**ЛЕКЦИЯ 9 КЛАСТЕРНЫЙ АНАЛИЗ**

*1 Понятие о кластерном анализе*

*2 Методы кластерного анализа*

3 *Алгоритм кластерного анализа*

*4 Реализация кластерного анализа в системе STATISTICA*

**1 Понятие о кластерном анализе**

Кластерный анализ объединяет различные процедуры, используемые для проведения классификации. В результате применения этих процедур исходная совокупность объектов разделяется на *кластеры* или *группы* (классы) схожих между собой объектов. *Кластер* – это группа объектов, обладающую свойством плотности (плотность объектов внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой (например, кластер может иметь очертания гиперсферы или эллипсоида), размером. Конечно, данное определение не является строгим (строгого определения не существует вообще). Если вы взглянете на географическую карту и увидите на ней горы или созвездия на звездном небе, то поймете, что такое кластеры.

Наиболее часто методы кластерного анализа используются в социологии, маркетинговых исследованиях, экономике, биологии, медицине, археологии.

Сложность задач кластерного анализа состоит в том, что реальные объекты являются многомерными, то есть описываются не одним, а несколькими параметрами (представьте, что объекты – это персональные компьютеры), и объединение объектов в группы проводится в пространстве многих измерений, что весьма нетривиально. Кроме того, данные могут носить нечисловой характер.

В целом **методы кластеризации** делятся на *агломеративные* (от слова агломерат – скопление) и *итеративные* *дивизивные* (от слова division –деление, разделение).

В агломеративных, или объединительных методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде дендрограммы, или дерева объединения. Это удобное представление позволяет наглядно представить кластеризацию агломеративными алгоритмами.

Исходными данными для анализа могут быть собственно объекты и их параметры. Данные для анализа могут быть также представлены матрицей расстояний между объектами, в которой на пересечении строки с номером i и столбца с номером j записано расстояние между i-м и j-м объектом.

Если расстояния не даны сразу, то агломеративные алгоритмы начинаются с вычисления расстояний между объектами.

Переход от объектов к расстояниям между объектами – важный момент.

Расстояние между объектами – одна из мер сходства (!). Интуитивно понятно, что, чем меньше расстояние между объектами, тем они более схожи. Но как выбрать естественную метрику, то есть, как естественно для данной задачи измерить расстояние между объектами?

Часто используют обычную **евклидову метрику**, например, если объект описывается двумя параметрами, то он может быть изображен точкой на плоскости, а расстояние между объектами – это расстояние между точками, вычисленное по теореме Пифагора. Вы просто возводите в квадрат расстояния по каждой координате, суммируете их и из полученной суммы извлекаете квадратный корень. Если вы не будете возводить в квадрат покоординатные расстояния, а просто возьмете их абсолютные значения и просуммируете, то получите так называемое **манхэттенское расстояние**, или «расстояние городских кварталов». Такое расстояние связано с перемещением человека по улицам города, а не с движением по ровной местности.

Представьте, что вы находитесь в городе. Здесь существуют определенные правила перемещения и, соответственно, правила вычисления пройденного расстояния. Перемещаться можно только по улицам (нельзя, например, пересечь квартал или дом по диагонали). Аналогия в декартовой плоскости приводит к перемещениям только по линиям, параллельным осям координат, и, соответственно, к манхэттенскому расстоянию.

**2 Методы кластерного анализа**

В практике обычно реализуются агломеративные методы кластеризации.

Обычно перед началом классификации данные стандартизуются (вычитается среднее и производится деление на корень квадратный из дисперсии). Полученные в результате стандартизации переменные имеют нулевое среднее и единичную дисперсию.

Можно выбрать следующие правила иерархического объединения кластеров:

- метод одиночной связи;

- метод полной связи;

- невзвешенный метод «средней связи»;

- взвешенный метод «средней связи»;

- взвешенный центроидный метод;

- метод Уорда.

Данные алгоритмы различаются правилами объединения объектов в кластеры.

В **методе одиночной связи** на первом шаге объединяются два объекта, имеющие между собой максимальную меру сходства. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с одним из объектов кластера. Таким образом, процесс продолжается далее. Итак, *для включения объекта в кластер требуется максимальное сходство лишь с одним членом кластера*. Отсюда и название метода одиночной связи, нужна только одна связь, чтобы присоединить объект к кластеру: связь нового элемента с кластером определяется только по одному из элементов кластера. Недостатком этого метода является образование слишком больших «продолговатых» кластеров.

