Дискриминантный анализ

Есть зверьки разного возраста, у которых измеряли 20 показателей. По каким из них лучше всего определяется возраст?

Собирали данные про школьников 11-го класса (20 разнокачественных переменных); после этого школьники поступили в ВУЗ, колледж или вообще никуда не поступили. Какие показатели лучше всего предсказывают судьбу школьника?

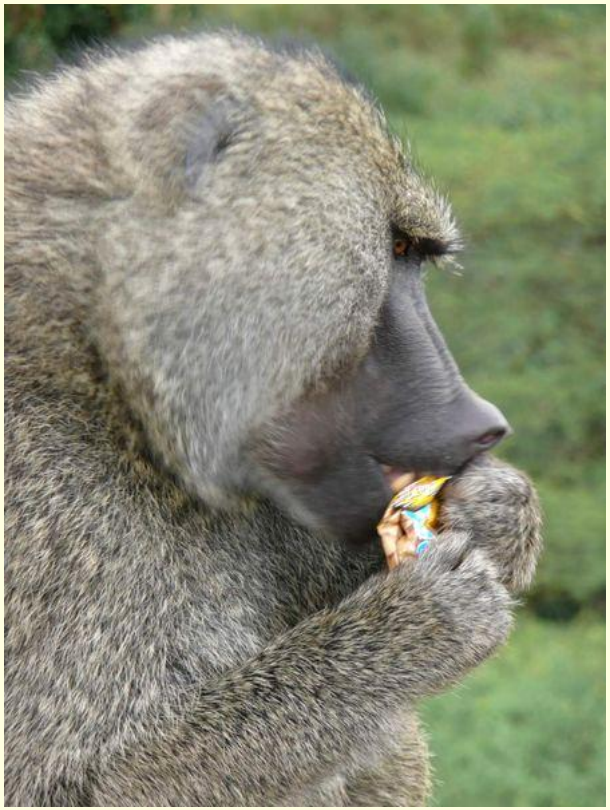
Дисперсионный факторный анализ

Прикладная задача: изучить пищевые предпочтения павианов и разработать комплексные оценки того, как они относятся к разным типам пищи.

Пища павианов - разнообразная, типов пищи – 10.

Гипотеза: реальных факторов, определяющих предпочтения павианов, меньше.

Вопрос: сколько (и какие) факторы определяют пищевые предпочтения павианов.

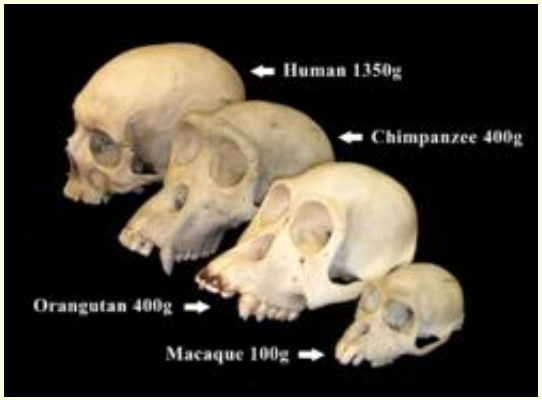


Павиа́ны (Papio) — род приматов из семейства мартышковых.

Дисперсионный факторный анализ

Задача анализа: *на*йти факторы, определяющие изменчивость (объясняющие действие) большого количества измеренных реальных переменных.

Подразумевается, что таких факторов гораздо меньше, чем исходных переменных.



Факторный анализ

Анализ главных компонент (principal component analysis)
Основная идея: получить факторы, объясняющие как можно больше общей изменчивости;
применение – сокращение числа переменных в анализе

Анализ главных факторов (principal factor analysis)
Основная идея: для каждой переменной используется только доля изменчивости, общая с другими переменными;
применение – поиск структуры переменных, определения их иерархии.

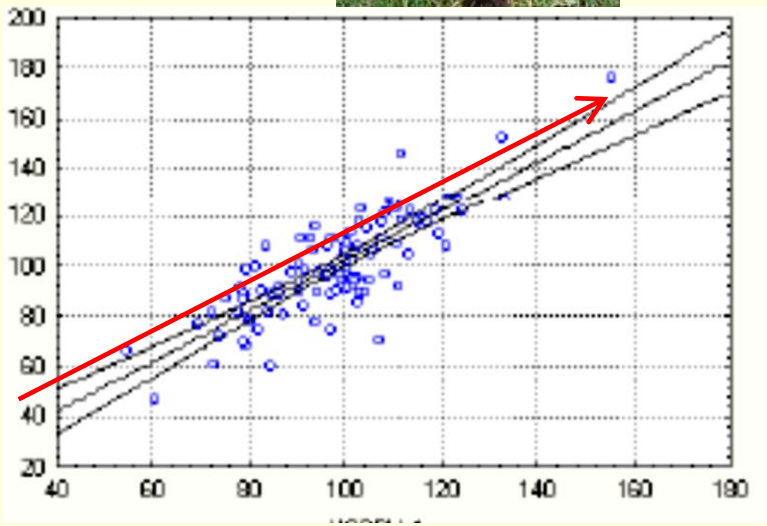
Анализ главных компонент

Гипотеза: измеренные переменные являются **линейными комбинациями** факторов.

Факторы (главные компоненты) находят на основании матрицы корреляции переменных – на основе **линий регрессии**.

Процедура анализа подобна вращению, максимизирующему дисперсию исходного пространства переменных.

После выделения первого фактора выделяют следующий, который также максимизирует оставшуюся дисперсию и т.д. – все факторы будут **ортогональны.**



Новая ось OX.

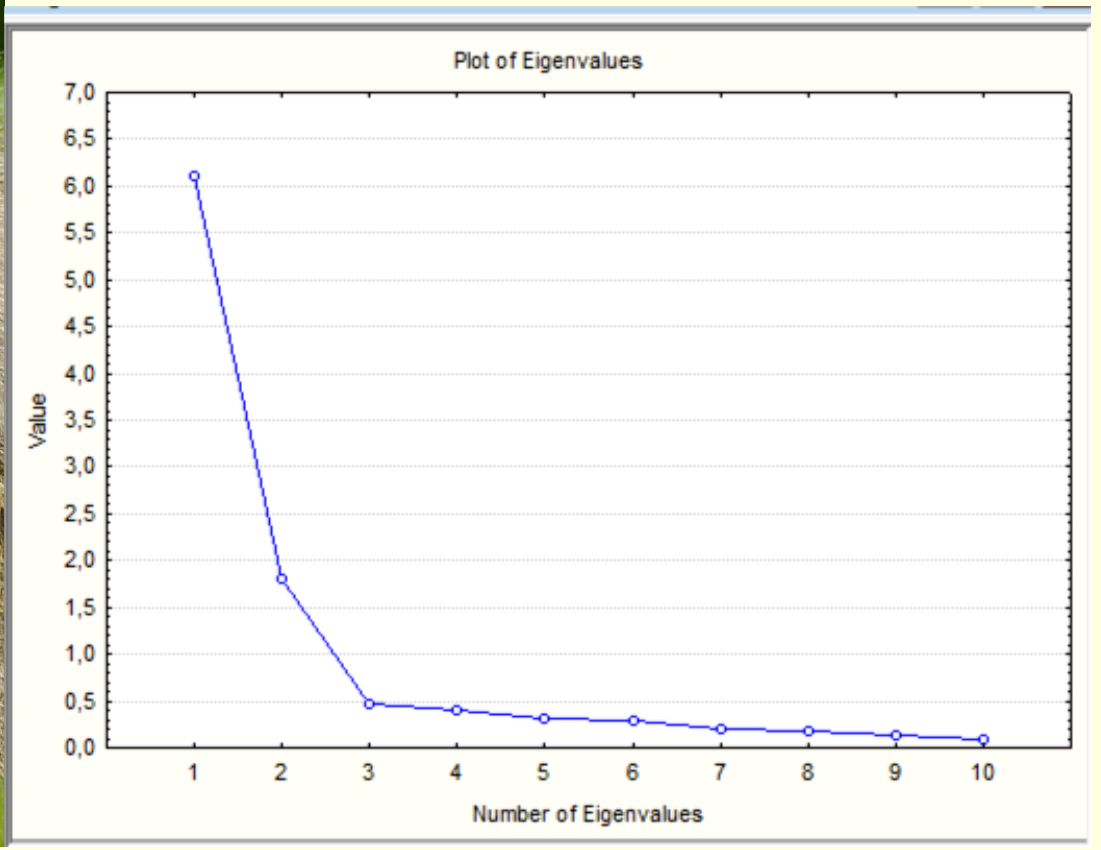
Анализ главных компонент

Типы пищи павианов:

апельсины,
бананы,
яблоки,
помидоры,
огурцы,
мясо,
курица,
рыба,
насекомые,
червяки.

