

Проведение кластерного анализа в пакете STATISTICA

Перед проведением кластерного анализа проведите нормировку исходных данных. Для этого выберите **Data/Standardize...**

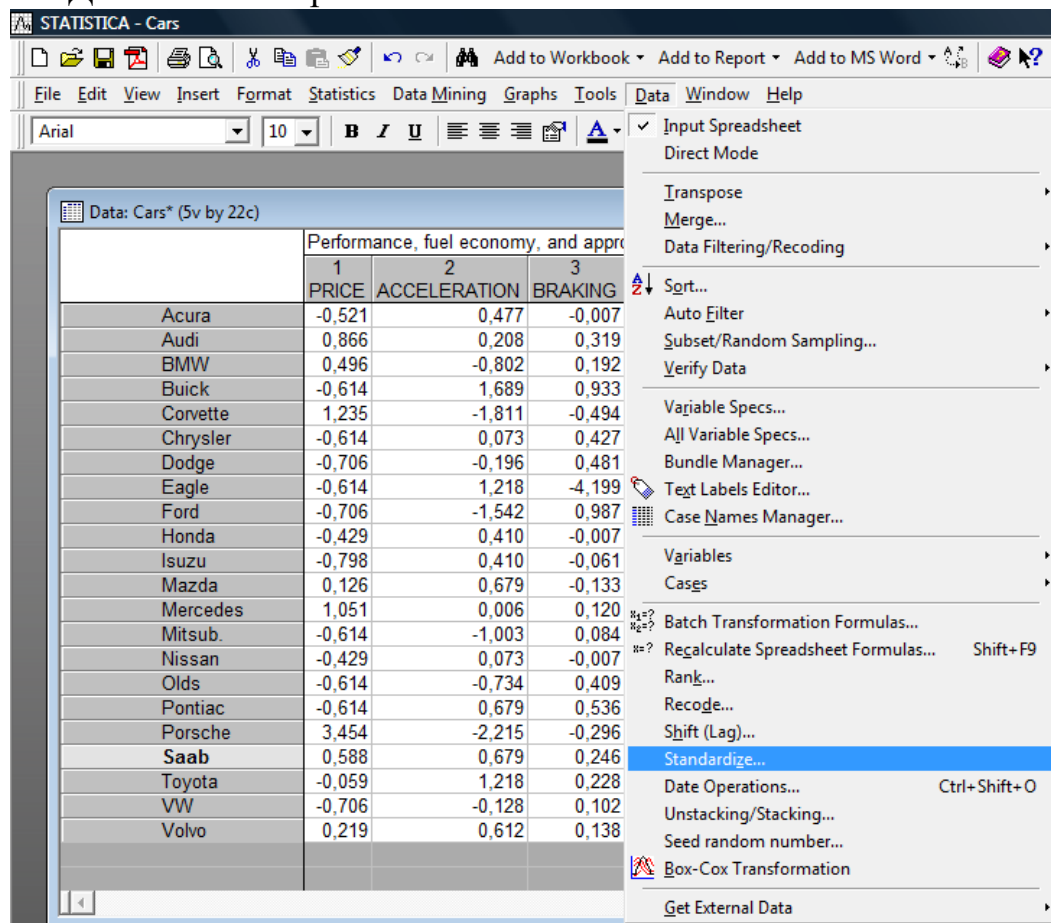


Рис.1.

В появившемся окне (рис. 2) выберите переменные, значения которых будут нормализованы.

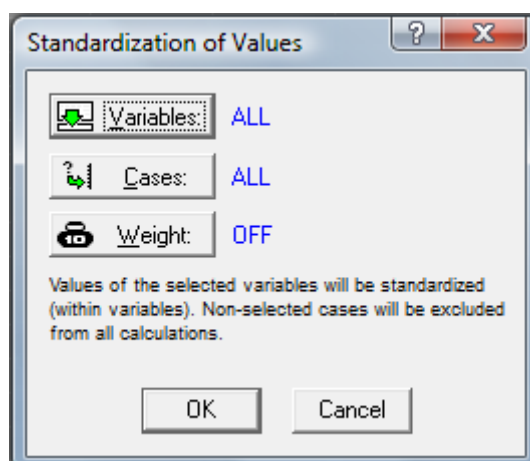


Рис. 2.

Кнопка **Variables** (Переменные) позволяет отобрать те показатели, которые будут нормированы. В моей задаче выбираются все.

С помощью кнопки **Cases** (Наблюдения), можно отобрать лишь часть наблюдений для стандартизации, по умолчанию выбираются все. Также по

умолчанию все наблюдения вносят одинаковый вклад в вычисляемые средние значения и стандартные отклонения.

Нажав кнопку **Weight** (Вес), возможно указать “весовую” переменную.

Для вызова модуля кластерного анализа выберете **Statistics/Multivariate Exploratory Techniques/Cluster Analysis** (рис. 3).

