Федеральное агентство по образованию
Нижегородский государственный университет им. Н.И. Лобачевского

Национальный проект «Образование»
Инновационная образовательная программа ННГУ. Образовательно-научный центр «Информационно-телекоммуникационные системы: физические основы и математическое обеспечение»

Н.Н. Буреева

Многомерный статистический анализ с использованием ППП “STATISTICA”

Учебно-методические материалы по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики»

Нижний Новгород
2007

Учебно-методические материалы подготовлены в рамках инновационной образовательной программы ННГУ: *Образовательно-научный центр «Информационно-телекоммуникационные системы: физические основы и математическое обеспечение»*

Буреева Н.Н. Многомерный статистический анализ с использованием ППП “STATISTICA”. Учебно-методический материал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики». Нижний Новгород, 2007, 112 с.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала. В пособии рассмотрены следующие методы многомерного статистического анализа: регрессионный анализ, факторный анализ, дискриминантный анализ. Излагается структура пакета прикладных программ «Statistica», а также реализация в данном пакете изложенных методов многомерного статистического анализа.

© Буреева Н.Н., 2007

ОГЛАВЛЕНИЕ

Введение 5

1. Кластерный анализ и его реализация в пакете STATISTICA 8

1.1. Введение в кластерный анализ 8

1.1.1. Измерение близости объектов 11

1.1.2. Характеристики близости объектов 13

1.2. Методы кластерного анализа 16

1.2.1. Иерархические алгоритмы 16

1.2.2. Процедуры эталонного типа 21

1.3. Пример решения задачи кластерным анализом в пакете STATISTICA 22

1.4. Задания для самостоятельной работы 43

2. Факторный анализ и его реализация в пакете STATISTICA 46

2.1. Введение в факторный анализ 46

2.1.1. Цели и сущность факторного анализа 46

2.1.2. Линейная модель факторного анализа 48

2.1.3. Решение модели факторного анализа 52

2.1.4. Задача о числе факторов 54

2.1.5. Статистическая оценка факторных нагрузок и остаточных дисперсий 55

2.2. Пример решения задачи факторным анализом в пакете STATISTICA 61

2.3. Задания для самостоятельной работы 72

3. Дискриминантный анализ и его реализация в пакете STATISTICA 74

3.1. Введение в дискриминантный анализ 74

3.1.1. Проблема классификации 74

3.1.2. Дискриминирующие функции. Классификация объектов с помощью функции расстояния 79

3.2. Пример решения задачи дискриминантным анализом в пакете STATISTICA 88

3.3. Задания для самостоятельной работы 99

Заключение 101

Литература 103

Приложения 105

ВВЕДЕНИЕ

Социально-экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обуславливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т.е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

К области приложения математической статистики могут быть отнесены задачи, связанные с исследованием поведения индивидуума, семьи или другой социально-экономической или производственной единицы, как представителя большой совокупности объектов.

Многомерный экономико-статистический анализ опирается на широкий спектр методов. В учебном пособии рассматриваются некоторые из наиболее используемых методов, а именно: факторный, кластерный и дискриминантный анализы

Методы многомерной классификации, которые предназначены разделять рассматриваемые совокупности объектов, субъектов или явлений на группы в определенном смысле однородные. Необходимо учитывать, что каждый из рассматриваемых объектов характеризуется большим количеством разных и стохастически связанных признаков. Для решения столь сложных задач классификации применяют кластерный и дискриминантный анализ. Наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Такое решение может быть обеспечено методами снижения размерности, куда относятся факторный анализ. Этот метод позволяет учитывать эффект существенной многомерности данных, дает возможность лаконичного и более простого объяснения многомерных структур. Вскрывает объективно существующие,

непосредственно не наблюдаемые закономерности при помощи полученных факторов или главных компонент.

Это дает возможность достаточно просто и точно описать наблюдаемые исходные данные, структуру и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число факторов или главных компонент – новых единиц измерения – используется значительно меньше, чем исходных признаков.

Все перечисленные методы наиболее эффективны при активном применении статистических пакетов прикладных программ. При помощи этих пакетов предоставляется возможным даже восстанавливать пропущенные данные и др.

Стандартные статистические методы обработки данных включены в состав электронных таблиц, таких как Excel, Lotus 1-2-3, QuattroPro, и в математические пакеты общего назначения, например Mathcad. Но гораздо большими возможностями обладают специализированные статистические пакеты, позволяющие применять самые современные методы математической статистики для обработки данных. По официальным данным Международного статистического института, число статистических программных продуктов приближается к тысяче. Среди них есть профессиональные статистические пакеты, предназначенные для пользователей, хорошо знакомых с методами математической статистики, и есть пакеты, с которыми могут работать специалисты, не имеющие глубокой математической подготовки; есть пакеты отечественные и созданные зарубежными программистами; различаются программные продукты и по цене.

Среди программных средств данного типа можно выделить узкоспециализированные пакеты, в первую очередь статистические - STATISTICA, SPSS, STADIA, STATGRAPHICS, которые имеют большой набор статистических функций: факторный анализ, регрессионный анализ, кластерный анализ, многомерный анализ, критерии согласия и т. д. Данные программные продукты обычно содержат и средства для визуальной интерпретации полученных результатов: различные графики, диаграммы, представление данных на географической карте.

При анализе данных пользователю статистического программного пакета приходится выполнять вычисления широкого спектра статистик, передавать и преобразовывать данные для их анализа, а также представлять полученные результаты в наглядном виде. Поэтому при выборе того или иного статистического пакета, для сравнения пакетов, необходимо прежде всего обращать внимание на такие характеристики, как:

- удобство управления данными (экспорт/импорт данных, их реструктуризация);
- статистическое разнообразие (количество статистических модулей);
- графические возможности (наличие встроенного графического редактора, возможность показа отдельных элементов графика, возможности экспорта графиков).

Кроме того, большое значение имеет удобство работы с пакетом, легкость его освоения (наличие встроенной системы помощи, руководства пользователя, степень удобства управления данными, результатами вычислений, таблицами и графиками), а также скорость произведения вычислений.

Существуют также нестатистические пакеты, решающие задачи классификации (PolyAnalyst, ДА-система, АРГОНАВТ, ЛОРЕГ, пакет ОТЭКС и разнообразные нейросетевые пакеты).

Для реализации рассмотренных методов многомерной классификации выбрана достаточно популярная в нашей стране система STATISTICA.

ГЛАВА 1. КЛАСТЕРНЫЙ АНАЛИЗ И ЕГО РЕАЛИЗАЦИЯ В ПАКЕТЕ STATISTICA

1.1. Введение в кластерный анализ

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом (Tryon). Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы социально-экономической информации, делать их компактными и наглядными.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие (например, общехозяйственной и товарной конъюнктуры). Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл здесь может давать информацию, которая способна сильно изменить

направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

В задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с регрессионным анализом).

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения: В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

В кластерном анализе считается, что:

а) выбранные характеристики допускают в принципе желательное разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Выбор масштаба играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве *X*, разбить множество объектов *G* на m (m – целое) кластеров (подмножеств) $Q_1, Q_2, \mathbf{K}, Q_m$, так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения. А объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W = s_n = \sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2,$$

где x_j - представляет собой измерения j-го объекта.

Кластер имеет следующие *математические характеристики*: *центр, радиус, среднеквадратическое отклонение, размер кластера*.

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от *центра кластера*. Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют *спорными*.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до *центра кластера* меньше *радиуса кластера*. Если это условие выполняется для двух и более кластеров, объект является *спорным*.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. В начале лекции мы уже упоминали о сравнимости шкал, это и есть второе предположение - правильность выбора масштаба или единиц измерения признаков.

Выбор масштаба в кластерном анализе имеет большое значение. Рассмотрим пример. Представим себе, что данные признака x в наборе данных A на два порядка больше данных признака y: значения переменной x находятся в диапазоне от 100 до 700, а значения переменной y - в диапазоне от 0 до 1.

Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная x, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной y. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

Эта проблема решается при помощи предварительной *стандартизации* переменных. *Стандартизация* (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.:

$z=(x-\bar{x})/s$, $z=x/\bar{x}$, $z=x/x_{\max}$, $z=(x-\bar{x})/(x_{\max}-x_{\min})$,

где $\bar{x}$, s - соответственно среднее и среднеквадратическое отклонение x; $x_{\max}$, $x_{\min}$ - наибольшее и наименьшее значение x.

Наряду со *стандартизацией* переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В ходе экспериментов возможно сравнение результатов, полученных с учетом экспертных оценок и без них, и выбор лучшего из них.

1.1.1. Измерение близости объектов

Проблема измерения близости объектов неизбежно возникает при любых трактовках кластеров и различных методах классификации.

Отметим основные трудности, возникающие при этом: неоднозначность выбора способа нормировки и определения расстояния между объектами. Рассмотрим результаты небольшого обследования. Студенты группы записывают свои данные (вес, рост), оформляют в таблицу и строят по ним корреляционное поле. Масштабы по осям выбираются произвольно (рис.1.1).

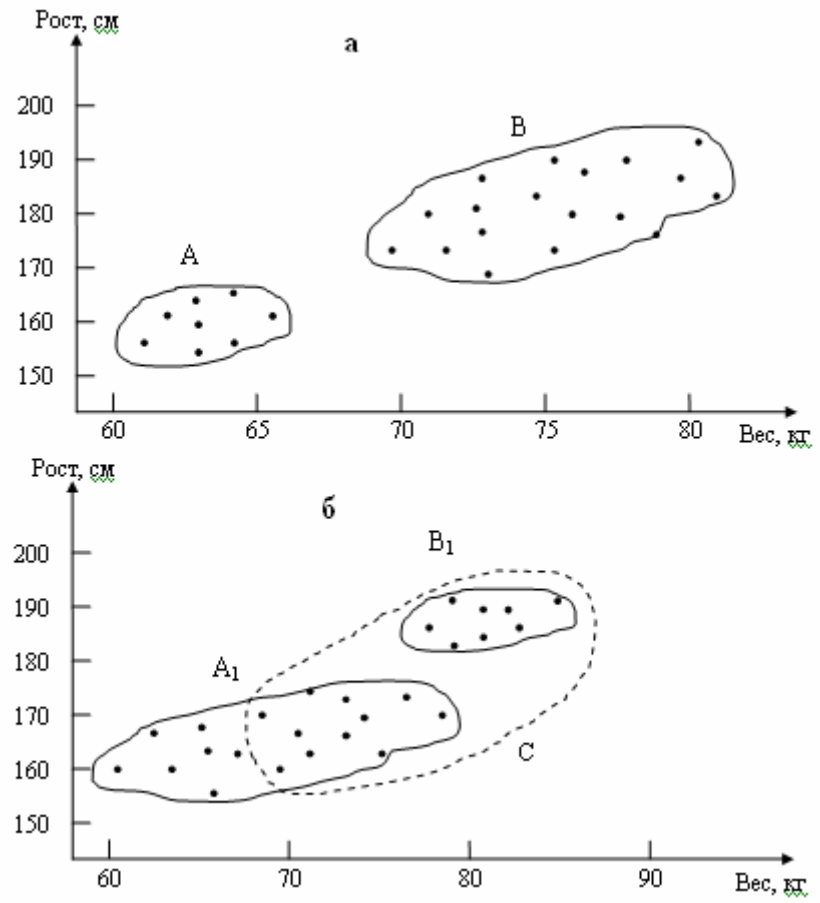


Рис. 1.1

На рис. 1.1а выделяются классы А – девушки, В – юноши. На рис. 1.1b выделяются классы A_1 (юноши и девушки) и В1(часть юношей). Класс юношей С (пунктирная линия) на рис. 1.1б не выделит, поскольку расстояния между ближайшими объектами классов A_1 и B_1 существенно больше, чем внутренние расстояния в A_1, юноши из А почти никакими алгоритмами к B_1 не присоединяются.

Однако определить расстояние между объектами в данном случае нельзя, поскольку признаки измерены в разных единицах измерения. Требуется нормировка показателей, переводящая их в безразмерные величины: тогда измерение близости объектов становится оправданным.

В кластерном анализе для количественной оценки сходства вводится понятие *метрики.* Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в *k*-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

Расстоянием (метрикой) между объектами в пространстве параметров называется такая величина d_{ab}, которая удовлетворяет аксиомам:

A1. $d_{ab} > 0$, $d_{ab} = 0$,

A2. $d_{ab} = d_{ba}$,

A3. $d_{ab} + d_{bc} \geq d_{ac}$.

Мерой близости (сходства) обычно называется величина m_{ab}, имеющая предел и возрастающая с возрастанием близости объектов.

B1. m_{ab} непрерывна,

B2. $m_{ab} = m_{ba}$,

B3. $1 \leq m_{ab} \leq 0$.

Существует возможность простого перехода от расстояний к мерам близости: $m = \frac{1}{1+d}$.

1.1.2. Характеристики близости объектов

Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Например, если вы должны кластеризовать типы еды в кафе, то можете принять во внимание количество содержащихся в ней калорий, цену, субъективную оценку вкуса и т.д. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если вы имеете двух- или трёхмерное пространство, то эта мера является реальным геометрическим расстоянием между объектами в пространстве (как будто расстояния между объектами измерены рулеткой). Однако алгоритм объединения не "заботится" о том, являются ли "предоставленные" для этого расстояния настоящими или некоторыми другими производными мерами расстояния, что более значимо для исследователя; и задачей исследователей является подобрать правильный метод для специфических применений.

Рассмотрим основные способы определения близости между объектами (Таблица 1.1)

Таблица 1.1

Показатели	Формулы
Для количественных шкал	
Линейное расстояние	$d_{l\,i\,j} = \sum_{l=1}^{m} \left| x_i^l - x_j^l \right|$
Евклидово расстояние	$d_{E\,i\,j} = \left(\sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^2 \right)^{\frac{1}{2}}$
Квадрат евклидово расстояния	$d^2{}_{E\,i\,j} = \sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^2$
Обобщенное степенное расстояние Минковского	$d_{P\,i\,j} = \left(\sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^P \right)^{\frac{1}{P}}$
Расстояние Чебышева	$d_{ij} = \max_{1 \le i,\, j \le l} \left| x_i - x_j \right|$
Расстояние городских кварталов (Манхэттенское расстояние)	$d_H \left(x_i, x_j \right) = \sum_{l=1}^{k} \left| x_i^l - x_j^l \right|$

Евклидово расстояние является самой популярной метрикой в кластерном анализе. Оно попросту является геометрическим расстоянием в многомерном пространстве. Геометрически оно лучше всего объединяет объекты в шарообразных скоплениях.

Квадрат евклидова расстояния. Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом *евклидова расстояния* путем возведения в квадрат стандартного *евклидова расстояния*.

Обобщенное степенное расстояние представляет только математический интерес как универсальная метрика.

Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Манхэттенское расстояние (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием.

Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании *евклидова расстояния*, поскольку здесь координаты не возводятся в квадрат.

Процент несогласия. Это расстояние вычисляется, если данные являются категориальными.

1.3. Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

1.3.1. Иерархические алгоритмы

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде *дендрограммы* показан на рис. 1.2.

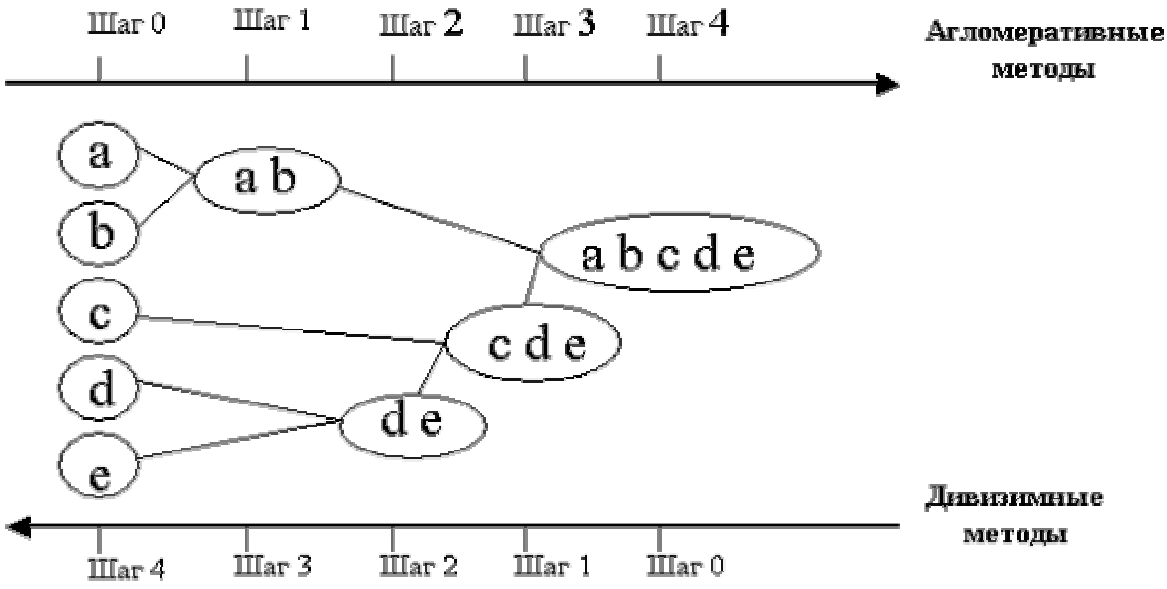


Рис. 1.2. Дендрограмма агломеративных и дивизимных методов

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных. Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением *дендрограмм* (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы приведен на рис. 1.3.

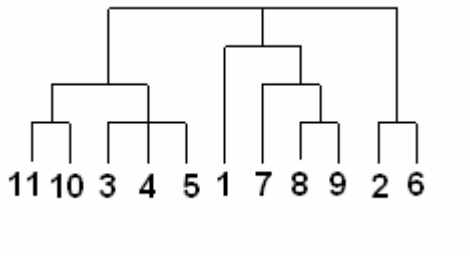


Рис. 1.3. Пример дендрограммы

Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Пусть

K_i – i-я группа (класс, кластер), состоящая из n объектов;

$\bar{x}_i$ – среднее арифметическое векторных наблюдений K_i группы, т.е. «центр тяжести» i - й группы;

$r(K_i, K_j) = r_{ij}$ – расстояние между группами K_i и K_j.

Обобщенная алгомеративная процедура. На первом шаге каждый объект считается отдельным кластером. На следующем шаге объединяются два ближайших объекта, которые образуют новый класс, определяются расстояния от этого класса до всех остальных объектов, и размерность матрицы расстояний D сокращается на единицу. На p-ом шаге повторяется та же процедура на матрице $D_{(n-p)(n-p)}$, пока все объекты не объединятся в один класс.

Если сразу несколько объектов (классов) имеют минимальное расстояние, то возможны две стратегии: выбрать одну случайную пару или объединить сразу все пары. Первый способ является классическим и реализован во всех процедурах (иногда его называют восходящей иерархической классификацией). Второй способ называют методом ближайших соседей (не путать с алг. *“Ближайшего соседа”*) и используют реже.

Результаты работы всех иерархических процедур обычно оформляют в виде так называемой *дендограммы* (рис. 1.4.1 – 1.4.3). В дендограмме номера объектов располагаются по горизонтали, а по вертикали - результаты кластеризации.

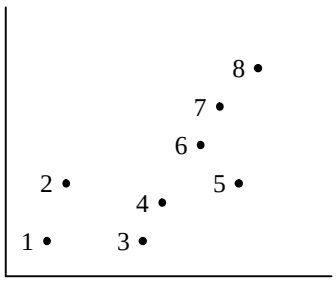


Рис. 1.4.1.

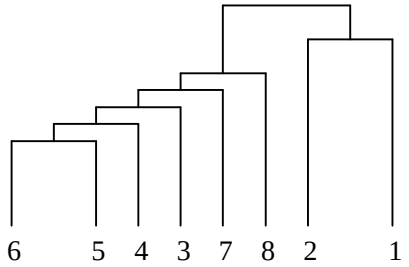


Рис. 1.4.2.

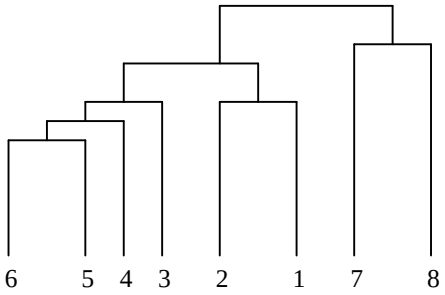


Рис. 1.4.3.

Расстояния между кластерами

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда связываются вместе несколько объектов, возникает вопрос, как следует определить расстояния между кластерами? Другими словами, необходимо правило объединения или связи для двух кластеров.

Здесь имеются различные возможности: например, вы можете связать два кластера вместе, когда любые два объекта в двух кластерах ближе друг к другу, чем соответствующее расстояние связи. Другими словами, вы используете правило ближайшего соседа для определения расстояния между кластерами; этот метод называется методом одиночной связи. Это правило строит волокнистые кластеры, т.е. кластеры сцепленные вместе только отдельными элементами, случайно оказавшимися ближе остальных друг к другу. Как альтернативу вы можете использовать соседей в

кластерах, которые находятся дальше всех остальных пар объектов друг от друга. Этот метод называется метод полной связи. Существует также множество других методов объединения кластеров, подобных тем, что были рассмотрены здесь, и модуль Кластерный анализ предлагает широкий выбор таких методов.

1. *Расстояние "Ближайшего соседа" (Одиночная связь).* Первый шаг 1.–7. совпадает с первым шагом алгоритма *Обобщенная алгомеративная процедура*. Расстояние равно расстоянию между ближайшими объектами классов.

$$r_{\min}(K_i, K_j) = \min_{x_i \in K_i,\, x_{ji} \in K_j} r(x_i, x_j)$$

2. *Расстояние "Дальнего соседа" (Полная связь).* Расстояние равно расстоянию между самыми дальними объектами классов.

$$r_{\max}(K_i, K_j) = \max_{x_i \in K_i,\, x_{ji} \in K_j} r(x_i, x_j)$$

3. *Невзвешенное попарное среднее.* В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные рощи, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. *Взвешенное попарное среднее.* Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован (скорее даже, чем предыдущий), когда предполагаются неравные размеры кластеров.

5. *Невзвешенный центроидный метод.* В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. *Взвешенный центроидный метод (медиана).* Тот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

7. *Метод Варда (Ward, 1963).* В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые

приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов (SS). Этот метод направлен на объединение близко расположенных кластеров.

1.3.2 Процедуры эталонного типа

Наряду с иерархическими методами классификации, существует многочисленная группа так называемых итеративных методов кластерного анализа (метод k - средних.). Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). Название метода было предложено Дж. Мак-Куином в 1967 г. В отличие от иерархических процедур метод k - средних не требует вычисления и хранения матрицы расстояний или сходств между объектами. Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы k выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты. Метод k - средних удобен для обработки больших статистических совокупностей.

Математическое описание алгоритма метода k - средних.

Пусть имеется n наблюдений, каждое из которых характеризуется m признаками $X_1, X_2, \mathbf{K}, X_n$. Эти наблюдения необходимо разбить на k кластеров. Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений k точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся $(n-k)$ объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2}, \ldots, x_{im})$ и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние. Проверяемый объект присоединяется к тому центру (эталону), которому соответствует минимальное из расстояний. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то i-ый объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку X_{i+1} и для нее повторяются все процедуры. Таким образом, через $(n-k)$ шагов все точки (объекты) совокупности

окажутся отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки $X_1, X_2, \mathbf{K}, X_n$ опять подсоединяются к полученным кластером, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить $C_1, C_2, \mathbf{K}, C_k$. При этом каждая точка X_i $(i = 1,2,\mathbf{K},n)$ будет относиться к такому кластеру (классу) l, для которого расстояние минимально. Возможны две модификации метода k - средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется.

1.3. Пример решения практической задачи кластерным анализом в пакете STATISTICA

Система STATISTICA является одним из наиболее известных в мировой практике пакетов статистической обработки в среде Windows.

Разработчиком пакета является фирма Stat Soft, Inc. (США).

Пакет STATISTICA отличается от большинства других программных продуктов для Windows тем, что состоит из отдельных программ-модулей, каждый из которых содержит конкретный метод обработки данных, например кластерный анализ, регрессионный анализ и т.д. Каждый такой модуль можно рассматривать как самостоятельную программу, независимую от остальных. Но такие операции, как ввод, корректировка данных, осуществляется в любом из модулей по желанию пользователя.

Таким образом, если перед пользователем стоит конкретная задача, к примеру, провести классификацию данных, то достаточно вызвать модуль *Cluster Analysis (Кластерный анализ)*, чтобы провести полноценную работу: ввод данных, их корректировку, построить различные диаграммы, а также осуществить непосредственно заданную классификацию.

Если же обработка данных предполагает использование нескольких групп методов, то есть требуется использование ряда модулей, то каждый модуль может быть вызван из

другого, при этом автоматически осуществляется переход файла с обрабатываемыми данными в новый модуль.

Для запуска пакета требуется войти в среду Windows, нажать на кнопку **Пуск**, в строке **Программы** выбрать **STATISTICA**. Далее либо подвести курсор к строке **Основные статистики**, либо щелкнуть левой кнопкой мыши на слове **STATISTICA**. В последнем случае появится меню **STATISTICA Module Switcher** (Переключатель модулей системы STATISTICA), предлагающее выбрать конкретный статистический модуль (рис.1.5).

Далее следует навести курсор на тот модуль, в котором собираемся работать, и щелкнуть дважды левой кнопкой мыши. Операцию вызова также можно осуществить через кнопку **Switch to…**(Переключиться в…)**.** Для отмены вызова меню предназначена клавиша **Cancel** (Отмена).

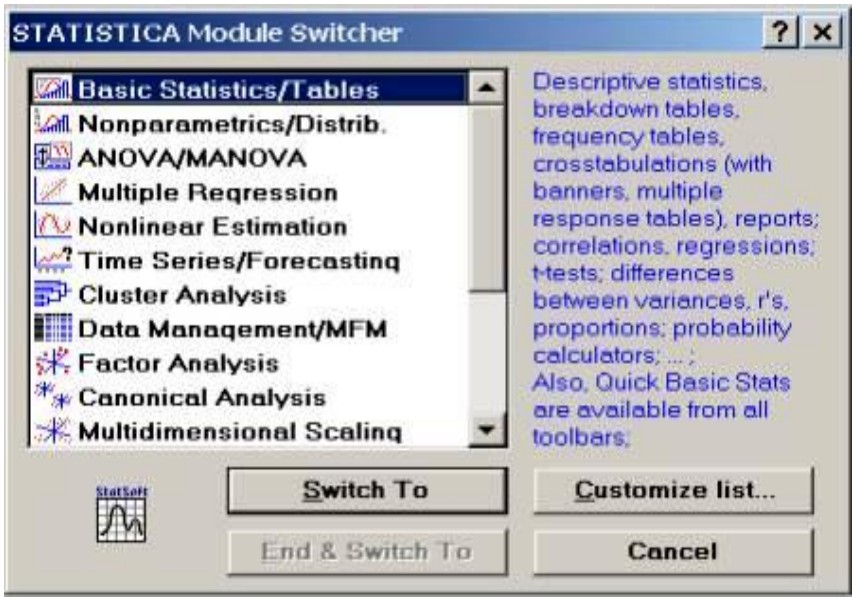


рис. 1.5

Рассмотрим основные этапы проведения кластерного анализа в системе STATISTICA на следующем примере.

В файле (рис.1.6) содержатся данные по 20 странам (импорт товаров за 2000–2002 г). Задача состоит в том, чтобы распределить объекты по однородным группам и установить качественные взаимосвязи между группами стран с близкими значениями показателей.

Data: Импорт товаров по странам (3v by 20c)

	1 2000	2 2001	3 2002
Австрия	2206	6560	5777
Аргентина	30	441	603
Бельгия	1658	4346	5303
Болгария	2008	7119	9286
Бразилия	639	1114	1245
Великобритания	151	608	2412
Венгрия	286	91	1586
Германия	4473	6959	11436
Дания	1241	23	751
Индия	83	39	84
Испания	4724	4283	3331
Италия	5979	5871	6611
Канада	81	7	207
Китай	102	129	338
США	386	2297	36128
Турция	6	46	101
Франция	2698	6437	7999
Швейцария	438	1027	532
ЮАР	1265	1228	2139
Япония	162	166	1803

Рис. 1.6

Алгоритм выполнения

Предварительно нормируем исходные данные в целях устранения различий в единицах измерения показателей. После вызова строки **Standardize Variables** (Стандартизация значений) на экране компьютера появится диалоговое окно задания параметров нормирования данных (рис. 1.7).

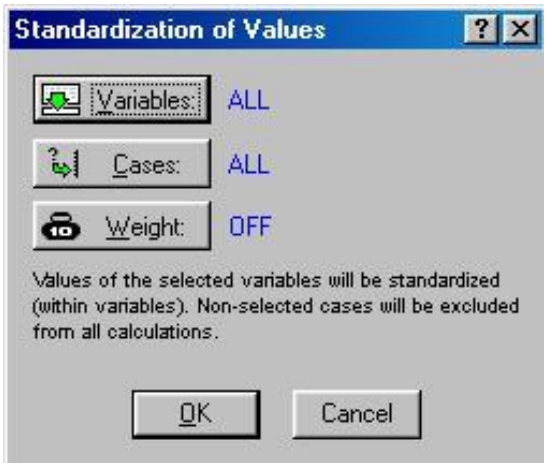


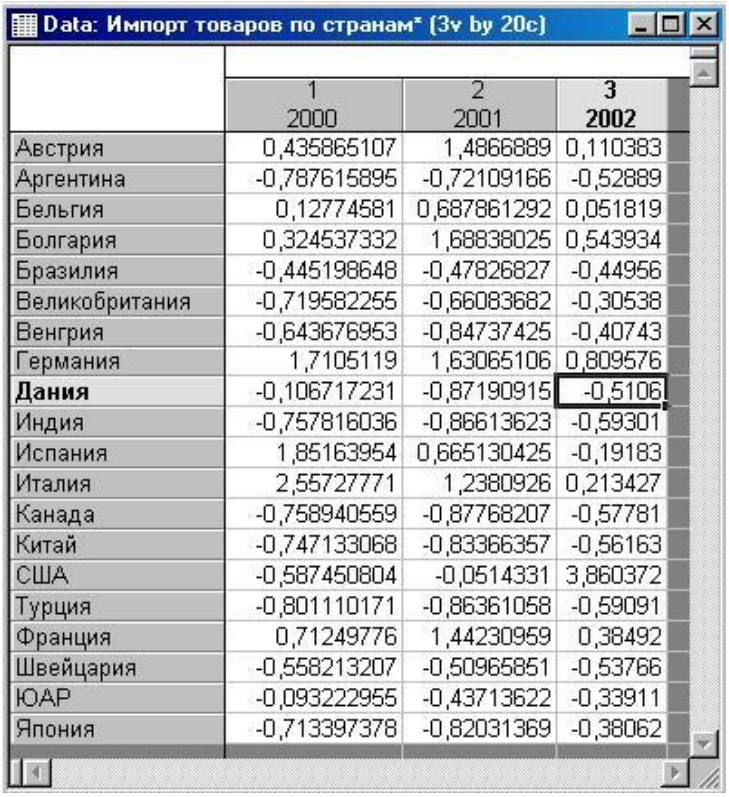
рис. 1.7

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 в диалоговом окне **Define Weight** (Задание веса).

