STATISTICS AND DATA ANALYSIS IN GEOLOGY

Second edition

John C. Davis

Kansas Geological Survey

John Wiley and Sons
New York – Chichester – Brisbane – Toronto – Singapore

Дж. С. Дэвис

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ГЕОЛОГИИ

В ДВУХ КНИГАХ

КНИГА 2

МОСКВА "НЕДРА" 1990

Перевод с английского доктора физико-математических наук В. А. Голубевой
Под редакцией доктора геолого-минералогических наук Д. А. Родионова

Содержание:

ПРЕДИСЛОВИЕ

В 1973 г., когда вышло первое издание этой книги, использование геологами вычислительных средств находилось на качественно ином уровне, чем сейчас. Это было время массивных ЭВМ, сосредоточенных в вычислительных центрах, доступ к которым осуществлялся через окошко в закрытых дверях. При этом исследователь получал результат в лучшем случае через несколько дней.

Теперь большинство геологов имеют доступ к ЭВМ через терминал или к мини-компьютеру, или имеют персональный компьютер. Компьютер стал обыденной вещью в жизни геолога. К нему обращаются как новички, так и профессионалы в надежде повысить эффективность своей работы.

К сожалению, легкий доступ к компьютерам не обеспечивает легкого получения знаний о том, что с ним делать. Для многих геологов анализ поверхностей тренда так же мало понятен, как и 10 лет назад. То же можно сказать и о факторном анализе. Более того, появились еще более экзотические и труднодоступные методы. Необходимость обучения геологов количественному анализу была очевидна уже в 1973 г., то же верно и сейчас. Вот почему написана эта книга.

В ответ на многие замечания, которые я получил после издания книги в 1973 г., а также, учитывая собственный опыт преподавания, я внимательно пересмотрел книгу для нового издания. Расположение материала сохранено, оно начинается с основных понятий и заканчивается анализом последовательностей, карт, многомерных наблюдений. Так как большинство студентов слушает один или более курсов по ФОРТРАНу, то глава о ФОРТРАНе в этом издании отсутствует.

Изложение начинается с основ теории вероятностей, очень важных в анализе данных. Добавлен раздел о непараметрических методах, представляющиеся более пригодными для геологических данных. Тема «Собственные значения и собственные векторы» остается трудной для геологов, и поэтому она затрагивается дважды: в разделе «Матричная алгебра» и в разделах, посвященных факторному анализу. Рассмотрена связь процедур нахождения собственных значений и собственных векторов с методом главных компонент, *R*- и *Q*-факторным анализами, анализом соответствия.

Некоторые темы анализа данных в последние несколько лет приобретали все большее значение в науках о Земле. Теория регионализованных переменных привлекается сейчас для объяснения пространственных свойств геологических переменных многими исследователями. Центральную роль в этой теории играют полувариограммы и крайгинг. Эти методы представлены в настоящем издании. Геофизики поняли важную роль спектрального анализа. Полезность этих методов очевидна при решении многих других задач, начиная от предсказания землетрясений и кончая описанием формы ископаемых остатков. Раздел о рядах Фурье излагается с учетом этих изменений.

Ряд таблиц и рисунков в этой книге воспроизведены с разрешения авторов – владельцев авторских прав. Источник для каждой таблицы и рисунка указан в квадратных скобках, а полная ссылка приводится в списках литературы к каждой главе. Таблицы 2.10, 2.22, 2.25 и 2.26 являются собственностью издательства Джон Уайли и Сыновья Inc., таблицы 2.11, 2.14 и 2.18 – собственностью Пингвин-Бук Ltd, а таблицы 4.30 и 4.31 – Американского химического общества. Часть таблиц 5.6 – собственность Американской статистической ассоциации и другая часть – Американского института биологических наук; комбинирование таблиц сделано с их разрешения. Таблицы 5.7 и 5.9 представляют собственность Академик Пресс Inc (Лондон) Ltd и. Рис. 5.24 – собственность Американской статистической ассоциации, а рис. 5.25 – собственность Харкурт Брейс Йованович, Inc.

В дополнение к тем, кто был назван в первом издании, приношу мою благодарность доктору Паулю Брокинстону, доктору Джиму Кемпбеллу и доктору Кейту Терперу за их помощь. Мои рецензенты, доктор Дейв Бест, профессор Франк Этридж и профессор Джиэн Фэнг сделали много ценных исправлений в окончательном тексте.

Многие добавления были сделаны коллегами из Канзаской геологической службы, включая доктора Дэйвида Коллинза, доктора Калина Фергюсона и моего помощника по первому изданию мистера Роберта Сэмпсона. Трое моих коллег приняли активное участие в написании книги: доктор Рикардо Олеа – раздел по регионализованным переменным, доктор Жу Ди – раздел о собственных значениях и доктор Джон Доветон, который предложил многие из упражнений и примеров, приведенных в книге, и который помогал мне на всех стадиях проверки. Я особенно признателен моему ассистенту, исследователю и компаньону миссис Джо Энн Де-Греффенрайд, которая издавала, корректировала и проверяла, и без поддержки которой выход этой книги оказался бы невозможным.

Джон С. Дэвис

Глава 5. АНАЛИЗ КАРТ

УСЛОВНЫЕ И ДРУГИЕ ГЕОЛОГИЧЕСКИЕ КАРТЫ

Геологи проводят свои исследования в реальном трехмерном мире, однако их представления о нем в значительной степени остаются двумерными. Это является следствием того, что третье измерение, роль которого в геологии обычно играет глубина, нередко бывает лишь частично доступно для изучения по сравнению с двумя другими измерениями. Кроме того, наши представления во многом обусловлены средствами их выражения, которыми могут быть карты, фотографии, разрезы, напечатанные на плоских листах бумаги. Предметом исследования могут быть геологические характеристики пород, доступных для изучения на глубине в горных выработках на различных уровнях, в штольнях и восстающих. Эти характеристики образуют сложную трехмерную сеть, которую для наглядного представления нужных зависимостей требуется изобразить в виде плоской проекции. Геологи уделяют большое внимание чтению, использованию и построению карт и, возможно, являются самыми сведущими в наглядном выражении и изучении пространственных зависимостей. В науках о Земле карты играют ту же роль, что и ноты в музыке, будучи компактным и эффективным средством выражения зависимостей и различных деталей.

Карты – это средство обучения и работы геологов. Однако теоретические аспекты автоматизации их построения развиты слабо. Большинство способов сравнения карт между собой введено географами, хотя геологам постоянно приходится сравнивать карты, отыскивая в них элементы сходства. С внедрением электронной вычислительной техники роль автоматического построения карт в геологической практике стала возрастать. В настоящее время самые активные потребители графопостроителей – нефтяные компании. До сих пор почти ничего не опубликовано об алгоритмах, используемых в работе графопостроителей, о сравнительных преимуществах возможных различных подходов к их разработке. Недостаточно изучен также вопрос надежности или эффективности карты. Анализ поверхностей тренда, пожалуй, единственный широко применяемый способ анализа карт, но те исследователи, которые используют существенные программы этого метода для ЭВМ, обычно не осведомлены о присущих ему ограничениях. Следствие этого – неожиданное появление как осмысленных, так и с трудом поддающихся интерпретации результатов применения данного метода, что в свою очередь порождает как сторонников метода, так и скептиков.

Карта – это двумерное представление некоторой области. Обычно это четырехугольник, ограничивающий горный район или страну, построенный в результате масштабного сокращения реальных пространственных соотношений с целью их более легкого восприятия. Однако такое представление пространственных соотношений может в равной степени относиться и к тому, что мы привыкли называть словом «карта», и к петрографическому шлифу, и к снимку, сделанному на электронном микроскопе, когда взаимоотношения между характеристиками изучаемого объекта представляются в увеличенном виде (масштабе). В самом общем определении картами считаются обычные геологические и географические карты, аэрофотоснимки, маркшейдерские планы, карты поверхностей отдельных стратиграфических подразделений, микрофотографии и снимки электронного микроскопа. Таким образом, любой из видов двумерного пространственного построения можно рассматривать как карту.

Зависимости, изучаемые на карте, почти всегда изображаются с помощью точек. При этом обычно рассматриваются расстояния между точками, их плотность и значения, приписанные каждой точке. Подавляющее большинство карт представляет собой оценки некоторых непрерывных функций по результатам наблюдений в дискретных контрольных точках. Типичным примером может служить обычная топографическая карта, которая, несмотря на непрерывность изображенных на ней контуров, построена на основе дискретной триангуляционной сети с узлами в контрольных точках. Еще более наглядный пример – это структурная карта в изолиниях. В связи с тем, что мы наблюдаем структурную поверхность только в точках пересечения ее скважинами, неясно, является ли она непрерывной в промежутках между этими точками, недоступными для наблюдения. Тем не менее, в подобной ситуации предполагается, что поверхность непрерывна, и ее форма устанавливается по результатам наблюдений в контрольных точках, причем учитывается, что ее реконструкция неточна и не отражает ряда деталей ввиду отсутствия данных между скважинами.

При аэрологическом картировании пустынных районов можно установить простирание, измерить углы падения пород и довольно точно нанести на карту границы между формациями, так как их удается проследить во всем регионе. В районах же с повышенным развитием растительности и мощной корой выветривания мы вынуждены довольствоваться разобщенными выходами коренных пород и весьма плохими обнажениями, что сказывается на качестве карты, которое зависит от плотности точек наблюдения. Вопрос о влиянии распределения точек наблюдения на качество карты представляет существенный интерес для геологов, но по данной теме существует очень немного опубликованных работ. Почти все исследования, связанные с этим вопросом, были выполнены географами. В данной главе мы рассмотрим некоторые результаты этих исследований и возможности их применения при изучении геологических карт, а также для решения таких задач, как распределение зерен минералов в петрографических шлифах.

Построение геологической карты – своего рода искусство, в котором проявляется талант исследователя. В ряде случаев дополнительная геологическая интерпретация первичных результатов наблюдения в значительной степени повышает качество карты. Однако иногда на геологических выводах сказывается влияние персонального фактора, отражающего индивидуальные взгляды исследователя, что значительно снижает качество карты. Методы построения карт в изолиниях на ЭВМ препятствуют действию персонального фактора при интерпретации. Конечно, субъективные суждения необходимы при выборе алгоритма построения карты, но сам выбор между конкурирующими процедурами осуществляется с помощью соответствующих критериев. Главная причина развития машинных методов построения карт – экономическая, вызванная попыткой использования огромной стратиграфической информации, накопленной в нефтяной промышленности. Однако одно из главных достижений этих методов заключается в том, что они заставляют сосредоточить внимание на проблеме надежности карты. Методы построения карт в изолиниях изложены в специальном разделе этой главы, где приведены также примеры построения и применения некоторых простых программ для ЭВМ.

Анализ поверхностей тренда – один из наиболее широко применяемых в геологии математических методов, и поэтому мы рассмотрим его детально, останавливаясь на различных вариантах, в частности, на четырехмерных поверхностях тренда и на поверхностях, описываемых гармониками Фурье. Несмотря на то что этот метод широко применяется в геологии, им нередко и злоупотребляют. В связи с этим нам придется рассмотреть вопросы распределения точек наблюдения, недостаточного приближения, вычислительных «срывов» и неправильных применений.

Если построение поверхности тренда рассматривать как задачу множественной регрессии, то при этом необходимо использовать соответствующие статистические критерии, причем эта необходимость заранее постулируется. Существуют различные точки зрения на модели поверхности тренда (а следовательно, и на их применение). Разногласия в данном случае вызваны тем, что ряд авторов отстаивают преимущества построения поверхностей тренда по сравнению с методами сглаживания с помощью скользящего среднего. Обе эти точки зрения будут проанализированы, и будет сделана попытка дать соответствующее заключение.

Двумерные методы – это, как правило, обобщение методов, изложенных в гл. 4. Тренд-анализ – раздел регрессионного анализа. В свою очередь, крайгинг связан с анализом временных рядов, а построение изолиний – обобщение теории интерполяции. Мы просто расширяем общность наших задач путем введения новых пространственных переменных. Конечно, некоторые задачи, возникающие при анализе карт, уникальны, но в целом эта глава посвящена многомерным методам математической геологии. Важность одно- и двумерных задач в науках о Земле заставила выделить эти разделы в самостоятельные главы.

СИСТЕМАТИЧЕСКИЕ МЕТОДЫ ПОИСКА

Большинство геологов посвятили свою профессиональную карьеру исследованию чего-то скрытого от взора человека. Обычно объект геологического исследования – это неоткрытое нефтяное поле или рудное тело, но иногда это может быть трещина на некоторой глубине, ископаемые остатки приматов при раскопках или температурный скачок на океанском дне. Очень часто такие находки делаются случайно во время осмотров исследуемой площади, подобно тому, как это делали старые разведчики, следующие к месту своей службы. Все чаще, однако, геологи и другие ученые, исследующие Землю, применяют более систематические процедуры поисков, в частности, пользуются различными инструментами для обнаружения искомых объектов.

Наиболее систематические исследования проводятся вдоль одного или более множеств параллельных линий. Рудные тела, имеющие особые радиоактивные или магнитные свойства, находятся при помощи воздушной разведки вдоль равномерно расположенных линий – трасс полета. Сейсмические данные основываются на регулярном множестве траверсов. Разведка со спутников в силу ее специфической природы производится по параллельным орбитам.

Вероятность обнаружения цели при исследовании вдоль множества линий может быть определена из геометрических соображений. Обычно вероятность открытия связана с относительным размером цели сравнительно с пространственной схемой опробования. На вероятность также оказывают влияние вид цели и расположение линий поисков. Если предположить, что цель имеет эллиптическую форму, а поиск ведется по параллельным линиям, то можно вычислить вероятность того, что линия пересечет скрытую цель заданного размера, невзирая на то, где она расположена внутри исследуемой площади. Такие предположения не кажутся необоснованными для большей части разведочных служб. Заметим, что вероятности связаны только с пересечением цели линией и не касаются проблемы поиска скрытой цели.

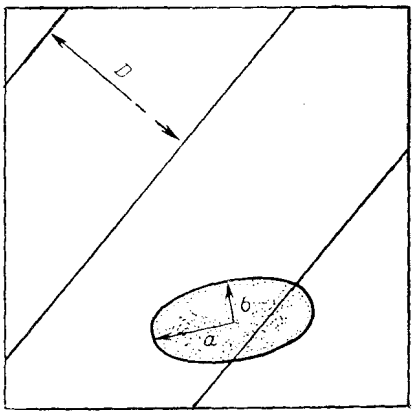


Рис. 5.1. Поиск эллиптического объекта с помощью параллельных линий:
a – большая полуось; *b* – малая полуось; *D* – расстояние между линиями

Мак-Кеммон [53] приводит метод вычисления геометрических вероятностей для круглых и линейных целей при параллельной схеме поиска. Его работа основана главным образом на математических методах Кендалла и Морана [43]. Ранее в книге Успенского [76] была получена более общая формула для эллиптического случая, используемая нами.

Предположим, что цель, которую требуется найти, является эллипсом, полуоси которого равны a и b. (Если цель круговая, то $a=b=r$, где r – радиус круга). Схема поиска состоит в проведении серии параллельных прямых, расположенных друг от друга на расстоянии D (см. рис. 5.1). Вероятность того, что цель (меньшая, чем расстояние между линиями) будет пересечена некоторой линией, равна

$$P = p / \pi D \qquad (5.1)$$

$$p = 2\pi\sqrt{\frac{a^2 + b^2}{2}}$$

где p – периметр эллиптической цели. Периметр эллипса вычисляется по формуле
где a и b – его полуоси. Подставляя значение p в формулу (5.1), получим

$$P = \frac{2\pi}{\pi D}\sqrt{\frac{a^2 + b^2}{2}} = \frac{2}{D}\sqrt{\frac{a^2 + b^2}{2}} \qquad (5.2)$$

Определим Q по формуле

$$Q = 2\sqrt{\frac{a^2 + b^2}{2}} \qquad (5.3)$$

Тогда вероятность пересечения эллиптической цели одной из параллельных прямых равна

$$P = Q / D \qquad (5.4)$$

В частном случае круговой цели, когда a и b равны радиусу, получаем

$$P = 2r / D \qquad (5.5)$$

В другом крайнем случае одна из осей эллипса может быть настолько короткой, что цель становится случайно ориентированной линией. Это геометрическое соотношение известно как задача Бюффона. Она состоит в вычислении вероятности того, что игла длиной l, брошенная на множество параллельных прямых, расположенных на расстоянии D одна от другой, пересечет хотя бы одну из этих прямых. Эта вероятность равна

$$P = 2l /(\pi D) \quad (5.6)$$

где l – длина цели.

Аналогичное соотношение, известное как проблема Лапласа, встречается в теории систематического поиска. Задача состоит в нахождении вероятности того, что игла длиной l, брошенная на поле, покрытое сетью прямоугольников, попадет целиком в один из них. Один из вариантов этой задачи – вычисление вероятности того, что монета, брошенная на шахматную доску, падает целиком в один из квадратов. В разведке представляют интерес также дополнительные вероятности, т.е. вероятности того, что случайно расположенная цель будет пересечена один или большее число раз множеством линий, как, например, сейсмические разрезы, вложенные в прямоугольную сеть (рис. 5.2).

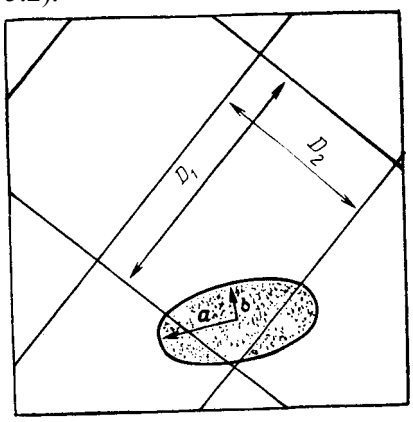


Рис. 5.2. Поиск эллиптического объекта с помощью сети, характеризуемой интервалом D_1 в одном направлении и интервалом D_2 в другом: a – большая полуось; b – малая полуось

Общее уравнение имеет вид

$$P = \frac{Q(D_1 + D_2 - Q)}{D_1 D_2} \quad (5.7)$$

Здесь D_1 – расстояние между параллельными линиями первого сейсмического разреза и D_2 – расстояние между перпендикулярным множеством разрезов. В частном случае поисков при квадратной сети равенство упрощается

$$P = \frac{Q}{D}\left(2 - \frac{Q}{D}\right) \qquad (5.8)$$

Ламби в неопубликованном докладе 1981 г. [48] указывает, что эти уравнения для геометрической вероятности являются аппроксимациями интегральных уравнений. Сравнивая точные вероятности, найденные численным интегрированием, с вероятностями, полученными приближенными уравнениями, он нашел, что значимые различия случаются только для очень удлиненных целей, которые велики по сравнению с расстоянием между линиями поисков. Кроме того, уравнения (5.4) и (5.7) сильно завышают вероятности обнаружения.

Вероятности пересечения цели, вычисленные по приближенным уравнениям, удобно представить на графике. Мак-Кеммон [53] представил такие графики для различных наборов вида и размеров по отношению к расстоянию между линиями поиска в очень удобной безразмерной форме. На рис. 5.3 представлена вероятность обнаружения эллиптической цели, вид котором изменяется от круга до линии, причем метод поиска – схемы параллельных линий. Размер цели выражается как отношение (максимальное направление цели)/(расстояние между линиям поиска). На рис. 5.4 представлен аналогичный график-схема поиска с помощью квадратной сетки линий.

Если вид цели задан, то вероятности пересечения можно нанести на график для различных поисковых схем. На рис. 5.5 например, представлены вероятности пересечения некоторой круговой цели поисковыми схемами, изменяющимися от квадратной сети через прямоугольную сеть до параллельных линий. На рис. 5.6 представлен эквивалентный график для линейной цели. Можно рассмотреть всевозможные промежуточные случаи всех возможных эллиптических целей и всех возможных поисковых схем вдоль двух перпендикулярных семейств линий.

Из приближенных уравнений мы можем вычислить графики частного вида, пригодные в конкретных приложениях. На рис. 5.7, например, приведена вероятность обнаружения некоторой

1000-акровой эллиптической цели сетью параллельных линий с расстоянием между ними в одну милю. Этот график был использован сейсмической службой для оценки возможности обнаружения сейсмической аномалии минимального размера.

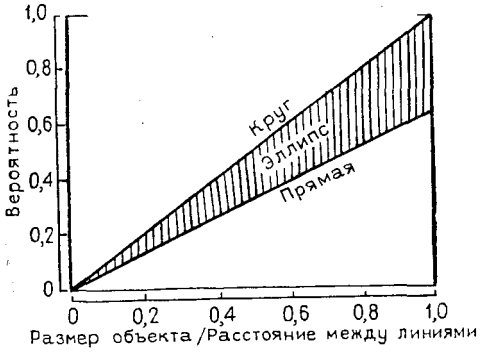


Рис. 5.3. Вероятность обнаружения объекта с помощью схемы поиска параллельными линиями. Форма объекта может изменяться от круга до линии; эллиптические объекты с различными осевыми отношениями попадают в заштрихованную область. Горизонтальная ось – это отношение наибольшего размера объекта к расстоянию между линиями поиска [53]

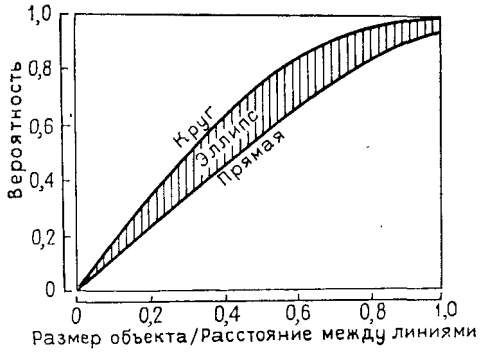


Рис. 5.4. Вероятность обнаружения объекта квадратной сетью поиска. Форма объекта может изменяться от круга до линии; эллиптические объекты с различными отношениями осей попадают в заштрихованную область. Горизонтальная ось – это отношение наибольшего размера объекта к расстоянию между линиями поиска [53]

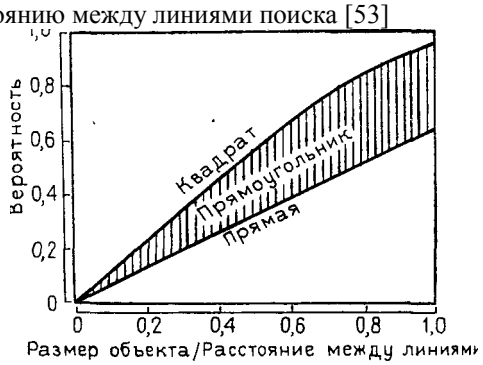


Рис. 5.5. Вероятность обнаружения кругового объекта с прямоугольной схемой поиска, изменяющейся от квадратной схемы до множества параллельных линий. Прямоугольные схемы с различными отношениями D_1/D_2, попадающими в заштрихованную область. Горизонтальная ось – это отношение наибольшего размера объекта к минимальному расстоянию между линиями поиска [53]

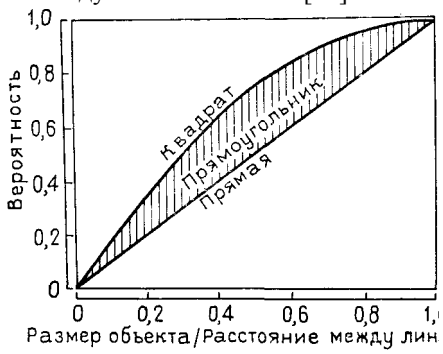


Рис. 5.6. Вероятность обнаружения линейного объекта равномерной схемой, изменяющейся от поиска множеством квадратов до поиска множеством параллельных линий. Схемы прямоугольного поиска с различными отношениями D_1/D_2 попадают в заштрихованную область. Горизонтальная ось – это отношение длины объекта к минимальному расстоянию между линиями поиска [53]

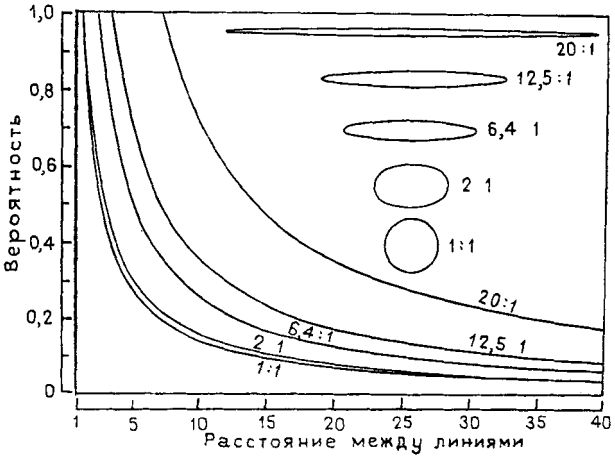


Рис. 5.7. Вероятность пересечения 1000-акрового объекта параллельными линиями поиска с расстоянием между линиями одна миля

РАСПОЛОЖЕНИЕ ТОЧЕК

Одна из распространенных геологических задач заключается в изучении способа распределения точек на двумерной поверхности или карте. Эти точки могут соответствовать местам взятия проб, получения наблюдений или быть точками проекции. Задача может состоять в изучении однородности распределения точек наблюдения, плотности распределения или в изучении связи точек друг с другом. Все эти вопросы возникают как у географов, так и у геологов, а полевые наблюдения, связанные с анализом положения точек, всегда приводят к этим или сходным задачам. Хотя географы уделяют большое внимание результатам культурной деятельности человека, разработанная ими методология применима непосредственно при изучении природных явлений.

Существующие схемы расположения точек на картах удобно разделить на три категории: равномерные, случайные и групповые. Примеры этих трех типов расположения приведены на рис. 5.8. Конечно, для большинства карт характерны схемы распределения точек, занимающих промежуточное положение между перечисленными крайними типами, и обычно задача заключается в классификационном отнесении наблюдаемой схемы к одному из этих типов. Например, большинство читателей отнесли бы распределение точек на рис. 5.9 к случайному типу, что было бы неверно, так как на карту предварительно нанесена регулярная сеть, в каждой из ячеек которой затем помещалась наудачу одна точка. Таким образом, это распределение обладает как случайными, так и регулярными свойствами.

Схема расположения точек на карте называется равномерной, если плотность точек в любой подобласти равна плотности точек во всех других подобластях. Схема называется регулярной, если точки образуют какой-либо вид сети. Это значит, что расстояния между точками i и j, лежащими на некотором направлении сети, остаются постоянными для всех пар i и j на карте. Случайная схема возникает в том случае, если любая подобласть одного размера характеризуется одной и той же вероятностью появления в ней точки, и появление одних точек не влияет на появление других.

Равномерность расположения точек – важное условие, необходимое для применения многих видов анализов, в частности, анализа поверхностей тренда, который мы рассмотрим несколько позднее. Достоверность карты находится в прямой зависимости от плотности и равномерности расположения точек наблюдения. Однако большинство геологов оценивают распределение точек наблюдения лишь с качественных позиций. Даже несмотря на то что часто подчеркивается желание получить равномерное распределение точек наблюдения, степень равномерности крайне редко измеряется. Критерии, применяемые для определения равномерности, очень просты, но, к сожалению, многие геологи не подозревают об их существовании и о том, что ими можно пользоваться. Однако эти критерии широко используются географами, например Кингом [44], Коулом и Кингом [17], Хеггетом [34] и др.

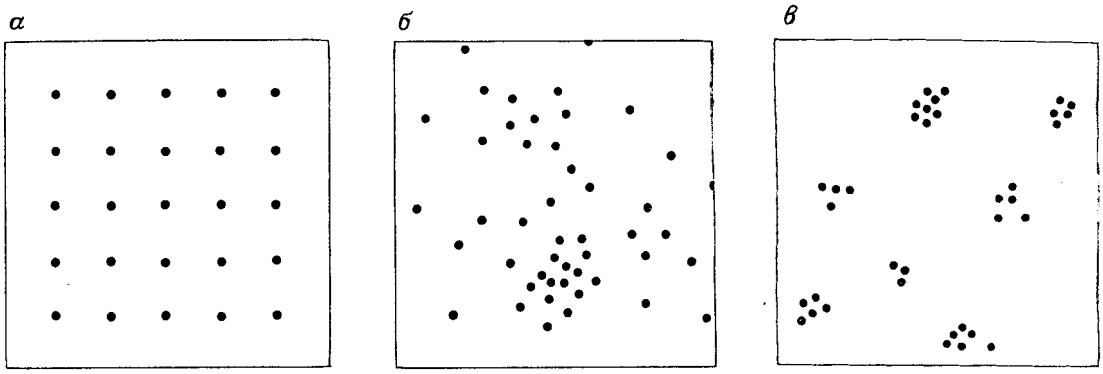


Рис. 5.8. Возможные схемы точек на картах: *а* – точки, регулярно расположенные на сетке; *б* – точки, размещенные случайно; *в* – сгруппированные точки

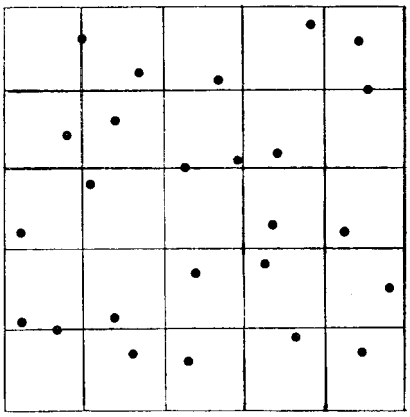


Рис. 5.9. Случайное расположение, полученное случайным выбором точки в пределах регулярной схемы ячеек. Распределение точек более равномерное, чем при полностью случайном расположении точек

Равномерные схемы

Всю карту можно разделить на множество подобластей равного размера (иногда это бывают квадраты) так, что каждая подобласть будет содержать некоторое множество точек. Если точки наблюдения расположены равномерно, то следует ожидать, что каждая подобласть будет содержать одно и то же число точек. Эту гипотезу об отсутствии существенных различий в числе точек для каждой подобласти можно проверить с помощью критерия χ^2, который теоретически не зависит от формы или ориентировки подобластей. Однако критерий будет наиболее эффективным, если число подобластей сделать по возможности большим (что приводит к увеличению числа степеней свободы), при условии, что все подобласти содержат не менее пяти точек. Ожидаемое число точек для каждой подобласти будет равно

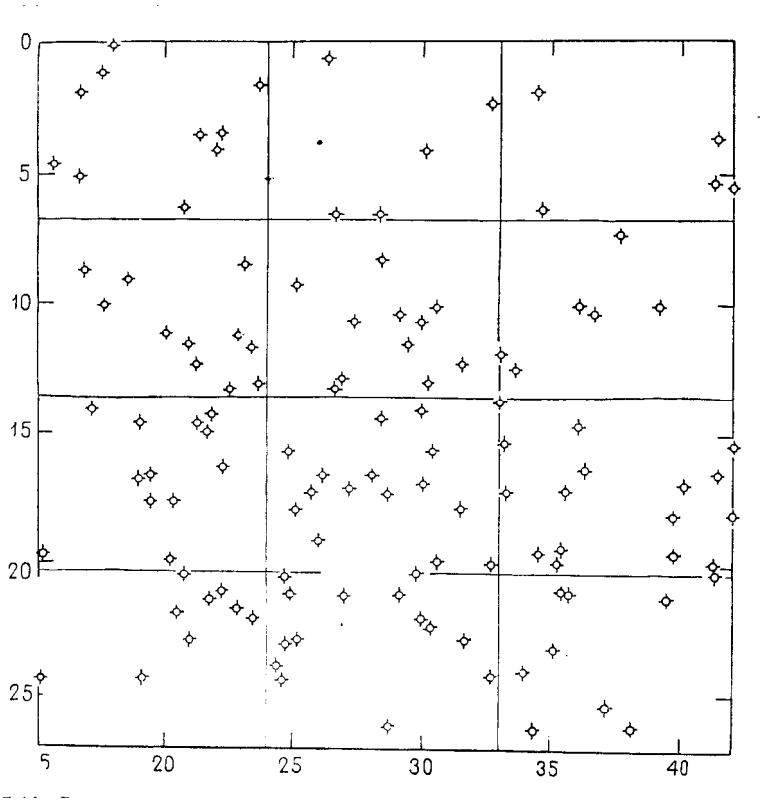
$$E = \frac{\text{общее число точек наблюдения}}{\text{число подобластей}} \quad (5.9)$$

Критерий χ^2 для проверки гипотезы о равномерном распределении точек будет определен следующим образом:

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)}{E_i} \qquad (5.10).$$

где O_i – наблюдаемое число точек в подобласти с номером i; E_i – ожидаемое число, определяемое выражением (5.9). Критерию χ^2 соответствует $\nu = m - 2$ степеней свободы, где m – число подобластей.

В качестве примера применения этого критерия рассмотрим данные, приведенные на рис. 5.10, которые показывают расположение 123 нефтяных скважин в Центральном Канзасе. Этими данными мы воспользуемся несколько позднее при построении поверхности тренда для кровли ордовикских отложений этого региона. На рис. 5.10 вся площадь карты разделена на 12 равных участков и число точек для каждого участка равно приблизительно 10. В табл. 5.1 приведены наблюдаемые значения числа точек в каждом участке, а также показана процедура вычисления критерия χ^2. Так как в данном случае число степеней свободы $\nu = 10$, то критическое значение χ^2, соответствующее 5%-ному уровню значимости, равно 18,3. Вычисленное значение критерия, равное 15,23, не превышает 18,3, что дает основание сделать вывод о несущественном отклонении распределения точек от равномерного. Заметим, что этот вывод касается только однородности распределения точек по участкам определенного размера. Вполне возможно, что существует такой вариант размера квадрата (особенно если он меньше, чем выбранный), при котором гипотеза о равномерности будет отклонена.



с. 5.10. Расположение 123 скважин, пробуренных в верхних слоях ордовикских отложений в Центральном Канзасе. Карта разделена на 12 ячеек одинакового размера

Таблица 5.1. Число скважин по 12 клеткам карты Центрального Канзаса

Наблюдаемое число точек	$\frac{(O-E)}{E}$
10	0,00
5	2,60
5	2,60
11	0,06
12	0,32
6	1,73
12	0,32
16	3,30
15	2,26
9	0,14
14	1,42
8	0,48
Сумма 123	χ^2=15,23*

* Значение критерия несущественно при уровне значимости $\alpha = 0{,}05$.

Р

и

Случайные схемы

Установление факта равномерности расположения точек на карте ни в коей мере не определяет природу равномерности, и в данном случае можно ожидать как регулярный, так и случайный типы однородности. Однако для большинства задач такое выявление равномерности распределения достаточно. Если же нам потребуются дополнительные сведения об изучаемой схеме расположения точек, то для их получения придется обратиться к другому критерию. Если точки равномерно распределены по карте, не следует ожидать, что их число, приходящееся на каждый участок, будет одинаковым для всех участков. Скорее всего, мы увидим, что имеется число, наблюдаемое чаще других, тогда как отклонения от него в ту или иную сторону наблюдаются реже. Это отчетливо видно из только что рассмотренного примера. Несмотря на то что ожидаемое значение числа точек в одном участке равно 10, в действительности мы наблюдаем отклонения от этого числа в ту или иную сторону.

Напомним, что распределение Пуассона является предельным случаем биномиального распределения, когда p (вероятность успеха) очень мала и $1–p$ приближается к 1,0. Пуассоновское распределение может использоваться как модель встречаемости редких, случайно происходящих во времени событий, как это показывалось в гл. 4, или же оно может использоваться для моделирования случайного размещения точек в пространстве. Хотя в определении пуассоновского распределения, как и в определении биномиального распределения, при вычислении вероятностей используется одинаковая терминология (число успехов, неудач и испытаний), его можно представить в таком виде, что ни число неудач, ни общее число испытаний не требуются. В нем используется число точек в квадрате и плотность точек на всей площади для предсказания того, как много квадратов должно содержать заданное число точек. Это предсказанное или ожидаемое число квадратов может быть использовано в критерии χ^2 для определения того, является ли распределение точек внутри заданной площади случайным.

В качестве приложения рассмотрим задачу, состоящую в проверке того, случайно ли откры-

тие продуктивных нефтяных скважин в некотором бассейне или они распределены иначе. Интуитивно не очевидно, что пуассоновское распределение пригодно для решения этой задачи, поэтому мы рассмотрим ее подробнее.

Предположим, что некоторый бассейн имеет площадь a, и в нем случайным образом расположено m разведочных скважин. Обозначим через λ плотность расположения скважин в бассейне, т.е. $\lambda = m/a$. Изучаемый бассейн можно разделить на небольшие участки площади A (термин «участок» эквивалентен термину «квадрат»). В свою очередь каждый участок можно разделить на n крайне малых равных площадок, которые можно рассматривать как потенциальные места бурения.

Вероятность того, что какая-либо из этих крайне малых площадок содержит скважину, стремится к нулю, так как п становится бесконечно большим:

$$p = \lambda \frac{A}{n}$$

а вероятность того, что она не содержит скважины, равна

$$1 - p = 1 - \lambda \frac{A}{n}$$

Надо найти вероятность того, что r из n «возможных мест бурения» внутри участка содержит скважины, а $(n–r)$ таких мест скважин не содержат. Вероятность появления конкретной комбинации точек со скважинами и без них внутри участка равна

$$p = \left(\lambda \frac{A}{n}\right)^r \left(1 - \lambda \frac{A}{n}\right)^{n-r}$$

Однако имеется $\binom{n}{r}$ комбинаций n «буровых точек», из которых r содержат продуктивные скважины в пределах участка, и все варианты равновероятны. Вероятность того, что некоторый участок содержит в точности г продуктивных скважин, равна

$$P(r) = \binom{n}{r}\left(\lambda \frac{A}{n}\right)^r \left(1 - \lambda \frac{A}{n}\right)^{n-r}$$

Заметим, что это просто биномиальная вероятность появления r скважин на n «возможных местах бурения». Приведенную выше формулу можно выразить через факториалы

$$P(r) = \frac{n(n-1)(n-2)....(n-r+1)}{r!} \frac{(\lambda A)^r}{n^r} \left(1 - \frac{\lambda A}{n}\right)^n \left(1 - \frac{\lambda A}{n}\right)^{n-r}$$

Переставляя и сокращая члены, получаем

$$P(r) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)..\left(1 - \frac{r-1}{n}\right)\left(1 - \frac{\lambda A}{n}\right)^{n-r}\left[\left(1 - \frac{\lambda A}{n}\right)^n \frac{(\lambda A)^r}{r!}\right]$$

При неограниченном возрастании n все дроби в скобках, которые содержат n в знаменателе, бесконечно малы и стремятся к нулю, так как все члены в скобках становятся просто равными 1. Члены внутри скобок упрощаются и принимают вид

$$p(r) = e^{(-\lambda A)} \frac{(\lambda A)^r}{r!} \quad (5.11)$$

Заметим, что число «потенциальных мест бурения» n равно нулю для уравнения, имеющего только плотность скважин λ, число скважин r и площадь участка A. Это как раз выражение распределения Пуассона, как и положено для вероятности редких случайных событий (появление скважин), происходящих на географических площадях. Заметим также, что λA – это просто среднее число скважин на участок, так как оно есть произведение плотности скважин на площадь участка. На практике мы оцениваем λA по общему числу скважин m и общему числу участков T:

$$\lambda A = m/T \quad (5.12)$$

Мы можем теперь применить критерий χ^2 для проверки того, согласуется ли ожидаемое число скважин на участок с числом скважин, вычисленным в предположении, что они распределены случайно в соответствии с распределением Пуассона. Число участков, которое содержит в точности

r скважин, может быть найдено по формуле

$$n_r = mP_{(r)} = me^{(-\lambda A)}\frac{(\lambda A)^r}{r!} \quad (5.13)$$

Если λA оценивается через m/T, то уравнение принимает вид

$$n_r = me^{(-m/T)}\left(\frac{m}{T}\right)^r\frac{1}{r!} \quad (5.14)$$

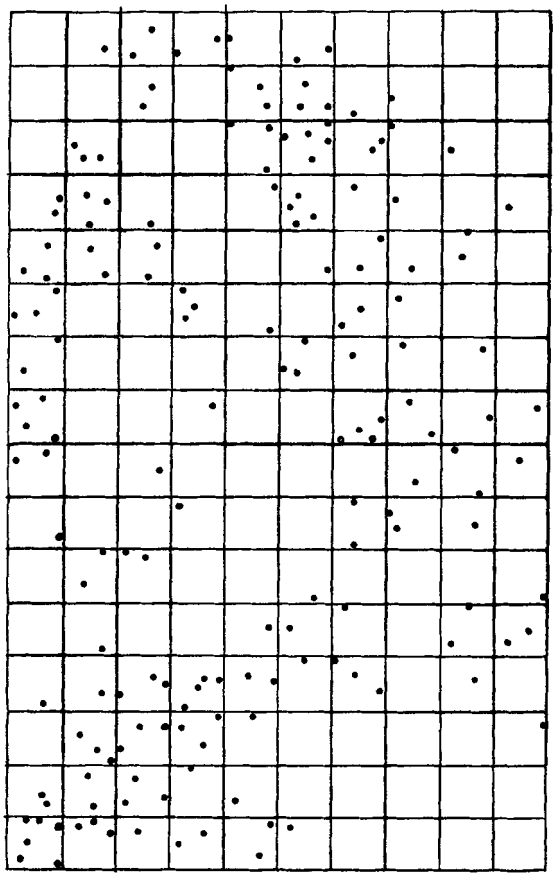


Рис. 5.11. Расположение нефтяных поисковых скважин в восточной части шельфа пермского бассейна, графства Фишер и Ноланд, Техас

Рис. 5.11 указывает расположение скважин в части восточного шельфа пермского бассейна в Техасе. Площадь была разделена на 160 участков (квадратов), площадь каждого из которых равна приблизительно 0,16 км^2. Так как на площади имеется 168 скважин, то их среднее число на участок

$$m/T = 168/160 = 1{,}05.$$

Мы можем подсчитать число участков на карте, которые не содержат ни одной скважины, содержат в точности одну, две и так далее. Мы также можем вычислить математическое ожидание участков, которые содержат те же самые числа скважин, используя уравнение (5.14). Для площади Пермского бассейна ожидаемые и наблюденные числа участков приведены в табл. 5.2.

Таблица 5.2 содержит все числа, необходимые для вычисления критерия соответствия χ^2, который есть не что иное как сравнение двух гистограмм, представленных на рис. 5.12. Последние три категории должны быть скомбинированы так, чтобы наблюдаемое число участков было равно 5:

$$\chi^2 = \frac{(70-56{,}0)^2}{56{,}0}56{,}0 + \frac{(42-58{,}8)^2}{58{,}8}58{,}8 + \frac{(26-30{,}9)^2}{30{,}9}30{,}9 + \frac{(17-10{,}8)^2}{10{,}8} + \frac{5-3{,}5)^2}{3{,}5} = 13{,}28$$

Проверяемая статистика имеет $c–2$ степеней свободы, где c – число категорий (первая степень свободы теряется потому, что ожидаемые частоты подчинены условию равенства их суммы 160 и вторая степень свободы требуется для оценки параметра λ. Для $c=5$ имеется три степени свободы. Критическое значение χ^2 для $\nu = 3$ и $\alpha = 0{,}05$ равно 7,81. Проверяемая статистика значительно превышает это значение, поэтому мы должны отклонить гипотезу о равенстве между наблюденным и ожидаемым распределениями и заключить, что пуассонова модель неправомерна. Скважины не разбросаны случайно на этой площади пермского бассейна.

Таблица 5.2 Вычисление ожидаемых чисел участков, содержащих r открытий, в восточной части пермского бассейна, Техас. Предполагается, что распределение пуассоново

Число открытий на участок (r)	Уравнение Пуассона	Вероятность того, что участок содержит r открытий	Число участков	
			ожидаемое	наблюденное
0	$P_{(0)} = e^{(-1{,}05)} \frac{(1{,}05)^0}{0!}$	0,3499	56,0	70
1	$P_{(1)} = e^{(-1{,}05)} \frac{(1{,}05)^1}{1!}$	0,3674	58,8	42
2	$P_{(2)} = e^{(-1{,}05)} \frac{(1{,}05)^2}{2!}$	0,1929	30,9	26
3	$P_{(3)} = e^{(-1{,}05)} \frac{(1{,}05)^3}{3!}$	0,0675	10,8	17
4	$P_{(4)} = e^{(-1{,}05)} \frac{(1{,}05)^4}{4!}$	0,0177	2,8	3
5	$P_{(5)} = e^{(-1{,}05)} \frac{(1{,}05)^5}{5!}$	0,0038	0,6	1
6	$P_{(6)} = e^{(-1{,}05)} \frac{(1{,}05)^6}{6!}$	0,0007	0,1	1
Суммы		0,9998	160,0	160

В процессе подбора модели Пуассона к этим данным мы получили некоторую информацию, которая могла бы пролить дополнительный свет на природу пространственного распределения. Среднее число скважин на участок оценивается уравнением (5.12). Дисперсия числа скважин на участок есть

$$s^2 = \frac{\sum_{i=1}^{r}\left(r_i - \frac{m}{T}\right)^2}{T-1} \qquad (5.15)$$

где r_i – число скважин в i-м участке. Суммирование распространяется на все T участков. При сравнении среднего и дисперсии альтернативы таковы:

$m/T > s^2$ – схема ближе к равномерной, чем к случайной;

$m/T = s^2$ – случайная схема;

$m/T < s^2$ – схема ближе к кластеризованной, чем к случайной.

глубина структуры = $b_0 + b_1$ (время возвращения сейсмической волны).

глубина структуры = 1389,6–1692,5 (время возвращения сейсмической волны).

Конечно, некоторые различия между m/T и s^2 могут возникнуть в силу случайного изменения выбора конкретного множества участков. Статистическая значимость наблюденной разности может быть проверена с помощью t-критерия, основанного на стандартном отклонении среднего, равном корню квадратному из дисперсии, которая могла бы быть получена в точке m/T, если бы бассейн был повторно опробован другим множеством участков того же размера. Стандартное отклонение среднего числа скважин, приходящихся на участок, есть

$$s_e = \sqrt{\frac{2}{T-1}} \qquad (5.16)$$

С помощью t-критерия сравнивается отношение m/T к s^2, которое должно быть равным 1,0 при условии, что две статистики одинаковы:

$$t = \left[\left(m/T\right)/s^2 - 1{,}0\right]/s_e \qquad (5.17)$$

Этот критерий имеет T–1 степеней свободы.

Для площади восточной части пермского бассейна дисперсия числа скважин на участок есть

$$s^2 = 231{,}6 : 159 = 1{,}46$$

стандартное отклонение среднего числа скважин на участок может быть оценено так:

$$s_2 = \sqrt{2/159} = 0{,}112$$

t-статистика для проверки эквивалентности среднего и дисперсии есть

$$t = \left[(1{,}05 : 1{,}46) - 1{,}0\right] : 0{,}112 = -8{,}86$$

Для уровня значимости $\alpha = 0.05$ и 159 степеней свободы критическое значение t для двустороннего критерия равно ±1,96; вычисленное значение значительно превышает это значение, и потому мы можем заключить, как мы это делали в разделе о критерии χ^2, что пространственное распределение неслучайно. Так как дисперсия значительно больше среднего, то мы должны заключить, что скважины образуют группы на изучаемой площади.

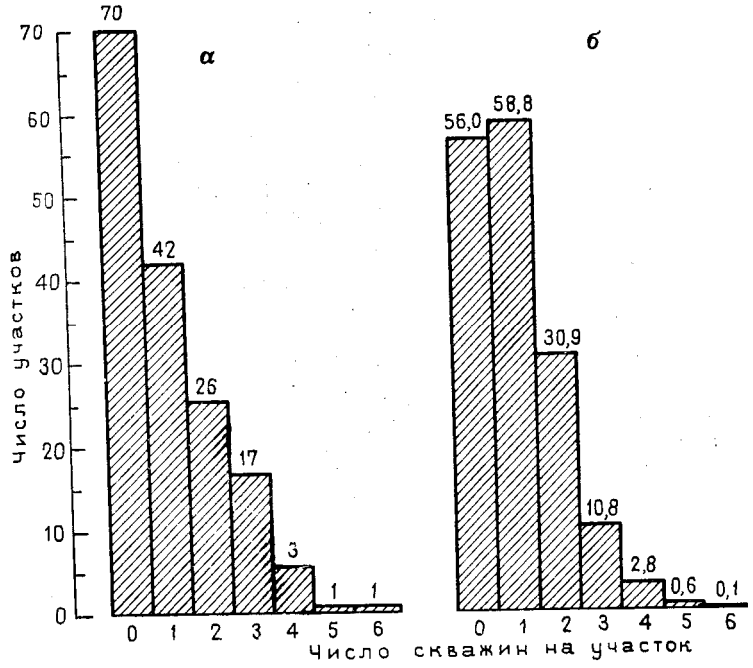


Рис. 5.12. Гистограммы, показывающие наблюдаемые числа скважин на участок на площади пермского бассейна (я) и ожидаемые числа, если поля распределены случайно в соответствии с пуассоновской моделью (б)

Схемы группирования

Многие встречающиеся в природе распределения обнаруживают ярко выраженную тенденцию к группированию. Это особенно верно для некоторых биологических переменных, таких, как наличие специфических организмов или присутствие инфекции. Потомки древних, непередвигающихся родителей, как, например, кораллы или деревья, имеют тенденцию к росту в ближайшей окрестности, что приводит к развитию плотно заселенных площадей, окруженных относительно пустынными регионами. Схемы группирования точек могут быть представлены различными моделями распределений, большинство из которых можно рассматривать как комбинации двух или более простых распределений. Одно из этих распределений описывает положения центров групп, в то время как другие описывают расположение точек вокруг этих центров.

Отрицательное биномиальное распределение можно использовать для моделирования расположения групп в пространстве таким же образом, каким пуассоновское распределение использовалось для моделирования случайного расположения точек. Подробное обсуждение этого распределения с примерами его приложений во многих отраслях приводится в [65]. Гриффитс [31, 32] приводит длинное обоснование пригодности отрицательного биномиального распределения в качестве модели частоты появления нефтяных полей и рудных тел.

Один из способов получения отрицательного биномиального распределения – это компози-

ция пуассоновского и логарифмического распределений, когда группы точек случайно расположены в пределах некоторого региона и индивидуальные точки внутри групп подчиняются логарифмическому распределению. В формулировке, пригодной для описания пространственных схем, отрицательное биномиальное распределение есть

$$P(r) = \left(\frac{k+r-1}{r}\right)\left(\frac{p}{1+p}\right)^r\left(\frac{1}{1+p}\right)^k \quad (5.18)$$

В терминах проблем разведки нефти, как мы уже видели, г есть число скважин на участок, p – вероятность того, что некоторая заданная разведочная площадка содержит скважину, и k есть мера степени группирования скважин. Если k велико, то группирование менее ярко выражено и пространственное распределение приближается к пуассоновскому или случайному. Если k стремится к нулю, то схемы группирования становятся более явными. Плотность равна

$$\lambda = kp \quad (5.19)$$

Если k не есть целое (и вообще оно не должно быть таким), то это комбинаторное уравнение не может быть решено. Тогда следует использовать следующую аппроксимацию

$$P(0) = \frac{1}{(1+p)^k} \quad (5.20)$$

$$P(r) = \frac{(k+r-1)\left(\frac{p}{1+p}\right)}{r}P(r-1) \quad (5.21)$$

Как и в пуассоновском распределении, λ оценивается как средняя плотность скважин на участок m/T. Параметр группирования k оценивается по формуле

$$k = \frac{(m/T)^2}{s^2 - (m/T)} \quad (5.22)$$

где s^2 – дисперсия числа скважин на участок. Тогда вероятность оценивается по формуле

$$p = \lambda / k = (m/T)/k \quad (5.23)$$

Можно применить отрицательную биномиальную модель к данным по скважинам в восточной части пермского бассейна (см. рис. 5.11) для того, чтобы установить, может ли это распределение адекватно описывать их пространственное распределение. Среднее значение и дисперсия числа скважин на участок уже были найдены и равны m/T=1,05 и s^2=1,46. Эффект группирования можно оценить, используя уравнение (5.22), как

$$k = 1{,}05^2/(1{,}46 - 1{,}05) = 2{,}69$$

В свою очередь вероятность встретить скважину на некотором участке равна

$$p = 1{,}05/2{,}69 = 0{,}390.$$

Используя приближенные уравнения, получаем вероятность того, что данный участок не будет содержать скважин:

$$P(0) = \frac{1}{(1+0{,}390)^{2{,}69}} = 0{,}4124$$

Вероятность того, что некоторый участок будет содержать в точности одну скважину, есть

$$P(1) = \frac{(2{,}69+1-1)(0{,}390/1{,}390)}{1}0{,}4124 = 0{,}3112$$

Вероятности того, что участок будет содержать в точности две, три или больше скважин, могут быть вычислены аналогично. Тогда ожидаемое число участков, содержащих r скважин, может быть определено просто умножением этих вероятностей на 160 – общее число участков. В табл. 5.3 представлены вероятности, соответствующие числам скважин вплоть до шести открытий на участок.

Числа участков, содержащих в точности г скважин, вычисленные с помощью отрицательного биномиального закона распределения, сравниваются с соответствующими наблюденными числами участков, приведенными на рис. 5.13. Соответствие отрицательного биномиального распределения может быть проверено с помощью критерия χ^2 в точности аналогично тому, как проверялось

соответствие пуассоновской модели. Если необходима комбинация окончательных трех категорий, то получается пять частот. Проверяемая статистика есть $\chi^2 = 4{,}82$ с $5-2$ степенями свободы. Это меньше, чем критическое значение χ^2 для $\alpha = 0{,}05$ и $\nu = 3$, поэтому мы не можем отклонить гипотезу об отрицательном биномиальном распределении как модели пространственного распределения скважин в восточной части пермского бассейна. Необходимо иметь в виду, что полученный вывод не эквивалентен доказательству того, что скважины подчиняются отрицательной биномиальной модели. Вполне возможно, что некоторые другие модели группирования могут давать более точную аппроксимацию. Однако отрицательное биномиальное распределение генерирует некоторое пространственное распределение, которое статистически неотличимо от наблюденного распределения.

Таблица 5.3 Ожидаемое число участков, содержащих *r* открытий, в восточной части пермского бассейна. Предполагается отрицательное биномиальное распределение

Число открытий на участок (*r*)	Вероятность того, что участок содержит *r* открытий	Число участков	
		ожидаемое	наблюденное
0	0,4124	66,0	70
1	0,3112	49,8	42
2	0,1611	25,8	26
3	0,0706	11,3	17
4	0,0281	4,5	3
5	0,0106	1,7	1
6	0,0038	0,6	1
Сумма	0,9988	159,7	160

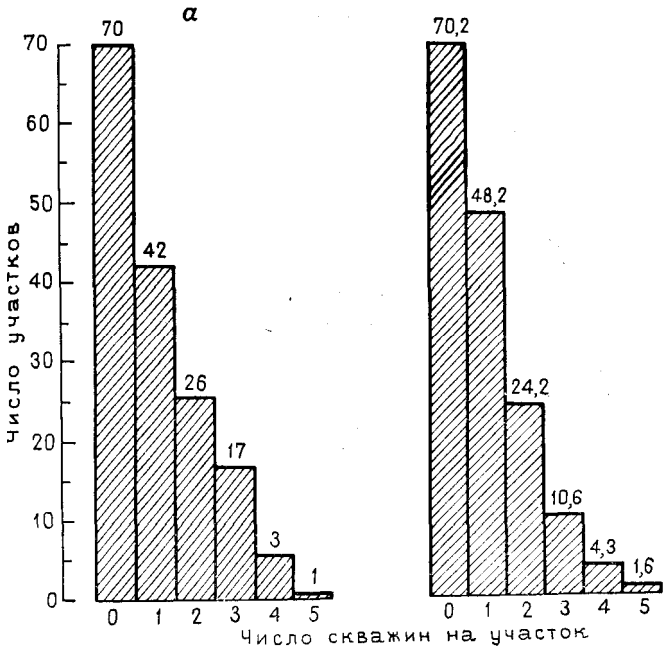


Рис 5 13 Гистограммы, показывающие наблюдаемые числа разведочных скважин на участок на одной из площадей пермского бассейна (*а*) и ожидаемые числа в отрицательной биномиальной модели (*б*)

Метод ближайшего соседа

Существует еще один метод исследования подмножеств, на которые разбита некоторая область – метод ближайшего соседа. Анализируемые данные в этом случае представляют собой не множество точек, расположенных внутри некоторой заданной области, а расстояния между наиболее близкими парами точек. Так как не обязательно выбирать размеры квадрата, процедуры поиска ближайших соседей исключают возможность получения схемы, которая является случайной в од-

ном масштабе, и не является случайной в другом. Поскольку обычно имеется намного больше пар ближайших соседей, чем квадратов, этот анализ более чувствителен. Хорошее введение в методы ближайшего соседа дают Джетис и Бутс [29]; Риплай [65], а также Клифф и Орд [15] приводят обзор теории с применениями в разных областях.

Метод ближайшего соседа основан на сравнении наблюдаемого множества расстояний между парами ближайших точек с характеристиками, которые ожидались бы в том случае, если бы точки были случайно распределены. Характеристики теоретической случайной схемы можно вычислить из пуассоновского распределения. Если мы игнорируем эффект краев нашей карты, то ожидаемое расстояние между ближайшими соседями есть

$$\overline{\rho} = 0{,}5\sqrt{A/n} \qquad (5.24)$$

где A – площадь карты; n – число точек. Напомним, что – есть плотность точек λ. Выборочная дисперсия величины $\overline{\delta}$ задается по формуле

$$\sigma^{\frac{2}{\delta}} = (4-\pi)A/(4\pi n^2) \qquad (5.25)$$

Проведя действия с константами, получим

$$\sigma^{\frac{2}{\delta}} = 0{,}06831A/n^2 \qquad (5.26)$$

Стандартное отклонение среднего расстояния между ближайшими соседями есть квадратный корень из $\sigma^{\frac{2}{\delta}}$:

$$s_e = 0{,}26136/\sqrt{A/n^2} \qquad (5.27)$$

Распределение $\overline{\delta}$ нормально при условии, что n больше 6, так что мы можем использовать простой Z-критерий, приведенный в гл. 2, для проверки гипотезы о том, что наблюденное среднее расстояние между ближайшими соседями $\overline{d}$ равно значению δ для случайной схемы точек той же плотности. Значение критерия есть

$$Z = (\overline{d} - \overline{\delta})/s_e \qquad (5.28)$$

Это – общепринятый вид критерия ближайшего соседа, однако, к сожалению, он имеет значительный дефект в большинстве практических случаев. При построении ожидаемого значения $\overline{\delta}$ предполагается, что краевой эффект полностью отсутствует, а это означает, что наблюденные схемы точек могут быть распространены неограниченно во всех направлениях, если $\overline{d}$ и $\overline{\delta}$ обоснованно сравниваются. Так как карта не распространяется неограниченно, то ближайшие окрестности точек вблизи краев должны лежать внутри поля карты и потому $\overline{d}$ смещено в направлении больших значений. Имеется несколько поправок в решении этой задачи. Если данные за пределами исследуемой площади доступны, то карту можно окружить охранной областью. Тогда расстояния, вычисленные по методу ближайшего соседа между точками внутри карты и точками в охранной области, можно включить в вычисление $\overline{d}$. Другой способ состоит в том, что мы можем считать нашу карту вычерченной не на плоскости, а на торе. Это значит, что правый край карты считается склеенным с левым краем, а нижний край – склеенным с верхним. Тогда ближайшая соседняя точка к точке, лежащей у правого края, может быть расположена вблизи левого края (такое использование точек хорошо известно всякому, кто имел дело с построением изолиний плотностей по стереосетям). Еще один способ построения поправок состоит в построении повторений во всех направлениях, подобно мозаике. Для любой точки, примыкающей к краю карты, имеется точка, которую можно с большим основанием считать ближайшим соседом, чем ближайшую точку внутри заданной карты.

Третий способ построения поправок состоит в изменении $\overline{\delta}$ так, чтобы граничный эффект был включен в ожидаемое значение. Используя численное моделирование, Доннелли [24] нашел эти альтернативные выражения для теоретического значения средних расстояний по методу ближайшего соседа и его выборочного среднего:

$$\overline{\delta} = \frac{1}{2}\sqrt{\frac{A}{n}} + \left(0{,}514 + \frac{0{,}412}{\sqrt{n}}\right)\frac{p}{n} \qquad (5.29)$$

$$s_{\bar{\delta}}^2 \approx 0{,}070\frac{A}{n^2} + 0{,}035p\frac{\sqrt{A}}{n^{3/2}} \qquad (5.30)$$

В этой аппроксимации p есть периметр прямоугольной карты. Заметим, что если карта не имеет краев, как это имело место на торе, то p равно нулю, и эти уравнения тождественны уравнениям (5.24) и (5.26).

Ожидаемые и наблюденные средние значения расстояний по методу ближайшего соседа могут быть использованы для построения индекса пространственной схемы. Отношение

$$R = \bar{d} / \bar{\delta} \qquad (5.31)$$

есть статистика ближайшего соседа и изменяется от $0{,}0$ для распределения, в котором все точки совпадают и разделены расстояниями, равными нулю, до значения $1{,}0$ для случайного распределения точек, т. е. до максимального значения $2{,}15$. Последнее значение характеризует распределение, в котором среднее расстояние до ближайшей окрестности максимизировано. Распределение имеет форму регулярной шестиугольной схемы, в которой каждая точка равноудалена от шести других точек. На рис. 5.14 представлены множества точек с различными значениями статистики ближайшего соседа, причем все имеют одну и ту же точечную плотность.

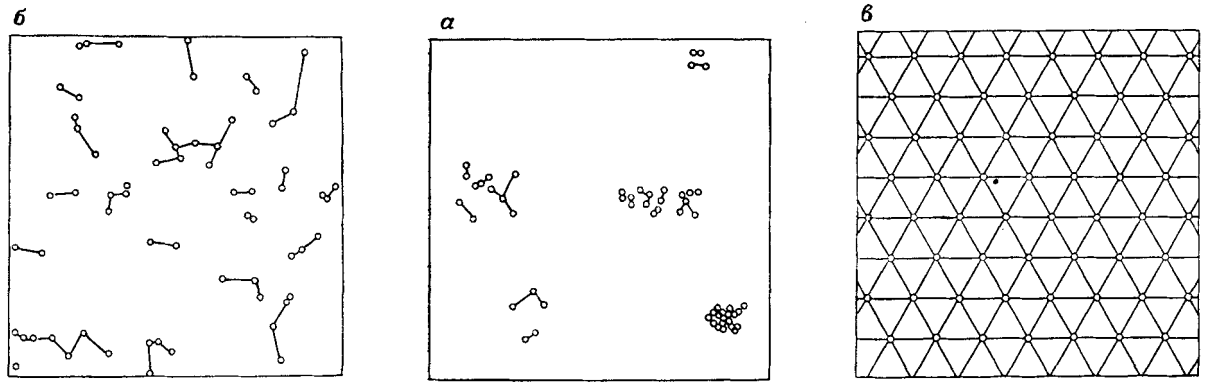


Рис. 5.14. Значения статистик метода ближайшего соседа R для различных схем расположения точек на картах (по Р. А. Ома, 1984): *а* – точки сгруппированы в пять групп, R=0,34; *б* – точки рассеяны случайно, R=0,91; *в* – точки размещены по шестиугольной сети, R=2,15. Плотность точек λ одинакова для всех схем

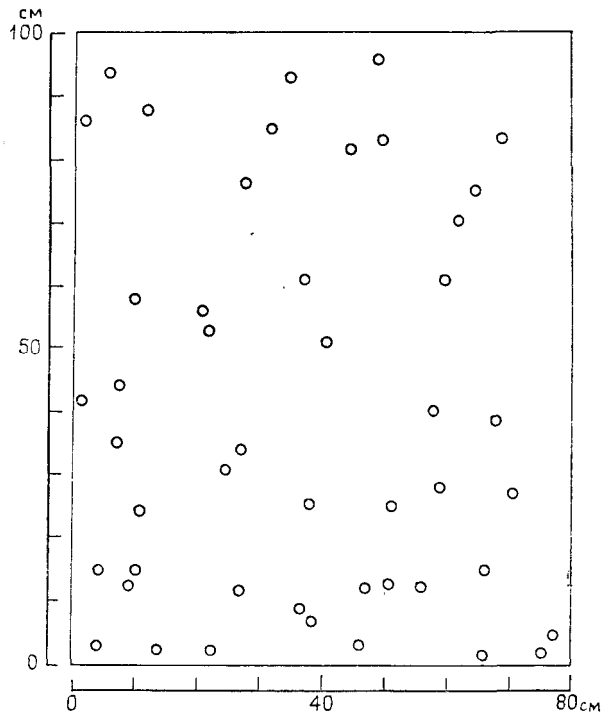


Рис. 5.15. Схематическое представление полированной плиты анортозита, на которой показано расположение кристаллов магнетита

Проиллюстрируем применение метода ближайшего соседа, используя «карту», изображенную на рис. 5.15. Эта «карта» в действительности представляет собой полированную поверхность каменной облицовки фасада здания банка в университетском городке и служит интересным примером петрологического изучения изверженных пород. Это черный анортозит, который содержит мелкие рассеянные кристаллы магнетита. Преподаватели нередко используют эти плиты как наглядное пособие для различной тематики, включая и применение математических методов в петрографии. При этом обычно постулируется, что положение плиты совпадает с ее природной ориентацией. Таким образом, рассматривается вертикаль-

ная поверхность, нижней части которой соответствует нижний обрез плиты. Так называемая «карта» характеризует положение всех наблюдаемых на этой поверхности зерен магнетита. В табл. 5.4 приведены значения координат этих точек в сантиметрах, отсчитываемых от левого нижнего угла плиты. В данном случае задача заключается в получении ответа на вопрос: равномерно ли распределены зерна на поверхности плиты или же они обладают тенденцией к группировке? Кроме того, если распределение зерен неравномерно, спрашивается, можно ли считать плотность зерен у нижнего края плиты превышающей плотность у верхнего края. Получение ответа на подобные вопросы играет существенную роль при решении различных задач петрогенезиса изверженных пород. Рассматриваемые в этой главе методы анализа данных могут оказаться весьма полезными. Проверку гипотезы о равномерном случайном распределении зерен магнетита можно провести с помощью одного из рассматриваемых ранее способов – разделения на более дробные участки или же с помощью метода ближайшего соседа. Вычисления можно проделать вручную, измерив расстояния прямо на рис. 5.15, или вычислив их с помощью координат в табл. 5.4. Риплай [65] (стр. 175 – 181) дает исчерпывающий анализ этих данных, используя различные методы.

Таблица 5.4 Координаты магнетитовых зерен на полированном анортозитовом слое

Расстояние от нижнего левого угла слоя, см		Расстояние от нижнего левого угла слоя, см		Расстояние от нижнего левого угла слоя, см	
Горизонтальное	Вертикальное	Горизонтальное	Вертикальное	Горизонтальное	Вертикальное
1	86	27	34	50	13
2	41	28	76	51	25
4	3	37	14	56	12
4	15	37	61	58	40
8	95	27	85	59	28
9	13	11	25	60	61
7	35	15	15	62	70
8	44	35	93	66	0
10	58	38	25	66	15
12	88	38	7	65	75
14	2	41	51	69	33
22	2	46	2	69	83
21	56	47	12	71	27
22	53	45	82	76	1
24	31	50	83	77	4
27	12	49	96		

РАСПРЕДЕЛЕНИЕ ПРЯМЫХ

Некоторые естественно встречающиеся в природе схемы составлены из линий, таких, как линеаменты, видимые со спутников, следы трещин на выветренной поверхности гранитов, или микроскопические структуры, наблюдаемые в шлифах деформированной породы. Так же, как множество точек может образовывать схему, которая изменяется от равномерной до кластеризованной с прочными связями, так и множество прямых может обладать различными свойствами. Конечно, линии более сложны, чем точки, так как они обладают длиной, ориентацией и местоположением. Соответственно их анализ более труден, и статистические методы, пригодные для изучения множеств линий, менее развиты, чем методы, предназначенные для исследования схем точек. Изучению распределения длин линий посвящено мало работ. Исключение составляют работы по логнормальному распределению [2]. Очень небольшое число ученых исследовали распределение пространства между линиями в заданной схеме, т. е. решали задачу, аналогичную методу ближайшего соседа, в анализе точечных распределений [59, 19]. Значительно полнее освещен вопрос об ориентации линий, который будет рассмотрен в следующем параграфе.

Случайную схему линий можно определить как такую схему, в которой каждое положение

равновероятно может быть пересечено некоторой линией и любая ориентация секущей линии также равновероятна. Такие случайные схемы можно генерировать различными способами; одна из таких процедур состоит в выборе двух пар координат из таблицы случайных чисел и последующем проведении через соответствующие точки прямой. Другой способ состоит в выборе произвольного угла на плоскости, проведении произвольного радиуса внутри этого угла, проведении на произвольном расстоянии от вершины угла прямой, перпендикулярной к выбранному радиусу. Повторяя эту процедуру достаточное число раз, получаем в результате схему статистически неразличимых линий.

Мы можем определить меру плотности линий, аналогичную ранее определенной точечной плотности:

$$\lambda = \sum \frac{l}{A} \qquad (5.32)$$

Величина $\sum l$ – общая длина линий на карте, имеющей площадь A; λ – параметр, который определяет форму пуассоновского распределения; как и следовало ожидать, пуассоновская модель описывает распределение многих параметров схемы, образованной случайными линиями.

Распределение расстояний между парами линий можно исследовать с помощью метода ближайшего соседа. Сначала произвольно выберем по одной точке на линиях карты. Из каждой точки по перпендикуляру к ближайшей линии вычисляется расстояние до нее. Среднее этих расстояний до ближайших соседей $\overline{d}$ есть среднее этих измерений. Эта процедура проиллюстрирована на рис. 5.16.

Дэси [19] определил среднее расстояние $\overline{\delta}$ по методу ближайшего соседа для схемы случайных линий следующим образом:

$$\overline{\delta} = 0{,}31831/\lambda \qquad (5.33)$$

а также дисперсию этого среднего:

$$\sigma^{\frac{2}{\delta}} = 0{,}10132/\lambda^2 \qquad (5.34)$$

Используя эту дисперсию и число линий в схеме, можно найти стандартное отклонение нашей оценки среднего расстояния по методу ближайшего соседа. Стандартное отклонение определено формулой

$$s_e = \sqrt{\sigma^{\frac{2}{\delta}}/n} \qquad (5.35)$$

Она позволяет нам вычислить простую Z-статистику для проверки значимости разности между ожидаемым и наблюденным средним расстоянием, вычисленным по методу ближайшего соседа:

$$Z = \frac{\overline{d} - \overline{\delta}}{s_e} \qquad (5.36)$$

Критерий является двусторонним. Если величина Z не значима, то можно заключить, что наблюденная схема линий неотличима от схемы, порожденной случайным (пуассоновским) распределением. Мы можем также ввести индекс ближайшего соседа, идентичный тому, который использовался для точечных схем, взяв отношение наблюденного и ожидаемого среднего расстояний по методу ближайшего соседа, или $\overline{d}/\overline{\delta}$. Индекс интерпретируется совершенно аналогично индексу для точечных схем.

Этот критерий пригоден для множеств как прямых, так и искривленных линий при условии, что эти линии не часто меняют направление. Значит, линии должны быть по крайней мере в полтора раза длиннее, чем среднее расстояние между ними. Если число линий на одной карте мало, то оцененную плотность рекомендуется снабдить множителем $(n-1)/n$, где n – число линий в схеме. Оценка плотности линий поэтому такова:

$$\lambda = \frac{(n-1)\sum l}{nA} \qquad (5.37)$$

Простой метод исследования множества линий на карте состоит в преобразовании двумерной схемы в одномерную последовательность. Мы можем сделать это, проведя случайно линию на карте и отмечая точки, в которых эта линия пересекает линии заданного множества. Распределение интервалов между точками пересечения вдоль проведенной линии дает информацию о пространст-

венной схеме. Мы можем применить к исследованию этой одномерной последовательности методы, изложенные в гл. 4. Если одна секущая не даст достаточного количества пересечений для того, чтобы можно было применить одномерные критерии, то можно провести некоторую другую случайно ориентированную линию и т.д. (рис. 5.17). Извилистый путь секущей линии напоминает случайное блуждание, и последовательность пересечений может рассматриваться как последовательность, расположенная вдоль одной прямой линии. Этот и другие методы исследования плотности схемы линий рассмотрены в работе [29].

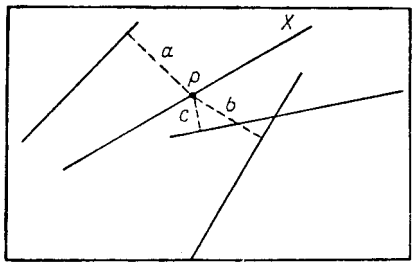


Рис. 5.16. Вычисление расстояний между линиями по методу ближайшего соседа. Точка *p* выбрана случайно на линии *X*. Пунктирные линии *a, b* и *c* – это перпендикуляры, проведенные из точки *p* к ближайшим линиям. Кратчайшая из них *c* – расстояние до ближайшего соседа линии *X*. Процесс повторяется с целью нахождения расстояний по методу ближайшего соседа для всех линий

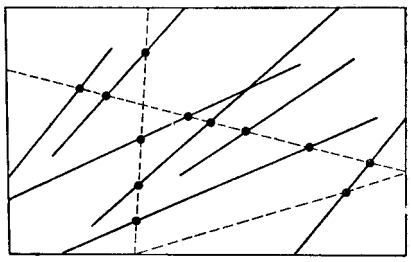


Рис. 5.17. Случайное расположение линий опробования (пунктир), секущих схему линий на карте. Пересечения последними линий опробования образуют последовательность интервалов, для которых можно проверить гипотезу о случайности

АНАЛИЗ НАПРАВЛЕННЫХ ДАННЫХ

Направленные данные – важная категория геологической информации. Все объекты, такие, как плоскости напластований, складчатые поверхности и трещины, характеризуются залеганием, выраженным простиранием пластов и их наклоном. Ледниковые трещины, маркирующие горизонты, ископаемые раковины и отшлифованная водой галька имеют преобладающие направления ориентации. Фотографии, полученные с воздуха и спутников, также показывают наличие ориентированных линейных схем. Указанные объекты могут быть исследованы количественными методами, наподобие измерений других геологических свойств, однако необходимо использовать специальные статистики, отражающие циклическую (или сферическую) природу направленных данных.

Учитывая практику географов, мы должны различать направленные и ориентированные объекты. Предположим, что автомобиль движется на север вдоль шоссе; движение автомобиля имеет направление, в то время как шоссе само имеет только ориентацию север – юг. Простирания слоев пород и прослеживание складок – примеры ориентированных геологических наблюдений, в то время как друмлины и некоторые ископаемые, такие, как сильно скрученные гастроподы, имеют ярко выраженные направленные характеристики.

Мы можем также выделить наблюдения, которые удобно изображать точками на окружности, или наблюдения, естественно представленные на сфере, например, метаморфические структуры. Первые из упомянутых данных удобно изображать на диаграмме, имеющей форму циклической гистограммы, вторые – изображаются точками в проекции на полусферу. Хотя геологи многие годы представляют направленные данные в таком виде, они не разработали формальных статистических процедур для проверки истинности заключений, которые извлекались из этих диаграмм. Такое положение достойно сожаления не только потому, что эти критерии полезны сами по себе, но и потому, что развитие многих из этих методов было инспирировано первоначально проблемами наук о Земле.

На рис. 5.18 представлена карта трещиноватости ледников, измеренной на малой площади в Южной Финляндии (табл. 5.5). Направления, указанные штрихами, можно изобразить единичными векторами или же на окружности единичного радиуса (рис. 5.19, *а*). Если окружность разбить на

сегменты и подсчитать число векторов в каждом сегменте, то результат можно представить в виде розы или циклической гистограммы (см. рис. 5.19, *б*). Однако для того, чтобы вычислить статистики, которые дают характеристики всего множества векторов, мы должны работать непосредственно с самими измерениями. (Необходимо отметить, что, следуя соглашению между геологами, мы измеряем углы против часовой стрелки от положительного направления оси *Y*, т.е. от направления на север). В большинстве работ по статистике направленных данных углы измеряются против часовой стрелки от положительного направления оси *X*, т.е. направления на восток.

Таблица 5.5. Вектор направлений штриховки ледника, измеренный на площади в Южной Финляндии (значения даны в градусах, отсчет велся против часовой стрелки, начиная с севера)

23	99	125	132	146	172
27	100	126	132	153	179
53	105	126	132	155	181
58	113	126	134	155	186
64	113	127	135	155	190
83	114	127	137	157	212
85	117	128	144	163	
88	121	128	145	165	
93	123	129	145	171	

Главное направление в некотором множестве векторов можно найти, вычисляя результирующий вектор – результант. Пусть X и Y – координаты конца единичного вектора, направление которого задается углом 9:

$$X_i = \cos\theta_i, \quad Y_i = \sin\theta_i \qquad (5.38)$$

Такие три вектора представлены на рис. 5.20. Также показан результирующий вектор, полученный суммированием синусов и косинусов индивидуальных векторов:

$$X_r = \sum_{i=1}^{n} \cos\theta_i, \quad Y_r = \sum_{i=1}^{n} \sin\theta_i \qquad (5.39)$$

Из результирующего вектора можно получить среднее направление θ, которое является угловым средним всех векторов выборки. Оно совершенно аналогично среднему значению множества скалярных измерений:

$$\overline{\theta} = \tan^{-1}(Y_r / X_r) = \tan^{-1}(\sum_{i=1}^{n} \sin\theta_i / \sum_{i=1}^{n} \cos\theta_i) \qquad (5.40)$$

Очевидно, величина или длина результирующего вектора зависит отчасти от вклада дисперсии выборки векторов, но также и от числа векторов. Для сравнения результирующих векторов выборок различного объема их необходимо привести к стандартному виду путем деления координат результирующего вектора на число наблюдений:

$$\overline{C} = \frac{X_r}{n} = \frac{1}{n}\sum_{i=1}^{n} \cos\theta_i, \quad \overline{S} = \frac{Y_r}{n} = \frac{1}{n}\sum_{i=1}^{n} \sin\theta_i \qquad (5.41)$$

Заметим, что эти координаты также определяют центроид конечных точек заданных единичных векторов.

Результирующий вектор дает информацию не только о среднем направлении некоторого множества векторов, но и о расположении векторов относительно среднего. На рис. 5.21,а представлены три вектора, которые только немного отклоняются от среднего направления. Результирующий вектор по длине почти равен сумме длин трех этих векторов. Наоборот, три вектора, представленные на рис. 5.21, б, сильно разбросаны; их результирующий вектор очень короткий. Длина результирующего вектора, обозначаемая R, определена по теореме Пифагора:

$$R = \sqrt{X_r^2 + Y_r^2} = \sqrt{(\sum_{i=1}^{n} \cos\theta_i)^2 + (\sum_{i=1}^{n} \sin\theta_i)^2} \qquad (5.42)$$

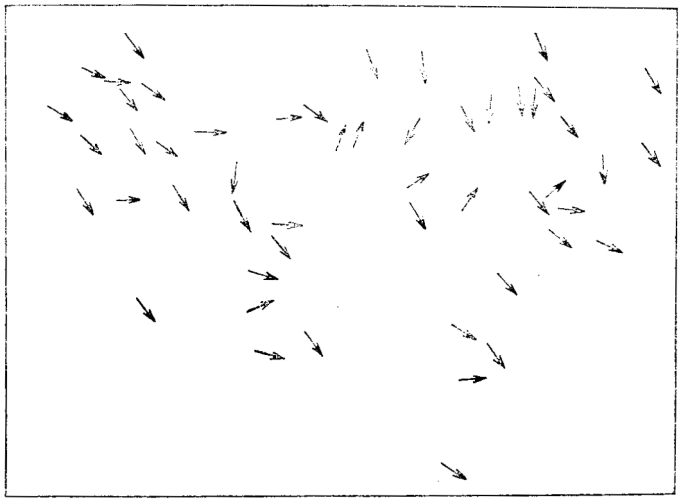


Рис. 5.18. Карта, представляющая направления штриховки ледника в южной части Финляндии (51 измерение) на площади 35 км^2

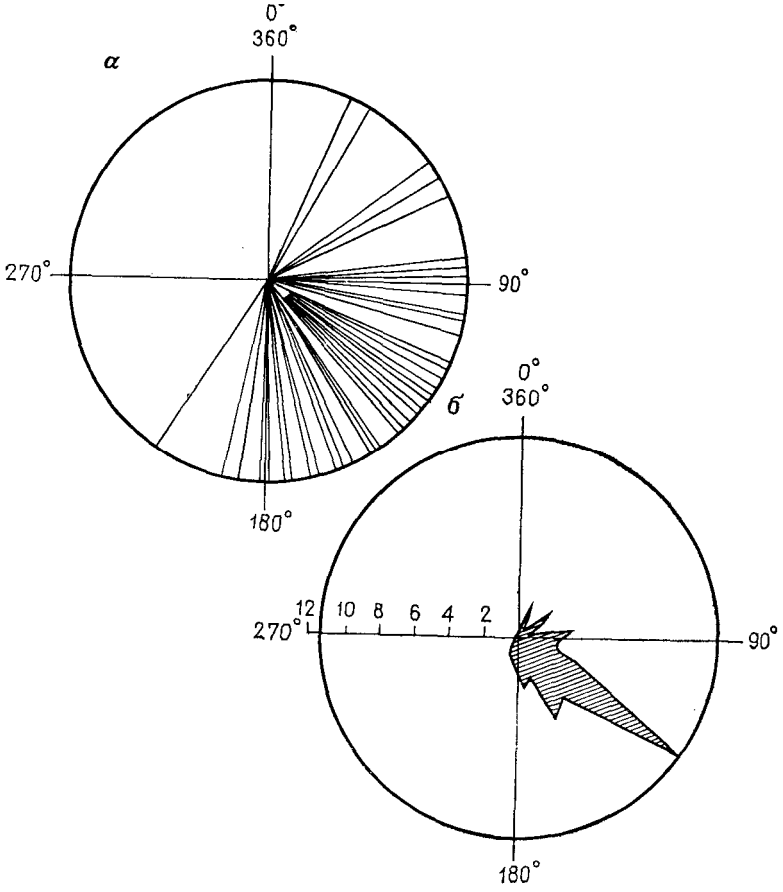


Рис. 5.19. Направления штриховки ледника, изображенные на рис. 5.18: *а* – направления, представленные как единичные векторы; *б* – направления, представленные как роза-диаграмма, на которой указаны числа векторов внутри последовательных 10-градусных интервалов

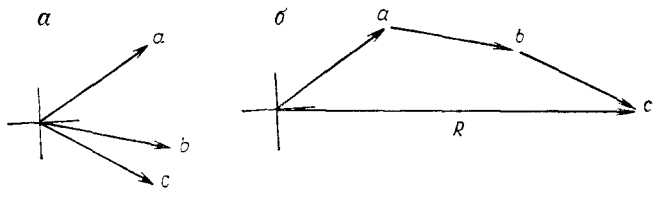


Рис. 5.20. Определение среднего направления множества единичных векторов: *а* – три вектора, взятые с рис. 5.18; *б* – вектор *R*, полученный комбинированием трех единичных векторов. Порядок комбинирования несуществен.

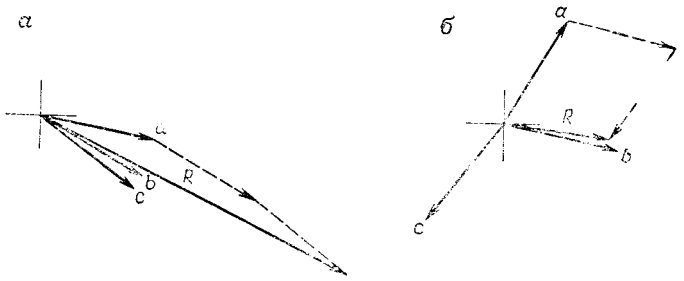


Рис. 5.21. Использование длины вектора R для выражения дисперсии некоторого набора единичных векторов: *а* – три вектора, расположенных в узком пучке вокруг общего направления; вектор *R* относительно длинный, близкий к числу *n*; *б* – три широко рассеянных вектора; *R* имеет длину, меньшую единицы

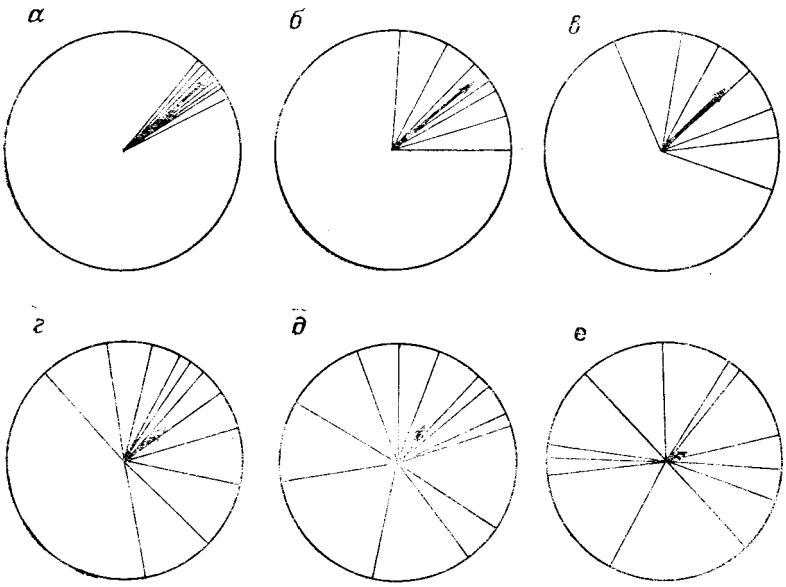


Рис. 5.22. Множества единичных векторов, иллюстрирующие значение, получаемое при различном расположении векторов во всех примерах, среднее направление – 52°: *а* – $\overline{R} = 0{,}997$; *б* – $\overline{R} = 0{,}90$; *в* – $\overline{R} = 0{,}75$; *г* – $\overline{R} = 0{,}55$; *д* – $\overline{R} = 0{,}40$; *е* – $\overline{R} = 0{,}10$

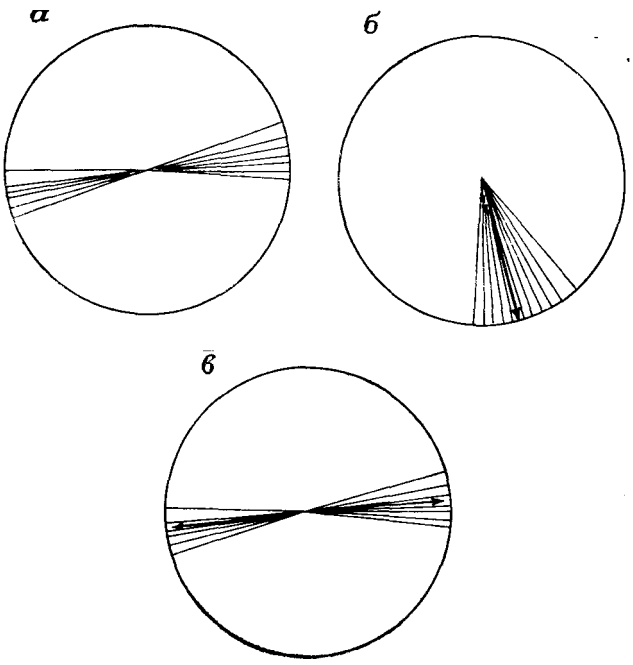


Рис. 5.23. Удвоение угла с целью вычисления средней ориентации: *а* – ориентация измерении, представленных как векторные направления; результант *R* среднего направления равен 285° и близок к нулевому по длине (*R*=0,08); *б* – ориентация измерений, представленных как векторы направлений после удвоения углов. Распределение более не является бимодальным; вектор R отражает правильный тренд удвоенных углов и близок к единице по длине (среднее направление – 120°; $\overline{R} = 0{,}97$); в – ориентации вновь нанесены как исходные углы и истинное направление вектора *R* (81°) находится делением пополам направления *R* в «б»

Длину результирующего вектора можно стандартизировать делением на число наблюдений. Стандартизованный результирующий вектор по длине равен также квадратному корню из суммы квадратов стандартизованных координат концевых точек:

$$\overline{R} = R / n = \sqrt{\overline{C}^2 + \overline{S}^2} \qquad (5.43)$$

Величина $\overline{R}$ называется средней длиной результирующего вектора и изменяется от нуля до единицы. Эта мера аналогична дисперсии, но в некотором смысле противоположна ей. Большие значения $\overline{R}$ указывают на то, что наблюденные векторы находятся в узком пучке с малой дисперсией, а значения $\overline{R}$, близкие к нулю, указывают на большой разброс векторов. На рис. 5.22 представлены множества векторов, имеющих различные значения $\overline{R}$. Для того, чтобы иметь меру дисперсии, которая увеличивается с увеличением рассеяния, R иногда заменяют на его дополнение, называемое циклической дисперсией:

$$s_0^2 = 1 - \overline{R} = (n - R)/n \qquad (5.44)$$

Можно вычислить и другие направленные статистики, включая циклические аналоги стандартного отклонения, моды, медианы. Их определения приведены в удобной таблице Гейлом и Бертом [28].

Ориентация данных может быть изменена до вычисления средних направлений или мер рассеяния. Так как ориентация может иметь одно из двух противоположных значений, то во избежание ошибок в определении дисперсии необходимо принять некоторые соглашения. На примере ориентации речной гальки Крамбейн [47] предлагает новое решение этой задачи. Если все измеренные углы удвоить, будет записан тот же угол, независимо от того, какая ориентация была использована. В качестве примера рассмотрим шарнир складки, которая простирается с северо-востока на юго-запад. Его ориентация будет одинаковой независимо от того, задать ли угол равным 45° или 225°. Если удвоить углы, мы получим $45^o \times 2 = 90^o$ и $225^o \times 2 = 450^o$, что составляет $450^o - 360^o = 90^o$.

Среднее направление, длина среднего результирующего вектора и циклическая дисперсия могут быть найдены обычным образом после того, как ориентированные утлы были удвоены. Для нахождения истинной средней ориентации разделим вычисленный угол среднего направления на два. Это проиллюстрировано на рис. 5.23.

Проверка гипотез о циклически распределенных данных

Для проверки статистических гипотез о циклически распределенных данных мы должны иметь некоторую вероятностную модель, соответствующую изучаемому параметру. Существуют циклические аналоги одномерных распределений, которые мы обсудили в гл. 2, однако наиболее полезно из них распределение фон Мизеса. Это – циклический эквивалент нормального распределения, также обладающий двумя параметрами: средним направлением $\overline{\theta}$ и параметром концентрации k. Распределение фон Мизеса унимодально и симметрично относительно среднего направления. По мере увеличения параметра концентрации вероятность получения направленного измерения, очень близкого к среднему направлению, увеличивается. Если к равно нулю, все направления равновероятны и распределение становится циклическим равномерным. На рис. 5.24, а представлена форма распределения фон Мизеса для некоторых значений k. Это распределение может быть также представлено в условной форме (рис. 5.24, 6).

Прямое определение параметру концентрации затруднительно, но его можно оценить через $\overline{R}$, если допустить, что данные являются выборкой из совокупности, имеющей распределение фон Мизеса. В табл. 5.6 приведены оценки максимального правдоподобия для параметра k, соответствующие некоторому вычисленному $\overline{R}$. В некоторых приводимых ниже статистических критериях мы будем использовать эти оценки параметра k.

Критерии проверки случайности.

Простейшая гипотеза, которую можно проверить статистическими методами, это гипотеза о случайности направленных наблюдений. Это эквивалентно утверждению о том, что нет предпочтительных направлений или же что вероятность любого направления одинакова. Если предположить, что наблюдения представляют выборку из совокупности с распределением фон Мизеса, то соответствующая гипотеза эквивалентна утверждению, что параметр концентрации к равен нулю, так как при k=0 распределение циклически равномерно. Иными словами, нулевая гипотеза и альтернатива таковы: H_0:k=0, H_1:k>0.