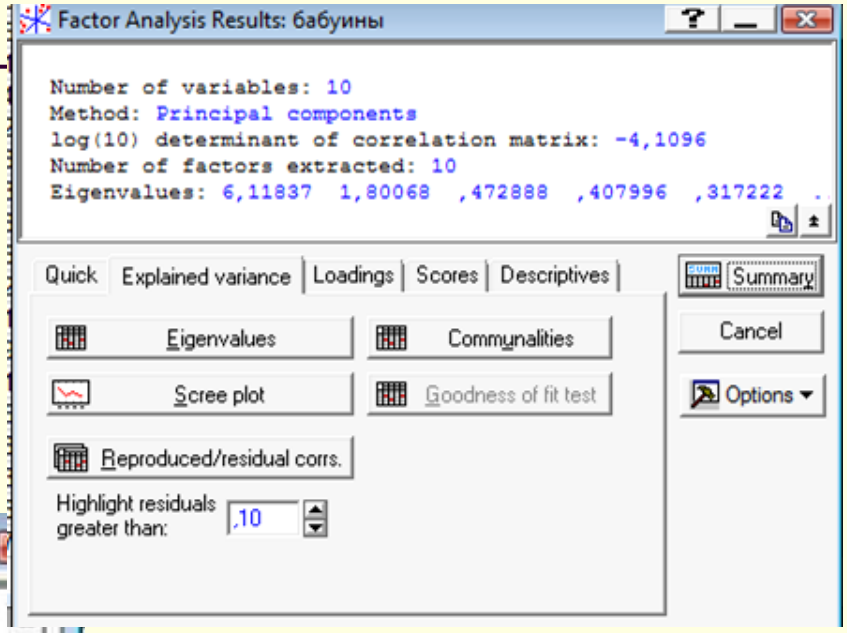
Сколько факторов скрывается за разными предпочтениями павианов в еде?

Анализ главных компонент



Анализ главных компонент

Собственные значения (eigenvalues)– определяют, какую долю общей дисперсии объясняет данный фактор.



nvalues (бабуины)

Eigenvalues (бабуины)
Extraction: Principal components

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6,118369	61,18369	6,11837	61,1837
2	1,800682	18,00682	7,91905	79,1905
3	0,472888	4,72888	8,39194	83,9194
4	0,407996	4,07996	8,79993	87,9993
5	0,317222	3,17222	9,11716	91,1716
6	0,293300	2,93300	9,41046	94,1046
7	0,195808	1,95808	9,60626	96,0626
8	0,170431	1,70431	9,77670	97,7670
9	0,137970	1,37970	9,91467	99,1467
10	0,085334	0,85334	10,00000	100,0000

Требования к выборкам для проведения факторного анализа

1. Внутри групп должно быть многомерное *нормальное распределение* (оценка – на основе построения гистограмм частот).
2. Гомогенность *дисперсий* (для метода главных компонент; не очень критичное требование).
3. Связь переменных должна быть *линейной.*
4. Размер выборки не должен быть меньше 50, оптимальный – ≥100 наблюдений.
5. Между переменными должна быть *ненулевая корреляция*, но коэффициентов корреляции, близких *единице*, тоже быть не должно.

Кластерный анализ

описательная математическая процедура группировки и классификации данных

Проверка статистической значимости неприменима

Понятие «кластер»

Термин "кластер" происходит от английского "cluster" - рой, гроздь, груда, скопление.

Кла́стер (англ. *cluster* скопление) — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

Кластерный анализ

Термин ***кластерный анализ*** (впервые ввел Tryon, 1939) в действительности включает в себя набор различных алгоритмов классификации.

Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как *организовать* наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии.

Техника кластеризации применяется в самых разнообразных областях. Хартиган (Hartigan, 1975) дал прекрасный обзор многих опубликованных исследований, содержащих результаты, полученные методами кластерного анализа.

Кластерный анализ

- **Кластерный анализ** (англ. *Data clustering*) — задача разбиения заданной выборки *объектов* (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.
- Кластерный анализ – один из методов многомерного анализа, предназначенный для группировки (кластеризации) совокупности элементов, которые характеризуются многими факторами, и получения однородных групп (кластеров).

Цели кластеризации

- ***Понимание данных*** путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).
- ***Сжатие данных.*** Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.
- ***Обнаружение новизны*** (англ. *novelty detection*). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Задачи кластерного анализа

- Проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение задачиприводит к углублению знаний о совокупности классифицируемых объектов;

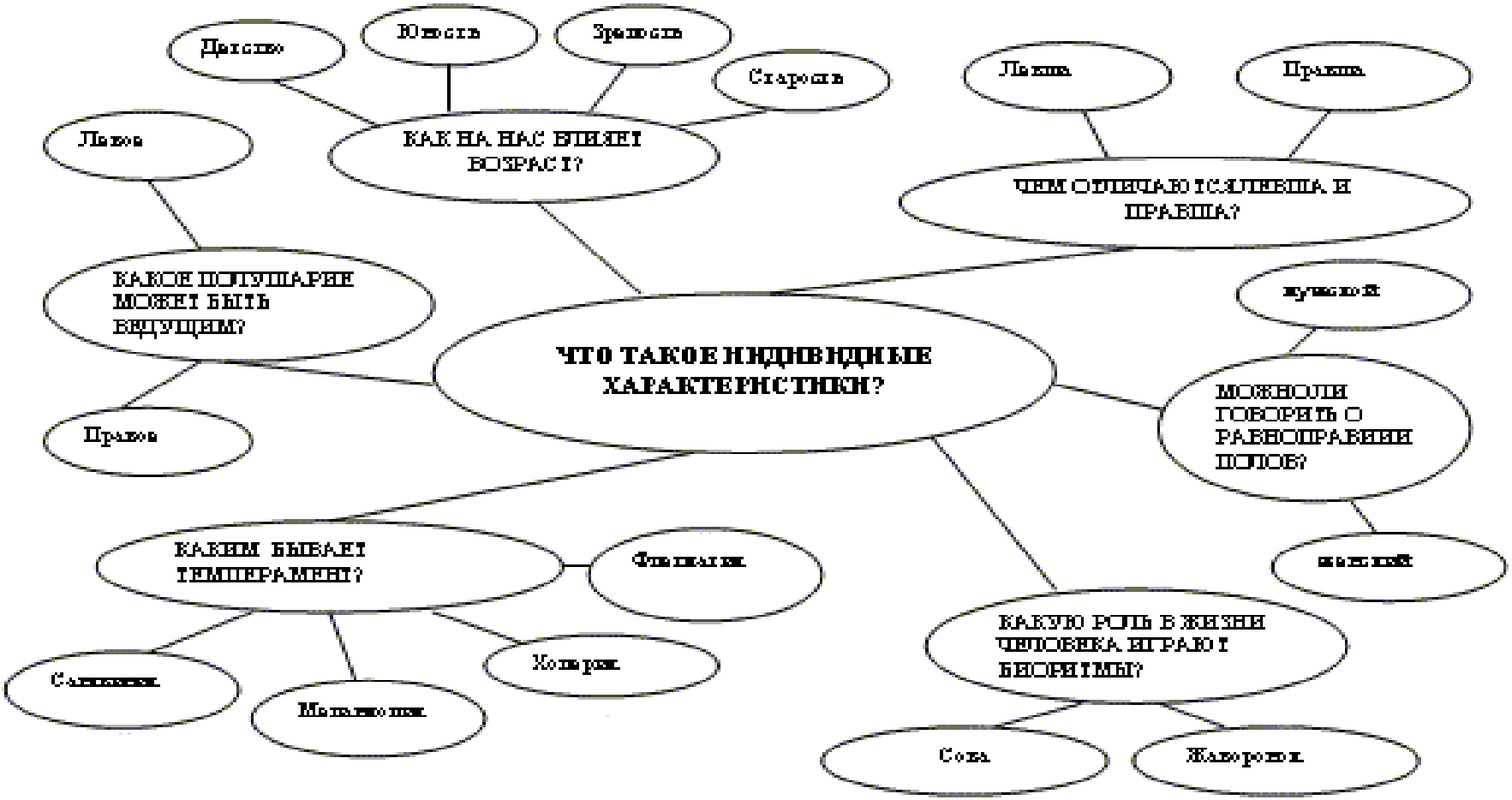
- Проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т.е. поиск существующей структуры;
- Построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Построение «кластеров»

В центральное овале располагается ключевое слово, понятие, фраза, в дополнительных слова, раскрывающие смысл ключевого.

С помощью кластеров можно в систематизированном виде представить большие объемы информации (ключевые слова, идеи).

