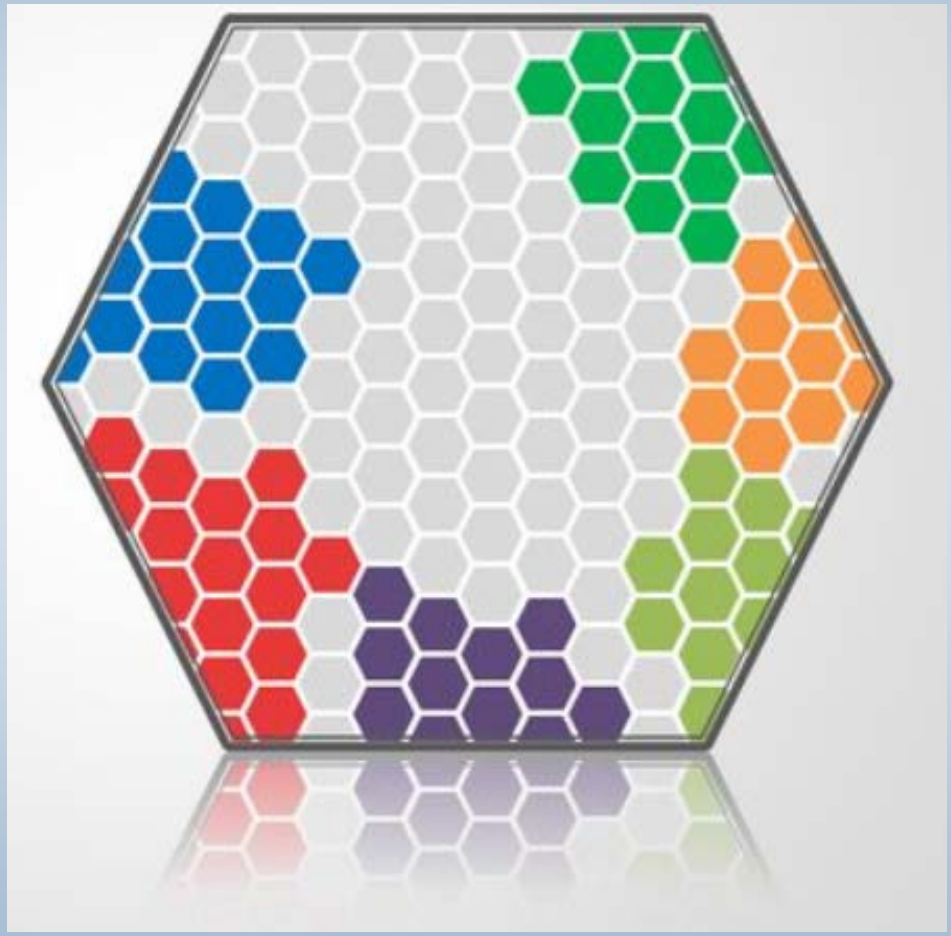
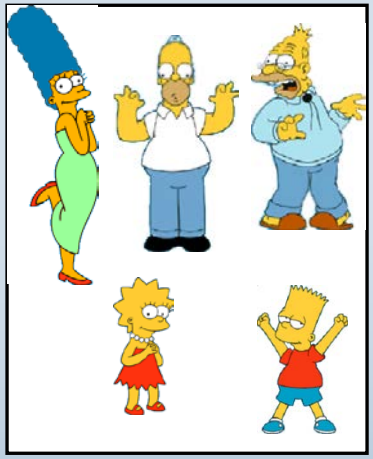
Cluster analysis



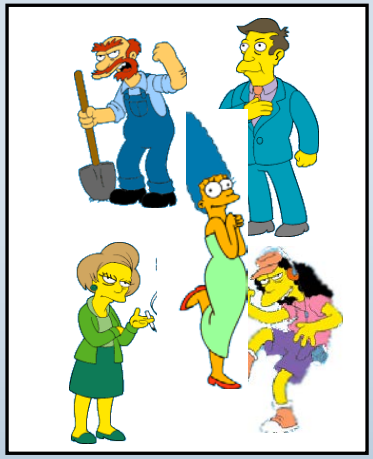
Кольцов С.Н.

Задачи и методы кластер - анализа

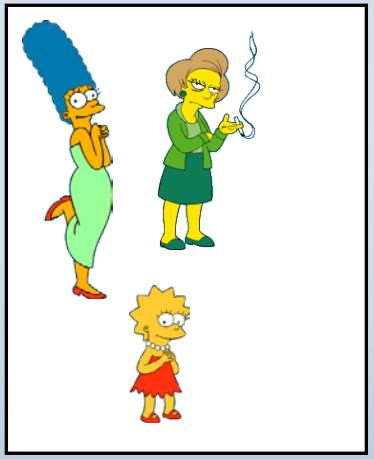
Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров.



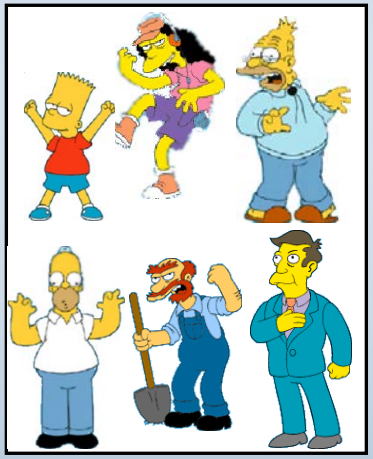
Семья



Сотрудники



Женщины



Мужчины

Лейбелинг групп – то что нужно найти

Кластеризация достаточно субъективна и зависит от цели пользователя

Задачи и методы кластер - анализа

Процедура кластеризации – зависит от меры сходства или не сходства. Такие меры выражаются виде функций расстояний, выраженных в виде той или иной функции.

Сходство тяжело определить

“We know it when we see it”

Задача определения сходства является задачей Machine learning.

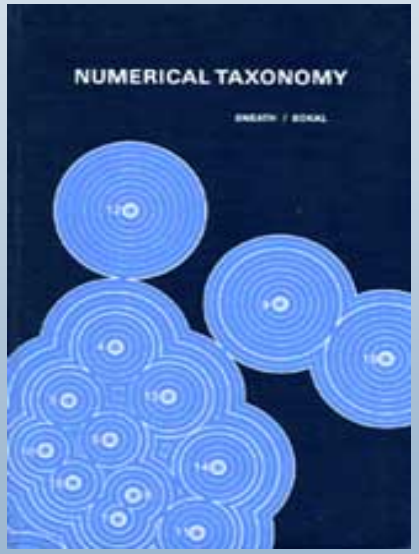
История кластер - анализа

Первые работы, описывающие методы кластерного анализа относятся к концу 30-х годов.

Считается, что термин «кластерный анализ» первым в употребление ввёл американский психолог из университета Беркли Роберт Трайон (Robert C. Tryon) в 1939.

Однако активный интерес к данной теме пришёлся на период 60-80 гг.

Импульсом для разработки многих кластерных методов послужила книга «**Начала численной таксономии**», опубликованная в 1963 г. двумя биологами — **Робертом Сокэлом** и **Петером Снитом**



Современные исследователи: Миркин Б.Г.
МЕТОДЫ КЛАСТЕР-АНАЛИЗА ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ (2011).
Everitt B.S., Landau S., Leese M. et al. (2011) Cluster Analysis. 5th ed. Wiley, 2011
García-Escudero L., Gordaliza A., Matrán C. et al. (2010) A review of robust clustering methods, Advances in Data Analysis and Classification,
4, 2–3: 89–109.

Применение кластер - анализа

1. Статистика
2. Распознавание образов
3. Машинное обучение
4. Финансовая математика
5. Автоматическая классификация в различных областях науки (например, в археологи, биологии (кластеризация видов животных и растений))
6. Маркетинг. Маркетологи выделяют группы с целью оптимизации рекламной деятельности, оптимизации логистической деятельности.
7. Исследование свойств ДНК
8. Страхование (цель выделения групп населения и соотнесение групп с геогрф. расположением, заработком, семейным статусом и другой..)
9. Городское планирование.
10. Финансовое планирование города, района….

