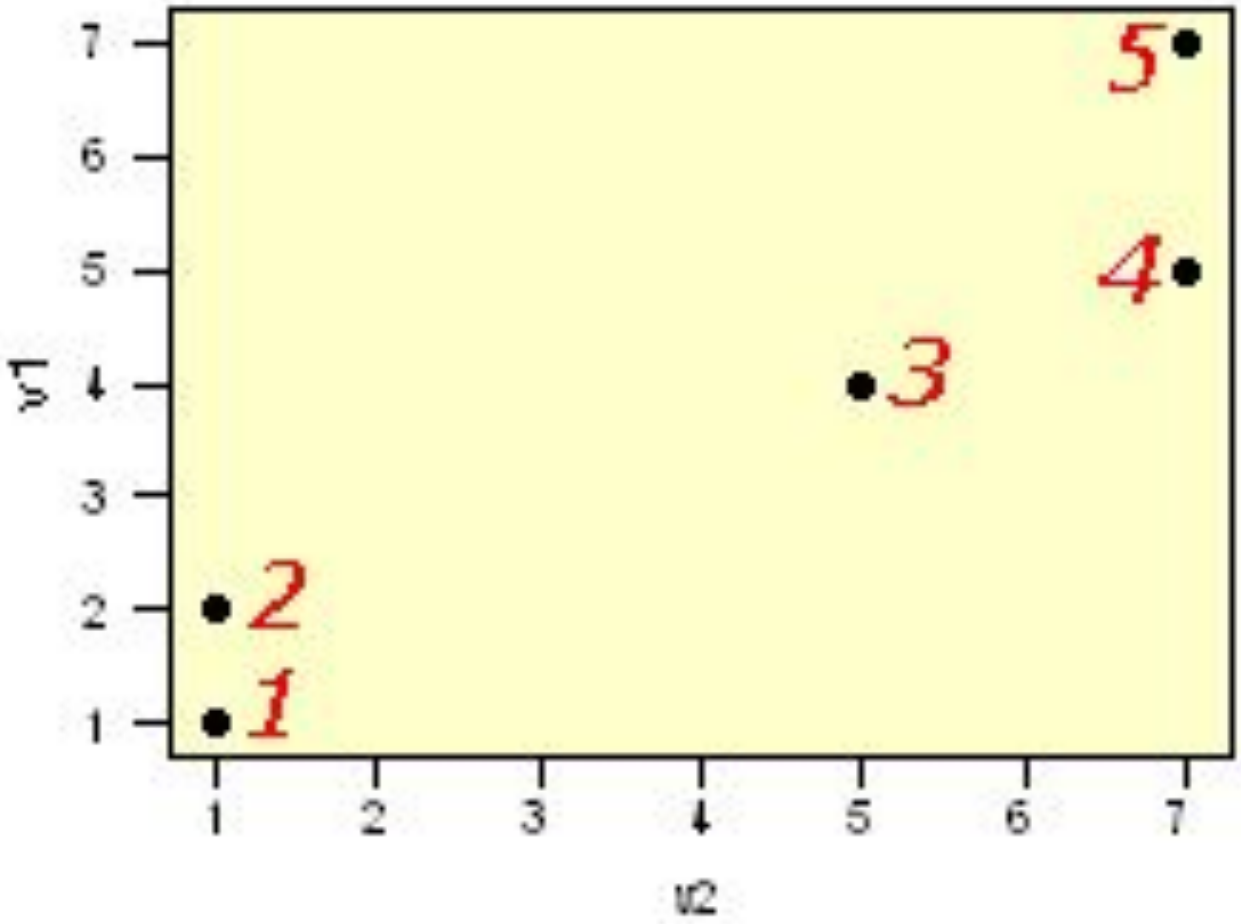
- Метод Варда (WARD).
- Предполагается использование квадрата евклидова расстояния

Начинающим рекомендуем

-
- – метод Варда;
- – метод ближнего соседа (Complete linkage clustering);
- – среднее невзвешенное расстояние (Average linkage clustering).