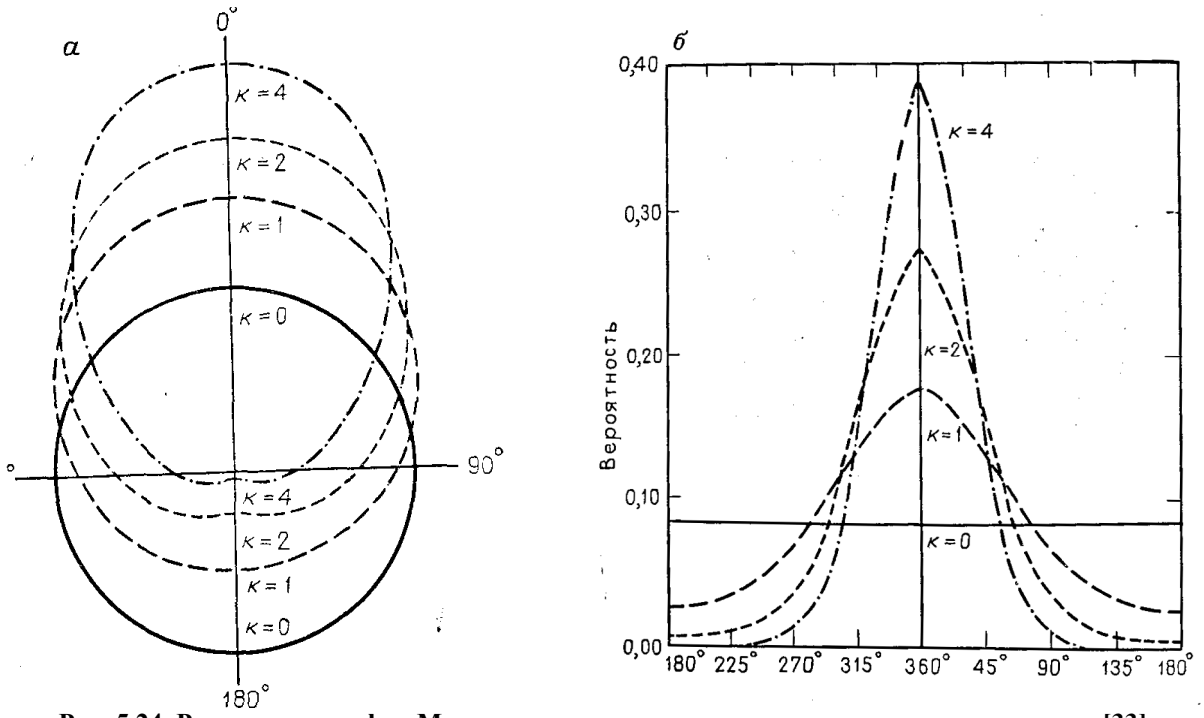


Рис. 5.24. Распределения фон Мизеса, имеющие различные параметры концентрации [33]:
а – распределение, представленное в полярной форме; *б* – распределение, представленное как условное. Интервал измерения соответствует полной окружности

Таблица 5.6. Оценка максимума правдоподобия параметра концентрации k для вычисленных значении R [5, 33]

$\overline{R}$	k	$\overline{R}$	k	$\overline{R}$	k	$\overline{R}$	k
0,00	0,00000	0,25	0,51649	0,50	1,15932	0,75	2,36930
0,01	0,02000	0,26	0,53863	0,51	1,19105	0,76	2,45490
0,02	0,04001	0,27	0,56097	0,52	1,22350	0,77	2,54686
0,03	0,06003	0,28	0,58350	0,53	1,25672	0,78	2,64613
0,04	0,08006	0,29	0,60625	0,54	1,29077	0,79	2,75382
0,05	0,10013	0,30	0,62922	0,55	1,32570	0,80	2,87129
0,06	0,12022	0,31	0,65242	0,56	1,36156	0,81	3,00020
0,07	0,14034	0,32	0,67587	0,57	1,39842	0,82	3,14262
0,08	0,16051	0,33	0,69958	0,58	1,43635	0,83	3,30114
0,09	0,18073	0,34	0,72356	0,59	1,47543	0,84	3,47901
0,10	0,20101	0,35	0,74783	0,60	1,51574	0,85	3,68041
0,11	0,22134	0,36	0,77241	0,61	1,55738	0,86	3,91072
0,12	0,24175	0,37	0,79730	0,62	1,60044	0,87	4,17703
0,13	0,26223	0,38	0,82253	0,63	1,64506	0,88	4,48876
0,14	0,28279	0,39	0,84812	0,64	1,69134	0,89	4,85871
0,15	0,30344	0,40	0,87408	0,65	1,73945	0,90	5,3047
0,16	0,32419	0,41	0,90043	0,66	1,78953	0,91	5,8522
0,17	0,34503	0,42	0,92720	0,67	1,84177	0,92	6,5394
0,18	0,36599	0,43	0,95440	0,68	1,89637	0,93	7,4257
0,19	0,38707	0,44	0,98207	0,69	1,95357	0,94	8,6104
0,20	0,40828	0,45	1,01022	0,70	2,01363	0,95	10,2716
0,21	0,42692	0,46	1,03889	0,71	2,07685	0,96	12,7661
0,22	0,45110	0,47	1,06810	0,72	2,14359	0,97	16,9266
0,23	0,47273	0,48	1,09788	0,73	2,21425	0,98	25,2522
0,24	0,49453	0,49	1,12828	0,74	2,28930	0,99	50,2421

Критерий крайне прост и требует лишь вычисления значения $\overline{R}$ по формуле (5.43). Затем эта статистика сравнивается с критическим значением $\overline{R}$ для заданного уровня значимости. Если наблюдения извлечены из циклического равномерного распределения, то можно ожидать, что $\overline{R}$ будет мало (см. рис. 5.22, *е*). Однако если вычисленная статистика превышает критическое значение, то нулевая гипотеза должна быть отклонена, и можно предположить, что наблюдения получены из совокупности, имеющей предпочтительную ориентацию. Этот критерий был получен лордом Релеем в начале столетия; современное изложение соответствующих вопросов принадлежит Мардиа [51]. В табл. 5.7 приведены критические значения $\overline{R}$ для различных объемов выборки и уровней значимости.

Таблица 5.7. Критическое значение $\overline{R}$ критерия Релея о наличии предпочтительного тренда [51]

Число наблюдений *n*	Уровень значимости α, %			
	10	5	2,5	1
4	0,768	0,847	0,905	0,960
5	0,677	0,754	0,816	0,879
6	0,618	0,690	0,753	0,825
7	0,572	0,642	0,702	0,771
8	0,535	0,602	0,660	0,725
9	0,504	0,569	0,624	0,687
10	0,478	0,540	0,594	0,655
11	0,456	0,516	0,567	0,627
12	0,437	0,494	0,544	0,602
13	0,420	0,475	0,524	0,580
14	0,405	0,458	0,505	0,560
15	0,391	0,443	0,489	0,542
16	0,379	0,429	0,474	0,525
17	0,367	0,417	0,460	0,510
18	0,357	0,405	0,447	0,496
19	0,348	0,394	0,436	0,484
20	0,339	0,385	0,425	0,472
21	0,331	0,375	0,415	0,461
22	0,323	0,367	0,405	0,451
23	0,316	0,359	0,397	0,441
24	0,309	0,351	0,389	0,432
25	0,303	0,344	0,381	0,423
30	0,277	0,315	0,348	0,387
35	0,256	0,292	0,323	0,359
40	0,240	0,273	0,302	0,336
45	0,226	0,267	0,285	0,318
50	0,124	0,244	0,270	0,301

Напомним, что критерий Релея основан на предположении, что наблюденные векторы извлечены из совокупности с распределением фон Мизеса, т.е. совокупность векторов либо имеет равномерное распределение, если k=0, либо имеет единственную моду или предпочтительное направление. Если же в действительности векторы извлечены из бимодального распределения, такого, например, как изображено на рис. 5.23, то этот критерий дает ошибочные результаты.

Проверим, не имеют ли предпочтительного направления штриховки ледника в Финляндии при уровне значимости 5%. Так как имеется 50 наблюдений, то табл. 5.7 дает критическое значение $\overline{R}_{50;0,05} = 0{,}244$. Проверяемая статистика – это просто нормализованное значение $\overline{R}$. Сумма косинусов векторов равна X_2= –25,793 и сумма синусов равна Y_2=31,637. Длина есть

$$R = \sqrt{(-25{,}793)^2 + (31{,}637)^2} = 40{,}819$$

что, деленное на объем выборки, дает среднюю длину

$$\overline{R} = 40{,}819/51 = 0{,}800$$

Так как вычисленное значение $\overline{R}$ значительно превышает критическое, то мы отклоняем нулевую гипотезу о том, что параметр концентрации равен нулю. Штриховки должны иметь предпочтительный тренд.

Критерий проверки специфического тренда.

В некоторых случаях мы можем столкнуться с необходимостью проверить гипотезу о том, что наблюдения имеют некоторый специфический тренд. Например, территория Финляндии, где проводились измерения направлений штриховки ледников, расположена в пределах обширной топографической депрессии, вытянутой с северо-запада на юго-восток примерно на 105°. Совпадает ли среднее направление движения льда, показываемое штриховкой, с осевым направлением этой депрессии?

Точная проверка гипотезы о том, что выборка векторов была извлечена из совокупности, имеющей некоторое заданное направление, требует использования обширных таблиц для определения критических значений. Более простой метод состоит в определении доверительного угла вокруг среднего направления выборки и в проверке того, достаточен ли этот угол по величине для того, чтобы гипотетическое среднее вошло в него. Определение этого угла основано на стандартном отклонении оценки среднего направления 9, и потому для его вычисления используют полный объем выборки и ее дисперсию.

Прежде чем вычислять доверительный угол, применим критерий Релея для подтверждения того, что статистически значимое среднее направление существует. Затем, используя данные табл. 5.6, вычислим среднюю длину вектора $\overline{R}$ и оценим параметр концентрации к. Приближенное значение стандартного отклонения среднего направления в радианах есть $s_e = 1/\sqrt{n\overline{R}k}$.

Так как эта величина является мерой ожидаемого случайного отклонения оценки среднего направления от выборки к выборке, мы может использовать ее для определения вероятностных пределов положения истинного среднего направления совокупности. Предполагая, что ошибки оценок нормально распределены, заключаем, что интервал $\theta \pm Z_\alpha s_e$ включает в себя истинное среднее направление совокупности в α% случаев. Например, если мы собрали наудачу 100 выборок одного объема из совокупности векторов и вычислили средние направления и 95%-ные доверительные интервалы вокруг каждого из них, то можно ожидать, что почти все интервалы, кроме пяти, будут содержать истинное среднее направление. Конечно, мы можем не знать, какие из пяти интервалов не будут охватывать среднее значение, но мы можем дать вероятностную характеристику этого события для каждого из них. Мы можем, например, полагать, что интервал вокруг среднего значения выборки частного вида содержит истинное среднее направление. Вероятность получить противоположный результат равна 5%.

Мы уже применили критерий Релея и отклонили гипотезу о том, что не имеется никакого тренда в наблюдениях штриховки ледника. Приближенное значение стандартного отклонения среднего направления в радианах теперь можно найти по формуле

$$s_e = \frac{1}{\sqrt{51 \cdot 0{,}8004 \cdot 2{,}87129}} = \frac{1}{10{,}826} = 0{,}0924\,(3{,}14^\circ)$$

Поэтому вероятность того, что интервал 129,2°±1,96x3,14° содержит среднее направление совокупности, равна 95%. Другими словами, $126{,}1^o \le \theta \le 132{,}3^o$. Так как этот интервал не включает направление вытягивания в линию топографической депрессии, мы должны заключить, что он не совпадает со средним направлением штриховки.

Критерий соответствия.

Простая непараметрическая альтернатива к критерию Релея проверки равномерности состоит в подразделении единичной окружности на подходящее число угловых сегментов. Если эти сегменты одинаковы по размеру и наблюденные векторы распределены случайно, то следует ожидать, что в каждом сегменте число векторов примерно одинаково. Число действительно наблюденных векторов сравнивается с ожидаемым с помощью критерия χ^2. Ожидаемая частота в каждом сег-

менте должна быть, самое меньшее, равна 5, и они должны быть расположены между $n/15$ и $n/5$ сегментами. Критерий χ^2 вычисляется обычным образом с помощью уравнения (2.45) и имеет k–1 степеней свободы, где k – число сегментов.

Та же процедура может быть использована для проверки соответствия наблюденных векторов другим теоретическим моделям, таким, как распределение фон Мизеса с заданным параметром концентрации k, большим нуля и заданным средним направлением θ. Вычисление ожидаемых частот однако может оказаться непростым. Примеры приводятся в работах [33] и [5].

Проверка равенства двух множеств направленных векторов

Иногда бывает необходимо проверить гипотезы об эквивалентности двух выборок или наборов направленных измерений. Предположим, что мы имеем наборы направленных палеоизмерений в двух стратиграфических единицах. Задача состоит в том, чтобы сравнить их средние направления для определения того, не являются ли они одинаковыми. Мы хотим знать, не совпадают ли ориентации линеаментов, изображений, видимых со спутников, с ориентациями складок, существование которых установлено по фотографиям соответствующих площадок. В несколько меньшем масштабе мы можем сравнить направление удлиненных зерен в тонких сечениях двух образцов керна песчаника из нефтяного месторождения.

Совпадение двух средних направлений может быть установлено с помощью сравнения двух векторов, полученных по двум группам, с вектором, полученным при объединении двух множеств измерений. Если две выборки действительно извлечены из одной и той же совокупности, вектор объединенных выборок должен быть приблизительно равным сумме двух векторов. Если средние направления двух выборок значимо различаются, то объединенный вектор будет короче, чем сумма.

Если k – некоторое большое (>10) значение, можно вычислить значение F-критерия:

$$F_{1,n-2} = \frac{(n-2)(R_1 + R_2 - R_p)}{n - R_1 - R_2} \qquad (5.45)$$

где n – общее количество наблюдений; R_1 и R_2 – векторы по двум выборкам; R_p – вектор объединенной выборки.

Используя данные табл. 5.6, мы можем оценить значение k из $\overline{R}_p$, результирующего длины вектора двух объединенных выборок. Если k меньше 10, но больше 2, то более точное значение F-критерия находят следующим образом:

$$F_{1,n-2} = \left(1 + \frac{3}{8k}\right)\frac{(n-2)(R_1 + R_2 - R_p)}{n - R_1 - R_2} \qquad (5.46)$$

Если k меньше 2, то необходимы специальные таблицы, аналогичные таблицам, приводимым Мардиа [51].

Можно также проверить равенство параметров концентрации двух множеств векторов, однако вычисления довольно сложны. За деталями отсылаем к [51], а также к [56] и [28], где приведен пример из геоморфологии.

Складчатый пояс, топографически выраженный как Нага Хиллс и их отроги, занимает промежуточное положение между Индийским субконтинентом и Индокитайским полуостровом. Очевидно, связанный со сжимающими движениями, которые создали Гималаи, складчатый пояс включает ряд субпараллельных антиклиналей вдоль восточной границы Бангладеш. В этом регионе в структурных ловушках были найдены нефть и газ, поэтому определение направления складок имеет как экономический, так и научный интерес. Есть основания предположить, что складки меньшей величины расположены к западу от Нага Хиллс, но они скрыты современными осадками, приносимыми Гангом и его притоками. К сожалению, сейсмические данные, которые могли бы нам дать информацию о погребенных структурах, очень скудны.

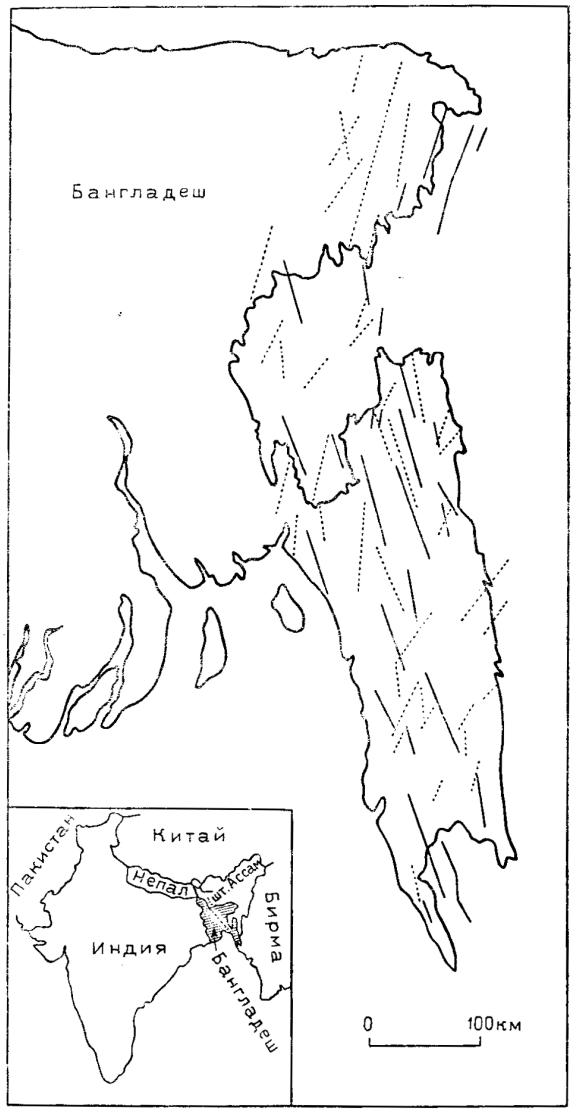


Рис 5.25. Карта Восточного Бангладеш, на которой указаны осевые плоскости больших антиклиналей (жирные линии) и обширные линеаменты, полученные со спутников (пунктир). (Границы государств приведены по оригиналу. – *Прим. перев.*)

Интерпретация снимков со спутников этого региона указывает на многочисленные штрихи неизвестного происхождения. Вполне возможно, что штрихи отражают погребенные складки, и если так, то они дают ценные сведения о структурной геологии и возможных месторождениях нефти.

На рис. 5.25 представлена карта Восточного Бангладеш, на которой показаны следы осевых плоскостей наибольших антиклиналей и больших штрихов, замеченных со спутников. Ориентации этих двух множеств показаны на рис. 5.26. Так как эти линии не направленные, то графики являются бимодальными, и для получения правильного распределения векторов мы должны удвоить наблюденные углы. В табл. 5.8 приведены ориентации обеих осевых плоскостей и линеаментов. Очевидно различие между этими двумя множествами, но является ли это различие статистически значимым или возникает из-за ошибок опробования.

Таблица 5.8. Ориентация осевых плоскостей антиклиналей и линеаментов Ландсата в Восточном Бангладеш (значения даны в градусах, отсчет велся против часовой стрелки, начиная с севера)

Ось антиклинали n=32	Линеаменты Ландсата n=40	Ось антиклинали n=32	Линеаменты Ландсата n=40	Ось антиклинали n=32	Линеаменты Ландсата n=40
103,5	248,98	301,0	233,7	265,6	255,3
288,5	283,8	281,3	293,8	259,7	235,4
282,2	247,8	291,2	246,8	257,9	281,0
265,7	258,7	294,6	266,4	104,8	239,4
256,8	275,6	287,2	279,4		261,9
253,6	238,9	299,4	257,4		257,1
249,8	228,7	300,1	256,4		100,8
250,0	239,5	273,4	232,0		247,0
107,9	277,2	294,4	275,8		110,2
287,9	245,5	291,6	244,8		109,0
291,7	277,9	113,8	254,4		251,4
287,1	236,3	290,9	230,5		230,9
283,5	235,9	290,1	252,0		
290,9	266,1	279,2	257,7		

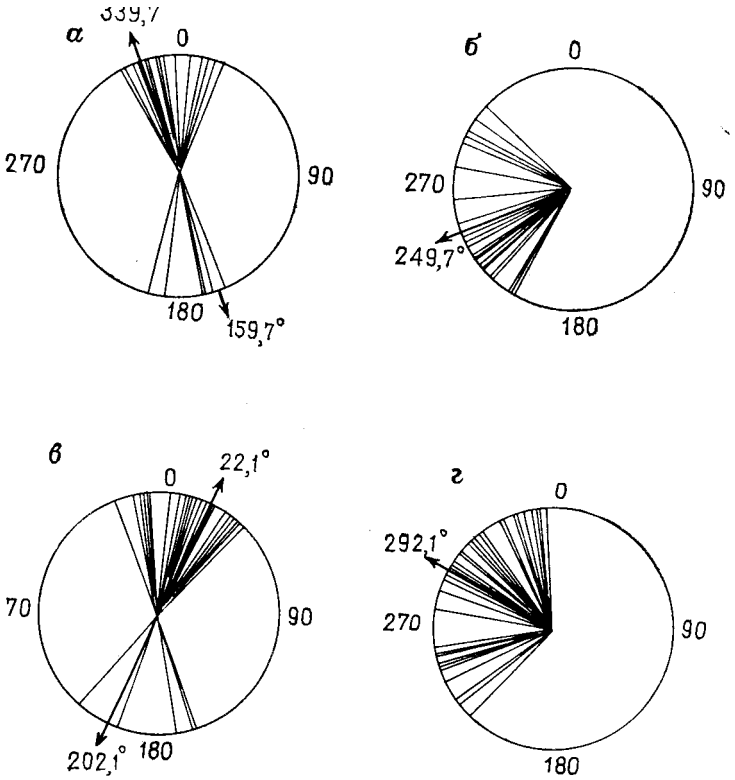


Рис. 5.26. Круговое представление ориентации по данным Восточного Бангладеш. Указана средняя ориентация. *а* – ориентации осевых плоскостей больших антиклиналей; *б* – углы осевых плоскостей удваиваются; *в* – ориентации линеаментов со спутников; *г* – углы простирания удвоены

Для проверки гипотезы о том, что средние направления осей антиклинали и штриховки одинаковы, мы можем сначала построить векторы для каждой из двух групп и затем представить результант их объединения. Так, для 32 удвоенных измерений осевых плоскостей R_1=32,09, а для 40 удвоенных измерений линеаментов Ландсата R_2=27,81. Объединяя обе группы в множество из 72 наблюдений, получаем $\overline{R}_p = 54{,}00$. Среднее значение для объединенной выборки есть $\overline{R}_p = 54/72 = 0{,}7$, и с помощью табл. 5.6 мы можем оценить фактор концентрации.

Так как k больше 2 и меньше 10, то подходящая статистика дается формулой (5.46). Подставляя вычисленные значения, получим

$$F = \left(1 + \frac{3}{8(2{,}3693)}\left(\frac{(72-2)(32{,}09+27{,}81-54{,}00)}{(72-32{,}09-27{,}81}\right)\right) = 39{,}59$$

Этот критерий имеет $\nu_1 = 1$ и $\nu_2 = 72-2$ степеней свободы. Из табл. 2.14 (см. кн. 1) методом интерполяции находим критическое значение F при 5%-ном уровне значимости ($\alpha = 0{,}05$) и с 1 и 70 степенями свободы; это значение равно 3,96. Так как проверяемое значение значительно превышает критическое, мы можем уверенно заключить, что линеаменты Ландсата и оси складок нельзя считать извлеченными из общей совокупности. Хотя линеаменты Ландсата и полезны при геологических исследованиях, в этом регионе они не отражают тренда структурных складок.

СФЕРИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ

Статистические критерии для проверки гипотез о направленных данных в трех измерениях были созданы лишь в последние годы частично потому, что математические проблемы, относящиеся к этой теории, очень сложны. Однако геологические проблемы, содержащие гипотезы о трехмерных векторах, являются общими, и мы не уклоняемся от использования подходящей статистической техники для их интерпретации. Некоторые из этих методов требуют использования аппарата матричной алгебры для матриц невысокого порядка, а также нахождения собственных векторов и собственных значений. Геометрическая интерпретация собственных векторов, данная в гл. 3 (см. кн. 1),

имеет здесь непосредственное приложение. Соответствующие математические методы тесно связаны с многомерными процедурами, описанными в главе 6. Здесь мы имеем дело с тремя физическими измерениями; позже мы применим те же операции для анализа многомерных данных, в котором каждое «измерение» – это самостоятельная геологическая переменная.

Примеры трехмерных направленных данных в науках о Земле: измерения простирания и падения пласта, используемые для структурного анализа; векторные измерения геомагнитного поля; направленная проницаемость, измерения в пробах, взятых из пород нефтяных месторождений; измерения ориентации осей в петротектонических исследованиях.

Как и в двумерном случае, введем сначала обозначения. Мы можем считать трехмерные направленные наблюдения векторами. Поскольку нас будут интересовать главным образом углы между ними, мы можем предполагать, что они имеют единичную длину. Если все направленные измерения с некоторой площади имеют общую исходную точку, то концы соответствующих векторов располагаются на единичной сфере; отсюда термин «сферическое распределение».

Некоторые ориентированные изображения нельзя трактовать как направления, их удобнее считать осями. В качестве примеров можно назвать линии пересечения множеств плоскостей падения пластов, оси вращения и перпендикуляры к плоскостям. В добавление к этому, иногда удобно игнорировать направления векторов, предпочтительнее считать их осями.

Для описания вектора в трехмерном пространстве принято использовать тройку декартовых координат (рис. 5.27,*а*). Направление вектора OP характеризуется косинусами углов между вектором и каждой из координатных осей. Координаты точки P равны

$$X=\cos a, \quad Y=\cos b, \quad Z=\cos c.$$

Так как рассматриваемый вектор имеет единичную длину, то

$$X^2+Y^2+Z^2=1$$

Используя сферические углы, мы можем определить направление вектора OP углом Φ между осью X и проекцией заданного вектора на плоскость XY и углом θ между данным вектором и осью Z (см. рис. 5.27,*б*). Действительно, θ определяет широту вектора, в то время как Φ определяет его долготу. Соотношение между этими сферическими полярными координатами и декартовыми координатами имеет вид

$$X = \sin\theta\cos\Phi; \qquad Y = \sin\theta\sin\Phi, \qquad Z = \cos\theta$$

Результаты измерения геологических свойств часто даются в терминах простирания и падения, а не в терминах косинусов углов между направлениями или через декартовы координаты. Кроме того, координатные обозначения, используемые геологами, отличаются от обычно используемых математиками. Если считать положительное направление оси X соответствующим направлению на север, положительное направление оси Y соответствующим направлению на восток и положительное направление оси Z соответствующим направлению вертикально вниз, то мы получим декартову систему координат, в которой падение выражается положительными углами [51].

Эти обозначения проиллюстрированы на рис. 5.28 для вектора OP, определенного простиранием и падением заключавшей его плоскости. Линия ON есть азимут, или проекция OP на горизонтальную плоскость XY; она перпендикулярна к линии простирания. Угол A есть угол простирания, измеренный в направлении против часовой стрелки от 0° к северу. D – падение, измеренное как положительный угол от ON вниз. Координаты X, Y и Z точки P

$$X= -\sin A \sin D, \qquad Y= \cos A \sin D, \qquad Z= \sin D. \qquad (5.47)$$

Формулы становятся более сложными, если простирание измеряется в квадранте, тогда требуется более точно задать направление падения. См. пояснения в статье Ватсона [78].

Если сферические измерения представлены в виде координат X, Y и Z конца вектора, то очень просто вычислить среднее направление и сферическую дисперсию. Это делается аналогично вычислению циклического среднего и дисперсии. Среднее направление дается в виде единичного вектора R. Его длина есть

$$R = \sqrt{\left(\sum X_i\right)^2 + \left(\sum Y_i\right)^2 + \left(\sum Z_i\right)^2} \qquad (5.48)$$

В нормализованной форме $\overline{R} = R/n$.

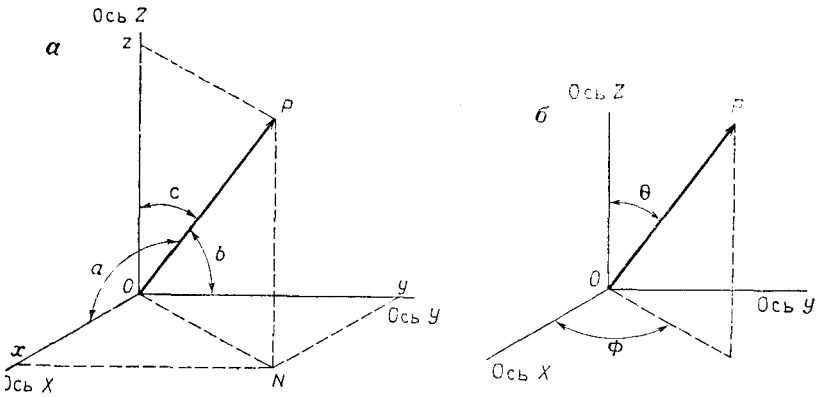


Рис. 5.27. Система обозначений векторов в трехмерном пространстве: *а* – *OP*-вектор в пространстве, определенный декартовыми координатами X, Y и Z. Углы между *OP* и осями равны *a*, *b* и *c*; *б* – вектор в пространстве, определенный сферическими углами Ф и θ

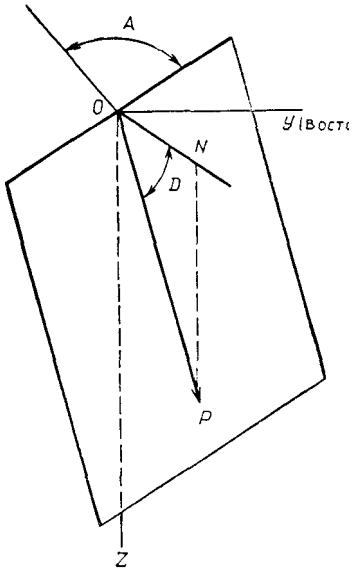


Рис. 5.28. Обозначения для трехмерного вектора, характеризующего простирание и падение. Угол *A*, измеренный против часовой стрелки от направления на север, есть угол простирания поверхности, содержащей вектор *OP*. Плоскость *ONP* перпендикулярна к поверхности падения. Угол *D* характеризует падение

Направление *R* по отношению к трем координатным осям дается косинусами углов между *R* и этими осями:

$$\cos\overline{X} = \sum X_i / R; \qquad \cos\overline{Y} = \sum Y_i / R; \qquad \cos\overline{Z} = \sum Z_i / R \tag{5.49}$$

Если наблюдения плотно расположены вокруг общего направления, то *R* будет большим числом, стремящимся к *n*. Если наблюдения рассеяны, *R* будет мало. Как и в случае циклического распределения, *R* можно использовать как меру концентрации, и она может быть представлена как сферическая дисперсия

$$s_2^2 = (n - R)/n = (1 - \overline{R}) \tag{5.50}$$

Эти методы определения среднего направления и сферической дисперсии пригодны тогда, когда векторы не очень сильно разбросаны. При некоторых условиях, однако, среднее направление может оказаться ложным. Предположим, что были измерены падения полого падающих слоев; одни из них полого падают на запад, другие – на несколько градусов на восток. Так как падение считается вектором, конец которого лежит на нижней полусфере, то вектор *R* восточного и западного падений будет направлен вертикально вниз! Конечно, длина *R* будет близка к нулю, так что сферическая дисперсия будет большой, указывая на крайне высокую дисперсию среди векторов.

Если эти падения рассматривать не как векторы, а как ненаправленные оси, то два конца будут проектироваться в верхнюю и нижнюю полусферы; очевидно, что линии, представляющие восточное и западное падения, тесно связаны. Средняя ось, вычисленная с помощью методов теории собственных векторов, изложенной ниже, будет горизонтальной и будет проходить через пучок осей падения.

Матричное представление векторов

В гл. 3 (см. кн. 1) уже отмечалось, что строки матрицы графически могут быть представлены векторами. Обратно угловые характеристики векторов можно представить в матричной форме. Собственные значения и собственные векторы такой матрицы дают информацию о размещении векторов в пространстве. Однако прежде чем охарактеризовать представление множества векторов в матричной форме, приведем обзор положений геометрии, начиная с двумерного случая.

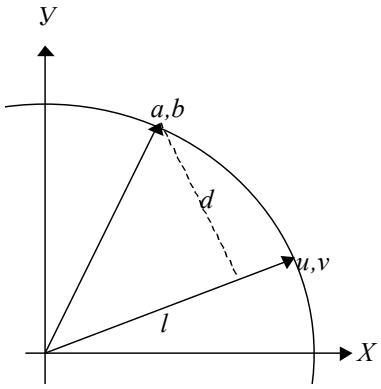


Рис. 5.29. Проекция вектора *a, b* на вектор *u, v*. Длина проекции *a, b* на *u, v* равна *l*. Расстояние вектора концевой течки *a, b* от вектора *u, v* равно *d*

Геометрическое соотношение, дающее величину проекции одного вектора на направление другого, – это скалярное произведение двух векторов (рис. 5.29). Если предположить, что оба вектора единичные, то декартовы координаты их концевых точек будут такие же, как их направляющие косинусы по отношению к осям *X* и *Y*. Проекция равна

$$l=au+bv, \qquad (5.51)$$

где *l* – длина проекции вектора *a, b* на вектор *u, v*. Это также есть длина проекций *u, v* на *a, b*.

На рис. 5.29 видно, что вектор *a, b* есть гипотенуза прямоугольного треугольника, стороны которого есть проекция *l* на вектор *u, v* и перпендикуляр *d*. По теореме Пифагора получаем соотношение

$$d^2=1-l^2=1-(au+bv)^2 \quad (5.52)$$

Любое число векторов можно спроектировать на линию *u, v* с помощью уравнения (5.51), и квадраты их расстояний от линии *u, v* определяются уравнением (5.52). Сумма квадратов расстояний будет

$$M=\sum_{i=1i}^{n} d_i^2 = n-\sum_{i=1}^{n}\left(a_i u+b_i v\right)^2 \qquad (5.53)$$

ее можно рассматривать как момент инерции конечных точек векторов относительно линии *u, v*. Это уравнение можно обобщить на трехмерный случай с помощью введения третьей пространственной координаты

$$M=\sum_{i=1i}^{n} d_i^2 = n-\sum_{i=1}^{n}\left(a_i u+b_i v+c_i w\right)^2 \qquad (5.54)$$

Уравнение (5.54) можно выразить в матричной форме. Сначала координаты линии задаются вектор-столбцом *U*:

$$[U]=\begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

Мы также определим матрицу B:

$$[B]=n[I]-[T],$$

где *T* – матрица размера 3x3 сумм квадратов и попарных произведений направляющих косинусов векторов

$$[T]=\begin{bmatrix} \sum a_i^2 & \sum a_i b_i & \sum a_i c_i \\ \sum b_i a_i & \sum b_i^2 & \sum b_i c_i \\ \sum c_i a_i & \sum c_i b_i & \sum c_i^2 \end{bmatrix}$$

Поэтому матрица B имеет вид

$$[B]=\begin{bmatrix} n-\sum a_i^2 & \sum a_i b_i & \sum a_i c_i \\ \sum b_i a_i & n-\sum b_i^2 & \sum b_i c_i \\ \sum c_i a_i & \sum c_i b_i & n-\sum c_i^2 \end{bmatrix}$$

Момент инерции векторов относительно направления [*u*] есть просто $M=[U]'[B][U]$.

Прежде чем определять момент относительно некоторой произвольной линии [*U*], мы можем найти единственную линию, относительно которой момент инерции будет максимальным. Координаты этой линии задаются первым собственным вектором матрицы [*B*]. Если λ – первое собственное значение [*B*] и [b_1] – соответствующий ему собственный вектор-столбец, то, как отмеча-

лось в гл. 3, $\lambda_ш = [b'_1]\cdot[B]\cdot[b_1]$, т.е. λ есть момент инерции векторов относительно первого собственного вектора. Это значит, что сумма квадратов расстояний от концов данных векторов до первого собственного вектора максимально возможная, или что собственный вектор одновременно приблизительно перпендикулярен ко всем из данных векторов, насколько это возможно.

Момент инерции второго собственного вектора – наибольший из возможных для любой линии, ортогональной первому собственному вектору. Третий собственный вектор должен быть ортогональным двум другим и должен также учитываться для всех остальных квадратов расстояний до вершин векторов. Так как эти три собственных вектора определяют ортогональный базис, полностью эквивалентный исходному множеству декартовых осей, то третий собственный вектор должен определять линию, вдоль которой момент инерции минимален. То есть, он будет ориентирован так, как будто он одновременно близок настолько, насколько это возможно, ко всем этим векторам.

Если два вектора диаметрально противоположны (рис. 5.30), они оба будут иметь одинаковые длины перпендикуляров вне собственного вектора $[b]$ и одинаковое влияние на расположение собственного вектора. Это значит, что направление векторов теряет смысл; они неотличимы от осей. По этой причине методы теории собственных векторов предпочтительнее при исследовании данных, распределенных на сфере, в тех случаях, где неоднозначность соответствует различию между векторами с концами в верхней и нижней полусферах.

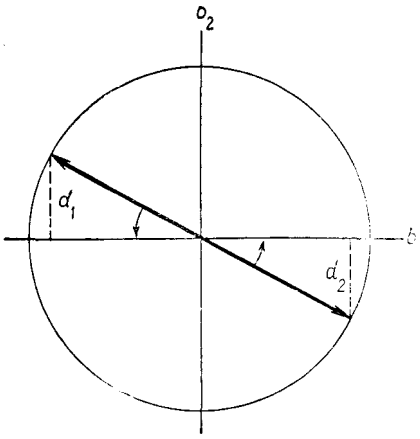


Рис. 5.30. Проекция двух диаметрально противоположных векторов на собственный вектор b_1. Расстояния d_1 и d_2 идентичны и имеют одинаковый вращательный момент

Собственные векторы обеспечивают прямую информацию о распределении этих векторов. Мардиа [51] различает четыре случая:

1. λ_1 велико, в то время как λ_2 и λ_3 малы. Это значит, что сумма квадратов перпендикуляров между концевыми точками векторов и осью, соответствующей первому собственному вектору, очень велика. Большинство наблюдений должно лежать в плоскости, содержащей собственные векторы с номерами 2 и 3, и образовывать опоясывающее распределение (рис. 5.31, *а*).
2. λ_1 и λ_2 оба велики, в то время как λ_3 мало. Расстояние по перпендикуляру от концевых точек до первого и второго собственных векторов должно быть очень большим, а расстояние до третьего собственного вектора должно быть малым. Наблюдения собираются в пучок вокруг конца третьего собственного вектора (см. рис. 5.31, *б, в*). Как бимодальное, так и унимодальное распределения дают одинаковый результат; они могут различаться значением $\overline{R}$, которое для унимодального случая будет большим.
3. Два собственных значения совпадают. Это на самом деле некоторый частный случай случая 1. Наблюдения образуют симметрический пояс вокруг оси, соответствующей единственному собственному значению (см. рис. 5.31, *г*).
4. Все три собственных значения одинаковы. Распределение равномерное, так как перпендикулярные направления для трех точек одинаковы для всех трех ортогональных осей. На единичной сфере нет предпочтительного размещения точек (см. рис. 5.31, *д*).

Вудкок [81] обобщил эту классификацию, представив графически логарифмы отношений собственных значений $(\ln\lambda_1)/\lambda_2$ в зависимости от $(\ln\lambda_2)/\lambda_3$. На его диаграмме все возможные схемы точек на сфере попадают в специфические области. Эта форма графического анализа может быть особенно полезна при работе с петротектоническими данными. На рис. 5.32 представлена одна из диаграмм Вудкока.

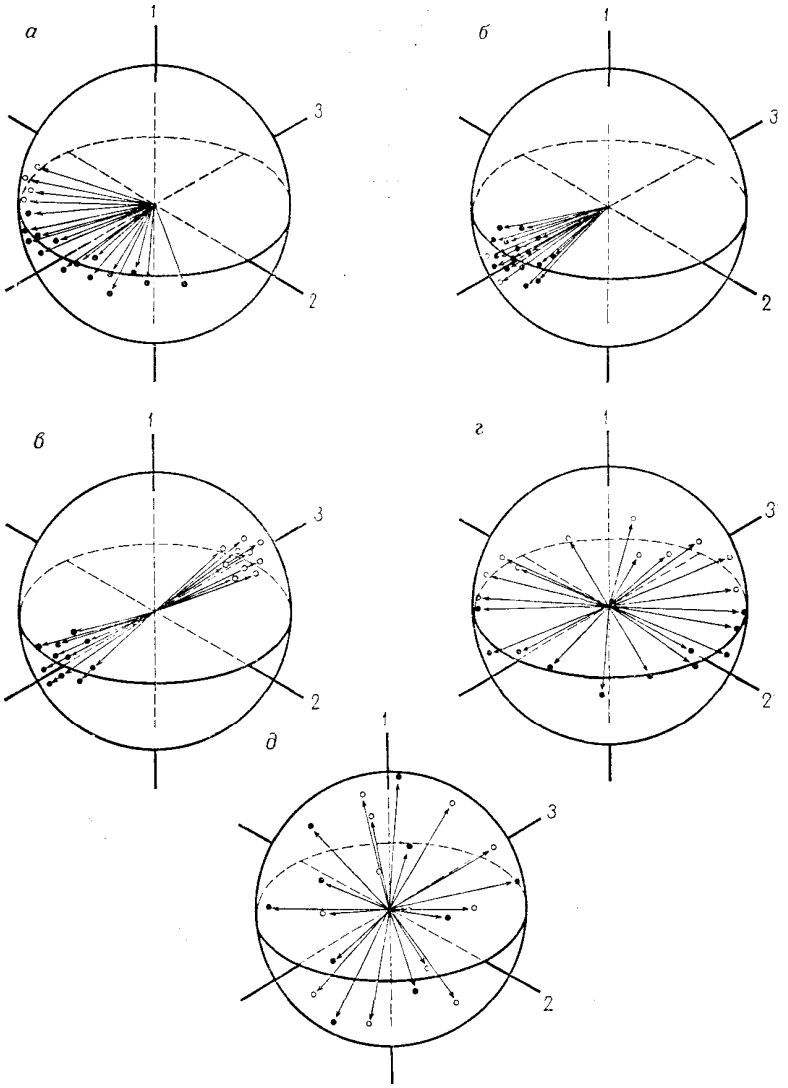


Рис. 5.31. Схемы векторов на единичной сфере: *а* – схема частичного опоясывания в плоскости, содержащей собственные векторы 2 и 3; *б* – унимодальное распределение векторов относительно собственного вектора 3; *в* – бимодальное распределение векторов относительно собственного вектора 3; *г* – схема полного опоясывания в плоскости, содержащей собственные векторы 2 и 3; их собственные значения идентичны или близки к этому; *д* – равномерное распределение. Все собственные значения приблизительно равны

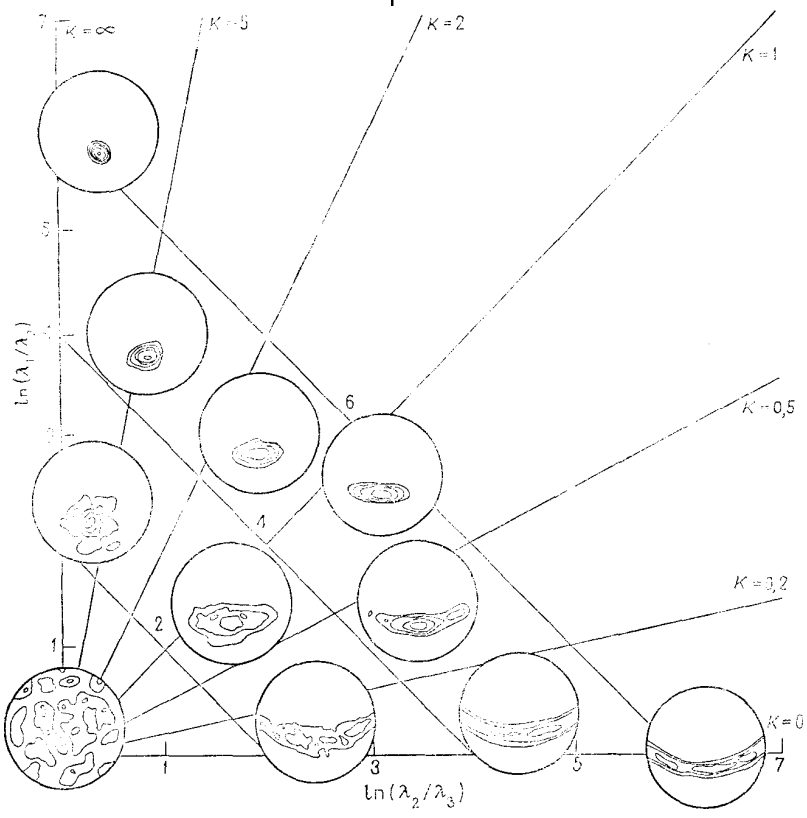


Рис. 5.32. Классификация схем векторов на единичной сфере, соответствующая логарифмам отношений их собственных значении [81]. Для различных отношений указаны типичные петротектонические диаграммы. K – отношение $\ln(\lambda_1/\lambda_2)/\ln(\lambda_2/\lambda_3)$

Представление сферических данных

Хотя изображение в перспективе на единичной сфере полезно для целей иллюстрации, оно не может передать детальную информацию о распределении векторов. По принятому соглашению трехмерные векторы показаны в проекции их концов на плоскость. Так как эти точки действительно лежат на поверхности сферы, представление их в двух измерениях требует использования уравнения проекции. Геологи традиционно используют равноплощадную полярную проекцию Ламбера, которая относится к сети Шмидта. Кристаллографы предпочитают полярную стереографическую проекцию, сохраняющую углы, или сеть Вульфа.

На рис. 5.33 представлены набор векторов на единичной сфере и их проекция на равноплощадную диаграмму. Необходимо отличать векторы, которые имеют концы на нижней полусфере, от векторов с концами на верхней полусфере. Так как геологи часто описывают векторы в терминах их «погружения», то это соответствует их изображению на нижней полусфере единичной сферы.

В добавление к векторам иногда необходимо нанести трехмерные ориентации, такие, как складки и поверхности разломов. Если плоскость проходит через центр единичной сферы, то ее пересечение со сферой образует большой круг (рис. 5.34, *а*). Однако легче представить плоскость осью, называемой полярной, которая перпендикулярна к плоскости в начале координат. Геологи изображают пересечение полярной оси с нижней полусферой, хотя кажется более логичным изображать ее пересечение с верхней полусферой. Тогда проекция полярной оси на плоскость, имеющей наклон к западу, например, будет изображена на левой стороне западной части диаграммы (см. рис. 5.34,*б*).

Иногда на диаграмме бывают представлены очень большие множества трехмерных данных, так что общую схему расположения точек нельзя охватить единым взглядом. В таких случаях локальные скопления точек можно оконтурить, пересчитав число точек, лежащих внутри некоторой малой площади диаграммы. Это может быть сделано лишь при использовании равноплощадной проекции. Проекция покрывается регулярной схемой узлов сети и подсчитывается число точек внутри окружности фиксированного радиуса. Обычно радиус выбирается так, чтобы площадь описываемого круга заключала 10% общей площади. Так как расстояния между узлами сети меньше радиуса, последовательные площади перекрываются, и плотности точек постепенно изменяются от одной части диаграммы к другой. Малые диаграммы, представленные на рис. 5.32, – это типичные контурные схемы, встречающиеся в петротектонических исследованиях.

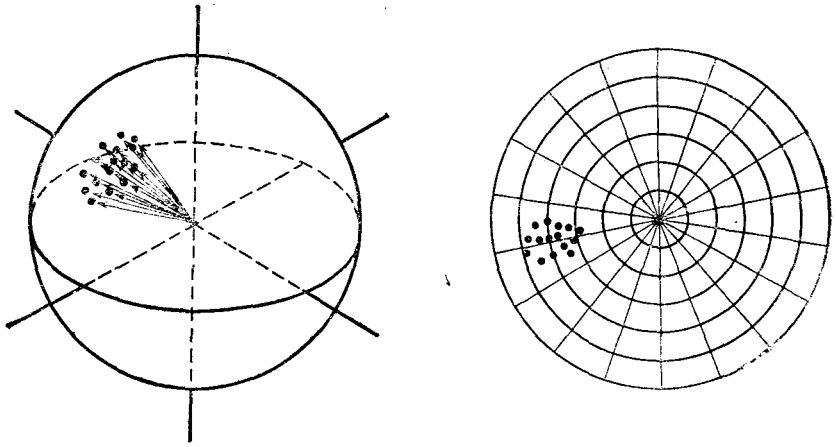


Рис. 5.33. Векторы внутри единичной сферы и их проекции на равноплощадную полярную диаграмму

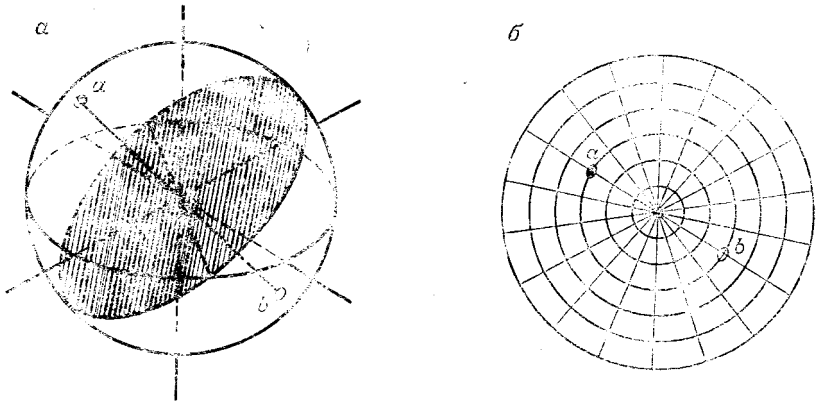


Рис. 5.34. Проектирование полюсов на плоскости: *а* – плоскость и ее полюсы на единичной сфере; *б* – полюсы плоскости, спроектированные на равноплощадную диаграмму. Точка *a* – проекция на верхнюю полусферу, точка *b* – проекция на нижнюю полусферу

Проверка гипотез о сферически направленных данных

Простейшие критерии, относящиеся к ориентации в трехмерном пространстве, – это обобщения критериев, используемых в теории циклических данных. Как и в этом случае, мы нуждаемся в вероятностной модели известной характеристики, гипотезы о значениях которой мы проверяем. Широко используемая модель – распределение Фишера – это обобщение распределения фон Мизеса и сферический эквивалент нормальной кривой. Распределение Фишера характеризуется двумя параметрами: вектором среднего направления 9 и дисперсией k. Так как мы имеем дело с тремя измерениями, то вектор среднего направления имеет три направляющих косинуса по отношению к трем координатным осям.

Вектор среднего оценивается через направляющие косинусы (уравнение 5.49). Дисперсия может быть аппроксимирована выражением

$$k=(n-2)/(n-R), \qquad (5.55)$$

которое достаточно точно, если k велико, например больше 10. Мардна [51] дает таблицу более точных оценок, в зависимости от величины нормализованного $\overline{R}$.

Критерий случайности.

Как и в циклическом случае, можно проверить гипотезу о том, что данные распределены равномерно во всех направлениях. Это эквивалентно утверждению о том, что параметр концентрации равен нулю.

Нулевая гипотеза и альтернатива таковы: $H_0: k=0$, $H_1: k>0$. Проверяемая статистика вычисляется так же, как и для циклических данных; она равна нормализованному $\overline{R}$. Эта статистика затем сравнивается с критическим значением для выбранного уровня значимости (табл. 5.9). Если вычисленное значение $\overline{R}$ превышает табличное, то гипотеза о том, что наблюдения извлечены из равномерно распределенной совокупности, отклоняется при заданном уровне значимости.

Также можно проверить гипотезу о заданной ориентации среднего вектора и построить конус доверия относительно этого вектора. Эти критерии, однако, требуют обширных таблиц, аналогичных опубликованным Стефенсом [73] и Мардиа [51]. Там же приводятся двухвыборочные критерии эквивалентности средних направлений двух множеств наблюдений и необходимые таблицы.

Таблица 5.9. Критические значения $\overline{R}$ для проверки равномерности сферического распределения [51]

Число наблюдений n	Уровень значимости α, %			
	10	5	2	1
5	0,637	0,700	0,765	0,805
6	0,583	0,642	0,707	0,747
7	0,541	0,597	0,659	0,698
8	0,506	0,560	0,619	0,658
9	0,478	0,529	0,586	0,624
10	0,454	0,503	0,558	0,594
11	0,433	0,408	0,533	0,568
12	0,415	0,460	0,512	0,546
13	0,398	0,442	0,492	0,526
14	0,384	0,427	0,475	0,507
15	0,371	0,413	0,460	0,491
16	0,359	0,400	0,446	0,476
17	0,349	0,388	0,443	0,463
18	0,339	0,377	0,421	0,450
19	0,330	0,367	0,410	0,438
20	0,322	0,358	0,399	0,428
21	0,314	0,350	0,390	0,418
22	0,307	0,342	0,382	0,408
23	0,300	0,334	0,374	0,400
24	0,294	0,328	0,366	0,392
25	0,288	0,321	0,359	0,384
30	0,26	0,29	0,33	0,36
35	0,24	0,27	0,31	0,33
40	0,23	0,26	0,29	0,31
45	0,22	0,24	0,27	0,29
50	0,20	0,23	0,26	0,28
100	0,14	0,16	0,18	0,19

ФОРМА

Форма – очень нелегкое для измерения свойство, которому даже нельзя дать какое-либо точное определение. Возможно, по этой причине имеется так много предложенных мер формы, что ни одну из которых нельзя считать вполне удовлетворительной. Мера формы должна обладать некоторыми обязательными свойствами. Очевидно, объекты различной формы должны иметь различные меры, а подобные формы должны давать подобные значения, независимо от размера и ориентации объектов. К сожалению, меры формы, обладающие этими свойствами, оказываются нереальными. Математическими методами показано, что никакая конкретно выбранная мера не может быть единственной только для одной формы [49].

Ученые, изучающие Землю, пытались охарактеризовать широкий спектр форм, начиная с относительно простых, например, проекции контуров песчаных зерен, до более сложных форм, соответствующих ископаемым остаткам организмов. Геоморфологи особенно преуспели в образовании мер формы, применив их к изучению дренажных бассейнов, друмлинов, коралловых атоллов и некоторых других форм ландшафта. Были исследованы формы нефтеносных полей по отношениям их осей, а также по форме некоторых типов структурных ловушек. Обширную литературу по этим вопросам приводят Меллеринг и Рейнер [61] и Кларк [14]; в обеих работах обсуждаются также теоретические аспекты измерения формы.

В табл. 5.10 приведены меры форм, характеризуемых одним значением, взятым из геологической и географической литературы. Этот перечень не является исчерпывающим. Большинство из

Таблица 5.10. Меры форм, используемые в геологической и географической литературе; перечислены только безразмерные меры [26], [60]

1. Меры, основанные на отношении осей	
Форма	$F = 1/w$
Удлиненность	$E = w/l$
Округлость	$C_1 = \sqrt{lw/l^2}$
2. Меры, основанные на периметрах	
Индекс формы зерна	$GSI = p/l$
Фактор формы	$SF_1 = \frac{p_c}{p};\quad SF_2 = \frac{p}{p_c}\times 100$
3. Меры, включающие как периметры, так и площади	
Округлость	$C_2 = 4A/p^2;\quad C_3 = 4A/lp$
Компактность	$K_1 = \frac{2\sqrt{\pi A}}{p};\quad K_2 = \frac{p^2}{4\pi A}$
Тонкость	$TR = 4\pi\left(\frac{A}{p^2}\right)$
4. Меры, основанные на площадях	
Цикличность	$C_4 = \sqrt{A/A_c}$
Фактор формы	$SF_3 = \frac{A_i}{A};\quad SF_4 = \frac{A_c - A_i}{A};\quad SF_5 = \frac{A}{A_c}\times 100$
5. Меры, основанные на площадях и длинах ареалов	
Отношение формы	$FR = A/l^2$
Индекс эллиптичности	$EI = \frac{\pi(1/2l)l}{A}$
6. Другие меры	
Цикличность	$C_5 = \sqrt{D_i/D_c}$
Радиальная дисперсия	$s_R^2 = \frac{\sum\left(R_j - \overline{R}\right)^2}{n}$
Среднее сечение	$\overline{S} = \frac{\sum S_j}{n}$
Дисперсия сечения	$s_s^2 = \frac{\sum\left(S_j - \overline{S}\right)^2}{n}$
Средний радиус	$\overline{R} = \frac{\sum R_j}{n}$

A – площадь объекта; A_c – площадь наименьшего объемлющего круга; A_i – площадь наибольшего вписанного круга; D_c – диаметр наименьшего объемлющего круга; D_i – диаметр наибольшего вписанного круга; l – длина длинной оси; p – периметр объекта; p_c – периметр окружности круга, имеющего ту же площадь, что и объект; R_j – j-й радиус объекта, измеренный от центроида до края; S_j – длина j-й стороны объекта, рассмотренной как многоугольник; n – число сторон объекта, рассматриваемого как многоугольник; w – ширина объекта по перпендикуляру к длинной оси.

них вычислены по таким основным измерениям, как длины осей, периметры и площади. Некоторые меры форм содержат сравнения со стандартными формами, например, окружностью.

Любая из этих мер формы может быть использована таким же образом, как и какой-либо другой дескриптор. Хотя и нет гарантий того, что измерения подчиняются нормальному распределению, по мерам формы для собранных объектов могут быть вычислены обобщающие статистики, такие, как средние и дисперсии.

Измерения формы с помощью преобразований Фурье

Относительная польза методов исследования совокупностей измерений, характеризующих форму, иногда отстаивается с некоторой горячностью и, возможно, это плодотворная область для одной или двух кандидатских диссертаций. Однако мы сейчас обратим наше внимание на более перспективные способы описания формы, не являющиеся однозначными. Среди них находятся различные модификации преобразований Фурье, уже использовавшихся в гл. 4 (см. кн. 1) для анализа временных рядов.

Координаты замкнутой линии, например проекция контура зерна песка или ископаемой раковины, можно выразить в полярных координатах, как это изображено на рис. 5.35. Одна из двух координат – это угол радиуса-вектора точки контура; другая – расстояние вдоль этого радиуса до точки от центра.

В связи с выбором центра сразу возникает вопрос, как его выбрать внутри контура объекта. Если центр сдвигается, то расстояния вдоль всех радиусов изменяются и, конечно, соответствующие преобразования Фурье будут различными. Если мы пожелаем сравнить несколько различных форм, мы должны будем отождествить эквивалентные точки внутри каждой из них, считая их центрами соответствующих координатных систем. Если этого не сделать, то мы не сможем сказать, являются ли видимые различия между спектрами Фурье следствием различий формы или же они проистекают из-за нашего выбора центра.

В некоторых приложениях имеется единственная точка внутри каждой формы, которая может служить началом полярной системы координат. Такие примеры приводят Кеслер и Уотерс [42]. Они измеряют радиусы от рубца отпечатка мускула ракушковых рачков до контура их раковины. К сожалению, большинство форм (зерна песка, галька, контуры соляных копей) не имеют ярко выраженных точек, которые можно было бы считать центрами отсчета. Мы можем, однако, внутри каждой замкнутой формы произвольно выбрать точку, которая будет служить началом системы координат.

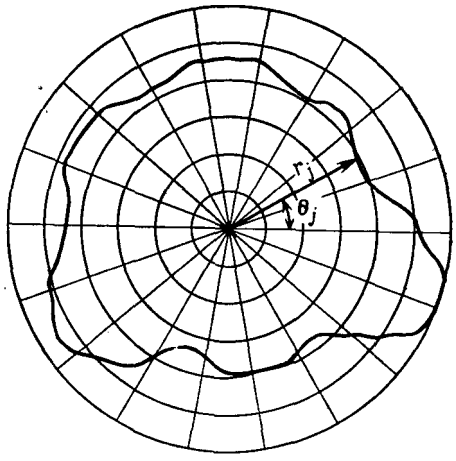


Рис. 5.35. Проекция контура зерна песка, выраженная в полярных координатах. Координатные пары образованы длиной r_j и углом θ_i радиуса, проведенного из центроида к краю

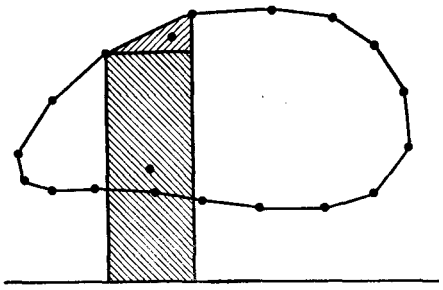


Рис. 5.36. Определение центроида трапециевидной аппроксимации. Оцифрованная форма подразделяется на трапеции, состоящие из прямоугольника и прямоугольного треугольника. Точками представлены их центроиды

Центроид – это центр тяжести формы, он единственен для каждого объекта. Если контур некоторого объекта представляется набором декартовых координат, аналогичным порожденному компьютером, то центроид может быть найден интегрированием координат X и Y. Простая процедура интегрирования построена на использовании метода трапеций.

Ряды точек размещаются на границе объекта (рис. 5.36). Для определения трапеций используются пары этих точек в комбинации с осями. Выбрав произвольную начальную точку, построим серию трапеций, расположенных на рисунке против часовой стрелки. Можно найти центр тяжести каждой трапеции, и комбинация этих центров тяжести даст центр тяжести рассматриваемой фигуры. Координаты центроида фигуры есть

$$\overline{X} = \sum \overline{X}_t A_t \Big/ \sum A_t \quad (5.56)$$

$$\overline{Y} = \sum Y_i A_i \Big/ \sum A_i \quad (5.57)$$

где $\overline{X}_i$ – координаты X центроида 1-й трапеции, $\overline{Y}_i$ – ее координаты Y, A_i – площадь трапеции.

Центроиды площадей трапеций в свою очередь находятся с помощью простых геометрических процедур. Каждая трапеция разбивается на прямоугольник и прямоугольный треугольник. Центроид прямоугольника расположен посередине, в точке пересечения параллелей сторонам, проходящих через их середины. Центроид треугольной части расположен на расстоянии 1/3 пути между прямым углом и двумя острыми углами. Центроиды двух частей комбинируются с весами, соответствующими их относительным площадям. Эти операции могут быть скомбинированы и упрощены до выражений следующего вида:

$$A_t = \frac{1}{2}(Y_{i+1} + Y_i)(X_i - X_{i+1})$$

$$\overline{X}_t A_t = \frac{1}{6}\left(X_{i+1}^2 + X_{i+1}X_i + X_i^2\right)\left(Y_{i+1} - Y_i\right)$$

$$\overline{Y}_t A_t = \frac{1}{6}\left(Y_{i+1}^2 + Y_{i+1}Y_i + Y_i^2\right)\left(X_{i+1} - X_i\right)$$

Эта процедура может быть легко проделана ЭВМ, которая даст положение центроида сложной фигуры. Точность положения зависит от числа точек, размещенных по периметру.

Теперь мы установили положение центроида фигуры и можем провести радиусы из центроида к точкам периметра фигуры. Длины этих радиусов вместе с их угловой ориентацией дают пары полярных координат, которые можно анализировать методами Фурье, рассмотренными в предыдущей главе. Анализ даст спектр замкнутой фигуры, и из этого спектра мы можем вывести много интересных свойств формы фигуры. Циклический спектр Фурье имеет все желаемые свойства, как и обычный спектр Фурье: он содержит всю информацию, содержащуюся в исходной фигуре, последовательные гармоники не зависят друг от друга и каждое спектральное значение есть мера вклада соответствующей гармонической формы в общую дисперсию.

Для того, чтобы сделать циклический анализ Фурье приемлемым, радиусы должны выбираться с равными угловыми приращениями. Как в обычном анализе Фурье, пробы равномерно размещаются во времени или пространстве. К сожалению, невероятно, чтобы координаты исходных точек на периметре фигуры, используемые для определения центроида, размещались так, чтобы соответствующие им радиусы образовывали равные углы. Мы должны найти вдоль периметра новое множество точек, которые определяют равные углы относительно данного центроида либо осуществить новый ряд измерений или интерполяцию между этими точками.

Однако каждая процедура вводит новые усложнения, так как центроид, вычисленный по новому ряду точек, не обязательно в точности совпадает с центроидом, определенным для исходного множества точек. Если радиусы измеряются не от истинного центра, то в спектр Фурье вводится смещение. Эффект от этого аналогичен нецентрированному колесу, которое дает восьмерку при каждом вращении. Эта восьмерка дает вклад в первую гармонику, которая иначе равнялась бы нулю. Так как обычно для сравнения спектр стандартизуется, то наличие ложной первой гармоники уменьшает относительные величины остальных гармоник. В статье [9] описана итерационная процедура, которая дает на каждом шаге все более близкую аппроксимацию к истинному центроиду ряда точек, расположенных вдоль периметра объекта так, что соответствующие им радиусы образуют равные углы.

Преобразования прямоугольной системы координат в полярную дают «развертку» замкнутого контура (рис. 5.37). Методы обычного анализа Фурье очевидно применимы для неразвернутой фигуры.

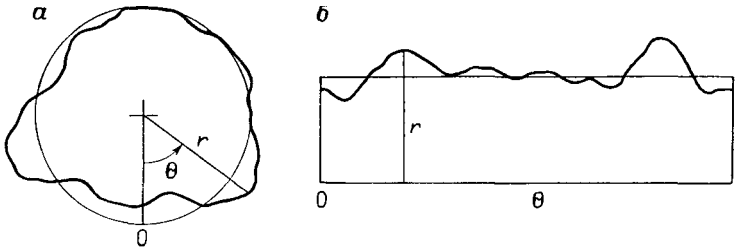


Рис. 5.37. Эквивалентность между полярными и декартовыми представлениями координат: *а* – контур зерна, представленный в полярных координатах; *б* – полярные координаты нанесены на график как функция r от аргумента θ

Напомним, что в гл. 4 разложение в ряд Фурье записывалось в виде

$$Y = \sum_{k=0}^{\infty} \alpha_k \cos(k^{\theta}) + \beta_k \sin(k^{\theta}) \qquad (5.58)$$

причем коэффициенты находились по формулам

$$\alpha_k = \frac{2}{k} \sum_{j=1}^{n} Y_j \cos\left(\frac{2\pi jk}{n}\right) \qquad (5.59)$$

$$\beta_k = \frac{2}{k} \sum_{j=1}^{n} Y_j \sin\left(\frac{2\pi jk}{n}\right) \qquad (5.60)$$

Эти уравнения можно модифицировать так, чтобы они позволяли получать разложение в ряд Фурье для замкнутых фигур. Если расположить концы N радиусов вдоль всей окружности на одинаковом удалении друг от друга, то угол между каждой парой радиусов будет равен $2\pi / n$. Поэтому угол от исходного радиуса до j-го равен $2\pi j / n$. Обозначая угловое направление j-го радиуса через θ_j, а его длину через r_j и подставляя их в уравнения (5.59) и (5.60), получаем

$$\alpha_k = \frac{2}{k} \sum_{j=1}^{n} r_j \cos k_j^{\theta} \qquad (5.61)$$

$$\beta_k = \frac{2}{k} \sum_{j=1}^{n} r_j \sin k_j^{\theta} \qquad (5.62)$$

Как и в обычном анализе Фурье, эти выражения можно упростить: β_0 равно нулю, а α_0 равно среднему радиусу:

$$\bar{r} = \sum_{j=1}^{n} \frac{r_j}{n} \qquad (5.63)$$

Эффект размера может быть исключен из анализа делением всех радиусов на средний радиус, тогда ко будет всегда равно единице. По причинам, указанным ранее, коэффициенты первой гармоники, α_1 и β_1, должны равняться нулю в случае, если радиусы измеряются от центроида.

Как только для ряда гармоник определены коэффициенты α и β, интерпретация циклического спектра Фурье проводится аналогично интерпретации обычного спектра Фурье. Амплитуды фазовых углов полярных гармоник можно выразить через коэффициенты Фурье:

$$A_k = \sqrt{\alpha_k^2 + \beta_k^2} \qquad (5.64)$$

$$\Phi_k = arctg\left(\beta_k / \alpha_k\right) \qquad (5.65)$$

Эти соотношения в точности такие же, как соотношения, данные в предыдущих разделах для спектра дискретного временного ряда.

Обычно для построения энергетического или дисперсионного спектра приходится комбинировать коэффициенты α и β. Это позволяет прямо оценить вклад каждой гармоники в форму фигуры. Последовательные приближения анализируемой формы можно построить с помощью подстановки коэффициентов α и β в уравнение Фурье. По мере увеличения числа гармоник исследуемая форма воспроизводится со все большей детальностью, пока не будет получен исходный оцифрованный образ. Это случается при частоте Найквиста, когда $k = n - 1$. На рис. 5.38 представлена рекон-

струкция песчинки. Нулевые гармоники дают окружность, диаметр которой равен среднему диаметру исходного контура. Здесь первая гармоника отсутствует, так как песчинка соответствующим образом центрирована относительно своего центроида. Вторая гармоника превращает прежнюю реконструированную форму в эллипс, третья гармоника добавляет треугольную компоненту, четвертая – квадратичную компоненту и т. д. (рис. 5.39).

Пропорция вариаций формы, вызываемых последовательными гармониками, дается энергетическим спектром. Обычно малое число низших гармоник учитывается для почти всех вариаций простых форм, таких, как проекции контуров зерен песка (рис. 5.40). Высшие гармоники отражают все более мелкие детали контура и потому используются седиментологами как меры поверхностной округлости. Говорят, что высшие гармоники контуров зерен кварца отражают происхождение или историю переноса (многие работы на эту тему цитируются Эрлихом, Брауном и Ярусом [26], однако такая интерпретация частей спектра требует предосторожностей. Стандартные отклонения оценок коэффициентов Фурье очень велики и имеют тот же порядок, что и сами оценки. Высшие гармоники контуров зерен имеют очень низкую степень влияния, возможно, на пять или более порядков ниже степени влияния низших гармоник. Одного взгляда на стандартные отклонения достаточно, чтобы понять, что им нельзя придавать большое значение. Добавим к этому, что ложные эффекты наиболее ярко выражены при высоких частотах, что затрудняет определение истинных спектральных значений.

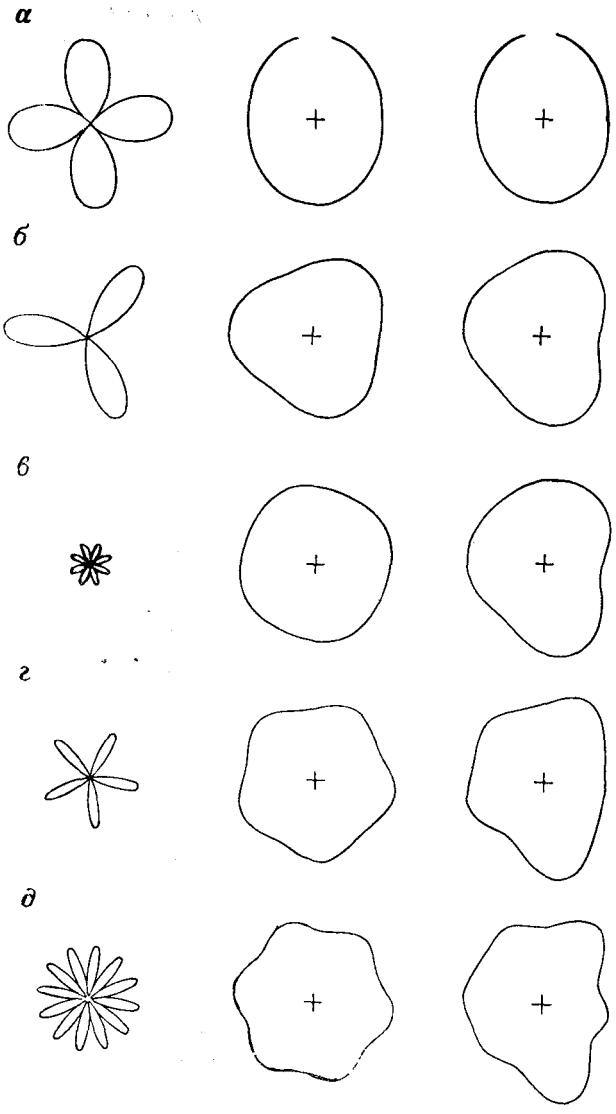


Рис. 5.38. Восстановление формы песчинки с помощью полярных рядов Фурье: *а* – вторая гармоника (слева), вторая гармоника плюс окружность, соответствующая среднему радиусу, или нулевая гармоника (центр), кумулятивная сумма гармоник (справа); *б* – *д* – соответственно третья, четвертая, пятая и шестая гармоники

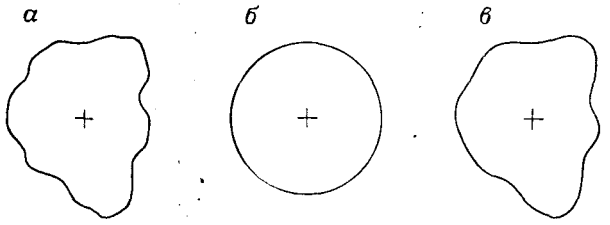


Рис. 5.39. Сравнение оцифрованного контура песчинки (*а*) с нулевой никой или окружностью среднего радиуса (*б*) и с контуром, построенным не шести гармоникам (*в*)

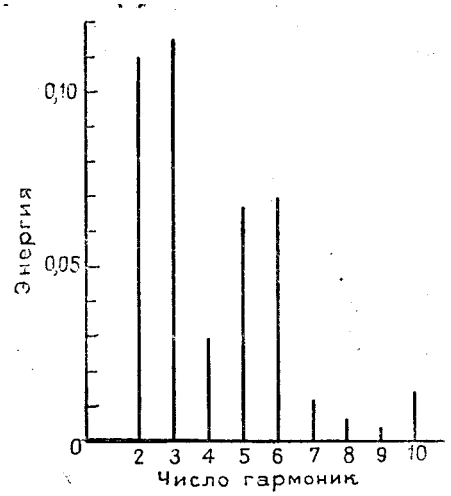


Рис. 5.40. Спектр оцифрованной песчинки, представленной на рис. 5.39

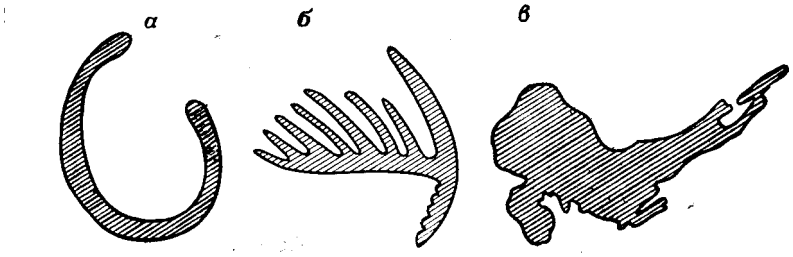


Рис. 5.41. Примеры форм, которые нельзя анализировать с помощью полярного метода Фурье из-за наличия двузначных радиусов: *а* – береговая линия атолла в центральной части Тихого океана; *б* – проекция конодонта *Ligonodina*; *в* – контур гранитного плутона в южной части Онтарио

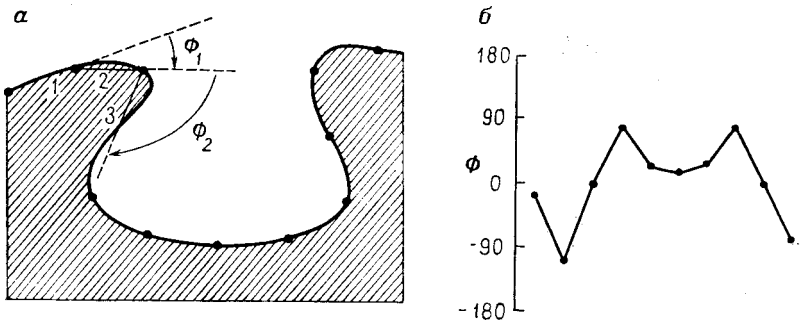
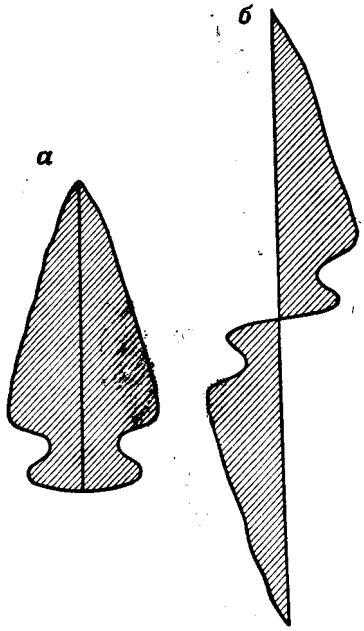


Рис. 5.42. Преобразование цифровых координат в угловые отклонения: *а* – прямые линии представляют собой аппроксимацию, образованную оцифрованной исходной формой в равномерно расположенных точках. Изменения угла между отрезками 1 и 2 (2 и 3) задаются значениями $\Phi_1(\Phi_2)$; *б* – изменение угла между последовательными отрезками. Это – регистрация результатов анализа методом Фурье

Полярное преобразование Фурье применимо при некоторых ограничениях, наиболее сильным из которых является однозначность анализируемого контура. Это значит, что радиус, проведенный из центроида, должен пересекать периметр лишь однажды. Следовательно, этот метод не может быть использован для анализа очень сложных форм (рис. 5.41).

Другие процедуры Фурье позволяют проводить анализ более сложных форм, которые могут быть двузначными в полярных координатах. Один из методов состоит в превращении исходного контура в ряд угловых отклонений. Сначала равномерно распределенные точки размещаются вокруг периметра объекта. Находится направление от первой ко второй точке и затем угловая разность между этим направлением и направлением от второй к третьей точке (рис. 5.42). Повторяя этот процесс, заменяем исходный ряд координат *X, Y*, определяющий заданную форму, рядом угловых отклонений последовательных точек. Этот новый ряд можно анализировать с помощью обычного ме-

тода Фурье, хотя спектр может оказаться трудно интерпретируемым, так как здесь переменная Y не расстояние, а приращение угла. Это и аналогичные преобразования обсуждаются в [14].



Некоторые замкнутые формы обладают двусторонней симметрией или различительной линией, или складкой, которая позволяет ориентировать их вдоль оси. Такую форму можно расщепить или сложить пополам с обращением знаков одной координаты вдоль оси симметрии для одной половины формы (рис. 5.43). Тогда объект выглядит как синусоидальная форма, которая может быть исследована методами обычного анализа Фурье. Эти методы полезны при изучении форм некоторых ископаемых беспозвоночных и также для характеристики заостренных форм.

Рис. 5.43. Контур наконечника кремниевой стрелы позднего палеолита: *а* – форма, представленная в обычном виде; *б* – преобразование, состоящее во вращении половины фигуры вдоль оси симметрия

ПОСТРОЕНИЕ КАРТ В ИЗОЛИНИЯХ С ПОМОЩЬЮ ЭВМ

Цель построения карт в изолиниях – представление портрета формы на поверхности. Такое построение представляет собой нечто вроде трехмерного графа или диаграммы, изображенных на плоскости, двумерном куске бумаги. X_1 – ось, направленная вдоль страницы, X – ось, направленная вверх или вниз по странице. Оси соответствуют географическим направлениям с востока на запад и с севера на юг. Вертикальное направление Y может также представлять возвышения над уровнем моря, мощность или некоторые другие характеристики, изображаемые изолиниями или линиями равных значений. (Многие авторы обозначают эти три оси через X, Y и Z в соответствии с двумя географическими координатами и изображаемой переменной. Чтобы сохранить связь с последующим изложением, мы будем продолжать использовать Y как переменную, которая является функцией других переменных X_i.) Изолинии, строго говоря, есть линии, характеризующие возвышения, либо кровлю, либо подошву изучаемой формации. Геологи казуистически относятся к терминологии, однако обычно называют изолинию контуром, независимо от того, изображает ли она абсолютные отметки, мощность, пористость, состав или какое-либо другое свойство.

Изолинии на карте связывают точки равных значений, и пространство между двумя последовательными изолиниями содержит только точки, значения в которых попадают внутрь интервала, определенного этими изолиниями. Во многих случаях значение на поверхности нельзя определить в каждой из возможных точек и их также нельзя измерить в какой-либо конкретной точке, которую мы могли бы выбрать. Обычно проводится серия контрольных измерений в довольно произвольном небольшом множестве точек, таких, как положение скважин, точки сейсмических исследований или положение точек опробования.

Проведение изолиний может осуществляться вручную, как это делалось прежде, или же с помощью ЭВМ. Построение карты с помощью ЭВМ обычно содержит промежуточный этап – построение математической поверхности. Он должен предшествовать построению изолиний.

Соответствующая программа для ЭВМ позволяет вычерчивать изолинии на основе математических строгих соотношений, вытекающих из геометрии контрольных точек. При этом геолог обычно использует не только контрольные точки, но и свою концепцию и представления о том, как эта поверхность должна выглядеть. Если эти предварительные догадки окажутся на самом деле корректными, то компетентный геолог окажется в состоянии создать карту, превосходящую по качест-

ву карту, вычерченную ЭВМ. С другой стороны, если догадки геолога окажутся ошибочными, то окончательная карта будет содержать серьезные ошибки.

Была проведена серия экспериментов, состоящих в проверке карт, вычерчиваемых ЭВМ [20]. Ряд опытных нефтяников-геологов выступили против широко используемых программ автоматического построения карт в изолиниях. Проверяемые данные состояли из структурных абсолютных отметок в наборе скважин, пробуренных вокруг Девонского рифа в Альберте. Информация из небольшого числа действительно доступных скважин была представлена участникам эксперимента, цель которого состояла в оценке сравнительных способностей человека и машины в создании отвечающей действительности структурной карты в изолиниях.

Все построенные карты были очень похожи вблизи контрольных точек и в корне отличались на неконтролируемых площадях. Некоторые геологи построили структурные карты, качество которых было выше, чем качество карт, построенных по программе для ЭВМ, в том смысле, что их представления оказались более близкими к структурным конфигурациям, соответствующим полному множеству данных. Однако другие геологи допустили серьезные ошибки. Самым интересным оказалось то, что карта, построенная ЭВМ, почти точно совпадала с усредненными картами, построенными вручную, т. е. часть геологов проводила линии между контрольными точками в одном направлении, в то время как другая часть проводила их в противоположном направлении. Большинство провели свои .линии через общую среднюю точку, и только некоторые допустили отклонение в ту или другую сторону. Контуры, проведенные ЭВМ, почти в точности проходили через среднюю точку этого пучка линий. В этом примере программа для ЭВМ вела себя как «усредненный геолог».

Однако наиболее сильный аргумент в пользу построения карт в изолиниях с помощью ЭВМ состоит в том, что при этом создается математическая модель картируемой поверхности, которая может быть использована также и для других целей. Например, могут быть выполнены такие операции, как картирование отклонений от поверхности, вычисление объемов ниже поверхности, различные операции между поверхностями (построение изопахит – вычитание одной поверхности из другой).

Необходимо отметить, что большинство процедур построения карт в изолиниях с помощью ЭВМ являются неожиданными; в их основе лежит слабая теоретическая проработка связей различных используемых в них методологий. Скорее они основаны на допущениях здравого смысла и свойствах поверхности, вытекающих из практического опыта.

Некоторые допущения должны быть заложены в любой вычислительный алгоритм, используемый для создания математической модели, для которой строится карта в изолиниях. Законченная карта отражает эти допущения; модель оправдана и карта является соответствующим действительности представлением поверхности только при условии, если эти допущения обоснованы. В общем случае программа предназначена для картирования поверхности, которая:

a) однозначна во всякой точке;
b) непрерывна всюду в пределах картируемой площади;
c) автокоррелирована на некотором расстоянии, большем, чем типичное расстояние между контрольными точками.

Если имеется только одно возможное значение, которое картируемое свойство может иметь в некоторой заданной географической точке, то это свойство называется однозначным.

В качестве примера можно привести измеренную в скважине абсолютную отметку стратиграфического горизонта, которая используется при построении структурных карт в изолиниях. Предполагается, что нет других причин неопределенности измерений кроме тех, которые возникают из-за несовершенства инструментов. Только при очень редких геологических ситуациях (например, перевернутая складка) в стратиграфической единице можно предположить наличие более одной абсолютной отметки в одной точке.

Совсем не очевидна однозначность некоторых геологических переменных. Например, измерения пористости или химического состава – статистические по своей природе величины, а повторное опробование и анализ в одной и той же точке могут привести к некоторой последовательности значений. Это обусловлено как ошибками измерений, так и случайными изменениями в небольших образцах анализируемых пород. Большинство автоматических программ построения карт в изолиниях не приспособлено к таким повторяющимся данным, хотя множественные наблюдения могут быть приведены к единственному представительному значению, такому, как среднее или математическое ожидание, которые затем можно нанести на карту.

Автоматические процедуры построения карт в изолиниях содержат интерполяцию между

контрольными точками и экстраполяцию вне их пределов. Как следует из используемых математических методов, все значения, полученные в результате интерполяции, лежат на непрерывной наклонной поверхности между контрольными точками. Если вещественная поверхность содержит разрывные нарушения, то они не будут выявлены программой, а будут закартированы просто как области очень крутого наклона. Для представления разломов или других дизъюнктивных нарушений, о которых имеется предварительная информация, выбираются специальные процедуры, в результате применения которых на карте появляются границы. ЭВМ по программе картирования вычертит поверхности на противоположных сторонах от границ так, как если бы они были совершенно отдельными картами. Соответствующая математическая модель будет иметь разрыв в численных значениях вдоль границ. Однако невозможно создать программу построения карты в изолиниях, которая автоматически распознает разрывные нарушения.

Программа построения карт в изолиниях основана на допущении, что значения поверхности в одной точке тесно связаны со значениями в соседних точках и менее тесно связаны со значениями в более отдаленных точках. Допущение о том, что картируемая переменная положительно автокоррелирована по меньшей мере на малых расстояниях, в алгоритме соответствует выбору вблизи оцениваемой точки всех близлежащих контрольных точек, после чего оценка поверхности в точке осуществляется с помощью некоторого способа усреднения. Если поверхность высоко автокоррелирована, то все эти соседние контрольные точки будут иметь примерно одно и то же значение, и их среднее будет обоснованной оценкой в промежуточных точках. Наоборот, если поверхность слабо автокоррелирована, то соседние контрольные точки будут мало связанными друг с другом; они также не будут иметь связи со значением в оцениваемой точке. При таких условиях сделать обоснованное предположение о природе этой поверхности между контрольными точками оказывается невозможным.

Триангуляция как метод построения карт в изолиниях

Первая программа построения карт в изолиниях с помощью ЭВМ была простым применением методов, использовавшихся для построения карт топографами вручную [39].

Принимая, что положения контрольных точек произвольны и не подчиняются никакому правилу, их соединяют сначала прямыми линиями. Таким образом получается сеть треугольников, покрывающих карту (рис. 5.44). При интерполяции ниже сторон треугольников находят точки, в которых абсолютная отметка имеет заданную постоянную величину. Соединяя точки равных абсолютных отметок, мы получаем некоторый контур. Действительно, поверхность можно представить как ряд плоских треугольных пластинок, каждая из которых имеет углы в контрольных точках. Построение карты сводится к проведению горизонтальных линий через эти наклонные пластинки. Почти каждый студент-геолог или инженер имел дело с ручным эквивалентом этого процесса при выполнении упражнений по картированию или съемке.

Очевидно, что если контрольные точки связаны различным образом, то будут определены различные множества треугольных пластинок и различные множества контурных линий. В некоторых ранних программах построения карт в изолиниях простой ввод данных точек в различной последовательности приводил к заметному на глаз различию изолиний на карте. Чтобы избежать этого, были сделаны попытки выбрать единственное оптимальное множество треугольников для картирования. Обычно это означает, что треугольники выбираются настолько близкими к равносторонним, насколько это возможно, или что треугольники должны иметь наименьшую возможную высоту, или что наибольшая сторона треугольника должна быть настолько короткой, насколько это возможно [30]. К сожалению, не известно ни одного алгоритма, который бы обеспечивал построение треугольной сети таким образом, чтобы достигалась оптимальная конфигурация. Это часто приводит к чрезвычайно большому времени выполнения алгоритма, результатом чего оказался полный отказ от метода треугольников. Его место заняли процедуры, которые использовали контрольные точки только для получения оценок поверхности в узлах регулярной сети и затем для проведения изолиний по ее узлам, а не по самим контрольным точкам.

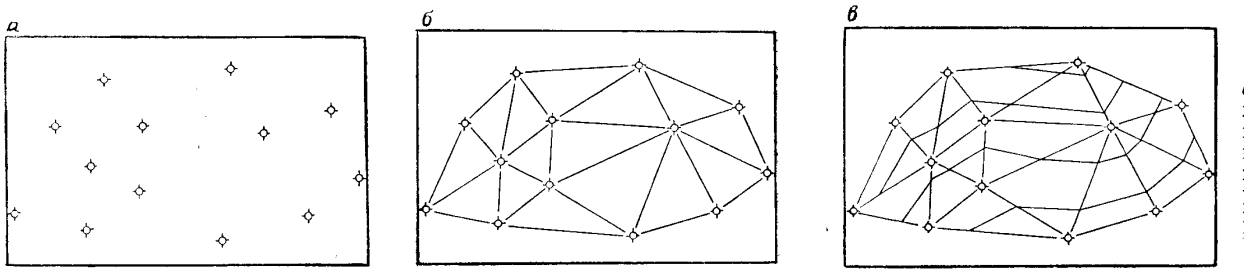


Рис. 5.44. Оценка положения изолинии методом триангуляции [39]: *а* – данные, неравномерно расположенные в пространстве; *б* – треугольники картируемой площади с вершинами в заданных точках; *в* – изолинии, проведенные через стороны треугольников, причем конкретные значения отметок в заданных точках были найдены методом линейной интерполяции

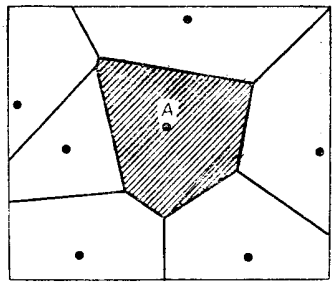


Рис. 5.43. Многоугольник Тиссена (заштрихован) вокруг точки *A*.

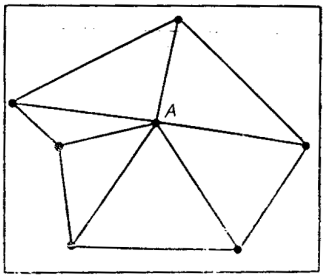


Рис. 5.46. Сеть треугольников Делоне вокруг точки *A*. Все точки внутри многоугольника ближе к точке *A*, чем любая другая точка. Все точки, соединенные с точкой *A*, являются ее Тиссеновскими соседями

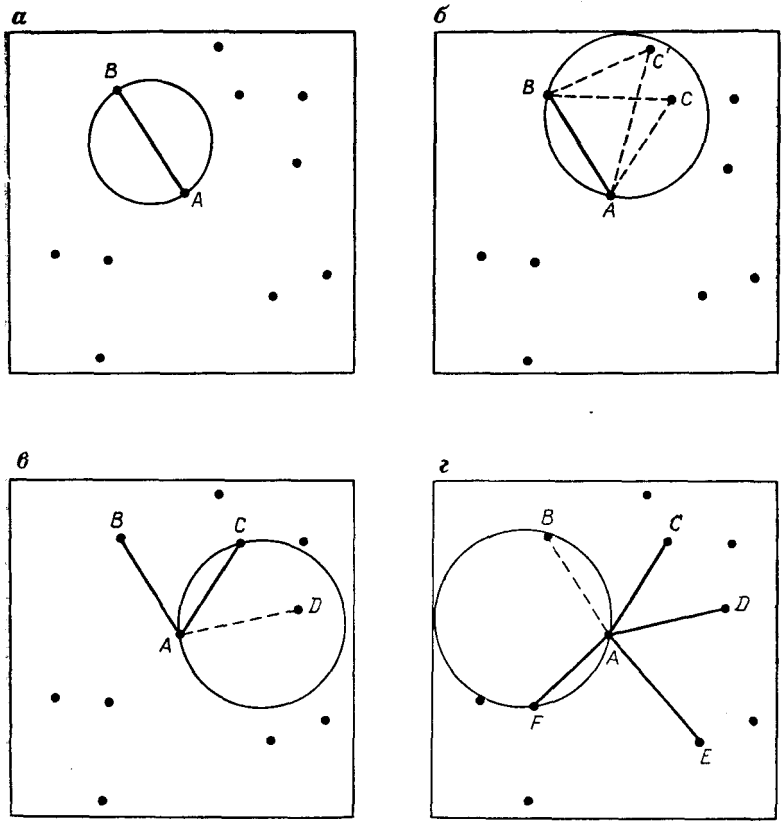


Рис. 5.47. Определение Тиссеновских соседей точки *A*: *а* – в качестве возможного соседа выбрана точка *B* и построена окружность с диаметром *AB*. Если окружность не содержит никаких других точек, то *B* есть сосед; *б* – большая окружность с точками *A* и *B* на ней используется для поиска ближайшего соседа в направлении против часовой стрелки. *C* является соседом, так как угол *BCA* превышает угол *BCA*; *в* – исходя из окружности для хорды *AC*, находим соседа *D*; *г* – окончательный поиск круга приводит снова к точке *B*

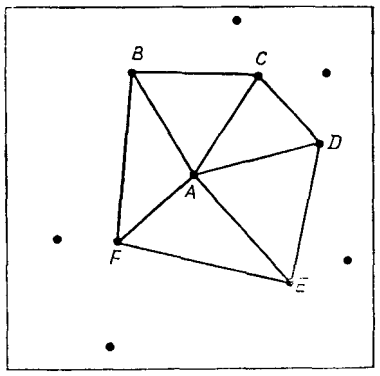


Рис. 5.48. Соединяя Тиссеновских соседей точки *A*, получаем треугольную сеть вокруг точки *A*. Процесс затем повторяют для одной из точек *B* – *F* и затем опять повторяют

В последние годы наметился возврат к триангуляционным процедурам благодаря созданию алгоритма, который позволяет получить почти оптимальную сеть на первом шаге [39], [55]. Эти сети, называемые триангуляцией Делоне, определены единственным образом для данного множества точек. В дополнение к этому образуемые треугольники настолько близки к равносторонним, насколько это возможно. Это значит, что наибольшие расстояния, на которых должна производиться интерполяция для нахождения уровней изолиний меньше, чем в другой треугольной сети.