Направления в кластер - анализе

- **Partitioning approach**: плоская кластеризация - предполагает разделение объектов на кластеры сразу, причем один объект относится только к одному кластеру.

  Typical methods: **k-means**, k-medoids, CLARANS
- **Fuzzy approach**: Метод нечеткой кластеризации позволяет разбить имеющееся множество объектов p на заданное число нечетких множеств, то есть один и тот же объект может принадлежать разным классам. Принадлежность характеризуется степенью принадлежности, например вероятностью.
- Typical methods:C-means (C-средних)
- **Hierarchical approach**:
  - Восходящая/нисходящая кластеризации: Иерархическая кластеризация (восходящая) - допускаем наличие подкластеров, осуществляется в несколько приемов, в результате образуется в иерархическое дерево (дендрограмму).
  - Typical methods: **Hierarchical**, Diana, Agnes, BIRCH, ROCK, CAMELEON
- **Density-based approach**:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue

Данные в кластер - анализе

Табличные данные:

города	Население т.ч	Плотность 1кв. Км на 1ч	среднемесячная зароботная плата руб	кол. Преступлений на 100 чел.	кол. Образов. Учржд. На 100 чел.
Москва	11514	10588	38410	16	68
Санкт-Петербург	4848	3480	27189	13	72
Новосибирск	1473	2947	23374	29	83
Екатеринбург	1350	2489	23216	33	147
Н. Новгород	1250	3153	21821	35	84
Самара	1164	2152	20690	27	82
Омск	1154	1923	19317	17	86
Казань	1143	1865	19410	20	88
Челябинск	1130	2258	20510	26	87
Ростов на дону	1064	3127	21053	19	78
Уфа	1089	1518	22089	21	93
Волгоград	1021	1791	18294	17	81
Пермь	991	1239	22678	29	93
Красноярск	973	2754	25159	29	86
Воронеж	890	1633	18178	14	87
Саратов	837	2192	18107	18	155
Краснодар	744	991	22587	18	103

Кластеризация может помочь ответить на такие вопросы как:

1. Будут ли группы компаний (кластеры) отражать параметры которые не вошли в кластерный анализ.
2. На основание анализа можно сформулировать правила соотнесения между компаниями, продуктами потребителями и применить эти правила к компаниям, которые не вошли в анализ.

Данные в кластер - анализе

Текстовые данные: Документы представляются в векторной модели, то есть, коллекцию документов предоставляется в виде матрицы термин-документ. Строки будут обозначать отдельные документы (тексты), а колонки – словарь выборки, заключающий в себе все слова в коллекции документов. Кроме того проводится препроцессинг, который заключается в удалении html тэгов, лематизации и удалении стоп слов.

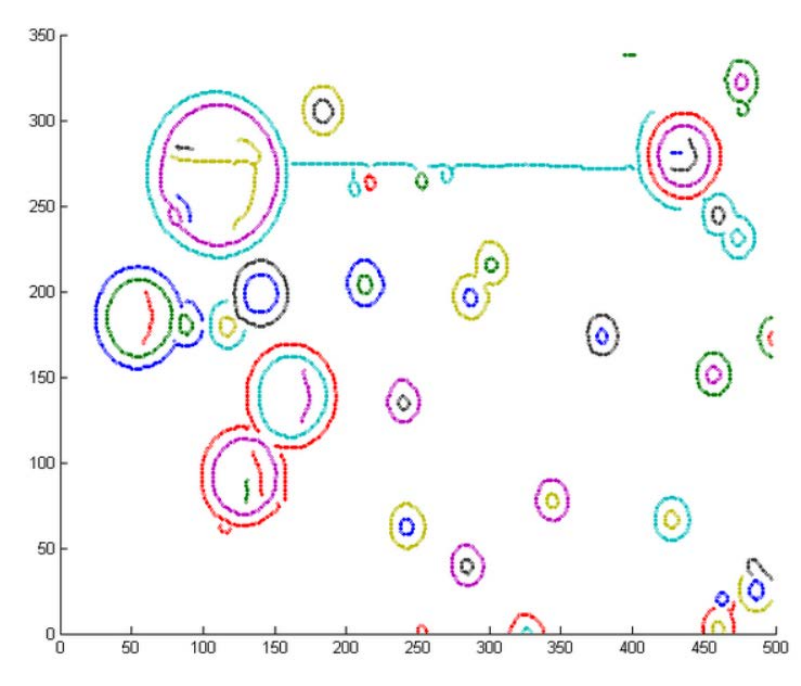
Таблица: матрица термин-документ

	Jane	***likes***	***coffee***	***and***	***tea***	***also***	***cookies***
1ый текст	**1**	**1**	**1**	**1**	**1**	**0**	**0**
2ой текст	**1**	**1**	**0**	**0**	**0**	**1**	**1**

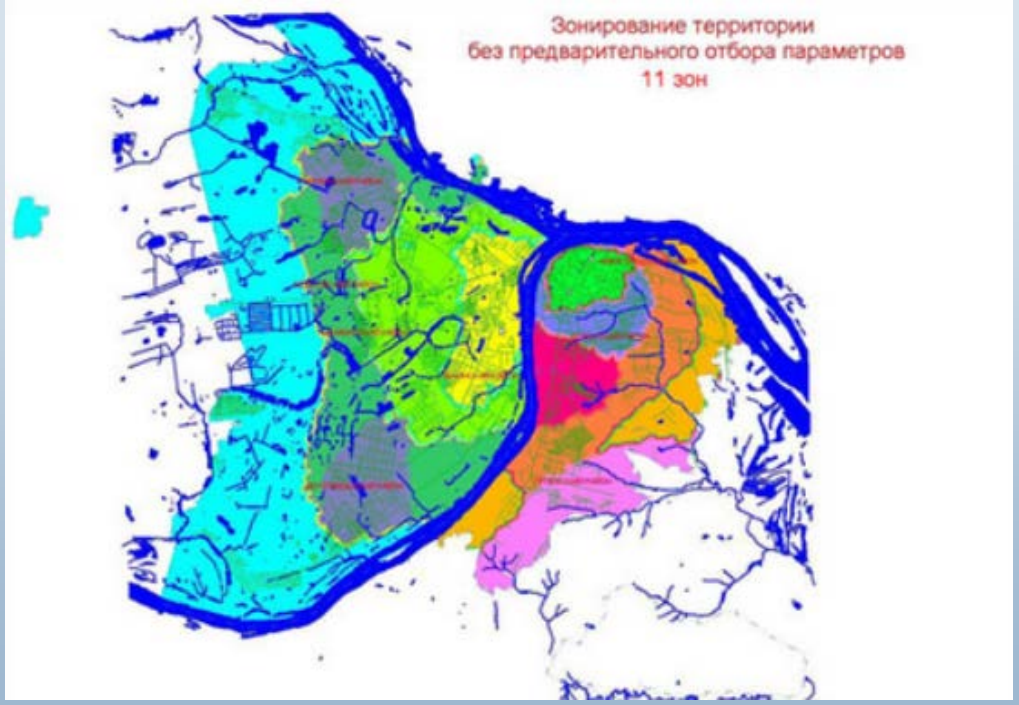
Текстовые данные в виде совокупности бинарных векторов:
d_1 = [1,1,1,1,1,0,0] и d_2 = [1,1,0,0,0,1,1].

Данные в кластер - анализе

Картографические данные: пространственные данные (например, фото, гео - карты) могут быть представлены в виде совокупности точек.



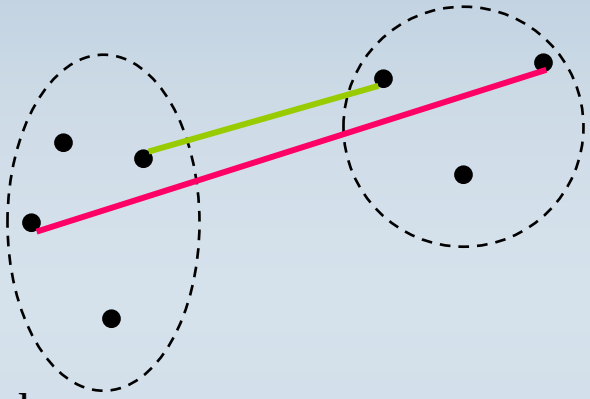
Такие данные часто встречаются при решении проблемы распознавания образов.