**Метод полных связей** позволяет устранить указанный недостаток. Здесь *мера сходства между объектом – кандидатом на включение в кластер и всеми членами кластера не может быть меньше некоторого порогового значения*.

**В методе средней связи** *мера сходства между кандидатом и членами кластера усредняется*, например, берется просто среднее арифметическое мер сходства.

Идея **метода Уорда** состоит в том, чтобы *проводить объединение, дающее минимальное приращение внутригрупповой суммы квадратов отклонений*. Замечено, что метод Уорда приводит к образованию кластеров примерно равных размеров и имеющих форму гиперсфер.

Рассмотрим еще **итеративный метод группировки k-средних**. Данный метод работает непосредственно с объектами, а не с матрицей сходства.

В методе k-средних объект относится к тому классу, расстояние до которого минимально. Расстояние понимается как евклидово расстояние, то есть объекты рассматриваются как точки евклидова пространства.

Как определить евклидово расстояние, мы уже знаем. Но как определить расстояние от объекта до совокупности объектов? Это можно сделать следующим способом**: каждый класс объектов имеет центр тяжести** (рассмотрите, как и ранее, простейший случай – представьте, что объект имеет только два параметра, тогда его можно изобразить точкой на плоскости, а группа объектов – это просто группа точек).

Расстояние между объектом и классом есть расстояние между объектом и центром класса. Но как вычислить центр класса? Например, взять средние по каждому параметру. Тогда расстояние между объектом и группой объектов вполне определено и алгоритм может работать.

Представьте, что число объектов в группе равно 2. Соедините эти точки отрезком прямой и найдите его середину. Это и будет центр тяжести группы, состоящей из двух точек. Расстояние от этого центра до исходной точки будет искомым расстоянием.

Принципиально метод k-средних «работает» следующим образом:

1) задается некоторое разбиение данных на кластеры (число кластеров определяется заранее); вычисляются центры тяжести кластеров;

2) происходит перемещение точек: каждая точка помещается в ближайший к ней кластер;

3) вычисляются центры тяжести новых кластеров;

4) шаги 2, 3 повторяются, пока не будет найдена стабильная конфигурация (то есть кластеры перестанут изменяться) или число итераций не превысит заданное пользователем. Итоговая конфигурация и является искомой.

**3 Алгоритм кластерного анализа**

Кластерный анализ – это совокупность методов классификации многомерных наблюдений или объектов, основанных на определении понятия расстояния между объектами с последующим выделением из них групп, «сгустков» наблюдений (кластеров, таксонов). При этом не требуется априорной информации о распределении генеральной совокупности.

Выбор конкретного метода кластерного анализа зависит от цели классификации.

Кластерный анализ используется при исследовании структуры каких–либо совокупностей.

От матрицы исходных данных



переходят к матрице нормированных значений Z с элементами:

,

где: j = 1, 2, 3, 4 – номер показателя, i = 1,2,..., n – номер наблюдения;

;



В качестве расстояния между двумя наблюдениями zi и zν используется «взвешенное» евклидово расстояние, определяемое по формуле:



Полученные значения удобно представить в виде матрицы расстояний:

,



Так как матрица R симметрическая, т.е. , то достаточно ограничиться записью наддиагональных элементов матрицы.

Используя матрицу расстояний, можно реализовать агломеративную иерархическую процедуру кластерного анализа. Расстояния между кластерами определяют по принципу «ближайшего соседа» или «дальнего соседа». В первом случае за расстояние между кластерами принимают расстояние между ближайшими элементами этих кластеров, а во втором – между наиболее удаленными друг от друга.

Принцип работы иерархических агломеративных процедур состоит в последовательном объединении групп элементов сначала самых близких, а затем все более отдаленных друг от друга.

На первом шаге алгоритма каждое наблюдение zi (*i* = 1, 2,..., *n*) рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и вновь строится матрица расстояний, размерность которой снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс.

**4 Реализация кластерного анализа в системе STATISTICA**

В программе STATISTICA доступны следующие меры сходства объектов: евклидова метрика, квадрат евклидовой метрики, манхэттенское расстояние, или «расстояние городских кварталов», метрика Чебышева, метрика Минковского, пирсоновский коэффициент корреляции (точнее, 1–пирсоновский коэффициент корреляции), коэффициент совстречаемости (точнее, 1–коэффициент совстречаемости).