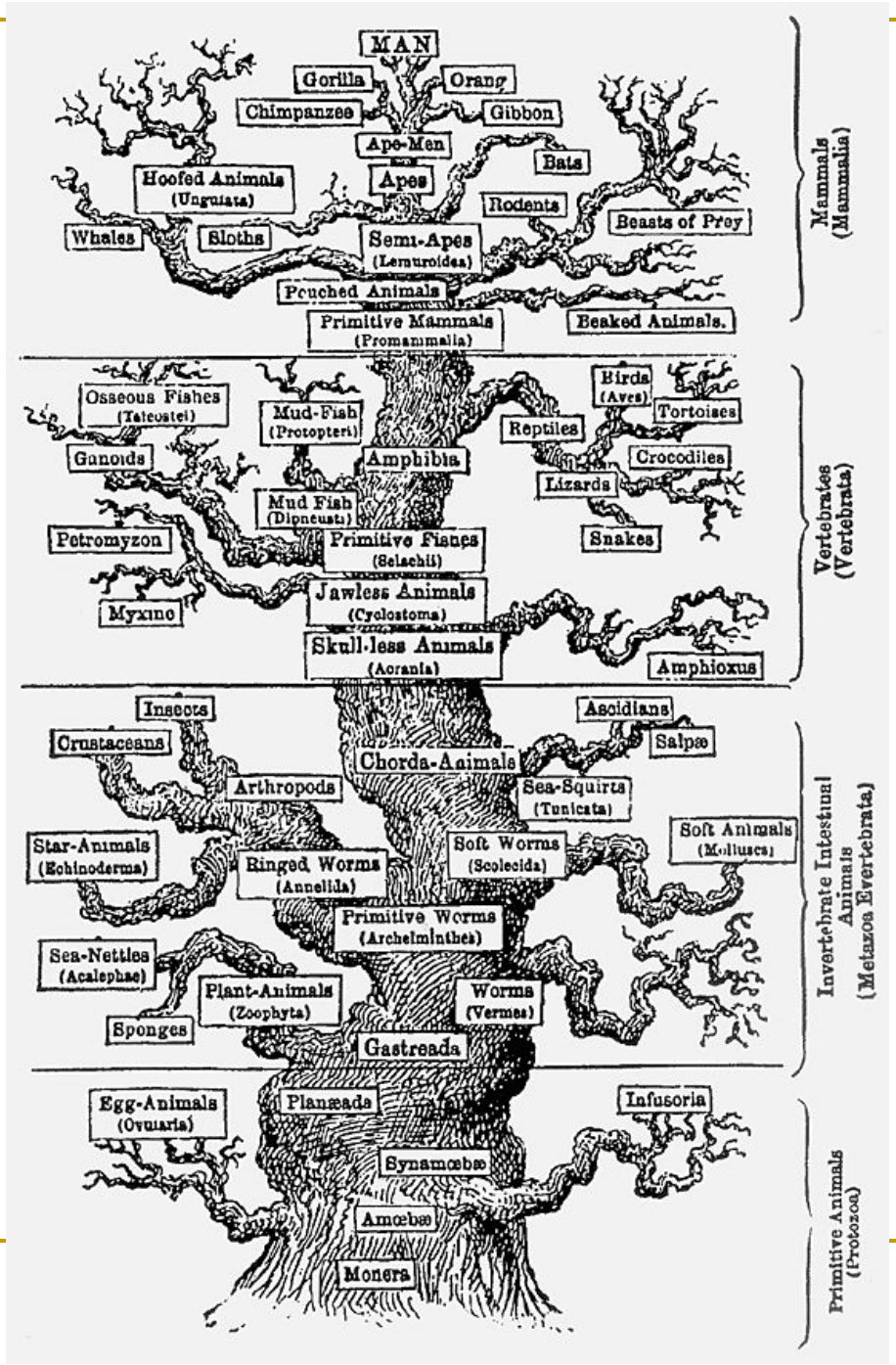
Алгоритм кластерного анализа

- Разберемся с процедурой иерархического кластерного анализа на примере

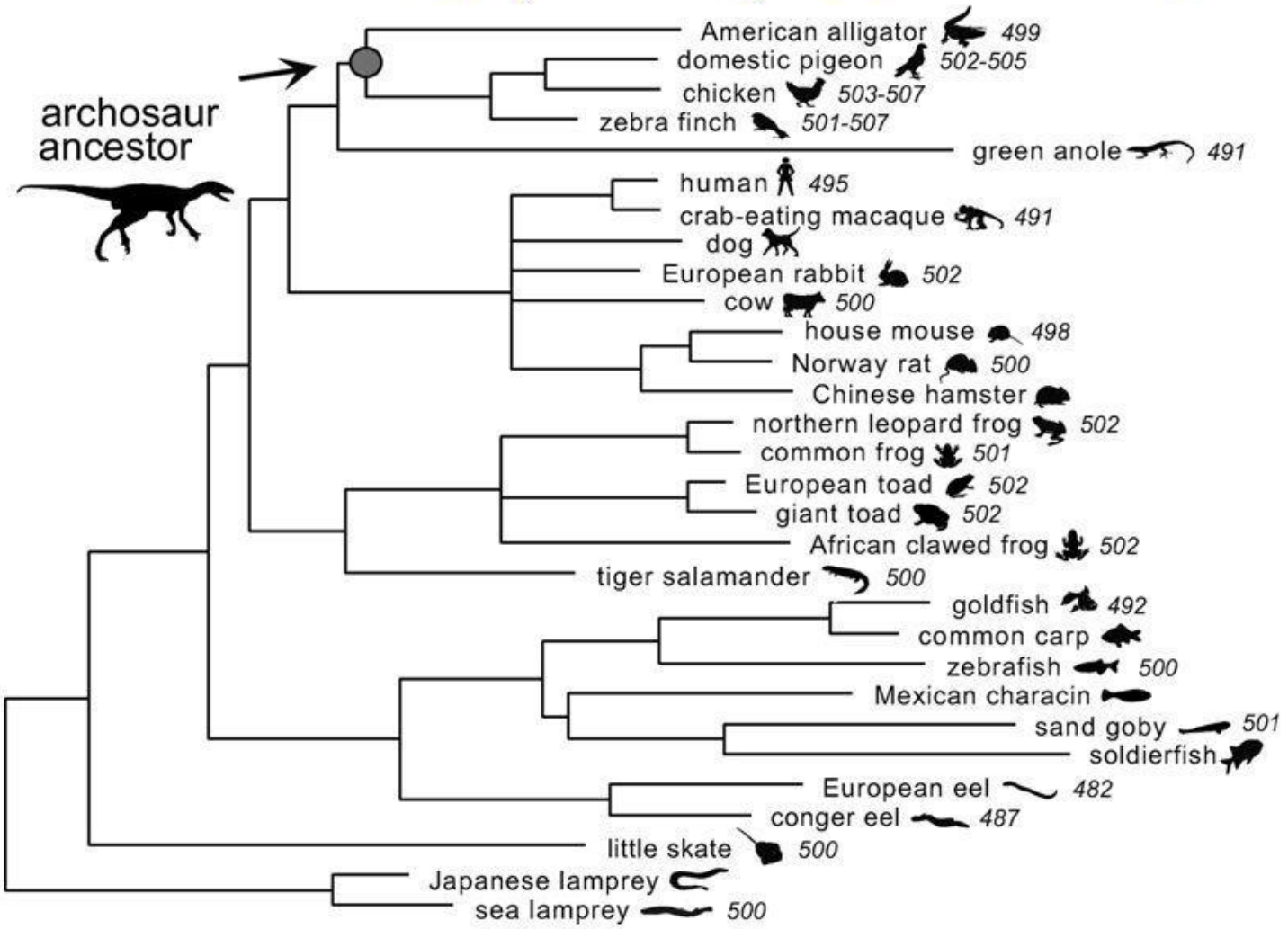

7
6
5
4
3
2
1
v1
5
4
3
2
1
1
2
3
4
5
6
7
v2

- Алгоритм построения дендрограммы

- Ernst Haeckel
- Tree of Life
- The Evolution of Man (1879)
-
- Но он не был первым…
- Древо Порфирия (300+ год)


MAN
Gorilla
Orang
Chimpanzee
Gibbon
Ape-Men
Bats
Hoofed Animals
(Ungulata)
Apes
Rodents
Whales
Sloths
Semi-Apes
(Lemuroidea)
Beasts of Prey
Pouched Animals
Primitive Mammals
(Promammalia)
Beaked Animals.
Mammals
(Mammalia)
Osseous Fishes
(Teleostei)
Mud-Fish
(Protopteri)
Birds
(Aves)
Tortoises
Reptiles
Ganoids
Amphibia
Crocodiles
Lizards
Mud Fish
(Dipneusti)
Petromyzon
Primitive Fishes
(Selachii)
Snakes
Jawless Animals
(Cyclostoma)
Myxine
Skull-less Animals
(Acrania)
Amphioxus
Vertebrates
(Vertebrata)
Insects
Ascidians
Crustaceans
Salpæ
Chorda-Animals
Arthropods
Sea-Squirts
(Tunicata)
Soft Animals
(Molluscs)
Star-Animals
(Echinoderms)
Soft Worms
(Scolecida)
Ringed Worms
(Annelida)
Primitive Worms
(Archelminthes)
Sea-Nettles
(Acalephæ)
Plant-Animals
(Zoophyta)
Worms
(Vermes)
Sponges
Gastreada
Invertebrate Intestinal
Animals
(Metazoa Evertebrata)
Egg-Animals
(Ovularia)
Planæada
Infusoria
Synamœbæ
Amœbæ
Monera
Primitive Animals
(Protozoa)

Типичное дерево (родопсины)



Recreating a Functional Ancestral Archosaur Visual Pigment

Belinda S. W. Chang, Karolina Jönsson,* Manija A. Kazmi,* Michael J. Donoghue,† and Thomas P. Sakmar**

