Кластерный анализ

Стукач Олег Владимирович
Каф. КИСМ, пр. Ленина, д. 2, оф. 204
☏(3822)-701777*2754
tomsk@ieee.org
http://ieee.tpu.ru/pages/stukach.htm

Кластерный анализ

– это один из способов классификации объектов по их признакам

1 Общие сведения о кластерном анализе

2 Методы кластеризации

3 Пример кластеризации методом К-средних в программе STSTISTICA

Стукач О.В. Программный комплекс Statistica в решении задач управления качеством: учебное пособие / О. В. Стукач ; Национальный исследовательский Томский политехнический университет (ТПУ). - 1 компьютерный файл (pdf; 2.4 MB). - Томск : Изд-во ТПУ, 2011. - ТПУ. - Adobe Reader. - <URL:http://www.lib.tpu.ru/fulltext2/m/2011/m426.pdf>.

Стр. 134, глава 7

Кластерный анализ

Достоинство кластерного анализа состоит в том, что он работает даже тогда, когда данных мало и не выполняются требования нормальности распределений случайных величин и другие требования классических методов статистического анализа

Два типа задач кластерного анализа

в разбиении множества, состоящего из *n* объектов на *к* кластеров, то есть число кластеров может быть выбрано априорно и наперед задано

число кластеров определяется в процессе разбиения множества на кластеры, исходя из оптимизации некоторой целевой функции. Классифицируемы могут быть как параметры, так и объекты

Правило трёх частей

Плэтт В. Информационная работа стратегической разведки. – М. : Изд-во иностранной литературы, 1958.

В первые годы после второй мировой войны в колледж принимали слишком много студентов. При таком большом количестве студентов было трудно, а зачастую невозможно хорошо организовать преподавание. Профессор сказал мне: «В таких условиях вы не можете уделить достаточное внимание каждому студенту. Приходится расходовать своё время с наибольшей пользой. В самом начале работы выявите наиболее способных студентов, составляющих первую четверть группы. Этими студентами в дальнейшем можно не заниматься. Они в состоянии все усвоить сами и наверняка сдадут экзамены. Затем *как можно скорее выявите самых слабых студентов, составляющих последнюю четверть группы. На них не следует тратить время. Они не принесут славы ни вам, ни университету. Вероятно, им не удастся получить диплом инженера.*

Чтобы расходовать своё время с максимальным эффектом, тратьте его почти целиком на студентов со средними способностями, составляющими половину группы. Они нуждаются в вашей помощи и обладают достаточными способностями, чтобы извлечь из неё пользу».

Разделение любой изучаемой группы людей или организаций на три части практически полезно. С помощью этого метода мы выделяем в рамках любой группы людей, которые занимают положение в середине группы и со временем воспримут передовые методы, и, наконец, людей, плетущихся в хвосте. ***Рано или поздно в силу экономической или интеллектуальной конкуренции последняя часть группы исчезает***

Кластерный анализ

Результаты, полученные методами кластерного анализа, применяют :

в медицине: кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям.

в маркетинге: сегментация конкурентов и потребителей;

в менеджменте: классификация поставщиков, выявление схожих производственных ситуаций, при которых возникает брак;

в социологии: разбиение респондентов на однородные группы;

в других областях

Методы кластерного анализа

Иерархические методы:

метод ближней связи,

метод средней связи,

метод Уорда

Итеративные методы группировки:

метод к-средних

Алгоритмы типа разделения графа:

метод корреляционных плеяд Терентьева,

вроцлавская таксономия

Метод k-средних

Метод k-средних – это метод кластерного анализа, цель которого является разделение m наблюдений на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центу (центроиду) которого оно ближе всего.

В качестве меры близости используется Евклидово расстояние:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n} (x_p - y_p)^2}$$

Метод k-средних

Итак, рассмотрим ряд наблюдений $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}),\ x^{(j)} \in R^n$

Метод k-средних разделяет m наблюдений на k групп (или кластеров) ($k \le m$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min \left[\sum_{i=1}^{k} \sum_{x^{(j)} \in S_i} \left\| x^{(j)} - \mu_i \right\|^2 \right] \qquad x^{(j)} \in R^n, \mu_i \in R^n$$

где μ_i - центроид для кластера S_i

Метод k-средних

Итак, если мера близости до центроида определена, то разбиение объектов на кластеры сводится к определению центроидов этих кластеров. Число кластеров k задается исследователем заранее.

На первом этапе центроиды кластеров выбираются случайно или по определённому правилу (например, выбрать центроиды, максимизирующие начальные расстояния между кластерами).
Относим наблюдения к тем кластерам, чье среднее (центроид) к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам.