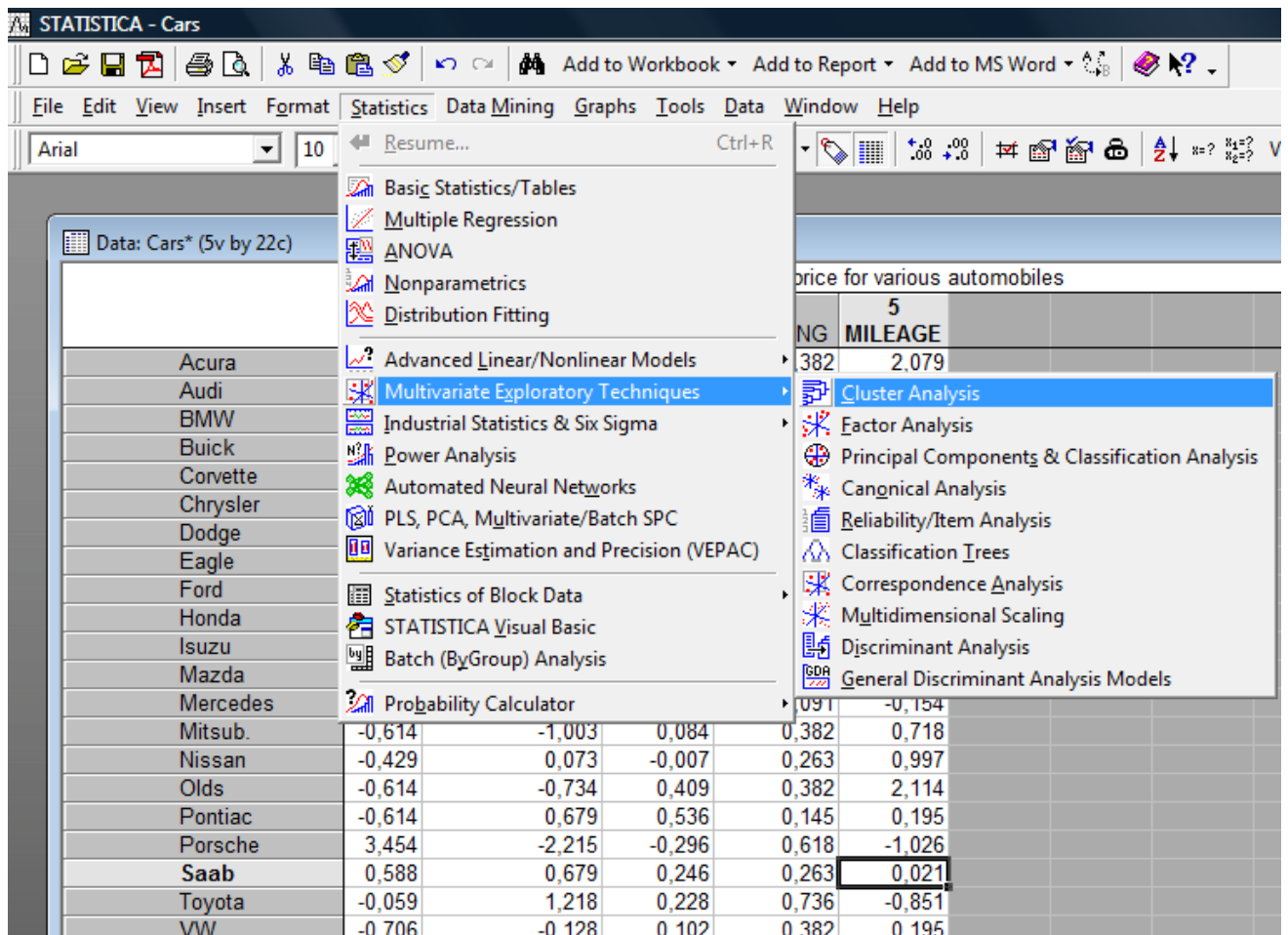


Рис.3.

Появившееся диалоговое окно содержит следующие методы (рис.4):

- **Joining (tree clustering)** - Объединение (двевовидная кластеризация);
- **K – means clustering** - Кластеризация методом К– средних;
- **Two-way joining** – Двухходовое объединение.

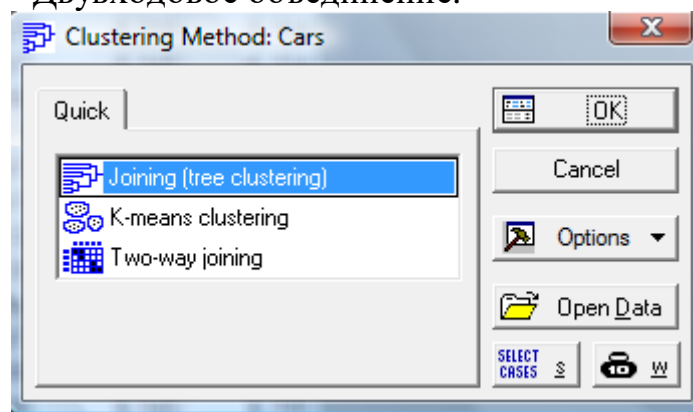


Рис. 4.

Объединение (древовидная кластеризация) – Joining (tree clustering)

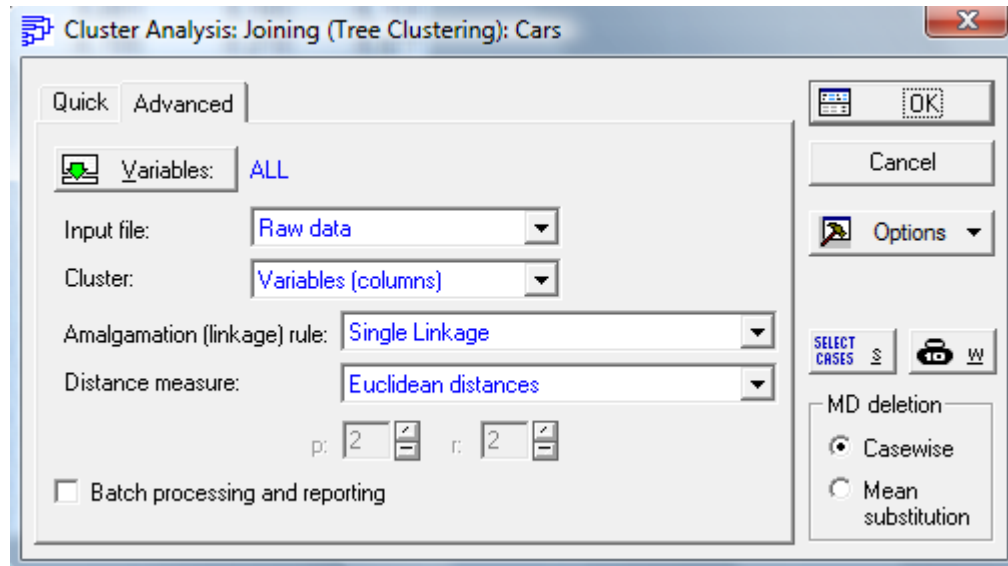


Рис. 5.

Input (Исходные данные) (рис.5) представляет собой раскрывающееся меню. Строка **Distance matrix** (Матрица расстояний) предусмотрена на тот случай, если входная информация представлена в виде мер сходства. Если представлены значения переменных, то выбираем **Raw data** (Исходные данные).

В поле **Cluster** (Кластер) (рис. 5) задается направление классификации. При кластеризации самих переменных помечаются **Variables [Columns]** (Переменные [столбцы]). При кластеризации объектов выбирается **Cases [rows]** (Наблюдения [строки]).

Строка **Amalgamation [linkage] rule** (Правило объединения [связи]) содержит установки для выбора следующих мер сходства:

- **Single Linkage** (Метод одиночной связи “принцип ближайшего соседа”).
- **Complete Linkage** (Метод полной связи “принцип дальнего соседа”).
- **Unweighted pair-group average** (Невзвешенное попарное среднее).
- **Weighted pair-group average** (Взвешенное попарное среднее).
- **Unweighted pair-group centroid** (Невзвешенный центроидный метод).
- **Weighted pair-group centroid** (Взвешенный центроидный метод).
- **Ward’s method** (Метод Варда).

В окошке **Distance measure** (Мера расстояния) (рис. 5) предлагаются различные виды расстояний:

- **Squared Euclidean distances** (квадрат Евклидова расстояния).
- **Euclidean distances** (Евклидово расстояние).
- **City-block (Manhattan) distance** (Расстояние городских кварталов (Манхэттенское расстояние)).

- **Chebyshev distance metric** (Расстояние Чебышева).
- **Power: $\text{SUM}(\text{ABS}(x-y)**p)**1/r$** (Степенное расстояние).
- **Percent disagreement** (Процент несогласия).

После установки всех необходимых параметров для проведения кластеризации щелчком на **Ok** и рассмотрим окно с результатами классификации (рис. 6).