Mol. Biol. Evol. 19(9):1483–1489. 2002

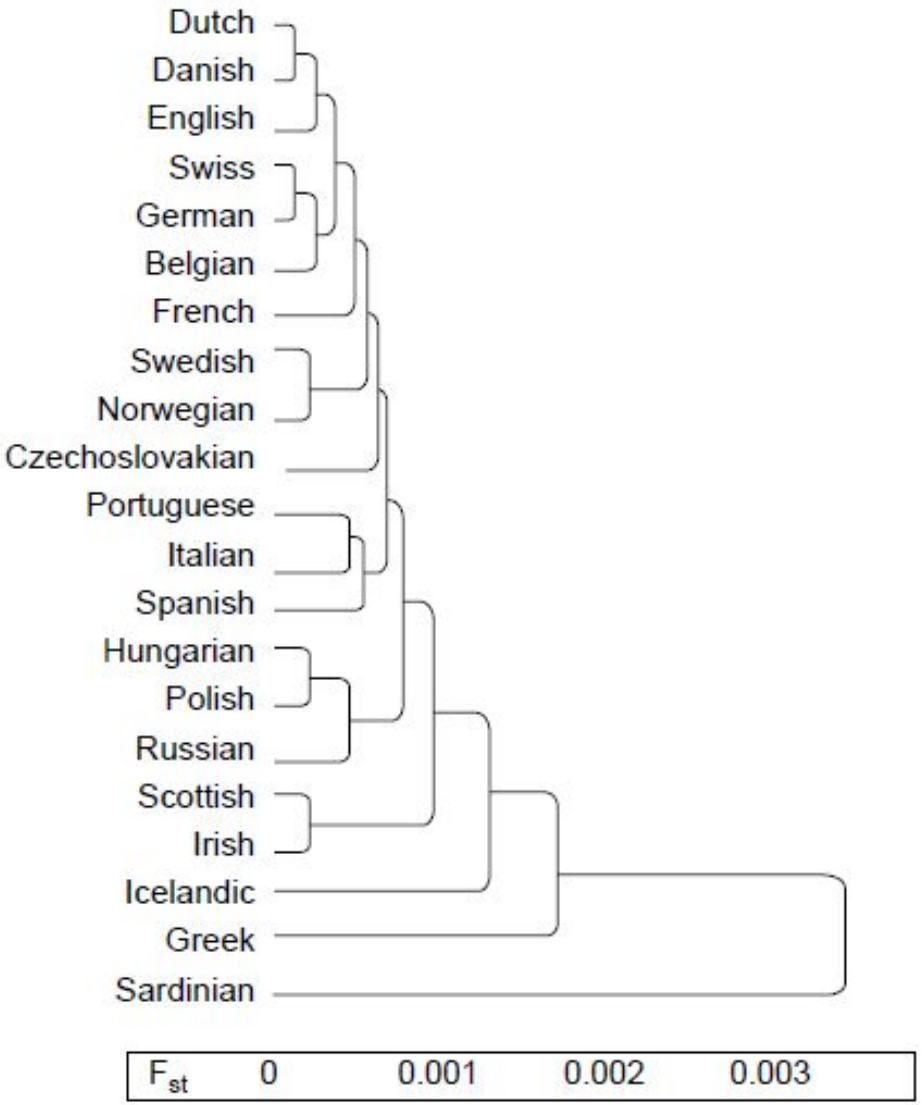
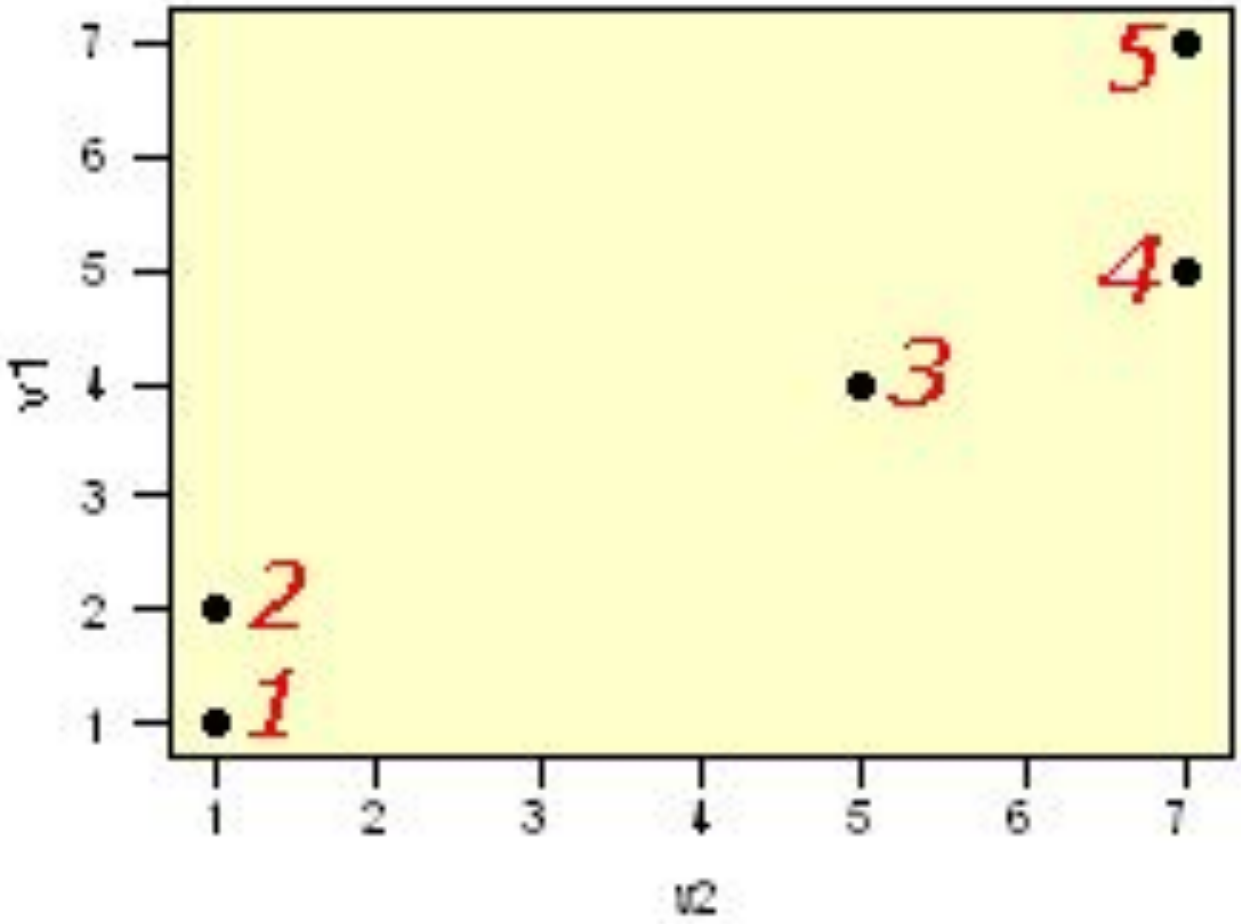
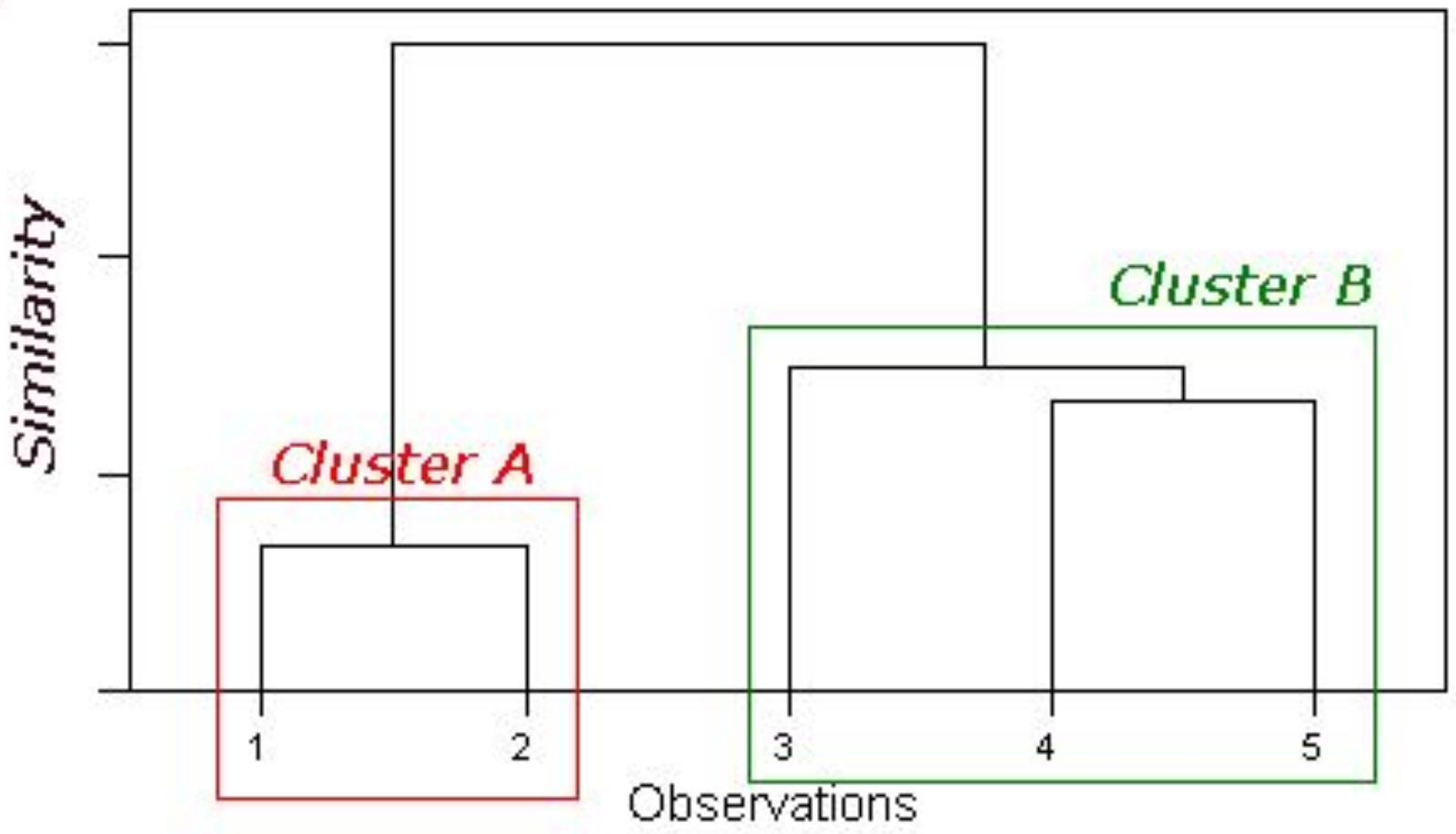
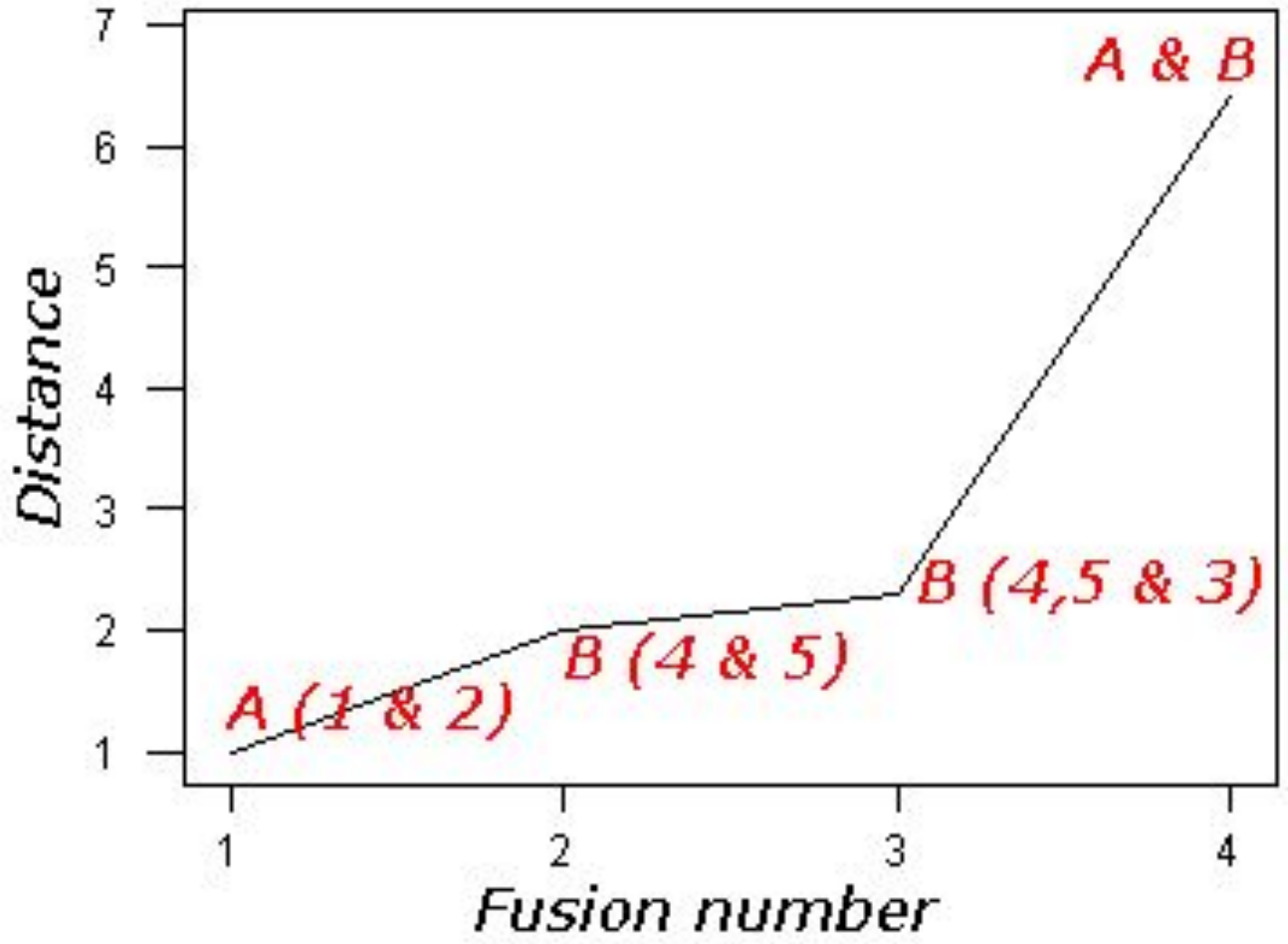


Fig. 5.2 Genetic tree of European populations from genetic distances (= F_{st}) between populations, based on 88 genetic polymorphisms from data in Cavalli-Sforza et al. (1994)