Меры близости

Евклидово расстояние - наиболее общий тип расстояния. Является геометрическим расстоянием между точками в многомерном пространстве:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]^{1/2}$$



где: $\mathbf{X_i}$, $\mathbf{X_j}$ - координаты **i**-го и **j**-го объектов в k-мерном пространстве;

$\mathbf{x_{il}}$ - $\mathbf{x_{jl}}$ - величина **l**-той компоненты у **i**-го (**j**-го) объекта (**l=1,2,...,k; i,j=1,2,...,n**).

Квадрат евклидова расстояния - используется, чтобы придать большие веса более отдаленным друг от друга объектам:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]$$

Меры близости

Расстояние city-block (городских кварталов) или манхэттенское расстояние - по сравнением с евклидовым расстоянием влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат:

$$\rho_{ij} = \sum_k | x_{ik} - x_{jk} |$$

где: $\mathbf{X_i}$, $\mathbf{X_j}$ - координаты **i**-го и **j**-го объектов в k-мерном пространстве;
$\mathbf{x_{il}}$ - $\mathbf{x_{jl}}$ - величина **l**-той компоненты у **i**-го (**j**-го) объекта (**l=1,2,...,k; i,j=1,2,...,n**).

Расстояние Минковского (Minkowski Metric)

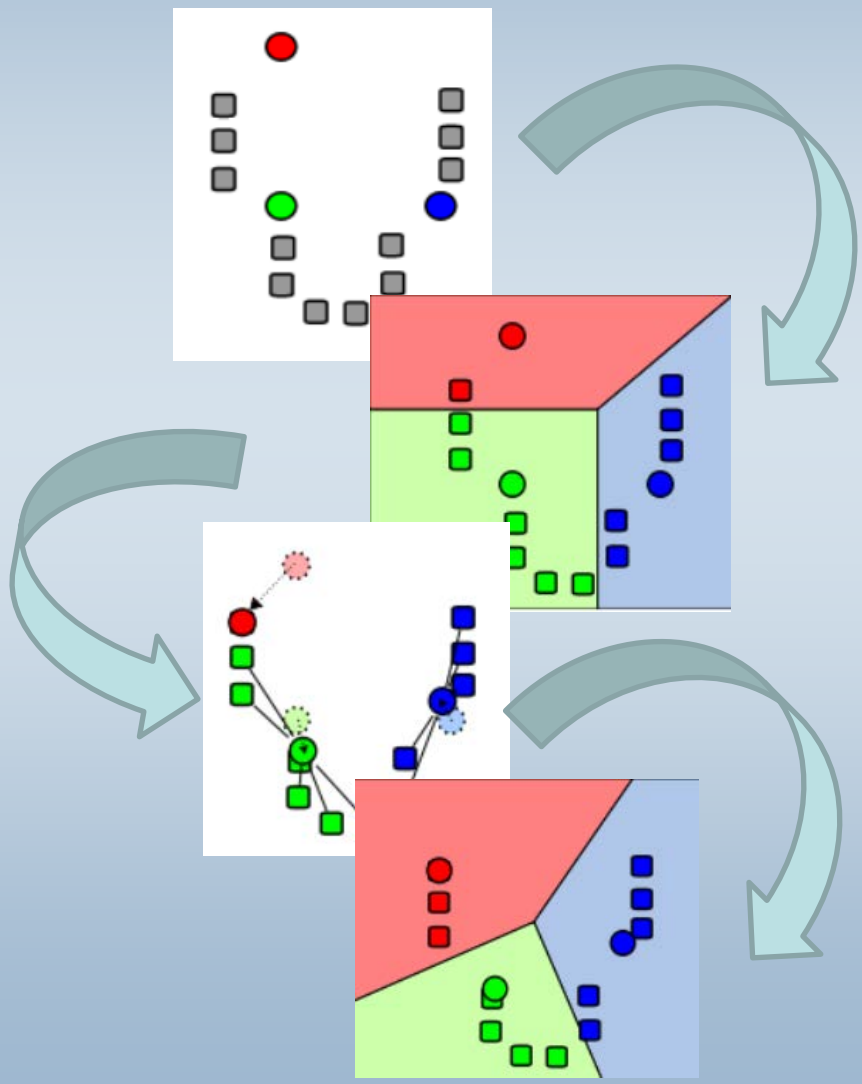
$$\rho_{ij} = \left[\sum_k \left| x_{ik} - x_{jk} \right|^p \right]$$

Если P=1 – расстояние городских кварталов,
Если P=2 – Евклидово расстояние

Алгоритм K means.

Основная суть кластеризации заключается в следующем: Пусть у нас есть совокупность объектов.

1. Выбираем начальные точки для кластеров.
2. Привязать ближайшие точки к центрам кластеров.
3. Пересчитать центры кластеров, исходя из того, что в кластер были добавлены новые объекты.
4. После того как нашли новые центры кластеризации, снова перераспределяем ближайшие точки по кластерам.

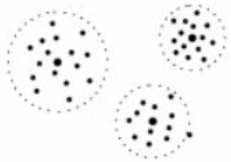


Под центром кластера понимается средне арифметическое значение кластеров.

Виды кластеров.

внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

Разные виды кластеров ведут к проблеме выбора оптимального числа кластеров и проблеме выбора нужного минимума.

кластеры могут соединяться перемычками

кластеры могут накладываться на разрежен-ный фон из редко расположенных объектов

кластеры могут перекрываться

Проблемы K means.

1. Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
3. Число кластеров надо знать заранее.

Как можно преодолеть эти проблемы?

1. Запускать алгоритм много раз (с разными центрами кластеров), после чего выбрать результат с минимальной величиной ошибки.

2. Использовать дополнительные модели дл оценки количества кластеров.

Выбор числа кластеров – проблема остановки расчета

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров y_i объектам x_i , чтобы значение выбранного функционала качества приняло наилучшее значение. Существует много разновидностей функционалов качества кластеризации, но нет «самого правильно го» функционала

Среднее внутрикластерное расстояние должно быть как можно меньше:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min.$$

Среднее межкластерное расстояние должно быть как можно больше:

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max.$$

Выбор числа кластеров – проблема остановки расчета

Если алгоритм кластеризации вычисляет центры кластеров μ_y, $y \in Y$, то можно определить функционалы, вычислительно более эффективные.
Сумма средних внутрикластерных расстояний должна быть как можно меньше:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\, y_i = y} \rho^2(x_i, \mu_y) \to \min,$$

где $K_y = \{xi \in X\ell \mid y_i = y\}$ — кластер с номером y. В этой формуле можно было бы взять не квадраты расстояний, а сами расстояния. Однако, если ρ — евклидова метрика, то внутренняя сумма в Φ_0 приобретает физический смысл момента инерции кластера K_y относительно его центра масс, если рассматривать кластер как материальное тело, состоящее из $|K_y|$ точек одинаковой массы.

Сумма межкластерных расстояний должна быть как можно больше:

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max,$$

Теория скачков - *Алгоритм Сьюгер-Джеймса*

Для решения этой проблемы существует непараметрический метод, предложенный Sugar and James, который позволяет трансформировать функцию качества так, чтобы перегиб или скачок стал явно видимым. Метод основан на использовании понятия «искажений» (**distortion),** которые являются оценками дисперсии внутри класса (кластера).