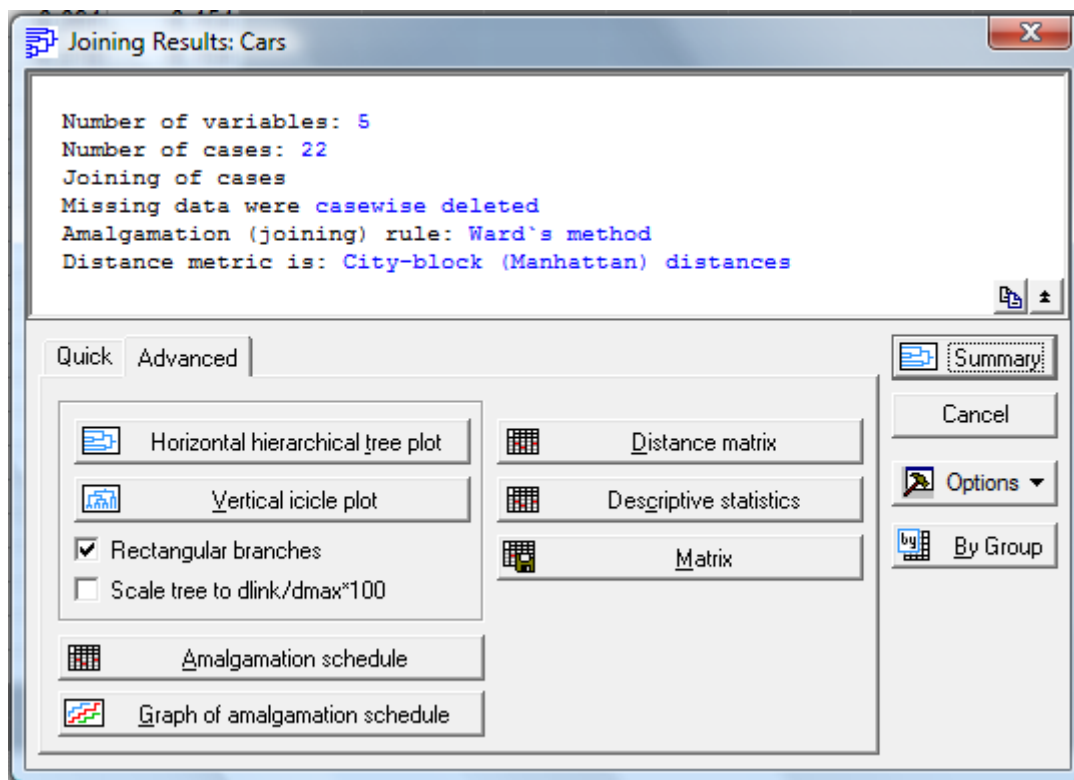


Рис. 6.

Рассмотрим **Vertical icicle plot** (Вертикальную древовидную дендограмму) (рис. 7).

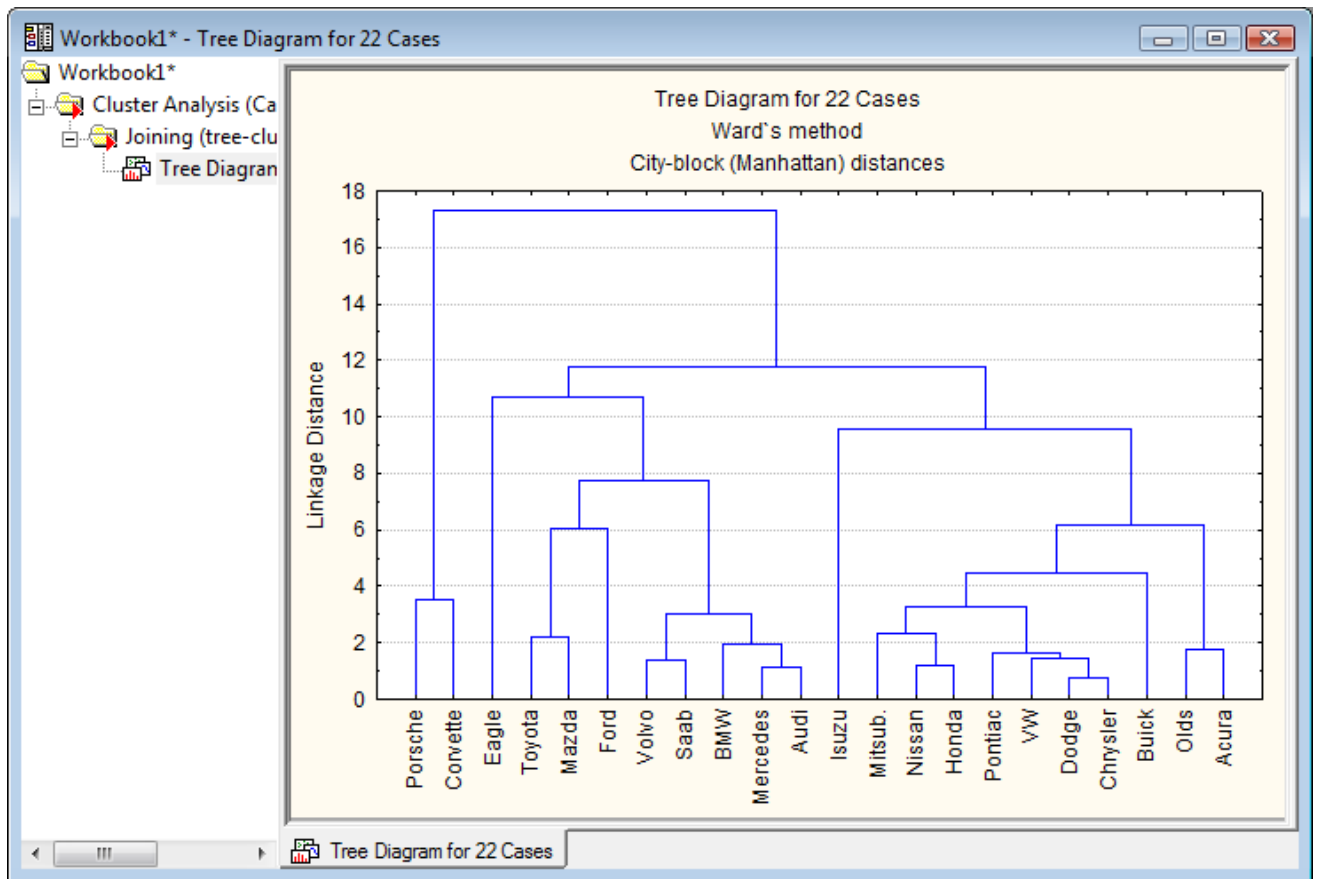


Рис. 7.

Щелкнув по кнопке **Amalgamation schedule** (Схема объединения), можно выбрать таблицу результатов со схемой объединения (рис. 8). Первый столбец таблицы содержит расстояния для соответствующих кластеров. Каждая строка показывает состав кластера на данном шаге классификации.

Workbook2* - Amalgamation Schedule (Cars)

Amalgamation Schedule (Cars)
Ward's method
City-block (Manhattan) distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10
,7706233	Chrysler	Dodge								
1,108841	Audi	Mercedes								
1,200928	Honda	Nissan								
1,367350	Saab	Volvo								
1,471008	Chrysler	Dodge	VW							
1,618842	Chrysler	Dodge	VW	Pontiac						
1,754438	Acura	Olds								
1,939549	Audi	Mercedes	BMW							
2,193332	Mazda	Toyota								
2,360809	Honda	Nissan	Mitsub.							
3,040594	Audi	Mercedes	BMW	Saab	Volvo					
3,301799	Chrysler	Dodge	VW	Pontiac	Honda	Nissan	Mitsub.			
3,524945	Corvette	Porsche								
4,466608	Buick	Chrysler	Dodge	VW	Pontiac	Honda	Nissan	Mitsub.		
6,040308	Ford	Mazda	Toyota							
6,180648	Acura	Olds	Buick	Chrysler	Dodge	VW	Pontiac	Honda	Nissan	Mits
7,757560	Audi	Mercedes	BMW	Saab	Volvo	Ford	Mazda	Toyota		
9,578758	Acura	Olds	Buick	Chrysler	Dodge	VW	Pontiac	Honda	Nissan	Mits
10,71485	Audi	Mercedes	BMW	Saab	Volvo	Ford	Mazda	Toyota	Eagle	
11,79834	Acura	Olds	Buick	Chrysler	Dodge	VW	Pontiac	Honda	Nissan	Mits

Рис. 8.

Щелкнув по кнопке **Graph of amalgamation schedule** (График схемы объединения) (рис. 9), посмотрим результаты древовидной кластеризации в графическом виде.

