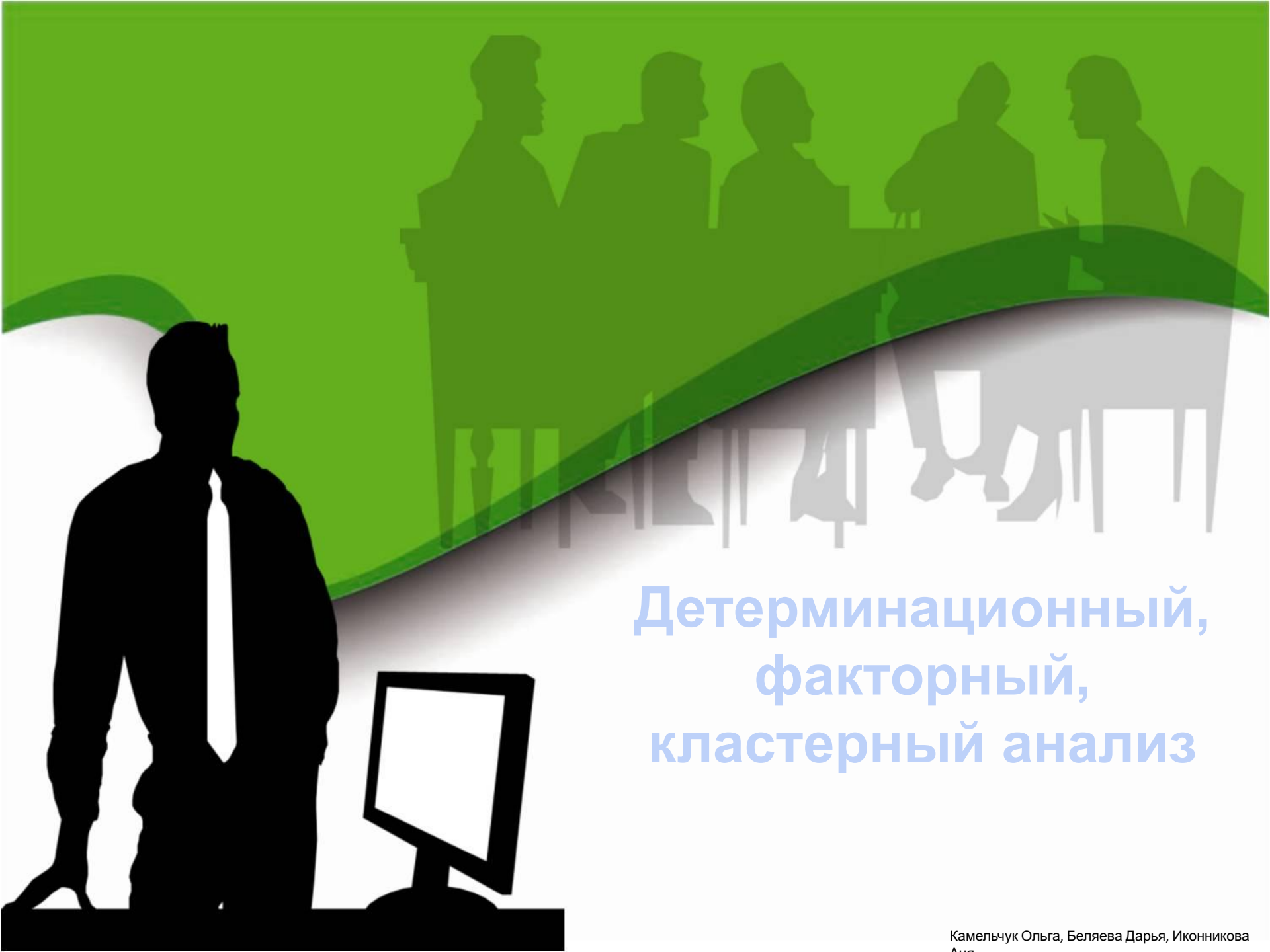


Детерминационный, факторный, кластерный анализ

Камельчук Ольга, Беляева Дарья, Иконникова Аня

Детерминационный анализ (теория правил) — это, с одной стороны, математическая теория детерминаций, а с другой — практический метод анализа правил, который позволяет искать и анализировать правила, обрабатывая данные опыта.

Термин «детерминация» происходит от латинского determinatio — определение, ограничение.

Правило — это особый математический объект, представляющий суждение вида «Если , то » , где , — соответственно, объясняющий и объясняемый признаки.