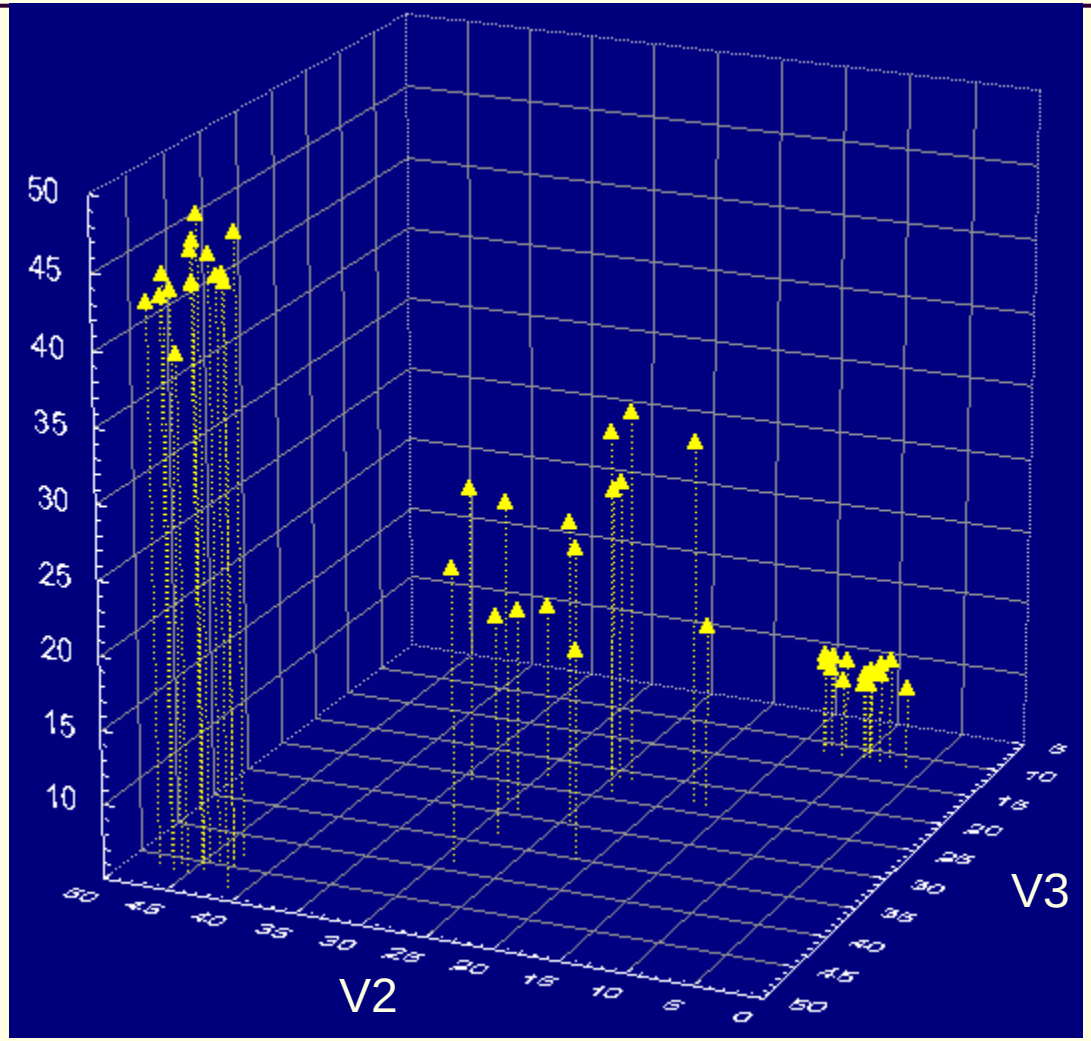
Пример кластеров для обучения



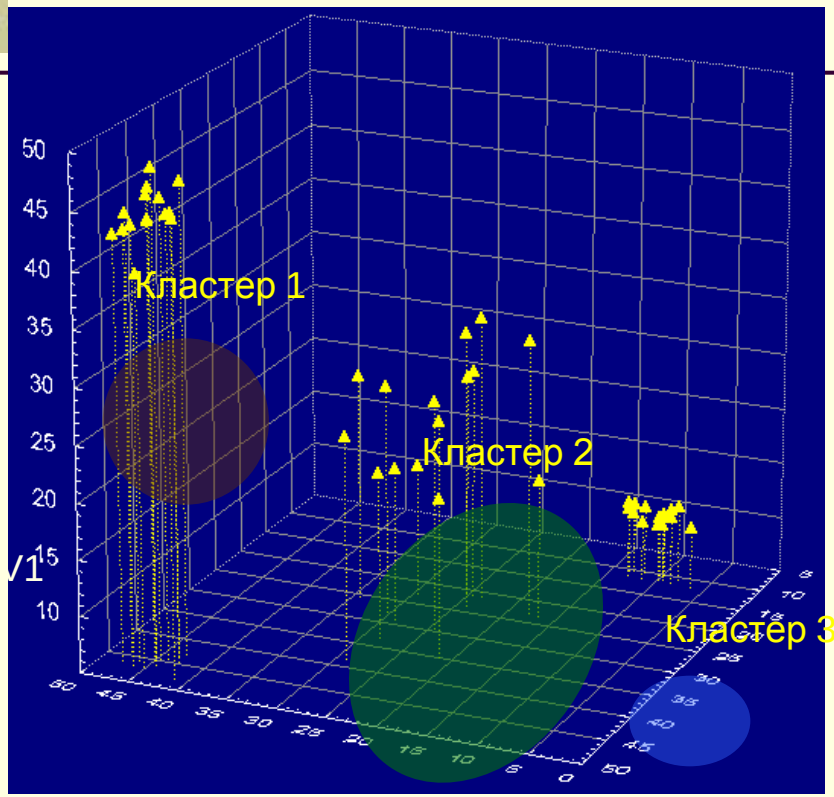
Пример кластеров для обучения

Пример кластеров для обучения

Кластерный анализ



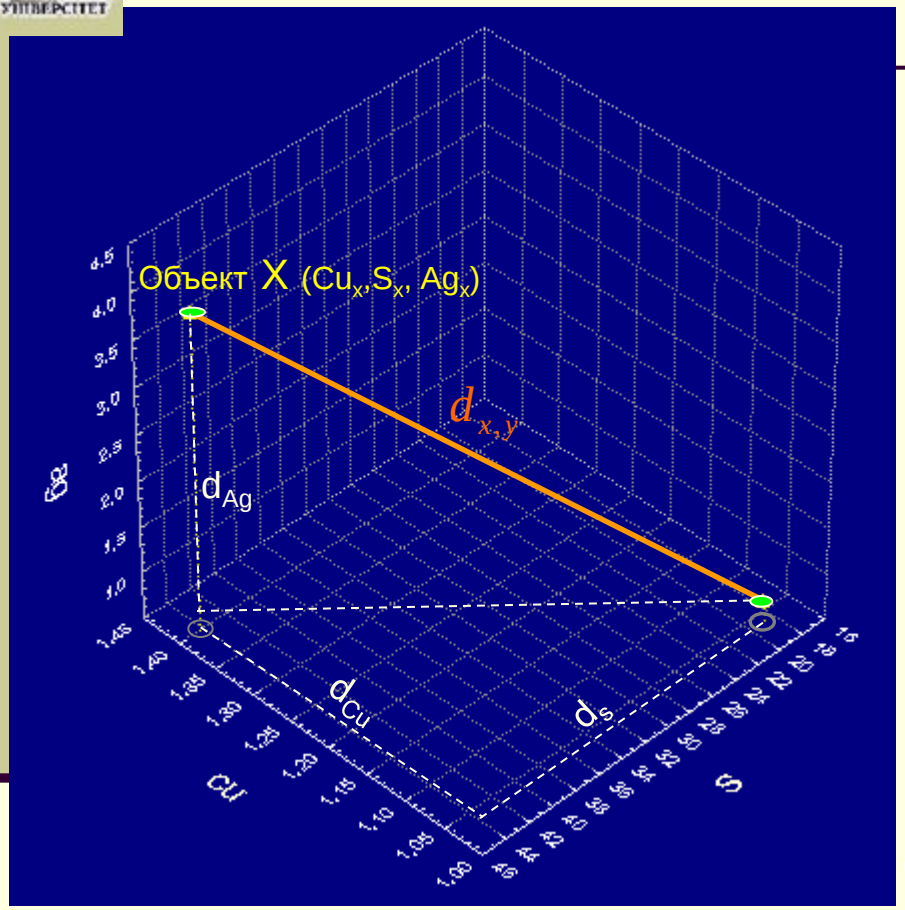
Кластерный анализ



Кластер No. 1	Кластер No. 2	Кластер No. 3
44.83276	25.50245	10.41274
45.40516	26.96871	10.97025
44.25457	23.77148	10.86401

Кластерный анализ – разбиение множества объектов на подмножества, называемые ***кластерами***, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Кластерный анализ



$d_{x,y}$ - расстояние между объектами x и y;

x_i, y_i - значения *i*-ого свойства объектов x и y

- количество свойств

Объект Y (Cu_y,S_y, Ag_y)