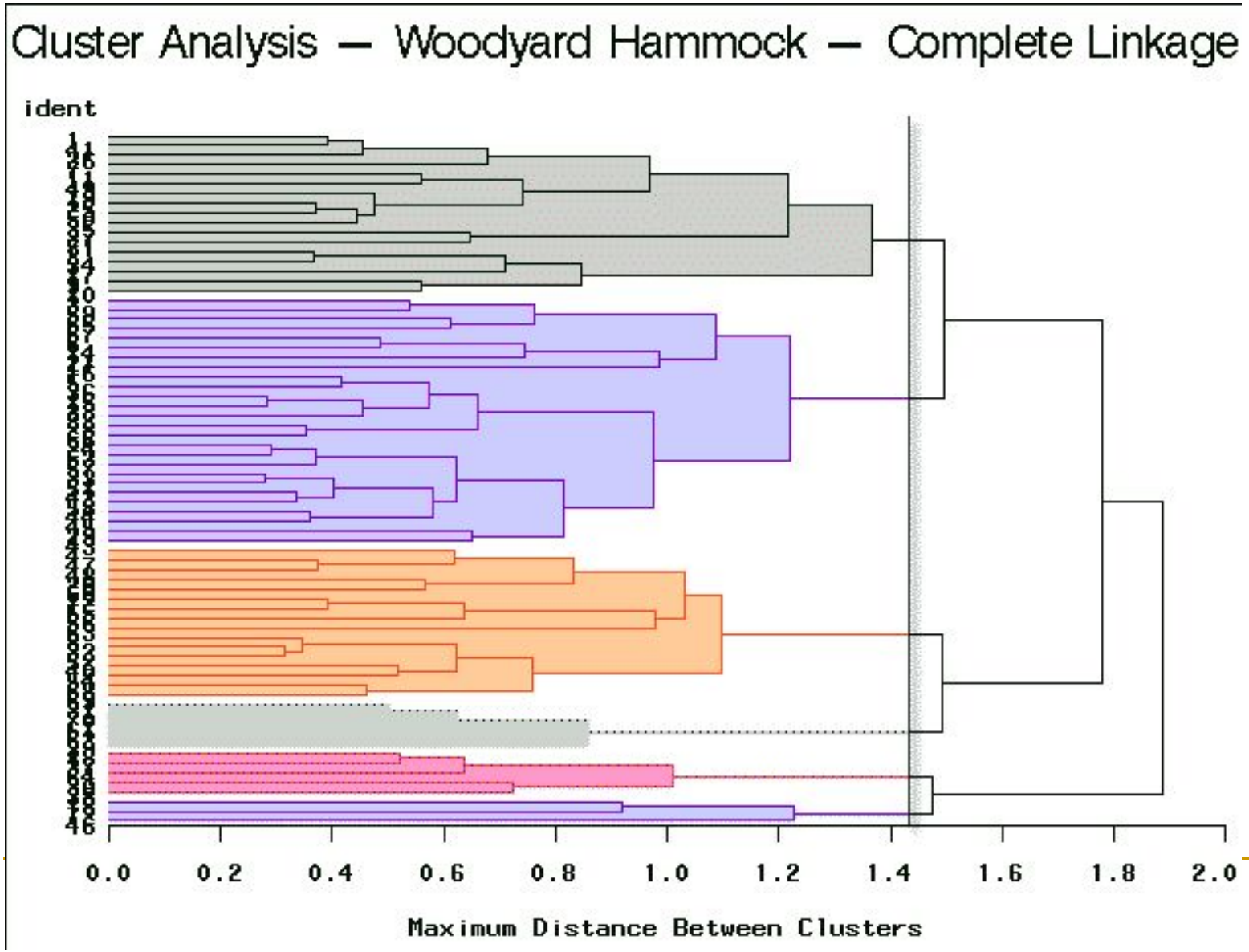

7
6
5
4
3
2
1
v1
5
4
3
2
1
1
2
3
4
5
6
7
v2


Similarity
Cluster A
Cluster B
1
2
3
4
5
Observations

каменистая осыпь / локоть

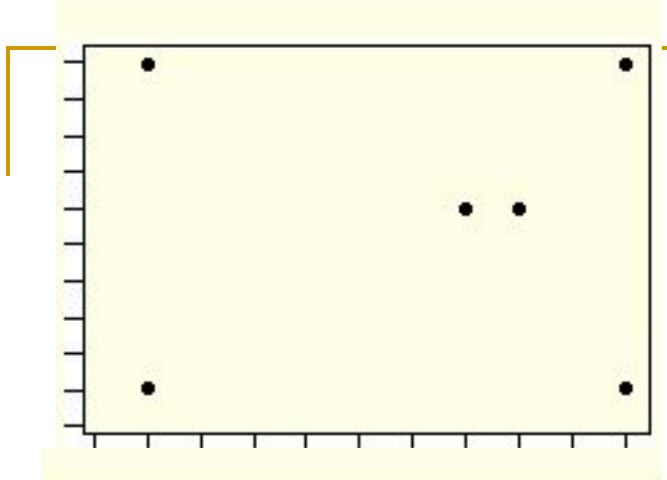


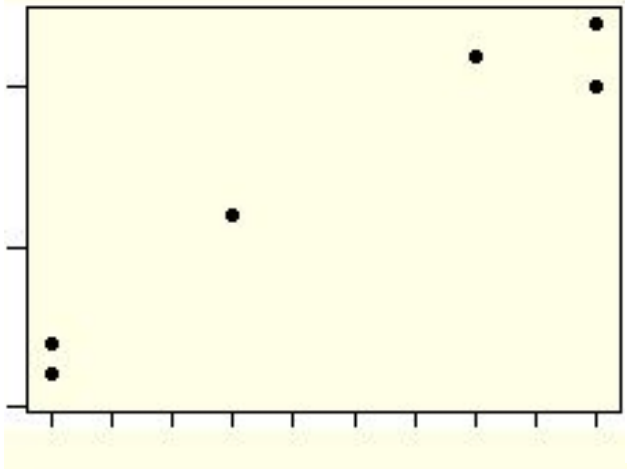
Где на дендрограмме кластеры?

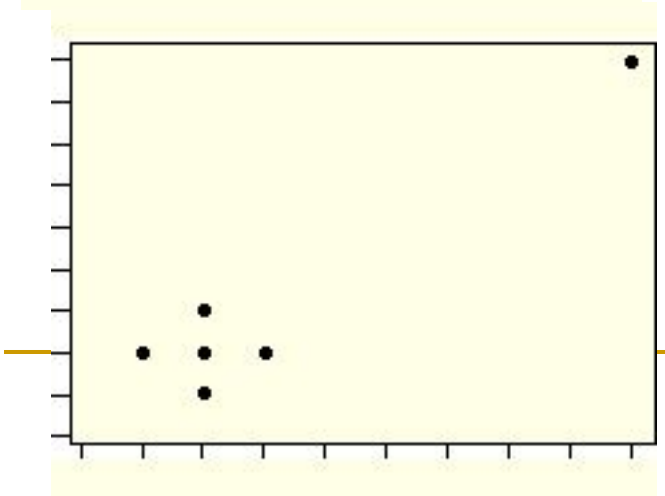


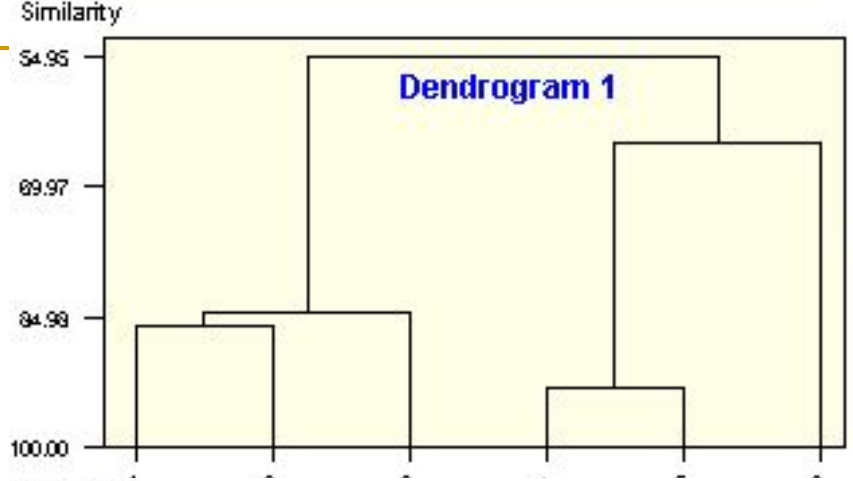
Упражнение

- Разбить на пары:
- Каждой диаграмме рассеивания поставить в соответствие дендрограмму

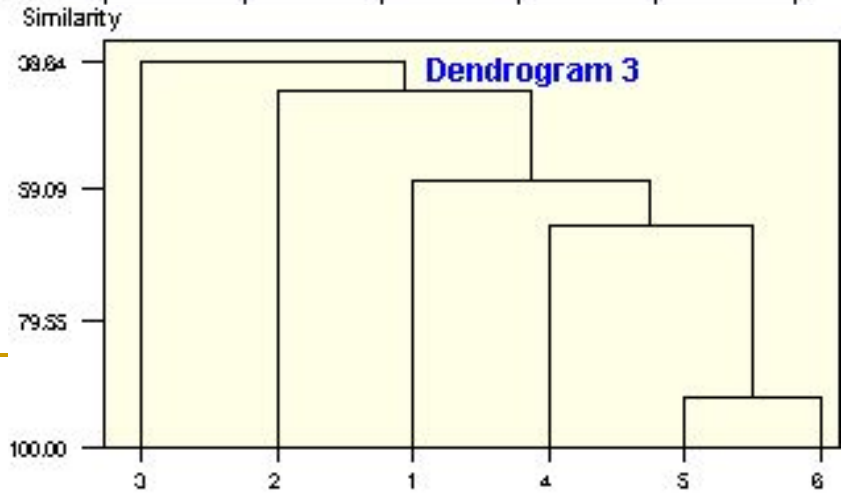







Similarity
54.95
69.97
84.98
100.00
Dendrogram 1


Similarity
13.27
42.18
71.09
100.00
Dendrogram 2


Similarity
39.64
59.09
79.55
100.00
Dendrogram 3
3
2
1
4
5
6

Участие аналитика

1. Отбор переменных
2. Метод стандартизации
3. Расстояние между кластерами
4. Расстояние между объектами

Отбор переменных

- 1. Какие переменные будут использоваться при анализе?

- Все?
- Как влияет цвет глаз покупателя на средний объем выпиваемого пива?
- Распознавание танков

С другой стороны

- если нам неизвестны зарплаты/доходы покупателей, но для каждого из них известны профессия, образование и стаж работы, исключение этих трех переменных влечет за собой исключение из рассмотрения платежеспособность покупателей.
- Если классифицируются школы, и не включены ни переменная «число школьников», ни переменная «число учителей», то кластеры будут формироваться без учета размера школ.

Вывод

- Правильный выбор переменных очень важен.
- Критерием при отборе переменных для анализа является в первую очередь ясность интерпретации полученного результата, во вторую – интуиция исследователя.

Надо ли стандартизировать переменные?

- Правило для новичка:

- если Вы не знаете, стандартизировать или нет, стандартизируйте.