Правило как детерминация — это условное суждение вида: Если a, то b вместе с двумя своими характеристиками: **точностью** и **полнотой**. Признак **a** называется **объясняющим**. Признак b называется **объясняемым**.

Точность правила — это доля случаев, когда правило подтверждается, среди всех случаев его применения (доля случаев b среди a случаев).

$$\text{Точность правила } \{\text{Если } a, \text{ то } b\} = N(a,b)/N(a) = P(b|a)$$

Полнота правила — это доля случаев, когда правило подтверждается, среди всех случаев, когда имеет место объясняемый признак (доля случаев a среди случаев b).

$$\text{Полнота правила } \{\text{Если } a, \text{ то } b\} = N(a,b)/N(b) = P(a|b)$$

$$\text{Точность правила } \{\text{Если } a, \text{ то } b\} = \text{Полнота правила } \{\text{Если } b, \text{ то } a\}$$

$$\text{Полнота правила } \{\text{Если } a, \text{ то } b\} = \text{Точность правила } \{\text{Если } b, \text{ то } a\}$$

a	b	«Если a, то b»
Истина	Истина	**Истина**
Истина	Ложь	**Ложь**
Ложь	Истина	**Истина**
Ложь	Ложь	**Истина**

Примеры правил:

Примеры иллюстрируют правила вида «Если , то » с различным содержанием признаков и . Приведенные примеры демонстрируют четыре правила со значениями точности и полноты, близкими или равными единице либо нулю:1) точное, но неполное;2) неточное, но полное;3) точное и полное;4) неточное и неполное.

Пример 1. Правило точное и полное: *В прямоугольном треугольнике из трех углов имеется два, сумма которых составляет прямой угол (= «прямоугольный треугольник», = «в треугольнике из трех углов имеется два, сумма которых составляет прямой угол)».*

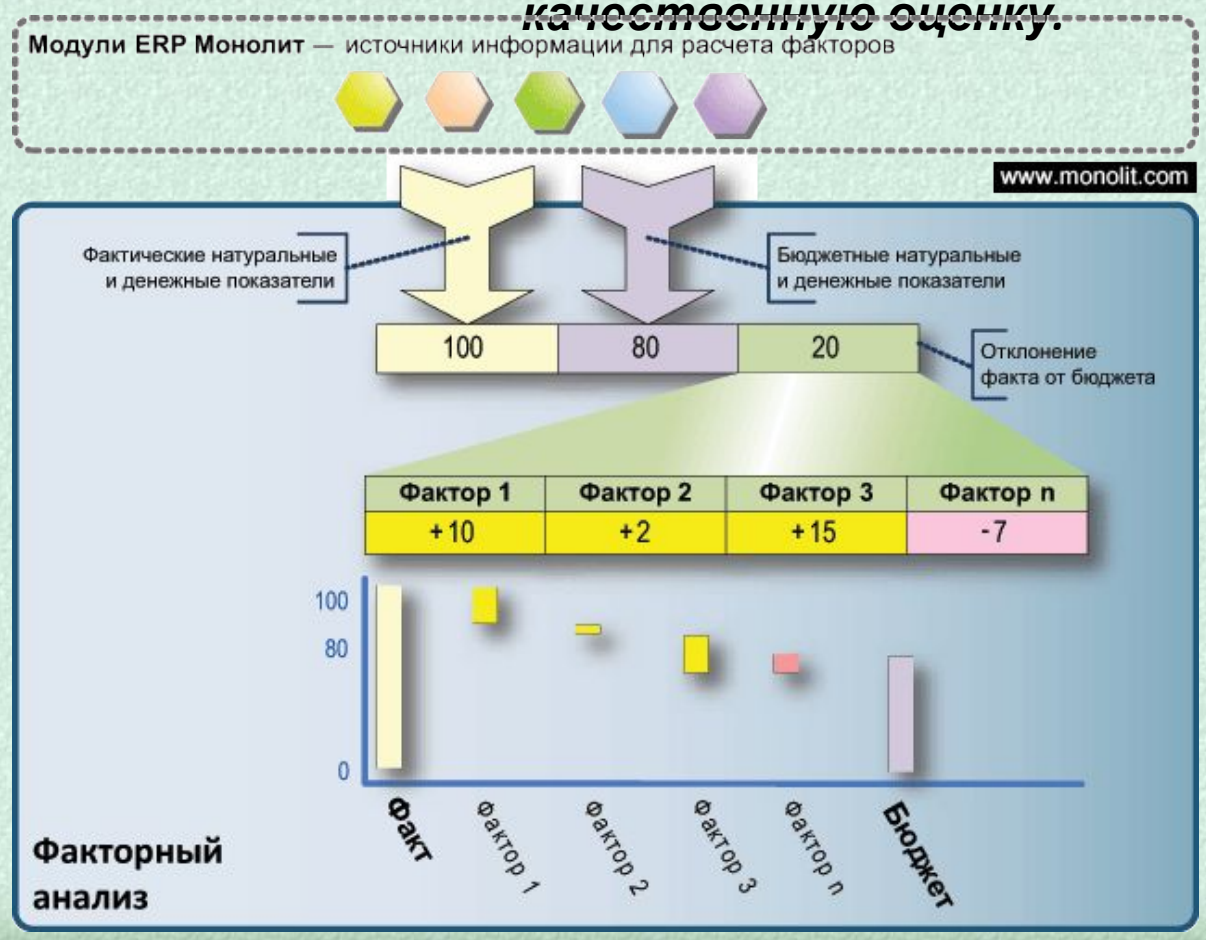
В мире не слишком больших масштабов, где справедлива геометрия Евклида, это правило имеет точность, равную единице (среди прямоугольных треугольников все обладают свойством). Полнота правила также равна единице (среди треугольников, которые обладают свойством , все прямоугольные).