Если мы имеем дело с рассеянным множеством точек, то можно себе представить, что каждая точка, лежащая внутри многоугольника, расположена ближе к любой другой точке, содержащейся в нем, чем к точке вне его (рис. 5.45). Наоборот, каждая точка, находящаяся вне заданного многоугольника, ближе к некоторой другой, чем к точке, лежащей внутри этого многоугольника. Это – наиболее компактное подразделение пространства из всех возможных. Множества многоугольников с такими свойствами называются многоугольниками Тиссена, Дирихле, Вороного и возникают во многих областях.

Географы используют многочлены Тиссена для моделирования зон влияния конкурирующих городов. Модель роста кристаллов в твердеющем растворе в металлургии основана на многогранниках Вороного, являющихся трехмерным обобщением многоугольников. Совокупность мыльных пузырей образует легко наблюдаемую сеть многогранников Вороного.

Многоугольники, непосредственно примыкающие к многоугольнику Тиссена, заключающему заданную точку А, также являются многоугольниками Тиссена, каждый из которых заключает единственную точку. Эти точки называются тиссеновскими соседями точки А. Если эти точки соединить прямыми линиями, то получится треугольная сеть Делоне (рис. 5.46). Для любого размещения точек как многоугольники Тиссена, так и треугольники Делоне единственны. Процесс триангуляции состоит в определении тиссеновских соседей последовательных точек на карте. На рис. 5.47, а найдены соседи точки *A*.

Сначала предположим, что вблизи точки *B* имеется сосед и построим круг, диаметр которого определяется отрезком *AB*. Если внутри круга нет других точек, то *B* действительно будет соседом *A*. Если некоторая точка внутри круга будет найдена, то она заменит точку В. Поиск следующего соседа производится против часовой стрелки относительно точки *A*. Круг разлагается указанным образом так, чтобы точки *A* и *B* были расположены на его периметре. Затем внутренность круга проверяется на наличие каких-либо заключенных в нем точек. Если будет найдена одна точка, то она будет вторым тиссеновским соседом. Если будут найдены две или более точки, надо правильно определить вторую окрестность. Это делается с помощью вычисления угла, образованного точкой *б*, точкой-кандидатом и точкой *A*. Истинный тиссеновский сосед будет образовывать наибольший угол (рис. 5.47, *б*). Поиск третьего тиссеновского соседа проводится с помощью вычерчивания такой окружности, что точка *A* и точка *C*, второй сосед, лежат на ее периметре. Проверяют, нет ли внутри этого круга какой-либо точки, которую можно было считать третьим соседом *D* (см. рис. 5.47, *б*). Далее строится круг, который на своем периметре содержит точки *A* и *D* и внутри которого производится поиск четвертого тиссеновского соседа. Может случиться, что точка *B* будет снова объявлена тиссеновским соседом; тогда все соседи *A* должны быть отождествлены (рис. 5.47, *г*). Соединив этих соседей, получим треугольную сеть вокруг *A* (рис. 5.48).

Один из Тиссеновских соседей теперь обозначается через новую точку *A*, вокруг которой снова будет производиться поиск, и весь процесс начнется снова. Сеть разрастается, распространяясь, подобно волне, по карте до тех пор, пока каждая точка не будет в нее включена. Для достижения эффективности вычислений целесообразно предварительно осуществить сортировку координат точек для того, чтобы при поиске сначала рассматривались наиболее вероятные кандидаты на роль соседа. Хотя этот процесс довольно трудно описать, МакКаллах и Росс [55] определяют число необ-

ходимых шагов для осуществления триангуляции некоторого множества из n точек и показывают, что оно пропорционально $nlog(n)$. Наоборот, число операций, требуемых для построения треугольной сети методом проб и ошибок, пропорционально n^3, а число шагов при построении сети пропорционально n^2.

Допущение, что треугольники представляют наклоненные плоские пластинки, очевидно, следует из очень грубой аппроксимации поверхности. Более точная аппроксимация может быть достигнута, если использовать изогнутые треугольные пластинки, в частности, если удастся соединить их гладко вдоль краев треугольников. Для этой цели использовались различные методы. Один из наиболее ранних состоит в нахождении трех ближайших соседей к сторонам треугольника, и затем в построении полиномиальной тренд-поверхности второй степени по этим точкам и по вершинам треугольника. Поверхность тренда второй степени куполо- или бассейнообразная, она определяется шестью коэффициентами. Это означает, что искомая поверхность проходит в точности через все шесть точек. Полученное уравнение затем может быть использовано для нахождения ряда точек, имеющих данную отметку. Эти точки соединяются, образуя не прямые линии, а искривленные.

Даже несмотря на то, что примыкающие пластинки подбираются с помощью использования одних и тех же точек, их поверхности тренда не совпадают точно вдоль линии перекрытия. Это значит, что направления могут меняться скачком, когда изолинии переходят с одной треугольной пластинки на другую. Один из путей исправления положения – это совмещать изолинии двух поверхностей, усредняя их.

Наиболее элегантная процедура основана на использовании трехмерных эквивалентов сплайн-функций, введенных в предыдущей главе. В деталях они рассмотрены в [74]. Поверхностные интерполяционные уравнения, используемые в анализе конечных элементов, также могут быть применены для оконтуривания [30], [54]. Процедуры, используемые МакКаллахом для моделирования формы поверхности внутри каждой треугольной пластинки, слишком сложны, чтобы здесь описывать их в деталях. Интересующихся можно отослать к его статье или к полному математическому изложению [8]. Интерполяционное уравнение называется трикубическим многочленом, и получение оценки каждой точки внутри треугольника в форме, используемой при проведении изолиний, требует девяти параметров. Первые три из этих параметров – это на самом деле попарные произведения сторон треугольника. Второе множество трех параметров – это по существу координаты оцениваемой точки, выраженные по отношению к каждой из трех вершин. Последнее множество из трех координат – это первые производные поверхности в каждой вершине. Производная в некоторой вершине оценивается аппроксимацией плоскости к ее тиссеновским соседям с помощью метода наименьших квадратов. Эта плоскость подчиняется условию принимать значение Y в этой вершине. Тогда координаты X_1 и X_2 плоскости комбинируются так, чтобы образовать общий угол наклона.

Эти девять параметров линейно комбинируются таким образом, чтобы дать оценку $\overline{Y}$ в заданной точке. Все оценки внутри треугольника лежат на гладкой искривленной поверхности, которая изменяется непрерывно с аппроксимирующей поверхностью в примыкающих треугольниках. Непрерывность вдоль сторон двух примыкающих треугольников обеспечивается тем, что они имеют две общие вершины и разделенные поровну обобщенные наклоны.

При проведении изолиний уравнение оценивания обращается. Значение $\overline{Y}$ полагается равным заданному уровню изолиний, и для нескольких выбранных координат X_1 находятся координаты X_2 (или наоборот). В результате получается ряд точек, имеющих постоянные значения $\overline{Y}$. Изолиния может быть проведена простым соединением этих точек.

Набор значений, по которому строится карта в изолиниях, вводится в машину в виде матрицы порядка $n\times3$, в которой каждая строка содержит три элемента: X_1 и X_2 – координаты и Y – картируемая характеристика, заданная как функция на множестве значений координат. На рис. 5.49 приведен типичный набор точек с соответствующими им значениями результатов измерения абсолютных отметок топографической поверхности. Эти данные получены при мензульной съемке и равномерно распределены на изучаемой площади с учетом заданного масштаба карты. Все эти данные с соответствующими им координатами приведены в табл. 5.11. Для удобства за начало координат принят левый нижний угол карты, а значения координатных отсчетов выражены в произвольных единицах (одна единица – 50 футов). Положение точек наблюдения можно было бы выразить и в любых других единицах, что не повлияло бы на результаты.

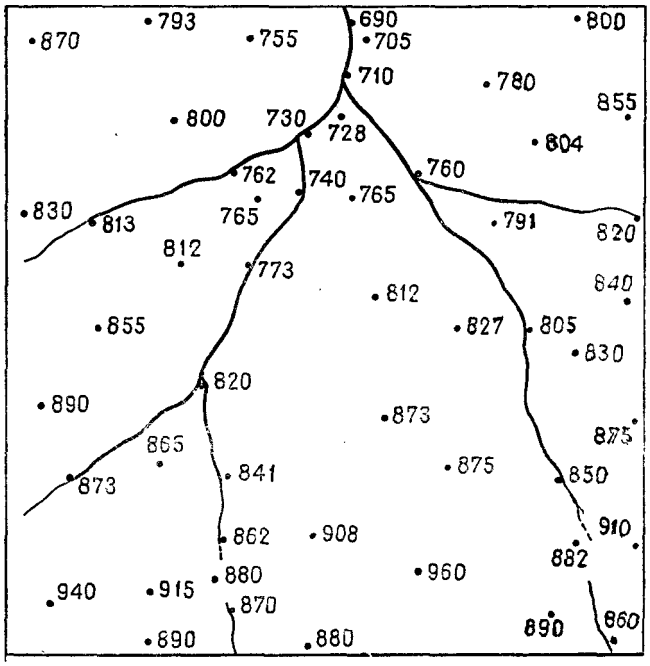


Рис. 5.48. Изображение контрольных точек для задачи топографического картирования. За единицу масштаба выбрали 50 футов, начало отсчета – левый нижний угол; высота над уровнем моря – в футах

Таблица 5.11. Географические координаты и возвышения контрольных точек в проблемах разведки

Координаты север – юг	Возвышение восток–запад	Возвышение над уровнем моря, футы	Координаты восток–запад	Координаты север–юг	Возвышение над уровнем моря, футы
0,3	6,1	870	5,2	3,2	805
1,4	6,2	793	6,3	3,4	840
2,4	6,1	755	0,3	2,4	890
3,6	6,2	690	2,0	2,7	820
5,7	6,2	800	3,8	2,3	873
1,6	5,2	800	6,3	2,2	875
2,9	5,1	730	0,6	1,7	873
3,4	5,3	728	1,5	1,8	865
3,4	5,7	710	2,1	1,8	841
4,8	5,6	780	2,1	1,1	862
5,3	5,0	804	3,1	1,1	908
6,2	5,2	855	4,5	1,8	855
0,2	4,3	830	5,5	1,7	850
0,9	4,2	813	5,7	1,0	882
2,3	4,8	762	6,2	1,0	910
2,5	4,5	765	0,4	0,5	940
3,0	4,5	740	1,4	0,6	915
3,5	5,5	765	1,4	0,1	890
4,1	4,6	760	2,1	0,7	880
4,9	4,2	790	2,3	0,3	870
6,3	4,3	820	3,1	0,0	880
0,9	3,2	855	4,1	0,8	960
1,7	3,8	812	5,4	0,4	890
2,4	3,8	773	6,0	0,1	860
3,7	3,5	812	5,7	3,0	830
4,5	3,2	827	3,7	6,0	705

На рис. 5.50 представлено множество треугольников Делоне, построенных МакКаллахом по его программе. Для проведения изолиний вне ограничивающего многоугольника, который заключает большинство контрольных точек, необходимо разместить вдоль границы карты ряд псевдоточек. В этом примере одна псевдоточка размещается в вершине каждого угла, и еще две размещаются вдоль каждой стороны карты.

Окончательная карта, построенная методом треугольников, приведена на рис. 5.51. Она очень напоминает карту (см. рис. 5.58), построенную методом сеток, но имеются различия в деталях. Наиболее очевидные из них – это узкие области, на которых наклон поверхности резко изменяется, как на юго-западном и центральном северном участках карты. Это как раз площади, где треугольники Делоне очень острые. Вдоль полей карты эти черты можно исправить, вставив разумным образом псевдоточки, но они не могут быть изменены внутри поля карты до тех пор, пока большее количество данных не окажется доступным.

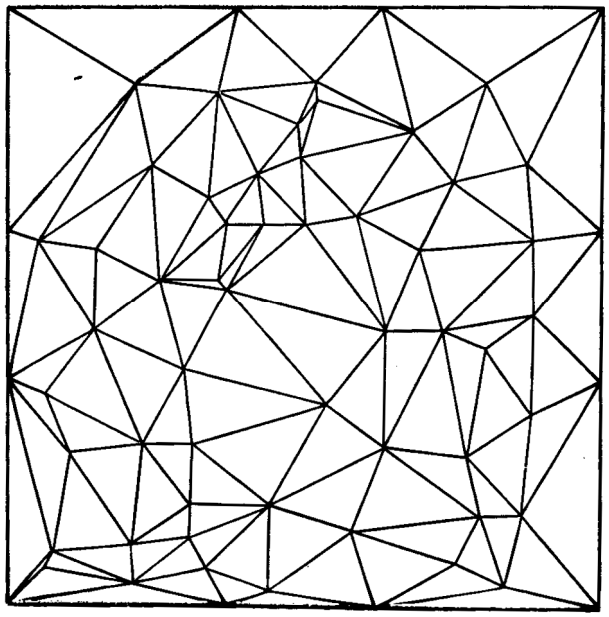


Рис. 5.50. Треугольная сеть Делоне для контрольных точек, изображенных на рис. 5.49. Добавлены ложные точки по краю карты

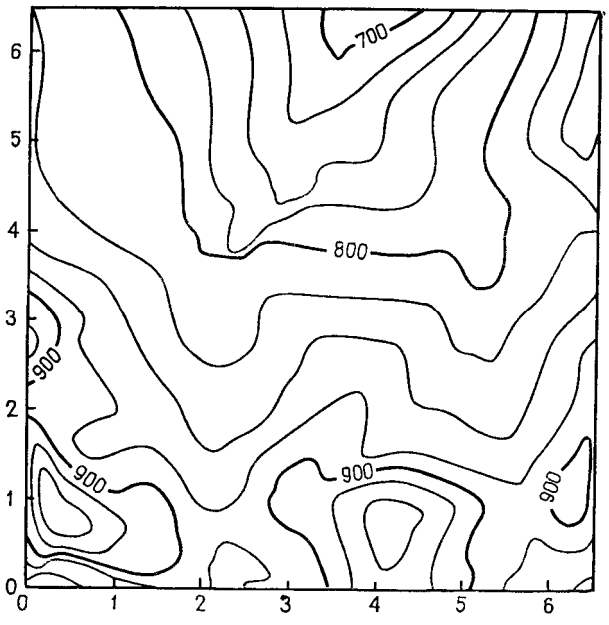


Рис. 5.51. Топографическая карта в изолиниях, полученная по программе, в которой в качестве математической модели поверхности использована треугольная сеть. Интервал между изолиниями – 25 футов (0,3 м)

Построение карт в изолиниях методом сетей

Построение методом сетей – это процесс определения значений поверхности в некотором множестве точек, которые размещены по регулярной схеме, обычно квадратной, которая полностью покрывает картируемую площадь. В общем случае значения поверхности в этих равномерно расположенных в пространстве точках неизвестны, и задача состоит в их оценке по известным значениям поверхности в неправильно расположенном множестве контрольных точек. Точки, в которых производится оценка, называются узлами сети.

В этом методе сначала строится математическая модель поверхности, которая имеет форму наклонной квадратной пластинки. В простом алгоритме эти пластинки плоские. В более сложных алгоритмах они искривлены и каждая гладко переходит в примыкающие пластинки. Математическая модель строится для чисто практических целей. Много легче провести изолинии через сеть регулярно расположенных точек, чем проводить их через иррегулярную сеть исходных точек. Все возможные способы, которыми изолиния входит в квадрат и покидает его и которые определены четырьмя равномерно расположенными узлами сети, известны. Легко написать алгоритм вычерчивания линий с учетом всех этих возможностей. Изолиния может быть проведена просто вычерчиванием пути при ее переходе из одного квадрата сети в следующий. Определить путь изолинии через иррегулярную схему контрольных точек, как это делается в алгоритме метода треугольников, намного труднее. Отдельные точки не могут быть связаны так, чтобы образовать регулярную сеть, поэтому возможные пути изолинии заранее неизвестны. Значит, явные координаты X_1 и X_2 всех промежуточных точек, содержащихся в вычислениях пути изолиний, должны сохраняться в памяти ЭВМ. В регулярной схеме оцененных значений координаты X_1 и X_2 определяются положением в схеме. Это позволяет сэкономить память ЭВМ и время вычислений.

Узлы сети или промежуточные точки, в которых должны быть оценены значения поверхности, обычно располагаются в квадратную схему, в которой расстояния между узлами в одном направлении такие же, как и в перпендикулярном к нему. В большинстве программ построения карт в изолиниях эти расстояния находятся под контролем пользователя и их величина является одним из многих параметров, который должен быть выбран прежде, чем поверхность будет покрыта сетью и картирована. Площадь, заключенная между четырьмя вершинами, называется ячейкой сети; если ячейка сети выбрана большой, полученная карта будет иметь низкую разрешающую способность и грубый вид, зато может быть быстро построена. Наоборот, если ячейки сети малы, то карта будет иметь больше деталей, однако потребует больше усилий для ее построения.

Так как алгоритм построения сети позволяет оценить только одно значение по набору близких контрольных точек, то процедура оценки должна быть повторно применена в пределах всего поля карты до тех пор, пока вся карта не покроется регулярной сетью оцененных значений. Как только регулярная сеть оценок построена, изолинии могут быть проведены.

В некоторых пакетах программ начальный шаг построения оценок узлов сети дополняется одним или более дополнительными шагами, в которых оценки узлов сети уточняются. Обычно узлы сети непосредственной окрестности каждой контрольной точки вычисляются повторно, при этом используются как исходные контрольные точки, так и первоначальные оценки в окрестности узлов сети. Это позволяет получить карту поверхности, которая находится ближе к контрольным точкам, чем на предыдущем шаге.

Построение сети или вычисление регулярной схемы оцененных значений содержит три существенных шага. Первый: контрольные точки должны быть рассортированы в соответствии с их географическими координатами. Второй: контрольные точки, окружающие оцениваемый узел сети, должны быть выбраны из рассортированных файлов. Третий: алгоритм должен дать оценку значения в этом узле сети через некоторую математическую функцию значений в соседних точках. Сортировка значительно влияет на скорость выполнения операций и, следовательно, на цену выполнения программы построения изолиний. Однако она не изменяет точность оценок, и потому мы не будем рассматривать ее далее. Как выбор программы поиска, так и выбор математической функции имеет значительное влияние на окончательную форму карты.

Большинство известных функций, которые позволяют оценить значение поверхности в заданной точке карты, это просто вычисление среднего известных значений поверхности по близким контрольным точкам. Действительно, оно сводится к горизонтальному проектированию всех этих окружающих известных значений на оцениваемое положение (рис. 5.52). Затем проводится сложная

оценка с помощью усреднения этих точек, обычно взвешенных с большим весом для самых близких точек, чем для более удаленных. Если это сделано на регулярной сети по всему полю карты, полученная карта будет иметь определенные характеристики. Наибольшие и наименьшие площади на поверхности будут содержать контрольные точки и большинство узлов сети интерполяции будет лежать в промежуточных значениях, так как среднее не может лежать вне области влияния значений, по которым оно было вычислено. В узлах сети, расположенных вне большинства известных точек, оценки получаются экстраполяцией и будут близкими по величине к значениям в ближайших контрольных точках.

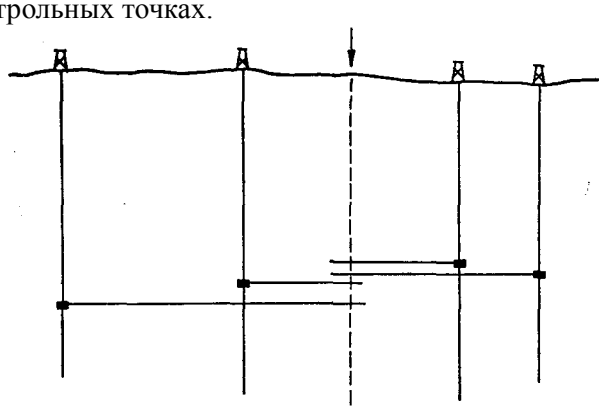


Рис. 5.52. Вид разреза, на котором представлены контрольные точки (скважины) с известными отметками поверхности (кровли формации).
Оценки в узле сети (стрелка) сделаны в результате горизонтального проектирования и последующего усреднения

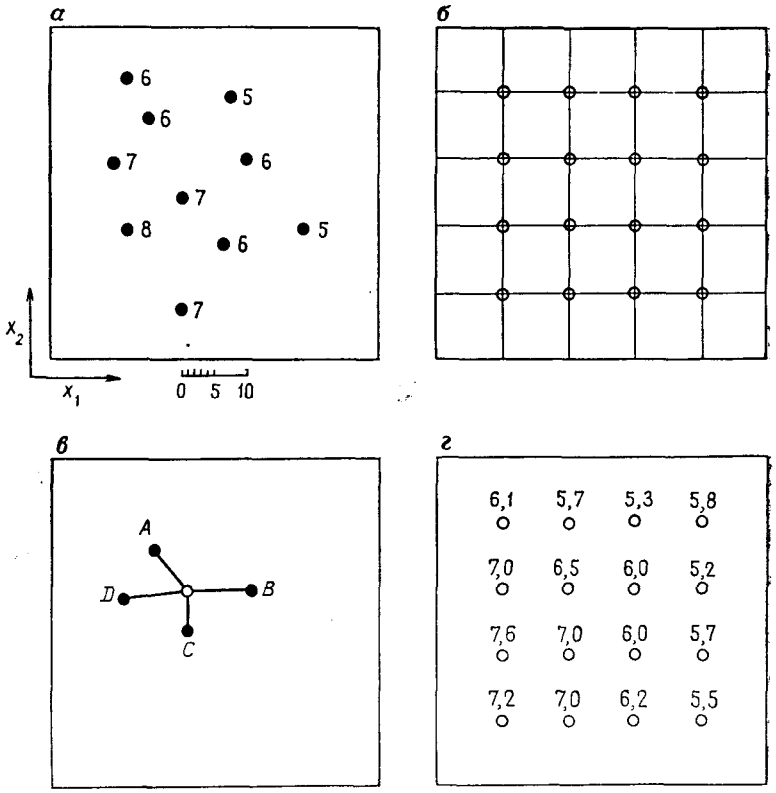


Рис. 5.53. Последовательность вычислений для построения изолиний при нахождении значений в узлах сети: *а* – исходный набор неравномерно расположенных контрольных точек на карте; числа указывают абсолютные отметки; *б* – равномерная сеть, в узлах которой вычисляются значения; *в* – расположение четырех ближайших контрольных точек по отношению к узлу равномерной сети; эти четыре ближайших значения используются для вычисления значения в данном узле; *г* – окончательный результат вычисления отметок в каждом узле равномерной сети

На рис. 5.53, *а* изображена серия наблюдений, причем каждая точка охарактеризована значениями координат X_1 и X_2, а также значением высоты над уровнем моря, которое приведено справа от каждой точки.

Набор точек можно перенумеровать, т.е. приписать каждой точке номер i. Следовательно, в новых обозначениях точка с номером i будет обладать координатами X_{1i} и X_{2i}, а также абсолютной отметкой Y_i. На рис. 5.53, *б* приведена выбранная правильная сеть точек, по которой будут строиться изолинии. Каждой из этих точек можно приписать соответствующий номер k. Таким образом, точка этой сети с номером k будет обладать координатами X_{1k}, X_{2k} и вычисленным значением $\overline{Y_k}$.

Нам нужно вычислить оценку $\overline{Y_k}$ по n ближайшим к ней исходным точкам наблюдения. Следовательно, сначала нужно найти эти n ближайших точек и подсчитать соответствующие им расстояния от точки с номером k заданной сети. Процедура поиска может быть простой или сложной; позже мы рассмотрим различные методы. Теперь мы предположим, что с помощью некоторого метода мы определим n ближайших точек к заданной точке с номером k. Согласно теореме Пифагора расстояние D_{ik} от точки с номером i до точки с номером k будет равно

$$D_{ik} = \sqrt{(X_{ik} - X_{1i})^2 + (X_{2k} - X_{2i})^2} \quad (5.66)$$

Вычислив расстояние D_{ik} для всех n ближайших точек, можно подсчитать значение $\overline{Y_k}$ по следующей формуле:

$$\overline{Y}_k = \frac{\sum_{i=1}^{n}(Y_i / D_{ik})}{\sum_{i=1}^{n}(1/D_{ik})} \qquad (5.67)$$

Процесс этих вычислении можно показать на примере данных, приведенных на рис. 5.53, в. Мы произвольно выберем четыре ближайшие точки (т.е. $n = 4$) и подсчитаем $\overline{Y}_k$. На рис. 5.53, в числа 1, 2, 3, 4 являются номерами точек. Тогда

$$D_{1k} = \sqrt{(2{,}0-1{,}5)^2 + (3{,}0-3{,}6)^2} = \sqrt{0{,}61} = 0{,}78$$
$$D_{2k} = \sqrt{(2{,}0-3{,}0)^2 + (3{,}0-3{,}0)^2} = \sqrt{1{,}0} = 1{,}00$$
$$D_{3k} = \sqrt{(2{,}0-2{,}0)^2 + (3{,}0-2{,}4)^2} = \sqrt{0{,}36} = 0{,}60$$
$$D_{4k} = \sqrt{(2{,}0-1{,}0)^2 + (3{,}0-2{,}9)^2} = \sqrt{1{,}01} = 1{,}00$$

Используя полученные расстояния, можно вычислить $\overline{Y}_k$. Числитель выражения (5.67) будет равен

$$\frac{6{,}0}{0{,}78} + \frac{6{,}0}{1{,}00} + \frac{7{,}0}{0{,}60} + \frac{7{,}0}{1{,}00} = 32{,}36,$$

соответственно знаменатель определяется как сумма

$$\frac{1}{0{,}78} + \frac{1}{1{,}00} + \frac{1}{0{,}60} + \frac{1}{1{,}00} = 4{,}95$$

так что

$$\overline{Y}_k = \frac{32{,}36}{4{,}96} = 6{,}54$$

Точно так же можно выполнить эту процедуру и для остальных точек заданной сети, которая со всеми вычисленными значениями $\overline{Y}_k$ приведена на рис. 5.53, *г*.

Этот тип алгоритма иногда называют «скользящим средним», так как каждый узел сети оценивается как среднее значение в контрольных точках внутри окрестности, т. е. оценка движется от узла сети к другому узлу. Такие алгоритмы могут рассматриваться как специальные случаи более общего множества процедур, которые содержат подбор плоскостей или искривленных поверхностей к контрольным точкам в пределах некоторой окрестности. Сначала находятся все контрольные точки вокруг оцениваемого узла сети, расположенные внутри заданной окрестности. Мы можем затем вообразить, что значения в этих точках приближенно определяют наклонную плоскость. Эта плоскость может быть охарактеризована как поверхность линейного тренда, коэффициенты уравнения которого могут быть найдены методом наименьших квадратов. Коэффициенты плоскости вычисляются в точности так же, как и коэффициенты поверхности тренда, только, конечно, по точкам внутри используемой окрестности.

После того как найдено уравнение плоскости, в него подставляются значения координат X_{1k} и X_{2k}, соответствующие узлу сети. Это даст значение $\overline{Y}_k$, которое является оценкой поверхности для этого узла сети. Процесс подбора плоскости и получения с помощью ее уравнения оценок по-

верхности повторяется для каждого узла сети. Эта плоскость задает общий наклон поверхности в окрестности узла сети. Действительно, значения поверхности в контрольных точках внутри окрестности проектируются параллельно этой наклонной плоскости, затем усредняются в узле сети (рис. 5.54). Если подбираемая плоскость не наклонная, а совершенно горизонтальная, то оценка, полученная этим методом, будет такой же, что и оценка, полученная методом скользящего среднего.

Процедуры построения сетей, основанные на таком подборе плоскостей, иногда называются кусочно-линейными, основанными на методе наименьших квадратов. Ее разновидность, называемая кусочно-квадратичной, основанной на методе наименьших квадратов, отличается лишь тем, что подбираемая поверхность квадратичная, а не плоскость. Квадратичная, или тренд-поверхность второй степени имеет форму соляного купола, бассейна или седла и определяется уравнением, содержащим квадраты переменных X_1 и X_2 и их попарные произведения.

Так как эти алгоритмы используют наклон плоскости в некоторой окрестности, то они работают лучше, чем простые методы скользящего среднего, основанные на использовании методов интерполяции между контрольными точками. В точках сети получаются значения, которые могут быть либо больше, либо меньше, чем значения в контрольных точках. Однако вне этих зон контроля экстраполяция может дать экстремальные значения, которые ничем не оправданы. Это случается потому, что любые наклоны, которые существуют вблизи полей контролируемой части карты, продолжаются неограниченно за пределы данных. Использование квадратичной поверхности в этом случае нецелесообразно.

Несколько более сложный алгоритм позволяет вычислить локальный наклон или угол падения картируемой поверхности в каждой контрольной точке. С помощью линейного метода наименьших квадратов к контрольным точкам вокруг узлов сети подбирается плоскость, т.е. определяются коэффициенты уравнения плоскости, которая проходит через точки поверхности в окружающих контрольных точках настолько близко, насколько это возможно.

Алгоритм осуществляется в два шага. На первом шаге определяется окрестность вокруг каждой контрольной точки и находятся все точки внутри этой окрестности. Затем к известным значениям поверхности в этих точках методом наименьших квадратов подбирается плоскость. Однако плоскость подчинена ограничению: она обязана проходить в точности через значения в центральной контрольной точке. Коэффициенты этой плоскости, которые определяют наклон поверхности в центральной точке, вдоль нее сохраняются вместе со значением в этой точке.

На втором шаге определяется окрестность вокруг каждого оцениваемого узла сети. Находятся контрольные точки, внутри этой окрестности и вычисляются уравнения плоскостей в каждой из контрольных точек для этого положения узла сети. Различные оценки из этих плоскостей затем взвешиваются и комбинируются. Действительно, наклоны поверхности в контрольных точках проектируются на узел сети и затем усредняются (рис. 5.55).

Различные варианты этого алгоритма, иногда называемого «линейным проектированием», наиболее популярны из используемых в пакетах алгоритмов, предназначенных для коммерческого построения карт в изолиниях. В некоторых пакетах содержатся модификации этой процедуры, в которых к контрольным точкам подбираются не плоские, а квадратичные поверхности. Эти алгоритмы особенно эффективны в пределах площадей, которые плотно заполнены равномерно заданными в пространстве точками. Подобно кусочно-линейным методам наименьших квадратов, они имеют неприятное свойство давать оценки узлов сети вне географических пределов данных.

Контрольные точки, используемые в оценке узла сети, независимо от того, проектируются они или нет, обычно являются взвешенными. Веса изменяются в соответствии с расстояниями между оцениваемыми узлами сети и контрольными точками. На рис. 5.56 представлено несколько часто используемых весовых функций. Большинство программ позволяет пользователю выбрать требуемую среди этого множества функций.

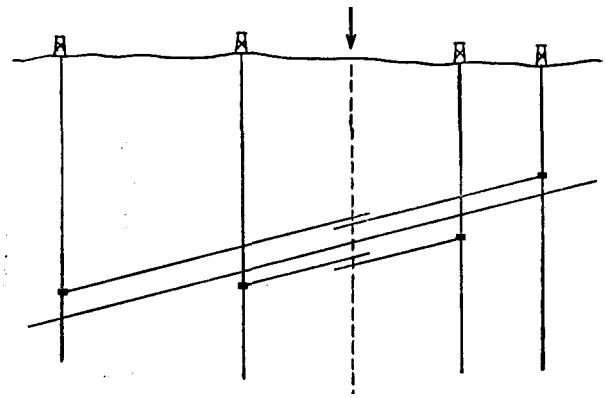


Рис. 5.54. Вид разреза, на котором представлены контрольные точки (скважины) с известными отметками поверхности (кровли формации). Оценки в узле сети (стрелка) сделаны путем подбора линейной функции (плоскости) *A–A′* к известным значениям и последующего вычисления ее значений

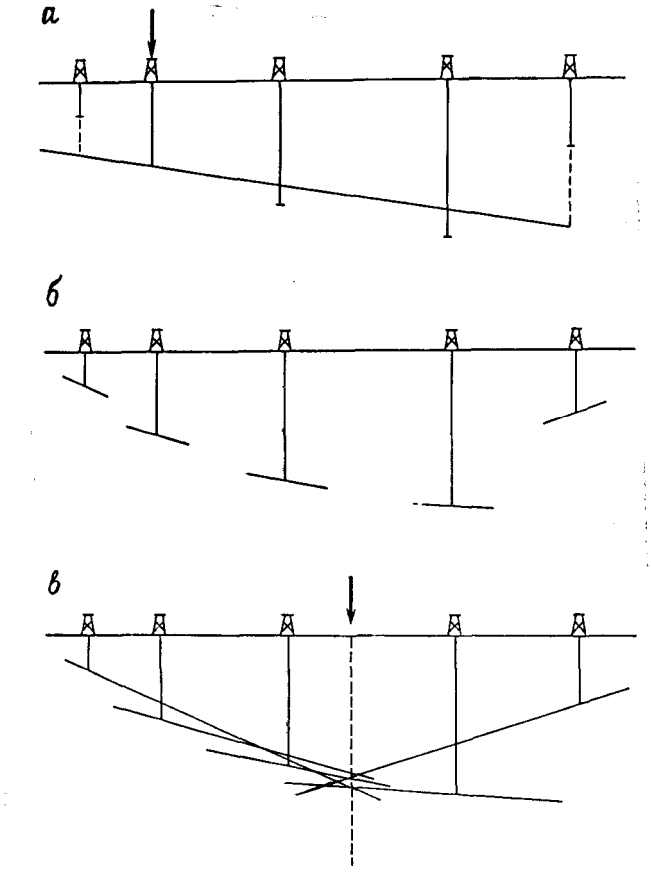


Рис. 5.55. Интерполяция для равномерно расположенных точек сети с помощью метода проектирования падения, который требует вычисления локального падения в каждой данной точке (изображено в разрезе) [69]: *а* – локальное падение поверхности в данной точке (стрелка) находится с помощью подбора плоскости по близлежащим точкам методом наименьших квадратов; подбираемая плоскость должна пройти через данную точку; *б* – локальное падение становится частью данных для каждой контрольной точки; *в* – локальное падение в контрольных точках проектируется на узлы сети (стрелка). Значения, приписанные узлам, являются взвешенными средними этих проекций

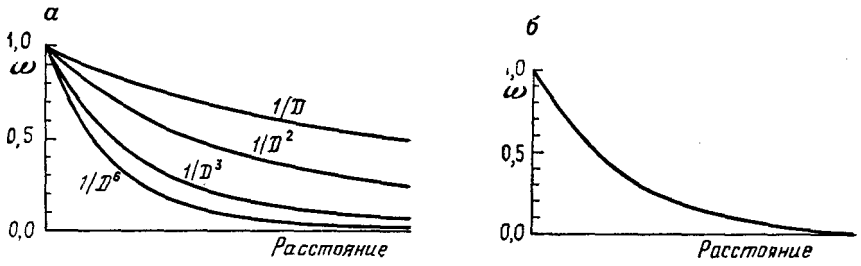


Рис. 5.56. Примеры взвешивающих функций от расстояния, используемых в программах построения карт в изолиниях: *а* – функции обратных степеней расстояния; *б* – шкалированная функция обратного квадрата расстояния

Веса, приписанные контрольным точкам в соответствии с некоторыми функциями, подчиняются условию: их сумма должна равняться единице. Поэтому весовые функции в действительности снабжаются пропорциональными весами и выражают относительное влияние на каждую контрольную точку. Широко используемая версия процесса взвешивания приводит к функции, точная форма которой зависит от расстояния от оцениваемой точки и от наиболее удаленной точки, используемой в оценивании, или в одном из вариантов, от расстояния до внешних границ окрестности. Весовая функция обратных значений квадратов расстояний затем шкалируется, так что принимаемые ею значения на этом промежутке оказываются между нулем и единицей. Этот процесс можно

записать единственным уравнением

$$\omega_p = \left(1 - \frac{D}{D_{\max}}\right)^2 \Big/ \left(\frac{D}{D_{\max}}\right) \qquad (5.68)$$

Подобно другим весовым функциям, сумма весов полагается равной 1,0.

Наиболее очевидные отличия между различными программами построения карт в изолиниях заключаются в применяемом методе поиска. Они основаны на алгоритмах, используемых для выбора данных точек внутри локальной окрестности вокруг оцениваемой точки сети. Простейший метод выбора носит название поиска ближайшего соседа (рис. 5.57,*а*). Он основан на локализации некоторого определенного числа контрольных точек или скважин, которые находятся на ближайшем расстоянии от оцениваемого узла сети. Множество возможных ближайших контрольных точек выбирается из полного набора данных сортировкой координат X_1 и X_2 этих точек. Затем вычисляются евклидовы расстояния от оцениваемого узла до каждой из этих точек и находится заданное число ближайших точек.

Возражение против простого метода поиска ближайшего соседа состоит в том, что при этом может случиться, что все близкие точки лежат в узкой полосе по одну сторону оцениваемого узла сети. Окончательная оценка в сущности не ограничена, исключая одно направление. Оно оказывается выброшенным из-за некоторого ограничения поиска, обеспечивающего равномерное распределение контрольных точек вокруг оцениваемой точки. На рис. 5.57,*в* проиллюстрирован один способ ограничения, называемый квадрантным поиском. Из каждого из четырех квадрантов вокруг оцениваемого узла сети должно быть извлечено некоторое минимальное число контрольных точек. Усложнение квадрантного поиска – это октантный поиск (см. рис. 5.57,*в*), который требует введения нового ограничения на радиальное распределение точек, используемых в уравнении оценки. Заданное число контрольных точек должно быть найдено в каждом из октантов, окружающих оцениваемый узел сети. Этот метод поиска – один из наиболее изящных в настоящее время, на нем часто основываются коммерческие программы.

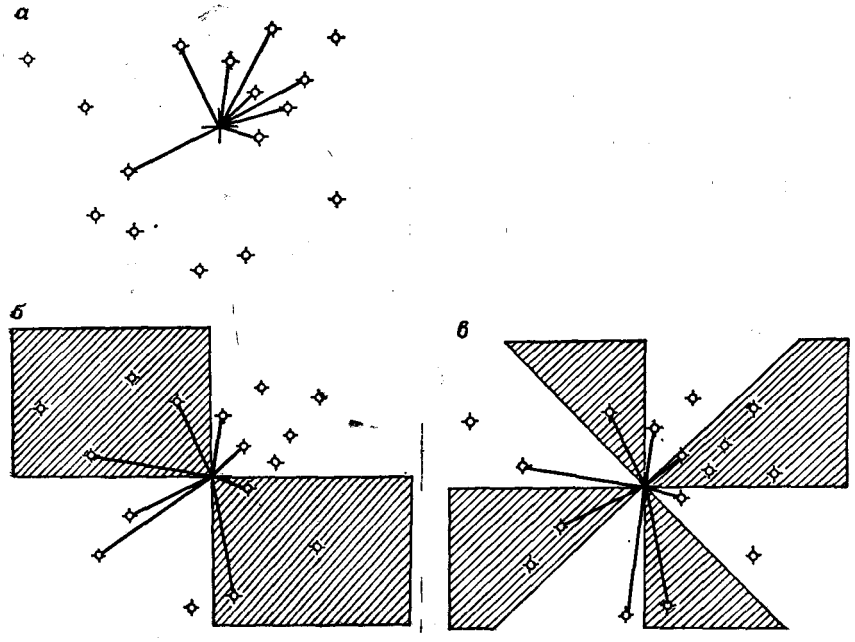


Рис. 5.57. (*а*) Метод поиска, локализующий n ближайших соседей вокруг оцениваемого узла сети. На радиальное распределение контрольных точек не накладывается никаких ограничений. (*б*) Схема поиска по квадратам, используемая для установления истинного распределения контрольных точек вокруг оцениваемых узлов сети. (*в*) Октантный вариант схемы поиска вокруг оцениваемых узлов сети

Любые ограничения на форму области поиска ближайших контрольных точек, например, таких, как квадрант или октант, очевидно, приводят к уменьшению окрестности вокруг оцениваемого узла сети. Вот почему некоторые близлежащие контрольные точки могут перейти в разряд более удаленных точек в связи с необходимостью удовлетворения требования того, что только некоторые из них могут быть взяты из единичного объема. К сожалению, автокорреляция типичной геологической поверхности уменьшается с увеличением расстояния, так что эти более удаленные контрольные точки менее тесно связаны с оцениваемой точкой. Это означает, что оценивание будет хуже, чем при использовании простой процедуры поиска ближайшего соседа.

При использовании программы проведения изолиний весьма важно, насколько хорошо ре-

зультаты согласуются с данными, т.е. насколько хорошо математическая модель согласуется с контрольными точками, которые были использованы при ее построении. Так как изолинии были проведены на основании. соотношений между узлами сети, а не на самих контрольных точках, то правильный способ построения изолинии состоит в использовании значений сети, а неправильный – в использовании данных в исходных точках. Это особенно важно, если сеть относительно груба. Обычно ошибки очень малы и легко обнаруживаются на тех площадях, где поверхность круто изменяется, так что изолинии расположены тесно. Однако на площадях с очень плавным наклоном малое различие между значениями в узле сети и контрольной точке может быть достаточно для смещения какой-либо изолинии на некоторое расстояние из точки, с которой она по предположению должна согласовываться. Это сразу бросается в глаза на примере изолинии, проходящей через контрольную точку не с той стороны. Такого, конечно, не случается в методе треугольников, так как модель поверхности образуется с помощью самих контрольных точек. Вид изолиний не является надежным критерием того, насколько хорошо рассматриваемая математическая модель представляет исходные контрольные точки. Не существует формальной статистической теории, которая позволяет на чисто теоретическом уровне определить, какая процедура проведения изолиний лучше. В любой конкретной ситуации реализация конкретного алгоритма зависит от сложности картируемой поверхности, плотности и размещения контрольных точек, размера сети и, конечно, от самого алгоритма. Эмпирические критерии качества различных сетевых алгоритмов, использующих типичные подповерхностные данные, были опубликованы Дэвисом в 1976 г. [22].

Однако представление известных данных настолько точно, насколько это возможно, но основная цель большинства задач картирования в изолиниях. Скорее всего, задача состоит в получении оценки (с наименьшими возможными ошибками) значений поверхности в точках, в которых измерения еще не были проведены. Возможность с помощью различных алгоритмов получить точные оценки в недоступных контролю точках была проверена эмпирическими критериями для данных, в которых небольшая доля доступных точек была опущена до начала картирования. Затем сравнивались истинные значения и полученные теоретически с помощью программы картирования. За тем пропущенные точки были возвращены в множество данных, снова были выброшены другие точки, и процесс продолжался снова и снова [22].

Разочаровывающий (хотя и не удивительный) вывод состоит в том, что различные объективные данные, которые использовались для построения изолиний, не являются взаимозаменяемыми. Для того, чтобы успешно воспроизвести и учесть исходные контрольные точки с помощью алгоритма сетевого типа, необходимо использовать весовую функцию, очень быстро убывающую с расстоянием и вычисленную на основе использования небольшой доли ближайших соседей. Такой алгоритм обеспечивает плохое предсказание или оценку в точках, в которых никакой контроль невозможен. Для получения наилучшего предсказания поверхности в неопробованных точках, в каждом вычислении нужно использовать много контрольных точек узлов сети и придать большие веса удаленным точкам. К сожалению, полученная таким образом гладкая обобщенная поверхность довольно плохо согласуется с исходными контрольными точками. В некоторых программах построения изолиний делается попытка преодолеть этот тупик в два этапа. На первом этапе по некоторой схеме строится сеть, предназначенная для достижения хорошего предсказания на неконтролируемых площадях, а на втором – изменение сети в непосредственной окрестности контрольных точек. Полученная сложная поверхность по меньшей мере частично удовлетворяет различным поставленным условиям. Однако она имеет специфические черты в малых окрестностях каждой контрольной точки, которые не встречаются больше нигде на этой карте.

Карта топографических данных, приведенная в табл. 5.11, представлена на рис. 5.58, построенном по программе, которая дает регулярную сеть. В этом примере математическая модель поверхности содержит 63 строки и 63 столбца; каждый узел сети был оценен с помощью проекционного алгоритма, учитывающего примерно 16 контрольных точек. Контрольные точки взвешивались в соответствии с их расстоянием от узлов сети с помощью функций, заданных уравнением (5.68). Сравним эту карту с двумя картами рис. 5.51, полученными по методу треугольников, и картой рис. 5.59, построенной вручную работником картографической службы. Имеются расхождения между двумя картами, построенными ЭВМ, и картой, построенной вручную, которые отражают различия в соответствующих моделях и также тот факт, что человек может учесть информацию о текущих эффектах на топографической карте. Эта дополнительная информация недоступна никакой программе построения изолиний. Риплай [63] дает несколько дополнительных примеров построения карт по тем же данным с применением других алгоритмов.

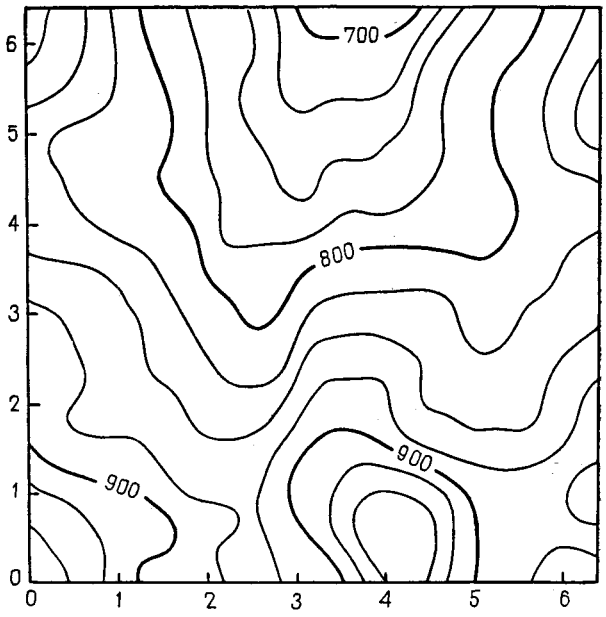


Рис. 5.58. Карта в изолиниях, построенная по топографическим данным с помощью программы, использующей равномерную сеть. Изолинии проведены через 25 футов (7,5 м)

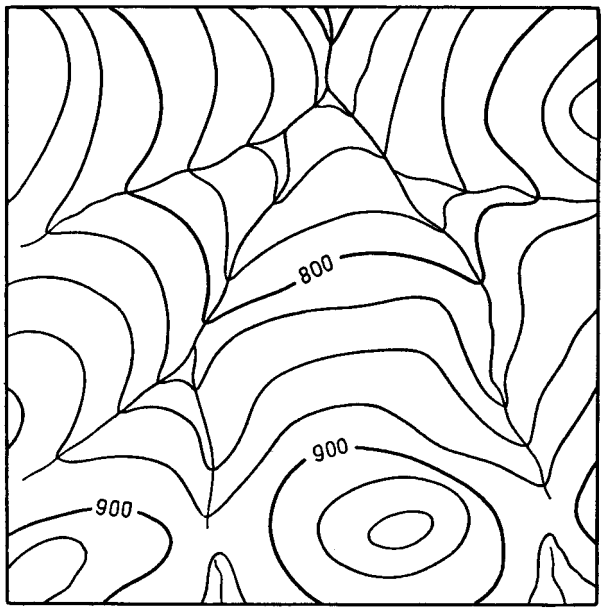


Рис. 5.59. Карта в изолиниях, построенная по топографическим данным вручную. Отметим увеличение влияния русел на форму изолиний

СКОЛЬЗЯЩИЕ СРЕДНИЕ

Несмотря на то, что некоторые из рассмотренных методов применялись в геологии рудных месторождений, оценка запасов полезных ископаемых и контроль качества разработок представляют специальные задачи. Для их решения были разработаны математические и статистические методы оценки запасов, стоящие несколько в стороне от основного направления развития математической геологии. В США статистика, используемая в геологии рудных месторождений, следовала общей линии развития традиционного статистического анализа; особенно дисперсионного (несколько примеров и подробная библиография приводятся в книге Коха и Линка [45]. Однако в Южной Африке и Франции теория оценки запасов и их предсказание развивались по независимому пути, который мы рассмотрим позже.

В пластовом осадочном рудном теле содержание рудного компонента распределено нормально. Это значит, что месторождение характеризуется определенным средним содержанием полезного компонента, а его содержание в отдельных пробах распределено более или менее симметрично относительно этого среднего значения с убыванием относительной частоты появления к крайним значениям. Для анализа проб из таких месторождений можно использовать обычные параметрические статистики, а регрессионные процедуры оказываются ценным методом предсказания и оценки запасов. Однако характеристики месторождений драгоценных минералов или редких металлов чаще всего не подчиняются этому закону распределения. Пробы, взятые с поверхности или из буровой скважины, характеризуются крайними значениями, которые являются необычными в их пространственном распределении. Вообще «хорошим» статистическим переменным свойственно отсутствие закономерности поведения.

На рис. 5.60,*а* представлены значения содержаний серебра в пробах по штольне одного из мексиканских серебряных рудников. Две характерные черты видны сразу: значения растут быстрее, чем линейная функция (богатые участки во много раз богаче, чем бедные), и изменчивость увеличивается по мере увеличения содержания. В богатой части штольни изменения значений более сильные, чем это характерно для интервала значений в бедных частях той же штольни. Суммируя, можно сказать, что эта зависимость сортности руды от расстояния характеризуется экспоненциальным законом, а дисперсия является гетероседастичной.

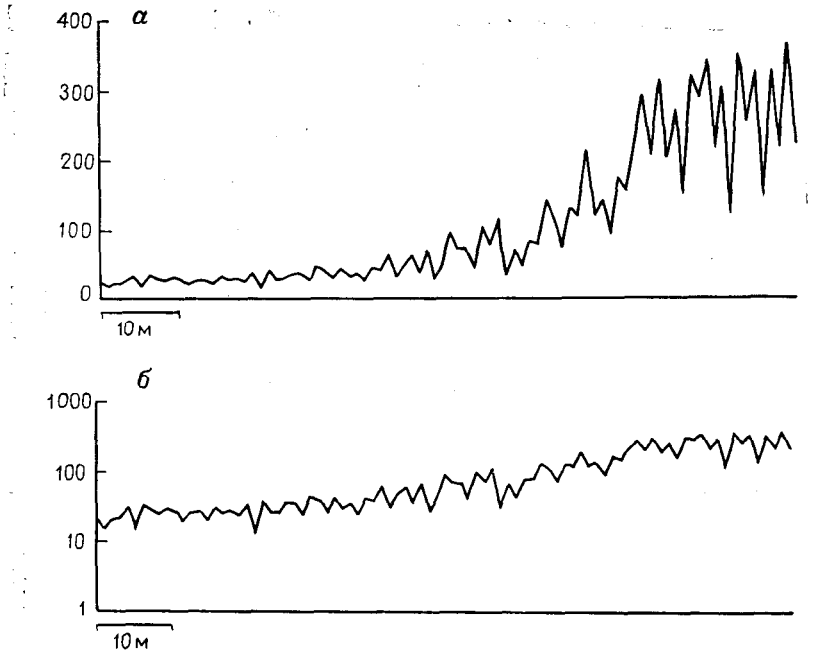


Рис. 5.60. Изменчивость содержаний серебра вдоль штольни рудника в Мексике: *а* – содержания, представленные в обычном масштабе; *б* – содержания, представленные в логарифмическом масштабе

В этом случае данные можно преобразовать, взяв логарифмы значений содержаний рудного компонента. Тогда экспоненциальная кривая тренда превратится в прямую, а гетероседастичная дисперсия относительно тренда станет постоянной. Логарифмы тех же данных представлены на рис. 5.60,*б*.

Другая особенность многих рудных месторождений заключается в том, что среднее содержание полезного компонента в блоке чаще всего не зависит от размера блока, если превышен некоторый минимальный объем. Однако часто оказывалось, что дисперсия содержаний в пробах руды находится в обратной связи с размером блока, уменьшаясь по мере увеличения объема пробы. Вероятно, это приводит к задаче оценки запасов, так как объемы проб, на основании которых должны быть получены оценки, во много раз меньше, чем блоки руды, которые затем будут отработаны. (Этот вопрос рассматривают Кох и Линк [45]. Поэтому дисперсия содержаний в пробах может оказаться настолько высокой, что реальную оценку содержаний руды в блоках дать нельзя. Для получения оценок с меньшей дисперсией можно воспользоваться методом скользящего среднего по выборке в надежде усреднить большую изменчивость, связанную с отдельными пробами.

Двумерные методы скользящего среднего являются обобщением процедур сглаживания данных, рассмотренных в гл. 4 (см. книгу 1). В общем случае требуется оценить переменную в ряде точек сети или приписать значения последовательно примыкающим друг к другу квадратам или прямоугольникам на карте. Данные, на основании которых делаются оценки, разбросаны на площади карты и могут лежать или не лежать в узлах сети. Фигура, аналогичная сглаживающему интервалу в анализе временного тренда, располагается так, чтобы ее центр лежал в точке, в которой должна быть получена оценка. Всем данным точкам, лежащим внутри этой фигуры, например квадрата или круга, приписываются некоторым образом веса, которые затем используются при получении оценки в центральной точке. Простейшая схема метода скользящего среднего состоит в том, что оцениваемой точке приписывается значение, равное среднему арифметическому всех наблюдений, лежащих внутри рассматриваемой фигуры. Фигура затем передвигается в следующий узел сетки, и процесс повторяется снова. Когда будут вычислены оценки для одной строки или столбца сети, переходят к следующей строке или столбцу, и так до тех пор, пока вся площадь карты не будет покрыта полностью.

Общую модель любого метода скользящего среднего можно записать в следующем виде:

$$\overline{Y}_{ij} = \sum_{k=1}^{n} W_k Y_k \qquad (5.69)$$

т.е. оцениваемое значение $\overline{Y}_{ij}$ строится на основании взвешенной суммы Y_k соседних наблюдений. Вид весовой функции изменяется от одной схемы скользящего среднего к другой. Например, в программе построения изолиний использовано скользящее среднее с весами, равными обратным величинам расстояний точек от $\overline{Y}_{ij}$. Можно использовать столь же простую схему выборочного средне-

го, как схема сглаживания функций, изложенная в гл. 4. Очевидно, возможны также и другие схемы взвешивания.

Большинство методов скользящего среднего так или иначе использует расстояния от оцениваемой точки до оценивающих точек. В программе построения изолиний измеряется расстояние до n ближайших точек и каждому приписываются соответствующие веса. Методы, аналогичные одномерным сглаживающим процедурам, требуют, чтобы данные были расположены по сетке. Тогда пространственное соотношение между $\overline{Y}_{ij}$ и каждым значением Y внутри скользящего интервала оказывается известным. В этом случае веса остаются постоянными для эквивалентных точек Y_k по мере того, как поверхность скользящего среднего дает последовательные оценки значений $\overline{Y}_{ij}$. В третьем методе определяются ряды областей или блоков, прилегающих к оцениваемой точке. Все наблюдения в пределах каждого из них усредняются, затем средним по блокам Y приписываются веса, которые используются для получения оценки $\overline{Y}_{ij}$. Если в каждом блоке содержится много наблюдений, то, совершая лишь незначительную ошибку, можно считать, что среднее значение соответствует центру блока. Так как центры блоков всегда расположены на фиксированном расстоянии от оцениваемой точки, то можно использовать постоянные весовые функции. Этот метод имеет значительные преимущества в том случае, если оценки строятся на основании крайне нерегулярной сети контрольных значений. Мы детально остановимся на рассмотрении метода скользящего среднего третьего типа, так как это – один из наиболее перспективных методов, который широко использовался для оценки некоторых крупных рудных месторождений Северной Америки.

Скользящие взвешенные средние, полученные по средним значениям в блоках

Даже в том случае, если минерализированная жила очень мала, в процессе разработки месторождения должен быть выбран блок породы соответствующего размера. Минимальный размер этого блока определяется методом отработки и зависит от размеров оборудования и строения породы. Хотя в процессе отработки блоков содержание полезного компонента уменьшается за счет разубоживания пустой породой, технология разработок делает это неизбежным. Все перечисленные факторы определяют минимальный размер блока, содержание в котором мы хотим оценить. Например, бессмысленно оценивать содержание руды в одной кубической единице объема, если размер блока выработки составляет 100 кубических единиц. В разработках, в которых используется метод взвешенного усреднения по блокам, наименьшая длина в каждом измерении, используемая на практике, составляет примерно 100 футов. Эта наименьшая практическая единица, используемая в разработках, меняется от одного месторождения к другому и, вероятно, является наибольшей при разработке тел вкрапленных руд и наименьшей в очень богатых гидротермальных жильных месторождениях.

К счастью, крупные блоки менее подвержены большим изменениям содержаний, чем малые пробы. Действительно, фундаментальное допущение в методе скользящих взвешенных блоков состоит в том, что минимальный размер разработки так велик по сравнению с расстоянием, на котором наблюдается быстрое изменение содержаний, что дисперсия не изменяется при увеличении размера блоков. Поэтому наименьшие практические единицы отработки приходится комбинировать в последовательном порядке для получения больших блоков. Если изменчивость в большинстве единиц оказывается много меньше, чем изменчивость практически наименьшего блока, то эти малые блоки будут иметь дисперсию, не превосходящую дисперсию больших комбинированных блоков. Теоретическая кривая зависимости дисперсии от объема выборки указана на рис, 5.61: она иллюстрирует предположение, что дисперсия устойчива при превышении минимального критического размера пробы.

Чтобы определить скользящее среднее для средних значений блоков, мы должны сначала определить веса, соответствующие блокам. Необходимо задать также размещение самих блоков, но оно является более или менее произвольным. Предположим, что мы решили воспользоваться планом скользящего среднего, представленным на рис. 5.62. Мы хотим получить оценку Y содержаний компонента в руде, которая будет добыта из блока 1 на основании разведочных проб, взятых в блоках 1–9. (Заметим, что Y – оценка для центра блока, а $\overline{Y}_i$ – среднее значение разведочных проб в том же блоке.) Задав уравнение, с помощью которого вычисляются оценки для $\overline{Y}$, мы будем передвигать

схему опробования на неизвестную область, получая при этом оценки содержания в последовательности примыкающих блоков до тех пор, пока не покроем всю площадь.

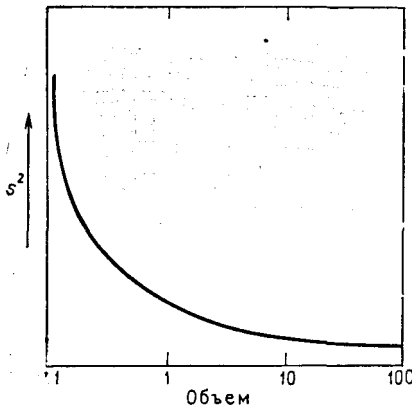


Рис. 5.61. Теоретическое изменение дисперсии содержаний полезного компонента в зависимости от изменений объема пробы или разрабатываемого блока

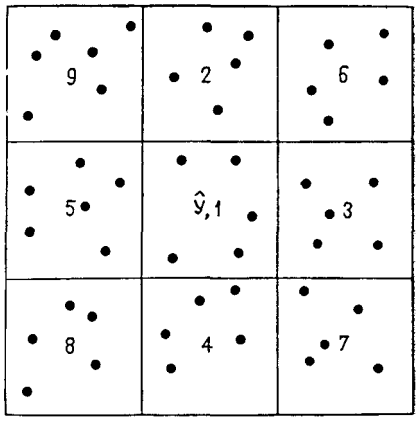


Рис. 5.62. Схема скользящего среднего для оценки содержаний полезного компонента в центре блока $\overline{Y}$ по средним для буровых скважин в блоках 1-9

Уравнение скользящего среднего определяется по значениям содержаний рудного компонента, извлеченного из горной выработки в изучаемом регионе. Предположим, что наш план скользящего среднего нанесен на карту рудного тела, которое уже выработано. Используя полученные пробы и количественные анализы по каждому из блоков, можно вычислить средние значения для блоков, которым соответствуют наши переменные $\overline{Y}_1$, $\overline{Y}_2$, ..., $\overline{Y}_9$. Записи добычи дают количество полезного компонента, в действительности добытое из интересующего нас блока. Если обозначить его через Y, то можно связать содержание в добываемой руде с количественными анализами разработки с помощью следующего линейного уравнения:

$$Y = \beta_0 + \sum_{i=1}^{9} \beta_i \overline{Y}_i + \varepsilon \quad (5.70)$$

которое, очевидно, является уравнением регрессии Y относительно Y_1, ..., Y_9. Применяя эту схему последовательно ко многим блокам, получим множество наблюдений Y. Это даст возможность решить (5.70) с помощью методов наименьших квадратов, почти в точности совпадающих с теми, которые использовались для построения поверхностей тренда.

Теперь у нас есть оценка Y содержания полезного компонента в центральном блоке, которая основана на средних пробах, взятых из окружающих блоков. Прогнозирующее уравнение содержит постоянный член и девять весовых коэффициентов. Однако можно определить постоянную β_0 как произведение новой постоянной на среднее по всем полученным пробам для всех блоков:

$$\beta_0 = \beta_0' \overline{\overline{Y}} \qquad \text{или} \qquad \beta_0' = \frac{\beta_0}{\overline{\overline{Y}}} \qquad (5.71)$$

Новая постоянная β'_0 находится простым делением β_0 на обобщенное среднее по всем пробам. Под термином «обобщенное среднее» мы подразумеваем

$$\overline{\overline{Y}} = \frac{1}{9} \sum_{i=1}^{9} \overline{Y}_i .$$

Таким образом, уравнение регрессии может быть записано так:

$$\overline{Y} = \beta_0' \overline{\overline{Y}} + \sum_{i=1}^{9} \beta_1 \overline{Y}_i + \varepsilon \qquad (5.72)$$

Все коэффициенты уравнения регрессии, включая постоянный член, можно теперь считать весовыми функциями. Приравнивая β_0 к $\beta_0' \overline{\overline{Y}}$, получим регрессию, не зависящую от какого-либо тренда, который имеется в данных, и уравнение будет применимо к другим площадям, даже несмотря на то, что среднее значение по этим новым площадям будет выше или ниже, чем в области, для которой уравнение было построено.

Теперь можно использовать уравнение регрессии в качестве скользящего среднего для неотработанной части месторождения. Поместив схему на карту разработки, нужно будет только взять пробы в каждом из девяти блоков. Используя значения количественных анализов, можно вычислить значения для каждого блока, которые затем станут $\overline{Y}$-ами в (5.72) и будут использованы для оценки значения Y. Затем эта схема применяется последовательно к различным блокам в неразрабатываемой части месторождения, давая оценки содержаний для разведки и подсчета запасов.

Успех применения метода зависит от двух условий. Во-первых, уравнение регрессии должно точно выражать содержание полезного компонента в уже освоенной части месторождения. Если качество аппроксимации с помощью линейной регрессионной модели плохое, то скользящее среднее Y в применении к неразработанным частям месторождения будет малоэффективной оценкой для Y. Однако благодаря устойчивости дисперсий в пределах больших блоков обыкновенно удается получить хорошую аппроксимацию. Плохой подбор указывает на изменчивость, масштаб которой сравним с размерами самих блоков. Второе важное условие состоит в том, что распределение значений в отработанной части месторождения очень сильно напоминает распределение в неотработанной части, хотя для оценок часто приходится использовать уравнения, построенные по данным совершенно другого рода. Однако оценка месторождений – это динамический процесс, и техника оценки совершенствуется по мере развития науки. Обычно же площадь, по которой строится прогнозное уравнение, выбирается от прогнозируемой настолько близко, насколько это возможно.

Как вы, вероятно, догадались, схемы взвешивания скользящих блоков являются «изготовленными по заказу» рецептами для данного месторождения или даже для его части. Выбор схемы скользящего среднего определяется наличием сильного тренда значений, применяемой системой разработки месторождения, а также многими другими факторами. По этой причине мы даже не пытаемся написать программу скользящего среднего общего назначения, хотя это относительно простая задача. Развитие этого метода и примеры его приложения к большим месторождениям даны Криге [46].

КРАЙГИНГ

Понятие регионализованной переменной было введено в гл. 4 как естественная характеристика, промежуточная между полностью случайной и полностью детерминированной переменными. Многие геологические поверхности, как существующие, так и воображаемые, можно рассматривать как регионализованные переменные. Эти поверхности непрерывны от точки к точке и, следовательно, могут коррелироваться на коротких расстояниях. Однако точки на нерегулярной поверхности,. которые отстоят далеко друг от друга, являются статистически независимыми. Степень пространственной непрерывности регионализованной переменной может быть выражена вариограммой, как это было показано в гл. 4. Если измерения были сделаны в рассеянном множестве точек опробования и форма вариограммы известна, то можно оценить значение поверхности в любой точке, не принадлежащей выборке. Процедура оценки называется крайгингом в честь Д. Г. Криге, южноафриканского горного инженера и пионера применения статистических методов при подсчете запасов.

Крайгинг можно использовать для построения карт в изолиниях, но, в отличие от обычных алгоритмов оконтуривания, он имеет статистически оптимальные свойства. Возможно, наиболее важным является то, что этот метод обеспечивает измерение ошибки или неопределенности поверхности изображаемой изолиниями. Крайгинг использует информацию из полувариограммы для нахождения оптимального множества весов, для оценки поверхности в точках, отличных от точек опробования. Так как полувариограмма является функцией расстояния,. то веса изменяются в соответствии с географическим положением точек опробования.

Точечный крайгинг

Точечный Крайгинг – простейшая форма крайгинга, в котором наблюдения состоят из измерений, взятых в безразмерных точках и оценки проводятся в других местах, которые сами также являются безразмерными точками. Точечный крайгинг используется, например, в построении карты в изолиниях для наблюдений, являющихся абсолютными отметками кровли формации, измеренными в ряде разведочных скважин. Построение структурной карты в изолиниях требует, чтобы оценки абсолютных отметок кровли формации были сделаны в близко расположенных точках на картируемой площади. Проделав это, можно провести изолинии через эти оценки так, как описано в предыдущем разделе.

Для упрощения задачи можно допустить, что картируемая переменная статистически стационарна или свободна от дрифта. Значение в точке, не принадлежащей выборке, может быть оценено как взвешенное среднее известных наблюдений, т.е. значение в точке p основано на ограниченном множестве близлежащих контрольных точек:

$$\overline{Y}_p = \sum W_i Y_i \, .$$

Следует ожидать, что оценка &&Y_p будет отличаться от истинного (но неизвестного) значения $\overline{Y}_p$ на величину, которую можно назвать ошибкой оценки:

$$\varepsilon_p = \left(\overline{Y}_p - Y_p\right) \qquad (5.73)$$

Если сумма весов, используемых в оценке, равна единице, то полученная оценка называется несмещенной при условии, что дрифта нет. Это значит, что для большого множества оценок средняя ошибка будет равна нулю, так как положительные и отрицательные отклонения взаимно компенсируют друг друга. Однако даже если средняя ошибка оценки оказывается нулевой, оценки могут быть широко рассеянными относительно истинных значений. Это рассеяние можно охарактеризовать дисперсией ошибки:

$$s_\varepsilon = \frac{1}{n}\sum\left(\overline{Y}_p - Y_p\right)^2 \qquad (5.74)$$

или после извлечения квадратного корня стандартной ошибкой оценки:

$$s_\varepsilon = \sqrt{s_\varepsilon^2} \qquad (5.75)$$

$$\overline{Y_p} = W_1 Y_1 + W_2 Y_2 + W_3 Y_3 \, .$$

Как уже отмечалось в разделе по картированию в изолиниях, интуитивно представляется правдоподобным, что ближайшие контрольные точки оказываются наиболее влияющими на оценку значения в точке поверхности, не являющейся точкой опробования, и что более удаленные контрольные точки оказывают меньшее влияние. Мы также вправе ожидать, что используемые веса в процессе оценивания и ошибки оценок некоторым образом должны быть связаны с полувариограммой поверхности. В простом примере Кларк [13] показывает, что это так.

Предположим, что требуется оценить значение Y в точке p по трем близким точкам, используя в качестве оцениваемого параметра взвешенное среднее трех известных значений:

Веса подчинены условию, что сумма их равна единице, поэтому в отсутствие тренда оценка является несмещенной. Предположим, что вес W_1 выбран равным 1,0. Тогда веса W_2 и W_3 должны быть равны нулю, и оценка в точке p есть

$$\overline{Y_p} = 1{,}0 Y_1 + 0{,}0 Y_2 + 0{,}0 Y_3$$

или

$$\overline{Y_p} = Y_1$$

Очевидно, ошибка оценки есть просто $\varepsilon = Y_p - Y_1$, так как Y_1 есть оценка $\overline{Y}_p$. Если многие другие значения, подобные Y_p, оцениваются по точкам, размещенным в пространстве подобно Y_1, то оценка дисперсии может быть вычислена как средняя квадратичная разность между этими парами точек. Для удобства можно обозначить эти оцененные значения через Y_{pi} и другие оценки через Y_{1i}. Тогда

$$s_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{pi} - Y_{1i}\right)^2 .$$

Если это уравнение сравнить с (4.71), то ясно, что оценка дисперсии равна удвоенной полувариограмме для расстояния, равного интервалу, разделяющему точки Y_{pi} и $Y_{1\text{i}}$.

В данном случае выбрана одна частная комбинация весов для получения оценки *Yp* и для определения ошибки оценки. Имеется бесконечное множество способов комбинирования весов, которые должны быть выбраны, и каждый из них дает различную оценку и различную ошибку оценки. Однако имеется только одна комбинация, которая будет давать минимум ошибки оценивания. Именно эту комбинацию весов и позволяет найти крайгинг.

Получение уравнений крайгинга требует вычислений и здесь не рассматривается. Простое изложение этого вопроса содержится в [13], а полный вывод уравнений дан Олеа [62] для случая точечного крайгинга. Оптимальные значения для весов можно найти решением системы совместных уравнений, которые включают значения из вариограммы оцениваемой переменной. Эти веса оптимальны в том смысле, что окончательные оценки являются несмещенными и имеют минимальную оценку дисперсии. Никакая другая линейная комбинация наблюдений не может дать оценки, которые имеют меньшее рассеяние относительно их истинных значений.

В простейших случаях задача крайгинга состоит в оценке значения Y в точке p по трем известным наблюдениям. Для ее нахождения требуется решить систему трех уравнений:

$$W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) = \gamma(h_{1p})$$
$$W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) = \gamma(h_{2p})$$
$$W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) = \gamma(h_{3p})$$

Здесь $\gamma(h_{\text{ij}})$ – полувариограмма на расстоянии h, соответствующем интервалу между контрольными точками i и j. Например, $\gamma(h_{13})$ – полувариограмма для расстояния между известными точками 1 и 3; $\gamma(h_{1\text{p}})$ – полувариограмма для расстояния между известной точкой 1 и точкой p, в которой производится оценка. Матрица в левой части системы симметрична, так как $h_{\text{ij}} = h_{\text{ji}}$. Диагональные элементы этой матрицы равны нулю, так как h_{ii} представляет расстояние точки от себя самой, которое равно нулю. В предположении, что полувариограмма проходит через начало координат, полувариограмма для нулевого расстояния равна нулю. Значения полудисперсии взяты из полувариограммы, которая должна быть известна или оценена до крайгинга.

Однако, для того чтобы обеспечить несмещенность решения, необходимо наложить ограничение на веса: их сумма должна быть равна единице. Четвертое уравнение

$$W_1 + W_2 + W_3 = 1{,}0.$$

В итоге получается набор четырех уравнений для трех неизвестных. Так как уравнений больше, чем неизвестных, то для того чтобы обеспечить минимально возможную ошибку оценки, нужно использовать дополнительные степени свободы. Это делается добавлением в систему уравнений немой переменной λ, называемой множителем Лагранжа. Полная система уравнений имеет следующий вид

$$\begin{aligned}
W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) + \lambda &= \gamma(h_{1p})\\
W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) + \lambda &= \gamma(h_{2p})\\
W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) + \lambda &= \gamma(h_{3p})\\
W_1 + W_2 + W_3 + 0 &= 1
\end{aligned} \qquad (5.76)$$

или в матричной форме

$$\begin{bmatrix}
\gamma(h_{11}) & \gamma(h_{12}) & \gamma(h_{13}) & 1\\
\gamma(h_{12}) & \gamma(h_{22}) & \gamma(h_{23}) & 1\\
\gamma(h_{13}) & \gamma(h_{23}) & \gamma(h_{33}) & 1\\
1 & 1 & 1 & 0
\end{bmatrix} \times \begin{bmatrix} W_1\\ W_2\\ W_3\\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(h_{1p})\\ \gamma(h_{2p})\\ \gamma(h_{3p})\\ 1 \end{bmatrix} \qquad (5.77)$$

В общем виде нужно решить матричное уравнение

$$[A] \bullet [W] = [B]$$

для вектора неизвестных коэффициентов $[W]$. Члены матрицы $[A]$ и вектора $[B]$ берутся непосред-

ственно из полувариограммы или из математических функций, описывающих ее вид.

Определив неизвестные веса, значение оцениваемой переменной в точке p представим в виде

$$\overline{Y}_p = W_{1_1} Y_1 + W_2 Y_2 + W_3 Y_3 . \qquad (5.78)$$

Оценка дисперсии имеет вид

$$s_\varepsilon^2 = W_1 \gamma (h_{1p}) + W_2 \gamma (h_{2p}) + W_2 \gamma (h_{2p}) + \lambda \quad (5.79)$$

Иными словами, дисперсия оценки есть в сущности взвешенная сумма полудисперсий для расстояний до точек, использованных в оценивании, плюс вклад от коэффициента K, который эквивалентен постоянному члену. Крайгинг имеет два больших преимущества перед обычными процедурами оценивания, которые используются при построении карт в изолиниях. Оценки процедур крайгинга в среднем имеют наименьшую возможную ошибку, и также обеспечивают явное выражение величины этой ошибки.

Для иллюстрации точечного крайгинга мы приведем оценку уровня воды в точке p на карте, представленной на рис. 5.63. Оценка будет проведена по известным уровням, измеренным в трех наблюдательных скважинах. Координаты карты скважин и расстояния между ними приведены в табл. 5.12. Предварительный структурный анализ позволил получить линейную полувариограмму (рис. 5.64). Значения полудисперсии, соответствующие расстоянию между скважинами, даны в табл. 5.12; они могут быть получены прямо из полувариограммы или вычислены из наклона.

Уравнения, которые должны быть решены для определения весов, в этом примере имеют вид

$$W_1\gamma(0) + W_2\gamma(12{,}2) + W_3\gamma(11{,}5) + \lambda = 4{,}0$$
$$W_1\gamma(12{,}2) + W_2\gamma(0) + W_3\gamma(18{,}1) + \lambda = 12{,}1$$
$$W_1\gamma(11{,}5) + W_2\gamma(18{,}1) + W_3\gamma(0) + \lambda = 7{,}9$$
$$W_1 + W_2 + W_3 + 0 = 1{,}0$$

или в матричной форме

$$\begin{bmatrix} 0 & 12{,}2 & 11{,}5 & 1{,}0 \\ 12{,}2 & 0 & 18{,}1 & 1{,}0 \\ 11{,}5 & 18{,}1 & 0 & 1{,}0 \\ 1{,}0 & 1{,}0 & 1{,}0 & 0 \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 4{,}0 \\ 12{,}1 \\ 7{,}9 \\ 1{,}0 \end{bmatrix}$$

Матрицу, обратную к матрице левой части, можно найти с помощью метода, изложенного в гл. 3 (см. кн.1), но в целях упрощения вычислений целесообразно поменять порядок уравнений для того, чтобы избежать нулей на диагонали. Обратная матрица есть

$$\begin{bmatrix} -0{,}0680 & 0{,}0326 & 0{,}0354 & 0{,}1932 \\ 0{,}0326 & -0{,}0433 & 0{,}0106 & 0{,}4072 \\ 0{,}0354 & 0{,}0106 & -0{,}0461 & 0{,}3995 \\ 0{,}1932 & 0{,}4072 & 0{,}3995 & -9{,}5851 \end{bmatrix}$$

Теперь можно найти неизвестные веса умножением справа на транспонированную матрицу вектора правой части, состоящего из полудисперсий. В результате получим

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0{,}5954 \\ 0{,}0975 \\ 0{,}3071 \\ -0{,}7298 \end{bmatrix}$$

оценка уровня воды в точке p находится подстановкой подходящих весов в линейное уравнение (5.78):

$$\overline{Y}_p = 0{,}5954(120) + 0{,}975(103) + 0{,}3071(142) = 125{,}1 \text{ м.}$$

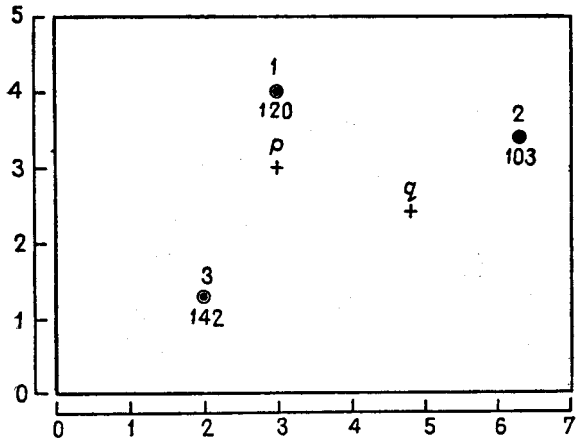


Рис. 5.63. Карта уровней воды (в м) в трех наблюдательных скважинах. Оценки уровня воды сделаны в точках p и q. Координаты указаны (в км) для произвольного начала

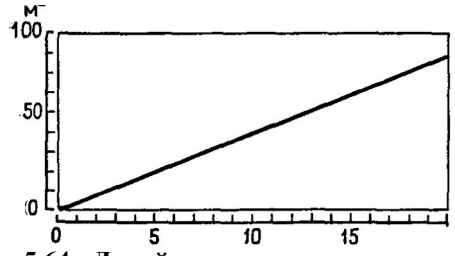


Рис. 5.64. Линейная полувариограмма уровней воды на площади, включающей карту рис. 5.63. Полувариограмма имеет наклон 4,0 м^2/км внутри 20-километровой зоны.

Таблица 5.12

Наблюдения, проведенные в скважинах, используемых для оценки уровня воды в точке p				
Скважина	Координата X_1	Координата X_2	Уровень воды	
1	3,0	4,0	120	
2	6,3	3,4	103	
3	2,0	1,3	142	
Точка p	3,0	3,0		
Расстояния между скважинами и точкой p				
Скважина	1	2	3	P
1	0	3,35	2,88	1,00
2		0	4,79	3,32
3			0	1,97
Полудисперсии для расстояний между скважинами и точкой p				
Скважина	1	2	3	p
1	0	13,42	11,52	4,00
2		0	19,14	13,30
3			0	7,89

Аналогично находится и ошибка дисперсии во взвешенной сумме полувариограмм для расстояний от контрольных точек до оцениваемой точки. В матричных обозначениях $s_\varepsilon^2=[W][B]$:

$$s_\varepsilon^2=0{,}5954(4) + 0{,}0975(12{,}1) + 0{,}3071(7{,}9) - 0{,}7298(1) = 5{,}25 \text{ м}^2.$$

Стандартная ошибка оценки есть просто квадратный корень из оценки дисперсии или $s_\varepsilon = \sqrt{5{,}25}=2{,}3$ м. Если мы предположим, что ошибки оценивания распределены нормально относительно истинного среднего значения, то мы можем использовать стандартную ошибку для определения доверительного интервала этой оценки. Вероятность того, что истинный уровень воды в точке p находится в пределах одной стандартной ошибки выше или ниже оцениваемого значения, равна 68%, а вероятность того, что истинный уровень лежит в пределах двух стандартных ошибок, равна 95%. Иными словами, уровень воды в точке должен быть $Y_p = 125{,}1 \pm 4{,}6$ м с вероятностью 95%. В каждой точке этой карты мы можем оценить уровень воды и можем также определить стандартные ошибки этих оценок. Из них мы можем построить две карты; первая основана на самих оценках и является наилучшим образом предсказанной конфигурацией картируемых переменных, вторая – это ошибка карты, показывающая доверительную обертывающую поверхность, которая окружает оцениваемую поверхность; она выражает относительную надежность первого отображения. На площа-

дях слабого контроля ошибка карты может принимать большие значения, показывая, что оцениваемый параметр подвергается большой изменчивости. На площадях слабого контроля ошибка карты будет показывать низкие значения, и в самих контрольных точках ошибка оценки будет равна нулю.

Система уравнений, используемая для нахождения весов крайгинга, должна решаться для каждой оцениваемой точки до тех пор, пока пробы расположены по регулярной схеме так, что расстояния между точками остаются одинаковыми. Если мы пожелаем оценить уровень воды в точке q на рис. 5.63, то необходимо рассмотреть расстояния между q и тремя наблюдаемыми скважинами. Эти расстояния и соответствующие полудисперсии, взятые из рис. 5.64, следующие:

$$\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}2{,}4\\1{,}6\\3{,}0\end{bmatrix}; \qquad \begin{bmatrix}9{,}6\\6{,}2\\12{,}0\end{bmatrix}$$

Так как расположение наблюдаемых скважин остается тем же самым, то все расстояния между ними одинаковы и левая часть системы совместных уравнений неизменна. Обратная матрица тоже не изменяется. Поэтому, умножив ее на новый вектор полудисперсий, мы получим веса для оценки уровня воды в точке q. Новое множество весов таково:

$$\begin{bmatrix}W_1\\W_2\\W_3\\\lambda\end{bmatrix}=\begin{bmatrix}0{,}1676\\0{,}5796\\0{,}2528\\-0{,}3711\end{bmatrix}$$

Оценим уровень воды $\overline{Y}_q$ и дисперсию $s_\varepsilon{}^2$:

$$\overline{Y}_q = 0{,}1676(120) + 0{,}5796(103) - 0{,}2528(142) = 115{,}7 \text{ м};$$

$$s_\varepsilon{}^2 = 0{,}1676(9{,}6) - 0{,}5796(6{,}3) + 0{,}2528(12{,}0) - 0{,}3711(1) = 7{,}91 \text{ м}^2.$$

Стандартная ошибка оценки в точке $q - s_\varepsilon = \sqrt{7{,}91} = 2{,}8$ м, поэтому уровень воды в этой новой точке может быть выражен в виде $Y_q = 115{,}7 \pm 5{,}6$ м с вероятностью 95%. Поверхность подземных вод в точке q ниже, чем в точке p, и стандартная ошибка больше, что отражает большое общее расстояние до контрольных скважин.

Если одна из контрольных точек изменяется, то некоторые из расстояний также изменяются, и система уравнений должна быть решена заново. На рис. 5.65 наблюдаемая скв. 2А была пробурена в точке, более близкой к точке p, и для регионализованной переменной, характеризующей уровень воды, было замерено значение 115 м. Новые расстояния между точками и соответствующие полудисперсии приведены в табл. 5.13. Система уравнений теперь имеет вид

$$\begin{bmatrix}0 & 7{,}2 & 11{,}5 & 1{,}0\\7{,}2 & 0 & 8{,}4 & 1{,}0\\11{,}5 & 8{,}4 & 0 & 1{,}0\\1{,}0 & 1{,}0 & 1{,}0 & 0\end{bmatrix}\times\begin{bmatrix}W_1\\W_2\\W_3\\\lambda\end{bmatrix}=\begin{bmatrix}4{,}0\\4{,}0\\7{,}9\\1{,}0\end{bmatrix}$$

а решение этой системы

$$\begin{bmatrix}W_1\\W_2\\W_3\\\lambda\end{bmatrix}=\begin{bmatrix}0{,}4545\\0{,}3858\\0{,}1598\\-0{,}6001\end{bmatrix}$$

Новая оценка (в м) уровня воды в точке p, основанная на информации по скв. 2А, имеет следующий вид:

$$\overline{Y}_p = 0{,}4545(120) + 0{,}3858(115) + 0{,}1598(142) = 121{,}6$$

Дисперсия этой новой оценки (в м^2)

$$s_\varepsilon{}^2 = 0{,}4545(4{,}0) + 0{,}3858(4{,}0) + 0{,}1598(7{,}9) - 0{,}6001(1) = 4{,}02.$$

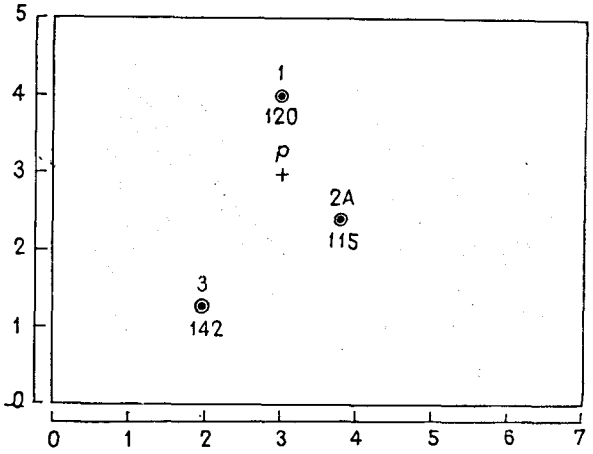


Рис. 5.65. Карта, показывающая уровни воды (в м) в трех наблюдательных скважинах. Скважина 2А ближе к оцениваемой точке p, чем скважина 2 на рис. 5.63

Таблица 5.13. Второй ряд наблюдений, проведенных в скважинах, используемых для оценки уровня воды в точке p

Наблюдения, проведенные в скважинах, используемых для оценки уровня воды в точке p				
Скважина	Координата X_1	Координата X_2	Уровень воды	
1	3,0	4,0	120	
2A	3,8	2,4	115	
3	2,0	1,3	142	
Точка p	3,0	3,0		
Расстояния между скважинами и точкой p				
Скважина	1	2A	3	P
1	0	1,79	2,88	1,00
2		0	2,11	1,00
3			0	1,97
Полудисперсии для расстояний между скважинами и точкой p				
Скважина	1	2A	3	p
1	0	7,16	11,52	4,00
2		0	8,44	4,00
3			0	7,89

Стандартная ошибка (в м) в точке p теперь будет равна $s_\varepsilon^2 = \sqrt{4{,}02} = 2{,}0$, что несколько ниже, чем было найдено на основе наблюдений скважины 2, а не 2А. Это иллюстрирует тот факт, что ошибки оценки уменьшаются, если контрольные точки располагаются ближе к оцениваемой точке.

Предположим, что одна из контрольных точек совпадает с точкой, в которой производится оценка. Тогда одно из значений правой части матричного уравнения равно нулю; оставшиеся значения становятся равными некоторым из значений матрицы, стоящей в левой части. Предполагая, что наблюдаемая скв. 2В пробурена в точке p, можно определить эффект от этой замены в данном примере: оцененный уровень воды оказывается равным 125 м. Расстояние между любой точкой i и точкой 2В теперь то же, что и расстояние между любой точкой i и точкой p. Аналогично полудисперсии будут теми же, и система уравнений примет вид

$$\begin{bmatrix} 0 & 4{,}0 & 11{,}5 & 1{,}0 \\ 4{,}0 & 0 & 7{,}9 & 1{,}0 \\ 11{,}5 & 7{,}9 & 0 & 1{,}0 \\ 1{,}0 & 1{,}0 & 1{,}0 & 0 \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 4{,}0 \\ 0{,}0 \\ 7{,}9 \\ 1{,}0 \end{bmatrix}$$

Можно вычислить вектор весов и, как и следует ожидать,

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0{,}0000 \\ 1{,}0000 \\ 0{,}0000 \\ 0{,}0000 \end{bmatrix}$$

Если эти веса использовались для оценки точки p, то оцениваемая отметка в точности равна измеренному значению уровня воды (в м) в скв. 2В:

$$\overline{Y}_p = 0{,}0000(120) + 1{,}0000(115) + 0{,}0000(142) = 125{,}0$$

Значит, как и следовало ожидать,

$$s_\varepsilon{}^2 = 0{,}0000(4{,}0) + 1{,}0000(0) + 0{,}0000(7{,}9) + 0{,}0000(1) = 0{,}00 \text{ м}^2.$$

Это объясняет, что подразумевается под часто используемой характеристикой крайгинга как точного интерполяционного метода; он действительно позволяет предсказать измеренные значения в известных точках и делает это с ошибкой, равной нулю. Конечно, мы обычно не занимаемся оценкой в точках, уже известных, но это может понадобиться в случае, если точечный крайгинг будет использоваться для построения карты в изолиниях. Если случится, что какая-либо из контрольных точек совпадает с узлом сети, то крайгинг даст правильные, свободные от ошибок значения. Мы также должны быть уверены, что оцениваемая поверхность пройдет в точности через все контрольные точки и что доверительные области вокруг оценки поверхности проходят через нуль в контрольных точках.

С целью максимального упрощения математических выкладок в этих примерах предполагается, что каждая оценка сделана на основе трех контрольных точек. На практике при получении оценок приходится использовать больше точек, а возможно, значительно больше. Каждую контрольную точку в оценке можно взвесить, и определение каждого веса требует решения своей системы уравнений. Большинство программ для получения оценки в каждом узле сети для построения карты в изолиниях использует 16 контрольных точек или более, что приводит к необходимости решать по меньшей мере 17 систем уравнений для каждой точки. Таким образом, использование крайгинга для построения карт в изолиниях приводит к очень трудоемким вычислениям.

В теории число точек, необходимых для получения оценки в точке, изменяется с изменением локальной плотности контроля. Для получения оценки в точке должны быть учтены все контрольные точки, расположенные в окрестности этой точки. На практике многие из этих точек оказываются избыточными, и их применение лишь незначительно улучшает оценку. При использовании крайгинга для целей картирования в изолиниях следует руководствоваться практическим правилом, которое ограничивает число действительно необходимых контрольных точек, в пределах зоны влияния или в окрестности. Оптимальное число контрольных точек определяется полувариограммой и пространственной схемой расположения точек [63]. Структурный анализ, таким образом, играет двойную роль в крайгинге: он обеспечивает получение полувариограммы, необходимой для построения системы уравнений крайгинга, и также позволяет определить размер окрестности, внутри которой для получения каждой оценки выбираются контрольные точки.

Универсальный крайгинг

Важное свойство точечного крайгинга состоит в том, что он перестает работать в случае, когда картируемая регионализованная переменная не является стационарной. В присутствии тренда или медленного изменения среднего значения линейная оценка не будет несмещенной. Вычисленные оценки будут систематически сдвигаться вверх и вниз от истинных значений, зависящих от размещения контрольных точек и направления наклона поверхности.

Выражаясь языком геостатистики, нестационарная регионализованная переменная рассматривается как состоящая из двух компонент. Дрифт состоит из среднего или ожидаемого значения регионализованной переменной в пределах окрестности и медленно изменяется, представляя нестационарную часть поверхности. Остаток представляет собой разность между действительными изменениями и дрифтом. Очевидно, если из регионализованной переменной устранить дрифт, то остатки будут стационарными, и к ним можно применить крайгинг. Таким образом, универсальный край-

гинг можно считать состоящим из трех операций: первая – оценка и устранение дрифта; затем стационарные остатки крайгируются с целью получения необходимых оценок. Наконец, оцененные остатки комбинируются с дрифтом с целью получения истинной поверхности.