После процедуры нормирования данные выглядят следующим образом (рис. 1.8):



Data: Импорт товаров по странам* (3v by 20c)

	1 2000	2 2001	3 **2002**
Австрия	0,435865107	1,4866889	0,110383
Аргентина	-0,787615895	-0,72109166	-0,52889
Бельгия	0,12774581	0,687861292	0,051819
Болгария	0,324537332	1,68838025	0,543934
Бразилия	-0,445198648	-0,47826827	-0,44956
Великобритания	-0,719582255	-0,66083682	-0,30538
Венгрия	-0,643676953	-0,84737425	-0,40743
Германия	1,7105119	1,63065106	0,809576
Дания	-0,106717231	-0,87190915	-0,5106
Индия	-0,757816036	-0,86613623	-0,59301
Испания	1,85163954	0,665130425	-0,19183
Италия	2,55727771	1,2380926	0,213427
Канада	-0,758940559	-0,87768207	-0,57781
Китай	-0,747133068	-0,83366357	-0,56163
США	-0,587450804	-0,0514331	3,860372
Турция	-0,801110171	-0,86361058	-0,59091
Франция	0,71249776	1,44230959	0,38492
Швейцария	-0,558213207	-0,50965851	-0,53766
ЮАР	-0,093222955	-0,43713622	-0,33911
Япония	-0,713397378	-0,82031369	-0,38062

рис. 1.8

Для проведения кластеризации вызовем модуль **Cluster Analysis** (Кластерный анализ) с помощью переключателя модулей **STATISTICA Module Switcher** (рис.1.9)

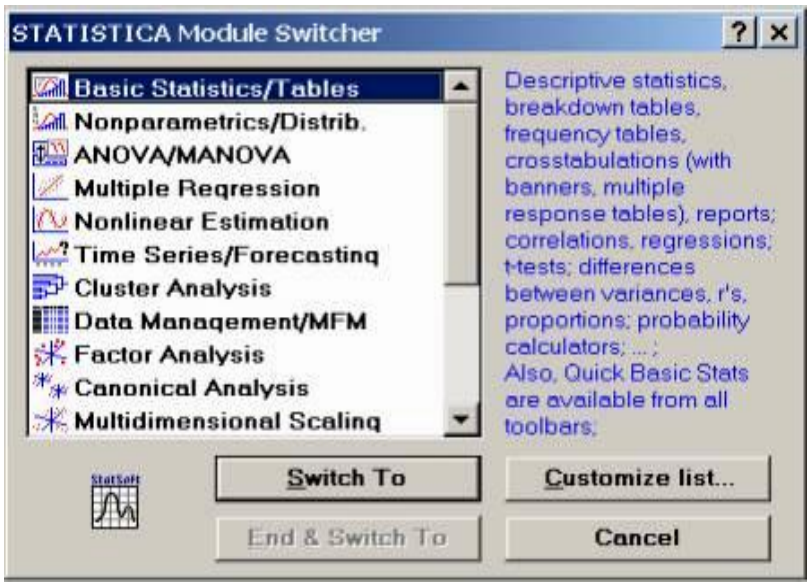


рис. 1.9.

или через меню **Статистика \ Многомерные исследовательские методы \ Анализ кластера**.[1]

Появившееся диалоговое окно содержит следующие методы (рис.1.10):

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



рис. 1.10

[1] В зависимости от версии программы STATISTICA

Объединение (древовидная кластеризация) – Joining (tree clustering)



рис.1.11

Щелкнув по кнопке **Variables** (Переменные), выберем показатели для анализа с помощью мыши или же в белом нижнем окошке через тире наберем их номера с клавиатуры, если они идут подряд, иначе через пробел (рис.1.12). Кнопка **Select All** (Выбрать все) отбирает весь список.

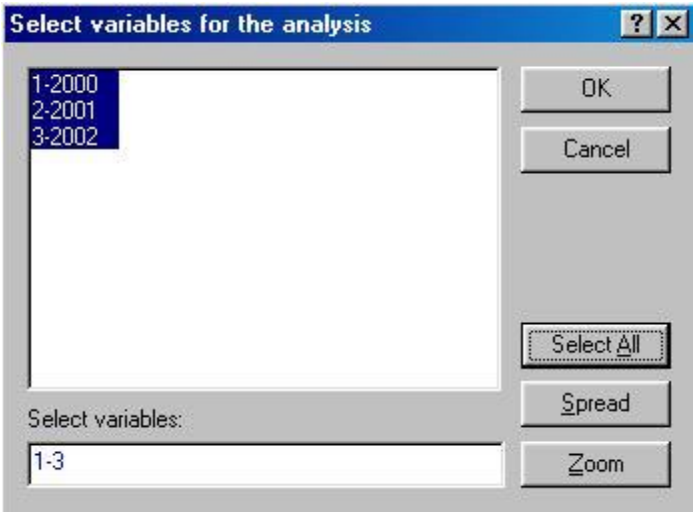


рис. 1.12

Input (Исходные данные) (рис.1.11) представляет собой раскрывающееся меню. Строка **Distance matrix** (Матрица расстояний) предусмотрена на тот случай, если входная информация представлена в виде мер сходства. В данной задаче выбираем второй тип – **Raw data** (Исходные данные).

В поле **Cluster** (Кластер) (рис. 1.11) задается направление классификации. При кластеризации самих переменных помечаются **Variables [Columns]** (Переменные [столбцы]), в данной задаче **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).

Для решения данной задачи выбираем метод Варда.

В окошке **Distance measure** (Мера расстояния) (рис. 1.11) предлагаются различные виды расстояний:

- **Squared Euclidean distances** (квадрат Евклидова расстояния)
- **Euclidean distances** (Евклидово расстояние)
- **City-block (Manhattan) distance** (Расстояние городских кварталов (Манхэттенское расстояние))
- **Chebychev distance metric** (Расстояние Чебышева)
- **Power: SUM(ABS(x-y)**p)**1/r** (Степенное расстояние)
- **Percent disagreement** (Процент несогласия)

Выберем **City-block (Manhattan) distance** (Манхэттенское расстояние [расстояние городских кварталов]).

Перейдем к строке **Missing data [MD deletion]** (Пропущенные данные) (рис. 1.11).

В системе STATISTICA в модуле **Cluster Analysis** (Кластерный анализ) предусмотрено два способа обработки некомплектных наблюдений, содержащих пропуски хотя бы одной переменной:

- **Casewise deleted** (Построчное удаление) - обработки некомплектные наблюдения полностью исключаются из дальнейшего анализа. Однако подобный метод приводит к смещенности и несостоятельности полученных статистических оценок, а также к искажению эмпирического распределения.

- **Substituted by means** (Замена средними значениями) – пропущенные данные заменяются средними значениями показателя, полученным по комплектным (полным) данным, что также имеет свои недостатки.

В данном примере пропуски отсутствуют, поэтому выбор того или иного метода не актуален.

После установки всех необходимых параметров для проведения кластеризации щелкнем на **Ok** и рассмотрим окно с результатами классификации (рис. 1.13).

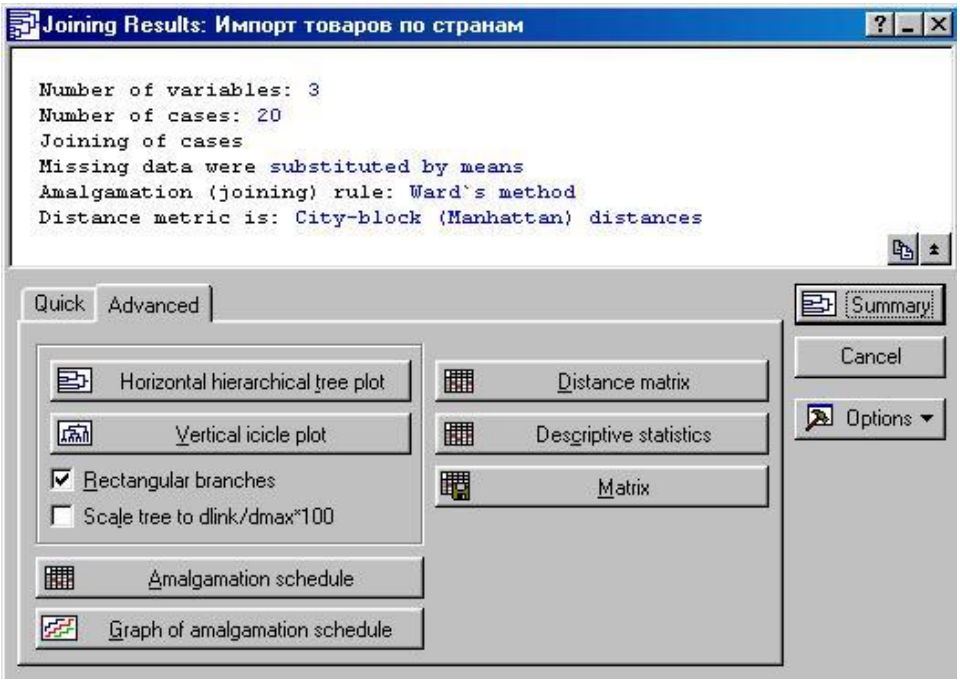


рис. 1.13

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

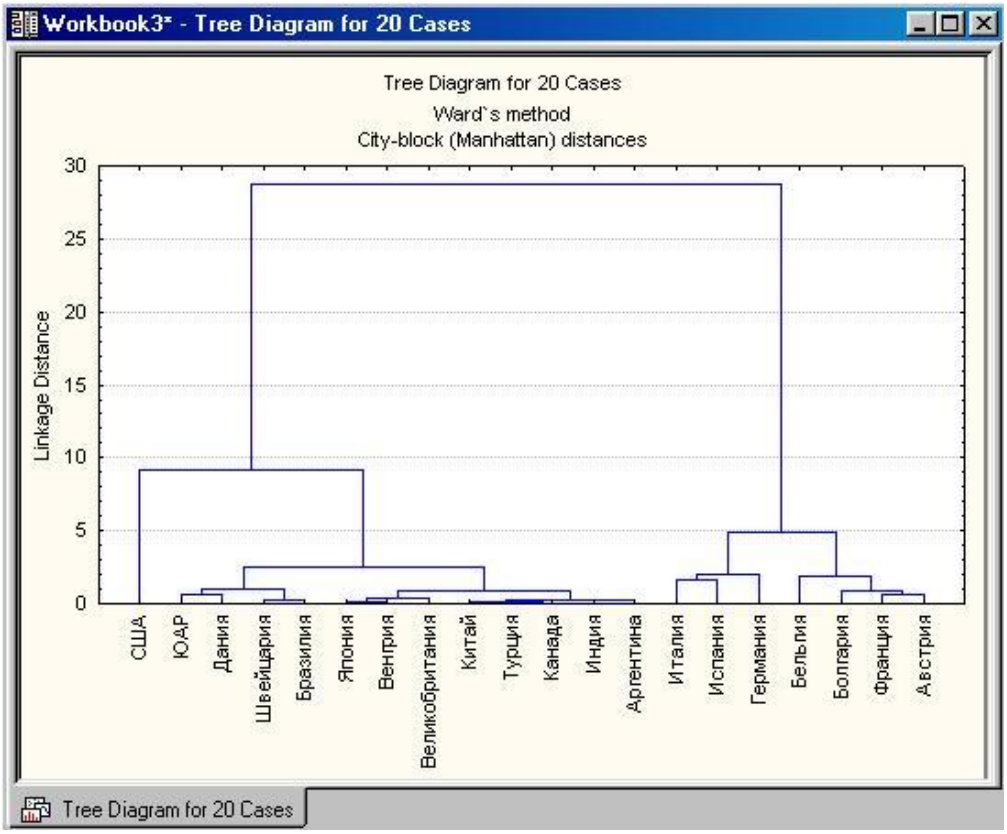
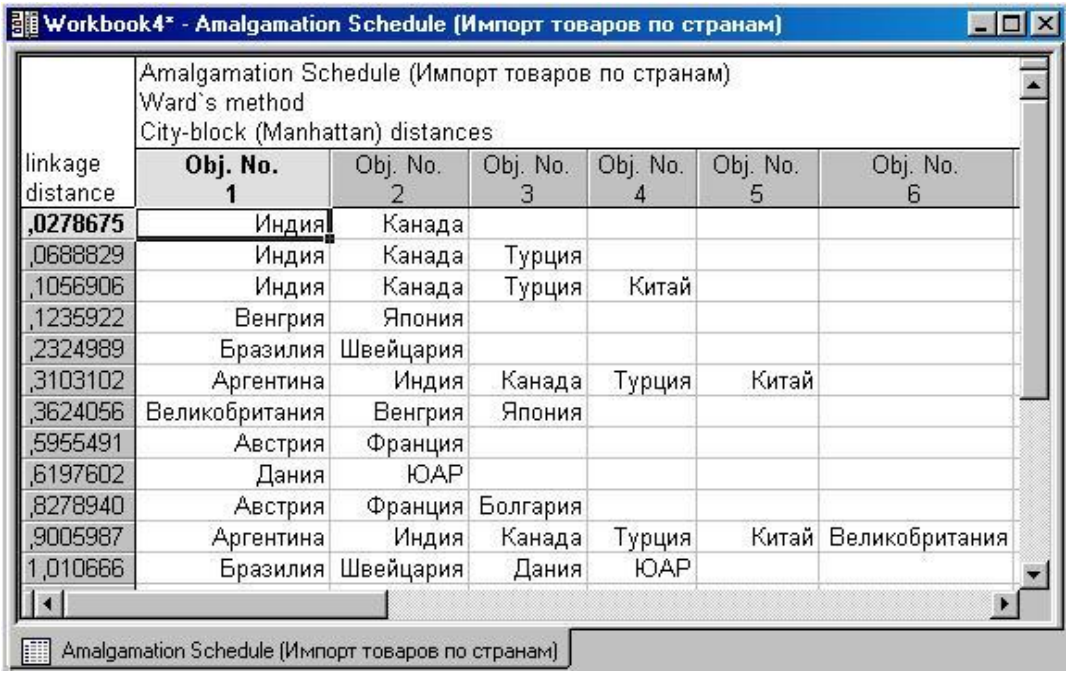


рис. 1.14

На этой дендограмме (рис. 1.14) горизонтальная ось представляет наблюдения, вертикальная – расстояние объединения. Таким образом, на первом шаге были объединены Индия и Канада, как имеющие минимальное расстояние, а на последнем – все, уже объединенные в какие-либо кластеры. На основе анализа диаграммы посмотрим, на каком шаге кластеризации следует принять полученную классификацию как окончательную. При числе кластеров равным двум К=2: в первый войдут: США, ЮАР, Дания, Швейцария, Бразилия, Япония, Венгрия, Великобритания, Китай, Турция, Канада, Индия, Аргентина, а во второй – Австрия, Франция, Болгария, Бельгия, Германия, Испания, Италия; при К=3 первый кластер будет состоять из США, второй - ЮАР, Дания, Швейцария, Бразилия, Япония, Венгрия, Великобритания, Китай, Турция, Канада, Индия, Аргентина, третий – все остальные; при К=4 первый кластер будет состоять из США, второй - ЮАР, Дания, Швейцария, Бразилия, Япония, Венгрия, Великобритания, Китай, Турция, Канада, Индия, Аргентина, третий - Италия, Испания, Германия, четвертый - Бельгия, Болгария, Франция, Австрия; при К=5 первый кластер будет состоять из США, второй - ЮАР, Дания, Швейцария, Бразилия, третий - Япония, Венгрия, Великобритания, Китай, Турция, Канада, Индия, Аргентина, четвертый – Германия, Испания, Италия,

пятый – Австрия, Франция, Болгария, Бельгия. Предпочтительнее взять количество кластеров равное 4 или 5, так как при большем их числе теряется наглядность классификации.

Щелкнув по кнопке **Amalgamation schedule** (Схема объединения), можно выбрать таблицу результатов со схемой объединения (рис. 1.15). Первый столбец таблицы содержит расстояния для соответствующих кластеров. Каждая строка показывает состав кластера на данном шаге классификации. Например, на первом шаге (1 строка) объединились Индия и Канада, на втором (2 строка) – Индия, Канада и Турция, на третьем (3 строка) - Индия, Канада, Турция и Китай, на четвертом (4 строка) – Венгрия и Япония, на пятом - Бразилия и Швейцария, и т.д.



linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6
,0278675	Индия	Канада				
,0688829	Индия	Канада	Турция			
,1056906	Индия	Канада	Турция	Китай		
,1235922	Венгрия	Япония				
,2324989	Бразилия	Швейцария				
,3103102	Аргентина	Индия	Канада	Турция	Китай	
,3624056	Великобритания	Венгрия	Япония			
,5955491	Австрия	Франция				
,6197602	Дания	ЮАР				
,8278940	Австрия	Франция	Болгария			
,9005987	Аргентина	Индия	Канада	Турция	Китай	Великобритания
1,010666	Бразилия	Швейцария	Дания	ЮАР		

рис. 1.15

Щелкнув по кнопке **Graph of amalgamation schedule** (График схемы объединения) (рис. 1.16), просмотрим результаты древовидной кластеризации в графическом виде.

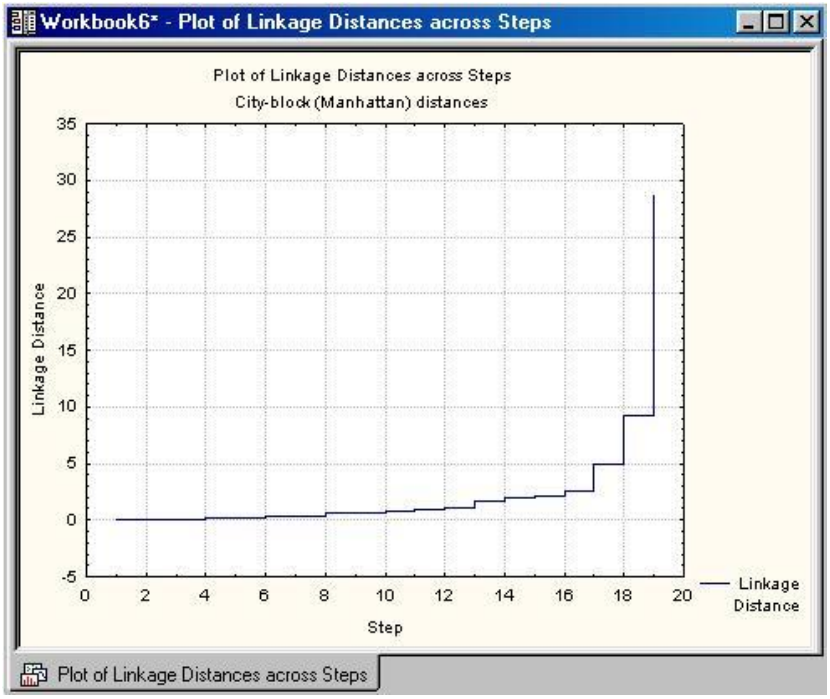


рис. 1.16

Просмотр матрицы расстояний осуществляется через кнопку **Distance matrix** (Матрица расстояний) (рис. 1.17) , которая затем сохраняется с помощью опции **Save distance matrix** (Сохранить матрицу расстояний).

Workbook7* - City-block (Manhattan) distances (Импорт товаров по странам)

Case No	City-block (Manhattan) distances (Импорт товаров по странам)					
	Австрия	Аргентина	Бельгия	Болгария	Бразилия	Вели
Австрия	0,00	4,07	1,17	0,75	3,41	
Аргентина	4,07	0,00	2,91	4,59	0,66	
Бельгия	1,17	2,91	0,00	1,69	2,24	
Болгария	0,75	4,59	1,69	0,00	3,93	
Бразилия	3,41	0,66	2,24	3,93	0,00	
Великобритания	3,72	0,35	2,55	4,24	0,60	
Венгрия	3,93	0,39	2,77	4,46	0,61	
Германия	2,12	6,19	3,28	1,71	5,52	
Дания	3,52	0,85	2,36	4,05	0,79	
Индия	4,25	0,24	3,08	4,77	0,84	
Испания	2,54	4,36	1,99	3,29	3,70	
Италия	2,47	6,05	3,14	3,01	5,38	
Канада	4,25	0,23	3,08	4,77	0,84	
Китай	4,18	0,19	3,01	4,70	0,77	
США	6,31	5,26	5,26	5,97	4,88	

City-block (Manhattan) distances (Импорт товаров по странам)

рис. 1.17

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

зависимости от установки, выбранной в списке **Cluster** (Кластер) стартовой панели (рис. 1.18).

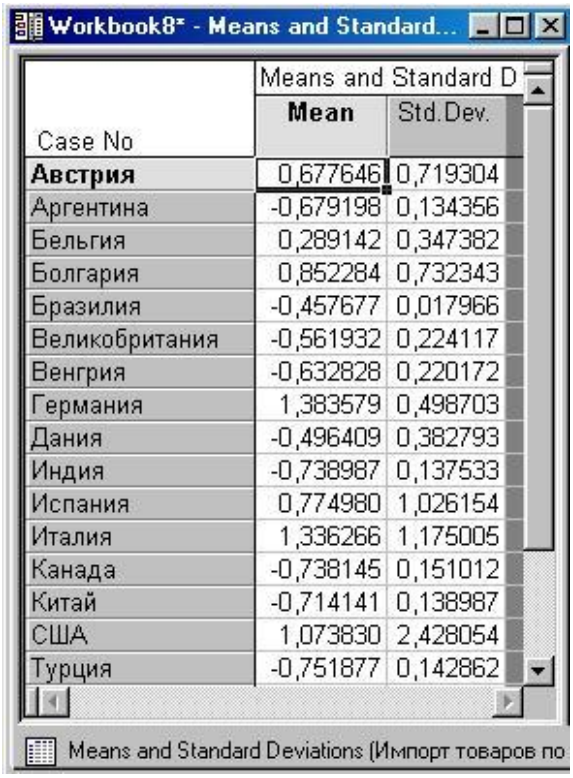


рис. 1.18

Кластеризация методом K – средних – K – means clustering

Этот метод кластеризации существенно отличается от иерархических агломеративных методов. Он применяется, если пользователь уже имеет представление относительно числа кластеров, на которые необходимо разбить наблюдения. Тогда метод k – средних строит ровно k различных кластеров, расположенных на возможно больших расстояниях друг от друга.

Рассмотрим работу метода k-средних на данных нашего примера.

Щелкнем по строке – **K – means clustering** (Кластеризация методом k-средних) стартовой панели модуля **Cluster analysis** (Кластерный анализ) (рис. 1.10). На экране появится окно настройки параметров кластеризации (рис. 1.19).

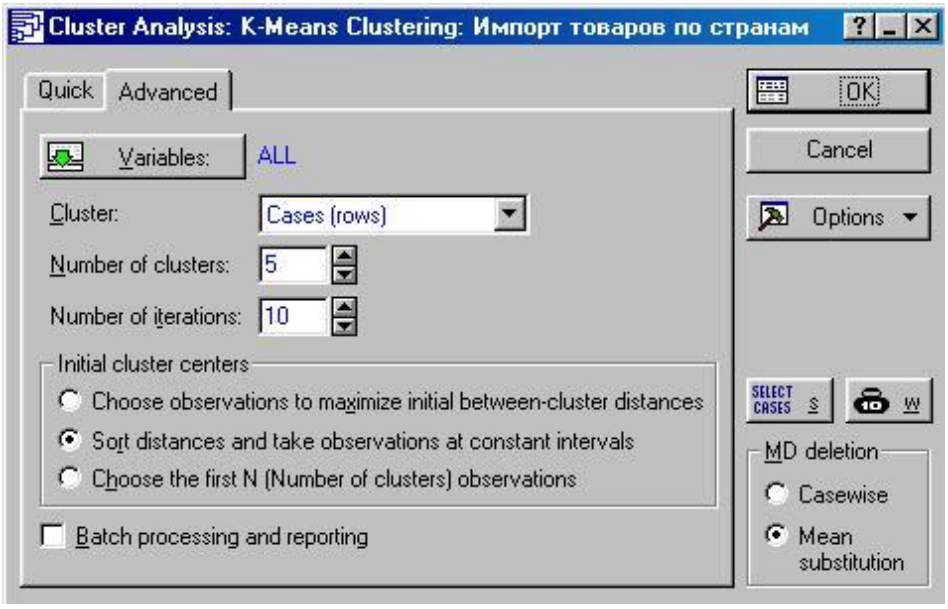


рис. 1.19

С помощью кнопки **Variables** (Переменные) выберем показатели, по которым будет происходить кластеризация. В строке **Cluster** (Кластер) укажем объекты для классификации **Cases [rows]** (Наблюдения [строки]).

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 набл, максимиз. 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.

Для нашего примера сделаем установку - **Sort distances and take observations at constant intervals** (Сортировать расстояния и выбрать наблюдения на постоянных интервалах).

Опция **MD deletion** (Пропущенные данные) устанавливает режим работы с теми наблюдениями (или переменными, если установлен режим **Variables (columns)**) в строке **Cluster**, в которых пропущены данные. По умолчанию установлен режим **Casewise** (Случай удаления). Тогда наблюдения просто исключаются из рассмотрения. Если установить режим **Mean subsitution** (Заменять на среднее), то вместо пропущенного числа будет использовано среднее по этой переменной (или наблюдению).

Если установлена опция **Batch processing and reporting** (Пакетная обработка и сообщение), тогда система STATISTICA автоматически осуществит полный анализ и представит результаты в соответствии с установками.

После соответствующего выбора нажмем кнопку **OK.** STATISTICA произведет вычисления и появится новое окно**: "K - Means Clustering Results"** (рис. 1.20).

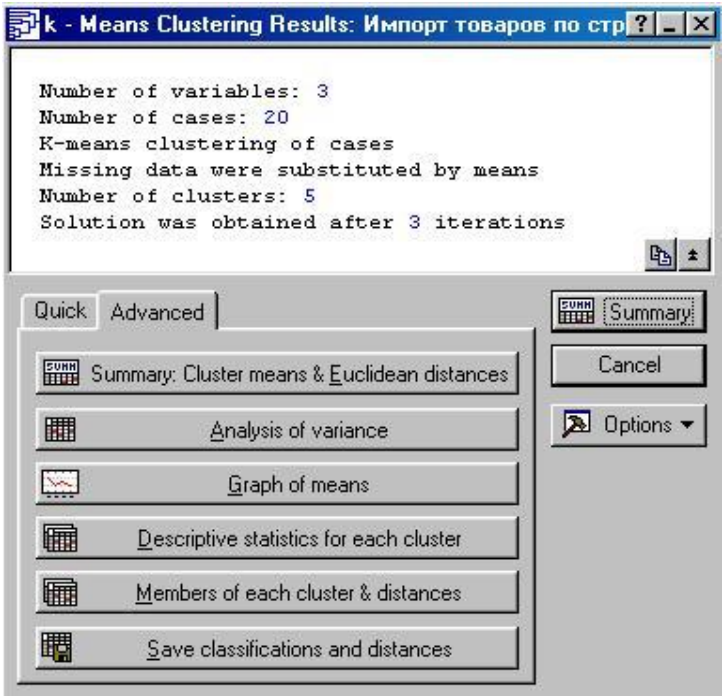


рис. 1.20

Вывод результатов и их анализ

Появившееся окно с результатами классификации условно разделено на две части. 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 - 3;
- Количество наблюдений - 20;
- Классификация наблюдений (или переменных, зависит от установки в предыдущем окне в строке **Cluster**) методом K - 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: 3 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. 1.21). 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. Проверка данной гипотезы похожа на проверку гипотезы в дисперсионном анализе, когда делается предположение о том, что уровни фактора не влияют на результат.

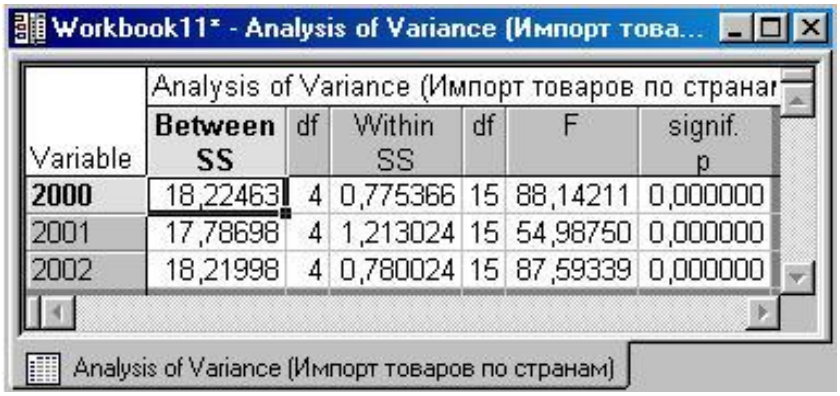


рис. 1.21

Cluster Means & Euclidean Distances (средние значения в кластерах и евклидово расстояние). Выводятся две таблицы. В первой (рис. 1.22) указаны средние величины

класса по всем переменным (наблюдениям). По вертикали указаны номера классов, а по горизонтали переменные (наблюдения).



рис. 1.22

Во второй таблице (рис. 1.23) 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.

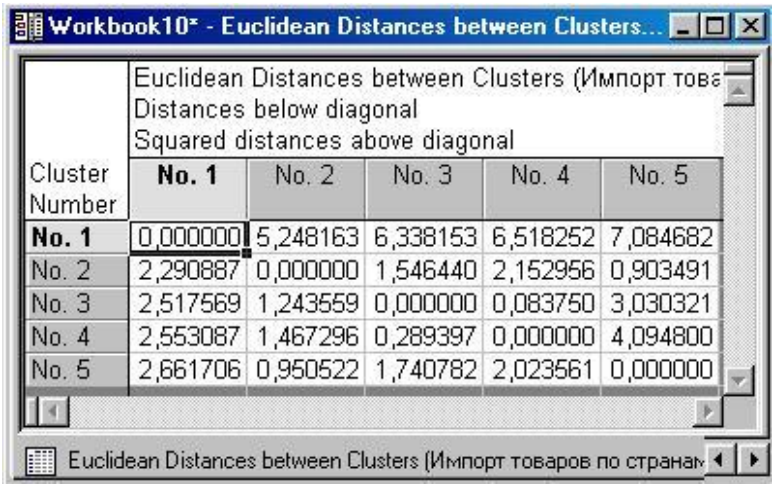


рис. 1.23

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. 1.24).