Пример 2. Правило неточное и неполное: *Если у человека родинка на щеке, то он альбинос*

(= «человек имеет родинку на щеке», = «альбинос»).

Среди людей, у которых родинка на щеке, доля альбиносов заведомо невелика. Среди альбиносов также, по всей видимости, не так много имеют родинку на щеке. Это означает, что и точность и полнота такого правила будут значительно меньше единицы.

ФАКТОРНЫЙ АНАЛИЗ - ***это методика комплексного и системного изучения и измерения воздействия факторов на величину результативного показателя. Факторы в результате анализа получают количественную и качественную оценку.***



Различают следующие противоположные типы факторного анализа:

Детерминированный факторный анализ - представляет собой методику исследования влияния факторов, связь которых с результативным показателем носит функциональный характер, т.е. когда результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов.

Стохастический факторный анализ - это методика исследования влияния факторов, связь которых с результатом является неполной. Носит характер вероятностной, корреляционной зависимости, поскольку изменение фактора может дать несколько значений результата в зависимости от сочетания других факторов.

Прямой факторный анализ - ведется дедуктивным способом - от общего к частному.

Обратный факторный анализ - осуществляет исследование причинно-следственных связей способом логической индукции - от частных, отдельных факторов к обобщающим, от причин к следствиям с целью установления чувствительности изменения многих результативных показателей к изменению изучаемого фактора.

Факторный анализ может быть одноуровневым и многоуровневым.

Одноуровневый факторный анализ - используется для исследования факторов только одного уровня (одной ступени) подчинения без их детализации на составные части.

Многоуровневый, многоступенчатый факторный анализ - проводит детализацию факторов a и b на составные элементы с целью изучения их сущности.

Статический факторный анализ - применяется при изучении влияния факторов на результативные показатели на соответствующую дату.

Динамический факторный анализ - представляет собой методику исследования причинно-следственных связей в динамике.

Ретроспективный факторный анализ - изучает причины изменения результатов хозяйственной деятельности за прошлые периоды.

Перспективный факторный анализ - исследует поведение факторов и результативных показателей в перспективе.

Основные задачи факторного анализа:

- Выявление, поиск факторов.
- Отбор факторов для анализа исследуемых показателей.
- Классификация и систематизация их с целью обеспечения системного подхода.
- Моделирование взаимосвязей между результативными и факторными показателями.
- Расчет влияния факторов и оценка роли каждого из них в изменении величины результативного показателя.
- Работа с факторной моделью (практическое ее использование для управления экономическими процессами).

Коэффициент взаимосвязи между некоторой переменной и общим фактором, выражающий меру влияния фактора на признак, называется факторной нагрузкой данной переменной по данному общему фактору. Значение фактора у отдельного объекта называется факторным весом объекта по данному фактору.

Процесс стохастического факторного анализа состоит из трех больших этапов:

1- Подготовки ковариационной матрицы;

2- Выделения первоначальных ортогональных векторов;

3- Вращение с целью получения окончательного решения.