Надо стандартизировать

5296782.7	0.5	1
7400381.4	0.7	0
9362870.2	0.1	0
7594038.5	0.4	0
6455034.1	0.4	1

Стандартизация

- Для каждого столбца.
- Линейное преобразование

- 1. Максимальное значение =1, минимальное = 0 (-1)
- 2. z-метки. Среднее равно 0, выборочная дисперсия равна 1.

- Иногда решением будет преобразование данных

- Если кластеров нет
- Они все равно будут найдены

Результаты кластерного анализа нуждаются в интерпретации

- какой вариант кластеризации даст лучшие результаты?
- тот, который вы смогли понять и проинтерпретировать

Еще раз об участии аналитика

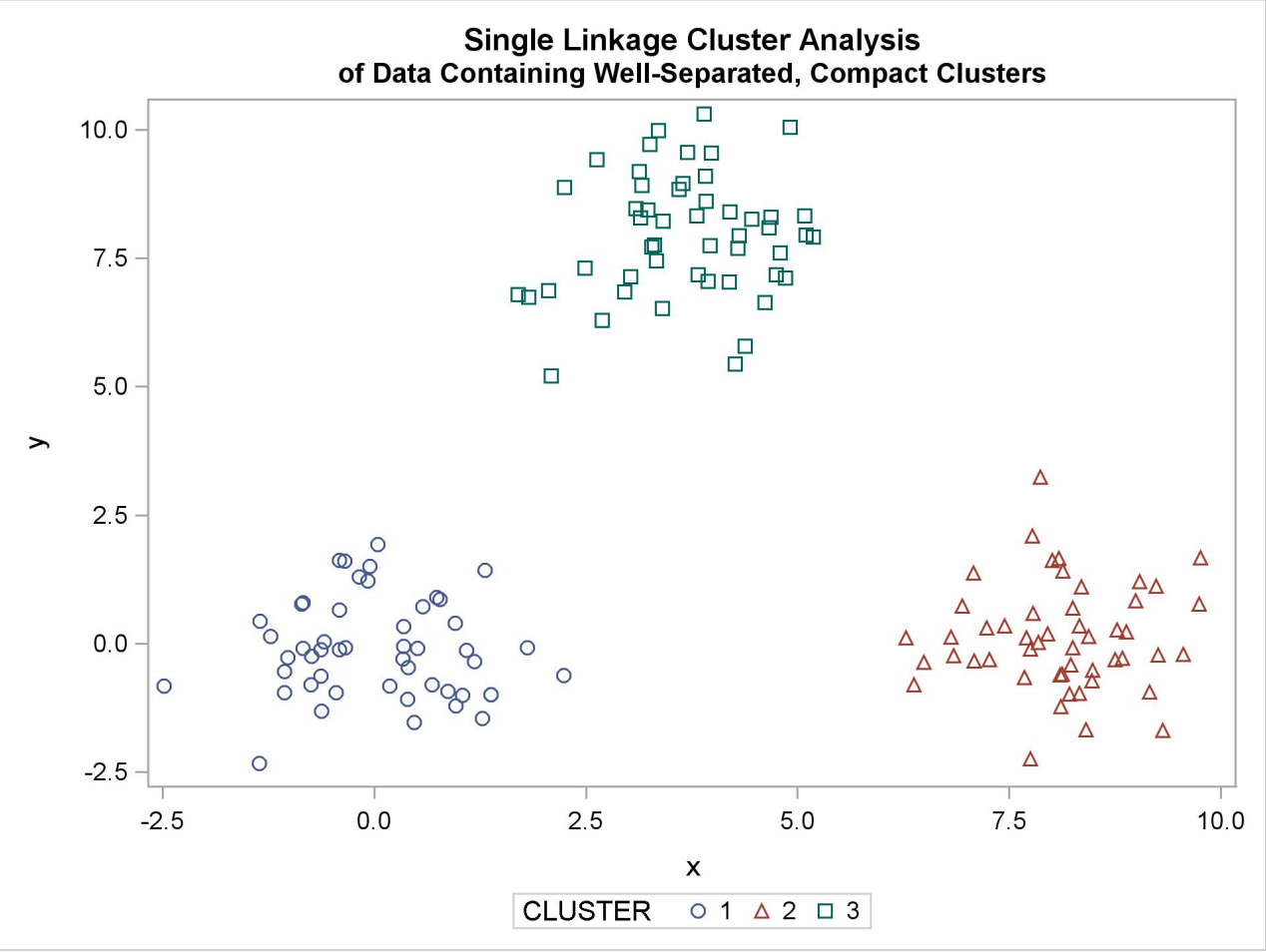
Иерархический кластерный анализ требует вдохновенного выбора способа подсчета расстояния между объектами и расстояния между кластерами. Кроме того, надо угадать число кластеров. Потом останется неясной геометрия кластеров. Таким образом, многое надо угадать и осмыслить. Не всегда это удается.

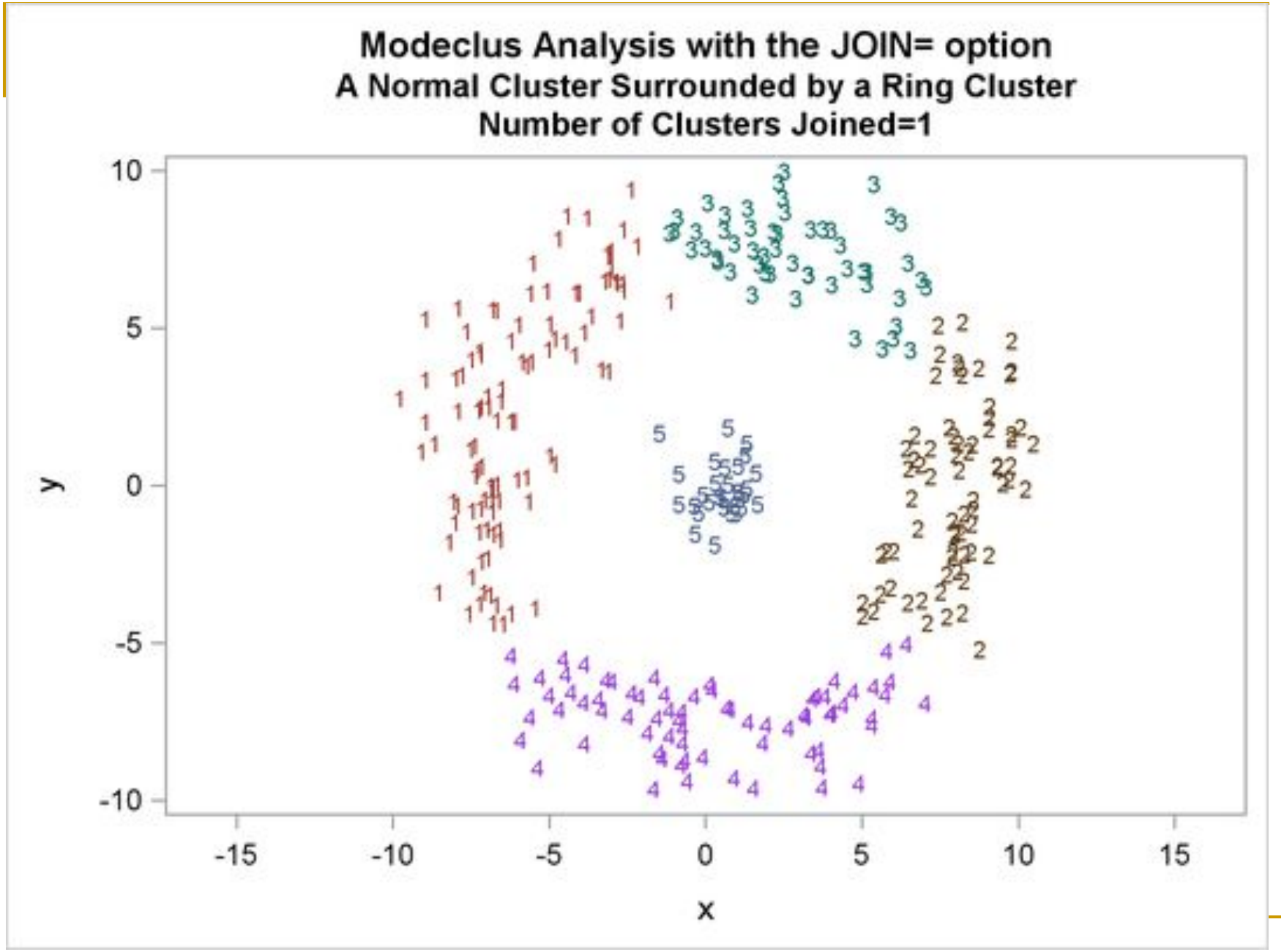
Типы кластеров

- Шаровые
- Ленточные
- ...