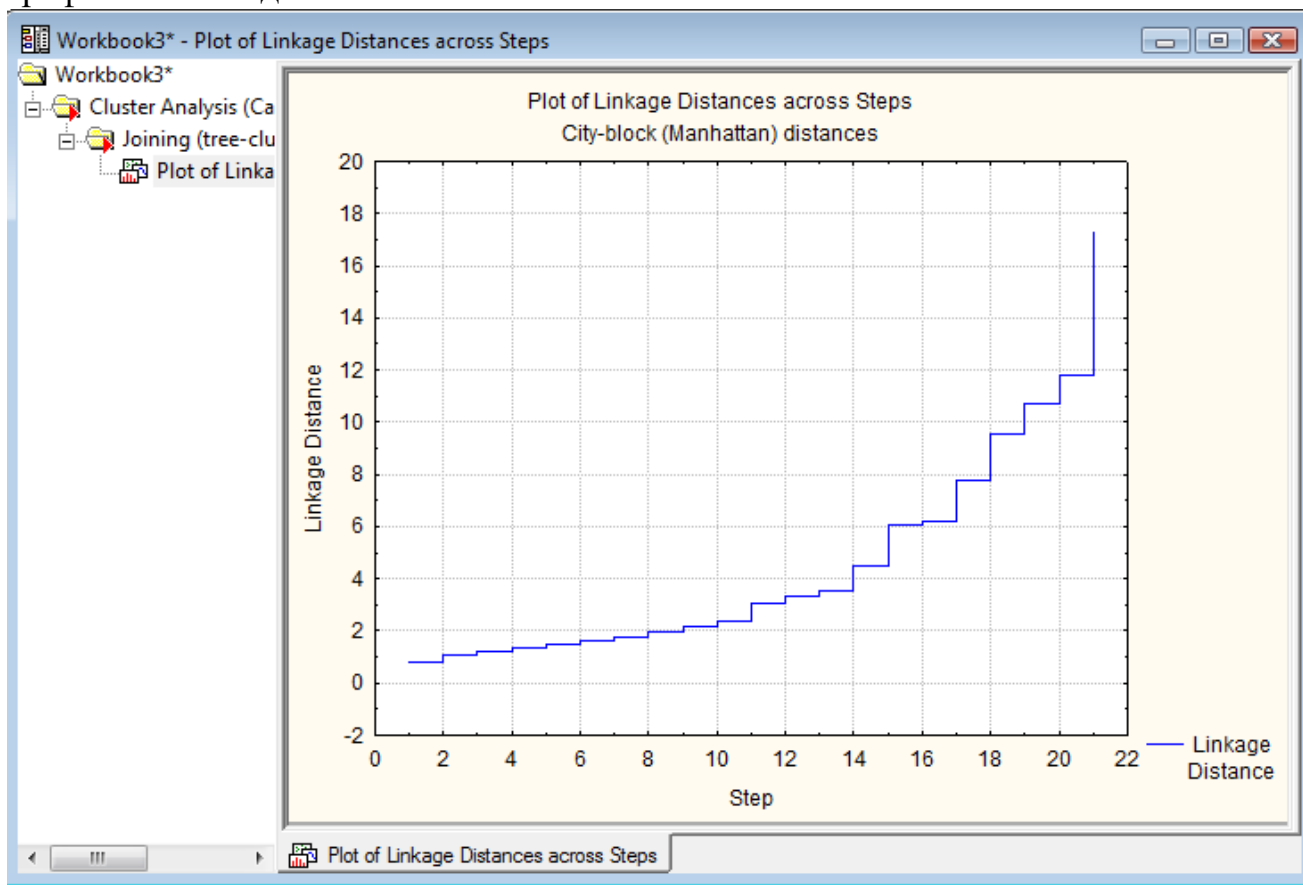


Рис. 9.

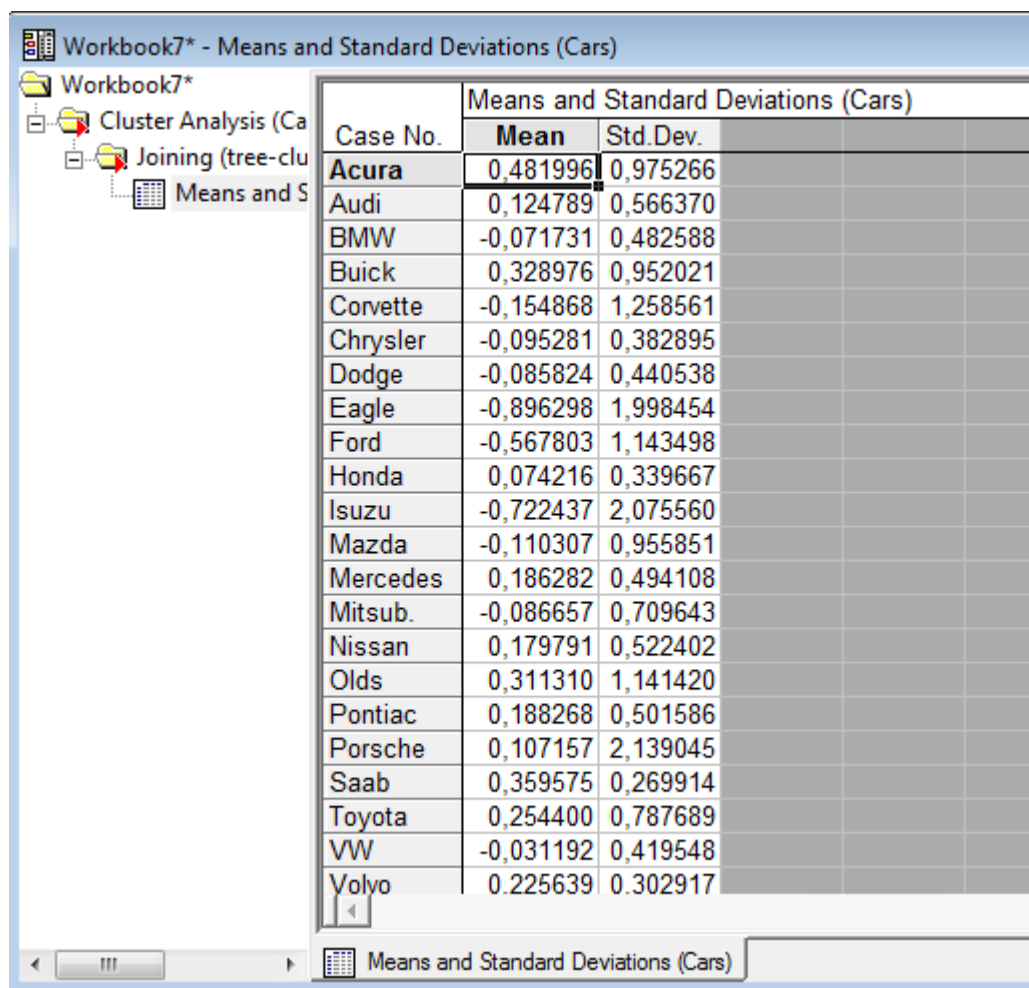
Просмотр матрицы расстояний осуществляется через кнопку **Distance matrix** (Матрица расстояний) (рис. 10).