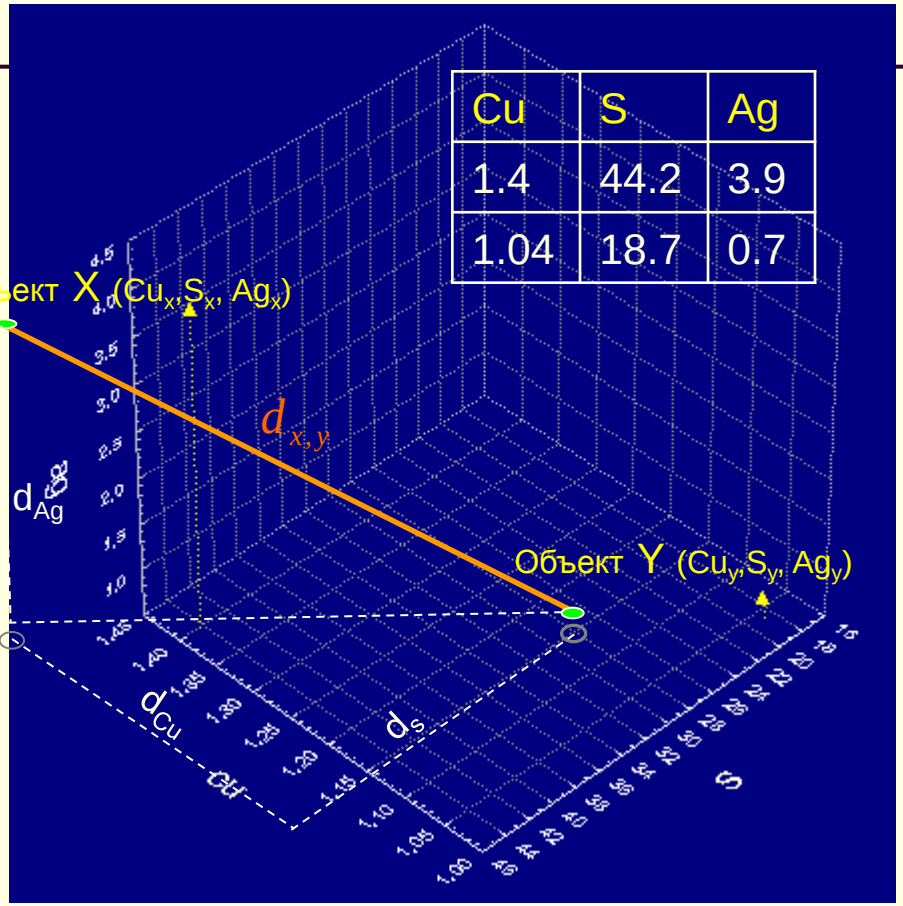
$$d_{x,y} = \sqrt{d_{Cu}^2 + d_S^2 + d_{Ag}^2}$$

Группировка наблюдений осуществляется на основе мер сходства.

$$l_{x,y} = \sqrt{(Cu_x - Cu_y)^2 + (S_x - S_y)^2 + (Ag_x - Ag_y)^2}$$

Чаще всего в качестве такой меры используется Евклидово расстояние (геометрическое расстояние в многомерном пространстве):

Кластерный анализ



$$d^2_{Ag} = (3,9 - 0,7)^2 = 10,24$$

$$d^2_{Cu} = (1,04 - 1,4)^2 = 0,13$$

$$d^2_{S} = (44,2 - 18,7)^2 = 650,25$$

$$650,25 / 660,62 * 100 = 98\%$$

Расстояние между объектами зависит от масштаба по осям.

В этом примере расстояние между объектами будет определяться в основном разницей содержаний серы.

Медь и серебро практически не будут учтены при выделении кластеров поскольку их содержания на порядок меньше, чем содержания серы.

Алгоритмы кластеризации

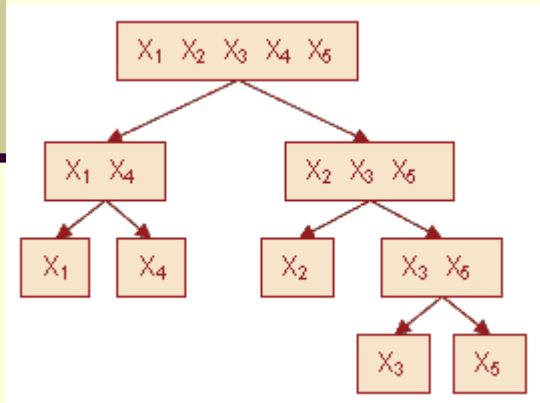
По способу разделения множества наблюдений на классы иерархический и итерационный.

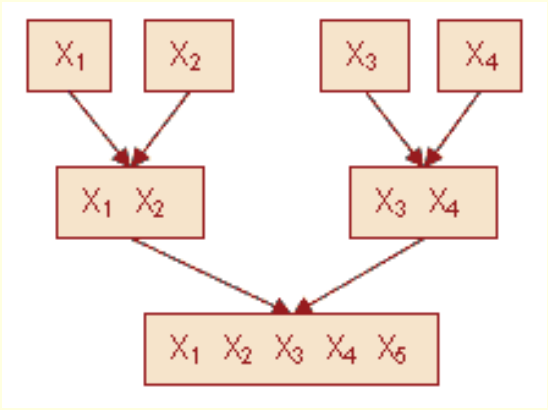
Иерархическая кластеризация

Иерархические методы – комплекс алгоритмов, использующих разделение крупных кластеров на более мелкие или объединение мелких в более крупные.

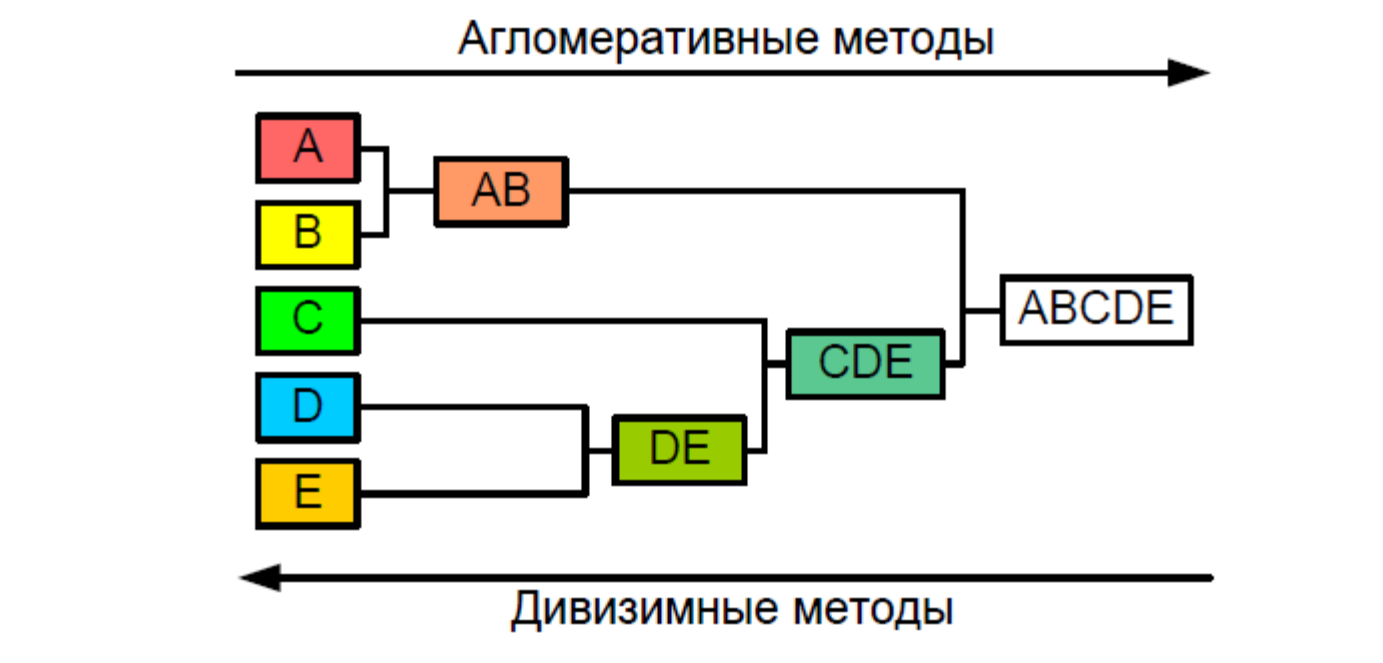
Соответственно, выделяют кластеризации

разделительную (дивизивную) и агломеративную (объединительную)