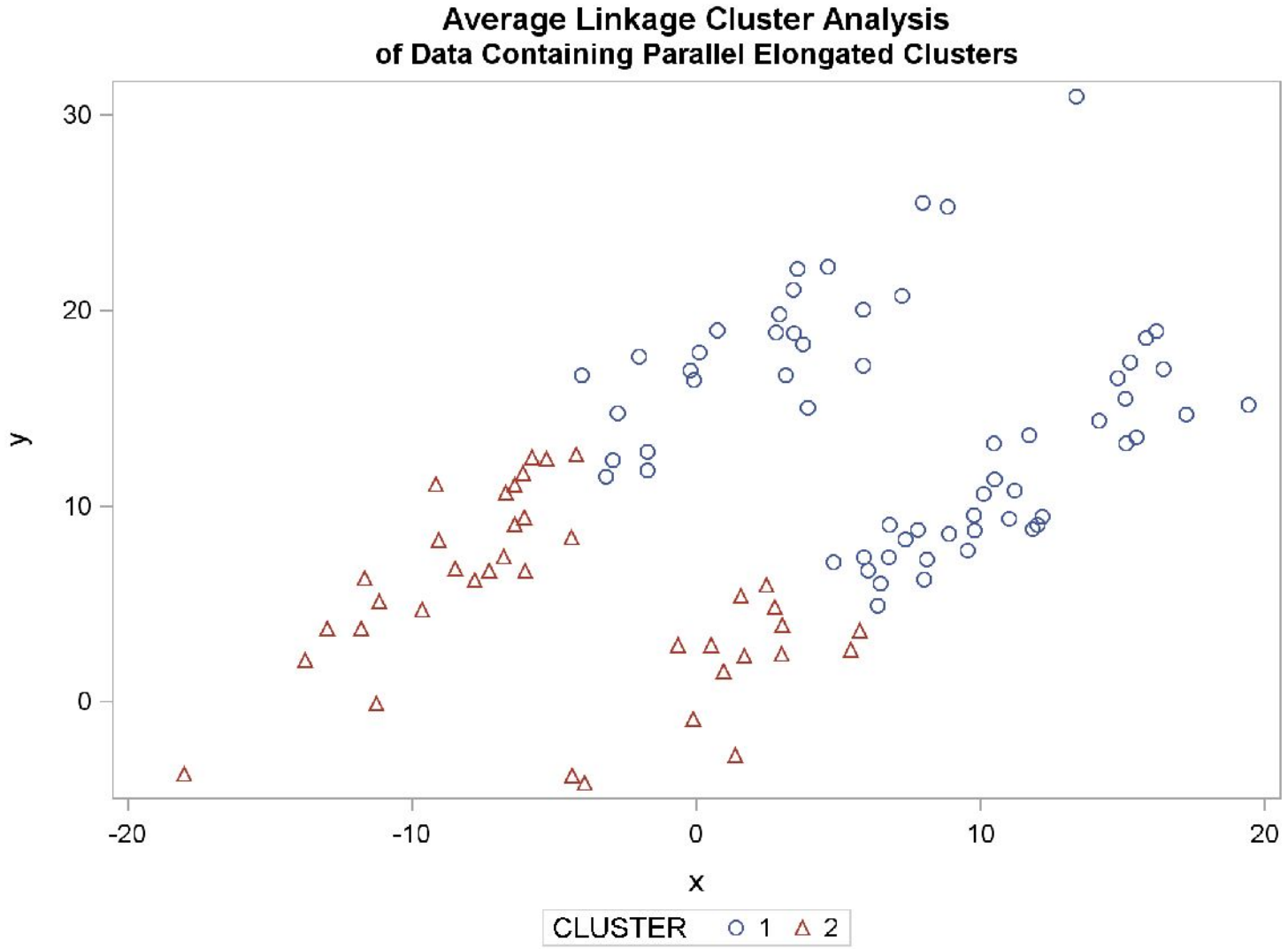
- Выбор расстояния между кластерами

Выраженные кластеры – все равно какой метод

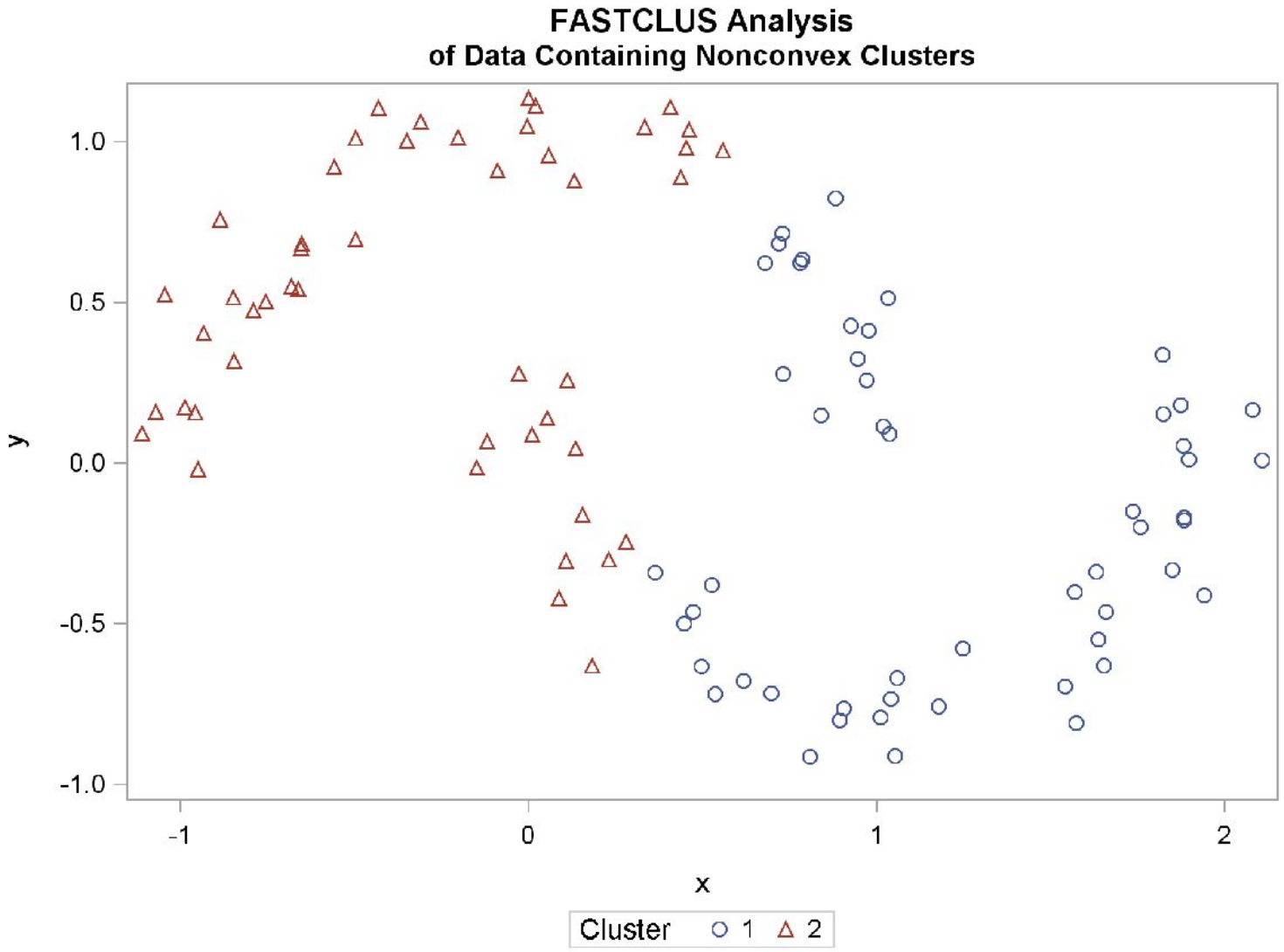



Modeclus Analysis with the JOIN= option
A Normal Cluster Surrounded by a Ring Cluster
Number of Clusters Joined=1
y
x
10
5
0
-5
-10
-15
-10
-5
0
5
10
15

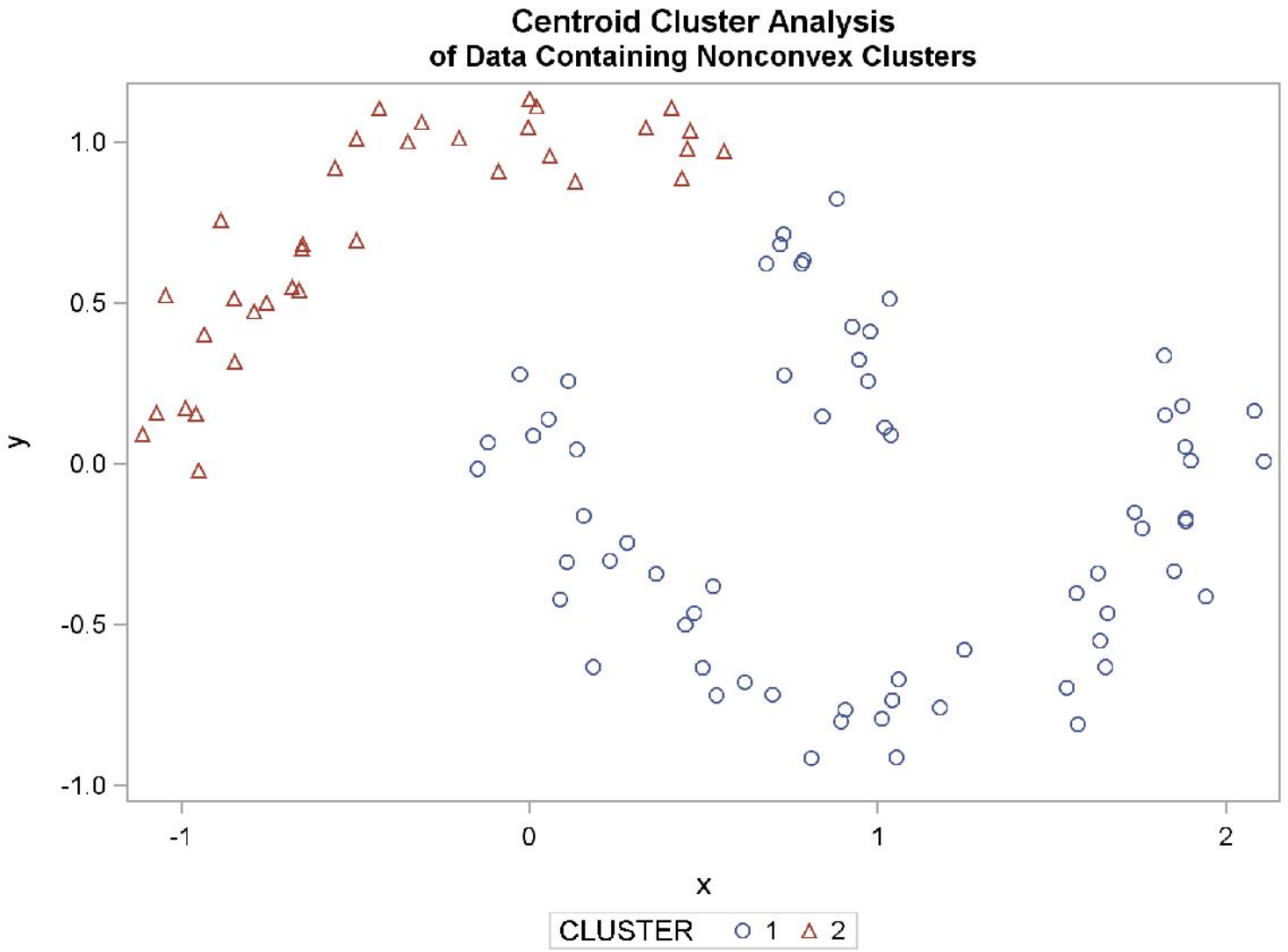
Какой метод будет лучше?