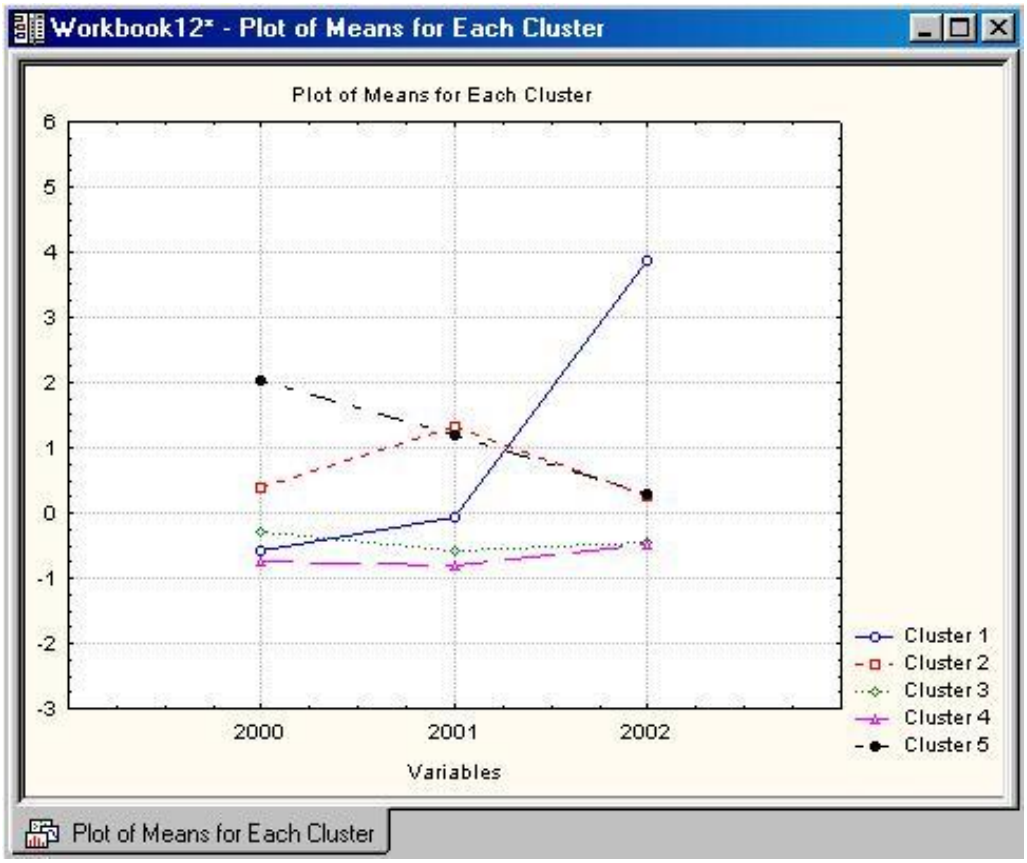


рис. 1.24

По горизонтали отложены участвующие в классификации переменные, а по вертикали - средние значения переменных в разрезе получаемых кластеров.

Descriptive Statistics for each cluster (Описательная статистика для каждого кластера). После нажатия этой кнопки выводятся окна, количество которых равно количеству кластеров (рис. 1.25). В каждом таком окне в строках указаны переменные (наблюдения), а по горизонтали их характеристики, рассчитанные для данного класса: среднее, несмещенное среднеквадратическое отклонение, несмещенная дисперсия.

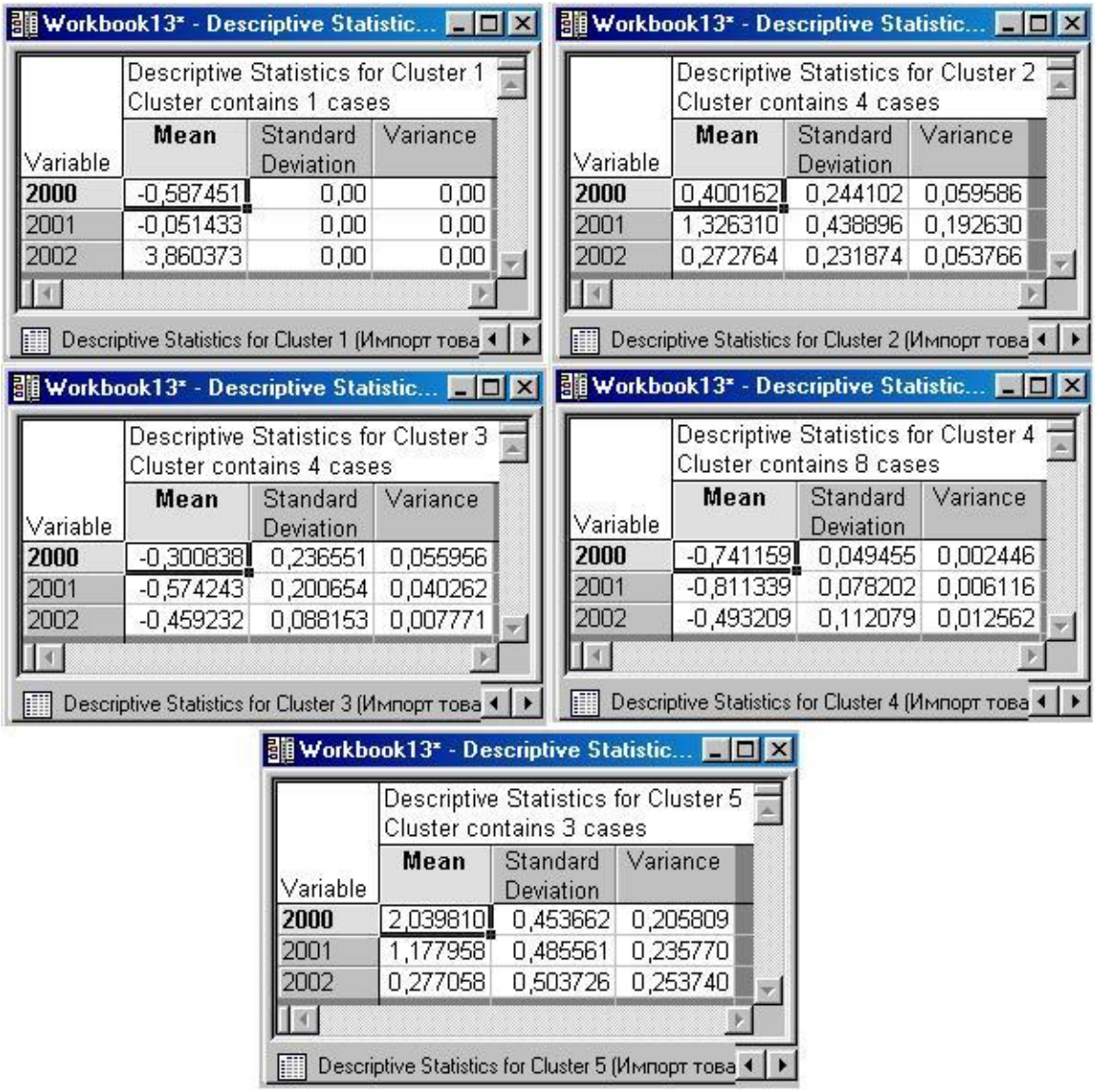


рис. 1.25

Members for each cluster & distances (Члены каждой группы и расстояния)**.** Выводится столько окон, сколько задано классов (рис. 1.26). В каждом окне указывается общее число элементов, отнесенных к этому кластеру, в верхней строке 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.

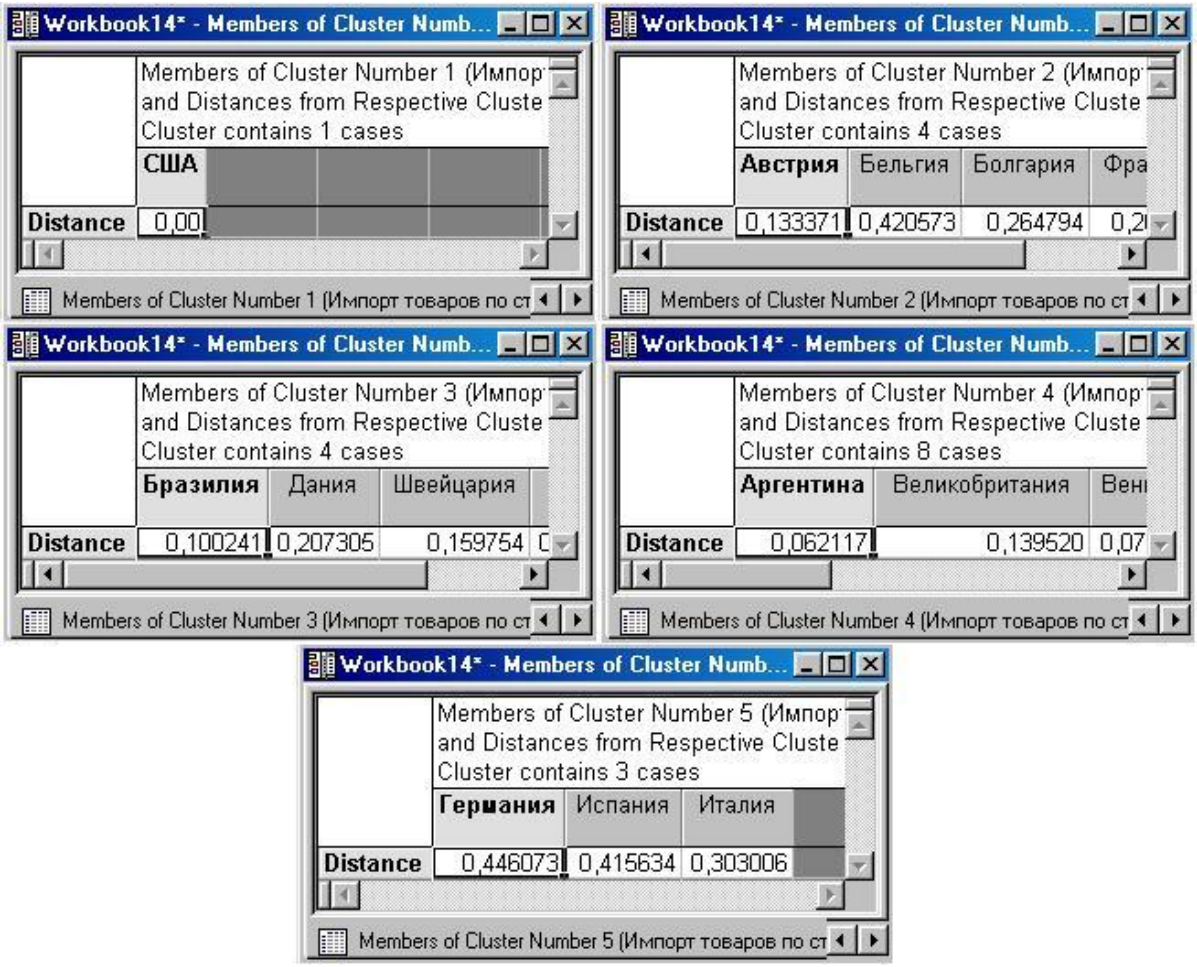


рис. 1.26

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.

Двухвходовое объединение - Two-way joining

Идея метода **Two-way joining** (Двухвходовое объединение) состоит в том, чтобы одновременно классифицировать как наблюдения, так и переменные.

Трудность с интерпретацией полученных результатов возникает вследствие того, что получающиеся кластеры являются по своей природе часто неоднородными.

Вызовем исходное меню **Cluster analysis** (Кластерный анализ) (рис.1.10) и выберем строку **Two-way joining** (Двухвходовое объединение). В появившемся диалоговом окне (рис. 1.27) через кнопку **Variables** (Переменные) укажем переменные для анализа.

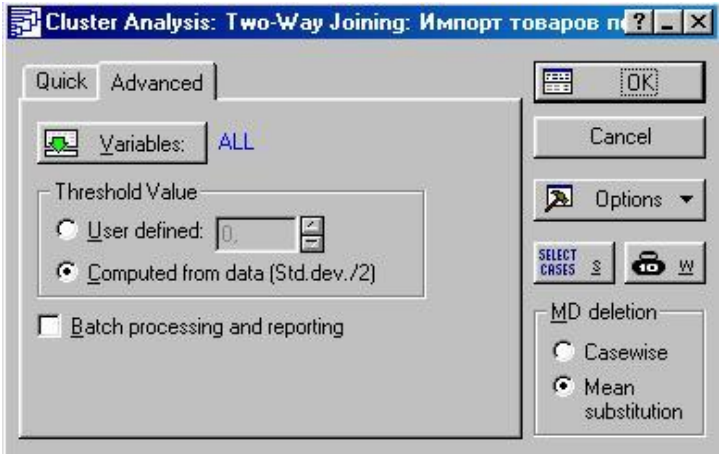


рис. 1.27

Группа операций **Threshold Value** (Значение порога) содержит два режима: **User defined** (Заданное пользователем) и **Computed from data (Std.Dev./2)** (Вычисленное по данным).

Пороговый параметр определяет принадлежность элементов матрицы данных к формируемым кластерам. Если эта величина слишком велика по сравнению со значениями элементов в матрице, то будет сформирован только один кластер; если очень мала, то кластером будет являться каждая точка данных.

Для большинства случаев берут пороговое значение, равное половине величины общего стандартного отклонения (режим **Computed from data (Std.Dev./2)** (Вычисленное по данным)). После задания всех параметров нажмем **Ok**.

Окно с результатами вычислений имеет следующий вид:

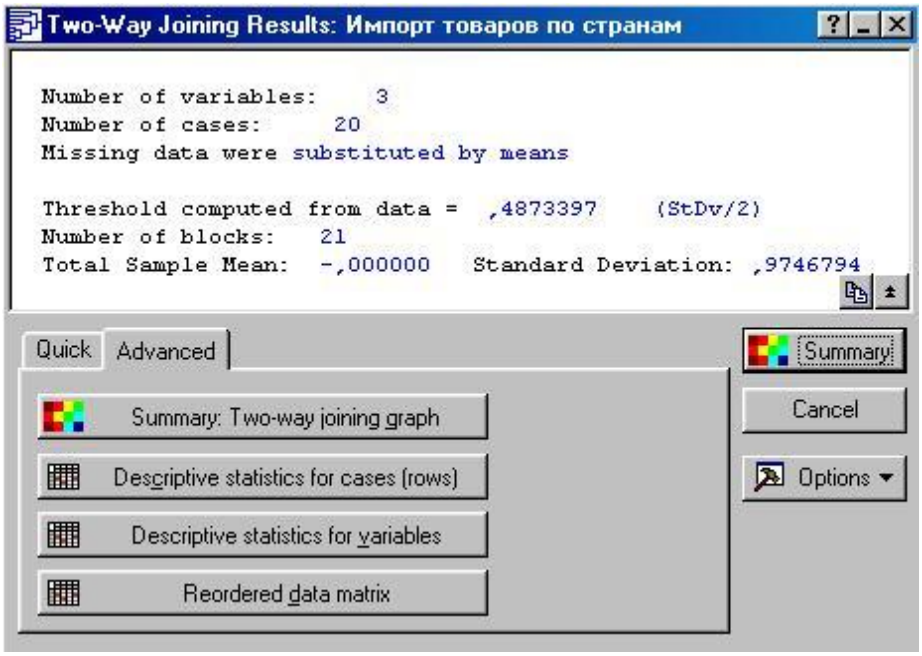


рис. 1.28

Две средние кнопки – **Descriptive statistics for cases (rows)** (Описательные статистики [строк]) и **Descriptive statistics for variables** (Описательные статистики для переменных) – выводят на экран таблицы со статистическими характеристиками данных: средними значениями и стандартными отклонениями (рис. 1.29).

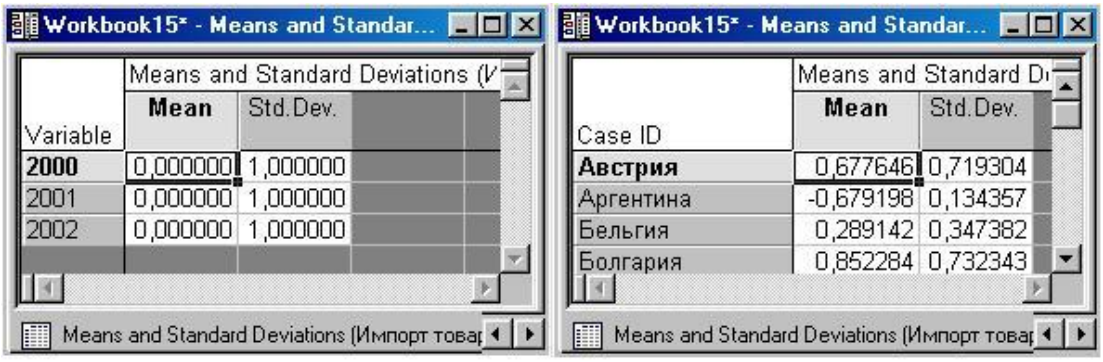


рис. 1.29

Опция **Reordered statistics for variables** (Переупорядоченная матрица данных) формирует таблицу с наблюдениями в соответствии с результатами двухвходового объединения (рис. 1.30).

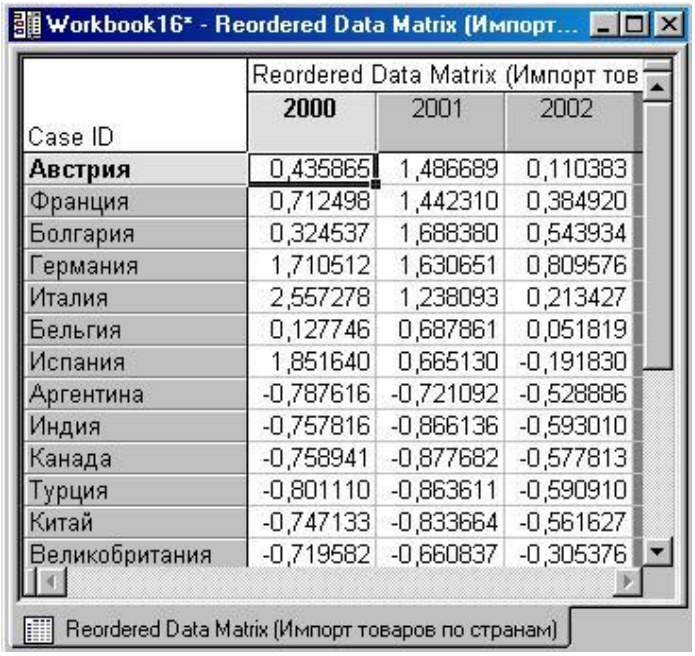


рис. 1.30

Кнопка **Summary: Two-way joining graph** выводит графическое изображение результатов двухвходового объединения. В данном случае переупорядоченная матрица данных отображается в виде карты линий уровня (рис. 1.31).

На графике по горизонтали отложены участвующие в классификации переменные, а по вертикали – наблюдения. Цвета ячеек, находящихся на пересечении, указывают на принадлежность элементов матрицы к определенному кластеру.

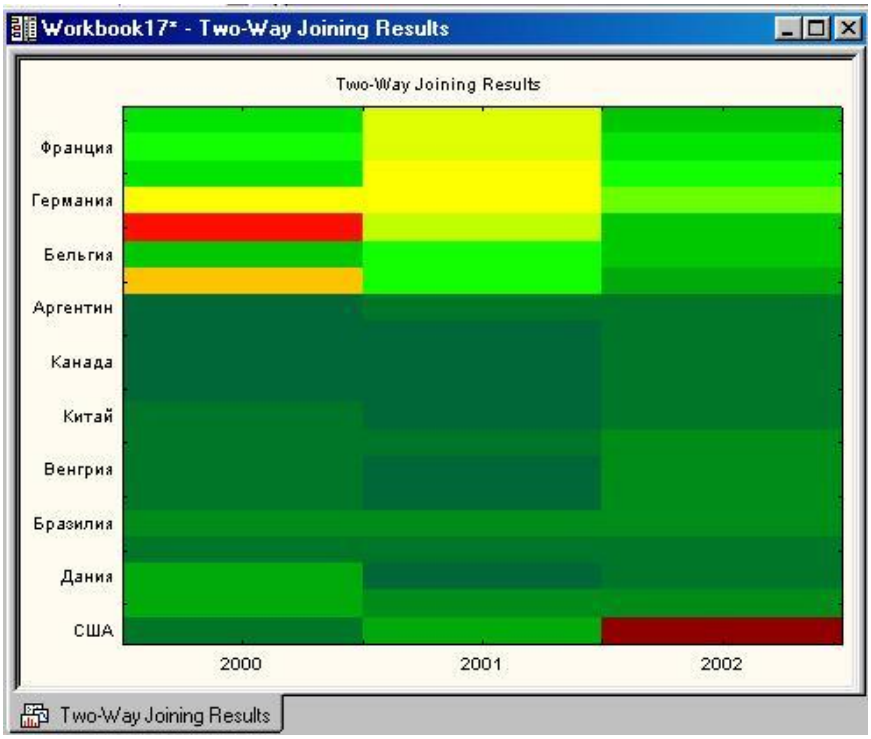


рис. 1.31

Такая сложная интерпретация результатов и спорная их практическая ценность делает рассматриваемую процедуру **Two-way joining** (Двухвходового объединения) не столь привлекательной.

1.4. Задания для самостоятельной работы

Задание 1.

По данным Таблицы 1.2 и алгоритму кластерного анализа провести классификацию объектов иерархическим методом (древовидная кластеризация).

Таблица 1.2

№ 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.

Задание 6.

По данным Таблицы 1.3 и алгоритму кластерного анализа провести классификацию объектов иерархическим методом (древовидная кластеризация).

Таблица 1.3

№ п/п	Страны	Продукты					
		Мясо, кг	Масло животное, кг	Сахар, кг	Алкоголь, л	Фрукты, кг	Хлебопродукты, кг
		X1	X2	X3	X4	X5	X6
1	Россия	55	3,9	30	5	28	124

2	Австралия	100	2,6	47	8,2	121	87
3	Австрия	93	5,3	37	12	146	74
4	Азербайджан	20	4,1	12,4	7,9	52	141
5	Армения	20	3,7	4,3	6,5	72	134
6	Беларусь	72	3,6	28	5,4	38	120
7	Бельгия	85	6,9	48	11	83	72
8	Болгария	65	3	18	9,5	92	156
9	Великобритания	67	3,5	39	8,8	91	91
10	Венгрия	73	1,7	40	10,9	73	106
11	Германия	88	6,8	35	8,1	138	73
12	Греция	83	1	24	8,8	99	108
13	Грузия	21	3,8	36	9,8	55	140
14	Дания	98	5	38	10,3	89	77
15	Ирландия	99	3,3	31	9,6	87	102
16	Испания	89	0,4	26	8,95	103	72
17	Италия	84	2,2	27	9,6	169	118
18	Казахстан	61	4,2	19,2	7,2	10	191
19	Канада	98	3,1	44	7,4	123	77
20	Киргизия	46	4,1	23,5	6,7	20	134

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 7.

Решить Задание 6, предварительно нормировав исходные данные.

Задание 8.

Решить Задание 6 при условии, что расстояния между кластерами измеряются по принципу “дальнего соседа”, предварительно нормируя исходные данные.

Задание 9.

Решить Задание 6, 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.

ГЛАВА 2. ФАКТОРНЫЙ АНАЛИЗ И ЕГО РЕЛИЗАЦИЯ В ПАКЕТЕ STATISTICA

2.1. Введение в факторный анализ

2.1.1 Цели и сущность факторного анализа

Основное понятие факторного анализа (фактор) определяется следующим образом:

факторы - гипотетические непосредственно не измеряемые, скрытые (латентные) переменные в той или иной мере связанные с измеряемыми характеристиками – проявлениями этих факторов.

Идея факторного анализа основана на предположении, что имеется ряд величин, не известных исследователю, которые заставляют проявляться различные соотношения между переменными. То есть структура связей между p анализируемыми признаками $x^{(1)}, \mathbf{K}, x^{(p)}$ может быть объяснена тем, что все эти переменные зависят (линейно или как-то еще) от меньшего числа других, непосредственно не измеряемых факторов $f^{(1)}, \mathbf{K}, f^{(m)}$ $(m < p)$, которые принято называть общими. Такая взаимозависимость может быть расценена как своего рода базис взаимосвязи между рассматриваемыми переменными. Таким образом, *факторный анализ* (в широком смысле) - совокупность моделей и методов, ориентированных на выявление, конструирование и анализ внутренних факторов по информации об их “внешних” проявлениях. В узком смысле под *факторным анализом* понимают методы выявления гипотетических (ненаблюдаемых) факторов, призванных объяснить корреляционную матрицу количественных наблюдаемых переменных.

Существуют 2 уровня факторного анализа: разведочный (эксплораторный), когда не известно ни количество факторов, ни структура связи; проверочный (конфирматорный) – осуществляется проверка гипотезы о влиянии факторов.

Большинство моделей конструируется так, чтобы общие факторы оказались некоррелированными. При этом в общем случае не постулируется возможность однозначного восстановления значений каждого из наблюдаемых признаков $x^{(j)}$ по соответствующим значениям общих факторов $f^{(1)}, \mathbf{K}, f^{(m)}$: допускается, что любой из исходных признаков $x^{(j)}$ зависит также и от некоторой своей (“специфической”) остаточной случайной компоненты $e^{(j)}$ - характерного фактора, который и

обуславливает статистический характер связи между $x^{(j)}$ с одной стороны и $f^{(1)}, \mathbf{K}, f^{(m)}$ с другой.

Конечная цель статистического исследования, проводимого с привлечением факторного анализа, как правило, состоит в выявлении и интерпретации латентных общих факторов с одновременным стремлением минимизировать их число и степень зависимости $x^{(j)}$ от своих характерных факторов $e^{(j)}$. Как и в любой модельной схеме, эта цель может быть достигнута лишь приближенно. Принято считать статистический анализ такого рода успешным, если большое число переменных удалось объяснить малым числом факторов. Являются ли факторы причинами или просто агрегированными теоретическими конструкциями зависит от интерпретации модели.

Методы и модели факторного анализа нацелены на сжатие информации или, что тоже, на снижение размерности исходного признакового пространства. При этом методы факторного анализа базируются в основном на возможности снижения размерности с помощью использования взаимной коррелированности исходных признаков (могут использоваться также малая “вариабельность” некоторых из них, агрегирование).

Факторный анализ позволяет выявить зависимость между явлениями, обнаружить скрытую основу нескольких явлений, ответить на вопрос, почему связаны явления.

Как метод статистического исследования факторный анализ включает следующие основные этапы:

1) Формирование цели

Цели могут быть:

a) Исследовательские (выявление факторов и их анализ)

b) Прикладные (построение агрегированных характеристик для прогнозирования и управления)

2) Выбор совокупности признаков и объектов

3) Получение исходной факторной структуры

4) Корректировка факторной структуры исходя из целей исследования.

При проверочном факторном анализе критерий качества - соответствие структуры заданной исследователем, при разведочном - достижение "простой структуры", когда связь переменных с каким-либо фактором выражена максимально чётко.

Иногда коррекция приводит к коррелированным факторам. Это позволяет повторно применить факторный анализ, используя в качестве исходной информации значения факторов. В связи с этим возможен такой промежуточный этап, как

5) Выявление факторов второго порядка

Поучаем факторы второго порядка - более общие и глубокие категории исследуемого явления.

6) Интерпретация и использование результатов.

Основные идеи факторного анализа в рамках традиционных представлений сводятся к следующему:

- Сущность вещей заключена в их простых и многообразных проявлениях, которые могут быть объяснены с помощью комбинации нескольких основных факторов.
- Общая сущность наблюдаемых вещей постигается через бесконечное приближение.

Реализуя эти принципы, факторный анализ позволяет получить простую модель взаимосвязи явлений на более высоком, причинном уровне, что очень ценно как в теоретических, так и в практических исследованиях.

2.1.2. Линейная модель факторного анализа

Как и любая задача, задача факторного анализа имеет свои условия, выполнение которых обязательно для эффективного применения аппарата:

- Необходимо установить границы области, наделённой структурой, провести анализ объёма данных, установить уровень различения переменных.
- При выборе переменных важно сохранить возможность их классификации.
- Число переменных p должно соответствовать числу наблюдений n: $n >> p$.

Модель факторного анализа определяет некоторые количественные соотношения между переменными. Она может быть выражена в коэффициентах или процентных отношениях, которые указывают, до какой степени рассматриваемые переменные подвержены влиянию некоторых общих факторов.

Факторный анализ можно применять к:

§ Вектору отклонений $y = x - \bar{x}$ с $E(x - \bar{x}) = E(y = 0)$. В этом случае исходный материал – матрица ковариаций $K = E(yy')$;

§ Нормированному вектору $z = S^{-1}(x - \bar{x}) = S^{-1}y$, где S-диагональная матрица стандартных отклонений, $E(z = 0)$. Основа – ковариационная матрица R, которая в этом случае равна корреляционной;

§ Наблюдаемому вектору X без поправки на среднее. Рассматриваем матрицу $E(xx')$.

Для удобства будем полагать исследуемые наблюдения $X_1, \mathbf{K}, X_n$ нормированными.

Традиционная модель основана на представлении матрицы наблюдений $X = (x_{ik})$, где x_{ik} - значение k-го признака для i-го объекта, в виде линейных комбинаций значений f_{it} факторов f_t на объектах с невязками e_{ik}:

$$x_{ik} = a_{1k} f_{i1} + \mathbf{K} + a_{tk} f_{it} + e_{ik},\ i = 1, \mathbf{K}, n;\ k = 1, \mathbf{K}, p;\ t = 1, \mathbf{K}, m.$$

a_{tk} - нагрузки факторов на признак k. Выбор a_{tk}, f_{it} осуществляется по критерию минимизации корреляций между векторами невязок $e_k = (e_{ik})$ – характерными факторами.

Соотношения факторного анализа формально воспроизводят запись модели множественных регрессий, в которой под $f^{(i)}$ $(i = 1,2, \mathbf{K}, m)$ понимаются так называемые *объясняющие переменные* (факторы-аргументы). Однако принципиальное отличие модели факторного анализа от регрессионных схем состоит в том, что переменные $f^{(j)}$, выступающие в роли аргументов в моделях регрессии, не являются непосредственно наблюдаемыми в моделях факторного анализа, в то время как в регрессионном анализе значения $f^{(j)}$ измеряются на статистически обследованных объектах.