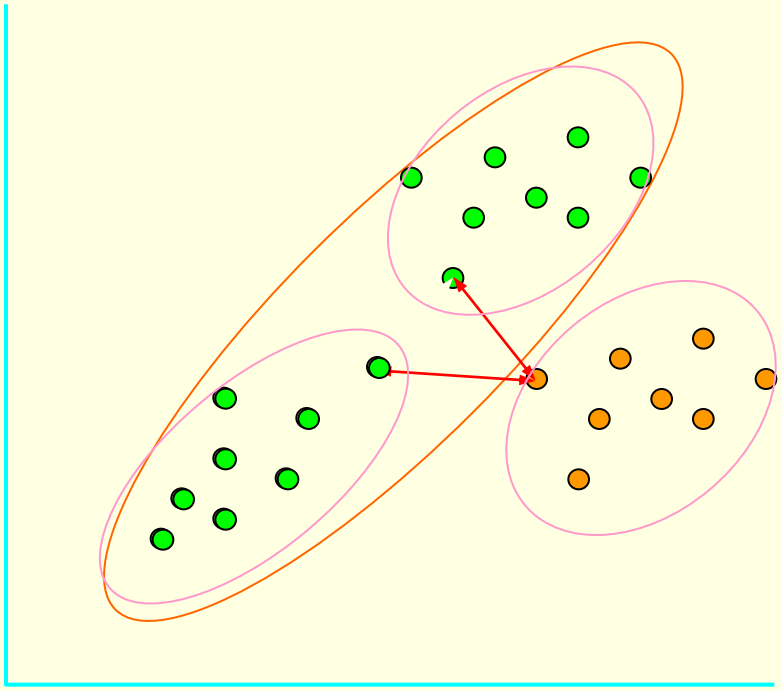
Алгоритмы кластеризации

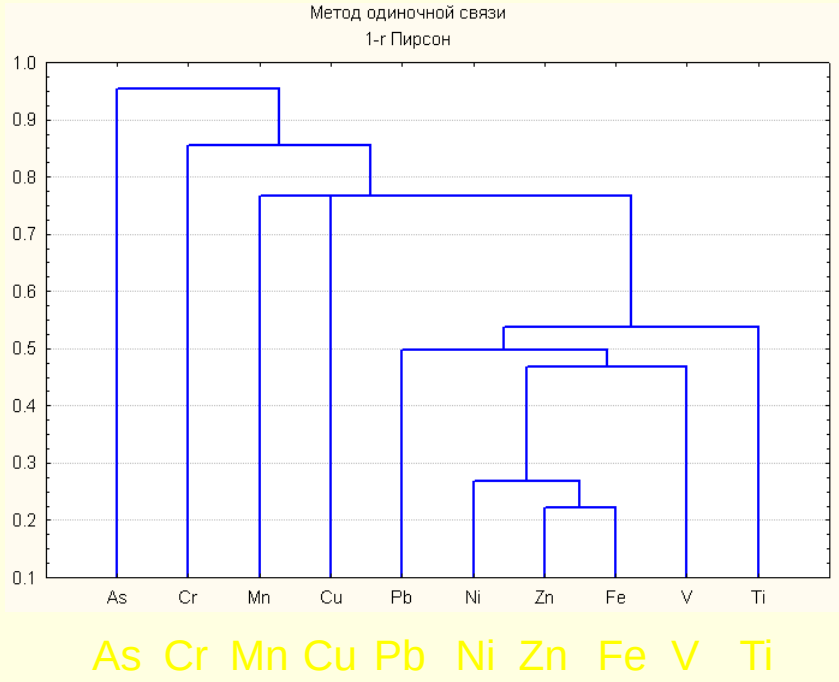


Методы связывания кластеров

метод ближайшего соседа

Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Результирующие кластеры имеют тенденцию быть представленными длинными "цепочками."

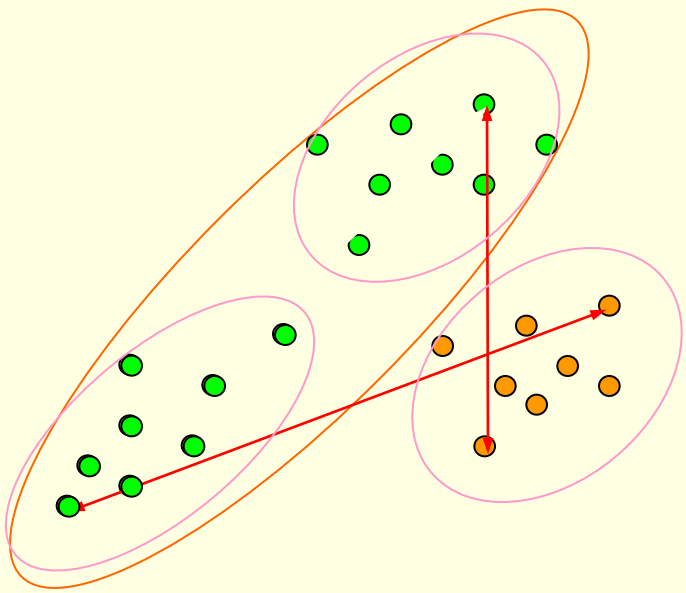


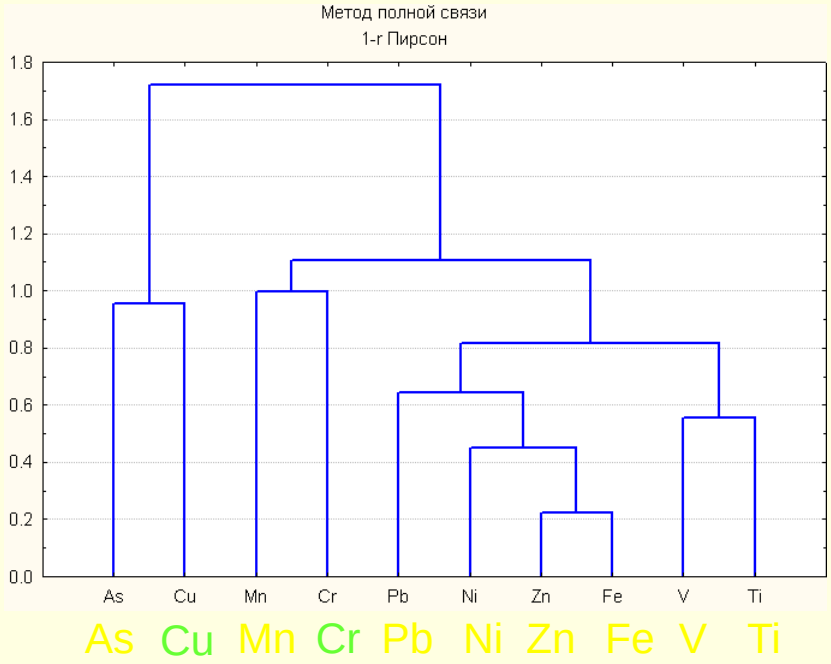


As Cr Mn Cu Pb Ni Zn Fe V Ti

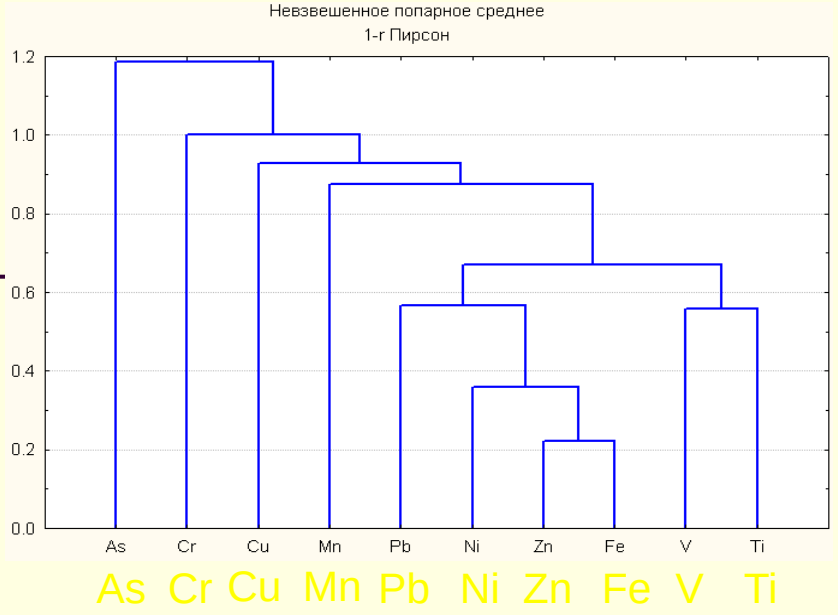
метод наиболее удаленных соседей

Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями ").

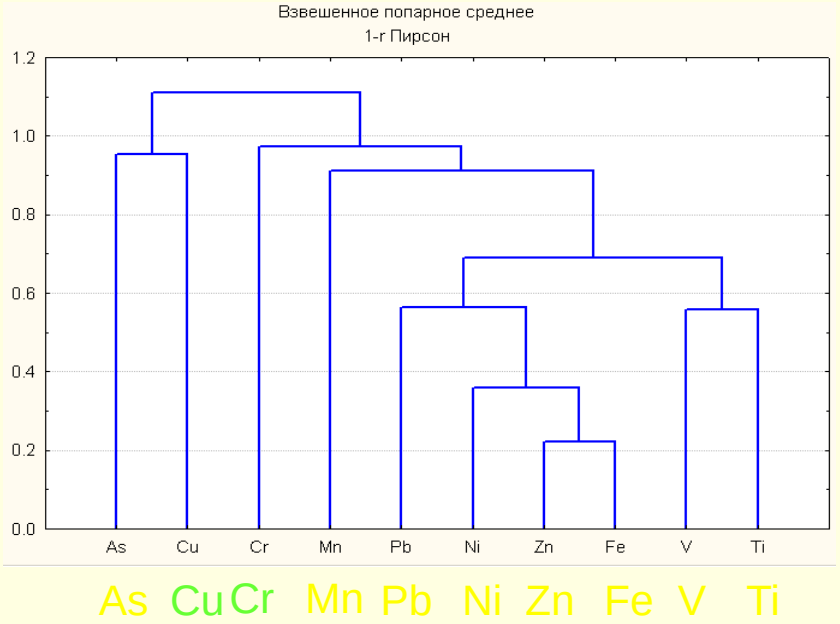




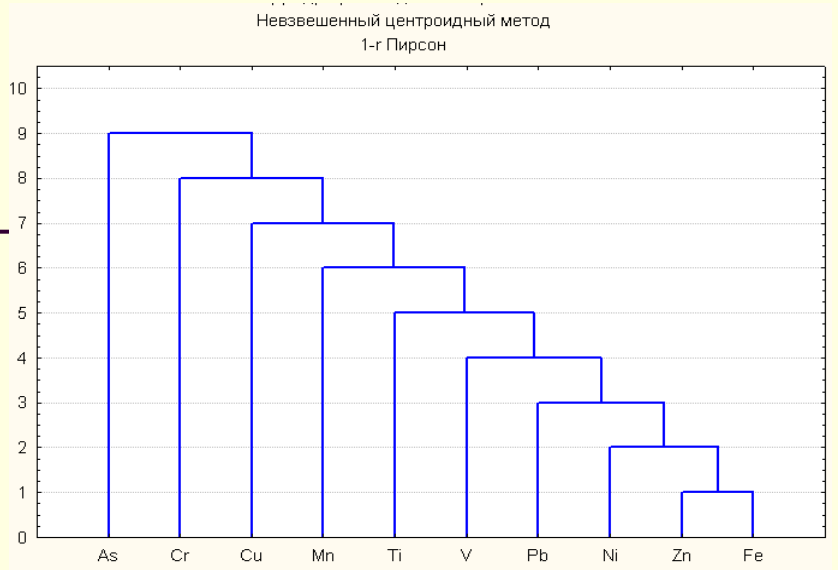
Расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.
Метод применим как в случаях хорошо обособленных кластеров, так и в случаях протяженных (цепочного типа) кластеров.