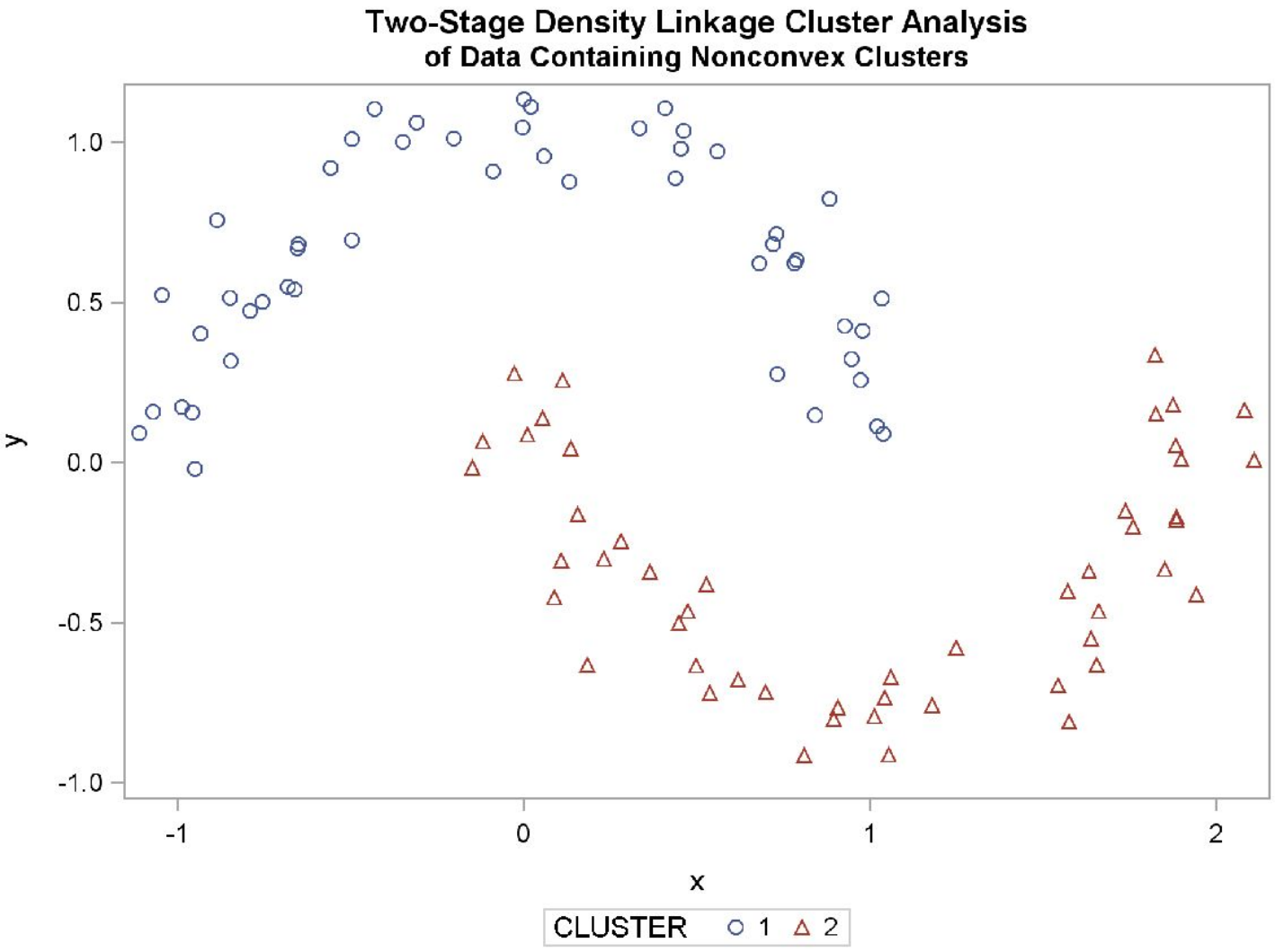
Неудача


FASTCLUS Analysis
of Data Containing Nonconvex Clusters
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
2
x
Cluster 1 2

Неудача


Centroid Cluster Analysis
of Data Containing Nonconvex Clusters
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
2
x
CLUSTER 1 2

Метод ближайшего соседа



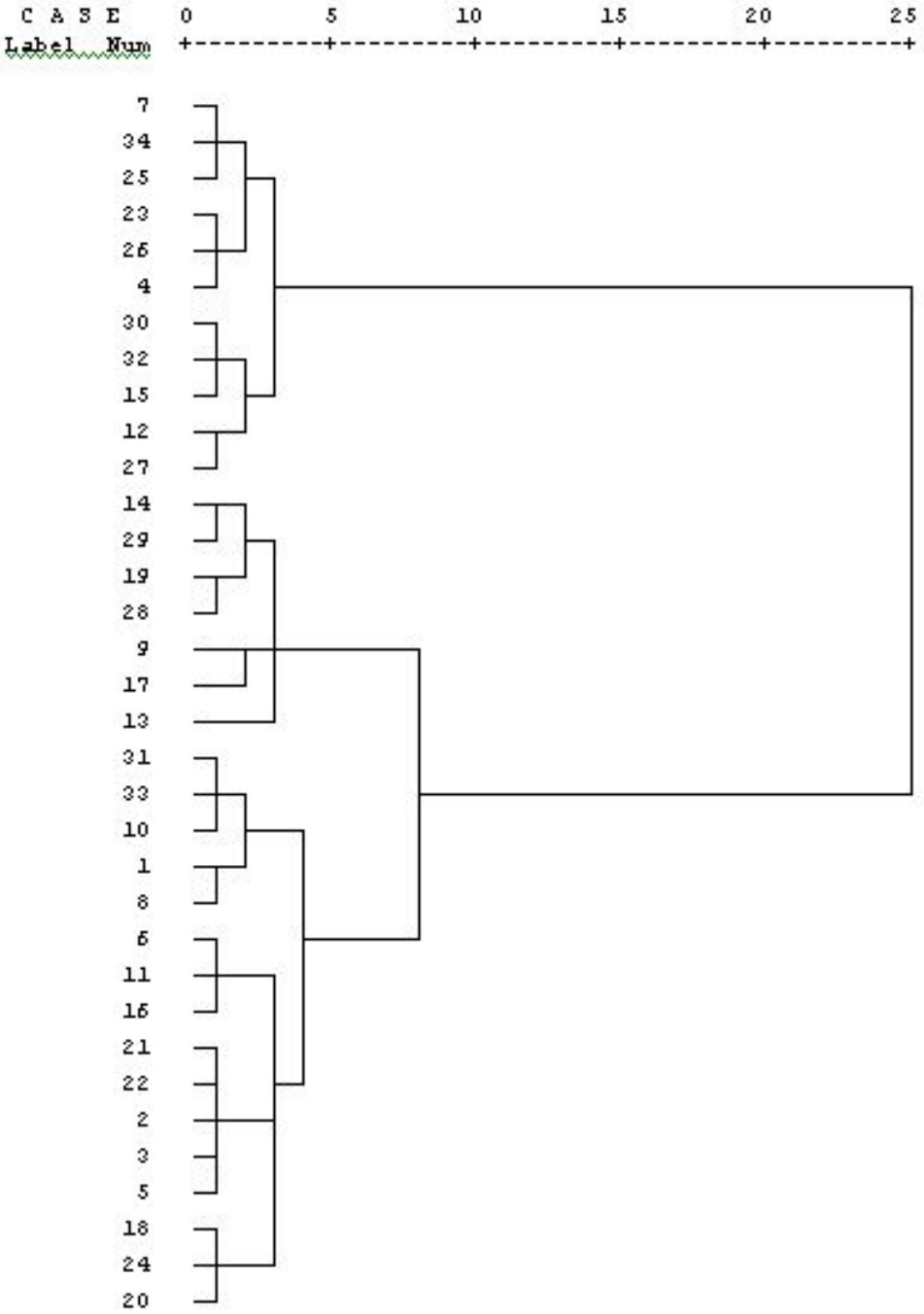
Пример

- Сегментация потребителей безалкогольных напитков

- Компания провела опрос с целью выявить, какие напитки предпочитают респонденты. Опрошенные указывали, какие напитки из предложенного списка они пьют регулярно.

В списке присутствовали

- Coca-Cola,
- диетическая Coca-Cola,
- Pepsi-Cola,
- диетическая Pepsi-Cola,
- 7-Up
- диетический 7-Up,
- Спрайт,
- минеральная вода


C A S E
Label Num
0
5
10
15
20
25
7
34
25
23
26
4
30
32
15
12
27
14
29
19
28
9
17
13
31
33
10
1
8
6
11
16
21
22
2
3
5
18
24
20

Решение для трех кластеров

- перечисляя сверху вниз на дендрограмме,
- В верхний кластер войдут респонденты с номерами от 7-го до 27-го,
- в средней группе – от 14-го до 13-го,
- в нижний – от 31-го до 20-го.

- R нумерует кластеры не сверху вниз!
- Как ему захочется!