В ходе реализации данного метода рассчитывается специальная функция (distortion function), которая определяется на основе трех параметров:

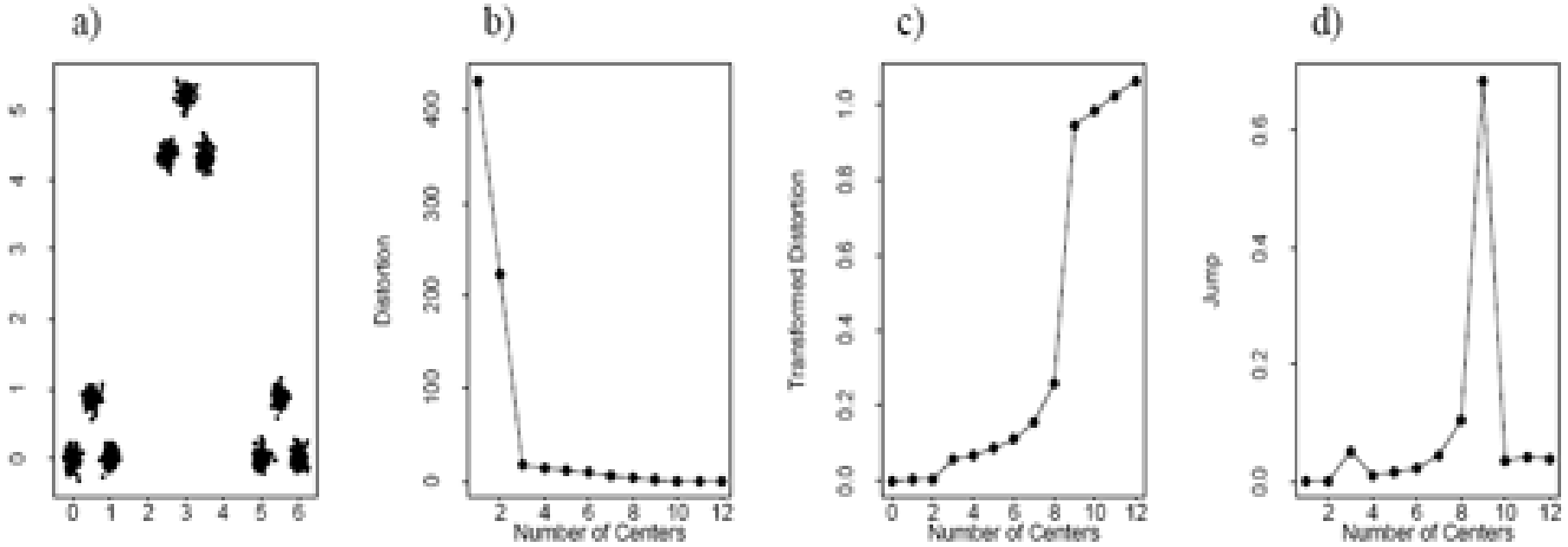
1. **Distortion** для заданного кластерного решения.
2. **Количество кластеров** для заданного кластерного решения.
3. **Коэффициент преобразования**. Затем производится преобразование **distortion** в **transformed distortion**, на основе коэффициента преобразования. В конце производится анализ поведения **transformed distortion function** в зависимости от числа кластеров. На основании анализа делается вывод о том, какое кластерное решение наилучшее.

Теория скачков - *Алгоритм Сьюгер-Джеймса*

1. **Определение минимального искажения (distortion).** В качестве минимального искажения берется минимальное значение внутрикластерной дисперсии, встречающееся в данном кластерном решении. Это значит следующее: для заданного кластерного решения (например, для 5 кластеров) рассчитываются дисперсии внутри каждого кластера (среднее внутрикластерное расстояние). Из этого множества чисел (5 штук в 5-кластерном решении) выбирается минимальное значение d.
2. **Коэффициент трансформации**. Согласно разработчикам метода, в качестве коэффициента трансформации можно взять величину Y=P/2, где P – размерность векторного пространства. В качестве коэффициента также можно взять величину 1/K, где K – число кластеров.
3. **Transformed distortion**. Данная величина рассчитывается следующим образом:

$$D_t(K) = d^Y(K)$$

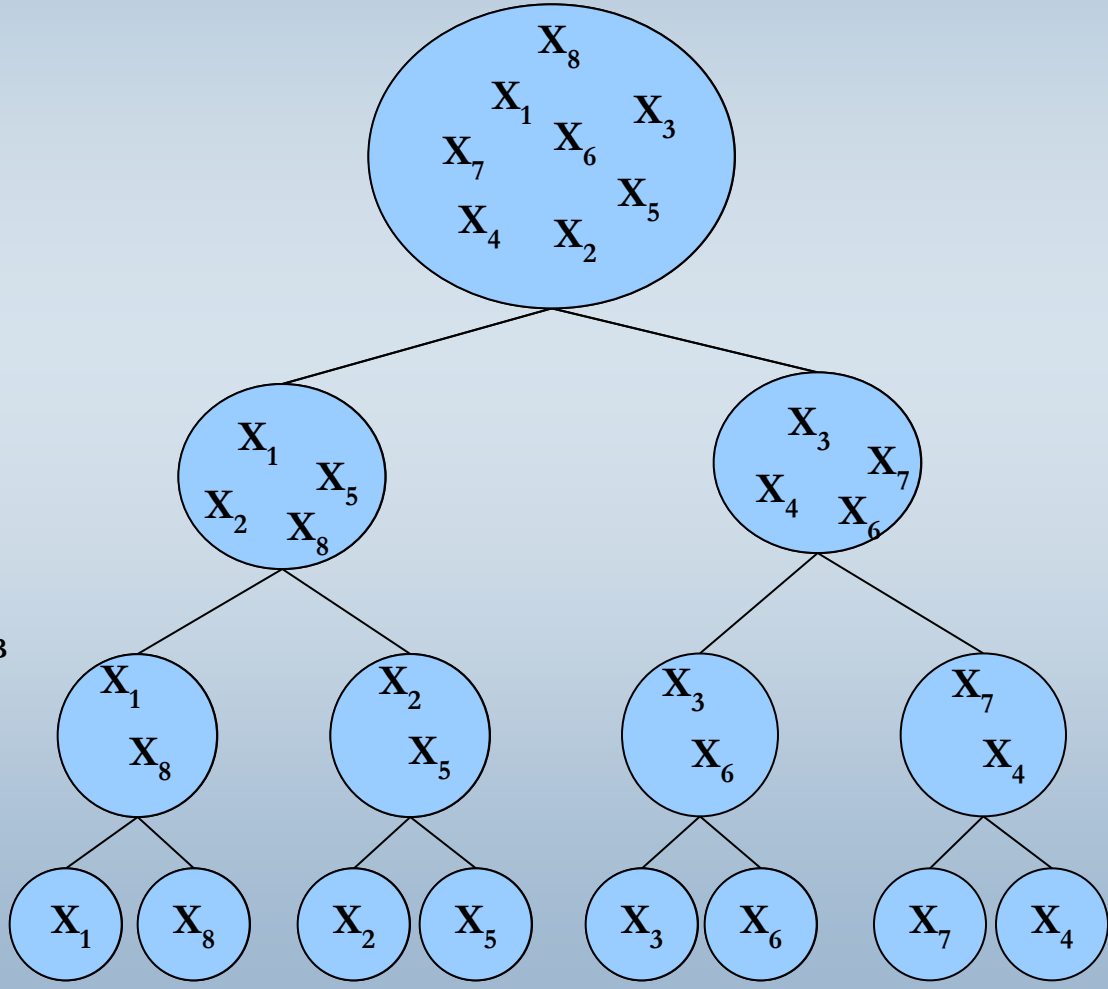
Теория скачков - *Алгоритм Сьюгер-Джеймса*