Дрифт аналогичен поверхности тренда, исключая случай, когда оценка дрифта основана только на контрольных точках внутри окрестности оцениваемой точки. В общем случае каждая окрестность каждой оцениваемой точки характеризуется различным расположением контрольных точек, поэтому уравнение, определяющее дрифт, должно быть решено столько раз, сколько имеется точек. Это не так трудно, как кажется, так как уравнения крайгинга должны быть решены для каждой из этих точек, и эти две операции можно комбинировать.

Определим дрифт как некоторую произвольную функцию координат контрольных точек, например многочлен низкого порядка. Дрифт M в точке p может быть определен как многочлен либо первого (уравнение 5.80), либо второго (уравнение 5.81) порядка:

$$M_{\text{p}} = \alpha_1 X_{1\text{i}} + \alpha_2 X_{2\text{i}} \quad (5.80)$$

$$M_{\text{p}} = \alpha_1 X_{1\text{i}} + \alpha_2 X_{2\text{i}} + \alpha_3 X^2_{1\text{i}} + \alpha_4 X_{1\text{i}} X_{2\text{i}} + \alpha_5 X^2_{2\text{i}} \quad (5.81)$$

Пусть $X_{1\text{i}}$ и $X_{2\text{i}}$ – географические координаты i-й контрольной точки в пределах окрестности и α – неизвестные коэффициенты дрифта, которые должны быть найдены. Однако прежде чем это делать, необходимо провести структурный анализ с целью определения наилучшей комбинации размера окрестности и выражения дрифта. Как отмечалось в разделе, посвященном полувариограммам, это нетривиальная задача, так как модель дрифта и размер окрестности взаимозависимы.

Выражения для дрифта в качестве дополнительных ограничений можно ввести в систему уравнений, используемую для нахождения весов крайгинга. Решая эту расширенную систему уравнений, мы получим оценки весов крайгинга, которые включают эффект от заданного дрифта в пределах данной локальной окрестности. Выражения дрифта связывают географические координаты каждой контрольной точки с географическими координатами крайгируемой точки. Форма модели дрифта, размер окрестности и форма полувариограммы остатков от дрифта тесно связаны друг с другом. Это значит, что дисперсия остатков частично зависит от несколько произвольного задания дрифта. В универсальном крайгинге должны быть определены веса, приписываемые контрольным точкам, а также коэффициенты дрифта. Так как оценивается большее число членов,. то в пределах рассматриваемой окрестности следует использовать выборки контрольных точек большего объема. Простейший пример представляет крайгинг некоторой точки, имеющий линейный дрифт и линейную полувариограмму остатков от дрифта. Линейная модель дрифта, заданная уравнением (5.80), имеет два коэффициента, поэтому в процессе оценки дрифта должны быть использованы минимум три точки, или же мы выйдем за рамки степеней свободы.

Если мы хотим оценить методом универсального крайгинга. как дрифт, так и регионализованную переменную, то для обеспечения необходимых степеней свободы при оценке коэффициентов крайгинга потребуются дополнительные контрольные точки. В противном случае процесс крайгинга приведет к одинаковым оценкам как для дрифта, так и для крайгируемой поверхности. Целесообразно выбрать пять контрольных точек, три из которых дают степени свободы для определения дрифта, а дополнительные две – степени свободы для оценки, самой поверхности.

Пусть координаты контрольной точки с номером i обозначаются через $X_{1\text{i}}$ (направление с востока на запад) и $X_{2\text{i}}$ (.направление с севера на юг). Мы должны определить множество из пяти весов и коэффициенты единственного ограничения (сумма весов равна 1), плюс еще двух ограничений для линейного дрифта. Для этого требуется следующая система из восьми уравнений:

$$W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) + W_4\gamma(h_{14}) + W_5\gamma(h_{15}) + \lambda + \alpha_1 X_{11} + \alpha_2 X_{21} = \gamma(h_{1\text{p}})$$
$$W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) + W_4\gamma(h_{24}) + W_5\gamma(h_{25}) + \lambda + \alpha_1 X_{12} + \alpha_2 X_{22} = \gamma(h_{2\text{p}})$$
$$W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) + W_4\gamma(h_{34}) + W_5\gamma(h_{35}) + \lambda + \alpha_1 X_{13} + \alpha_2 X_{23} = \gamma(h_{3\text{p}})$$
$$W_1\gamma(h_{14}) + W_2\gamma(h_{24}) + W_3\gamma(h_{34}) + W_4\gamma(h_{44}) + W_5\gamma(h_{45}) + \lambda + \alpha_1 X_{14} + \alpha_2 X_{24} = \gamma(h_{4\text{p}})$$
$$W_1\gamma(h_{15}) + W_2\gamma(h_{25}) + W_3\gamma(h_{35}) + W_4\gamma(h_{45}) + W_5\gamma(h_{55}) + \lambda + \alpha_1 X_{15} + \alpha_2 X_{25} = \gamma(h_{5\text{p}})$$
$$W_1 + W_2 + W_3 + W_4 + W_5 + 0 + 0 + 0 = 1$$
$$W_1 X_{11} + W_2 X_{12} + W_3 X_{13} + W_4 X_{14} + W_5 X_{15} + 0 + 0 + 0 = X_{1\text{p}}$$
$$W_1 X_{21} + W_2 X_{12} + W_3 X_{13} + W_4 X_{24} + W_5 X_{25} + 0 + 0 + 0 = X_{2\text{p}}$$

или в матричной форме

$$\begin{bmatrix} \gamma(h_{11}) & \gamma(h_{12}) & \gamma(h_{13}) & \gamma(h_{14}) & \gamma(h_{15}) & 1 & X_{11} & X_{21} \\ \gamma(h_{12}) & \gamma(h_{22}) & \gamma(h_{23}) & \gamma(h_{24}) & \gamma(h_{25}) & 1 & X_{12} & X_{22} \\ \gamma(h_{13}) & \gamma(h_{23}) & \gamma(h_{33}) & \gamma(h_{34}) & \gamma(h_{35}) & 1 & X_{13} & X_{23} \\ \gamma(h_{14}) & \gamma(h_{24}) & \gamma(h_{34}) & \gamma(h_{44}) & \gamma(h_{45}) & 1 & X_{14} & X_{24} \\ \gamma(h_{15}) & \gamma(h_{25}) & \gamma(h_{35}) & \gamma(h_{45}) & \gamma(h_{55}) & 1 & X_{15} & X_{25} \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ X_{11} & X_{12} & X_{13} & X_{14} & X_{15} & 0 & 0 & 0 \\ X_{21} & X_{22} & X_{23} & X_{24} & X_{25} & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \\ \lambda \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \gamma(h_{1p}) \\ \gamma(h_{2p}) \\ \gamma(h_{3p}) \\ \gamma(h_{4p}) \\ \gamma(h_{5p}) \\ 1 \\ X_{1p} \\ X_{1p} \end{bmatrix}$$

Дополнительная процедура, позволяющая упростить вычисления, состоит в переносе начала системы координат в крайгируемую точку. Тогда координаты X_{1p} и X_{2p} становятся равными нулю. Это приводит к изменению всех координат X_{1i} и X_{2i}, но не изменяет расстояний между точками, поэтому веса крайгинга остаются неизменными.

С целью демонстрации последовательности шагов при выполнении универсального крайгинга мы рассмотрим еще один пример, использующий данные по водоносному пласту в западном Канзасе. На рис. 5.66 изображены положения пяти наблюдательных скважин, которые будут использованы для получения оценки дрифта и крайгируемого значения уровня воды в точке p. Предположим теперь, что на рис. 5.64 представлена оценка полувариограммы остатков и что она линейна по форме с наклоном 4,0 м^2/км. Вся основная необходимая информация представлена в табл. 5.14, которая также включает необходимые полувариограммы.

Для оценки уровня воды в точке p должно быть решено уравнение

$$\begin{bmatrix} 0 & 13{,}4 & 11{,}5 & 7{,}2 & 8{,}9 & 1 & 0 & 1{,}0 \\ 13{,}4 & 0 & 19{,}1 & 10{,}8 & 21{,}3 & 1 & 3{,}3 & 0{,}4 \\ 11{,}5 & 19{,}1 & 0 & 8{,}4 & 7{,}9 & 1 & -1{,}0 & -1{,}7 \\ 7{,}2 & 10{,}8 & 8{,}4 & 0 & 11{,}5 & 1 & 0{,}8 & -0{,}6 \\ 8{,}9 & 21{,}3 & 7{,}9 & 11{,}5 & 0 & 1 & -2{,}0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 3{,}3 & -1{,}0 & 0{,}8 & 0 & 0 & 0 & 0 \\ 1{,}0 & 0{,}4 & -1{,}7 & -0{,}6 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \\ \lambda \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 4{,}0 \\ 13{,}3 \\ 7{,}9 \\ 4{,}0 \\ 8{,}0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Решая это уравнение, получаем восемь коэффициентов, первые пять из которых представляют собой веса крайгинга:

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \\ W_6 \\ W_7 \\ W_8 \end{bmatrix} = \begin{bmatrix} 0{,}4119 \\ -0{,}0137 \\ 0{,}0934 \\ 0{,}4126 \\ 0{,}0957 \\ -0{,}7245 \\ 0{,}0660 \\ 0{,}0229 \end{bmatrix}$$

Оценка уровня воды (в м) в точке p следующая:

$$\overline{Y} = 0{,}4119(120) - 0{,}0137(103) + 0{,}0934(142) + 0{,}4126(115) + 0{,}0957(148) = 122{,}9\,,$$

что лишь немногим отличается от результатов, полученных из трех наблюдений без предположения о дрифте. Оценка ошибки дисперсии может быть вычислена в точности так же, как если бы дрифт отсутствовал, т.е. с помощью умножения слева вектора правой части $[B]$ на транспонированный вектор решения $[W]$. Оценка дисперсии ошибки равна 8,1 м^2.

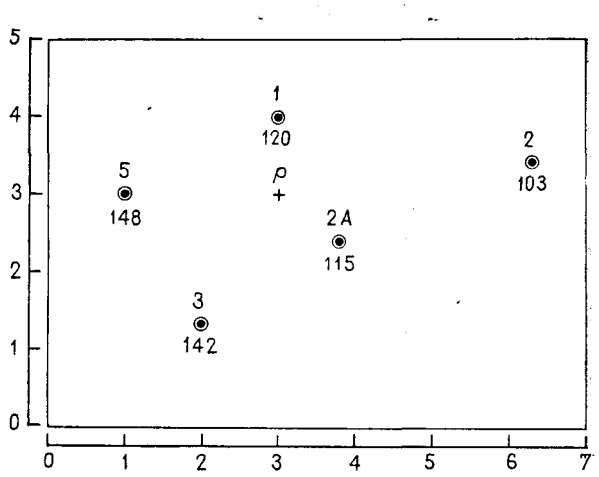


Рис. 5.66. Карта, на которой представлены уровни воды (в м) в 5 наблюдательных скважинах. Оценки уровней воды получены методом универсального крайгинга в точке p и в юго-западном углу карты

Таблица 5.14. Наблюдения, проведенные в скважинах, используемых для оценки уровня воды и дрифта в точке *p*

Скважина	Координата X_1	Координата X_2	Уровень воды
1	3,0	4,0	120
2	6,3	3,4	103
3	2,0	1,3	142
2A	3,8	2,4	115
5	1,0	3,0	148
Точка *p*			

Расстояния между скважинами и точкой *p*

Скважина	1	2	3	2A	5	*p*
1	0	3,35	2,88	1,79	2,24	1,00
2		0	4,79	2,69	5,32	3,32
3			0	2,11	1,97	1,97
2A				0	2,86	1,00
5					0	2,00

Полудисперсии для расстояний между скважинами и точкой *p*

Скважина	1	2	3	2A	5	*p*
1	0	13,42	11,52	7,16	8,94	4,00
2		0	19,14	10,77	21,26	13,30
3			0	8,44	7,89	7,89
2A				0	11,45	4,00
5					0	8,00

Этот пример показывает, что нет большого различия между простым точечным крайгингом и универсальным крайгингом с дрифтом, так как в этом примере эти две процедуры дают почти идентичные оценки. Однако точечный крайгинг, как и другие методы взвешенного усреднения, не может экстраполироваться за область влияния множества контрольных точек. Это значит, что большая часть оцененных значений будет лежать на наклонных участках поверхности, и точки наивысших и наинизших значений на поверхности обычно будут определены контрольными точками. Предположим, что мы оценили уровень воды в точке, причем оказалось, что поверхность вышла за пределы интервала, определяемого наблюдаемыми скважинами. Оказывается, что уровень воды изменяется с запада на восток, падая почти на 40 м между наблюдениями в скважинах 2 и 3. Если это падение продолжить, то мы можем ожидать уровни воды выше 142 м в точках западнее наблюдае-

мой скв. 3, и уровни ниже 103 м в точках восточнее скв. 2.

Мы можем оценить уровень воды в юго-западном углу карты в точке с координатами $X_1 = 0$, $X_2 = 0$. Сначала используем простой точечный крайгинг и множество из пяти скважин. Это дает следующую систему весов:

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \end{bmatrix} = \begin{bmatrix} -0{,}1221 \\ 0{,}0110 \\ 0{,}7523 \\ -0{,}0307 \\ 0{,}3895 \end{bmatrix}$$

Оценим уровень воды (в м):

$$\overline{Y} = 0{,}1221(120) + 0{,}0110(103) + 0{,}7523(142) - 0{,}0307(115) + 0{,}3895(148) = 174{,}4 .$$

Как следовало ожидать, оценка базируется полностью на ближайших скважинах, но находится в интервале, определенном наибольшим и наименьшим наблюденными значениями. Оценка дисперсии ошибки равна $s_\varepsilon^2 = 17{,}3$ м^2.

Предполагая дрифт первой степени, получаем коэффициенты универсального крайгинга

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \\ \lambda \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} -0{,}5594 \\ -0{,}3020 \\ 1{,}3133 \\ 0{,}1451 \\ 0{,}4030 \\ 26{,}3832 \\ -1{,}7940 \\ -4{,}1795 \end{bmatrix}$$

Используя эти веса, получаем оценку уровня воды (в м):

$$\overline{Y} = -0{,}5594(120) - 0{,}3020(103) + 1{,}3133(142) + 0{,}1451(115) + 0{,}4030(148) = 164{,}6 ,$$

что намного превышает наибольшее контрольное значение. Универсальный крайгинг предназначен для определения изменения уровня воды или дрифта в пределах локальной окрестности и проектирует его в крайгируемую точку. Оценка дисперсии ошибки равна $s_\varepsilon^2 = 26{,}8$ м^2, что много больше чем неопределенность как в оценке дрифта, так и в оценке самой регионализованной переменной.

Вычисление дрифта

Наблюдения, выбираемые для получения оценки крайгинга с некоторой точке p, располагаются в зоне влияния вокруг точки p. Крайгинг дает множество весов, которые позволяют приравнять взвешенную сумму полудисперсий между наблюдениями и полудисперсии между наблюдениями и точкой p. Предположим, однако, что все доступные наблюдения вышли за пределы области влияния. Полувариограммы γ_{ip}^2 между точкой p и некоторым удаленным наблюдением i будут идентичны для всех наблюдений и равны дисперсии регионализованной переменной (или, при наличии дрифта, равны дисперсии остатков от дрифта), т. е. первые n элементов вектора правой части системы крайгинга будут содержать значения S_o^2, а не значения γ_{ip}^2. Так как регионализованная переменная в точке p была статистически независимой от ее значений в каждой из точек наблюдения, то мы не могли предсказать локальное значение поверхности. Вместо этого оценка, которую мы получали, решая систему уравнений крайгинга, была основана на глобальных или усредненных свойствах регионализованной переменной. Иными словами, мы оценивали только дрифт.

Мы можем вычислить дрифт в точке p даже в том случае, если используемые наблюдения расположены в некоторой окрестности, меньшей чем область влияния точки p. Все, что необходимо,

– это замена полудисперсий в правой части уравнений крайгинга на полудисперсии, которые наблюдались бы в том случае, если бы точка p была расположена так далеко от контрольных точек, что была бы независимой от них. Так как полудисперсия для всех расстояний вне области влияния равна дисперсии остатков, то члены в правой части можно приравнять дисперсии остатков, S_o^2. К сожалению, мы снова зацикливаемся, так как мы не знаем значения S_o^2 до тех пор, пока дрифт не вычислен. К счастью, структурный анализ позволяет нам сделать априорную оценку дрифта, так как он равен полудисперсии в точке максимума или области влияния.

Следует отметить, что дрифт – это произвольная, но удобная конструкция, необходимая для того, чтобы удовлетворить требованиям стационарности регионализованной переменной. Может существовать много различных комбинаций модели дрифта, размера окрестности и оценки полувариограммы, которые удовлетворительно представят структуру регионализованной переменной. Выбор некоторой специфической комбинации зависит от степени доступности данных, удобства вычислений и других факторов. Продолжим рассмотрение нашего простого примера, предполагая, что дрифт в данных по уровню воды линеен и что полувариограмма остатков также линейна. Из структурного анализа можно заключить, что область влияния регионализованной переменной распространяется на 30 км. Так как наклон полувариограммы равен 4 м^2/км, то дисперсия вне области влияния должна быть около 4х30 или 120 м^2.

При вычислении дрифта в точке p правая часть матрицы крайгинга имеет следующий вид

$$[C]=\begin{bmatrix}120\\120\\120\\120\\120\\1\\0\\0\end{bmatrix}$$

При вычислении крайгинга левая часть $[A]$ остается неизменной, поэтому все, что требуется для оценки дрифта, – это умножить обратную к $[A]$ матрицу на $[C]$. Это дает множество из пяти весов M_i, которые используются для вычисления дрифта, плюс три постоянных члена, которые дают вклад в оценку ошибки дисперсии дрифта. Вектор решения $[M]$ есть

$$[M]=\begin{bmatrix}0{,}1311\\0{,}3702\\0{,}2202\\-0{,}1587\\0{,}4372\\109{,}2283\\-0{,}9048\\0{,}4935\end{bmatrix}$$

Дрифт (в м) в точке p находится умножением уровней в наблюдательных скважинах на подходящие коэффициенты дрифта и суммированием:

$$\overline{M}=0{,}1311(120)+0{,}3702(103)+0{,}2202(142)-0{,}1587(115)+0{,}4372(148)=131{,}6\,.$$

Ошибка оценки находится как дисперсия умножением слева $[C]$ на транспонированную матрицу $[M]$. Например, $s_m^2=229{,}3$ м^2. Как и ранее, стандартная ошибка оценки равна квадратному корню из оценки дисперсии, или $s_m=\sqrt{229{,}3}=15{,}1$ м.

Снова предполагая нормальность, можно сделать вероятностные утверждения о доверительном интервале для истинного значения дрифта. Например, вероятность того, что линейный дрифт лежит внутри интервала 131,6±30,2 м или между 199,0 и 259,4, равна 95%. (Необходимо напомнить,

что дрифт не имеет никакого физического смысла. Доверительный интервал просто указывает ошибку опробования и дает наиболее вероятное значение для дрифта регионализованной переменной в случае, если бы она подвергалась повторно опробованию, и дрифт вычислялся бы повторно.)

Для того, чтобы численные примеры этого параграфа оказались доступными, было сделано много упрощений.

Например, используемое число наблюдений является наименьшим из возможных. В действительных приложениях рассматривается значительно большее число контрольных точек, так как это увеличивает точность крайгинга и уменьшает ошибку оценки. Аналогично, предполагается линейность полувариограммы, так как это простейшая из возможных моделей, имеющая только один параметр. Строго линейная полувариограмма не имеет максимума, а может быть продолжена до бесконечных значений дисперсии. В действительности же мы имеем линейную полувариограмму вплоть до некоторого значения, и только после излома она становится постоянной. Армстронг и Жабин [3] указали, что полувариограммы, обладающие внезапными изменениями наклона, могут приводить к неустойчивым решениям и отрицательным дисперсиям. Для представления полувариограммы оказывается удобнее использовать непрерывную функцию, такую, как, например, сферическая модель, хотя это несколько усложняет вычисления. Неизвестно, не является ли слишком жестким требование линейности полувариограммы, но так как оно довольно широко используется на практике, то, очевидно, им можно пренебречь. Неопределенность в правильном определении априорной оценки дисперсии остатков также не слишком обременительна, так как только оценка ошибки дрифта зависит от этого параметра. Сам дрифт будет вычислен правильно, независимо от выбранного в качестве S_0^2 значения.

Пример

Вы уже могли заметить, что крайгинг, даже в очень упрощенном варианте, который мы рассмотрели, требует выполнения множества скучных арифметических действий. Практическое применение крайгинга к настоящим задачам оказывается возможным лишь при использовании ЭВМ, так как для того чтобы охарактеризовать изменения регионализованной переменной на некоторой площади, оценки должны производиться многократно для различных точек. В качестве примера рассмотрим карту, изображенную на рис. 5.67 и представляющую собой модификацию соответствующей карты [62]. На ней представлены значения уровня воды в Экус Бедз, водоносного горизонта на юге Центрального Канзаса. Структурный анализ, выполненный на значительно большей площади, показывает, что значения уровня воды могут рассматриваться как нестационарная регионализованная переменная, имеющая дрифт первого порядка. Область изменения полувариограммы остатков от дрифта есть 28 миль. В качестве модели полувариограммы можно выбрать линейную функцию с наклоном 60 квадратных футов на 1 милю.

Для построения карты уровней воды с помощью универсального крайгинга была использована программа построения карты в изолиниях, с помощью которой получались оценки уровней воды в точках, расположенных на равных интервалах вдоль картируемой площади. На карте рис. 5.67 представлено $31 \times 61=1891$ оценок крайгинга, каждая из которых основана на восьми ближайших контрольных скважинах, выбранных в октанте поиска вокруг оцениваемой точки. Каждая оценка в свою очередь требует решения системы одиннадцати уравнений.

В дополнение к карте уровней воды крайгинг был также использован для построения карты стандартной ошибки оценок, представленной на рис. 5.68. Стандартная ошибка в каждой из 47 наблюдательных скважин равна нулю, но увеличивается с расстоянием от известных контрольных точек (рис. 5.68). С вероятностью 95% истинная поверхность уровня воды лежит внутри интервала, определенного с точностью плюс или минус два указанных значения. Например, в точке А уровень воды оценивается примерно в 1480 футов. Так как имеется относительно немного наблюдательных скважин вблизи от этой точки, то карта стандартной ошибки указывает некоторое значение более 6 футов. Поэтому истинное значение в этой точке должно быть 1480±12 футов, т. е. между 1468 и 1492 футами с доверительной вероятностью 95%.

В этом примере мало геологического смысла в отношении самого дрифта, но при необходимости его можно изобразить на карте. На рис. 5.69 показан дрифт первого порядка уровней воды. На

рис. 5.70 представлена стандартная ошибка дрифта. Остатки от дрифта, найденные вычитанием карты дрифта из карты крайгинга, представлены на рис. 5.71. Площади отрицательных остатков, где уровни воды ниже, чем дрифт, указаны штриховкой.

Оценка значения переменной в точке по точечным наблюдениям – это лишь одно из применений крайгинга. Этот метод можно обобщить для построения оценок значений площади по выборкам, состоящим из площадей или же для построения оценок значений по выборкам, соответствующих объемам. Последнее применение особенно важно в горном деле, где оцениваемые величины являются содержанием руды в блоке и наблюдения представляют собой таблицы содержаний в керне скважины. Процедура оценки по существу такая же, как было описано выше, однако здесь возникают дополнительные сложности из-за изменчивости в пределах площадей или объемов. Отличное введение в использование крайгинга для оценки руды приводится в работе И. Кларк [13]. Более подробное изложение, включающее более трудные разделы госстатистики, можно найти в книгах Давида [21] и Журнеля и Юбре [41].

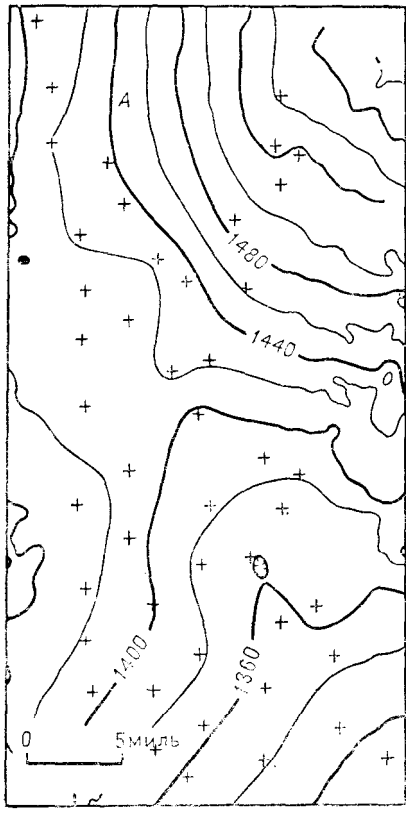


Рис. 5.67. Карта уровней воды в Экус Бедз, большом водоносном слое в южной части Центрального Канзаса. Карта построена методом универсального крайгинга в предположении дрифта первого порядка. Уровни изолинии указаны в футах выше уровня моря. Крестик – наблюдательные скважины

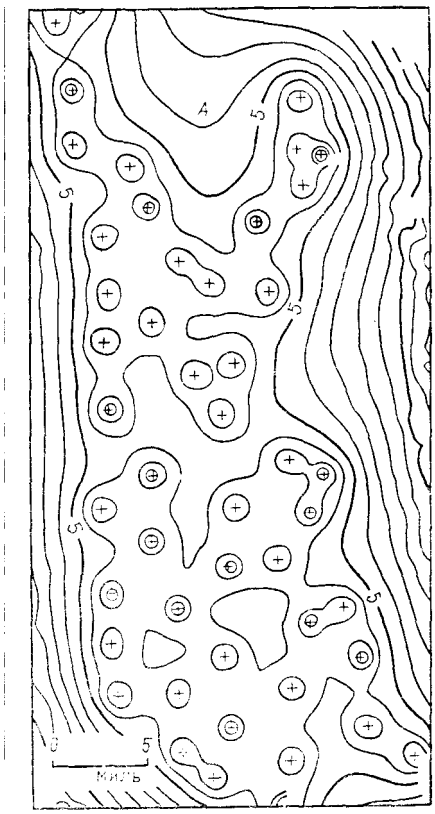


Рис. 5.68. На карте представлена стандартная ошибка оценок уровней воды в Экус Бедз. Интервал между изолиниями 1 фут

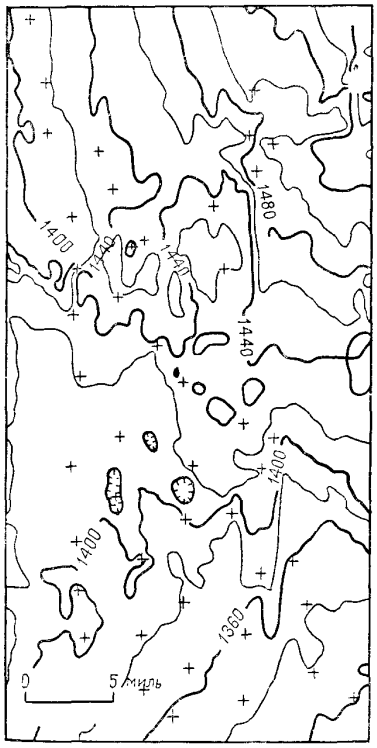


Рис. 5.69. Карта, представляющая дрифт первого порядка уровня воды в Экус Бедз. Уровни изолиний указаны в футах выше уровня моря

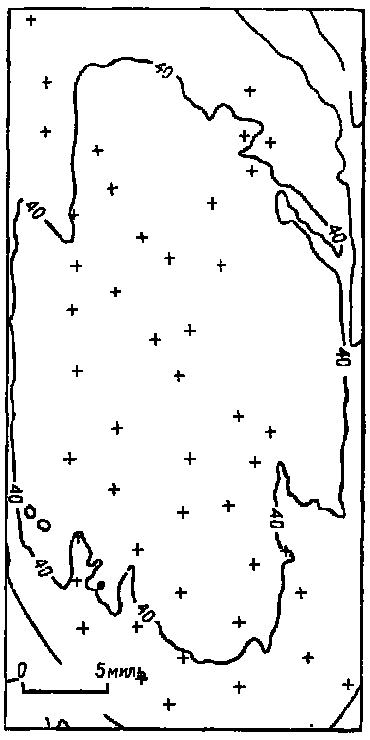


Рис. 5.70. Карта представляет стандартную ошибку дрифта первого порядка уровней воды в Экус Бедз. Интервал между изолиниями равен 10 футам

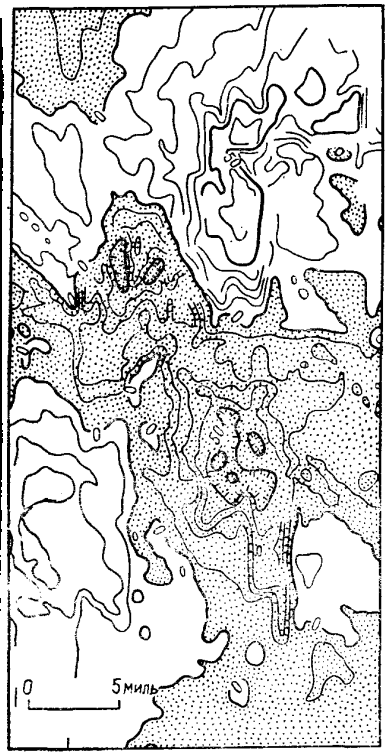


Рис. 5.71. Карта, на которой представлены остатки от дрифта первого порядка уровня воды в Экус Бедз. Интервал между изолиниями 10 футов. Заштрихованы площади отрицательных остатков, на которых уровни воды ниже дрифта

ПОВЕРХНОСТИ ТРЕНДА

Тренд-анализ – это чисто геологическое название математического метода разделения двух компонент: систематической и случайной по эмпирическим данным. Это разделение всегда проводилось геологами интуитивно или с помощью некоторых графических построений. Так, например, геологи-нефтяники обычно противопоставляют понятия «региональный прогиб», или «конфигурация бассейна», термину «локальная структура». Петрографы могут, например, говорить о «региональной зернистости» некоторой области метаморфизма. Геофизики же привыкли к понятиям «региональный тренд» и «локальные аномалии». Все эти выражения характеризуют ситуацию, ко гда наблюдаемый результат является следствием двух взаимодействующих геологических факторов или групп факторов, один (одна) из которых отражает региональную, или общую, геологическую обстановку, а второй (вторая) – мелкие локальные отклонения от региональных закономерностей. Весьма наглядные примеры для иллюстрации этих соотношений можно заимствовать из структурной геологии. Так, третичный бассейн Вайоминга сформировался в результате движений земной коры по разломам глубокого заложения, тогда как складчатые структуры внутри бассейна возникли под действием гравитационного скольжения, мелких дизъюнктивных нарушений и т.п. В подобной ситуации форма бассейна характеризует региональную структуру, а более мелкие структуры можно рассматривать как локальные отклонения.

Понятия «региональный» и «локальный» весьма субъективны. Они в значительной степени зависят и от размеров изучаемого региона. Так, если мы будем рассматривать всю поверхность докембрия в США, то по отношению к ней бассейны и разделяющие их горные хребты Вайоминга будут локальными отклонениями, или аномалиями, как и Блэл-Хилс, купол Озарк, Мичиганский бас-

сейн и др. Внутри же одного бассейна Вакоминг понятия «региональный» и «локальный» имеют совершенно иной смысл.

Имеющиеся данные также оказывают влияние на характер устанавливаемого регионального тренда и локальных отклонений. Так, например, бесполезно искать какой-либо смысл локальных закономерностей, если области их проявления близки по размерам и участкам опробования. Такие закономерности независимо от того, существуют они или нет, в подобных условиях просто нельзя установить. Меру зависимости между размерами участков проявления устанавливаемой закономерности и пространственным размещением точек равномерной сети можно вычислить для правильной сети точек [71], но не для случая нерегулярно расположенных точек, который значительно хуже поддается математической обработке.

Цель геологического исследования также влияет на рассматриваемые нами два понятия пространственных соотношений. Так, например, для золоторудного месторождения в Южной Африке представляют интерес только те «отклонения» содержаний золота, которые превышают заданное значение, заранее определенное экономистами. С другой стороны, при повторных поисках нефти в какой-либо области могут представить интерес небольшие структурные аномалии, так как заранее известно, что более крупные структуры данной области уже изучены. В этих условиях закономерности, выявляемые на таких мелких структурах, следует рассматривать как «региональный тренд».

Для иллюстрации рассмотрим график, изображенный на рис. 5.72,*а*. В представленной ситуации можно различными способами выделить «региональную» и «локальную» компоненты. Допустим, что региональный тренд характеризуется прямой линией, проходящей через совокупность точек наблюдения. Тогда все наши данные можно разделить на линейный тренд и три локальные большие аномалии, что и показано на рис. 5.72,*б*. Однако может оказаться, что для описания тренда параболическая функция будет более представительна, чем уравнение прямой линии. На рис. 5.72,*в* показано такое разделение на компоненту параболического тренда и локальные отклонения, что значительно отличается от ситуации, изображенной на рис. 5.72,*б*. Можно принять и более сложную функцию для описания тренда, например кубическую, которая приведет к еще меньшей величине аномалий (рис. 5.72,*г*). Возможна и такая ситуация, когда результаты опробования и кривая тренда будут совпадать, так что остаток будет отсутствовать. Конечно, в этом случае нельзя провести разделение на «региональную» и «локальную» компоненты и такое исследование потеряет смысл.

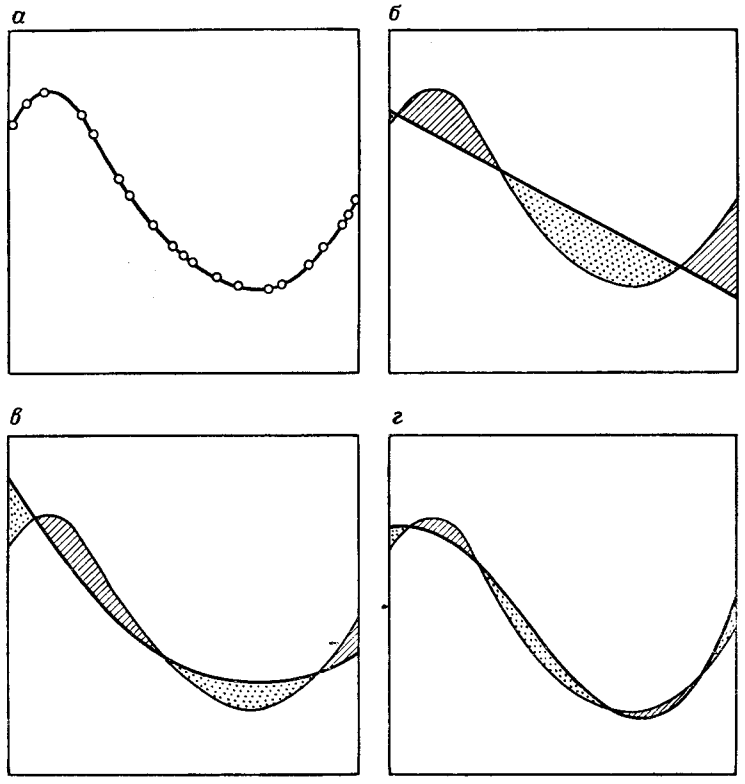


Рис. 5.72. Двумерная иллюстрация понятия тренда: *а* – множество данных точек и линия, на которой они расположены; *б* – прямая линия, подобранная к наблюдениям; *в* – параболический тренд; *г* – кубический тренд. Заштрихованная область соответствует отрицательному и положительному отклонениям от линии тренда

После всего того, что было сказано, вполне естественен вопрос: можно ли по результатам наблюдения получить объективное выделение двух компонент, если само определение этих компонент в значительной степени субъективное. Ответ на этот вопрос будет положительным, если вместо геологического определения тренда и отклонения воспользоваться операционным определением, которое фиксирует способ обработки данных. Например, тренд можно определить как линейную функцию географических координат, построенную по набору наблюдений так, что сумма

квадратов отклонений их от тренда минимальна. Рассмотрим более подробно три части этого определения.

1. Определение основано на понятии географических координат. Это значит, что результат наблюдения (абсолютная отметка местности, содержание золота в жиле и др.) рассматривается как функция положения наблюдения в пространстве.
2. Тренд рассматривается как линейная функция. Это значит, что уравнение, описывающее тренд, имеет форму $Y = b_1X_1 + b_2X_2 + ...$, где b – коэффициенты, а X – географические координаты. Уравнение будет включать значения Y, которые будут результатами наблюдения.
3. Требование минимизации суммы квадратов отклонений от тренда подробно описано в гл. 2 (см. кн. 1) применительно к дисперсии. Дело в том, что сумма квадратов отклонений результатов наблюдений от среднего характеризует выборочную дисперсию. Если вместо среднего подставить уравнение прямой линии или плоскости, то тогда, рассматривая это уравнение как функцию дисперсии, можно выбрать такой его вариант, который бы минимизировал сумму квадратов отклонений. Необходимо отметить, что определение уравнения линейной регрессии весьма сходно с только что приведенным определением. Вообще тренд-анализ можно рассматривать как один из вариантов статистического метода множественной регрессии, и поэтому все приемы обработки данных взяты непосредственно из регрессионного анализа. В некоторых случаях при решении геологических задач будут проверяться гипотезы, связанные с множественной регрессией.

В гл. 4 (см. кн. 1) была построена линия регрессии Y на X, которая являлась линией наилучшей оценки Y для любого заданного значения X. Уравнение прямой $Y = b_0 + b_1X$ находилось путем решения системы нормальных уравнений:

$$\Sigma Y = b_0 n + b_1 \Sigma X$$
$$\Sigma XY = b_0 \Sigma X + b_1 \Sigma X^2 \qquad (5.82)$$

относительно неизвестных коэффициентов b_2, b_1. Суммирование в этих и последующих уравнениях проводится от i=1 до n. Для простоты запись пределов суммирования опущена.

Эту систему уравнений легко приспособить, если имеется два аргумента, например такие, как географические координаты, в результате чего получим уравнение линейной поверхности тренда:

$$Y = b_0 + b_1X_1 + b_2X_2. \quad (5.83)$$

В данном случае результат геологического наблюдения рассматривается как линейная функция двух координат X_1 и X_2 с коэффициентами b_0, b_1, b_2, оценить которые можно с помощью системы следующих уравнений:

$$\begin{aligned} \sum Y &= b_0 n + b_1 \sum X_1 + b_2 \sum X_2 \\ \sum X_1 Y &= b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 \\ \sum X_2 Y &= b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 \end{aligned} \qquad (5.84)$$

Решив уравнения (5.84) относительно b_0, b_1, b_2, найдем их оценки. Этот метод нахождения оценок называется методов наименьших квадратов.

Уравнения (5.84) можно записать в матричной форме:

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{bmatrix} \qquad (5.85)$$

Сходство между матричным уравнением (5.85) и уравнением (4.33) очевидно. Оба эти уравнения можно рассматривать как приближенные характеристики функции двух аргументов, которые в данном случае являются географическими координатами X_1 и X_2. При подборе кривой второго порядка в рассмотренном выше примере также рассматривались две переменные X и X^2, и задача была сведена к решению системы линейных уравнений. Таким образом, между этими двумя процедурами нет принципиальной разницы. В качестве примера построения линейной поверхности тренда рассмотрим следующую задачу.

Одна английская нефтяная компания приобрела концессию в весьма удаленной части Северо-Восточной Африки. Территория была крайне необжитой, труднодоступной и почти полностью геологически неизученной. По условиям концессии компания должна была пробурить в течение года десять скважин или в случае невыполнения этого условия потерять свои права. Руководство ком-

пании приняло решение бурить серию далеко расположенных друг от друга разведочных скважин, предназначенных для создания геологической основы, необходимой для продолжения поисков. На рис. 5.73 показано расположение этих скважин на территории концессии, общая площадь которой составляла 100 км^2. Координаты скважин, отсчитываемые в километрах от юго-западного угла территории, и абсолютные отметки подошвы меловых отложений, зафиксированные в скважинах, приведены в табл. 5.15. Задача в данном случае заключается в построении линейной поверхности тренда и выявлении областей положительных отклонений от нее, которые можно рассматривать как заслуживающие внимания для проведения дополнительных исследований.

Для построения поверхности тренда мы должны сначала подсчитать суммы значений X_1, X_2 и Y, суммы квадратов X_1 и X_2, а также суммы соответствующих смешанных произведений, т.е. числа, требуемые формулой (5.84)

$$\sum X_1 = 539 \qquad \sum X_2 = 482 \qquad \sum Y = -4579$$
$$\sum X_1^2 = 36934 \qquad \sum X_2^2 = 31692 \qquad \sum X_1 Y = -211098$$
$$\sum X_1 X_2 = 27030 \qquad \sum X_2 Y = -232342$$

Подставив эти значения в формулу (5.85), получим

$$\begin{bmatrix} 10 & 539 & 482 \\ 539 & 36943 & 27030 \\ 482 & 27030 & 31692 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -4579 \\ -211098 \\ -232342 \end{bmatrix}$$

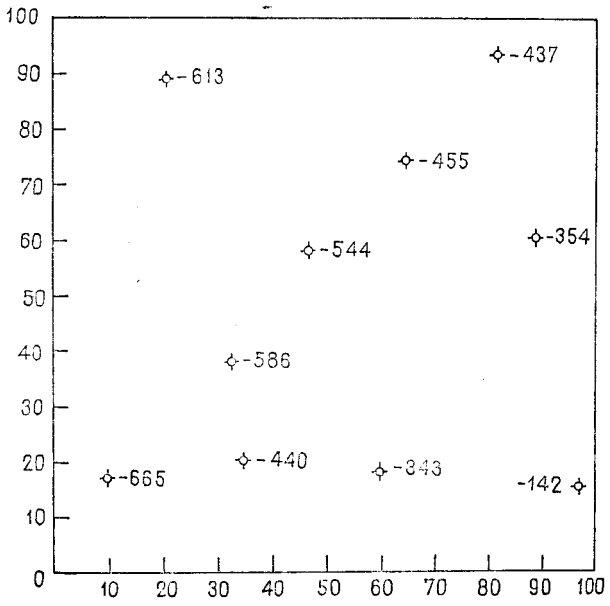


Рис. 5.73. Карта расположения скважин и абсолютных отметок подошвы меловых отложении в Северо-восточной Африке. По данным нефтяной компании Англо-Баррен Ойл Кампани. Отметки даны в метрах ниже уровня моря. За единицу выбран километр; отсчет координат ведется от юго-западного угла карты

Таблица 5.16. Координаты, абсолютные отметки подошвы меловых отложений, их оценки Y и разность $\left(Y - \overline{Y}\right)$

X_1, км	X_2, км	Y, м	$\overline{Y}$, м	$\left(Y - \overline{Y}\right)$,м
10,0	17,0	–665,0	–606,6	–58,3
21,0	89,0	–613,0	–695,7	82,7
33,0	38,0	–586,0	–537,8	–48,1
35,0	20,0	–440,0	–492,8	52,8
47,0	58,0	–544,0	–510,2	–33,7
60,0	18,0	–343,0	-369,3	26,2
65,0	74,0	–455,0	–455,5	0,5
82,0	93,0	–437,0	–411,5	–25,4
89,0	60,0	–354,0	–313,0	–40,9
97,0	15,0	–142,0	–186,0	44,1

Таблица 5.15 Координаты скважин и возвышений оснований мелового периода в пределах нефтяной концессии

X_1, км	X_2, км	Y, м
10,0	17,0	–665,0
21,0	89,0	–613,0
33,0	38,0	–586,0
35,0	20,0	–440,0
47,0	58,0	–544,0
60,0	18,0	–343,0
65,0	74,0	–455,0
82,0	93,0	–437,0
89,0	60,0	–354,0
97,0	15,0	–142,0

Решить это матричное уравнение можно методом, описанным в гл. 3 (см. кн. 1). Решение будет следующим:

$$b_0 = -621{,}0, \quad b_1 = 4{,}8, \qquad b_2 = -2{,}0.$$

Подставив эти значения коэффициентов в уравнение

$$Y = b_0 + b_1 + b_2X_2,$$

можно вычислить теоретические значения Y для каждой из десяти скважин, которые вместе с разностями $\left(Y_1 - \overline{Y}_1\right)$ приведены в табл. 5.16.

Кроме того, можно охарактеризовать качество приближения поверхности тренда к наблюдаемым результатам, используя формулы с (4.18) по (4.24), введенные для случая линии. В частности, мы можем охарактеризовать общую изменчивость, вычислив сумму квадратов для Y, т.е.

$$SS_T = \sum Y^2 - \frac{1}{n}\left(\sum Y\right)^2 = 215324{,}9\,.$$

Для того чтобы получить аналогичную характеристику для Y в табл. 5.16, мы должны вычислить сумму квадратов, возникающую из уравнения регрессии

$$SS_R = \sum \overline{Y}^2 - \frac{1}{n}\left(\sum Y^2\right) = 193861{,}4\,.$$

Разность между этими величинами будет характеризовать изменчивость отклонений от поверхности тренда:

$$SS_D = SS_T - SS_R = 21463{,}5.$$

Таким образом, поверхность тренда учитывает следующий процент общей изменчивости:

$$100\% \times R^2 = (SS_R)/(SS_T) = 90{,}0\%.$$

Коэффициент множественной корреляции в данном случае составит $R = \sqrt{R^2} = 0{,}95$.

Из полученных крайне высоких значений можно сделать вывод о том, что подошва меловых отложений изучаемой территории является почти ровной и описывается постепенно погружающейся плоскостью. Как видно из табл. 5.16, отклонения от теоретической плоскости весьма малы. Все это наглядно представлено на рис. 5.74, где приведена карта подошвы меловых отложений.

Хотя этот простейший вид анализа является удовлетворительным в данном примере, вполне возможны ситуации, когда плоскости будет недостаточно для описания геологического тренда, который может быть весьма сложным. Более того, мы очень редко располагаем априорными сведениями о форме функции, описывающей тренд. Физики, например, могут заранее сказать, что брошенный камень полетит по параболе, так как они располагают некоторыми сведениями о факторах, контролирующих этот процесс, т.е. об ускорении свободного падения и др. Геологи крайне редко могут говорить априори о наилучшей форме функции, описывающей поверхность тренда. Самое лучшее, что они могут сделать, – это проводить последовательные приближения к неизвестной функции, начиная с некоторой функции произвольной формы. В частности, они расширяют возможности представления линейной поверхности с помощью полиномов, вводя степени выше первой и смешанные произведения географических координат. Полиномы исключительно чувствительны, и, если использовать их достаточно высокие степени, с их помощью можно описывать весьма сложные поверхности.

Необходимо отметить, что полиномиальные функции используются в тренд-анализе главным образом как удобное средство описания полученных данных. При этом уравнения, по которым отыскиваются полиномиальные коэффициенты, легко строятся и решаются на ЭВМ.

Применение полиномов может привести к мнению, что геологические процессы являются полиномиальными функциями и даже их линейными вариантами. Нужно помнить, что природа этих процессов остается неизвестной и с помощью полиномов может быть описана только приближенно. В отдельных примерах могут быть более приемлемы и другие варианты приближения, что будет рассмотрено в данном разделе.

Как уже отмечалось в гл. 4, метод наименьших квадратов может быть применен не только к уравнению прямой, но и к кривой второго (и более высокого) порядка путем добавления соответствующих компонент:

$$Y = b_0 + b_1X_1 + b_2X_1^2. \qquad (5.86)$$

Поверхность тренда второго порядка будет описываться уравнением

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2 + b_5X_1X_2. \qquad (5.87)$$

Заметим, что эти уравнения содержат такие компоненты, как квадраты географических координат и их смешанное произведение. Перейти от этого уравнения к уравнению более высокого порядка сравнительно легко. Для этого каждая географическая координата просто возводится в заданную степень и добавляются соответствующие смешанные произведения. Например,

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2 + b_5X_1X_2 + b_6X_1^3 + b_7X_2^3 + b_8X_1^2X_2 + b_9X_1X_2^2 \qquad (5.88)$$

представляет собой уравнение поверхности тренда третьей степени. В этом уравнении членам первой степени соответствуют коэффициенты b_1 и b_2. Коэффициенты b_3, b_4 и b_5 соответствуют членам второй степени и стоят перед переменными, которые имеют следующую структуру:

$$X_3 = (X_1 \times X_1), \quad X_4 = (X_2 \times X_2), \quad X_3 = (X_1 \times X_2).$$

Таким образом, новые переменные представляют собой различные варианты произведений исходных переменных. Аналогично коэффициенты b_6, b_7, b_8 и b_9 соответствуют компонентам третьей степени, которые имеют следующую структуру:

$$X_6 = (X_1 \times X_1 \times X_1), \quad X_7 = (X_2 \times X_2 \times X_2), \quad X_8 = (X_1 \times X_2 \times X_2) \quad \text{и} \quad X_0 = (X_1 \times X_2 \times X_2).$$

Разведчики нефти нашли хорошее применение тренд-анализа поверхностей при поисках нефти и газа в Центральной Альберте (Канада). Нефтеносный бассейн в Альберте обширен, причем нефть локализуется как в нижнемеловых отложениях, так и в более глубоких карбонатных рифах верхнего девонского возраста. Естественно, большее количество скважин достигает меловых образований, меньшее – девонских, особенно вблизи от передней гряды Скалистых гор, где прогнозы глубинности рифов составляют 15000 футов или больше. Девонские рифы представлены мощной толщей карбонатов ледукской формации, переслаивающихся и фациально замещающихся сланцами лагунного и морского генезиса – кластическими образованиями. Карбонатные рифы не были уплотнены, в то время как переслаивающиеся с ними мелкозернистые пластические образования – сланцы Айртона были уплотнены до их первоначальной мощности по мере опускания бассейна и продолжения образования месторождения. Это дифференциальное уплотнение создано складчатыми структурами над погребенными рифами; эти структуры сохраняются в вышележащих породах, хотя их величины уменьшаются на более мелких горизонтах.

Глубокозаложенные черты уплотнения не являются очевидными на меловых горизонтах из-за пугающего эффекта сильного регионального наклона пласта (рис. 5.75). Замкнутость на глубине выражается только как слабые изменения локального градиента, который на структурных картах меловых горизонтов выглядит как слабое изменение в расположении изолиний. Однако если эта компонента сильного регионального наклона может быть устранена на этих картах, то гипотетические нижележащие складчатые структуры выглядят замкнутыми. Так как плотность скважин в основаниях меловых отложений относительно высока, их анализ может дать ценную информацию о возможных прогнозах в девонских пластах, хотя некоторые скважины и проникают в более глубокие горизонты.

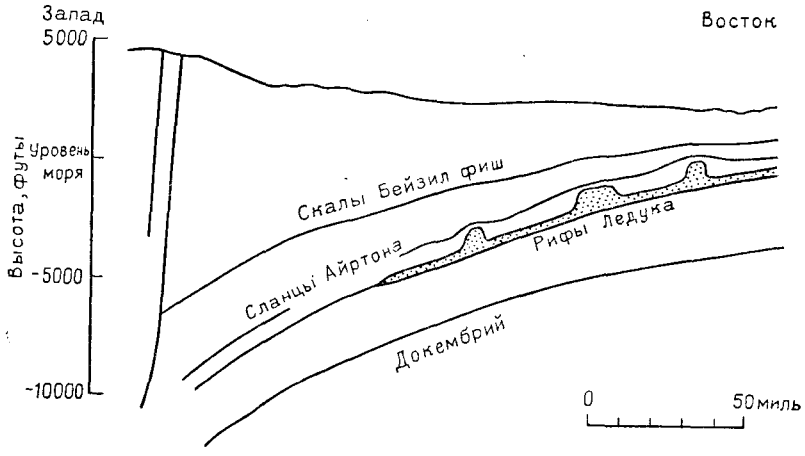


Рис. 5.75. Изображение разреза западного поля бассейна Альберта в виде диаграммы. Рифы Ледука и сланцы Айртона верхнего девона; скалы Бейзил-Фиш сложены нижнемеловыми породами

Скалы Бейзил-Фиш состоят из черного сланца, встречающегося вблизи границы между нижним и верхним меловыми пластами. Это место характеризуется многочисленными бентонитовыми образованиями, которые дают загадочные пики при каротаже в гамма лучах из-за их высокой радиоактивности. Бентониты синхронны и потому образуют отличные маркеры для региональной корреляции. Вершина скал Бейзил-Фиш может быть отмечена с исключительной точностью по каротажу благодаря заметному отражению от мощных бентонитов. Карта, представленная на рис. 5.76, охватывает площадь около 3500 квадратных миль на западе Центральной Альберты и была построена при использовании пика этих бенто-

нитов в верхней точке скал Бейзил-Фиш в 360 разведочных скважинах. Картируемая площадь расположена на западном крае бассейна Альберта непосредственно в передней части складчатой зоны, отделяющей фронт Скалистых гор. Аппроксимируется погружение оснований вниз к юго-западу с увеличением погружения в западной ветви бассейна. В пределах картируемой площади глубины скал Бейзил-Фиш изменяются от слабого с эксцессом в 1000 футов ниже уровня моря на северо-востоке, вплоть до 5000 футов ниже уровня моря на северо-западе.

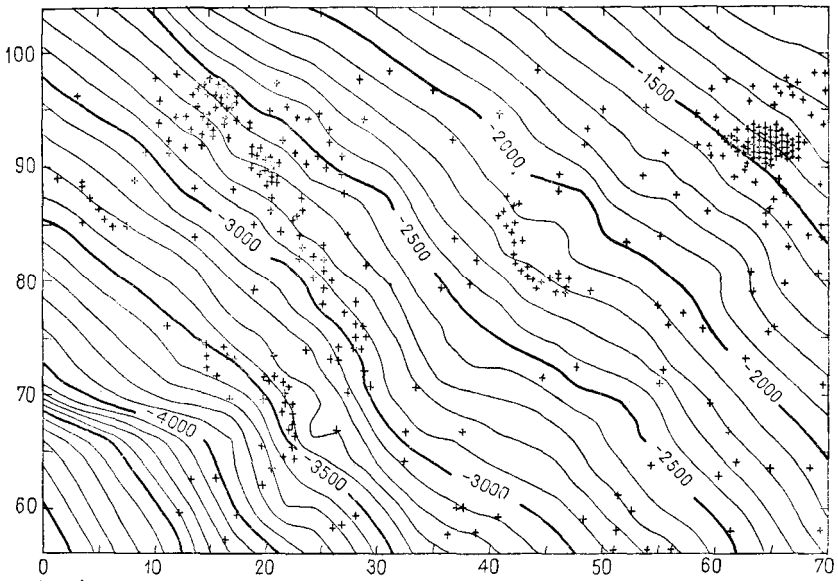


Рис. 5.76. Структурная карта вершины скал Бейзил-Фиш, бассейн Альберта. Изолинии указаны в футах ниже уровня моря. Крестики – контрольные скважины

Таблица 5.17 Статистики полиномиальной тренд-поверхности первой и второй степеней, подгоняемые к возвышениям Бейзил-Фиш в центре Западной Альберты

Тренд-поверхность первой степени:	
процент соответствия (R^2)	98,6%
Коэффициент корреляции (R)	0,993
Уравнение поверхности тренда:	$Y = -635{,}4 + 29{,}3X_1 + 33{,}8X_2$
Тренд-поверхность второй степени:	
процент соответствия (R^2)	99,7%
Коэффициент корреляции (R)	0,999
Уравнение поверхности тренда:	$Y = -7993{,}3 + 63{,}4X_1 + 59{,}2X_2 - 0{,}1X_1^2 - 0{,}3X_2^2 - 0{,}1X_1X_2$

На рис. 5.77 и 5.78 представлены полиномиальные поверхности тренда первого и второго порядков для скал Бейзил-Фиш; значения статистик, характеризующих подбор поверхностей тренда, приведены в табл. 5.17. Обе тренд-поверхности обеспечивают очень высокую точность аппроксимации наблюдений, хотя поверхность второй степени дает значительно более точное приближение, чем поверхность первой степени. В этом примере для выбора степени поверхности тренда статистические критерии значимости неприменимы, так как проблема состоит не только в статистической оценке. Цель скорее состоит в более точном моделировании региональных черт структурной контурной карты, и потому вычитание тренда помогает устранить региональные компоненты структуры. Действительно, поверхность тренда используется как фильтр высокой пропускной способности, устраняющий крупномасштабные структурные вариации из карты и сохраняющий мелкомасштабные характерные черты.

На рис. 5.79 изображены остатки от тренд-поверхности второй степени. Рис. 5.80 – это палеографическая карта верхнего девона, реконструированная по скважинам и сейсмической информации. Заметно сильное совпадение между положительными остатками тренда на скалах Бейзил-Фиш и положениями Больших девонских рифов, в частности рифов Вайндфолл. После устранения регионального тренда остаются последовательные изолированные более мелкие компоненты структур, представляющих складки над этими рифами. Несколько крупных нефтяных полей на этой площади были открыты с помощью данного метода.

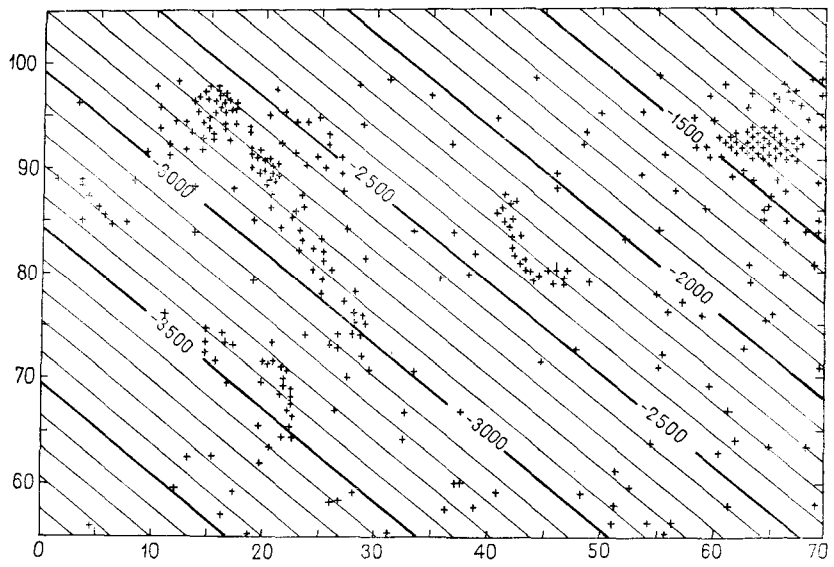


Рис. 5.77. Поверхность тренда первой степени, построенная для скал Бейзил-Фиш. Изолинии указаны в футах ниже уровня моря

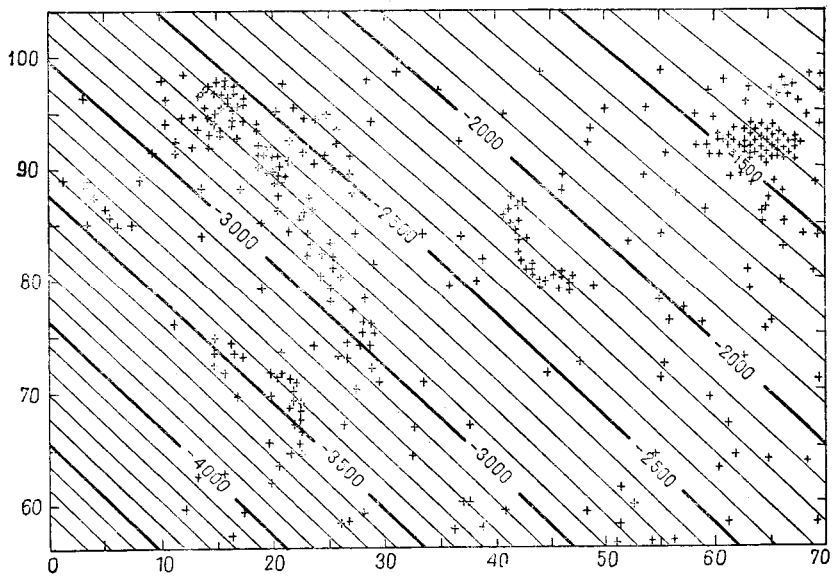


Рис. 5.78. Поверхность тренда второй степени, построенная для скал Бейзил-Фиш. Изолинии указаны в футах ниже уровня моря

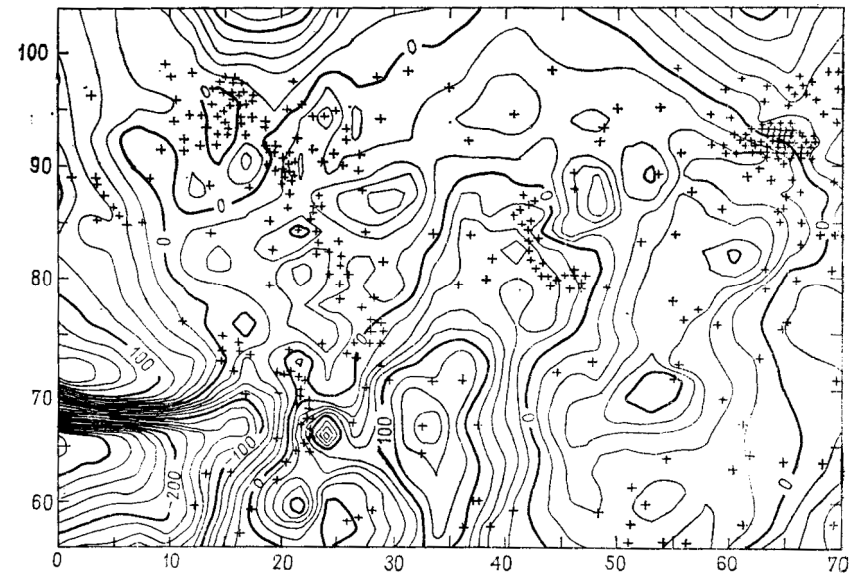


Рис. 5.79. Карта разностей значений структурной карты скал Бейзил-Фиш (см. рис. 5.78) и тренда второй степени (см. рис. 5.75). Крестиками заштрихованы области, соответствующие положительным разностям. Интервал между изолиниями 20 футов

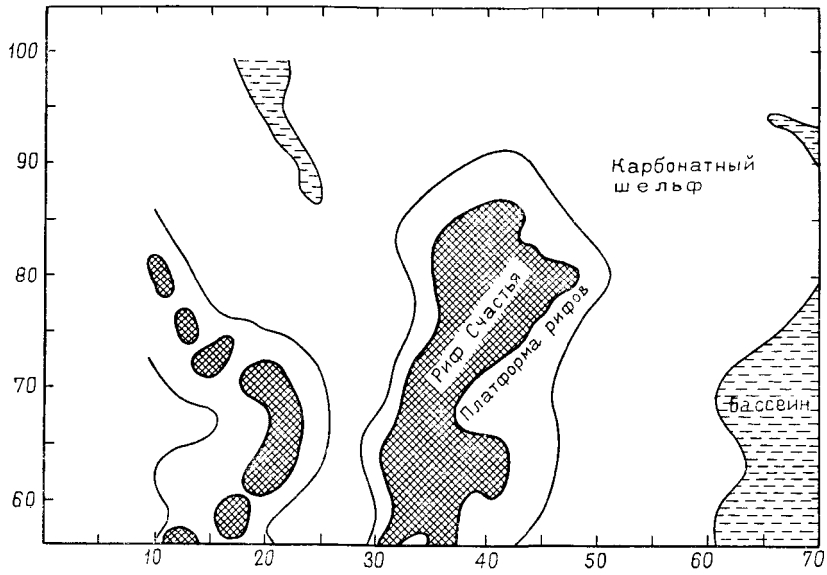


Рис. 5.80. Палеогеографическая карта верхнего девона рифов Ледука, развитых на карбонатной платформе вдоль западного побережья Девонского моря

Теперь нужно получить ответ на вопрос: что такое высокое (или низкое) соответствие поверхности тренда результатам наблюдений, т.е. что такое корреляция? В геологических исследованиях, чтобы получить ответ на какой-либо вопрос, часто приходится доверяться опыту и интуиции. Например, структурные анализы, полученные по данным наблюдений, выполненных в Канзасе, Оклахоме, Техасе, Вайоминге, Калифорнии (США), Англии и других местах, показал, что приближение плохое, если коэффициент корреляции меньше 0,3. Если же он принимал значение в интервале от 0,4 до 0,6, то это интерпретировалось как необходимость составить наряду с картой тренда карту остатков, а если коэффициент корреляции превышал 0,7, то делался вывод о хорошей согласованности поверхности тренда и исходных данных. Необходимо отметить, что при рассмотрении степени приближения поверхности тренда к исходным данным нужно обязательно учитывать цель проводимого исследования. Во всех этих структурных исследованиях мы изучали бассейны, имеющие относительно простую форму, и выбирали те из них, которые характеризовались небольшими отклонениями от поверхности тренда по сравнению с общими размерами бассейна. При этом весьма хорошее приближение (коэффициенты корреляции около 0,8) обеспечивали полиномы третьей и четвертой степени. Заметим, что в моделях, построенных с помощью случайных чисел, принимающих значения в том же интервале, что и реальные данные, для полиномов четвертой степени коэффициенты корреляции оказывались близкими к 0,3. Таким образом, реально существующий тренд можно отделить от случайных отклонений, что и делается геологами при интерпретации тренда и карты остатков.

На рис. 5.81,*а* приведена карта, построенная по данным табл. 5.18, характеризующей абсолютные отметки кровли верхней формации ордовика Центрального Канзаса. На рис. 5.81,*б* приведена поверхность тренда первой степени, представленная плоскостью, построенной по этим же данным, а на рис. 5.81,*в* изображена карта отклонений (остатков) от этой плоскости. Поверхность тренда второй степени и соответствующая ей карта остатков приведены на рис. 5.81,*г* и *д*. В качестве упражнения по данным табл. 5.18 постройте поверхность тренда, описываемую полиномом третьей степени, для кровли пород ордовика. Сравните полученную вами карту с рис. 5.81,*г*. Постройте для вашей поверхности карту остатков и сравните ее с рис. 5.81,*д*. Изменилась ли конфигурация карты остатков? На рис. 5.82 изображена карта основных нефтегазоносных полей Центрального Канзаса. Сравните ее с картой распределения положительных отклонений от поверхности тренда третьего порядка.

Таблица 5.18. Географические координаты точек наблюдения и абсолютные отметки кровли пород ордовика Центрального Канзаса

X_1	X_2	Y, футы	X_1	X_2	Y, футы	X_1	X_2	Y, футы
23,36	22,10	–2961,0	25,60	17,18	–2337,0	31,60	22,82	–2711,0
29,80	10,58	–2240,0	28,60	26,15	–4373,0	35,90	9,90	–1537,0
22,18	16,24	–1872,0	29,32	11,48	–2104,0	21,54	14,90	–1667,0
22,91	3,36	–2584,0	32,60	24,15	–2923,0	33,46	12,31	–1694,0
39,33	21,14	–1119,0	25,90	19,05	–2607,0	19,40	16,57	–2300,0
15,10	18,50	–3062,0	20,12	19,80	–2751,0	35,18	19,79	–1465,0
28,96	10,30	–2540,0	33,87	23,97	–1626,0	21,20	3,50	–2340,0
18,90	16,69	–2300,0	34,28	1,78	–2305,0	23,50	13,02	–1564,0
21,11	12,26	–1505,0	39,60	19,40	–1135,0	19,09	24,40	–3657,0
29,92	4,00	–1921,0	39,01	9,93	–197!,0	34,45	19,45	–1257,0
41,99	5,31	–2056,0	24,60	20,45	–2483,0	20,94	22,90	–3044,0
21,86	4,01	–2466,0	25,10	22,84	–3095,0	17,53	10,02	–1657,0
41,30	16,38	–1077,0	24,48	24,45	–3589,0	28,27	8,16	–2540,0
22,75	21,67	–2780,0	26,47	13,19	–1490,0	17,83	0,10	–1647,0
35,93	14,53	–707,0	20,23	17,58	–2307,0	32,58	19,86	–2140,0
21,59	21,16	–2677,0	35,27	19,23	–1037,0	25,45	17,03	–883,0
28,20	6,40	–2801,0	25,00	9,20	–1407,0	21,21	14,57	–1695,0
30,50	19,83	–2678,0	22,95	8,42	–2133,0	26,20	16,25	–1746,0
41,20	3,50	–2586,0	35,00	23,13	–3090,0	28,52	17,20	–2440,0
31,38	17,74	–2190,0	30,07	15,18	–1890,0	29,80	13,96	–2346,0
24,30	23,90	–3367,0	40,00	16,82	–1366,0	30,40	10,02	–2182,0
24,80	21,10	–2959,0	18,45	9,01	–1651,0	37,45	7,18	–1934,0
16,75	8,66	–1709,0	38,01	26,05	–1857,0	27,91	16,48	–2481,0
22,72	11,15	–1431,0	29,05	21,08	–2998,0	19,92	11,10	–1599,0
20,40	21,82	–3022,0	25,40	20,90	–1902,0	26,62	20,27	–2875,0
26,49	6,43	–2431,0	41,18	20,14	–998,0	36,50	10,21	–1353,0
33,12	17,04	–1792,0	26,70	12,80	–1765,0	16,58	5,12	–1608,0
24,67	15,61	–2146,0	25,58	4,63	–1593,0	34,45	6,22	–1499,0
30,10	12,92	–2131,0	20,70	20,39	–2722,0	30,26	22,27	–3029,0
26,12	0,50	–2295,0	20,80	11,50	–1477,0	35,60	20,95	–1860,0
28,25	14,35	–2421,0	41,11	5,22	–2274,0	23,23	11,59	–1412,0
27,21	10,58	–2204,0	22,11	21,00	–2598,0	27,02	17,01	–2407,0
19,38	17,60	–2348,0	26,86	21,17	-2951,0	25,02	17,82	–2063,0
42,00	17,90	–885,0	29,93	16,80	–2435,0	34,20	26,20	–2834,0
32,48	2,24	–1618,0	24,62	23,12	–3177,0	16,60	1,85	–1583,0
18,96	14,54	–1834,0	26,04	16,52	-2241,0	17,40	1,10	–1521,0
32,97	11,79	–1884,0	32,86	13,52	–1775,0	41,12	19,86	–835,0
23,50	1,57	–2308,0	15,10	25,85	–5400,0	29,92	22,02	–3061,0
41,89	15,25	–1108,0	36,99	25,28	–1852,0	17,10	14,05	–2402,0
31,40	12,20	–2058,0	22,40	13,23	–1500,0	21,70	14,21	–1608,0
20,60	6,28	–1613,0	30,32	15,55	–2195,0	39,62	18,03	–1421,0

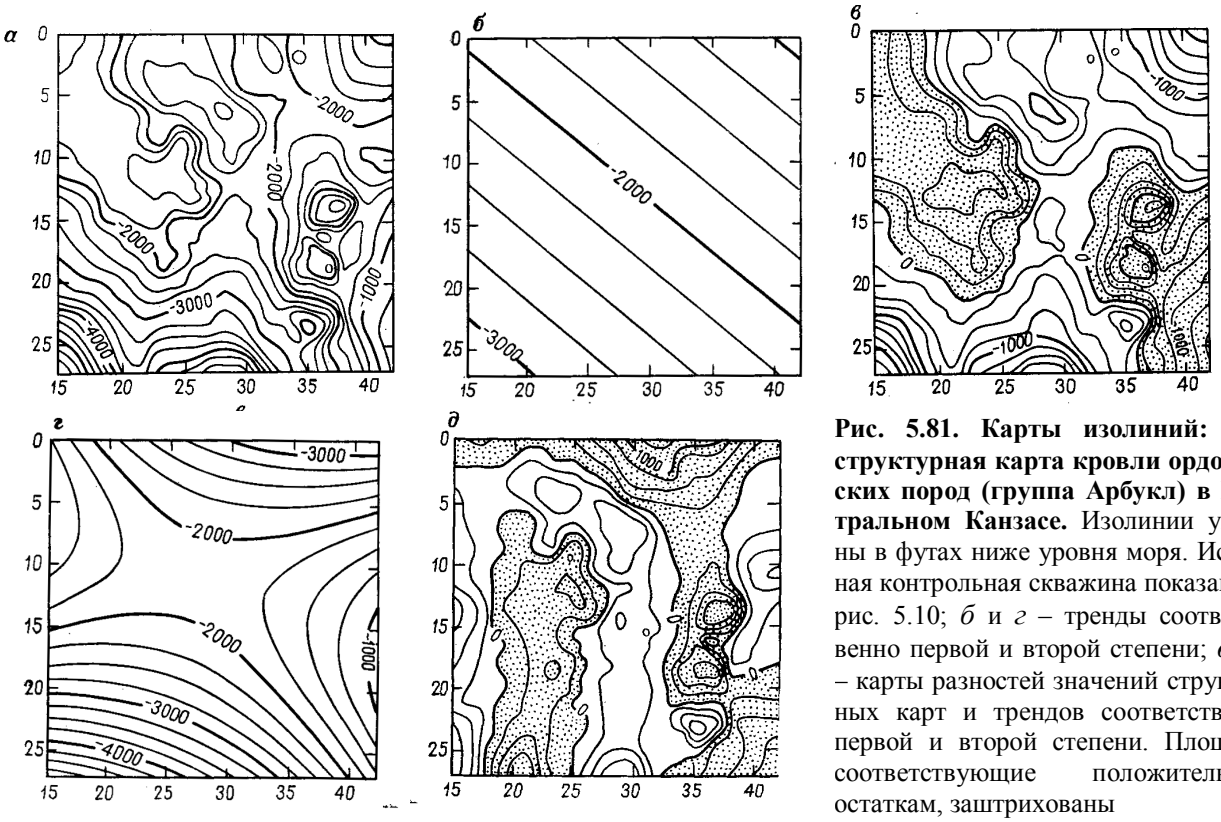


Рис. 5.81. Карты изолиний: *а* – структурная карта кровли ордовикских пород (группа Арбукл) в Центральном Канзасе. Изолинии указаны в футах ниже уровня моря. Исходная контрольная скважина показана на рис. 5.10; *б* и *г* – тренды соответственно первой и второй степени; *в* и *д* – карты разностей значений структурных карт и трендов соответственно первой и второй степени. Площади, соответствующие положительным остаткам, заштрихованы

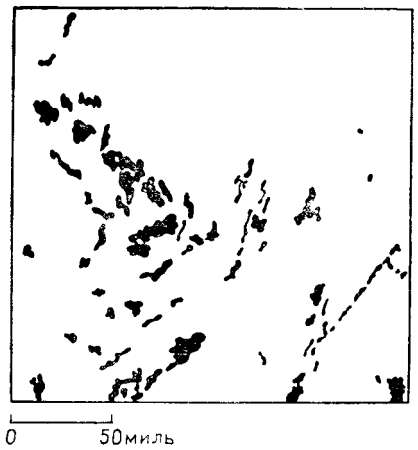


Рис. 5.82. Расположение основных нефтегазоносных полей в центральном Канзасе

Статистические критерии в тренд-анализе

Согласованность построенной поверхности тренда и исходных данных можно проверить статистически, путем сравнения дисперсии этой поверхности с дисперсией отклонений от нее. Вернувшись к гл. 2 (см. кн. 1), отметим, что проверка гипотезы о равенстве дисперсий с помощью F-критерия действительна только при выполнении определенных условий по отношению к выборочным данным. Если эти предположения сделаны обоснованно, то мы можем рассматривать коэффициенты b_i, найденные с помощью метода наименьших квадратов, как оценки истинных значений коэффициентов регрессии β_i и проверить сформулированные относительно них гипотезы. Мы должны допустить, что случайная величина, рассматриваемая в качестве функции, распределена нормально с математическим ожиданием, равным истинному значению регрессии, и что дисперсия функции не изменяется с изменением аргумента. Кроме того, отбор проб из генеральной совокупности должен быть беспристрастным. Последнее условие обычно трудно выполнимо, особенно в структурном анализе, базирующемся на данных скважин, так как их положение, как правило, не яв-

ляется случайным. Проверка статистических гипотез о характере поверхности тренда будет наиболее простой в случае, когда данные представляют собой результаты химического анализа проб, собранных по заданному плану.

Значимость поверхности тренда, или уравнения регрессии, можно проверить с помощью дисперсионного анализа, который основан на разделении общей дисперсии набора наблюдений на компоненты, соответствующие определенным источникам изменчивости. Конечно, именно это и делается при разделении общей изменчивости величины Y на тренд (или регрессию) и остаток (или отклонение). Число степеней свободы, соответствующее общей изменчивости в тренд-анализе, равно n–1, где n – число наблюдений. Число степеней свободы, соответствующее уравнению регрессии, равно числу членов полинома, используемого для построения этого уравнения. Число степеней свободы, соответствующее отклонениям, равно разности между числами степеней свободы для упомянутых двух типов изменчивости, т.е. $\nu_D = \nu_T - \nu_R$. Формальная процедура дисперсионного анализа показана в табл. 5.19.

Таблица 5.19. Общие **ANOVA** для проверки значимости полиномиальной регрессии тренд-поверхности k-й степени (n – число наблюдений; m – число членов полинома, не считая b_0)

Источник изменчивости	Сумма квадратов	Число степеней свободы	Среднее квадратов	F-критерий
Регрессионный полином	SS_n	m	MS_R	MS_R/MS_D
Отклонение от полинома	SS_D	$n–m–1$	MS_D	
Общая изменчивость	SS_T	$n–1$		

Значения средних квадратов вычисляются путем деления соответствующих сумм квадратов на определенное число степеней свободы. Средние квадраты в свою очередь являются оценками дисперсий, и при их сравнении можно воспользоваться F-распределением Фишера. Так, MS_D – оценка дисперсии, возникающей как следствие отклонений отдельных наблюдений от поверхности регрессии, а MS_R – оценка дисперсии самой поверхности регрессии. Если регрессия играет существенную роль, то дисперсия отклонений от поверхности регрессии будет мала по сравнению с дисперсией самого уравнения регрессии.

При общей проверке предположения о наличии или отсутствии тренда рассматривается частное от деления оценки дисперсии уравнения регрессии на оценку дисперсии отклонений. Значение F дает вероятный ответ на вопрос: можно ли рассматривать две упомянутые оценки дисперсий как несущественно отличающиеся одна от другой, т. е. регрессия не дает какого-либо эффекта по сравнению со случайными отклонениями. Утвердительный ответ на этот вопрос можно интерпретировать так, что распределение случайной величины не зависит от значений $X_1, ..., X_n$ или величина Y частично зависит от $X_1, ..., X_m$, а математическая модель для описания этой зависимости выбрана неверно.

В большинстве вариантов формальной постановки задачи о пригодности уравнения регрессии для описания зависимости используются следующая проверяемая гипотеза (H_0) и соответствующее ей множество альтернатив (H_1):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$
$$H_0 : \beta_1, \beta_2, \ldots, \beta_m \neq 0 \qquad (5.89)$$

Проверяемая гипотеза заключается в том, что все коэффициенты регрессии равны нулю, или, иными словами, что регрессии нет. Если вычисленное значение F превысит допустимое, соответствующее заданному уровню значимости и числу степеней свободы, то проверяемая гипотеза отвергается кал противоречащая выборочным данным и принимается альтернатива H_1.

Некоторые исследователи, применяя полиномы в тренд-анализе, последовательно увеличивают их степень, что приводит к постепенному увеличению числа слагаемых. В подобной ситуации дисперсионный анализ можно распространить на изучение тех вкладов в изменчивость, которые дают добавляемые регрессионные компоненты, что позволит ввести меру эффективности увеличения порядка уравнения. Такой критерий строится как разность между суммами квадратов уравнений регрессии высшего и предшествующего порядков. Разделив эту разность на разность соответст-

вующих чисел степеней свободы, получим средний квадрат регрессии, обусловленный увеличением степени полинома. Частное от деления полученного среднего квадрата на средний квадрат отклонения от регрессионной поверхности более высокой степени будет иметь F-распределение. Если вычисленное значение отношения F превысит допустимое при заданном уровне значимости и соответствующем числе степеней свободы, то из этого следует, что увеличение степени полинома дает эффект. Если значение отношения F незначимо, то делаем вывод, что увеличение степени многочлена не дает эффекта. Общая схема проверки значимости полиномиального тренда более высокой степени приведена в табл. 5.20.

Таблица 5.20. Общая схема проверки эффективности увеличения степени полинома в уравнении регрессии

Источник изменчивости	Сумма квадратов	Число степеней свободы	Среднее квадратов	F-критерий
Уравнение регрессии степени (p–1)	$SS_{\text{RP+1}}$	M	$MS_{\text{RP+1}}$	$\frac{MS_{RP+1}}{MS_{DP+1}}$*
Отклонения от уравнения регрессии степени (p+1)	$SS_{\text{DP+1}}$	N–m–1	$MS_{\text{DP+1}}$	
Уравнение регрессии степени p	SS_{RP}	K	MS_{RP}	$\frac{MS_{RP}}{MS_{DP}}$**
Отклонения от уравнения регрессии степени p	SS_{DP}	N–k–1	MS_{DP}	
Увеличение степени уравнения регрессии от p до (p+1)	SS_{RT}=–$SS_{\text{RP+1}}$–SS_{RP}	M–k	MS_{RT}	$\frac{MS_{RP}}{MS_{DP+1}}$***
Общая изменчивость	SS_{T}	N–1		

* Проверка значимости поверхности тренда (p+1) степени.

** Проверка значимости поверхности тренда p-степени.

*** Проверка эффективности увеличения степени полинома от p до p–1.

Таким образом, проводится проверка следующей гипотезы:

$$H_0 : \beta_{\text{k+1}} = \beta_{\text{k+2}} = \ldots = \beta_{\text{m}} = 0 \qquad (5.90)$$

при альтернативе

$$H_1 : \beta_{\text{k+1}}, \beta_{\text{k+2}}, \ldots, \beta_{\text{m}} = 0.$$

Согласно нулевой гипотезе, все коэффициенты регрессии, начиная с номера k+1, равны нулю, поэтому введение в уравнение регрессии членов с номерами, превышающими k, не дает никакого эффекта (не следует забывать, что полиному степени p соответствует k коэффициентов регрессии, а полиному степени $(p+1) - m$ коэффициентов). Если вычисленное значение F превышает табличное, то гипотеза отклоняется. Проверочная процедура для нелинейного случая подробно описана в работе Ли [50].

В ряде геологических задач возникает потребность оценить эффект, обусловленный одним коэффициентом регрессии в уравнении, описывающем поверхность тренда. Такую оценку можно провести путем простого устранения данного члена полинома с последующим повторным вычислением сумм квадратов для регрессии и отклонения. Вклад исключенного члена является разностью двух сумм квадратов. Значимость этого члена можно проверить с помощью вычисления отношения среднего квадрата для уравнения с исключенным членом и среднего квадрата для полного уравнения регрессии. F-отношение имеет числа степеней свободы, равные 1 и (n–m–1). В табл. 5.21 приведена схема дисперсионного анализа (**ANOVA**) для проверки значимости одного исключенного коэффициента.

Таблица 5.21. Дисперсионный анализ. Проверка значимости одного исключенного коэффициента; полное уравнение полиномиальной регрессии содержит m коэффициентов, не считая члена b_0; после исключения коэффициента с номером k уравнение регрессии содержит m–1 коэффициентов; число наблюдений равно n

Источник изменчивости	Сумма квадратов	Число степеней свободы	Среднее квадратов	F-критерий
Регрессия всех членов	SS_R	m	MS_R	$\frac{MS_R}{MS_D}$*
Отклонение	SS_D	$n–m–1$	MS_D	
Регрессия после исключения k-го члена	SS_{R-1}	$m–1$	MS_{R-1}	$\frac{MS_{R-1}}{MS_{D-1}}$**
Отклонение	SS_{D-1}	$n–m–2$	MS_{D-1}	
Регрессия только k-го члена	$SS_{Rk} = SS_R – SS_{R-1}$	1	MS_{RK}	$\frac{MS_{RK}}{MS_D}$***
Сумма	SS_T	$n–1$		

* Критерий значимости тренд-поверхности p-й степени.

** Критерий значимости тренд-поверхности p-й степени без k-го члена.

*** Критерий значимости одного k-го коэффициента.

При добавлении новых переменных к каждому члену уравнения регрессии можно применить тот же критерий, вычисляя приращение SS_R. Однако этот прием не рекомендуется применять, так как имеется тенденция после появления нескольких последовательных незначащих коэффициентов считать все следующие члены незначащими, хотя это не всегда так. В анализе поверхностей тренда после прибавления полного набора членов более высокого порядка для их исключения требуется индивидуальная проверка каждого из них. Сокращенные совокупности членов высокого порядка нельзя приписать слепо, если на то нет особых оснований. В одном примере из-за ограничений, связанных с машиной, для данных по нефти была построена «гиперповерхность» третьего порядка, уравнение которой не содержало членов третьей степени с разными переменными. Это уравнение сильно отличалось от уравнения, полученного по тем же данным на более мощной ЭВМ с использованием программы построения полного кубического уравнения. Более того, если коэффициенты корреляции членов низкого порядка малы, то добавленные члены имеют тенденцию быть значимыми.

Два вышеприведенных множества данных (см. табл. 5.15 и 5.18) характерны для задач структурного тренд-анализа. Цель обоих исследований состоит в нахождении площадей, на которых структурные поверхности можно представить полиномиальными уравнениями. В этих задачах распределение ошибки таково, что пригодность критериев значимости для коэффициентов уравнения регрессии вызывает подозрение. Однако в следующем примере условия сбора данных и проведения эксперимента, по-видимому, хорошо согласованы с условиями применимости регрессионных критериев.

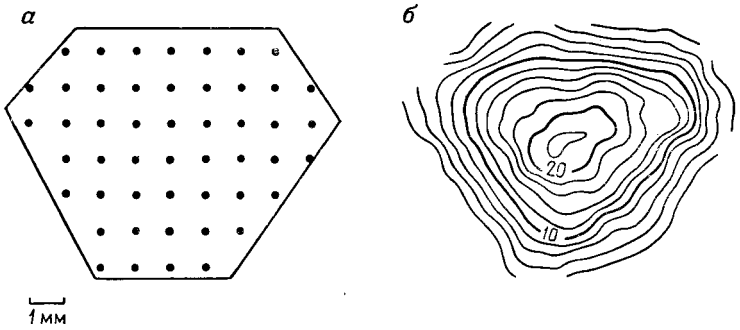


Рис. 5.83. (*а*) Расположение точек анализа в кристалле сфалерита (анализ выполнен под электронным микроскопом); (*б*) изолинии, характеризующие содержание железа

На рис. 5.83 изображена плоская проекция одного кристалла сфалерита, найденного в шахте на севере Мексики. Исследователя интересовало содержание железа в кристалле. Кристалл был ос-

торожно расколот по центру, а поверхность отполирована. С помощью электронного микроанализатора определялось содержание железа на участках размером в 1 нм в поперечнике через интервал в 1 мм. Сеть точек анализа изображена на рис. 5.83,*а*; изолинии полученных значений представлены на рис. 5.83,*б*; соответствующие данные приведены в табл. 5.22. Хотя использование данных о содержании железа в сфалерите в качестве температурного индикатора и подвергалось критике из-за возможного существования неравновесных условий во время кристаллизации, все же многие исследователи считают, что растущая грань кристалла все время находится в равновесии с рудообразующим раствором. Поэтому средний состав кристалла может лишь неадекватно служить температурным указателем, но состав последовательных срезов кристалла позволяет определить кривую температурного изменения. Простейшая модель распределения железа в кристалле отражает постепенное симметричное изменение (увеличение или уменьшение) его содержания при удалении от центра. Построение квадратичной полиномиальной регрессии значений содержания железа по координатам точек опробования является хорошим методом проверки этой модели. Используя данные табл. 5.22, постройте ряд поверхностей тренда для содержания железа. Вычислив необходимые суммы квадратов, примените аналогичную приведенной в табл. 5.20 схему **ANOVA** для проверки значимости уравнений полиномиальной регрессии.

Таблица 5.22. Содержание железа (в %) в участках диаметром 1 нм, расположенных по сетке 1×1 мм на плоскости кристалла сфалерита

X_1	X_2	Y	X_1	X_2	Y
2,0	1,0	3,1	2,0	4,0	6,4
3,0	1,0	4,6	3,0	4,0	14,6
4,0	1,0	5,8	4,0	4,0	17,6
5,0	1,0	7,2	5,0	4,0	21,2
6,0	1,0	8,4	6,0	4,0	21,0
7,0	1,0	6,3	7,0	4,0	13,4
8,0	1,0	2,4	8,0	4,0	7,5
1,0	2,0	2,5	9,0	4,0	0,4
2,0	2,0	10,2	2,0	5,0	3,1
3,0	2,0	12,8	3,0	5,0	8,6
4,0	2,0	16,1	4,0	5,0	15,0
5,0	2,0	14,2	5,0	5,0	16,2
6,0	2,0	15,1	6,0	5,0	14,8
7,0	2,0	12,8	7,0	5,0	9,8
8,0	2,0	9,0	8,0	5,0	3,1
9,0	2,0	5,3	3,0	6,0	5,0
1,0	3,0	4,3	4,0	6,0	7,2
2,0	3,0	14,1	5,0	6,0	12,3
3,0	3,0	15,6	6,0	6,0	10,6
4,0	3,0	20,2	7,0	6,0	4,5
5,0	3,0	20,6	3,0	7,0	0,6
6,0	3,0	18,5	4,0	7,0	2.4
7,0	3,0	16,2	5,0	7,0	3,5
8,0	3,0	10,2	6,0	7,0	4,7
9,0	3,0	4,6			

Две модели поверхностей тренда

Читатель, вероятно, отметил, что выше было рассмотрено два в корне различных типа геологических задач, решаемых с использованием методов тренд-анализа. С одной стороны, целью построения поверхностей тренда по структурным данным является выявление «локальных структур». Эмпирически было доказано, что в бассейне осадконакопления эти отклонения от поверхности тренда могут быть структурно или гидродинамически ассоциированы с нефтяными ловушками. С другой стороны, регрессионные поверхности использовались для определения регионального тренда по петрологическим и геохимическим данным. В этих двух приложениях различны как задачи, так и основные допущения, но метод по-прежнему остается общим.

Поверхности тренда, подбираемые к структурным данным, можно представить уравнением

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_{2^n} + (\gamma_i + \varepsilon_i) \qquad (5.91)$$

которое показывает, что данное наблюдение (абсолютная отметка кровли изучаемого слоя) равно сумме постоянного члена, связанного со средними значениями географических координат, плюс полиномиальное разложение степени p этих координат, плюс локальная компонента, плюс случайная компонента. Обычно последние два члена совмещаются и исследуются в совокупности.

Наоборот, поверхность тренда, подбираемая для петрографических или подобных данных, обычно описывается следующей простой моделью, называемой уравнением поверхности отклика

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_{2^p} + \varepsilon_i \qquad (5.92)$$

которое во всех отношениях похоже на уравнение поверхности тренда, но в нем отсутствует локальная компонента γ_i. В этом случае представляет интерес исследование природы тренда, т.е. получение оценок для коэффициентов β_i полинома.

Петрографические и геохимические переменные обычно характеризуются высокой дисперсией между повторениями. Эта изменчивость возникает в силу неоднородности в пределах анализируемых выборок, локальной или мелкомасштабной изменчивости в составе (в масштабе, большем, чем проба, но меньшем, чем интервал между пробами), а также из-за наличия аналитических или инструментальных ошибок. Обычно последние совмещаются, а это в свою очередь приводит к тому, что ошибку наблюдения можно рассматривать как нормально распределенную случайную величину. Хотя каждый источник изменчивости можно изолировать и измерения провести повторно, этого обычно не делается из соображений экономии, а также по ряду других причин.

Поверхность тренда или регрессии, построенная по географическим переменным, хорошо подходит к таким данным, если выполнены основные допущения. В последние входят требования, чтобы случайные компоненты ε были нормально распределены относительно регрессии и имели нулевое среднее и постоянную дисперсию. В свою очередь это означает, что компоненты ε независимы друг от друга. Если эти условия выполнены, то можно проверить значимость регрессии и затем сделать выводы относительно тренда. Подходящие для этих целей статистические критерии представлены в таблицах 5.19–5.21. Имеется также множество других статистических критериев, которые широко используются в сельском хозяйстве и инженерной химии: введение в эти методы дано Менденхоллом [56]. Кох и Линк [45] рассматривают вопрос о применимости одного из этих критериев в геологии. Важное свойство таких выводов состоит в том, что они применимы к тренду. Это изображено на рис. 5.84, где показано, что наблюдаемые значения У, попадают внутрь интервала значений отклонений относительно регрессии.