1 кластер 16 наблюдений

COKE	15
D_COKE	4
D_PEPSI	1
D_7UP	0
PEPSI	16
SPRITE	5
TAB	0
SEVENUP	5

2 кластер 11 наблюдений

COKE	0
D_COKE	11
D_PEPSI	6
D_7UP	6
PEPSI	0
SPRITE	0
TAB	10
SEVENUP	0

3 кластер 7 наблюдений

COKE	5
D_COKE	2
D_PEPSI	1
D_7UP	1
PEPSI	0
SPRITE	6
TAB	1
SEVENUP	4

Потребление протеинов в Европе

- Переменные
- redmeat Мясо
- whitemeat Птица
- eggs Яйца
- milk Молоко
- fish Рыба
- cereals Хлебо-булочные
- starch Крахмал: картофель, макароны
- nuts Орехи
- fruits_v Фрукты и овощи

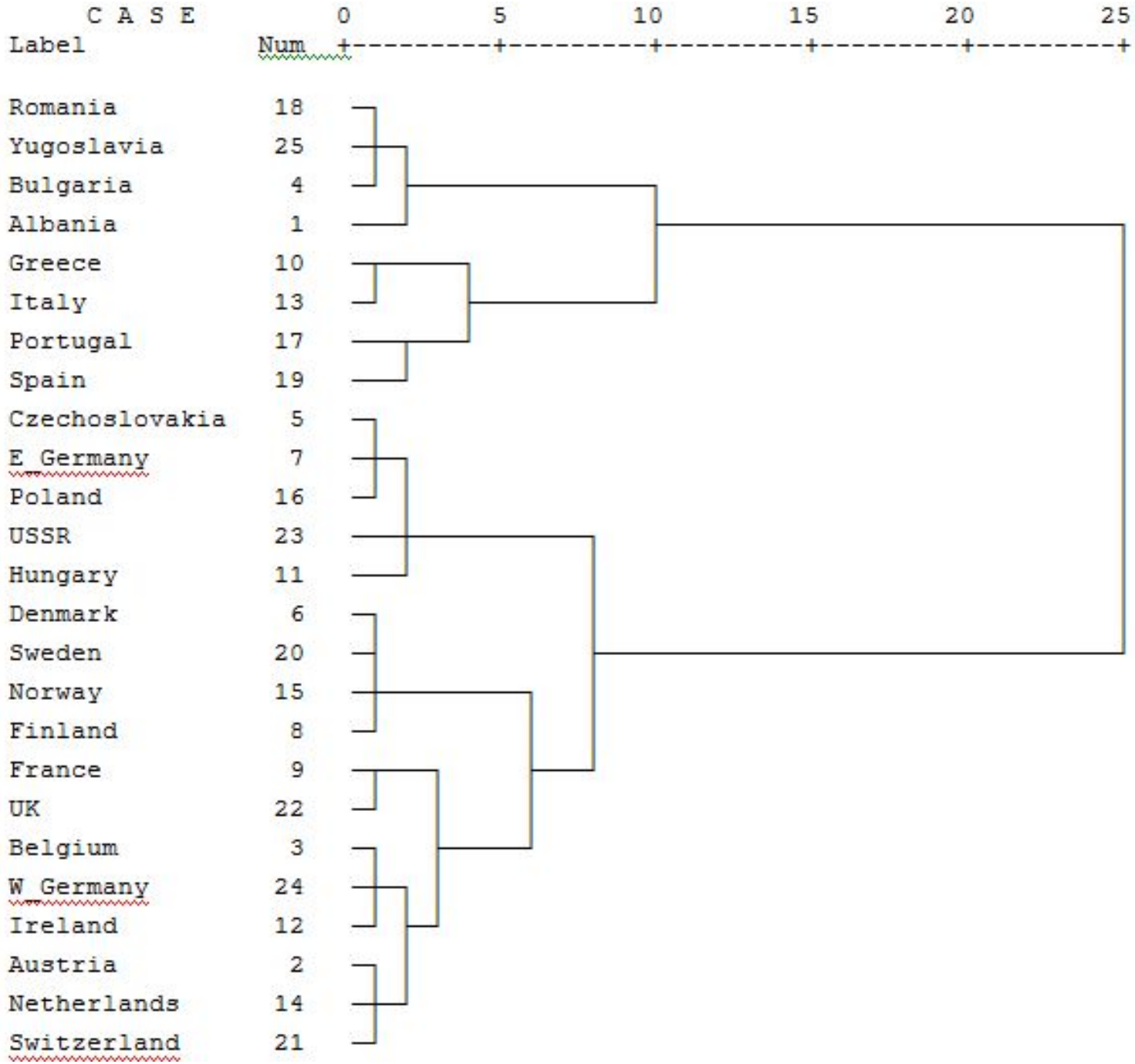
Задача:

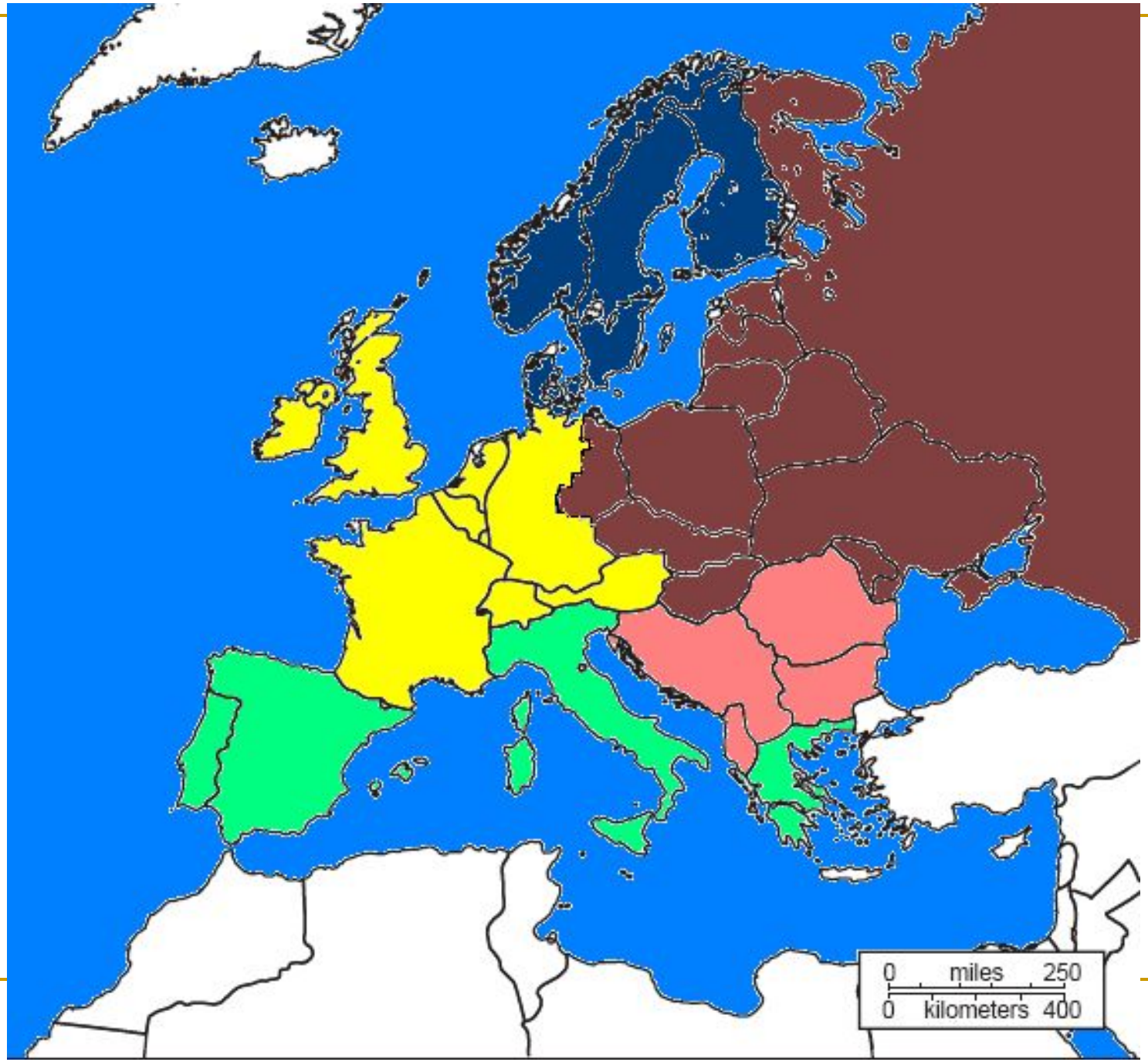
- Разбить страны на группы.

- Надо ли проводить стандартизацию?
- Как отличаются кластеры?

- (Использовалось решение Тропинина.)

- Стандартизация обязательна, так как средние значения некоторых переменных отличаются в десятки раз.
- Из всех методов иерархического кластерного анализа наиболее понятную картину дал
- метод Варда + стандартизация [0, 1]


CASE
Label
Num
0
5
10
15
20
25
Romania 18
Yugoslavia 25
Bulgaria 4
Albania 1
Greece 10
Italy 13
Portugal 17
Spain 19
Czechoslovakia 5
E Germany 7
Poland 16
USSR 23
Hungary 11
Denmark 6
Sweden 20
Norway 15
Finland 8
France 9
UK 22
Belgium 3
W Germany 24
Ireland 12
Austria 2
Netherlands 14
Switzerland 21


0 miles 250
0 kilometers 400

1.  Страны Балканского полуострова (кроме Греции), социализм

2.  Западная Европа

3.  Восточная Европа (остальные страны соц.лагаря)

4.  Северная Европа

5.  Страны Южной Европы, капитализм

- особенности питания зависят от
- географического положения и от
- экономического строя,

- что вполне естественно

Далее, сравниваем потребление в разных кластерах

- 1 кластер: большое потребление злаков и орехов (Pulses, nuts, and oil-seeds);
- маленькое потребление мяса (Red meat, White meat), рыбы, крахмалистых продуктов (Starchy foods) и яиц.
- 2 кластер: большое потребление мяса, яиц, молока; небольшое потребление злаков и орехов.
- 3 кластер: большое потребление птицы (White meat), крахмалистых продуктов; небольшое потребление орехов.
- 4 кластер: большое потребление яиц, молока, рыбы; маленькое потребление злаков, орехов, фруктов и овощей.
- 5 кластер: большое потребление рыбы, орехов, фруктов и овощей; маленькое потребление птицы.

Мы узнали что-то новое? Или результат тривиальный?

- Англия и Франция
- Социализм или капитализм: две германии в разных кластерах
- Два социализма