В STATISTICA реализованы следующие методы кластеризации – агломеративные методы: **joining (tree clustering)**, **two-way joining**, а также метод **k-средних (k-means clustering)**.

Обычно перед началом классификации данные стандарти-зуются (вычитается среднее и производится деление на корень квадратный из дисперсии). Полученные в результате стандарти-зации переменные имеют нулевое среднее и единичную диспер-сию. Рассматриваемые далее данные уже стандартизованы.

В STATISTICA можно выбрать следующие правила иерархи-ческого объединения кластеров:

− **Single linkage** – метод одиночной связи;

− **Complete linkage** – метод полной связи;

− **Unweighted pair group average** – невзвешенный метод «средней связи»;

− **Weighted pair group average** – взвешенный метод «средней связи»;

− **Weighted centroid pair group (median)** – взвешенный центроидный метод;

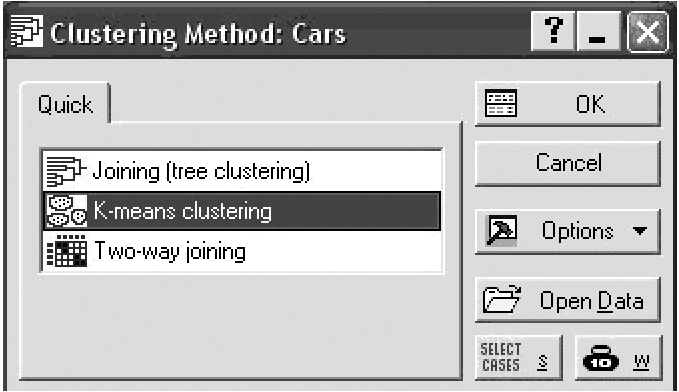
− **Ward method** – метод Уорда.

|  |
| --- |
|  |
| *Рис. 1 – Запуск модуля Кластерный анализ* |

В примере рассма-триваются автомобили разных марок, которые различаются ценой, рас-ходом горючего и неко-торыми техническими характеристиками.

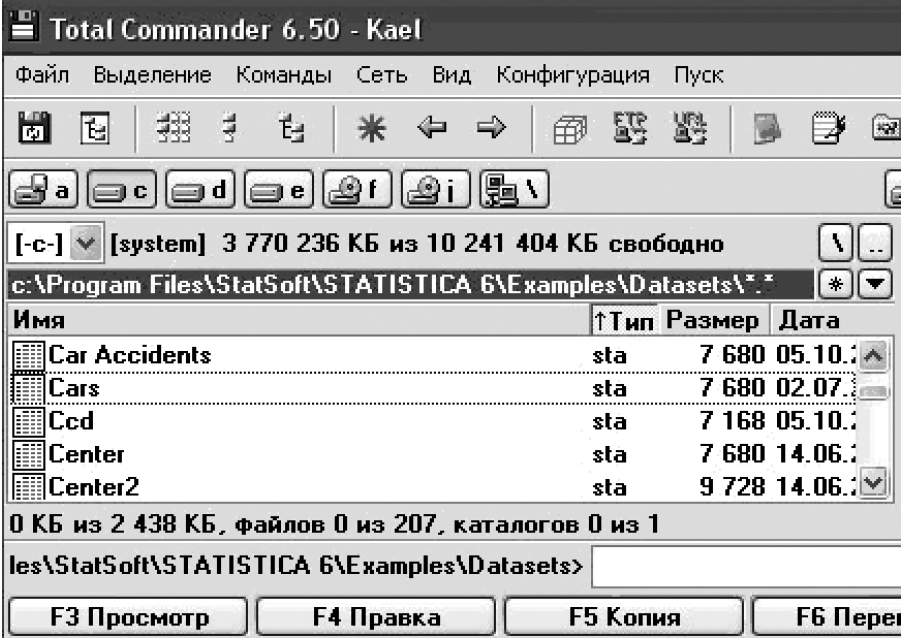
В рабочем окне STATISTICA выберете название модуля – **Cluster Analysis (клас-терный анализ)**, высветите его имя и щелкните на его имени (рис. 1).

На экране появится стартовая панель модуля **Claster Analysis (кластерный анализ)** (рис. 2).