Метод k-средних

Затем центроид каждого i-го кластера перевычисляется по следующему правилу:

$$\mu_j = \frac{1}{s_j} \sum_{x^{(j)} \in S_i} x^{(j)}$$

Таким образом, алгоритм k-средних заключается в перевычислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге.
Алгоритм останавливается, когда значения μ_i не меняются:

$$\mu_i^{\text{mar } t} = \mu_i^{\text{mar } t+1}$$

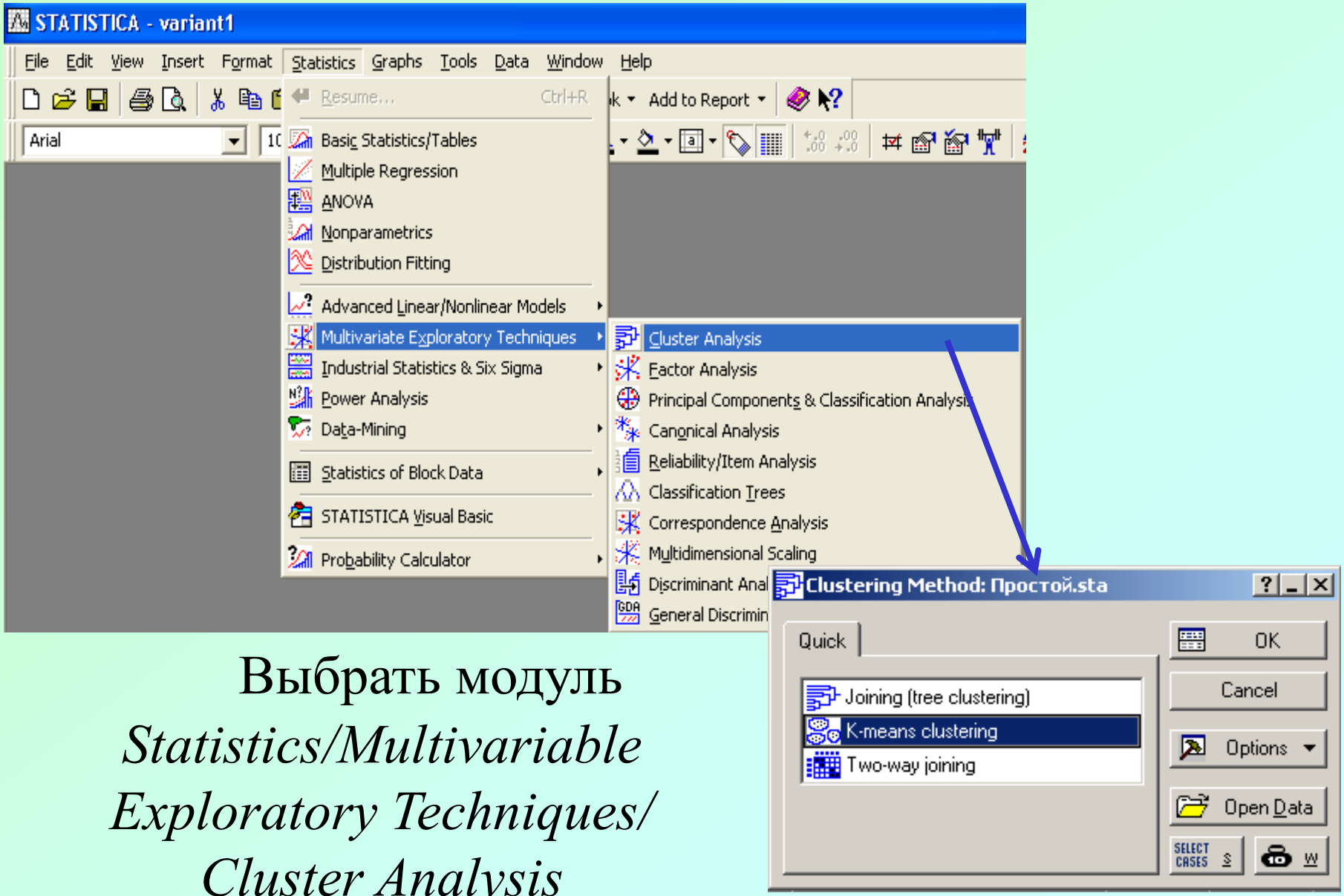
Важно: Неправильный выбор первоначального числа кластеров k может привести к некорректным результатам. Именно поэтому при использовании метода k-средних важно сначала провести проверку подходящего числа кластеров для данного набора данных.

Метод k-средних

Итак, еще раз подчеркнем некоторые особенности метода k-средних:

- В качестве метрики используется Евклидово расстояние
- Число кластеров заранее не известно и выбирается исследователем заранее
- Качество кластеризации зависит от первоначального разбиения