Workbook4* - City-block (Manhattan) distances (Cars)

Case No.	Acura	Audi	BMW	Buick	Corvette	Chrysler	Dodge	Eagle	Ford	Honda
Acura	0.0	5.21	5.20	5.1	7.9	3.75	3.81	8.4	7.2	2.2
Audi	5.2	0.00	2.03	4.2	4.3	2.36	2.90	7.1	5.3	2.9
BMW	5.2	2.03	0.00	4.5	4.0	2.34	2.33	8.2	4.5	2.9
Buick	5.1	4.22	4.46	0.0	8.5	2.12	2.78	6.1	5.3	3.1
Corvette	7.9	4.27	4.02	8.5	0.0	6.36	5.88	9.8	5.6	6.3
Chrysler	3.8	2.36	2.34	2.1	6.4	0.00	0.77	6.3	4.2	1.7
Dodge	3.8	2.90	2.33	2.8	5.9	0.77	0.00	7.1	3.4	2.0
Eagle	8.4	7.12	8.16	6.1	9.8	6.29	7.06	0.0	9.4	6.0
Ford	7.2	5.27	4.54	5.3	5.6	4.19	3.42	9.4	0.0	5.4
Honda	2.2	2.99	2.98	3.2	6.4	1.71	2.01	6.5	5.4	0.0
Isuzu	6.0	8.13	8.12	7.7	11.6	6.25	6.84	10.9	10.3	5.3
Mazda	4.9	3.30	4.34	5.1	5.5	4.18	4.25	7.1	4.5	3.5
Mercedes	4.9	1.11	1.43	4.3	4.2	2.16	2.56	7.8	6.0	2.6
Mitsub.	3.0	4.79	2.76	5.0	5.2	2.88	2.41	8.5	4.2	2.3
Nissan	1.7	3.78	3.50	4.4	6.4	2.24	2.30	7.7	5.7	1.2
Olds	1.8	5.78	4.13	5.8	7.2	3.68	3.21	9.9	5.6	3.8
Pontiac	3.0	3.28	3.52	2.1	7.1	1.42	1.37	6.5	4.7	1.2
Porsche	10.3	6.68	6.44	10.9	3.5	8.78	8.30	12.6	7.3	8.7
Saab	3.7	1.87	2.16	3.5	5.3	2.64	2.70	7.4	6.1	2.1
Toyota	4.7	3.03	4.14	3.4	5.5	3.54	3.60	6.1	5.6	3.3
VW	2.8	3.47	2.79	3.7	5.7	1.56	1.03	7.2	4.5	1.4
Volvo	3.3	2.40	2.39	3.2	6.3	2.18	2.95	6.8	6.4	1.2

Рис. 10.

Строка **Descriptive statistics** (Описательные характеристики) открывает таблицу результатов со средними значениями и стандартными отклонениями для каждого объекта, включенного в кластерный анализ (рис. 11).



Case No.	Mean	Std.Dev.
Acura	0,481996	0,975266
Audi	0,124789	0,566370
BMW	-0,071731	0,482588
Buick	0,328976	0,952021
Corvette	-0,154868	1,258561
Chrysler	-0,095281	0,382895
Dodge	-0,085824	0,440538
Eagle	-0,896298	1,998454
Ford	-0,567803	1,143498
Honda	0,074216	0,339667
Isuzu	-0,722437	2,075560
Mazda	-0,110307	0,955851
Mercedes	0,186282	0,494108
Mitsub.	-0,086657	0,709643
Nissan	0,179791	0,522402
Olds	0,311310	1,141420
Pontiac	0,188268	0,501586
Porsche	0,107157	2,139045
Saab	0,359575	0,269914
Toyota	0,254400	0,787689
VW	-0,031192	0,419548
Volvo	0,225639	0,302917

Рис. 11.

Кластеризация методом К – средних – K – means clustering

Щелчком по строке – **K – means clustering** (Кластеризация методом k-средних) стартовой панели модуля **Cluster analysis** (Кластерный анализ) (рис. 4). На экране появится окно настройки параметров кластеризации (рис. 12).

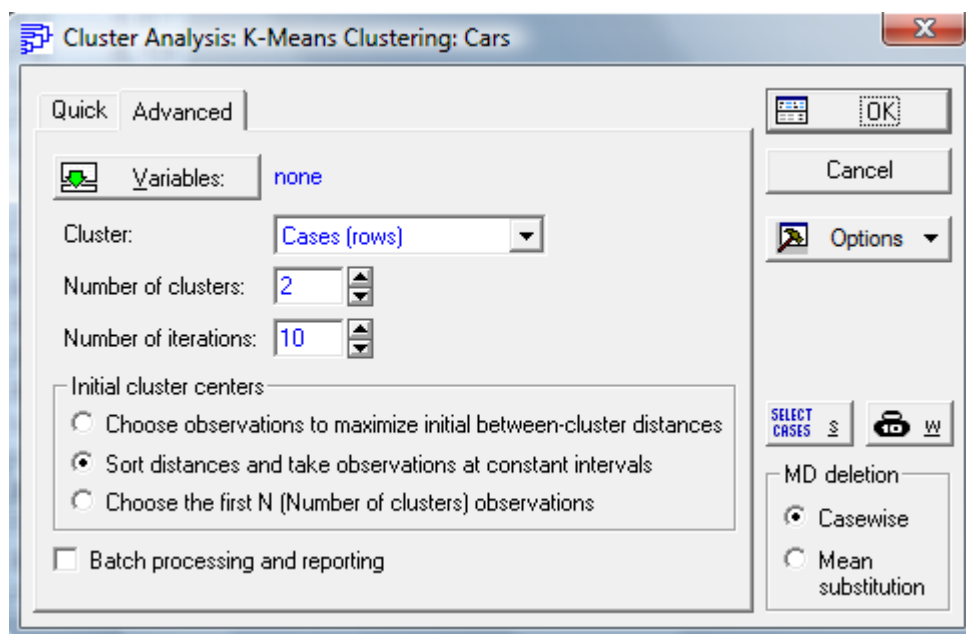


Рис. 12.

Поле **Number of clusters** (Число кластеров) позволяет ввести желаемое число кластеров, которое должно быть больше 1 и меньше чем количество объектов.

Метод k-средних является итерационной процедурой, в результате которой на каждой итерации объекты перемещаются в различные кластеры. Поле **Number of iterations** (Число итераций) предназначено для указания их максимального числа.

Важным моментом при настройке параметров является выбор **Initial cluster centers** (Начальных центров кластеров), так как конечные результаты зависят от начальной конфигурации.

Опция **Choose observations to maximize initial between-cluster distances** (Выбрать наблюдения, максимизирующие начальные расстояния между кластерами) выбирает первые k в соответствии с количеством кластеров, наблюдений, которые служат центрами кластеров. Последующие наблюдения заменяют ранее выбранные центры в том случае, если наименьшее расстояние до любого из них больше, чем наименьшее расстояние между кластерами. В результате этой процедуры начальные расстояния между кластерами максимизируются.

Если выбрана опция **Sort distances and take observations at constant intervals** (Сортировать расстояния и выбрать наблюдения на постоянных интервалах), то сначала сортируются расстояния между всеми объектами, а затем в качестве начальных центров кластеров выбираются наблюдения на постоянных интервалах.

Choose the first N (Number of cluster) (Выбрать первые N [количество кластеров] наблюдений). Эта опция берет первые N (количество кластеров) наблюдений в качестве начальных центров кластеров.

После соответствующего выбора нажмем кнопку **OK**. STATISTICA

произведет вычисления и появится новое окно: "**К - Means Clustering Results**" (рис. 13).