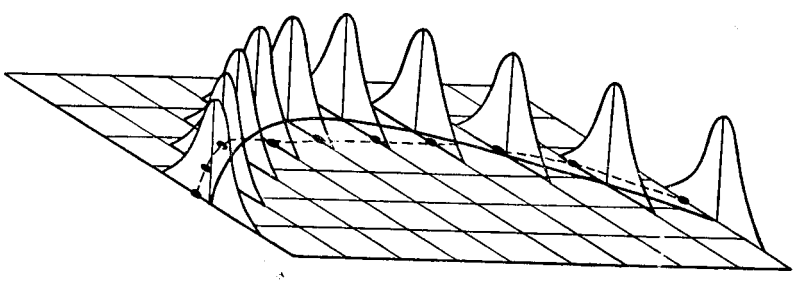


Рис. 5.84. Распределение случайной компоненты относительно линии регрессии в полиномиальной модели. Наблюдения, указанные пунктиром, предполагаются лежащими внутри области отклонении относительно линии регрессии

При подборе поверхностей тренда к структурным данным наблюдения (обычно это абсолютная отметка кровли пласта) не повторяются. Действительно, если пробурена одна скважина, то

обычно вблизи нее нецелесообразно бурить другую, так как последняя даст те же результаты. Повторные измерения в скважине могут колебаться в зависимости от глубины и общей протяженности скважины, но этот источник экспериментальной ошибки будет на один (или более) порядок меньше, чем отклонения в анализе поверхностей тренда. Отсутствие повторных наблюдений означает, что локальную изменчивость нельзя учесть. Однако необходимо отметить, что этот источник ошибки будет также немаловажен, так как бурение скважины нельзя считать выборкой из совокупности поверхностей. Существует только одна поверхность кровли, слоя, и возникает только одна дисперсия, связанная со столь малыми значениями ошибки, что ею можно пренебречь. Таким образом, дисперсию остатка в анализе поверхностей тренда можно объяснить недостаточным приближением.

В терминах нашей модели эквивалентное высказывание заключается в том, что ε_i пренебрежимо мало по сравнению с γ_i Хотя случайная компонента ε_i имеет нулевое среднее значение и является независимой для всех значений Y_i, ее нельзя отделить, так как мы не делаем повторных измерений. Совмещенный член $\gamma_i + \varepsilon_i$ также имеет нулевое среднее, однако в общем случае не является независимым для всех значении Y_i. На самом деле цель анализа заключается в определении областей данного размера по X_1 и X_2, над которыми член $(\gamma_i + \varepsilon_i)$ коррелирован. На рис. 5.85 указано теоретическое распределение величин ε_i относительно структурной поверхности. В большинстве случаев отклонение поверхности от полиномиальной модели отражает не величину ошибки, а поведение локальной компоненты γ_i.

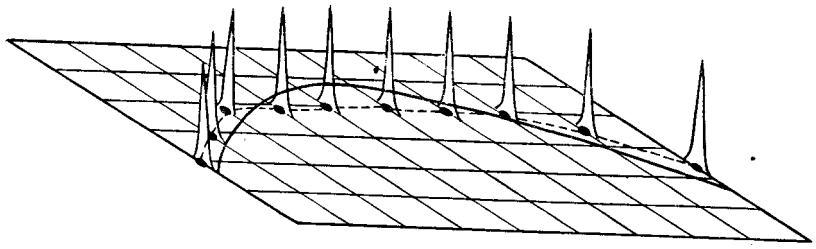


Рис. 5.83. Распределение случайной компоненты относительно наблюдаемой поверхности, соответствующей структурной модели поверхности тренда. Отклонения в повторениях точек концентрируются около средних значений, а не около линии, определенной полиномиальной моделью

Различие между этими двумя уравнениями для поверхностей тренда отражено в методе, с помощью которого изучается автокорреляция между остатками. В полиномиальном регрессионном анализе автокорреляция рассматривается как усиление ограничений, принятых в модели, и вводится для обоснования (или серьезного ослабления) выводов, полученных в результате анализа. Такое положение характерно для петрографических и геохимических данных, так как повторения приводят к тому, что ошибки распределены нормально при сравнительно простом уравнении регрессии. Хотя отсутствие согласования заставляет предположить наличие более сложных уравнений регрессии, все же ошибка достаточно велика для того, чтобы можно было считать, что все отклонения учтены.

В противоположность этому геологи, используя тренд-анализ структурных данных, находят области автокоррелированных остатков. Как было установлено выше, почти все структурные отклонения можно охарактеризовать отсутствием согласования, а наличие автокоррелированных остатков указывает на некоторую область, более широкую, чем интервал опробования, в котором поверхность отклоняется от полиномиальной модели в требуемом направлении. Как большие площади автокоррелированных остатков, так и одиночные точки со значительными отклонениями представляют интерес при разведке нефти, потому что указывают области, где локальные структуры (γ_i) имеют большое влияние. Отклонения не являются случайными величинами, поэтому обычные критерии значимости регрессии в этом случае неприменимы.

Особенности тренд-анализа

Теперь представляется целесообразным указать факторы, которые могут значительно испортить результаты тренд-анализа, т.е. любой тип анализа карт [10].

Ясно, что необходимо иметь некоторый способ контроля полученных результатов. Число данных точек должно как минимум превосходить число коэффициентов в полиномиальном уравнении, в противном случае построенное уравнение регрессии нельзя использовать. Если в качестве

статистических критериев использовать критерии скачков, то число контрольных точек определяет число степеней свободы и последнее должно быть довольно значительным для того, чтобы обоснованно применять F-критерий. Если число степеней свободы для отклонений невелико (вследствие того, что число полиномиальных коэффициентов близко к числу данных точек), то только крайне высокие значения коэффициентов корреляции могут быть приняты как значимые. Далее, мощность критерия (вероятность отсутствия ошибки второго рода) сильно убывает с уменьшением объема выборки. Конечно, число и расположение контрольных точек имеют прямое влияние на величину локальных отклонений, которые можно обнаружить при тренд-анализе структурных данных, и имеют связь с допуском в их определении.

Обыкновенно мы не рассматриваем контрольные точки, лежащие за пределами границ нашей карты. Зачастую, когда область карты немного выходит за пределы действительных границ данных точек, может быть несколько контрольных точек (конечно, необязательно), расположенных в точности на границах карты. В таких случаях не существует почти никаких ограничений на форму поверхности тренда вблизи от краев карты. Какой бы наклон ни был в контролируемой области, он экстраполируется без ограничений вдоль границ карты. Это явление называется «краевым эффектом». Если к имеющимся данным подбирается поверхность тренда высокого порядка, то экстраполируемые значения вблизи краев карты могут достигать астрономических размеров. Более слабые краевые эффекты возникают даже тогда, когда все поле карты вплоть до ее границы равномерно покрыто контрольными точками. Поэтому желательно иметь данные по площади за пределами карты. Последние образуют вокруг карты «буферную область», в которой сконцентрированы краевые эффекты; контрольные точки в этой области определяют форму поверхности тренда внутри поля карты. Ширина буферной области зависит в первую очередь от допустимой плотности контроля. Если карта содержит много контрольных точек, то достаточно узкой граничной полосы. Если контрольная плотность низкая, то для поглощения краевых эффектов нужен значительно более широкий пояс вокруг карты. Отметим, что краевые эффекты свойственны не только поверхностям тренда, но также встречаются при построении карт в изолиниях, поверхностей скользящего среднего, а также других типов аппроксимирующих поверхностей.

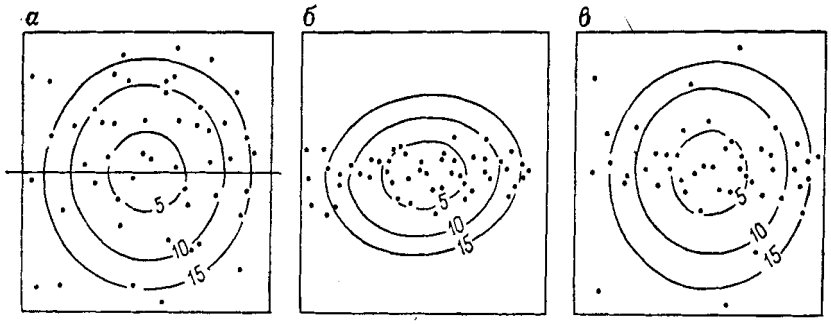


Рис. 5.86. Влияние распределения контрольных точек на поверхность тренда: *а* – исходная поверхность со случайным расположением контрольных точек; *б* – возмущенный тренд, полученный путем опробования исходной изучаемой поверхности в узкой полоске; *в* – почти правильный тренд, полученный путем добавления нескольких дополнительных точек за пределами узкой полосы наблюдений

Расположение данных точек в пределах карты также влияет на форму регрессии. Примеры на рис. 5.86 приведены для того, чтобы показать влияние распределения данных на полиномиальные поверхности тренда [25]. Множество точек было случайно размещено на поверхности, имеющей форму бассейна, и для них была найдена поверхность тренда второго порядка. Регрессионное уравнение затем было использовано для вычисления значений зависимой переменной в точках, размещенных в соответствии с различными выборочными планами. В идеале поверхности, построенные по этим точкам, должны быть идентичными поверхностям, для которых эти данные были получены. На рис. 5.86,*а* показана поверхность, построенная по случайно распределенным точкам; точность аппроксимации выше 95%, и тренд в сущности идентичен тренду оригинала. Однако на рис. 5.86,*б* выборочные точки распределены вдоль узкой полосы. Точность аппроксимации еще высокая (93%), но сама поверхность регрессии сильно смещена в направлении, параллельном выборке. Для того чтобы исправить это смещение, достаточно лишь нескольких контрольных точек вне этой полосы, как показано на рис. 5.86,*в*, где точность аппроксимации также 93%. Эти примеры показывают, что вид полиномиального уравнения сильно зависит от формы площади, занимаемой картой. Если данные не распределены приблизительно равномерно, то поверхность тренда вытягивается в направлении расположения точек. Напомним, что эти примеры составлены для идеализированных моделей,

т. е. моделей, не содержащих локальных или случайных компонент; смещения будут более заметными, если имеется хотя бы незначительный «шум».

Отметим, что в ряде работ содержатся предостережения, касающиеся плохого влияния групповых скоплений точек на поверхность тренда. Группирование контрольных точек причиняет особое беспокойство при разведке нефти, так как скважины наиболее густо расположены в пределах уже известных нефтяных полей. Такие площади могут оказывать значительное влияние на региональный тренд, хотя очевидно, что этот эффект не столь страшен, как это иногда кажется [25]. На рис. 5.87,*а* изображена модель поверхности тренда третьей степени, используемая для вычисления значений в данной точке с целью изучения эффекта группирования. Были взяты точки из различных групп и была сделана попытка воссоздать первоначальную поверхность. На рис. 5.87,*б* указана поверхность тренда, построенная по данным точкам, довольно явно сгруппированным. Построенная поверхность учитывает 99% исходной изменчивости. Еще более отчетливая группировка указана на рис. 5.87,*в*, но кубическая поверхность учитывает 100% суммарной изменчивости. Рисунки 5.87,*б* и 5.87,*в* в сущности идентичны исходной поверхности. Эти эксперименты показывают, что методы тренд-анализа, по-видимому, более устойчивы относительно группирования, чем это обычно предполагается. Напомним еще раз, что указанные критерии по существу не содержат шума и что при наличии локальных изменений возможны более серьезные изменения поверхности тренда.

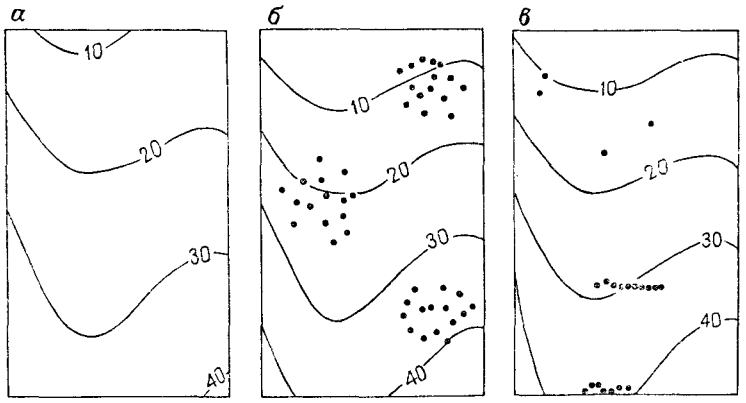


Рис. 5.87. Влияние группировки контрольных точек на поверхности тренда: *а* – исходная поверхность тренда третьей степени; *б* – поверхность тренда, полученная по контрольным точкам, расположенным группами; *в* – поверхность тренда, полученная по контрольным точкам, расположенным по частичным пересечениям проверяемой поверхности

ЧЕТЫРЕХМЕРНЫЙ ТРЕНД-АНАЛИЗ

Логическое обобщение полиномиального тренд-анализа требует введения третьей географической координаты (а также ее степеней и смешанных произведений с другими пространственными координатами) в качестве независимой переменной. Получаемая регрессия имеет несколько названий: «гиперповерхность», «изоплетная огибающая», «*U-V-W*-тренд» или просто четырехмерная поверхность тренда. В этом методе зависимая переменная, обычно процентное содержание некоторой компоненты, изменяется с востока на запад, с севера на юг и по высотной координате, и изолинии представляют собой линии уровня поверхностей более высоких размерностей. Результаты анализа изображаются в виде тела (в трехмерном пространстве), содержащего вложенные в него поверхности уровня. Эти поверхности интерпретируются так же, как и изолинии на обычной карте; они заключают равные значения состава. Так же как площади между двумя последовательными изолиниями на топографической карте заключают точки, имеющие почти одинаковую высоту, так и объем между двумя последовательными изоповерхностями четырехмерного тренда заключает точки, имеющие почти одинаковые значения состава.

В качестве простого примера рассмотрим данные, приведенные в табл. 5.23. В ней представлено процентное содержание оксида урана в небольшом карнотитовом теле в юрских отложениях плато Колорадо. Такие небольшие, но богатые рудные тела сформировались в результате замещения скоплений органических остатков в осадочной толще урановыми и ванадиевыми минералами. Рудное тело имеет форму эллипсоида; содержание урана увеличивается по направлению к центру. Протяженность рудного тела, грубо говоря, определяется общей длиной скоплений органических остатков, которые образуют ядро. Рудное тело было тщательно опробовано, а анализ собранных проб дал прекрасный пример данных, пригодных для построения четырехмерной поверхности тренда.

Таблица 5.23. Содержание оксида урана Y (в %) в карнотитовом теле

X_1	X_2	X_3	Y,%	X_1	X_2	X_3	Y,%
4,0	3,0	5,0	1,2	24,0	5,0	11,0	0,8
5,0	3,0	8,0	12,4	25,0	3,0	6,5	23,2
5,0	5,0	8.0	0,6	26,0	2,5	9,0	12,6
6,0	2,0	9,0	1,1	27,0	1,5	6,5	1,5
8,0	3,0	10,5	11,0	28,0	1,5	7,0	3,1
9,0	2,0	8,0	6,7	29,0	1,5	11,0	4,0
10,0	3,0	5,5	12,4	29,0	3,0	9,0	17,2
10,0	5,0	8,5	1,4	30,0	4,0	13,0	6,0
13,0	1,0	6,0	3,2	31,0	4,5	7,0	4,8
14,0	3,0	8,5	17,4	34,0	1,0	10,5	0,3
15,0	3,0	6,0	9,8	35,0	4,0	8,0	4,9
16,0	4,5	9,0	3,3	36,0	3,0	10,5	17,7
17,0	2,0	7,5	1,8	37,0	3,0	13,0	8,1
19,0	3,0	4,5	4,7	38,0	1,5	10,0	1,6
20,0	3,0	7,0	21,4	40,0	2,0	14,5	4,1
20,0	3,0	10,0	7,6	40,0	4,0	15,5	2,3
22,0	5,0	8,0	2,9	42,0	4,0	13,0	8,7

Координаты являются расстояниями (в футах) от произвольно выбранной точки в северо-западном углу рудного тела (X_1 – направление север–юг; X_2 – глубина; X_3 – направление восток–запад).

Легко убедиться, что как построение стереоскопической или перспективной карты, так и создание пространственной модели переменной, изменяющейся в трехмерном пространстве, являются нелегкой задачей. Трудность увеличивается, если значения переменной в контрольных точках содержат случайную компоненту. Традиционный метод построения состоит в создании карт для различных уровней или ряда сечений, оконтуривании их и совместном их изображении. К сожалению, метод изолиний оперирует с градиентами лишь одномерных или самое большее двумерных векторных полей. Почти весь процесс построения и сглаживания изображений субъективен, и модель, построенная по контурным уровням, может значительно отличаться от модели, построенной по оконтуренным сечениям, за исключением лишь самых простых примеров.

Четырехмерный тренд-анализ особенно полезен при построении таких моделей. Точки опробования могут располагаться в пространстве нерегулярно, и их не нужно проектировать на плоскости. Градиенты векторных полей могут рассматриваться независимо от направления. По подобранной методом наименьших квадратов поверхности тренда строится сглаженное изображение или простая форма, которая затем может быть расположена бесчисленным множеством способов. Вычислительные программы предусматривают получение «слоистых карт» (сечений, перпендикулярных любой из координатных плоскостей), которые затем собираются в модель «яичной корзины». Используя графопостроитель, многие более сложные программы выдают перспективные, изометрические или стереоскопические проекции.

Так как мы добавили одну независимую переменную, то для построения поверхности тренда первой степени мы должны решить систему четырех совместных уравнений. Нормальные уравнения приведены ниже. Для упрощения обозначений снова опущены пределы суммирования. Суммирование распространяется на все наблюдения по i от 1 до n.

$$\begin{aligned}
\sum Y &= b_0 n + b_1 \sum X_1 + b_2 \sum X_2 + b_3 \sum X_3 \\
\sum X_1 Y &= b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + b_3 \sum X_1 X_3 \\
\sum X_2 Y &= b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + b_3 \sum X_2 X_3 \\
\sum X_3 Y &= b_0 \sum X_3 + b_1 \sum X_1 X_3 + b_2 \sum X_2 X_3 + b_3 \sum X_3^2
\end{aligned} \qquad (5.93)$$

Если сравнить эту систему уравнений с системой (5.82), то легко убедиться, что она является прямым обобщением (4.13). Эту систему в матричном виде можно переписать следующим образом:

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 & \sum X_3 \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 & \sum X_1X_3 \\ \sum X_2 & \sum X_1X_2 & \sum X_2^2 & \sum X_1 \\ \sum X_3 & \sum X_1X_3 & \sum X_2X_3 & \sum X_3^2 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1Y \\ \sum X_2Y \\ \sum X_3Y \end{bmatrix} \qquad (5.94)$$

Построение аналогичной матрицы для получения уравнений поверхности тренда более высоких порядков требует тех же операций, какие используются в тренд-анализе, лишь с добавлением в смешанных произведениях дополнительной координаты. В качестве упражнения получите матричное уравнение для полиномиальной поверхности тренда второй степени с тремя независимыми переменными. Сколько членов содержит матрица системы? Этот вопрос затрагивает одно из наиболее серьезных ограничений этого метода, так как матрицы систем уравнений для поверхностей высокого порядка содержат очень много элементов. Следовательно, решения становятся неустойчивыми, и выход состоит в стандартизации данных с учетом шкалы матричных коэффициентов, после чего требуется использование наиболее эффективных и сложных процедур обращения матриц. Даже с этими предосторожностями часто невозможно получить надежное решение задачи построения поверхности тренда высокого порядка по заданному множеству точек.

На рис. 5.88,*а* изображена полиномиальная поверхность тренда первой степени, построенная по результатам определения содержаний урана. Тренд имеет форму параллельных пластин, между которыми заключены объемы равных содержаний. Как и следовало ожидать, точность линейной модели невысокая, т.е. только 2,3%. Поверхность второй степени изображена на рис. 5.88,*б*; точность этой модели значительно выше по сравнению с простой линейной моделью, и соответствующий ей вклад в суммарную изменчивость содержаний урана составляет примерно 59,9%. На рис. 5.88,*в* изображена поверхность регрессии третьей степени, имеющая приблизительно ту же форму, что и поверхность второй степени, однако суммарная изменчивость увеличилась до 79,3%.

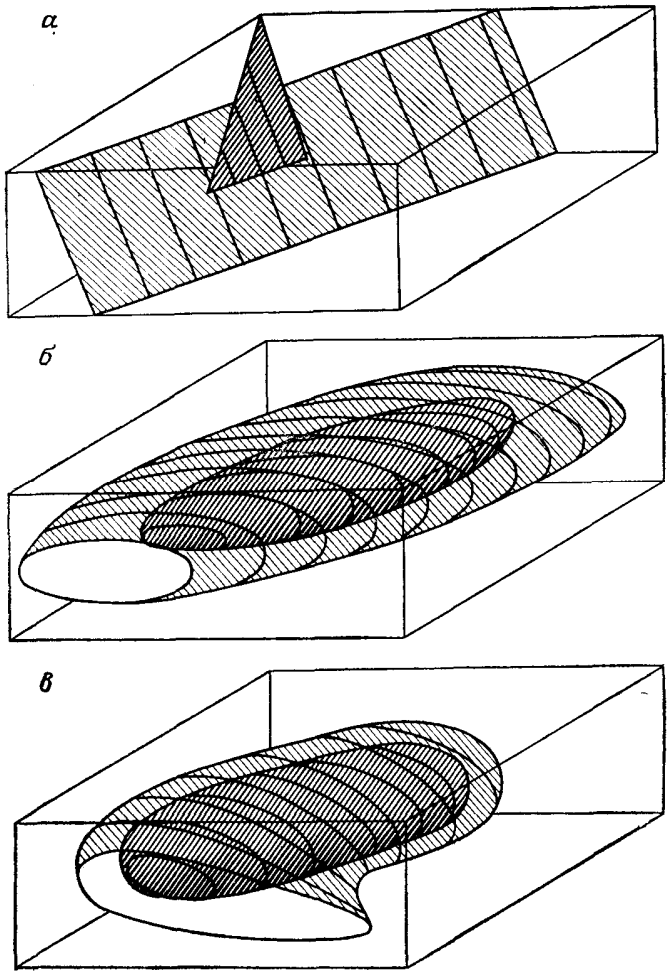


Рис. 5.88. Четырехмерные полиномиальные поверхности тренда для содержаний урана. Более темные внутренние тела – руда, содержащая свыше 10% оксида урана. В области, заключенной между внутренней и внешней огибающими поверхностями, руда содержит 5–10% оксида урана. В области, внешней к затемненному телу, руда содержит менее 5% оксида урана, *а* – поверхность первой степени; *б* – поверхность второй степени; *в* – поверхность третьей степени

Описанные выше эффекты, связанные с распределением исходного множества данных и имеющие значительное влияние на анализ поверхностей тренда, нельзя устранить и в четырехмерном случае, где они играют даже большую роль, так как нередко возможны затруднения в получении проб с требуемых интервалов глубины. В идеале выборка должна быть равномерно распределенной внутри куба или, в худшем случае, внутри толстой призмы, как это имело место в нашем примере. Хотя иногда и возможно реализовать равномерные схемы опробования, но чаще это очень трудно осуществить, так как интересующий нас район бывает в двух измерениях намного больше, чем в глубину. Такова ситуация, например, в задачах исследования распределения стратиграфических составляющих в пределах заданных интервалов. Даже если исследуемая область распространяется на большую глубину, недоступность проб часто заставляет нас ограничиваться рассмотрением тонкой плиты, что приводит к сокращению интервала на глубину. Это трехмерный аналог задачи, иллюстрация к которой приведена на рис. 5.88,*б*.

При изучении таких стратиграфических единиц, как формации, указанное затруднение преодолевается с помощью изменения масштаба, вертикальные расстояния обычно измеряются в футах, в то время как горизонтальные – в милях. Это приводит к растяжению в вертикальном направлении, и нарушения выборочной схемы несколько скрадываются на чрезмерно вытянутой вертикальной шкале. Конечно, если окончательную регрессию привести к истинной шкале измерений, то эти нарушения вновь появляются. Существует лишь один компенсирующий фактор, а именно: изменения в составе в пределах осадочной стратиграфической единицы более ярко выражены в направлении, параллельном стратификации, причем изменения вблизи кровли и подошвы слоя уменьшают влияние смещения. Однако при изучении состава тел, которые распространяются на большие глубины, например граниты магматического происхождения, смещения, вызванные ограниченным выборочным пространством, могут быть очень сильными. В этих случаях большие градиенты на поверхности вблизи нижней границы контрольных данных возникают в основном за счет смещений, обусловленных выборочной схемой.

Метод наименьших квадратов при построении уравнения регрессии представлен на рис. 5.90. Типичные тренды для каждой категории указаны на рис. 5.89. В следующей главе мы снова рассмотрим обобщение метода построения линейной регрессии для k независимых переменных. Конечно, это не географические переменные, хотя часть из них и является таковыми.

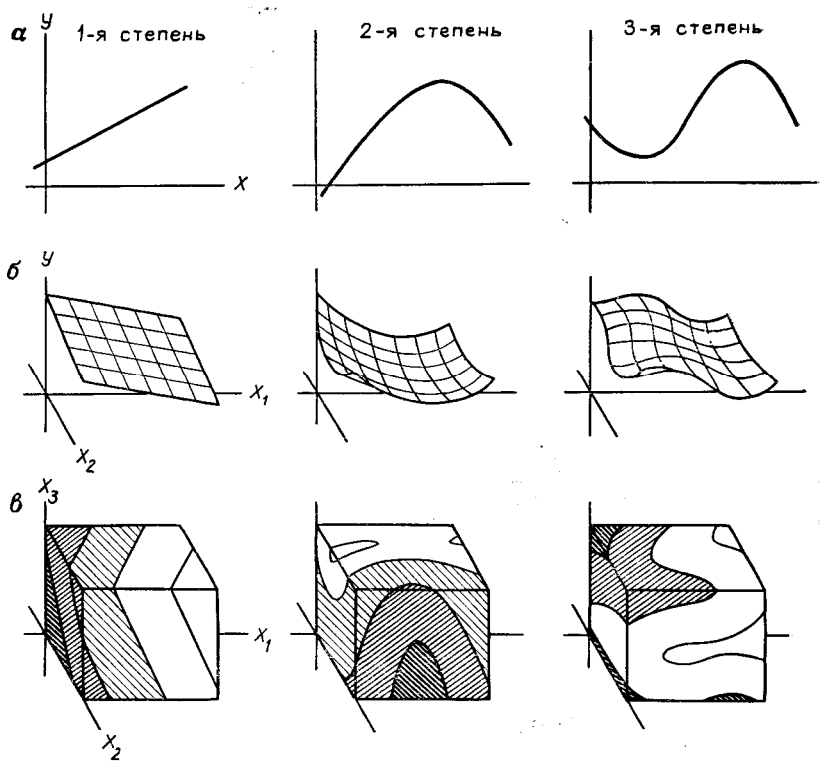


Рис. 5.89. Тренды от одной (*а*), двух (*б*) и трех (*в*) независимых переменных, полученные по полиномиальным уравнениям первой, второй и третьей степеней [37]

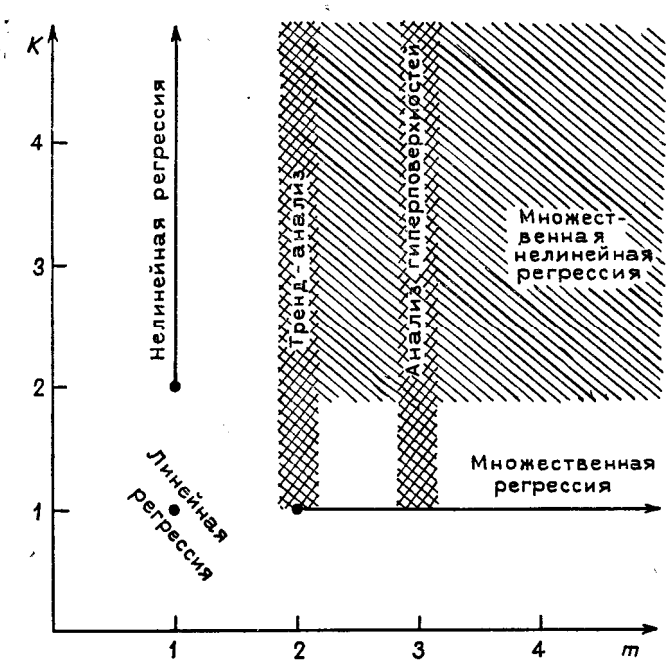


Рис. 5.90. Соотношение между нелинейной регрессией, анализом поверхностей тренда, четырехмерными поверхностями тренда и множественной регрессией: k – степень уравнения регрессии; m – число независимых переменных (размерность пространства)

ДВОЙНЫЕ РЯДЫ ФУРЬЕ

Уже отмечалось, что использование полиномиальных выражений в качестве аппроксимирующих функций для поверхности тренда объясняется главным образом той легкостью, с которой их можно вычислить. Опыт показывает, что распределение многих геологических переменных, особенно абсолютных отметок структурных горизонтов, хорошо описывается поверхностями тренда сравнительно небольшого порядка. Однако нет никаких особых оснований для использования в качестве аппроксимирующих функций многочленов, так как возможны и другие варианты выбора таких функций, которые нередко лучше подходят для тренд-анализа. В качестве таких функций широко используются двойные ряды Фурье.

Терминология, используемая в теории двойных рядов Фурье, заимствована главным образом из электротехники и анализа временных рядов. Необходимые определения и понятия были даны в гл. 4, где мы рассмотрели простые ряды Фурье; здесь мы дадим соответствующие обобщения этих понятий. Напомним, что сложный осциллирующий или периодический сигнал;. подобный электрическому, может быть представлен в виде суммы большого числа простых синусоидальных волн. Амплитуды и фазовые углы этих простых волновых форм можно определить исходя из условия близости рядов гармоник синусоидальных и косинусоидальных волн к исходным данным. Аналогичным образом сложную поверхность можно рассматривать как сумму двух взаимодействующих множеств двумерных синусоидальных волновых форм, каждая из которых содержит набор гармоник с различными амплитудами и фазовыми углами. В простейшем примере все гармоники в одном из направлений имеют нулевую амплитуду, и лишь одна гармоника в другом направлении имеет амплитуду больше нуля. Результирующая поверхность напоминает гофрированную железную крышу или ряд параллельных волн на гладкой поверхности воды (рис. 5.91,*а*). Кроме того, можно рассмотреть поверхность, в которой волны двух гармоник имеют одинаковое направление и нет ни одной волны в других направлениях (рис. 5.91,*б*). Очевидно, эти поверхности являются прямыми обобщениями одномерных сигналов, аналогичных тем, которые изображены на рис. 4.53.

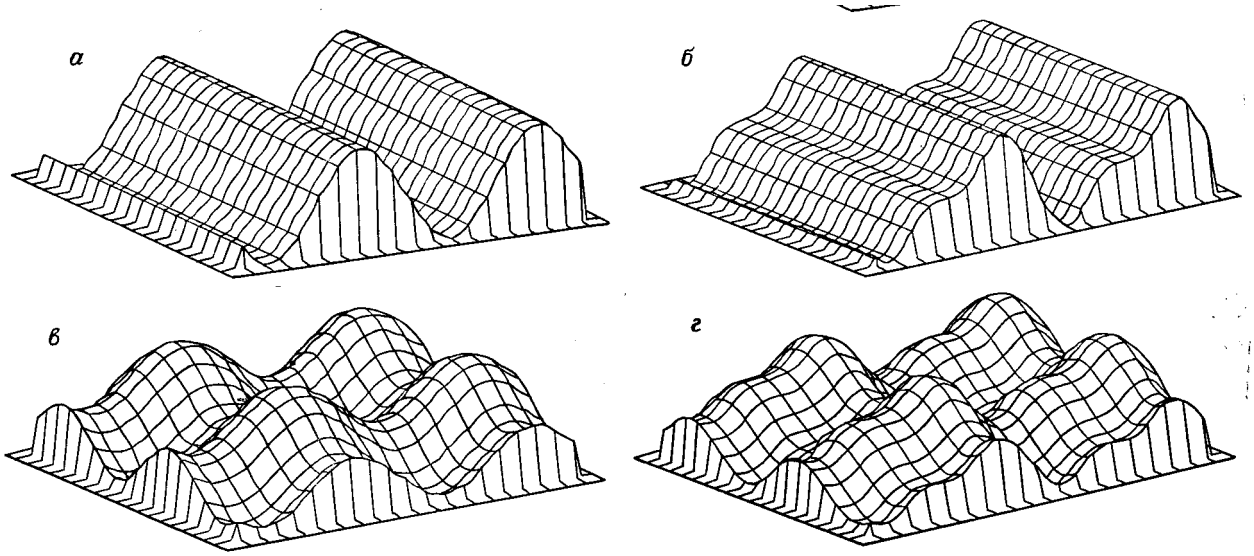


Рис. 5.91. Двумерные синусоидальные волны: *а* – единственная гармоника в направлении оси X_1; *б* – две гармоники в направлении оси X_1; *в* – единственная гармоника в направлении обеих осей X_1 и X_2; *г* – две гармоники в двух направлениях

Более сложная поверхность получается в результате наложения волн одной гармоники в одном направлении на волны другой, соответствующей иному направлению. Если две волны имеют одну и ту же амплитуду и длину волны, то результирующая поверхность напоминает вафлю или картон для упаковки яиц (рис. 5.91,*в*). Еще более, сложная поверхность изображена на рис. 5.91,*г*. Очевидно, что, используя двойные ряды Фурье, можно построить сколь угодно сложную поверхность.

Если основные длины волн в двух взаимно перпендикулярных направлениях меньше, чем размеры картируемой площади, то поверхность Фурье несколько раз уложится на этой площади. Такая процедура применяется лишь в очень редких случаях, например при изучении гармонических складок. Обычно же длина основной волны выбирается таким образом, чтобы она превышала длину карты, и повторение оказывалось ненужным. Выбор длины двух основных волн λ_1 и λ_2 и начала отсчета для двумерных рядов, вообще говоря, произволен. Если подобрать достаточное число гармоник двойного ряда Фурье, то при разных длинах основных волн и началах отсчета можно получить одинаковые карты тренда. Несмотря на то, что вклады отдельных гармоник будут меняться от карты к карте, форма поверхности тренда в сущности остается неизменной.

При построении двойных рядов Фурье используется в сущности та же модель, что и в анализе поверхностей тренда. Картируемая переменная Y_{ij} считается функцией линейного тренда среднего значения Y_{ij} по всей карте некоторой региональной компоненты, а также локальной компоненты, совмещенной со случайной компонентой. Именно эта модель неявно подразумевается при большинстве анализов, выполняемых на основании двойных рядов Фурье, особенно тех, которые посвящены исследованию структурной деформации слоистых осадочных пород [37]. Другие исследования, такие, как изучение распределения рудных тел в некотором регионе или выяснение химического или минерального состава пород, основываются на более известных моделях теории временных рядов [1]. При этом распределение Y_{ij} рассматривается как функция линейного тренда, различных периодических компонент и случайной компоненты. Как и в первой модели, линейный тренд лучше устранить до выполнения анализа с помощью двойных рядов Фурье. Эти две модели отличаются тем, что в первой просто стремятся разделить изменчивость на крупномасштабную и мелкомасштабную, в то время как во второй делаются попытки дать физическую интерпретацию более значимым гармоникам. Следовательно, модель временных рядов более предпочтительна, и так как в ней делаются. попытки извлечь больше информации, она соответственно требует более точных данных.

В другой модели мы предполагаем, что распределение в пределах карты может быть представлено двойным степенным рядом. Этот ряд является прямым обобщением ряда (4.87) и имеет вид

$$Y_{ij} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\alpha_{nm}\cos\frac{2k_1\pi X_{1i}}{n_1}\cos\frac{2k_1\pi X_{2j}}{n_2} + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\beta_{nm}\cos\frac{2k_1\pi X_{1i}}{n_1}\sin\frac{2k_1\pi X_{2j}}{n_2} + \\ + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\gamma_{nm}\sin\frac{2k_1\pi X_{1i}}{n_1}\cos\frac{2k_1\pi X_{2j}}{n_2} + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\delta_{nm}\sin\frac{2k_1\pi X_{1i}}{n_1}\sin\frac{2k_1\pi X_{2j}}{n_2} \qquad (5.95)$$

Обозначения такие же, как в (4.88) и (4.89); k_1 и k_2 – числа гармоник в направлениях X_1 и X_2, n_1 и n – числа равномерно расположенных наблюдений в каждом из двух направлений, X_{1i} и X_{2j} обозначают наблюдения в точках $X_1 = i$, $X_2 = j$. Как и в простом ряде Фурье, если предположить, что числа наблюдений от 0 до n_1 и от 0 до n_2 составляют полные множества доступных выборочных данных, то верхние пределы суммирования будут n_1 и n_2. Действительно, они определяют число фундаментальных гармоник. Хотя это равенство кажется громоздким, дробные выражения в каждом члене есть не что иное как простое преобразование X_{ij} в радианы.

Если ввести те же сокращенные обозначения для членов ряда, которые были использованы в гл. 4, то запись двойного ряда Фурье можно упростить. Пусть

$$C_{k_1} = \cos\frac{2k_1\pi X_{1i}}{n_1} \quad C^*_{k_1} = \cos\frac{2k_2\pi X_{2i}}{n_2} \quad S_{k_1} = \sin\frac{2k_1\pi X_{1i}}{n_1} \quad S^*_{k_1} = \sin\frac{2k_2\pi X_{2i}}{n_2}$$

Уравнение (5.95) можно теперь переписать следующим образом:

$$Y_{ij} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left(\alpha_{k_1k_2}C_{k_1}C^*_{k_1} + \beta_{k_1k_2}C_{k_1}S^*_{k_1} + \gamma_{k_1k_2}S_{k_1}C^*_{k_1} + \delta_{k_1k_2}S_{k_1}S^*_{k_1}\right) \qquad (5.96)$$

Записав двойные ряды Фурье в таком виде, можно получить систему нормальных уравнений для неизвестных коэффициентов и решить ее точно таким же образом, как это делалось для уравнений поверхности тренда [40]. Однако так как в уравнение входят дважды индексированные коэффициенты и в каждый член входит произведение двух рядов гармоник (один в направлении X_1, другой в направлении X_2), то даже в этих упрощенных обозначениях полученные уравнения оказываются сложными. Структура матрицы сумм и смешанных произведений станет яснее, если мы введем обозначения для ее строк и столбцов, как это было сделано для простого ряда Фурье в гл. 4. Матрица сумм и произведений $[A]$ входит в матричное уравнение $[A] \times [\beta] = [C]$, где $[\beta]$ – неизвестные коэффициенты, а матрица $[C]$ содержит суммы смешанных произведений между Y и различными гармониками. Мы находим неизвестную матрицу $[\beta]$ с помощью матричных операций обращения и умножения: $[\beta] = [A]^{-1} \times [C]$.

Матрица сумм и смешанных произведений вычисляется на основании разложения в ряд Фурье с любым заранее заданным числом гармоник n в направлении X_1 и m в направлении X_2. Столбцы и строки матрицы сумм и произведений тогда имеют вид

$$\begin{array}{c} \\ C_0C^*_0 \\ C_1C^*_0 \\ \vdots \\ C_3S^*_1 \\ \vdots \\ S_{k_1}S^*_{k_2} \end{array}\begin{array}{c} \begin{array}{ccccccc} C_0C^*_0 & & C_1C^*_0 & \cdots & C_3S^*_1 & \cdots & S_{k_1}S^*_{k_2} \end{array} \\ \begin{bmatrix} \sum(C_0C^*_0) & \sum C_0C^*_0C_1C^*_0 & \cdots & \sum C_0C^*_0C_3S^*_1 & \cdots & \sum C_0C^*_0S^*_{k_1}S^*_{k_2} \\ \sum(C_1C^*_0C_0C^*_0) & \sum(C_1C^*_0)^2 & \cdots & \sum C_1C^*_0C_3S^*_1 & \cdots & \sum C_1C^*_0S_{k_1}S_{k_2} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sum C_3S^*_1C_0C^*_0 & \sum C_3S^*_1C_1C^*_0 & \cdots & \sum(C_1S^*_1)^2 & \cdots & \sum C_3S^*_1S_{k_1}S^*_{k_2} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sum S_{k_1}S^*_{k_2}C_0C^*_0 & \sum S_{k_1}S^*_{k_2}C_1C^*_0 & \cdots & \sum S_{k_1}S^*_{k_2}C_3S^*_1 & \cdots & \sum(S_{k_1}S^*_{k_2})^2 \end{bmatrix} \end{array} \times \begin{bmatrix} \alpha_{00} \\ \alpha_{10} \\ \cdots \\ \beta_{31} \\ \cdots \\ \delta_{k_1k_2} \end{bmatrix} = \begin{bmatrix} Y \\ \sum YC_0C^*_0 \\ \sum YC_1C^*_0 \\ \vdots \\ \sum YC_3S^*_1 \\ \vdots \\ \sum YS_{k_1}S^*_{k_2} \end{bmatrix} \qquad (5.97)$$

Так как для каждой гармоники необходимо вычислить четыре коэффициента, то матричное уравнение при большом числе гармоник становится очень сложным. Например, вычисление коэффициентов для пяти гармоник в двух направлениях требует определения примерно сотни коэффициентов.

При обсуждении одномерных рядов Фурье (см. гл. 4) указывалось, что синус угла, равного нулю градусов, есть нуль; это приводит к тому, что члены

$$\sin\frac{2n\pi X_{1i}}{\lambda_1} \quad \text{и} \quad \sin\frac{2m\pi X_{2i}}{\lambda_2}$$

обращаются в нуль, если n или m равно нулю. Поэтому все члены ряда, содержащие синус нулевой гармоники, равны нулю, а один столбец и одна строка матрицы сумм и произведений обращаются в

нуль. Столбец и строка, соответствующие нулевым элементам матрицы, заштрихованы на рис. 5.92. C_0 и C_0^* равны 1,0, так как косинус нуля градусов равен единице. Это позволяет упростить выражение одной строки и одного столбца матрицы, содержащей эти члены. При расположении коэффициентов, изображенном на рис. 5.92, блок 0 содержит только один член, порождающий горизонтальную плоскость, соответствующую значению этого коэффициента. Блок 1 состоит из восьми членов, каждый из которых представляет поверхности, соответствующие длинам основных волн. Блоки 0 и 1 в совокупности составляют коэффициенты первой гармоники поверхности тренда. Блок 2 содержит 16 дополнительных членов, которые представляют поверхность гармоники, имеющей длины волн, равные половинам длин основных волн. Полная вторая гармоника поверхности тренда содержит коэффициенты блоков 0, 1 и 2. Каждая последующая гармоника поверхности тренда строится с помощью членов следующего блока.

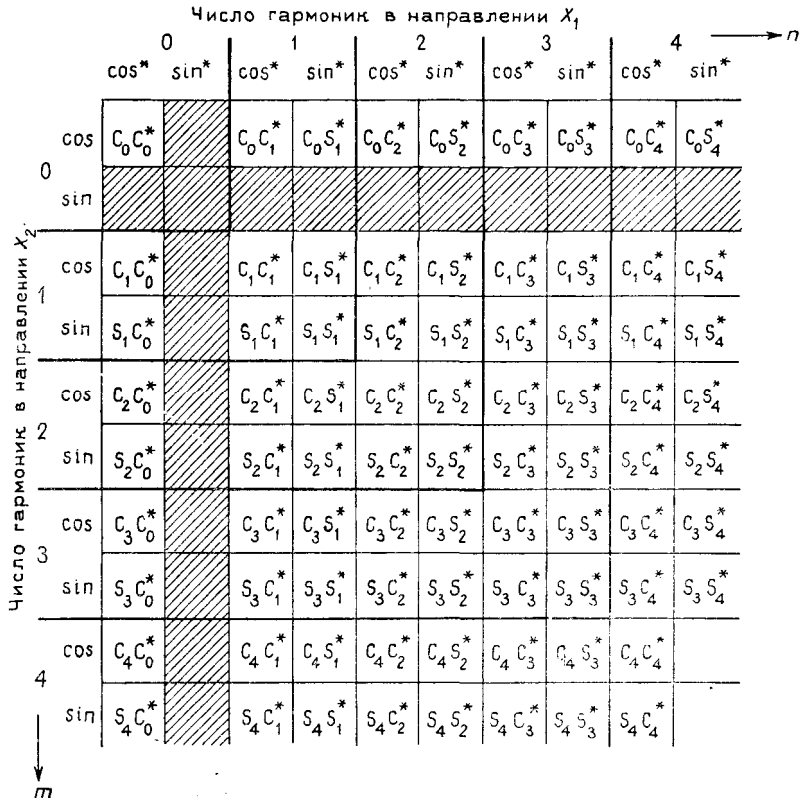


Рис. 5.92. Коэффициенты двойного ряда Фурье, расположенные в соответствии с длинами волн [40]. Заштрихованные квадраты соответствуют нулевым коэффициентам

Двойные ряды Фурье обычно применяются в классе задач, аналогичных тем, для которых использовался полиномиальный тренд-анализ. Аппроксимирующая функция, построенная по географическим координатам заданных точек, подбирается к нерегулярно расположенным в пространстве данным. При обобщении представления аппроксимирующей функции наша цель – разделить изменчивость данных на две компоненты: региональный тренд, представленный подобранной функцией, и локальные остатки, характеризующие отклонения от поверхности тренда. В тех случаях, когда данные содержат пространственно повторяющиеся элементы, ряды Фурье оказываются более подходящими аппроксимирующими функциями, чем степенные ряды. Так как точный подбор поверхности к выборочным данным совсем не обязателен (даже нежелателен), то для проведения анализа достаточно ограниченного числа гармоник.

Обычно в задачах тренд-анализа данные располагаются нерегулярно, и поэтому для вычисления коэффициентов Фурье нужно решить общую систему нормальных уравнений (5.97). В связи с тем что порядки матриц, с которыми приходится оперировать, оказываются очень большими, на малых ЭВМ удается вычислить лишь несколько гармоник. Например, на рис. 5.92 показано, что для получения двойного ряда Фурье всего лишь с тремя гармониками требуется вычислить 49 коэффициентов, а матрица, обращение которой требуется для нахождения этих коэффициентов, содержит 2450 элементов! При этом мы вынуждены учитывать не только допустимый объем памяти ЭВМ, но и ошибки округления, которые устрашающе растут при обращении матриц очень больших порядков. Из-за этих ограничений мы не пытались писать программу анализа с применением двойных рядов Фурье. Основные вопросы, связанные с построением такой программы, были рассмотрены при написании программы полиномиального тренд-анализа, и требуемая программа явилась бы лишь громоздким, но прямым ее обобщением.

Однако если наши наблюдения получены в результате опробования по регулярной сети, из этого вычислительного тупика можно найти выход. В этой ситуации все недиагональные члены

(смешанные произведения) нормальных уравнений обращаются в нуль. Матрица, которую требуется обратить, становится диагональной, и решение матричного уравнения осуществляется просто. Более того, в этом случае можно вычислить много гармоник, а не ограниченное их число, как это приходится делать при решении общих уравнений для двойного ряда Фурье. Это открывает возможности осуществления двумерного гармонического анализа и содержательного изучения спектров карт, фотографий и вообще различных изображений.

Методы двойных рядов Фурье обычно используются в двух больших группах геологических задач. Одна из них включает поиск значимых периодичностей в поведении концентраций металлов в пределах рудных зон и в размещении рудных тел в пределах металлогенической провинции. Многие месторождения твердых полезных ископаемых тесно связаны с системами разломов, которые в свою очередь являются следствием региональных деформаций. В расположении этих систем в больших регионах земной коры может проявиться некоторая периодичность или регулярность. Если это верно и удается получить некоторые оценки спектра разломов, то можно сделать и некоторые предсказания. Используя результаты анализа пространственных длин волн между известными рудными месторождениями, можно указать местоположения других возможных месторождений. Та же идея была использована, но в меньшем масштабе, при изучении отдельных рудных месторождений с целью определения перспектив развития рудника.

Широкое распространение в нефтяной геологии получило совместное применение двойных рядов Фурье и двумерных методов фильтрации [67]. При этом вычисляется двумерный энергетический спектр структурной карты в изолиниях, по которому определяются пространственные длины волн. Затем фильтры (являющиеся не чем иным как разновидностью скользящих средних) используются для отделения этих установленных длин волн и для выявления структурных характеристик заданного типа и ориентации для структурной карты. Выявленные положительные аномалии могут быть связаны с нефтью. Опытные исследователи могут обнаружить в осадочных толщах структуры, которые отражают деформацию подстилающих пород, расположенных на глубине многих тысяч футов. Вообще эти периодические компоненты приглушены и завуалированы. на обычной структурной карте, но там, где их можно выявить, они оказывают значительную помощь при разведке и предсказании месторождений на более глубоких горизонтах.

Вторая группа геологических задач связана с применением двумерного анализа Фурье при изучении размещения зерен минералов в шлифах и для исследования пористости пород в пределах нефтеносного бассейна [23]. В последней работе авторы стремились к получению простых численных характеристик для описания чрезвычайно сложных картин пористости в песчаниках и известняках и в целях их использования для оценки проницаемости этих пород. Такие характеристики были получены из четкого спектра изучаемой породы. Рассматриваемые данные представляют собой замеры оптической плотности в точках на фотографии шлифа, на которой зерна породы выглядят черными, а поры – белыми. Фотографии были подвергнуты числовой. обработке с помощью электронного устройства, которое измеряет оптическую плотность в тысячах точек, расположенных на регулярной пространственной сетке. Эти данные затем подвергаются преобразованию Фурье, в результате чего получаем двумерный спектр. Дальнейшие преобразования позволяют определить нужные параметры спектра, а последние используются в моделях, предсказывающих поведение флюида в изучаемых породах.

Недавно создан метод, обобщающий анализ Фурье для одной и двух переменных, названный быстрым преобразованием Фурье (*FFT*), а также вычислительный алгоритм, позволяющий значительно сократить время вычислений и облегчить практический анализ больших массивов данных. В сочетании с вычислительным устройством специального назначения этот алгоритм, обеспечивающий численное представление сейсмической записи, получил большое распространение в геофизической разведке. Методы, использующие быстрое преобразование Фурье, нашли применение в других областях анализа временных рядов и пространственного анализа, даже несмотря на то что потребовались дополнительные ограничения на объем и структуру рассматриваемого набора данных. Этот метод позволяет достичь большого быстродействия благодаря ряду специальных матричных преобразований, которые можно сравнить с построением «циклической» матрицы, в которой начало ряда значений помещается в центр, а данная последовательность «вращается вокруг» центра. Применяемые при этом операции состоят в умножении со сдвигом и сложении соседних элементов. Вычислительная схема этого метода, необходимая для составления программы *FFT*, выходит за рамки данной книги, и поэтому мы не будем на ней останавливаться. Однако как одномерная, так и двумерная программа *FFT* допускают очень широкую область использования и могут быть включены в

программу, которую мы изложим ниже, если объем данных окажется подходящим. Хорошее описание алгоритма *FFT* приводится в книге Кокрана и др. [16].

В гл. 4 указывалось, что в случае равномерного пространственного размещения данных можно построить систему уравнений, которая позволяет вычислить коэффициенты ряда Фурье непосредственно, без каких-либо матричных преобразований. Эти уравнения можно распространить на двумерный случай, хотя при этом вместо двух коэффициентов для каждой гармоники придется определять четыре. Тем не менее метод имеет значительное преимущество по сравнению с общим методом, так как позволяет вычислять большое число гармоник. Это значит, что мы можем произвольно выбирать множество длин основных волн и затем исследовать большое число гармоник для отыскания пространственных периодичностей. Введение в двойной анализ Фурье для сетей при геологическом опробовании дано в книге Харбуха и Мерриэма [37J. Математические основы двумерного анализа Фурье содержатся в книгах Коута и др. [18] и Райнера [64].

Если данные собраны по регулярной сети, мы можем произвольно определить длину фундаментальной волны в направлении оси X_1, а именно принять ее равной длине набора данных в этом направлении или n_1. Аналогично длину основной волны в направлении оси X_2 можно взять равной n_2. Далее легко вычислить коэффициенты любой гармоники с помощью уравнений

$$\begin{aligned}
\alpha_{k_1k_2} &= \frac{k}{n_1n_2}\sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} Y_{ij}\cos\frac{2\pi k_1X_{1i}}{n_1}\cos\frac{2\pi k_2X_{2j}}{n_2}\\
\beta_{k_1k_2} &= \frac{k}{n_1n_2}\sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} Y_{ij}\cos\frac{2\pi k_1X_{1i}}{n_1}\sin\frac{2\pi k_2X_{2j}}{n_2}\\
\gamma_{k_1k_2} &= \frac{k}{n_1n_2}\sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} Y_{ij}\sin\frac{2\pi k_1X_{1i}}{n_1}\cos\frac{2\pi k_2X_{2j}}{n_2}\\
\delta_{k_1k_2} &= \frac{k}{n_1n_2}\sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} Y_{ij}\sin\frac{2\pi k_1X_{1i}}{n_1}\sin\frac{2\pi k_2X_{2j}}{n_2}
\end{aligned} \qquad (5.98)$$

где k=1, если k_1=0 и k_2=0; k=2, если k=0 или k_2=0, но не одновременно; k=4, если k_1>0 и k_2>0; k_1 – гармоническое число в направлении X_1; k_2 – гармоническое число в направлении X_2; n_1 – число точек сети в направлении X_1; n_2 – число точек сети в направлении X_2. В этих формулах предполагается, что начало гармонического ряда соответствует началу координат системы переменных X_1 и X_2. Гармоники могут быть вычислены вплоть до значений $k_1=n_1/2$ и $k_2=n_2/2$, а гармоники с большими номерами могут быть оценены лишь по нескольким точкам на волну. Поэтому результаты для гармоник с большими номерами оказываются менее надежными, чем для гармоник с малыми номерами.

Вычислив коэффициенты Фурье, мы можем найти энергетический спектр по формуле, являющейся двумерным обобщением формулы (4.96):

$$S^2_{k_1k_2} = \left(\alpha^2_{k_1k_2} + \beta^2_{k_1k_2} + \gamma^2_{k_1k_2} + \delta^2_{k_1k_2}\right)\Big/4 \qquad (5.99)$$

Читатель, наверное, помнит, что s^2_{nm} было выражением для дисперсии и характеризовало вклад k_1-й и k_2-й гармоник в общую дисперсию величины Y. Если мы обозначим общую дисперсию через S^2_{Y}, то процентный вклад каждой гармоники выразится формулой

$$S^2_{k_1,k_2}\Big/S^2_2\cdot 100\% \qquad (5.100)$$

Однако исходный спектр является только оценкой вкладов гармоник в дисперсию, и, подобно всем выборочным оценкам, они могут значительно отклоняться от спектра истинной совокупности. Исходный спектр становится более устойчивым или лучше оценивает спектр совокупности при стремлении объема выборки к бесконечности, но мы не можем бесконечно брать пробы и на некотором этапе вынуждены закончить анализ. Нашу оценку выборочного спектра можно улучшить путем сглаживания, так как соседние гармоники оказываются близкими, и процедура усреднения приводит к тому, что они сходятся к общему значению. Сглаживание выполняется с помощью метода, рассмотренного в разделе о двумерном скользящем среднем. Веса в формуле для $\overline{S}^2_{k_1k_2}$, используемой для сглаживания, выбираются следующим образом:

$$\overline{S}^2_{k_1k_2} = \frac{1}{16}\left(S^2_{k_1-1,k_2-1} + S^2_{k_1-1,k_2+1} + S^2_{k_1+1,k_2-1} + S^2_{k_1+1,k_2+1}\right) + \frac{1}{8}\left(S^2_{k_1-1,k_2} + S^2_{k_1+1,k_2} + S^2_{k_1,k_2-1} + S^2_{k_1,k_2+1}\right) + \frac{1}{4}S^2_{k_1,k_2} \qquad (5.101)$$

Рассмотрим пример использования двойных рядов Фурье, в котором представляют интерес как тренд Фурье, так и спектр. Волноприбойные знаки, сохранившиеся на плоскостях напластования песчаников, могут содержать ценную палеогеографическую информацию. В стратиграфических исследованиях принято измерять главное направление ориентации, длину волны и амплитуду волноприбойных знаков. Иногда измерения, сделанные на многих обнажениях, можно обобщить для построения карты направления древних течений с целью получения выводов о береговых линиях. Было сделано несколько попыток детального анализа схемы волноприбойных знаков, изученных на большой площади одного обнажения. На рис. 5.93 представлена схема расположения волноприбойных знаков на плите песчаника докембрийского возраста из восточной части штата Айдахо. Типичные волноприбойные знаки на этой плите имеют амплитуду около 3 см и длину волны около 15 см. Диаграмма была получена в результате оконтуривания измерений, сделанных по сети 2х2 см. На поверхность породы была помещена деревянная рама, по которой передвигалась линейка для определения последовательности пересечений. Через каждые 2 см определялось расстояние от линейки до поверхности породы. Так как полученные данные оказались весьма объемистыми, то они представлены здесь лишь в форме контурной диаграммы.

Анализ этих данных с помощью двойных рядов Фурье дает спектр, изображенный на рис. 5.94, который был получен как двумерное обобщение диаграмм спектра, приведенных в гл. 4. На диаграмме (см. рис. 5.94) видны два больших пика, представляющих значимые гармоники.

Используя информацию о спектре, мы можем определить индивидуальные волновые формы, дающие наибольший вклад в схему волноприбойных знаков. На рис. 5.95,*а* представлена объемная диаграмма наибольшей спектральной составляющей, полученная с помощью ЭВМ; рис. 5.95,*б* – аналогичная диаграмма второй по величине волновой формы. Их комбинация изображена на рис. 5.95,*а*. Карта поверхности, полученная как комбинация этих двух гармоник, изображена на рис. 5.96. Если сравнить карту с исходной картой, представленной на рис. 5.93, то можно заметить, что наиболее существенные черты исходной карты в этой простой модели, использующей лишь две гармоники, сохранились.

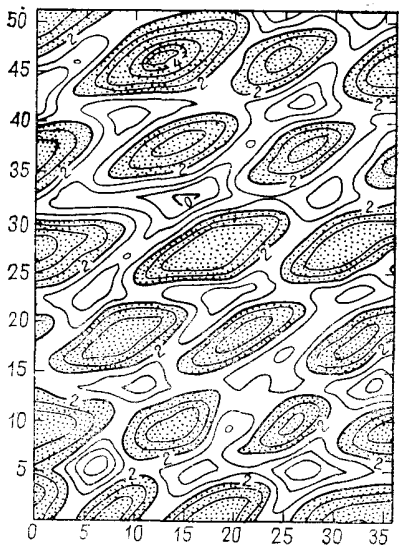


Рис. 5.93. Диаграмма в изолиниях волноприбойных знаков на плите докембрийского песчаника из Айдахо. Заштрихованы углубления более 2 см. Масштаб на осях – 2 см в единице

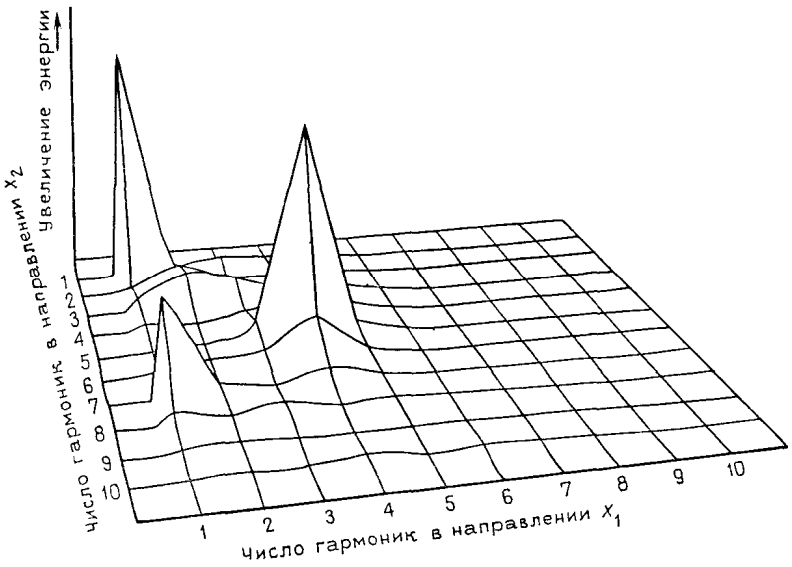


Рис. 5.94. Двумерный спектр волноприбойных знаков. Выделен значительный вклад двух доминирующих совокупностей длин волн

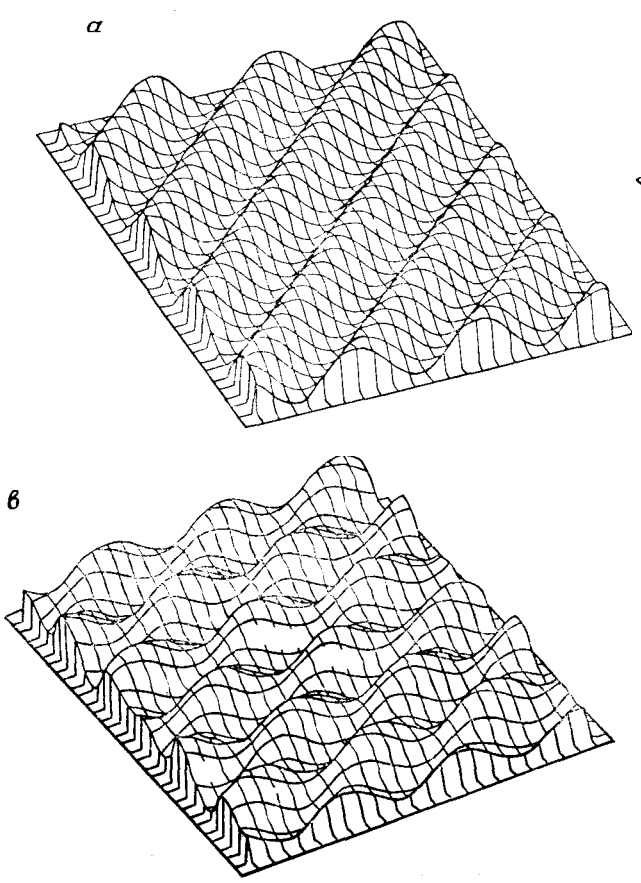


Рис. 5.95. Восстановление формы волноприбойных знаков в песчанике: *а* – перспективная диаграмма самой большой спектральной компоненты; *б* – перспективная диаграмма второй главной волновой формы; *в* – поверхность, образованная комбинацией двух доминирующих волновых форм

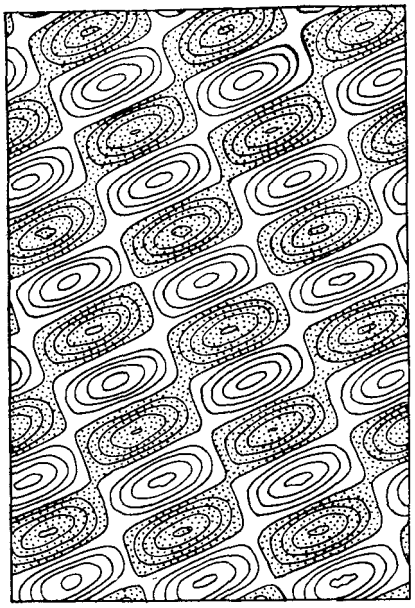


Рис. 5.96. Реконструкция изолиний двух доминирующих волновых форм для волноприбойных знаков. Сравните с исходными данными, изображенными на рис. 5.93

СРАВНЕНИЕ КАРТ

Обычная задача большинства геологических исследований заключается в сравнении друг с другом двух или более карт некоторой территории. Это могут быть карты различных петрографических и минералогических составляющих, различных параметров, характеризующих размер зерен, или структурные карты различных геологических горизонтов. Прямые сравнения поверхностей тренда различных горизонтов делались для выявления периодов структурной деформации. Связанная с этой, но более трудная задача заключается в сравнении карт одного параметра для двух и более регионов. Геологи постоянно сравнивают карты различных районов и выделяют характерные геологические особенности. Конечно, искусство интерпретации геолога основано на его предшествующем опыте и его способности мысленно сравнить новый район с тем, который он изучал в прошлом. Несмотря на то что количественные сравнения картированных переменных оказываются потенциально полезными для геологов, существует весьма ограниченное число попыток измерять сходство между картами. К счастью, много лет назад этой задачей заинтересовались географы, которые создали несколько полезных методов сравнения пространственных распределений, например, метод вычисления усредненной меры сходства и метод построения карты расхождений между двумя поверхностями.

По-видимому, роль сравнения карт должна значительно возрасти в будущем, так как интерпретация большого числа данных, получаемых со спутников Земли, требует развития схем автоматического распознавания и анализа карт. Соответствующие алгоритмы должны быть созданы геологами и другими учеными, исследующими Землю, которые обладают достаточными знаниями для интерпретации этих данных. В свою очередь геологи должны овладеть методами количественной обработки и их систематизации с тем, чтобы машины смогли взять часть работы на себя. Если этого не сделать, то мы будем буквально погребены под грудой карт, фотографий, полученных со спутников, с орбитальной геофизической аппаратуры и с других экзотических средств будущего.

Общее сходство

По-видимому, простейший путь сравнения двух карт состоит в вычислении коэффициентов корреляции между картируемыми переменными. Если две карты построены для переменных, измеренных в одних и тех же точках, то этот метод состоит в вычислении коэффициента корреляции между картированной переменной 1 и картированной переменной 2, при этом совсем не принимается во внимание расположение точек. Коэффициент корреляции вычисляется по формуле (2.24) и дает меру общего соответствия между двумя переменными.

На побережье Северной Каролины (рис. 5.97,*а*) был собран ряд донных проб из осадка лимана. С целью изучения распределения размеров зерен осадка, вычислены обычные статистические характеристики (среднее, стандартное отклонение, асимметрия и эксцесс). Многие авторы высказали предположение, что различные комбинации моментов распределения – эффективное средство установления условий осадконакопления. При исследовании связей между различными статистиками была сделана попытка установить, достаточны ли они для определения в лимане областей, в которых условия осадконакопления существенно различаются.

На рис. 5.97,*б* представлена карта стандартного отклонения размеров зерен осадка в лимане, на рис. 5.97,*в* – карта асимметрии распределения размеров зерен для той же выборки. Высказывалось предположение, что асимметрия и стандартное отклонение, рассматриваемые вместе, являются эффективным средством классификации осадков, поэтому важно, чтобы была получена оценка соотношения между ними. Так как обе переменные были вычислены для исходных данных, собранных в одних и тех же точках, коэффициент корреляции может быть вычислен прямо по исходным данным. Коэффициент корреляции между двумя картами r равен 0,52. Заметим, что географические координаты точек не рассматривались. Это – самый большой недостаток метода общего сравнения, так как данная корреляция может отражать степень соответствия двух картируемых площадей целиком или же быть результатом большого отклонения в малой области карты.

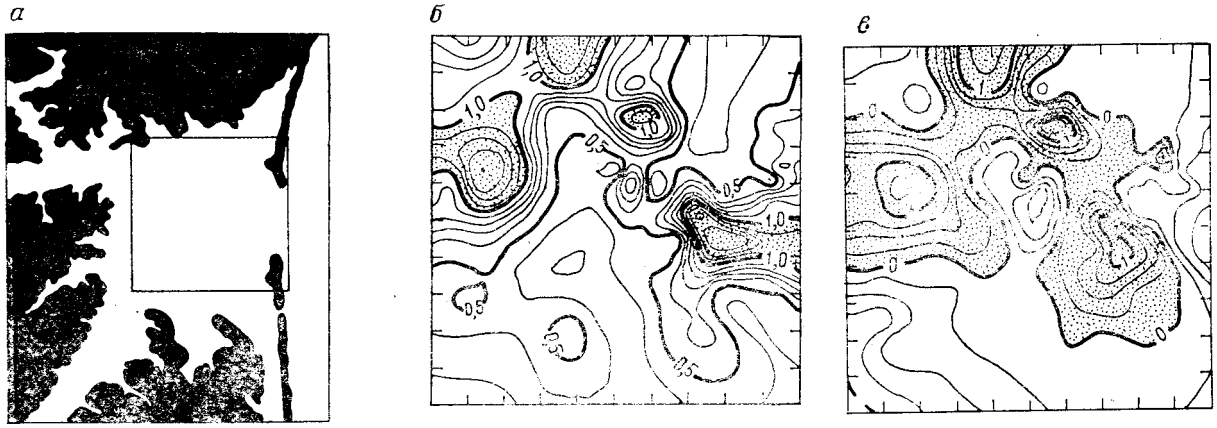


Рис. 5.97. Изменчивость размеров частиц донных осадков на побережье залива в Северной Каролине: *а* – карта, показывающая связь изучаемой площади с побережьем залива; *б* – стандартизованное отклонение размеров зерен (в φ-единицах); заштрихованы площади, имеющие стандартные отклонения размеров зерен, большие 1,0; *в* – асимметрия распределения размеров зерен; заштрихованы площади, соответствующие осадкам, имеющим положительную асимметрию

Предположим, что теперь мы хотим сравнить две карты, в которых переменные измерены не в одних и тех же точках. Один из возможных способов получения меры общего сходства – оценка значений двух переменных на множестве точек сетки, общей для двух карт. Так как процедура оценок связана с появлением ошибок, величины которых мы не в состоянии оценить, то желательно разместить точки получения оценок настолько близко к точкам опробования, насколько это возможно. Действительно, одна из наиболее эффективных схем состоит в получении оценок одной переменной по ее карте в изолиниях в каждой точке расположения пробы второй переменной. После этого можно оценить корреляцию между двумя переменными. Однако никакой статистической значимости приписать полученному коэффициенту корреляции мы не можем, так как он целиком основан на интерполированных значениях. Очевидно, надежность корреляции повышается с увеличением плотности контрольных точек.

Карты сходства

Пример простейшего сравнения двух карт переменной одного типа в одной и той же области – карта изопахит, полученная по двум структурным картам в изолиниях. Карта изопахит – это карта мощностей или разностей абсолютных отметок двух поверхностей. Аналогичные карты можно построить для разностей таких переменных, как размер зерен или процентное содержание компонента, измеренное для двух горизонтов. Простые примеры представлены на рис. 5.98, *а* и *б* картами уровня подземных вод в штате Небраска. Карта *а* изображает уровень подземных вод по данным, собранным в 1950г. Карта *б* построена по данным, собранным через 10 лет после возведения большой дамбы недалеко от середины площади. Изменение уровня подземных вод значительно и представлено картой разностей (рис. 5.98, *в*) Так же, как и исходные карты, построенные с помощью автоматических программ построения карт в изолиниях, карта разности была построена путем вычитания двух матриц значений в узлах сети друг из друга и последующего проведения изолиний по данным полученной матрицы. Конечно, если для обеих карт были использованы одни и те же наблюдения в скважинах, то разности между двумя множествами исходных данных легко вычислить и затем провести изолинии. Однако, как вы, наверно, заметили, не все точки являются общими для двух карт.

В этом примере обе карты построены в одних п тех же единицах и сравнение их проводится с помощью простого вычитания значений, соответствующих разным картам. Однако задача сравнения значительно усложняется, если картированы две разные переменные. Обратимся теперь к рассмотрению случая, когда требуется сравнить две карты, выраженные в разных единицах и построенные по различным контрольным точкам.

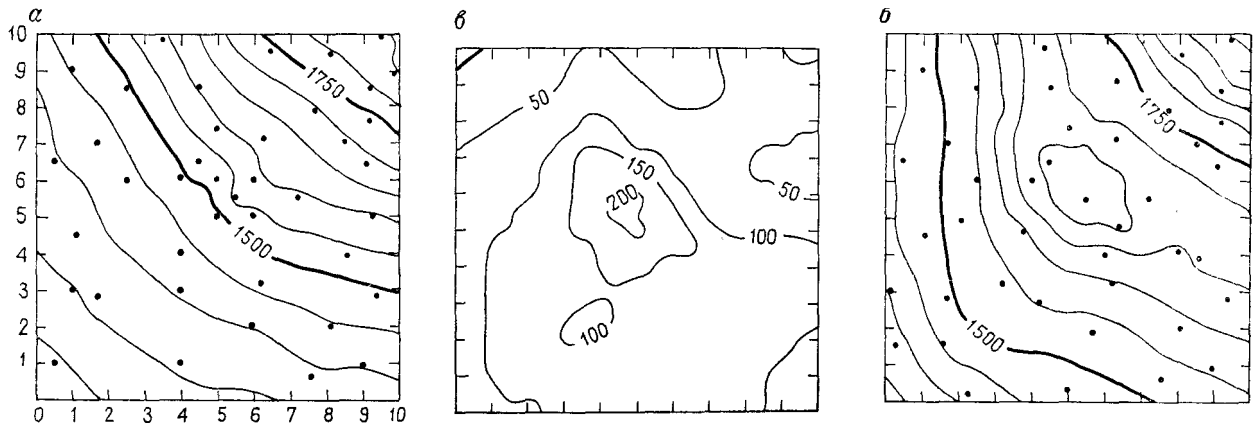


Рис. 5.98. Отметки уровня воды в области Небраска (по наблюдениям в скважинах): *а* – уровень воды в 1950 г.; *б* – уровень воды в 1960 г. после сооружения водохранилища; *в* – изменения уровня с 1950 по 1960г., полученные по разностям отметок карт *а* и *б*. Места наблюдения отмечены точками. Отметки указаны в футах над уровнем моря

Наиболее перспективными методами предсказания геологических структур, которые могут содержать нефть, являются методы отраженных сейсмических волн. Время, требуемое сейсмическим волнам для того, чтобы пройти от поверхности до отражающего горизонта, можно точно измерить. По этим измерениям можно построить сейсмический разрез, который даст конфигурацию отражающих горизонтов вдоль линии геофонов. По ряду сейсмических профилей вдоль подозреваемой структуры можно построить сейсмическую карту в изолиниях, как это было сделано на рис. 5.99. Хотя в эти данные были внесены поправки на все геометрические факторы, которые влияют на время возврата волн, нет точного способа превратить эти измерения в оценки глубины. Это происходит по той причине, что скорость, с которой сейсмическая волна проходит через породы, изменяется в зависимости от их состава, глубины и многих других причин. Однако карта в изолиниях для времени возврата, аналогичная изображенной на рис. 5.99, будет соответствовать по форме структурной карте отражающего горизонта.

На рис 5.100 изображена структурная карта в изолиниях на той же площади, построенная по измерениям в скважина, пересекающих отражающий горизонт. Это дает несколько более детальную картину структурной конфигурации, чем сейсмическая карта, хотя обе очень сильно напоминают друг друга, нам нужно сравнить эти две карты, чтобы определить, где сейсмическая оценка структурных отклонений сильнее отличается от картины полученной при бурении. Однако сейсмические наблюдения проводились не в буровых скважинах, где измерялись отметки кровли отражающего горизонта, и обе карты выражены в разных единицах. Для того чтобы прямо сравнить их, мы должны одну из них выразить в единицах другой либо перевести обе в стандартизованную, безразмерную форму.

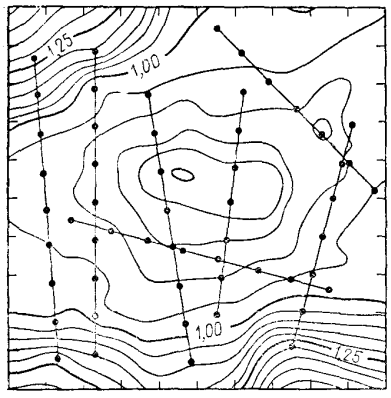


Рис. 5.99. Структурная конфигурация отражающей поверхности, определенная с помощью времени возвращения сейсмических сигналов. Прямыми линиями указаны сейсмические профили. Изолинии соответствуют исправленному времени возврата волн, выраженному в секундах

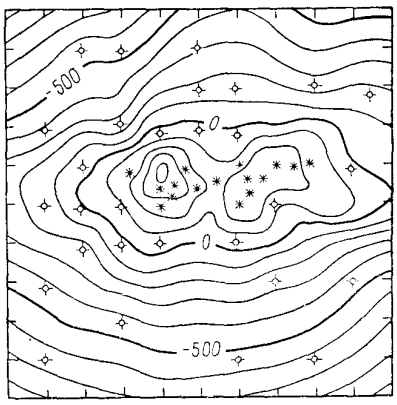


Рис. 5.100. Структурная карта кровли отражающего горизонта, построенная по данным бурения. Отметки указаны в футах над уровнем моря

Выражение одной переменной в терминах другой имеет определенное преимущество, так как сравнение будет проводиться в единицах одной из исходных карт, что позволяет нам указать площади, на которых картированная переменная «больше, чем она должна быть», или «меньше, чем значение, предсказанное на основании значений другой переменной». Оценка одной переменной через другую осуществляется с помощью методов наименьших квадратов, после чего на карту наносятся отклонения предсказанных значений переменной от действительных значений. Обозначим одну картируемую оцениваемую переменную через Y, а другую через X. Если наблюдения обеих переменных X и Y осуществлялись в одних и тех же точках, то по этим наблюдениям можно найти регрессию Y на X. Используя уравнение регрессии, можно предсказать значение переменной $\overline{Y}$ в каждой точке, т. е. фактически составить карту X в терминах Y. Иными словами, мы вычислим

$$\overline{Y} = \beta_0 + \beta 4_1 X_1 \qquad (5.102)$$

для всех точек карты, где $\overline{Y}$ является линейным преобразованием X в единицы Y, полученным по методу наименьших квадратов. Хотя обычно для нахождения значений используется линейная регрессия Y на X, для этой цели можно использовать также полиномиальную регрессию низших порядков. Процедура ее получения в точности такая же, как в системе уравнений (4.13)–(4.16), с тем лишь изменением, что аргумент – это одна из двух картируемых переменных, а не пространственная координата. Нахождение коэффициентов уравнения (5.102) осуществляется с помощью решения нормальных уравнений:

$$\sum_{i=1}^{n} Y_i = b_0 + b_1 \sum_{i=1}^{n} X_i ; \qquad \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (5.103)$$

которые в матричной форме можно записать так:

$$\begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} \qquad (5.104)$$

Попутно можно получить лучшую оценку $\overline{Y}$ с помощью полиномиальной модели; соответствующие нормальные уравнения даны в (4.32).

После того как найдено уравнение, определяющее одну из переменных через другую, можно вычертить карту предсказанных значений $\overline{Y}$. Так как значения $\overline{Y}$ получаются только на основании значений второй переменной X и соотношения между Y и X, которое мы определили, карту $\overline{Y}$ можно считать картой X, выраженной в единицах Y. Карту разностей $Y - \overline{Y}$ можно считать картой разностей между X и Y, выраженных в единицах Y.

В этой частной задаче мы интересуемся тем, насколько хорошо сейсмические данные предсказывают структуру. Следовательно, мы можем определить Y как глубину структуры, а X – как время возвращения сейсмической волны. Таким образом, с помощью уравнения регрессии

$$\text{глубина структуры} = b_0 + b_1 \text{ (время возвращения сейсмической волны).}$$

можно получить оценку структуры по сейсмическим данным. Так как две переменные были измерены не в одинаковых точках, мы должны строить наше уравнение на основании оценок X и Y, полученных в промежутках между контрольными точками по программе построения изолиний. По этой причине невозможны никакие оценки качества регрессии Y на X.

Рис. 5.101, *а* представляет собой карту оценок структурных отметок, полученных из уравнений регрессии. Так как уравнение регрессии выражает прямую зависимость, форма оцениваемой структуры идентична форме карты времени возвращения сейсмических волн. Однако шкала времени была преобразована в шкалу абсолютных отметок. Уравнение, связывающее глубину залегания структуры и время возвращения сейсмических волн, можно рассматривать как способ оценки скорости распространения сейсмических сигналов. Таким образом,

$$\text{глубина структуры} = 1389{,}6 - 1692{,}5 \text{ (время возвращения сейсмической волны).}$$

На рис. 5.101, *б* изображена карта разности $Y - \overline{Y}$. Площади, на которых сейсмические данные оценивают меньшую глубину структурного горизонта, чем реально существующая, имеют вид положительных отклонений. Там, где сейсмические методы предсказывают большую глубину, отклонение будет отрицательным.

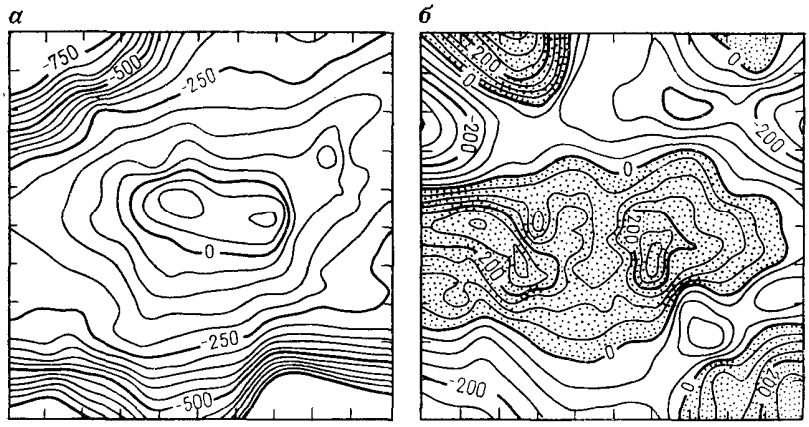


Рис. 5.101. Структурная карта, полученная на основании сейсмических данных (*а*) и карта разностей отметок изолиний карты *а* и структурной карты 5.100*(б)*. Заштрихованные площадки, на которых предсказанные значения положительных отклонений ниже истинных. Изолинии указаны в футах над уровнем моря

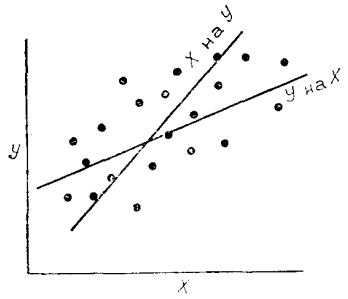


Рис 5.102. Две прямые регрессии *Y* на *X* и *X* на *Y*, построенные по набору выборочных точек

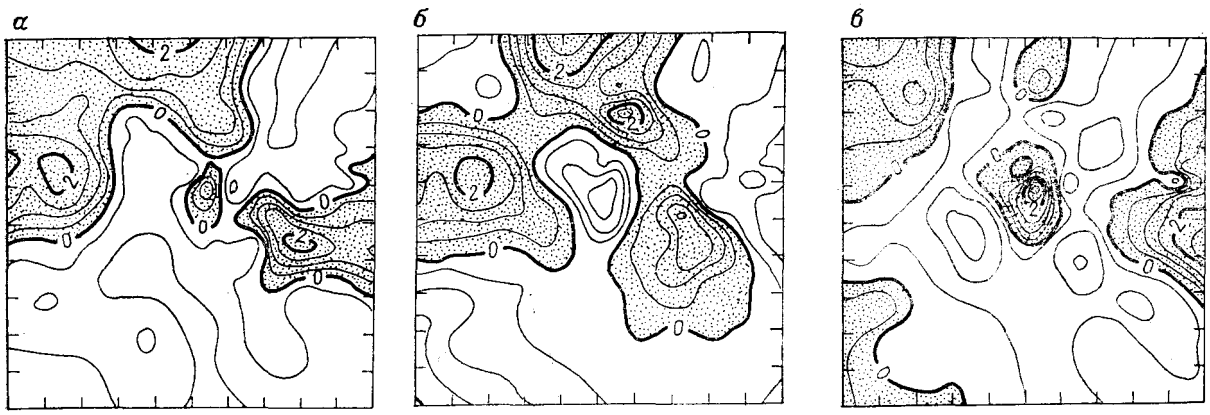


Рис. 5.103. Изменчивость стандартизованных параметров размера частиц донных осадков, собранных на побережье Северной Каролины (карты исходных параметров представлены на рис. 5.97): *а* – стандартизованное стандартное отклонение; *б* – стандартизованная асимметрия; *в* – карта разности отметок изолиний карт *а* и *б*. Единицей измерения карт является стандартное отклонение от среднего значения исходных параметров. Площади, соответствующие положительным отклонениям, заштрихованы

При использовании этого метода могут возникнуть некоторые затруднения. Во-первых, оценки $\overline{Y}$ можно получить лишь для части значений переменной *Y*. Если корреляция между *X* и *Y* невысокая, то при подстановке оценки $\overline{Y}$ могут возникнуть серьезные ошибки. В этом примере $R=0,87$ и поэтому оценки получаются хорошими.

Несмотря на то что выбор единиц измерения глубины для построения карты разностей кажется вполне оправданным, вопрос о том, какую переменную надо выбрать для получения оценок, решить удается не всегда. Как показано на рис. 5.102, в этом случае возможны две линии регрессии: *X* на *Y* и *Y* на *X*. Если корреляция между *X* и *Y* высокая, то две линии приблизительно совпадают, но если корреляция не ярко выражена, то две линии оценок могут привести к совершенно различным результатам [60]. Возможный способ преодоления этой трудности – использовать приведенную главную ось для выражения одной переменной через другую (см. гл. 4). Конечно, сомнительно, что при отсутствии высокой корреляции между двумя переменными имеет смысл пытаться сравнить их, используя прогнозные методы.

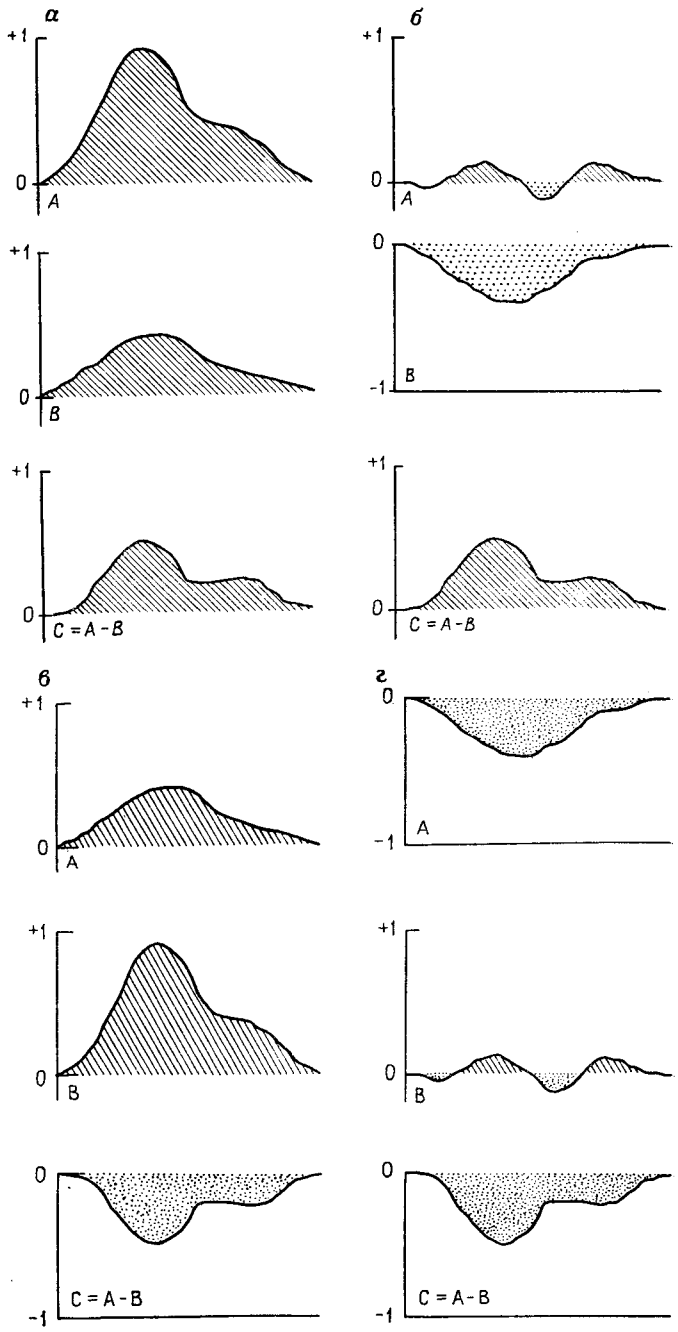


Рис. 5.104. Разрезы карт, показывающие случаи, возникающие при построении карты разностей методом вычитания значении одной карты из другой: *а* – вычитание значении малой положительной площади из значений большой положительной площади; *б* – вычитание значений малой отрицательной площади из значений малой положительной площади; результат такой же, как и в случае *а*; *в* – вычитание значений большой положительной площади из значений малой положительной площади; *г* – вычитание значений малой положительной площади из значений малой отрицательной площади; результат такой же, как и в случае *в*.

Трудностей, возникающих при изучении карт разностей, построенных на основании оценок или предсказанных значений переменных, удастся избежать, если две исходные карты преобразовать в стандартизированную форму. С этой целью из значения переменной в каждой контрольной точке вычитается среднее значение этой переменной и полученная разность делится на стандартное отклонение. Иными словами, выполняется то же преобразование, которое использовалось в гл. 2 для преобразования данных в стандартную нормальную форму

$$Z_i = \left(X_i - \overline{X}\right)/s \qquad (5.105)$$

После того как данные по каждой карте стандартизированы, по ним можно построить изолинии по обычной схеме. Однако значения изолиний выражены в единицах стандартного отклонения и принимают как положительные, так и отрицательные значения. Так как единицы измерения карты неизвестны, это может представить некоторые затруднения при интерпретации. Однако следует отметить, что тот, кто использует статистические процедуры, не будет иметь больших затруднений, связанных со стандартизированной шкалой измерений.

В качестве примера карты стандартных отклонений и асимметрии распределения осадков в лимане (см. рис. 5.97,*б* и *в*) приведем к стандартному виду (рис. 5.103,*а* и *б*).

Теперь обе карты вычерчены в одном и том же масштабе, и их можно сравнить между собой, если «вычесть» одну из другой. В результате получится карта разностей, аналогичная карте изопахит. Карта разности стандартного отклонения и асимметрии представлена на рис. 5.103,*в*. Однако, так как карта разностей может содержать области неопределенности, могут возникнуть различного рода вопросы. Например, рассмотрим ряды разрезов двух карт поверхностей, изображенных на рис. 5.104. В этих иллюстрациях разрез *B* вычитался из разреза *A*. На рис. 5.104,*а* ожидаемая положительная разность есть результат вычитания малой положительной площади из большой положительной площади. На рис. 5.104, *б*, однако, видно, что положительная разность также может быть результатом вычитания большой отрицательной площади из малой положительной площади (и даже

из нулевой площади или малой отрицательной площади). Аналогичные случаи, соответствующие отрицательным разностям, изображены на рис. 5.104,*б* и *г*. Хотя разности указаны верно, мы должны считать две поверхности, изображенные на рис. 5.104,*а* и *в*, более похожими, чем поверхности, изображенные на рис. 5.104,*б* и *г*. В частях а и в обе поверхности отклоняются от среднего значения в одну и ту же сторону и могут рассматриваться как близкие к параллельным. Корреляция между рядами точек, общих для обеих поверхностей, будет положительной. Если две поверхности близки по форме, то между ними имеется более высокая корреляция. Наоборот, пары поверхностей, изображенных на рис. 104,*б* и *г* будут отрицательно коррелированы, так как они имеют противоположный наклон.

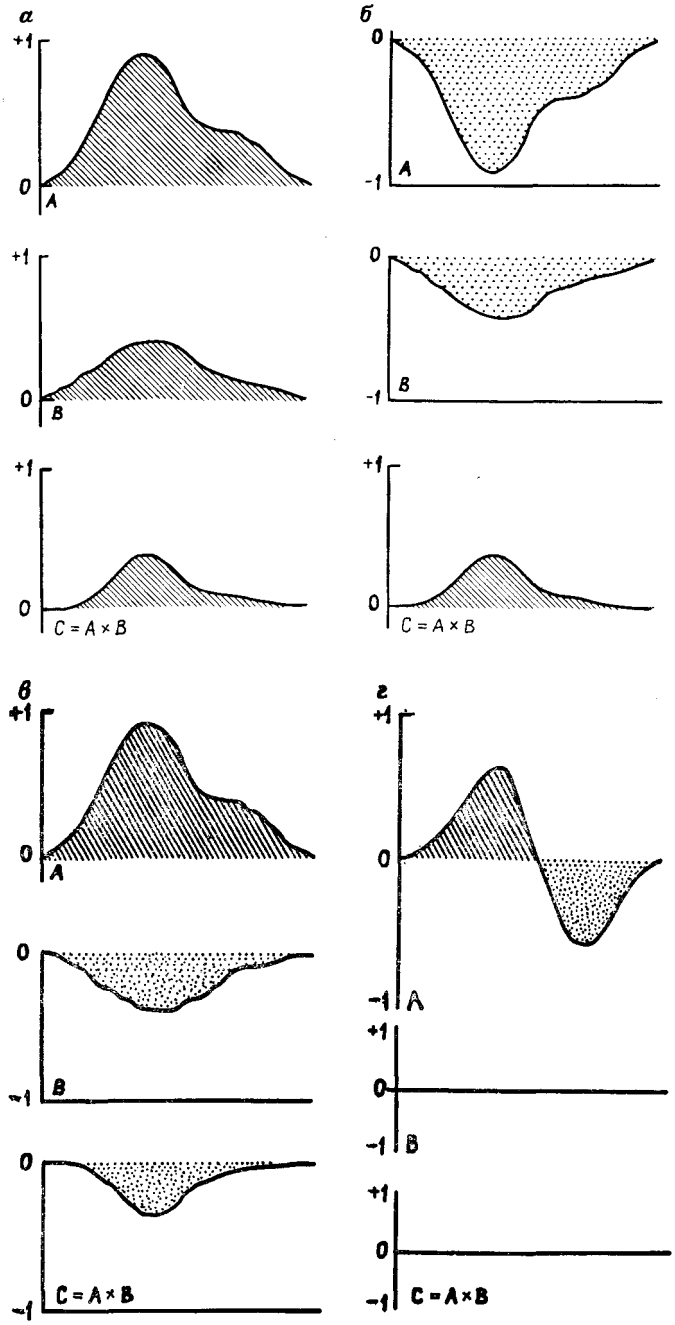


Рис. 5.105. Разрезы карты, показывающие результаты сравнения карт при их перемножении: *а* – умножение значений двух положительных площадей; *б* – умножение значений двух отрицательных площадей; *в* – умножение значений положительной площади на значения отрицательной; *г* – умножение значений либо положительной, либо отрицательной площади на значения нулевой площади

Поскольку поверхности стандартизированы, в действительности для получения меры сходства между двумя поверхностями нет необходимости вычислять коэффициент корреляции. Вместо этого поверхности можно перемножить. Если обе поверхности отклоняются от среднего значения в одном и том же направлении, их произведение будет положительным. На рис. 5.105 приведены примеры, представляющие обе указанные возможности. Единственный неясный случай возникает тогда, когда одна из поверхностей является плоскостью, характеризующей среднее значение. После стандартизации все точки такой поверхности будут нулевыми, поэтому карта произведения будет также нулевой независимо от формы второй карты. Этот случай изображен на рис. 5.105, *г*.

Карта произведения двух стандартизированных статистик размера зерен изображена на рис. 5.106. Площади высокой степени близости двух поверхностей ясно видны, так же как и области, где они заметно отличаются друг от друга.

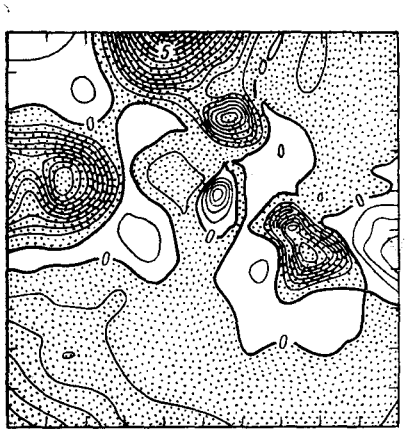


Рис. 5.106. Карта произведении значений карты стандартного отклонения размеров зерен и карты асимметрии размеров зерен в осадках, собранных на побережье Северной Каролины. Первоначальные параметры размеров стандартизованы так, как это показано на рис. 5.103. Положительные площади (заштрихованы) указывают на высокую корреляцию между картами поверхностей; области отрицательных значении указывают на обратную корреляцию между картами.

Коэффициенты сравнения карт

Выше для сравнения поверхностей тренда различных стратиграфических горизонтов была использована очень простая процедура. Так как предсказание значений $\overline{Y}_i$, осуществляется исключительно на основании полиномиального уравнения, степень сходства между поверхностями можно установить с помощью сравнения одних только полиномиальных коэффициентов β_i. Конечно, поверхности должны иметь одну и ту же степень, так как сравниваемые уравнения должны содержать одинаковое число членов. Кроме того, для получения уравнений сравниваемых поверхностей нужно пользоваться пробами, имеющими то же местоположение. Эти ограничения уменьшают область применимости данного метода к таким задачам, как изучение структурных изменений в регионе, которые отражаются в сходстве поверхностей тренда на последовательных стратиграфических горизонтах. Другая возможная область применения – сравнение поверхностей тренда различных минералогических и геохимических составляющих при условии, что все переменные получены на основании одного и того же множества проб и все поверхности имеют одну и ту же степень.

Коэффициенты можно сравнить с помощью одной из нескольких мер сходства. Для этой цели подходит коэффициент корреляции, определяемый по формуле

$$r = \frac{\operatorname{cov} \beta_{12}}{\sqrt{\operatorname{var} \beta_1 \cdot \operatorname{var} \beta_2}} \quad (5.106)$$

которая получена прямо из соотношения (2.24). Однако вместо сравнения наблюдений X_1 и X_2 мы сравниваем коэффициенты β_i уравнений поверхностей тренда 1 и 2.

Другая мера сходства, используемая для этих целей, – таксономическое расстояние d, определяемое по формуле

$$d = \sqrt{\sum_{i=1}^{n} (\beta_{1i} - \beta_{2i})^2} \Big/ n \qquad (5.107)$$

т.е. таксономическое расстояние равно квадратному корню из среднего арифметического квадратов расстояний между эквивалентными коэффициентами двух сравниваемых поверхностей. Мы рассмотрим эту меру в гл. 6 при изучении методов классификации. Такая мера не более эффективна, чем коэффициент корреляции, но ее иногда легче интерпретировать, так как она всегда положительна и не может, как коэффициент корреляции, принимать значения, меньшие 1,00. Аналогичные множества данных дают таксономические расстояния от нуля; множества данных с увеличивающимся разбросом имеют возрастающие таксономические расстояния между точками.

На рис. 5.107 изображен ряд поверхностей тренда третьей степени, подобранный для последовательности формаций на одной из территорий в Северо-Западном Канзасе. Исследование этих двух матриц, представленных в табл. 5.24 и 5.25, показывает, что относительное сходство, выявляемое этими двумя мерами, в обоих случаях одинаковое.

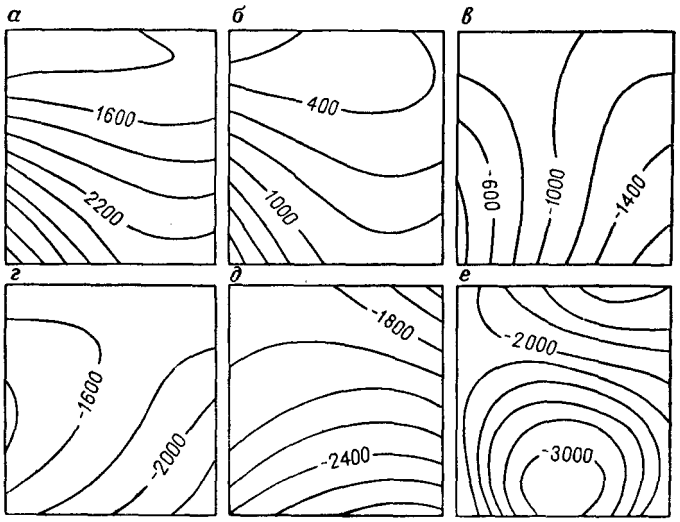


Рис. 5.107. Поверхности тренда третьей степени для стратиграфических формаций в Северо-Западном Канзасе [57]. Возраст: *а* – меловой, *б* – пермский, *в* – пенсильванский, *г* – миссисипский, *д* – ордовикский, *е* – докембрийский

Уравнение поверхности тренда, полученное на основании одного множества контрольных точек в регионе, может сильно отличаться от уравнения той же степени, полученного для других точек в том же регионе. Однако форма двух поверхностей тренда может быть одной и той же, и обе могут давать одинаково хорошую аппроксимацию. Это вытекает из того факта, что уравнение тренда – это уравнение регрессии на географических координатах, построенное по некоторому множеству данных с указанным местоположением, поэтому любое изменение последних отражается в коэффициентах. По этой причине сравнения между коэффициентами тренда обоснованны, если для построения сравниваемых поверхностей тренда использовались одни и те же выборочные данные. Далее, абсолютные значения коэффициентов регрессии иногда подвержены сильным флуктуациям, в особенности для высоких степеней, что является результатом машинного округления и отбрасывания избыточных разрядов. Эти факторы ограничивают применимость данного метода сравнения карт поверхностей.

Таблица 5.24. Матрица коэффициентов корреляции между коэффициентами поверхностей тренда третьей степени (коэффициент β_0 исключен из рассмотрения) [58]

Возраст	*а*	*б*	*в*	*г*	*д*	*е*
а. Меловой	1,000					
б. Пермский	0,684	1,000				
в. Пенсильванский	0,672	0,250	1,000			
г. Миссисипский	–0,547	–0,032	–0,036	1,000		
д. Ордовикский	–0,002	–0,010	0,521	0,256	1,000	
е. Верхний докембрий	–0,470	–0,486	–0,838	–0,248	–0,693	1,000

Таблица 5.25. Матрица таксономических расстояний между коэффициентами поверхностей тренда третьей степени (коэффициент β_0 исключен из рассмотрения) [58]

Возраст	*а*	*б*	*в*	*г*	*д*	*е*
а. Меловой	0,000					
б. Пермский	0,510	0,000				
в. Пенсильванский	0,565	0,691	0,000			
г. Миссисипский	1,316	0,975	1,108	0,000		
д. Ордовикский	0,943	0,789	0,629	0,936	0,000	
е. Верхний докембрий	2,096	1,839	2,796	2,057	2,205	0,000

СПИСОК ЛИТЕРАТУРЫ К ГЛАВЕ 5 «АНАЛИЗ КАРТ»

1. *Agterberg F. P.* Computer techniques in geology. Earth Science Reviews, 3. 1967, p. 47–77.
   Общий обзор, включающий несколько примеров двумерного анализа.
2. *Aitchinson J., Brown I. A. C*. The lognormal distribution, with special reference to its uses in economics. Cambridge Univ. Press Cambridge, 1969, 176 p.
3. *Armstrong M., Jabin R*. Variogram models must be positive-definite, Jour. Int'l Assoc. Mathematical Geology, 13, № 5, 1981, p. 455–459.
4. *Barfels C. P. A., Keteliapper R. H.* (eds) Exploratory and explanatory statistical analysis of spatial data. Martinus Nijhoff Publishing Boston, 1979, 268 p.
   Критерии, относящиеся к точечным распределениям и пространственной зависимости, обсуждаются в этом сборнике статей, посвященных «региональному анализу».
5. *Batschelet E*. Statistical methods for the analysis of problems in anumal orientation and certain biological rhythms. American Institute of Biological Sciences Monograph. Washington, D. C., 1965, 57 p.
6. *Bennett R. J*. Spatial time series, analysis–forecasting–control. Pion Ltd., London, 1979, 674 p.
   Новейшая трактовка временной и пространственной изменчивости, рассматриваемая с точки зрения классического анализа временных рядов.
7. *Berry B. J. L., Marble D. F*. (eds.). Spatial analysis, a reader in statistical geography. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1968, 512 p.
   Сборник основополагающих статей, представляющий интерес как для геологов, так и для географов. Многие из них извлечены из недоступных источников.
8. *Birkhoff G., Mansfield L*. Compatible trianguiar finite elements. Jour. Mathematical Analysis and Applications, 47, № 3, 1974, p. 531–553.
   Подробное обсуждение математических принципов, лежащих в основе алгоритма оконтуривания методом триангуляции.
9. *Boon J. D., Evans D. A., Henningar H. F*. Spectral information from Fourier analysis of digitized quartz grain profiles Jour Int'l Assoc Mathematical Geology, 14, № 6, 1982, p. 589–605.
   В статье указывается на важность определения истинного центроида в полярном анализе Фурье, обращается особое внимание на его применение к изучению формы зерен.
10. *Chayes F*. On deciding whether trend surfaces of progressively higher order are meaningful. Geol. Soc. Amer. Bull., 81, 1970, p. 1273–1278.
11. *Cheeney R. F*. Statistical methods in geology, George Alien and Unvin Ltd., London, 1983, 169 p.
    Имеется русский перевод. Чини Р. Ф. Статистические методы в геологии. М., Мир, 1986. В главах 8 и 9 этой книги содержится исключительно хорошо изложенный анализ двумерных и трехмерных ориентированных геологических данных.
12. *Charley R. J., Haggett P*. Trend-surface mapping in geographical research. Trans. Inst, British Geographers. Publ. № 37, 1965, p. 47–67.
    Ясное изложение тренд-анализа поверхностей с нестатистических позиций. Большинство рассматриваемых примеров геологические.
13. *Clark I*. Practical geostatistics. Applied Science Publishers. London, 1979, 129 p.
    В этой компактной книжке наиболее ясно изложены теория регионализованных переменных и практическая теория крайгинга.
14. *Clark M. W*. Quantitative shape analysis. A review. Int'l Assoc. Mathematical geology, 13, № 4, 1981, p. 303–320.
    Подробный обзор анализа частных форм, основанного на методе разложения в ряды Фурье.
15. *Cliff A. D., Ord I. K*. Spatial processes, models and applications Pion Ltd, London, 1981, 266 p.
    Описаны новейшие методы пространственной автокорреляции и анализа образов из точек. Большинство приложений касаются дискретных географических переменных.
16. *Cochran W. T. and others*. What is the Fast Fourier Transfom Proc. Inst. Electrical and Electronics Engineers, 55, 1967, p. 1664–1674.
17. *Cole I. P, King C. A. M*. Quantitative geography. John Wiley and Sons. Inc., New York, 1962. 692 p.
    Ценная точка зрения географов на пространственный анализ. Часть 2, озаглавленная «Пространственные распределения и соотношения», имеет наибольшее отношение к материалу этой главы.
18. *Cote L. J., Davis J. O., Marks W., McGough R. J., Mehr E., Pierson W. J., Roper J. F., Stephenson G., Vetter R. C*. The directional spectrum of wind generated sea as determined from data obtained by the

stereowave observation project. New York Univ. Meteorological Papers, 2, № 6, 1960, 88 p.
19. *Dacey M F.* Description of line patterns. Northwestern Studies in Geography, 13, 1967, p. 277–287.
Одно из немногих изданий, в которых рассмотрены схемы линии на плоскости.
20. *Dalberg E. C.* Relative effectiveness of geologists and computers in mapping potential hydrocarbon exploration targets. Jour. Int'l Assoc. Mathematical Geology, 7, № 5/6, 1975, p. 373–394.
Детальное сравнение контурных карт, вычерченных геологами и компьютерами.
21. *David M.* Geostatistical ore reserve estimation. Elsevier Publ. Co., Amsterdam. 1977, 364 p.
Имеется русский перевод: Давид М. Геостатистические методы при оценке запасов руд. М., Недра, 1980. Подробное изложение теории регионализованных переменных и использования крайгинга для подсчета запасов. Несколько программ на ФОРТРАНе включены в текст.
22. *Davis J. C.* Conturing algorithm, in AUTOCARTO II – Proceedings of the International Symposium on Computer-assisted Cartography, Sept. 21–25, 1976. U. S Bureau of the Census/Amer. Congress on Survey and Mapping, p. 352–359.
Поясняются выбор алгоритма опробования и его преобразование.
23. *Davis I. C., Preston F. W.* Optical processing – An alternative to digital computing, in Fenner P. (ed.). Quantitative Geology. Geol. Soc. America, Spec. Paper, 146, 1972, p. 49–68.
24. *Donnelly K. P.* Simulations to determine the variance and edge effect of total nearest neighbor distance, in Hodder, I. Simulation studies in archaeology: Cambridge Univ. Press, Cambridge, 1978, p. 91–95.
Важная статья, описывающая изменения, необходимые для компенсации краевого эффекта в методе ближайшего соседа.
25. *Doveton I. H., Parsley A. I.* Experimental evaluation of trend surface distortions induced by inadequate data-point distributions: Inst. Min. and Met., Trans. Sec. B. 1970, p. B197–B208.
Раздел, относящийся к влиянию заданного распределения точек на поверхности тренда, в этой главе целиком основан на большой части этих экспериментов. Рис. 5.86 и 5.87 адаптированы из рисунков этой работы.
26. *Ehrllch R., Brown I. P., Yarus J. M.* The origin of shape frequency distributions and the relationship between size and shape. J. Sed. Petrol., 50, № 2, 1980, p. 475–484.
Одна из серии статей этих авторов по использованию циклических рядов Фурье в изучении формы зерен.
27. *Folk R. L.* Petrology of sedimentary rocks. Hemphill's, Austin, Texas, 1968, 184 p.
Обсуждаются традиционные меры формы зерен.
28. *Gaile G. L., Burt J. E.* Directional statistics. Concept and Techniques in Modern Geography, № 25, Geo Abstracts, Univ. East Anglia, Norwich, 1980, 39 p.
Компактная монография, в которой суммированы многие процедуры проверки гипотез о направленных данных.
29. *Getis A., Boots B.* Models of spatial processes, an approach to the study of point, line and area patterns. Cambridge Univ. Press, Cambridge, 1978, 198 p.
30. *Gold C. M., Charters T. D., Ramsden J.* Automated contour mapping triangular element data structures and an interpolant over each irregular triangular domain. Computer Graphics, 11, № 2, 1977, p. 170–175.
31. *Griffiths J. C.* Frequently distributions of some natural recources materials. Pennsylvania State Univ., Mineral Industries Experiment Station Circular, 63, 1962, p. 174–198.
В этой и следующей ссылке рассматривается отрицательное биноминальное распределение как модель встречаемости минеральных залежей и нефтяных полей.
32. *Griffiths J. C.* Exploration for natural resources. J. Oper. Res. Soc. Amer., 14, № 2, 1966, p. 189–209.
33. *Cumber E. J., Greenwood J. A., Durand D.* The circular normal distribution. Tables and theory. J. Amer. Stat. Soc., 48, 1953, p. 131–152.
34. *Haggvl T., Cliff A. D., Frey A.* Locational analysis in human geography, 2nd ed. John Wiley and Sons, Inc., New York, 1977, 605 p.
Современная книга, содержащая много методов, которые могут оказаться новыми для геологов. Часть 2 «Методы пространственного анализа» посвящена выборочным схемам классификации регионов и проверке статистических гипотез о пространственных соотношениях.
35. *Harbaugh J. W.* A computer method for four-variable trend analysis illustrated by a study of oil-gravity variations in southeastern Kansas. Kansas Geological Survey Bull., 171, 1964, 58 p.
36. *Harbaugh S. W., Preston F. W.* Fourier series analysis in geology. Computers and Computer Applications in Mining and Exploration. School of Mines, Univ. Arizona, Tucson, 1966, p. R1–R46.

Рассмотрены вопросы, связанные с построением поверхностей тренда по геологическим данным на основе простых и двойных рядов Фурье. Эта статья входит в сборник [7] под номером IV–8.

37. *Harbaugh S. W., Meriam D. F.* Computer applications in stratigraphic analysis. John Wiley and Sons, Inc., New York, 1968, 282 p.
38. *Hodder Ian, Orion C.* Spatial analysis in archaeology. Cambridge Univ, Press, Cambridge, 1979, 270 p.
Непритязательное переиздание книги 1976 г. в бумажном переплете содержит теорию точечных распределений, тренд-анализа поверхностей и методы сравнения карт, используемых в археологии.
39. *IBM.* Numerical surface techniques and contour map plotting. International Business Machines, Data Processing Applications, White Plains, № Y, 1965, 35 p.
Введение в методы построения карт.
40. *James W. R.* FORTRAN IV program using double Fourier series for surface fitting of irregularity spaced data Kansas Geological Survey Computer Contribution, 1, 1966, 19 p.
Краткое изложение метода наименьших квадратов при аппроксимации геологических данных отрезком двойного ряда Фурье. Рис. 5.92 взят из этой статьи.
41. *Journel A. G., Huijbregts Ch. J.* Mining geostatistics. Academic Press, London, 1978, 600p.
(Перепечатана с исправлениями в 1981 г.) Наиболее полное изложение теории регионализованных переменных, причем затронуты вопросы оценки месторождений. Характерен высокий научный уровень.
42. *Kaesler R. L., Waters J. A.* Fourier analysis of the ostracode margin Geol. Soc. America Bulletin, 83, 1972, p. 1169–1178.
43. *Kendall M. G., Moran P. A. P.* Geometric probability. Hafner Publ Co, New York, 1963, 125 p.
Глава 2 касается распределения точек на плоскости.
44. *King L. J.* Statistical analysis in geography. Prentice-Hall, Inc., Engle-wood Cliffs, N. J., 1969, 288 p.
Курс статистики повышенной трудности во многом соответствует изложенному в этой книге. Проверка гипотез о распределении рассматривается в гл. 5, а тренд-анализ поверхностей – в гл. 6.
45. *Koch G. S. Jr., Link R. F.* Statistical analysis of geological data Dover Publications Inc., New York, 1980, 850 p.
Тренд-анализ поверхностей рассматривается в гл. 9, а систематическое изложение теории и применения поверхностей отклика – в гл. 12.
46. *Krige D. G.* Two-dimentional weighted moving average trend surfaces for ore valuation., in Proc. Symposium on Mathematical Statistics and computer Applications in Ore Valuation. Mar. 7–8. Jour. South African Inst., Mining and Metalurgy, Jochannesbourg, 1966, p. 13–38.
Критика тренд-анализа поверхностей и обсуждение схем скользящего Среднего для предсказания рудных содержаний. В статье, следующей за статьей Криге, указываются пункты, по которым имеются разногласия между двумя лагерями, стоящими на противоположных точках зрения.
47. *Krumbein W. C.* Preffered orientation of pebbles in sidementary deposits. Jour. Geology, 47, 1939, p. 673–706.
48. *Lambie F.* An analysis of the probability of hitting an arbitrary elliptical target with sets of parallel search fines. Terrasciences Inc., Unpub. Report, San Ramon, Calif., 1981, 17 p.
49. *Lee D. R., Sallee G. T.* A method of measuring shape Geographical Review, 60, № 4, 1970, p. 555–563.
50. *Li J. C. R.* Statistical inference, v. 2. Edwards Bros, Inc., Ann Arbor, Mich., 1964, 575 p.
В гл. 30 излагаются критерии криволинейной регрессии, которые допускают непосредственное обобщение на поверхности тренда.
51. *Mardia K. V.* Statistics of directional data. Academic Press Ltd, London, 1972, 357 p.
Полный обзор статистических методов обработки дву- и трехмерных ориентированных данных. Многие примеры взяты из геологии.
52. *Mafheron G.* Principles of geostatistics. Economic Geology, 58, 1963, p. 1246–1266.
Этот обзор теории регионализованных переменных и применения крайгинга написан человеком, внесшим наибольший вклад в эту область прикладной статистики.
53. *McCammon R. B.* Target intersection probalities for parallel-line and continuous-grid types of search. Jour. Int'l. Assoc. Mathematical Geology, 9, № 4, 1977, p. 369–383.
Содержит уравнения и графы для определения вероятностей подгонки эллиптической схемы к исследуемой схеме, состоящей из регулярно расположенных в пространстве линий.
54. *McCullagh M. J.* Creation of smooth contours over imagularly distributed data using local surface patches. Geographical Analysis, 13, № 1, 1981, p. 51–63.

В этой и следующих статьях рассмотрены вопросы вычисления треугольном сети, связывающей нерегулярно расположенные в пространстве точки, а также вопросы оконтуривания поверхности на основе этой сети.

55. *McCullagh M. J., Ross C. G*. Delaney triangulation of a random data set for isarithmic mapping. Cartographical Journal, 17, № 2 1980 p. 93-99.
56. *Mendenhall W*. Introduction to linear models and the design and analysis of experiments Wadsworth Publ. Co., Inc., Belmot, Calif., 1968, 465 p.
    В главе 10 обсуждается подгонка поверхности отклика к экспериментальным схемам.
57. *Merrian D. F., Harbaugh J. W*. Trend-surface analysis of regional and residual components of geologic structure in Kansas. Kansas Geological Survey Spec. Distribution Publ., 11, 1964, 27p.
58. *Merrian D. F., Sneath P. H. A*. Quantitative comparison of contour maps Journal Geophysical Research, 71, 1966, p. 1105–1115.
    В этой статье описывается метод сравнения карт с помощью коэффициентов корреляции поверхностей тренда. Использованы данные из [57]. Табл. 5.24 и 5.25 взяты из этой статьи.
59. *Miles R. E*. Random polygons determined by lines in a plane, I and II. Proceedings of the National Academy of Sciences, 52, 1964, p. 901-907, 1157-1160.
60. *Mills F. C*. Statistical methods, 3rd ed. Holt, Rinehart and Winston, New York, 1955, p. 842.
61. *Moellering H., Rayner J. N*. Measurement of shape in geography and cartography. Oslo State Univ., Report of the Numerical Cartography Laboratory, NSF Grant No. SOC77–11318. 1979. 109 D.
    Подробный обзор анализа форм, включая полярный метод Фурье.
62. *Olea R. A*. Optimum mapping techniques using redionalized variable theory. Kansas Geological Survey Series on Spatial Analysis, № 2, 1975, 137 p.
    В этой монографии содержится полный вывод уравнений, используемых для осуществления точечного и универсального крайгинга как методов построения контурных карт.
63. *Olea R. A*. Optimization of the High Plains aquifer observation network. Kansas. Kansas Geological Survey Ground Water Series, № 7, 1982, 73 p.
64. *Rauner J. N*. An introduction to spectral analysis. Pion Ltd., London. 1971. 174 p.
    Одно- и двумерный анализ Фурье, изложенный с точки зрения географа. Изложение быстрого преобразования Фурье особенно хорошее.
65. *Ripley B. D*. Spatial statistics. John Wiley and Sons, New York, 1981. 252 D.
    Полный обзор пространственного анализа. В различных местах обсуждаются вопросы тренд-анализа поверхностей и квадрантного анализа. Риплай детально осуществил повторный анализ нескольких множеств данных из первого издания этого текста, включая данные по магнетитовым кристаллам, приведенным в табл. 5.4.
66. *Robinson J. E*. Computer applications in petroleum geology Hutchinson Ross Publ. Co., New York, 1982, 164 p.
    Элементарная книга, посвященная обработке нефтяных данных и методам картирования. Главы 7 и 8 посвящены контурным картам, поверхностям тренда и пространственной фильтрации.
67. *Robinson J. E., Charlesworth H. A. K; Ellls M. J*. Structural analysis using spatial filtering in interior plains of south-central Algebra. Bull. American Assoc. Petroleum Geologists, 53, 1969, p. 2341–2367.
    Описание двумерного анализа Фурье и фильтрации структурных данных.
68. *Rogers A*. Statistical analysis of spatial dispersion, the auadrat method, Pion Ltd., London, 1974, 164 p.
    Современное изложение квадратных методов анализа точечных схем. Примеры взяты из географии.
69. *Sampson R. I*. The SURFACE II graphics system, in *Davis J. C. and McCullagh M. J*. (eds.). Display and analysis of spatial data. Wiley Inter-scnence. London. 1975. p. 244–266.
70. *Silk J*. Statistical Concepts in Geography. George Alien and Unwin Ltd London, 1979, 276 p.
    Очень легко читаемый вводный текст, предназначенный для географов. В книге содержатся главы, посвященные точечным и площадным схемам, а также опробованиям на плоскости. Продается в бумажном переплете.
71. *Singer D. A., Wickman F. E*. Probability tables for locating elliptical targets with square, rectangular, and hexagonal point-nets. Pensylvania State Univ. Mineral Sciences Experiment Station Spec. Publ., 1969, 1–69, 100 p.
72. *Stephens M. A*. Tests for randomness of directions against two circular alternatives. Jour. American Statistical Association, 64, № 325, 1969, p. 280– 289.
73. *Stephens M. A*. Tests for the dispersion and for the nodal vector of a distribution on a sphere. Bio-

metrika, 54, 1967, p. 211–223.

74. *Tipper J. C.* Surface modelling techniques. Kansas Geological Survey on Spatial Analysis, № 4, 1979, 108 p.
Обзор методов исследования точечных схем, таких, как подгонка сплайнов, потенциально применимых в науках о Земле. Примеры взяты из палеонтологии и моделирования нефтяных резервуаров.

75. *Unwin D.* Introductory spatial analysis. Methuen and Co., Ltd, London, 1981,212 p.
Компактное изложение распределения точек, линий и площадей на картах. Включены также главы, посвященные оконтуриванию и сравнению карт.

76. *Uspensky J. V.* Introduction to mathematical probability. McGraw-Hill. Inc., New York, 1937, 411 p.

77. *Walters R F.* Contouring by machine. A user's guide. Bull. American Assoc. Petroleum Geologists, 53, 1969, p. 2324–2340.
Устаревшее, но полезное обсуждение контурного картирования.

78. *Walson G. S.* Orientation statistics in the Earth sciences. Bull. Geological Institute of Uppsala, 2, № 9, 1970, p. 73–89.
Эта статья посвящена компактному изложению применений направленных статистик к исследованию данных из геологии и географии. Многие из примеров вошли в [80].

79. *Watson G. S.* Trend-surface analysis. Jour. Int'l Assoc. Mathematical, Geology, 3, № 3, 1971, p. 215–226.

80. *Watson G. S.* Statistics on spheres. John Wiley and Sons, Inc., New York, 1983, 238 p.
Это – современная трактовка статистики точечных распределений на сферах. Рассмотрены также направленные данные, так как точки могут представлять концы векторов на единичной сфере. Эта книга, к сожалению, содержит многочисленные типографские ошибки.

81. *Woodcock N. H.* Specification of fabric shapes using an eigenvalue method. Geol. Soc. America Bulletin, 88, 1977, p. 1231–1236.
Классификация петротектонических диаграмм, основанная на отношении их собственных значений.

82. *Wrigley N.* (ed.). Statistical applications in the spatial sciences Pion: Ltd., London, 1979, 310 p.
Собрание статей, в которых рассмотрено множество географических тем. Глава 2, посвященная применениям науки об окружающей среде, представляет особый интерес.

Глава 6. АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ

В предыдущих главах мы рассмотрели анализ данных, представляющих измерения одной переменной в каждом образце или наблюдаемом объекте. В гл. 4 и 5 мы изучили влияние географических или временных координат на выборочные характеристики. Теперь остановимся на методах анализа многомерных данных, в которых каждый наблюдаемый объект характеризуется множеством переменных. Многомерные методы позволяют одновременно изучать изменение набора характеристик. Существует ряд примеров геологических данных, к которым применимы методы многомерного анализа. Среди них можно назвать химические анализы, в которых переменные представляют собой содержания редких элементов, выраженные в процентах или частях на миллион; такие измерения в руслах рек, как сток, число взвешенных частиц, глубина, содержание диссоциированных твердых веществ, содержание кислорода. Примером могут служить также палеонтологические характеристики, такие, как набор измерений, сделанных на особях некоторого ископаемого вида. Можно привести и ряд других примеров. Одни из них – это простые обобщения рассмотренных ранее задач, другие принадлежат к совершенно новому классу проблем.

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 геологических исследованиях. В большинстве геологических 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 всестороннем исследовании, которое позволяет выделить наиболее важные факторы. Методы, изложенные в данной главе, могут оказать в этом существенную помощь.

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Мы начнем эту главу с изложения известных вещей, которые, однако, будут представлены в несколько нетрадиционном и более общем виде. Это вопросы регрессии, которые включают теорию аппроксимации полиномиальными кривыми (см. гл. 4) и анализ поверхностей тренда (см. гл. 5). Теперь, однако, мы не будем ограничиваться рассмотрением функций только расстояния или пространственных координат. Любую наблюдаемую переменную можно рассматривать как функцию любой другой переменной измеренной на тех же образцах. В гл. 4 мы изучали зависимость влажности осадка от глубины его залегания. Таким образом можно определить процентное содержание монтмориллонита в осадке и содержание в нем воды. Мы могли бы измерить еще несколько переменных, таких, как содержание органических веществ, средний размер зерен и общую плотность, а также изучать зависимость содержания воды от изменения каждой или всех вместе взятых переменных. В некотором смысле переменные можно считать пространственными и изучать изменения, происходящие в направлении, определенном переменной. Это обычный прием; мы пользуемся им всякий раз, когда изображаем графически зависимость одной переменной. от другой, используя при этом пространственную шкалу вместо первоначальной, по которой эти переменные были измерены.. Такая замена с использованием p-мерного пространства широко применяется в литературе по многомерному статистическому анализу. Так же как поверхность тренда является двумерным аналогом аппроксимирующей кривой, множественная регрессия является дальнейшим обобщением этих методов для многомерных случаев.

Мы не будем останавливаться на множественной регрессии очень подробно, так как теоретические и вычислительные аспекты этой теории были изложены в предыдущих главах. Мы напомним только (см. гл. 4), что полиномиальную регрессию от одной независимой переменной) можно представить с помощью уравнения следующего вида:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_1^m + \varepsilon \qquad (6.1)$$

Эта модель утверждает, что наблюдение Y складывается из. постоянного члена, степенного ряда некоторой независимой переменной и случайной компоненты. Нахождение коэффициентов β в этом линейном уравнении по методу наименьших квадратов осуществляется путем решения системы нормальных уравнений, которую в матричной форме можно записать следующим образом:

$$\left[\sum X\right] \times [\beta] = \left[\sum Y\right] \quad (6.2)$$

с решением

$$[\beta] = \left[\sum X\right]^{-1} \times \left[\sum Y\right] \qquad (6.3)$$

где $\left[\sum Y\right]$ – матрица, столбцы которой состоят из сумм квадратов и смешанных произведений переменной Y с переменными X_1, X_1^2, ... X_1^m; $\left[\sum X\right]$ – матрица сумм квадратов и смешанных произведений степеней переменных X_i; [β] – столбец неизвестных коэффициентов. В гл. 4 мы находили элементы таких матриц, перемножая строки и столбцы исходных матриц.

Хотя в рассмотренном примере мы имели дело только с одной независимой переменной (или с двумя, как в случае анализа поверхностей тренда, рассмотренном в гл. 5), эту же задачу можно трактовать и как многомерную, содержащую т независимых переменных. Это станет очевидным, если мы запишем наше уравнение в следующем виде:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \qquad (6.4)$$

и определим переменные как $X_1 = X_1$, $X_2 = X_1^2$, $X_3 = X_1^3$ и т. д. Тогда регрессионную модель, рассмотренную раньше, можно будет считать частным случаем многомерной модели, в которой независимые переменные определены некоторым специфическим образом. Соотношения между различными типами регрессионного анализа представлены на рис. 5.89.

Общий вид уравнения регрессии зависимой переменной относительно m независимых переменных представлен формулой (6.4). Система нормальных уравнений для нахождения решения по методу наименьших квадратов снова получается в результате попарного перемножения строк и столбцов матрицы уравнения. Для трех независимых переменных мы получим

$$\begin{bmatrix} & X_0 & X_1 & X_2 & X_3 \\ X_0 & & & & \\ X_1 & & & & \\ X_2 & & & & \\ X_3 & & & & \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} Y \\ \\ \\ \\ \end{bmatrix}$$

где X_0 снова принимает значения 1 для всех наблюдений. Матричное уравнение после вычисления смешанных произведений выглядит следующим образом:

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 & \sum X_3 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 & \sum X_1 X_3 \\ \sum X_2 & \sum X_2 X_1 & \sum X_2^2 & \sum X_2 X_3 \\ \sum X_3 & \sum X_3 X_1 & \sum X_3 X_2 & \sum X_3^2 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \sum X_3 Y \end{bmatrix} \qquad (6.5)$$

Коэффициенты p, регрессионной модели оцениваются с помощью выборочных частных коэффициентов регрессии b_i. Они носят название частных коэффициентов регрессии по той причине, что каждый из них характеризует скорость изменения (или наклон) по отношению к одной независимой переменной при условии, что все остальные переменные фиксированы. В некоторых руководствах для отражения этого факта используется следующая запись:

$$Y = b_0 + b_{1,23} X_1 + b_{2,13} X_2 + b_{3,12} X_3 + \varepsilon$$

где коэффициент $b_{1,23}$ называется коэффициентом регрессии переменной Y на X_1 при фиксирован-

ных переменных 2 и 3. Эти коэффициенты в общем случае отличаются от общих регрессионных коэффициентов, которые характеризуют простую регрессию переменной Y на каждой отдельной переменной X. Как и следовало ожидать, множественная регрессия вносит в общую изменчивость больший вклад, чем любой из общих регрессионных коэффициентов. Это происходит по той причине, что множественная регрессия строится на основе учета всех возможных взаимодействий между переменными и их комбинациями.

В качестве типичного примера использования уравнения множественной регрессии, мы рассмотрим задачу из геоморфологии. Для этой цели некоторый район восточной части штата Кентукки был разделен на относительно однородные в геологическом отношении области. Изучаемый район охватывает ряд дренажных бассейнов различных размеров, из которых были отобраны все бассейны третьего порядка, и в каждом из них измерены значения некоторых переменных. Порядок бассейна определяется числом последовательных уровней слияния в его русле от истоков до точки, в которой поток сливается с другим потоком того же или более высокого порядка. Таким образом, бассейном третьего порядка считается бассейн, имеющий два уровня. Однако размер бассейна можно определить различными методами. Один из них по существу сводится к подсчету числа русел в бассейне. Множеству бассейнов данного порядка могут соответствовать различные значения величины бассейна. Соотношение между величиной и порядком бассейнов представлено на рис. 6.1.

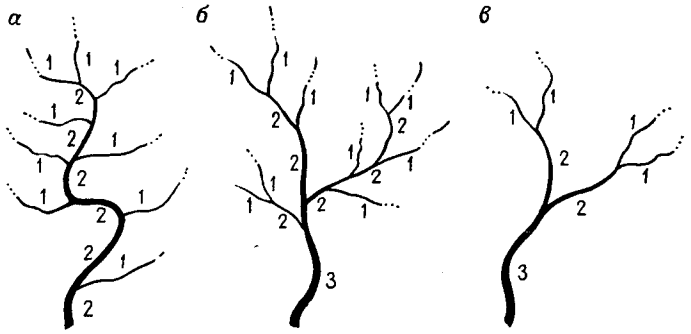


Рис. 6.1. Примеры потоков различных величин и порядков:
а – величина 10, порядок 2; *б* – величина 10, порядок 3; *в* – величина 4, порядок 3. Величиной называется число слияний потоков, порядком – число последовательных уровней слияния

На множестве бассейнов третьего порядка были измерены следующие семь переменных:

1) высота истоков бассейна над уровнем моря (в футах);
2) характеристика рельефа бассейна (в футах);
3) площадь бассейна (в квадратных милях);
4) общая протяженность русел в бассейне (в милях);
5) плотность дренажа, измеряемая как отношение общей длины русел в бассейне к его площади;
6) характеристика формы бассейна, измеряемая как отношение площадей вписанного и описанного кругов;
7) величина бассейна, определяемая числом истоков.

Наша задача состоит в изучении влияния первых шести переменных на седьмую. Для этой цели подходящим оказывается метод множественной регрессии, причем величина бассейна используется в качестве зависимой переменной. Уравнение регрессии позволяет оценить влияние всех переменных на величину бассейна. Измерения значений этих переменных для 92 бассейнов третьего порядка в изучаемом районе приведены в табл. 6.1, которая взята из книги Крамбейна и Шрива [38]. Значимость полученного линейного соотношения можно проверить методами дисперсионного анализа, описанными в гл. 4. Например, табл. 4.12, в которой приведена схема ANOVA для простой линейной регрессии, можно расширить до схемы множественной регрессии, если сделать соответствующие изменения в числах степеней свободы с учетом дополнительных переменных. Схема модифицированного ANOVA приведена в табл. 6.2. Результаты ANOVA для множественной регрессии величины бассейна указаны в табл. 6.3. Коэффициенты регрессии также приведены.

В задачах множественной регрессии нас обычно интересует относительная эффективность предсказания зависимой переменной по набору аргументов. Однако мы не можем сделать этого на основании прямого исследования выборочных коэффициентов регрессии, так как их величины зависят от значений самих переменных. Эта зависимость легко прослеживается при построении поверхностей тренда, где коэффициенты при членах высшего порядка дают больший вклад в тренд, чем члены более низкого порядка. Это вытекает из того, что высокая степень переменной имеет значительно больший порядок, чем первоначальная переменная. Соответственно коэффициенты при регрессионных членах более высокого порядка уменьшаются.

Таблица 6.1

Результаты измерения семи геоморфологических переменных в речных бассейнах третьего порядка штата Кентукки

Y	X_1	X_2	X_3	X_4	X_5	X_6
14	720	570	07	154	2200	61
6	670	610	03	80	2667	62
5	860	550	11	84	763	62
7	870	610	11	122	1110	63
11	730	570	14	185	1321	52
14	690	590	12	200	1667	50
12	880	640	11	170	1545	41
18	760	690	28	340	1215	57
6	820	600	5	100	2000	41
5	720	480	3	80	2667	60
17	670	670	19	290	1526	51
5	660	600	5	90	1800	53
22	830	660	18	260	1444	57
7	780	620	17	111	652	57
15	750	740	15	184	1227	67
17	770	630	21	227	1080	59
5	750	570	4	60	1500	55
18	750	580	50	259	1295	39
14	740	760	9	62	689	64
21	750	740	6	95	1583	53
22	750	760	11	105	954	64
23	740	770	32	350	1094	55
28	940	510	21	232	1105	52
42	700	600	32	266	1156	34
22	810	580	44	390	886	29
10	920	500	13	142	1092	65
11	920	490	12	145	1208	72
12	790	605	33	253	766	59
13	860	550	23	241	1048	76
31	860	630	87	702	807	55
18	880	520	37	288	778	51
13	780	460	17	162	953	40
4	720	440	8	67	838	60
5	780	300	3	52	1733	57
9	700	460	10	121	1210	50
13	680	520	26	220	846	41
10	820	520	8	123	1537	51
13	710	520	24	238	992	41
13	800	440	19	231	1216	51
11	700	510	16	178	1113	76
12	675	570	18	168	933	42
4	740	510	8	65	812	49
17	740	520	31	334	1078	67
9	770	600	21	184	876	47
8	820	520	11	136	1237	56
13	850	490	22	233	1059	74
22	820	629	34	410	1206	39
10	820	510	11	149	1354	60
19	680	640	46	348	757	55
27	660	789	55	382	695	38

Переменные: Y – величина бассейна; X_1 – абсолютная отметка истоков бассейна (в футах); X_2 – характеристика рельефа бассейна (в футах); X_3 – площадь бассейна (в квадратных милях); X_4 – общая длина русел бассейна (в милях): X_5 – плотность дренажа (отношение общей длины русел к площади бассейна); X_6 – отношение площадей наибольшего вписанного круга и наименьшего описанного круга [16].

Таблица 6.2

Дисперсионный анализ (ANOVA) для множественной регрессии с m независимыми переменными

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние значения	F-критерий
Регрессия	SS_R	m	MS_R	MS_R/MS_D
Отклонение	SS_D	$n–m$-1	MS_D	
Сумма	SS_T	n-1		

Таблица 6.3

Результаты определения значимости регрессионной зависимости величины бассейна от шести геоморфологических переменных с помощью дисперсионного анализа*

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние значения	F-критерий
Регрессия	1800,70	6	300,12	11,38
Отклонение	1134,12	43	26,37	
Сумма	2934,82	49		

Уравнение регрессии: $Y= –2{,}24 + 0{,}01X_1 + 0{,}02X_2 – 23{,}28X_3 + 6.26X_4 – 0{,}20X_5 –11{,}66X_6$. Коэффициент множественной регрессии: R=0,78.

К счастью, частные коэффициенты регрессии легко выразить в единицах стандартного отклонения (Ли [43]). Стандартные коэффициенты частной регрессии B_k находятся по формуле

$$B_k = b_k s_k / s_y \qquad (6.6)$$

где S_k – оценка стандартного отклонения переменной X_k и S_y – оценка стандартного отклонения величины Y. Так как все стандартные частные коэффициенты регрессии выражаются в единицах стандартного отклонения, то их можно прямо сравнить друг с другом и определить наиболее эффективные из них.

Вычислив элементы матрицы сумм квадратов и произведений, необходимых для построения нормальных уравнений, найдем диагональные элементы $\sum X_k^2$ и по ним вычислим исправленные суммы квадратов SS_k и затем стандартные отклонения, необходимые для вычисления частных коэффициентов корреляции. Однако можно получить решение нормальных уравнений в таком виде, что при этом прямо получаются значения стандартизованных частных коэффициентов регрессии, в результате чего получается значительный выигрыш в вычислительном процессе.

Большая часть ошибок при построении множественной регрессии возникает при вычислении элементов матрицы $\sum X$ и в процессе ее обращения. Суммы квадратов переменных X_k могут возрасти настолько, что при отбрасывании разрядов, выходящих за пределы разрядной сетки, могут быть потеряны значащие цифры. Далее, если элементы матрицы $\sum X$ сильно различаются по величине, то при ее обращении может произойти дополнительная потеря знаков, в особенности в тех случаях, когда между переменными имеется высокая корреляция. Некоторые вычислительные программы в состоянии сохранить только одну или две значащих цифры в коэффициентах, а с некоторыми данными дело может обстоять еще хуже. Исследования показали, что вычислений с использованием двойной точности недостаточно для того, чтобы преодолеть эту трудность. Однако несколько простых изменений в программе позволят сохранить в процессе вычислений от двух до шести значащих цифр и значительно повысить точность уравнения регрессии [44].

Во-первых, все наблюдения заменим на их отклонения от среднего значения. Это преобразование позволяет уменьшить абсолютную величину переменных и приводит к переменным, имеющим общее среднее значение, равное нулю. При этом преобразовании коэффициент b_0 обращается в нуль, так что порядок матрицы системы уравнений снижается на единицу. В результате такого преобразования сохраняется несколько значащих цифр. Однако порядок величин элементов матрицы можно еще уменьшить, если использовать вместо них соответствующие коэффициенты корреляции.

Это преобразование соответствует записи исходных переменных в стандартной нормальной форме с нулевым средним и единичным стандартным отклонением. Матричное уравнение для определения коэффициентов регрессии тогда примет вид

$$[r_{xx}]\times[B]=[r_{xy}] \qquad (6.7)$$

а его решение запишется так:

$$[B]=[r_{xx}]^{-1}\times[r_{xy}] \qquad (6.8)$$

Здесь $[r_{xy}]$ – вектор-столбец коэффициентов корреляции между переменной Y и независимыми переменными X_k. Матрица коэффициентов корреляции между переменными X_k порядка $m\times m$ обозначается через $[r_{xy}]$. Например, нормальное уравнение для трех независимых переменных имеет вид

$$\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = \begin{bmatrix} r_{x,y} \\ r_{x,y} \\ r_{x,y} \end{bmatrix} \qquad (6.9)$$

Отметим, что в этом уравнении на одну строку и один столбец меньше, чем в эквивалентном уравнении (6.5).

Однако этот метод, основанный на вычислении корреляционной матрицы и получении стандартизованного уравнения регрессии, имеет тот недостаток, что он увеличивает объем вычислений. Для сохранения точности коэффициенты корреляции рекомендуется вычислять не по формуле (2.24), а на основании определяющего уравнения. Использование формулы (2.24) нецелесообразно по той причине, что она содержит квадраты величин $\sum X_j$ и $\sum X_k$. Если эти суммы велики, то их квадраты могут оказаться неточными за счет отбрасывания разрядов, выходящих за пределы разрядной сетки. Этой проблемы не возникает, если до вычисления сумм квадратов из каждого наблюдения вычесть среднее значение. Суммы квадратов находятся по формулам (2.16) и (2.19). Для осуществления этой операции требуется использовать исходные данные дважды – первый раз для вычисления среднего значения, а затем при вычитании полученного значения из наблюдений. В то же время как при вычислениях вручную это приводит к значительному увеличению объема работы, на вычислительной машине такая операция проводится очень просто. Вычисленные коэффициенты должны выдаваться в «нестандартизированном» виде, так как они затем используются для построения уравнения прогноза вместе с необработанными данными. Однако этот недостаток окупается преимуществами возрастающей устойчивости и точности матричного решения, а стандартизованные коэффициенты дают возможность оценить величины вкладов отдельных переменных в уравнение регрессии. Коэффициенты частной регрессии можно получить из стандартизированных коэффициентов частной регрессии с помощью преобразования

$$b_k = B_k s_y / s_k \qquad (6.10)$$

Постоянный член b_0 находится по формуле

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - \ldots - b_m\overline{X}_m \qquad (6.11)$$

Несмотря на то, что суммы квадратов изменяются при стандартизации данных или при использовании матричного уравнения в корреляционной форме, отношения сумм квадратов остаются неизменными. Поэтому критерии значимости, основанные на стандартизованной регрессии, идентичны критериям, основанным на нестандартизованной регрессии. Такие величины, как коэффициент множественной корреляции (R) и процентное выражение точности аппроксимации (100% R^2), также остаются неизменными.

Данные, приведенные в табл. 6.4, представляют собой характеристики нефтегазоносного бассейна в Арканзасе. Зависимой переменной является оценка запасов нефти в некотором участке бассейна, вычисленная на основании метода материального баланса. Уравнение материального баланса в сущности является соотношением между добычей нефти, добычей газа и давлением. В него включаются также допущения об объеме резервуара и начальных объемах нефти, газа и воды. Независимыми переменными являются время заполнения резервуара, давление в нем, общая добыча нефти, кумулятивное отношение добычи газа к добыче нефти. Так как между зависимой переменной и аргументами в уравнении материального баланса имеется неявная связь, то мы вправе ожидать необычно высокую внутреннюю корреляцию. Действительно, если модель материального баланса

выбрана удачно и наши представления о начальном состоянии и объеме резервуара правильны, то корреляция будет высокой. Неудачные попытки полностью оценить размеры нефтяных запасов могут быть связаны с ошибками в начальных допущениях или с неполным исследованием всех факторов, входящих в уравнение материального баланса.

Таблица 6.4

Числовые характеристики одной из залежей нефтегазоносного поля в Арканзасе

Y	X_1	X_2	X_3	X_4
110273,0	1,0	3520,0	0,0	760,0
111105,0	4,0	3125,0	29183,0	853,0
114992,0	8,0	2910,0	46536,0	906,0
119437,0	12,0	2785,0	60302,0	939,0
118961,0	16,0	2650,0	73604,0	960,0
116968,0	20,0	2505,0	87513,0	990.0
119663,0	24,0	2425,0	98738,0	1018,0
117514,0	28,0	2290,0	112587,0	1070,0
117292,0	32,0	2125,0	126192,0	1200,0
114776,0	36,0	1950,0	139981,0	1310,0
113969,0	40,0	1785,0	153419,0	1440,0
111881,0	44,0	1670,0	161327,0	1500,0
114455,0	48,0	1601,0	173485,0	1516,0
116196,0	52,0	1537,0	185832,0	1520,0

Переменные: Y – оцениваемые запасы нефти в исследуемом районе (x10^3 баррелей); X_1 – время после завершения полевых работ (месяцы); X_2 – давление в залежи (фунт/дюйм^2); X_3 – суммарная добыча нефти (x10^2 баррелей); X_4 – отношение добытого количества газа к добытому объему нефти (фут^3/баррель).

Эти данные содержат некоторые характеристики, представляющие трудности для анализа. Так как порядки значений изучаемых переменных сильно различаются, то элементы матрицы смешанных произведений также сильно отличаются по величине. Эти данные образуют многомерный временной ряд. Так же, как и в других рядах этого типа, таких, как кривые роста экономики или использования трудовых ресурсов, переменные сильно коррелированны. Сохранить достаточное количество цифр в матричных вычислениях или сохранить точность в процессе обращения оказывается затруднительным. Полезно вычислить коэффициенты регрессии, используя матрицу $\sum X$ и матрицу r_{xx}. Для сравнения стандартизованные частные коэффициенты регрессии должны быть преобразованы к обычному виду (6.10) и (6.11). Различия, которые можно обнаружить, возникают из-за ошибок округления при использовании матрицы $\sum X$.

Несмотря на то что стандартизованные частные коэффициенты регрессии позволяют находить наиболее важные переменные, входящие в уравнение регрессии, они не могут служить непогрешимым указанием на то, что это уравнение выбрано наилучшим образом. Предположим, что, исследуя уравнение регрессии, мы пришли к выводу, что две переменные дают несущественный вклад в регрессию и их можно отбросить. Если одну из переменных устранить и снова построить уравнение регрессии, то качество подбора и само уравнение, конечно, изменятся. Если мы решили устранить вторую переменную, уравнение регрессии снова изменится, но изменение может быть совсем иным по сравнению с изменением, которое произойдет в том случае, если первая переменная сохранится в регрессии. Это происходит по той причине, что эффекты взаимодействия двух отбрасываемых переменных с другими переменными нельзя оценить без повторного построения регрессионного уравнения. Если необходимо провести исследование большого числа переменных и отбросить те переменные, которые несущественны для данной задачи, то мы не должны ограничиваться простым исследованием частных коэффициентов регрессии.

Увеличение числа независимых переменных в уравнении регрессии всегда ведет к увеличению SS_R (исключая те случаи, когда новые переменные полностью коррелированны со старыми). Однако это увеличение не может быть значительным. Потерю степеней свободы отклонений можно

компенсировать уменьшением SS_D, что в действительности приводит к увеличению среднего значения квадратов отклонений. Если это происходит, то F-отношение уменьшается, что приводит к сокращению числа членов в уравнении регрессии. Для определения наилучшей возможной регрессии (наиболее значимого F-отношения) приходится исследовать всевозможные комбинации переменных; если переменных немного, это сделать легко, так как число их возможных комбинаций равно 2^m. Однако если m велико, эта процедура требует значительных затрат машинного времени. Существуют другие процедуры, которые позволяют получать оптимальную регрессию со значительно меньшими затратами времени. Среди них можно назвать обратную процедуру исключения, прямую процедуру выбора, методы пошаговой и многошаговой регрессии. При большом количестве исходных переменных эти методы не всегда приводят к одинаковым уравнениям регрессии, однако результаты, полученные на их основании, все же эквивалентны. Изложение этих методов не входит в наши задачи, и мы приведем лишь краткое описание одного из них. Эти методы хорошо изложены в некоторых руководствах, например, в книгах Дрейпера и Смита [14] и Мараскило и Левина [46].

Обратная процедура исключения сводится к построению уравнений регрессии, включающих все возможные переменные, и в последующем отборе наименее значимых аргументов. Отбор проводится путем исследования стандартизированных коэффициентов частной регрессии с наименьшими значениями и последующего построения уравнения регрессии, из которого удалены эти переменные. Значимость отбрасываемых переменных проверяется с помощью приемов дисперсионного анализа, аналогичных представленным в табл. 4.16. Если переменная не дает значимого вклада в регрессию, то она обыкновенно отбрасывается. Затем стандартизированные коэффициенты частной регрессии приведенного уравнения анализируются снова, и процесс повторяется. На каждом шаге число переменных в уравнении регрессии уменьшается на единицу до тех пор, пока все оставшиеся переменные не окажутся значимыми.

Весьма полезно исследование набора семи переменных, представляющих характеристики бассейна рек (см. рис. 6.1), с целью возможного исключения каких-либо из них. Исследуя стандартизованные коэффициенты частной регрессии, и отбрасывая наименьшие из них и снова вычисляя регрессию, мы можем найти минимальное множество аргументов регрессии.

Повторное применение программы множественной регрессии, очевидно, менее эффективно, чем использование пошагового вычислительного алгоритма, но оно имеет то преимущество, что каждый шаг процесса может быть тщательно проанализирован. После того как будет достигнуто понимание процессов исключения и изменения, происходящих при вычислении коэффициентов регрессии, можно обратиться к более автоматизированным алгоритмам.

Хотя по внешним признакам теорию множественной регрессии можно отнести к «многомерным» теориям, так как в ней участвует несколько переменных, измеренных на каждом объекте наблюдения, все же по существу своему она является одномерной, так как мы имеем дело с дисперсией только одной зависимой переменной Y, а поведение независимых переменных X анализу не подвергается.

Следующая тема нашего изложения – дискриминантный анализ, цель которого – идентификация или распределение объектов в заранее заданные группы. Разделение на две взаимно исключающие друг друга группы – это процесс, который в вычислительном плане является промежуточным между одномерными процедурами и настоящими многомерными методами, в которых много переменных рассматриваются одновременно. Две группы, каждая из которых характеризуется некоторым множеством многомерных переменных, можно разделить с помощью решения некоторого множества совместных уравнений, почти таких же, как те, которые используются в множественной регрессии. Вектор правой части матричного уравнения, однако, не содержит степеней и попарных произведений единственной зависимой переменной, а содержит разности между многомерными средними этих двух групп.

Критерии теории дискриминантных функций включают многомерные обобщения простых одномерных статистических критериев проверки гипотез о равенстве. Они будут рассмотрены позже, после многомерных методов классификации или распределения объектов в однородные группы. Затем мы рассмотрим методы, в которых используются собственные значения, включая метод главных компонент и факторный анализ. Последние параграфы содержат многомерные обобщения дискриминантного анализа и множественной регрессии.

Этот перечень, очевидно, не является исчерпывающим. Однако рассматриваемые методы были выбраны по той причине, что они нашли применение в науках о Земле. Они включают множество вычислительных методов и оперируют с рядом фундаментальных понятий. Понимание теории

этих методов и соответствующих ей вычислительных процедур обеспечивает достаточную базу для оценки других многомерных методов.

ДИСКРИМИНАНТНЫЕ ФУНКЦИИ

Один из наиболее широко используемых в науках о Земле многомерных методов – дискриминантный анализ. Мы рассматриваем его здесь потому, что, во-первых, он является мощным статистическим методом, и, во-вторых, его можно поставить в один ряд с одномерными задачами, связанными с множественной регрессией или рассмотренными выше многомерными задачами проверки статистических гипотез. Поэтому он позволяет установить дополнительную связь между одномерной и многомерной статистикой.

Определим сначала понятие разделения (дискриминации) и покажем, чем оно отличается от близкого к нему понятия классификации. Предположим, что имеются две группы проб сланцев, о которых заранее известно, что они образовались в пресноводном и морском бассейнах. Это можно определить на основании исследования остатков ископаемых организмов. В пробах измерено некоторое число геохимических характеристик, а именно, содержания ванадия, бора, железа и других элементов. Задача состоит в нахождении такой линейной комбинации этих переменных, которая даст максимально возможное различие между двумя ранее определенными группами. Если нам удастся найти такую функцию, то мы сможем использовать ее для отнесения новых образцов к той или другой исходной группе. Иными словами, новые образцы сланца, не содержащие диагностических ископаемых остатков, можно будет разделить на морские и пресноводные на основе линейной дискриминантной функции, построенной по их геохимическим компонентам. (Эта задача рассматривалась Поттером, Шимпом и Уиттерсом [55]).

Задачу классификации можно проиллюстрировать на аналогичном примере. Предположим, что мы собрали большую коллекцию образцов сланцев, каждый из которых был подвергнут геохимическому анализу. Можно ли на основе значений измеренных переменных осуществить разделение выборки на относительно однородные группы (кластеры), отличающиеся друг от друга? Численные методы решения такого рода задач достаточно хорошо разработаны и принадлежат к разделу науки, называемому таксономией. Они будут рассмотрены в следующем разделе. Существует несколько явных различий между этими методами и методами дискриминантного анализа. Классификация внутренне замкнута, т.е., в отличие от дискриминантного анализа, она не зависит от априорных сведений о соотношении между пробами. В дискриминантном анализе число групп задается заранее, в то время как число кластеров, которые получаются в результате классификации, не может быть заранее определено. Каждая проба из исходного множества в дискриминантном анализе принадлежит к одной из заданных групп. В большинстве задач классификации проба может войти в любую из групп, возникающих в результате классификации. Другие различия станут очевидными при рассмотрении этих двух процедур. В результате кластерного анализа сланцев пробы распределяются по группам. Представляет интерес проведение геологического осмысливания найденных таким образом групп.

Простая линейная дискриминантная функция осуществляет преобразование исходного множества измерений, входящих в выборку, в единственное дискриминантное число. Это число, или преобразованная переменная, определяет положение образца на прямой, определенной дискриминантной функцией. Поэтому мы можем представлять себе дискриминантную функцию. как способ преобразования многомерной задачи в одномерную.

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 совокупностями точек в многомерном пространстве, то легко найти такое направление, вдоль которого эти совокупности явно разделяются и в то же время имеют наименьшую выпуклость. Графически эта картина представлена на рис. 6.2. Если использовать переменные X_1 и X_2, то провести удовлетворительное разделение групп A и B не удается. Однако можно найти направление, вдоль которого разделение совокупностей очевидно, а выпуклость минимальна. Координаты точек этого направления задаются уравнением линейной дискриминантной функции.

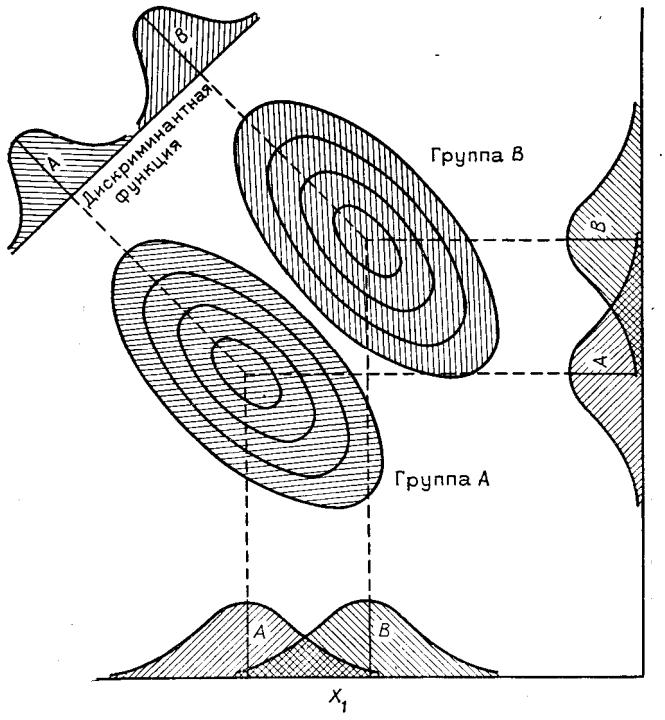


Рис. 6.2. Графическое представление двух двумерных распределений. Указаны перекрытия распределений для групп A и B по осям X_1 и X_2; проектирование на дискриминантную линию позволяет различить две группы

Один из методов нахождения линейной дискриминантной функции – построение уравнения регрессии, в котором зависимыми переменными являются разности между многомерными средними двух групп. В матричном обозначении мы должны решить уравнение вида

$$\left[s_p^2\right] \times [\lambda] = [D] \qquad (6.12)$$

где $\left[s_p^2\right]$ – $m \times m$ – матрица дисперсий и ковариаций объединенной выборки. Коэффициенты дискриминантной функции представляются вектором-столбцом неизвестных, который обозначен буквой λ. Его коэффициенты играют ту же роль, что и коэффициенты β в уравнении регрессии. Не надо путать эти коэффициенты λ с теми, которые используются для обозначения собственных значений матриц в компонентном и факторном анализах.

В правой части уравнения (6.12) стоит вектор-столбец разностей между средними значениями двух групп. Такое уравнение решается с помощью операций обращения и умножения матриц, т.е.

$$[\lambda] = \left[s_p^2\right]^{-1} \times [D] \qquad (6.13)$$

Чтобы определить дискриминантную функцию, мы должны определить величины, входящие в матричное уравнение. Разности средних находятся по формуле

$$D_j = \overline{A}_j - \overline{B}_j = \frac{\sum_{i=1}^{n_a} A_{ij}}{n_a} - \frac{\sum_{i=1}^{n_b} B_{ij}}{n_b} \qquad (6.14)$$

где A_{ij} – i-е наблюдение j-й переменной в группе A; A_j – среднее значение j-й переменной группы A, или среднее по n_a наблюдениям. Те же обозначения используются для группы B. Многомерные средние переменные групп A и B можно считать двумя векторами. Поэтому разность между ними снова образует вектор

$$\left[D_j\right] = \left[\overline{A}_j\right] - \left[\overline{B}_j\right]$$

или в расширенной записи

$$\begin{bmatrix} D_1 \\ D_2 \\ \dots \\ D_m \end{bmatrix} = \begin{bmatrix} \overline{A}_1 \\ \overline{A}_2 \\ \dots \\ \overline{A}_m \end{bmatrix} - \begin{bmatrix} \overline{B}_1 \\ \overline{B}_2 \\ \dots \\ \overline{B}_m \end{bmatrix}$$

Для построения ковариационной матрицы объединенной выборки нужно вычислить матрицу сумм квадратов и смешанных произведений (*SP*) для всех переменных в группе A и аналогичную матрицу для группы B. Например, если рассмотреть только группу A, то

$$SP\left(A_{jk}\right)=\sum_{ii=1}^{n_a}\left(A_{jk}A_{ik}\right)-\frac{\sum_{i=1}^{n_a}A_{ij}\sum_{i=1}^{n_a}A_{ik}}{n_a}$$

Здесь, как и ранее, A_{ij} – 1-е наблюдение j-й переменной в группе A; A_{ik} – i-е наблюдение k-й переменной в той же группе. Конечно, при $j=k$ эта величина даст сумму квадратов переменной с номером k. Аналогично можно найти матрицу сумм квадратов и смешанных произведений для группы B:

$$SP\left(B_{jk}\right)=\sum_{ii=1}^{n_a}\left(B_{jk}B_{ik}\right)-\frac{\sum_{i=1}^{n_b}B_{ij}\sum_{i=1}^{n_b}B_{ik}}{n_b}$$

Обозначим для простоты записи матрицу сумм произведений для группы A через [SPA], а для группы B – через [SPB]. Ковариационную матрицу объединенной выборки теперь можно записать в виде

$$\left[s_p^2\right]=\frac{\left[SPA\right]+\left[SPB\right]}{n_a+n_b-2} \quad (6.15)$$

Легко видеть, что это определение дисперсионной матрицы объединенной выборки в точности такое же, как и использованное при рассмотрении критерия T^2 для проверки гипотезы о равенстве многомерных средних. Хотя объем вычислений, которые необходимо провести для того, чтобы получить коэффициенты дискриминантной функции, на первый взгляд и кажется большим, фактически он значительно меньше. В качестве примера построим дискриминантную функцию для двух групп данных, приведенных в табл. 6.5. 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 уреза воды. Переменные в этом случае такие же, как и для группы A. Точечная диаграмма исходных наблюдений представлена на рис. 6.3. Хотя две группы точек и перекрываются, совершенно очевидно, что разделяющая их линия проходит между ними так, что большинство наблюдений группы А находится по одну сторону от нее, а большинство наблюдений группы B – по другую.

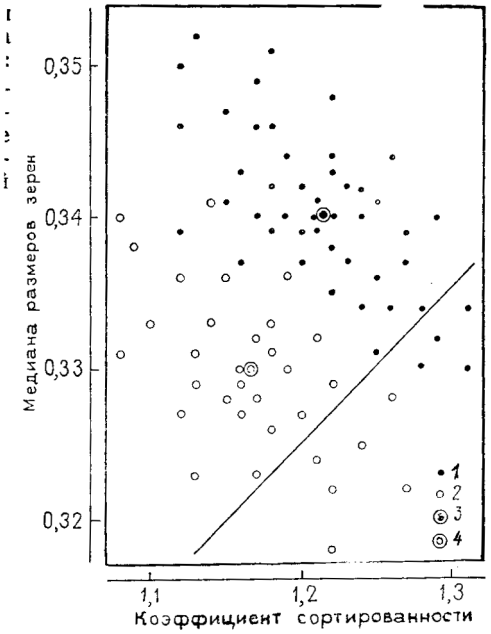


Рис. 6.3. Зависимость медианы размеров зерен от коэффициента отсортированности в пробах песка:
1 – пробы пляжного песка; *2* – пробы, взятые в отдалении от берега; *3* и *4* – двумерные средние двух групп функции. Прямая линия – график дискриминантной функции

Таблица 6.5

Результаты измерения среднего размера зерен и коэффициента отсортированности в двух группах проб песка, взятых у уреза воды (*A*) и в удалении от него (*B*)

Средний размер зерен	Коэффициент отсортированности	Средний размер зерен	Коэффициент отсортированности	Средний размер зерен	Коэффициент отсортированности	Средний размер зерен	Коэффициент отсортированности
Группа *A*		0.332	1,17	0,350	1,12	0,339	1,20
0,333	1,08	0,331	1,18	0,352	1,13	0,342	1,20
0,340	1,08	0.326	1,18	0,341	1,15	0,339	1,21
0,338	1,09	0,333	1,18	0,347	1,15	0,340	1,21
0,333	1,10	0.330	1,19	0,337	1,16	0,341	1,21
0,323	1,13	0,335	1,19	0,343	1,16	0,335	1,22
0,327	1,12	0,327	1,20	0,340	1,17	0,337	1,22
0,329	1,13	0,324	1,21	0,346	1,17	0,340	1,22
0,331	1,13	0,332	1,21	0,349	1,17	0,343	1,22
0,336	1,12	0,322	1,22	0,339	1,18	0,337	1,22
0,333	1,14	0,329	1,22	0,342	1,18	0,342	1,23
0,341	1,14	0,325	1,24	0,346	1,18	0,334	1,24
0,328	1,15	0,328	1,26	0,351	1,18	0,340	1,24
0,336	1,15	0,322	1,27	0,339	1,27	0,342	1,24
0,327	1,16	Группа *B*		0,330	1,28	0,331	1,25
0,329	1,16	0,340	1,19	0,334	1,27	0,336	1,25
0.330	1,16	0,344	1,19	0,332	1,28	0,341	1,25
0,323	1,17	0,339	1,12	0,334	1,22	0,334	1,26
0,318	1,22	0,333	1,20	0,330	1,31	0,340	1,21
0,330	1,17	0,337	1,20	0,334	1,31	0,337	1,27
0.328	1,17	0,346	1,12	0,348	1,22		

В табл. 6.6 приведены результаты вычислений двух векторов многомерных средних и двух матриц сумм квадратов и смешанных произведений. На основании этих данных вычисляется ковариационная матрица объединенной выборки. Теперь у нас есть данные для нахождения дискриминантной функции

$$\left[s_p^2\right]^{-1} \times [D] = [\lambda]$$

$$\begin{bmatrix} 59112{,}280 & 4312{,}646 \\ 4312{,}646 & 747{,}132 \end{bmatrix} \times \begin{bmatrix} -0{,}010 \\ -0{,}043 \end{bmatrix} = \begin{bmatrix} -783{,}63 \\ -75{,}62 \end{bmatrix}$$

Полученное множество коэффициентов λ используется для построения дискриминантной функции вида

$$R = \lambda_1\psi_1 + \lambda_2\psi_2 + \ldots + \lambda_n\psi_m \quad (6.16)$$

Это – линейная функция; суммируя ее слагаемые, получим число, называемое дискриминантной меткой. В двумерном случае мы можем изобразить дискриминантную функцию прямой линией на точечной диаграмме двух исходных переменных. Это прямая с угловым коэффициентом

$$\alpha = \lambda_2 / \lambda_1 \qquad (6.17)$$

Такая линия и изображена на рис. 6.3.

Подставляя в уравнение дискриминантной функции среднее арифметическое, полученное из средних для двух выборок, мы получаем значение дискриминантного индекса R_0. Иными словами, каждое значение ψ_J в формуле (6.16) мы полагаем равным

$$\psi_j = \left(\overline{A}_j + \overline{B}_j\right)/2 \qquad (6.18)$$

В нашем примере

$$R_0 = \lambda_1\psi_1 + \lambda_2\psi_2 = -783{,}63(0{,}335) - (-75{,}62)(1{,}189) = -352{,}22$$

Таблица 6.6

Матрицы, используемые при вычислении дискриминантной функции для двух групп наблюдений, приведенных в табл. 6.5

Вектор средних значений группы A	$[0{,}330 \quad 1{,}167]$
Вектор средних значений группы B	$[0{,}340 \quad 1{,}210]$
Вектор разностей средних	$[-0{,}010 \quad -0{,}043]$
Исправленная матрица сумм квадратов A	$\begin{bmatrix} 0{,}00092 & -0{,}00489 \\ -0{,}00489 & 0{,}07566 \end{bmatrix}$
Исправленная матрица сумм квадратов B	$\begin{bmatrix} 0{,}00138 & -0{,}00844 \\ -0{,}00844 & 0{,}10700 \end{bmatrix}$
Ковариационная матрица объединенной выборки	$\begin{bmatrix} 0{,}00003 & -0{,}00017 \\ -0{,}00017 & 0{,}00231 \end{bmatrix}$
Матрица, обратная к ковариационной матрице объединенной выборки	$\begin{bmatrix} 59112{,}280 & 4312{,}646 \\ 4312{,}646 & 747{,}132 \end{bmatrix}$

Дискриминантный индекс R_0 соответствует точке разделяющей прямой, которая лежит строго посередине между центром группы A и центром группы B. Мы можем подставить многомерное среднее группы A в уравнение, т.е. принять, что $\psi_j = \overline{A}_j$.

Это даст нам значение R_A. Аналогично, подстановка среднего группы B даст нам значение R_B (при $\psi_j = \overline{B}_j$). Эти значения определяют центры двух исходных групп на разделяющей прямой:

$$R_A = \lambda_1 \overline{A}_1 + \lambda_2 \overline{A}_2 = -783{,}63(0{,}330) - (-75{,}62)(1{,}167) = -346{,}64;$$

$$R_A = \lambda_1 \overline{B}_1 + \lambda_2 \overline{B}_2 = -783{,}63(0{,}340) - (-75{,}62)(1{,}210) = -357{,}81$$

Эти три точки изображены на рис. 6.4. Аналогично каждое наблюдаемое значение можно подставить в дискриминантное уравнение и затем нанести полученное число на график. Все это можно сделать на одной диаграмме; заметим, что несколько точек группы A попали в группу B, т. е. расположены по правую сторону от R_0, а несколько точек группы B попали в группу A. Это – точки, неправильно расклассифицированные с помощью дискриминантной функции.

Критерии значимости

Если поставить некоторые условия для данных, используемых при построении дискриминантной функции, можно провести проверку значимости разделения на две группы. Основными условиями являются:

a) наблюдения в каждой группе проводятся наудачу;
b) вероятности того, что неизвестное наблюдение принадлежит любой из групп, равны между собой;
c) внутри каждой из групп переменные рассматриваются как случайные величины, распределенные нормально;
d) ковариационные матрицы различных групп имеют одинаковый порядок;
e) ни одно из наблюдений, используемых для построения дискриминантной функции, не было ложно расклассифицировано.

Наиболее трудно удовлетворимы условия «b – d». К счастью, дискриминантная функция изменяется незначительно при малых отклонениях от нормальности или при малых отклонениях дисперсий. Выполнение условии «b» зависит от априорно заданного уровня относительных вкладов исследуемых групп. Если условие о равенстве вкладов относительных содержаний не выполняется, можно сделать некоторые другие допущения, которые приводят к смещению значения R_0. (Подробное изложение вопросов принятия альтернативных решений в дискриминантном анализе содержится в книге Андерсона [2], глава 12.) 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, я обычно более удобна, если использовать ее в стандартизованной форме. Разделив разность на объединенное стандартное отклонение, мы получаем стандартизованную разность

$$d = \left(\overline{X}_1 - \overline{X}_2\right)/s_p \qquad (6.19)$$

Возведя обе части (6.19) в квадрат и обозначив знаменатель, являющийся объединенной дисперсией двух выборок, через s_p^2, получим

$$d^2 = \left(\overline{X}_1 - \overline{X}_2\right)^2/s_p^2 \quad (6.20)$$

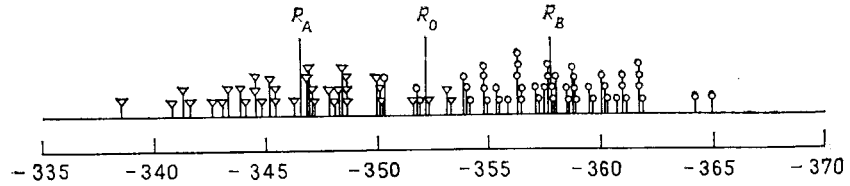


Рис. 6.4. Проекция выборок, представленных в табл. 6.5, на дискриминантную прямую, изображенную на рис. 6.3: R_A – проекция двумерного среднего для пляжного песка; R_B – проекция двумерного среднего для песка, удаленного от берега; R_0 – дискриминантный индекс

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

$$\text{евклидово расстояние} = \sqrt{\left(\overline{A}_1 - \overline{B}_1\right)^2 - \left(\overline{A}_2 - \overline{B}_2\right)^2} \quad (6.21)$$

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

$$\text{евклидово расстояние} = \sqrt{\sum_{i=1}^{m}\left(\overline{A}_i - \overline{B}_i\right)^2} \qquad (6.22)$$

Квадрат евклидова расстояния есть $\sum_{i=1}^{m}\left(\overline{A}_i - \overline{B}_i\right)^2$; легко проверить, что это то же, что и матричное произведение

$$\left(\text{евклидово расстояние}\right)^2 = \left[\overline{A}_i - \overline{B}_i\right]' \times \left[\overline{A}_i - \overline{B}_i\right] \quad (6.23)$$

Евклидово расстояние и его квадрат, к сожалению, выражаются в единицах, составленных из исходных единиц измерений. Для того чтобы иметь возможность их интерпретировать, их надо стандартизировать. Сравнение с формулой (6.19) позволяет предположить, что стандартизация должна содержать деление на многомерный эквивалент дисперсии, которым является ковариационная матрица $\left[s_p^2\right]$. Конечно, деление – операция, не определенная в матричной алгебре, но ей эквивалентно умножение на обратную матрицу. Умножая вектор-строку $\left[\overline{A}_i - \overline{B}_i\right]$ справа на матрицу, обратную

ковариационной матрице, т. е. $\left[s_p^2\right]^{-1}$ и затем на вектор-столбец $\left[\overline{A}_i - \overline{B}_i\right]''$, получаем стандартизованный квадрат расстояния

$$D^2 = \left[\overline{A}_i - \overline{B}_i\right]'' \times \left[s_p^2\right]^{-1} \times \left[\overline{A}_i - \overline{B}_i\right] \quad (6.24)$$

Эта мера расстояния между средними двух многомерных групп называется расстоянием Махалонобиса. Подставляя данные табл. 6.6 в формулу (6.24), получим

$$D^2 = \begin{bmatrix} -0{,}010 & -0{,}043 \end{bmatrix} \times \begin{bmatrix} 59112{,}280 & 4312{,}646 \\ 4312{,}646 & 747{,}132 \end{bmatrix} \times \begin{bmatrix} -0{,}010 \\ -0{,}043 \end{bmatrix} = 11{,}15$$

Интересно, что мы можем получить ту же меру расстояния (в пределах ошибок округления), подставляя вектор средних значений в уравнение самой дискриминантной функции:

$$D^2 = -783{,}63(0{,}010) - 75{,}62(-0{,}043) = 11{,}17$$

Расстояние Мехалонобиса графически представлено на рис. 6.4, 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 Хотелинга T^2. Этот критерий более подробно рассмотрен в следующем параграфе. Здесь просто отметим, что он имеет вид

$$T^2 = \frac{n_a n_b}{n_a + n_b} D^2 \qquad (6.25)$$

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

$$F = \left(\frac{n_a + n_b - m - 1}{(n_a + n_b - 2)m}\right)\left(\frac{n_a n_b}{n_a + n_b}\right) D^2 \quad (6.26)$$

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 $H_0 : [D_i] > 0$.

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. В качестве примера поучительно проверить значимость дискриминантной функции, которую мы только что строили.

Поскольку не все переменные, включенные в дискриминантную функцию, в равной степени полезны при отделении групп друг от друга, эти бесполезные переменные желательно найти и исключить из дальнейшего рассмотрения. Выбор наименее эффективного множества переменных дискриминантной функции аналогичен выбору наиболее эффективных зависимых переменных в уравнении множественной регрессии. Эта задача, однако, более сложная, так как зависимые или предсказываемые переменные в дискриминантной функции составлены из разностей между двумя множествами одинаковых переменных, которые используются как независимые переменные классификации. В отличие от регрессии, в которой суммы квадратов Y не изменяются при добавлении к уравнению различных переменных X_i, суммы квадратов разностей между группами A и B претерпевают изменения при добавлении или устранении переменных.

Некоторое представление об эффективности независимых переменных в дискриминантной функции можно получить, вычисляя стандартизованные разности

$$D_i = \left(\overline{A}_i - \overline{B}_i\right) / s_{pi} \qquad (6.27)$$

Это – просто разность между средними двух групп A и B для i-й переменной, деленная на объединенное стандартное отклонение переменной i. Так как эта мера не учитывает взаимосвязи переменных, то она полезна только как некоторый общий ориентир эффективности классификации. Пошаговые алгоритмы дискриминантного анализа используют стандартизованные разности в выборе порядка, в котором переменные добавляются к дискриминантной функции. Однако значимость раз-

личных комбинаций переменных может быть проверена только с помощью вычисления различных функций и определения относительных вкладов в процесс разделения на два класса, которые дают различные уравнения.

Дискриминантный анализ обеспечивает естественный переход между двумя большими классами многомерных статистических методов. С одной стороны, он тесно связан с множественной регрессией и тренд-анализом. С другой, он может трактоваться как задача определения собственных значений типа метода главных компонент, факторного анализа и других аналогичных многомерных методов. Имеются преимущества в нс-пользовании собственных векторов при вычислении дискриминантной функции, так как она позволяет одновременно осуществить разделение более чем на две группы. Однако отложим рассмотрение этого вопроса до момента, когда будут изложены элементарные сведения из теории собственных значений и собственных векторов.

Следующая задача – дать пример применения дискриминантного анализа в геологии – поможет лучше понять этот метод.

Правительственная разведочная партия проводила поиски месторождений тяжелых металлов в густо залесенных горах северной части Швеции. Данные, собранные аэромагнитометром, оказались недостаточными, и поэтому было проведено геохимическое исследование, основанное на анализах водных потоков. Было выбрано семь переменных и проведено две последовательности измерений. Группа *А* состоит из измерений, сделанных в потоках, дренирующих площади, на которых имеются действующие шахты и подтвержденные рудные тела. Группа *Б* состоит из аналогичных измерений на площадях, на которых оруденение не обнаружено. Данные по этим площадям приведены в табл. 6.7. Вычислите дискриминантную функцию для продуктивного и непродуктивного районов. Определите, являются ли различия между двумя группами значимыми, и исследуйте относительное влияние используемых переменных. Для удобства в этом примере предполагается, что изучаемые совокупности обеих групп подчиняются многомерному нормальному распределению. В табл. 6.7 приведен также ряд измерений (группа *B)*, сделанных на площадях, относительно которых не известно, разведывались ли они когда-нибудь. Используя дискриминантную функцию, ответьте на вопрос, можно ли рекомендовать какую-либо из этих площадей в качестве перспективной для разведки.

ПЕРЕХОД ОТ ОДНОМЕРНОЙ СТАТИСТИКИ К МНОГОМЕРНОЙ

В гл. 2 были рассмотрены некоторые простые геологические задачи, которые можно решать с помощью элементарных статистических методов.

Начнем изложение многомерных методов в геологии с рассмотрения прямых обобщений этих простых критериев. Напомним, что характеристики изменчивости, связанные с наиболее естественными явлениями природы, обычно описываются нормальным распределением. Это отражает так называемую центральную предельную теорему, которая гласит, что распределение случайной величины, являющейся суммой n независимых случайных величин при больших n, можно приближенно считать нормальным. Именно это свойство позволяет использовать нормальное распределение в качестве основы статистических критериев и считать его отправным пунктом при построении других распределений, таких, как *t*-, *F*- и χ^2 -распределения. Понятие нормального распределения можно распространить на ситуации, в которых наблюдения содержат много переменных.

Предположим, что мы отбираем образцы пород на некоторой площади и измеряем некоторый набор характеристик каждого из них. Измерения могут представлять собой значения химических и минералогических характеристик, удельного веса, характеристик магнитных и радиоактивных свойств, а также любые из бесчисленного множества возможных переменных. Набор измерений, сделанных на индивидуальном образце породы, можно записать в виде вектора $[X] = [X_1 \quad X_2 \quad \ldots \quad X_m]$, где m – число измеренных характеристик или переменных. Если множество измерений, представляющих вектор $[X]$, случайно извлечено из совокупности, которая возникла в результате воздействия многих независимых факторов, то наблюдаемые векторы приближенно можно считать имеющими многомерное нормальное распределение. Каждая переменная, рассматриваемая отдельно, имеет нормальное распределение со средним значением μ_k и дисперсией ν^2.

Таблица 6.7.

Результаты анализа проб руслового аллювия на редкие элементы, собранных в двух районах Швеции.

Элементы												
Ti[a]	Mn[b]	Ag[c]	Ba[a]	Co[d]	Cr[a]	Cu[c]	Ni[a]	Pb[a]	Sr[a]	V[a]	Zn[a]	Au[c]
Группа А – продуктивная площадь (пробы, взятые из аллювия в старом рудоносном районе)												
7280	1300	30,0	720	30	150	73	50	70	60	70	190	0,02
10300	1200	0,7	1 280	20	160	25	50	70	90	50	50	0,02
6500	700	1,0	1070	20	200	48	70	100	210	50	170	0,01
7000	1500	0,7	760	30	160	70	40	110	240	40	250	0,01
5100	1000	0,5	740	20	140	39	50	80	50	60	130	0,02
10600	2100	0,3	980	30	50	25	30	70	150	160	110	0,01
14200	2000	0,2	690	30	70	25	50	60	160	70	180	0,01
9700	900	0,2	680	35	70	38	30	70	80	110	250	0,01
2300	1500	0.2	710	5	110	50	20	70	80	30	120	0,01
12100	6300	0,1	1520	30	30	24	30	80	320	160	190	0,02
3000	1100	0,2	510	5	30	15	30	30	240	30	50	0,02
3000	1100	0,2	510	5	30	15	30	30	240	30	50	0,02
7500	2400	0,7	690	30	30	31	10	100	210	40	280	0,03
7800	1800	4,0	730	55	40	24	30	20	90	320	90	0,01
6900	1500	1,0	326	30	50	25	10	90	70	200	70	0,04
11200	3100	1,5	660	50	40	20	40	50	140	280	90	0,01
5200	1400	0,8	680	35	50	42	20	50	30	150	150	0,01
5100	1500	0,9	700	25	60	67	40	80	40	190	90	0,01
10600	2900	0,4	1640	25	20	21	30	30	320	90	200	0,01
11500	3200	0,7	710	30	30	15	20	20	260	270	180	0,01
7100	1800	0,9	490	75	50	8	10	30	80	180	100	0,02
Группа Б – непродуктивная площадь (пробы, взятые из аллювия, предполагаемого бесперспективного района)												
4820	500	0,1	160	20	70	30	10	0	720	140	200	0,01
3040	500	0,2	150	20	30	82	10	20	1 580	160	70	0,01
890	600	0,1	50	10	10	61	10	0	340	40	50	0,02
2100	500	0,1	100	15	30	77	10	0	650	90	80	0,02
5060	700	0,3	140	20	50	154	20	0	1240	140	80	0,01
1980	700	0,1	80	15	20	63	20	0	720	80	110	0,00
3200	600	0,2	160	20	30	45	20	10	1 100	120	60	0,01
3280	800	0,2	90	15	10	40	30	20	1480	70	40	0,00
2020	700	0,1	80	15	20	104	20	0	420	80	70	0,00
4600	700	0,3	160	20	60	48	10	20	780	150	50	0,02
3100	500	0,2	100	15	30	65	10	20	710	100	40	0,01
3020	600	0,2	90	15	10	69	0	30	1310	110	30	0,02
1860	500	0,1	70	10	20	63	0	10	480	80	50	0,00
2800	700	0,1	110	15	20	58	10	20	730	120	80	0,01
1040	1600	0,1	20	5	10	37	0	10	140	30	80	0,01
4640	800	0,3	220	15	20	121	20	20	1200	210	160	0,00
4990	900	0,3	190	20	40	59	20	30	480	230	120	0,02
2830	800	0,2	120	15	20	40	10	20	690	140	60	0,00
4500	700	0,2	140	20	30	82	20	10	710	170	70	0,00
2900	600	0,1	80	15	10	99	0	0	760	80	90	0,01
Группа В – непредсказанные площади (пробы, взятые на площадях, которые нужно классифицировать на перспективные и бесперспективные)												
4260	800	0,3	180	20	60	128	30	30	460	110	80	0,02
6500	1200	0,5	380	30	40	72	50	20	320	90	160	0,01
12200	5200	1,5	630	25	80	39	40	90	210	200	180	0,01
1080	1600	0,2	80	5	10	102	0	10	160	30	80	0,00
3820	500	0,2	170	25	40	60	20	10	1100	160	40	0,02
1020	2400	0,1	20	0	10	28	0	0	1320	20	60	0,00

Примечание: Буквы в индексах к символам элементов означают точность анализа (в г/т): a – 10; b – 100; c – < 1; d – 5.

Совместное вероятностное распределение является n-мерным эквивалентом нормального распределения, имеющего вектор среднего $[\mu]=[\mu_1 \quad \mu_2 \quad \ldots \quad \mu_m]$ и, если компоненты вектора $[X]$ независимы, обобщенную дисперсию, которая представлена в виде диагональной матрицы:

$$[\Sigma^2]=\begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ 0 & 0 & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_m^2 \end{bmatrix}$$

Кроме этих очевидных обобщенных характеристик одномерного нормального распределения, если компоненты вектора $[X]$ зависимы, для многомерного случая следует упомянуть коэффициенты ковариации cov_{jr}, которые занимают все недиагональные позиции в матрице $[\Sigma^2]$. Таким образом, многомерное нормальное распределение характеризуется вектором среднего и матрицей дисперсий и ковариации, являющейся многомерным аналогом дисперсии одномерной нормальной случайной величины. В простом случае при m=2 форма поверхности нормального распределения напоминает колокол (см. рис. 2.13), «карта в изолиниях» которого представлена на рис. 6.5. Хотя распределения переменных X_1 и X_2 изображены вдоль соответствующих осей, наиболее существенные черты совместного распределения лучше характеризуются большой и малой осями эллипса. Многие из многомерных задач, которые будут рассмотрены ниже, связаны с относительной ориентацией этих полуосей.

Одним из самых простых критериев, которые рассматривались в гл. 2, был t-критерий, используемый для проверки предположений, что случайную выборку из n наблюдений можно считать извлеченной из нормальной совокупности с некоторым средним μ_0 и дисперсией σ^2. Критерий, заданный формулой (2.32), может быть переписан в виде

$$t=\left[\left(\overline{X}-\mu_0\right)\sqrt{n}\right]/\sqrt{[s^2]} \qquad (6.28)$$

Очевидное обобщение этого критерия на многомерный случай состоит в замене X вектором выборочного среднего $[X]$, μ_0–вектором среднего совокупности $[\mu]$ и s^2 – матрицей дисперсий и ковариаций.