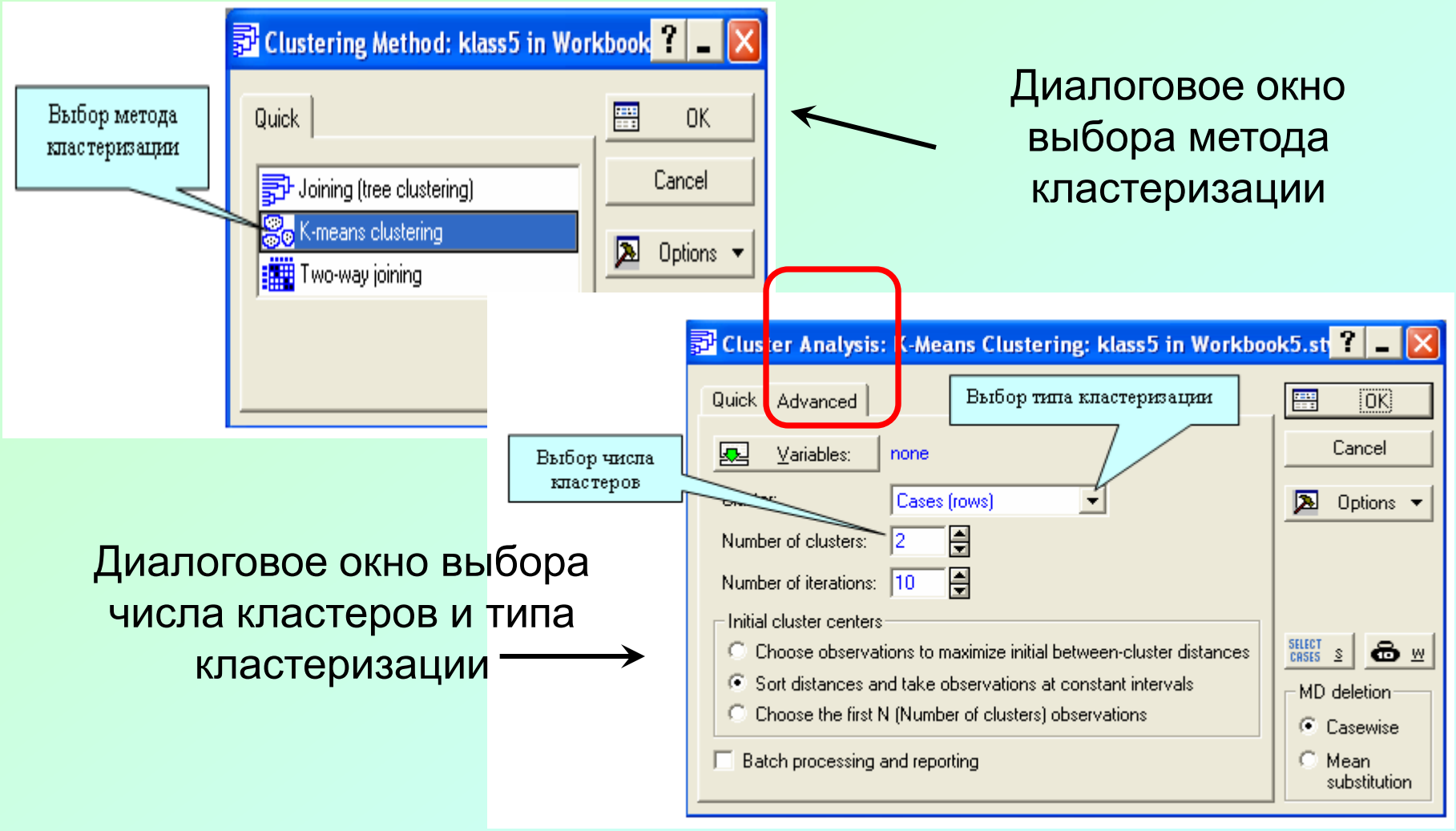
Модуль «Кластерный анализ»



Выбрать модуль

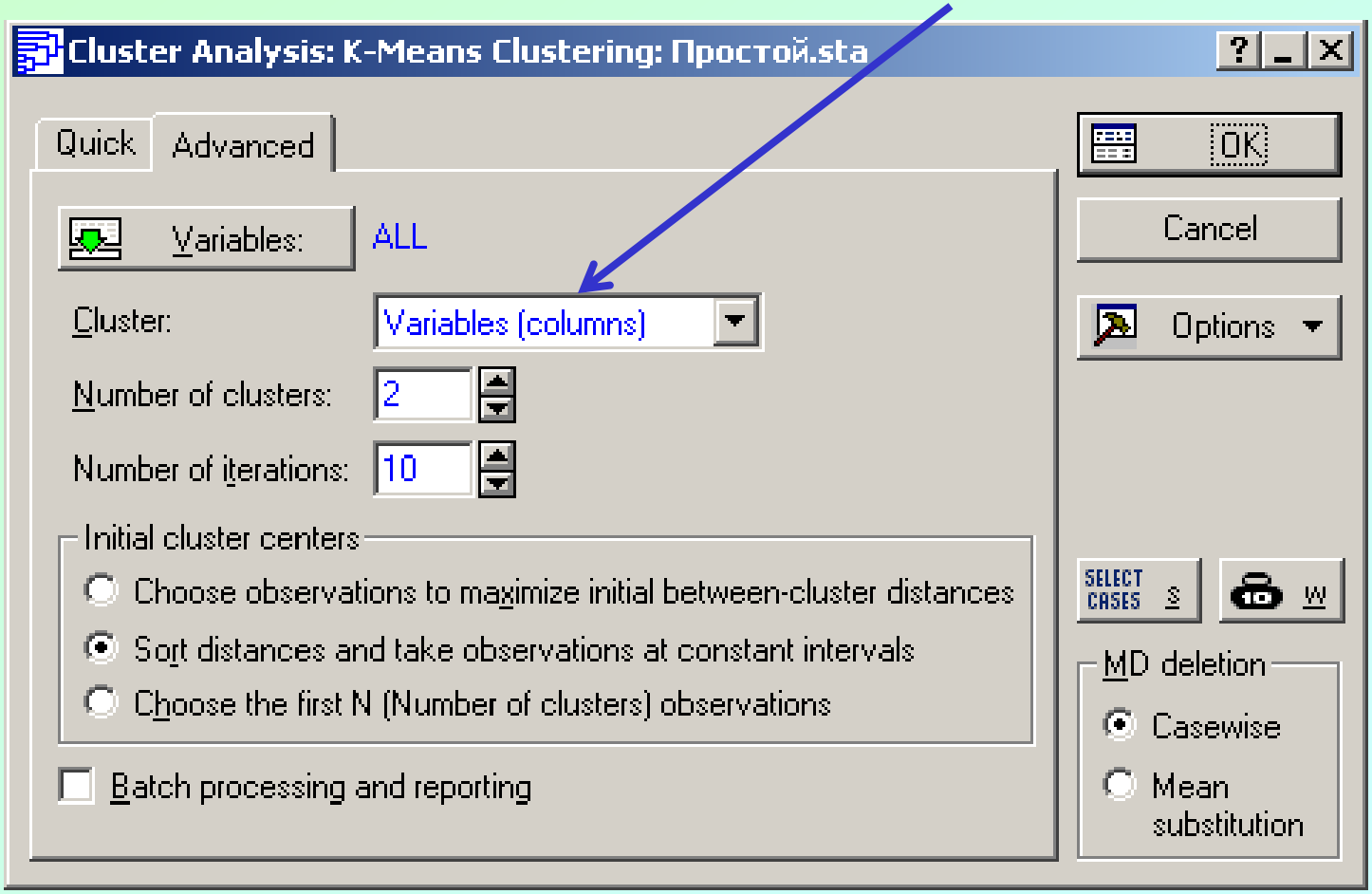
Statistics/Multivariable Exploratory Techniques/ Cluster Analysis