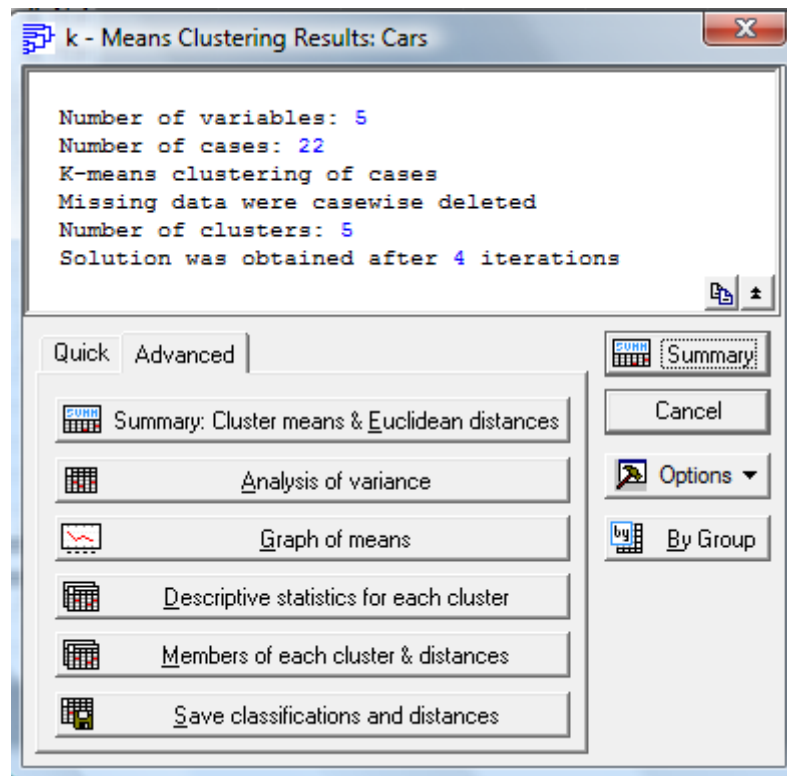


Рис. 13.

В верхней части содержатся значения параметров, по которым проводится анализ, а в нижней – кнопки для вывода результатов.

В верхней части окна (в том же порядке, как они идут на экране):

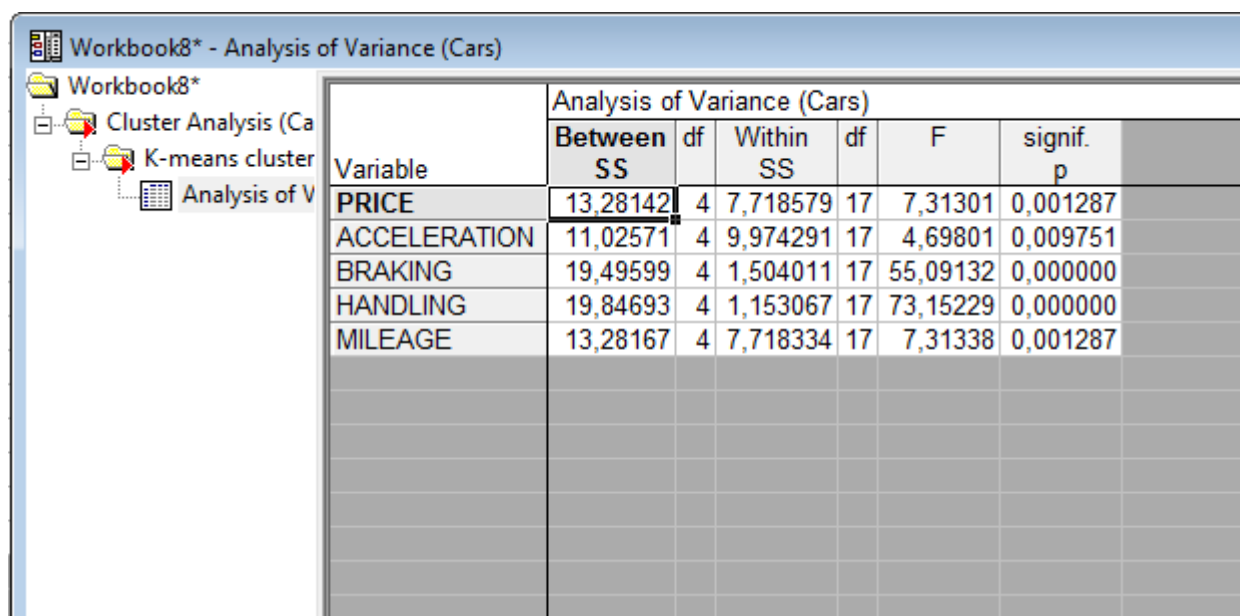
- Количество переменных - 5;
- Количество наблюдений - 22;
- Классификация наблюдений (или переменных, зависит от установки в предыдущем окне в строке **Cluster**) методом К - средних;

- Наблюдения с пропущенными данными удаляются или изменяются средними значениями. Зависит от установки в предыдущем окне в строке **MD deletion**.

- Количество кластеров - 5;
- Решение достигнуто после: 4 итераций.

В нижней части окна расположены кнопки для вывода различной информации по кластерам.

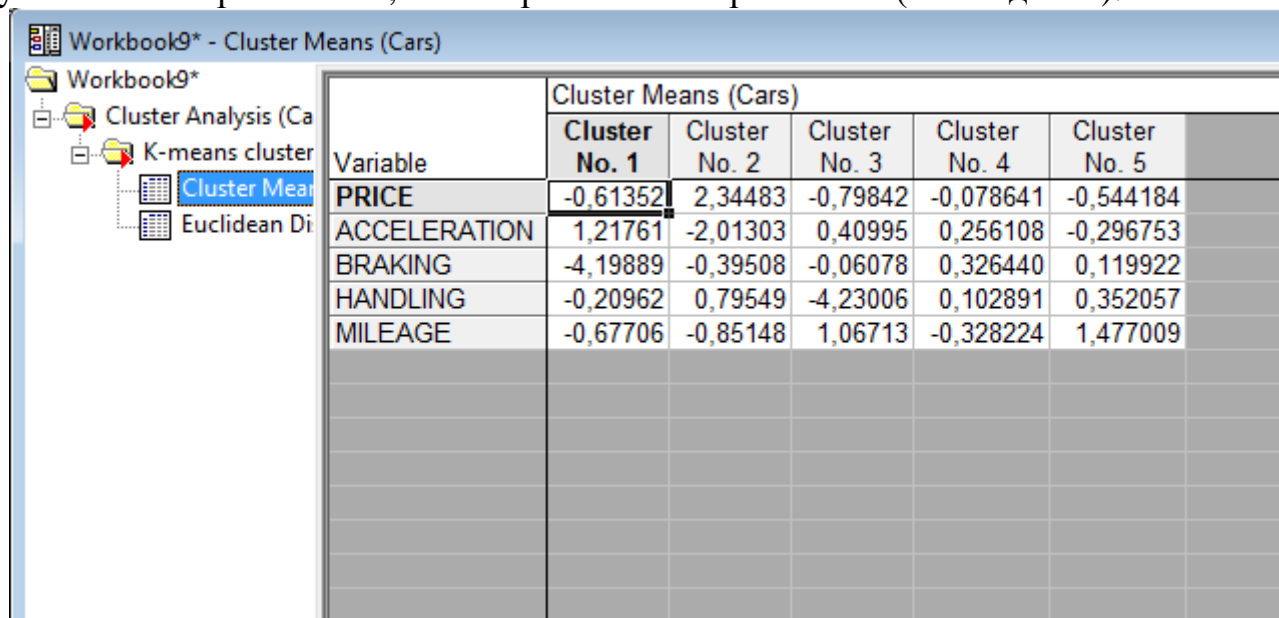
Analysis of Variance (Дисперсионный анализ). После нажатия появляется таблица, в которой приведена межгрупповая и внутригрупповая дисперсии (рис. 14). Где строки – переменные (наблюдения), столбцы – показатели для каждой переменной: дисперсия между кластерами, число степеней свободы для межклассовой дисперсии, дисперсия внутри кластеров, число степеней свободы для внутриклассовой дисперсии, F-критерий для проверки гипотезы о неравенстве дисперсий.