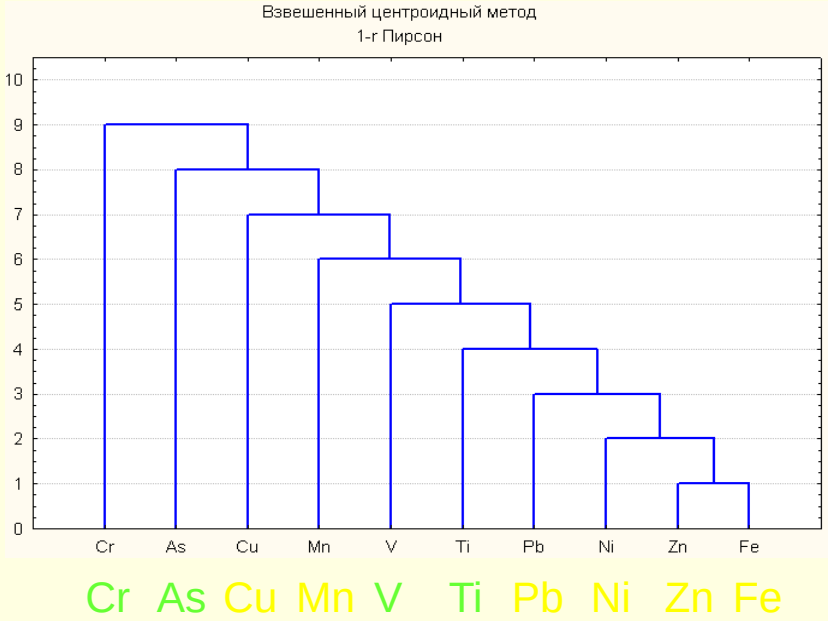
Взвешенное попарное среднее.
Отличается от невзвешенного тем, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Используется, когда предполагаются неравные размеры кластеров.

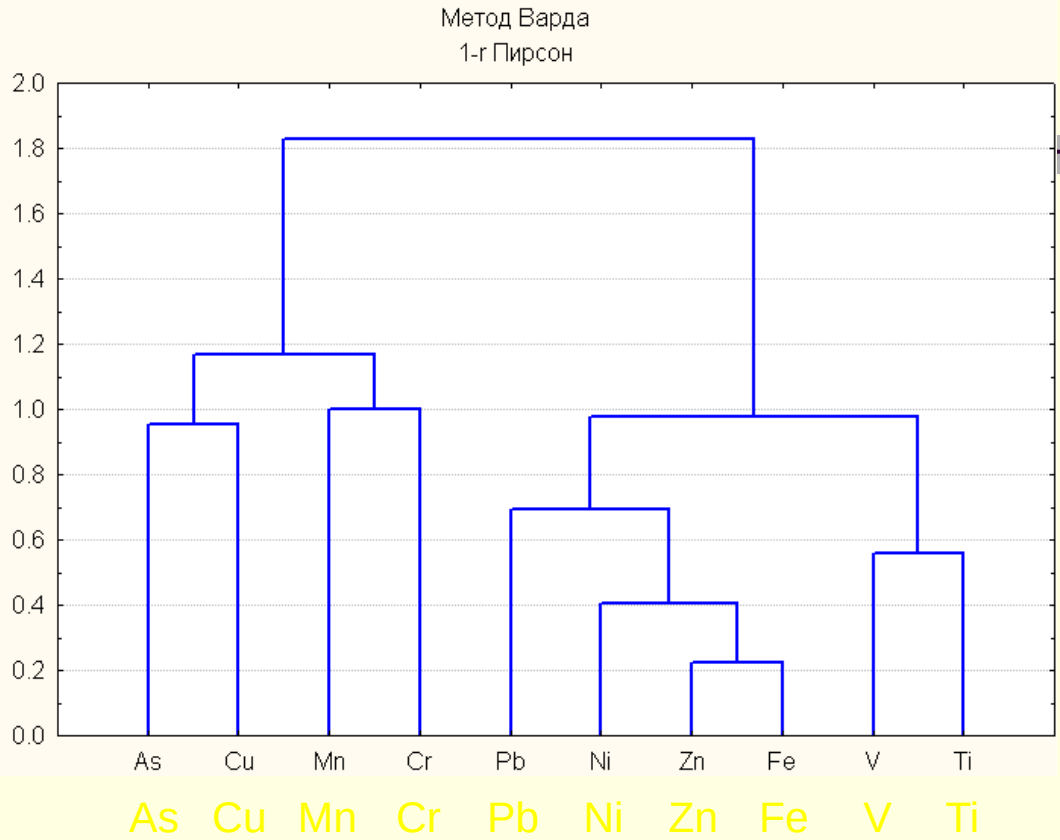


Невзвешенный центроидный метод Расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.



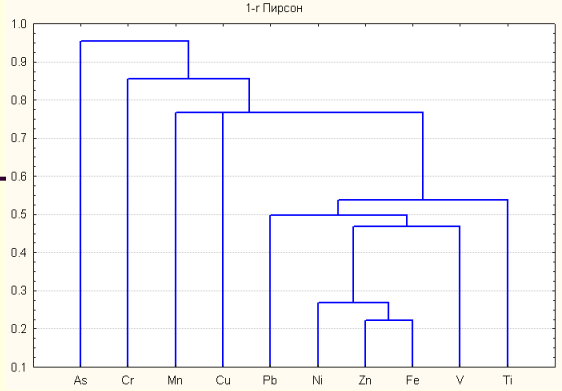
Взвешенный центроидный метод. Отличается от невзвешенного тем, что при вычислениях используются веса для учёта разницы между размерами кластеров (количество объектов).



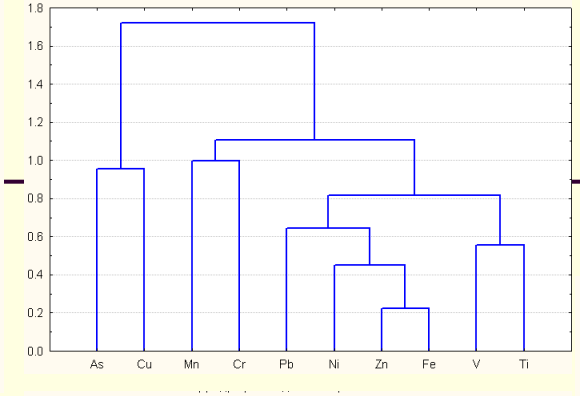


Этот метод использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (SS) для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

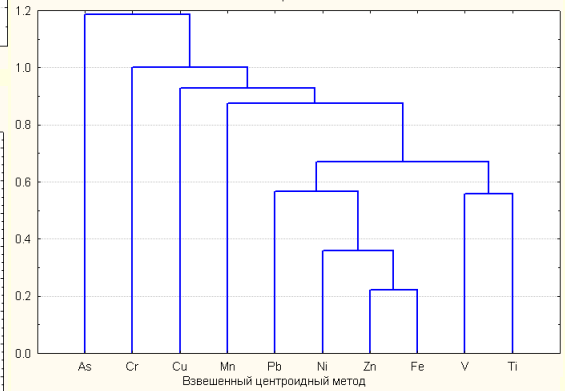
Метод одиночной связи
1-г Пирсон



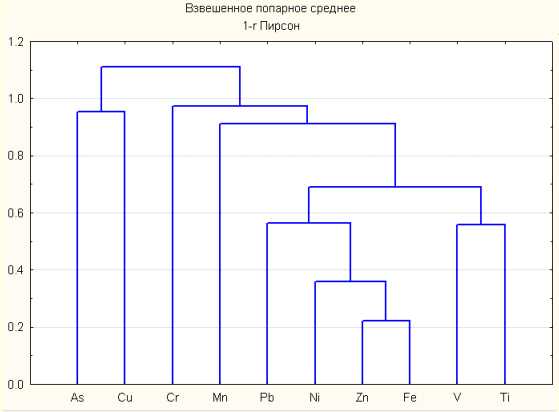
Метод полной связи
1-г Пирсон



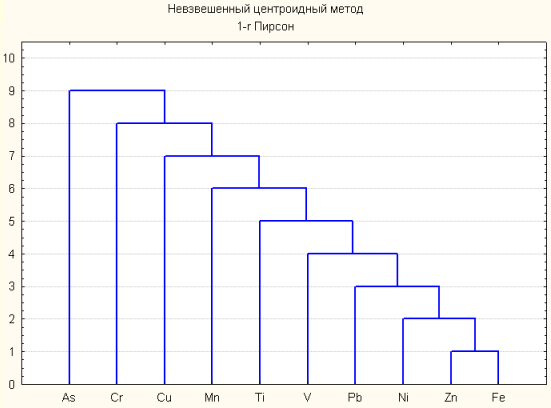
Невзвешенное попарное среднее
1-г Пирсон