Этапы проведения кластерного анализа в программе STATISTICA



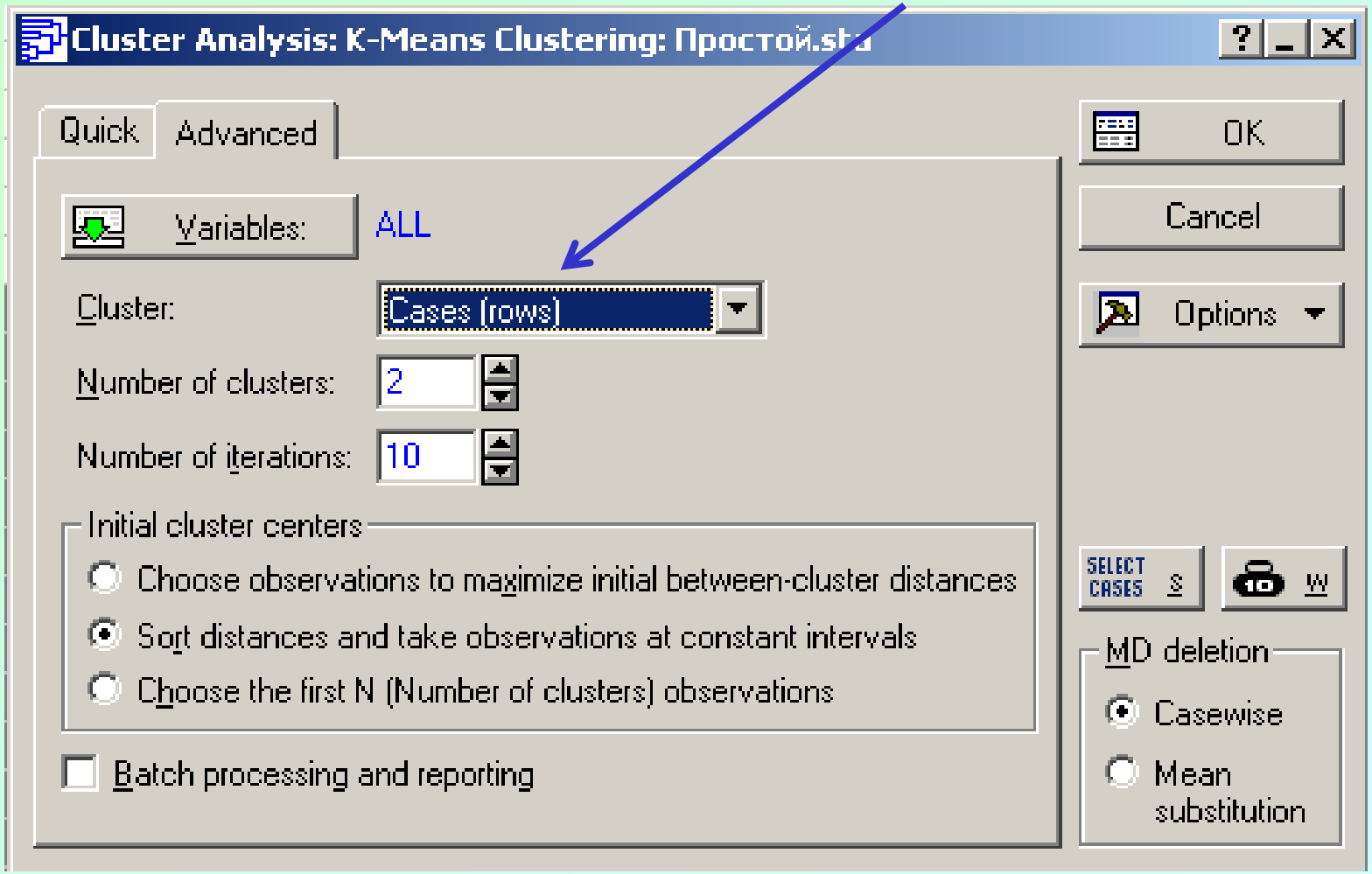
Выбор переменных

Кластеризация по переменным

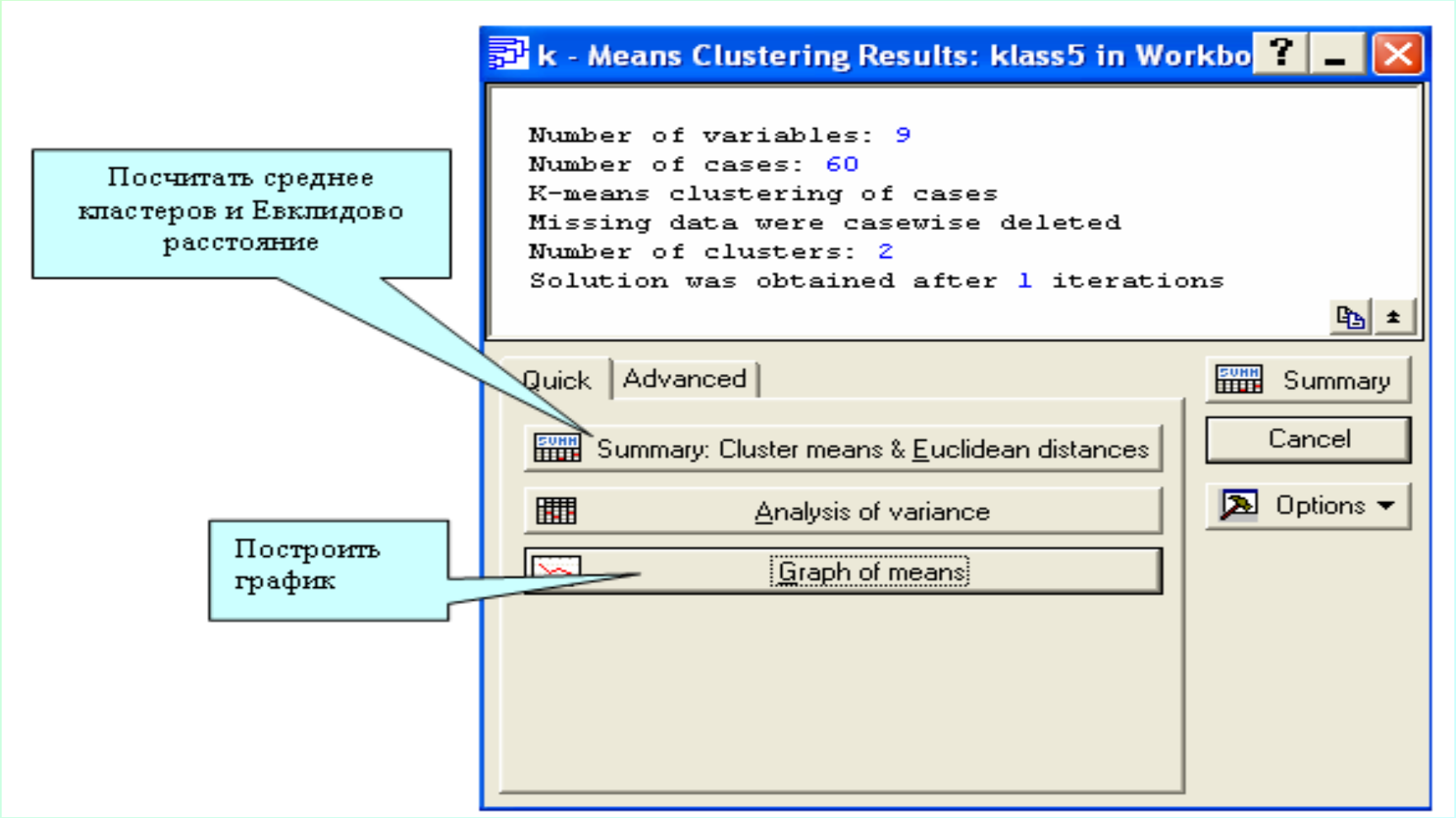


Выбор переменных

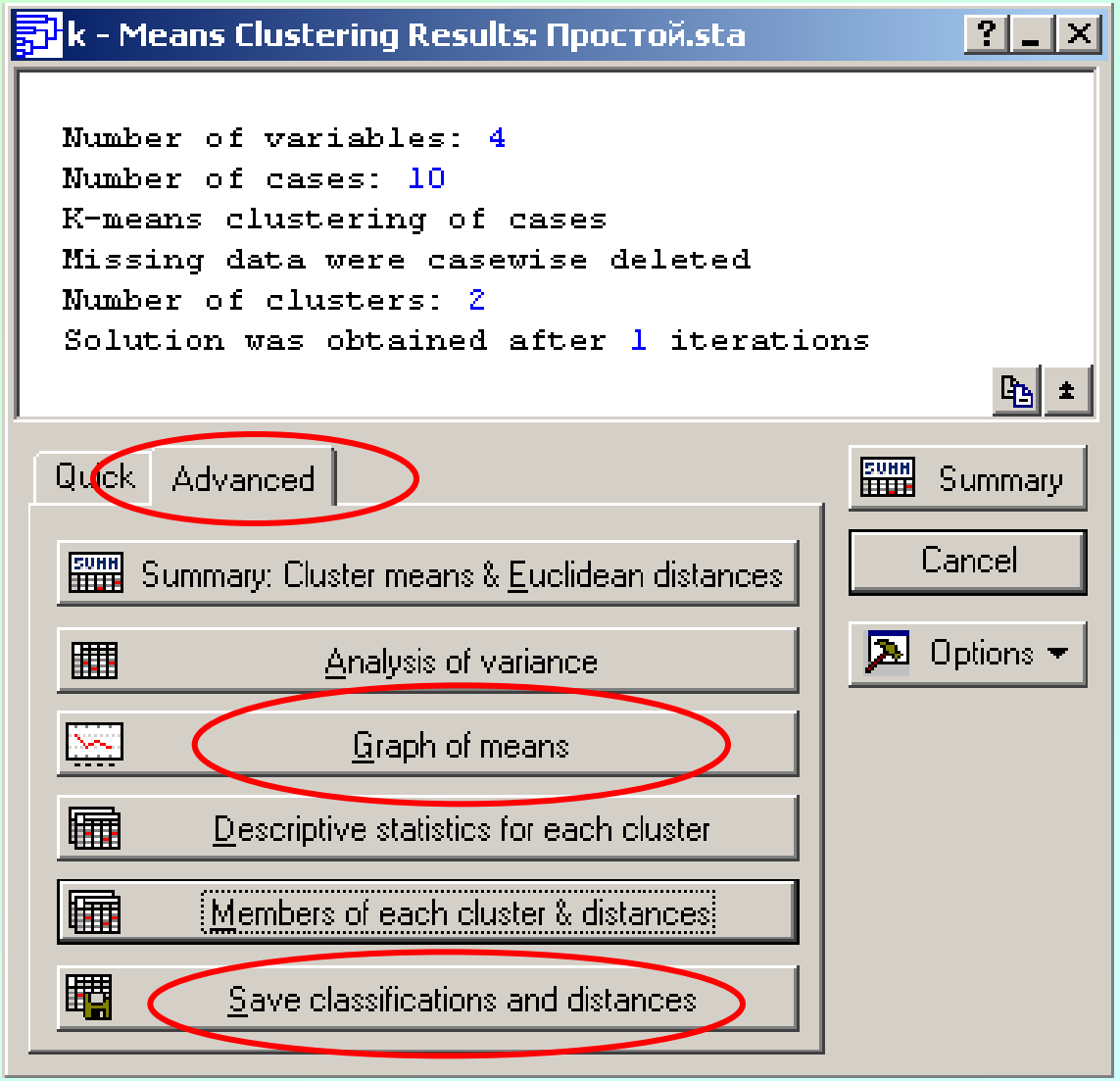
Кластеризация
по наблюдениям



Построение графика координат центров кластеров



Просмотр результатов

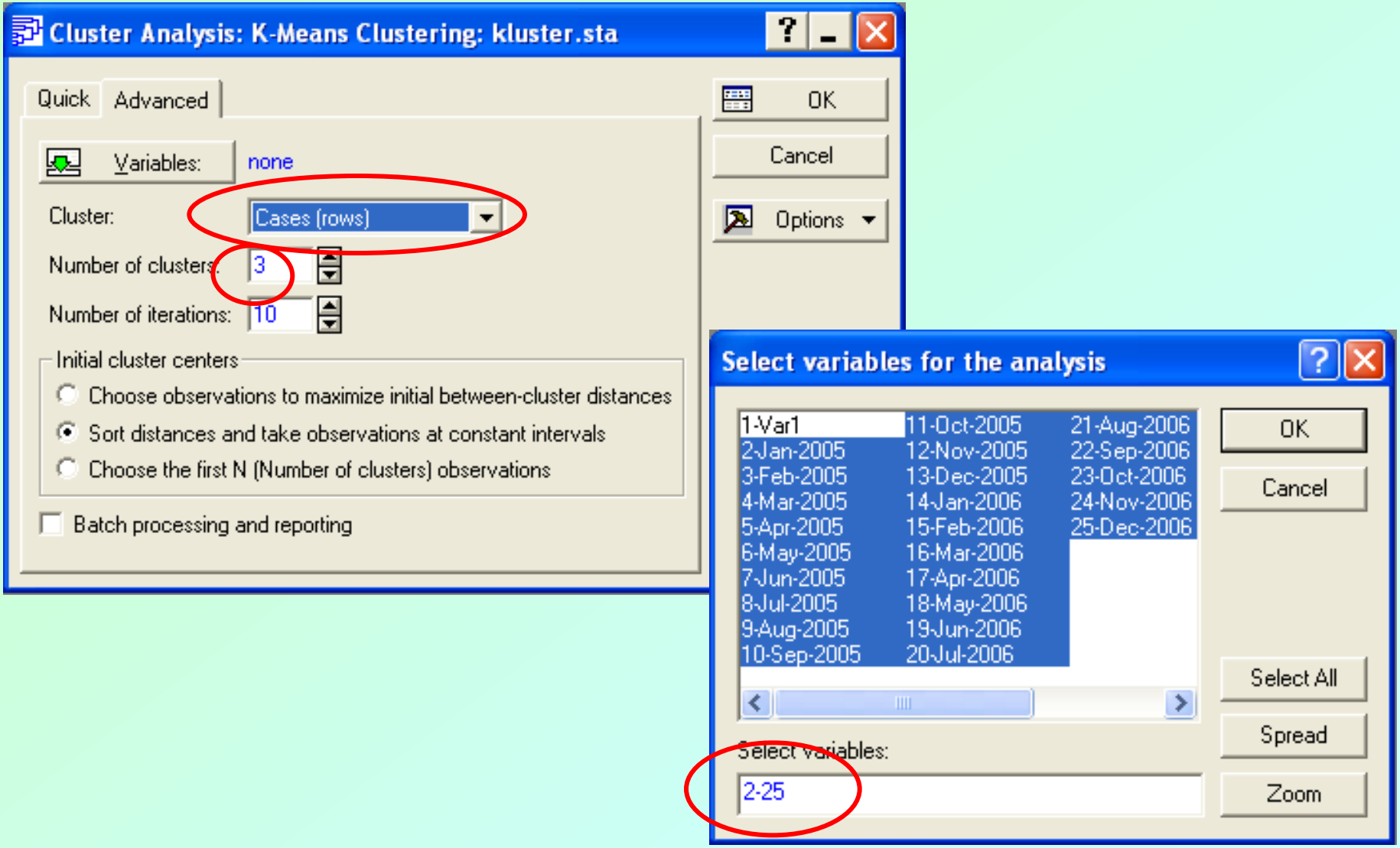


Данные

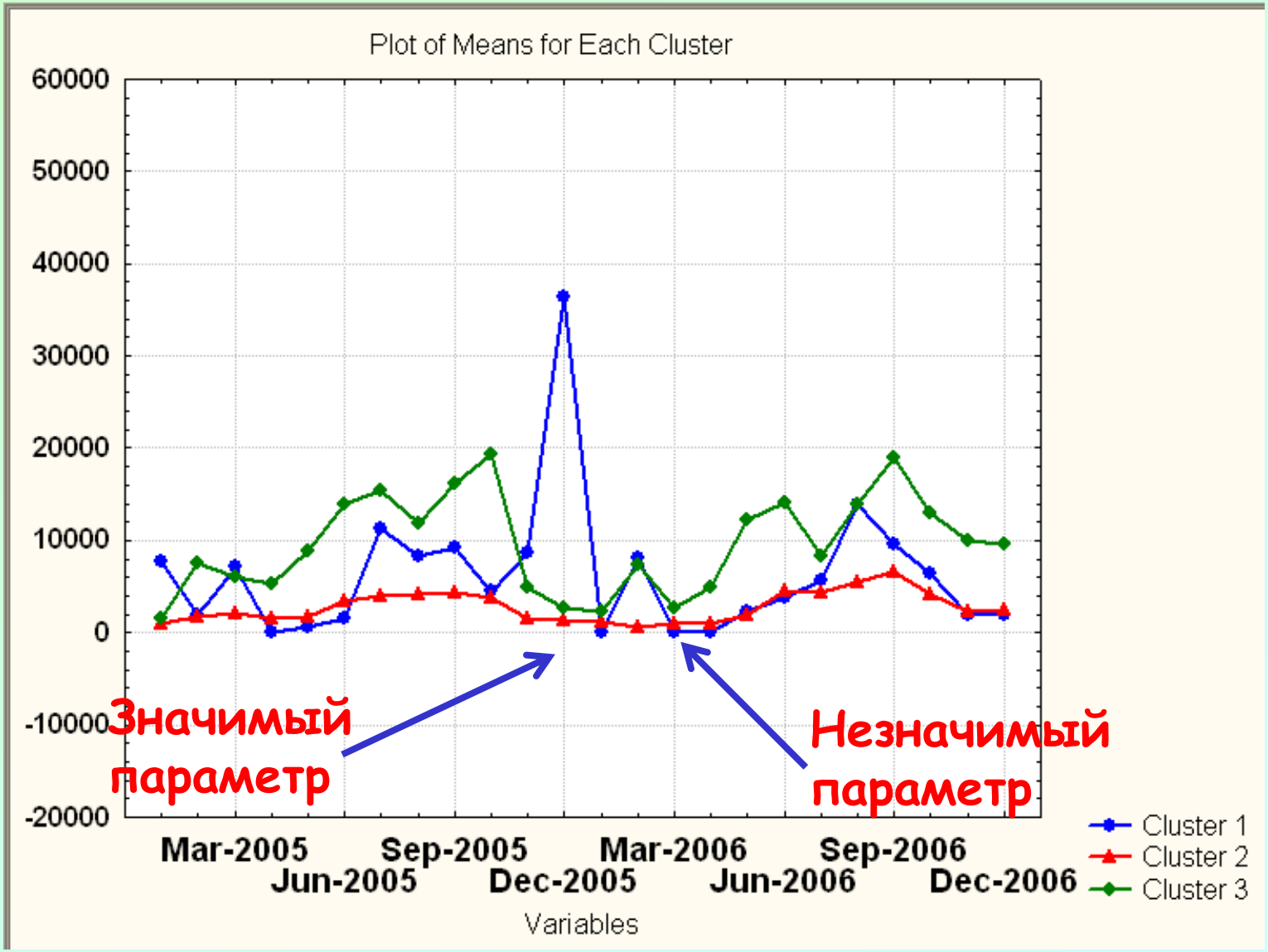
Data: kluster.sta* (25v by 13c)

	1 Var1	2 Jan-2005	3 Feb-2005	4 Mar-2005	5 Apr-2005	6 May-2005	7 Jun-2005	8 Jul-2005	9 Aug-20
1	0-20	20,0	2 606,1	1 821,0	322,1	1 243,6	3 563,5	4 162,4	2 6
2	0-25 (ЩПС)	0	636,0	9 914,0	25 470,0	3 004,0			2 4
3	0-5 (М 800)	0	0	0	871,6	4 694,0	6 947,4	2 221,3	4 4
4	0-70	7 760,0	1 866,1	7 082,0	61,9	571,1	1 596,7	11 363,5	8 2
5	20-40	1 550,0	7 586,1	6 105,0	5 288,1	8 863,7	13 973,0	15 488,9	11 8
6	20-70	74 260,0	43 076,0	0	0	9,1	1 855,9		7
7	40-70	2 200,0	736,1	296,0	888,0	2 545,9	4 714,3	7 233,4	8 2
8	25-60	3 537,0	636,0	12 570,0	9 061,0	1 436,0	505,0	4 332,0	4 8
9	5-20	0	636,1	0	7,4	0	2 076,7	4 074,8	5 1
10	5-25 М 1000	2 980,0	1 276,1	1 651,0	2 740,9	3 978,3	4 387,30	6 133,6	7 9
11	ВКСВ	0	1 272,2	0	0	0	0	0	
12	Вскрыша	0	636,10	0	294,60				
13	С-10	0	7 066,1	2 898,0	553,9				

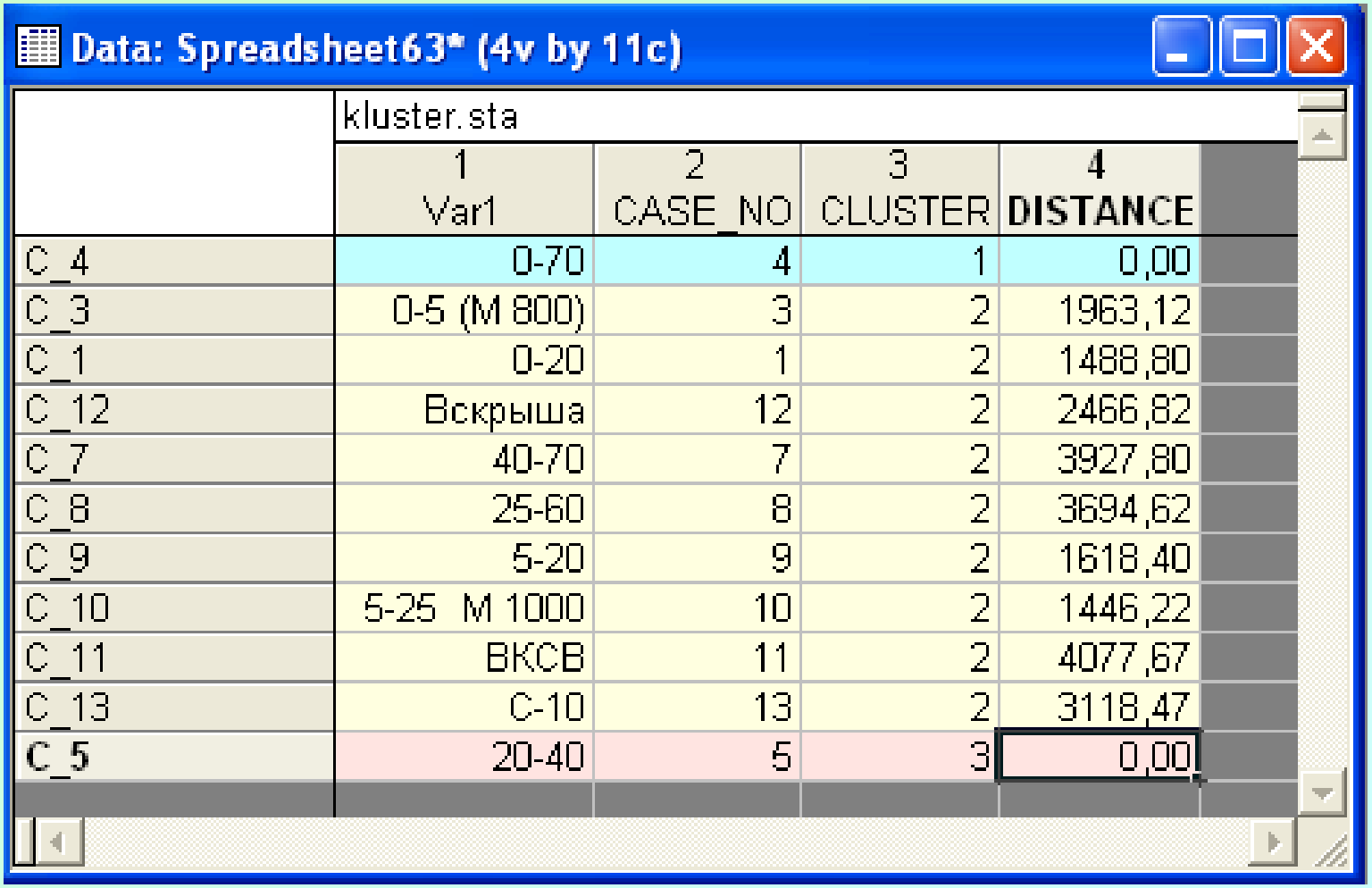
Начинаем анализ



Просмотр результатов



Просмотр результатов



	kluster.sta			
	1 Var1	2 CASE_NO	3 CLUSTER	4 DISTANCE
C_4	0-70	4	1	0,00
C_3	0-5 (M 800)	3	2	1963,12
C_1	0-20	1	2	1488,80
C_12	Вскрыша	12	2	2466,82
C_7	40-70	7	2	3927,80
C_8	25-60	8	2	3694,62
C_9	5-20	9	2	1618,40
C_10	5-25 M 1000	10	2	1446,22
C_11	ВКСВ	11	2	4077,67
C_13	С-10	13	2	3118,47
C_5	20-40	5	3	0,00