Подготовка к факторному анализу:

Ковариация двух векторов:

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{M}((\mathbf{X} - \mathbb{M}(\mathbf{X}))(\mathbf{Y} - \mathbb{M}(\mathbf{Y}))),$$

$\mathbb{M}()$ – математическое ожидание $\mathbb{M}(X) = \sum_i x_i p_i, P(X = x_i) = p_i, \sum_i p_i = 1$

Корреляция двух векторов:

$$\varrho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathrm{D}[X]} \cdot \sqrt{\mathrm{D}[Y]}},$$

$D[X] = M\left[(X - M[X])^2\right]$, - дисперсия.

Условия применения факторного анализа:

- все признаки должны быть количественными;

- число наблюдений должно быть не менее чем в два раза больше числа переменных;

- выборка должна быть однородна;

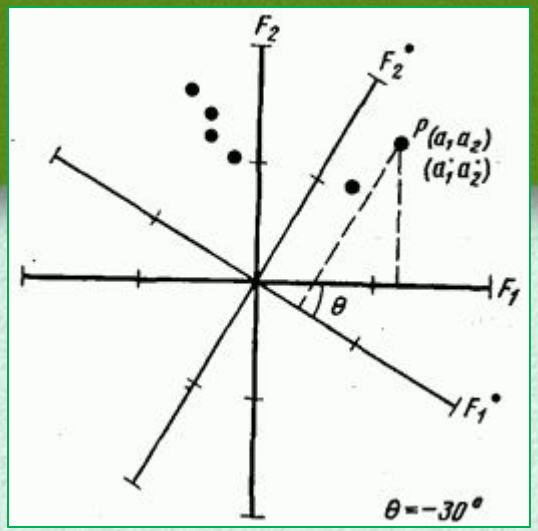
- исходные переменные должны быть распределены симметрично;

- факторный анализ осуществляется по коррелирующим переменным.

Процедура вращения. Выделение и интерпретация факторов

Простая структура соответствует требованиям:

- В каждой строке матрицы вторичной структуры V должен быть хотя бы один нулевой элемент;
- Для каждого столбца k матрицы вторичной структуры V должно существовать подмножество из r линейно-независимых наблюдаемых переменных, корреляции которых с k-м вторичным фактором — нулевые.
- У одного из столбцов каждой пары столбцов матрицы V должно быть несколько нулевых коэффициентов (нагрузок) в тех позициях, где для другого столбца они ненулевые.
- При числе общих факторов больше четырёх в каждой паре столбцов должно быть некоторое количество нулевых нагрузок в одних и тех же строках.
- Для каждой пары столбцов матрицы V должно быть как можно меньше значительных по величине нагрузок, соответствующих одним и тем же строкам.



ВРАЩЕНИЕ БЫВАЕТ:

- ОРТОГОНАЛЬНЫМ

При первом виде вращения каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, поэтому факторы оказываются независимыми, некоррелированными друг от друга.

- КОСОУГОЛЬНЫМ.

Второй вид — это преобразование, при котором факторы коррелируют друг с другом.

Кластерный анализ — **многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.**

Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры.

Различные приложения кластерного анализа можно свести к четырем основным задачам:

1- разработка типологии или классификации;

2- исследование полезных концептуальных схем группирования объектов;

3- порождение гипотез на основе исследования данных;

4- проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Недостатки кластерного анализа:

Многие методы кластерного анализа — довольно простые процедуры, которые, как правило, не имеют достаточного статис тического обоснования.

Методы кластерного анализа разрабатывались для многих научных дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.

Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.

Цель кластерного анализа заключается в поиске существующих структур.

Методы кластеризации

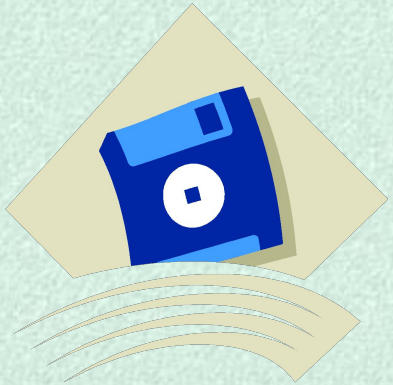
1. ***Вероятностный подход***. **Предполагается, что каждый рассматриваемый объект относится к одному из k классов.**
   1. **K-средних (K-means)**
   2. **K-medians**
   3. **EM-алгоритм**
   4. **Алгоритмы семейства FOREL**
   5. **Дискриминантный анализ**
2. **Подходы на основе систем искусственного интеллекта: весьма условная группа, так как методов очень много и методически они весьма различны.**
   1. **Метод нечеткой кластеризации C-средних (C-means)**
   2. **Нейронная сеть Кохонена**
   3. **Генетический алгоритм**
3. **Логический подход. Построение дендрограммы осуществляется с помощью дерева решений.**
4. **Теоретико-графовый подход.**
   1. **Графовые алгоритмы кластеризации**
5. **Иерархический подход. Предполагается наличие вложенных групп (кластеров различного порядка).**
   1. **Иерархическая дивизивная кластеризация или таксономия. Задачи кластеризации рассматриваются в количественной таксономии.**
6. **Другие методы. Не вошедшие в предыдущие группы.**
   1. **Статистические алгоритмы кластеризации**
   2. **Ансамбль кластеризаторов**
   3. **Алгоритмы семейства KRAB**
   4. **Алгоритм, основанный на методе просеивания**
   5. **DBSCAN и др.**