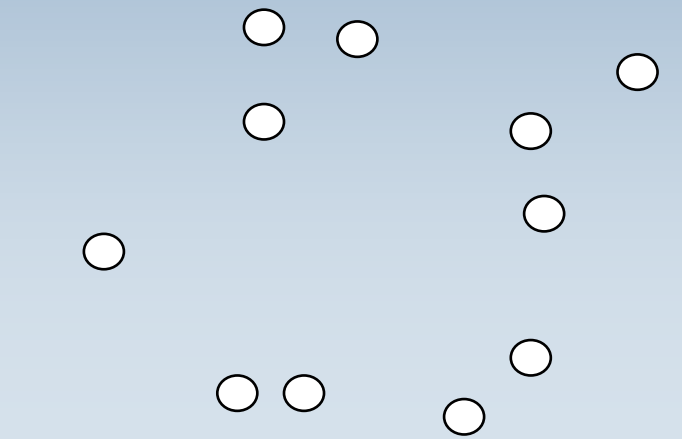
Восходящая / нисходящая кластеризации

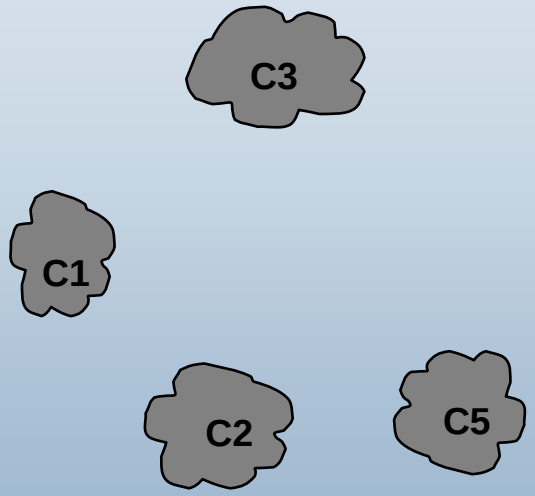
Восходящая кластеризация (agglomerative): В рамках данного алгоритма предполагается что каждый элемент нашего множества является отдельным кластером. Процесс образования новых кластеров заключается в объединение некоторых кластеров в один новый кластер. Объединение осуществляется на основе заданного расстояния между кластерами. Производя такое итеративное объединение мы получаем дерево кластеров, которое в итоге сходится к одному кластеру.

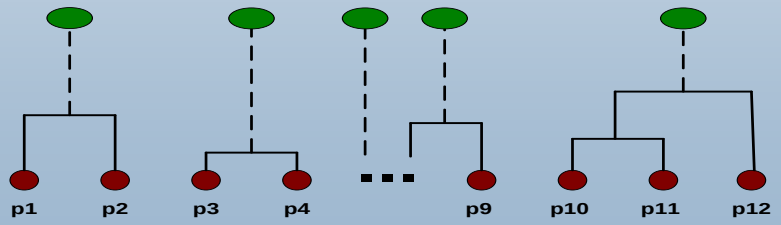
Нисходящая кластеризация (divisive): Данный вид кластеризация заключатся в следующем. Мы предполагаем, что все объекты принадлежат одному кластеру. В ходе итеративного процесса мы разделяем кластеры на несколько разных кластеров. Соответственно при этом, получаем дерево кластеров (дендрограмма).