Модель (1) в матричном виде выглядит как **X**=**AF**+**E**. Здесь **A** = a_{ij} — прямоугольная $p \times m$ *матрица нагрузок* общих факторов на исследуемые признаки, связывающих исследуемые признаки $x^{(i)}$ с ненаблюдаемыми общими факторами $f^{(1)}, \mathbf{K}, f^{(m)}$: если a_{ij} достаточно большая, то фактор $f^{(j)}$ хорошо описывает поведение наблюдаемой переменной $x^{(i)}$. Вектор-столбец E определяет ту часть каждой из наблюдаемых величин, которая не может быть объяснена общими факторами. Невязка **E** разложения факторного анализа включает характерную часть исследуемых переменных - E_s и ошибку измерений E_n. Предполагается, что компоненты **F** и **E** некоррелированы; без ограничения общности можно рассматривать $D(e_i) = 1,\ \ i = 1,2, \mathbf{K}, p$ и $D(f_j) = 1,\ \ j = 1,2, \mathbf{K}, m$. Тогда математическое ожидание $E(X) = 0$ и матрица ковариаций

$$D(X)=D(\mathbf{AF}+\mathbf{E})=\mathbf{AA}^{\mathrm{T}}+\mathbf{L}^2,$$

где $\mathbf{AA}^{\mathrm{T}}$ называется *матрицей общности* и отражает изменение переменных под влиянием общих факторов, диагональные элементы этой матрицы - *общности* - h_i^2. $\mathbf{L}^2$ называется *характерной матрицей* и указывает на специфическую связь переменных, её диагональные элементы $l_i^2 = 1 - h_i^2$ определяют *характерность*.

Фактически, мы разделили факторы, влияющие на переменные, на три группы:

a) *Общие факторы*, $f_j \ \ (j = 1,2,\mathbf{K},m)$: факторы, которые влияют на несколько переменных $x_i \ \ (i = 1,2,\mathbf{K},p)$ одновременно (фактор, входящий во все наблюдения называется *главным*);

b) *Характерные факторы*, $e_{si} \ \ (i = 1,2,\mathbf{K},p)$: факторы, которые одновременно влияют только на одну переменную;

c) *Факторы погрешности*, $e_{ni} \, (i = 1,2,\mathbf{K},p)$: факторы, к которым относится погрешность в наблюдениях, они могут быть случайными компонентами.

Основные различия между общими и характерными факторами: общий фактор воздействует на несколько переменных $x_i \, (i = 1,2,\mathbf{K},p)$ сразу, определяя одну модель поведения переменных, а специфические - только на одну переменную; переменная x_i может одновременно зависеть от нескольких общих факторов, но только от одного характерного и одного фактора погрешности.

Аналогично дисперсию наблюдаемых переменных можно разбить следующим образом: $s_i^2 = h_i^2 + l_{si}^2 + l_{ni}^2$, где $h_i^2 + l_{si}^2$ - надёжность, а $l_{si}^2 + l_{ni}^2$ - характерность или ненадёжность.

Итак, модельная оценка переменных **AF** воспроизводит исходные данные с точностью до остатков **E**, представляющих невязку (характерные и ошибочные факторы). **AF** представляет собой общую (редуцированную) часть, а **E** - специфическую часть. Таким образом, если дана выборка наблюдаемого вектора X с n элементами, то главная численная задача линейной факторной модели в том, чтобы оценить матрицу нагрузок **A**. Часто необходимо оценить и характерную матрицу $\mathbf{L}^2$. $S = \{A, L^2\}$ представляет *структуру факторного анализа.*

Обычно, в факторном анализе мало внимания уделяется характерному фактору и фактору погрешности, чтобы связать применяемый факторный анализ исключительно с общими факторами. Однако пренебрежение специфическими факторами не всегда

оправдано. Присутствие переменной с высокой характерной дисперсией или высокой компонентой дисперсии ошибки может быть сигналом, что данная переменная, вероятно, выпадает из общего ряда и связана с переменными, еще не включенными в рассмотрение. Если эта переменная более важна, чем остальные, то должна быть введены новые переменные.

Тем не менее, факторный анализ нацелен, прежде всего, на анализ общих факторов f_j и соответствующих факторных нагрузок. Практическая модель факторного анализа, выраженного на нормированном наблюдении часто записывается так:

$$x_{it} = a_{i1} f_{1t} + \mathbf{K} + a_{im} f_{mt},$$

где e_{si} и $e_{ni}\ (i = 1,2,\mathbf{K}, p)$ полагаются равными 0.

В матричном обозначении это:

$$\mathbf{X} = \mathbf{AF}.$$

Особенность модели (1) - существенная неоднозначность: вместе с любым решением $\mathbf{A}=(a_{tk})$ и $\mathbf{F}=(f_{it})$ найденной системе невязок $\mathbf{E}=(e_{ik})$ соответствуют и всевозможные решения вида $\mathbf{A}'=\mathbf{BA}$, $\mathbf{F}'=\mathbf{FB}^{-1}$, где **B** - произвольная невырожденная матрица $m \times m$. В принятых матричных обозначениях эта модель имеет вид матричного равенства $\mathbf{X}=\mathbf{FA}+\mathbf{E}$, которое не нарушится при $\mathbf{A}'=\mathbf{BA}$ **и** $\mathbf{F}'=\mathbf{FB}^{-1}$. Эта неопределённость сохраняется и для более простой модели, включающей только общие факторы.

Геометрически совокупность факторов $f_1, \mathbf{K}, f_m$ задаёт базис пространства факторов, а **B** - преобразование (в общем случае *косоугольное вращение*) этого базиса, не меняющее факторное пространство.

Фактически имеем 2-х этапную процедуру:

1) ищем факторное решение по критерию, учитывающему только невязки

2) с помощью вращения приводим к виду, наиболее отвечающему целям исследования.

Идеальной матрицей нагрузок считаем ту, которая позволяет максимально чётко разделить переменные по тому, какой фактор проявляется в них наиболее сильно.

Итак, сформулированная общая модель факторного анализа позволяет эффективно достигать целей исследования и уже исходя из построения, задаёт определённое направление интерпретации результатов.

Один из необходимых моментов исследования составляет проверка гипотез, связанных с природой и параметрами используемой модели факторного анализа. Теория статистических критериев применительно к моделям факторного анализа разработана

слабо. Пока существуют лишь так называемые критерии адекватности модели, т.е. критерии, предназначенные для проверки гипотезы H_0, заключающейся в том, что исследуемый вектор наблюдений *X* допускает представление с помощью модели факторного анализа с данным (заранее выбранным) числом общих факторов *m*. На языке ковариационных матриц гипотеза H_0 означает, что элементы матрицы $\hat{\Sigma} - (\hat{A}\hat{A}^T + \hat{L}^2)$ должны лишь статистически незначимо отличаться от нуля, или, что эквивалентно, матрица $\hat{\Sigma} - \hat{L}^2$ должна иметь ранг, равный m. А это в свою очередь означает что последние $p - m$ характеристических корней $\hat{l}_{m+1}, \ldots, \hat{l}_p$ уравнения $\left|\hat{\Sigma} - \hat{L}^2 - l\hat{L}^2\right| = 0$ должны лишь незначимо отличаться от нуля.

2.1.3. Решение модели факторного анализа

Исследователю не известны коэффициенты a_{ij}, но он хочет научиться наилучшим (в некотором смысле) образом аппроксимировать признаки $x^{(1)}, \mathbf{K}, x^{(p)}$ с помощью линейных функций от небольшого (заранее определенного) числа *m* факторов $f^{(1)}(m), \mathbf{K}, f^{(m)}(m)$, которые поэтому естественно назвать *общими*. Аппроксимация признаков *X* с помощью $f^{(1)}(m), \mathbf{K}, f^{(m)}(m)$ означает представление *X* в виде $X(m) = A_m F(m)$, где A_m — матрица порядка $p \times m$, составленная из первых m столбцов матрицы A, а $F(m) = \left(f^{(1)}(m), \mathbf{K}, f^{(m)}(m)\right)^T$. Общая же задача сводится к получению оценок общей и характерной компонент дисперсии исходных переменных *X* (*A* и L^2) и самих *X*. Это эквивалентно определению структуры факторного анализа $S = \{A, L^2\}$.

При введении модели факторного анализа было показано, что матрицу ковариаций исходных наблюдений Σ можно представить в виде (4.2): $\Sigma = AA^T + L^2$, где *A*-матрица факторных нагрузок (AA^T - матрица общности), а $\mathbf{L}^2$- диагональная характерная матрица.

По-разному формулируя критерий оптимальности аппроксимации *X* с помощью $F(m)$, придем к различным способам решения модели: методу главных компонент или общих факторов. Так, например, если определение элементов матрицы A_m подчинить идее минимизации отличия ковариационной матрицы Σ исследуемого вектора *X* от

ковариационной матрицы $\Sigma_{\hat{X}} = A_m A_m^T$ аппроксимирующего вектора $X(m)$ (в смысле минимизации евклидовой нормы $\left\|\Sigma - \Sigma_{\hat{X}}\right\|$), то $f^{(i)}(m)$ определяется пропорционально *i*-й главной компоненте вектора X, в частности $f^{(i)}(m) = l_i^{-1/2} z^{(i)}$, где l_i — *i*-й по величине характеристический корень ковариационной матрицы Σ, а $z^{(i)}$ - i-я главная компонента *X*; при этом 1-й столбец матрицы A_m, $(i = 1,2,\mathbf{K},m)$ есть $\sqrt{l_i} P_i^T$, где P_i, — собственный вектор матрицы $\sum$, соответствующий характеристическому корню l_i. Здесь мы пользуемся методом главных компонент.

Если же определение аппроксимирующего вектора $X(m) = A_m F(m)$ подчинить идее максимального объяснения корреляции между исходными признаками $x^{(i)}$ и $x^{(j)}$ с помощью вспомогательных (ненаблюдаемых) факторов $f^{(1)}(m), \mathbf{K}, f^{(m)}(m)$ и, в частности, идее минимизации величины

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left(\frac{\operatorname{cov}(x^{(i)}, x^{(j)}) - \operatorname{cov}(\hat{x}^{(i)}(m), \hat{x}^{(j)}(m))}{(s_{ii} s_{jj})^{1/2}} \right)^2$$

при условии неотрицательности величин $s_{ii} - D\hat{x}^{(i)}$, то можно показать, что 1-я строка оптимальной в этом смысле матрицы преобразования A_m, состоит из *m* факторных нагрузок общих факторов $f^{(1)}(m), \mathbf{K}, f^{(m)}(m)$ на i-й исходный признак $x^{(i)}$ в модели факторного анализа. Другими словами, сущность задачи минимизации (по A_m и $F(m)$) величины (4) состоит в следующем. Первый из m общих факторов $f^{(1)}(m)$ находится из условия, чтобы попарные корреляции между исходными признаками были как можно меньше, если влияние на них этого фактора $f^{(1)}(m)$ учтено. Следующий общий фактор $f^{(2)}(m)$ находится из условия максимального ослабления попарных корреляционных связей между исходными признаками, оставшихся после учета влияния первого общего фактора $f^{(1)}(m)$, и т.д. Этот метод носит название метода главных факторов.

Из сказанного, в частности, следует, что методы главных компонент и главных факторов должны давать близкие результаты в тех случаях, когда главные компоненты

строятся по корреляционными матрицами исходных признаков, а остаточные дисперсии e_i^2 сравнительно невелики.

Подробнее остановимся на описании практической реализации метода главных компонент:

Найдём собственные числа l_i и собственные векторы P_i ковариационной матрицы K (K - выборочный аналог Σ, который как правило доступен исследователю), такие что $KP_i = lP_i$. Так как ковариационная матрица симметрична, её собственные числа действительны; упорядочим их по убыванию. Будем считать, что P_i – нормированный вектор-столбец. Собственные вектора, соответствующие разным собственным значениям ортогональны, то есть P_i ортогонален P_j при $i \neq j$. Определим матрицу $P = (P_1, \mathbf{K}, P_p)$, тогда $KP = PC$, где C – диагональная матрица с соответствующими собственными числами на диагонали. Преобразуем полученное соотношение следующим способом: $KPP^T = PCP^T$, пользуясь свойством ортогональности различных собственных векторов и предположением о нормированности P ($PP^T = I$), получим $K = PCP^T$. Положив Ah=PC$^{1/2}$, получим K=AhAh^{T}, то есть одно из приемлемых фундаментальных (неповёрнутых) решений, дающее искомую матрицу факторных нагрузок A. Преобразовав исходные данные X как $Y=PX$, получим максимально простую линейную комбинацию факторов (диагональную матрицу нагрузок). Y_i называются главными компонентами.

2.1.4 Задача о числе факторов

До сих пор не удалось построить многомерной решающей процедуры для неизвестного числа общих факторов m. При использовании только что описанного метода для отбора числа факторов используется один из следующих критериев:

- *Критерий Кайзера*: в рассмотрение принимаются только факторы, которым соответствуют собственные числа ковариационной матрицы больше 1.
- *Критерий каменистой осыпи*: отбрасываем все факторы, соответствующие собственные числа которых мало отличаются.

В общем же случае приходится ограничиваться последовательной эксплуатацией критериев адекватности $H_0 : m = m_0$ при альтернативе $H_1 : m > m_0$. Если гипотеза H_0 отвергается, то переходят к проверке гипотезы $H_0' : m' = m_0 + 1$ при альтернативе

$H_1' : m' > m_0 + 1$ и т.д. Однако по уровням значимости a каждой отдельной стадии такой процедуры трудно сколько-нибудь точно судить о свойствах всей последовательной процедуры в целом.

В разделе постановки модели факторного анализа уже описывалась возникающая неопределённость решения, когда мы получаем бесконечное число матриц нагрузок, отличающихся некоторым ортогональным преобразованием, удовлетворяющих всем условиям. Геометрически это означало, что систему координат, определённую осями-факторами, можно вращать в общем факторном пространстве, не нарушая структуру модели. Тем не менее, проблема остаётся: надо понять, при каких дополнительных условиях на искомую матрицу нагрузок A и на соотношение между p и m не может существовать двух различных решений $A^{(1)}$ и $A^{(2)}$ таких, чтобы одно из них нельзя было бы получить из другого с помощью соответствующим образом подобранного ортогонального преобразования B (единственность с точностью до ортогонального преобразования или с точностью до вращения факторов). Оказывается, достаточным условием единственности такого рода является требование к матрице A, чтобы при вычеркивании из нее любой строки оставшуюся матрицу можно было бы разделить на две подматрицы ранга m, откуда автоматически следует требование $m < (p-1)/2$. Далее будем полагать выполнение этого условия (существование единственного с точностью до ортогонального преобразования решения факторной модели).

Выбор определённого решения в этом случае связан с достижением простой структуры факторного анализа. Структура проста, если любая строка матрицы факторных нагрузок A имеет максимальное (меньшее $m-1$) число нулевых элементов. Это позволяет выделить множество наблюдаемых переменных с наибольшей нагрузкой на латентном факторе и реально объяснить значение фактора.

Как правило, исследователю известна лишь оценка K ковариационной матрицы Σ - ее выборочное значение. Необходимый результат: оценки факторных нагрузок, остаточных дисперсий и самих факторов.

2.1.5 Статистическая оценка факторных нагрузок и остаточных дисперсий

Логическая схема итераций следующая:

• задаемся некоторым нулевым приближением $L^{2(0)}$ матрицы L^2;

• используя соотношение между K, A и L^2 получаем нулевое приближение $K^{(0)} = K - L^{2(0)}$ матрицы $K^* = AA^T = K - L^2$;

• по *K** последовательно определяем нулевые приближения $a_1^0, \mathbf{K}, a_m^0$ для столбцов *A*. Затем определяем следующее (первое) приближение $L^{2(1)}$ и т. д.

Для факторных нагрузок a_{ij} и остаточных дисперсий l_{it}^2 оценивание производится либо методом максимального правдоподобия, либо так называемым *центроидным методом*. Первый метод хотя и дает эффективные оценки для a_{ij} и l_{it}^2, но требует постулирования закона распределения исследуемых величин (разработан лишь в нормальном случае), а также весьма обременительных вычислений. Центроидный метод, давая оценки, близкие к оценкам максимального правдоподобия, как и всякий непараметрический метод, является более «устойчивым» по отношению к отклонениям от нормальности исследуемых признаков и требует меньшего объема вычислений. Однако из-за определенного произвола в его процедуре, которая приведена ниже, статистическая оценка метода, исследование его выборочных свойств (в общем случае) практически невозможны.

Общая схема реализации метода максимального правдоподобия следующая. Составляется логарифмическая функция правдоподобия как функция неизвестных параметров a_{ij} и l_{it}^2, отвечающая исследуемой модели, с учетом нормальности $X_1, \mathbf{K}, X_p$. С помощью дифференцирования этой функции правдоподобия по каждому из неизвестных параметров и приравнивания полученных частных производных к нулю получается система уравнений, в которой известными величинами являются выборочные ковариации $\hat{s}_{ij}$, а также числа *p* и *m,* а неизвестными — искомые параметры a_{ij} и l_{it}^2. И наконец, предлагается вычислительная (как правило, итерационная) процедура решения этой системы. Отметим, что при достаточно общих ограничениях доказана асимптотическая нормальность оценок максимального правдоподобия *A* и L^2, что дает основу для построения соответствующих интервальных оценок.

Как выше отмечено, центроидный метод является одним из способов реализации вычислительной схемы, приспособленной для выявления структуры модели факторного анализа и оценки неизвестных параметров. Этот метод поддается весьма простой геометрической интерпретации. Отождествим исследуемые признаки $x^{(1)}, \mathbf{K}, x^{(p)}$ с векторами, выходящими из начала координат некоторого вспомогательного *p*-мерного пространства, построенными таким образом, чтобы косинусы углов между $x^{(i)}$ и $x^{(j)}$

равнялись бы их парным корреляциям (r_{ij}), а длины векторов $x^{(i)}$ — стандартным отклонениям соответствующих переменных ($s_{ii}^{1/2}$). Далее изменим, если необходимо, направления, т. е. знаки отдельных векторов так, чтобы как можно больше корреляций стало положительными. Тогда векторы будут иметь тенденцию к группировке в одном направлении в пучок. После этого первый общий фактор $f^{(1)}$ определяется как нормированная (т. е. как вектор единичной длины) сумма всех исходных векторов пучка, которая будет проходить каким-то образом через середину (центр) этого пучка; отсюда название «центроид» для общего фактора в этом случае.

Переходя затем к остаточным переменным $x^{(i1)} = x^{(i)} - a_{i1} f^{(1)}$. Подсчитывая ковариационную матрицу $K^{(1)} = K - a_1 a_1^T$ для этих остаточных переменных и проделывая относительно $x^{(i1)}$ и $K^{(1)}$ ту же самую процедуру построения пучка и т.п., выделяем второй общий фактор («второй центроид») $f^{(2)}$ и т. д.

Формализация этих соображений приводит к следующей итерационной схеме вычислений по определению факторных нагрузок a_{ij} и остаточных дисперсий l_{it}^2. Задаемся некоторым начальным приближением $L^{2(0)}$ для дисперсии остатков L^2. Обычно полагают $l_{ii}^{2(0)} = \hat{s}_{ii}\left[1 - \max_{\substack{1 \le j \le p \\ (j \ne i)}} |r_{ji}|\right]$

Подсчитываем $K^{(0)} = K - L^{2(0)}$. Выбираем в качестве нулевого приближения $b_1^{(0)}$ первого столбца b_1 вспомогательной матрицы **B** столбец, состоящий из одних единиц. Далее определяем нулевое приближение $a_1^{(0)}$ первого столбца матрицы нагрузок

$$a_1^{(0)} = \frac{K^{(0)} b_1^{(0)}}{(b_1^{(0)T} K^{(0)} b_1^{(0)})^{1/2}}$$

Затем вычисляется матрица $K_1^{(0)} = K^{(0)} - a_1^{(0)} a_1^{(0)T}$ определяется нулевое приближение $a_2^{(0)}$ второго столбца матрицы нагрузок

$$a_2^{(0)} = \frac{K^{(0)} b_2^{(0)}}{(b_2^{(0)T} K_1^{(0)} b_2^{(0)})^{1/2}}$$

где вектор $b_2^{(0)}$ состоит только из ± 1, а знаки подбираются из условия максимизации знаменателя правой части и т.д. Получив, таким образом, нулевое приближение $A^{(0)} = \left(a_1^{(0)}, \mathbf{K}, a_{m'}^{(0)}\right)$ для матрицы нагрузок **A,** вычисляем $L^{2(1)} = K - A^{(0)}A^{(0)T}$ и переходим к следующей итерации. При этом матрица $B^{(1)}$ не обязана совпадать с $B^{(0)}$. i-й столбец матрицы B задает веса, с которыми суммируются векторы одного пучка для образования i-го общего фактора («центроида»). Поскольку смысл центроидной процедуры в простом суммировании векторов пучка, она иногда так и называется — *«процедура простого суммирования»,* то исследователю остается определить лишь нужное направление каждого из векторов пучка) т.е. знаки единиц, образующих столбцы b_i. Непосредственная ориентация (при подборе знаков у компонент вектора b_i) на максимизацию выражений $b_i^{(v)T} K_{i-1}^{(v)} b_i^{(v)}$ хотя и несколько сложнее реализуема, чем некоторые приемы, опирающиеся на анализ знаков элементов остаточных матриц K_{i-1}, но быстрее и надежнее приводит к выделению именно таких центроидов, которые при заданном m будут обусловливать возможно большую часть обшей дисперсии исходных признаков, т. е. минимизировать дисперсию остаточных компонент l_{it}^2. Недостатком центроидного метода является зависимость центроидных нагрузок a_{ij} от шкалы, в которой измерены исходные признаки. Поэтому исходные признаки $x^{(i)}$ обычно нормируют с помощью среднеквадратических отклонений $\sqrt{s_{ii}}$, так что выборочная ковариационная матрица заменяется во всех рассуждениях выборочной корреляционной матрицей.

Анализируя описанную выше процедуру центроидного метода, нетрудно понять, что построенные таким способом общие факторы могут интерпретироваться как первые m «условных» главных компонент матрицы K—L^2, найденные при дополнительном условии, что компоненты соответствующих собственных векторов могут принимать лишь два значения: ± 1, то есть мы видим явную аналогию с методом главных компонент.

*Оценка значений общих факторов***.** Это одна их основных задач исследования. Действительно, мало установить лишь сам факт существования небольшого числа скрыто действующих общих факторов $f^{(1)}, \mathbf{K}, f^{(p')}$, объясняющих природу взаимной коррелированности исходных признаков и основную часть их дисперсии. Желательно

непосредственно определить эти общие факторы, описать их в терминах исходных признаков и постараться дать им удобную содержательную интерпретацию.

Приведем здесь идею и результат 2 наиболее распространенных методов решения этой задачи:

Метод Бартлетта (1938). Этот метод рассматривает отдельно для каждого фиксированного номера наблюдения $v\,(v=1,2,\mathbf{K},n)$ модель факторного анализа как регрессию признака x_i по аргументам $a_{i1},\mathbf{K},a_{im}$; при этом индекс $(i=1,2,\mathbf{K},p)$ у признака и нагрузок играет в данном случае роль номера наблюдений в этой регрессионной схеме, так что

$$x_n^{(i)}=\sum_{j=1}^{m} f_n^{(j)}a_{ij}+e_n^{(i)} \qquad (i=1,2,\mathbf{K},p).$$

Таким образом, величины $f_n^{(1)},\mathbf{K},f_n^{(m)}$ интерпретируются как неизвестные коэффициенты регрессии x_v по факторным нагрузкам. В соответствии с известной техникой метода наименьших квадратов (с учетом того, что, вообще говоря, $Dx^{(i1)}\neq Dx^{(i2)}$ при $i1\neq i2$), определяющей неизвестные коэффициенты регрессии $\hat{F}_n=(\hat{f}_n^{(1)},\ldots\hat{f}_n^{(m)})^T$ из условия

$$\sum_{i=1}^{p}\frac{1}{s_{ii}}(x_n^{(i)}-\sum_{j=1}^{m}\hat{f}_n^{(j)}a_{ij})^2=\min_{F_n}\sum_{i=1}^{p}\frac{1}{s_{ii}}(x_n^{(i)}-\sum_{j=1}^{m}f_n^{(j)}qa_{ij})^2,$$

Получаем $\hat{F}_n=(A^TL^{2-1}A)^{-1}A^TL^{2-1}X_n,(n=1,..,n)$.

Очевидно, если исследуемый вектор наблюдений X нормален, то эти оценки являются одновременно и оценками максимального правдоподобия. Нестрогость данного метода — в замене истинных (неизвестных нам) величин a_{ij} и l_{it}^2 их приближенными (оценочными) значениями.

Модель Томсона рассматривает модель факторного анализа как бы «вывернутой наизнанку», а именно как регрессию зависимых переменных $f^{(1)},\mathbf{K},f^{(m)}$ по аргументам $x^{(1)},\mathbf{K},x^{(p)}$. Тогда коэффициенты $\hat{c}_{ij}$ в соотношениях $\hat{f}^{(i)}=\sum_{j=1}^{p}\hat{c}_{ij}x^{(j)}$ $(i=1,2,\mathbf{K},m)$,

или в матричной записи **F**=**CX**, где **C** — матрица коэффициентов c_{ij} размера $m \times p$, находят в соответствии с методом наименьших квадратов из условия

$$\sum_{n=1}^{n}\sum_{i=1}^{m}(\hat{f}_n^{(i)} - \sum_{j=1}^{p}\hat{c}_{ij}x_n^{(j)})^2 = \min_{c_{ij}}\sum_{n=1}^{n}\sum_{i=1}^{m}(\hat{f}_n^{(i)} - \sum_{j=1}^{p}c_{ij}x_n^{(i)})^2$$

Поскольку решение экстремальной задачи выписывается в терминах ковариаций $x^{(i)}$ и $f^{(i)}$, то отсутствие наблюдений по зависимым переменным $f^{(j)}$ можно компенсировать знанием этих ковариаций, так как легко подсчитать, что

$$E\left\{\begin{bmatrix} x^{(1)} \\ \mathbf{K} \\ x^{(p)} \\ f^{(1)} \\ \mathbf{K} \\ f^{(m)} \end{bmatrix}(x^{(1)},..,x^{(p)}, f^{(1)},.., f^{(m)})\right\} = \begin{pmatrix} AA^T + L^2 & A \\ A^T & I_m \end{pmatrix}.$$

Отсюда, используя известные формулы метода наименьших квадратов, получаем

$F_v = (I + \Gamma)^{-1}\hat{A}^T\hat{L}^{2^{-1}}X_v \quad (v = 1,2,\mathbf{K},n)$, где матрица Γ (размера $p \times p$) определяется соотношением $\Gamma = \hat{A}^T\left(\hat{L}^2\right)^{-1}\hat{A}$

Если элементы матрицы $\hat{A}^T\left(\hat{L}^2\right)^{-1}\hat{A}$ достаточно велики, то эти два метода будут давать близкие решения.

Цель факторного анализа не только, чтобы определить, но также и дать существенное истолкование матрице нагрузок **A**. Каждый столбец содержит весь общий фактор или коэффициенты факторной нагрузки, которые указывают степень зависимости переменных $x_i \ (i = 1,2,\mathbf{K},p)$ от факторов $f_j \ (j = 1,2,\mathbf{K},m)$. Истолкование факторов основано на коэффициентах факторной модели a_{ij}, в особенности использовании общностей a_{ij}^2 $(i = 1,2,\mathbf{K},p)$ для фактора $f_j \ (j = 1,2,\mathbf{K},m)$. Таблица, которая иллюстрирует матрицу возведенных в квадрат факторных нагрузок, может интерпретироваться двумя способами:

a) в столбцах: матрица отражает характер связи наблюдаемых переменных;

b) в строках: матрица указывает степень связи общей дисперсии $x_i \ (i = 1,2,\mathbf{K},p)$ с каждым отдельным фактором.

Не смотря на большую работу в данной области, содержательная интерпретация скрытых факторов, как основная творческая часть, остаётся одной из главных проблем факторного анализа. Также к нерешённым проблемам следует отнести разработку эффективных общих методов отбора числа факторов и системы гипотез о параметрах модели.

При разработке моделей большое внимание следует уделить анализу исходных данных и характерных факторов, которые, имея большие значения, могут нарушить структуру факторного анализа и привести к ошибочным результатам.

2.2. Пример решения практической задачи факторным анализом в системе STATISTICA

Рассмотрим основные этапы проведения кластерного анализа в системе STATISTICA на следующем примере.

Исходными показателями послужили:

X1 – численность населения (тыс.)

X2 – количество человек, приходящихся на одного врача

X3 - Расходы на здравоохранение на душу населения ($)

X4 – Уровень детской смертности

X5 – ВВП, рассчитанный по паритету покупательной способности на душу населения (млн. $)

X6 - Смертность на 1000 человек

В файле (рис. 2.1) содержатся данные по 12 странам (по медицинской характеристике).

Data: Медицинская характеристика (6v by 12c)

	1 Числ_Насел	2 Кол_чел_на_1врача	3 Расх_на_здрав	4 Ур_детск_смерт	5 ВВП_на_душу_насел	6 Смертность
Россия	145491	235	159	16,8	7700	13,9
Азербайджан	8041	256	99	29,3	3000	9,6
Армения	3787	198	152	15,4	3000	9,7
Белоруссия	10187	222	157	12,5	7500	14
Грузия	5262	182	152	17,6	4600	14,6
Казахстан	16172	265	154	42,1	5000	10,6
Киргизия	4921	301	118	37	2700	9,1
Молдавия	4295	251	143	20,5	2500	12,6
Таджикистан	6087	439	100	53,3	1140	8,6
Туркмения	4737	320	125	48,6	4300	9
Узбекистан	24881	299	116	36,7	2400	8
Украина	49568	224	131	15,3	3850	16,4

рис. 2.1

Задачей факторного анализа является объединение большого количества показателей, признаков, которыми характеризуется экономический процесс или объект, в меньшее количество искусственно построенных на их основе факторов, чтобы полученная в итоге система факторов (столь же хорошо описывающая выборочные данные, что и исходная) была наиболее удобна с точки зрения содержательной интерпретации.

Алгоритм выполнения

Модуль **Factor Analysis** (Факторный анализ) содержит широкий набор методов, снабжающих пользователя исчерпывающими средствами выделения факторов и представления результатов. Для вызова данного модуля можно использовать **STATISTICA Module Switcher** (Переключатель модулей), который содержит список всех доступных модулей (рис. 2.2)

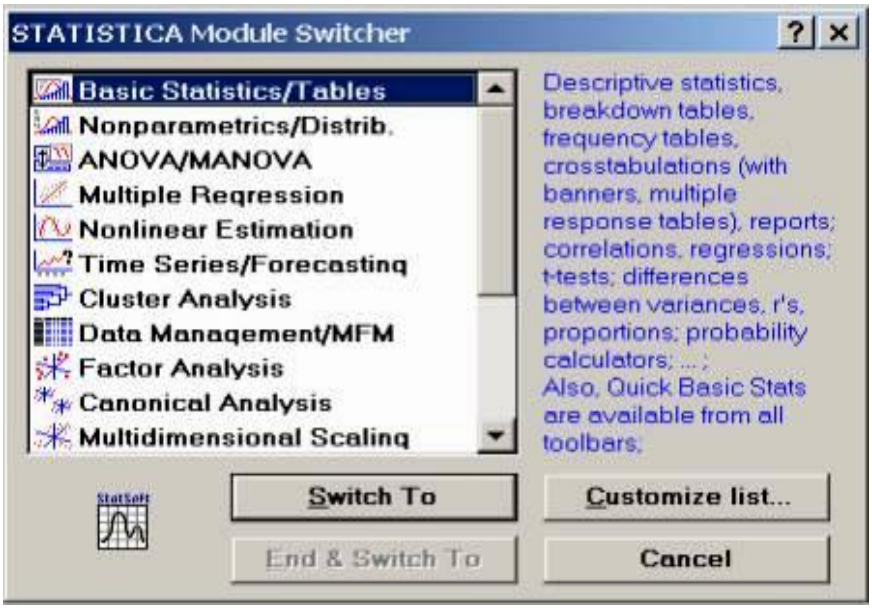


рис. 2.2

или через меню **Статистика \ Многомерные исследовательские методы \ Анализ фактора**.[2]

На экране появится диалоговое окно (рис. 2.3) **Factor Analysis**:

[2] В зависимости от версии программы STATISTICA

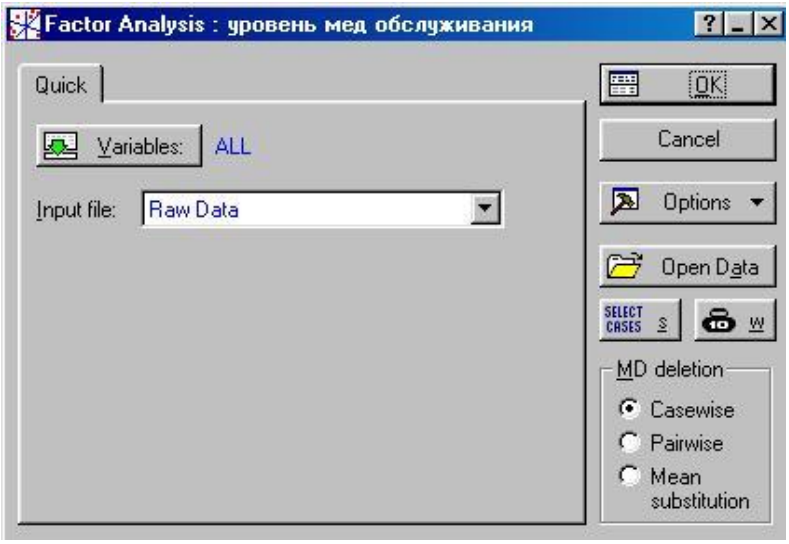


рис. 2.3

Кнопка **Variables** (Переменные) позволяет отобрать все переменные из файла данных, которые должны быть включены в факторный анализ (рис. 2.4). Если при анализе будут использованы все переменные, то можно воспользоваться кнопкой **Select All** (Выбрать все).

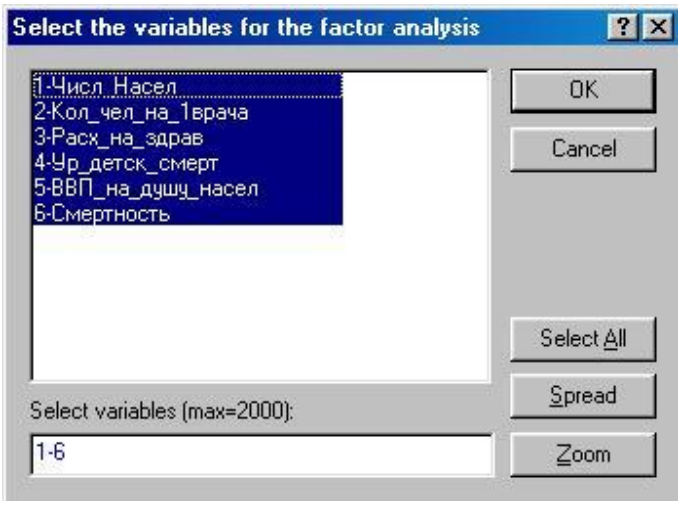


рис. 2.4

Input File (Входной файл) (рис.4.3) должен содержать или необработанные данные или матрицу корреляций, и быть предварительно созданным в модуле **Factor Analysis** (Факторный анализ) или другом статистическом модуле.

В модуле возможны следующие типы исходных данных:

- **Correlation Matrix (**Корреляционная матрица)
- **Raw Data** (Исходные данные)

Выберите, например, **Raw Data**. Это обычный файл данных, где по строкам записаны значения переменных.

MD deletion (Замена пропущенных переменных) (рис. 2.3). Способ обработки пропущенных значений:

- **Casewise** (способ исключения пропущенных случаев) - состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки (случаи), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В таблице остаются только случаи, в которых нет ни одного пропуска.
- **Pairwise** (Парный способ исключения пропущенных значений) - игнорируются пропущенные случаи не для всех переменных, а лишь для выбранной пары. Все случаи, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляционной матрицы, когда последовательно рассматриваются все пары переменных. Очевидно, в способе Pairwise остается больше наблюдений для обработки, чем в способе Casewise. Тонкость, однако, состоит в том, что в способе Pairwise оценки различных коэффициентов корреляции строятся по разному числу наблюдений.
- **Mean Substitution** (подстановка среднего вместо пропущенных значений).

Щелкнув в стартовом окне модуля на кнопку **OK**, вы начнете анализ выбранных переменных.

STATISTICA обработает пропущенные значения тем способом, какой вы ей указали, вычислит корреляционную матрицу и предложит на выбор несколько методов факторного анализа.

Вычисление корреляционной матрицы (если она не задается сразу) – первый этап факторного анализа.

После щелчка по Ok можно перейти к следующему диалоговому окну.

Define Method of Factor Extraction (Определить метод выделения факторов) (рис. 2.5).

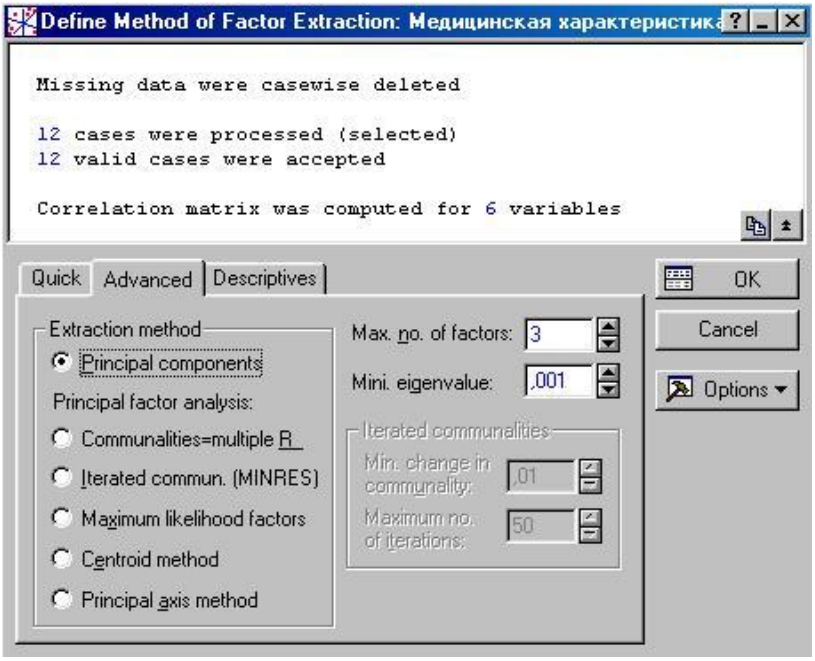


рис. 2.5

Данное окно имеет следующую структуру.

Верхняя часть окна является информационной: здесь сообщается, что пропущенные значения обработаны методом Casewise. Обработано 12 случаев и 12 случаев приняты для дальнейших вычислений. Корреляционная матрица вычислена для 6 переменных.

Группа опций, объединенных под заголовком **Extraction method** (Методы выделения факторов) – позволяет выбрать метод обработки:

- **Principal components** (метод главных компонент) – позволяет выделить компоненты, работая с первоначальной матрицей корреляций.
- **Communalities=multiple R** (общности как множественный R-квадрат) – на диагонали матрицы корреляций будут находиться оценки квадрата коэффициента множественной корреляции- R^2 (соответствующей переменной со всеми другими переменными).
- **Iterated communalities (MINRES)** (метод минимальных остатков) – выполняется в два этапа. Сначала оценки квадрата коэффициента множественной корреляции- R^2 используются для определения общностей, как в предыдущем методе. После первоначального выделения факторов метод корректирует их нагрузки с помощью метода наименьших квадратов с целью минимизировать остаточные суммы квадратов.
- **Maximum likelihood factors** (метод максимального правдоподобия) – в этом методе считается заранее известным число факторов (оно устанавливается в поле ввода максимального числа факторов, см. ниже). STATISTICA оценит нагрузки и общности,

которые максимизируют вероятность наблюдаемой в таком случае матрицы корреляций. В диалоговом окне результатов анализа доступен C-квадрат тест для проверки справедливости принятой гипотезы о числе общих факторов.

- **Centroid method** (центроидный метод) – основан на геометрическом подходе.
- **Principal axis method** (метод главных осей) – основан на итеративной процедуре вычисления общностей по текущим собственным значениям и собственным векторам. Итерации продолжаются до тех пор, пока не превышено максимальное число итераций или минимальное изменение в общностях больше, чем это определено в соответствующем поле (см. ниже).

Max. no. of factors (Максимальное число факторов). Заданное в этом поле число определяет, сколько факторов может быть выделено при работе рассмотренных выше методов. Это поле работает вместе с полем **Min. eigenvalue** (Минимальное собственное значение). Часто при заполнении этого поля руководствуются критерием Кайзера, который рекомендует использовать лишь те факторы, для которых собственные значения не меньше 1.

Остальные поля доступны только при выбранном методе **Centroid method** (Центроидный метод) или **Principal axis method** (Метод главных осей), и определяют необходимые для успешного выполнения последовательных итераций параметры минимального изменения в общностях и максимального числа итераций.

В окне **Define Method of Factor Extraction** (Определить метод выделения факторов) (рис. 2.6) щелкните на кнопке **Review correlations, means, standart deviations** (Просмотреть корреляции /средние/стандартные отклонения).

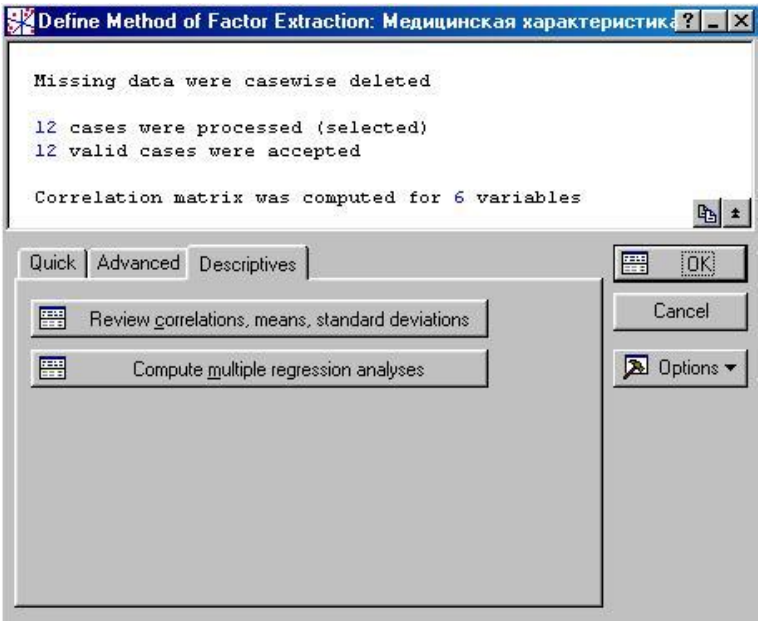


рис. 2.6

Перед вами появилось окно просмотра описательных статистик для анализируемых данных (рис. 2.7), где можно посмотреть средние, стандартные отклонения, корреляции, ковариации, построить различные графики. Здесь можно провести дополнительный анализ текущих данных, проверить соответствие выборочных переменных нормальному закону распределения и существование линейной корреляции между переменными.

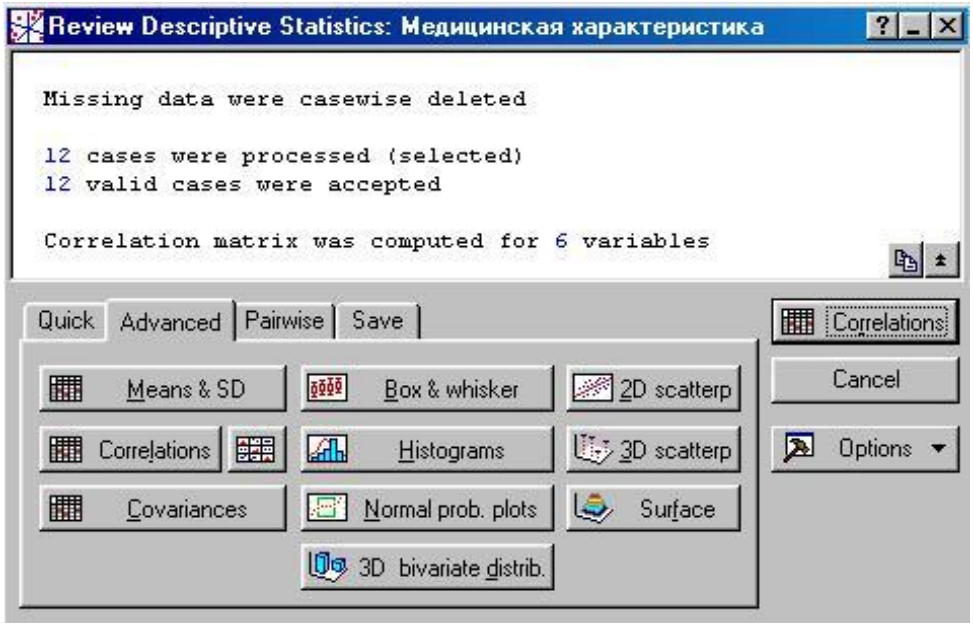


рис. 2.7

Щелкните на кнопке **Correlations** (**Корреляции**) (рис. 2.8). Вы увидите на экране корреляционную матрицу выбранных ранее переменных.



рис. 2.8

Выберите опцию **Principal components** (Главные компоненты) и щелкните по кнопке **OK**.

Система быстро произведет вычисления, и на экране появится окно **Factor Analysis Results** (Результаты факторного анализа).

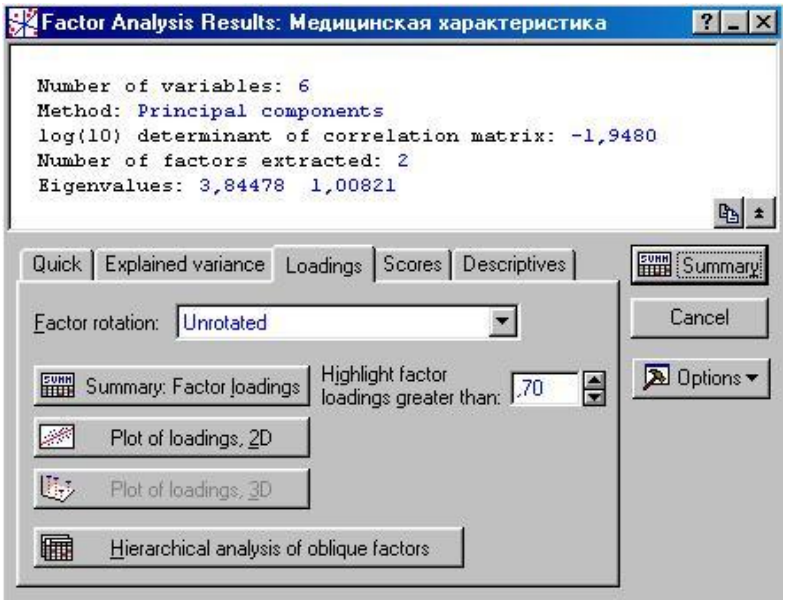


рис. 2.9

В верхней части окна **Результаты факторного анализа** дается информационное сообщение:

Number of variables (число анализируемых переменных) – 6;

Method (метод анализа) – главные компоненты;

log(10) determination of correlation matrix (десятичный логарифм детерминанта корреляционной матрицы) – -1,9480;

Number of Factor extraction (число выделенных факторов) – 3;

Eigenvalues (собственные значения) – 3,84478; 1,00821.

В нижней части окна находятся подразделы, позволяющие всесторонне просмотреть результаты анализа численно и графически.

Plot of loadings, 2D и **Plot of loadings, 3D** (Графики нагрузок) – эти опции построят графики факторных нагрузок в проекции на плоскость любых двух выбранных факторов (рис.2.10) и в проекции в пространство трех выбранных факторов (для чего необходимо наличие как минимум трех выделенных факторов).

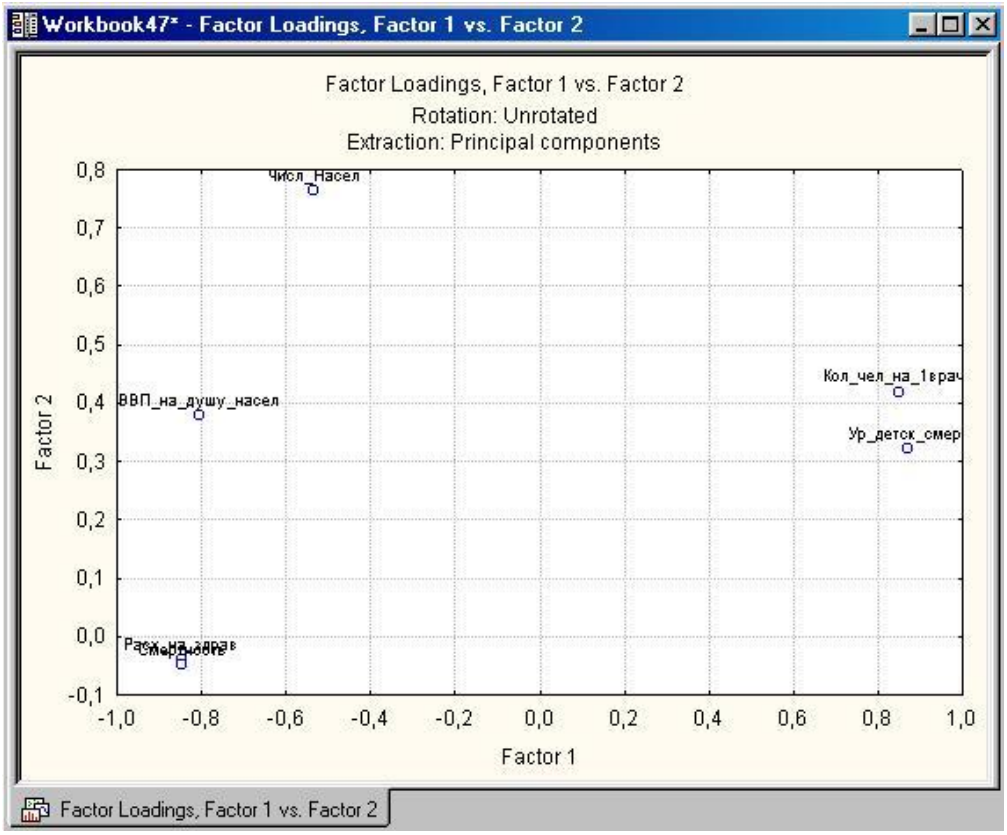


рис. 2.10

Summary. Factor loadings (Факторные нагрузки). Эта опция вызывает таблицу с текущими факторными нагрузками (рис. 2.11), т.е. вычисленными для данного метода вращения факторов, который указан справа от соответствующей кнопки. В этой таблице факторам соответствуют столбцы, а переменным – строки и для каждого фактора указывается нагрузка каждой исходной переменной, показывающая относительную величину проекции переменной на факторную координатную ось. Факторные нагрузки могут интерпретироваться как корреляции между соответствующими переменными и факторами – чем выше нагрузка по модулю, тем больше близость фактора к исходной переменной; т.о., они представляют наиболее важную информацию для интерпретации полученных факторов. В сгенерированной таблице для облегчения трактовки будут выделены факторные нагрузки по абсолютной величине больше 0,7.

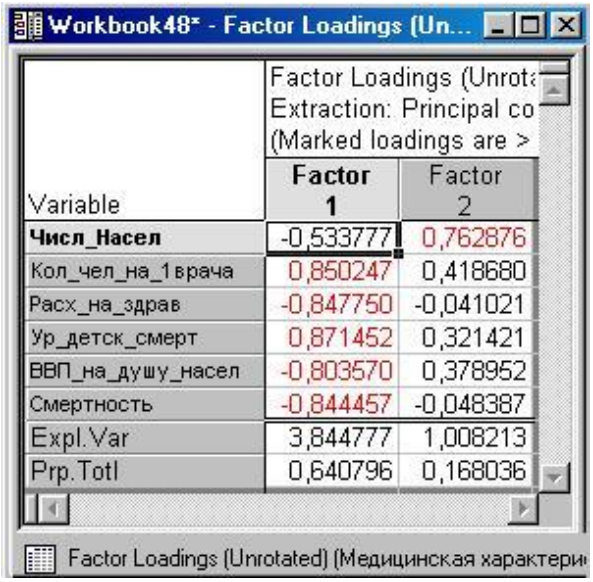


рис. 2.11

По-видимому, первый фактор более коррелирует с переменными, чем второй.

Их трудно проинтерпретировать, возникает вопрос, какой смысл придать второму фактору. В этом случае целесообразно прибегнуть к повороту осей, надеясь получить решение, которое можно интерпретировать в предметной области.

Щелкните по меню **Factor rotation** (Вращение факторов) (рис. 2.12).

Цель вращения – получение простой структуры, при которой большинство наблюдений находится вблизи осей координат. При случайной конфигурации наблюдений невозможно получить простую структуру.

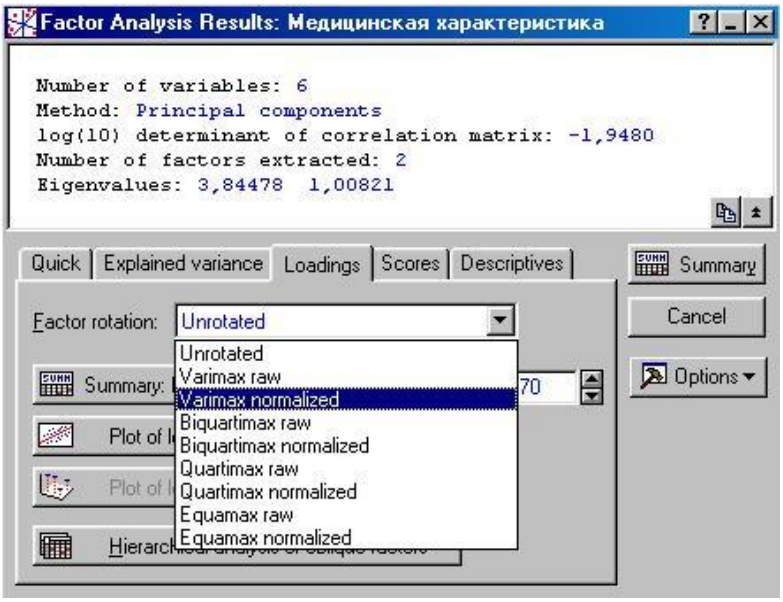


рис. 2.12

В данном раскрывающемся меню вы можете выбрать различные повороты оси. Окно предлагает несколько возможностей оценить и найти нужный поворот следующими методами: **Varimax** – Варимакс;

Biquartimax – Биквартимакс;

Quartimax – Квартимакс;

Equamax – Эквимакс.

Дополнительный термин в названии методов – **normalized** (нормализованные) – указывает на то, что факторные нагрузки в процедуре нормализуются, т.е. делятся на корень квадратный из соответствующей дисперсии. Термин **raw** (исходные) показывает, что вращаемые нагрузки не нормализованы.

Инициируйте кнопку **Varimax normalized** (Варимакс нормализованный).

Система произведет вращение факторов методом нормализованного варимакса, и окно **Factor Analysis Results** (Результаты факторного анализа) снова появится на мониторе. Вновь инициируйте в этом окне кнопку **Plot of Loadings 2D** (Двумерный график нагрузок). Вы опять увидите график нагрузок (рис. 2.13).

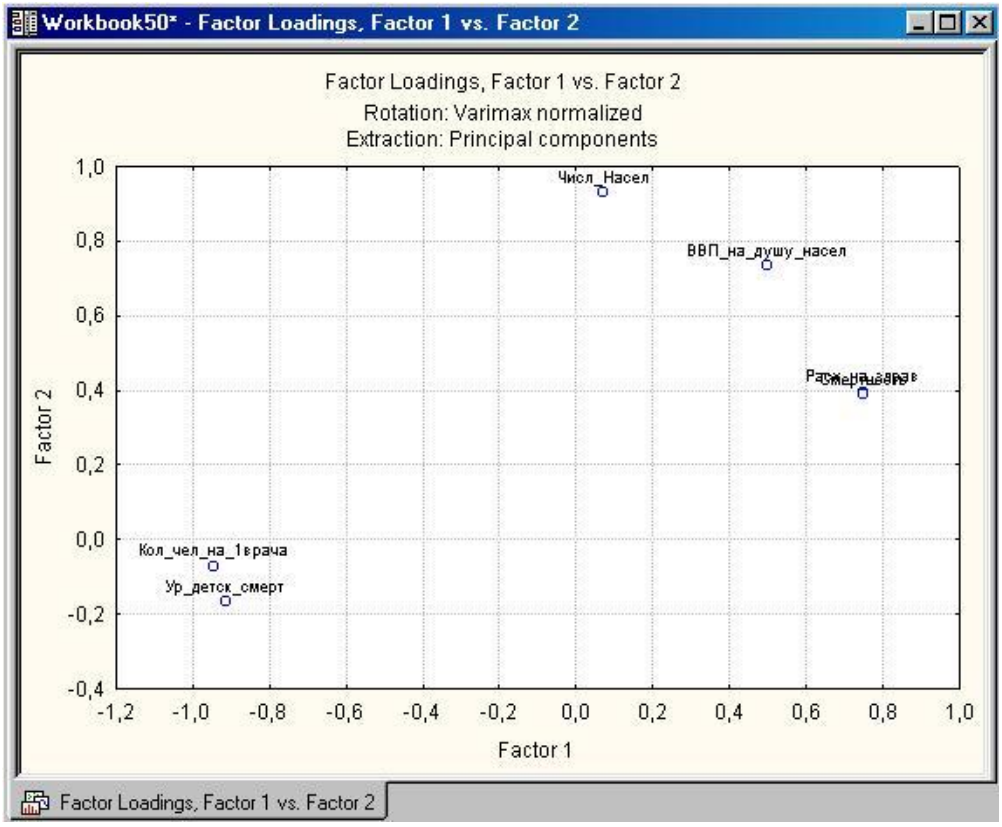


рис. 2.13

Конечно, этот график немного отличается от предыдущего. Посмотрим еще нагрузки численно, инициировав кнопку **Факторные нагрузки** (Factor loadings). Щелкните на кнопке **Summary. Factor loadings**, вы откроете окно:



рис. 2.14

Теперь найденное решение уже можно интерпретировать. Факторы чаще интерпретируют по нагрузкам. Первый фактор теснее всего связан с X2 , X3, X4, X6. Второй фактор – X1 и X5. Таким образом, произвели классификацию переменных на две группы. Возникает вопрос: сколькими же факторами следует ограничиваться на практике? Для этого в программном пакете STATISTICA существует критерий **Scree plot** (Критерий каменистой осыпи). В окне **Factor Analysis Results** нажмите кнопку **Scree plot** получите следующий график собственных значений.

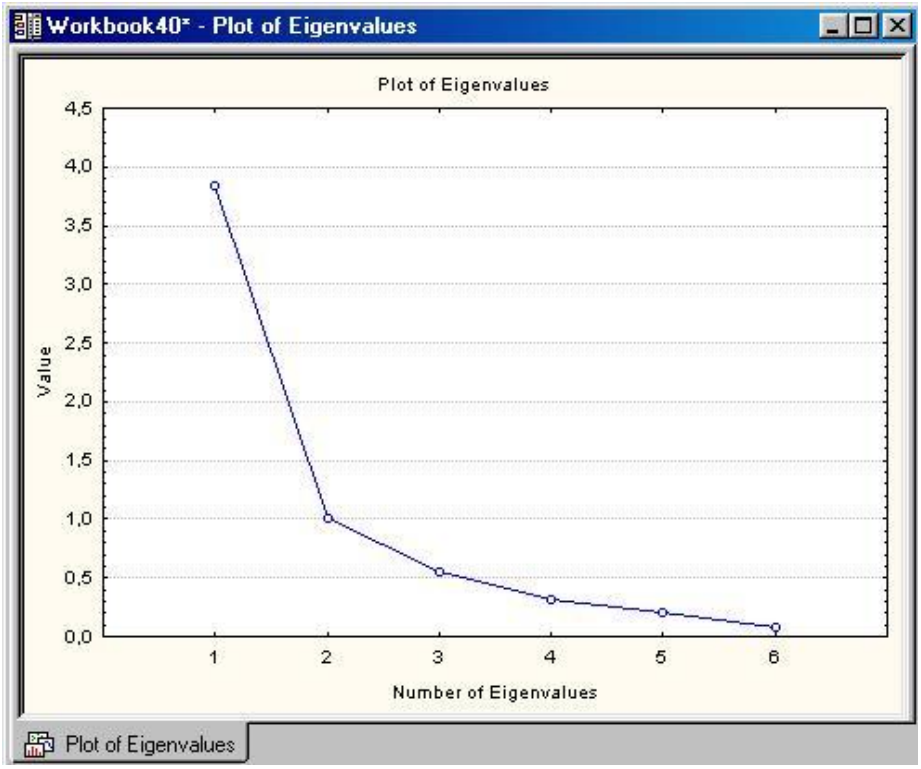


рис. 2.15

В точках с координатами 1, 2 осыпание замедляется наиболее существенно, следовательно, теоретически можно ограничиваться двумя факторами.

2.3. Задания для самостоятельной работы

Задание 1

По исходным данным Приложения 3 за исключением переменных таблицы 6.1 согласно номеров вариантов провести факторный анализ. Не совершая поворот факторов.

Получить матрицу парных коэффициентов корреляции.

Получить первые три общие фактора и дать им экономическую интерпретацию по матрице факторных нагрузок.

Таблица 2.1

Номер варианта	Переменные	Номер варианта	Переменные	Номер варианта	Переменные
1	*Y*1,*Y*2,*X*4,*X*5	6	*Y*1,*Y*2,*X*14,*X*15	11	*Y*2,*Y*3,*X*10,*X*11
2	*Y*1,*Y*2,*X*6,*X*7	7	*Y*1,*Y*2,*X*16,*X*17	12	*Y*2,*Y*3,*X*12,*X*13
3	*Y*1,*Y*2,*X*8,*X*9	8	*Y*2,*Y*3,*X*4,*X*5	13	*Y*2,*Y*3,*X*14,*X*15
4	*Y*1,*Y*2,*X*10,*X*11	9	*Y*2,*Y*3,*X*6,*X*7	14	*Y*2,*Y*3,*X*16,*X*17
5	*Y*1,*Y*2,*X*12,*X*13	10	*Y*2,*Y*3,*X*8,*X*9	15	*Y*1,*Y*3,*X*4,*X*5

Y1 – производительность труда;

Y2 – индекс снижения себестоимости продукции;

Y3 – рентабельность;

X4 – трудоемкость единицы продукции;

X5 – удельный вес рабочих в составе ППП;

X6 – удельный вес покупных изделий;

X7 - коэффициент сменности оборудования;

X8 – премии и вознаграждения на одного работника;

X9 – удельный вес потерь от брака;

X10 – фондоотдача;

X11 – среднегодовая численность ППП;

X12 – среднегодовая стоимость ОПФ;

X12 – среднегодовой фонд заработной платы;

X14 – фондовооруженность труда;

X15 – оборачиваемость нормированных оборотных средств;

X16 – оборачиваемость ненормированных оборотных средств;

X17 – непроизводственные расходы.

Задание 2

На основе решения полученного в Задании 1 произвести поворот факторов методом Варимакс нормализованный.

Получить первые три общие фактора и дать им экономическую интерпретацию по матрице факторных нагрузок.

Задание 3

На основе решения полученного в Задании 1 произвести поворот факторов методом Варимакс исходный.

Получить первые три общие фактора и дать им экономическую интерпретацию по матрице факторных нагрузок.

ГЛАВА 3. ДИСКРИМИНАНТНЫЙ АНАЛИЗ И ЕГО РЕАЛИЗАЦИЯ В ПАКЕТЕ STATISTICA

3.1. Введение в дискриминантный анализ

3.1.1 Проблема классификации

Дискриминантный анализ является разделом многомерного статистического анализа, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно. Дискриминантный анализ – это общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Эти процедуры можно разделить на методы *интерпретации межгрупповых различий – дискриминации* и методы *классификации наблюдений* по группам. При интерпретации нужно ответить на вопрос: возможно ли, используя данный набор переменных, отличить одну группу от другой, насколько хорошо эти переменные помогают провести дискриминацию и какие из них наиболее информативны?

Методы классификации связаны с получением одной или нескольких функций, обеспечивающих возможность отнесения данного объекта к одной из групп. Эти функции называются классифицирующими и зависят от значений переменных таким образом, что появляется возможность отнести каждый объект к одной из групп.

Задачи дискриминантного анализа можно разделить на три типа. Задачи *первого типа* часто встречаются в медицинской практике. Допустим, что мы располагаем информацией о некотором числе индивидуумов, болезнь каждого из которых относится к одному из двух или более диагнозов. На основе этой информации нужно найти функцию, позволяющую поставить в соответствие новым индивидуумам характерные для них диагнозы. Построение такой функции и составляет задачу дискриминации.

Второй тип задач относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить. Примером может служить определение пола давно умершего человека по его останкам, найденным при археологических раскопках.

Задачи *третьего типа* связаны с предсказанием будущих событий на основании имеющихся данных. Такие задачи возникают при прогнозе отдаленных результатов лечения, например, прогноз выживаемости оперированных больных.

Целью задачи дискриминации является изучение основных процедур дискриминантного анализа: дискриминации и классификации, построение и определение количества

дискриминантных функций и их разделительной способности, нахождение классифицирующих функций.

Проблема классификации возникает, когда исследователь делает некоторое число измерений, связанных с каким-то индивидуумом, и на основе этих измерений хочет отнести его к одной из нескольких категорий. Он не может непосредственно определить категорию, к которой относится индивидуум, и вынужден использовать эти измерения. Во многих случаях можно предположить, что имеется конечное число категорий или генеральных совокупностей, из которых мог быть взят рассматриваемый индивидуум, причем каждая из этих категорий характеризуется определенным законом распределения вероятностей для измерений. Таким образом, индивидуум рассматривается как случайное наблюдение над этой генеральной совокупностью. Вопрос ставится так: как по результатам измерений определить, из какой генеральной совокупности взят данный индивидуум.

Проблему классификации можно рассматривать как проблему *статистических решающих функций*. Имеется несколько гипотез, каждой из которых соответствует свое распределение вероятностей для наблюдений. Мы должны принять одну из этих гипотез и отвергнуть остальные. Если множество допустимых генеральных совокупностей состоит лишь из двух генеральных совокупностей, то рассматриваемая задача является элементарной задачей проверки одной гипотезы, соответствующей определенному распределению вероятностей, при одной конкурирующей гипотезе.

В одних случаях категории определены заранее тем, что полностью известны распределения вероятностей измерений. В других вид каждого распределения известен, но неизвестны параметры этих распределений, которые и должны быть оценены по выборке из генеральной совокупности.

Ставя задачу отнесения каждого из классифицируемых наблюдений $X_i = (x_{i1}, x_{i2}, \dots x_{ip})^{\mathrm{T}}$, $i = 1, 2, \dots, n$, к одному из классов, необходимо четко определить понятие *класса*. Мы будем понимать под *«классом» генеральную совокупность, описываемую унимодальной плотностью* $f(x)$ *(или унимодальным полигоном вероятностей в случае дискретных признаков* X*)*. Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют, классифицируют) две или более возникающие совокупности (классы).

При построении процедуры классификации желательно сделать минимальной вероятность неправильной классификации , точнее – добиться того, чтобы в среднем неправильные выводы делались как можно реже.

Введем величину $c(j|i)$ потерь, которые мы несем при отнесении *одного* объекта i- го класса к классу j (при $i = j$, очевидно, $c(i|i) = 0$). Следовательно, если в процессе классификации мы поступили таким образом $m(j|i)$ раз, то потери, связанные с отнесением объектов i- го класса к классу j составят $m(j|i)\,c(j|i)$. Для того чтобы подсчитать общие потери C_n при такой процедуре классификации, надо просуммировать величину произведения $m(j|i)\,c(j|i)$ по всем $i = 1, 2, \ldots, k$ и $j = 1, 2, \ldots, k$, т.е.

$$C_n = \sum_{i=1}^{k}\sum_{j=1}^{k} m(j|i)c(j|i).$$

Величина

$$C^{(i)} = \sum_{i=1}^{k} c(j|i)P(i|j)$$

определяет средние потери от неправильной классификации объектов i-го класса, так что *средние удельные потери* от неправильной классификации всех анализируемых объектов будут равны

$$C = \sum_{i=1}^{k} p_i C^{(i)}.$$

В достаточно широком классе ситуаций полагают, что потери $c(j|i)$ одинаковы для любой пары i и j.

В этом случае *стремление к минимизации средних удельных потерь* C *будет эквивалентно стремлению минимизации вероятности правильной классификации объектов, равной* $\sum_{i=1}^{k} p_i P(i|i)$. Поэтому часто при построении процедур классификации говорят не о потерях, а о *вероятностях неправильной классификации* $1 - \sum_{i=1}^{k} p_i P(i|i)$.

Решение задачи байесовских процедур классификации

Сформулируем постановку задачи построения оптимальной процедуры классификации *p*-мерных наблюдений X_1, K, X_n при наличии обучающих выборок. Классифицируемые наблюдения (5.3) интерпретируются в данной задаче как выборка из генеральной совокупности, описываемой смесью k классов (унимодальных генеральных совокупностей) с плотностью вероятности

$$f(x) = \sum_{j=1}^{k} p_i f_i(x),$$

где p_j - априорная вероятность появления в этой выборке элемента

из класса (генеральной совокупности) j с плотностью $f_j(x)$ или, другими словами, p_j - это удельный вес элемента j-го класса в обще генеральной совокупности (5.4).

Введем понятие *процедуры классификации* (*решающего правила, дискриминантной функции*) $d(X)$. Функция $d(X)$ может принимать только целые положительные значения 1, 2, … , k, причем те X, при которых она принимает значение, равное j, мы будем относить к классу j, т.е.

$$S_j = \{X : d(X) = j\}, \quad j = 1, 2, \ldots, k.$$

Очевидно, S_j - это *p*-мерные области в пространстве $\text{P}(X)$ возможных значений анализируемого многомерного признака X, причем функция $d(X)$ строится таким образом, чтобы их объединение $S_1 \text{U} S_2 \text{U} \ldots \text{U} S_k$ заполняло все пространство $\text{P}(X)$ и чтобы они попарно не пересекались. Таким образом, решающее правило $d(X)$ может быть задано разбиением

$$S = (S_1, S_2, \text{K}, S_k)$$

всего пространства $\text{P}(X)$ на k непересекающихся областей.

Процедура классификации (дискриминантная функция) $d(X)$ *(или* S*) называется оптимальной (байесовской), если она сопровождается минимальными потерями среди всех других процедур классификации.*

Другими словами, наблюдение $X_n (n = 1, 2, \ldots, n)$ будет отнесено к классу j тогда, когда средние удельные потери от его отнесения именно в этот класс окажутся минимальными по сравнению с аналогичными потерями, связанными с отнесением этого наблюдения в любой другой класс.

Введем необходимые обозначения. Если цена ошибочной классификации наблюдения, произведенного над i-й совокупность, как наблюдения над j-й, равна $c(j|i)$, то вероятность ошибочной классификации равна

$$P(j|i,S)=\int_{S_j} f_i(x)dx.$$

ТЕОРЕМА: *Пусть* p_1 *и* p_2 - *априорные вероятности того, что наблюдение производится над первой генеральной совокупностью с плотностью распределения вероятностей* $f_1(x)$ *и над второй генеральной совокупностью с плотностью* $f_2(x)$ *соответственно. Пусть, далее, цена ошибочной классификации наблюдения, производимого над первой совокупностью, равна* $c(2|1)$, *а цена ошибочной классификации наблюдения, производимого над второй совокупностью, равна* $c(2|1)$. *Тогда области* S_1 *и* S_2, *определяемые из условия:*

$$S_1: \frac{f_1(x)}{f_2(x)} \text{ i } \frac{c(1|2)p_2}{c(2|1)p_1},$$

$$S: \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)p_2}{c(2|1)p_1},$$

дают минимум математического ожидания потерь.

3.1.2 Определение числа и вида дискриминирующих функций. Классификация объектов с помощью функции расстояния

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть *дискриминантными переменными*), которая бы оптимально разделила рассматриваемые группы. Линейная функция

$$d_{km} = b_o + b_1 x_{1km} + ... + b_p x_{pkm}, m = 1, ..., n, k = 1, ..., g,$$

называется *канонической дискриминантной функцией* с неизвестными коэффициентами b_i. Здесь d_{km} – значение дискриминантной функции для *m*-го объекта в группе *k*; x_{ikm} – значение дискриминантной переменной для *m*-го объекта в группе *k*. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в *p*-мерном пространстве. В частном случае при *p* = 2 она является прямой, а при *p* = 3 – плоскостью.

Коэффициенты b_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция d имеет нулевую внутригрупповую корреляцию с $d_1, \ldots, d_{g-1}$. Если число групп равно g, то число канонических дискриминантных функций будет на единицу меньше числа групп. Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображение объектов будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Для получения коэффициентов b_i канонической дискриминантной функции нужен статистический критерий различения групп. Очевидно, что классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Разумеется, что большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$l = B(d)/W(d) \qquad (*)$$

где B – межгрупповая и W – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних. Иногда в (*) вместо W используют матрицу рассеяния T объединенных данных.

Рассмотрим максимизацию отношения (*) для произвольного числа классов. Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в k-й группе;

$n-$ общее число наблюдений по всем группам;

$x_{ikm}-$ величина переменной *i* для *m*-го наблюдения в *k*-й группе;

$\overline{x}_{ik}-$ средняя величина переменной *i* в *k*-й группе;

$\overline{x}_i-$ среднее значение переменной *i* по всем группам;

$T(u,v)-$ общая сумма перекрестных произведений для переменных *u* и *v*;

$W(u,v)-$ внутригрупповая сумма перекрестных произведений для переменных *u* и *v*;

$$t_{ij}=T(x_i,x_j),\ w_{ij}=W(x_i,x_j).$$

В модели дискриминации должны соблюдаться следующие условия:

1. число групп: $g\geq 2$;

2. число объектов в каждой группе: $n_i\geq 2$;

3. число дискриминантных переменных: $0<p<(n-2)$;

4. дискриминантные переменные измеряются в интервальной шкале;

5. дискриминантные переменные линейно независимы;

6. ковариационные матрицы групп примерно равны;

7. дискриминантные переменные в каждой группе подчиняются многомерному нормальному закону распределения.

Рассмотрим задачу максимизации отношения (*) когда имеются g групп. Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычислим матрицу рассеяния T, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних $\overline{x}_i, i=1,\dots,p$ по каждой переменной. Элементы матрицы T определяются выражением

$$t_{ij}=\sum_{k=1}^{g}\sum_{m=1}^{n_k}(x_{ikm}-\overline{x}_i)(x_{jkm}-\overline{x}_j),$$

где

$$\overline{x}_i=\frac{1}{n}\sum_{k=1}^{g}n_i\overline{x}_{ik},\ i=1,\dots,p;\ \ \overline{x}_{ik}=\frac{1}{n_i}\sum_{m=1}^{n_i}x_{ikm},\ i=1,\dots,p;\ k=1,\dots,g.$$

Запишем это выражение в матричной форме. Обозначим *p*-мерную случайную векторную переменную *k*-й группы следующим образом

$$X_k=\{x_{ikm}\},\ i=1,\dots,p,\ \ k=1,\dots,g,\ \ m=1,\dots,n_k.$$

Тогда объединенная p-мерная случайная векторная переменная всех групп будет иметь вид

$$X = \left[X_1 X_2 \ldots X_g \right].$$

Общее среднее этой p-мерной случайной векторной переменной будет равен вектору средних отдельных признаков

$$\overline{x} = \left[\overline{x}_1 \overline{x}_2 \ldots \overline{x}_g \right].$$

Матрица рассеяния от среднего при этом запишется в виде

$$T = \sum_{k=1}^{g} (X_k - \overline{x})(\overline{X}_k - \overline{x})^{\mathsf{T}}$$

Если использовать векторную переменную объединенных переменных X, то матрица T определится по формуле

$$T = (X - \overline{x})(X - \overline{x})^{\mathsf{T}}$$

Матрица T содержит полную информацию о распределении точек по пространству переменных. Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Если разделить матрицу T на $(n-1)$, то получим ковариационную матрицу. Для проверки условия линейной независимости переменных полезно рассмотреть вместо T нормированную корреляционную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу W, которая отличается от T только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы внутригруппового рассеяния определятся выражением

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \overline{x}_{ik})(x_{jkm} - \overline{x}_{jk}),$$

Запишем это выражение в матричной форме. Данным g групп будут соответствовать векторы средних

$$\overline{x}_1 = \left[\overline{x}_{11} \overline{x}_{21} \ldots \overline{x}_{p1} \right]$$

. . . .

$$\overline{x}_g = \left[\overline{x}_{1g} \overline{x}_{2g} \ldots \overline{x}_{pg} \right].$$

Тогда матрица внутригрупповых вариаций запишется в виде

$$W = \sum_{k=1}^{g} (X_k - \overline{x}_k)(\overline{X}_k - \overline{x}_k)^{\mathsf{T}}$$

Если разделить каждый элемент матрицы W на $(n-g)$, то получим оценку ковариационной матрицы внутригрупповых данных.

Когда центроиды различных групп совпадают, то элементы матриц T и W будут равны. Если же центроиды групп различные, то разница $B = T - W$ будет определять межгрупповую сумму квадратов отклонений и попарных произведений. Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса. Отметим, что элементы матрицы B можно вычислить и по данным средних

$$b_{ij} = \sum_{k=1}^{g} n_k (\overline{x}_{ik} - \overline{x}_i)(\overline{x}_{jk} - \overline{x}_j),\ i,\ j = 1, \ldots,\ p.$$

Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами. Для лучшего разделения наблюдений на группы нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей. Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах . Это утверждение можно сформулировать так: если спроектировать g групп p-мерных выборок на $(g-1)$ пространство, порожденное собственными векторами $(n_{1k}, \ldots, n_{pk}), k = 1, \ldots, g-1$, то отношение (5.7) будет максимальным, т. е. рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании. Если бы мы захотели спроектировать g выборок на прямую при условии максимизации наибольшего рассеивания между группами, то следовало бы использовать собственный вектор $(n_{11}, \ldots, v_{p1})$), соответствующий максимальному собственному числу l_1. При этом дискриминантные функции можно получать: по *нестандартизованным* и *стандартизованным* коэффициентам.

Нестандартизованные коэффициенты. Пусть $l_1 \geq l_2 \geq \ldots \geq l_p$ и $n_1, \ldots, n_p$ соответственно собственные значения и векторы. Тогда условие (5.7) в терминах

собственных чисел и векторов запишется в виде $l = \frac{\sum_k b_{jk} n_j n_k}{\sum_k w_{jk} n_j n_k}$, что влечет равенство $\sum_k (b_{jk} - l\ w_{jk}) n_k = 0$, или в матричной записи

$$(B - l_i W) n_i = 0, n_i^{\mathrm{T}} W n_j = d_{ij}, \quad (**)$$

где символ Кронекера. Таким образом, решение уравнения позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям. Если B и W невырожденные матрицы, то собственные корни уравнения такие же, как и у . Решение системы уравнений (**) можно получить путем использования разложения Холецкого LL^{T} матрицы W^{-1} и решения задачи о собственных значениях

$$(L^{\mathrm{T}} BL - l_i I) n_i = 0,\ n_i^{\mathrm{T}} n_j = d_{ij}.$$

Каждое решение, которое имеет свое собственное значение l_i и собственный вектор n_i, соответствует одной дискриминантной функции. Компоненты собственного вектора n_i можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того, чтобы начало координат совпало с главным центроидом нужно нормировать компоненты собственного вектора

$$b_i = n_i \sqrt{n - g},\ b_o = -\sum_{i=1}^{p} b_i \bar{x}_i$$

Нормированные коэффициенты получены по нестандартизованным исходным данным, поэтому они называются нестандартизованными. Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты можно получить двумя способами:

1) β_i, если исходные данные были приведены к стандартной форме;

2) преобразованием нестандартизованных коэффициентов к стандартизованной форме:

$$c_i = b_i \sqrt{\frac{w_{ij}}{n - g}},$$

где w_{ii} – сумма внутригрупповых квадратов i-й переменной. Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных. Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Структурные коэффициенты определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты полезны при классификации групп. Структурный коэффициент можно вычислить и для переменной в пределах отдельно взятой группы. Тогда получаем *внутригрупповой структурный коэффициент*, который вычисляется по формуле

$$s_{ij} = \sum_{k=1}^{p} r_{ik} c_{kj} = \sum_{k=1}^{p} \frac{w_{ik} c_{kj}}{\sqrt{w_{ii} w_{jj}}},$$

где s_{ij} – внутригрупповой структурный коэффициент для i-й переменной и j-й функции; r_{ik} – внутригрупповые структурные коэффициенты корреляции между переменными i и k; c_{kj} – стандартизованные коэффициенты канонической функции для переменной k и функции j.

Структурные коэффициенты по своей информативности несколько отличаются от стандартизованных коэффициентов. Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных. Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.

*Число дискриминантных функций***.** Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на 1 меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной

способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу l_1, вторая обеспечивает максимальное различение после первой и т. д. Различительную способность *i*-й функции оценивают по относительной величине в процентах собственного числа l_i от суммы всех l.

Коэффициент канонической корреляции. Другой характеристикой, позволяющей оценить полезность дискриминантной функции является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для i-й дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{l_i}{1+l_i}}.$$

Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

Остаточная дискриминация. Так как дискриминантные функции находятся по выборочным данным, они нуждаются в проверке статистической значимости. Дискриминантные функции представляются аналогично главным компонентам. Поэтому для проверки этой значимости можно воспользоваться критерием, аналогичным дисперсионному критерию в методе главных компонент. Этот критерий оценивает остаточную дискриминантную способность, под которой понимается способность различать группы, если при этом исключить информацию, полученную с помощью ранее вычисленных функций. Если остаточная дискриминация мала, то не имеет смысла дальнейшее вычисление очередной дискриминантной функции. Полученная статистика носит название "$\Lambda-$*статистика Уилкса*" и вычисляется по формуле:

$$\Lambda_k = \prod_{i=k+1}^{g}(1/(1+l_i)),$$

где k – число вычисленных функций. Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция. Величина

$$c^2 = -(n-((p+g)/2)-1)\ln\Lambda_k,\ k=0,1,\ldots,g-1$$

имеет хи-квадрат распределение с $(p-k)(g-k-1)$ степенями свободы.

Вычисления проводим в следующем порядке.

1. Находим значение критерия c^2 при k = 0. Значимость критерия подтверждает существование различий между группами. Кроме того, это доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.

2. Определяем первую дискриминантную функцию, и проверяем значимость критерия при k = 1. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая информация.

*Классификация объектов с помощью функции расстояния***.** Выбор функций расстояния между объектами для классификации является наиболее очевидным способом введения меры сходства для векторов объектов, которые интерпретируются как точки в евклидовом пространстве. В качестве меры сходства можно использовать евклидово расстояние между объектами. Чем меньше расстояние между объектами, тем больше сходство. Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонения, трудно четко определить понятие "расстояния". В этом случае полезнее применить не евклидовое расстояние, а *выборочное расстояние Махаланобиса*

$$D^2(x/G_k)=(n-g)\sum_{v=1}^{p}\sum_{j=1}^{p}\left(w^{-1}\right)_{vj}\left(x_{iv}-\bar{x}_{vk}\right)\left(x_{ij}-\bar{x}_{jk}\right), k=1,\ldots,g$$

или в матричной записи

$$D^2(x/G_k)=(n-g)(x-\bar{x}_k)^{\mathrm{T}}W^{-1}(x-\bar{x}_k),\ k=1,\ldots,g\ ,$$

где x представляет объект с p переменными, $\bar{x}_k-$ вектор средних для переменных k-й группы объектов. Если вместо W использовать оценку внутригрупповой ковариационной матрицы $W^{-1}=W/(n-g)$, то получим стандартную запись выборочного расстояния Маханалобиса

$$D^2\left(x/G_k\right)=\left(x-\bar{x}_k\right)^T W^{-1}\left(x-\bar{x}_k\right), k=1,\mathbf{K},g.$$

При использовании функции расстояния, объект относят к той группе, для которой расстояние D^2 наименьшее. Относя объект к ближайшему классу в соответствии с D^2, мы неявно приписываем его к тому классу, для которого он имеет наибольшую вероятность принадлежности $\mathbf{P}(x/G_k)$. Если предположить, что любой объект должен принадлежать одной из групп, то можно вычислить вероятность его принадлежности для любой из групп

$$\mathbf{P}(G_k / x) = \frac{\mathbf{P}(x / G_k)}{\sum_{i=1}^{g} \mathbf{P}(x / G_i)}$$

Объект принадлежит к той группе, для которой апостериорная вероятность $\mathbf{P}(G_k / x)$ максимальна, что эквивалентно использованию наименьшего расстояния.

До сих пор при классификации по D^2 предполагалось, что априорные вероятности появления групп одинаковы. Для учета априорных вероятностей нужно модифицировать расстояние D^2, вычитая из выражения Маханалобиса удвоенную величину натурального логарифма от априорной вероятности p_k. Тогда, вместо выборочного расстояния Махаланобиса, получим

$$D_p^2(x / G_k) = (x - \bar{x}_k)^T W^{-1}(x - \bar{x}_k) - 2\ln(p_k).$$

Отметим, тот факт, что априорные вероятности оказывают наибольшее влияние при перекрытии групп и, следовательно, многие объекты с большой вероятностью могут принадлежать ко многим группам. Если группы сильно различаются, то учет априорных вероятностей практически не влияет на результат классификации, поскольку между классами будет находиться очень мало объектов.

V-статистика Рао. В некоторых работах для классификации используется обобщенное расстояние Махаланобиса *V* – обобщение величины D^2. Эта мера, известная как *V*-статистика Рао, измеряет расстояния от каждого центроида группы до главного центроида с весами, пропорциональными объему выборки соответствующей группы. Она применима при любом количестве классов и может быть использована для проверки гипотезы $H_o : m_1 = ... = m_g$. Если гипотеза H_o верна, а объемы выборок n_i стремятся к ∞, то распределение величины *V* стремится к c^2 с $p(g-1)$ степенями свободы. Если наблюдаемая величина $c^2 > c_{1-a}^2(p(g-1))$, то гипотеза H_o отвергается. *V*-статистика вычисляется по формуле

$$V = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p}(w^{-1})_{ij}\sum_{k=1}^{g} n_k(\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j).$$

Матричное выражение оценки *V* имеет вид

$$V = (n-g)\sum_{k=1}^{g} n_k(\bar{x}_k - \bar{x})W^{-1}(\bar{x}_k - \bar{x}).$$

Отметим, что при включении или исключении переменных V-статистика имеет распределение хи-квадрат с числом степеней свободы, равным $(g-1)$, умноженное на число переменных, включенных (исключенных) на этом шаге. Если изменение статистики не значимо, то переменную можно не включать. Если после включения новой переменной V-статистика оказывается отрицательной, то это означает, что включенная переменная ухудшает разделение центроидов.

3.2. Пример решения практической задачи дискриминантным анализом в системе STATISTICA

Рассмотрим основные этапы проведения дискриминантного анализа в системе STATISTICA на следующем примере.

Исходными показателями послужили:

X1 – количество человек, приходящихся на одного врача

X2 - Смертность на 1000 человек

X3 – ВВП, рассчитанный по паритету покупательной способности на душу населения (млн. $)

X4 - Расходы на здравоохранение на душу населения ($)

В файле (рис. 3.1) содержатся данные по 10 странам, которые были выбраны и отнесены к соответствующим группам экспертным способом (по уровню медицинского обслуживания):

- страны с высокими уровнем медицинского обслуживания
- страны со средним (удовлетворительным) уровнем медицинского обслуживания
- страны с низким уровнем медицинского обслуживания

Задача состоит в том, чтобы на основе аналогичных показателей классифицировать страны: Молдавия, Украина.

Data: уровень мед обслуживания123a (5v by 10c)

	1 Кол_чел_на_1врача	2 Расх_на_здрав	3 ВВП_на_душу_насел	4 Смертность	5 Class
Азербайджан	256	99	3000	9,6	низкий
Армения	198	152	3000	9,7	низкий
Белоруссия	222	157	7500	14	высокий
Грузия	182	152	4600	14,6	удовлетв
Казахстан	265	154	5000	10,6	удовлетв
Киргизия	301	118	2700	9,1	низкий
Россия	235	159	7700	13,9	высокий
Таджикистан	439	100	1140	8,6	низкий
Туркмения	320	125	4300	9	удовлетв
Узбекистан	299	116	2400	8	низкий

рис. 3.1

Алгоритм выполнения

Для вызова данного модуля можно использовать **Module Switcher** (Переключатель модулей) (рис. 3.2), который содержит список всех доступных модулей. Высветите название модуля **Discriminant Analysis** (**Дискриминантный Анализ**) и нажмите кнопку **Switch to** (Переключиться в),

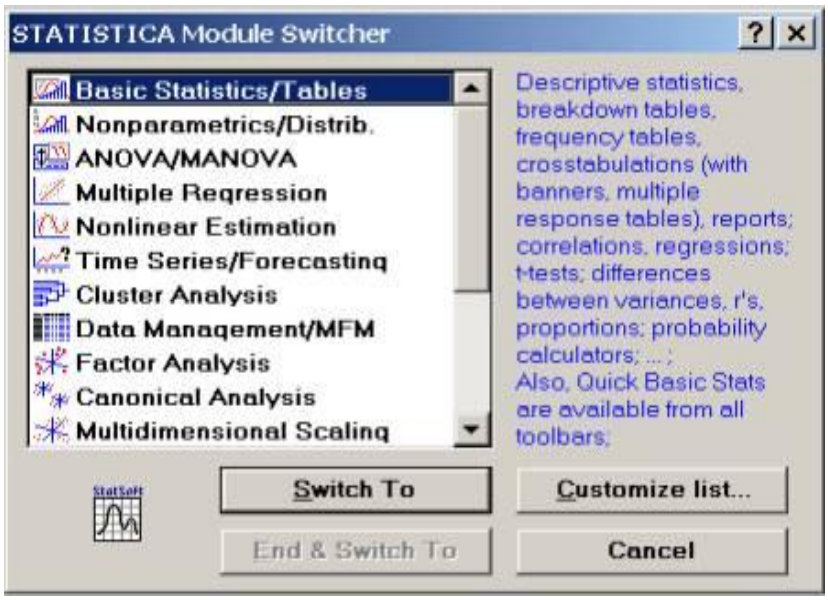


рис. 3.2

или через меню **Статистика \ Многомерные исследовательские методы \ Дискриминантный анализ.** [3]

На экране появится стартовая панель модуля **Discriminant Function Analysis** (Анализ дискриминантных функций) (рис. 3.3),

[3] В зависимости от версии программы STATISTICA

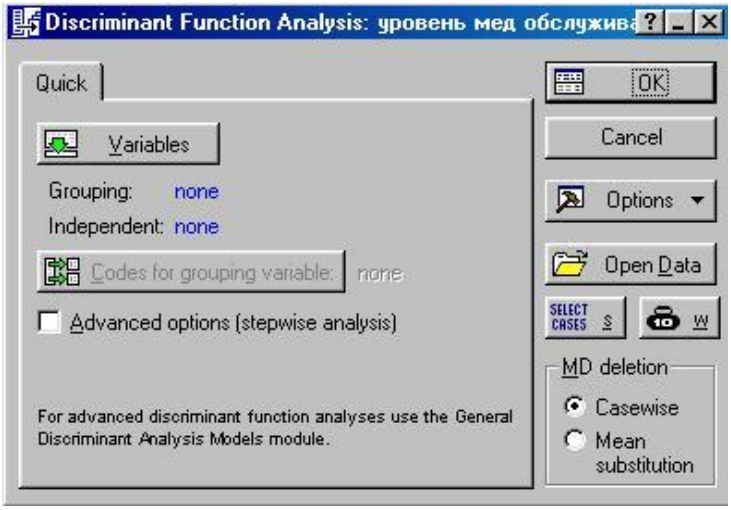


рис. 3.3

в котором:

- **Variables** позволяет выбрать **Grouping** (Группируемую переменную) и **Independent** (Независимые переменные).
- **Codes for grouping variable** (Коды для групп переменной) указывают количество анализируемых групп объектов.
- **Missing data** (пропущенные переменные) позволяет выбрать **Case wise** (построчное удаление переменных из списка), либо **Mean substitution** (заменить их на средние значения).
- **Open Data** – открывает файл с данными.
- Можно указать условия выбора наблюдений из базы данных – кнопку **Select Cases** и веса переменных, выбрав их из списка – кнопку **W**.

При нажатии кнопки **Variables** открывается диалоговое окно выбора переменных (рис. 3.4).

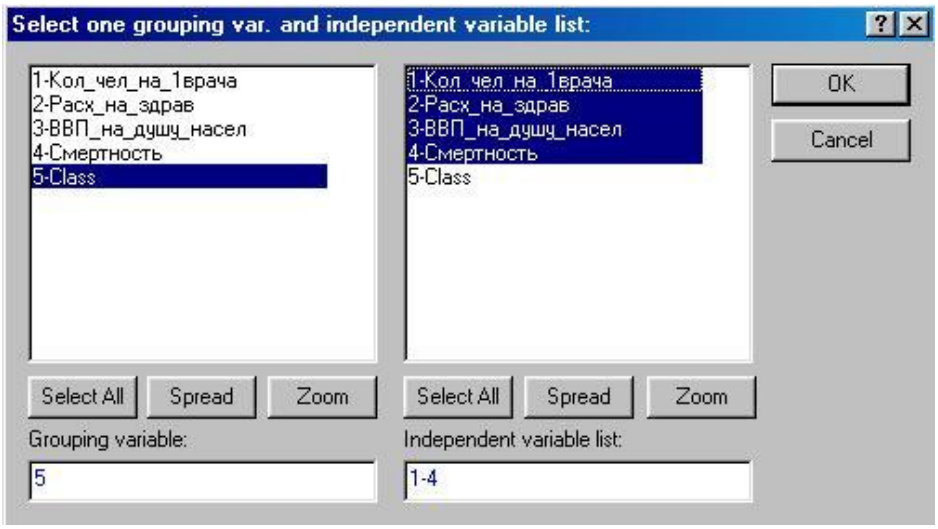


рис. 3.4

В левой части выбирается группирующие переменные, в правой – независимые переменные. Имена переменных в левой и правой части не должны пересекаться. В данном примере в качестве группирующей переменной выбрана переменная *CLASS*, а в качестве группирующих переменных *X1–X4*

Select All (Выделить все) выделяет все переменные, **Spread** (Подробности) – для просмотра длинного имени, **Zoom** (Информация о переменной) позволяет просмотреть информацию о переменной: ее имя, формат числового значения, описательные статистики: номер в группе, среднее значение, статистическое отклонение.

После нажатия кнопки **OK** откроется диалоговое окно **Model Difinition** (Определение модели) (рис. 3.5).

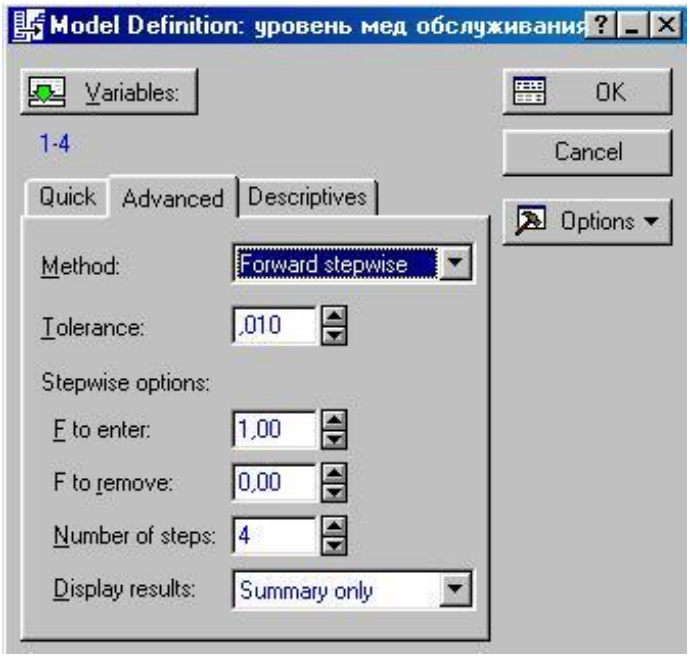


Рис. 3.5

В диалоговом окне **Model Definition** предложен выбор метода выбора значимых переменных. **Method** может быть задан **Standart** (Стандартный), **Forward stepwise** (Пошаговый с включением) и **Backward stepwise** (Пошаговый с исключением).

- Если выбран **Standart** (Стандартный метод), то все переменные будут одновременно включены в модель.
- В методе **Forward stepwise** (Пошаговый с включением) на каждом шаге в модель выбирается переменная с наибольшим F – значением. Процедура заканчивается, когда все переменные, имеющие F – значение больше значения, указанного в поле **F to enter** (F - включить), вошли в модель.
- Если выбран метод **Backward stepwise** (Пошаговый с исключением), то в уравнение будут включены все выбранные пользователем переменные, которые затем

удаляются в зависимости от величины F – значения. Шаги заканчиваются, когда нет переменных, имеющих F – значение меньше определенного пользователем в поле **F to remove** (F - исключить).

Если при проведении анализа пользователь хочет включить все переменные, то следует установить в поле **F to enter** (F - включить) очень маленькую величину, например 0.0001, а в поле **F to remove** (F – исключить) – 0.0.

Если же требуется исключить все переменные, то в поле **F to enter** (F - включить) следует установить большое значение, например 0.9999, а в поле **F to remove** (F – исключить) – 9998.

Поле **Number of steps** (число шагов) определяет максимальное число шагов анализа, по достижении которых процедура заканчивается.

Поле **Tolerance** (толерантность) позволяет исключить из модели неинформационные переменные. Значение толерантности вычисляется как $1-R^2$ переменной со всеми другими переменными в модели. Если толерантность имеет значение меньшее, чем значение по умолчанию 0.01 (или установленное специально пользователем), то эта переменная признается не информативной и не включается в модель.

В отличие от стандартного метода для пошаговых процедур предусмотрено два режима **Display of results** (вывода результатов) анализа:

- **At each step** (на каждом шаге) – программа выводит на экран диалоговое окно полных результатов на каждом шаге, начиная с нулевого.
- **Summary only** (на заключительном шаге) выводит окно с результатами только на последнем шаге, однако оно содержит опцию для просмотра основных итоговых статистик и для пошаговой процедуры.

Descriptives \ Review Descriptive Statistics (Обзор описательных статистик) позволяет получить описательные статистики для выбранных переменных:

- **Pooled within-groups covariances & correlations** (*объединенные внутригрупповые ковариации и корреляции*);
- **Total covariances & correlations** (*полные ковариации и корреляции*);
- **Graph** (*графики корреляционных функций для всех переменных*);
- **Means & number of cases** (*средние значения для каждой переменной*);
- **Box & wh** (*диаграммы размаха*);
- **Standart deviations** (*стандартные отклонения переменных в каждой группе*);

- **Categjrized histogram (by group)** (*категоризованные гистограммы по группам для каждой переменной*);
- **Box & whisker plot (by group)** (*диаграммы размаха по группам –категоризованную диаграмму рассеяния* (*по группам*));
- **Categorized scatterplot (by group)** (*для двух любых переменных*);
- **Categorized normal probability plot (by group)** (*категоризованный нормальный график для любой переменной по группам*).

Выберем в качестве метода (**Method**) – **Standard** и нажмем **OK**. В ходе вычислений системой получены результаты, которые представлены в окне **Discriminant Function Analisis Results** (Результаты анализа дискриминантных функций) (рис. 3.6).

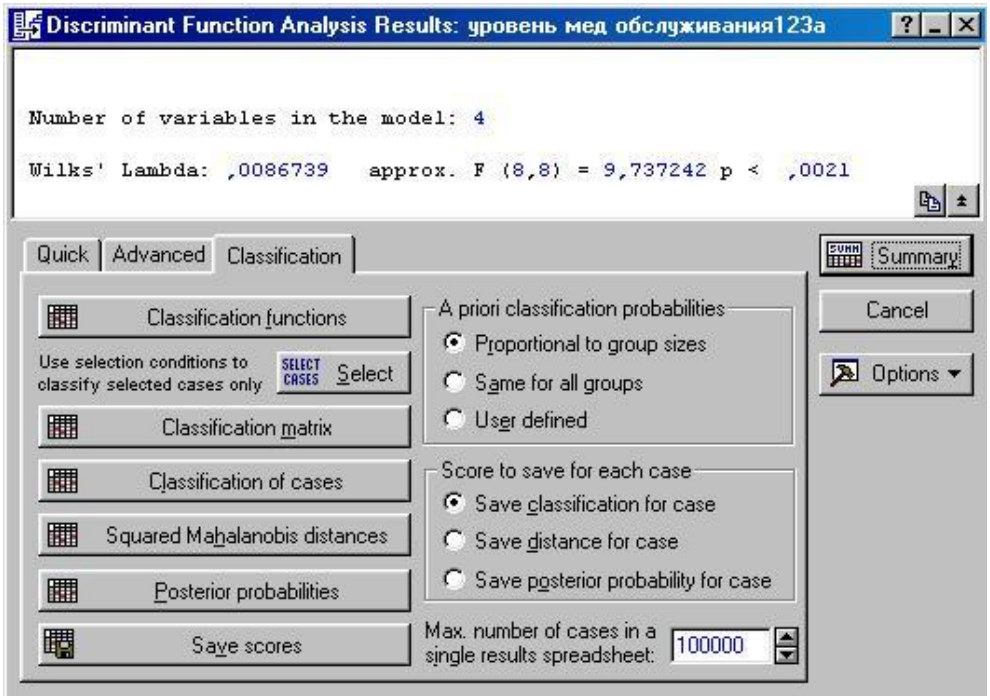


Рис. 3.6

Вывод результатов и их анализ:

Информационная часть диалогового окна **Discriminant Function Analisis Results** (Результаты Анализа Дискриминантных Функций) сообщает, что:

- **Number of variables in the model** (число переменных в модели) - 4;
- **Wilks' Lambda** (значение лямбды Уилкса) - 0,0086739;
- **Approx. *F* (8, 8)** (приближенное значение *F* – статистики, связанной с лямбдой Уилкса) - 9,737242;
- ***p* < *0.0021*** – уровень значимости *F* – критерия для значения 9,737242.

Значение статистики Уилкса лежит в интервале [0,1]. Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации, а значения, лежащие около 1, свидетельствуют о плохой дискриминации.

По данным показателя *Wilks' Lambda* (значение лямбды Уилкса), равного 0,0086739 и по значению *F* – критерия равного 9,737242, можно сделать вывод, что данная классификация корректная.[4]

В качестве проверки корректности обучающих выборок посмотрим результаты классификационной матрицы, нажав кнопку **Classification matrix** (Классификационная матрица**)** (рис. 3.7), предварительно выбрав **Same for all groups** (см. ниже) в правой части окна **Discriminant Function Analisis Results** (рис. 3.6).

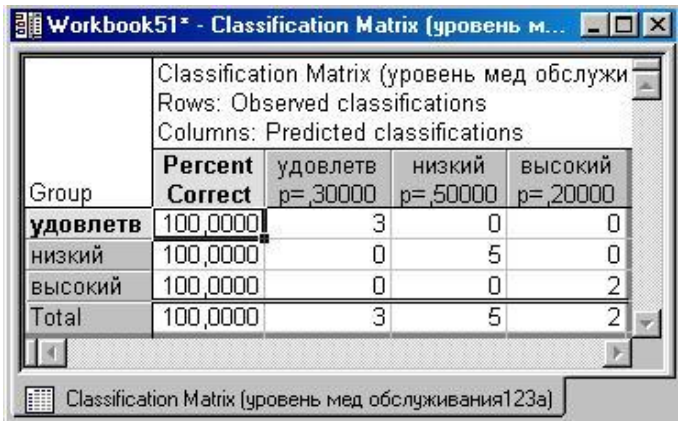


Рис. 3.7

Из классификационной матрицы можно сделать вывод, что объекты были правильно отнесены экспертным способом к выделенным группам. Если есть предприятия, неправильно отнесенные к соответствующим группам, можно посмотреть **Classification of cases** (Классификация случаев) (рис. 3.8).

[4] Значение статистики *F*–критерия: F(8,8) = 9,737242 больше табличного значения *F*–распределения: $F_{0,05}(8,8) = 3,4381$, т.е. нулевая гипотеза о том, что наблюдения принадлежат к одному классу, отвергается. Поэтому, дискриминантный анализ возможен.

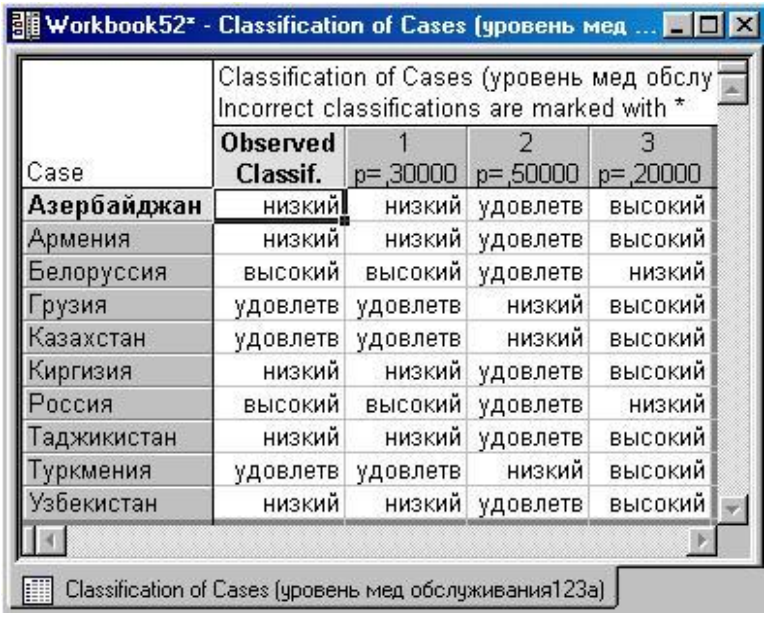


Рис. 3.8

В таблице классификации случаев некорректно отнесенные объекты помечаются звездочкой (*). Таким образом, задача получения корректных обучающих выборок состоит в том, чтобы исключить из обучающих выборок те объекты, которые по своим показателям не соответствуют большинству объектов, образующих однородную группу.

Для этого с помощью метрики Махаланобиса определятся расстояние от всех n объектов до центра тяжести каждой группы (вектор средних), определяемых по обучающей выборке. Отнесение экспертом i-го объекта в j-ю группу считается ошибочным, если расстояние Махаланобиса от объекта до центра его группы значительно выше, чем от него до центра других групп, а апостериорная вероятность попадания в свою группу ниже критического значения. В этом случае объект считается некорректно отнесенным и должен быть исключен из выборки.

Процедура исключения объекта из обучающих выборок состоит в том, что в таблице исходных данных у объекта, который должен быть исключен из выборки (он помечен "*"), убирается номер принадлежности к этой группе, после чего процесс тестирования повторяется. По предположению, сначала убирается тот объект, который наиболее не подходит к определенной группе, т.е. у которого наибольшее расстояние Махаланобиса и наименьшая апостериорная вероятность.

При удалении очередного объекта из группы нужно помнить, что при этом смещается центр тяжести группы (вектор средних), так как он определяется по оставшимся наблюдениям. После удаления очередного объекта из списка обучающих выборок не исключено, что появятся новые некорректно отнесенные объекты, которые до удаления

были учтены как правильно отнесенные. Поэтому данную процедуру нужно проводить, удаляя на каждом шаге лишь по одному объекту и возвращая его обратно в обучающие выборки.

Процедура исключения наблюдений продолжается до тех пор, пока общий коэффициент корректности в классификационной матрице достигнет 100%, т.е. все наблюдения обучающих выборок будут правильно отнесены к соответствующим группам.

Результаты полученных обучающих выборок, представлены в окне **Discriminant Function Analisis Results** (Результаты Анализа Дискриминантных Функций). В результате проведенного анализа общий коэффициент корректности обучающих выборок должен быть равен 100%.

Классификация объектов.

На основе полученных обучающих выборок можно проводить повторную классификацию тех объектов, которые не попали в обучающие выборки, и любых других объектов, подлежащих группировке. Для решения данной задачи, существуют два варианта: первый – провести классификацию на основе дискриминантных функций, второй – на основе классификационных функций.

В первом случае необходимо, не закрывая диалогового окна **Discriminant Function Analisis Results**, добавить в таблицу исходных скорректированных данных новые случаи. Для того чтобы понять, к какому классу относится этот объект, нажмите кнопку **Posterior probabilities** (Апостериорные вероятности). В системе имеется три способа задания априорной вероятности:

- **Proportional to group sizes** (Пропорциональные размерам групп)
- **Same for all groups** (Одинаковые для всех групп)
- **User defined** (Заданные пользователем)

После этого вы увидите таблицу с апостериорными вероятностями. К тем группам (классам), которые будут иметь максимальные вероятности, можно отнести новые случаи.

Во втором варианте необходимо в окне диалогового окна **Discriminant Function Analisis Results** нажать кнопку **Classification functions** (Классификационные функции). Появится окно (рис. 3.9), из которого можно выписать классификационные функции для каждого класса.

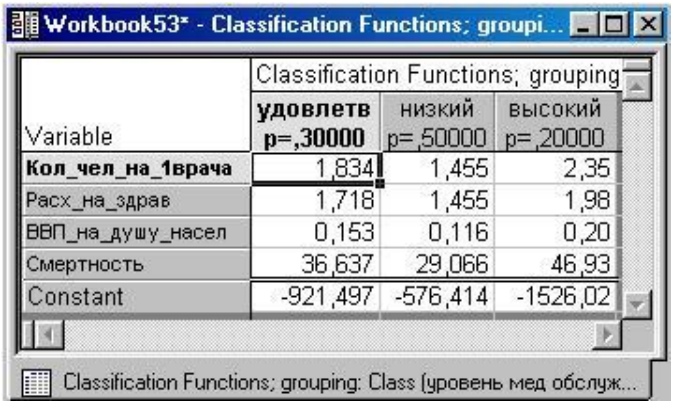


Рис. 3.9

$$высокий = -1526.02 + 2{,}35 * КолЧел_на_врача + 1{,}98 * Расх_здрав + \\ + 0{,}20 * ВВП + 46{,}93 * Смертность$$

$$удовлетв = -921.497 + 1{,}834 * КолЧел_на_врача + 1{,}718 * Расх_здрав + \\ + 0{,}153 * ВВП + 36{,}637 * Смертность$$

$$низкий = -576.414 + 1{,}455 * КолЧел_на_врача + 1{,}455 * Расх_здрав + \\ + 0{,}116 * ВВП + 29{,}066 * Смертность$$

С помощью этих функций можно будет в дальнейшем классифицировать новые случаи. Новые случаи будут относиться к тому классу, для которого классифицированное значение будет максимальное. Выбор метода окончательной классификации зависит от количества новых объектов, подлежащих классификации. Если количество новых случав невелико, можно применить метод, основанный на статистических критериях. Если же количество новых случаев велико, то рациональнее по обучающим выборкам получить классификационные функции и затем, настроить формулы и провести окончательную классификацию.

Определим принадлежность стран Молдавия и Украина, подставив значения соответствующих показателей в формулы (Таблица 3.1).

Таблица 3.1

Страна	Кол-во человек на 1 врача	Расходы на здравоохранение	ВВП на душу населения	Смертность	*Высокий*	*Удовлетворительный*	*Низкий*	Class
Молда	251	143	2500	12,6	438,2	628,63	**653,08**	Низк

вия					88	72	**76**	ий
Украина	224	131	3850	16,4	880,232	**904,2738**	863,3934	Удовл

Дополнительно можно просмотреть результаты канонического анализа, нажав кнопку **Perform canonical analysis** (выполнить канонический анализ) (рис. 3.10), если для анализа были выбраны, по крайней мере, три группы и есть хотя бы две переменные в модели.

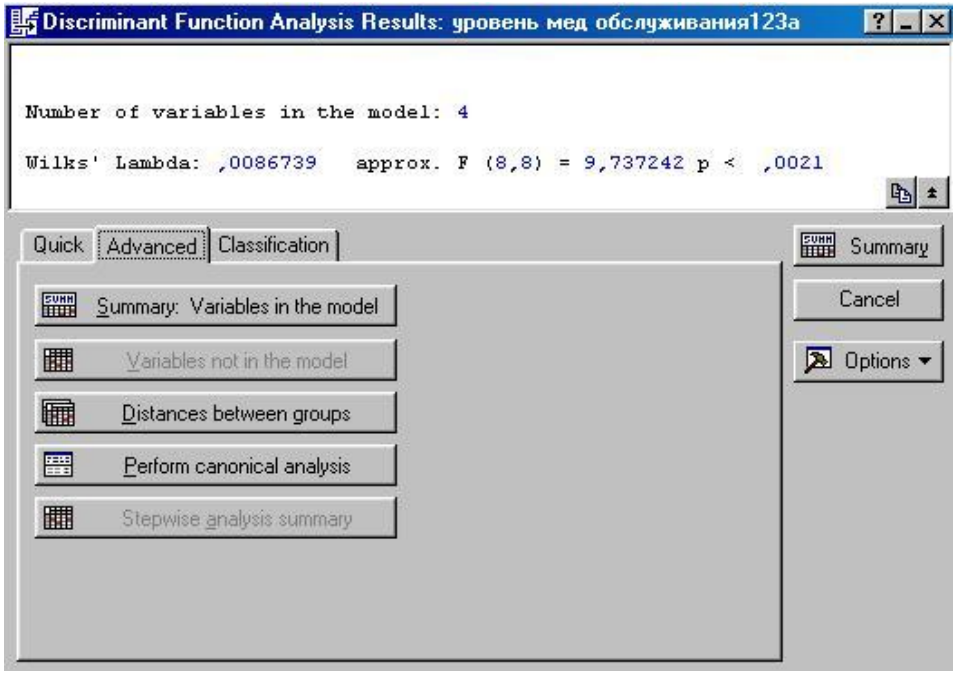


рис.3.10

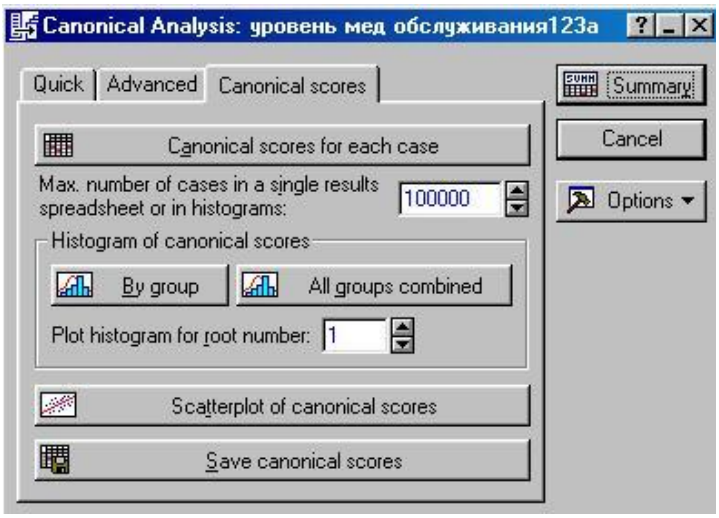


Рис 3.11

В выводном на экран окне **Canonical analysis** (Канонический анализ) (рис. 3.11) опция **Scatterplot of canonical scores** (Диаграмма рассеяния для значений) строит график

рассеяния канонических значений для канонических корней (рис. 3.12). С его помощью можно определить вклад, который вносит каждая дискриминантная функция в разделение между группами.

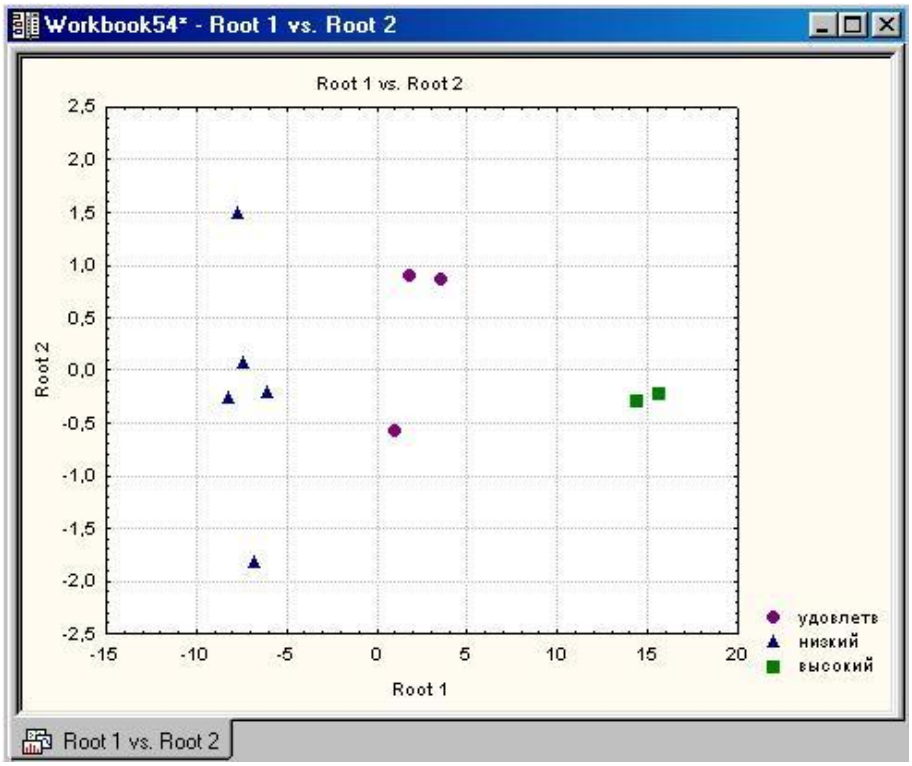


рис.3.12

3.3 Задания для самостоятельной работы

Задание 1

В таблице 3.2 представлены 9 машиностроительных предприятий, которые были выбраны и отнесены к соответствующим группам экспертным способом (с высоким и низким уровнем организации управления производством). Требуется на основе аналогичных показателей классифицировать предприятия 10-12 (Таблица 3.3).

Таблица 3.2

№ предприятия	Рентабельность, %	Производительность труда, млн руб/чел	Уровень организации управления производством

1	23,4	9,1	Высокий
2	19,1	6,6	Высокий
3	17,5	5,2	Высокий
4	17,2	10,1	Высокий
5	5,4	4,3	Низкий
6	6,6	5,5	Низкий
7	8,0	5,7	Низкий
8	9,7	5,5	Низкий
9	9,1	6,6	Низкий

Таблица 3.3

№ предприятия	Рентабельность, %	Производительность труда, млн руб/чел
10	9,9	7,4
11	14,2	9,4
12	12,9	6,7

Задание 2

В таблице 3.4 представлены 9 регионов, которые были выбраны и отнесены к соответствующим группам экспертным способом (с высоким и низким уровнем использования земли). Требуется с помощью дискриминантного анализа провести классификацию 10-12 районов (Таблица 3.5) по показателям объема реализованной продукции растениеводства и животноводство с 1 га посевной площади.

Таблица 3.4

№ района	Объем реализованной продукции		Уровень использования земли
	Растениеводство	Животноводство	
1	0,25	0,41	Низкий
2	0,51	0,51	Низкий
3	0,27	0,42	Низкий
4	0,33	0,56	Низкий

5	1,17	0,28	Высокий
6	4,99	0,67	Высокий
7	5,18	0,45	Высокий
8	2,49	0,38	Высокий
9	2,73	0,33	Высокий

Таблица 3.5

№ района	Объем реализованной продукции	
	Растениеводство	Животноводство
10	0,32	0,45
11	0,67	0,32
12	4,6	0,56

Задание 3

Эффективность деятельности 9 фермерских хозяйств оценивалась по трем показателям: производительность труда (млн. руб./чел.), Объем реализованной продукции растениеводства и животноводство с 1 га посевной площади (млн. руб./га). В результате были выделены хозяйства с высоким и низким уровнем использования земли

(Таблица 3.6). Необходимо с помощью дискриминантного анализа произвести классификацию 10-12 хозяйств (Таблица 3.7).

Таблице 3.6

№ хозяйства	Производительность труда	Объем реализованной продукции с 1 га посевной площади		Уровень использования земли
		Растениеводство	Животноводство	
1	8,22	0,25	0,41	Высокий
2	6,43	0,51	0,51	Высокий
3	6,39	0,27	0,42	Высокий
4	5,84	0,33	0,56	Высокий

5	3,14	1,17	0,28	Низкий
6	4,02	4,99	0,67	Низкий
7	4,19	5,18	0,45	Низкий
8	4,33	2,49	0,38	Низкий
9	4,58	2,73	0,33	Низкий

Таблице 3.7

№ хозяйства	Производительность труда	Объем реализованной продукции	
		Растениеводство	Животноводство
10	4,92	0,32	0,45
11	5,22	0,67	0,32
12	5,62	4,60	0,56

Задание 4

В Приложении 2 представлены данные по 65 предприятиям (n=65) и указаны номера групп, к которым они были отнесены экспертным способом. Задача состоит в том, чтобы проверить корректность экспертного отнесения предприятий к группам и получить корректные обучающие выборки (коэффициент корректности в классификационной матрице должен быть равен 100%). Выпишите линейные классификационные функции.

Задание 4

На основе корректных обучающих выборок и классификационных функций, полученных в результате выполнения задания 4, провести группировку предприятий по пяти группам на основе данных, представленных в Приложении 3, и соответствующих вариантов в таблице 3.7.

Таблице 3.7

Номер варианта	Номера предприятий	Номер варианта	Номера предприятий	Номер варианта	Номера предприятий
1	1-50	10	46-95	19	91-140
2	6-55	11	51-100	20	96-145
3	11-60	12	56-105	21	101-150

4	16-65	13	61-110	22	106-155
5	21-70	14	66-115	23	110-160
6	26-75	15	71-120	24	116-165
7	31-80	16	76-125	25	121-170
8	36-85	17	81-130		
9	41-90	18	86-135		

X1 – прибыль (тыс. р.);

X2 – валовая продукция на 1 работника, занятого в сельском хозяйстве (тыс. р.);

X3 – валовая продукция на 1 га сельхозугодий (тыс. р.);

X4 – производство молока на 1 га сельхозугодий (кг);

X5 – производство мяса на 1 га сельхозугодий (кг);

X6 – выручка от реализации продукции на 1 работника (тыс. р.);

X7 – выручка на 1 га сельхозугодий (тыс. р.).

ЗАКЛЮЧЕНИЕ

В учебном пособии были рассмотрены некоторые методы многомерного статистического анализа: кластерный, факторный и дискриминантный.

В частности, в кластерном анализе рассмотрены: меры близости объектов, способы определения кластеров и расстояния между ними.

В факторном анализе рассмотрены: основные понятия факторного анализа, анализ главных компонент. Представлены два критерия о выборе числа факторов: критерий Кайзера, критерий каменистой осыпи.

В дискриминантном анализе рассмотрены: основные понятия, цели и задачи дискриминантного анализа. А также определение числа и вида дискриминирующих функций, и классификация объектов с помощью функции расстояния.

Для каждого метода приведены примеры решения задач с использованием ППП STATISTICA.

Приведены задания для самостоятельного решения с использованием пакета STATISTICA.

ЛИТЕРАТУРА

1. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. – М.:Статистика, 1974. – 240 с.

2. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.:ЮНИТИ, 1998.

3. Айвазян С.А., Мхитарян В.С. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.

4. Салин В.Н., Чурилова Э.Ю. Практикум по курсу “статистика” (в системе STATISTICA), М., 2002.

5. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988.

6. Дуброва Т.А., Павлов Д.Э., Осипова Н.П. Факторный анализ с использованием пакета “STATISTICA”. Учебное пособие / МГУ экономики, статистики и информатики, - М., 2002.

7. Дуброва Т.А., Бажин А.Г., Бакуменко Л.П. Методы многомерной классификации. Дискриминантный анализ в системе STATISTICA. Учебное пособие / МГУ экономики, статистики и информатики; М., 2002.

8. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000.

Приложение 1

	Y1	Y2	Y3	X4	X5	X6	X7	X8
1	9.26	204.20	13.26	0.89	0.34	1.73	0.31	0.28
2	9.44	209.60	10.16	0.93	0.33	0.99	0.15	0.25
3	12.11	223.54	13.72	1.33	0.17	1.73	0.14	0.47
4	10.81	236.70	12.83	0.68	0.32	0.47	0.18	1.53
5	9.33	62.00	10.63	0.89	0.36	1.73	0.31	0.21
6	9.87	53.10	9.12	1.53	0.33	1.33	0.17	0.13
7	8.17	172.10	25.95	1.12	0.15	0.97	0.26	0.38
8	9.12	56.50	23.39	0.99	0.32	1.82	0.29	0.38
9	5.88	52.60	14.68	1.65	0.31	0.68	0.26	0.20
10	6.30	46.60	10.05	0.56	0.15	1.80	0.28	0.35
11	6.19	53.20	13.89	0.58	0.17	1.19	0.25	0.20
12	5.46	30.10	9.68	1.53	0.15	0.97	0.49	0.20
13	6.50	146.40	10.03	0.70	0.16	1.15	0.26	0.17
14	6.61	18.10	9.13	1.77	0.15	0.02	0.28	0.25
15	4.32	13.60	5.37	0.74	0.17	0.06	0.17	0.16
16	7.37	89.80	9.86	1.08	0.34	1.39	0.17	0.21
17	7.02	62.50	12.62	1.15	0.34	0.08	0.31	0.19
18	8.25	46.30	5.02	0.97	0.34	0.77	0.18	1.24
19	8.15	103.47	21.18	1.12	0.19	0.77	0.31	0.43
20	8.72	73.30	25.17	0.99	0.19	1.08	0.18	0.14
21	6.64	76.60	19.40	0.58	0.34	0.93	0.31	0.29
22	8.10	73.01	21.00	1.03	0.34	0.10	0.15	0.43
23	5.52	32.30	6.57	1.24	0.15	0.11	0.28	0.17
24	9.37	198.54	14.19	0.89	0.19	1.44	0.18	0.21
25	13.17	598.12	15.81	0.68	0.34	0.48	0.14	0.42
26	6.67	71.69	5.20	1.03	0.19	1.24	0.18	1.19
27	5.68	90.63	7.96	0.73	0.32	0.77	0.29	1.87
28	5.19	82.10	17.50	0.73	0.19	0.93	0.30	0.15
29	10.02	76.20	17.16	0.85	0.33	0.13	0.27	0.03
30	8.16	119.47	14.54	1.03	0.34	1.73	0.29	0.24
31	3.78	21.83	6.21	0.47	0.36	0.77	0.14	0.93
32	6.45	48.40	12.08	0.56	0.33	0.16	0.29	0.13
33	10.38	173.50	9.39	0.89	0.32	0.74	0.44	0.27
34	7.65	74.10	9.28	0.99	0.15	1.95	0.14	0.17
35	8.77	68.60	11.44	1.95	0.16	0.58	0.29	0.24
36	7.00	60.80	10.31	1.03	0.16	1.77	0.18	0.19
37	11.06	355.60	8.65	0.01	0.20	0.70	0.44	0.29
38	9.02	264.81	10.88	0.02	0.15	0.74	0.31	0.25
39	13.28	526.62	9.87	0.60	0.33	1.15	0.18	0.36
40	9.27	118.60	6.14	0.97	0.33	1.19	0.14	0.17
41	6.73	37.10	12.99	1.12	0.19	1.03	0.31	0.23
42	6.72	57.97	9.78	1.77	0.15	1.08	0.18	0.17
43	9.44	51.84	13.22	0.93	0.32	0.13	0.27	0.24
44	7.21	64.70	17.29	1.12	0.16	0.74	0.26	0.26
45	5.39	48.30	7.11	0.74	0.31	0.99	0.49	0.13
46	5.61	15.00	22.49	0.47	0.32	0.64	0.28	0.28
47	5.59	87.47	12.14	1.12	0.15	1.87	0.31	0.34

48	6.57	108.40	15.25	0.58	0.16	1.12	0.26	0.26
49	6.54	267.31	31.34	0.64	0.20	0.08	0.40	4.47
50	4.20	34.20	11.56	0.60	0.19	0.17	0.26	0.25
51	5.19	26.92	30.14	1.53	0.20	0.03	0.44	2.13
52	18.00	43.84	19.71	1.73	0.32	0.02	0.30	0.27
53	11.03	72.00	23.56	1.12	0.15	0.85	0.27	2.20

	X9	**X10**	**X11**	**X12**	**X13**	**X14**	**X15**	**X16**	**X17**
1	0.89	0.14	112216.00	166.19	9889.98	6.40	167.29	10.08	17.72
2	1.80	0.30	37631.94	186.10	22123.47	7.80	92.88	14.76	18.39
3	1.53	0.31	45178.00	220.45	10787.25	9.76	159.01	6.45	26.46
4	0.60	0.18	76688.00	169.30	10272.25	7.90	93.96	21.83	22.37
5	1.39	0.37	7361.00	39.93	55268.00	5.35	173.88	11.94	28.13
6	1.24	0.19	84496.00	40.41	45322.00	9.96	162.30	12.60	17.55
7	1.77	0.41	114132.00	102.96	12657.25	4.50	88.56	11.52	21.79
8	0.09	0.36	7801.00	37.02	57564.00	4.88	101.16	8.28	19.52
9	0.52	0.41	84504.00	45.94	118239.88	3.46	167.29	11.52	23.85
10	0.80	2.06	35852.00	40.07	64362.00	3.62	140.76	32.40	21.88
11	0.74	0.41	43244.00	45.44	69647.88	3.56	128.52	11.52	25.68
12	0.05	0.24	6358.00	41.08	49844.00	5.65	177.84	17.28	18.13
13	1.03	0.40	47378.00	136.14	22497.50	4.28	114.48	16.20	25.74
14	1.48	0.21	4210.00	42.39	6920.00	8.83	93.24	13.36	21.21
15	0.73	0.36	3572.50	37.39	5736.00	8.52	126.72	17.28	22.86
16	0.36	0.49	54544.00	101.78	47266.00	7.22	91.27	9.72	16.38
17	0.13	0.43	91264.00	47.91	72080.00	4.82	69.12	16.20	13.21
18	0.46	0.44	5975.00	32.61	83704.00	5.47	66.24	24.88	14.41
19	0.29	0.18	64044.00	103.73	107636.00	6.23	67.16	14.76	13.44
20	1.87	2.24	34328.00	38.95	67592.00	4.25	50.40	7.56	13.69
21	0.47	0.30	58424.00	81.32	99812.00	5.38	70.89	8.64	16.66
22	0.34	0.15	83240.00	67.75	75680.00	5.88	72.00	8.64	15.06
23	0.27	0.17	6462.00	59.66	44196.00	9.27	97.20	9.00	20.09
24	0.80	2.30	114896.00	107.81	20898.50	4.36	80.28	14.76	15.91
25	0.97	0.31	21791.47	512.62	28946.00	10.31	51.48	10.08	18.27
26	1.39	0.44	83568.00	53.53	74687.88	4.72	105.12	14.76	14.44
27	0.16	0.18	68976.00	80.83	8631.25	4.18	128.52	10.38	22.88
28	0.15	0.39	67663.88	59.42	31314.00	3.13	94.68	14.76	15.50
29	1.15	2.60	34428.00	36.96	64752.00	4.02	85.32	20.52	19.35
30	0.21	0.45	127256.00	91.88	8206.25	5.20	76.32	14.46	16.95
31	0.89	0.45	6265.00	17.16	44676.00	2.72	153.00	24.88	30.53
32	1.15	2.25	33192.00	27.29	65188.00	3.12	107.34	11.16	17.78
33	0.13	0.49	127983.88	184.33	22697.00	10.38	90.72	6.45	22.09
34	0.33	0.14	41368.00	58.42	68104.00	5.65	82.44	9.72	18.29
35	0.64	0.18	33556.00	59.31	65616.00	6.67	79.12	3.24	26.05
36	0.93	0.29	124560.00	49.87	127344.00	5.93	120.96	6.45	26.20
37	0.14	0.50	110548.00	391.27	7919.00	11.89	84.60	5.40	17.26
38	0.13	0.26	95968.00	258.61	14314.75	8.30	85.32	6.12	18.95
39	0.16	0.21	21182.50	75.14	9277.13	0.18	101.52	8.64	19.66
40	0.49	0.49	53284.00	123.16	122072.00	8.88	107.34	11.94	16.97
41	1.80	0.28	6338.00	37.21	85792.00	5.82	85.32	7.92	14.63
42	0.15	0.15	44460.00	53.37	79631.88	4.80	131.76	10.08	22.17
43	1.99	2.03	6555.00	32.87	40516.00	5.01	116.64	18.72	22.62

44	0.10	0.40	44340.00	45.87	72580.00	4.12	138.24	13.68	26.44
45	0.17	0.27	35888.00	48.41	72296.00	5.10	156.96	16.56	22.26
46	0.21	0.38	3947.50	13.58	7612.00	3.47	137.52	14.76	19.13
47	1.65	0.45	59356.00	63.69	49072.00	4.22	134.06	7.92	18.28
48	0.64	0.14	75304.00	104.55	22896.00	5.01	155.52	18.36	28.23
49	0.20	0.43	68712.00	223.70	10949.73	11.38	48.60	8.28	12.44
50	1.92	0.14	3351.00	25.88	6235.00	7.67	42.84	14.04	11.64
51	1.65	0.17	6369.00	29.52	84000.00	4.68	143.64	16.79	8.62
52	0.09	0.18	90336.00	41.69	65050.00	4.30	145.80	11.16	20.10
53	0.19	0.17	38988.00	78.11	80708.00	6.62	120.52	14.76	19.41

Приложение 2

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*CLASS1*
1	2174.000	9658.000	466.000	386.000	35.000	36.400	1756.000	1
2	274.000	10477.000	2321.000	767.000	56.000	35.600	7884.000	1
3	-146.000	6567.000	713.000	581.000	74.000	13.800	1501.000	1
4	-338.000	10282.000	499.000	764.000	51.000	30.200	1466.000	1
5	-716.000	9316.000	677.000	533.000	109.000	20.500	1486.000	1
6	892.800	6425.000	944.000	1390.000	78.000	13.200	1936.000	1
7	191.000	5367.000	786.000	819.000	104.000	13.700	2011.000	1
8	0.000	6342.000	486.000	261.000	52.000	24.100	1841.000	1
9	-107.000	5868.000	531.000	450.000	63.000	22.300	1608.000	1
10	-903.000	6330.000	636.000	401.000	69.000	17.600	1768.000	1
11	-765.000	12573.000	669.000	713.000	47.000	33.900	1806.000	1
12	326.000	4110.000	600.000	373.000	74.000	8.000	1172.000	2
13	150.000	7832.000	288.000	336.000	49.000	20.000	736.000	2
14	-18.000	6793.000	620.000	487.000	104.000	19.400	1775.000	2
15	1.300	4731.000	447.000	405.000	64.000	10.400	979.000	2
16	-380.000	5564.000	565.000	400.000	48.000	14.900	1517.000	2
17	-790.000	5470.000	432.000	509.000	85.000	11.800	935.000	2
18	-666.800	3988.000	364.000	213.000	35.000	10.300	943.000	2
19	-204.500	5121.000	495.000	628.000	77.000	16.700	1616.000	2
20	-094.000	3900.000	420.000	359.000	53.000	9.600	1034.000	2
21	-034.000	5871.000	495.000	353.000	92.000	14.300	1206.000	2
22	-784.000	4352.000	429.000	197.000	62.000	10.900	1070.000	2
23	-403.800	4635.000	378.000	221.000	46.000	10.500	856.000	3
24	-717.000	6056.000	247.700	150.000	24.100	15.600	640.000	3
25	-458.000	5180.000	433.600	429.000	44.400	10.500	880.000	3
26	-908.000	6295.000	206.000	127.000	17.000	22.800	743.000	3
27	-514.000	5340.000	364.000	411.000	17.000	14.400	984.000	3
28	-205.000	5357.000	583.000	716.000	87.000	14.800	1606.600	3
29	403.100	2969.000	382.000	274.000	29.000	5.700	728.000	3
30	-205.000	4924.000	284.000	292.000	35.000	17.500	1010.000	3
31	-256.000	7622.000	342.000	223.000	26.000	14.100	634.000	3
32	-314.000	4394.000	471.000	396.000	68.000	9.900	1065.000	3
33	-027.000	3312.000	284.000	229.000	39.000	11.100	948.000	3
34	1779.000	5001.000	304.400	286.000	37.600	12.000	732.000	3
35	-842.000	4247.000	233.000	189.000	28.000	12.800	757.000	3
36	-542.000	4025.000	199.300	145.000	14.400	12.000	596.000	4
37	-298.000	3429.000	184.000	105.000	18.000	6.700	357.300	4
38	-446.000	3047.000	310.000	244.000	47.000	5.500	560.000	4
39	-236.000	3410.000	181.000	147.000	20.000	10.900	576.000	4
40	-493.000	4551.000	212.000	169.000	22.000	13.800	645.000	4
41	-900.000	4573.000	284.000	254.000	37.000	11.300	698.000	4
42	-586.000	3924.000	212.000	154.000	17.000	13.000	704.000	4
43	-634.000	3751.000	212.000	125.000	17.000	5.400	303.000	4
44	-142.000	4318.000	257.000	151.000	33.000	16.500	985.000	4
45	-394.000	3140.000	218.000	241.000	47.000	8.500	592.000	4
46	-571.000	4617.000	171.000	137.000	13.000	13.100	484.000	4

47	-728.300	5448.000	348.000	215.000	28.000	5.700	367.000	4
48	-796.000	2902.000	161.000	182.000	22.000	11.400	631.000	4
49	-955.200	3634.000	334.000	361.000	59.000	10.100	925.000	4
50	-294.000	3499.000	204.000	129.000	27.000	6.800	398.000	4
51	-500.000	6368.000	288.000	169.000	27.000	13.300	601.000	4
52	-961.000	4194.000	328.000	312.000	44.000	9.500	744.000	4
53	-934.000	6322.000	510.000	548.000	41.000	14.700	1187.000	4
54	-161.600	3196.000	288.000	149.000	55.000	7.600	684.000	5
55	-004.000	3666.000	168.000	131.000	19.000	8.300	382.000	5
56	-879.000	3058.000	169.000	86.000	23.000	5.600	307.000	5
57	-197.000	5110.000	82.000	57.000	11.000	1.100	174.000	5
58	-310.700	4166.000	207.000	183.000	32.000	9.800	487.000	5
59	-437.000	5168.000	151.000	96.000	8.000	10.700	359.000	5
60	-482.000	2061.000	78.000	47.000	4.000	2.900	110.300	5
61	-855.000	3483.000	109.000	90.000	16.000	7.600	237.000	5
62	-892.200	1917.000	98.000	64.000	9.000	4.000	174.000	5
63	-766.000	2001.000	95.000	87.000	18.000	5.000	239.000	5
64	-950.000	1728.000	87.000	75.000	13.000	3.400	172.300	5
65	-369.000	1094.000	38.000	1.200	3.200	3.300	114.000	5

Приложение 3

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
1	-790.000	5470.000	432.000	509.000	85.000	11.800	935.000
2	-1205.000	3698.000	187.500	156.000	21.600	10.000	507.000
3	-751.000	3448.000	278.000	206.000	25.000	7.400	596.000
4	-107.000	5868.000	531.000	450.000	63.000	22.300	1608.000
5	-903.000	6330.000	636.000	401.000	69.000	17.600	1768.000
6	-1204.500	5121.000	495.000	628.000	77.000	16.700	1616.000
7	-717.000	6056.000	247.700	150.000	24.100	15.600	640.000
8	-1027.000	3312.000	284.000	229.000	39.000	11.100	948.000
9	-625.000	2806.000	201.000	207.000	29.000	6.000	429.000
10	-1784.000	4352.000	429.000	197.000	62.000	10.900	1070.000
11	-1147.000	1849.000	158.000	45.000	13.000	3.100	266.600
12	-614.000	2168.000	191.000	168.000	22.000	9.100	804.000
13	-1080.000	3130.000	213.000	236.000	33.000	11.800	807.000
14	403.100	2969.000	382.000	274.000	29.000	5.700	728.000
15	139.000	3264.000	340.000	316.000	28.000	6.800	711.000
16	-444.000	3920.000	139.000	120.000	9.000	13.100	464.000
17	-833.000	5563.000	271.000	148.000	15.700	8.700	426.000
18	-380.000	5564.000	565.000	400.000	48.000	14.900	1517.000
19	-634.000	3751.000	212.000	125.000	17.000	5.400	303.000
20	-482.000	2061.000	78.000	47.000	4.000	2.900	110.300
21	-1892.000	4823.000	254.000	269.000	33.000	12.800	673.000
22	-256.000	7622.000	342.000	223.000	26.000	14.100	634.000
23	-987.000	4412.000	249.400	236.000	26.400	10.300	582.000
24	-255.000	2806.000	262.000	147.000	11.000	5.000	468.000
25	-908.000	6295.000	206.000	127.000	17.000	22.800	743.000
26	-1012.000	3912.000	102.000	87.000	9.000	11.200	290.000
27	-1500.000	6368.000	288.000	169.000	27.000	13.300	601.000
28	-321.000	1339.000	71.000	50.000	11.000	3.100	161.600
29	-1872.000	3883.000	262.000	230.000	26.000	8.300	557.000
30	-1161.600	3196.000	288.000	149.000	55.000	7.600	684.000
31	307.700	16355.000	939.000	83.000	25.000	8.700	501.000
32	-1146.000	6567.000	713.000	581.000	74.000	13.800	1501.000
33	-718.000	2027.000	126.000	105.000	12.000	7.700	474.000
34	-1197.000	5300.000	485.000	412.000	51.000	10.000	924.000
35	-497.800	5056.000	445.000	536.000	60.000	15.400	1388.100
36	-314.000	4394.000	471.000	396.000	68.000	9.900	1065.000
37	-373.000	5216.000	321.000	311.000	43.000	11.900	734.000
38	-854.300	5602.000	856.000	999.000	120.000	15.000	2285.000
39	-2004.000	3666.000	168.000	131.000	19.000	8.300	382.000
40	-900.000	4573.000	284.000	254.000	37.000	11.300	698.000
41	-219.800	3619.000	333.000	358.000	31.000	7.400	679.800
42	-523.600	3586.000	383.000	442.000	41.000	11.700	1252.300
43	-1731.300	2785.000	388.000	366.000	48.000	5.500	771.000
44	-205.000	5357.000	583.000	716.000	87.000	14.800	1606.600
45	-867.000	3495.000	150.700	98.000	14.200	9.700	419.000
46	-950.000	1728.000	87.000	75.000	13.000	3.400	172.300
47	-2408.000	3714.000	253.000	229.000	24.000	7.600	518.000
48	-226.200	2863.000	252.000	203.000	31.000	6.600	582.600

49	-1597.000	2721.000	157.000	133.000	19.000	6.700	388.000
50	-1148.000	5126.000	468.000	340.000	50.000	12.200	1116.000
51	-504.600	3624.000	250.000	235.000	28.000	15.500	1071.000
52	-1796.000	2902.000	161.000	182.000	22.000	11.400	631.000
53	-1879.000	3058.000	169.000	86.000	23.000	5.600	307.000
54	-503.000	4634.000	336.000	271.000	31.000	11.100	802.000
55	-1394.000	3140.000	218.000	241.000	47.000	8.500	592.000
56	653.000	4791.000	245.000	202.000	37.000	17.300	881.000
57	-1094.000	3900.000	420.000	359.000	53.000	9.600	1034.000
58	-1949.000	3294.000	322.000	347.000	84.000	10.600	1034.000
59	-1064.000	2610.000	203.000	237.000	43.000	7.400	578.000
60	-1571.000	4617.000	171.000	137.000	13.000	13.100	484.000
61	-447.000	3390.000	282.000	258.000	33.000	7.100	592.000
62	-768.000	2845.000	332.000	241.000	43.000	5.500	637.000
63	-848.000	5197.000	269.200	163.000	14.200	9.900	513.000
64	-205.000	4924.000	284.000	292.000	35.000	17.500	1010.000
65	-1961.000	4194.000	328.000	312.000	44.000	9.500	744.000
66	-767.200	5863.000	622.200	277.000	95.000	13.800	1779.000
67	-1294.000	3499.000	204.000	129.000	27.000	6.800	398.000
68	-375.000	4291.000	187.000	165.000	18.000	11.300	493.000
69	-18.000	6793.000	620.000	487.000	104.000	19.400	1775.000
70	-526.000	3338.000	247.200	216.000	15.700	6.800	501.000
71	-745.000	3262.000	214.000	103.000	29.000	5.600	370.000
72	-113.400	7444.000	506.000	669.000	50.000	16.200	1100.200
73	-1332.000	4455.000	184.000	325.000	54.000	10.500	825.000
74	-1493.000	4551.000	212.000	169.000	22.000	13.800	645.000
75	-446.000	3047.000	310.000	244.000	47.000	5.500	560.000
76	714.600	7655.000	687.000	226.000	36.000	16.400	1474.000
77	179.000	4097.000	474.000	345.000	46.000	8.000	536.000
78	274.000	10477.000	2321.000	767.000	56.000	35.600	7884.000
79	-361.100	2754.000	385.000	499.000	62.000	10.700	1493.000
80	-1765.000	2977.000	281.000	249.000	34.000	6.500	612.000
81	1.300	4731.000	447.000	405.000	64.000	10.400	979.000
82	-3636.000	4129.000	133.000	57.000	14.000	13.900	446.000
83	-1016.000	4028.000	217.000	236.000	23.000	10.600	568.000
84	-2766.000	2001.000	95.000	87.000	18.000	5.000	239.000
85	-1008.000	2954.000	179.000	133.000	17.000	7.600	461.000
86	-452.000	2034.000	251.000	185.000	48.000	3.700	457.000
87	-419.500	1982.000	237.000	242.000	66.000	11.100	1327.000
88	-1306.800	2400.000	270.000	212.000	29.000	4.800	538.000
89	-809.000	4590.000	180.000	147.000	23.000	12.500	488.000
90	-1765.000	12573.000	669.000	713.000	47.000	33.900	1806.000
91	24.000	4666.000	268.000	344.000	16.000	8.900	512.000
92	-454.000	4799.000	433.000	421.000	41.000	9.200	831.000
93	-599.300	3535.000	287.000	287.000	40.000	9.000	752.000
94	-2142.000	4318.000	257.000	151.000	33.000	16.500	985.000
95	-1514.000	5340.000	364.000	411.000	17.000	14.400	984.000
96	-1458.000	5180.000	433.600	429.000	44.400	10.500	880.000
97	-337.000	3017.000	187.000	191.000	29.000	6.600	411.000
98	-279.000	4211.000	294.000	259.000	41.000	6.400	451.000
99	-1376.000	4961.000	292.000	294.000	25.000	17.000	999.000

100	-403.800	4635.000	378.000	221.000	46.000	10.500	856.000
101	-2716.000	9316.000	677.000	533.000	109.000	20.500	1486.000
102	-298.000	3429.000	184.000	105.000	18.000	6.700	357.300
103	191.000	5367.000	786.000	819.000	104.000	13.700	2011.000
104	-728.300	5448.000	348.000	215.000	28.000	5.700	367.000
105	-404.000	2988.000	161.000	97.000	20.000	7.900	426.200
106	-473.700	5850.000	454.000	644.000	68.000	17.600	1365.500
107	-444.200	2351.000	297.000	183.000	35.000	4.500	564.000
108	-1236.000	3410.000	181.000	147.000	20.000	10.900	576.000
109	-715.000	2980.000	199.000	182.000	42.000	8.900	606.000
110	-1166.000	3798.000	273.000	267.000	42.000	8.600	619.000
111	-1141.000	3395.000	123.000	105.000	18.000	12.400	449.000
112	-486.000	1304.000	167.000	109.000	20.000	2.300	289.000
113	-667.000	2935.000	154.000	136.000	20.000	5.800	302.000
114	1365.000	5833.000	272.000	145.000	17.000	9.300	434.000
115	1553.000	2922.000	214.000	190.000	47.000	7.900	582.000
116	-706.000	3878.000	230.300	121.000	31.800	7.000	417.000
117	-1615.100	3779.000	216.000	264.000	18.000	13.300	763.000
118	-916.000	2362.000	173.000	144.000	18.000	6.900	505.000
119	-854.000	4532.000	278.000	233.000	26.000	11.600	714.000
120	-314.000	2788.000	191.000	105.000	21.000	5.900	406.000
121	-2128.000	3904.000	203.000	152.000	13.000	8.500	440.000
122	1779.000	5001.000	304.400	286.000	37.600	12.000	732.000
123	360.000	5490.000	315.000	214.000	30.000	14.900	854.000
124	2174.000	9658.000	466.000	386.000	35.000	36.400	1756.000
125	-662.000	3257.000	180.000	94.000	12.000	5.500	303.000
126	-288.300	3678.000	399.000	485.000	43.000	9.900	1074.000
127	1173.000	3426.000	202.000	214.000	32.000	8.500	502.000
128	-374.200	2823.000	216.000	62.000	13.000	4.400	337.000
129	-1527.000	3273.000	164.000	136.000	16.000	14.300	719.000
130	326.000	4110.000	600.000	373.000	74.000	8.000	1172.000
131	-542.000	4025.000	199.300	145.000	14.400	12.000	596.000
132	-745.000	2529.000	151.400	117.000	12.000	7.100	427.000
133	-409.600	2442.000	191.000	113.000	19.000	5.300	413.000
134	-892.200	1917.000	98.000	64.000	9.000	4.000	174.000
135	-2253.000	6344.000	359.000	321.000	62.000	22.800	1292.000
136	-563.700	4248.000	343.000	260.000	42.000	10.800	868.900
137	-101.000	4962.000	523.000	537.000	66.000	11.300	1190.000
138	37.000	4725.000	148.000	115.000	17.000	11.900	371.200
139	-393.300	5373.000	325.000	135.000	22.000	7.600	462.000
140	-1050.000	4004.000	189.000	187.000	43.000	11.600	544.000
141	-478.000	1709.000	165.000	132.000	21.000	4.800	465.000
142	-1279.100	3469.000	337.000	375.000	74.000	9.000	871.300
143	-2034.000	5871.000	495.000	353.000	92.000	14.300	1206.000
144	-453.000	5507.000	296.000	281.000	36.000	17.200	925.000
145	0.000	6342.000	486.000	261.000	52.000	24.100	1841.000
146	-2033.000	2908.000	289.000	254.000	28.000	9.500	941.000
147	-1738.000	4906.000	169.200	149.000	27.900	13.400	464.000
148	-1900.000	5603.000	246.600	181.000	16.300	14.700	648.000
149	-1437.000	5168.000	151.000	96.000	8.000	10.700	359.000
150	-1978.000	3354.000	442.000	400.000	59.000	7.500	1098.000

151	-3161.000	3641.000	260.400	201.000	37.300	9.900	711.000
152	-1369.000	1094.000	38.000	1.200	3.200	3.300	114.000
153	-531.000	3714.000	211.600	189.000	20.800	10.900	622.000
154	-884.300	2866.000	237.000	146.000	30.000	6.800	566.000
155	-1955.200	3634.000	334.000	361.000	59.000	10.100	925.000
156	-1004.000	4948.000	230.000	144.000	32.000	11.800	548.000
157	-1855.000	3483.000	109.000	90.000	16.000	7.600	237.000
158	-1338.000	10282.000	499.000	764.000	51.000	30.200	1466.000
159	-878.000	4395.000	275.000	208.000	20.000	7.200	447.000
160	-1078.000	4367.000	216.000	176.000	25.000	11.500	567.000
161	-666.800	3988.000	364.000	213.000	35.000	10.300	943.000
162	-277.600	1744.000	202.000	222.000	31.000	5.700	665.000
163	-350.200	2549.000	246.000	161.000	34.000	6.300	610.000
164	-412.000	4331.000	416.000	253.000	64.000	10.900	1045.000
165	-573.000	5163.000	407.000	285.000	43.000	12.900	1017.000
166	-197.000	5110.000	82.000	57.000	11.000	1.100	174.000
167	-1842.000	4247.000	233.000	189.000	28.000	12.800	757.000
168	-666.000	4222.000	284.700	198.000	44.300	14.000	944.000
169	892.800	6425.000	944.000	1390.000	78.000	13.200	1936.000
170	-1597.000	3396.000	162.000	146.000	23.000	10.600	506.000