Variable	Between SS	df	Within SS	df	F	signif. p
PRICE	13,28142	4	7,718579	17	7,31301	0,001287
ACCELERATION	11,02571	4	9,974291	17	4,69801	0,009751
BRAKING	19,49599	4	1,504011	17	55,09132	0,000000
HANDLING	19,84693	4	1,153067	17	73,15229	0,000000
MILEAGE	13,28167	4	7,718334	17	7,31338	0,001287

Рис. 14.

Cluster Means & Euclidean Distances (средние значения в кластерах и евклидово расстояние). Выводятся две таблицы. В первой (рис. 15) указаны средние величины класса по всем переменным (наблюдениям). По вертикали указаны номера классов, а по горизонтали переменные (наблюдения).



Variable	Cluster No. 1	Cluster No. 2	Cluster No. 3	Cluster No. 4	Cluster No. 5
PRICE	-0,61352	2,34483	-0,79842	-0,078641	-0,544184
ACCELERATION	1,21761	-2,01303	0,40995	0,256108	-0,296753
BRAKING	-4,19889	-0,39508	-0,06078	0,326440	0,119922
HANDLING	-0,20962	0,79549	-4,23006	0,102891	0,352057
MILEAGE	-0,67706	-0,85148	1,06713	-0,328224	1,477009

Рис. 15.

Во второй таблице (рис. 16) приведены расстояния между классами. И по вертикали и по горизонтали указаны номера кластеров. Таким образом, при пересечении строк и столбцов указаны расстояния между соответствующими классами. Причем выше диагонали (на которой стоят нули) указаны квадраты, а ниже просто евклидово расстояние.

Workbook9* - Euclidean Distances between Clusters (Cars)

Cluster Analysis (Ca)

K-means cluster

Cluster Mean

Euclidean Di

Euclidean Distances between Clusters (Cars)					
Distances below diagonal					
Squared distances above diagonal					
Cluster Number	No. 1	No. 2	No. 3	No. 4	No. 5
No. 1	0,000000	6,939689	7,403314	4,381713	5,181152
No. 2	2,634329	0,000000	8,959942	2,459250	3,435140
No. 3	2,720903	2,993316	0,000000	4,282629	4,352102
No. 4	2,093254	1,568200	2,069451	0,000000	0,777197
No. 5	2,276214	1,853413	2,086169	0,881588	0,000000

Рис. 16.

Щелкнув по кнопке **Graph of means** (График средних), можно получить графическое изображение информации содержащейся в таблице, выводимой при нажатии на кнопку **Analysis of Variance** (Дисперсионный анализ). На графике показаны средние значения переменных для каждого кластера (рис. 17).

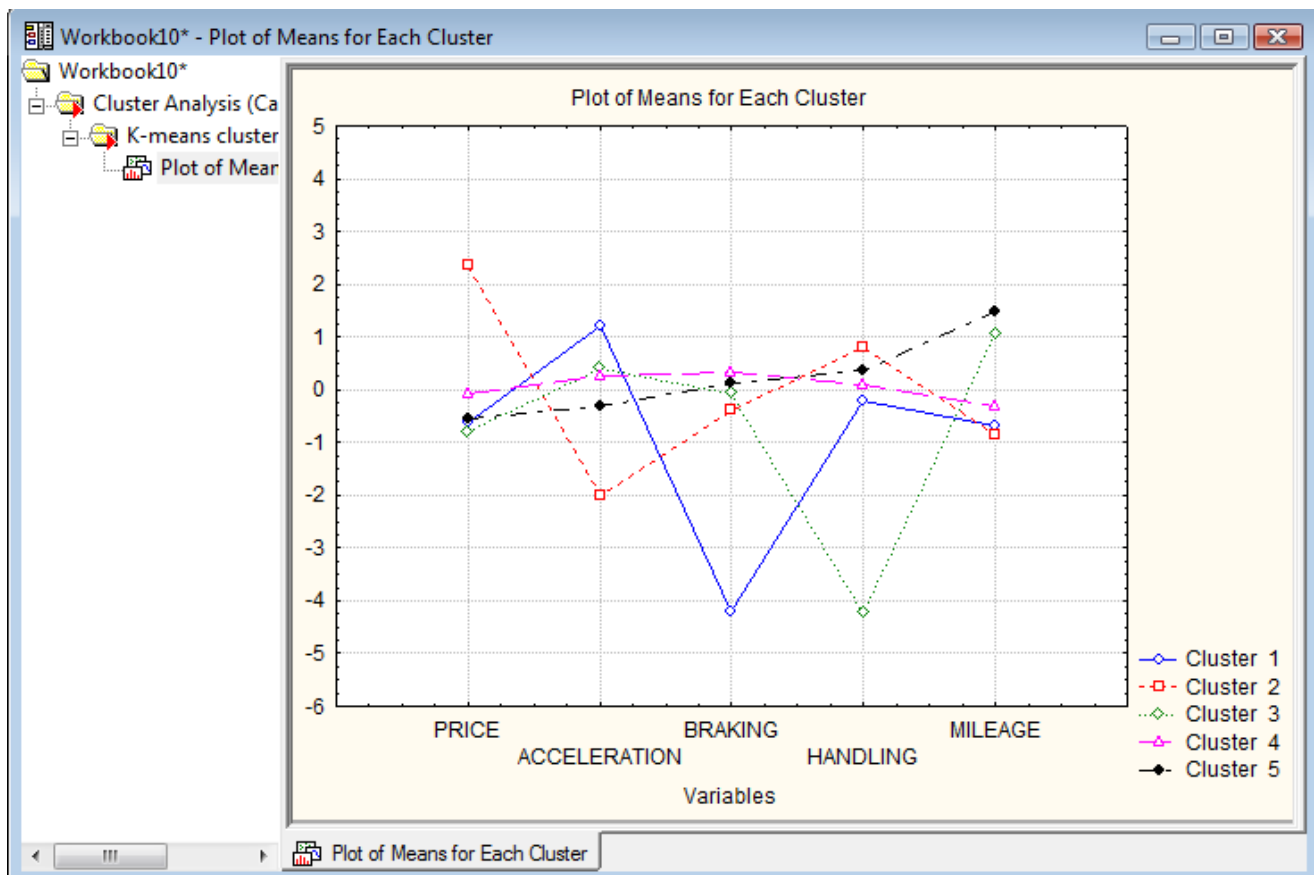
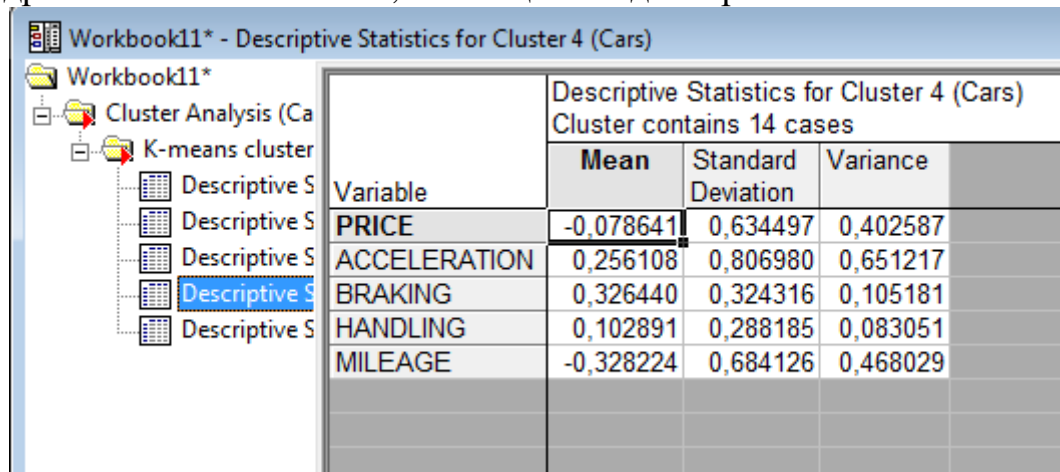


Рис. 17.