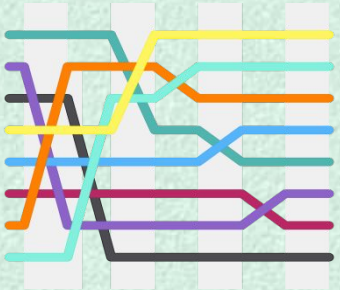
Формальная постановка задачи кластеризации

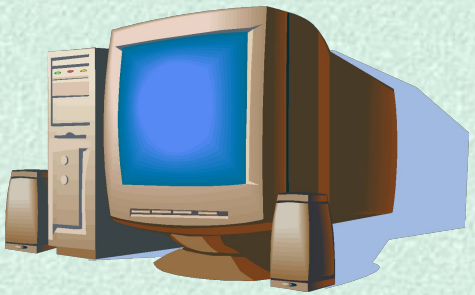
Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^{m} = \{x_1, \ldots, x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^{m}$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.

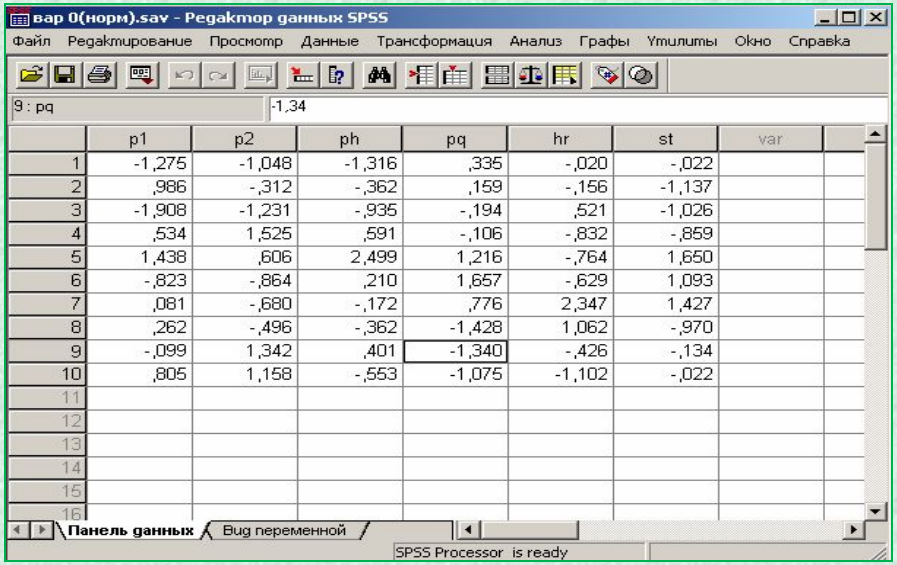


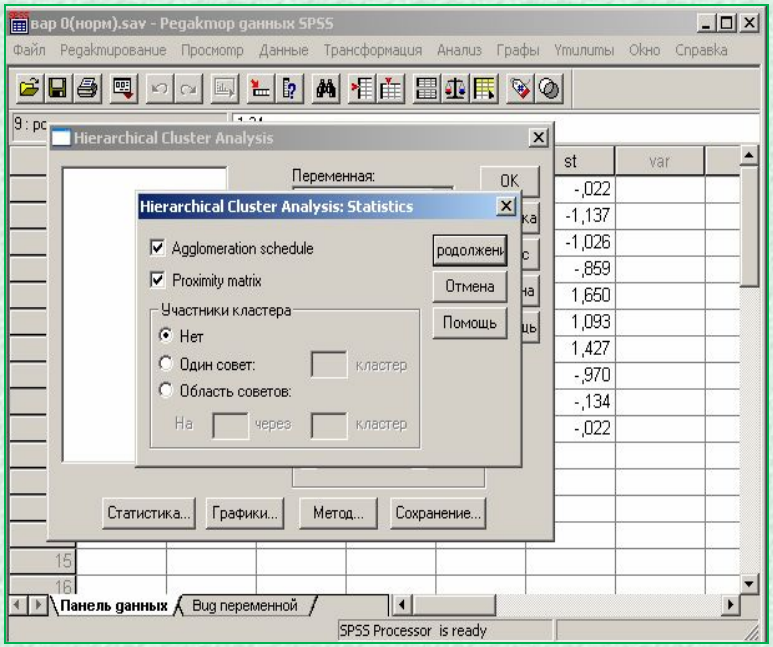




Пример решения в программе SPSS 11.0

1) Запустите программу SPSS 11.
2) Выберите в меню Fail… (файл)New…(новый)Data… (данные)
3) Заполните матрицу данных предварительно нормированными значениями в соответствии с вариантом. В панели данных введите заданные данные, а в панеливид переменнойзадайте имя и тип переменной.





Пример решения в программе SPSS 11.0

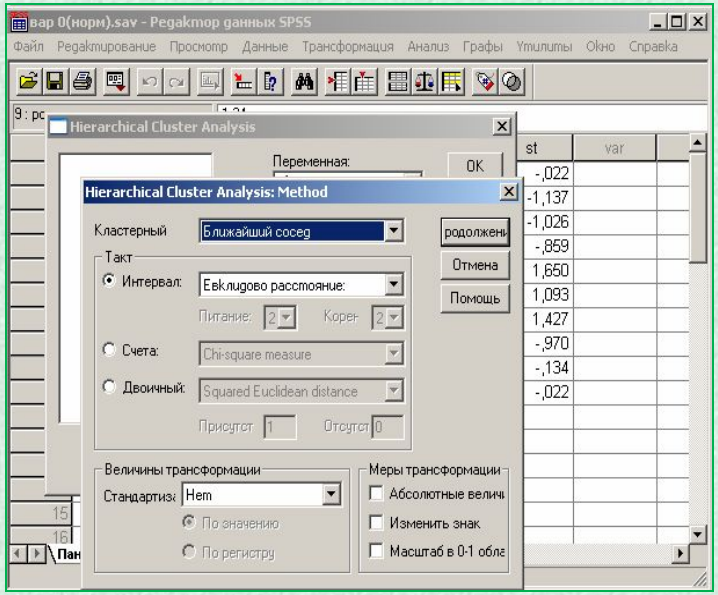
4) Выберите в меню Analyze (Анализ) Classify (Классифицировать) Hierarchical Cluster... (Иерархический кластерный анализ). Перенесите значения (p1, p2, ph, pq, hr, st) в поле переменных.

5) В меню статистика поставьте галочку в полеproximitymatrixи нажмите кнопку продолжить.