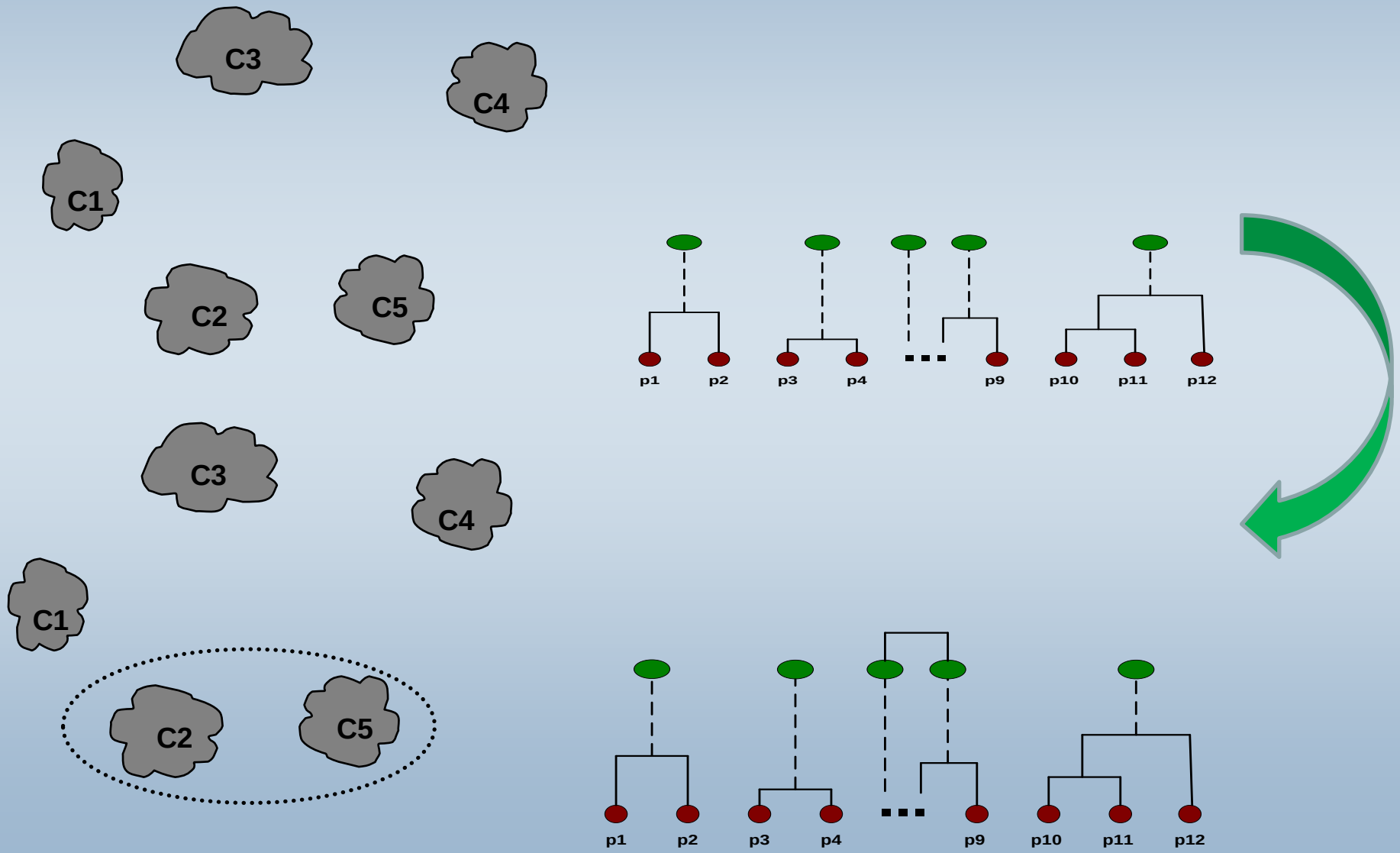
Метод средней связи Кинга



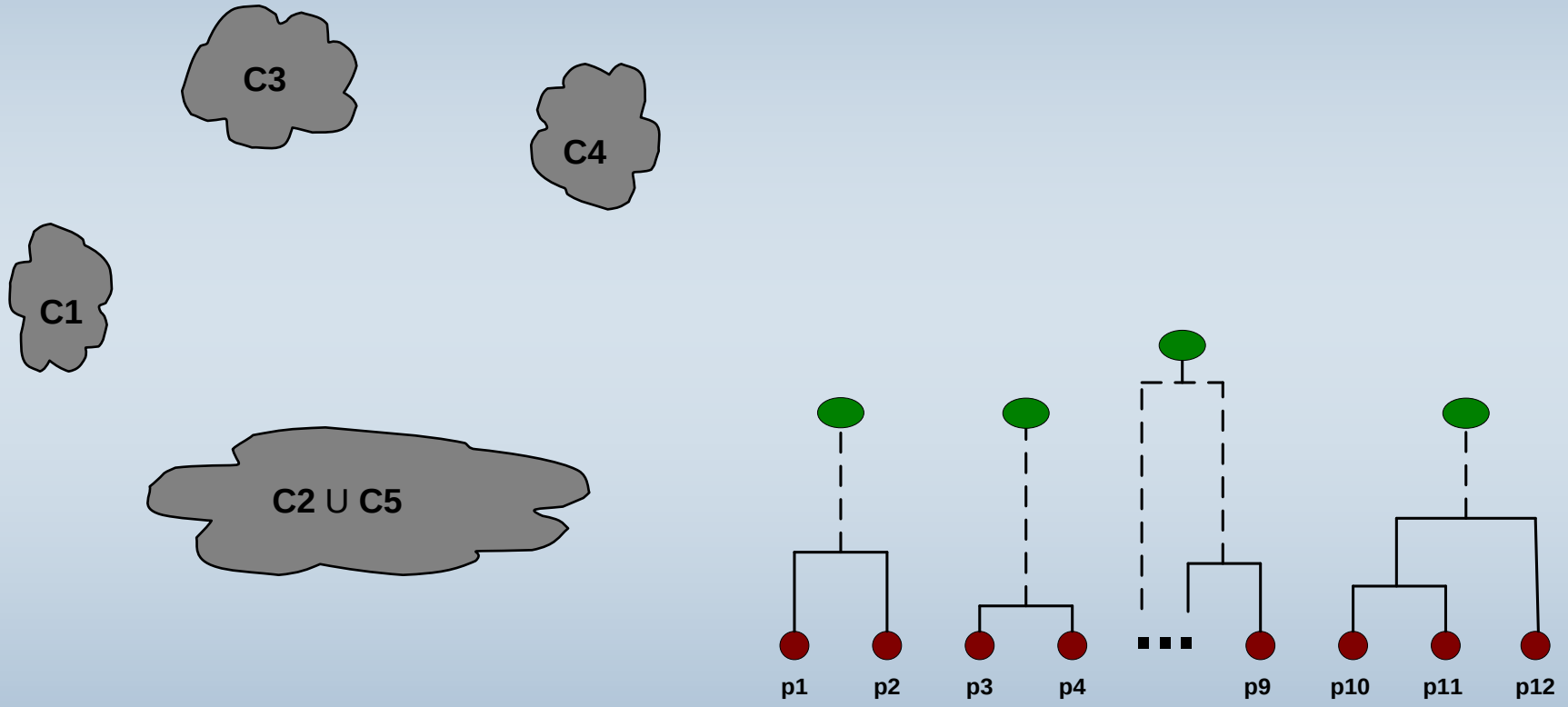




Метод средней связи Кинга



Метод средней связи Кинга



Меры качества кластерного анализа

Можно выделит два типа меры качества:

1. **Внешняя мера качества (External Measures)**: Внешние меры основаны на сравнении автоматического разбиения данных с полученным от экспертов «эталонным» разбиением этих же данных. Кроме того, в качестве эталона может использоваться математическая величина, которая выбирается на основе каких то теоретических рассуждений.

Энтропия одного кластера:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r},$$

Энтропия кластерного решения:

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r).$$

$\mathbf{n_r}$ – число элементов в заданном кластере,
q – число классов во всей коллекции,
$\mathbf{n_r^i}$ – число элементов i-того класса внутри кластера r. **Очевидно**, что если в кластере все документы относятся к одному кластеру, то log(1)=0, то есть мы достигаем минимальное значение энтропии.

Таким образом, хаос (то есть отсутствие кластеров) соответствует максимальной величине энтропии, хорошее кластерное решение соответствует минимуму энтропии.

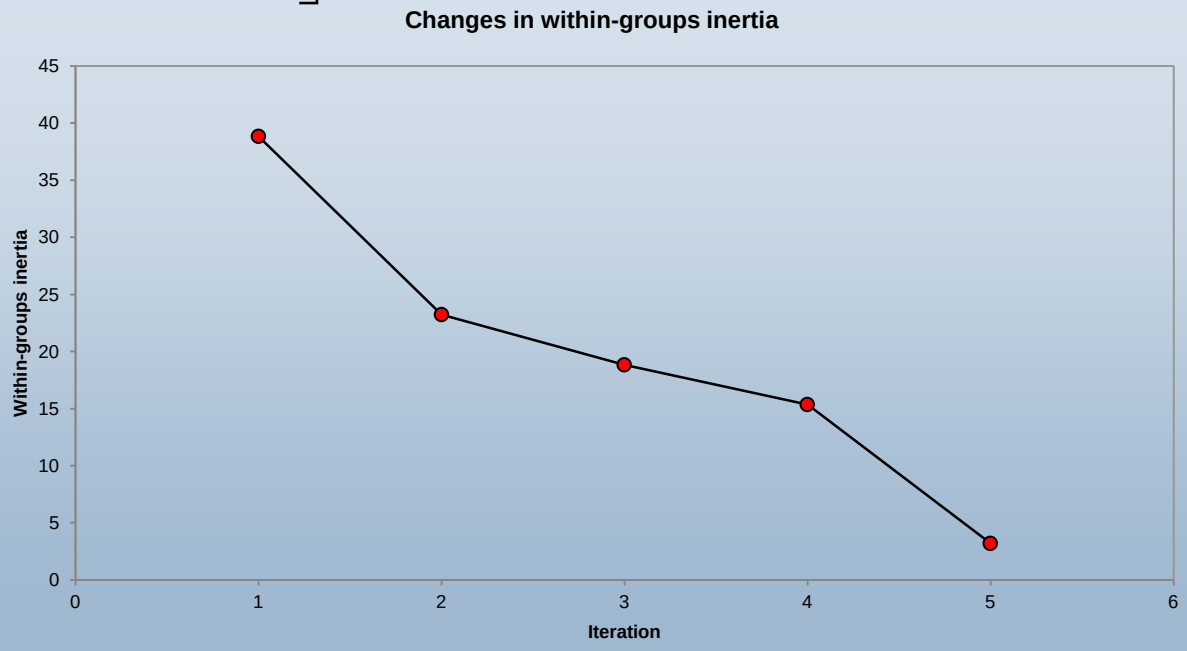
Меры качества кластерного анализа

2. **Внутренняя Measures**: Внутренние меры основаны на оценке свойств отделимости и компактности полученного разбиения данных.

Например, в качестве такой меры, например, используется функция суммы квадратов отклонений объектов от центра кластеров (**inertia**).

$$\rho_{ij} = \left[\sum_k (x_{ik} - A_k)^2 \right]$$

где A_{jk} центр к кластера. В ходе итераций ищется минимальное значение функции ρ_{ij}