Descriptive Statistics for each cluster (Описательная статистика для

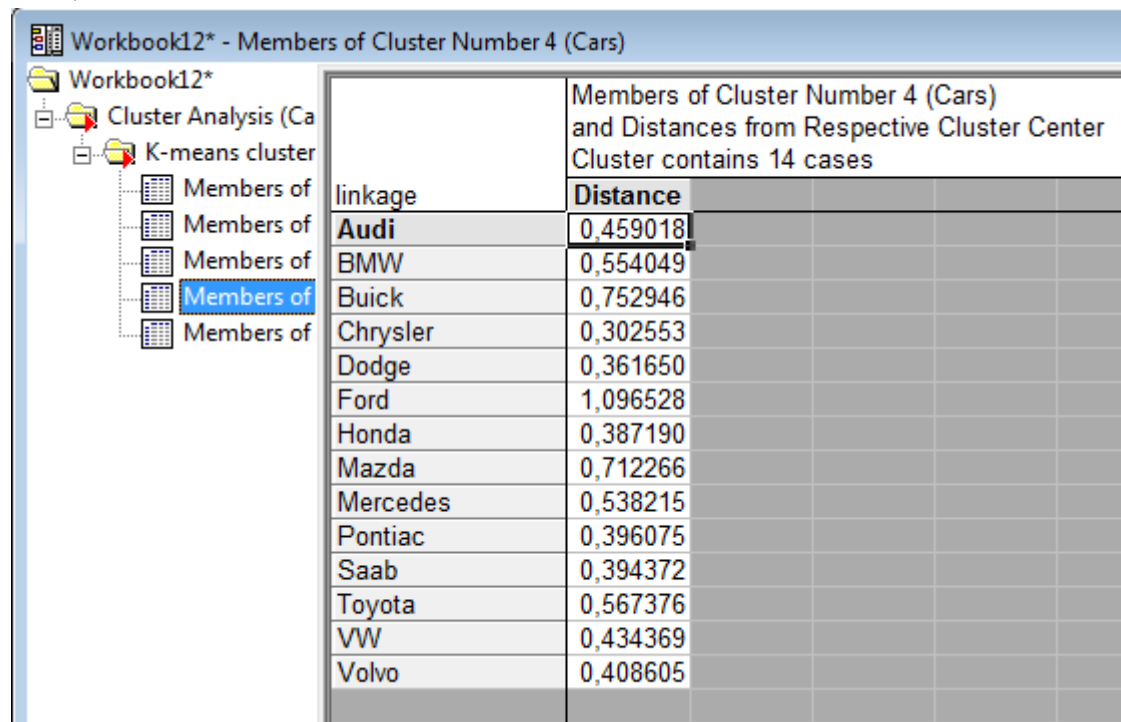
каждого кластера). После нажатия этой кнопки выводятся окна, количество которых равно количеству кластеров (рис. 18). В каждом таком окне в строках указаны переменные (наблюдения), а по горизонтали их характеристики, рассчитанные для данного класса: среднее, несмещенное среднеквадратическое отклонение, несмещенная дисперсия.



Descriptive Statistics for Cluster 4 (Cars) Cluster contains 14 cases			
Variable	Mean	Standard Deviation	Variance
PRICE	-0,078641	0,634497	0,402587
ACCELERATION	0,256108	0,806980	0,651217
BRAKING	0,326440	0,324316	0,105181
HANDLING	0,102891	0,288185	0,083051
MILEAGE	-0,328224	0,684126	0,468029

Рис. 18.

Members for each cluster & distances (Члены каждой группы и расстояния). Выводится столько окон, сколько задано классов (рис. 19). В каждом окне указывается общее число элементов, отнесенных к этому кластеру, в первом столбце указан номер наблюдения (переменной), отнесенной к данному классу и евклидово расстояние от центра класса до этого наблюдения (переменной). Центр класса - средние величины по всем переменным (наблюдениям) для этого класса.



Members of Cluster Number 4 (Cars) and Distances from Respective Cluster Center Cluster contains 14 cases	
linkage	Distance
Audi	0,459018
BMW	0,554049
Buick	0,752946
Chrysler	0,302553
Dodge	0,361650
Ford	1,096528
Honda	0,387190
Mazda	0,712266
Mercedes	0,538215
Pontiac	0,396075
Saab	0,394372
Toyota	0,567376
VW	0,434369
Volvo	0,408605

Рис. 19.

Save classifications and distances. Позволяет сохранить в формате

программы статистика таблицу, в которой содержатся значения всех переменных, их порядковые номера, номера кластеров к которым они отнесены, и евклидовы расстояния от центра кластера до наблюдения. Записанная таблица может быть вызвана любым блоком или подвергнута дальнейшей обработке.

Задание.

Задание 1.

По данным Таблицы 1 и алгоритму кластерного анализа провести классификацию объектов иерархическим методом (древовидная кластеризация).

Таблица 1.

№ п/п.	Страны	Число врачей на 10000 населения	Смертность на 100000 населения	ВВП по паритету покупательной способности, в % к США	Расходы на здравоохранение, в % к США
		X1	X2	X3	X4
1	Россия	44.5	84.98	20.4	3.2
2	Австралия	32.5	30.58	71.4	8.5
3	Австрия	33.9	38.42	78.7	9.2
4	Азербайджан	38.8	60.34	12.1	3.3
5	Армения	34.4	60.22	10.9	3.2
6	Беларусь	43.6	60.79	20.4	5.4
7	Бельгия	41	29.82	79.7	8.3
8	Болгария	36.4	70.57	17.3	5.4
9	Великобритания	17.9	34.51	69.7	7.1
10	Венгрия	32.1	64.73	24.5	6
11	Германия	38.1	36.63	76.2	8.6
12	Греция	41.5	32.84	44.4	5.7
13	Грузия	55	62.64	11.3	3.5
14	Дания	36.7	34.07	79.2	6.7
15	Ирландия	15.8	39.27	57	6.7
16	Испания	40.9	28.46	54.8	7.3
17	Италия	49.4	30.27	72.1	8.5
18	Казахстан	38.1	69.04	13.4	3.3
19	Канада	27.6	25.42	79.9	10.2
20	Киргизия	33.2	53.13	11.2	3.4

В качестве расстояния между объектами принять “обычное евклидово расстояние”, а расстояния между кластерами измерять по принципу: “ближайшего соседа”.

Исходные данные не нормировать.

Номер варианта соответствует номеру строки исключаемой из таблицы данных. Т.е. исследования проводятся для всех стран, кроме той, номер строки которой соответствует вашему варианту.

Задание 2.

Решить Задание 1, предварительно нормировав исходные данные.

Задание 3.

Решить Задание 1 при условии, что расстояния между кластерами измеряются по принципу “дальнего соседа”, предварительно нормируя исходные данные.

Задание 4.

Решить Задание 1, но в качестве расстояния между объектами принять “расстояние городских кварталов (Манхэттенское расстояние)”, а расстояния между кластерами измерять по методу Варда. Не нормируя предварительно исходные данные.

Задание 5.

Решить Задание 1 методом К-средних. Предварительно нормируя исходные данные.