6) В меню графики поставьте галочку в полеdendrogramи выберите положение дендрограммы (вертикальное или горизонтальное) и нажмите кнопку продолжить.

7) В меня метод выберите способ расчета расстояния(в нашем случаеевклидово растояние), и метод кластерного анализа (в нашем случаеближайший сосед) и нажмите кнопку продолжить.

8) Ничего больше не меняя, начните расчет нажатием кнопки ОК.



Proximity Matrix

	Euclidean Distance									
Case	1	2	3	4	5	6	7	8	9	10
1	,000	2,803	1,470	3,884	5,364	2,435	3,344	2,959	3,609	3,596
2	2,803	,000	3,185	2,256	4,300	3,368	3,770	2,142	2,797	2,427
3	1,470	3,185	,000	4,214	6,096	3,444	3,890	2,720	3,832	4,199
4	3,884	2,256	4,214	,000	3,654	3,828	4,665	3,227	1,638	1,797
5	5,364	4,300	6,096	3,654	,000	3,610	4,533	5,289	4,140	4,264
6	2,435	3,368	3,444	3,828	3,610	,000	3,277	4,275	3,994	4,031
7	3,344	3,770	3,890	4,665	4,533	3,277	,000	3,515	4,365	4,633
8	2,959	2,142	2,720	3,227	5,289	4,275	3,515	,000	2,648	2,962
9	3,609	2,797	3,832	1,638	4,140	3,994	4,365	2,648	,000	1,517
10	3,596	2,427	4,199	1,797	4,264	4,031	4,633	2,962	1,517	,000

This is a dissimilarity matrix

Вывод основных результатов выглядит следующим образом: