Федеральное агентство по образованию
ГОУ ВПО «Иркутский государственный университет»

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКИХ МЕТОДОВ ПРИ АНАЛИЗЕ ГЕОЛОГИЧЕСКОЙ ИНФОРМАЦИИ (с использованием компьютерных технологий)

Часть III

Учебное пособие

Иркутск
2006

УДК 518.+550 (07)
ББК 26.3:22.1

Рецензенты:
канд. физ.-мат. наук, доц**. В. Н. Докин,**
канд. физ.-мат. наук, доц. **Н. Г. Коновалова**

Применение математических методов при анализе геологической информации (с использованием компьютерных технологий) / сост.: *И. М. Михалевич, С. П. Примина* : учеб. пособие. Ч. III. – Иркутск : Иркут. гос. ун-т, 2006, – 115 с.

Данное учебное пособие является продолжением I и II частей пособий, составленных в 2001 г. и 2004 г. /5, 10/ и предназначено для дальнейшего изучения и применения распространенных многомерных статистических методов при анализе данных, полученных при геологоразведочных работах.

Описание количественных методов сопровождается примерами и решением их с помощью широко известного статистического пакета программ **Statistica**.

Рассчитано на студентов геологических специальностей, может быть использовано аспирантами, научными сотрудниками и практическими геологами.

Библиогр. 49 назв. Ил. 75. Табл. 15.

ОГЛАВЛЕНИЕ

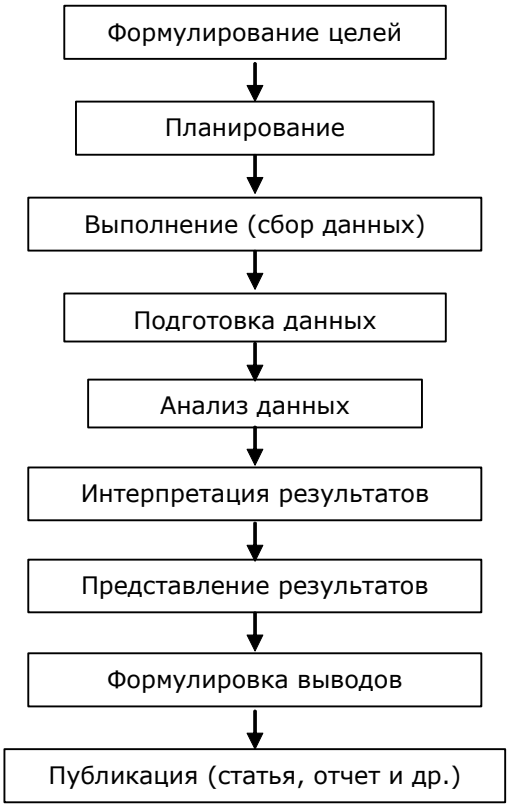
ВМЕСТО ВВЕДЕНИЯ 3
Цель исследования 5
Планирование исследования 5
Типы исследований 6
Объемы выборок 6
Типы экспериментальных данных 6
Статистический анализ данных 7
Описательная статистика 8
Статистические выводы 8
Статистические критерии и методы 8
Многомерные методы (наиболее часто используемые) 10
1. КЛАСТЕРНЫЙ АНАЛИЗ 12
1.1.Виды группирования объектов в программе *STATISTICA* *23*
1.1.1. Методы кластеризации 24
1.1.2. Меры сходства, используемые в программе STATISTICA..... 26
1.1.3. Метод k-средних — k-means clastering 27
1.2. Применение кластерного анализа в программе *STATISTICA* 28
1.2.1. Варианты группирования 34
1.2.2. Пример использования программы STATISTICA при группировании методом k-средних — k-means clastering 38
2. МЕТОД ГЛАВНЫХ КОМПОНЕНТ 44
2.1. Сущность метода главных компонент 44
2.2. Применение метода главных компонент в пакете *STATISTICA* (пример) 49
2.3. Принцип факторного анализа 52
3. ДИСКРИМИНАНТНЫЙ АНАЛИЗ 54
3.1. Критерии значимости 59
3.2. Дискриминантный анализ в пакете STATISTICA, интерпретация результатов 61
3.2.1. Демонстрационный пример 61
3.2.2. Критерий Хотеллинга T^2 пакете STATISTICA 71
3.2.3. Применение дискриминантного анализа при количестве групп более двух 72
4. ПРИЛОЖЕНИЯ 85
4.1. Задачи и упражнения 96
4.2. Ответы и решения 96
Заключение 105
Библиографический список 111

ВМЕСТО ВВЕДЕНИЯ
Краткий обзор использования математических методов при исследованиях в геологии

Любые исследования, в том числе и исследования в геологии, делятся на этапы. Кратко остановимся на этапах исследования (см. схему).

Схема

Этапы исследования



Рассмотрим подробнее структуру исследования и роли в нем математико-статистического анализа. Следует подчеркнуть, что для получения надежных, научно обоснованных результатов исследования необходимы два компонента:

а) правильное планирование структуры исследования (обеспечивающей возможность получения ответов на поставленные вопросы) и

б) грамотный статистический анализ.

Ошибки в планировании исследования первичны. Если структура исследования неадекватна задачам исследования и чревата систематическими ошибками, то даже самый совершенный статистический анализ не обеспечит научно обоснованных результатов.

Аналогичная ситуация возникает и в том случае, если исследование спланировано правильно, но статистический анализ проведен плохо. Ошибки в статистическом анализе ведут к неверным выводам.

Отметим сразу, что участие специалиста по прикладной математике в исследовании весьма желательно не только на этапе анализа данных, но и практически на всех других этапах (см. схему), которые мы кратко рассмотрим [10].

Цель исследования

Цель исследования рекомендуется формулировать максимально кратко и ясно. Несмотря на то, что цель исследования иногда сформулирована достаточно просто, это не обязательно означает, что его структура может быть простой.

При формулировании цели необходимо дать точное определение каждого используемого понятия.

По возможности должны использоваться объективные методы измерения и стандартного представления типов данных [30].

Планирование исследования

Этап планирования исследования является оптимальным для начала совместной работы со специалистом в области математической обработки данных.

Планирование исследования в общем виде можно разбить на 2 этапа:

1. Определение типа исследования, обеспечение достоверности и обобщаемости результатов планируемого исследования, применение способов сведения к минимуму систематических и случайных ошибок.

2. Определение объемов выборок.

Типы исследований

Для того чтобы выбрать тип исследования, необходимо представить себе весь спектр существующих типов структуры исследования.

Классификация исследований может проводиться по нескольким принципам.

По цели исследования:

- выдвигающие гипотезу (относительно менее высокая научная ценность исследования);
- проверяющие гипотезу (относительно более высокая научная ценность исследования).

По временным параметрам:

- однократное обследование объектов исследования;
- многократное обследование объектов исследования.

По соотношению времени сбора данных и формирования выборок:

- проспективное (изучаемые группы формируют до сбора данных);
- ретроспективное (изучаемые группы формируют после сбора данных).

Объемы выборок

Значение объема выборки в 30 элементов как пограничное между малыми и большими выборками состоит в том, что при $n \leq 30$

– распределение очень тесно аппроксимируется («описывается») нормальным и поэтому вариациями вследствие объема выборки можно пренебречь [23].

Для объема выборки, меньшего, чем 30 элементов, доверительные границы шире и вероятная ошибка больше, чем для выборок с объемом больше 30. *При уменьшении объема выборки доверительные границы расширяются и вероятная ошибка возрастает.*

В конце концов, для очень малых выборок доверительные границы столь широки, а вероятные ошибки столь велики, что практическая ценность любого статистического вывода незначительна.

Типы экспериментальных данных

В ходе исследований получают разнообразные данные. При

статистическом анализе их называют признаками или переменными. Именно типом данных и организацией исследования часто определяется выбор статистического критерия.

(Подробно о типах данных можно посмотреть в [29, 24, 12, 30].)

Здесь же хотелось бы отметить, что практически в каждой задаче (особенно при многомерных исследованиях) среди переменных присутствуют переменные, измеряемые в различных шкалах измерений, или так называемые «*качественные» факторы* (марка материала, наименование оборудования, шкала твердости материалов, кодировка цвета сланцев и т. д.). Часто можно встретить рекомендации по особым процедурам их обработки. Доказано, что применения специальных методов для этого не требуется [26]. Построение, например, регрессионной модели с «качественными» факторами математически не отличается от таковой с наличием только количественных переменных.

Статистический анализ данных

Подчеркнем сразу, что результат анализа данных зависит, в первую очередь, не от правильности статистических вычислений, а от того, насколько высоко методологическое качество исследования (т. е. насколько правильно оно организовано). Статистический анализ, каким бы он ни был сложным, не способен компенсировать дефекты организации эксперимента или наблюдения. Однако некорректное использование самых простых статистических методов для анализа доброкачественных данных часто приводит к ложным заключениям и выводам.

Для этого ниже будут кратко представлены общие описания некоторых статистических методов и принципы их применения. Настоящий «краткий обзор» касается только наиболее распространенных статистических процедур.

Правильное описание методов статистического анализа должно содержать указание на применяемые статистические критерии и их конкретные варианты; если применялся не общепринятый критерий или метод расчета, должно быть объяснение, почему использован именно он. Поскольку результат применения многих критериев может зависеть от используемого алгоритма вычислений, в отчетах по конкретным исследованиям должны быть указаны программа или пакет статистических программ, с помо-

щью которых проводились вычисления. Большинство коммерческих пакетов статистических программ имеют отличные характеристики, и нет заметной разницы в том, какой из них применяется для большинства задач.

Описательная статистика

Оценки описательной статистики можно посмотреть в [29, 24, 47, 33, 12].

Статистические выводы

Наиболее распространенный тип статистического вывода – доказательство того, что найденные различия между выборками неслучайны и, следовательно, отражают действительные различия между генеральными совокупностями.

Различие групп означает их неравенство по некоторому показателю. Чтобы доказать этот вывод, применяют формальный логический прием. Сначала предполагают, что группы неразличимы (нулевая гипотеза), затем доказывают, что эта гипотеза может быть отвергнута без совершения большой ошибки. Отрицание нулевой гипотезы и означает неравенство групп, т. е. принятие альтернативной гипотезы, которая состоит в том, что различия между группами не случайны, а отражают истинные различия между совокупностями.

Для проверки нулевой гипотезы используют разнообразные статистические критерии. Никакой статистический критерий не дает абсолютной уверенности в различии или в идентичности групп. Напротив, все статистические критерии позволяют утверждать что-либо лишь с некоторой вероятностью ошибки.

Чаще всего исследователя интересует различие, и соответственно речь идет об отклонении нулевой гипотезы. Поэтому обычно оценивают именно риск принятия ошибочного решения о том, что различие существует (ошибка первого рода, или α – ошибка).

Для риска и устанавливается пороговая величина вероятности ошибки ***p*** на уровне, традиционно равном 0,05 или 0,01.

Статистические критерии и методы

Выбор того или иного статистического критерия для проверки нулевой гипотезы диктуется характером данных, полученных в ходе исследования, его организацией и вопросом, поставленным для исследования.

Критерии, которые могут применяться к различного типа данным, а в случае количественных данных используются независимо от формы распределения, называют непараметрическими критериями [19, 12]. Столкнувшись с применением неизвестного Вам вида статистического анализа, обратите внимание на его изложение. Если оно понятно, Вы обогатите свои познания, в противном случае у Вас имеется основание для скептического отношения и к представленным результатам. Необходимо предупредить, что использование редкого или нового метода при неясном его изложении или с претензией на нечто новое в статистическом анализе нередко сочетается с ошибочным применением фундаментальных принципов [10]. В приведенной ниже таблице перечислены часто используемые критерии и методы математического анализа данных.

Таблица

Некоторые часто используемые статистические критерии и методы

Параметрический критерий	**Непараметрический аналог критерия**	**Назначение критерия**
Две независимые выборки (непарный) *t*-критерий	U-тест Манна–Уитни	Сравнивает две независимые выборки
Две зависимые выборки (парный *t*-критерий)	Критерий Уилкоксона	Сравнивает наблюдения за одними и теми же объектами (проверяет гипотезу, что среднее различие между двумя измерениями равно нулю)
Однофакторный дисперсионный анализ.	Дисперсионный анализ рангов	Обобщение парного t-критерия или критерия Уилкоксона, где 3 выборки наблюдений или более делаются на одном объекте
Двухфакторный дисперсионный анализ с повторениями и без	Двусторонний анализ расхождения по рангу	То же, что и выше, но тестирует влияние (и взаимодействие) двух различных факторов

Параметрический аналог отсутствует	χ^2 - критерий	Проверяет нулевую гипотезу, что пропорции переменных, изменяющихся на двух уровнях (или более), независимы от второй переменной
Коэффициент корреляции Пирсона	Коэффициент корреляции рангов Спирмена	Оценивает силу линейной взаимозависимости между двумя переменными
Регрессия методом наименьших квадратов	Непараметрический аналог отсутствует	Описывает численную связь между двумя переменными, позволяя предсказывать одну переменную через значение другой
Множественная регрессия методом наименьших квадратов	Непараметрический аналог отсутствует	Описывает численную связь между зависимой переменной и несколькими независимыми, предсказывающими переменными (предикторами)

Подробно с критериями и методами, приведенными в таблице, можно познакомиться по работам [29, 24, 7, 33, 12, 30, 44, 16, 17, 14, 26].

Многомерные методы (наиболее часто используемые)

Остановимся на методах анализа многомерных данных, в которых каждый наблюдаемый объект характеризуется множеством переменных. Многомерные методы позволяют одновременно изучать изменение набора характеристик (множественный регрессионный анализ также относится к этим методам).

Многомерные методы являются необычайно мощными, так как они позволяют исследователю работать с большим числом переменных, чем он может осознать сам. Однако они сложны как с теоретической, так и с методологической точек зрения. Статистические критерии и процедуры большей части этих методов разработаны лишь при очень сильных ограничениях. Вид этих критериев и их поведение при более слабых допущениях (которые обычно используются при решении большинства реальных задач)

плохо изучены. В самом деле, некоторые из рассмотренных ниже процедур совсем не имеют теоретического обоснования, а критерии проверки соответствующих гипотез для них еще не созданы. Тем не менее, эти методы кажутся наиболее перспективными и многообещающими в исследованиях. В большинстве задач приходится иметь дело со сложными комбинациями действующих факторов, которые не удается выделить в чистом виде и изучить изолированно. Зачастую бывает трудно принять обоснованное правильное решение относительно какой-либо из переменных. В этом случае лучший способ решения задачи состоит в ее многомерной реализации.

Наиболее часто из многомерных методов используют кластерный анализ, метод главных компонент, дискриминантный анализ и т. д. Эти методы и принцип факторного анализа описаны в данном пособии.

В заключение отметим, что представленный здесь обзор, прежде всего, касается этапа «Анализ данных» (см. схему). В связи с этим мы не будем рассматривать этапы научного исследования, а начнем с «Интерпретации результатов», показанных на схеме.

Работу по этим этапам можно «посмотреть» в литературе [29, 24, 7, 47, 19, 23, 12, 48, 30, 42, 49, 25, 44, 5, 20, 1, 46, 10].

1. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ предназначен для классификации наблюдений в более или менее однородные группы. Под кластером обычно понимают группу объектов, обладающую свойством плотности (плотность объектов внутри кластера выше, чем вне его), отделимостью от других кластеров, формой (например, кластер может иметь очертания гиперсферы или эллипсоида), размером.

Хотя имеются альтернативные классификации классификаций, большинство из них может быть сгруппировано в четыре общих типа [12].

1. Методы разделения на части, применяемые к самим многомерным наблюдениям или к проекциям этих наблюдений на плоскости более низкой размерности. В их основе лежит правило объединения областей в пространстве, определенном m-переменными, которые бедно представлены наблюдениями, и отделения от них тех областей, которые плотно представлены наблюдениями. Математически «разбиения» помещаются в разреженных районах, подразделяя пространство в дискретные классы.

2. Произвольные исходные методы основываются на сходстве между наблюдениями и множеством произвольных исходных точек. Если n наблюдений подразделяются на k групп, то необходимо вычислить асимметрию – матрицу порядка $n \times k$ сходства между пробами и k произвольными точками, которые играют роль исходных центроидов (один из вариантов – центры тяжести групп по средним значениям переменных) групп. Самое близкое наблюдение или наиболее сходное с начальной точкой комбинируется с нею и образует кластер. Наблюдения последовательно добавляются к ближайшему кластеру, после чего центроид для расширен-

ного кластера вычисляется заново.

3. Процедуры взаимного сходства соединяют вместе наблюдения, которые имеют общее сходство с другими наблюдениями. Сначала вычисляется матрица сходства порядка $n \times n$ между всеми парами наблюдений. Затем итерационным методом оценивается сходство между столбцами этой матрицы. Столбцы, представляющие члены одиночного кластера, имеют высокие внутренние корреляции, в то время как их корреляции с другими элементами значительно ниже.

4. Иерархическая кластеризация состоит в объединении наиболее сходных наблюдений, затем последовательно к ним присоединяются следующие наиболее близкие наблюдения. Сначала вычисляется матрица сходства порядка $n \times n$ между всеми парами наблюдений. Пары, имеющие наивысшее сходство, затем выделяются, и матрица пересчитывается. Это делается усреднением коэффициентов сходства, которые имеют с другими наблюдениями комбинированные наблюдения. Этот процесс итерационным путем повторяется до тех пор, пока матрица сходства будет приведена к матрице 2×2. Уровни сходства, при которых наблюдения устраняются, используются для построения дендрограммы (древовидного дерева).

Предположим, что мы располагаем некоторым множеством объектов, которые желательно иерархически расклассифицировать. На каждом объекте мы проводим ряд измерений, которые составляют наше множество данных. Если у нас n объектов и измерено m характеристик, то множество данных образует матрицу порядка $n \times m$. Далее между каждой парой объектов вычисляется некоторая мера сходства или подобия. Коэффициенты сходства могут быть разными, как, например, коэффициент корреляции или стандартизованное m-мерное евклидово расстояние d_{ij}. Последнее вычисляется по формуле

$$d_{ij} = \sqrt{\frac{\sum_{i=1}^{m} (X_{ik} - X_{jk})^2}{m}} \qquad (1.1)$$

где X_{ik} – значение k-й переменной на i-м объекте и X_{jk} – значение k-й переменной на j-м объекте. Естественно ожидать, что малое значение этого расстояния указывает на то, что объекты подобны

или близки друг другу, в то время как большое значение указывает на отсутствие подобия. Обычно матрица исходных данных до вычисления расстояний подвергается стандартизации [33, 19, 12]. Это позволяет учитывать каждую переменную с одинаковым весом.

Множество мер сходства между всеми парами объектов можно представить в виде симметричной матрицы порядка $n \times n$. Для вычисления элементов этой матрицы транспонировать матрицу исходных данных, порядок которой $n \times m$, в матрицу порядка $m \times n$. В результате получим матрицу сходства порядка $m \times m$ между переменными (в отличие от корреляционной матрицы сходства между наблюдениями порядка $n \times n$). Элемент c_{ij} матрицы дает характеристику сходства между i-м и j-м объектами. Следующая задача – получение иерархической группировки объектов, при которой объекты с наивысшими коэффициентами сходства размещаются вместе. Затем группы объектов соединяются в другие группы, с которыми они наиболее тесно связаны, и так, продолжается до тех пор, пока не будет получена полная классификация объектов. Существует много методов анализа групп, рассмотрение всех разновидностей этих методов выходит за рамки учебного пособия. Здесь рассмотрим метод, называемый методом взвешенной парной группировки с арифметическими средними [12].

В табл. 1.1 приведена матрица коэффициентов корреляции между шестью объектами, названными *A, B, ..., F*. Объекты — это пациенты, а переменные – характеристики, включающие показатели различных исследований. В этом примере в качестве меры сходства взят коэффициент корреляции.

Таблица 1.1

Матрица коэффициентов корреляции для шести объектов

	A	**B**	**C**	**D**	**E**	**F**
A	1,00	0,57	0,29	-0,59	-0,59	-0,59
B		1,00	0,29	-0,59	-0,59	-0,59
C			1,00	-0,59	-0,59	-0,59
D				1,00	0,66	0,41
E					1,00	0,41
F						1,00

Первый шаг анализа групп методом попарного объединения состоит в нахождении в корреляционной матрице наибольших

коэффициентов корреляции с целью выделения центров групп. Объекты *A* и *B* (0,57) образуют пару с высокой мерой сходства, так как *A* наиболее близок к *B* и *B* наиболее близок к *A*. Однако *C* и *B* (0,29*)* не образуют пары с высокой мерой сходства, так как хотя *C* близок к *B, B* ближе к *A*, чем к С. Для выделения пары с высокой мерой сходства коэффициенты c_{ij} должны иметь наибольшие значения в соответствующих столбцах.

Пары с наивысшими мерами сходства изображены на рис. 1.1, а. Объект *A* связан с *B* на уровне 0,57, указывающем меру их взаимного сходства. Таким же образом связаны *D* и *E* (0,66). Это первый шаг в построении дендрограммы, или «дерева», позволяющего наглядно изобразить результаты разбиения.

Далее матрицу сходства вычисляют снова, причем сгруппированные элементы при этом считаются одним элементом. Существует несколько методов выполнения этой процедуры. Здесь используется наиболее простой из них, состоящий в том, что новые коэффициенты корреляции между всеми группами и объектами, не включенными в группы, вычисляются заново с помощью простого усреднения.

Например, новый коэффициент корреляции между группой *AB* и объектом *C* равен сумме коэффициентов корреляции элементов, входящих как в *AB,* так и в *C*, деленной на 2. В табл. 1.2 приведены результаты этих вычислений. Процедура образования групп снова повторяется: находим пары с сильными связями и объединяем.

На этом этапе объект С присоединяется к группе *AB,* а объект *F* присоединяется к группе *DE* (рис. 1.1, б). Процесс продолжается до тех пор, пока все группы не объединятся вместе. Окончательная матрица сходства, как показано в табл. 1.3, будет иметь порядок 2×2 и соответствовать двум последним группам. Очевидно, что группа *ABC* имеет с группой *DEF* коэффициент сходства — 0,59. На этом построение дендрограммы заканчивается (рис. 1.1, в).

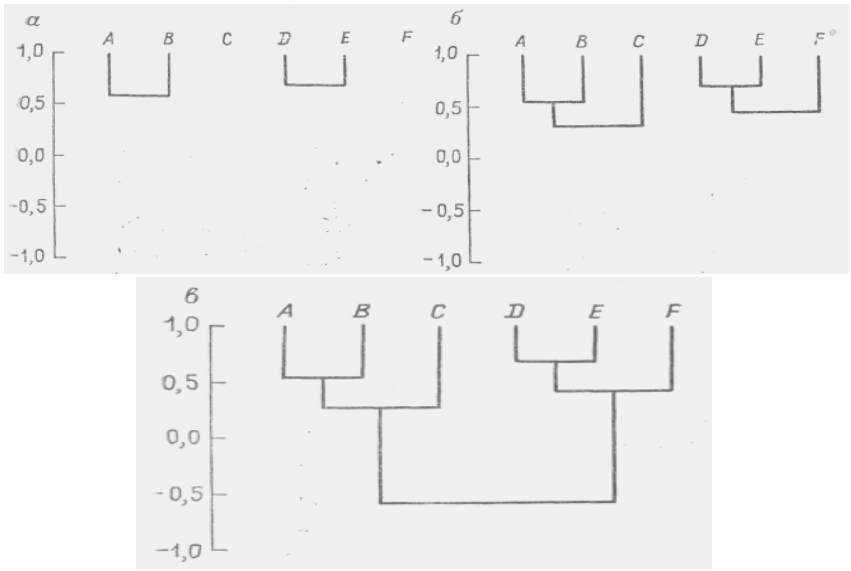


Рис 1.1. а – исходные группы дендрограммы; б – построение групп для остальных объектов; в – окончание построения дендрограммы; две группы связываются между собой

Таблица 1.2

Матрица коэффициентов корреляции между двумя усредненными группами и двумя пациентами

	AB	**C**	**DE**	**F**
AB	1,00	0,29	-0,70	-0,55
C		1,00	-0,59	-0,52
DE			1,00	0,41
F				1,00

Таблица 1.3

Матрица усредненных коэффициентов корреляции между двумя окончательными группами

	ABC	**DEF**
ABC	1,00	-0,59
DEF		1,00

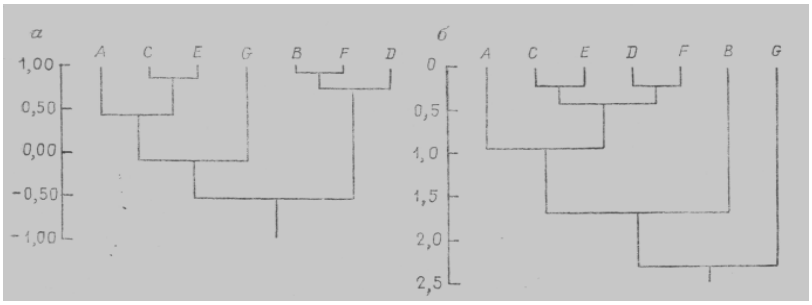


Рис. 1.2. а – дендрограмма, построенная по методу группового объединения, основанного на усреднении коэффициентов корреляции. Исходная матрица приведена в табл. 1.4; б – дендрограмма, построенная тем же методом, но основанная на расстоянии

Наиболее существенные черты этого метода анализа групп заключаются в следующем:

1) коэффициент корреляции используется в качестве меры сходства;

2) объединение в группы начинается с объектов, имеющих наиболее высокие значения коэффициентов корреляции, характеризующих сходство;

3) два объекта можно объединить только в том случае, если они имеют наивысшее значение коэффициента корреляции друг с другом;

4) после того как два объекта объединены в группу, их коэффициенты корреляции со всеми другими объектами усредняются.

Введение иных мер сходства приводит к очевидным модификациям этой схемы. Хотя меры могут быть разными, широко используются только две из них: коэффициент корреляции и расстояние. Если провести стандартизацию исходных данных до вычисления коэффициента сходства, то коэффициент корреляции и расстояние можно непосредственно преобразовать друг в друга. Вообще дендрограммы, построенные на основании этих двух мер, подобны (коэффициент корреляции и расстояние). Однако в отличие от коэффициента корреляции расстояние не обязательно принимает значение в пределах ±1, и поэтому оно может привести к более наглядным дендрограммам в тех случаях, когда несколько объектов сильно отличаются от других. В табл. 1.4 приведены как расстояния, так и коэффициенты корреляции для семи объектов. В качестве переменных выбраны некоторые характеристики.

Таблица 1.4

Меры сходства между семью объектами (над диагональю в скобках указаны расстояния, под диагональю – коэффициенты корреляции)

	A	**B**	**C**	**D**	**E**	**F**	**G**
A		(2,15)	(0,70)	(1,07)	(0,85)	(1,16)	(1,56)
B	-0,93		(1,53)	(1,14)	(1,88)	(1,01)	(2,83)
C	0,59	-0,44		(0,43)	(0,21)	(0,55)	(1,86)
D	-0,55	0,67	0,31		(0,29)	(0,22)	(2,04)
E	0,26	0,02	0,85	0,63		(0,41)	(2,02)
F	-0,79	0,94	-0,20	0,80	0,30		(2,05)
G	0,37	-0,64	-0,38	-0,90	-0,79	-0,82	

Дендрограммы, построенные для каждой матрицы сходства изображены на рис. 1.2. Хотя общие черты группирования очевидны, все же можно отметить два существенных различия. Наиболее очевидными из них являются замена *B* одной из центральных групп на *D* и перемещение *B* в более дальнюю позицию в иерархической структуре. Полезно исследовать причины этого изменения.

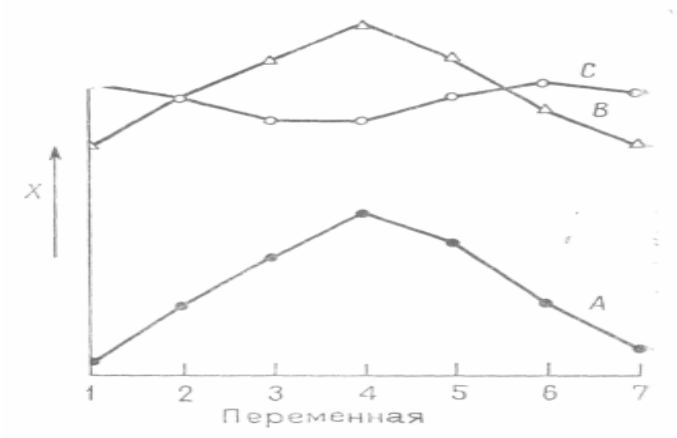


Рис.1.3. Графики переменных, измеренных на трех объектах: кривые *A и B* сильно коррелированы, но разделены большим расстоянием. Кривые *B* и *C* отрицательно коррелированы, но «близки» в смысле расстояния

Предположим, что измерено семь переменных на каждом из трех объектов. Ими могут быть, например, определенные химические анализы у трех пациентов. Если нанести каждое измерение на график так, как это показано на рис. 1.3, можно убедиться в том, что соотношения между переменными в двух объектах сходны. Им соответствуют более или менее параллельные графики *A* и *B* на диаграмме. У третьего графика другой вид, но он значительно ближе к графическому представлению множества измерений, соответствующего одному из двух других объектов. В этом примере *A* и *B* сильно коррелированы, т. е. имеют высокие линейные связи, но зато расстояние между *B* и *C* минимально. Если бы в качестве переменных были выбраны исходные данные пациентов (рост, вес, объемы различные и т. п.), то это привело бы к выводу, что *A* и *B* имеют близкую форму, а *B* и *C* – сходные размеры. Если бы в качестве переменных были выбраны содержания тяжелых элементов в пробах, то можно сделать вывод, что образцы *A* и *B* аналогичны по составу, но *A* обладает пониженными содержаниями по сравнению с *B*. Содержания элементов в *B* и *C* близки, но их отношения различны.

Необходимо пояснить, что коэффициент корреляции указывает на наибольшее сходство в тех случаях, когда он имеет высокое положительное значение, в то время как расстояние указывает на наибольшее сходство в тех случаях, когда оно наименьшее. Поэтому коэффициент корреляции выявляет наличие связи при его высоких значениях, а расстояние – при низких.

Критерий объединения двух объектов в группу требует, чтобы оба они имели наибольшую корреляцию относительно друг друга. Возможны также и другие критерии. Так, известен простой метод образования групп, называемый простым объединением и основанный на использовании наивысшего коэффициента сходства между некоторым фиксированным объектом и любым объектом группы. Результаты анализа групп этим методом по корреляционной матрице, приведенной в табл. 1.4, изображены на рис. 1.4. Так как объекты вводятся в группу на основании наивысшего значения коэффициента корреляции с любым объектом, уже принадлежащим группе, то теснота связи в этом случае оказывается более высокой, чем в методах группового объединения. При этом кроме сжатия дендрограммы, возникают и другие отличия. На-

пример, группа *CE* прямо соединена с группой *BFD* в силу наличия высокой корреляции между *E* и *D.* Если корреляцию с *C* и *E* усреднить, то наивысшей будет корреляция между *CE* и *A*.

Простое объединение прямо приводит к окончательной характеристике, среднему арифметическому мер сходства объектов, которые уже определены по группам. При использовании этого метода образования групп никакого усреднения совсем не делается. Методы, проиллюстрированные на рис. 1.2, *а* и *б* и в предыдущем примере (см. рис 1.1), называются взвешиванием, хотя на самом деле их следовало бы назвать методами равного взвешивания. На рис. 1.2, *а C* и *E* соединены в начале образования групп. Корреляции новой группы *CE* находятся комбинированием строк и столбцов *C* и *E* и делением каждого из элемента на 2. Далее в группу вводится объект *A*, и коэффициент корреляции новой группы *ACE* находится комбинированием строк и столбцов группы *CE* со строками и столбцами *A* и делением их на 2. Иными словами, *CE* считается единственным объектом, в то время как на самом деле он состоит из двух объектов. Новый объект *A* имеет двойное влияние на коэффициент корреляции группы *ACE,* так же, как *E* или *C*. Объекты, присоединенные к группе позже, больше влияют на матрицу сходства, чем объекты, присоединенные ранее.

Методы усреднения без учета весов стремятся избежать этого, приписывая в процессе усреднения каждой группе веса, пропорциональные числу объектов в ней. Например, образовав группу *CE,* можно присоединить к ней объект *A* с целью образования новой группы *ACE.* Однако меры сходства этой новой группы находятся в результате суммирования коэффициентов корреляции *A* со всеми элементами, исключая *C* и *E,* коэффициентов корреляции C со всеми элементами, исключая *A* и *E*, а также коэффициентов корреляции *E* со всеми элементами, исключая *A* и *C*. Таким образом, нужно сложить коэффициенты корреляции всех исходных элементов в группе, а затем каждую сумму разделить на 3. Эта процедура позволяет каждому объекту группы одинаково влиять на характеристики сходства всей группы. Такой метод по сравнению с обычными методами взвешивания имеет противоположное свойство: объекты, введенные в группу позже, почти не оказывают влияния на меры сходства внутри нее. На рис. 1.5 по

данным табл. 1.4 приведена дендрограмма, построенная на основе метода не взвешенного усреднения.

Можно проиллюстрировать эффект четырех различных стратегий установления связей, рассматривая очень простую задачу кластеризации, в которой на каждом объекте измерены только две переменные. Тогда все соотношения между объектами могут быть изображены на плоскости, как это представлено на рис. 1.6. Расстояния между объектами на диаграмме попросту пропорциональны мере расхождения между ними.

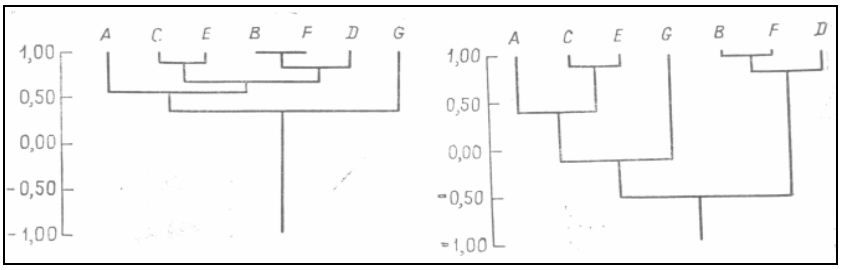


Рис. 1.4. Дендрограмма корреляционной матрицы, приведенной в табл. 1.4. Группы построены по методу прямой связи

Рис. 1.5. Дендрограмма корреляционной матрицы, приведенной в табл. 1.4. Группы построены на основании не взвешенного усреднения

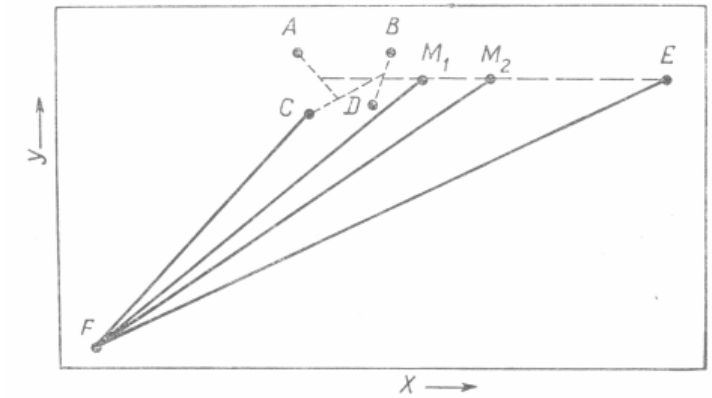


Рис. 1.6. Диаграмма, которая показывает, как объекты, характеризуемые двумя переменными *X* и У, входят в группу:

Объекты *A, B, C* и *D* образуют группу. Объект *E* присоединен к этой группе, а объект *F* является кандидатом на присоединение на следующем шаге итерационного процесса. M_1 – центроид объектов от *A* до *E*. M_2 – среднее объекта *E* и последнего среднего объектов от *A* до *D*

Четыре объекта, от *A* до *D*, образуют связанный пучок. Пунктирные линии указывают порядок, в котором эти четыре объекта были соединены вместе. Несколько менее сходный объект *E* также был присоединен к этому пучку. Шестой объект, обозначенный *F*, теперь рассматривается в качестве кандидата на возможное включение в расширенный пучок. Точка M_1 является центроидом точек от *A* до *E*, а M_2 – средняя для объекта *F* и среднего предыдущего пучка.

Используя единственный критерий связывания, объект *F* присоединяют к этому пучку, если расстояние *CF* меньше, чем расстояние до любого другого объекта в любом другом пучке. При невзвешенном усреднении или центроидной связи объект *F* будет присоединен к пучку, если расстояние *EF* меньше расстояния до центроида в любой другой группе. Во взвешенной парагрупповой или усредненной процедуре связывания объект будет присоединен, если расстояние M_2F меньше, чем расстояние до среднего в любом другом пучке. (Заметим, что точка находится посередине между средним пучка *ABCD* и объектом *E*, который участвовал в первом цикле.) Наконец, при полном связывании объект *F* присоединяется к пучку, если расстояние *EF* меньше, чем расстояние до большинства точек в любом другом пучке.

Столкнувшись с таким множеством методов, каждый из которых дает несколько отличающийся от других результат, исследователь вправе спросить о том, какой из них лучше. К сожалению, на этот важный вопрос нет четкого ответа. Опыт показывает, что методы взвешенного группового объединения обычно дают результаты лучше, чем любой из методов простого объединения или не взвешенного усреднения. В анализе групп матрицы расстояний обычно используются с большим успехом, чем матрицы коэффициентов корреляции. По-видимому, матрицы расстояний также менее чувствительны к замене метода при анализе групп. Большинство исследователей, использующих методы анализа групп, применяют различные меры сходства и процедуры по-

строения групп, а затем выбирают те из них, которые дают наиболее удовлетворительные результаты для их данных. Тщательный предварительный анализ может определить выбор процедуры кластеризации. Вероятно, что наиболее широко применяемый метод — это процедура k-средних [12, 8, 22]. Здесь *k* точек, характеризуемых *m* переменными, объявляются (либо пользователем, либо программой) исходными «центроидами» групп. Вычисляется матрица сходства между этими *k* «центроидами» и *n* наблюдениями, и затем ближайшие или наиболее сходные наблюдения объединяются в группы с этими «центроидами». Затем вычисляются новые центроиды, и процесс многократно повторяется в точности как иерархическая процедура. В принципе этот центроид по мере роста группы быстро сдвигается в направлении истинного центроида, так как влияние истинных наблюдений оказывает все более существенное влияние на произвольный выбор исходной точки. Недостаток метода k-средних состоит в том, что при неудачном выборе произвольных начальных точек может получиться неоптимальная кластеризация, что приведет к преждевременному сдвигу центроидов и к ошибке в обнаружении аномальных кластеров.

Польза кластерного анализа состоит в том, что он обеспечивает относительно простой и прямой путь классификации объектов и позволяет представить результаты в удобном для понимания виде.

1.1. Виды группирования объектов в программе *STATISTICA*

Как уже отмечалось выше, довольно сложно определить выбор метрики сходства и способ объединения объектов в кластеры. Наша задача существенно упрощается, так как мы ограничимся демонстрацией некоторых методов кластеризации и правилами объединения, предоставленные программой *STATISTICA*. Выбор способов объединения и метрик сходства будет зависеть от поставленной задачи и опыта исследователя (выбор этот всегда достаточно сложен).

1.1.1. Методы кластеризации

Рассмотрим, реализованные в пакете *STATISTICA* следующие методы кластеризации [8]:

joining (tree clustering) – агломеративный (агломерат – скопление) метод группировки с построением дендрограммы (иерархического объединения кластеров).

two-way joining – метод, в котором группируются наблюдения и переменные одновременно (в пособии этот метод не будет рассмотрен).

k-средних — k-means clastering – итеративный метод группировки.

В *STATISTICA* можно выбрать следующие правила иерархического объединения кластеров:

Single linkage – метод одиночной связи,

Complete linkage – метод полной связи,

Unweighted pair group average – не взвешенный метод «средней связи»,

Weighted pair group average – взвешенный метод «средней связи»,

Unweighted pair group centroid – не взвешенный центроидный метод,

Weighted centroid pair group (median) – взвешенный центроидный метод,

Ward method – метод Уорда (Варда).

Данные алгоритмы различаются правилами объединения объектов в кластеры.

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

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

В методе средней связи (Unweighted pair group average) мера сходства между кандидатом и членами кластера усредняется, например, берется просто среднее арифметическое мер сходства.

Взвешенный метод «средней связи» (Weighted pair group average) идентичен не взвешенному методу средней точки, за исключением того, что в вычислениях учитывается размер групп (количество объектов в кластере). Этот метод предпочтительнее использовать, когда предполагается «большое различие» по количеству объектов в группе.

Не взвешенный центроидный метод (Unweighted pair group centroid) использует так называемые «центры тяжести» соответствующих групп для расчета расстояний между группами и объектами. (Центр тяжести – средние значения признаков в группе).

Взвешенный центроидный метод (Weighted centroid pair group) – аналогичен не взвешенному центроидному методу. Отличие состоит в анализе количества объектов в кластерах при расчетах.

Идея еще одного агломеративного метода – иерархического метода Уорда (Варда) состоит в объединении кластеров, которые дают наименьший вклад в функцию качества [22]:

$$\sum_{K}\sum_{TJ}\sum_{M}(X_{ij}-\overline{X}_{ij})^2,$$

где K – число кластеров, TJ – число объектов в кластере, M – число признаков, i – индекс признака, j – номер объекта в кластере, X_{ij} – значение признака i для объекта j, $\overline{X}_{ij}$ – среднее значение признака i для кластера j.

Метод Уорда приводит к образованию кластеров примерно равных размеров и имеющих форму гиперсфер [8].

1.1.2. Меры сходства, используемые в программе *STATISTICA:*

1. Euclidean distances
2. Sguared Euclidean distances
3. City – block (Manhattan) distances
4. Chebychev distances metric
5. Power:Sum(ABS(x-y)**p)**1/r
6. Percent disagreement
7. 1 – Pearson R.

Euclidean distances и Sguared Euclidean distances (разновидности евклидовой метрики). Евклидово расстояние обычно применяется для переменных, измеренных в одних и тех же единицах измерения для каждого признака или стандартизированных данных [33, 19, 12].

По поводу использования евклидовой метрики существует иное мнение – евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным. Считается, что этот способ его вычисления имеет определенные преимущества (например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться «выбросом»).

$$\text{Расстояние } (x,y) = \{\sum_{i=1}^{n} (x_i - y_i)^2\}^{1/2},$$

$$\text{расстояние } (x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

City–block (Manhattan) distances – манхеттеновская метрика, как правило, применяется для номинальных или качественных переменных /25, 25/.

$$\text{Расстояние } (x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Chebychev distances metric – чебышевская мера расстояния. Эта мера может использоваться в случаях, когда нужно опреде-

лить два объекта «как разные «, если они различаются хотя бы по одному признаку.

$$\text{Расстояние } (x,y) = \text{Максимум } |x_i - y_i|$$

Степенное расстояние – расстояние Минковского. Это расстояние – одна из мер сходств для качественных признаков. В *Справке к программе STATISTICA* написано, что *r* и *p* – определенные значения параметров. Подбор разных значений этих параметров (*r* и *p* =1 до 4) для какого-либо примера дает представление как метрика Минковского «ведет себя».

$$\text{Расстояние } (x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/r}$$

Percent disagreement (процент несогласия) – здесь подсчитывается количество параметров, которые совпадают у объектов. Полученное число делят на общее число параметров и получают меру сходства. Используется для бинарных данных («0 – 1», «да – нет»). Эту меру называют простым коэффициентом совстречаемости [8].

$$\text{Расстояние } (x,y) = (\text{Количество } x_i\ y_i)/\ i$$

1–Pearson R – величина обратная коэффициенту корреляции Пирсона [33, 19].

1.1.3. Метод k-средних — k-means clastering

Данный метод работает непосредственно с объектами, а не с матрицей сходства. В методе **k-средних** объект относится к тому классу, расстояние до которого минимально. Расстояние понимается как евклидово. Как определить расстояние от объекта до совокупности объектов? Оказывается, это можно сделать следующим способом: каждый класс объектов имеет центр тяжести. Расстояние между объектом и классом есть расстояние между объектом и центром тяжести класса.

Принципиально метод **k-средних** «работает» следующим образом:

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

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

3) вычисляются центры тяжести новых кластеров;

4) шаги 2, 3 повторяются, пока не будет найдена стабильная конфигурация (т. е. кластеры перестанут изменяться) или число итераций не превысит заданное. (Несколько подробнее метод средних описан выше.)

1.2. Применение кластерного анализа в программе *STATISTICA*

Мы уже знаем, что термин «кластерный анализ» в действительности включает в себя набор различных алгоритмов классификации. В данном пособии покажем работу некоторых вариантов группирования на основе использования различных правил иерархического объединения кластеров и мер сходств (расстояний).

Для иллюстрации работы кластерного анализа воспользуемся «вкусным» примером о пиве (если пиво можно назвать вкусным) из работы [25]. В этом примере собраны данные по четырем определяющим переменным для 20 популярных сортов немецкого пива (табл. 1.5).

Таблица 1.5

сорт пива	калории	натрий	алкоголь	цена (у.е)
1_C	144	15	4,7	0,43
2_C	151	19	4,9	0,43
3_C	157	15	4,9	0,48
4_C	170	7	5,2	0,78
5_C	152	11	5	0,77
6_C	145	23	4,6	0,28
7_C	175	24	5,5	0,4
8_C	149	27	4,7	0,42
9_C	99	10	4,3	0,43
10_C	113	8	3,7	0,44
11_C	140	18	4,6	0,44
12_C	102	15	4,1	0,46
13_C	135	11	4,2	0,5

14_C	150	19	4,7	0,76
15_C	149	6	5	0,79
16_C	68	15	2,3	0,38
17_C	136	19	4,4	0,43
18_C	144	24	4,9	0,43
19_C	72	6	2,9	0,48
20_C	97	7	4,2	0,47

Приведем примеры кластеризации этих данных в *STATISTICA*. Открытие файла данных проводим стандартным образом:

Файл => Открытие …(адрес нахождения файла с данными о пиве).

Так как «мы имеем дело» с различными единицами измерения признаков, предварительно проводим стандартизацию исходных данных.

В программе *STATISTICA* стандартизация данных проводится следующим образом (один из способов):

выбираем пункт меню **Редактирование** => **Заполнение/ Стандартизация блока** => **Стандартизация столбцов/столбцов).**

В рабочем окне «высвечивается» файл со стандартизированными данными (рис. 1.7).

Для вызова метода группировки с построением дендрограммы выбираем пункт главного меню **Статистика => Многомерные исследующие методы => Групповой анализ => joining (tree clustering). OK.**

На экране монитора высвечивается диалоговое окно, после заполнения раскрывающихся пунктов которого:

variables,

входной файл,

klaster,

правило объединения,

измерения (расстояние) и нажатия кнопки **OK** программа проводит кластирование данных (рис. 1.8). По окончании кластирования на экране отображается диалоговое окно «Результаты соединения» (рис. 1.9).

пиво ст.исх.данные - Sheet1

пиво ст.исх.да
Sheet1
Sheet2

	1 КАЛОРИИ	2 НАТРИЙ	3 АЛКОГОЛЬ	4 ЦЕНА (У.Е)
1_C	0.383376131	0.00759731406	0.34220506	-0.47239217
2_C	0.614723796	0.615382439	0.605439722	-0.47239217
3_C	0.813021794	0.00759731406	0.605439722	-0.134969191
4_C	1.24266746	-1.20797294	1.00029172	1.88956868
5_C	0.647773462	-0.600187811	0.737057053	1.82208408
6_C	0.416425797	1.22316756	0.21058773	-1.48466111
7_C	1.40791579	1.37511385	1.39514371	-0.674845957
8_C	0.548624463	1.83095269	0.34220506	-0.539876766
9_C	-1.10385886	-0.752134092	-0.184264263	-0.47239217
10_C	-0.641163529	-1.05602665	-0.973968249	-0.404907574
11_C	0.251177465	0.463436158	0.21058773	-0.404907574
12_C	-1.00470986	0.00759731406	-0.447498925	-0.269938383
13_C	0.0859291327	-0.600187811	-0.315881594	3.74614760E-16
14_C	0.581674129	0.615382439	0.34220506	1.75459949
15_C	0.548624463	-1.35991922	0.737057053	1.95705328
16_C	-2.12839852	0.00759731406	-2.81661088	-0.809815149
17_C	0.118978799	0.615382439	-0.0526469324	-0.47239217
18_C	0.383376131	1.37511385	0.605439722	-0.47239217
19_C	-1.99619985	-1.35991922	-2.0269069	-0.134969191

Рис. 1.7. Стандартизированные данные табл.1.5

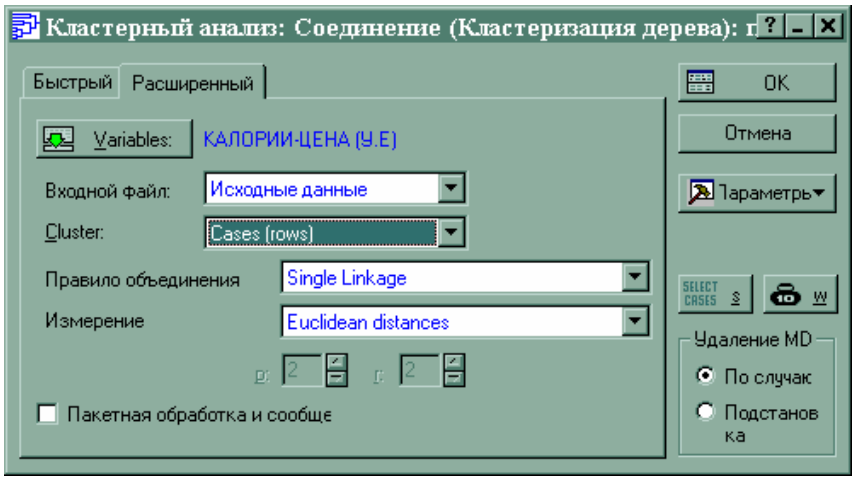


Рис.1.8. Стартовая модель модуля **joining (tree clustering**)

В информационной части окна данные о выборке (переменных – 4, наблюдений – 20), подвергшейся группированию, в строке «Joining of cases» – об объединении наблюдений (в примере – сорта пива), в строке «Missing data» задается способ обработки пропущенных значений в данных (в примере пропущенных дан-

ных нет). Далее показана информация о правиле объединения (в примере: Single linkage – метод одиночной связи), о выбранной мере сходства (в примере – евклидова метрика).

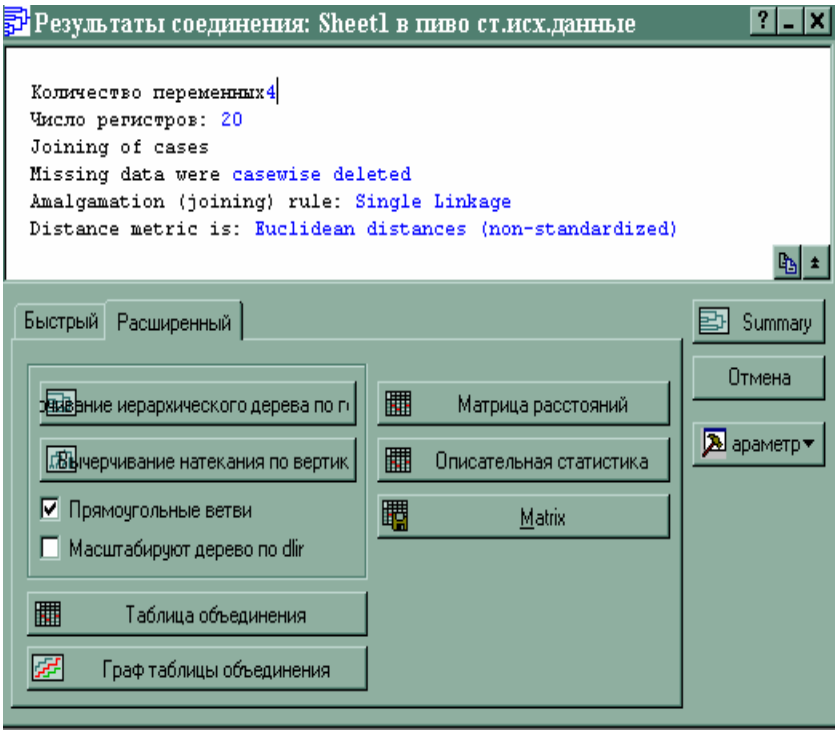


Рис.1.9. Диалоговое окно «Результаты соединения»

Для просмотра дендрограммы (дерева объединения) нажмите одну из кнопок «вычерчивание иерархического дерева по горизонтали (вертикали)», показанных на рис. 1.9. Результат см. на рис. 1.10.

Нажмите кнопку: «Таблица объединения». На экране отобразится таблица объединения наблюдений в кластер (рис. 1.11).

Нажмите кнопку: «Матрица расстояний». На экране отобразится матрица расстояний между наблюдениями (рис. 1.12).

С помощью диалогового окна «Результаты соединения» (рис. 1.9) можно вывести на экран другую информацию, полезную, возможно, для интерпретации результатов кластеризации.

На рис. 1.13 показан результат группирования сортов пива без стандартизации исходных данных. Сравните с результатом, показанным на рис. 1.10.

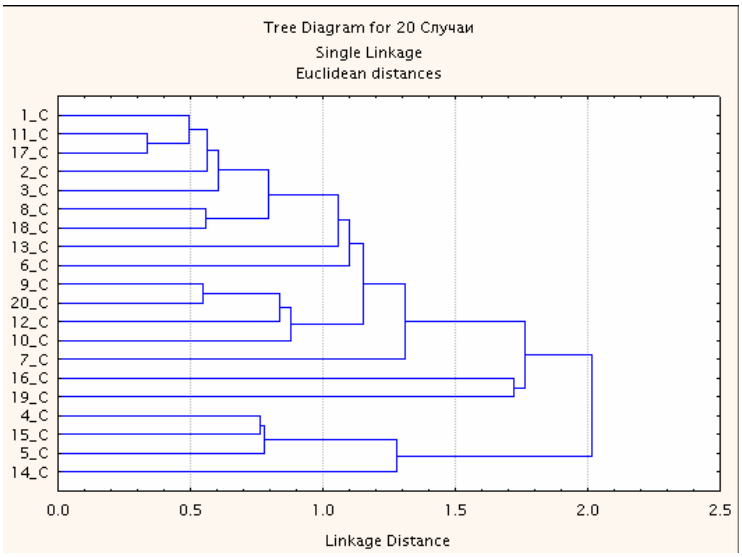


Рис.1.10. Дерево объединения различных сортов пива в кластер методом одиночной связи, в качестве меры сходства использована евклидова метрика. Данные стандартизированы

linkage distance	Amalgamation Schedule (пиво ст.исх.данные) Single Linkage Euclidean distances **Obj. No. 1**	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10
.4971346	C_1	C_11	C_17							
.5498618	C_9	C_20								
.5558262	C_8	C_18								
.5618860	C_1	C_11	C_17	C_2						
.6064175	C_1	C_11	C_17	C_2	C_3					
.7606773	C_4	C_15								
.7779711	C_4	C_15	C_5							
.7941748	C_1	C_11	C_17	C_2	C_3	C_8	C_18			
.8350463	C_9	C_20	C_12							
.8813496	C_9	C_20	C_12	C_10						
1.055514	C_1	C_11	C_17	C_2	C_3	C_8	C_18	C_13		
1.097623	C_1	C_11	C_17	C_2	C_3	C_8	C_18	C_13	C_6	
1.154764	C_1	C_11	C_17	C_2	C_3	C_8	C_18	C_13	C_6	C_9

Рис.1.11. Фрагмент таблицы объединения наблюдений в кластер

Случай	Euclidean distances (пиво ст.исх.данные)																	
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17	C_18
C_1	0.00	0.70	0.61	2.87	2.42	1.59	2.02	1.83	1.75	1.98	0.50	1.61	1.06	2.32	2.82	4.05	0.77	1.39
C_2	0.70	0.00	0.72	3.07	2.60	1.26	1.37	1.25	2.33	2.62	0.56	2.04	1.68	2.24	3.13	4.44	0.82	0.79
C_3	0.61	0.72	0.00	2.43	2.06	1.90	1.77	1.90	2.23	2.41	0.87	2.11	1.33	2.02	2.52	4.56	1.18	1.47
C_4	2.87	3.07	2.43	0.00	0.89	4.31	3.66	4.01	3.56	3.57	3.11	3.65	2.65	2.05	0.76	5.89	3.36	3.63
C_5	2.42	2.60	2.06	0.89	0.00	3.82	3.34	3.41	3.03	3.12	2.55	2.98	2.18	1.28	0.78	5.26	2.77	3.04
C_6	1.59	1.26	1.90	4.31	3.82	0.00	1.75	1.14	2.72	2.98	1.33	2.33	2.43	3.30	4.34	4.19	1.25	1.10
C_7	2.02	1.37	1.77	3.66	3.34	1.75	0.00	1.44	3.66	3.97	1.91	3.35	3.01	2.88	3.95	5.67	2.09	1.31
C_8	1.83	1.25	1.90	4.01	3.41	1.14	1.44	0.00	3.11	3.39	1.41	2.54	2.62	2.60	4.07	4.53	1.35	0.56
C_9	1.75	2.33	2.23	3.56	3.03	2.72	3.66	3.11	0.00	0.97	1.86	0.84	1.30	3.15	3.14	2.94	1.84	2.71
C_10	1.98	2.62	2.41	3.57	3.12	2.98	3.97	3.39	0.97	0.00	2.12	1.25	1.15	3.27	3.16	2.63	2.06	3.08
C_11	0.50	0.56	0.87	3.11	2.55	1.33	1.91	1.41	1.86	2.12	0.00	1.50	1.26	2.19	3.04	3.90	0.34	1.00
C_12	1.61	2.04	2.11	3.65	2.98	2.33	3.35	2.54	0.84	1.25	1.50	0.00	1.28	2.76	3.26	2.68	1.35	2.22
C_13	1.06	1.68	1.33	2.65	2.18	2.43	3.01	2.62	1.30	1.15	1.26	1.28	0.00	2.29	2.39	3.49	1.33	2.25
C_14	2.32	2.24	2.02	2.05	1.28	3.30	2.88	2.60	3.15	3.27	2.19	2.76	2.29	0.00	2.02	4.93	2.31	2.38
C_15	2.82	3.13	2.52	0.76	0.78	4.34	3.95	4.07	3.14	3.16	3.04	3.26	2.39	2.02	0.00	5.41	3.26	3.66
C_16	4.05	4.44	4.56	5.89	5.26	4.19	5.67	4.53	2.94	2.63	3.90	2.68	3.49	4.93	5.41	0.00	3.63	4.47

Рис.1.12. Матрица (не полная) расстояний между наблюдениями

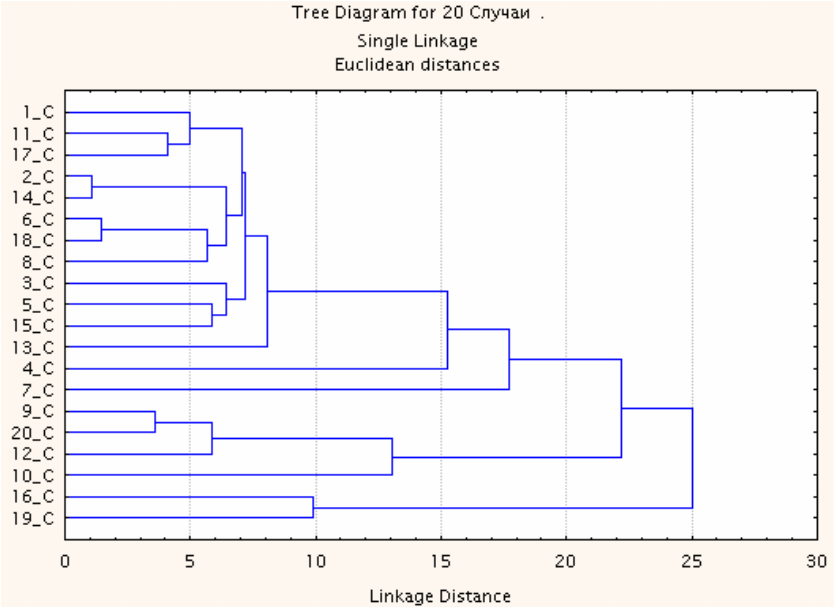


Рис.1.13. Дерево объединения различных сортов пива в кластер методом одиночной связи, в качестве меры сходства использована евклидова метрика. Данные не стандартизированы

1.2.1. Варианты группирования

Рассмотрим несколько вариантов объединений в кластеры (к сожалению, просмотр всех возможных вариантов практически нереален, так как будет занято много «печатного пространства» в пособии). В приведенных вариантах будут использованы стандартизированные данные (см. рис. 1.14 – рис. 1.19) «вкусного» примера о пиве и показаны результаты кластирования с различными правилами иерархического объединения и использования евклидовой метрики.

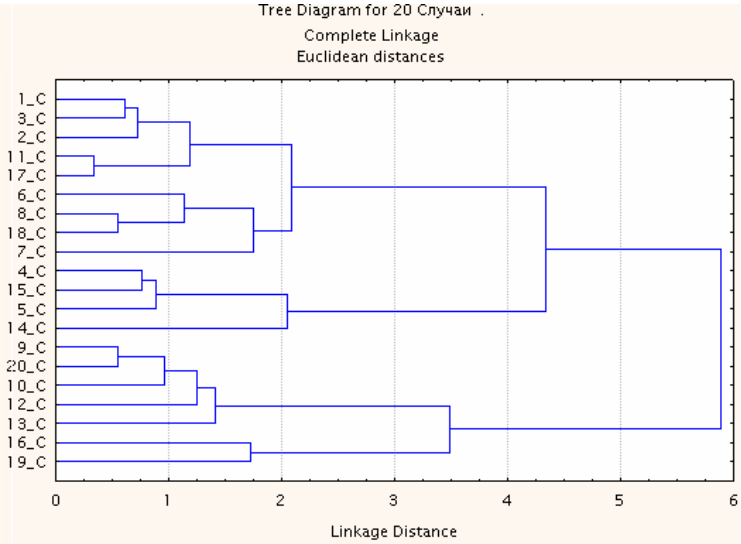


Рис.1.14. Дерево объединения различных сортов пива в кластер методом полной связи, в качестве меры сходства использована евклидова метрика

В завершении демонстрации вариантов объединений в кластеры и построения дендрограмм покажем еще два примера:

1. Кластирования объектов (сортов пива) методом одиночной связи, в качестве меры сходства использована метрика 1 – Pearson R. (R – величина обратная коэффициенту корреляции Пирсона). См. рис. 1.20.

2. Кластирования переменных методом одиночной связи, в качестве меры сходства использована метрика 1 – Pearson R. См. рис. 1.21.

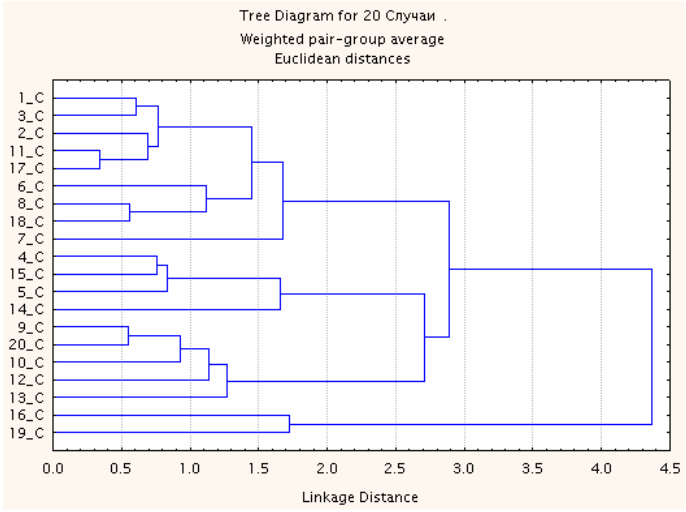


Рис. 1.15. Дерево объединения различных сортов пива в кластер методом не взвешенной «средней связи», в качестве меры сходства использована евклидова метрика

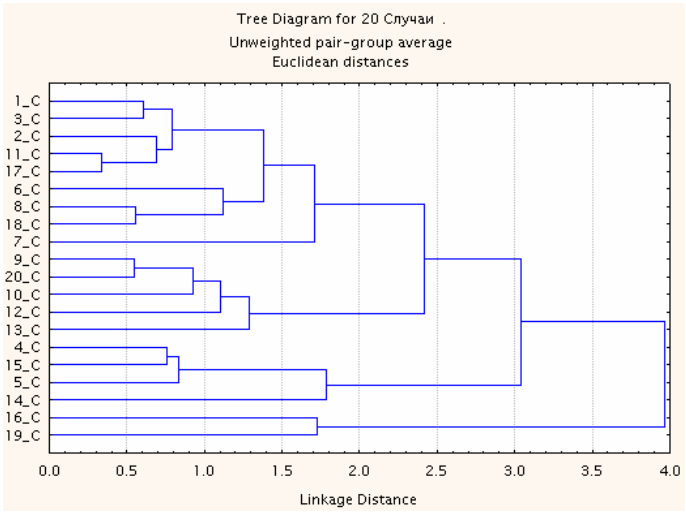


Рис.1.16. Дерево объединения сортов пива в кластер методом взвешенной «средней связи», в качестве меры сходства использована евклидова метрика

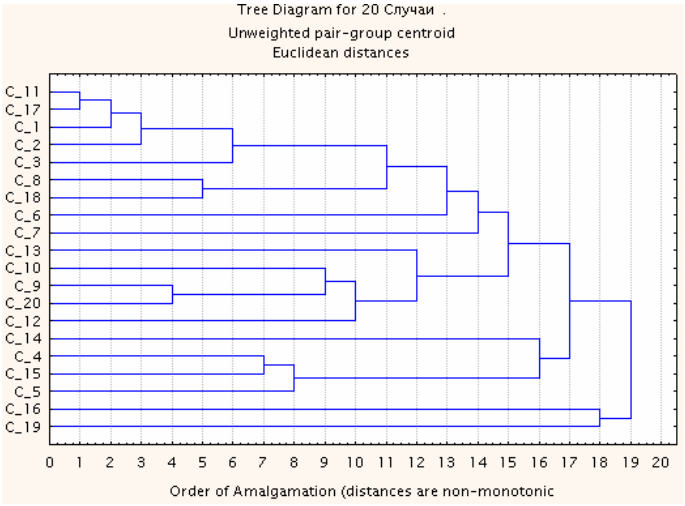


Рис.1.17. Дерево объединения сортов пива в кластер не взвешенным центроидным методом, в качестве меры сходства – евклидова метрика

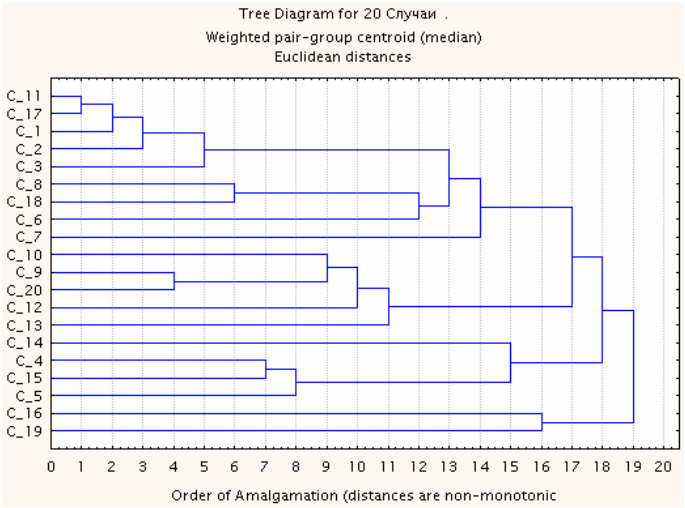


Рис.1.18. Дерево объединения сортов пива в кластер взвешенным центроидным методом, в качестве меры сходства – евклидова метрика

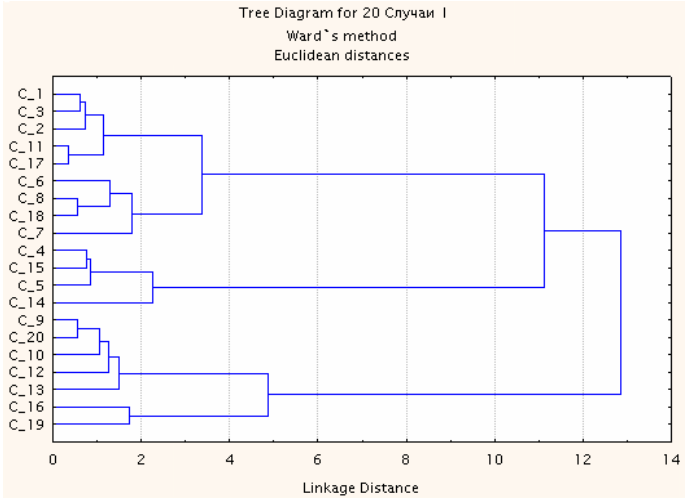


Рис.1.19. Дерево объединения сортов пива в кластер методом Уорда, в качестве меры сходства – евклидова метрика

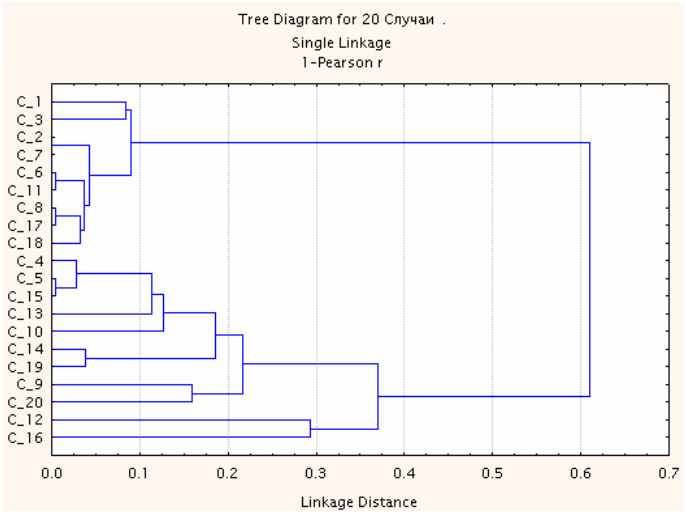


Рис.1.20. Дерево объединения различных сортов пива в кластер методом одиночной связи, в качестве меры сходства использована величина 1 – Pearson R

(Пример, где в качестве меры сходства использована метрика 1 – Pearson R (рис.1.20) нужно рассматривать только как «показа-

тельный», так как обычно коэффициенты корреляции не считаются при небольшом количестве переменных m. В примере m = 4.)

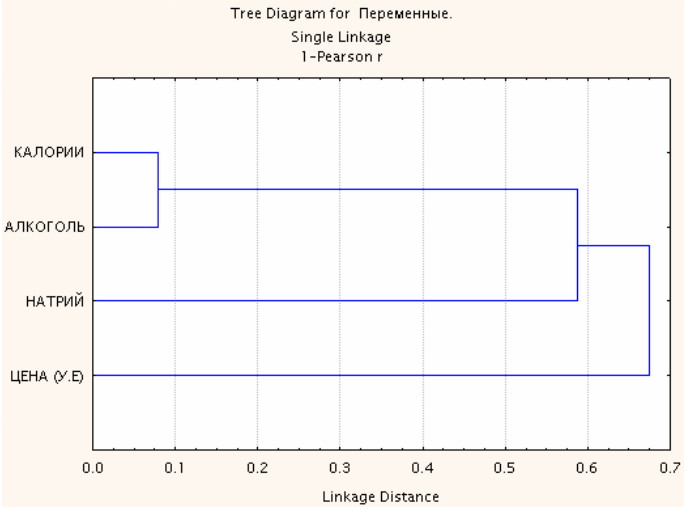


Рис.1.21. Дерево объединения четырех определяющих переменных для различных сортов пива в кластер методом одиночной связи, в качестве меры сходства использована величина 1 – Pearson R

Для получения дерева объединения (рис. 1.21) необходимо на стартовой модели модуля **joining (tree clustering)** в позиции KLASTER выбрать ‘’VARIABLES (Columns)» (рис. 1.8).

1.2.2 Пример использования программы STATISTICA при группировании методом k-средних — k-means clastering

При демонстрации работы метода воспользуемся снова примером с пивом.

Перед группированием надо провести открытие файла с исходными данными и стандартизировать их (см. выше).

Для вызова процедуры группировки методом **k-средних** выбираем пункт главного меню **Статистика => Многомерные исследующие методы =>Групповой анализ => k-means clastering. OK.**

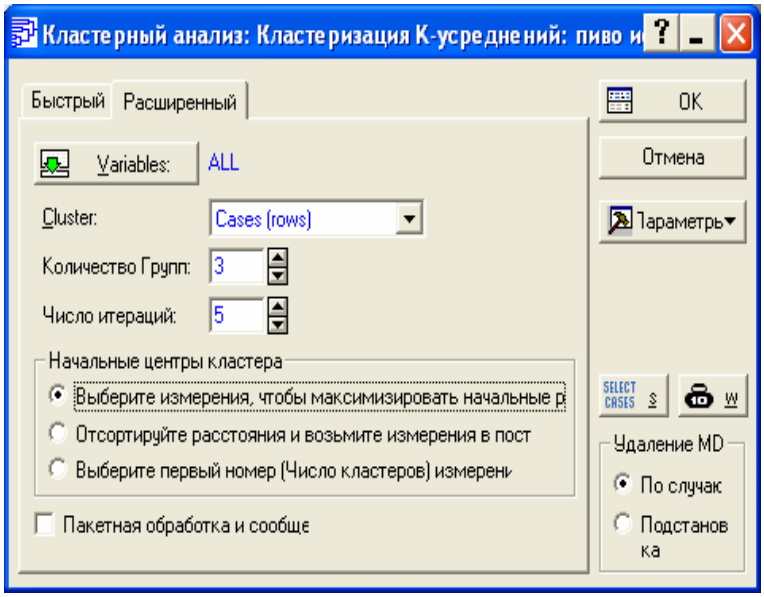


Рис.1.22. Кластерный анализ: Кластеризация К – усреднений: …

При правильном заполнении управляющих параметров программы **группировки методом k-средних** получим диалоговое окно, показанное на рис. 1.22.

Здесь необходимы некоторые пояснения.

В процедуре кластеризации предполагается определение количества групп пользователем, в примере задано деление на три группы из следующих соображений – пиво хорошее по вкусовым качествам, плохое, «так себе» или пиво дорогое, среднее по цене и дешевое. В разных задачах свои предпосылки определения количества групп.

В строке «Число итераций» задается число итераций, используемых при построении классов (как правило 5–10).

Группа опций «Начальные центры классов» позволяет задать разные условия для начальных «центров тяжести» групп. (В примере с пивом можно поэкспериментировать с заданием центров кластеров).

После заполнения диалогового окна (рис. 1.22) щелкаем по кнопке **OK**.

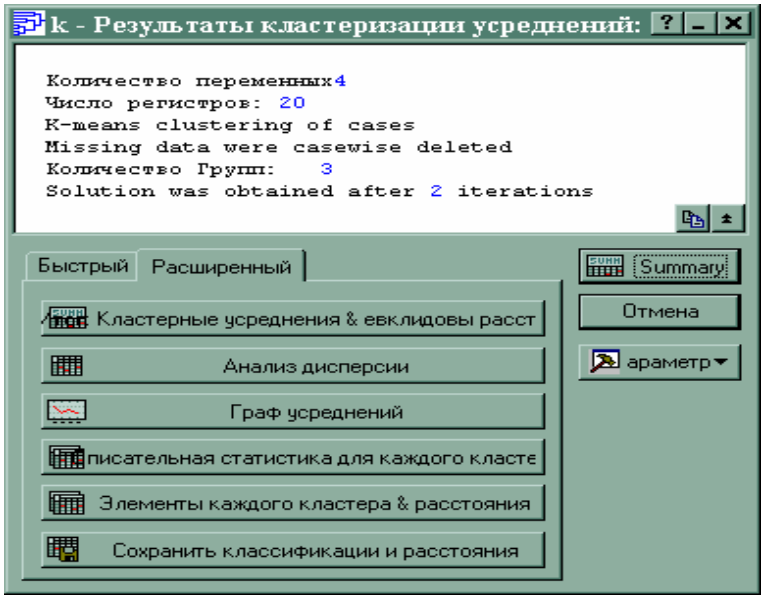


Рис.1.23. Окно результатов кластеризации сортов пива по методу k-средних

Происходит группирование «наших сортов пива» по трем классам. По окончании группирования на экране монитора появляется окно результатов (рис. 1.23), в информационной части которого отражена информация о выборке (m = 4, n = 20), о методе кластеризации (группировались сорта пива), о способе обработки пропущенных значений в матрице данных (таких нет), о количестве заданных групп (три группы), об итерациях (после скольких итераций найдено решение, в примере – две итерации).

Кнопки в нижней части окна (рис. 1.23) дают возможность провести анализ результатов группирования.

Кнопка (при нажатии): **«Кластерные усреднения и евклидовы расстояния «** позволяет вывести на экран таблицы, в первой из которых указаны средние значения переменных для каждого кластера (рис. 1.24), во второй (рис. 1.25) – евклидовы расстояния и их квадраты между кластерами (над диагональю – квадраты).

Кнопка: **«Граф усреднений «** позволяет посмотреть на экране средние значения для кластеров на графике (рис. 1.26).

Кнопка: **«Анализ дисперсий «** (рис. 1.27) позволяет проанализировать результаты дисперсионного анализа / 30, 44/.

Нажатие кнопки: **«Описательная статистика для каждого кластера»** дает возможность посмотреть средние значения, дисперсии, стандартные отклонения по переменным групп.

	Cluster Means (пиво исх.данные)		
Переменн	Кластер Нет.1	Кластер Нет.2	Кластер Нет.3
КАЛО	-1.13691	0.548624	0.755185
НАТР	-0.70872	0.834860	-0.638174
АЛКО	-1.01157	0.473822	0.704153
ЦЕНА	-0.32778	-0.569870	1.855826

Рис. 1.24. Средние значения переменных по кластерам

	Euclidean Distances between Clusters (пиво исх.данные) Distances below diagonal Squared distances above diagonal		
Кластер Номер	**Нет.1**	Нет.2	Нет.3
Нет.1	0.000000	1.872168	2.824214
Нет.2	1.368272	0.000000	2.037388
Нет.3	1.680540	1.427371	0.000000

Рис. 1.25. Евклидовы расстояния и их квадраты между кластерами

Для просмотра распределения сортов пива по группам используется кнопка **«Элементы каждого кластера и расстояния».** При ее нажатии на экране появится электронная таблица (рис. 1.28) для кластеров (поочередно, при нажатии в последней строке кнопки «**Members of Cluster Number**» … высвечивается таблица с нужным номером группы) с сортами пива, разнесенным по группам. В строках таблиц указано расстояние от каждого сорта пива до центра кластера.

Кнопка **«Сохранить классификации и расстояния»** позволяет сохранить результаты группирования для дальнейшего исследования (рис. 1.29).

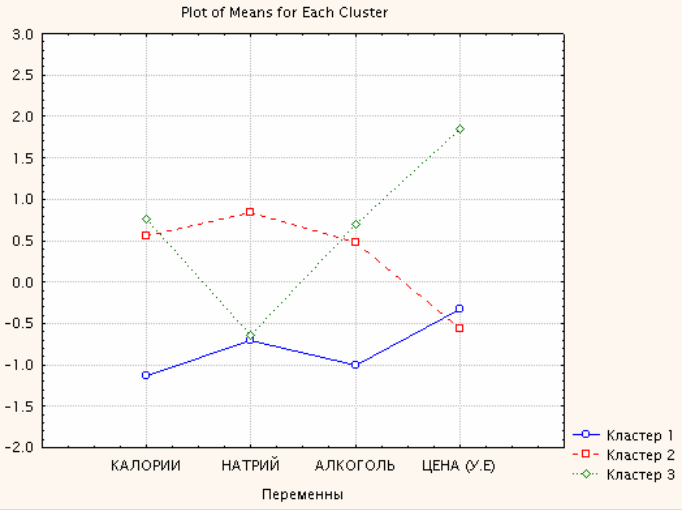


Рис. 1.26. График средних значений переменных для кластеров

Переменн	Дисперсионный анализ (пиво исх.данные)					
	между SS	df	Внутрен SS	df	F	signif. p
КАЛО	14.03804	2	4.961957	17	24.04764	0.000011
НАТР	11.41799	2	7.582009	17	12.80042	0.000406
АЛКО	11.16686	2	7.833144	17	12.11752	0.000536
ЦЕНА	17.45122	2	1.548779	17	95.77572	0.000000

Рис. 1.27. Таблица дисперсионного анализа

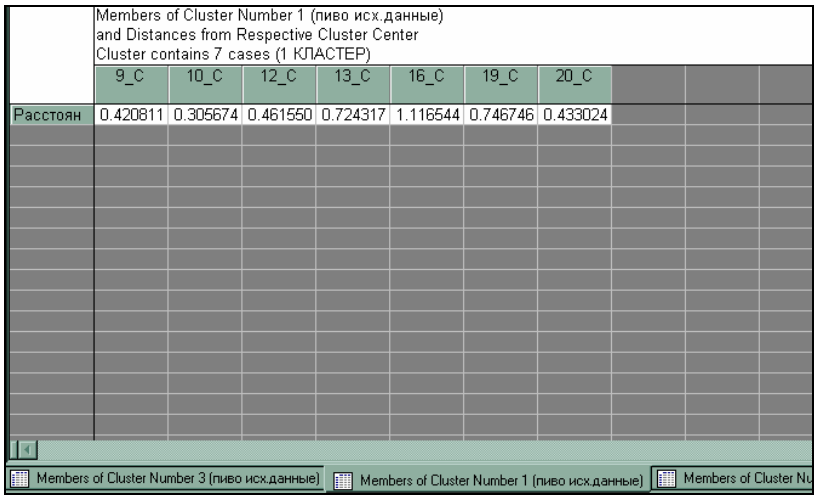


Рис. 1.28.Таблица с сортами пива, разнесенными по кластерам

	пиво исх.данные						
	1 КАЛОРИИ	2 НАТРИЙ	3 АЛКОГОЛЬ	4 ЦЕНА (У.Е)	5 CASE_NO	6 CLUSTER	7 DISTANCE
1_C	0.383	0.008	0.342	-0.472	1	2	0.43
2_C	0.615	0.615	0.605	-0.472	2	2	0.14
3_C	0.813	0.008	0.605	-0.135	3	2	0.49
4_C	1.243	-1.208	1.000	1.890	4	3	0.40
5_C	0.648	-0.600	0.737	1.822	5	3	0.06
6_C	0.416	1.223	0.211	-1.485	6	2	0.52
7_C	1.408	1.375	1.395	-0.675	7	2	0.69
8_C	0.549	1.831	0.342	-0.540	8	2	0.50
9_C	-1.104	-0.752	-0.184	-0.472	9	1	0.42
10_C	-0.641	-1.056	-0.974	-0.405	10	1	0.31
11_C	0.251	0.463	0.211	-0.405	11	2	0.28
12_C	-1.005	0.008	-0.447	-0.270	12	1	0.46
13_C	0.086	-0.600	-0.316	0.000	13	1	0.72
14_C	0.582	0.615	0.342	1.755	14	3	0.66
15_C	0.549	-1.360	0.737	1.957	15	3	0.38
16_C	-2.128	0.008	-2.817	-0.810	16	1	1.12
17_C	0.119	0.615	-0.053	-0.472	17	2	0.36
18_C	0.383	1.375	0.605	-0.472	18	2	0.29
19_C	-1.996	-1.360	-2.027	-0.135	19	1	0.75
20_C	-1.170	-1.208	-0.316	-0.202	20	1	0.43

Рис.1.29. Таблица с результатами классификации сортов пива

2. Метод главных компонент

Практически ни одно современное исследование многомерных данных не обходится без применения метода главных компонент (МГК). Это – классический метод снижения размерности данных путем определения незначительного числа линейных комбинаций исходных признаков, объясняющих большую часть изменчивости данных в целом, дающий однозначное решение.

Методу посвящено большое количество публикаций, он широко представлен в литературных источниках, обратившись к которым можно получить сведения об МГК с различной степенью детализации и математической строгости. Перечень литературы дан в конце главы.

2.1 Сущность метода главных компонент

МГК осуществляет переход к новой системе координат $y_1,....y_m$ в исходном пространстве признаков $x_1,...,x_m$, которая является системой ортонормированных линейных комбинаций. Линейные комбинации представляют собой собственные векторы корреляционной матрицы. Первая главная компонента – это линейная комбинация, обладающая наибольшей дисперсией. Геометрически выглядит как новая ось y_1, ориентированная вдоль направления наибольшей «вытянутости эллипсоида рассеивания объектов выборки» в исходном пространстве. Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеивания, перпендикулярное первой главной компоненте и т. п.

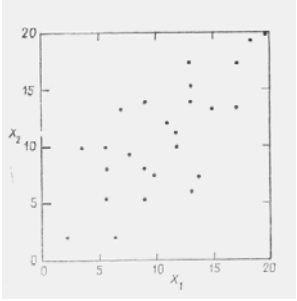


Рис. 2.1 Диаграмма рассеяния двумерных данных

По ковариационной матрице s^2 равной (расчет в пакете *STATISTICA* – см. рис. 2.5)

$\begin{bmatrix} 20{,}3 & 15{,}6 \\ 15{,}6 & 24{,}1 \end{bmatrix}$, вычисленной для двумерных данных (рис. 2.1) построен эллипс.

Вышесказанное выглядит следующим образом (рис. 2.2).

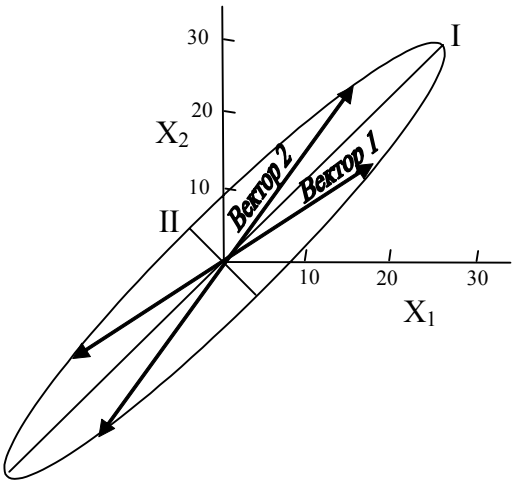


Рис. 2.2. Эллипс, определенный дисперсиями и ковариациями данных, представленных на рис. 2.1.
Главная компонента I соответствует 85.4 % суммарной дисперсии, главная компонента II – 14.6 % (см. рис. 2.6).
(Дисперсия переменной X_1=20,3, переменной X_2=24,1, ковариация равна 15,6)

МГК обладает рядом полезных свойств, делающих его эффективным для визуализации структуры многомерных данных. Все они касаются наименьшего искажения геометрической структуры точек (объектов) при их проектировании в пространство меньшей размерности $q < m$, «натянутое» на q первых главных компонент.

Приведенные свойства главных компонент обусловливают полезность МГК для изучения структуры распределения объектов в многомерном пространстве признаков. Как уже отмечалось, практически ни одно современное исследование не обходится без того, чтобы не рассмотреть проекции объектов в пространстве,

натянутом на первую, первые две и, реже, первые три главные компоненты.

Ценную информацию о структуре данных могут дать главные компоненты, полученные раздельно для различных классов объектов. В этом случае к интересным результатам может привести анализ не только первых главных компонент, но и главных компонент с высоким порядком, близким к *p*. По определению на такие главные компоненты приходится минимальный процент дисперсии распределения объектов. Поэтому они выражают устойчивые, стабильные свойства классов, инвариантные к изменчивости внутри классов.

Но вернемся к нашему примеру и посмотрим использование МГК на практике.

При определении главных компонент нам встретятся термины из матричной алгебры. К сожалению, большинство медиков не изучали высшую алгебру и, в связи с этим, некоторые понятия и термины будут незнакомы.

Так как нас больше будет интересовать использование главных компонент при многомерных исследованиях, а не их получение с помощью стандартных алгебраических приемов, мы в примере, рассчитанном с помощью пакета *STATISTICA*, отдадим предпочтение их интерпретации, а не расчетам.

Таблица 2.1

Двухмерные наблюдения с дисперсией X_1=20,3, дисперсией X_2=24,1 и ковариацией равной 15,6

X1	**X2**	**X1**	**X2**	**X1**	**X2**	**X1**	**X2**	**X1**	**X2**
3	2	7	2	9	14	13	6	15	13
4	10	7	13	10	7	13	14	17	13
6	5	8	9	11	12	13	15	17	17
6	8	9	5	12	10	13	17	18	19
6	10	9	8	12	11	14	7	20	20

Если проведены измерения переменных на некотором множестве объектов, то для них можно вычислить матрицу ковариаций порядка $m*m$ – [s^2]. Найдем m ее собственных векторов и m собственных значений по правилам матричной алгебры [14]. Так как ковариационная матрица всегда симметрична, то эти m собственных векторов будут ортогональными, т. е. углы между ними будут прямыми.

Первый собственный вектор имеет координаты (расчет – рис. 2.7)

$$I = \begin{bmatrix} 0{,}66 \\ 0{,}75 \end{bmatrix}.$$

Первое собственное значение равно 37,9 и является длиной главной полуоси.

Второй собственный вектор имеет координаты (расчет – рис.2.7)

$II = \begin{bmatrix} 0{,}75 \\ -0{,}66 \end{bmatrix}$ и образует прямой угол с первым.

Собственное значение, соответствующее этому вектору, т. е. длина II главной полуоси, равна 6,5. Эти геометрические соотношения показаны на рис. 2.2. Обратите внимание на то, что на диаграмму нанесены векторы ковариационной матрицы и поэтому измерения на диаграмме даны в тех же единицах, что и в дисперсии, или, как в этом примере, в квадратах единиц длины.

Определим суммарную дисперсию рассматриваемых данных как сумму вкладов от индивидуальных дисперсией. В данном примере суммарная дисперсия равна 20,3 + 24,1 = 44,4. Вклад первой переменной составляет 20,3/44,4, или около 44 % суммарной дисперсии, а вклад второй – примерно 56 % (рис. 2.8).

Сумма собственных значений матрицы также равна 37,9+ 6,5 = =44,4 (рис. 2.6). Так как эти собственные значения определяют длину двух главных осей (рис 2.2), то последние также характеризуют суммарную дисперсию множества данных, и вклад каждой из них в суммарную дисперсию равен соответствующему собственному значению, деленному на сумму собственных значений. Первая главная ось составляет 37,9/44,4, или 85,5 % суммарной дисперсии, в то время как вторая ось – только 14,5 %. Иными словами, изменчивость множества данных по первой главной оси равна 4/5 общей изменчивости наблюдений. Как правило, оказывается, что по крайней мере одна из главных осей эффективнее (по вкладу в суммарную дисперсию), чем любая из первоначальных переменных. С другой стороны, по меньшей мере одна из осей должна оказаться менее эффективной, чем любая из исходных переменных.

Если сделать преобразование вида $Y_1 = \alpha_1 X_1 + \alpha_2 X_2$, где α_1 и α_2 – координаты первого собственного вектора, то в результате получим новое множество данных, с дисперсией 37,9. Аналогичное преобразование $Y_2 = \beta_1 X_1 + \beta_2 X_2$, где β_1 и β_2 – координаты второго собственного вектора, приведет к преобразованию данного множества точек в множество с дисперсией, равной только 6,5. Так как эти новые переменные определены на осях, образующих прямой угол друг с другом, то ковариация между ними равна нулю. В табл. 2.2 представлены данные табл. 2.1, преобразованные таким образом – каждое исходное наблюдение заменено его проекцией на главные оси.

Проектирование на первую главную ось осуществляется по формуле

$$Y_{1i} = 0{,}66 X_{1i} + 0{,}75 X_{2i},$$

где коэффициенты при X_1 и X_2 являются координатами первого собственного вектора.

Проектирование на вторую главную ось осуществляется по формуле

$$Y_{2i} = 0{,}75 X_{1i} - 0{,}66 X_{2i},$$

Координаты собственных векторов, используемые для вычисления проекций наблюдений, называются нагрузками. Они являются коэффициентами линейного уравнения, которое используется для определения собственного вектора.

Таблица 2.2

Главные компоненты для данных табл. 2.1, вычисленные с помощьюпроектирования исходных данных на главные оси; дисперсия Y_1 равна 37,9, дисперсия Y_2 равна 6,5

Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
3,49	0,92	3,49	0,92	3,49	0,92	3,49	0,92	3,49	0,92
10,13	-3,64	10,13	-3,64	10,13	-3,64	10,13	-3,64	10,13	-3,64
7,72	1,17	7,72	1,17	7,72	1,17	7,72	1,17	7,72	1,17
9,96	-0,82	9,96	-0,82	9,96	-0,82	9,96	-0,82	9,96	-0,82
11,46	-2,14	11,46	-2,14	11,46	-2,14	11,46	-2,14	11,46	-2,14

Вернемся к нашему множеству данных. Мы определили собственные векторы матрицы и нашли, что первый собственный вектор дает вклад в суммарную дисперсию около 85,5 %. Допустим, что нужно свести систему только к одной переменной. Это можно сделать, отбросив любую из переменных X_1 или X_2, что приведет к потере либо 44 %, либо 56 % изменчивости (см. выше), в зависимости от того, какую переменную мы сохраним. Однако если спроектировать все наблюдения на первую главную ось, то потеря составит только 14,5 % от изменчивости данных.

Теперь покажем нахождение главных компонент на примере (табл. 2.1) с помощью пакета *STATISTICA* и используемых в пакете терминов.

2.2 Применение метода главных компонент в пакете *STATISTICA* (пример)

Приведем пример использования пакета *STATISTICA* для МГК.

Открытие файла данных проводим стандартным образом:

Файл => Открытие …(адрес нахождения файла с данными табл. 2.1).

Проводить стандартизацию исходных данных не будем (у переменных одни и те же единицы измерения).

Для вызова модуля с методом главных компонент выбираем пункт главного меню **Статистика => Многомерные исследующие методы => Основные компоненты и классификационный анализ. OK.**

На экране монитора высвечивается диалоговое окно, с помощью которого выбираем переменные для нахождения главных компонент (рис. 2.3).**OK.**

По окончании расчетов на экране отображается диалоговое окно «Основные компоненты и результаты анализа классификации» (рис. 2.4).

Рис. 2.3. Стартовая модель модуля «Основные компоненты и классификационный анализ»

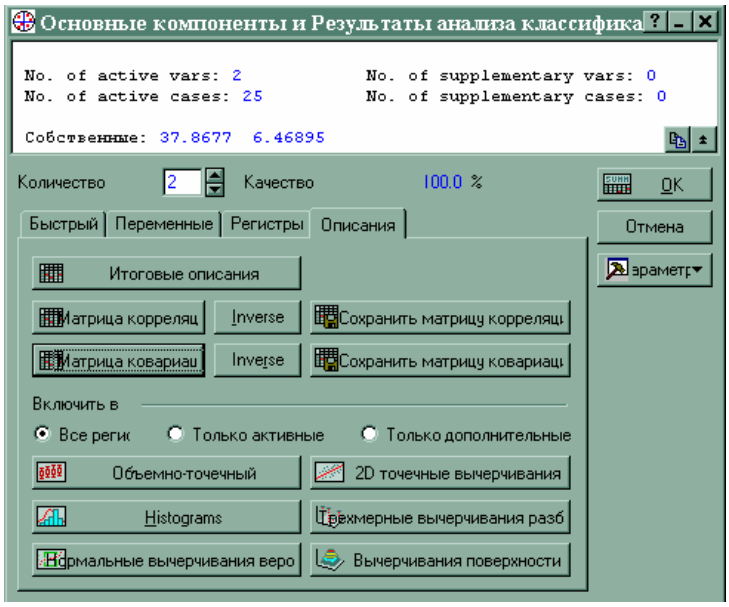


Рис. 2.4. Диалоговое окно «Основные компоненты и результаты анализа классификации»

Для просмотра результатов воспользуемся кнопками этого окна.

Нажмите кнопку «Матрица ковариаций». На экране высветится ковариационная матрица s^2 (рис. 2.5).

Переменная	Переборы (Мгк дев) X1	X2
X1	20.27667	15.58500
X2	15.58500	24.06000

Рис. 2.5. Матрица ковариаций (дисперсия переменной X_1=20,3, переменной X_2=24,1, ковариация равна 15,6

Value number	Eigenvalues of covariance matrix, and related statistics (Мгк дев) Active variables only **Eigenvalue**	% Total variance	Совокупный Eigenvalue	Совокупный %
1	37.86772	85.40948	38	85.4095
2	6.46895	14.59052	44	100.0000

Рис. 2.6. Результат расчета собственных значений главных компонент, их процентное содержание, накопленные суммы

При нажатии кнопок «Переменные» (рис. 2.4), далее «EIGENVALUES» отображаются результаты расчетов собственных значений (первое значение равно 37,86, второе – 6,48), главная компонента I соответствует ≈85,4 % суммарной дисперсии, главная компонента II – ≈14,5 % (рис. 2.6).

Переменная	Eigenvectors of covariance matrix (Мгк дев) Active variables only **Фактор 1**	Фактор 2
X1	0.663139	0.748496
X2	0.748496	-0.663139

Рис. 2.7. Собственные векторы (I и II вектор) матрицы S^2

При нажатии кнопок «Переменные» (рис. 2.4), «EIGENVECTORS» высвечиваются координаты векторов (рис. 2.7).

При нажатии кнопок «Переменные» (рис. 2.4) и «Контрибуция переменных» отображается вклад в общую дисперсию переменных (рис. 2.8)

Переменная	Variable contribution, based on covariances (Мгк дев)	
	Фактор 1	Фактор 2
X1	0.439753	0.560247
X2	0.560247	0.439753

Рис. 2.8. Вклад в общую дисперсию переменных (в %)

В этой главе нами показан, пожалуй, только принцип работы с методом главных компонент. Для того чтобы подробно изучить его, воспользуйтесь литературой [12, 16, 17, 35, 18, 15, 20, 2 и др.].

Здесь же нам хотелось бы, в связи со знакомством в следующей главе с дискриминантным анализом, отметить, что канонические переменные, которые рассчитываются и используются в нем, являются ни чем иным, как главными компонентами, полученными при наличии дополнительной априорной информации групповой принадлежности объектов [18]. Применяются же, в основном, канонические переменные для визуализации данных [18].

2.3 Принцип факторного анализа

В отличие от метода главных компонент факторный анализ основан не на дисперсионном критерии системы признаков, а ориентирован на объяснение имеющихся между признаками корреляций с последующим сокращением размерности [17, 46].

Задачу факторного анализа нельзя решить однозначно. Существует много методов факторного анализа. Здесь сошлемся на слова одного из основоположников современного факторного анализа Г. Хартмана: «Ни в одной из работ не было показано, что какой-либо один метод приближается к «истинным» значениям ... Выбор среди группы методов наилучшего производится в основном с точки зрения вычислительных удобств, а также склонностей и «привязанностей» исследователя, которому тот или иной метод казался более адекватным …» [3].

В настоящее время одними из наиболее популярных являются три метода вращения факторов: варимакс, квартимакс и эквимакс [17]. Кроме перечисленных трех методов нередко осуществляют вращение факторов до тех пор, пока не получатся результаты, поддающиеся содержательной интерпретации. Можно, например, потребовать, чтобы один фактор был нагружен преимущественно признаками одного типа, а другой – признаками другого типа. Или, скажем, можно потребовать, чтобы исчезли какие-то трудно интерпретируемые нагрузки с отрицательными знаками.

В целом по факторному анализу можно отметить следующее. С помощью такого анализа снижение размерности достигается за счет существования групп взаимосвязанных признаков, которые агрегируются в строящихся факторах. Как и при использовании метода главных компонент, полезные сведения о структуре данных можно почерпнуть на основании визуального анализа проектов объектов в одно-, двух- и трехмерных пространствах, образованных комбинациями различных факторов. Также информацию о структуре исследуемой выборки могут дать результаты факторного анализа, проведенного раздельно в различных подгруппах объектов.

Подробно с факторным анализом можно ознакомиться по следующей литературе [12, 42, 22, 17, 35, 15, 20, 1, 2, 3, 34, 46 и др.].

3. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Один из наиболее широко используемых в статистике многомерных методов – дискриминантный анализ. Его можно поставить в один ряд с одномерными задачами – задачами проверки статистических гипотез. Он позволяет также установить дополнительную связь между одномерной и многомерной статистикой.

Понятие разделения (дискриминации) отличается от близкого к нему понятия классификации.

Предположим, что имеются две группы точек (группы А и В), о которых заранее известно, что они образованы на основании их расположения по осям X_1 и X_2 (рис. 3.1). Задача состоит в нахождении такой линейной комбинации этих переменных (оси X_1 и X_2), которая даст максимально возможное различие между двумя ранее определенными группами. Если удастся найти такую функцию, то мы сможем использовать ее для отнесения новых точек к той или другой исходной группе. Иными словами, новые точки, не имеющие диагностических признаков отнесения их к той или иной группе, можно будет разнести по группам на основе линейной дискриминантной функции, построенной по переменным (оси X_1 и X_2) [12].

Дискриминантный анализ основан на нахождении преобразования, которое дает минимум отношения разности многомерных средних значений для некоторой пары групп к многомерной дисперсии в пределах двух групп. Если мы изобразим наши две группы точек в многомерном пространстве, то легко найти такое направление, вдоль которого эти совокупности явно разделяются. Графически эта картина представлена на рис. 3.1. Если использовать переменные (оси) X_1 и X_2, то провести удовлетворительное разделение групп *A* и *B* не удается. Однако можно найти направ-

ление, вдоль которого разделение совокупностей очевидно. Координаты точек этого направления задаются уравнением линейной дискриминантной функции.

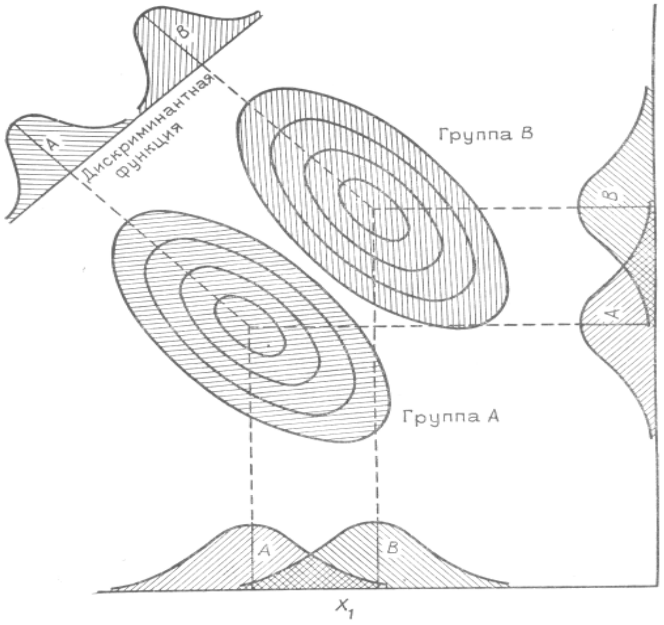


Рис. 3.1. Графическое представление двух групп**.** Указаны перекрытия распределений для групп A и B по осям X_1 и X_2. Проектирование на дискриминантную линию позволяет различить две группы

Задачу классификации (группирования) можно проиллюстрировать на аналогичном примере. Предположим, что есть большое количество точек. Можно ли на основе значений измеренных переменных осуществить разделение их на относительно однородные группы (кластеры), отличающиеся друг от друга? Численные методы решения такого рода задач достаточно разработаны и рассмотрены в первой главе. Здесь же отметим, что существуют различия между этими методами и дискриминантным анализом.

Классификация внутренне замкнута, т. е., в отличие от дискриминантного анализа, она не зависит от априорных сведений о принадлежности точек к различным группам. В дискриминантном

анализе число групп задается заранее, в то время как число кластеров, которые получаются в результате классификации, не может быть, как правило, заранее определено. Каждая точка из исходного множества в дискриминантном анализе принадлежит к одной из заданных групп. В большинстве задач классификации точка может войти в любую из групп, возникающих в результате классификации.

Остановимся более подробно на дискриминантном анализе.

Линейная дискриминантная функция осуществляет преобразование исходного множества измерений, входящих в выборку, в единственное дискриминантное число. Это число, или преобразованная переменная, определяет положение точки на прямой, определенной дискриминантной функцией. Поэтому мы можем представлять себе дискриминантную функцию как способ преобразования многомерной задачи в одномерную. Мы не будем подробно рассматривать математические выкладки расчетов и нахождения необходимых параметров и коэффициентов, используемых в дискриминантном анализе, а покажем работу его на примере. (Подробно с этими расчетами можно познакомится [12, 48, 5, 17].)

В качестве примера построим дискриминантную функцию для двух групп данных, приведенных в табл. 3.1 [12].

Таблица 3.1

Результаты измерения среднего размера зерен и коэффициента отсортированности в двух группах проб песка, взятых у берега (А) и в удалении от него (В)

Средний размер зерен	**Коэффициент отсортированности**	**Средний размер зерен**	**Коэффициент отсортированности**
X1*	**X2***	**X1**	**X2**
Группа А		**Группа В**	
0,333	1,08	0,339	1,12
0,340	1,08	0,346	1,12
0,338	1,09	0,350	1,12
0,333	1,10	0,352	1,13
0,323	1,13	0,341	1,15
0,327	1,12	0,347	1,15
0,329	1,13	0,337	1,16
0,331	1,13	0,343	1,16

0,336	1,12	0,340	1,17
0,333	1,14	0,346	1,17
0,341	1,14	0,349	1,17
0,328	1,15	0,339	1,18
0,336	1,15	0,342	1,18
0,327	1,16	0,346	1,18
0,329	1,16	0,351	1,18
0,330	1,16	0,340	1,19
0,323	1,17	0,344	1,19
0,328	1,17	0,333	1,20
0,332	1,17	0,337	1,20
0,331	1,18	0,339	1,20
0,326	1,18	0,342	1,20
0,333	1,18	0,339	1,21
0,330	1,19	0,340	1,21
0,336	1,19	0,341	1,21
0,327	1,20	0,335	1,22
0,324	1,21	0,337	1,22
0,332	1,21	0,340	1,22
0,322	1,22	0,343	1,22
0,329	1,22	0,334	1,22
0,325	1,24	0,348	1,22
0,328	1,26	0,337	1,22
0,322	1,27	0,342	1,23
0,318	1,22	0,334	1,24
0,330	1,17	0,340	1,24
-	-	0,342	1,24
-	-	0,331	1,25
-	-	0,336	1,25
-	-	0,341	1,25
-	-	0,334	1,26
-	-	0,337	1,27
-	-	0,339	1,27
-	-	0,330	1,28
-	-	0,334	1,28
-	-	0,332	1,29
-	-	0,330	1,31
-	-	0,334	1,31
-	-	0,340	1,21

* Переменные X1 и X2 имеют разные единицы измерения.

Группа *A* представлена пробами современного песка, взятого с морского пляжа; две переменные – это средний размер зерен и коэффициент отсортированности. Группа *B* представлена пробами песка, взятого в отдалении от берега. Переменные в этом случае такие же, как и для группы *A*.

Точечная диаграмма исходных наблюдений представлена на рис. 3.2. Хотя две группы точек и перекрываются, совершенно очевидно, что разделяющая их линия проходит между ними так, что большинство наблюдений группы *A* находится по одну сторону от нее, а большинство наблюдений группы B – по другую.

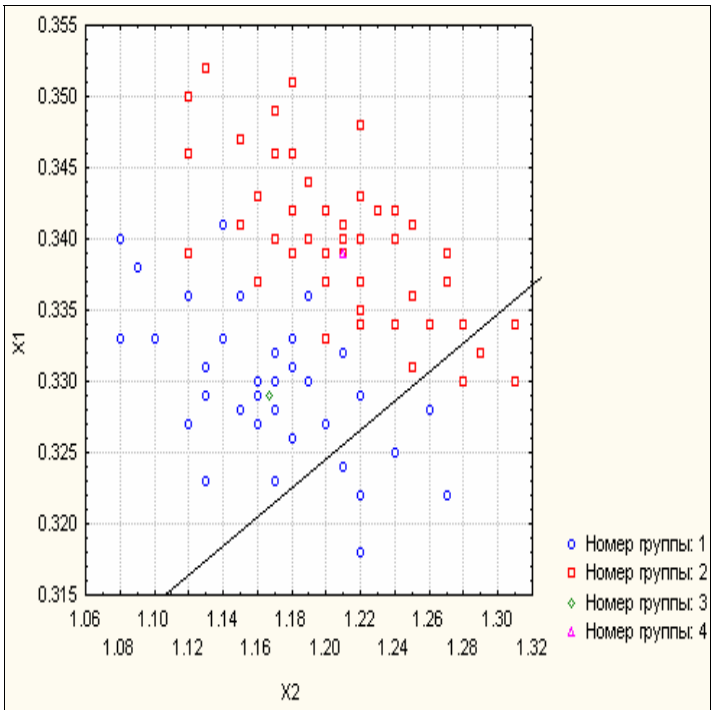


Рис. 3.2. Зависимость медианы размеров зерен от коэффициента отсортированности в пробах песка:
1 — пробы пляжного песка; 2 — пробы, взятые в отдалении от берега; 3 и 4 — двумерные средние (центроиды) двух групп функций. Прямая линия – график дискриминантной функции; X1 – средний размер зерен, X2 – коэффициент отсортированности.

3.1. Критерии значимости

Первый шаг в применении критерия значимости дискриминантной функции – оценка различия между группами. Это можно сделать с помощью вычисления расстояния между центроидами или многомерными средними групп. Мера расстояния получается прямо из многомерных статистик. Мы можем получить меру различия между средними двух одномерных выборок, $\overline{X_1}$ и $\overline{X_2}$ просто вычитая одно значение, из другого. Однако разность выражается в тех же единицах, что и исходные наблюдения, и обычно более удобна, если использовать ее в стандартизованной форме. Разделив разность на объединенное стандартное отклонение, мы получаем стандартизованную разность

$$d = \frac{(\overline{X_1} - \overline{X_2})}{s_p}. \qquad (3.1)$$

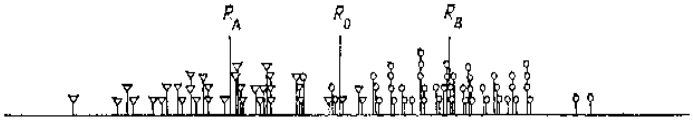


Рис. 3.3. Проекция выборок, представленных в табл. 3.1, на дискриминантную прямую, изображенную на рис. 3.2: R_a– проекция двумерного среднего для группы *A*; $R_в$ – проекция двумерного среднего для группы *B*; R_0 – дискриминантный индекс, соответствует точке, находящейся посередине между R_a и $R_в$. Разность между R_a R_a и $R_в$ – расстояние Махалонобиса D^2.

Возведя обе части формулы (3.1) в квадрат и обозначив знаменатель, являющийся объединенной дисперсией двух выборок, через $s_p^{2,}$ получим

$$d^2 = (\overline{X}_1 - \overline{X}_2)^2 / s_p^{\,2}.$$

Предположим, что вместо единственной переменной на каждом наблюдении двух групп измеряются две переменных. Разность между двумерными средними двух групп может быть выражена как обыкновенное евклидово расстояние или расстояние по прямой между ними. Обозначая эти две группы через A и B, получаем

$$\text{евклидово расстояние} = \sqrt{(\overline{A}_1 - \overline{B}_1)^2 - (\overline{A}_2 - \overline{B}_2)^2}.$$

В общем случае, если на каждом наблюдении измеряется *m* переменных, то расстояние по прямой между многомерными средними двух групп есть

$$\text{евклидово расстояние} = \sqrt{\sum_{i=1}^{m} (\overline{A}_i - \overline{B}_i)^2}.$$

Евклидово расстояние и его квадрат выражаются в единицах, составленных из исходных единиц измерений. Для того чтобы иметь возможность их интерпретировать, их надо стандартизировать.

Пропуская некоторые промежуточные математические расчеты (их можно посмотреть в [12]) получаем стандартизованный квадрат расстояния

$$D^2 = \left[\overline{A}_i - \overline{B}_i\right]' \left[s_p^2\right]^{-1} \left[\overline{A}_i - \overline{B}_i\right]. \tag{3.2}$$

Эта мера расстояния между средними двух многомерных групп называется *расстоянием Махалонобиса* [22].

Расстояние Махалонобиса графически представлено на рис. 3.3, где оно равно расстоянию между R_A и R_B..

Значение расстояния Махалонобиса состоит в том, что оно является многомерным аналогом t-критерия для проверки гипотезы о равенстве двух средних, называемого критерием *Хотеллинга* T^2. (Работу этого критерия в пакете STATISTICA рассмотрим позднее.) Здесь отметим, что он имеет вид

$$T^2 = \frac{n_a n_b}{n_a + n_b} D^2 \tag{3.3}$$

и может быть преобразован в F – критерий. Этот критерий проверки гипотезы о равенстве двух многомерных параметров, использующий более известную статистику, определен выражением

$$F = \left(\frac{n_a + n_b - m - 1}{(n_a + n_b - 2)m}\right)\left(\frac{n_a n_b}{n_a + n_b}\right) D^2. \tag{3.4}$$

Числа степеней свободы полученной статистики равны m и $(n_a + n_b - m - 1)$. Проверяемая с помощью этой статистики нулевая гипотеза заключается в том, что два неизвестных многомерных средних равны между собой или что расстояние между ними равно нулю, т. е. $H_0 : [D_i] = 0$ при множестве альтернатив $H_1 : [D_i] > 0$.

Пригодность метода дискриминантного анализа для проверки этой гипотезы не вызывает сомнений. Если средние значения двух групп очень близки друг к другу, то их трудно разделить, особенно если обе группы имеют большой разброс. Наоборот, если два средних значения легко разделяются и рассеяние вокруг средних мало, разделение осуществляется относительно просто.

Поскольку не все переменные, включенные в дискриминантную функцию, в равной степени полезны при отделении групп друг от друга, то эти «бесполезные» переменные желательно найти и исключить из дальнейшего рассмотрения.

Процедура исключения малозначимых переменных для разделения групп по известным статистическим методам [12, 17, 18] реализована в статистическом пакете STATISTICA.

Вернемся к демонстрационному примеру.

3.2. Дискриминантный анализ в пакете STATISTICA, интерпретация результатов

3.2.1. Демонстрационный пример

Воспользуемся модулем ***Дискриминантный анализ*** статистического пакета STATISTICA [7, 42, 8].

Предварительно данные (см. табл. 3.1) могут быть внесены в таблицу, показанную на рис. 3.4. Подготовку табл. 3.1 можно провести как в Excel [4], так и в STATISTICA [7, 8].

	1 номер группы	2 средний размер зерен	3 коэффициент отсортированности
28	1	0,322	1,22
29	1	0,329	1,22
30	1	0,325	1,24
31	1	0,328	1,26
32	1	0,322	1,27
33	1	0,318	1,22
34	1	0,33	1,17
35	2	0,339	1,12
36	2	0,346	1,12
37	2	0,35	1,12
38	2	0,352	1,13

Рис. 3.4. Переменные табл. 3.1 (фрагмент), подготовленные для расчетов

На рис. 3.4 Столбец «номер группы» – группировочный признак, где»1» –объекты группы *A* , «2» – объекты группы *B*. Объектов (проб, случаев) всего 81, в том числе в группе *A* – 34 , в группе *B* – 47 .

Так как используемые переменные имеют разные единицы измерения, то необходима их стандартизация [33, 19, 12]. Для этого «выделяем» столбцы, которые необходимо стандартизировать (X1, X2). Один из вариантов стандартизации: в главном меню программы STATISTICA 6.0 выбираем пункт **Редактирование,** в раскрывшемся меню – пункт **Заполнение / Стандартизация блока,** высвечивается каскадное меню, в котором устанавливаем указатель манипулятора «мышь» на пункте **Стандартизация столбцов** и щелкаем левой кнопкой мыши. Выделенные переменные стандартизируются (рис. 3.5).

	1 номер группы	2 средний размер зерен	3 коэффициент отсортированности
28	1	-1,84554512	0,53382604
29	1	-0,895114527	0,53382604
30	1	-1,43821772	0,916480104
31	1	-1,03089033	1,29913417
32	1	-1,84554512	1,4904612
33	1	-2,38864832	0,53382604
34	1	-0,759338728	-0,42280912
35	2	0,462643463	-1,37944428
36	2	1,41307406	-1,37944428
37	2	1,95617725	-1,37944428
38	2	2,22772885	-1,18811725

Рис. 3.5. Стандартизированные данные табл. 3.1

Для вызова модуля ***Дискриминантный анализ*** работаем по схеме:

пункт главного меню **Статистика => Многомерные исследующие методы => Дискриминантный анализ.**

На экране высвечивается Стартовая модель модуля (рис. 3.6). Предварительно выбираем переменные для анализа. **(Файл => Открытие => путь к каталогу с файлом** *Пример***).**

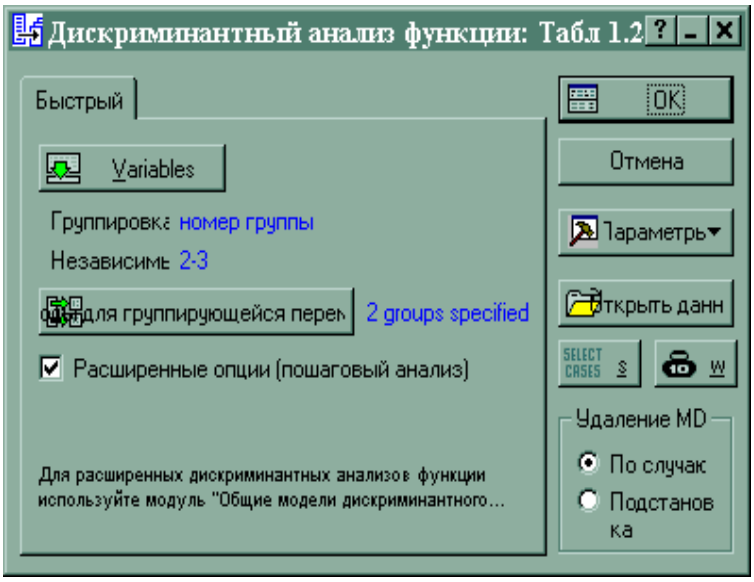


Рис. 3.6. Стартовая модель модуля Дискриминантный анализ

На стартовой модели модуля (рис. 3.6) нажимаем кнопку VARIABLES (переменные). Выбираем в качестве группирующей переменной «номер группы», в качестве независимых переменных (INDEPENT VARIABLES) выбираются «X1» и «X2». Нажимаем кнопку OK. Указатель мыши устанавливаем на кнопке «для группирующей переменной». Нажимаем OK. Высвечивается окно, в котором нажмите на кнопки «все» и OK. Результат проделанных выше действий со стартовой моделью модуля ***Дискриминантный анализ*** показан на рис. 3.6**.**

Если не устанавливать флажок (√) в опции «Расширенные опции (пошаговый анализ)», запускается расчет стандартного

дискриминантного анализа, иначе используется «пошаговый» вариант дискриминантного анализа.

Результат высвечивается в диалоговом окне «Результаты дискриминантного анализа функции» (рис. 3.7).

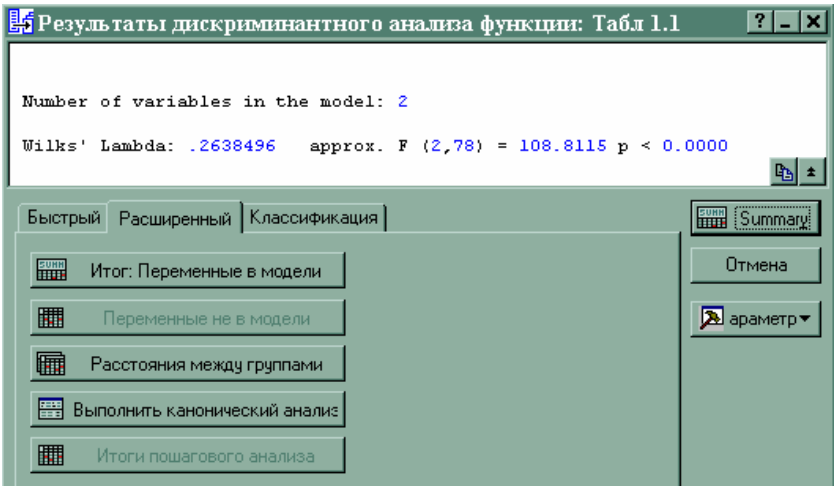


Рис. 3.7. Диалоговое окно «Результаты дискриминатного анализа»

В информационной части окна дана следующая информация:

- NUMBER OF VARIABLES IN THE MODEL (число переменных в модели): 2
- WILKS LAMBDA (значение лямбды Уилкса): 0,26
- APPROX (приближенное значение) F (2,78)=108,81
- p – уровень значимости F – критерия для F = 108,81 ($p<0,00001$).

Значения статистики лямбда Уилкса лежат в интервале [0,1], при этом значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации. Иными словами, это можно выразить следующим образом: если значения лямбда Уилкса близки к 0, то мощность дискриминации (мощность = 1 – вероятность ошибки) близка к 1, если лямбда Уилкса близка к 1, то мощность близка к 0 [8].

Нажмите кнопку: «Итог. Переменные в модели» (рис. 3.7). На экране монитора отображается окно с итоговой таблицей стандартного дискриминантного анализа данных из табл. 3.1 (рис. 3.8).

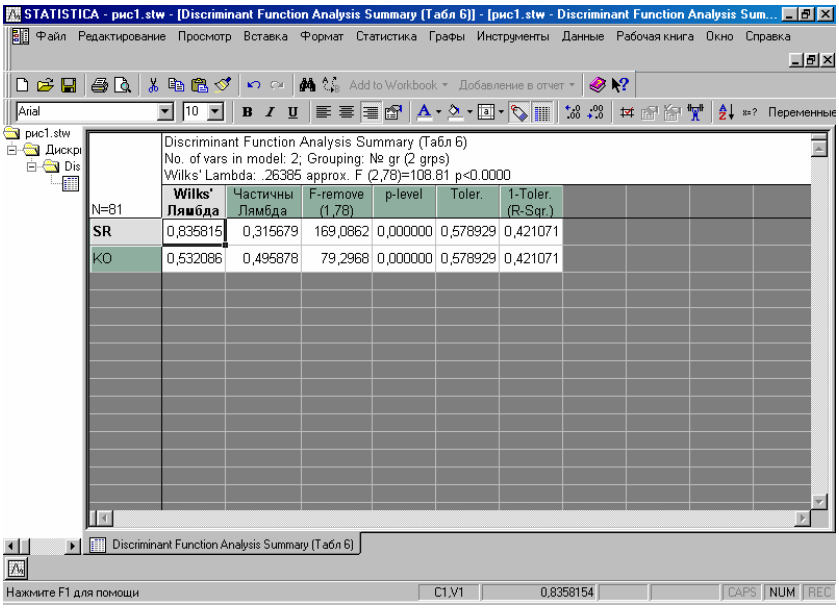


Рис. 3.8. Стандартный дискриминантный анализ данных

На рис. 3.8 показаны оценки информативности переменных, включенных в линейные дискриминантные функции (ЛДФ). По данным таблицы (рис. 3.8) видно, что обе переменные являются информативными параметрами с уровнями значимости $p<0,0000001$. Наиболее информативен признак X1, так как имеет большее значение $F = 169,08$.

Для возращения к диалоговому окну «Результаты дискриминатного анализа» (рис. 3.7) необходимо нажать на кнопку **«Результаты дискрим …»,** находящуюся в нижней части окна «Стандартный дискриминантный анализ данных» (рис. 3.8).

При нажатии на кнопку **Классификация** (рис. 3.7) Вы увидите на экране окно (рис. 3.9) с классифицирующей информацией. Рассмотрим наиболее важную информацию.

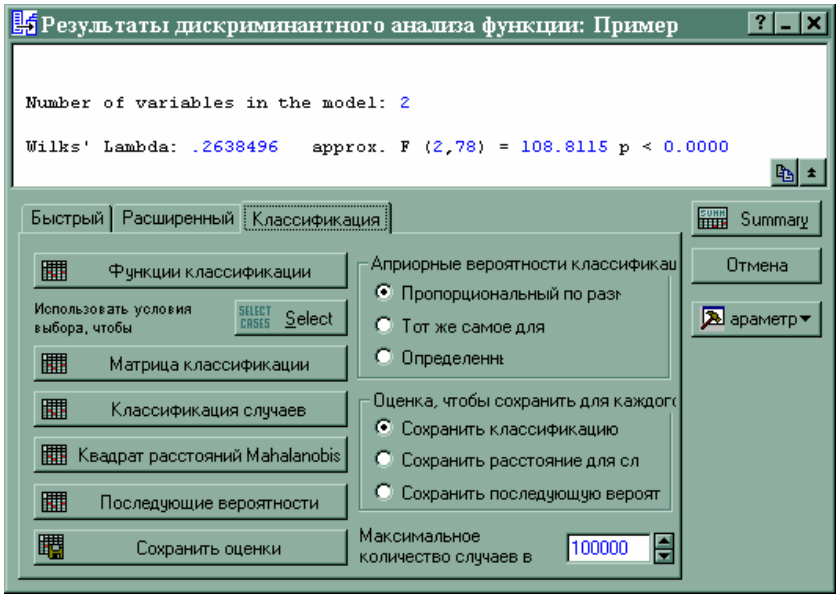


Рис. 3.9. Диалоговое окно «Результаты дискриминантного анализа. Классифицирующая информация»

Таблица с коэффициентами линейных дискриминантных (или классификационных) функций (ЛДФ или ЛКФ) отображается на экране при нажатии кнопки **Функции классификации** (рис. 3.9)**.** Результат на рис. 3.10.

	Classification Functions; grouping: № g			
Переменная	G_1:1 p=,41975	G_2:2 p=,58025		
SR	-3,34809	2,42202		
KO	-2,29283	1,65864		
Постоянн	-2,74888	-1,52855		

Рис. 3.10. Коэффициенты линейных классификационных функций (ЛКФ)

Линейные классификационные функции (ЛКФ) рассчитываются по формулам (рис. 3.10):

$$F1 = -2{,}74 - 3{,}34 \cdot X1 - 2{,}29 \cdot X2$$
$$F2 = -1{,}52 + 2{,}42 \cdot X1 + 1{,}65 \cdot X2 \qquad (3.5)$$

(X1 – первая переменная, X2 – вторая переменная).

На рис. 3.11 показан фрагмент классификации проб по ЛКФ (3.5).

Случай	Классификация случаев (Пример) Incorrect classifications are marked with *		
	Измеренн Classif.	1 p=.41975	2 p=.58025
28	G_1:1	G_1:1	G_2:2
29	G_1:1	G_1:1	G_2:2
30	G_1:1	G_1:1	G_2:2
* 31	G_1:1	G_2:2	G_1:1
32	G_1:1	G_1:1	G_2:2
33	G_1:1	G_1:1	G_2:2
34	G_1:1	G_1:1	G_2:2
* 35	G_2:2	G_1:1	G_2:2
36	G_2:2	G_2:2	G_1:1
37	G_2:2	G_2:2	G_1:1
38	G_2:2	G_2:2	G_1:1

Рис. 3.11. Фрагмент таблицы классификации по ЛКФ

Столбец «Измерен Classif» (рис. 3.11) содержит априорную информацию о принадлежности объекта к одной из двух групп *A* (G_1:1) или *B*(G_2:2).

В столбце «1» показана информация о классификации объектов (проб, случаев) по формулам (3.5) для группы *A*, в столбце «2» – классификация для группы *B* по (3.5). По таблице классификации случаев (рис. 3.11) видно несовпадение априорно заданной информацией с результатами расчета объектов (случаев) под номерами «31 и 35».

Нажатие кнопки **Матрица** классификации высвечивает окно (рис 3.12) с таблицей по результатам классификации с применением ЛКФ. (В работе [48] эта таблица называется «Оценка чувствительности решающих правил».)

Группа	Classification Matrix (Пример) Rows: Observed classifications Columns: Predicted classifications **Процент Исправле**	G_1:1 p=.41975	G_2:2 p=.58025
G_1:1	91.17647	31	3
G_2:2	93.61702	3	44
Итог	92.59259	34	47

Рис. 3.12. Оценка чувствительности решающих правил

Из этой таблицы видно, что при проверке линейными классифицирующими функциями (ЛКФ) предварительно проведенной «разбивки» объектов на две группы было неверно разнесено шесть объектов (три в первой группе и три во второй). Точность группирования – 92,6 %. Номера неверно разнесенных объектов можно увидеть в таблице классификации (рис. 3.11) в столбце «случай» (пример – 31 и 35 объекты).

Как рассчитывается и определяется принадлежность объекта к той или иной группе?

Рассмотрим эту процедуру на тридцать третьем и тридцать пятом объектах.

Стандартизированные значения переменных (данные взяты из таблицы, фрагмент которой показан на рис. 3.5):

объект №33 – X1= -2,38864832 X2 = 0,53382604
объект №35 – X1 = 0,462643463 X2 = -1,37944428

Для объекта №33 уравнения ЛКФ имеют вид:

F1 = – 2,74888 – 3,34809 · (-2,38864832) -2,29283 · 0,53382604 = 4,024557

F2 = – 1,52855+2,42202 · (-2,38864832) +1,65864 · 0,53382604 = -6,42848

Для объекта №35 уравнения ЛКФ имеют вид:

F1 = – 2,74888-3,34809 · 0,462643463 -2,29283 · (-1,37944428) = -1,13502

F2 = – 1,52855+2,42202 · 0,462643463+1,65864 · (-1,37944428) = -2,69602

Отнесение объекта к определенной группе выполняется по максимальному значению ЛКФ после их расчета по набору переменных для каждой группы.

Объект №33 относится к первой группе (4,024557 > -6,42848), объект №35 относится к первой группе (-1,13502 > -2,69602). Этот результат отражен на рис.3.11. Изначально же объект 35 относился ко второй группе или группе *B*.

На рис. 3.13. показана таблица квадрата расстояния Махалонобиса (для получения на экране таблицы надо нажать кнопку **Расстояния между группами**, в нижней части высветившегося окна нажать на кнопку **Квадрат расстояний Mahalonobi**). Высветится таблица с квадратами расстояний **Махалонобиса**. Эта таблица наиболее информативна, когда анализируется более двух групп (такой пример рассмотрим ниже). При анализе данных, расклассифицированных по двум группам, этот параметр интересен при перераспределении случаев (объектов, анализов) по этим группам без изменения количества признаков классификации.

Например, если при «передаче» трех объектов в группу *A* из группы *B* (табл. 3.1) и наоборот согласно результатам, полученным по формулам (3.5) и показанным на рис. 3.11 и рис. 3.12 квадрат расстояния Махалонобиса превысит значение 11,455 (см. рис. 3.13), то эту «передачу» можно считать удачной.

Здесь же можно посмотреть (при нажатии кнопки **Расстояния между группами)** таблицы F-values и p-levels – таблицы полезны при оценке «различимости групп».

	Квадрат расстояний Mahalanobi (Пример)				
Номер группы	**G_1:1**	G_2:2			
G_1:1	0.00000	11.45522			
G_2:2	11.45522	0.00000			

Рис. 3.13. Квадрат расстояний Махалонобиса между группами

При нажатии на кнопку **Квадрат расстояний Mahalonobis** (рис. 3.9) на экране монитора отображается таблица с квадратами расстояний Махалонобиса от объектов (пациентов) до центров групп.

Случай	Squared Mahalanobis Distances from Group Centroids (Пример) Incorrect classifications are marked with *					
	Измеренн Classif.	G_1:1 p=.41975	G_2:2 p=.58025			
28	G_1:1	2.08153	17.36765			
29	G_1:1	1.77961	6.09755			
30	G_1:1	2.30339	9.86478			
* 31	G_1:1	5.22147	5.05811			
32	G_1:1	4.55972	12.28562			
33	G_1:1	4.85438	26.40804			
34	G_1:1	0.01706	10.32834			
* 35	G_2:2	2.98494	6.75448			
36	G_2:2	10.71212	3.51348			
37	G_2:2	17.72798	4.26181			
38	G_2:2	23.23478	5.12279			

Рис. 3.14. Расстояния Махалонобиса от проб до центров групп

Случай (проба, объект) относится к группе, до которой расстояние Махаланобиса минимально.

После того как дискриминантный анализ выполнен, можно посмотреть вероятности принадлежности каждого случая к группам. Нажав кнопку **Последующие вероятности (***Апостериорные вероятности***)** (рис. 3.9), мы увидим таблицу с апостериорными вероятностями принадлежности объекта (случая) к определенной группе (рис. 3.15).

Интерпретация данной таблицы проста. Столбец «Измерен Classif» (рис.3.15) содержит априорную информацию о принадлежности пробы к одной из двух групп *A* (G_1:1) или *B* (G_2:2).

В столбце «(G_1:1)» и в столбце «(G_2:2)» даны апостериорные вероятности отнесения каждой объекта к определенной группе. Объект относится к группе с максимальной апостериорной вероятностью.

В таблицах (рис. 3.11, рис. 3.14, рис. 3.15) знаком * отмечаются неправильно расклассифицированные объекты относительно исходно заданного группирования.

Случай	Последующие вероятности (Пример) Incorrect classifications are marked with * Измеренн Classif.	G_1:1 p=.41975	G_2:2 p=.58025	
28	G_1:1	0.999338	0.000662	
29	G_1:1	0.862378	0.137622	
30	G_1:1	0.969436	0.030564	
* 31	G_1:1	0.400000	0.600000	
32	G_1:1	0.971782	0.028218	
33	G_1:1	0.999971	0.000029	
34	G_1:1	0.992091	0.007909	
* 35	G_2:2	0.826495	0.173505	
36	G_2:2	0.019396	0.980604	
37	G_2:2	0.000861	0.999139	
38	G_2:2	0.000084	0.999916	

Рис. 3.15. Таблица апостериорных вероятностей

3.2.2. Критерий Хотеллинга T^2 пакете STATISTICA

Рассчитаем критерий *Хотеллинга* T^2 для примера, показанного в табл. 3.1. Критерий считается по формуле (3.3) .

В пакете *STATISTICA* выбираем пункт **Статистика,** далее подпункт **Основная статистика/таблицы => t- test, indeperent, by groups. OK.**

Высвечивается диалоговое окно, которое после заполнения необходимых пунктов (рис. 3.16) и нажатия кнопки **SUMMARY** проводится расчет критерия *Хотеллинга* (рис. 3.17).

Из таблицы расчетов (рис. 3.17) видно, что T^2 = 220,41 значительно превышает критическое значение F = 108,81 при ν_1=2 и ν_2 =78, p < 0,00001.

Вывод: Критерий T^2 *Хотеллинга* дает основание предполагать, что группы (табл. 3.1) «различимы».

Этот вывод не противоречит результатам дискриминантного анализа, где точность равна 92,59 % (рис. 3.12).

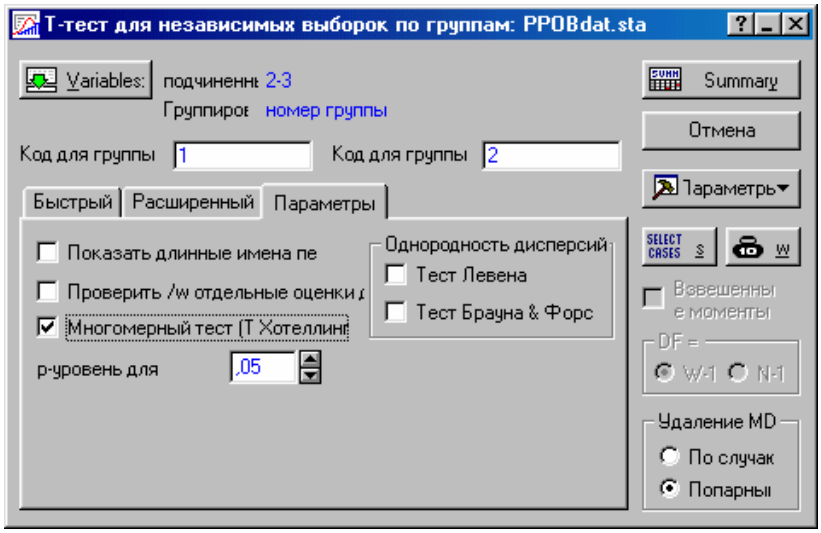


Рис. 3.16. Диалоговое окно для подготовки к расчету T^2 критерия Хотеллинга

T-tests; Группированомер группы (PPOBdat.sta)
Group 1: 1; Group 2: 2
T?(casewise MD)=220,413 F(2,78)=108,81 p<0,0000

Переменная	Mean 1	Mean 2	t-value	df	p	Valid N 1	Valid N 2
средний размер зерен	0,329706	0,339851	-8,33500	79	0,000000	34	47
коэффициент отсортированности	1,167353	1,210000	-3,93935	79	0,000175	34	47

Рис. 3.17. Результат расчета t–критерия Стьюдента [33] и критерия Хотеллинга.

3.2.3. Применение дискриминантного анализа при количестве групп более двух

В этом параграфе рассмотрим пример, в котором количество групп более двух.

Он интересен тем, что в нем используется значительно широкий набор признаков, достаточное количество объектов, которые довольно уверенно классифицируются. Исходные данные заимствованы из работы [17].

Для проверки классификации четырех сформированных групп был привлечен дискриминантный анализ.

В матрице обучающей информации (табл. 3.3) содержатся значения в баллах 8 признаков и девятый – группировочный признак, указывающий, к какой группе относится объект. Группа 1 состоит из 28 объектов, ее группировочный признак – 1. Группа 2 – из 25 объектов, объединенных в группу с признаком 2. Группа 3 содержит 26 наблюдений, ее группировочный признак – 3. Группа 4 – 24 объекта, группировочный признак – 4.

Таблица 3.3

Массив обучающей информации

Группировочный признак	**ПРИЗНАКИ**							
	x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**
1	2	3	1	2	1	2	2	2
1	2	2	2	2	1	2	0	2
1	2	3	1	3	1	2	2	2
1	2	2	3	1	1	0	2	2
1	2	3	2	2	1	2	2	0
1	2	3	1	3	1	0	2	2
1	2	2	2	2	0	2	0	2
1	2	4	1	3	1	2	2	2
1	1	2	2	3	1	2	2	2
1	2	3	2	2	1	2	2	2
1	2	1	1	3	1	2	2	0
1	2	3	2	2	1	2	2	2
1	2	3	1	3	1	2	0	2
1	2	3	2	2	0	0	2	2
1	2	4	2	2	1	2	2	2
1	2	2	1	3	1	2	2	2
1	2	3	3	2	1	2	0	2
1	1	1	2	2	1	2	2	2
1	2	3	2	3	0	2	2	2
1	2	1	1	3	1	0	2	2
1	2	3	3	2	1	2	2	2
1	2	3	2	3	1	2	2	2
1	2	2	1	2	1	2	0	2
1	2	3	2	2	0	2	2	2
1	2	3	1	2	1	2	2	2
1	2	3	2	3	1	2	2	2
2	2	3	1	3	1	2	2	2
2	2	3	1	2	1	2	2	0
2	2	3	1	2	1	2	2	2
2	1	4	2	1	0	2	0	2

2	2	3	1	3	1	0	2	2
2	1	4	2	2	1	2	2	2
2	2	4	1	2	0	2	2	0
2	2	4	2	2	1	2	0	2
2	1	2	1	2	1	2	2	2
2	2	4	2	3	0	0	2	2
2	1	3	1	1	1	2	0	0
2	2	4	1	2	1	2	2	2
2	2	4	1	3	0	2	2	2
2	1	3	1	2	1	0	0	0
2	2	4	1	3	1	2	2	2
2	1	4	1	1	1	2	2	2
2	2	3	1	2	0	2	0	0
2	2	4	1	2	1	0	2	2
2	1	4	2	2	1	2	2	0
2	2	4	1	3	0	2	2	2
2	2	3	1	2	1	2	0	2
2	1	4	2	1	1	0	2	0
2	2	3	1	2	0	2	2	2
2	2	4	1	2	1	2	2	2
2	2	4	2	2	1	2	2	2
3	2	4	2	3	1	0	2	2
3	1	3	2	2	1	2	2	2
3	1	3	1	2	1	0	2	2
3	2	4	1	1	0	2	0	2
3	2	3	1	2	1	0	2	2
3	2	4	2	2	1	2	0	0
3	1	2	1	1	0	0	2	2
3	2	3	1	3	1	2	2	0
3	2	4	1	2	1	2	2	2
3	2	1	1	1	1	2	2	0
3	1	4	1	2	0	0	0	2
3	2	1	2	2	1	2	2	0
3	2	3	1	1	1	2	0	2
3	2	4	1	2	1	0	0	0
3	1	3	1	1	0	2	2	0
3	1	4	1	2	1	0	2	2
3	2	3	2	2	1	2	2	2
3	2	4	1	1	0	0	2	0
3	1	3	1	2	1	2	2	0
3	2	4	2	2	1	2	0	2
3	2	3	1	3	0	0	2	2
3	2	4	1	2	1	0	0	0

3	2	3	1	1	1	2	2	0
3	1	3	1	2	0	2	2	2
3	2	4	1	2	1	2	2	2
3	2	4	2	1	1	2	2	0
4	1	3	1	2	1	2	2	0
4	2	4	1	2	1	2	2	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0
4	1	3	1	1	1	0	0	0
4	2	1	1	2	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	1	0	0	0
4	1	2	1	2	0	0	0	0
4	2	1	1	1	0	0	0	0
4	1	2	1	2	1	2	0	0
4	1	2	1	2	1	0	0	0
4	1	1	1	2	0	0	0	2
4	1	1	2	1	0	0	2	0
4	1	4	1	1	0	0	0	0
4	1	3	1	1	0	0	0	0
4	2	1	1	2	1	0	0	0
4	1	4	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	2	1	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0

Перед использованием дискриминантного анализа (табл.3.3) рассчитаем значения критерия *Хотеллинга* T^2 (см. параграф 3.2.2).

Таблица 3.4

Парные значения T^2 критерия *Хотеллинга* для групп табл.3.3

ГРУППЫ	T^2	***F***	***p***
1 – 2	76,49	8,24	< 0,000001
1 – 3	74,20	8,02	< 0,000001
1 – 4	635,11	68,75	< 0,000001
2 – 3	12,62	1,35	0,24
2 – 4	251,77	26,78	< 0,000001
3 – 4	246,99	26,37	< 0,000001

При анализе табл. 3.4 видно, что несколько обособленно по T^2 критерию расположена группа 4, при проверке на «делимость» могут возникнуть «проблемы» у групп 2 и 3 (по всей видимости, эти группы близки по выбранным показателям).

Д искриминантный анализ выполняется в три этапа.

На первом этапе формируется обучающая информация. Определение групп классификации, классифицирующих признаков осуществляется специалистом. Отбор объектов в матрицу наблюдений производится для этой цели из первичного геолого-геофизического и геохимического материала. Достоверность обучающей информации определяет надежность решающих правил классификации.

На втором этапе вырабатываются решающие правила и дается оценка их информативности. Программа *STATISTICA* обеспечивает отбор информативных признаков и получение решающих правил в виде линейных классификационных функций (ЛКФ) и канонических линейных дискриминантных функций (КЛДФ). ***Качество выработанных правил оценивается сопоставлением результатов классификации с исходной классификацией объектов в обучающей матрице.***

Итак, с помощью дискриминантного анализа попробуем определить по обучающей выборки:

1. Информативность признаков, включенных и не включенных в линейные дискриминантные функции при F для ввода = 2,00 , F для удаления = l,90 (обычно F для ввода и F для удаления задаются исследователем).

2. Коэффициенты линейных классификационных функций (ЛКФ).

3. Вклад ЛДФ в дисперсию признаков.

4. Коэффициенты канонических ЛДФ.

5. Факторную структуру канонических ЛДФ.

6. Координаты центроидов групп.

7. График положения объектов групп.

8. Классификационную матрицу с оценками чувствительности классификации групп обучающей информации.

Устанавливаем флажок (√) в опции «Расширенные опции (пошаговый анализ)» для использования «пошагового» варианта дискриминантного анализа (рис. 3.6). **OK.** Определяем схему ре-

шения как «Вперед пошагово» [14].Устанавливаем в строках *F* для ввода величину = 2,00 и в *F* для удаления – 1,90. Проводим решение по схеме, предложенной в параграфе 3.2.1 для демонстрационного примера.

Здесь отметим лишь, что оценка информативности признаков оценивается по *F*-критерию Фишера:

$$F = S_b^2 \Big/ S_w^2 ,$$

где S_b^2 – межгрупповая дисперсия признаков; S_w^2 – внутригрупповая дисперсия признака. Очевидно, чем больше S_b^2 и меньше S_w^2 тем больше классификационная информативность признака.

В модель включаются признаки, для которых уровень значимости по *F*-критерию $p < 0{,}05$ (уровень *p* может так же задать исследователь).

Оценки информативности признаков, включенных в ЛДФ, показаны на рис. 3.18.

Оценку информативности признаков, не включенных в ЛДФ можно получить с помощью диалогового окна и кнопки «Переменные не в модели» (рис. 3.7). Не информативным оказался признак X5.

Коэффициенты линейных классификационных функций приведены на рис. 3.19.

Discriminant Function Analysis Summary (append)
Шаг
Wilks' Lambda: .14241 approx. F (21,267)=12.379 p<0.0000

N=103	**Wilks' Лямбда**	Частичны Лямбда	F-remove (3,93)	p-level	Toler.	1-Toler. (R-Sqr.)
X2	0.206628	0.689195	13.97998	0.000000	0.987019	0.012981
X8	0.164277	0.866874	4.76068	0.003926	0.940174	0.059826
X6	0.178106	0.799563	7.77118	0.000110	0.985165	0.014835
X7	0.175280	0.812453	7.15604	0.000224	0.959078	0.040922
X1	0.159012	0.895578	3.61452	0.016099	0.928921	0.071079
X3	0.163186	0.872669	4.52319	0.005250	0.928465	0.071535
X4	0.158660	0.897564	3.53792	0.017703	0.818008	0.181992

Рис. 3.18. Оценки информативности признаков, включенных в ЛДФ

Расчет проводится аналогично демонстрационному примеру (параграф 3.2.1, формулы (3.5)), т. е.

$F1 = – 38{,}13+3{,}99 \cdot X2+1{,}56 \cdot X8+3{,}12 \cdot X6+$ (3.6)
$2{,}63 \cdot X7+9{,}62 \cdot X1+9{,}08 \cdot X3+7{,}05 \cdot X4$

$F2 = – 34{,}32 +$ и т. д.

$F3 = – 28{,}95 +$ и т. д.

$F4 = – 13{,}97 +$ и т. д.

Объект будет относиться к той группе, где $max F_i (i = 1, k)$, $к$ – количество групп. В примере $к = 4$.

На рис. 3.20 показано сопоставление результатов классификации по ЛКФ (формулы 3.6, рис. 3.19) с исходной классификацией объектов в обучающей выборке (табл. 3.3).

	Classification Functions; grouping: код (Spread			
Переменная	**G_1:1 p=.27184**	G_2:2 p=.24272	G_3:3 p=.25243	G_4:4 p=.23301
x2	3.9903	5.5440	5.0520	2.7377
x8	1.5559	1.1522	0.4412	-0.9636
x6	3.1234	2.7864	2.3093	0.0605
x7	2.6317	2.5840	2.6036	0.0284
x1	9.6156	7.9309	8.6685	5.8046
x3	9.0799	7.4986	6.6053	6.3172
x4	7.0495	6.1204	4.9821	4.6856
Постоянн	-38.1305	-34.3198	-28.9478	-13.9699

Рис. 3.19. Коэффициенты линейных классификационных функций

	Classification Matrix (Аппендицит 28.04 и КАН ПЕР И ЦЕ Rows: Observed classifications Columns: Predicted classifications				
Группа	**Процент Исправле**	G_1:1 p=.27184	G_2:2 p=.24272	G_3:3 p=.25243	G_4:4 p=.23301
G_1:1	82.1429	23	4	1	0
G_2:2	64.0000	3	16	4	2
G_3:3	61.5385	3	7	16	0
G_4:4	100.0000	0	0	0	24
Итог	76.6990	29	27	21	26

Рис. 3.20. Оценка чувствительности решающих правил для табл. 3.3

Более подробно рассмотрим *канонический анализ*, предусмотренный в пакете *STATISICA* (в 3.2.1 канонический анализ не проводился).

С помощью канонического анализа рассчитываются канонические переменные, суть которых показана в главе 2.

Решение классификационной задачи по каноническим линейным дискриминантным функциям (КЛДФ)

Для решения классификационной задачи по каноническим уравнениям надо нажать кнопку «Выполнить канонический анализ» в диалоговом окне «Результаты дискриминантного анализа» (рис. 3.7). На экране монитора высветиться окно стартовой модели модуля « Канонический анализ «, используя это окно можно рассчитать все элементы канонического анализа, начиная с таблицы на рис. 3.21 и т. д.

Для всех групп определяются канонические ЛДФ, обобщающие данные обо всех признаках, включенных в модель, по всем объектам, находящимся в обучающей матрице наблюдений. Первая КЛДФ1 охватывает наибольшую часть дисперсии признаков, вторая КЛДФ2 – наибольшую часть из оставшихся дисперсий признаков и т. д.

Вклад КЛДФ в межгрупповую дисперсию симптомов (Eigenvalue) оценивается по χ^2 – критерию Пирсона. Этот вклад признается значимым при уровне значимости $p < 0{,}05$.

В рассматриваемом примере значимыми получены две КЛДФ, обозначенные K1 и K2 (в отличие от F1 и F2 для ЛДФ), о чем свидетельствуют данные собственных вкладов функций (рис. 3.21).

На рис. 3.22 даны коэффициенты КЛДФ, их собственные вклады и кумулятивный вклад (Cum.Prop). Так, K1 и K2 (Корен1 и Корен2 в пакете *STATISTICA*) обобщили дисперсию всех признаков на 98,07 % (0,9807). Там же приведены формулы для расчета K1 и K2. По этим формулам (рис. 3.22) программой предусмотрен расчет K1 и K2 для всех объектов обучающей информации.

По таблице факторной структуры КЛДФ (рис. 3.23), судят о корреляционной связи наблюдавшихся признаков (переменных), включенных в модели с каноническими ЛДФ. С первой канонической ЛДФ более тесно связаны признаки X8, X6 и X7, со второй

канонической ЛДФ – признаки X2 и X3. Данные о факторной структуре канонических ЛДФ могут использоваться для оценки коэффициентов «весомости» признаков при решении классификационной задачи.

Roots Удален	Chi-Square Tests with Successive Roots Removed (Spreadsheet2					
	Eigen-value	Canonicl R	Wilks' Лямбда	Chi-Sqr.	df	p-level
0	3.228870	0.873802	0.142407	188.0848	21	0.000000
1	0.546061	0.594302	0.602221	48.9381	12	0.000002
2	0.074032	0.262543	0.931071	6.8920	5	0.228797

Рис. 3.21. Оценки вкладов канонических ЛДФ в дисперсию признаков

По данным о координатах объектов в группах производится расчет координат центроидов для каждой группы (Means of Canonical Variables – рис. 3.24). По этим координатам центроиды наносят на график. От них измеряется удаление до точки обследуемого объекта, которую наносят на график после расчета К1 и К2 по признакам обследуемого объекта. Объект относят к той группе, от центра которой получено **наименьшее** удаление.

График положения объектов четырех групп в координатах первой и второй КЛДФ показан на рис. 3.25.

На третьем этапе непосредственно решается задача классификации по выработанным решающим правилам.

После обследования объекта (определения признаков, включенных в ЛКФ или КЛДФ) рассчитываются эти функции и по их величинам дается решение об отнесении объекта к той или иной группе из заданных. Если используются ЛКФ, то отнесение больного к определенной группе выполняется по **максимальному** значению ЛКФ после их расчета по набору признаков для каждой группы.

В примере из табл. 3.3 используются формулы (3.6).

Переменная	Raw Coefficients (Аппендицит 2 лист) for Canonical Variables		
	Корен1	Корен2	Корен3
x2	0.44336	0.96548	-0.410132
x8	0.52762	-0.21859	-0.361847
x6	0.66641	-0.02595	0.052705
x7	0.60120	0.28604	0.583331
x1	0.72677	-0.35793	1.812263
x3	0.45638	-1.01108	-0.508685
x4	0.42646	-0.72975	-0.927936
Постоянн	-5.97444	0.53874	0.324398
Eigenval	3.22887	0.54606	0.074032
Cum.Prop	0.83889	0.98077	1.000000

Рис. 3.22. Коэффициенты канонических ЛДФ

Переменная	Factor Structure Matrix (append) Correlations Variables - Canonical Roots (Pooled-within-groups correlations)		
	Корен1	Корен2	Корен3
X2	0.363348	0.741838	-0.338580
X8	0.462416	-0.204541	-0.383504
X6	0.457627	-0.010467	0.106020
X7	0.451776	0.110535	0.339704
X1	0.363104	-0.184618	0.584948
X3	0.162064	-0.405078	-0.190685
X4	0.379866	-0.321412	-0.270646

Рис. 3.23. Факторная нагрузка канонических ЛДФ

Группа	Means of Canonical Variables (append)		
	Корен1	Корен2	Корен3
G_1:1	1.45775	-1.00819	0.063801
G_2:2	1.05004	0.59759	-0.385115
G_3:3	0.28135	0.75831	0.361884
G_4:4	-3.09929	-0.26776	-0.065314

Рис. 3.24. Координаты центроидов

При применении КЛДФ также производится расчет КЛДФ по значению признаков конкретного объекта. Отнесение объекта к определенной группе выполняется после нанесения значений рассчитанных КЛДФ на классифицирующий график. Объект относят к той группе, для которой его удаление от соответствующего центроида окажется **минимальным**.

Остановимся на классификации (третий этап) каноническими ЛДФ подробнее.

Применение решающих правил классификации объектов с помощью КЛДФ К1 и К2 и диагностического графика (рис. 3.25) покажем на примере для двух новых (не входящих в обучающую выборку – табл. 3.3) обследованных объектов. Для этого представим графики с положением центроидов четырех групп. Координаты центроидов возьмем из рис. 3.24.

В результате обследования двух объектов получены оценки в баллах по всем семи переменным (табл. 3.5), включенным в модели для двух канонических ЛДФ (рис. 3.22).

Таблица 3.5

Оценка признаков в баллах

№ обследованного объекта	**X2**	**X8**	**X6**	**X7**	**X1**	**X3**	**X4**
1	3	2	2	2	2	1	2
2	2	0	0	0	1	1	1

По формулам К1 и К2 (рис. 3.22) рассчитаны координаты обследованных:

обследованного объекта № 1

К1 = -5,97+0,44·3+0,52·2+0,66·2+0,60·2+0,72·2+0,45·1+0,42·2= = 1,70

К2 = 0,53+0,96·3- 0,21·2- 0,02·2+0,28·2- 0,35·2- 1,01·1+0,72·2 = = 0,33;

обследованного объекта №2

К1 = -5,97+0,44·2+0,52·0+0,66·0+0,60·0+0,72·1+0,45·1+0,42·1= = -3,47

К2 = 0,53+0,96·2- 0,21·0- 0,02·0+0,28·0- 0,35·1- 1,01·1+0,72·1 = = 0,37.

Данные по обследованным объектам нанесены по значениям К1 и К2 на график (рис. 3.25). По наименьшему удалению от центроидов установлено, что обследованные объекты следует отнести к группам:

- первого обследованного объекта – к группе 2;
- второго обследованного объекта – к группе 4 .

В заключение по результатам расчетов примера необходимо отметить, что:

1. Предварительная оценка групп по критерию *Хотеллинга* (табл. 3.4) не противоречит результатам, полученным с применением более мощного статистического аппарата – дискриминантного аппарата (рис. 3.20);

2. Информативными признаками являются Х2, Х8, Х6, Х7, Х1, Х3, Х4;

3. ЛКФ рассчитываются по уравнениям (рис. 3.19);

4. Отнесение объекта к определенной группе выполняется на основе расчета ЛКФ (формулы 3.6). Это наиболее точный способ классификации;

5. Для решения задачи классификации можно применить две первые канонические ЛДФ с суммарным вкладом в дисперсию 98,07 % (рис. 3.22);

6. КЛДФ рассчитываются по уравнениям (рис. 3.22);

7. Отнесение объекта к определенной группе можно выполнить на основе расчета КЛКФ (формулы расчета К1 и К2) и использования графика положения центроидов групп (рис. 3.25). Этот способ классификации более прост в расчетах (по сравнению с расчетами ЛКФ) и визуально более понятен;

8. Из данных таблицы (рис. 3.20) видно, что точность классификации по решающим правилам в среднем имеет достоверность 76,7 %, для первой группы – 82,1 %, второй – 64,0 %, третьей – 61,5 %, четвертой – 100 %. Недостаточность точности классификации для второй и третьей групп объясняется перекрытием признаков для этих групп.

Случай	Последующие вероятности (Пример) Incorrect classifications are marked with * **Измеренн Classif.**	G_1:1 p=.41975	G_2:2 p=.58025
28	G_1:1	0.999338	0.000662
29	G_1:1	0.862378	0.137622
30	G_1:1	0.969436	0.030564
* 31	G_1:1	0.400000	0.600000
32	G_1:1	0.971782	0.028218
33	G_1:1	0.999971	0.000029
34	G_1:1	0.992091	0.007909
* 35	G_2:2	0.826495	0.173505
36	G_2:2	0.019396	0.980604
37	G_2:2	0.000861	0.999139
38	G_2:2	0.000084	0.999916

Рис. 3.25. График положения центроидов для четырех групп (рис.3.24) и обследуемых №1 и №2 по рассчитанным значениям K1 и K2

4. ЗАДАЧИ И УПРАЖНЕНИЯ

1. *Модельный пример.* Здесь в качестве «модельного примера» для закрепления полученных знаний по многомерным статистическим методам и навыкам работы на PC использованы результаты исследования *ГРАНИТОИДОВ УНДИНСКОГО КОМПЛЕКСА* [13].

Прежде чем перейти к заданиям по модельному примеру, нужно будет ознакомиться с геологической основой примера, для того чтобы, помимо стандартных результатов, полученных с помощью статистических методов (в основном дискриминатного анализа), выдвинуть геологические гипотезы, которые можно было бы подтвердить или отвергнуть с помощью математического решения.

Краткое описание изученности петрохимии гранитоидов ундинского комплекса

На северо-западе Восточного Забайкалья широко распространены средне-верхнепалеозойские гранитоиды, слагающие крупный Ундино-Шилкинский батолит. В районе Газимурского Завода И. Н. Тихомиров [45] в составе этих гранитоидов выделяет три интрузивных комплекса:

1) среднепалеозойский раннеорогенный кручиненский комплекс габбро и диоритов;

2) более молодой среднепалеозойский газимурозаводский комплекс синорогенных полосчатых гранитов, гранодиоритов и кварцевых диоритов;

3) позднеорогенный каменноугольный верхнеундинский комплекс, к которому относится Ундинский плутон, являющийся крайним юго-западным окончанием Ундино-Шилкинского батолита.

Массив сформировался как ядерный геосинклинальный плутон в подзоне центрального прогиба палеозойской геосинклинали Восточного Забайкалья. Слагающие плутон породы обнажены в корродированном ядре Ундинского антиклинория и по периферии плутона трансгрессивно перекрыты морскими отложениями октагаинской и верхнегазимурской свит нижне-среднеюрского возраста. О характере экзо- и эндоконтактовых процессов можно судить лишь на основании изучения взаимоотношений гранитоидов с ксенолитами вмещающих пород, а ундуляция шарнира Ундинской антиклинали позволяет наблюдать на дневной поверхности как более насыщенные слабо переработанными ксенолитами апикальные части плутона, так и более глубинные фации с меньшим содержанием сильно измененных ксенолитов. Комплекс представлен двумя резко различными группами пород:

1. Гранитоиды пестрого состава — порфировидные граниты, грано-диориты, адамеллиты, плагиограниты, кварцевые диориты, диориты, тесно связанные *постепенными переходами друг с другом.* В них чрезвычайно широко распространены крупные и мелкие ксенолиты нижне-среднепалеозойских осадочных образований и магматических пород основного состава (долериты, габбро, габбро-диориты). Параксенолиты представлены мраморизованными известняками, двуслюдяными, андалузитовыми и кордиеритовыми кристаллическими сланцами, гнейсами, кварцитами, ороговикованными песчаниками. Ортоксенолиты встречаются чаще параксенолитов и по размерам не превышают первых сотен квадратных метров. Основная масса их связана с порфировидными гранитами и гранодиоритами серией постепенных по характеру контактов и по петрографическим и структурным особенностям разностей. Останцы метаморфизованных осадочных пород чаще образуют крупные, до нескольких квадратных километров по площади, блоки, контрастно отделяющиеся от окружающих их гранитоидов. *Постепенные переходы к гранитам характерны для небольших ксенолитов ороговикованных песчаников.*

2. Биотитовые и лейкократовые граниты, адамеллиты. Эти породы прорывают гранитоиды пестрого состава и отличаются от них четко выраженной штоко- или дайкообразной формой тел и постоянством состава.

Возрастные взаимоотношения названных двух основных групп обычно не вызывают сомнений. *Более сложно устанавливается последовательность образования отдельных разновидностей в группе гранитоидов пестрого состава*. Традиционные схемы расчленения предусматривают выделение *последовательных интрузивных фаз*, начиная с габброидов и кончая порфировидными гранитами. Детальное изучение минералогии, петрографии и геохимии пород ундинского комплекса позволило в качестве наиболее вероятной гипотезы принять, что, во-первых, габброиды и граниты не являются членами единого эволюционного ряда и кристаллизовались на разных тектоно-магматических этапах; интрузия основных пород предшествовала образованию гранитоидов и должна быть отнесена к более раннему, нижнесреднепалеозойскому интрузивному комплексу.

И, во-вторых, гранитоиды пестрого состава формировались в один этап в результате взаимодействия гранитного расплава с вмещающими толщами песчаников, сланцев и известняков, пронизанных телами габброидов.

При формировании пестрых гранитоидов ундинского комплекса вещество ксенолитов претерпело *три последовательных этапа преобразования*: метаморфизм, метасоматоз и дезинтеграцию. Метаморфизм пород, вмещающих ундинские гранитоиды, носит прогрессивный характер с образованием ассоциаций минералов, типичных для амфибол-роговиковой фации контактового метаморфизма. В непосредственной близости от гранитов в кварцево-полевошпатовых роговиках видны следы плавления, выражающиеся в образовании мелких изометричных изолированных участков крупнозернистого гранита гипидиоморфнозернистой структуры. Метасоматическое замещение ксенолитов и их дезинтеграция устанавливаются в основном при алюмосиликатной гибридизации гранитов основными магматогенными породами и рассмотрены выше.

Для выяснения петрохимических особенностей гранитоидов Ундинского массива совершенно недостаточно располагать анализами только наиболее типичных разновидностей. Разнообразие петрографического состава пород настолько велико, что едва ли не каждый штуф характеризуется своими особенностями состава и структуры. Применение квантометрического метода определе-

ния химического состава пород позволяет значительно расширить число анализируемых образцов. Непосредственно на квантометре определялось содержание семи главных породообразующих компонентов, а именно: кремния, титана, алюминия, железа, марганца и т. д. Группировка анализов, согласно петрографической номенклатуре, проведена путем сравнения химического состава пород со средними составами гранитоидов Советского Союза, по данным А. А. Беуса, А. А. Ситнина [6].

В табл. 4.1 приведены результаты анализа пород ундинского комплекса.

Таблица 4.1

Результаты квантометрического определения химического состава пород Ундинского комплекса

Номер пробы	**SiO_2**	TiO_2	**Al_2O_3**	**Fe_2O_3**	**MnO**	**Mg0**	**CaO**	**Na_2O**	**K_2O**	**код 1**	**код 2**
5-2	54,0	0,98	21,40	7,20	0,22	3,00	5,60	3,72	1,87	1	1
5-11	52,0	0,98	21,30	8,90	0,27	3,80	7,10	3,00	1,77	1	1
5-91	54,0	0,84	17,30	9,00	0,33	3,80	5,80	3,44	1,72	1	1
5-151	55,0	0,96	19,20	8,00	0,28	3,40	7,70	3,30	2,03	1	1
5-153	56,0	0,82	18,50	9,00	0,28	3,60	5,50	3,36	1,74	1	1
5-167	52,0	0,74	19,00	8,60	0,26	5,30	9,00	2,42	1,32	1	1
5-168	55, 0	0,40	18,80	7,20	0,22	4,20	8,70	3,05	1,09	1	1
5-193	55,0	0,80	18,80	11,60	0,32	2,80	5,20	3,59	2,16	1	1
5-223	56, 0	1,00	18,30	9,60	0,29	3,60	5,70	3,88	2,03	1	1
5-340	55, 0	0,74	16,10	8,70	0,28	3,60	6,50	3,33	1,52	1	1
5-341	55, 0	0,86	18,20	9,90	0,29	4,20	6,30	3,40	1,77	1	1
5-78	60,0	0,70	17,60	7,30	0,26	2,30	4,60	3,35	2,55	2	2
5-92	63,0	0,53	15,90	6,20	0,13	2,40	5,00	3,48	2,32	2	2
5-93	60,0	0,72	19,00	6,50	0,20	2,70	4,50	3,50	1,96	2	2
5-94	63,0	0,66	14,60	7,40	0,19	3,63	5,00	2,79	2,19	2	2
5-138	58,0	0,74	17,80	9,30	0,26	3,10	5,50	3,68	1,82	2	2
5-140	60,0	0,74	18,20	8,30	0,28	3,00	5,10	3,00	1,67	2	2
5-170	60,0	0,60	19,10	6,70	0,19	2,86	4,70	3,42	2,11	2	2
5-271	58,0	0,72	16,50	7,90	0,31	3,30	6,20	3,48	2,38	2	2

5-300	57,0	0,78	20,50	7,10	0,22	2,20	5,95	4,06	2,18	2	2
5-337	61,0	0,62	17,20	6,90	0,23	2,87	5,00	3,19	1,73	2	2
5-338	58,0	0,74	19,20	8,50	0,26	2,80	5,40	3,80	1,85	2	2
5 _ 20	64,0	0,56	17,6	5,8	0,16	1,9	4,25	4,51	1,33	3	3
5 _33	62,0	0.7	16.,6	6,9	0,19	3,1	4,5	3,6	2,18	3	3
5_59	67,0	0,44	16,80	5,50	0,15	1,70	3,60	3,23	1,38	3	3
5_61	62,0	0,52	20,50	6,30	0,17	1,60	3,80	3,81	1,96	3	3
5_69	60,0	0,70	16,30	7,80	0,30	2,25	4,00	4,31	2,42	3	3
5_79	62,0	0,52	17,20	8,70	0,24	2,20	3,70	3,44	2,55	3	3
5_97	61,0	0,62	19,90	5,90	0,18	2,30	3,80	3,20	2,42	3	3
5_127	63,0	0,54	18,20	6,75	0,12	1,80	3,60	4,13	1,76	3	3
5_134	62,0	0,46	18,50	7,10	0,22	1,60	4,00	4,40	1,82	3	3
5_144	64,0	0,60	17,60	5,20	0,20	2,40	3,50	4,05	1,94	3	3
5_159	63,0	0,58	18,80	5,10	0,18	1,80	3,60	3,90	2,51	3	3
5_183	61,0	0,88	18,10	7,20	0,21	2,40	4,00	4,10	2,31	3	3
5_219	56,0	0,70	21,60	6,70	0,19	2,20	5,20	4,65	2,47	3	3
5_227	62,0	0,65	17,70	6,70	0,20	2,30	4,40	4,20	1,92	3	3
5_230	65,0	0,54	16,10	5,95	0,22	1,85	3,46	3,77	3,15	3	3
5_269	61,0	0,72	18,10	6,90	0,26	2,20	4,50	3,78	2,41	3	3
5_01	66,0	0,44	14,70	4,80	0,15	2,00	2,80	3,44	2,35	4	4
5_02	65,0	0,50	18,40	4,60	0,16	1,80	3,50	3,44	2,26	4	4
5_03	65,0	0,56	17,50	5,70	0,22	1,60	3,20	4,10	2,37	4	4
5_04	65,0	0 48	17,30	5,70	0,17	1,50	2,90	3,65	2,55	4	4
5_05	66,0	0,54	16,50	5,30	0,18	1,60	3,20	3,68	2,13	4	4
5_06	66,0	0,40	16,80	4,30	0,12	1,60	3,10	3,92	1,95	4	4
5_07	67,0	0,54	16,00	5,23	0,20	2,20	3,10	3,45	1,90	4	4
5_08	66,0	0,48	16,50	5,10	0,15	1,45	2,60	3,28	3,02	4	4
5_09	63,0	0,54	18,90	5,20	0,18	1,20	3,10	3,60	3,49	4	4
5_010	65,0	0,46	19,10	4,70	0,15	1,33	3,40	3,65	3,28	4	4
5_011	65,0	0,40	18,20	3,70	0,14	1,30	3,30	3,47	3,69	4	4
5_012	66,0	0,52	16,80	5,20	0,21	1,90	2,90	3,75	2,67	4	4
5_013	66,0	0,54	15,00	6,10	0,20	2,00	3,50	3,39	2,94	4	4

5_014	67,0	0,22	15,60	4,80	0,15	1,50	2,80	3,59	3,54	4	4
5_015	66,0	0,52	16,20	6,40	0,20	1,90	3,00	2,92	2,34	4	4
5_016	64,0	0,36	17,60	4,30	0,11	1,20	2,70	3,60	3,83	4	4
5_017	63,0	0,58	18,50	6,50	0,19	1,60	3,20	3,85	3,33	4	4
5_018	55,0	0,60	17,60	5,30	0,21	1,70	2,90	3,92	3,09	4	4
5_019	69,0	0,74	17,70	7,60	0,21	2,80	3,30	4,67	2,55	4	4
5_020	66,0	0,36	17,50	4,90	0,14	1,00	3,10	4,21	2,28	4	4
5_021	74,0	0,22	14,30	3,40	0,12	0,70	2,90	3,50	1,43	4	4
5-253	62,0	0,60	18,10	7,80	0,28	2,00	3,20	3,85	2,63	4	4
5-260	63,5	0,38	19,00	4,30	0,80	1,25	2,80	5,00	2,17	4	4
5-274	62,0	0,50	19,60	4,80	0,20	1,70	3,00	3,70	1,85	4	4
5-296	72,0	0,20	14,20	2,60	0,11	1,00	3,20	3,19	1,93	4	4
5-328	61,0	0,68	18,20	7,23	0,23	1,90	3,30	3,64	3,23	4	4
5-354	67,0	0,42	16,70	5,80	0,20	1,10	2,90	3,48	2,92	4	4
5-28	66,0	0,36	19,00	3,74	0,12	1,10	2,40	3,66	3,49	5	5
5-30	66,0	0,42	17,10	4,60	0,13	1,50	2,70	4,02	2,86	5	5
5-40	70,0	0,30	16,70	3,05	0,15	0,80	2,50	3,27	3,33	5	5
5-54	67,0	0,44	15,60	4,50	0,13	1,50	2,30	3,68	2,75	5	5
5-55	64,0	0,56	16,80	6,60	0 15	1,65	2,60	3,71	2,42	5	5
5-58	71,0	0,18	15,30	3,60	0,09	0,80	2,30	3,11	3,20	5	5
5-60	70,0	0,24	17,20	3,00	0,13	0,70	1,40	2,87	4,37	5	5
5-129	68,0	0,34	17,70	4,05	0,14	0,90	2,00	3,50	4,15	5	5
5-130	67,0	0,32	18,20	4,90	0,11	1,00	2,10	3,45	3,92	5	5
5-131	66,0	0,48	17,30	6,10	0,21	1,30	2,60	3,45	3,75	5	5
5-156	69,0	0,50	14,50	6,00	0,25	1,20	2,20	3,28	3,36	5	5
5-162	65,0	0,46	18,00	6,30	0,16	1,20	2,70	3,59	3,51	5	5
5-196	65,0	0,46	18,10	5,70	0,19	1,30	2,60	3,59	3,38	5	5
5-221	68,0	0,32	16,10	4,25	0,10	1,10	2,40	3,59	3,62	5	5
5-275	70,0	0,44	15,20	5,00	0,21	1,60	2,20	3,55	1,65	5	5
5-284	68,0	0,30	18,30	4,10	0,15	0,80	2,20	3,66	3,08	5	5
5-286	63,0	0,22	15,60	4,15	0,15	7,60	2,35	3,48	3,42	5	5
5-288	62,0	0,36	15,90	4,20	0,23	7,10	2,60	3,64	4,27	5	5

5-301	68,0	0,39	18,50	3,50	0,17	1,20	2,10	3,52	4,ОЭ	5	5
5-349a	72,0	0,24	15,20	3,20	0,11	0,75	2,20	3,55	3,03	5	5
5-350a	67,0	0,38	16,90	5,40	0,13	1,50	2,70	3,84	2,74	5	5
5-351	69,0	0,28	17,20	4,10	1 0,10	1,00	2,13	3,33	3,20	5	5
5-111	66,0	0,42	15,10	6,70	0,15	1,60	1,70	3,44	3,61	6	6
5-126	67,0	0,32	17,60	5,00	0,14	0,90	1,80	3,60	4,17	6	6
5-205	66,0	0,34	14,70	9,10	0,23	1,00	1,70	3,07	4,30	6	6
5-206	70,0	0,20	15,20	4,80	0,11	1,40	1,35	2,92	3,23	6	6
5-212	68,0	0,34	16,30	4,80	0,20	0,73	1,30	3,65	4,54	6	6
5-255	73,0	0,08	12,80	6,05	0,11	0,20	0,55	2,90	4,56	6	6
5-289	72,0	0,31	15,20	5,35	0,15	1,00	1,35	3,55	4,10	6	6
5-302	67,0	0,50	15,80	5,80	0,21	1,60	2,15	3,64	3,65	6	6
5-102	69,0	0,26	14,80	4,80	0,16	0,50	1,35	3,11	5,10	7	6
5-113	73,0	0,24	13,30	3,70	0,16	0,50	1,10	2,87	4,42	7	6
5-128	73,0	0,22	13,10	4,40	0,09	0,60	1,20	2,48	4,54	7	6
5-182	72,0	0,20	15,00	3,30	0,12	0,50	1,30	3,42	4,10	7	6
5-200	70,0	0,32	14,70	4,40	0,15	0,75	1,70	3,25	4,31	7	6
5^207	74,0	0,18	13,40	2,73	0,13	0,50	1,30	3,45	4,45	7	6
5-218	72,0	0,18	13,70	4,00	0,10	0,40	1,50	3,05	4,80	7	6
5-234	70	0,30	13,80	3,60	0,15	0,75	2,30	3,25	3,82	7	6
5-290	72,0	0,18	14,60	3,60	0,14	0,70	1,50	3,64	4,21	7	6
5-65	73,0	0,08	15,10	2,20	0,08	0,25	0,60	2,72	5,21	8	6
5-163	74,0	0,C6	14,30	2,60	0,12	0,10	0,70	3,59	3,60	8	6
5-185	71,0	0,10	17,00	2,10	0,C9	0,25	0,80	3,42	5,38	8	6
5-252	72,0	0,14	12,80	5,60	0,12	0,30	0,70	2,80	4,85	8	6
5-282	72,0	0,14	14,60	3,85	0,11	0,40	0,80	3,33	4,25	8	6
5_84	68,0	0,36	15,4	6,35	0,12	1,7	2,5	4,57	1,1	9	6
5_114	66,0	0,7	14,2	6,5	0,18	2,2	2,4	3,96	1,74	9	6
5_249	67,0	0,22	18,1	4,1	0,15	2,4	2,3	4,55	1,69	9	6
5_356	70,0	0,26	15,9	4,1	0,13	0,8	1,7	4,55	2,76	10	6
5_360	72,0	0,26	13,8	4,9	0,12	0,80	1,5	3,48	3,35	10	6
5_36	74,0	0,16	14,50	3,4	0,13	0,16	0,6	2,4	5,75	11	11

5_7	72,5	0,22	14,8	3,6	0,14	0,3	0,8	2,48	4,88	11	11
5-77	73,0	0,20	13,20	4,10	0,13	0,50	1,20	2,60	4,35	11	11
5-109	74,0	0,16	13,45	2,40	0,11	0,50	1,10	3,60	4,23	11	11
5-115	72,0	0,24	13,80	3,85	0,14	0,60	1,70	3,45	4,17	11	11
5-120	65,0	0,50	16,30	6,00	0,22	1,10	2,90	3,50	4,62	11	11
5-157	74,0	0,12	16,10	2,00	0,10	0,10	1,00	3,45	4,19	11	11
5-220	74,0	0,05	14,10	3,10	0,07	0,20	0,60	2,75	5,77	11	11
5-224	68,0	0,24	17,00	3,70	0,07	0,60	1,30	3,34	5,15	11	11
5-244	72,0	0,18	14,80	3,60	0,13	0,55	1,30	3,42	4,54	11	11
5-245	72,0	0,06	15,10	2,50	0,08	0,20	1,35	2,73	7,02	11	11
5-246	73,0	0,13	14,30	2,83	0,10	0,30	1,30	2,90	5,10	11	11
5-280	71,0	0,24	15,70	3,10	0,16	0,50	1,10	3,40	4,27	11	11
5-292	72,0	0,12	14,30	3,90	0,05	0,40	1,10	2,44	5,87	11	11
5-303	72,0	0,14	14,60	3,80	0,16	0,40	1,20	3,40	3,98	11	11
5-304	70,0	0,14	18,00	2,90	0,09	0,50	1,60	3,48	3,85	11	11
5_18	73,0	0,20	16,40	2,93	0,12	0,35	1,6	3,56	2,4	12	12
5_90	73,0	0,14	15,20	2,50	0,07	0,6	2,0	3,75	2,37	12	12
5-119	69,5	0,32	15,5	4,8	0,12	0,9	2,03	3,53	3,12	12	12
5-137	70,0	0,28	16,03	3,33	0,08	0,25	2,03	3,92	3,75	12	12
5-169	69,0	0,32	15,03	4,23	0,12	1,0	2,3	3,75	3,23	12	12

Расшифровка данных столбца «код 1» таблицы 4.1 [13]
1 – Диориты.
2 – Кварцевые диориты.
3 – Кальциплетовые гранодиориты.
4 – Нормальные гранодиориты.
5 – Кальциптоховые гранодиориты.
6 – Фемиплетовые граниты первой интрузивной фазы.
7 – Нормальные граниты первой интрузивной фазы.
8 – Кальциптоховые граниты первой интрузивной фазы.
9 – Плагиограниты первой интрузивной фазы.
10 – Адамеллиты первой интрузивной фазы.
11 – Нормальные граниты второй интрузивной фазы.
12 – Адамеллиты второй интрузивной фазы.

Породы первой интрузивной фазы. К названным породам относятся как продукты кристаллизации интрудировавших расплавов, так и их эндоконтактовые гибриды.

Диориты. Распространены преимущественно в фации апикальных частей плутона. Выходы их на дневную поверхность образуют иногда крупные поля. Однако это не индивидуализированные тела выдержанного состава, а пестрая смесь ксенолитов, диоритов и гранодиоритов при преимущественном распространении диоритов. В фации относительно глубинных частей плутона диориты образуют изолированные участки в гранитах и гранодиоритах и *постепенно в них переходят.*

Кварцевые диориты. Чаще всего встречаются в фации апикальных частей плутона, где служат вмещающей средой для изолированных небольших участков основных пород и алюмосиликатных параксенолитов. Кварцевые диориты — *основная переходная разность от апикальных к более глубинным гранитоидам*.

В областях преимущественного развития последних кварцевые диориты чаще слагают изолированные участки, по площади выхода не превышающие первые сотни квадратных метров и не имеющие видимой связи с породами основного состава.

Переходы к гранодиоритам и порфировидным гранитам постепенные, что устанавливается по тесной перемежаемости небольших участков кварцевых диоритов с гранитами и гранодиоритами.

Характерной чертой петрохимии кварцевых диоритов ундинского комплекса является *разнонаправленное* изменение петрохимических характеристик у исследованных образцов, когда по значению одной характеристики порода может быть отнесена к основным, а по величине другой – к кислым гранитоидам.

Гранодиориты. Породы в виде небольших по площади выходов, не превышающих несколько десятков квадратных метров, участков тесно перемежаются с порфировидными гранитами и диоритами. В областях развития апикальных частей плутона гранодиориты являются преобладающей разностью кислых гранитоидов и во много раз по распространенности уступают диоритам. В более глубинных частях плутона выходы гранодиоритов образуют как относительно крупные поля, так и небольшие зонки в гранитах.

Для всех гранодиоритов очень характерна чрезвычайная неоднородность состава и структуры даже в пределах одного образца, что придает породам своеобразную пятнистость. Хорошо фиксируемой особенностью гранодиоритов является существенный и *разнонаправленный разброс* значений петрохимических характеристик.

По петрографическим и минералогическим особенностям гранодиориты разных форм залегания не различаются.

Колебания в содержаниях кварца и биотита наряду с особенностями распределения полевых шпатов приводят к тому, что состав основной массы *эвпорфировых гранитов изменяется от гранитов до гранодиоритов, адамеллитов и плагиогранитов. Выходы однородных гранитов образуют поля относительно выдержанного состава.*

Ксенолиты для них не характерны. Взаимоотношения их с другими гранитоидами первой интрузивной фазы определяются разно *проявленными постепенными переходами*.

Эвпорфировые неоднородные граниты – наиболее распространенная разновидность. В виде относительно небольших участков, насыщенных реликтами переработанных ксенолитов и разнообразными гибридными породами, они распространены повсеместно среди гранитоидов первой интрузивной фазы.

Все порфировидные граниты относятся к петрохимическому типу нормальных гранитов.

Плагиограниты и адамеллиты. Образец 5–114 и образцы адамеллитов являются *переходными разностями* от фемиплетовых эвпорфировых гранитов к кварцевым диоритам и гранодиоритам.

Породы второй интрузивной фазы. К породам данной фазы относятся продукты кристаллизации остаточных гранитоидных расплавов.

Нормальные граниты. По структурным и минералогическим свойствам названные граниты аналогичны слабопорфировидным гранитам первой интрузивной фазы, но отличаются от них морфологией слагаемых ими тел. Это небольшие штоки, дайки и жилы, прорывающие нормальные граниты магматических расплавов.

Адамеллиты. Слагают штоки и дайки подобно нормальным гранитам. Петрохимические особенности те же, что и у адамеллитов первой интрузивной фазы.

Итак, гранитоиды ундинского комплекса образуют *непрерывный петрохимический ряд известково-щелочных пород* от диоритов до гранитов.

Свойство *непрерывности* выражается в наличии *переходных петрохимических разновидностей* между любыми крайними точками области существования составов пород ундинского комплекса.

Другое проявление *непрерывности* – *разнонаправленное* изменение петрохимических характеристик большинства индивидуальных анализов и всех групповых средних, придавшее породам черты разных петрохимических типов. Более подробное описание гранитоидов Ундинского комплекса в работе [13].

Задания по модельному примеру

1. а) По данным (табл. 4.1) провести дискриминантный анализ групп, выделенных на основании главных породообразующих комплексов пород (код 1).

По результатам расчетов (по схеме интерпретации результатов в примере – параграф 3.2.3 данного учебного пособия) определить:

– информативность признаков;

– коэффициенты линейных классификационных функций (ЛКФ);

– коэффициенты канонических ЛДФ;

– факторную структуру канонических ЛДФ;

– квадрат расстояний Махалонобиса между группами;

– классификационную матрицу с оценками чувствительности диагностики групп обучающей информации.

б) С помощью дискриминатного анализа проверить гипотезы о плавных переходах групп пород и разнонаправленности изменения петрохимических характеристик у исследованных образцов, выдвинутые составителями пособия на основании описания изученности петрохимии гранитоидов Ундинского комплекса (В ОПИСАНИИИ ИЗУЧЕННОСТИ ПЕТРОХИМИИ ГРАНИТОИДОВ УНДИНСКОГО КОМПЛЕКСА *«плавность, непрерывность, постепенность, разнонаправленность и т. п. «* выделены курсивом).

2. По данным (табл. 4.1) провести дискриминантный анализ групп, выделенных на основании главных породообразующих

комплексов пород (**код 2**). (В комплексах по **коду 2** проведено практически произвольное укрупнение групп с целью повторения расчетов по дискриминантному анализу).

3. Найти для модельного примера (табл. 4.1) главные компоненты (глава 2 учебного пособия).

По главным компонентам, превышающим 95 % суммарной дисперсии провести дискриминантный анализ (**код 1**) и сравнить с результатом, полученным по заданию **1** пункт **а)** (т. е. провести полностью решение модельного примера по главным компонентам). Для решения *самостоятельно* подготовить набор главных компонент и **код 1** для анализа. Канонический анализ не проводить.

4. По химическому составу пород модельного примера (табл. 4.1), используя правило *Single Linkage* (глава 1 пособия) иерархического объединения в кластер и меры сходства *1 – Person R* построить кластер.

5. По данным модельного примера (табл. 4.1) для образцов пород комплекса, используя правило *Weighted pair group average* иерархического объединения и меры сходства *Euclidean distanse* построить кластер.

4.2. Ответы и решения

1. а)

Discriminant Function Analysis Summary (добр к дискрим)
Шаг
Wilks' Lambda: .01736 approx. F (55,554)=14.119 p<0.0000

N=135	Wilks' Лямбда	Частичны Лямбда	F-remove 11,119	p-level	Toler.	1-Toler. (R-Sqr.)
CaO	0.031970	0.543039	9.103380	0.000000	0.893086	0.106914
K2O	0.030016	0.578380	7.886086	0.000000	0.976662	0.023338
MGO	0.022469	0.772654	3.183142	0.000816	0.945432	0.054568
FE2O3	0.021282	0.815740	2.443624	0.008730	0.861010	0.138990
AL2O3	0.020705	0.838480	2.083953	0.026486	0.996164	0.003836

Рис. 4.1. Оценка информативности признаков (породообразующих компонент), включенных в ЛДФ

По данным таблицы (рис. 4.1) видно, что количество информативных признаков относительно исследуемых двенадцати групп – пять, при этом наиболее информативными являются CaO и K2O ($p < 0{,}0000001$), наименее информативен (среди информативных) Al2O3 ($p=0{,}0265$).

	Classification Functions; grouping: код 1 (добр к дискрим)							
Переменная	G_1:1 p=.08148	G_2:2 p=.08148	G_3:3 p=.11852	G_4:4 p=.20000	G_5:5 p=.16296	G_6:6 p=.05926	G_7:7 p=.06667	G_8:8 p=.03704
CaO	-20.4698	-15.0530	-9.0765	-3.00762	1.71573	10.15673	9.3874	13.9828
K2O	-6.8888	-5.5286	-4.5522	-1.93826	1.28703	3.70614	4.9306	5.8282
MGO	-9.7537	-7.2199	-4.8326	-1.52477	0.62832	1.27119	5.4894	7.3985
FE2O3	-2.5112	-1.6800	-2.0448	-0.86432	1.03932	-1.88940	-0.2606	0.6206
AL2O3	-3.0829	-2.1597	-1.4625	-0.72671	-0.34367	0.87128	1.7815	1.4452
Постоянн	-37.5174	-21.7414	-11.3914	-2.87019	-2.35408	-7.96810	-11.7000	-20.1789

Рис. 4.2. Фрагмент таблицы с коэффициентами ЛДФ

На рис. 4.2 приведены коэффициентами ЛДФ (расчет F1 – F12 производится по ЛДФ не будет, так как в модельном примере не преследуется задача определения типа породы математическим путем по геохимическим данным).

На рис. 4.3 показано сопоставление результатов классификации по уравнениям ЛКФ (рис. 4.2) с исходной кодировкой типов пород (табл. 4.1).

	Последующие вероятности (Пример) Incorrect classifications are marked with *			
Случай	**Измеренн Classif.**	G_1:1 p=.41975	G_2:2 p=.58025	
28	G_1:1	0.999338	0.000662	
29	G_1:1	0.862378	0.137622	
30	G_1:1	0.969436	0.030564	
* 31	G_1:1	0.400000	0.600000	
32	G_1:1	0.971782	0.028218	
33	G_1:1	0.999971	0.000029	
34	G_1:1	0.992091	0.007909	
* 35	G_2:2	0.826495	0.173505	
36	G_2:2	0.019396	0.980604	
37	G_2:2	0.000861	0.999139	
38	G_2:2	0.000084	0.999916	

Рис. 4.3. Оценка чувствительности решающих правил для данных табл. 4.1

Из таблицы (рис. 4.3) видно, что при проверке линейными классифицирующими функциями предварительно проведенной классификации образцов встречаются группы пород с безошибочной или высокой и средней долей вероятности отнесения образца к группе. К таким группам относятся:

группа 1 – 100 %,
группа 2 – 72,7 %,
группа 3 – 81,2 %,
группа 4 – 88,9 %,
группа 5 – 77,3 %,
группа 6 – 62,5 %,
группа 12 – 100 %.

Есть группы совсем с низкой долей совпадения предварительного определения типа породы у образцов группы с их математической проверкой – совсем малочисленная (три образца) группа 9 (33,3 %).

При анализе данных таблицы (рис. 4.3) хотелось бы обратить внимание на следующие факты.

В группах 2, 3, 4, 5, 6 и малочисленной группе 10 неверно разнесенные образцы распределялись по близким по геохимическому составу группам, так, например, образцы не вошедшие в группу 2 «попали» в группу 1(один образец) и группу 3 (два образца). Исключение составляет один образец из группы 5, который «разнесся» в группу 7 (на рис. 4.3 этот образец отмечен курсивом).

В группах 7, 8, 9, 11 неверно разнесенные образцы «размазаны» по нескольким близлежащим группам в разных направлениях. Здесь наиболее характерна группа 11.

В группе 7 обращает на себя внимание «скачок» трех образцов в группу 11. Это можно объяснить возможной сложностью определения различий нормальных гранитов первой и второй интрузивных фаз. Можно отметить, что при достаточно большом количестве образцов в группе 11, их разброс достиг группы 5. Объяснить это довольно сложно, разве что «разнонаправленностью» с большим диапазоном изменения петрохимических характеристик исследованных образцов (если нет ошибки в исходных данных).

На рис. 4.4 показана таблица квадратов расстояний Махалонобиса.

груп-па	1	2	3	4	5	6	7	8	9	10	11	12
1	0,0	5,2	18,9	51,2	93,8	140,1	172,8	219,8	239,2	259,0	215,0	308,6
2	5,2	0,0	4,4	23,9	55,2	92,3	118,7	158,0	174,0	191,3	154,2	233,8
3	18,9	4,4	0,0	8,2	30,0	57,6	79,2	111,6	124,9	140,4	110,1	176,4
4	51,2	23,9	8,2	0,0	7,0	23,5	36,8	59,4	71,0	81,9	58,7	110,6
5	93,8	55,2	30,0	7,0	0,0	7,4	13,4	27,1	39,4	43,6	25,9	68,2
6	140,1	92,3	57,6	23,5	7,4	0,0	4,1	11,7	24,8	21,7	14,3	45,9
7	172,8	118,7	79,2	36,8	13,4	4,1	0,0	3,4	13,7	11,9	4,2	29,0
8	219,8	158,0	111,6	59,4	27,1	11,7	3,4	0,0	9,3	6,0	2,3	17,4
9	239,2	174,0	124,9	71,0	39,4	24,8	13,7	9,3	0,0	9,2	11,6	4,6
10	259,0	191,3	140,4	81,9	43,6	21,7	11,9	6,0	9,2	0,0	5,8	10,9
11	215,0	154,2	110,1	58,7	25,9	14,3	4,2	2,3	11,6	5,8	0,0	19,7
12	308,6	233,8	176,4	110,6	68,2	45,9	29,0	17,4	4,6	10,9	19,7	0,0

Рис. 4.4. Квадрат расстояний Махалонобиса между группами

При анализе этой таблицы (рис. 4.4) можно отметить следующее:

а) наименьшее расстояние по квадрату расстояний Махалонобиса у групп 8 и 11 (2.3 усл. ед.), 7 и 8 (3.4), 6 и 7 (4.1), 7 и 11 (4,2), 2 и 3 (4.4) и т. д. Эти результаты не противоречат результатам таблицы (рис. 4.3). Так, например, у групп 8 и 11 «пересекаются « пять образцов (см. рис. 4.3), у групп 2 и 3 – четыре образца и т. п. Здесь хотелось бы отметить следующее обстоятельство – наблюдается наименьшее расстояние у близлежащих по коду групп, выделенных согласно группировке химического состава пород, исключение составляет группа 11 [6].

Закономерность «*близких расстояний*» у соседних по номерам групп подтверждает непрерывные переходы и разнонаправленность (пересечение образцов близких по кодам групп) петрохимических разновидностей пород.

На рис. 4.5 показаны коэффициенты канонических ЛДФ, их собственные вклады и кумулятивный вклад в дисперсию. Каноническая переменная 1 (на рис. Корен1) обобщила дисперсию всех признаков на 96,166 % (0,96166). Уравнения КЛДФ будут применены для расчета канонических переменных, по средним

значениям которых будет построен график расположения 12 групп в координатах двух первых канонических переменных (рис. 4.7).

Standardized Coefficients (добр к дискрим) for Canonical Variables

Переменная	Корен1	Корен2	Корен3	Корен4	Корен5
CaO	-0,61938	0,227173	-0,739741	-0,172762	0,327732
K2O	-0,35172	-0,924885	0,025131	0,200056	-0,064227
MGO	-0,41942	0,195079	0,218377	-0,306669	-0,837864
FE2O3	-0,15052	0,151633	0,878372	-0,025098	0,586183
AL2O3	-0,26031	0,333733	0,068767	0,894218	-0,142679
Eigenval	24,39052	0,567680	0,269356	0,073172	0,062304
Cum.Prop	0,96166	0,984039	0,994659	0,997544	1,000000

Рис. 4.5. Коэффициенты канонических ЛДФ

На рис. 4.6 дана таблица факторной структуры КЛДФ, в которой показана информация о корреляционной связи переменных, включенных в модель, с каноническими ЛДФ.

Factor Structure Matrix (добр к дискрим)
Correlations Variables - Canonical Roots
(Pooled-within-groups correlations)

Переменная	Корен1	Корен2	Корен3	Корен4	Корен5
CaO	-0,759969	0,206315	-0,431967	-0,202768	0,390090
K2O	-0,467735	-0,861991	0,080759	0,177655	-0,010534
MGO	-0,549387	0,185759	0,306660	-0,358785	-0,664003
FE2O3	-0,470969	0,128745	0,694317	-0,158041	0,504540
AL2O3	-0,243813	0,300025	0,023132	0,911896	-0,135835

Рис. 4.6 Факторная нагрузка канонических ЛДФ

С первой канонической переменной наиболее «тесно» коррелируют признаки CAO и MgO. Со второй – K2O

1. б) На рис. 4.8 показаны средние значения первых двух канонических переменных, по которым построен пототечный график (рис. 4.7).

Анализ расположения центроидов 12 групп показывает довольно плавный переход от «группы к группе», начиная с группы 1 и до, пожалуй, группы 9. Исключения составляют группа 10 и группа 11, особенно группа 11.Эта группа практически по всем оценкам «выбивается «из общего ряда при проверке гипотезы о

плавных переходах групп пород по петрохимическим характеристикам (см. рис. 4.3, рис. 4.4).

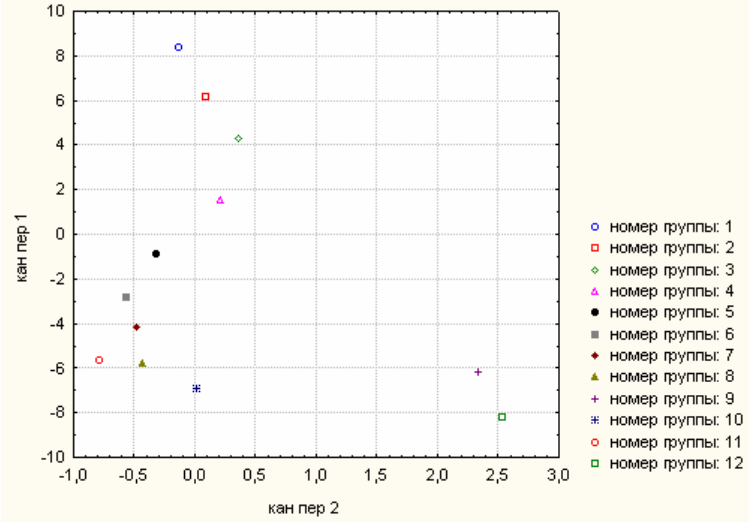


Рис. 4.7. График положения центроидов 12 групп в координатах двух первых канонических переменных

Данные: кан средн сгрупп* (10v by 12c)

	1 кан пер 1	2 кан пер 2	3 номер группы
1	8.3582	-0.133649	1
2	6.19358	0.088403	2
3	4.2813	0.357519	3
4	1.55856	0.215147	4
5	-0.85974	-0.321252	5
6	-2.79097	-0.564641	6
7	-4.15377	-0.48593	7
8	-5.75036	-0.434642	8
9	-6.19355	2.33055	9
10	-6.9342	0.015289	10
11	-5.61605	-0.787459	11
12	-8.19359	2.53536	12

Рис. 4.8. Координаты центроидов первых двух канонических переменных 12 групп

График расположения центроидов 12 групп (рис. 4.7) не противоречит результатам, полученным при определении расстояний между группами (рис. 4.4). Так, например, на графике визуально видно, что самые удаленные группы – это группа 1 и группа 12. Эти же группы имеют самое большое рассчитанное по метрике Махалонобиса расстояние – 308,6 стандартизированных единиц.

Резюме.

Полученные с помощью дискриминатного анализа расчеты, в основном, подтверждают гипотезы о плавных переходах групп пород и разнонаправленности изменения петрохимических характеристик у исследованных образцов, выдвинутые на основании описания изученности петрохимии гранитоидов Ундинского комплекса.

(Напоминаем – это модельный пример для изучения дискриминантного анализа и его реализации на персональном компьютере).

2. Решение должно быть проведено согласно схеме расчета, показанной в главе 3 пособия.

3. На рис. 4.9 показана таблица, из которой видно, что первые пять главных компонент превышают 95 % уровень суммарной дисперсии.

Дискриминантный анализ проводить по пяти первым главным компонентам.

4. На рис. 4.10 показано дерево объединения переменных химического состава гранитоидов Ундинского комплекса по заданному правилу объединения в кластер и заданной мере сходства.

5. На рис. 4.11 показан фрагмент дерева объединения образцов Ундинского комплекса по химическому составу гранитоидов.

Value number	Eigenvalues of covariance matrix, and related statistics (Active variables only			
	Eigenvalue	% Total variance	Совокупный Eigenvalue	Совокупный %
1	2511,995	74,44236	2511,995	74,4424
2	290,633	8,61283	2802,627	83,0552
3	178,968	5,30369	2981,596	88,3589
4	155,205	4,59947	3136,801	92,9584
5	110,864	3,28544	3247,665	96,2438
6	46,436	1,37611	3294,101	97,6199
7	34,510	1,02269	3328,611	98,6426
8	26,058	0,77223	3354,669	99,4148
9	19,746	0,58518	3374,415	100,0000

Рис. 4.9. Результат расчета собственных значений главных компонент, их процентное содержание, накопленные суммы

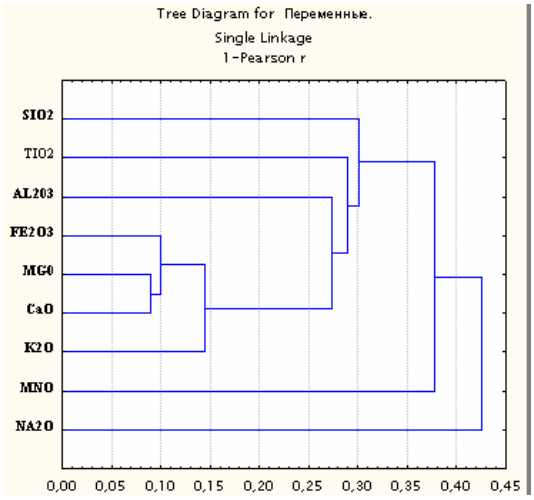


Рис. 4.10. Дерево объединения переменных химического состава гранитоидов Ундинского комплекса

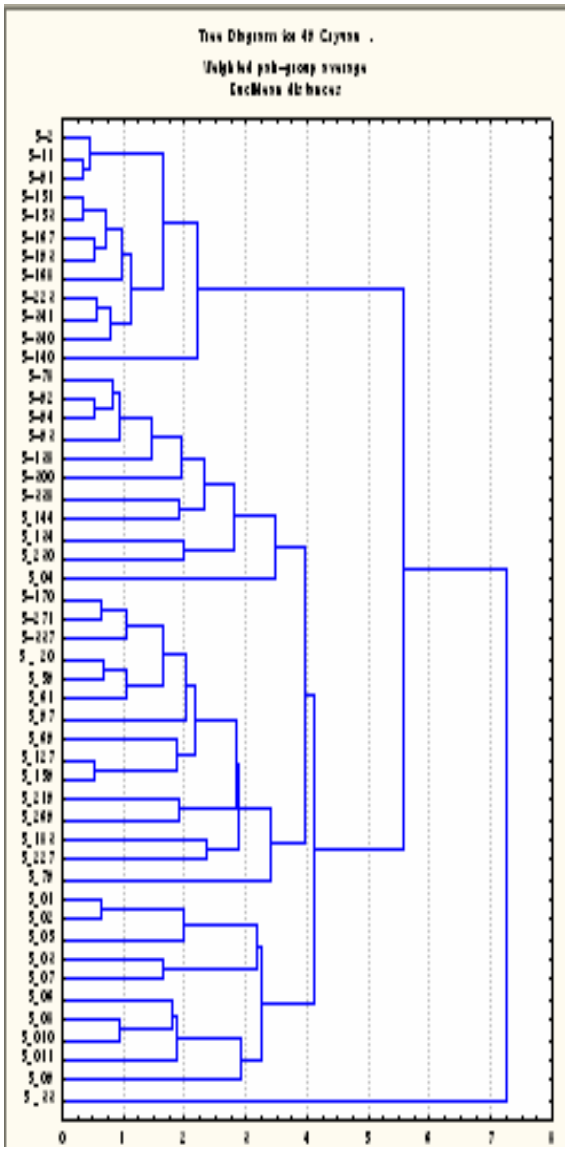


Рис. 4.11. Фрагмент дерева объединения образцов пород по химическому составу по правилу Weighted pair group average иерархического объединения и меры сходства Euclidean distance

ЗАКЛЮЧЕНИЕ

Приведенные в предыдущих двух частях [33, 30] и в этой части учебного пособия математические методы не являются полным спектром всех известных статистических методов и не охватывают всего многообразия форм применения численных оценок в геологических исследованиях, в том числе математико-статистического моделирования в поисково-разведочных работах на нефть и газ.

Почему именно в поисково-разведочных работах на нефть и газ?

Во-первых, разнообразие условий образования ловушек нефти и газа и, обычно, их сложное геологическое строение порождают трудности их прогнозирования. Повышение эффективности прогнозирования на стадиях поиска и разведки возможно при комплексном изучении строения перспективных территорий на основе обобщения имеющихся объемов разнородной геолого-геофизической информации – результатов полевых геофизических съемок, бурения, геофизических исследований скважин (ГИС) и т. п.

На стадии обобщения и анализа информации целесообразно привлечение математических методов. Использование численных методов при комплексном анализе геолого-геофизических данных применительно к геологическим условиям конкретных регионов – одно из направлений повышения достоверности геологического прогноза залежей нефти и газа.

Одними из сложных в методическом отношении задачами комплексного анализа геолого-геофизической информации в нефтегазопоисковой геологии является площадной прогноз структурных условий залегания нефтегазоносных толщ, прогноз пространственного изменения литолого-фациального состава пород и

фильтрационно-емкостных свойств продуктивных толщ на основе совместного анализа скважинных наблюдений, геофизических съемок, ГИС, определений на керне.

Во- вторых, на Вычислительном центре (директор В. Б. Манцивода) Иркутского университета, где работали долгое время составители пособия, с начала семидесятых и до конца девяностых годов прошлого столетия велись научно-исследовательские работа в ***лаборатории моделирования геологических процессов***. Исследования проходили в рамках программ союзных министерств высшего образования и геологии, а также на хоздоговорной основе с организациями нефтяного профиля. Цель работы лаборатории – адаптация известных математико-статистичеких методов применительно к геологии нефти и газа, разработка методик комплексного анализа геоинформации и создание пакетов прикладных программ, позволяющих проводить прогнозные построения. Эти программы дают возможность повышать эффективность работ, направленных на выявление зон, благоприятных для поисков и разведки нефти и газа, а так же ориентированы на сокращение временных и материальных затрат на геологический цикл подготовки месторождений.

Методы опробовались, как правило, с положительными результатами при прогнозных построениях, которые были проведены в разное время для территорий Западной и Восточной Сибири (в т. ч. для Ковыктинского и Верхнечонского месторождений), Туркмении, Республики Саха (Якутии), о. Сахалин и т. д. Результаты научно-исследовательских работ представлены в виде зарегистрированных в соответствующих фондах научно-исследовательских отчетов и в виде публикаций. Некоторые из них приведены в библиографическом списке [6, 11, 27, 38, 37, 39, 36, 41, 21, 40].

В качестве примера площадного прогноза покажем прогноз кровли баженовской свиты в пределах Уренгойского мегавала [21].

При прогнозе был использован регрессионно-интерполяционный метод (РИ – метод) [30, 31].

Кровля J^3b является региональным репером при проведении сейсморазведочных работ и привязывается к отражающему горизонту «Б» (Рудкевич М. Я., 1988). Прогнозные построения проводились в масштабе 1 : 200 000.

На первом этапе построения прогнозной модели кровли J^3 b была получена регрессионная зависимость

$$J^3 b = f(x,y, \Delta g, \Delta T),$$

где J^3 b – значения абсолютных отметок кровли баженовской свиты; $\Delta g, \Delta T$ – значения результатов геофизических исследований грави – и магниторазведки в точках расположения скважин глубокого бурения, участвующих в построении модели (скважины эталонной выборки); x, y – координаты площадного расположения скважин эталонной выборки.

Поля $\Delta g, \Delta T$ – в той или иной степени отражают строение геологических объектов (Элланский М. М. и др., 1972, Арабаджи М. С. и др., 1984). Исходя из этого была предпринята попытка использования полей Δg и ΔT в качестве параметров прогноза при построении модели кровли J^3 b.

Статистически использование потенциальных полей ($\Delta g, \Delta T$) обусловлено высокой корреляционной связью между отметками кровли J^3 b по данным бурения и значениями полей в точках эталонной выборки.

В построении модели участвовало 57 эталонных точек (скважин глубокого бурения). По ним была получена прогнозная модель:

$$J^3 b = 3665{,}32 - 14{,}03 \cdot \Delta g + 0{,}14 \cdot x \cdot y - 0{,}09 \cdot y^2 \cdot \Delta T + 2{,}21 \cdot \Delta T + 2{,}27 \cdot x + 0{,}46 \cdot y \cdot \Delta T.$$

Максимальная разница ε_{max} между прогнозными и исходными значениями кровли J^3 b оказалась равна 92 метрам. Среднеквадратичное отклонение – 40 м.

ε_{max} превысила заданную точность ε. Точность выбрана эквивалентной заданному сечению изолиний при построении прогнозной карты и составляет *50* м. В связи с тем, что $\varepsilon_{max} > \varepsilon$ проводилась корректировка прогнозной модели интерполяцией методом итераций.

В результате корректировки получена модель кровли J^3 b, удовлетворяющая условию $\varepsilon_{max} < \varepsilon$. Среднеквадратичное отклонение стало равно 28 м.

Надежность прогноза была проверена по 15 экзаменационным скважинам (см. табл.). Экзаменационные скважины (скважи-

ны, не участвующие в прогнозных построениях) выбраны случайным образом.

Таблица

Контроль построения структурного плана кровли J^3 b

№	Экзаменационные скважины. № скважины	**Кровля баженовской свиты, абсолютные отметки, м**		Абсолютная ошибка, м
		Исходные значения	Рассчитанные значения	
1	411 – У	- 3650	- 3660	10
2	445 – У	- 3714	- 3675	39
3	400 – У	- 3678	- 3628	50
4	446 – У	- 3679	-3727	48
5	674 – У	- 3616	- 3698	18
6	356 – Е	- 3700	- 3725	25
7	359 – ЮУ	- 3707	-3710	3
8	677 – У	- 3697	- 3620	***77***
9	281 – У	- 3716	- 3669	47
10	671 – У	- 3623	- 3574	49
11	655 – У	- 3647	- 3596	***51***
12	694 – У	- 3593	- 3555	38
13	279 – У	- 3591	- 3545	46
14	504 – С	- 3875	- 3900	15
15	500 – Е	- 3814	- 3778	36

(***Курсивом*** помечены в таблице ошибки, превысившие заданную точность ε .)

Вывод. Применение регрессионно-интерполяционного метода, как показывает контроль по экзаменационным скважинам, для геологических условий Западной Сибири довольно эффективен. РИ – метод при имеющихся объемах геофизических данных дает возможность оперативно и при небольших затратах проводить построения, используемые в последующем при проведении детальных площадных геофизических работ и постановке глубокого бурения.

И последнее. Составители пособия с удовольствием перечислят сотрудников, которые принимали в разное время участие в разработке математических методов геологии и апробации их на

геологическом материале в лаборатории моделирования геологических процессов (МГП) Вычислительного центра ИГУ:

Акимова Алина Андреевна,
Анисенко Владимир Павлович,
Анисенко Светлана,
Балюра Алексей Борисович,
Балюра Марина,
Белых Ольга Николаевна (в июле 2006 г. у О.Н. был восьмидесятилетний юбилей, мы все желаем ей, ветерану труда, проработавшей почти всю сознательную жизнь в нашем университете, отличного здоровья и всего самого хорошего),
Берковец Игорь Александрович,
Вассерман Татьяна Григорьевна,
Велижанина Лидия,
Грач Михаил,
Гусев Виктор Алексеевич,
Данильченко Леонид Григорьевич,
Ельконина Нина Семеновна,
Зимбалевский Николай Николаевич,
Иванов Александр,
Иванова Наталья Сергеевна,
Исаева Алла Николаевна,
Карпов Сергей Николаевич,
Кобелев Владимир Павлович,
Кобелева Галина Ивановна,
Королев Владимир Иванович,
Корсуков Виктор Михайлович,
Кузьмичева Татьяна Евгеньевна,
Кучерова Елена,
Лабутин Дмитрий Владимирович,
Лобанов Анатолий Дымбрендоржиевич,
Лохматов Геннадий Иванович,
Лузин Валентин Федорович,
Лысов Борис Антонович,
Матусевич Магарита Гавриловна,
Миткевич Наталья Ивановна,
Михалевич Исай Моисеевич,
Михеев Виталий Сергеевич,
Наумова Татьяна Трифоновна,

Плахова Галина Сергеевна (в апреле 2006 г. у Г. С. был юбилей, который она торжественно отметила со многими из этого списка),

Примина Светлана Павловна,
Розум Валентина,
Романенко Тамара Викторовна,
Рыбинская Татьяна Николаевна,
Сорокина Галина Семеновна,
Тененбаум Марина Павловна,
Третьтяков Игорь Викторович,
Филина Лидия Викторовна,
Ханыгина Маргарита Николаевна,
Шахов Николай Александрович,
Шипунова Ирина Борисовна,
Ширяева Наталья,
Юдина Светлана.

Основателем и, практически все время деятельности, научным руководителем лаборатории МГП был кандидат геолого-минералогических наук, доц. Г. И. Лохматов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

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

2. *Айвазян С. А.* Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков [и др.]. – М. : Финансы и статистика, 1989. – 607 с.

3. *Александров В. В.* Анализ данных на ЭВМ (на примере системы СИТО) / В. В. Александров, А. И. Алексеев, Н. Д. Горский. – М. : Финансы и статистика, 1990.

4. *Алферова М. А.* Примеры практической работы с Excel : учеб.-метод. пособие. Вып. 2 / М. А. Алферова, И. М. Михалевич, Н. Ю. Рожкова [и др.]. – Иркутск : ИГИУВ, 2006. – 41с. (изд. 6, стереотипное).

5. *Арабаджи М. С.* Математические методы и ЭВМ в поисково-разведочных работах / М. С. Арабаджи, Э. А. Бакиров, В. С. Мильничук [и др.]. – М. : Недра, 1984. – 264 с.

6. *Беус А. А.* Распределение элементов в гранитоидах / А. А. Беус, А. А. Ситнин // Проблемы геохимии. К 70-летию акад. А. П. Виноградова. – М. : Наука, 1965. – С. 429–435.

7. *Боровиков В.* STATISTIKA: искусство анализа данных на компьютере. Для профессионалов / В. Боровиков. – СПб. : Питер, 2001. – 656 с.

8. *Боровиков В. П.* Программа STATISTICA для студентов и инженеров. 2-е изд. / В. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

9. *Боровиков В. П.* STATISTICA – Статистический анализ в среде WINDOWS. Изд. 2-е, стереотип. / В. Боровиков, И. П. Боровиков. – М. : Инф.-изд. дом «Филинъ», 1998 . – 608 с.

10. *Власов В. В.* Введение в доказательную медицину / В. В. Власов. – М. : МедиаСфера, 2001 . – 392 с.

11. *Гусев В. А*. Численное моделирование геологических полей / В. А. Гусев, Г. И. Лохматов, В. П. Кобелев. – Иркутск : Изд-во Иркут. ун-та, 1984. – 144 с.

12. Девис Дж. С. Статистический анализ данных в геологии / Дж. Девис. – М. : Недра, 1990. – Т. 1. – 319 с., Т. 2. – 427с.

13. *Добрецов Н. Л.* Статистические методы в геологии / Н. Л. Добрецов, В. В. Зуенко, М. Л. Шемякин. – М. : Наука. Сиб. отд. 1974 . – 142 с.

14. *Драйпер Н.* Прикладной регрессионный анализ / Н. Драйпер, Г. Смит. – М. : Статистика, 1973. – 392 с.

15. *Дубров А. М.* Обработка статистических данных методом главных компонент / А. М. Дубров. – М. : Статистика, 1978 . – 135 с.

16. *Дюк В.* Обработка данных на ПК в примерах / В. Дюк. – М. : Изд-во «Питер Паблишинг», 1997. – 231с.

17. *Дюк В.* DataMining : учеб. курс / В. Дюк, А. Самойленко. – СПб. : Питер, 2001. – 368 с.

18. *Енюков И. С.* Методы, алгоритмы, программы многомерного статистического анализа / И. С. Енюков. – М. : Финансы и статистика, 1986. – 232 с.

19. *Закс Л.* Статистическое оценивание / Л. Закс. – М. : Статистика, 1976 . – 98.

20. *Иберла К.* Факторный анализ / К. Иберла. – М. : Статистика. 1980. – 398 с.

21. *Карпов С. Н.* Применение РИ-метода для прогноза кровли баженовской свиты в пределах Уренгойского вала. Информационный листок № 344-91 / С. Н. Карпов, О. А. Козлов, И. М. Михалевич.– Иркутск : ЦНТИ, 1991. – 4 с.

22. Классификация и кластер / ред. Дж. Вэн Райзин. – М. : Мир, 1980 . – 392 с.

23. Колкот Э. Проверка значимости / Э. Колкот. – М. : Статистика, 1978. – 128 с.

24. *Крамбейн У.* Статистические модели в геологии / У. Крамбейн, Ф. Грейбилл. – М. : Мир, 1969. – 397 с.

25. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows и STADIA. Т. 1. Изд. 3-е, перераб. и доп. / А. П. Кулаичев. – М. : «Информатика и компьютеры», 1999 . – 341 с.

26. *Лавач С. М.* Статистические методы в медико-биологических исследованиях с использованием Excel / С. М. Лавач, А. В. Чубенко, П. М. Бабич. – Киев : «Морион», 2000. – 320 с.

27. *Лохматов Г. И.* Особенности структурного плана подсолевых венд – нижнекембрийских отложений во внутренних районах Иркутского амфитеатра / Г. И. Лохматов, В. Ф. Лузин, И. М. Михалевич // Геология и полезные ископаемые Сибирской платформы и ее складчатого обрамления. – Иркутск : ИГУ ВИНИТИ № 7515-84, 1984. – С. 17–27.

28. *Лохматов Г. И.* Сочетание метода гиперсфер и потенциальных функций в задачах классификации геологических объектов / Г. И. Лохматов, И. М. Михалевич // Применение методов математического моделирования для прогноза рудных месторождений (на примере Восточной Сибири). – Иркутск, 1981. – С. 64–79.

29. *Миллер Р.* Статистический анализ в геологических науках / Р. Миллер, Дж. Кан. – М. : Мир, 1965. – 482 с.

30. *Михалевич И. М.* Методика построения прогнозных карт с применением интерполяционных процедур / И. М. Михалевич // Природные ресурсы, экология и социальная среда Прибайкалья. Т. 3. – Иркутск , 1995. – 5 с.

31. Михалевич И. М., Михалевич М. И. Интерполяция – прогнозные построения. Свидетельство об официальной регистрации программы для ЭВМ № 2004610613, РОСПАТЕНТ. 2004 г.

32. *Михалевич И. М.* Применение математических методов при анализе геологической информации (с использованием компьютерных технологий) : учеб. пособие. Ч. II / И. М. Михалевич, С. П. Примина. – Иркутск : Иркут. ун -т, 2004. – 120 с.

33. *Михалевич И. М.* Применение математических методов при анализе геологической информации (с использованием Excel) : учеб. пособие. Ч. I / И. М. Михалевич, С. П. Примина. – Иркутск : Изд. Иркут. ун-та, 2001. – 60 с.

34. *Налимов В. В.* Теория эксперимента / В. В. Налимов. – М. : Наука, 1971.

35. *Окунь Я.* Факторный анализ / Я. Окунь. – М. : Статистика, 1974. – 200 с.

36. Отчет «Разработка и внедрение нетрадиционных методов комплексного анализа геоинформации в целях изучения глубинного строения недр и ориентированных на повышение эффектив-

ности работ на нефть и газ», ВИНИТИ № гос. регистрации 01940000436, инф. номер 02.09.80 002452, 1998 г. – 49 с.

37. Отчет «Математическое моделирование и прогноз месторождений нефти и газа» ВИНИТИ № ГР 77016957. – Иркутск, 1985. – 343 с.

38. Отчет «Составление дежурных геофизических карт по основным сейсмическим горизонтам нефтегазоносных комплексов Юго-Восточной Туркмении». Геолфоды. № ГР 36-81-6/8. – Иркутск ; Ашхабад, 1983. – 129 с.

39. *Примина С. П.* Фациально-минералогический анализ венд – кембрийских отложений в связи с прогнозом зон нефтегазонакопления на юге Сибирской платформы : автореф. дис. … канд. геол.-минерал. наук. – М., 1991 . – 15 с.

40. Примина С. П. Детальные исследования при выявлении сложнопостроенных объектов поиска углеводородного сырья в Прибайкалье / С. П. Примина // Природные ресурсы, экология и социальная среда Прибайкалья. Т. III. – Иркутск, 1995. – С. 59–62.

41. *Примина С. П.* Прогноз размещения залежей нефти и газа на основе фациальной неоднородности и детальной корреляции базальной терригенной толщи на территории Непско-Ботуобинской антеклизы / С. П. Примина, С. Л. Арутюнов, И. М. Михалевич // Осадочные формации докембрия и их рудоносность : материалы Всероссийского совещания. – СПб., 1998. – 5 с.

42. *Реброва О. Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. – М. : МедиаСфера, 2003. – 312 с.

43. Рожкова Н. Ю., Алферова М. А., Михалевич И. М. Группирование объектов в признаковом пространстве методом совместного использования гиперсфер и потенциальных функций (PNCL). Свидетельство об официальной регистрации программ на ЭВМ № 2004610614, РОСПАТЕНТ. 2004 .

44. *Скрипченко Н. А.* Анализ данных в MICROSOFT EXCEL / Н. А. Скрипченко. – Иркутск : Изд-во ИГТУ, 1998. – 60 с.

45. *Тихомиров И. Н.* Интрузии каменноугольного возраста в бассейне среднего течения р. Газимур (Восточное Забайкалье) / И. Н. Тихомиров // Материалы по петрологии гранитоидов Забайкалья. – М., Госгеолтеиздат, 1962. С. 16–22.

46. *Тьюки Дж. У.* Анализ результатов наблюдений ; пер. с англ. / У. Дж. Тьюки– М. : Наука, 1971.

47. *Чини Р. Ф.* Статистические методы в геологии ; пер. с англ. / Р. Ф. Чини– М. : Мир, 1986 – 189 с.

48. Юнкеров В. И. Математико-статистические методы обработки данных медицинских исследований / В. И. Юнкеров, С. Г. Григорьев. – СПб. : ВМедА, 2002. – 266 с.

49. StatSoft, Ins (2001). Электронный учебник по статистике. – М., StatSoft. WEB: http://www.statsoft.ru/home/textbook/default.htm.

Учебное издание

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКИХ МЕТОДОВ ПРИ АНАЛИЗЕ ГЕОЛОГИЧЕСКОЙ ИНФОРМАЦИИ (с использованием компьютерных технологий) Часть III

Учебное пособие

Составители:
Михалевич Исай Моисеевич,
Примина Светлана Павловна

Редактор *Э. А. Невзорова*
Макет: *И. В. Карташова-Никитина*
Дизайн обложки: *М. Г. Яскин*

Темплан 2006 г. Поз. 116.

Подписано в печать 25.10.06. Формат 60x84 1/16. Печать трафаретная.
Усл. печ. л. 6,7. Уч.-изд. л. 3,5. Тираж 150 экз. Заказ 214.

РЕДАКЦИОННО-ИЗДАТЕЛЬСКИЙ ОТДЕЛ
Иркутского государственного университета
664003, Иркутск, бульвар Гагарина, 36