*Рис. 2 – Стартовая панель модуля Кластерный анализ*

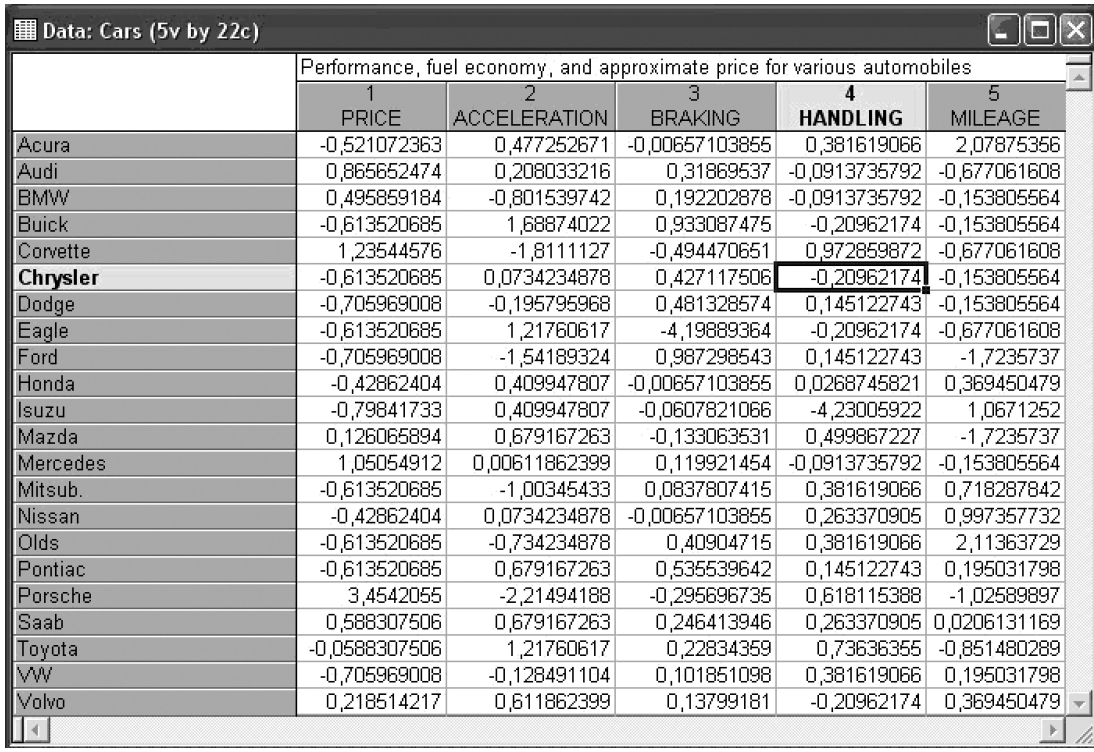
Стандартным образом, нажав кнопку Open Data, откройте окно выбора файла (рисунок 3).



*Рисунок 3 – Выбор файла с данными об автомобилях*

Выберите в этом окне файл **Cars.sta** и дважды нажмите левую кнопку мыши. Файл выбран, и вы вернетесь обратно, в стартовую панель модуля.

В рабочем окне, сзади стартовой панели, вы увидите открытый файл с данными (рисунок 4).



*Рисунок 4 – Файл cars.sta с данными автомобилей разных марок*

Из информации в верхней части окна видно, что в файле **Cars.sta** записаны цена автомобиля, технические характеристики, количество миль, пройденных на одном галлоне бензина.

Всего в файле содержатся данные о 22 машинах разных марок. Марки машин – это случаи.

Переменные в этом файле:

- **PRICE** – цена;

- **ACCELE** - **HANDLI** – технические характеристики;

- **MILAGE** – расход горючего (количество миль, пройденных на одном галлоне бензина).

Все характеристики машин уже стандартизованы (из значений переменной price вычтена средняя цена и разность поделена на корень квадратный из дисперсии).

Задача состоит в том, чтобы разбить автомобили на несколько групп, в которых автомобили мало отличаются друг от друга (существенно меньше, чем в целом в совокупности).

Задача эта сложна, так как сравниваются машины не по какому-то одному параметру, а по нескольким параметрам одновременно. По одним характеристикам одни машины близки друг к другу, по другим – другие. В конечном итоге разбиение на группы – тоже не самоцель. Конечно, число параметров можно увеличить. Очевидно, разбив машины на группы, можно лучше в целом представить их совокупность, с тем, чтобы затем более обоснованно принимать решение, например при покупке или обмене одной машины на другую.