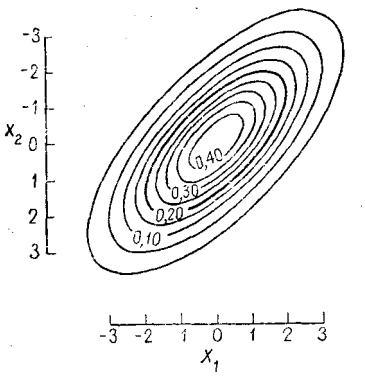


Рис. 6.5. Изолинии двумерного нормального распределения.
См. рис. 2.13 (в кн. 1), на котором представлена трехмерная диаграмма того же распределения

Обозначим вектор среднего совокупности через $[\mu]$, вектор выборочного среднего через $[X]$, матрицу дисперсий и ковариаций через $[\Sigma^2]$. Будем считать $[X]$ и $[\mu]$ векторами-столбцами, хотя можно было бы считать их и векторами-строками. Разность между вектором выборочного среднего и вектором среднего совокупности можно записать в виде $[\overline{X}-\mu]=[\overline{X}]-[\mu]$. Подставляя эти величины прямо в формулу (6.28), получаем

$$t=\left[\left(\overline{X}-\mu\right)\sqrt{n}\right]/\sqrt{[s^2]}$$

К сожалению, очевидных способов решения этого уравнения не существует, поэтому с его помощью получаем единственное значение t. Применяя этот критерий, и в числителе, и в знаменателевместо вектора и матрицы нужно иметь числа. Если столбец $[\overline{X}-\mu]$ умножить на вектор-строку, имеющую такое же число элементов, то в результате получится число. Определим произвольную вектор-строку $[A]$, в результате транспортирования которой получается вектор-столбец $[A]'$. Умножение вектора-строки $[A]$ на вектор-столбец разности $[\overline{X}-\mu]$ дает число, и умножение $[s^2]$ слева на $[A]$ и справа на $[A]$ снова дает число, т.е. наш критерий принимает вид

$$T = \frac{[A][\overline{X} - \mu]\sqrt{n}}{[A]\sqrt{s^2}[A]'}$$

Однако при этом мы изменили также запись проверяемой гипотезы. Нулевая гипотеза вместо предыдущей записи $H_0 : [\mu] = [\mu_0]$ теперь будет записываться так:

$$H_0^* : [A] \times [\mu_1] = [A][\mu_0]$$

Первоначальная гипотеза H_0 верна только в том случае, если новая гипотеза H_0^* верна для всех возможных значений $[A]$. Поэтому достаточно проверить только максимально возможное значение проверяемой статистики. Действительно, если гипотеза H_0^* отклоняется при любом $[A]$, то гипотеза H_0 тоже отклоняется. Проделав ряд несложных преобразований, можно определить условия, при которых достигается максимальное значение статистического критерия для произвольного вектора $[A]$. Это приводит к введению ограничения $[A][s^2][A]'=1$ и появлению определителей в основном уравнении. Далее, возводя обе части уравнения в квадрат, можно исключить причиняющие неудобство квадратные корни. Это приводит также к возведению в квадрат значения критерия, которое теперь обозначим T^2. Выполнив все преобразования, найдем, что статистический критерий можно представить в виде

$$T^2 = x[\overline{X} - \mu]'[s^2]^{-1}[\overline{X} - \mu] \quad (6.29)$$

т.е. произвольный вектор $[A]$ оказывается равным вектору разности между средними $[\overline{X} - \mu]$. Таким образом, для вычисления критерия находим матрицу, обратную ковариационной матрице, умножаем ее слева на вектор-строку $[\overline{X} - \mu]'$ и затем выполняем умножение справа на вектор-столбец той же разности. Полученная статистика является многомерным обобщением *t*-статистики и называется T^2-статистикой Хотеллинга в честь математика, который впервые ее использовал. Критические значения T^2 определяются с помощью соотношения

$$F = \frac{n-m}{m(n-1)} T^2 \qquad (6.30)$$

где n – число наблюдений, а m – число переменных. Эта формула позволяет для определения критических значений вместо специальных таблиц T^2-распределения использовать более удобные таблицы F-распределения. Более полное изложение вопросов, связанных с T^2-критерием, дается во многих руководствах, например в монографии Моррисона [51].

Хотя формула T^2-критерия (6.29) имеет очень простой вид, его вычисление для заданных числовых характеристик может оказаться очень трудоемким. Например, предположим, что измерены содержания четырех элементов в семи пробах лунного грунта. Требуется проверить гипотезу, что эти пробы извлечены из совокупности, имеющей то же среднее значение, как и континентальные базальта. Предположим, что взяты наши средние значения из «Справочника физических констант» [9]. Критерий Хотеллинга T^2 оказывается вполне пригодным для проверки гипотезы о том, что вектор средних значений лунных проб не отличается от вектора средних значений базальтов, приведенных в справочнике.

Сначала вычисляем вектор четырех выборочных средних и матрицу дисперсий и ковариаций порядка 4x4. Далее вычисляем вектор разностей между выборочными средними и средними значениями совокупности $[\overline{X} - \mu]$. Затем проводим обращение матрицы ковариаций и дисперсий, т.е. вычисляем $[s^2]^{-1}$. Наконец, после двух матричных умножений: $[\overline{X} - \mu]'[s^2]^{-1}[\overline{X} - \mu]$, умножение результата на n дает значение T^2. Из этого описания легко видеть, что объем вычислений быстро растет с увеличением числа переменных.

Данные по семи лунным пробам, так же как n вектор средних, приведены в табл. 6.8. В этой таблице даны также промежуточные результаты вычислений и окончательное значение T^2. Вычислите эквивалентное значение F-критерия с n и m степенями свободы и проверьте гипотезу, заключающуюся в том, что вектор средних значений лунных проб равен вектору средних значений совокупности континентальных базальтов.

Мы подробно остановились на рассмотрении T^2-критерия для проверки гипотезы о среднем не потому, что этот критерий используется в геологии чаще, чем другие многомерные критерии, а

чтобы показать тесную связь между обычными и многомерными статистическими методами. Для большинства одномерных критериев можно прямо указать соответствующие многомерные аналоги вместе с соответствующими основными допущениями. Однако при переходе от обычной алгебры к матричной связь между ними зачастую становится не столь явной. Так как многомерные статистические методы представляют собой обобщение одномерных, то последние являются частным случаем многомерных.

В оставшейся части этого раздела рассмотрим многомерные критерии, являющиеся m-мерными эквивалентами некоторых критериев, рассматривавшихся в гл. 2. Однако не будем указывать детали обобщенного перехода от одномерного варианта к многомерному, как это было сделано при построении T^2-критерия. Подобные рассуждения можно найти во многих руководствах по многомерному статистическому анализу. Некоторые из них указаны в списке литературы в конце этой главы.

Таблица 6.8.

Содержание четырех элементов в семи лунных пробах и средние значения содержаний тех же элементов в континентальных базальтах, % [64]

	Лунные пробы, %			
	Si	Al	Fe	Mg
1	19,4	5,9	14,7	5,0
2	21,5	4,0	15,7	3,7
3	19,2	4,0	15,4	4,3
4	18,4	5,4	15,2	3,4
5	20,6	6,2	13,2	5,5
6	19,8	5,7	14,8	2,8
7	18,7	6,0	13,8	4,6
Средние значения	19,66	5,31	14,69	4,19
Средние значения в базальтах	22,10	7,40	10,10	4,00
Разности	–2,44	–2,09	4,59	0,19
Ковариационная матрица				
$\begin{bmatrix} 1{,}1795350 & -0{,}3076159 & 0{,}0593007 & 0{,}0792885 \\ -0{,}3076159 & 0{,}8680964 & -0{,}6830902 & -0{,}3019053 \\ 0{,}0593007 & -0{,}6830902 & 0{,}8014425 & -0{,}5469023 \\ 0{,}0792885 & 0{,}3019053 & -0{,}5469023 & 0{,}8914289 \end{bmatrix}$				
Матрица, обратная ковариационной матрице				
$\begin{bmatrix} 1{,}0614130 & 0{,}9946304 & 0{,}8169364 & 0{,}0699353 \\ 0{,}9946304 & 5{,}2086160 & 5{,}3354270 & 1{,}4208490 \\ 0{,}8169364 & 5{,}3354270 & 7{,}6584370 & 2{,}8189000 \\ 0{,}0699353 & 1{,}4208490 & 2{,}8189000 & 2{,}3637960 \end{bmatrix}$				
$T^2 = 11{,}93$				

Проверка гипотез о равенстве двух векторов средних значении

Рассмотренный только что критерий, предназначен для проверки предположения о том, что вектор среднего значения, оцененный по одной выборке, равен заданному вектору. Рассмотрим теперь следующую задачу. Предположим, что мы располагаем двумя независимыми случайными выборками и хотим проверить гипотезу о равенстве истинных значений их векторов средних. Допустим, что две выборки взяты из многомерных нормальных совокупностей с одинаковыми неизвест-

ными ковариационными матрицами $\left[\sum^2\right]$, тогда нулевая гипотеза может быть записана в виде $H_0 : [\mu_1] = [\mu_2]$, а множество альтернатив – в виде $H_0 : [\mu_1] \neq [\mu_2]$.

Согласно гипотезе H_0 вектор средних значений совокупности, из которой взята первая выборка, такой же, как и аналогичный вектор второй совокупности.

Для проверки этой гипотезы используется многомерный эквивалент критерия (2.33). Ранее при вычислении t использовалась объединенная оценка дисперсии совокупности, построенная по двум выборкам. В многомерном же варианте нужно вычислить объединенную оценку $[s_p^2]$ ковариационной матрицы для двух многомерных выборок. Для этого вычисляется матрица сумм квадратов и смешанных произведений по каждой выборке. Будем использовать такую же терминологию, как и в гл. 2, и обозначим матрицу сумм квадратов и смешанных произведений первой выборки через $[SP_1]$, а аналогичную матрицу для второй выборки – через $[SP_2]$. Объединенная оценка ковариационной матрицы имеет следующий вид:

$$\left[s_p^2\right] = \frac{1}{n_1 + n_2 - 2}\left([SP_1] + [SP_2]\right) \qquad (6.31)$$

Таблица 6.8.

Содержание четырех элементов в семи пробах тихоокеанских базальтах, % [64]

	Тихоокеанские пробы			
	Si	Al	Fe	Mg
1	22,5	9,6	6,6	3,4
2	22,1	8,4	7,8	3,6
3	25,9	8,7	4,8	4,0
4	23,5	8,1	5,0	5,2
5	21,7	10,0	8,2	4,9
6	21,9	8,2	9,3	4,9
7	23,7	7,2	9,5	3,3
Средние значения	23,04	8,60	7,31	4,19
Разности между средними таблиц 6.8 и 6.9	–3,39	–3,29	7,37	0,00
Ковариационная матрица				
$\begin{bmatrix} 2{,}1828820 & -0{,}4399923 & -1{,}7223760 & -0{,}2409477 \\ -0{,}4399923 & 0{,}8966687 & -0{,}4199982 & 0{,}1266680 \\ -1{,}7223760 & -0{,}4199982 & 3{,}6547640 & -0{,}2480939 \\ -0{,}2409477 & 0{,}1266680 & -0{,}2480939 & 0{,}6380959 \end{bmatrix}$				
Объединенная ковариационная матрица для табл. 6.8 и 6.9				
$\begin{bmatrix} 1{,}6812080 & -0{,}3738041 & -0{,}8315379 & -0{,}0808296 \\ -0{,}3738041 & 0{,}8823825 & -0{,}5515442 & 0{,}2142866 \\ -0{,}8315379 & -0{,}5515442 & 2{,}2281030 & -0{,}3974981 \\ -0{,}0808296 & 0{,}2142866 & -0{,}3974981 & 0{,}7647624 \end{bmatrix}$				
Матрица, обратная объединенной ковариационной матрице				
$\begin{bmatrix} 1{,}1213670 & 0{,}8356105 & 0{,}6665248 & 0{,}2308193 \\ 0{,}8356105 & 1{,}9991410 & 0{,}7963902 & -0{,}0579050 \\ 0{,}6665248 & 0{,}7963902 & 0{,}9561165 & 0{,}3442555 \\ 0{,}2308193 & -0{,}0579050 & 0{,}3442555 & 1{,}5271490 \end{bmatrix}$				
$T^2 = 116{,}10$				

После того как найдена разность между двумя векторами средних $\left[\overline{X}_1\right]$ и $\left[\overline{X}_2\right]$ или $\left[\overline{X}_1 - \overline{X}_2\right]$, T^2–критерий примет следующий вид:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left[\overline{X}_1 - \overline{X}_2\right]' \times \left[s_p^2\right]^{-1} \times \left[\overline{X}_1 - \overline{X}_2\right] \qquad (6.32)$$

Для определения критического значения T^2-статистики можно воспользоваться F-распределением, т.е.

$$F = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m} T^2 \qquad (6.33)$$

с m и $(n_1 + n_2 - m - 1)$ степенями свободы [51].

В табл. 6.9 приведены данные семи анализов океанических базальтов Тихого океана на те же четыре элемента, что и в табл. 6.8. Предполагая, что обе выборки извлечены из многомерных совокупностей с нормальным распределением и с одинаковыми ковариационными матрицами, проверьте гипотезу о том, что векторы средних значений лунных и тихоокеанских образцов одинаковы. В таблице приведены необходимые вычисления. Является ли нулевая гипотеза о равенстве многомерных средних приемлемой при 5%-ном уровне значимости ($\alpha = 0{,}05$)?

Равенство ковариационных матриц

Основное допущение, на котором основаны два предыдущих критерия, заключается в том, что выборки считаются взятыми из совокупностей, имеющих одну и ту же ковариационную матрицу. Это многомерный аналог предположения о равенстве дисперсий, лежащем в основе применения t-критерия для проверки гипотезы о равенстве средних значений. На практике предположение о равенстве может нарушаться, так как выборки, имеющие высокое среднее значение, часто имеют также и большую дисперсию. Как указывалось в гл. 4, такое поведение характерно для многих геологических характеристик, таких, например как значения содержаний редких элементов. Равенство ковариационных матриц можно проверить с помощью «критерия обобщенных дисперсий», являющегося многомерным эквивалентом F-критерия [51].

Предположим, что имеется k групп наблюдений и в каждой из них m переменных. Для группы с номером i вычислим ковариационную матрицу $[S_i^2]$ и проверим нулевую гипотезу $H_0 : \left[\sum{}_1^2\right] = \left[\sum{}_2^2\right] = \ldots \left[\sum{}_k^2\right]$ при множестве альтернатив $H_1 : \left[\sum{}_i^2\right] \neq \left[\sum{}_j^2\right]$.

Согласно нулевой гипотезе все совокупности имеют одинаковые ковариационные матрицы. Альтернативы заключаются в том, что хотя бы две из них различны. Каждая матрица $[S_J^2]$ является оценкой соответствующей ковариационной матрицы $\left[\sum{}^2\right]$ изучаемой совокупности. Если исходные совокупности k выборок идентичны, то выборки можно объединить с целью получения обобщенной оценки ковариационной матрицы. Такая оценка вычисляется по формуле

$$\left[s_p^2\right] = \sum_{i=1}^{k} \frac{(n_i - 1)\left[s_i^2\right]}{\left(\sum_{i=1}^{k} n_i\right) - k} \qquad (6.34)$$

где n_i – объем i-й выборки, а $\sum n_i$ – общий объем k выборок.

Это определение алгебраически эквивалентно определений (6.31) при k=2, так как $\left[s_i^2\right] = [SP_i]/(n_i - 1)$.

Используя обобщенную оценку ковариационной матрицы вычислим статистику

$$M = \left[\left(\sum_{i=1}^{k} n_i\right) - k\right] \ln\left|s_p^2\right| - \sum_{i=1}^{k}\left[(n_i - 1)\ln\left|s_i^2\right|\right] \qquad (6.35)$$

В этом критерии за основу принята разность между логарифмом определителя обобщенной оценки ковариационной матрицы и средним значением логарифмов определителей выборочных ковариационных матриц. Если все выборочные матрицы одинаковы, то эта разность будет несущест-

венной. По мере. увеличения отклонений дисперсий и ковариаций выборок друг от друга эта статистика возрастает. Так как таблицы критических значений статистики M не очень доступны, то обычно используется ее аппроксимация χ^2-распределением с $\nu = \frac{1}{2}(k-1)m(m+1)$ степенями свободы:

$$\chi^2 = MC^{-1} \qquad (6.36)$$

где

$$C^{-1} = 1 - \frac{(2m^2+3m-1)(k+1)}{6(m+1)(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{\left(\sum_{i=1}^{k} n_i\right)-k}\right) \qquad (6.37)$$

Если выборки содержат одинаковое число наблюдений n, то формула (6.37) упрощается:

$$C^{-1} = 1 - \frac{(2m^2+3m-1)(k+1)}{6(m+1)k(n-1)} \qquad (6.38)$$

В тех случаях, когда k и m не превышают 5 и каждая оценка ковариационной матрицы строится на основании не менее 20 наблюдений, χ^2-аппроксимация наиболее точна. Используя этот критерий, проверьте гипотезы о равенстве ковариационных матриц по двум выборкам, приведенным в табл. 6.8 и 6.9. Для вычисления значения критерия необходимо подсчитать определители трех матриц четвертого порядка $[S_1^2]$, $[S_2^2]$ и $[Sp^2]$. Как только определители будут вычислены, получение значения критерия уже не вызовет затруднений. Для удобства предположим, что объемы выборок достаточно велики для того, чтобы χ^2-аппроксимация была точной.

С целью иллюстрации процедуры проверки гипотез с использованием многомерных статистик, решим следующую задачу. При этом заранее предположим, что число выборок допускает возможность применения некоторой аппроксимации.

В Восточном Канзасе на некоторой территории вся питьевая вода добывается из скважин. Одни из них пробурены в аллювиальных отложениях долин рек, другие пронизывают водоносный известняковый горизонт, являющийся источником многочисленных ключей в данном районе. Население предпочитает использовать воду из аллювиальных отложений, так как считает, что она вкуснее. Однако водные ресурсы аллювия ограничены, и иногда желательно было бы пополнить их за счет известнякового водоносного горизонта.

Для того чтобы доказать, что качество воды из этих двух источников одинаково, были взяты пробы в разных скважинах. Пробы анализировались на содержание тех химических компонентов, которые определяют качество воды. Некоторые из этих анализов приведены в табл. 6.10. В ней также представлены ковариационные матрицы, их обратные матрицы и определители для двух заданных совокупностей и для объединенной выборки. По этим данным можно проверить гипотезу о равенстве векторов средних значений. Предполагается, что выборки взяты случайно из многомерных совокупностей с нормальным распределением.

Используя статистику M, определенную по формуле (6.35), проверим сначала гипотезу о равенстве ковариационных матриц двух совокупностей:

$$M = (20+20-2)\ln 2{,}0725\cdot 10^8 - (19\ln 1{,}8838\cdot 10^8 + 19\ln 2{,}2582\cdot 10^8) = 0{,}1835$$

Для того чтобы использовать χ^2-аппроксимацию, вычислим множитель

$$C^{-1} = 1 - \frac{2\cdot 5^2+3\cdot 5-1}{6(5+1)(2-1)}\left(\frac{1}{19}+\frac{1}{19}-\frac{1}{40-2}\right) = 0{,}861$$

Значение χ^2-статистики равно 0,158, а число степеней свободы

$$\nu = \frac{1}{2}(2-1)(5)(5+1) = 15$$

Таблица 6.10. Катионный состав проб воды, взятых из скважин в Восточном Канзасе, % (X_1 – кремнезем, X_2 – железо, X_3 – магний, X_4 – натрий + кальций, X_5 – кальций)

X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
A. Пробы из аллювия					*B*. Пробы из известкового водоносного горизонта				
15,9	20,0	27,8	41,8	26,1	13,2	17,5	23,4	38,5	36,3
16,5	14,1	38,3	20,7	21,9	13,9	11,2	34,8	18,0	30,4
9,6	15,6	10,0	32,3	44,5	7,1	13,5	7,8	27,0	52,7
11,5	15,2	34,8	21,4	29,4	9,6	13,2	31,3	17,3	36,5
12,0	21,3	38,9	34,6	7,3	10,6	19,0	35,8	32,5	18,3
13,1	26,1	59,0	23,9	20,6	10,2	23,4	56,8	19,6	27,9
12,7	15,9	38,8	21,3	29,3	10,0	14,7	35,1	18,1	37,7
15,8	15,9	52,7	13,0	2,2	13,1	13,1	47,3	10,1	4,6
11,6	19,8	30,8	17,9	24,4	9,9	18,4	26,6	14,1	34,9
11,2	17,2	22,3	4!,0	28,0	9,8	14,8	17,2	37,4	34,4
8,2	19,5	31,6	26,6	33,6	6,0	16,6	26,1	21,1	40,8
10,8	16,8	39,2	20,9	7,7	8,2	14,6	34,9	17,7	15,8
10,8	8,2	29,1	27,1	33,1	8,3	6,6	27,1	21,6	43,5
15,5	15,5	35,3	29,6	62,6	13,4	13,9	33,3	24,9	70,4
10,2	6,1	32,6	35,0	60,3	8,0	4,6	30,0	29,4	70,5
9,6	18,1	30,6	19,2	11,6	7,4	16,1	28,2	17,2	17,5
12,2	15,1	32,3	42,6	7,4	9,4	13,6	28,3	37,4	16,6
8,9	18,0	40,3	42,9	25,3	6,8	15,6	36,6	37,8	30,7
123,0	6,4	37,3	35,5	18,3	10,3	4,7	33.9	33,2	27,7
12,7	16,8	27,6	50,9	7,5	10,0	14,0	24,2	45,7	18,5
Вектор средних значении для группы *A*					Вектор средних значений для группы *B*				
12,055	16,08	34,465	29,91	24,835	9,76	13,955	30,935	25,93	33,27

Ковариационная матрица A с определителем, равным 25822×10^3

$$\begin{bmatrix} 5{,}6394620 & 0{,}7332828 & 8{,}6868000 & -2{,}9821260 & -5{,}5767340 \\ 0{,}7332828 & 23{,}1732900 & 12{,}7656400 & -4{,}5592230 & -26{,}9460700 \\ 8{,}6868000 & 12{,}7656400 & 103{,}3983000 & -42{,}3948000 & -62{,}3459900 \\ -2{,}9821260 & -4{,}5592230 & -42{,}3948000 & 106{,}9526000 & 13{,}1360100 \\ -5{,}5767340 & -26{,}946070 & -62{,}3459900 & 13{,}1360100 & 286{,}8151000 \end{bmatrix}$$

Ковариационная матрица B с определителем, равным $1{,}88381\times10^3$

$$\begin{bmatrix} 5{,}1614800 & 0{,}5133812 & 7{,}3683410 & -1{,}4102910 & -3{,}4401890 \\ 0{,}5133812 & 21{,}0247300 & 10{,}6948600 & -4{,}0895870 & -25{,}3971700 \\ 7{,}3683410 & 10{,}6948600 & 102{,}8046000 & -38{,}5268000 & -58{,}1688200 \\ -1{,}4102910 & -4{,}0895870 & -38{,}5268000 & 98{,}8654500 & 7{,}2520690 \\ -3{,}4401890 & -25{,}3971700 & -58{,}1688200 & 7{,}2520690 & 290{,}8707000 \end{bmatrix}$$

Ковариационная матрица объединенной выборки с определителем, равным $2{,}07252\times10^3$

$$\begin{bmatrix} 5{,}4004710 & 0{,}6233320 & 8{,}0275700 & -2{,}1962090 & -4{,}5084610 \\ 0{,}6233320 & 22{,}0990100 & 11{,}7302500 & -4{,}3244050 & -26{,}1716200 \\ 8{,}0275700 & 11{,}7302500 & 103{,}1014000 & -40{,}4608000 & -60{,}2574000 \\ -2{,}1962090 & -4{,}3244050 & -40{,}4608000 & 102{,}9090000 & 10{,}1940400 \\ -4{,}5084610 & -26{,}1716200 & -60{,}2574000 & 10{,}19404000 & 288{,}8429000 \end{bmatrix}$$

Матрица, обратная ковариационной матрице объединенной выборки

$$\begin{bmatrix} 0{,}2100476 & 0{,}0029156 & -0{,}0176276 & -0{,}0023202 & 0{,}0000577 \\ 0{,}0029156 & 0{,}0518892 & -0{,}0036462 & 0{,}0004157 & 0{,}0039718 \\ -0{,}0176276 & -0{,}0036462 & 0{,}0148257 & 0{,}0050709 & 0{,}0023084 \\ -0{,}0023202 & 0{,}0004157 & 0{,}0050709 & 0{,}0115147 & 0{,}0006494 \\ -0{,}0000577 & 0{,}0039718 & 0{,}0023084 & 0{,}0006494 & 0{,}0042798 \end{bmatrix}$$

Критическое значение χ^2 при $\nu = 15$ и 5%-ном уровне значимости равно 25. Вычисленное значение намного ниже этой величины и потому мы можем заключить, что ничто не мешает нам принять гипотезу о равенстве ковариационных матриц изучаемых совокупностей. Далее, используя T^2-критерий (6.32), можно проверить гипотезу о равенстве многомерных средних:

$$T^2 = \frac{20 \times 20}{20 + 20}[1{,}566] = 15{,}66$$

В скобках указан результат, полученный после выполнения соответствующих матричных умножений, указанных в формуле (6.32). Используя формулу (6.33), вычислим соответствующее значение F-статистики:

$$F = \frac{20 + 20 - 5 - 1}{(20 + 20 - 2)^5} 15{,}66 = 2{,}80$$

Числа степеней свободы таковы: $\nu_1 = 5$; $\nu_2 = 20 + 20 - 5 - 1 = 34$. Критическое значение F-критерия при 5 и 34 степенях свободы .и 5%-ном уровне значимости 2,49. Так как вычисленная статистика превосходит это критическое значение, то можно заключить, что существует различие между векторами средних .двух изучаемых совокупностей. Иными словами, имеется статистически значимое различие в среднем составе воды из двух водоносных горизонтов. С помощью этого простого критерия не удалось выделить те переменные, которые обусловливают различие в химическом составе воды, но подтверждено мнение населения о том, что вода из этих двух источников различается по своим свойствам.

Можно также использовать многомерные методы, являющиеся обобщением процедур дисперсионного анализа, изложенных в гл. 2 (см. кн.1). Все они основаны на сравнении двух матриц порядка $m \times m$, являющихся многомерными эквивалентами внутригрупповых и межгрупповых сумм квадратов, используемых в обычном дисперсионном анализе. Проверяемая статистика основана на наибольшем собственном значении матрицы, используемой для сравнения. Мы не будем рассматривать эти критерии здесь, так как их формулировка сложна, а приложение к геологическим задачам весьма ограниченно. Из этого, однако, не следует делать вывод, что потенциальные возможности этих методов уже исчерпаны. Заинтересованные читатели могут обратиться к книге Кули и Лонеса [11], в которой имеются вычислительные программы многомерного дисперсионного анализа и примеры его использования, а также к книгам Моррисона [51], Коха и Линка [36], в которых содержится краткое изложение процедуры исследования геохимических данных с помощью многомерного дисперсионного анализа.

КЛАСТЕРНЫЙ АНАЛИЗ

Классификация – разделение объектов на более или менее однородные группы и установление соотношений между группами – важная особенность работы таксономистов, занимающихся определением происхождения живых организмов на основании их характеристик и сходства. Таксономия – в высшей степени субъективная наука, в которой выводы определяются интуицией ученого, выработанной годами опыта. В этом отношении таксономия очень сходна с многими разделами геологии. Ряд ученых, в том числе геологи, неудовлетворенные субъективностью и капризностью традиционных методов, разработали новые способы классификации, которые находятся в соответствии с возможностями современных вычислительных машин. Эта группа исследователей называет себя численными таксономистами, и им мы обязаны многими достижениями в численных методах классификации.

В настоящее время численная таксономия – предмет ожесточенных споров среди биологов, очень напоминающих острые дебаты психологов вокруг вопросов факторного анализа, имевших место в 30–40-х годах XX в. В этих обсуждениях некоторые практики рьяно отстаивают методы численной таксономии, заявляя, что они позволяют понять происхождение групп организмов лучше, чем любой другой метод классификации. Конечно, доказательств они представить не могут, так как в настоящее время теоретическое обоснование анализа групп недостаточно удовлетворительно, плохо исследованы статистические основы методов численной таксономии, нет соответствующих критериев значимости. По-видимому, здесь дело обстоит так же, как и в случае факторного анализа.

Однако уже многие методы численной таксономии нашли применение в геологических исследованиях, в особенности при классификации ископаемых беспозвоночных и при изучении палеобстановок.

Кластерный анализ предназначен для классификации наблюдений в более или менее однородные группы. Аналитического решения этой задачи, наподобие численной таксономии, которая была бы общей для всех областей классификации, не существует.

Хотя имеются альтернативные классификации классификаций [58], большинство из них может быть сгруппировано в четыре общих типа.

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. Хотя анализ делается в m-мерном пространстве, определенном этими переменными, а не в n-мерном пространстве, определенном наблюдениями, его итерационная реализация может оказаться весьма времяемкой [59].
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 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 расширенного 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 внутренние корреляции, близкие $k+1$, 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 по строения дендрограммы.

Иерархические методы наиболее широко применяются в науках о Земле, вероятно, потому, что развитие этих методов было тесно связано с численной таксономией ископаемых остатков, которые в силу их широкого распределения будут рассмотрены подробнее.

Предположим, что мы располагаем некоторым множеством объектов, которые желательно иерархически расклассифицировать. В биологии эти объекты обычно называются «операционными таксономическими единицами», или **OTE**. 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

$$d_{ij} = \sqrt{\frac{\left(\sum_{k=1}^{m} (X_{ik} - X_{jk})^2 \right)}{m}} \qquad (6.39)$$

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. Это позволяет учитывать каждую переменную с одинаковым весом. В противном случае расстояние определялось бы переменной, имеющей наибольшее значение. В некоторых случаях это даже желательно, однако неразумный выбор единиц измерения может иногда привести к нежелательным последствиям. Яркой иллюстрацией этой зависимости служит пример измерения трех осей образцов гальки. Если измерить две оси в сантиметрах, а третью – в миллиметрах, то третья ось будет иметь в десять раз большее влияние на расстояние, чем две другие переменные.

Множество мер сходства между всеми парами объектов можно представить в виде симметричной матрицы порядка $n \times n$. Для вычисления элементов этой матрицы с использованием уже написанных подпрограмм требуется 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 де будет получена полная классификация объектов. Существует много методов анализа групп; рассмотрение всех разновидностей этих методов и их сравнение выходят за рамки настоящей книги. Однако мы рассмотрим один простой метод, называемый методом взвешенной парной группировки с арифметическими средними, а затем укажем некоторые полезные разновидности этой схемы. Подробное изложение этого и других методов можко найти в книгах Трайопа и Бейли [64], а также Снита и Сокала [58]. В первой из них вопросы классификации излагаются с точки зрения экспериментальной физиологии, во второй – численной таксономии.

Таблица 6.11. Матрица коэффициентов корреляции для шести образцов песчаников. Измерения проводились в шлифах

	A	*B*	*C*	*D*	*E*	*F*
A	1,00	0,57	0,29	–0,59	–0,59	–0.59
B	0,57	1,00	0,29	-0,59	–0,59	–0,59
C	0,29	0,29	1,00	-0,59	–0,59	–0,59
D	–0,59	–0,59	–0,59	1,00	0,66	0,41
E	–0,59	–0,59	–0,59	0,66	1,00	0,41
F	–0,59	–0,59	–0,59	0,41	0,41	1,00

В табл. 6.11 приведена полная симметричная матрица коэффициентов корреляции между шестью объектами, названными *A, B, ..., F*. Объекты – это шлифы песчаника, а переменные – характеристики структуры породы, включающие размеры и показатели формы зерен, размеров и формы пор и плотности заполнения. В этом примере в качестве меры сходства взят коэффициент корреляции.

Первый шаг анализа групп методом попарного объединения состоит в нахождении в корреляционной матрице наибольших коэффициентов корреляции с целью выделения центров групп. Наивысшие коэффициенты корреляции в каждом столбце матрицы (см. табл. 6.11), выделены жирным шрифтом. Объекты *A* и *B* образуют пару с высокой мерой сходства, так как *A* наиболее близок к *B* и *B* наиболее близок к *A*. Однако *C* и *B* не образуют пары с высокой мерой сходства, так как хотя *C* близок к *B*, *B* ближе к *A*, чем к *C*. Для выделения пары с высокой мерой сходства коэффициенты c_{ij} и c_{ji} должны иметь наибольшие значения в соответствующих столбцах.

Пары с наивысшими мерами сходства изображены на рис. 6.6,*а*. Объект *A* связан с *B* на уровне 0,57, указывающем меру их взаимного сходства. Таким образом связаны *D* и *E*. Это первый шаг в построении дендрограммы, или «дерева», позволяющего наглядно изобразить результаты разбиения на группы.

Далее матрицу сходства вычисляют снова, причем сгруппированные элементы при этом считаются одним элементом. Существует несколько методов выполнения этой процедуры. Мы будем использовать 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. 6.12 приведены результаты этих вычислений. Наиболее высокие значения коэффициентов корреляции в каждом столбце выделены жирным шрифтом.

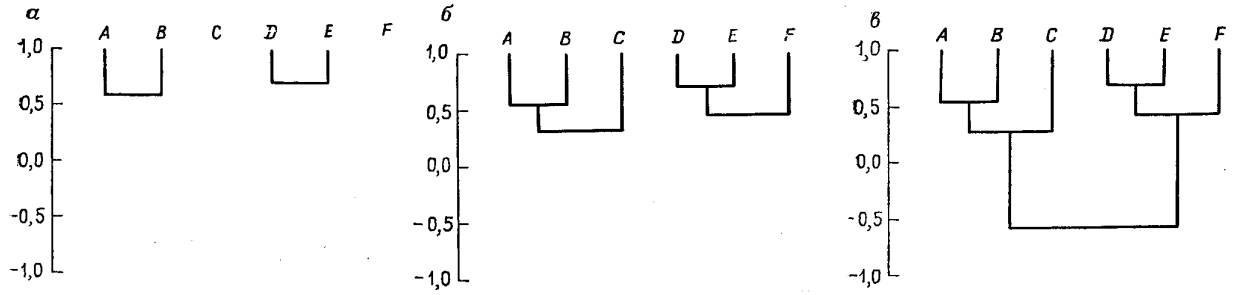


Рис. 6.6. (*а*) Исходные группы дендрограммы. (*б*) Построение групп для остальных объектов. (*в*) Окончание построения дендрограммы; две группы связываются между собой

Процедура образования групп снова повторяется: находим пары с сильными связями и объединяем. На этом этапе объект *C* присоединяется к группе *AB*, а объект *F* присоединяется к группе *DE* (рис. 6.6,*б*). Процесс продолжается до тех пор, пока все группы не объединятся вместе. Окончательная матрица сходства, как показано в табл. 6.13, 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. 6.6,*в*).

Таблица 6.12. Матрица коэффициентов корреляции между двумя усредненными группами и двумя образцами песчаника

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

Таблица 6.13. Матрица усредненных коэффициентов корреляции между двумя окончательными группами

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

Построение групп является эффективным способом представления сложных соотношений между объектами. Однако процесс усреднения по элементам группы и их трактовка в качестве единственного нового объекта приводит к изменениям дендрограммы. Это изменение становится все более очевидным по мере роста уровня усредняемых и объединяемых групп. Можно оценить степень этого изменения, исследуя матрицу, которая в таксономии носит название матрицы кофенетических значений. Это не что иное, как матрица коэффициентов корреляции дендрограммы. Например, коэффициенты корреляции между группами *D, E* и *F*, с одной стороны, и *A, B, C* – с другой, в дендрограмме на рис. 6.6 равны 0,59. Аналогично коэффициент корреляции между F и Z), а также между *F* и *E* равен 0,41. Наиболее сильные связи отмечаются только между парами *A* и *B*, а также *D* и *E*. В табл. 6.14 приведена полная матрица кофенетических значений, соответствующих дендрограмме. Можно получить наглядное представление о степени изменения в дендрограмме, сопоставив на графике каждый элемент исходной корреляционной матрицы с каждым элементом кофенетической матрицы (рис. 6.7). Если обе матрицы совпадут, то график этой зависимости будет представлен прямой линией. Отклонения от нее указывают на изменения в дендрограмме: если точка оказывается выше прямой, то корреляция, соответствующая дендрограмме, оказывается слишком высокой.

Наоборот, если точка попадет в область под прямой, то усреднение коэффициентов корреляции приводит к более низкому значению корреляции по сравнению с истинным. Численную меру сходства между двумя матрицами можно найти в результате простого вычисления коэффициентов корреляции между одинаково расположенными элементами. Так как обе матрицы симметричны относительно диагонали, то для этой цели достаточно использовать только одну половину элементов матрицы либо выше, либо ниже диагонали. В нашем случае коэффициент корреляции равен 0,98.

Таблица 6.14. Матрица кофенетических коэффициентов корреляции, полученных из дендрограммы рис. 6.6

A	1,00	0,57	0,12	–0,65	–0,62	–0,39
B	0,57	1,00	0,46	–0,79	–0,72	–0,72
C	0,12	0,46	1,00	–0,58	–0,61	–0,52
D	–0,65	–0,79	–0,58	1,00	0,66	0,41
E	–0,62	–0,72	–0,61	0,66	1,00	0,40
F	–0,39	–0,72	–0,52	0,41	0,40	1,00

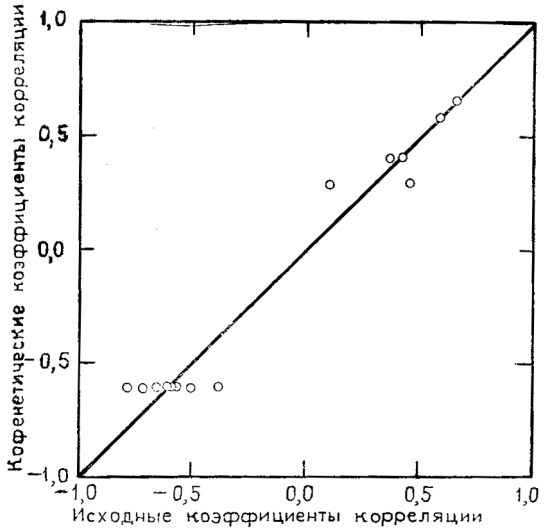


Рис. 6.7. Графическое представление зависимости кофенетических коэффициентов корреляции для дендрограммы, представленной на рис. 6.6, от эквивалентных им исходных коэффициентов корреляции, значения которых приведены в табл. 6.11. Если дендрограмма точно характеризует структуру корреляционной матрицы, то все точки попадают на диагональную линию. Отклонения от этой линии представляют неточности дендрограммы

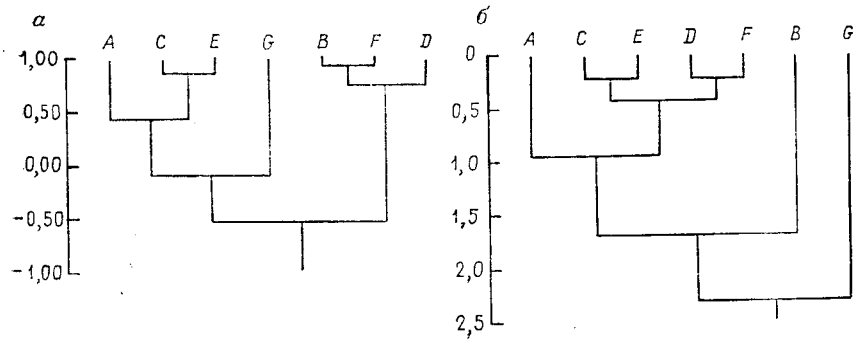


Рис. 6.8. (*а*) Дендрограмма, построенная по методу группового объединения, основанного на усреднении коэффициентов корреляции. Исходная матрица приведена в табл. 6.15. Кофенетическии коэффициент корреляции равен 0,77. (*б*) Дендрограмма, построенная тем же методом, но основанная на расстоянии. Кофенетический коэффициент корреляции равен 0,91

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. 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. 6.15 приведены как расстояния, так и коэффициенты корреляции для семи объектов, в данном случае для образцов карбонатных минералов. В качестве переменных выбраны некоторые физические характеристики.

Таблица 6.15. 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	

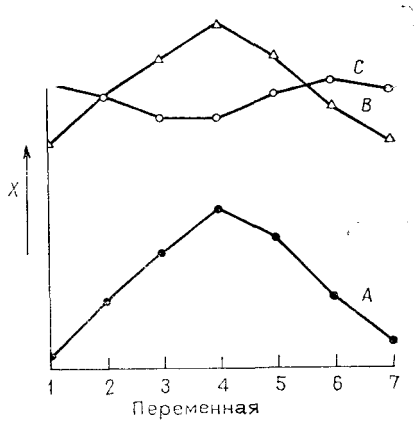


Рис. 6.9. Графики переменных, измеренных на трех объектах. Кривые А и В сильно коррелирова-ны, но разделены большим расстоянием. Кривые В и С отрицательно коррелированы, но «близки» в смысле расстояния

Дендрограммы, построенные для каждой матрицы сходства» изображены на рис. 6.8. 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 В в более дальнюю позицию в иерархической структуре. Полезно исследовать причины этого изменения.

Предположим, что измерено семь переменных на каждом из трех объектов. Ими могут быть, например, размеры трех ископаемых организмов или химические анализы трех пород. Если нанести каждое измерение на график так, как это указано на рис. 6.9, 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 размеры ископаемых организмов, например раковин брахиопод, 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 руды, 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. 6.15, изображены на рис. 6.10. 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 дендрограм-мы, возникают и другие отличия. Например, 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 нас 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. 6.8,*а* и *б* и *в* предыдущем примере (см. рис 6.6), называются взвешиванием, хотя на самом деле их следовало бы назвать методами равного взвешивания. На рис. 6.8,*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. 6.11 по данным табл. 6.15 приведена дендрограмма, построенная на основе метода невзвешенного усреднения.

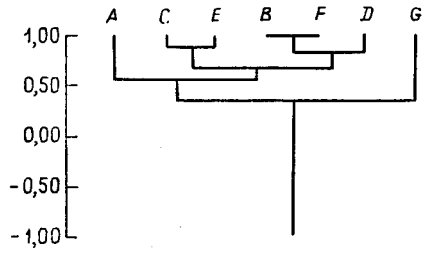


Рис. 6.10. Дендрограмма корреляционной матрицы, приведенной в табл. 6.15. Группы построены по методу прямой связи. Кофенетический коэффициент корреляции равен 0,71

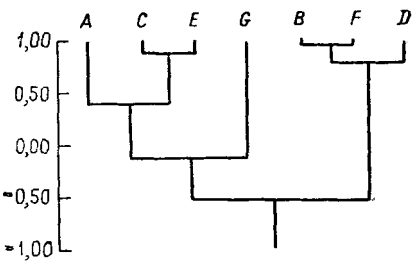


Рис. 6.11. Дендрограмма корреляционной матрицы, приведенной в табл. 6.15. Группы построены на основании невзвешенного усреднения. Кофенетический коэффициент коррелляции равен 0,72

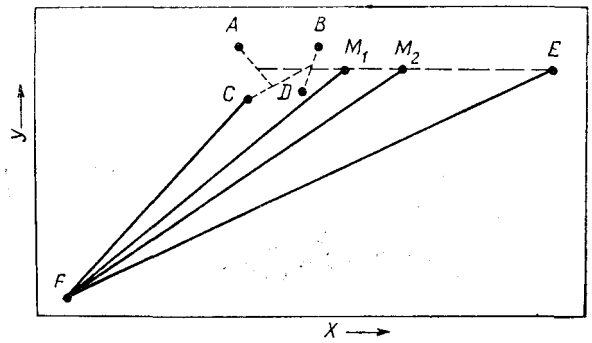


Рис. 6.12. Диаграмма, которая показывает, как объекты, характеризуемые двумя переменными *X* и *Y*, входят в группу. Объекты *A, B, C* и *D* образуют группу. Объект *E* присоединен к этой группе, а объект *F* является кандидатом на присоединение на следующем шаге итерационного процесса. M_1 – центроид объектов от *A* до *E*. M_2 – среднее объекта *E* и последнего среднего объектов от *A* до *D*

Мы можем 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. 6.12. Расстояния между объектами на диаграмме попросту пропорциональны мере расхождения между ними. 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 M_1F 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 невзвешенного усреднения. Относительное превосходство первых определяется тенденцией к получению наибольшего значения кофенетического коэффициента корргля-ции, который трактуется как индикатор малых изменений в дендрограмме. Значения кофенетических коэффициентов корреляции, меньшие 0,8, могут указывать на столь сильные изменения в дендрограмме для слабых связей, что она оказывается ошибочной. В анализе групп матрицы расстояний обычно используются с большим успехом, чем матрицы коэффициентов корреляции, так как дают более высокую кофенетическую корреляцию. По-видимому, матрицы расстояний также менее чувствительны к замене метода при анализе групп. Однако недостаток состоит в том, что они ограничивают использование каких-либо статистических методов. (Для других методов анализа групп имеются некоторые теоретические обоснования; см., например, [59].) Большинство исследователей, использующих методы анализа групп, применяют различные меры сходства и процедуры построения групп, а затем выбирают те из них, которые дают наиболее удовлетворительные результаты для их данных. Тщательный предварительный анализ может определить выбор процедуры кластеризации. Большинство иерархических методов, если число объектов велико, нуждается в вычислении и обработке очень больших матриц. (В экологии и археологии исследование тысяч, объектов является обычным делом.) Процедуры кластеризации, использующие ограниченное число произвольных центров групп, обычно сопровождаются приемами устранения этой вычислительной помехи. Вероятно, что наиболее широко применяемый метод – это процедура *k*-средних МакКвина [50]. 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. Преимущество процедуры k-средних состоит в том, что для нее необходима лишь матрица сходства порядка $k \times m$, а не матрица порядка $m \times m$. Если k мало (5 или 10), а n велико (1000 или больше), то процесс можно осуществить быстрее, чем иерархическим методом, получив меньше чем на 2 порядка число шагов. 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.

Многие методы кластеризации содержат субъективные процедуры, однако кофенетическая корреляция служит компасом в достижении объективной классификации. 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.

В качестве упражнения в кластерном анализе исследуем набор данных, представляющих измерения кембрийских трилобитов, собранных на западе США. В соответствии с требованиями таксономических процедур образцы были разделены на три рода. На десяти трилобитах, каждый из которых представлял определенный вид, были измерены 10 характеристик или переменных. Результаты этих измерений приведены в табл. 6.16. Вообще говоря, было установлено, что разные виды трилобитов плохо связаны между собой. Чтобы избежать недоразумений, проистекающих от того, что хвостовая часть больших индивидуумов может случайно ассоциироваться с передней частью малых, все измерения были преобразованы в отношения. Измерения, сделанные на осевой части головного щита, были разделены на его длину. Аналогично измерения, сделанные на хвостовой части щита, были разделены на его ширину. Части скелета, выбранные в качестве переменных, указаны на рис. 6.13. После подходящей стандартизации выполните анализ групп по данным измерений трилобитов и посмотрите, дают ли количественные методы ту же классификацию, которая получается методами обычной таксономии. Вычислите и используйте в качестве мер сходства коэффициенты корреляции и расстояния. Какая из этих мер дает лучший результат по сравнению с результатами, полученными методами обычной таксономии?

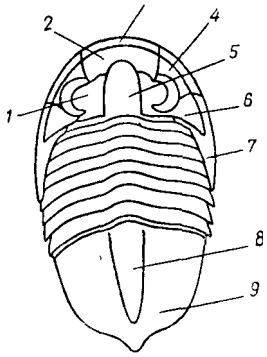


Рис. 6.13. Трилобит *Opisthoparian***.** Показана схема строения и измеряемые характеристики, приведенные в табл. 6.16. *1* – пальпебральная лопасть (глазная крышка); *2* – край; *3* – краевой валик; *4* – свободная щека; *5* – глабель; *6* – неподвижная щека; *7* – главный шип; *8* – ось пигидия; *9* – плевральная часть

Таблица 6.16. Десять отношений, полученных по результатам измерения десяти видов кембрийских трилобитов, собранных в штате Юта*

Виды	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Aphelaspis brachyphasis	0,208	0,250	0,540	0,237	0,875	0,292	0,284	0,925	0,343	0,373
A. haguei	0,318	0,318	0,545	0,428	1,000	0,318	0,296	0,796	0,444	0,537
A. subditus	0,174	0,304	0,391	0,375	0,913	0,304	0,297	0,946	0,401	0,486
Dicanfhopyge convergens	0,259	0,370	0,370	0,859	0,852	0,333	0,500	0,591	0,591	0,818
D. quadrata	0,250	0,350	0,500	0,615	0,900	0,351	0,434	0,783	0,478	0,652
D. reductus	0,316	0,421	0,474	0,736	1,158	0,421	0,500	0,675	0,500	0,775
Prehousia alata	0,136	0,409	0,273	0,469	1,000	0,136	0,269	0,769	0,327	0,423
P. indenta	0,192	0,308	0,269	0,628	0,923	0,154	0,308	0,795	0,308	0,436
P. prima	0,261	0,261	0,261	0,545	0,956	0,261	0,296	0,833	0,333	0,407
A. longispina	0,259	0,370	0,556	0,444	0,852	0,296	0,372	0,824	0,431	0,706

* C – длина глабели; X_1 – длина краевого валика/C; X_2 – длина края/C; X_3 – длина глазной крышки/C; X_4 – ширина глабели/C; X_5 – ширина неподвижной щеки/C; X_6 – длина главного шипа/длина свободной щеки; O – ширина пигидия; X_7 – ширина оси пигидия/O; X_8 – ширина плевральной оси/O; X_9 – длина оси пигидия/O; X_{10} – длина пигидия/O.

ВВЕДЕНИЕ В ТЕОРИЮ СОБСТВЕННЫХ ВЕКТОРОВ И ФАКТОРНЫЙ АНАЛИЗ

Некоторое множество вычислительных процедур часто небрежно называют «факторным анализом»; они обладают общими чертами, а именно, заранее предполагается, что в наборе многомерных наблюдений имеется скрытая простая структура. Эта структура выражается через дисперсии и ковариации переменных, а также с помощью мер сходства между наблюдениями. Все методы факторного анализа основаны на выделении собственных значений и собственных векторов из квадратной матрицы, получаемой умножением матрицы данных (или некоторым образом преобразованной матрицы данных) на ее транспозицию. Основные математические операции в точности те же, какие были указаны в гл. 3. На эти основные положения и методологию накладывается множество усовершенствований и изменений, часто произвольных, что в результате ведет к многообразию вычислительных методов, приводящему в уныние. Зачастую основные сходные черты этих методов сильно завуалированы сложными математическими обозначениями и терминологией, используемой различными практиками.

Факторный анализ первоначально развивался психологами в 1930-е годы, и многие используемые в нем термины имеют смысл лишь в пределах этой специфической области. Действительно, само название «фактор» относится к гипотетическим умственным способностям, которые можно еще характеризовать как «фактор интеллекта». Социологи и специалисты по биометрии также внесли свой вклад в богатство терминологии факторного анализа и помогли создать противоречивую и мало понятную методологию, которая увеличивает обманчивое впечатление о возможности получения мгновенного ответа у исследователя, поставленного перед числом данных, большим, чем то, которое можно осмыслить.

Создано множество методологических вариантов факторного анализа. Проанализируем некоторые наиболее распространенные из них, без глубокого изучения философских и математических аспектов, которые им сопутствуют. Мы используем некоторые математические соотношения, существующие между матрицей данных, соответствующей ей матрицей парных произведений, а также их собственных значений и собственных векторов.

Методы факторного анализа делятся на два больших класса, называемых *R*- и *Q*-факторным анализом. Первый связан с исследованием соотношений между переменными и основан на выделении собственных значений и собственных векторов из ковариационной или корреляционной матриц; второй – с исследованием соотношений между объектами и часто используется для исследования их внутренней структуры для представления в многомерном пространстве. Большинство *Q*-методов факторного анализа связано с нахождением собственных значений и собственных векторов матрицы сходства между всеми возможными парами объектов. Методы *R*-анализа – это статистические процедуры в том смысле, что данные рассматриваются как выборки, извлеченные из более крупных совокупностей, и результаты его применения обычно сохраняют общие черты, свойственные исходным переменным. Так как методы *R*-анализа связаны с исследованием исходного множества данных сходства между индивидуумами, то их нельзя свести к статистическому анализу.

Первый шаг как *R*-, так и *Q*-метода факторного анализа – это преобразование исходной матрицы данных в квадратную симметричную матрицу, которая выражает либо степени взаимосвязей между переменными, либо то же между объектами, на которых значения этих переменных определены. Это делается путем умножения слева или справа матрицы данных на транспонированную к ней. В простейшем случае, когда матрица необработанных данных $[X]$ состоит из n строк наблюдений и m столбцов переменных, умножение слева на транспонированную к ней матрицу $[X]'$ приводит к квадратной матрице $[R]$ порядка $m \times m$: $[R] = [X]'\,[X]$. Элементы $[R]$ состоят из сумм квадратов и попарных произведений те переменных, представленных в исходной матрице, т.е.

$$r_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \ ,$$

где j и k – номера двух столбцов матрицы данных. Если данные стандартизованы, т.е. каждая переменная имеет нулевое среднее и стандартное отклонение, равное единице, то матрица $[R]$ будет корреляционной матрицей m переменных.

Если теперь матрицу данных $[X]$ умножить справа на транспонированную к ней матрицу [X]', то получим квадратную симметричную матрицу $[Q]$, которая имеет n строк и n столбцов: $[Q] =$

$[X][X]'$.

Если $[X]$ содержит необработанные наблюдения, то $[Q]$ содержит квадраты и попарные произведения всех пар объектов, просуммированные по переменным. Действительно элементы матрицы $[Q]$ таковы:

$$q_{jk} = \sum_{j=1}^{m} x_{ij} x_{ij} ,$$

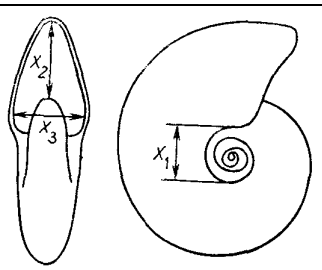
где i и j – номера двух строк матрицы данных. В большинстве исследований используется больше объектов, чем переменных, так что матрица $[Q]$ будет много большего порядка, чем матрица $[R]$, даже несмотря на то что они построены по одной исходной матрице данных $[X]$.

Применение факторного анализа в геологии, как правило, основано на нахождении собственных значений и собственных векторов либо для матрицы $[R]$, либо для матрицы $[Q]$. Однако очевидно, что имеется тесная связь между ними, так как обе матрицы порождены одним и тем же набором данных. Эта связь была установлена на заре развития факторного анализа, но ее использование подвергается пересмотру до сих пор. Отчасти это происходит потому, что психологи и социологи, которые были ответственны за большинство первых работ в области факторного анализа, пользовались исключительно R-методом. С некоторого времени биологи и геологи начали интересоваться теми методами факторного анализа, в которых широко используется техника Q-анализа. Большинство из них были простыми адаптациями методов R-анализа, в которых матрица $[Q]$ просто подставлялась вместо матрицы $[R]$. Так как $[Q]$ зачастую бывает очень большой матрицей, то до 1950 г., пока не появились мощные ЭВМ, эти способы применения Q-метода оказывались безрезультатными. К сожалению, основное соотношение между R- и Q-методами было упущено, и Q-метод факторного анализа часто оказывался чрезмерно сложным (см. например, [29], [52]). В последнее время ряд авторов [33, 12; 65] используют двойственность между R- и Q-методами, в результате чего достигается большое упрощение вычислений в Q-методе факторного анализа.

Теорема Эккарта–Юнга

Основное соотношение между матрицей данных и собственными значениями и собственными векторами двух матриц попарных произведений выражается теоремой Эккарта–Юнга. Впервые она была доказана этими двумя авторами в их классической работе, появившейся в первом томе журнала «Психометрика» (1936 г.). Теорема Эккарта–Юнга является краеугольным камнем ряда многомерных методов, включая факторный анализ. Она утверждает, что для всякой вещественной матрицы $[X]$ можно найти две ортогональные матрицы $[V]$ и $[U]$, для которых произведение $[V]'[X][V]$ есть вещественная диагональная матрица $[\Lambda]$ с неотрицательными элементами. Доказательство теоремы было дано Джонсоном [31]. Остановимся на следствиях из этой теоремы, которые важны в факторном анализе, используя числовой пример, первоначально рассмотренный Бертом [6].

Таблица 6.17. Измерения (в мм) четырех образцов гониатитовых аммоноидей, представляющих виды из рода *Manticoceras*

Вид	Ширина пупка, X_1	Высота последней камеры, X_2	Ширина последней камеры, X_3	
A	4	27	18	
B	12	25	12	
C	10	23	16	
D	14	21	14	
	$X_1 = 10$	$X_2 = 24$	$X_3 = 15$	

Данные, приведенные в табл. 6.17, типичны для простого геологического исследования. Вообразим, что эти измерения были сделаны палеонтологом на раковинах четырех видов гониатитовых аммоноидей. Рассматриваемые характеристики включают диаметр «пупка», т.е. части раковины

не закрытой последней камерой, а также высоту и ширину этой камеры в устье. Среднее каждой переменной можно вычесть из каждого наблюдения. В результате получим матрицу данных

$$[X]=\begin{bmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{bmatrix}$$

Используя теорему Эккарта–Юнга, покажем, что матрица данных может рассматриваться как произведение трех других матриц:

$$[X] = [V][\Lambda][U]' \qquad (6.40)$$

где V – ($n\times r$)-матрица, столбцы которой ортонормальны. Это означает, что $[V]'[V] = [I]$, где $[I]$ – $r\times r$ матрица. Аналогично, U – ($m\times r$)-матрица, столбцы которой ортонормальны, так что $[U]'[U] = [I]$, где $[I]$ также есть ($r\times r$) матрица. $[\Lambda]$ – квадратная ($r\times r$)-матрица, имеющая r положительных элементов вдоль диагонали. Последние называются сингулярными значениями матрицы $[X]$, недиагональные элементы матрицы $[\Lambda]$ равны нулю.

Меньшая матрица-произведение $[R] = [X]'[X]$ имеет порядок $m\times m$ и содержит r ненулевых собственных значений и $(m - r)$ собственных значений, равных нулю. Ненулевые собственные значения равны квадратам сингулярных значений матрицы $[\Lambda]$, т.е. $[\Lambda]^2 = [I][\lambda]'$, или эквивалентно

$$[\Lambda] = [I][\sqrt{\lambda}]' \qquad (6.41)$$

где $[\lambda]$ – вектор, содержащий r ненулевых собственных значений матрицы $[R]$. Большая матрица-произведение $[Q]= [X][X]'$ порядка $n\times n$ также имеет только t ненулевых собственных значений. Они идентичны собственным значениям, полученным по $[R]$, исключая те дополнительные собственные значения, которые равны нулю, если n (число объектов) превосходит m (число переменных).

Далее, столбцы матрицы $[U]$ содержат собственные векторы матрицы $[R]$, которые связаны с каждым собственным вектором λ. Столбцы $[V]$ содержат собственные векторы матрицы $[Q]$. Так как собственные значения обеих матриц $[R]$ и $[Q]$ одинаковы, то должно существовать соотношение между двумя наборами собственных векторов $[U]$ и $[V]$. Это соотношение таково:

$$\begin{aligned} [V] &= [X][U][\Lambda]^{-1} \\ [U] &= [X]'[V][\Lambda]^{-1} \end{aligned} \qquad (6.42)$$

В факторном анализе вектор, образованный умножением собственного вектора на соответствующее ему сингулярное значение, называется фактором. Напомним, что собственные векторы вычисляются так, чтобы сумма квадратов их элементов была равна 1,0, т.е. собственные векторы имеют единичную длину. Если их умножить на соответствующие им сингулярные значения (или квадратные корни из их собственных значений), то их длины будут пропорциональны величинам их сингулярных значений. Индивидуальные элементы некоторого фактора называются нагрузками и связывают факторы с исходными переменными. В матричном обозначении факторы R-метода имеют вид

$$[A^R] = [U][\Lambda] \qquad (6.43)$$

Нагрузки представляют пропорции или веса, которые должны быть в качестве нагрузок приписаны каждой переменной для того, чтобы спроектировать объекты на факторные оси. Они также представляют коэффициенты корреляции индивидуальных переменных с факторами. Соответствующее уравнение для Q-метола факторного анализа имеет вид

$$[A^Q] = [V][\Lambda] \qquad (6.44)$$

и нагрузки пропорциональны вкладу каждого индивидуального объекта, возникающему при проектировании переменных на факторные оси.

Факторные метки R-метода находятся умножением данных на факторные нагрузки или

$$[S^R] = [X][\Lambda^R] \qquad (6.45)$$

что соответствует проектированию n индивидуальных объектов на факторные оси. Для конкретного наблюдения получим

$$s_{ik} = \Sigma a_{mk} x_{mi}$$

или

$$s_{ik} = a_{1k}x_{1i} + a_{2k}x_{2i} + a_{3k}x_{3i} + \ldots + a_{mk}x_{mi}$$

где s_{ik} – нагрузка i-го наблюдения на k-й фактор; x_m – значение m-й переменной, измеренное на объекте i; a_{mk} – нагрузка m-й переменной на k-й фактор. В свою очередь, a_{mk} есть произведение элемента m на k-й собственный вектор, умноженный на квадратный корень из k-го собственного значения.

Аналогично, метки Q-метода находят умножением транспонированной матрицы данных на факторные нагрузки Q-метода;

$$[S^Q] = [X][A^Q] \qquad (6.46)$$

Это уравнение определяет проекции m переменных на факторные оси.

Некоторые алгебраические преобразования показывают связь между факторными нагрузками и метками в R- и Q-методах. Уравнение (6.43) определяет факторные нагрузки R-метода по формуле $[A^R] = [U]\,[\Lambda]$, по теореме Эккарта–Юнга матрица $[U]$ определяется как

$$[U] = [X]'[V][\Lambda]^{-1}$$

Умножая обе стороны на $[\Lambda]$, получаем

$$[U][\Lambda] = [X]'[V][\Lambda]^{-1}[\Lambda]$$

$$[A^R] = [X]'[V]$$

Метки Q-метода определяются уравнением (6.46):

$$[S^Q] = [X]'[A^Q]$$

нагрузки Q-метода $[A^Q]$ определяются по формуле:

$$[A^Q] = [V]\,[\Lambda]$$

После подстановки получаем

$$[S^Q] = [X]'[V][\Lambda]$$

Но $[X]'[V]=[A^R]$, поэтому

$$[S^Q] = [A^R][\Lambda] \qquad (6.47)$$

Аналогичные действия показывают, что

$$[S^R] = [A^R][\Lambda] \qquad (6.48)$$

Таким образом, метки Q-метода пропорциональны нагрузкам R-метода, и наоборот. Коэффициент пропорциональности равен $[\Lambda]$, т.е. сингулярным значениям. Эквивалентные выражения имеют вид

$$[A^R] = [S^Q][\Lambda]^{-1} \qquad (6.49)$$

$$[A^Q] = [S^R][\Lambda]^{-1} \qquad (6.50)$$

Это означает, что если провести R-метод факторного анализа, то автоматически будет проведен и Q-метод, так как и нагрузки, и метки Q-метода можно получить из результатов R-метода.

Проиллюстрируем эти соотношения, используя измерения приведенных ранее аммоноидей. Меньшая матрица произведений R-метода получается умножением матрицы данных слева на ее транспозицию

$$[X]'[X] = [R]$$

$$\begin{bmatrix} -6 & 2 & 0 & 4 \\ 3 & 1 & -1 & -3 \\ 3 & -3 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{bmatrix} = \begin{bmatrix} 56 & -28 & -28 \\ -28 & 20 & 8 \\ -28 & 8 & 20 \end{bmatrix}$$

Собственные значения $[R]$ есть $\lambda_1 = 84$, $\lambda_2 = 12$, $\lambda_3 = 0$. Так как последнее собственное значение равно нулю, то матрица имеет ранг 2, а не 3, т. е.

$$[\Lambda]^2 = \begin{bmatrix} 84 & 0 \\ 0 & 12 \end{bmatrix}, \quad [\Lambda] = \begin{bmatrix} 9{,}165 & 0{,}0 \\ 0{,}0 & 3{,}464 \end{bmatrix}$$

Собственные векторы [R} –

$$[U]=\begin{bmatrix} 0{,}8165 & 0{,}0 & 0{,}0 \\ -0{,}4082 & 0{,}7071 & 0{,}0 \\ -0{,}4082 & -0{,}7071 & 0{,}0 \end{bmatrix}$$

Так как последнее собственное значение равно нулю, то последний столбец матрицы $[U]$ исчезает, и мы получаем матрицу порядка 3×2:

$$[U]=\begin{bmatrix} 0{,}8165 & 0{,}0 \\ -0{,}4082 & 0{,}7071 \\ -0{,}4082 & -0{,}7071 \end{bmatrix}$$

Матрица факторных нагрузок R-метода $[A^R]$ задается уравнением (6.43):

$$\begin{bmatrix} 0{,}8165 & 0{,}0 \\ -0{,}4082 & 0{,}7071 \\ -0{,}4082 & -0{,}7071 \end{bmatrix}\times\begin{bmatrix} 9{,}165 & 0{,}0 \\ 0{,}0 & 3{,}464 \end{bmatrix}=\begin{bmatrix} 7{,}4832 & 0{,}0 \\ -3{,}7412 & 2{,}4494 \\ -3{,}7412 & -2{,}4494 \end{bmatrix}$$

Теперь можно спроектировать 4 вида на оси R-метода факторного анализа, вычисляя их факторные метки по формуле (6.45):

$$\begin{bmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{bmatrix}\times\begin{bmatrix} 7{,}4832 & 0{,}0 \\ -3{,}7412 & 2{,}4494 \\ -3{,}7412 & -2{,}4494 \end{bmatrix}=\begin{bmatrix} -67{,}3 & 0{,}0 \\ 22{,}4 & 9{,}8 \\ 0{,}0 & -4{,}9 \\ 44{,}9 & -4{,}9 \end{bmatrix}$$

Метки можно графически изобразить в пространстве, определенном ортонормальными факторными осями. На рис. 6.14 представлены четыре образца аммоноидей в проекции на плоскость первой и второй факторных осей.

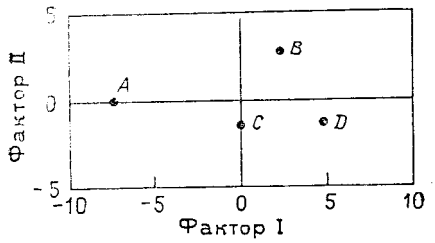


Рис. 6.14. Факторные метки *R*-метода для четырех видов аммоиоидей

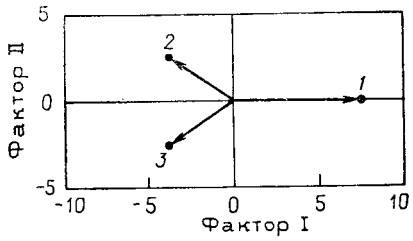


Рис. 6.15. Факторные метки *Q*-метода для трех переменных, измеренных на изученных видах аммоноидей

Q-метод факторного анализа можно начать с умножения справа на матрицу, являющуюся транспозицией матрицы исходных данных:

$$[X][X]'=[Q]$$

$$\begin{bmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{bmatrix}\times\begin{bmatrix} -6 & 2 & 0 & 4 \\ 3 & 1 & -1 & -3 \\ 3 & -3 & 1 & -1 \end{bmatrix}=\begin{bmatrix} 54 & -18 & 0 & -36 \\ -18 & 14 & -4 & 8 \\ 0 & -4 & 2 & 2 \\ -36 & 8 & 2 & 26 \end{bmatrix}$$

Матрицу V можно преобразовать в матрицу Q-метода факторного анализа с помощью уравнения (6.44):

$$[V][\Lambda]=[A^Q]$$

$$\begin{bmatrix} -0{,}8018 & 0{,}0 \\ 0{,}2673 & 0{,}8165 \\ 0{,}0 & -0{,}4082 \\ 0{,}5345 & -0{,}4082 \end{bmatrix} \times \begin{bmatrix} 9{,}165 & 0{,}0 \\ 0{,}0 & 3{,}464 \end{bmatrix} = \begin{bmatrix} -7{,}3485 & 0{,}0 \\ 2{,}4498 & 2{,}8284 \\ 0{,}0 & -1{,}4140 \\ 4{,}8987 & -1{,}4140 \end{bmatrix}$$

Метки Q-метода являются проекциями переменных на факторные оси и находятся умножением транспозиции матрицы данных на факторные нагрузки:

$$[X]'[A^Q] = [S^Q]$$

$$\begin{bmatrix} -6 & 2 & 0 & 4 \\ 3 & 1 & -1 & -3 \\ 3 & -3 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} -7{,}3485 & 0{,}0 \\ 2{,}4498 & 2{,}8284 \\ 0{,}0 & -1{,}4140 \\ 4{,}8987 & -1{,}4140 \end{bmatrix} = \begin{bmatrix} 68{,}8 & 0{,}0 \\ -34{,}3 & 8{,}5 \\ -34{,}3 & -8{,}5 \end{bmatrix}$$

Рис. 6.15 представляет три переменные образцов аммоноидей в проекции на плоскость, определенную первыми двумя факторами Q-метода.

Используя уравнение (6.47), можно утверждать, что метки R-метода пропорциональны нагрузкам Q-метода:

$$[A^R][\Lambda] = [S^Q]$$

$$\begin{bmatrix} 7{,}4832 & 0{,}0 \\ -3{,}7412 & 2{,}4494 \\ -3{,}7412 & -2{,}4494 \end{bmatrix} \times \begin{bmatrix} 9{,}165 & 0{,}0 \\ 0{,}0 & 3{,}464 \end{bmatrix} = \begin{bmatrix} 68{,}6 & 0{,}0 \\ -34{,}3 & 8{,}5 \\ -34{,}3 & -8{,}5 \end{bmatrix}$$

Окончательно можно продемонстрировать справедливость теоремы Эккарта–Юнга, восстановив матрицу данных $[X]$ порядка 7×3 по ее ортонормальным частям:

$$[X] = [V][\Lambda][U]'$$

$$\begin{bmatrix} -0{,}8018 & 0{,}0 \\ 0{,}2673 & 0{,}8165 \\ 0{,}0 & -0{,}4082 \\ 0{,}5345 & -0{,}4082 \end{bmatrix} \times \begin{bmatrix} 9{,}165 & 0{,}0 \\ 0{,}0 & 3{,}464 \end{bmatrix} \times \begin{bmatrix} 0{,}8165 & -0{,}4082 & -0{,}4082 \\ 0{,}0 & 0{,}7071 & -0{,}7071 \end{bmatrix} = \begin{bmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{bmatrix}$$

В этом простом численном примере многомерное множество наблюдений было преобразовано в меньшее число факторов. Также показано, что решения R-метода эквивалентны решениям Q-метода. Оба они представляют критические ситуации и будут повторяться в следующих параграфах, где придется рассматривать некоторые усложнения, накладывающиеся на относительно простые структуры, которые мы только что исследовали.

В начале этого раздела отмечалось, что «факторный анализ» – это привычный термин, включающий множество методов, основанных на нахождении собственных значений и собственных векторов матрицы парных произведений набора данных. Факторный анализ также использовался (более правильно) в строгом смысле как статистическая процедура, при которой матрица данных разлагается на заранее заданное число некоррелированных факторов и набор «уникальных» случайных компонент. Другие важные методы, основанные на собственных значениях, включают метод главных компонент (*МГК*), анализ соответствия и Q-метод, анализ главных векторов и главных координат.

Характеристики различных процедур вычисления собственных значений проиллюстрируем на искусственном примере, аналогичном приведенному Кули и Лонесом [11]. Мы все хорошо знаем, что значит определить «размеры» чего-либо. Можно ли измерить длину, ширину, площадь, объем или какое-либо отношение этих величин? Как установить различие между понятиями «размер» и «форма»? Чтобы получить ответы на эти вопросы, рассмотрим набор 25 объектов, имеющих форму параллелепипедов. Значения трех измерений этих тел выбирались случайно в пределах 10 единиц. В полученном наборе все размеры и формы были равновероятными, от куба со стороной, меньшей единицы, до призмы и плоской пластины или до куба размером 10×10×10 единиц. Совокупности измерений, сделанных на каждом из таких блоков, составили множество значений наших перемен-

ных. Они были выражены следующим образом:

X_1 – длинная ось;

X_2 – средняя ось;

X_3 – короткая ось;

X_4 – самая длинная диагональ;

X_5 – отношение суммы длинной и средней осей к короткой оси;

X_6 – отношение радиуса наименьшей описанной сферы к радиусу наибольшей вписанной сферы;

X_7 – отношение площади поверхности к объему.

Табл. 6.18 содержит 25 наблюдений семи переменных; набор блоков представлен на рис. 6.16. Заметим, что это множество данных обладает некоторыми интересными свойствами; оно имеет самое большее три независимых измерения, так как переменные X_4 и X_7 являются линейными комбинациями длины, ширины и высоты. Аналогично, данные содержат некоторый вклад наведенной корреляции, определяемой внутренними особенностями переменных. Длинная ось каждого блока по определению должна быть длиннее, чем промежуточная ось, которая в свою очередь длиннее, чем короткая ось. Это означает, что если, например, длина и ширина нанесены на график вместе, то представляющие их точки окажутся расположенными ниже диагонали (рис. 6.17). Это индуцирует положительную корреляцию, т.е. значительно больше ($r = 0{,}58$) ожидаемой корреляции для двух независимых переменных.

Таблица 6.18. Результаты измерения 25 параллелепипедов со случайной длиной сторон*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
a	3,760	3,660	0,540	5,275	9,768	13,741	4,782
b	8,590	4,990	1,340	10,022	7,500	10,162	2,130
c	6,220	6,140	4,520	9,842	2,175	2,732	1,089
d	7,570	7,280	7,070	12,662	1,791	2,101	0,822
e	9,030	7.080	2,590	11,762	4,539	6,217	1,276
f	5,510	3,980	1,300	6,924	5,326	7,304	2,403
g	3,270	0,620	0,440	3,357	7,62°	8,838	8,389
h	8,740	7,000	3,310	11,675	3,529	4,757	1,119
i	9,640	9,490	1,030	13,567	13,133	18,519	2,354
j	9,730	1,330	1,000	9,871	9,871	11,064	3,704
k	8,590	2,980	1,170	9,170	7,851	9,909	2,616
l	7,120	5,490	3,680	9,716	2,642	3,430	1,189
m	4,690	3,010	2,170	5,983	2,760	3,554	2,013
n	5,510	1,340	1,270	5,808	4,566	5,382	3,427
o	1,660	1,610	1,570	2,799	1,783	2,087	3,716
p	5,900	5,760	1,550	8,388	5,395	7,497	1,973
q	9,840	9,270	1,510	13,604	9,017	12,668	1.745
r	8,390	4,920	2,540	10,053	3,956	5,237	1,432
s	4,940	4,380	1,030	6,678	6,494	9,059	2,807
t	7,230	2,300	1,770	7,790	4,393	5,374	2,274
u	9,460	7,310	1,040	11,999	11,579	16,182	2,415
v	9,550	5,350	4,250	11,742	2,766	3,509	1,054
w	4,940	4,520	4,500	8,067	1,793	2,103	1,292
x	8,210	3,080	2,420	9,097	3,753	4,657	1,719
y	9,410	6,440	5,110	12,495	2,446	3,103	0,914

* Названия переменных см. в тексте.

Конечно, обычно измерения, о которых известно, что они зависят друг от друга, не изучаются. К сожалению, чаще случается, что не те геологические переменные взаимосвязаны; составные

переменные содержат индуцированные корреляции, так как они являются частью целого, и таксономические измерения могут быть связаны друг с другом в силу эффекта раз мера. В этом искусственном примере мы заранее знаем, что взаимозависимость существует, и можем надеяться, что в ре зультатах анализа эти связи обнаружатся. Этот пример может помочь понять результаты, которые получаются из анализа реальных данных, где существование подобных зависимостей нельзя установить заранее.

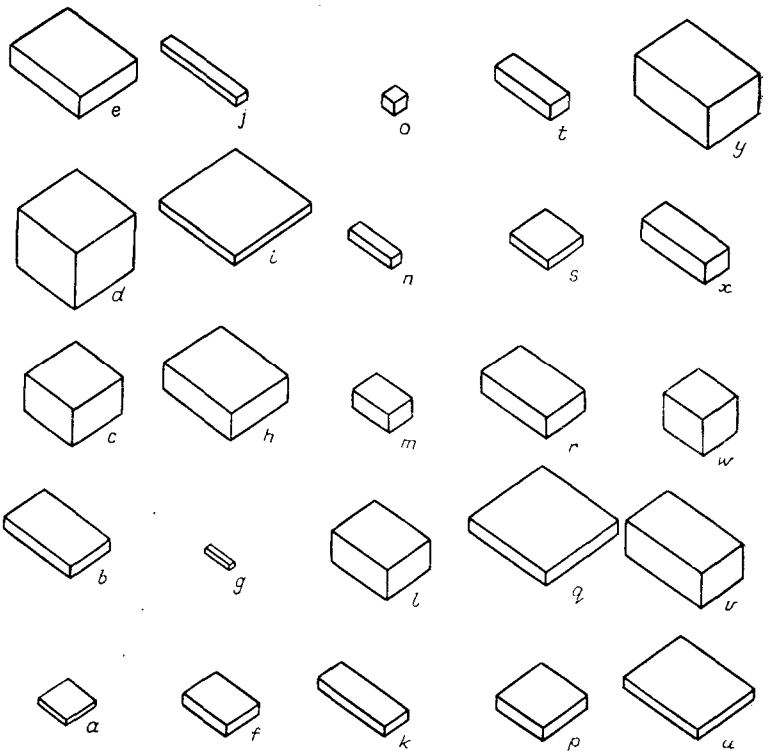


Рис. 6.16. Двадцать пять блоков со случайными значениями длины, ширины и высоты. *а–у* – см. в табл. 6.18

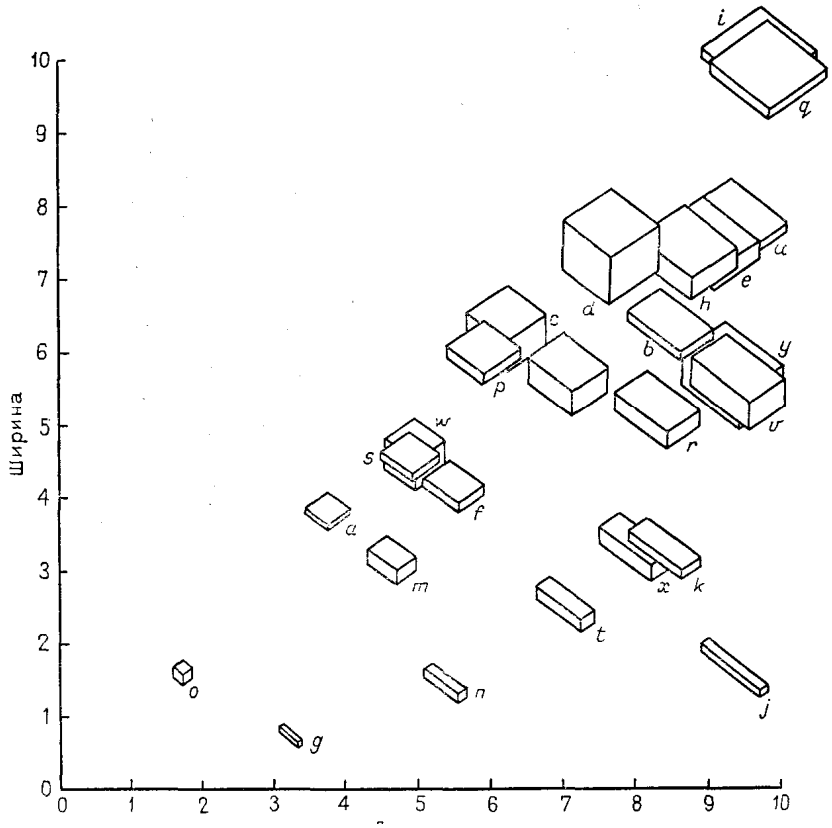


Рис. 6.17. Представление зависимости ширины случайных блоков от их длины.
Так как ширина меньше длины, то все блоки расположены в нижней части диаграммы под диагональю. Такое положение соответствует корреляции r=0,58 между длиной и шириной. Длина и ширина приведены в произвольных единицах

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Первая важная процедура, которую мы рассмотрим в этом разделе – это метод главных компонент (*МГК*). Главные компоненты – это не что иное как собственные векторы ковариационной матрицы. Сами по себе они позволяют глубже проникнуть в структуру матрицы, но часто их можно интерпретировать и как факторы. Многие современные схемы факторного анализа используют метод главных компонент в качестве отправного пункта для анализа. По этой причине, а также в силу того что доказательства и интерпретация метода главных компонент весьма просты, мы начнем с его рассмотрения.

Как уже отмечалось, геологи часто испытывают неудобство от пользования этой терминологией; большинство из опубликованных работ показывает, что геологи называют «факторным анализом» метод главных компонент. «Факторы», встречающиеся в этих работах, на самом деле нужно называть компонентами. Некоторые авторы испытывают столь большую неловкость от этой путаницы, что нашли выход в том, что ввели для факторного анализа, в отличие от компонентного, название «истинный факторный анализ» [33].

Предположим, что измерены две переменные на множестве объектов, например длина и ширина раковин брахиопод. Полученные данные приведены в табл. 6.19 и графически изображены на рис. 6.18. Дисперсия переменной X_1 равна 20,3, переменной X_2 – 24,1, а ковариация равна 1566. Это можно представить в виде ковариационной матрицы

$$s^2 = \begin{bmatrix} 20{,}3 & 15{,}6 \\ 15{,}6 & 24{,}1 \end{bmatrix}$$

В гл. 3 указывалось, что матрицу можно представить геометрически в многомерном пространстве как множество векторов. Будем считать, что каждая строка матрицы дает координаты концевых точек вектора, представляющего эту строку. Матрицу порядка 2×2 можно представить на плоской диаграмме, как это сделано на рис. 6.19. Более того, эти векторы можно считать произвольными осями m-мерного эллипсоида. Собственные векторы матрицы дают ориентацию главных осей эллипсоида, а собственные значения представляют длины каждой из последовательных главных полуосей. В гл. 3 так были интерпретированы произвольные матрицы, но очевидно, что ковариационные матрицы легко можно интерпретировать так же. Метод главных компонент сводится к нахождению этих осей и измерению их длин.

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

Вычислим собственные векторы и собственные значения нашей матрицы $[s^2]$ порядка 2×2 и изобразим полученные векторы графически. Первый собственный вектор имеет координаты

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

что означает, что первый собственный вектор, соответствующий I главной оси эллипсоида, отклоняется на 0,66 единиц по $S_1{}^2$ для каждых 0,75 единиц по $S_2{}^2$. Первое собственное значение равно 37,9 и является длиной главной полуоси. Второй собственный вектор имеет координаты

$$II = \begin{bmatrix} 0{,}75 \\ -0{,}66 \end{bmatrix}$$

Ясно, что он 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. 6.20. 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.

Таблица 6.19. Двухмерные наблюдения с дисперсией X_1, равной 20,3, дисперсией X_2, равной 24,1, и ковариацией 15,6

X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
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

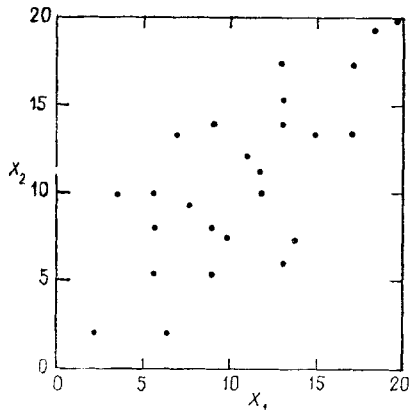


Рис. 6.18. Диаграмма рассеяния двумерных данных табл. 6.19

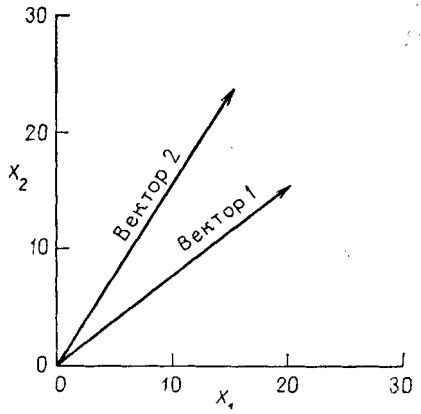


Рис. 6.19. Векторное представление дисперсий и ковариаций в виде матрицы порядка 2×2. Вектор 1 представлен первой строкой матрицы, а вектор 2 – второй строкой

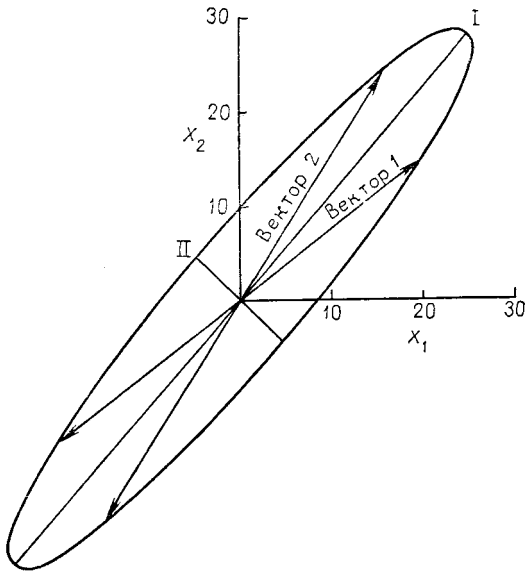


Рис. 6.20. Эллипс, определенный дисперсиями и ковариациями данных табл. 6.19 для ореола точек, представленных на рис. 6.18. Главная компонента *I* соответствует 86% суммарной дисперсии, главная компонента *II* – 14%

Определим суммарную дисперсию рассматриваемых данных как сумму вкладов от индивидуальных дисперсий. Так как последние расположены на диагонали ковариационной матрицы, то эта процедура эквивалентна нахождению следа матрицы. В данном примере суммарная дисперсия равна 20,3 + 24,1 = 44,4. Вклад первой переменной составляет 20,3/44,4, или около 46% суммарной дисперсии, а вклад второй – примерно 54%. Как указывалось в гл. 3, сумма собственных значений матрицы равна ее следу, поэтому сумма ее собственных значений также равна 37,9 + 6,5 = 44,4. Так как эти собственные значения определяют длину двух главных осей, то последние также характеризуют суммарную дисперсию множества данных, и вклад каждой из них в суммарную дисперсию равен соответствующему собственному значению, деленному на след матрицы. Первая главная ось составляет 37,9/44,4, или 86% суммарной дисперсии, в то время как вторая ось – только 14%. 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 $Y_1 = \beta_1 X_1 + \beta_2 X_2$, 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. 6.20 представлены данные табл. 6.19, преобразованные таким образом. В табл. 6.20 каждое исходное наблюдение заменено его проекцией на главные оси. Проектирование на первую главную ось осуществляется по формуле

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

где коэффициенты при X_1 и X_2 являются координатами первого собственного вектора. Проектирование на вторую главную ось осуществляется по формуле

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

Таблица 6.20. Главные компоненты для данных табл. 6.19, вычисленные с помощью проектирования исходных данных на главные оси; дисперсия Y_1 равна 37,9, дисперсия Y_2 равна 6,5

Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
3,49	0,92	11,96	1,43	19,85	–0,22
10,14	–3,64	16,45	-2,45	21,35	–1,54
7,72	1,18	11,87	2,84	14,52	5,84
9,97	–0,81	16,28	0,28	19,68	2,60
11,46	–2,14	15,44	2,35	21,00	4,10
6,14	3,91	16,19	1,69	24,00	1,45
14,37	–3,38	13,11	5,75	26,16	0,87
12,04	0,02	19,10	0,45	28,23	1,70
9,71	3,42				

Координаты собственных векторов, используемые для вычисления проекций наблюдений, называются нагрузками. Они являются коэффициентами линейного уравнения, которое используется для определения собственного вектора. В литературе по факторному анализу обычно используется термин «нагрузка переменной A на первый фактор», который означает, что речь идет о коэффициенте первого собственного вектора при переменной A. Эту операцию в матричной форме можно представить следующим образом $[X]\,[U] = [S^R]$, где $[S^R]$ – $(n \times m)$-матрица нагрузок на главные компоненты. $[X]$ – это $(n \times m)$-матрица исходных наблюдений, и $[U]$ –квадратная матрица, содержащая последовательные собственные векторы в m столбцах из m элементов и каждый из них соответствует некоторой исходной переменной. В нашем примере

$$\begin{bmatrix} 3 & 2 \\ 4 & 10 \\ 6 & 5 \\ \vdots & \vdots \\ 20 & 20 \end{bmatrix} \times \begin{bmatrix} 0{,}66 & 0{,}75 \\ 0{,}75 & -0{,}66 \end{bmatrix} = \begin{bmatrix} 3{,}49 & 0{,}92 \\ 10{,}14 & -3{,}64 \\ 7{,}72 & 1{,}18 \\ \vdots & \vdots \\ 28{,}23 & 1{,}70 \end{bmatrix}$$

Вернемся к нашему множеству данных. Мы определили собственные векторы матрицы и нашли, что первый собственный вектор дает вклад в суммарную дисперсию около 86%. 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 46%, либо 54% изменчивости, в зависимости от того, какую переменную мы сохраним. Однако если спроектировать все наблюдения на первую главную ось, то потеря составит только 14% от изменчивости данных.

На рис. 6.21 представлены проекции данных точек на главные оси. Необходимо отметить, что дисперсия, соответствующая первой главной оси, больше любой из дисперсий по какой-либо прямой, проходящей через точки заданного множества. Однако эта дисперсия не превосходит сумму дисперсий вдоль двух осей X_1 и X_2 и если второй собственный вектор исключается из рассмотрения,

то неизбежно происходит потеря изменчивости. В этом легко убедиться, если спроектировать вектор факторных значений Y_{1i} снова на оси X_1 и X_2. Хотя при этом некоторые точки и могут удаляться от своего среднего значения, все равно происходит уменьшение дисперсии по обеим переменным (рис. 6.22).

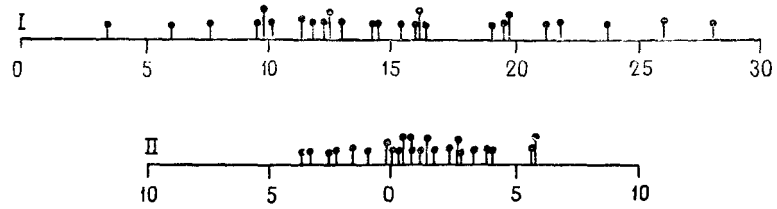


Рис. 6.21. Проекции точечных данных табл. 6.19 на главные компоненты. Дисперсия по компоненте *I* равна 37,9, а дисперсия по компоненте *II* – 6,5

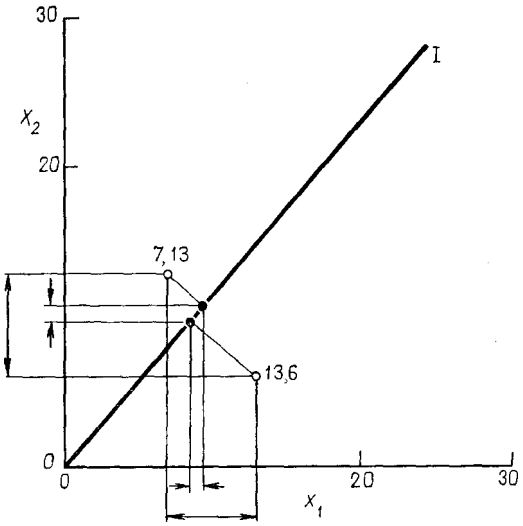


Рис. 6.22. Две первоначально заданные точки (светлые кружки) проектируются на главную компоненту *I*. Проекции их образов (темные кружки) на оси координат приводят к меньшей дисперсии. Потери дисперсии происходят из-за того, что не используется главная компонента *II*

Теперь рассмотрим данные табл. 6.19 и ранжируем каждое наблюдение по двум переменным. Тогда на графике эти данные будут выглядеть так, как изображено на рис. 6.23. Ранжирование приводит к тому, что переменные оказываются в высшей степени коррелированными, и этот факт отражен в том, что ковариация теперь равна 21,9. Так как использовались одни и те же наблюдения, то дисперсии остаются неизменными. Если теперь найти собственные векторы и собственные значения новой ковариационной матрицы, то можно обнаружить, что собственные векторы почти не уменьшились:

$$I = \begin{bmatrix} 0{,}68 \\ 0{,}74 \end{bmatrix} \quad , \quad II = \begin{bmatrix} 0{,}74 \\ -0{,}68 \end{bmatrix}$$

Однако новые собственные значения будут совершенно другими. Первое собственное значение равно 44,2, так что теперь первая главная ось дает вклад в суммарную дисперсию, равную 44,2/44,4, т.е. более 99%. Вторая ось так мала, что ее почти невозможно нанести на диаграмму (рис. 6.24). Очевидно, что можно отбросить вторую главную компоненту с очень малой потерей дисперсии множества данных. Сделав это, можно представить исходные двумерные данные с помощью одной единственной переменной (определенной первой главной компонентой), а это означает, что осуществлена редукция размерности исходных данных от двух к одному.