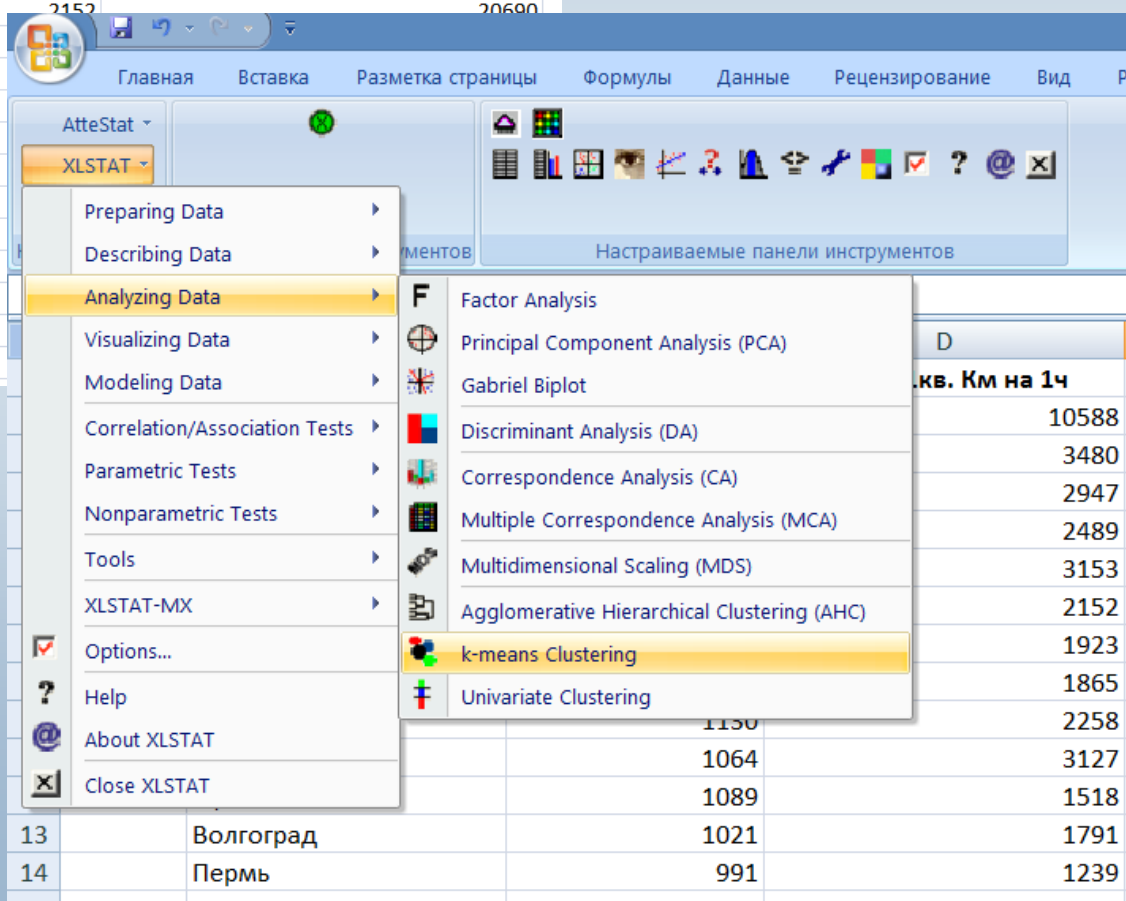


Кластеризация в пакете XlStat

Загрузка данных в Эксель

города	Население т.ч	Плотность 1кв. Км на 1ч	среднемесячная зароботная плата руб	к
Москва	11514	10588	38410	
Санкт-Петербург	4848	3480	27189	
Новосибирск	1473	2947	23374	
Екатеринбург	1350	2489	23216	
Н. Новгород	1250	3153	21821	
Самара	1164	2152	20690	
Омск	1154			
Казань	1143			
Челябинск	1130			
Ростов на дону	1064			
Уфа	1089			
Волгоград	1021			
Пермь	991			
Красноярск	973			
Воронеж	890			
Саратов	837			
Краснодар	744			

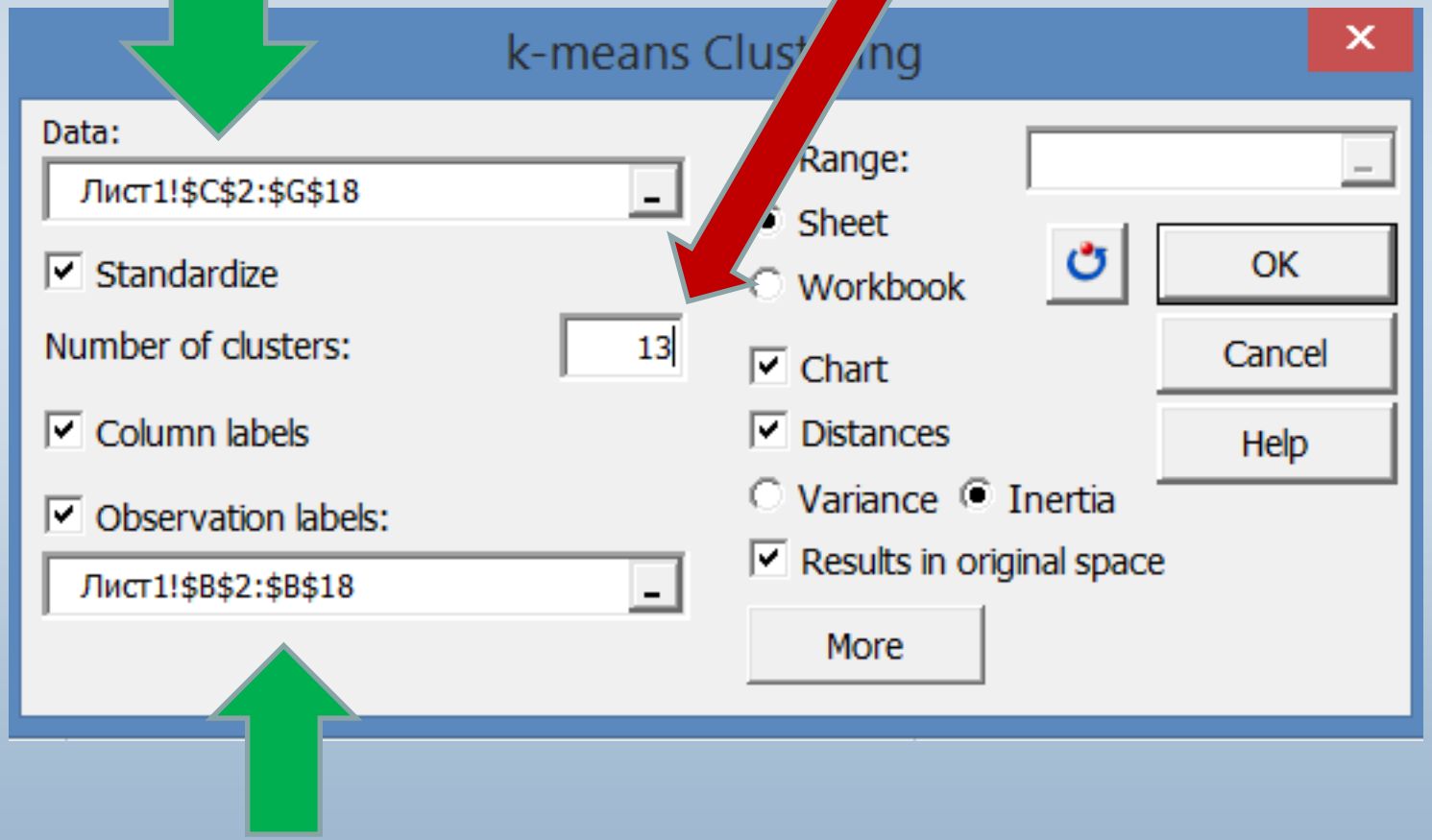
Запуска процедуры кластеризации



Кластеризация в пакете XlStat

Исходные главные для кластеризации

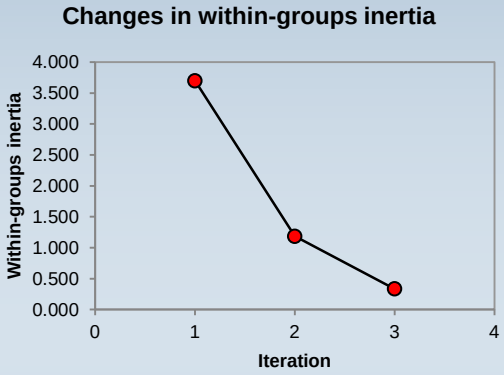
Число кластеров



Имена элементов, которые учувствуют в кластеризации

Результаты кластеризации в пакете XlStat

Сумма квадратов отклонений объектов от центра кластеров



Clusters composition:

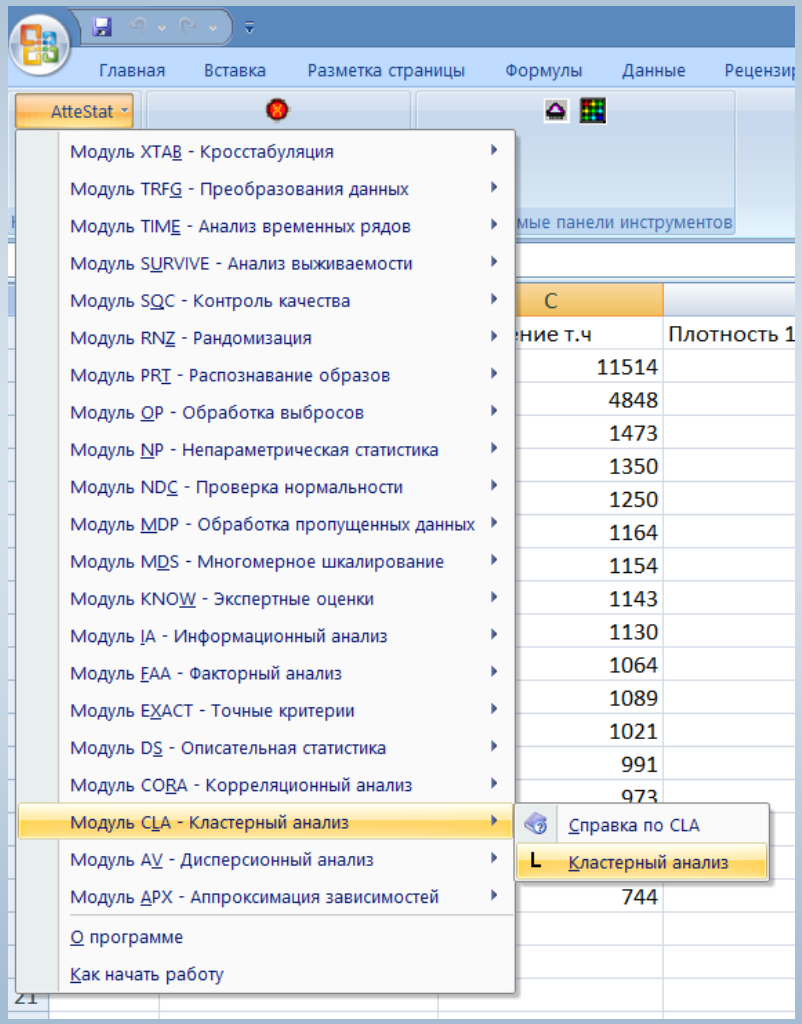
Cluster	1	2	3	4	5	6
Within-groups inertia	6073.500	27813.000	22409.000	0.000	0.000	0.000
Minimum distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Average distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Maximum distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Size	2	2	2	1	1	1
	Омск Казань	Волгоград Воронеж	Самара Челябинск	Ростов на дону	Новосибирск	Уфа