Если бы машины сравнивались по одному параметру, например по расходу горючего, то, наверное, следовало бы выбрать машину с меньшим расходом топлива на одну милю. Все машины были бы упорядочены в одну линию, и задача не представляла бы проблем.

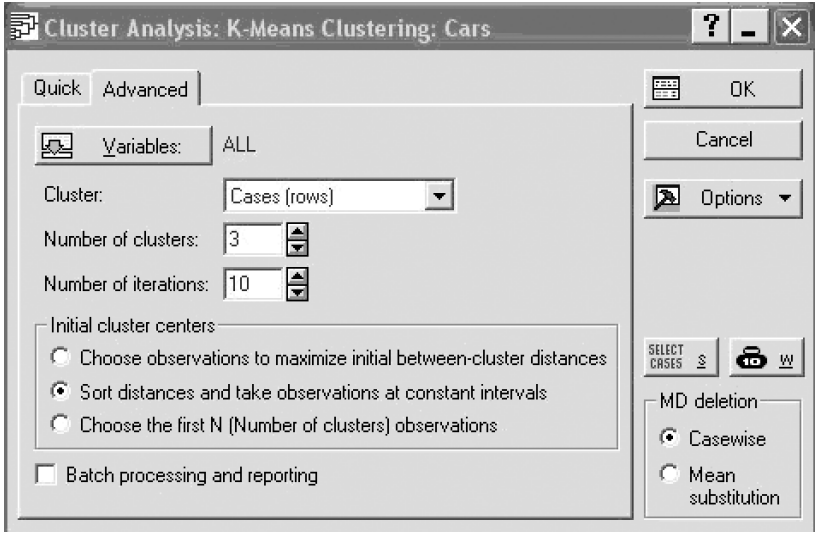
Однако параметров несколько, и ситуация существенно усложняется.

**Выбор метода**

Посмотрите на стартовую панель. В главной ее части находится список методов кластерного анализа, реализованных в STATISTICA.

В списке методов высветите **k-means (k-средних)** (рисунок 2) и нажмите кнопку **OK** в правом верхнем углу панели.

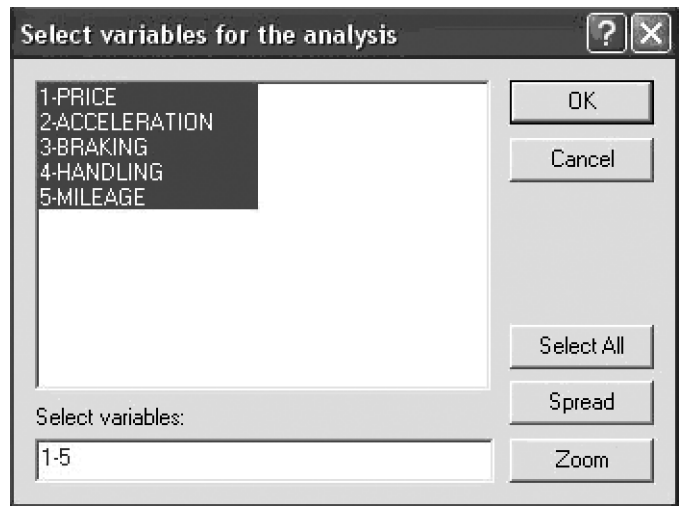
Диалоговое окно метода **k-means** появится на экране (рисунок 5).



*Рисунок 5 – Диалоговое окно метода k-means*

Начните работать в данном окне. Прежде всего, выберите переменные для анализа.

Нажмите кнопку **Variables (переменные)** в левом верхнем углу текущего окна и откройте диалоговое окно **Select variable for the analysis (выбрать переменные для анализа)** (рисунок 6).



*Рисунок 6 – Выбор переменных для кластерного анализа*

Так как машины разбиты на группы и учитываются все параметры, то нажмите вначале кнопку **Select All (выбрать все)**, а затем нажмите кнопку **OK**.

Посмотрите далее на поле **Cluster (кластер)**, находящееся ниже кнопки **Variables (переменные)** (рисунок 5). Нажав на стрелку в этом поле, выберите пункт меню **Cases (случаи)**. Альтернативный выбор был бы **Variables (переменные)**. Так следует поступить, если нужно кластеризировать переменные.

В данном примере кластеризируются машины, которые являются случаями в исходном файле данных, поэтому выбирается пункт **Cases**.

В поле **Number of clusters (число кластеров)** нужно определить число групп, на которые необходимо разбить автомобили. Запишите в это поле число 3 (таким образом, машины разбиваются на 3 кластера).

В строке **Number of iterations (Число итераций)** задается максимальное число итераций, используемых при построении классов. Задайте, например, число 11.

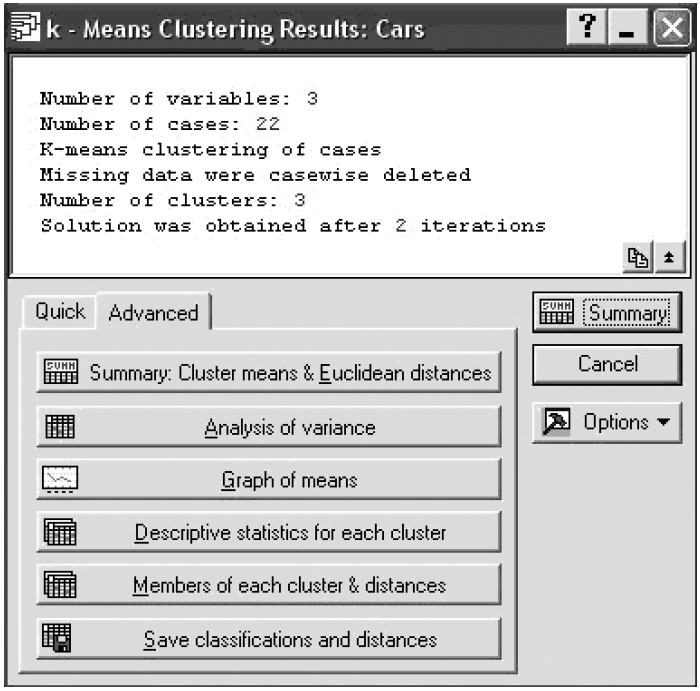
В строке **Missing data** задается способ обработки пропущенных значений в данных (например, для какой-то машины отсутствует значение некоторого параметра). В данном примере пропусков в данных нет и обработки пропущенных значений не происходит.

Группа опций **Initial cluster centers** позволяет задать начальные центры кластеров.

Сделайте установки, как показано на рисунке 5.

После того как все установки сделаны, нажмите кнопку **ОК** в верхнем правом углу окна **k-means Clustering** и запустите вычислительную процедуру.

Спустя несколько секунд после нажатия кнопки **ОК** в **k-means Clustering** окно результатов появится на экране (рисунок 7).



*Рисунок 7 – Окно результатов кластеризации машин по методу средних*

В верхней части окна записана информация: число переменных, число случаев, метод кластеризации, число кластеров, а также сообщение о том, после скольких итераций найдено решение: **Solution was obtained after 2 iterations (решение найдено после 2 итераций)**.

Кнопки в нижней части окна позволяют провести анализ результатов кластеризации.

Кнопка **Analysis of variation (дисперсионный анализ)** позволяет просмотреть таблицу дисперсионного анализа.

Кнопка **Cluster Means&Euclidean Distances** позволяет вывести таблицы, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй указаны евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

Кнопка **Graph of means** позволяет посмотреть средние значения для каждого кластера на линейном графике.

Кнопка **Descriptive Statistics for each clusters** открывает электронную таблицу с описательными статистиками для каждого кластера (среднее, дисперсия и т. д.)

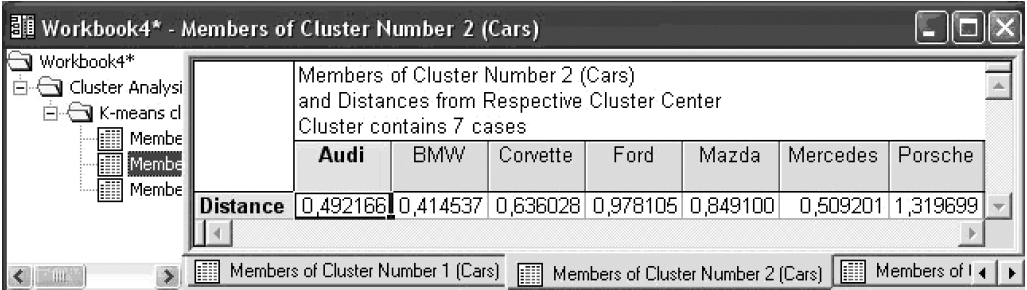
Кнопка **Save classifications and distances** позволяет сохранить результаты классификации в файле STATISTICA для дальнейшего исследования.

Следует посмотреть, как распределились машины по кластерам.

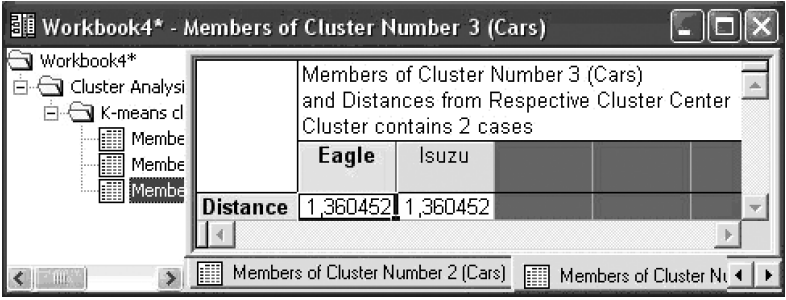
Нажмите для этого кнопку **Member of each cluster&distances**. На экране появятся 3 электронные таблицы с названиями машин, отнесенных к определенным кластерам (рисунки 8–10).



*Рисунок 8 – Первый кластер*



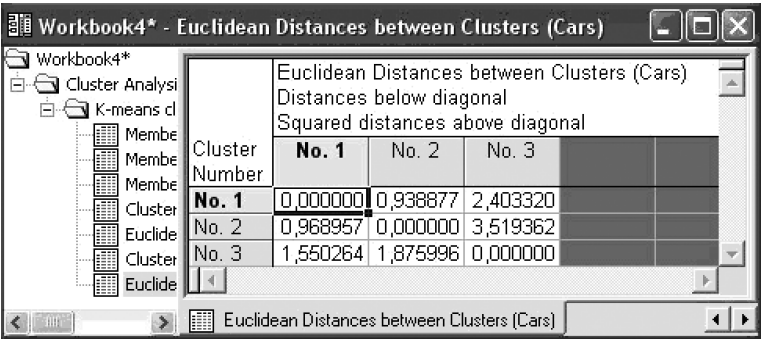
*Рисунок 9 – Второй кластер*



*Рисунок 10 – Третий кластер*

В строках таблиц указано расстояние от каждой машины до центра кластера.

Нажмите на кнопку **Cluster means&Euclidean distances**. На экране появится таблица, в которой даны евклидовы расстояния между средними кластеров (по каждому из параметров внутри кластера вычисляется среднее, получается 3 точки в пятимерном пространстве, и между ними находится расстояние) (рисунок 11).

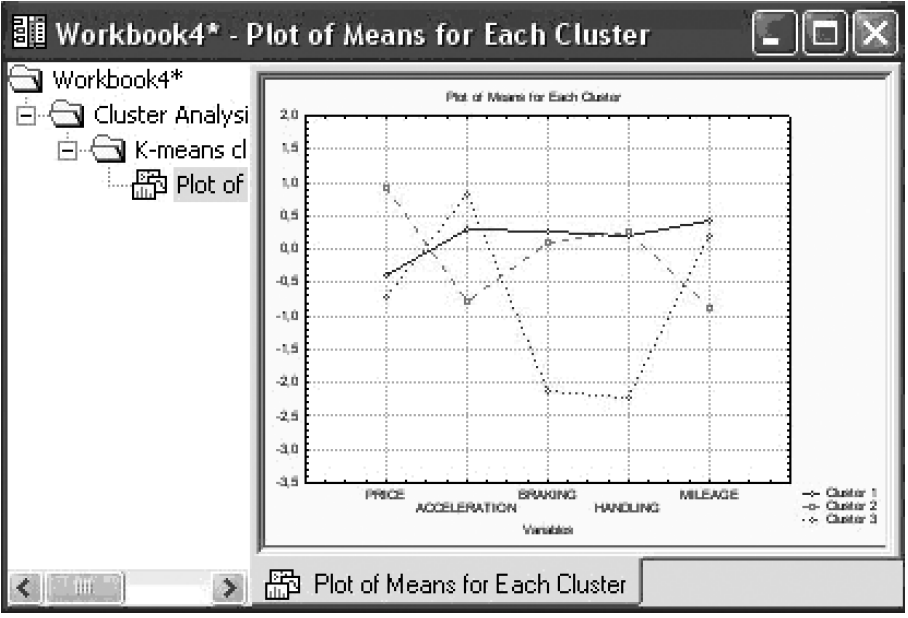


*Рисунок 11 – Расстояния между кластерами*

Из таблицы видно, что расстояние между кластерами даны под диагональю – между 1 и 2 кластером 0,969, а например, между вторым и третьим – 1,876.

Над диагональю в таблице даны квадраты расстояний между кластерами.

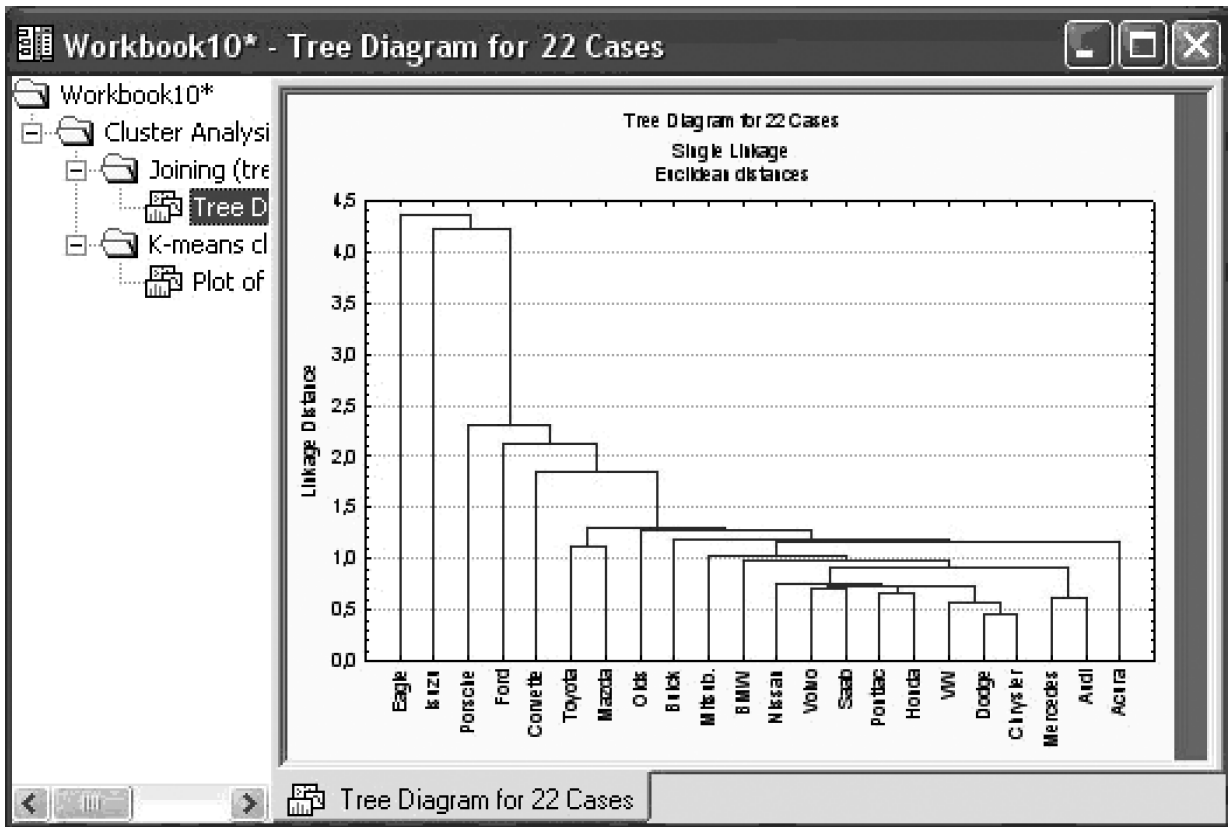
С помощью кнопки **Graph of means (график средних)** строятся следующие графики средних значений характеристик машин для каждого кластера (рисунок 12).



*Рисунок 12 – График средних для каждого кластера*

В системе реализованы также и другие методы кластеризации, в частности так называемый **two-way joining**, в котором кластеризируются случаи и переменные одновременно.

Если вы воспользуетесь **Joining (tree clustering)**, то сможете увидеть дендрограмму, или дерево объединения (рисунок 13), о котором говорилось вначале.



*Рисунок 13 – Дерево объединения машин разных марок в кластер*

*методом одиночной связи*