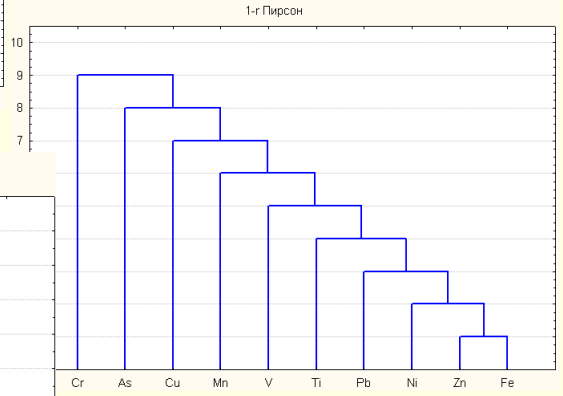
Взвешенное попарное среднее
1-г Пирсон



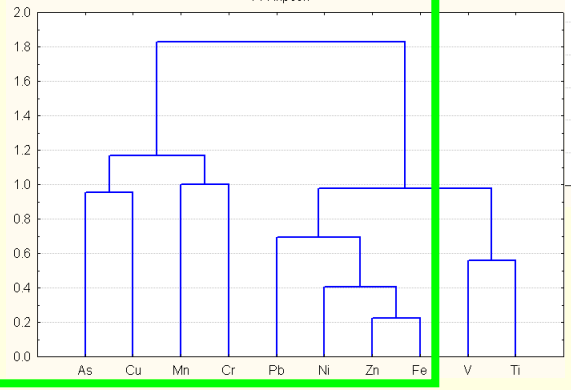
Невзвешенный центроидный метод
1-г Пирсон



Взвешенный центроидный метод
1-г Пирсон



Метод Варда
1-г Пирсон

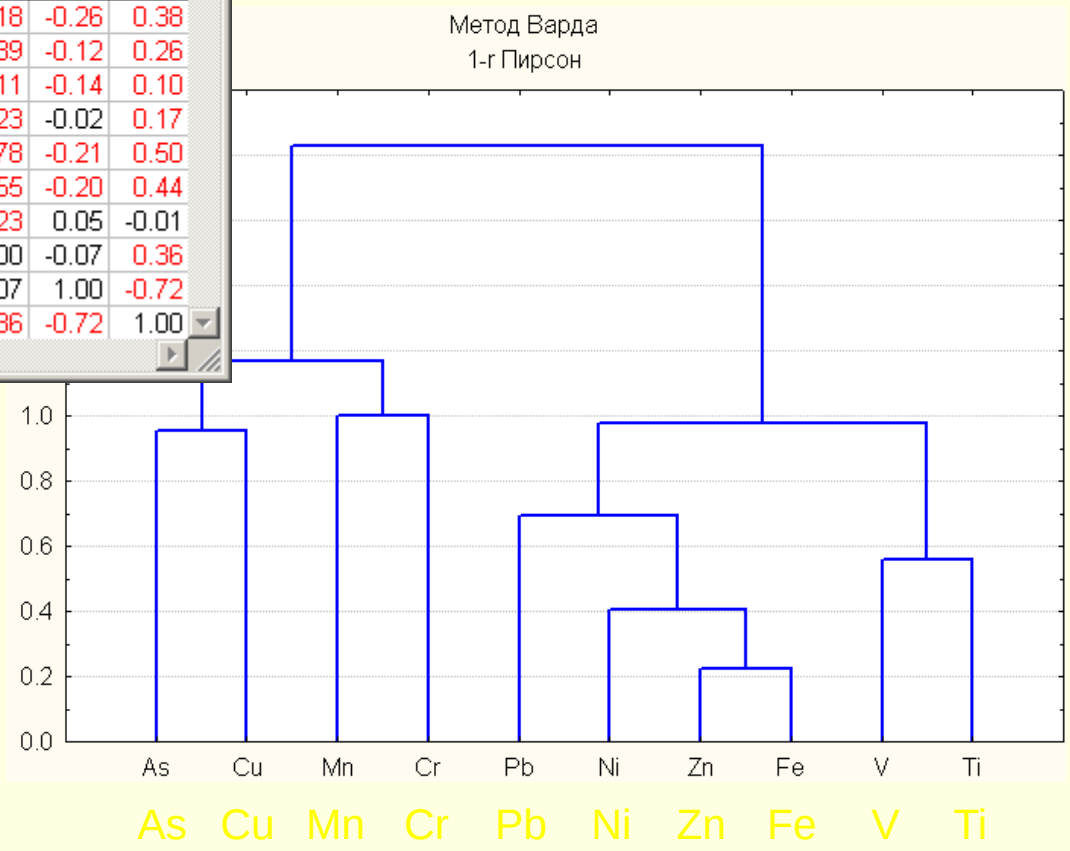


Данные: Корреляци...

Переменная	Fe	Ni	Zn
	Корреляции (Та		
	Отмеченные ко		
	N=500 (Построч		
Fe	1.00	0.73	0.78
Ni	0.73	1.00	0.55
Zn	0.78	0.55	1.00

Данные: Корреляции (Таблица_данных1_(Восстановлен))

Переменная	Ti	V	Cr	Mn	Fe	Ni	Cu	Zn	As	Pb
	Корреляции (Таблица_данных1_(Восстановлен))									
	Отмеченные корреляции значимы на уровне p <.05000									
	N=500 (Построчное удаление ПД)									
Ti	1.00	0.44	0.15	-0.08	0.46	0.20	-0.20	0.18	-0.26	0.38
V	0.44	1.00	-0.05	0.10	0.53	0.22	0.17	0.39	-0.12	0.26
Cr	0.15	-0.05	1.00	0.00	-0.03	0.09	-0.16	-0.11	-0.14	0.10
Mn	-0.08	0.10	0.00	1.00	0.20	0.12	0.03	0.23	-0.02	0.17
Fe	0.46	0.53	-0.03	0.20	1.00	0.73	0.21	0.78	-0.21	0.50
Ni	0.20	0.22	0.09	0.12	0.73	1.00	0.06	0.55	-0.20	0.44
Cu	-0.20	0.17	-0.16	0.03	0.21	0.06	1.00	0.23	0.05	-0.01
Zn	0.18	0.39	-0.11	0.23	0.78	0.55	0.23	1.00	-0.07	0.36
As	-0.26	-0.12	-0.14	-0.02	-0.21	-0.20	0.05	-0.07	1.00	-0.72
Pb	0.38	0.26	0.10	0.17	0.50	0.44	-0.01	0.36	-0.72	1.00



Итерационная кластеризация

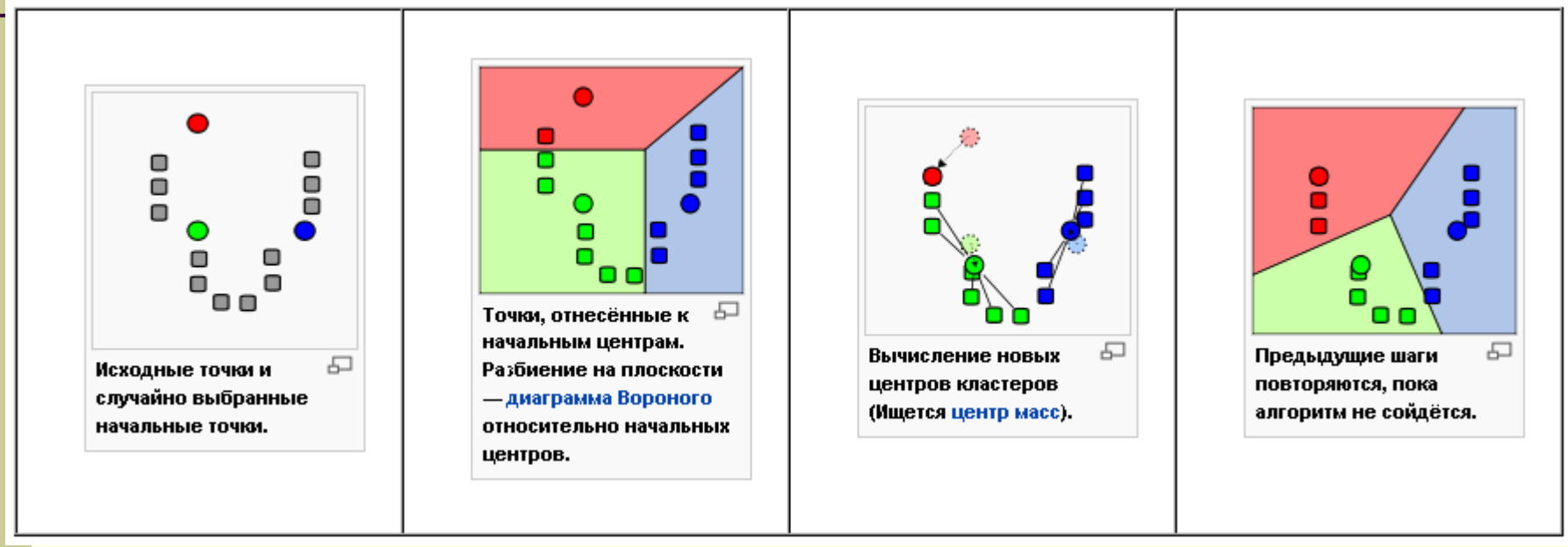
Метод К-средних

Метод К средних строит ровно *k* различных кластеров, расположенных на возможно больших расстояниях друг от друга. Вычисления начинаются с *k* случайно выбранных наблюдений, которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами.