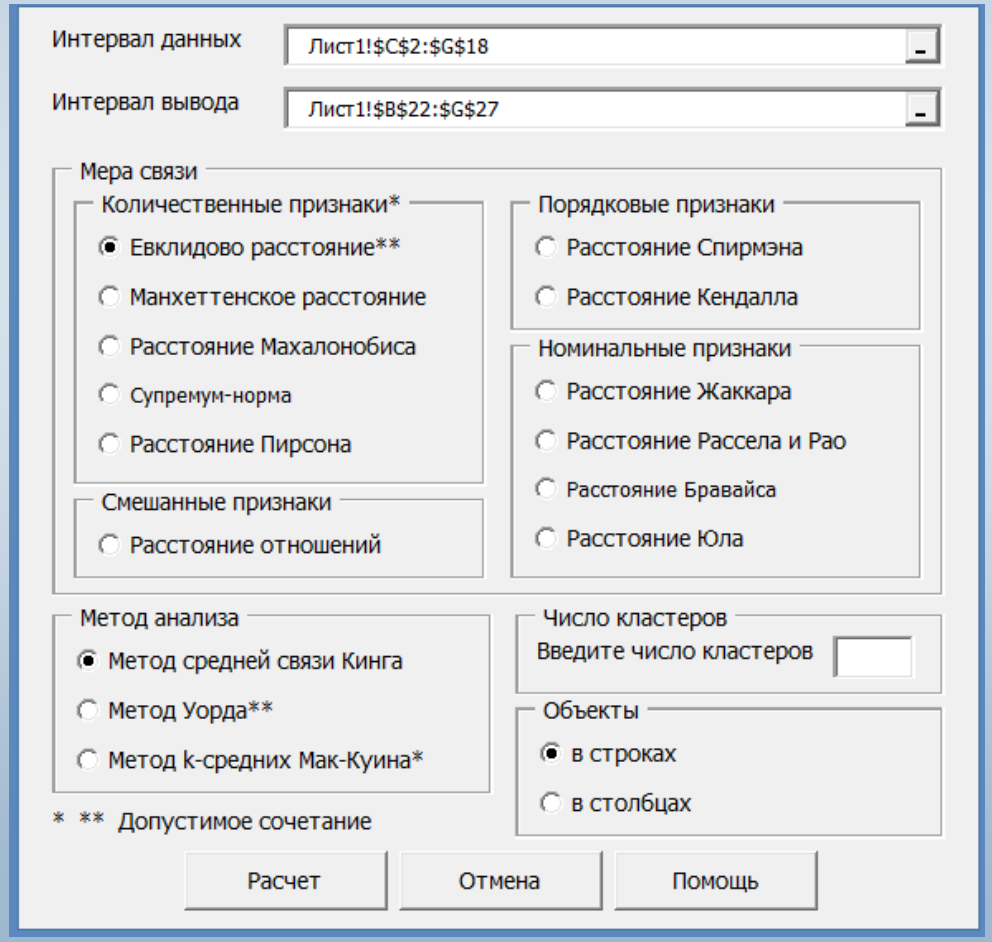
Результаты второй кластеризации в пакете XlStat

Clusters composition:

Cluster	1	2	3	4	5	6	7
Within-groups inertia	27813	1736742	22409	0	0	0	0
Minimum distance from centroid	117.926	931.864	105.851	0.000	0.000	0.000	0.000
Average distance from centroid	117.926	931.864	105.851	0.000	0.000	0.000	0.000
Maximum distance from centroid	117.926	931.864	105.851	0.000	0.000	0.000	0.000
Size	2	2	2	1	1	1	1
	Волгоград Воронеж	Новосибирск Красноярск	Самара Челябинск	Уфа	Саратов	Пермь	Н. Новгород

Кластеризация в пакете AttStat





Методы анализа в пакете AttStat.

Метод средней связи Кинга (иерархический метод).

Процесс классификации состоит из элементарных шагов:

• Поиск и объединение двух наиболее похожих объектов в матрице сходства.

• Основанием для помещения объекта в кластер является близость двух объектов, в зависимости от меры сходства.

• На каком–либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.

• На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

Метод оперирует не исходными объектами, а построенной матрицей сходства, по определению являющейся количественной. Координаты центра тяжести кластера вычисляются не по исходным данным – они являются продуктом манипуляций с матрицей сходства.

Матрица сходства — это квадратная матрица типа «объект-объект» (порядка n) содержащая в качестве элементов расстояния между объектами

$$\begin{bmatrix} m_{11} & \cdots & m_{1j} & \cdots & m_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ m_{i1} & \cdots & m_{ij} & \cdots & m_{in} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ m_{n1} & \cdots & m_{nj} & \cdots & m_{nn} \end{bmatrix}$$

Результат кластеризации в пакете AttStat

Метод Уорда (иерархический метод)

Метод Уорда (Ward) является одним из иерархических агломеративных методов кластерного анализа. Процесс классификации состоит из элементарных шагов:

1. Поиск и объединение двух наиболее похожих объектов в матрице сходства.
2. Основанием для помещения объекта в кластер является минимум дисперсии внутри кластера при помещении в него текущего классифицируемого объекта.
3. На каком–либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.
4. На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

В качестве меры различия для метода Уорда используется только евклидово расстояние. Этим фактом вызвано ограничение области применения программы только количественной шкалой.

Дисперсия:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\sigma = \sqrt{D}$$

D - дисперсия

Результат кластеризации в пакете AttStat

Метод k–средних Мак–Куина

Принцип классификации сводится к следующим элементарным шагам:

1. Некоторое, возможно, случайное, исходное разбиение множества объектов на заданное число кластеров (классов, групп, популяций). Расчет «центров тяжести» кластеров.
2. Отнесение остальных объектов к ближайшим кластерам.
3. Пересчет новых «центров тяжести» кластеров.
4. Переход к шагу 2, пока новые «центры тяжести» кластеров не перестанут отличаться от старых.
5. Получено оптимальное разбиение.

В качестве меры различия для метода средней связи используется любая из представленных в программе мер, предназначенных для количественных данных.

Результат кластеризации в пакете AttStat

Номер кластера, объекты	численность		
1	1	1	
2	1	2	
3	1	3	
4	1	4	
5	1	5	
6	2	6	9
7	2	7	8
8	1	10	
9	1	11	
10	2	12	15
11	2	13	17
12	1	14	
13	1	16	

Номер кластера,	численность		объекты
1	1	Москва	
2	1	Санкт-Петербург	
3	1	Новосибирск	
4	1	Екатеринбург	
5	1	Н. Новгород	
6	2	Самара	Челябинск
7	2	Омск	Казань
8	1	Ростов на дону	
9	1	Уфа	
10	2	Волгоград	Воронеж
11	2	Пермь	Краснодар
12	1	Красноярск	
13	1	Саратов	