Программа перемещает объекты (т.е. наблюдения) из одних групп (кластеров) в другие для того, чтобы получить наиболее значимый результат в соответствии с критерием дисперсионного анализа (максимальное отношение межгрупповой дисперсии к внутригрупповой).

Значения *F*-статистики, полученные для каждого измерения, является индикатором того, насколько хорошо соответствующая переменная дискриминирует кластеры.

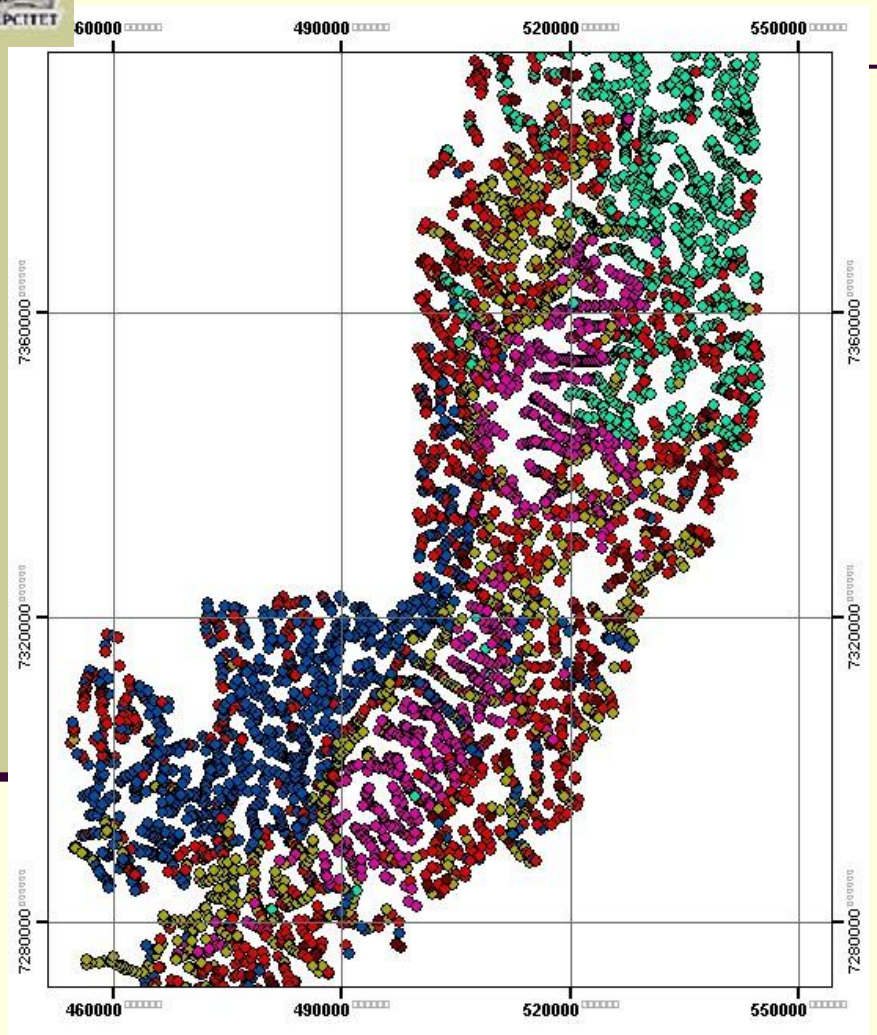
Схема выделения кластеров методом К-средних.

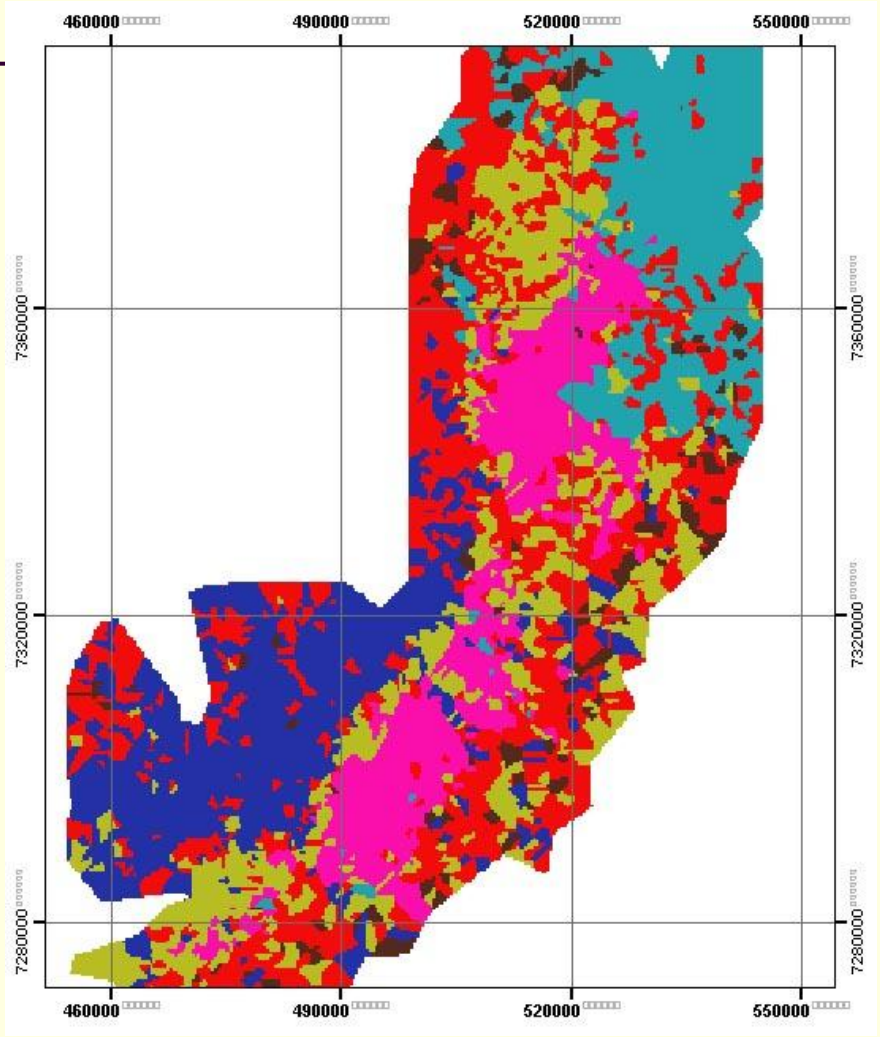


Недостатками метода К-средних являются:

непонятно, как выбирать исходные центры кластеров
число кластеров надо знать заранее

Критерием адекватной кластеризации может служить пространственная обособленность кластеров





Устойчивая типология сохраняется при изменении методов кластеризации. Результаты иерархического кластерного анализа можно проверять итеративным кластерным анализом по методу k-средних.

Если сравниваемые классификации групп респондентов имеют долю совпадений более 70 % (более 2/3 совпадений), то кластерное решение принимается.

Независимо от предмета изучения применение кластерного анализа предполагает *следующие этапы*:

- Отбор выборки для кластеризации.
- Определение множества переменных, по которым будут оцениваться объекты в выборке.
- Вычисление значений той или иной меры сходства между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения.

Кластерный анализ предъявляет следующие *требования к данным*:

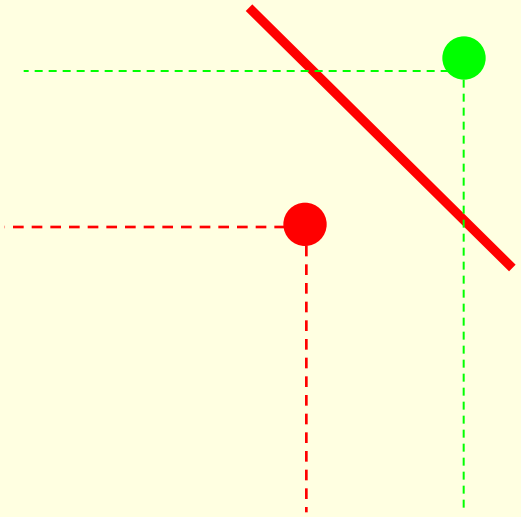
показатели не должны коррелировать между собой;

показатели должны быть безразмерными;

их распределение должно быть близко к нормальному;

показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов;

выборка должна быть однородна, не содержать «выбросов»



Проверка статистической значимости неприменима

Сравнение

Дискриминантный анализ

Есть *исходно существующие группы*. Ищем переменные, которые лучше всего их разделяют.

Кластерный анализ

Есть несколько *переменных*. На их основе классифицируем выборку – проверяем, не объединяются ли наблюдения в группы.

Факторный анализ; многомерное шкалирование

Есть несколько *переменных*. Классифицируем их или уменьшаем их число