Вместо ранжирования данных можно было бы провести рандомизацию, как показано на рис. 6.25. Рандомизация разрушает всю корреляцию между двумя переменными, т.е полученное в результате этой процедуры значение ковариации равно нулю. Дисперсии, естественно, остаются такими же, так как мы использовали те же данные и просто изменили их порядок. Если найти собственные значения и собственные векторы этой ковариационной матрицы, то получим два вектора:

$$I = \begin{bmatrix} -0{,}22 \\ 0{,}98 \end{bmatrix} \quad , \quad II = \begin{bmatrix} 0{,}98 \\ 0{,}22 \end{bmatrix}$$

Два собственных значения почти равны между собой: первое равно 24,3, а второе 20,1. Таким образом, найдены две главные оси эллипса, который почти совпадает с окружностью (рис. 6.26). Это находится в соответствии с исходными предположениями о том, что корреляция между переменными почти нулевая; следовательно, две исходные оси почти ортогональны. Так как две исходные переменные почти равны по величине, то найденные оси определяют эллипс, близкий к окружности. Никакие другие оси, даже найденные методом главных компонент не будут лучше, чем эти две исходные переменные. В указанной ситуации не существует такого преобразования исходных данных, которое позволило бы уменьшить число переменных без значительной потери информации. Конечно маловероятно, чтобы естественные множества данных имели нулевые ковариации для всех пар переменных.

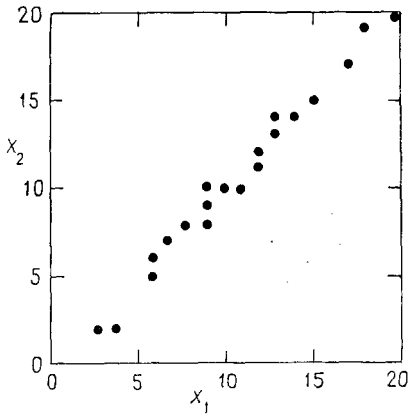


Рис. 6.23. Данные табл. 6.19 ранжированы и нанесены на график

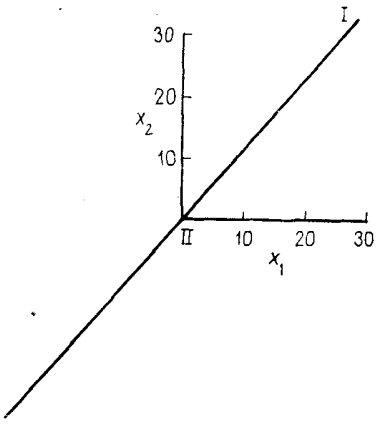


Рис. 6.24. Эллипс, определенный векторами дисперсий и ковариаций ранжированных данных рис. 6.23. Главная компонента *I* чрезвычайно вытянутого эллипса соответствует 99% общей дисперсии

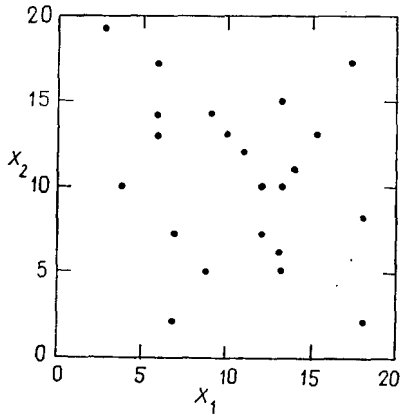


Рис. 6.25. Графическое представление рандомизированных данных табл. 6.19

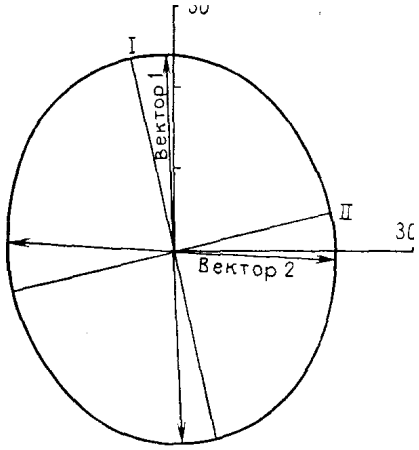


Рис. 6.26. Близкий к окружности эллипс, определенный векторами дисперсий и ковариаций для рандомизированных данных рис. 6.15. Главная компонента *I* дает лишь незначительно больший вклад в суммарную дисперсию, чем любая из исходных переменных

В этом простом примере сначала вычислялась ковариационная матрица, элементы которой отражают исходные единицы измерения. При условии, что все переменные выражены в одних и тех же или соизмеряемых единицах, главные компоненты будут отражать относительный вес различных переменных. Однако метод главных компонент чувствителен по отношению к величинам измерений, поэтому, если длины раковин брахиопод заданы в сантиметрах, а ширина – в миллиметрах, то переменная, представляющая ширину, будет оказывать в 10 раз большее влияние на результат, чем переменная, представляющая длину.

Очередной путь преодоления этого затруднения состоит в стандартизации всех переменных так, чтобы они имели среднее, равное 0,0, и дисперсии, равные 1,0. Тогда элементы ковариационной матрицы будут состоять их корреляций, и главные компоненты будут иметь безразмерный вид. Цель стандартизации выровнять влияние переменных, дисперсия которых велика. Это может оказаться нежелательной, но неизбежной процедурой, если исходные переменные выражены в различных и несовместимых единицах. Ле Мэтр [42] приводит интересные примеры, относящиеся к влиянию стандартизации на главные компоненты, вычисляемые по геохимическим и петрографическим данным.

В большинстве геологических примеров относительные величины переменных играют су-

щественную роль и поэтому предпочтительно проводить исследования с оригинальными переменными и исходной ковариационной матрицей. В тех условиях, когда требуется стандартизовать переменные для того, чтобы сделать их совместимыми, следует помнить, что свойство, которое кажется относительно незначимым, может оказать сильное влияние на анализ. Аналогично, если компоненты вычисляются из корреляционной матрицы, нагрузки должны быть вычислены из стандартизованных, а не исходных значений переменных.

Рассмотрим эффекты анализа главных компонент на более широком множестве данных, используя измерения на 25 случайно генерированных блоках, перечисленных в табл. 6.18. Ковариационная матрица для семи переменных приведена в табл. 6.21. Собственные векторы являются координатами осей главных компонент данных по блокам.

Таблица 6.21. Ковариационная матрица для семи переменных, измеренных на 25 параллелепипедах (приведена только нижняя часть симметричной матрицы)

	Переменная						
Переменная	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	5,400						
X_2	3,260	5,846					
X_3	0,7785	1,465	2,774				
X_4	6,391	6,083	2,204	9,107			
X_5	2,155	1,312	–3,839	1,610	10,710		
X_6	3,035	2,877	–5,167	2,782	14,770	20,780	
X_7	–1,996	–2,370	–1,740	–3,283	2,252	2,622	2,594
Матрица собственных векторов							
Переменная	Собственный вектор						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
X_1	0,164	0,422	0,645	–0,090	0,225	0,415	–0,385
X_2	0,142	0,447	–0,713	–0,050	0,395	0,066	–0,329
X_3	–0,173	0,257	–0,130	0,629	–0,607	0,280	–0,211
X_4	0,170	0,650	0,146	0,212	0,033	–0,403	0,565
X_5	0,546	–0,135	0,105	0,165	–0,161	–0,596	–0,513
X_6	0,768	–0,133	–0,149	–0,062	–0,207	0,465	0,327
X_7	0,073	–0,313	0,065	0,719	0,596	0,107	0,092
Собственные значения	34,490	19,000	2,540	0,810	0,340	0,033	0,003
Вклад каждого собственного значения (в %) в суммарную дисперсию	60,290	33,210	4,440	1,410	0,600	0,060	0,004

Первые два собственных вектора составляют примерно 93% общей дисперсии изучаемого множества данных. Первый, составляющий примерно 60% общей дисперсии, дает наибольшие вклады в переменные X_5 и X_6. Обе эти переменные являются отношениями, содержащими в знаменателе переменную X_3, т.е. длину короткой оси. Отсюда можно сделать вывод, что первая главная компонента позволяет устанавливать различия в толщине параллелепипедов. На рис. 6.27 видно, что это так: очень плоские тела размещаются далеко справа, изометричные – далеко слева. Первая компонента не позволяет осуществить разделение брусоподобных и плоских тел, так как каждое из них имеет хотя бы одну очень малую короткую ось. Отсюда мы заключаем, что первая компонента характеризует высоту по отношению к остальным размерам.

Второй собственный вектор имеет большие веса по всем трем осям и по длине главной диагонали. Мы можем интерпретировать его как фактор, обобщенно отражающий размеры тела. Эта интерпретация хорошо согласуется с рис. 6.27. Вдоль второй главной компоненты параллелепипеды рассортированы в соответствии с их размерами, причем наименьший расположен внизу, а наибольший – сверху.

Интерпретация смысла нагрузок на главные компоненты иногда называется апробацией. Возможно, некоторые специалисты считают, что использование этого термина делает процесс более приемлемым. Графическое представление нагрузок на главные компоненты оказывает существенную помощь в интерпретации (рис. 6.28). Напомним, что в гл. 3 собственные векторы подвергались стандартизации так, чтобы сумма квадратов их длин равнялась единице. Поэтому нагрузки отражают лишь относительную значимость переменной относительно некоторой главной компоненты, а не значимость самой этой компоненты.

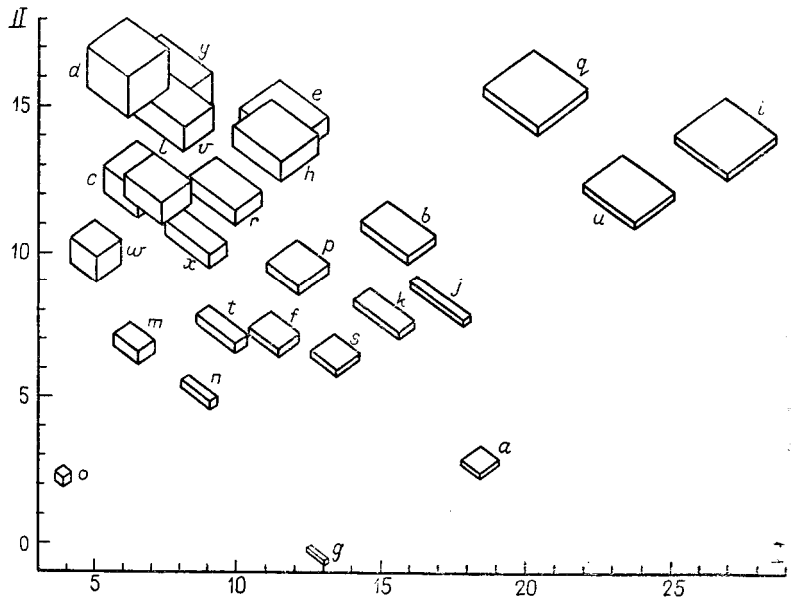


Рис. 6.27. Метки главных компонент данных по блокам представлены на плоскости главных компонент *I* и *II*

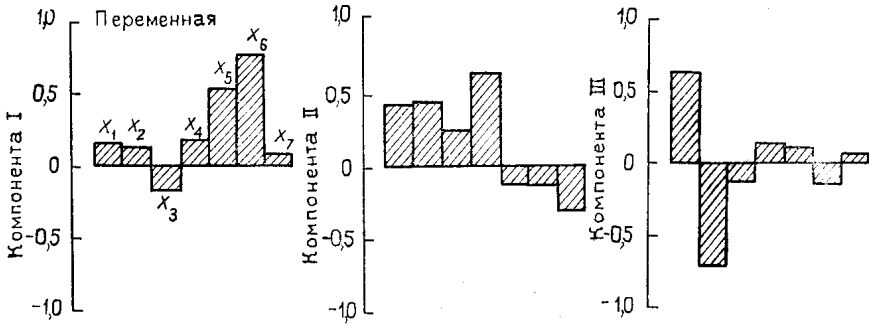


Рис. 6.28. Представление нагрузок на главные компоненты для данных по случайным блокам

Хотя *МГК* и привел к результату, который мы ожидали в этом эксперименте, разделение форм не является вполне определенным. Третья главная компонента дает вклад в дисперсию лишь около 4%, и ее можно рассматривать как фактор, отражающий длину средней по величине оси. По-видимому, в соединении с двумя первыми компонентами эта компонента позволяет осуществить полное разделение плоских и брусоподобных тел. Этот результат не является неожиданным, так как в качестве независимых переменных в эксперименте выбирались значения длины осей. Хотя двух компонент достаточно для характеристики большей части изменчивости изучаемых данных, все же третья компонента необходима для выделения существенных деталей. Этот пример показывает, что *МГК* – мощный метод определения истинного числа линейно независимых векторов, содержащихся в матрице. Поэтому он позволяет измерить избыточность множества переменных.

В качестве примера применения *МГК* к геологическим задачам рассмотрим данные, взятые у Крамбейна и Эбердина [37], приведенные в табл. 6.22. Они представляют собой результаты 50 гранулометрических анализов проб осадков, взятых со дна залива Баратария в западной части дельты Миссисипи (штат Луизиана). Эти пробы принадлежат различным донным фациям, соответствующим разным типам седиментации. Ситовой анализ проводился с помощью комплекса сит с интервалом 1–Ф. Полученные данные представляли собой весовые процентные отношения фракций различного размера.

Таблица 6.22. Пятьдесят гранулометрических анализов проб (в %) донных осадков, взятых в заливе Баратария, штат Луизиана*

Типы осадков**	Ф категории							Типы осадков**	Ф категории						
	1-2	*2-3*	*3–4*	*4-5*	*5-6*	*6-7*	*7-8*		*1-2*	*2-3*	*3–4*	*4-5*	*5-6*	*6-7*	*7-8*
I	0,6	70,2	29,2	0,0	0,0	0,0	0,0	*III*	0,5	5,9	32,2	32,7	4,9	5,4	2,7
	1,0	69 9	29,1	0,0	0,0	0,0	0,0		1,1	4,9	31,1	41,9	13,9	7,8	3,7
	0,8	73,7	25,5	0,0	0,0	0,0	0,0		7,9	8,5	21,0	19,9	8,9	5,9	6,3
	0,9	75,3	23,8	0,0	0,0	0,0	0,0		0,9	13,6	43,9	20,1	7,2	4,8	9,5
	0,3	62,5	36,9	0,0	0,0	0,0	0,0		2,9	15,5	37,0	30,3	5,1	1,9	2,2
	1,1	68,8	30,1	0,0	0,0	0,0	0,0		2,1	16,7	39,6	17,7	8,3	8,3	7,3
	0,8	10,2	79,2	9,8	0,0	0,0	0,0		0,3	20,6	55,4	16,6	6,2	6,1	5,5
	1,0	16,3	73,8	8,9	0,0	0,0	0,0	*IV*	1,2	1,6	15,3	38,4	13,0	9,5	5,6
	1,8	35,7	61,9	0,6	0,0	0,0	0,0		2,3	7,9	23,9	25,5	9,2	7,9	7,7
II	9,5	15,8	59,0	8,4	0,9	0,9	1,4		1,0	3,1	15,2	32,0	14,3	10,0	7,2
	2,4	14,5	53,9	12,2	5,5	1,6	2,5		0,0	11,5	28,4	19,1	7,3	7,8	4,8
	2,2	38,8	42,2	7,9	1,4	1,8	1,0		0,8	7,0	31,6	21,1	10,2	9,0	6,3
	1,7	30,4	44,5	11,2	3,0	1,9	2,9		0,5	2,1	14,0	37,2	19,9	11,4	6,1
	0,0	40,0	32,5	3,8	4,5	6,5	2,7		0,0	3,4	19,7	25,4	15,7	10,2	9,9
	0,0	37,0	45,4	7,3	3,8	3,3	3,8		1,4	1,9	14,4	40,2	8,5	8.4	7,1
	0,3	15,6	54,1	21,3	4,1	2,6	2,0		0,4	3,5	18,8	29,5	11,2	10,4	7,5
	0,3	24,4	56,0	15,1	4,2	0,0	0,0		0,8	6,3	18,2	28,0	9.1	9,7	9,9
	10,5	29,2	37,3	15,1	4,2	3,7	0,0	*V*	1,0	2,3	6,6	16,2	12,0	11,4	13,3
	0,3	13,3	63,5	14,2	4,0	3,4	1.3		3,2	3,9	10,5	24,1	14,2	15,4	13,5
	1,2	26,9	54,7	11,0	3,9	2,3	0,0		2,1	2,1	10,7	23,6	15,1	14,0	11,8
III	0,4	3,9	45,2	24,7	3,7	8,1	3,0		4,4	8,1	8,9	19,9	12,0	11,4	10,8
	0,0	13,8	39,3	15,4	9,1	4,5	6,4		0,6	3,6	4,2	17,8	12,4	10,8	9,9
	0,4	4,0	38,2	28,5	6,0	4,3	4,7		0,5	4,1	9,8	27,9	13,5	13,5	7,4
	1 9	11,5	49,5	22,4	5,7	4,5	2,0		0,7	2,3	5,2	23,2	19,4	14,1	10,1
	0,4	5,1	31,8	30,3	5,4	7,8	3,0		3,4	1,6	4,4	18,0	14,7	15,3	15,1

* Модифицированные данные Крамбеина и Эбердина [37].

** *I*–пляжевые и прибрежные пески; *II* – алевритовые русловые пески; *III* – алевритовые береговые пески; *IV* – органический донный алеврит; *V* – органические илы.

В качестве изучаемых переменных рассматривались процентные содержания фракций определенного размера в каждой пробе. Те же переменные использовались при вычислении таких статистических характеристик, как среднее значение, коэффициент отсортированности и асимметрия распределения размера зерен. С помощью *МГК* можно исследовать взаимосвязь между различными фракциями и найти наиболее эффективную их комбинацию, причем термин «наиболее эффективный» соответствует фактору, дающему наибольший вклад в суммарную дисперсию. Мы вправе ожидать, что нагрузки на первую главную компоненту некоторым образом можно считать аппроксимацией среднего значения, так как обычно эта статистика – наиболее эффективная их всех возможных статистик.

Анализ начинается с вычисления элементов ковариационной матрицы (табл. 6.23). Стандартизация в этом случае необязательна, так как исходные данные измерены в одних и тех же единицах для всех переменных. Необходимо отметить, что данные представляют собой приближенно замкнутую матрицу (т.е. в большинстве случаев сумма переменных составляет 100%), которая снова напоминает нам о теоретически интересном и важном вопросе, связанном с индуцированными отрицательными корреляциями. Ковариационная матрица «переопределена», т.е. она содержит больше строк и столбцов, чем это необходимо. Очевидно, если мы знаем *A, B* и *C* и сумму *A*+*B*+*C*, то имеем информации больше, чем нам нужно в действительности, и одна переменная является избыточ-

ной. Неизбежно, что одно собственное значение матрицы, построенной по таким данным, будет обязательно нулевым. В рассмотренном примере сумма по всем переменным не составляет в точности 100%, так как наблюдения меньше 8Ф отбрасываются. Последнее собственное значение ковариационной матрицы поэтому очень мало, но не равно нулю, как это было бы в случае замкнутой матрицы.

Таблица 6.23. Ковариационная матрица гранулометрических анализов осадков из залива Баратария, штат Луизиана (приведена только нижняя часть симметричной матрицы)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	4,8443						
X_2	–2,6234	468,8480					
X_3	–0,0011	81,3941	353,1255				
X_4	–1,5449	–200,2109	–84,6165	130,2741			
X_5	–0,5972	–84,2597	–73,0435	44,7616	30,4350		
X_6	–0,3805	–71,2097	–66,5433	34,9927	23,7565	22,4189	
X_7	–0,0222	–57,8578	–56,1533	23,9136	19,3907	17,9388	17,967

Ключ: X_1 = 1–2Ф, X_2 = 2–3Ф, X_3 = 3–4Ф, X_4 = 4–5Ф, X_5 = 5–6Ф, X_6 = 6–7Ф, X_7 = 7–8Ф.

Главные компоненты, т.е. собственные векторы по данным залива Баратария, приведены в табл. 6.24. Отметим, что две первые компоненты учитывают 90% изменчивости данных. Нагрузки по переменным для двух компонент представлены на рис. 6.29. Из этого графика видно, что первая главная компонента характеризует относительные доли тонких и очень тонких фракций в осадке, т.е. отношение песок/(глина+ил). Вторая компонента характеризуется отношением содержаний мелкого и очень мелкого песка, а все другие переменные имеют веса, близкие к нулю. Этих двух компонент вполне достаточно для описания почти всей изменчивости исходных данных, из которого вытекает, что разделение на илистую и глинистую фракции несущественно. Основные различия между осадками можно почти полностью описать только с помощью двух переменных.

Таблица 6.24. Собственные значения и собственные векторы (главные компоненты) ковариационной матрицы, приведенной в табл. 6.23

Вектор	Собственное значение	Вклад в дисперсию, %	Сумма вкладов в дисперсию
I	659,7759	64,18	64,19
II	318,4384	30,98	98,17
III	35,1959	3,42	98,59
IV	6,7528	0,66	99,25
V	3,8193	0,37	99,62
VI	2,3763	0,23	99,85
VII	1,5540	0,15	100,00

Переменная	Собственный вектор						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
X_1	–0,0019	0,0039	–0,0689	–0,5829	0,7554	0,2798	0,0818
X_2	0,7710	–0,4777	0,3194	0,1885	0,1169	0,1581	0,0326
X_3	0,4167	0,8647	0,0531	0,2116	0,1123	0,1294	0,0421
X_4	–0,3907	0,0761	0,8844	0,0704	0,0490	0,2280	0,0028
X_5	–0,1895	–0,0794	–0,0775	0,6308	0,6255	–0,3240	–0,2401
X_6	–0,1618	–0,0813	–0,1629	0,3330	0,0526	0,2570	0,8723
X_7	–0,1308	–0,0735	–0,2750	0,2570	–0,0815	0,8107	–0,4146

Ключ: X_1 = 1–2Ф, X_2 = 2–3Ф, X_3 = 3–4Ф, X_4 = 4–5Ф, X_5 = 5–6Ф, X_6 = 6–7Ф, X_7 = 7–8Ф.

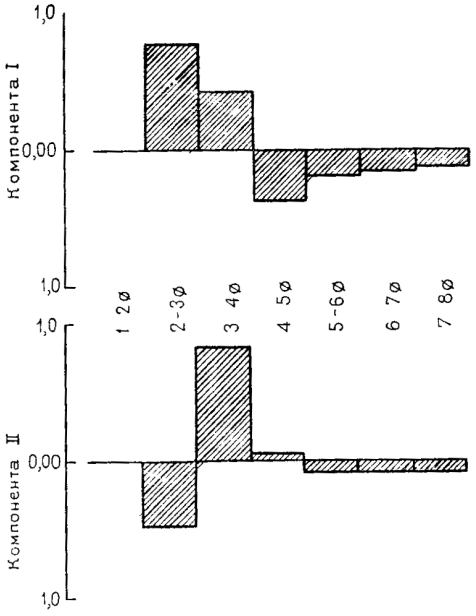


Рис. 6.29. Нагрузки переменных на первые две главные компоненты по данным изучения осадков из залива Баратария

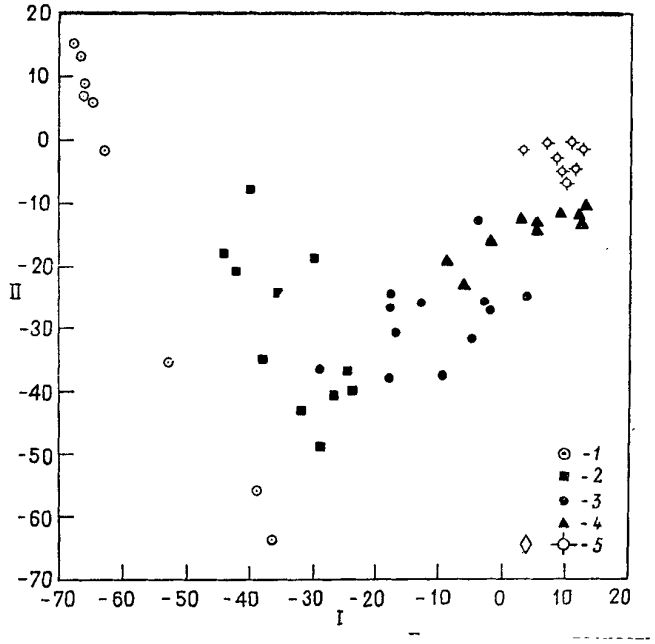


Рис. 6.30. Проекции анализов осадков залива Баратария на плоскость первых двух главных компонент. Различные символы соответствуют пяти различным типам осадочных пород [37]: 1 – пляжевые и прибрежные пески; 2 – илистые русловые пески; 3 – илистые береговые пески; 4 – органические донные илы; 5 – органические илы с подветренной стороны островов

Мы можем проверить результаты нашего анализа путем построения проекций наблюдений на первые две главные компоненты (рис. 6.30). Сравните различие между типами осадочных пород на рис. 6.30 и на рис. 6.31, на котором показана зависимость медианы размеров зерен от коэффициента отсортирован-ности (квартильное отклонение). Вероятно, что еще больший интерес представляет рис. 6.32, где изображено отношение содержания мелкой и очень мелкой песчаных фракций. Все эти диаграммы имеют приблизительно одинаковую эффективность с точки зрения разделения пяти типов осадков, хотя для построения диаграммы рис. 6.32 требуется больше экспериментальных данных, чем для построения рис. 6.31. Таким образом, для разделения образцов на семь разных классов по размеру достаточно только двух операций просеивания. Кроме того, результаты анализа с использованием *МГК* показывают, что осадки в бассейне можно рассматривать как смесь двух типов: песка и илисто-глинистой фракции. В этом примере метод главных компонент заставляет не только по-новому оценить состав изучаемых осадков, но и внести в методику исследования необходимые изменения, позволяющие значительно сократить расходы при минимальной потере информации. Такой эксперимент с незначительными изменениями был проведен Девисом [13]. Полезно сравнить эти результаты с результатами, полученными для тех же данных Клованом [35] с помощью Q-метода факторного анализа.

Интересно провести сравнение относительной эффективности среднего, первой главной компоненты и содержания песчаной фракции для различения пяти типов осадочных пород в заливе. Это можно сделать, применив однофакторный дисперсионный анализ к группам, образованным пятью типами осадочных пород. Отношение сумм квадратов между группами к общей сумме квадратов служит мерой того, насколько сильно группы связаны или отдалены друг от друга. Переменная, дающая наибольшее отношение SS_A/SS_T, оказывается наиболее эффективной для различения типов осадочных пород. Используя соответствующий из критериев **ANOVA**, указанных в гл. 2, определите, которая из трех переменных является наиболее эффективной.

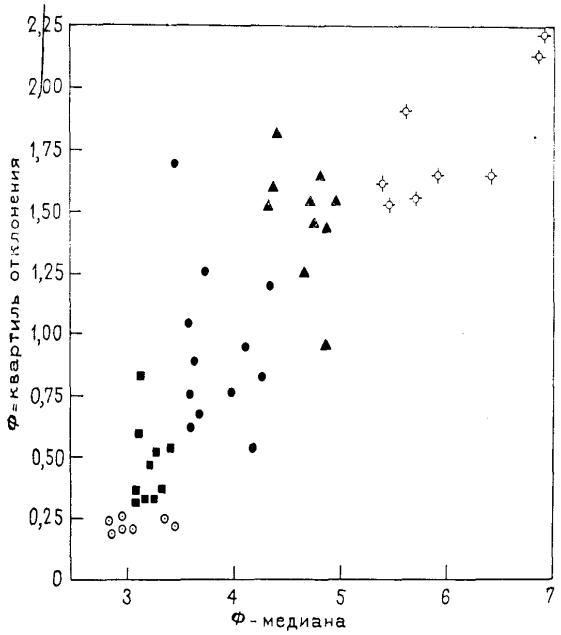


Рис. 6.31. Диаграмма зависимости квартиля от медианы в осадках залива Баратария в ϕ-единицах. Символы те же, что и на рис. 6.30

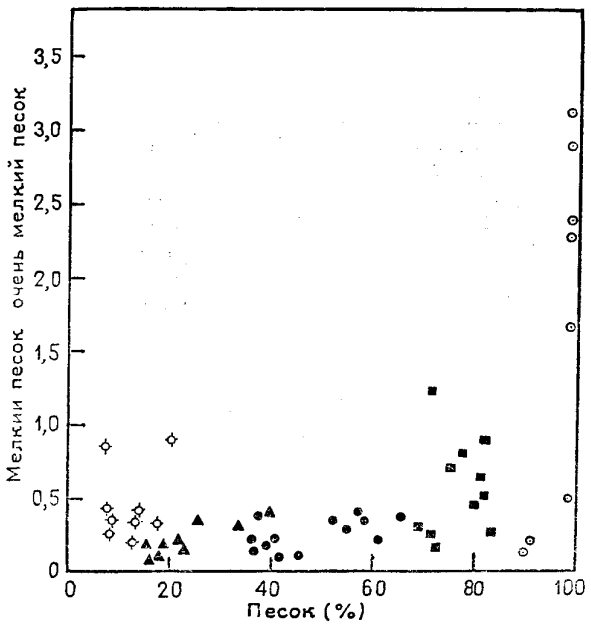


Рис. 6.32. Диаграмма зависимости содержания песка от отношения содержания тонкозернистого песка к очень тонкозернистому в осадках залива Баратария. Символы те же, что и на рис. 6.30

Возможны и другие методы исследования, которые подтверждают пользу набора статистик и квазистатистик, применяемых для характеристики осадочных отложений. Вычислив эти статистики, можно использовать их в качестве переменных в *МГК*, а также выбрать из них те комбинации, интерпретация которых позволит получить эффективную характеристику осадков. Определения различных статистик, применяемых при изучении размеров зерен, приведены во многих справочных изданиях по осадочной петрологии, например в работе Фолка [18]. Эти характеристики можно вычислить для необработанных данных, приведенных в табл. 6.22. Анализ главных компонент новых переменных оказывается очень поучительным. Аналогичные исследования были проведены Гриффитсом и Ондриком [21].

R-МЕТОД ФАКТОРНОГО АНАЛИЗА

Основой метода главных компонент является линейное преобразование т исходных переменных в т новых переменных,. где каждая новая переменная – линейная комбинация старых. Этот процесс осуществляется таким образом, чтобы каждая новая переменная давала возможно больший вклад в суммарную дисперсию. При вычислении новых переменных учитываются все исходные дисперсии. Так как *МГК*, вообще говоря, не статистический метод, мы ничего не можем сказать о вероятности, связанной с проверкой гипотез. Это просто математический метод. Однако при принятии решений об отбрасывании некоторых переменных или компонент, дающих очень малый вклад в суммарную дисперсию, приходится использовать некоторые статистические критерии, несмотря на то что последние имеют сильные ограничения и редко применимы (обзор этих критериев приводит Моррисон [51]. Метод главных компонент, как и анализ групп, относится к той категории методов, о пригодности которых судят после их применения, а не на основании теоретических рассуждений.

Факторный анализ, который принято считать статистическим методом, несколько отличается от этих методов, так как в его основе лежат некоторые предположения о природе изучаемой совокупности. Эти предположения позволяют указать те операции, которые должны быть выполнены, а также путь, по которому надо следовать при интерпретации результатов. Для некоторых процедур факторного анализа созданы даже критерии значимости [40], но они редко используются.

В факторном анализе предполагается, что связь между m переменными можно считать от-

ражением корреляционной зависимости каждой из переменных с p взаимно некоррелированными факторами. Обычное допущение состоит в том, что $p<m$. Поэтому дисперсию для m переменных можно вычислить с помощью дисперсии p-факторов плюс вклад, происхождение которого одинаково для всех m исходных переменных. В факторном анализе p независимых факторов носят название *общих факторов*, а независимая от них суммарная добавка обычно называется *фактором специфичности*. Факторная модель выражается в следующем виде:

$$X_j = \Sigma a_{jr} f_r + \varepsilon_j \qquad (6.51)$$

где f_r – r-й общий фактор; p – заранее заданное число факторов и ε – случайная компонента, присущая исходной переменной X_j. Так как имеется m исходных переменных X_j, то существует и m случайных переменных ε_j; рассматриваемые вместе, они составляют вектор факторов специфичности. Коэффициент a_{jr} называется нагрузкой j-й переменной на r-й фактор. В компонентном анализе этому понятию соответствуют нагрузки или веса на главные компоненты.

Предположим, что переменные X_j имеют многомерное нормальное распределение. Дисперсии и ковариации образуют матрицу порядка $m \times m$. Из формулы (6.51) вытекает, что диагональные элементы этой матрицы – дисперсии m переменных – можно выразить формулой

$$S_{jj}^2 = \sum_{r=1}^{p} a_{jr}^2 + \mathrm{var}\,\varepsilon_{jj} \qquad (6.52)$$

а недиагональные элементы, или ковариации, имеют вид

$$\mathrm{cov}_{jk} = \sum_{r=1}^{p} a_{jr} a_{kr} \qquad (6.53)$$

Основную гипотезу факторного анализа в матричной форме можно сформулировать следующим образом. Наблюдаемая ковариационная матрица, которую мы обозначим через $[s^2]$, является произведением матрицы порядка $m \times p$ факторных нагрузок (которую мы обозначим $[A^R]$ и ее транспозиции плюс диагональная матрица порядка $m \times m$ дисперсий факторов специфичности $[\mathrm{var}\varepsilon_{jj}]$:

$$[s^2] = [A^R]\,[A^R]' + [\mathrm{var}\,\varepsilon_{jj}] \qquad (6.54)$$

В результате умножения матрицы порядка $m \times p$ на ее транспонированную получим матрицу порядка $m \times m$, которая, однако, будет иметь только p положительных собственных значений и соответствующих им собственных векторов. Если $p=m$, то матрица$[\mathrm{var}\varepsilon_{jj}]$ оказывается тождественной и наша задача в точности эквивалентна *МГК*. В тех случаях, когда $p<m$, мы должны оценить матрицу параметров $[A^R]$, т.е. матрицу факторных нагрузок, и дисперсии факторов специфичности, т.е. матрицу $[\mathrm{var}\varepsilon_{jj}]$. Отметим, что в факторном анализе предполагается, что число факторов p известно до анализа, так как исследователь, исходя из некоторых предварительных рассуждений, в состоянии предсказать число факторов, от которых зависит изучаемая модель. Если число факторов p заранее предсказать нельзя, то разделение дисперсий между общими факторами и фактором специфичности становится неопределенным. Этот важный момент иногда остается незамеченным экспериментаторами, которые пытаются использовать факторный анализ для «ловли рыбы». Определенное другим способом число факторов p, $[A^R]$, матрица факторных нагрузок и дисперсии специфических факторов $[\mathrm{var}\varepsilon_{jj}]$ оказываются взаимосвязанными. Их нельзя оценить одновременно, поэтому для нахождения единственного решения необходимо вводить различные ограничения. Простейшее из них–это предположить число факторов равным некоторому априори заданному числу p. К сожалению, в большинстве геологических задач число факторов. неизвестно заранее и может быть даже важным объектом исследования. Другой путь – задать границу либо для $[A^R]\,[A^R]'$, либо для $[\mathrm{var}\varepsilon_{jj}]$ и затем извлекать факторы до тех пор, пока этот предел не будет достигнут.

Мы будем исследовать две из многих схем факторного анализа, начав с нахождения собственных значений и собственных векторов корреляционной матрицы и затем отбрасывая менее важные из них. Это не приводит к «истинному» факторному решению, однако математика слишком прямолинейна, и это приближение используется всюду в науках о Земле, в которых применяется факторный анализ. Мы приведем также краткий обзор метода максимального правдоподобия, который дает «истинные» факторы. К сожалению, соответствующие математические процедуры слишком сложны, чтобы их здесь описывать.

Хотя первый из рассматриваемых методов факторного анализа использует главные компоненты, все же вычисление собственных значений и собственных векторов в этом случае различается с двух точек зрения. Во-первых, собственные значения вычисляются для стандартизованной ковариационной, или корреляционной, матрицы. Это предполагает не только то, что все переменные имеют одинаковые веса, но также позволяет нам считать векторы главных компонент факторами. Во-вторых, собственные векторы, вычисленные в так называемой нормализованной форме, преобразуются так, что они определяют векторы, длины которых пропорциональны тем вариациям, которые они представляют. Эти преобразованные собственные векторы являются факторами множества данных.

Преобразование нормализованных, или единичных, собственных векторов в факторы не нарушает направлений векторов и не изменяет их длины. Это делается умножением каждого элемента нормализованного собственного вектора на соответствующее сингулярное значение, т.е. квадратный корень из соответствующего собственного значения. Полученный фактор есть вектор, взвешенный пропорционально вкладу в общую дисперсию, которую он представляет.

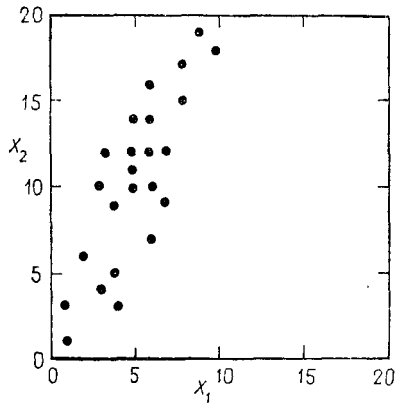


Рис. 6.33. Множество данных по стандартизации. Необработанные данные имеют средние значения $\overline{X}_1 = 5$ **и** $\overline{X}_2 = 10$

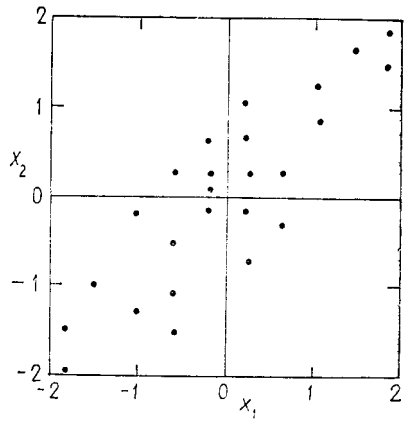


Рис. 6.34. Данные рис. 6.33 после стандартизации. Они имеют нулевые средние значения и единичное стандартное отклонение. Отметим, что области значений обеих переменных одинаковы

Влияние стандартизации обсуждалось в одном из предшествующих разделов, посвященных методу главных компонент. Здесь мы продемонстрируем его, используя данные, представленные на рис. 6.33. Необработанная ковариационная матрица имеет вид

$$\begin{bmatrix} 6{,}08 & 11{,}08 \\ 11{,}08 & 27{,}54 \end{bmatrix}$$

и следующие собственные векторы и собственные значения

собственный вектор $I = \begin{bmatrix} 0{,}39 \\ 0{,}92 \end{bmatrix}$, собственное значение 32,23, или 96%;

собственный вектор $II = \begin{bmatrix} 0{,}92 \\ -0{,}39 \end{bmatrix}$, собственное значение 1,39, или 4%.

Если эти данные стандартизовать, вычитая средние и осуществляя деление на стандартизованные отклонения (рис. 6.34), то ковариационная (или корреляционная) матрица стандартизированных данных будет иметь вид

$$\begin{bmatrix} 1{,}00 & 0{,}86 \\ 0{,}86 & 1{,}00 \end{bmatrix}$$

Ее собственные векторы и собственные значения:

собственный вектор $I = \begin{bmatrix} 0{,}707 \\ 0{,}707 \end{bmatrix}$, собственное значение 1,86, или 93%;

собственный вектор $II = \begin{bmatrix} -0,707 \\ 0,707 \end{bmatrix}$ собственное значение 0,14, или 7%.

Матрица собственных векторов [U]

$$U = \begin{bmatrix} 0,707 & -0,707 \\ 0,707 & 0,707 \end{bmatrix}$$

и собственные значения, или, скорее, квадратные корни из них, определяют матрицу сингулярных значений

$$[\Lambda] = \begin{bmatrix} \sqrt{1,86} & 0,0 \\ 0,0 & \sqrt{0,14} \end{bmatrix} = \begin{bmatrix} 1,36 & 0,0 \\ 0,0 & 0,37 \end{bmatrix}$$

Собственные значения по формуле (6.42) можно преобразовать в факторы $[A^R] = [U]\,[A]$.

Преобразование первого собственного вектора в факторы дает

$$\text{фактор } I = \begin{bmatrix} 0,707 & 1,86 \\ 0,707 & 1,86 \end{bmatrix} = \begin{bmatrix} 0,964 \\ 0,964 \end{bmatrix}$$

Компоненты фактора носят название *факторных нагрузок*. Если преобразование выполнено правильно, то сумма квадратов факторных нагрузок будет равна собственному значению:

$$(0,964)^2 + (0,964)^2 = 1,86$$

$$\text{фактор } II = \begin{bmatrix} -0,707 & 0,14 \\ 0,707 & 0,14 \end{bmatrix} = \begin{bmatrix} -0,264 \\ 0,264 \end{bmatrix}$$

Сумма квадратов факторных нагрузок также равна собственному значению:

$$(-0,264)^2 + (0,264)^2 = 0,14.$$

Эти два фактора графически представлены на рис. 6.35.

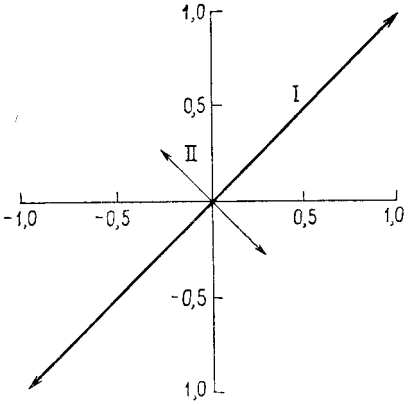


Рис. 6.35. Графическое представление двух факторов для двумерных данных, изображенных на рис. 6.34

Ориентации факторов такие же, как и исходных собственных векторов, а их длина равна квадратным корням из собственных значений. Собственные значения характеризуют относительный вклад каждого собственного вектора в суммарную дисперсию. Значения длины векторов, представляющих факторы, равны собственным значениям (точнее, квадратным корням из собственных значений), поэтому факторы отражают также дисперсии (или, более точно, стандартные отклонения). Так как собственные векторы сначала нормализовались, а затем умножались на квадратный корень из собственных значений, то каждая факторная нагрузка взвешена пропорционально квадратному корню из вклада в дисперсию, соответствующего данной переменной. В нашем примере первый фактор составляет примерно 1,86/2,00=93% общей дисперсии наших данных. Из них $(0,964)^2/1,86$=50% соответствует переменной 1 и $(0,964)^2/1,86$=50% – переменной 2. Аналогично второй фактор составляет 7% общей дисперсии, причем $(-0,264)^2/0,14$=50% соответствует переменной 1 и (0,264)2/0,14=50% – переменной 2. Сто процентов изменчивости переменной 1 распределяется на два фактора, так же как и 100% изменчивости переменной 2. (Взаимность влияния этих дисперсий на два фактора – неизбежное следствие работы с матрицей порядка 2×2. Это соотношение, вообще говоря, не имеет места для матриц более высоких порядков.)

Расположив факторные нагрузки в форме матрицы, мы получим матрицу факторных значе-

ний, которая для данных рис. 6.34 имеет вид

$$\text{Переменные}\begin{matrix}\\ 1\\ 2\end{matrix}\begin{matrix}\text{Факторы}\\ I \quad\quad II\end{matrix}\begin{bmatrix}0{,}964 & -0{,}264\\ 0{,}964 & 0{,}264\end{bmatrix}'$$

Если возвести в квадрат элементы матрицы факторных значений и произвести суммирование по каждой переменной, то суммы вкладов переменных в факторы будут одинаковыми, т.е.

$$\text{Переменные}\begin{matrix}\\ 1\\ 2\end{matrix}\begin{matrix}\text{Факторы}\\ I \quad\quad II\end{matrix}\begin{bmatrix}0{,}964 & -0{,}264\\ 0{,}964 & 0{,}264\end{bmatrix}'=\begin{matrix}\text{Суммы}\\ \\ \end{matrix}\begin{bmatrix}1{,}00\\ 1{,}00\end{bmatrix}$$

Эти суммы называются *общностями*, и их принято обозначать h_j^2, где j – номер переменной. Если определить m факторов из ковариационной матрицы порядка $m \times m$, то их общности будут равны исходным дисперсиям. Так как мы используем стандартизированные переменные, то эти общности равны 1,00. Однако если извлечь меньше факторов, то общности будут меньше исходных дисперсий, и мы получим показатель эффективности приведенного множества факторов. Например, если сохранить только один фактор из матрицы порядка 2×2, то общности будут равны $h_1^2 = (0{,}964)^2 = 0{,}93$, $h_2^2 = (0{,}964)^2 = 0{,}93$. Таким образом, сохранение только одного фактора позволяет учесть 93% дисперсии первой переменной и 93% дисперсии второй переменной.

Так как значения общностей зависят от числа сохраняемых факторов, то вопрос о последних ставит нас перед лицом одной из важнейших задач факторного анализа – какое число факторов должно быть сохранено? К сожалению, на этот вопрос нет простого ответа, и его отсутствие является одним из самых серьезных возражений против факторного анализа. Психологи-экспериментаторы на ранней стадии развития факторного анализа решали эту задачу прямолинейно: они определяли столько факторов, сколько требовала проверяемая ими теория. Другой столь же приближенный способ состоит в том, что определяется только два или три фактора, так как это максимальное число, которое можно изобразить графически на диаграмме рассеяния, и любое увеличение числа факторов ведет к увеличению размерности пространства, в котором решается поставленная задача, в результате чего ее трудности заметно возрастают.

Некоторые исследователи советуют сохранять столько факторов, сколько имеется собственных чисел, больших единицы. Иными словами, сохраняются все факторы, которые дают больший вклад в дисперсию, чем исходные стандартизированные переменные. В большинстве примеров лишь немногие факторы содержат большую часть вклада в дисперсию множества данных, и эта рекомендация оказывается полезной. Однако если исходные переменные оказываются слабо коррелированными или некоррелированными, половина или более факторов может иметь собственные значения, большие единицы. В результате не только получается очень большое число факторов, но и возможно, что ни один из них не допускает никакой интерпретации. Если данные таковы, что факторный анализ к ним применим (т.е. наблюдаемые дисперсии возникли благодаря корреляции между переменными и рассматриваемыми факторами), то лишь некоторые факторы дают большой процентный вклад в суммарную дисперсию и общности имеют высокие значения. Если для того чтобы учесть большую часть исходной дисперсии, требуется сохранение большого числа факторов или если значения общностей нескольких первых факторов низкие, то факторная модель чаще всего оказывается неподходящей.

Прежде чем переходить к рассмотрению следующей процедуры факторного анализа, а именно вращению факторных осей для получения «простой структуры», применим изложенные выше методы к рассмотренному примеру. Мы используем данные табл. 6.18 и сохраним два фактора, так как интуиция подсказывает нам, что в этом случае требуется только два фактора, а именно факторы размера и формы. Матрица стандартизированных дисперсий и ковариаций приведена в табл. 6.25. В табл. 6.26 дана матрица собственных векторов или главных компонент и приведены также соответствующие им собственные значения. Мы сохраним только первые два из них и преобразуем их в факторы. С этой целью умножим нормализованные собственные векторы на квадратный

корень из соответствующих собственных значений, в результате чего получим факторные нагрузки. Матрица факторных нагрузок $[A^R]$, имеющая в действительности порядок $m \times p$, с целью экономии места представлена в следующем виде:

$$\begin{matrix}\text{фактор I} \\ \text{факторII}\end{matrix}\begin{bmatrix} 0{,}747 & 0{,}795 & 0{,}710 & 0{,}910 & -0{,}235 & -0{,}178 & -0{,}886 \\ 0{,}491 & 0{,}373 & -0{,}596 & 0{,}389 & 0{,}963 & 0{,}971 & 0{,}218 \end{bmatrix}$$

Вектор общностей по всем переменным имеет вид

$$h_j^2 = \begin{bmatrix} 0{,}798 & 0{,}771 & 0{,}860 & 0{,}979 & 0{,}983 & 0{,}976 & 0{,}833 \end{bmatrix}$$

Остаточная дисперсия j-й переменной ($1{,}00 - h_j^2$) – единственная компонента, ассоциированная с этой переменной. Составляющие этой компоненты таковы:

$$[\operatorname{var}\varepsilon_{jj}] = \begin{bmatrix} 0{,}202 & 0{,}229 & 0{,}140 & 0{,}021 & 0{,}017 & 0{,}024 & 0{,}167 \end{bmatrix}$$

Если для множества m переменных приходится сохранять m факторов, то исходную ковариационную матрицу $[s^2]$ можно восстановить с помощью перемножения всевозможных пар факторных нагрузок и суммирования по факторам. Конечно, если сохраняется $p<m$ факторов, то исходную ковариационную матрицу точно воспроизвести нельзя. Получаемая таким образом оценка ковариации переменных j и k имеет вид

$$\bar{s}_{jk}^2 = a_{j1}a_{k1} + a_{j2}a_{k2} + \ldots + a_{jp}a_{kp} \qquad (6.55)$$

где a_{j1} – нагрузка j-й переменной на фактор i. Если $[A^R]$ – матрица факторных нагрузок, то эквивалентная матричная запись этой формулы имеет вид

$$[\bar{s}^2] = [A^R][A^R]'$$

если считать, что векторы факторных нагрузок являются столбцами матрицы. Стандартизированная матрица остаточных ковариации (или остаточная корреляционная матрица) находится как разность двух матриц

$$[\bar{s}_{res}^2] = [s^2] - [\bar{s}^2] \qquad (6.56)$$

Воспроизведенная и остаточная корреляционные матрицы для нашего примера даны в табл. 6.27. Остаточная матрица является мерой неспособности этих факторов учесть изменчивость исходного множества данных.

Таблица 6.25 Стандартизированные дисперсии и коэффициенты корреляции для семи переменных, измеренных на 25 параллелепипедах, указанных в табл. 6.18 (приведена только нижняя половина симметричной матрицы)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1,000						
X_2	0,580	1,000					
X_3	0,201	0,364	1,000				
X_4	0,911	0,834	0,439	1,000			
X_5	0,283	0,166	–0,704	0,163	1,000		
X_6	0,287	0,261	–0,681	0,202	0 990	1,000	
X_7	–0,533	–0,609	–0,649	–0,676	0,427	0,357	1,000

Таблица 6.26. Собственые значения и матрица собственных векторов для данных табл. 6.25 (сохранены лишь два собственных вектора с собственными значениями, превосходящими 1,000)

Вектор	Собственное значение	Вклад в дисперсию, %	Сумма вкладов в дисперсию, %
I	3,3946	48,4949	48,4949
II	2,8055	40,0783	88,5731
III	0,4373	6,2473	94,8204
V	0,2779	3.9707	98 7911
V	0,0810	1,1565	99,9476
VI	0,0034	0,0487	99,9963
VII	0,0003	0,0037	100,0000

Переменная	Собственный вектор						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
X_1	0,4053	–0,2929	–0,6674	0,0888	–0,2267	0,4088	–0,2782
X_2	0,4316	0,2224	0,6980	–0,0338	–0,4366	0,1443	–0,2540
X_3	0,3854	0,3559	0,1477	0,6276	0,5121	0,1875	–0,1081
X_4	0,4939	–0,2323	–0,1186	0,2103	–0,1054	–0,5878	0,5359
X_5	–0,1277	–0,5751	0,0294	0,1108	0,3890	0,1949	0,5562
X_6	–0,0968	–0,5800	0,1743	–0,0061	0,3549	0,5003	0,4975
X_7	–0,4809	–0,1303	0,0176	0,7353	–0,4553	0,0332	0,0489

Таблица 6.27. Корреляционная матрица для двух факторов, полученных по данным о параллелепипедах, и остаточная корреляционная матрица, в которой содержатся коэффициенты корреляции, неучтенные первыми двумя факторами (приведены лишь нижние половины симметричных матриц)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Корреляционная матрица							
X_1	0,7983						
X_2	0,7766	0,7711					
X_3	0,2379	0,3426	0,8596				
X_4	0,8704	0,8685	0,4143	0,9794			
X_5	0,2969	0,1718	–0,7413	0,1606	0,9833		
X_6	0,3434	0,2201	–0,7057	0,2157	0,9778	0,9756	
X_7	–0,5546	–0,6233	–0,7594	–0,7214	0,4187	0,3701	0,8328
Остаточная корреляционная матрица							
X_1	0,2017						
X_2	–0,1963	0,2289					
X_3	–0,0367	0,0212	0,1404				
X_4	0,0409	–0,0348	0,0243	0,0206			
X_5	–0,0135	–0,0060	0,0371	0,0024	0,0167		
X_6	–0,0569	0,0409	0,0252	–0,0134	0,0124	0,0244	
X_7	0,0214	0,0146	0,1105	0,0159	0,0085	–0,0129	0,1672

Вращение факторов

Несмотря на то что факторный анализ позволяет уменьшить число измерений в задаче, дать содержательную интерпретацию факторов бывает не очень легко. Возможно, это результат того, что положение p ортогональных факторных осей в m-мерном пространстве определяется положением $(m–p)$ ненужных ортогональных осей в выборочном пространстве. Однако для описания наших данных достаточно только p факторных осей. Если исключить из рассмотрения излишние ортогональные оси, то оставшиеся факторные оси можно подвергнуть дополнительному вращению, которое может помочь в нахождении наилучшего их расположения. Для этой цели можно использовать разнообразные схемы вращения. Мы будем использовать так называемый *варимакс Кайзера*, основой которого является изменение положения факторных осей до тех пор, пока проекции каждой переменной на факторные оси не окажутся близкими либо к нулю, либо к их максимальным значениям. Иными словами, в результате применения этого метода факторные нагрузки оказываются близкими либо к нулю, либо $к\pm1$. Обычно для каждого фактора мы получаем немного довольно больших значений факторных нагрузок и много незначимых нагрузок. В этом случае интерпретация в терминах исходных переменных проводится легко. Однако имеются случаи, когда вращение факторных осей не облегчает анализа и даже приводит к дальнейшему ухудшению результатов. Это связано либо с взаимной коррелированностью факторов, или указывает на то, что выбранная факторная модель оказывается плохой.

Метод варимакс сводится к максимизации дисперсии нагрузок на факторы. Определим дисперсию s_k^2 нагрузок на k-й фактор как

$$s_k^2 = \frac{1}{p^2}\left[p\sum_{j=1}^{m}\left(\frac{a_{jp}^2}{h_j^2}\right)^2 - \left(\sum_{j=1}^{m}\frac{a_{jp}^2}{h_j^2}\right)^2 \right] \qquad (6.57)$$

где, как и раньше, p – число факторов; m – число исходных переменных; a_{jp} – нагрузка j-й переменной на p-й фактор; h_j^2 – общность j-й переменной. Функция, которую мы хотим максимизировать, имеет вид

$$V = \sum_{k=1}^{p} s_k^2 \qquad (6.58)$$

Дисперсия вычисляется для факторных нагрузок a_{jp}, деленных на соответствующие общности h_j^2. Иными словами, рассматривается только общая часть дисперсии по каждой переменной и отбрасывается ее часть, соответствующая $m–p$ компонентам, необходимым для учета всех дисперсий каждой переменной. Максимизация дисперсии приводит к увеличению интервала изменения факторных нагрузок, которые для того, чтобы удовлетворить требованиям метода Кайзера, стремятся либо к своему экстремальному (положительному или отрицательному) значению, либо к нулю.

Никакой простой аналитической схемы для метода варимакс не существует. Обычно вращение факторных осей проводится итеративным методом. Вращению подвергаются две оси, в то время как другие оси остаются неподвижными. После того как все оси подвергнуты вращению, процесс повторяется снова до тех пор, пока уменьшение дисперсии нагрузок на каждом шаге не станет ниже некоторого заранее заданного уровня.

Этот метод вращения лучше всего проиллюстрировать на примере. Мы сделаем попытку «почистить» факторы, полученные для наших искусственно взятых данных по параллелепипедам, применяя метод вращения к двум оставленным факторам. На рис. 6.36 нагрузки семи исходных переменных на фактор *I* нанесены по отношению к нагрузкам на фактор *II*. Связывая построенные точки с началом координат, получаем диаграмму, на которой факторные нагрузки представлены как векторы. Ориентация векторов по отношению к факторным осям отражает степень их корреляции с факторами. Длина каждого вектора пропорциональна общности переменной, которую этот вектор представляет. Если два фактора нанесены с полным учетом всех вариаций исходной переменной, то общность переменной равна единице и на диаграмме она будет лежать на окружности единичного радиуса. В этом примере все общности высоки, поэтому все векторы, представляющие семь исходных переменных, лежат вблизи от единичной окружности.

Варимаксное вращение изменяет факторные нагрузки, поэтому исходные переменные имеют либо высокую положительную, либо отрицательную корреляцию (приблизительно ±1) с некото-

рым фактором или корреляцию, близкую к нулю. На рис. 6.37 представлены положения факторных осей после вращения. Заметим, что относительное положение переменных не изменяется при вращении; изменяются только их положения по отношению к факторным осям. Также заметим, что длины векторов остаются неизменными.

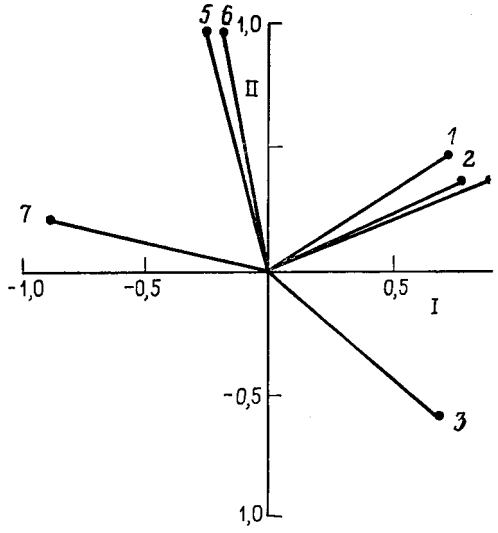


Рис. 6.36. Графическое представление нагрузок на два фактора для необработанных данных по 25 случайным блокам. Исходные данные для семи переменных приведены в табл. 6.18

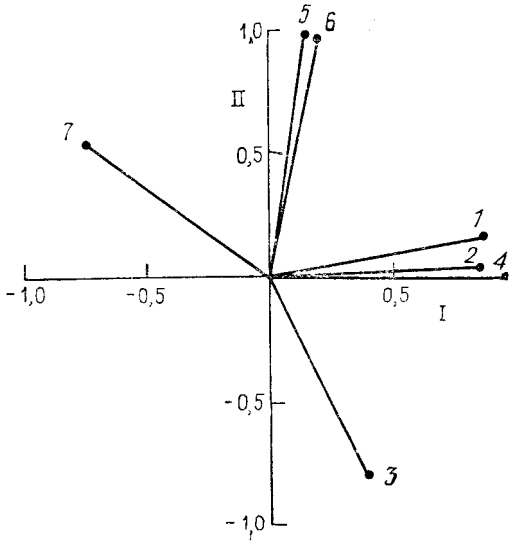


Рис. 6.37. Графически представление тех же нагрузок после вращения по методу варимакс. Использованы данные по 25 блокам, приведенные в табл. 6.18

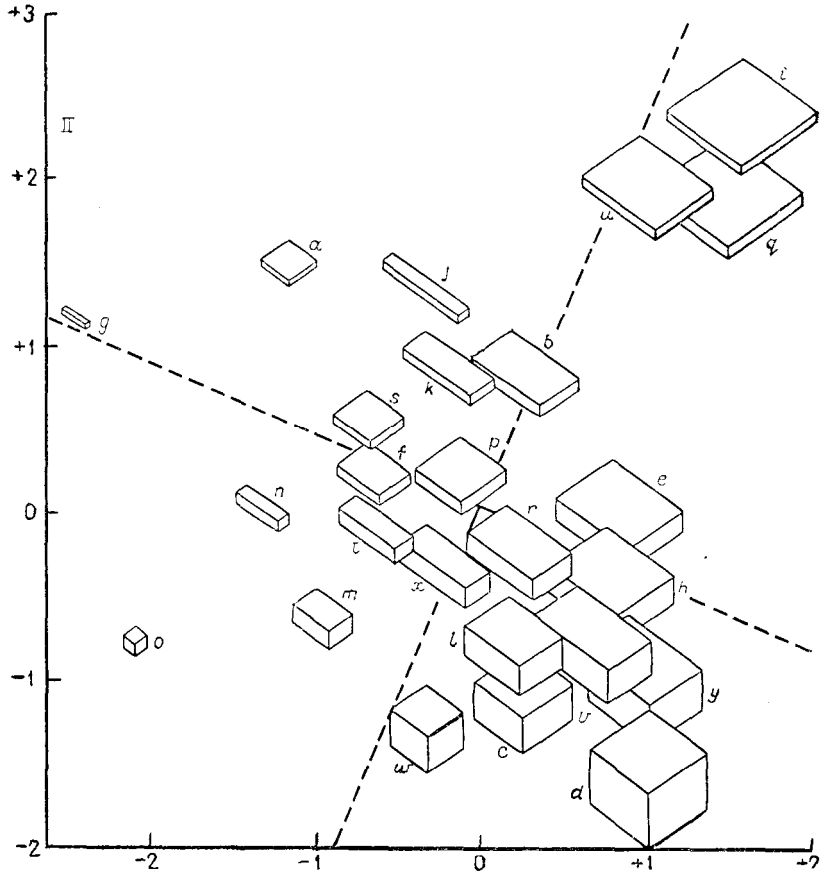


Рис. 6.38. Графическое представление факторных меток данных по случайным блокам для первых двух факторов (*I* и *II*). Указано положение параллелепипедов по отношению к двум факторам. Пунктирной линией представлены факторные оси после вращения по методу варимакс

Соотношения между самими переменными также не изменяются при вращении, хотя положение индивидуальных объектов в пространстве, определенном факторными осями, изменяется. На рис. 6.38 нанесены факторные метки, аналогично меткам главных компонент, представленным на рис. 6.27. Заметим, что два множества факторных осей представлены на диаграмме, одно для факторных меток до вращения и другое – для меток после варимаксного вращения. Оказывается, что первый фактор на самом деле отражает все размеры блоков, так что меньшие блоки располагаются слева, а большие – справа. Второй фактор разделяет формы одинакового размера на верхушке с уплощениями и удлинениями ниже относительно второго фактора. В этом случае варимаксное вращение, возможно, не даст вклад в нашу интерпретацию факторов. Никакая из схем факторных меток сильно не отличается от полученных методом главных компонент, хотя относительная важность первой и второй факторных осей меняются местами по сравнению с осями метода главных компонент.

Графическое изображение факторных меток (до вращения или после него) более сложно, чем в методе главных компонент. Главные компоненты получаются в результате применения линейных преобразований, в то время как факторные значения представляют оценки вкладов различных факторов в каждое исходное наблюдение. Так как факторы находятся по тем же данным, вычисление факторных нагрузок – в некотором роде циклический процесс, и потому результаты могут оказаться неоднозначными. Одно из наилучших изложений этого вопроса принадлежит Моррисону [51] (см. также процедуры, описываемые Харманом [22]). В психометрии факторы обычно представляют самостоятельный интерес, а факторные метки совсем не используются, поэтому вычислению факторных меток до сих пор уделялось мало внимания. Вероятно, факторные метки могут играть значительную роль в применениях факторного анализа в геологии и потому важно уметь их вычислять.

Предположим, что наши исходные данные представлены в виде матрицы $[X]$ порядка $n \times m$, где n – число строк, или наблюдений, а m – число столбцов, или переменных. По аналогии с *МГК* можно вычислить матрицу факторных меток $[S^R]$, умножая матрицу исходных данных на матрицу факторных нагрузок $[A^R]$, т.е. выполняя операцию $[X]\,[A^R] = [S^R]$. Если мы сохраняем p факторов, то матрица нагрузок $[A^R]$ будет матрицей порядка $m \times p$, а матрица факторных значений $[S^R]$ будет иметь порядок $n \times p$. Напомним, однако, что исходные данные представлены не только с помощью факторов, но и с помощью специфической переменной (см. формулу 6.51), поэтому матрица факторных значений, вычисленная таким образом, частично отражает ковариационную структуру заданного набора m переменных, а также структуру p факторов. Действительно, для того чтобы получить истинные факторные значения, специфическую компоненту исходных переменных необходимо отделить. Это делается с помощью умножения указанного уравнения на матрицу, обратную матрице ковариаций:

$$[X] \cdot [s^2]^{-1} \cdot [A^R] = [S^R] \qquad (6.59)$$

Обратная матрица имеет порядок $m \times m$, а матрица факторных нагрузок – порядок $m \times p$, так что матрица «истинных» факторных значений $[S^R]$ будет порядка $n \times p$, что и следовало ожидать. В результате выполнения этой операции мы получаем факторные метки, свободные от специфической компоненты; имеющейся в каждом из исходных наблюдений.

Несмотря на простоту, этот метод нахождения факторных меток не используется на практике. Матрица $[s^2]$ может быть очень большой, в особенности в Q-методе факторного анализа, на котором мы остановимся ниже, и ее обращение может оказаться очень сложным. Однако, используя алгебраические соотношения, можно обратить матрицу ковариаций порядка $p \times p$, построенную по факторам, и в результате получить тот же результат. Обычно p бывает значительно меньше m, что позволяет упростить вычисления, хотя число матричных преобразований при этом возрастает.

Вычислим сначала матрицу $[S]$, умножая матрицу факторных нагрузок на ее транспозицию

$$[A^R]' \cdot [A^R] = [S] \qquad (6.60)$$

Так как транспонированная матрица $[A^R]'$ имеет порядок $p \times m$, то в результате умножения получится квадратная матрица порядка $p \times p$. Эта матрица обращается и умножается на матрицу факторных нагрузок, что дает нам некоторую вспомогательную матрицу $[B]$:

$$[A^R] \cdot [S]^{-1} = [B] \qquad (6.61)$$

которая затем используется для вычислений истинной матрицы факторных значений по формуле

$$[X]\cdot[B]=[\overline{S}^{R}] \qquad (6.62)$$

Последовательность операций, приведенных в формулах (6.60)–(6.62), может быть представлена в терминах матрицы факторных нагрузок $[A^R]$:

$$\begin{aligned}[X]\cdot[B]&=[\overline{S}^{R}];\\ [X]\cdot[A^R]\cdot[S]^{-1}&=[\overline{S}^{R}];\\ [X]\cdot[A^R]\cdot\left([A^R]'\cdot[A^R]\right)^{-1}&=[\overline{S}^{R}]\end{aligned} \qquad (6.63)$$

Та же процедура используется для получения проекций факторных значений на факторные оси до или после выполнения вращения по методу Кайзера. Отметим, что в матрице данных $[X]$ представлены стандартизированные переменные, а не исходные, как в *МГК*. Это объясняется тем, что нагрузки в *МГК* вычисляются на основании ковариационной матрицы исходных данных, а факторные нагрузки – на основании стандартизированной, или корреляционной, матрицы. Конечно, если бы мы стандартизировали данные, используемые в *МГК*, то вычисляли бы факторные значения на главные оси для стандартизированных данных.

Теперь задача определения числа p сохраняемых факторов приобретает важное значение. Число факторов влияет на размеры воспроизведенной и остаточной корреляционной матриц, на общности и на нагрузки специфической компоненты. Сами факторные нагрузки при этом не изменяются. Это значит, что если по исходному множеству данных определить p равным 2, то факторные нагрузки *I* и *II* не изменятся при добавлении третьего фактора. Однако если подвергнуть вращению два фактора, то нагрузки могут сильно отличаться от нагрузок, полученных в случае, если бы мы определили их по тем же данным и подвергли вращению фактора. Вращение двух факторов при p=2 не ограничено никакими условиями. Вращение этих двух факторов не может выполняться так же свободно при p=3, из-за того что третья ортогональная ось накладывает ограничения, которые также должны быть согласованы с базисом m-мерного пространства исходных переменных.

Схема вращения по методу Кайзера сохраняет ортогональность факторных осей. Несмотря на то что после вращения факторные оси не совпадают больше с главными осями эллипсоида ковариаций, они образуют прямые углы друг с другом и, следовательно, некоррелированы. Существуют также схемы вращения, в которых не требуется выполнение условия ортогональности, и факторные оси могут образовывать друг с другом углы, отличные от прямых. В некоторых случаях такие факторы могут быть лучше проинтерпретированы, так как часто нагрузки на них оказываются более высокими. Однако при использовании этих схем возникают некоторые трудности теоретического характера. Во-первых, факторная модель основана на допущении, что наблюдаемая ковариационная матрица является результатом корреляции между m переменными и p взаимно некоррелированными факторами. Ослабление условия ортогональности ведет к возникновению взаимной корреляции между факторами и, по-видимому, расширяет исходное множество допущений. Если факторы оказываются коррелированными, то в силу существования взаимодействия между парами переменных и парами факторов соотношения между факторами и исходными переменными оказываются значительно более сложными, чем предполагает модель. Наличие взаимной корреляции наводит на мысль и о том, что, вероятно, неортогональные факторы сами по себе являются не чем иным как результатом корреляции между некоторыми «суперфакторами», еще более глубоко скрытыми от прямого наблюдения.

Первоначально факторный анализ предназначался для объяснений взаимных связей между большим количеством переменных при небольшом числе факторов. Это сопровождалось теорией, которая позволяла предсказать природу факторов, что облегчало их интерпретацию. Однако когда факторный анализ стал применяться в областях, в которых заранее никакой теории не существовало, то оказалось необходимым объяснить смысл получаемых факторов. Это не всегда возможно, так как теоретические основы факторного анализа слишком мало развиты для того, чтобы позволить во всех случаях давать адекватное объяснение явления. Однако вместо того чтобы признать свое поражение, факторный анализ рекомендует неортогональные схемы вращения, которые позволяют выразить факторы в терминах исходных переменных. Таким образом, исследователь проходит полный цикл от переменных к факторам (с целью сжатия данных) и затем снова к переменным (для интерпретации факторов).

Сказанное нельзя считать подтверждением того, что методы неортогональных факторов бесполезны: используя их, в некоторых случаях удалось получить хорошие результаты. Однако если

обычные методы факторного анализа дают плохие результаты, новичок приходит к выводу о том, что они неприменимы к данной задаче или что о причинных связях между переменными известно слишком мало для того, чтобы дать интерпретацию факторов. Неортогональные решения вносят элемент субъективности в уже и без того довольно произвольное решение задачи, которого надо тщательно избегать всем, за исключением экспертов. Интересующийся читатель может обратиться к книге Хармана [22] и ранней работе Тэрстоуна [63, гл. 15].

Рассмотрим теперь альтернативный *R*-метод факторного анализа, или метод максимального правдоподобия, разработанный Лоули [39] и затем модифицированный многими исследователями. Он не касается проблем, возникающих в других факторных процедурах. Метод максимального правдоподобия преодолевает их благодаря введению некоторых начальных предположений о природе факторов и дисперсии специфического фактора. Факторы предполагаются нормально распределенными с нулевыми средними и единичными дисперсиями. Элементы матрицы дисперсий специфических факторов также предполагаются нормально распределенными с нулевым средним и дисперсией $[\mathrm{var}\ \varepsilon_{jj}]$. Все факторы и элементы специфической дисперсии далее предполагаются независимыми. Таким образом, наблюденная матрица дисперсий и ковариаций $[s^2]$ предполагается достаточной для оценки $[\Sigma]$, ненаблюдаемой матрицы дисперсий и ковариаций между факторами.

Получение оценок максимального правдоподобия факторных нагрузок требует математических выкладок все возрастающей сложности; в согласии с другими авторами мы проследим развитие этих методов. Интересующихся читателей отсылаем к Лоули и Максвелу [40] и Йореского [32]; очень понятное и короткое изложение приводит Моррисон [51]. Мы ограничимся исследованием соответствующих вычислительных процедур в том виде, как они представлены в общедоступных фондах вычислительных программ.

Метод максимального правдоподобия исходит из того же модельного уравнения, что и другие формы факторного анализа:

$$[s^2] = [A^R] \cdot [A^R]' + [\mathrm{var}\,\varepsilon_{jj}]$$

Однако оценки максимального правдоподобия для факторных нагрузок получаются итерационным методом. Чтобы пояснить шаги итераций, введем обозначение $[{}_i a_{jr}]$, означающее *i*-ю итерацию в аппроксимации специфической дисперсии *j*-й переменной левой части после извлечения *r*-го фактора.

Начальные оценки нагрузок на первый фактор $[{}_0 a_{j1}]$ основываются на элементах первого собственного вектора, извлеченного из наблюденной матрицы дисперсий и ковариаций $[s^2]$. Шкалы элементов собственного вектора выбраны таким образом, что сумма их квадратов равна первому собственному значению. Начальная аппроксимация дисперсии определяется по формуле

$$[{}_0\,\mathrm{var}\,\varepsilon_{jj_1}] = \mathrm{diag}([s^2] - [{}_0 a_{j1}] \cdot [{}_0 a_{j1}]') \qquad (6.64)$$

(оператор *diag* означает, что сохраняются только диагональные элементы матрицы).

Далее вычисляются элементы матрицы

$$([{}_0\,\mathrm{var}\,\varepsilon_{jj_1}])^{-1/2}([\mathrm{s}^2] - [{}_0\,\mathrm{var}\,\varepsilon_{jj_1}])[{}_0\,\mathrm{var}\,\varepsilon_{jj_1}]^{-1/2} \qquad (6.65)$$

и определяются первое собственное значение и первый собственный вектор. Собственный вектор снова формируется так, чтобы сумма квадратов его элементов равнялась собственному значению; их мы обозначим $[{}_i a_{j1}]$. Оценка матрицы факторных нагрузок в конце первой итерации определена формулой

$$[{}_1 A_1^R] = [{}_0\,\mathrm{var}\,\varepsilon_{jj_1}]^{-1/2}[{}_1 a_{j1}] \qquad (6.66)$$

Оценка дисперсии в конце первой итерации находится по формуле

$$[{}_0\,\mathrm{var}\,\varepsilon_{jj_1}] = \mathrm{diag}([s^2] - [{}_1 A_1^R]') \qquad (6.67)$$

Это аналогично начальной оценке дисперсии, определенной формулой (6.64). Этот процесс повторяют, используя новые оцененные значения $[\mathrm{var}\ \varepsilon_{jj1}]$. Итерации продолжают до тех пор, пока $[{}_i A_1{}^{\mathrm{R}}]$ и $[{}_{i+1} A_1{}^{\mathrm{R}}]$ будут отличаться не более чем на заданную малую величину. Столбцы вектора $[{}_i A_1{}^{\mathrm{R}}]$ – это оценки максимального правдоподобия нагрузок на первый фактор.

Гипотезу о том, что данные содержат только один фактор, так что

$$[s^2] = [A_1^R] \cdot [A_1^R]' + [\mathrm{var}\,\varepsilon_{jj_1}]$$

можно проверить с помощью критерия χ^2 методом, описанным Моррисоном [51]. Если же гипотезу отклоняют, то добавляют новые факторы. Их вычисляют методом итераций, аналогичным тому, которым вычисляли первый фактор. Отличие состоит в том, что процесс начинался с матрицы остатков $[s^2_{\text{res}}]$:

$$[s^2_{res}] = [s^2] - [_i A_1^R] \cdot [_i A_1^R]'$$

Из этой матрицы определяют первые два собственных значения и собственных вектора. Собственный вектор [0^/2] является начальной аппроксимацией второго фактора, который комбинируется с вектором [iA^] с целью образования матрицы

$$[_0 A_2^R] = [_i A_{1\ 0}^R a_{j2}]$$

С помощью этой начальной оценки двухфакторной матрицы нагрузок получен аналог уравнения (6.64). Процесс, который был использован для оценки нагрузок на первый фактор, теперь применен для получения нагрузок на второй фактор. После того как устойчивая оценка найдена, проверяется ее значимость. В случае положительного результата процесс повторяется с целью определения третьего фактора, затем четвертого и так далее.

Может случиться, что все систематические источники изменчивости будут определены, и процесс выделения факторов закончится.

Обобщение метода максимального правдоподобия выделения факторов очень напоминает аналогичный процесс факторного анализа, основанного на методе главных компонент. Факторы можно подвергнуть вращению с целью получения простой структуры и даже до косых положений в поисках их осмысленной интерпретации. Решения задачи априорного определения числа факторов p можно избежать, и факторы оказываются свободными от смещения, присущего факторам, определенным более простыми методами. К сожалению, основные аргументы критиков факторного анализа применимы и здесь. Анализ очень полезен в тех областях, где хорошо известны причинные отношения явлений. В геологии, однако, положение таково, что справедливые сегодня истины завтра оказываются дискредитированными, и факторные интерпретации, по-видимому, находятся не в лучшем положении. Скептически настроенному читателю можно рекомендовать познакомиться с критикой применения факторного анализа в геологии Темпла [62] и с более обоснованным отрицательным мнением Метеле и Рейхера [49], относящимся к применению факторного анализа в гидрогеологии.

Q-МЕТОД ФАКТОРНОГО АНАЛИЗА

Теперь обратимся к *Q*-методу факторного анализа, в котором основное внимание уделяется исключительно интерпретации отношений множества данных внутри объекта, а не отношений между переменными (или ковариаций), используемых в *Q*-методе факторного анализа. Тот факт, что эти два метода в сущности эквивалентны, не был замечен большинством исследователей, что привело к созданию крайне запутанных и очень сложных в вычислительном плане алгоритмов *Q*-метода.

Первый шаг *Q*-метода факторного анализа состоит в формировании матрицы сходства порядка $n \times n$. Коэффициент корреляции, однако, можно рассматривать как меру сходства, так как для его определения требуется вычисление дисперсий по переменным. Такая мера имеет очень непонятный смысл.

Наиболее широко применяемая мера сходства в *Q*-методе факторного анализа – это косинус θ, пропорциональный коэффициенту сходства

$$\cos\theta_{ij} = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \qquad (6.68)$$

Эта мера выражает сходство между объектом с номером i и объектом с номером j, если рассматривать каждый из них как вектор в m-мерном пространстве. Косинус θ – косинус угла между этими двумя векторами. Заметим, что это уравнение очень напоминает по форме коэффициент корреляции (см. формулу 2.24); если переменные стандартизованы так, что их средние равны нулю и стандартные отклонения равны единице, то две меры будут численно совпадать.

Косинус θ изменяется от 1,0 для двух объектов, векторные представления которых совпадают, до 0,0 для объектов, векторы которых образуют угол в 90°. Так как косинус θ измеряет только угловое сходство, то он чувствителен к относительным пропорциям переменных, а не к их абсолютным величинам. Если, например, измерения были сделаны на двух брахиоподах, которые идентичны по виду, но не по размеру, то мера сходства cos θ между ними будет равна 1,0.

Матрицу сходства порядка $n \times n$ удобнее всего строить, вычисляя cos θ в два шага [33]. Сначала каждый элемент строки в матрице данных делится на квадратный корень из суммы квадратов элементов этого столбца:

$$w_{jk} = \frac{x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{jk}^2}} \qquad (6.69)$$

Это приводит к стандартизации объектов, так что квадраты переменных, измеренные на каждом объекте, в сумме дают единицу. Тогда cos θ определяется по формуле

$$\cos\theta_{ij} = \sum_{k=1}^{m} w_{ik} w_{jk} \qquad (6.70)$$

В матричных обозначениях, мы сначала определяем диагональную матрицу $[D]$, порядка $n \times n$, которая содержит суммы квадратов элементов каждой строки вдоль диагонали и нули в остальных местах.

Стандартизация осуществляется преобразованием

$$[W] = [D]^{-1/2}[X] \qquad (6.71)$$

и матрица сходства cos θ имеет вид

$$[\cos\theta] = [W] \cdot [W]' = [D]^{-1} \cdot [X] \cdot [X]' \cdot [D]^{-1} \quad (6.72)$$

Тот факт, что матрица [**cos** θ] порядка $n \times n$ имеет не более m собственных значений, вытекает из теоремы Эккарта–Юнга. Используя ее, достаточно найти собственные векторы $m \times m$ матрицы $[W]'[W]$, а не большей матрицы [**cos** θ], и затем преобразовать метки R-метода в нагрузки Q-метода и наоборот. Этот вопрос подробно рассмотрен в следующем разделе. В некоторых случаях соотношение взаимосвязи между факторами и метками R- и Q-методов справедливо точно, но это зависит от способа шкалирования, примененного к матрице данных $[X]$.

В табл. 6.28 приведена матрица сходства Q-метода для данных по случайным блокам, а в табл. 6.29 приведены собственные значения и три собственных вектора. Как и следовало ожидать, первые несколько собственных значений учитывают почти всю изменчивость между блоками. Их собственные векторы можно преобразовать в векторы факторных нагрузок, умножая каждую компоненту вектора на соответствующее сингулярное значение (или квадратный корень из соответствующего собственного значения). Это в точности та же процедура, которая используется в R-методе факторного анализа. Факторные оси Q-метода шкалируются таким образом, чтобы их длины были пропорциональны распределению вкладов содержащейся в них дисперсии между объектами. В табл. 6.29 также приводится перечень осей Q-метода, которые соответствуют первым трем собственным векторам.

В Q-методе факторного анализа если требуется установить связи между объектами в выборке, то нужно нанести нагрузки, а не факторные оси.

Таблица 6.28. Значения косинусов θ, которые пропорциональны коэффициентам сходства между 25 блоками, приведенным в табл. 6.18 (только нижняя половина симметрической матрицы имеет порядок 25×25)

	a	*b*	*b*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*
a	1,0000																								
b	0,8993	1,0000																							
c	0,5992	0,8554	1,0000																						
d	0,5096	0,7916	0,9920	1,0000																					
e	0,7406	0,9520	0,9658	0,9284	1,0000																				
f	0,9167	0,9959	0,8566	0,7917	0,9462	1,0000																			
g	0,9314	0,7963	0,5005	0,4152	0,6284	0,8276	1,0000																		
h	0,6782	0,9191	0,9848	0,9586	0,9951	0,9134	0,5708	1,0000																	
i	0,9458	0,9817	0,7881	0,7129	0,9021	0,9834	0,8130	0,8577	1,0000																
j	0,9137	0,9732	0,7591	0,6909	0,8790	0,9664	0,8628	0,8369	0,9472	1,0000															
k	0,9113	0,9935	0,8138	0,7477	0,9225	0,9877	0.8326	0,8850	0,9705	0,9929	1,0000														
l	0,6457	0,8963	0,9939	0,9766	0,9851	0,8927	0,5517	0,9965	0,8265	0,8154	0,8628	1,0000													
m	0,7873	0,9585	0,9529	0,9156	0,9830	0,9623	0,7258	0,9761	0,9042	0,9123	0,9423	0,9736	1,0000												
n	0,8868	0,9683	0,8183	0,7586	0,9074	0,9714	0,8724	0,8768	0,9259	0,9837	0,9824	0,8648	0,9531	1,0000											
o	0,7909	0,8171	0,7825	0,7385	0,7981	0,8579	0,8659	0,7860	0,7813	0,8098	0,8174	0,7943	0,8834	0,8924	1,0000										
p	0,8777	0,9887	0,8982	0,8392	0,9703	0,9924	0,7652	0,9449	0,9759	0,9332	0,9673	0,9251	0,9703	0,9442	0,8411	1,0000									
q	0,8766	0,9904	0,8826	0,8207	0,9664	0,9889	0,7448	0,9377	0,9820	0,9342	0,9687	0,9133	0,9561	0,9332	0,8003	0,9971	1,0000								
r	0,7393	0,9545	0,9603	0,9252	0,9960	0,9464	0,6462	0,9914	0,8934	0,8979	0,9335	0,9844	0,9888	0,9270	0,8081	0,9626	0,9573	1,0000							
s	0,9589	0,9828	0,7937	0,7193	0,9006	0,9915	0,8652	0,8579	0,9930	0,9598	0,9772	0,8319	0,9220	0,9538	0,8452	0,9772	0,9753	0,8972	1,0000						
t	0,8154	0,9758	0,8962	0,8495	0,9636	0,9685	0,7599	0,9456	0,9174	0,9624	0,9764	0,9359	0,9819	0,9801	0,8432	0,9602	0,9542	0,9794	0,9317	1,0000					
u	0,9460	0,9882	0,7900	0,7156	0,9061	0,9880	0,8244	0,8620	0,9984	0,9630	0,9819	0,8321	0,9129	0,9424	0,7893	0,9761	0,9816	0,9023	0,9943	0,9328	1,0000				
v	0,6095	0,8833	0,9833	0,9702	0,9759	0,8729	0,5244	0,9896	0,7987	0,8138	0,8561	0,9948	0,9639	0,8609	0,7651	0,9016	0,8917	0,9823	0,8041	0,9386	0,8091	1,0000			
w	0,5899	0,8396	0,9961	0,9943	0,9494	0,8438	0,5096	0,9715	0,7671	0,7526	0,8026	0,9869	0,9498	0,8182	0,8007	0,8812	0,8609	0,9487	0,7799	0,8906	0,7711	0,9776	1,0000		
x	0,7359	0,9503	0,9395	0,9059	0,9802	0,9394	0,6682	0,9747	0,8774	0,9166	0,9410	0,9704	0,9850	0,9446	0,8128	0,9450	0,9385	0,9939	0,8874	0,9900	0,8921	0,9769	0,9332	1,0000	
y	0,5749	0,8571	0,9925	0,9857	0,9676	0,8495	0,4835	0,9869	0,7739	0,7739	0,8227	0,9954	0,9520	0,8280	0,7547	0,8861	0,8740	0,9698	0,7784	0,9127	0,7814	0,9968	0,9877	0,9584	1,0000

Таблица 6.29. Собственные значения, первые три собственных вектора и векторы факторных нагрузок матрицы косинусов θ табл. 6.28. Собственные значения от VIII до XXV тождественно равны нулю

Вектор	Собственное значение	Вклад в общую дисперсию, %	Сумма вкладов в общую дисперсию, %
I	22,2999	89,20	89,20
II	1,9907	7,96	97,16
III	0,4546	1,82	98,98
IV	0,2192	0,88	99,86
V	0,0324	0,13	99,99
VI	0,0029	0,01	100,00
VII	0,0004	0,00	100,00

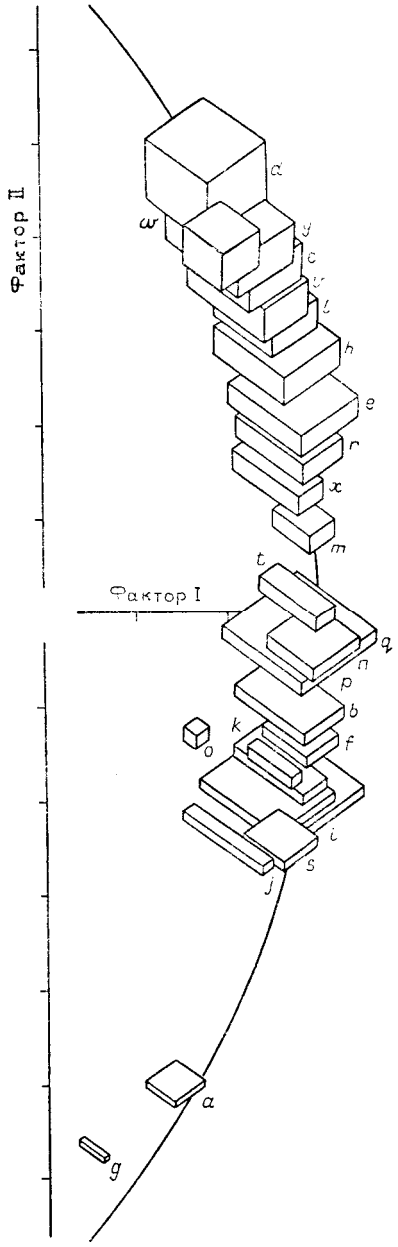
Блок	Собственный вектор			Фактор		
	I	*II*	*III*	*I*	*II*	*III*
a	0,1780	–0,3709	–0,0414	0,8407	–0,5234	–0,0414
b	0,2085	–0,0969	–0,1593	0,9845	–0,1368	–0,1593
c	0,1961	0,2574	0,0625	0,9261	0,3632	0,0625
d	0,1859	0,3260	0,1047	0,8780	0,4600	0,1047
e	0,2079	0,1155	–0,1101	0,9816	0,1629	–0,1101
f	0,2087	–0,1123	–0,0613	0,9858	–0,1585	–0,0613
g	0,1606	–0,4137	0,4273	0,7586	–0,5836	0,4273
h	0,2042	0,1804	–0,0635	0,9644	0,2545	–0,0635
i	0,2004	–0,1783	-0,2520	0,9462	–0,2516	–0,2520
j	0,1997	–0,1894	–0,0658	0,9431	–0,2672	–0,0658
k	0,2055	–0,1417	–0,1168	0,9706	–0,1999	–0,1168
l	0,2020	0,2126	0,0179	0,9537	0,2999	0,0179
m	0,2103	0,0604	0,1123	0,9933	0,0852	0,1123,
n	0,2045	-0,1301	0,1498	0,9658	–0,1835	0,1498
o	0,1833	–0,0913	0,7009	0,8655	–0,1288	0,7009
p	0,2095	–0,0459	–0,1089	0,9893	–0,0648	–0,1089
q	0,2078	–0,0535	–0,2168	0,9814	–0,0755	–0,2168
r	0,2085	0,1095	–0,0637	0,9847	0,1545	–0,0637
s	0,2026	–0,1917	–0,0826	0,9566	–0,2705	–0,0826
t	0,2090	–0,0069	0,0094	0,9870	–0,0097	0,0094
u	0,2018	–0,1792	–0,2294	0,9531	–0,2528	–0,2294.
v	0,1993	0,2327	0,0011	0,9411	0,3283	0,0011
w	0,1943	0,2603	0,1567	0,9176	0.3673	0,1567
x	0,2073	0,0925	–0,0103	0,9791	0,1305	–0,0103
y	0,1957	0,2697	0,0197	0,9243	0,3805	0,0197

На рис. 6.39 представлены первые две *Q*-факторные оси; блоки изображены в позициях, представляющих их нагрузки на факторы. Дуга на диаграмме – есть часть окружности, соответствующей области, равной 1,00; если некоторый объект попадает на окружность, то два фактора учитывают всю их изменчивость. Блоки, расположенные внутри окружности, характеризуются изменчивостью, которая не представляется двумя факторами.

На рис. 6.40 представлены вторая и третья *Q*-факторные оси. Эти и представленные на рис. 6.39 факторы вместе описывают 99% общей изменчивости блоков, которая в точности такова, как мы ожидали, зная заранее, какие блоки использовались. Хотя на рис. 6.39 представлена общая прогрессия больших блоков к меньшим, различие в форме не так важно, как в *R*-методе факторного

анализа (см. рис. 6.38). Можно заметить, что второй и третий факторы, представленные на рис. 6.40, дают распределение нагрузок, которые важнее для целей классификации, чем первый фактор, хотя первый собственный вектор по порядку величины больше, чем второй.

В *R*-методе факторного анализа данные центрируются относительно нуля вычитанием средних из каждого наблюдения до выделения факторов. Это не делается в *Q*-методе факторного анализа, поэтому первый *Q*-фактор есть просто вектор, выходящий из начала координат и кончающийся в центроиде заданного множества объектов. Действительно, этот первый фактор выражает «размер» или комбинированную величину переменных, измеренных на каждом объекте. Типично, что все нагрузки на первый фактор имеют положительный знак и примерно равные величины. Так как первый фактор дает очень мало информации о структуре данных, то он обычно характеризуется как «фактор рождения» и отбрасывается. Второй и третий факторы (см. рис. 6.40) считаются более важными.



Для иллюстрации рассмотрим геологический пример, взятый из петрологии изверженных пород. В табл. 6.30 представлены данные по главным химическим составляющим 20 образцов, взятых из сложного и отчетливо дифференцированного массива изверженных пород. С помощью *Q*-анализа мы рассчитываем поместить каждый образец на его настоящее место в ряду дифференцированной серии. Порядок следования наблюдений в пределах последовательности можно изобразить графически как нагрузки на пары факторов. Так как наши наблюдения были приведены к стандартному виду, то они изображаются векторами единичной длины, и потому их концы лежат на единичной окружности. Углы между различными радиусами можно рассматривать как меры сходства между наблюдениями, что следует из формулы (6.68). Матрица сходства, т.е. матрица [**cos** θ], приведена в табл. 6.31, а факторная матрица дана в табл. 6.32. Сохранены два фактора, и к ним применяется вращение: факторные нагрузки после вращения приведены в табл. 6.33. Графически эти нагрузки представлены на рис. 6.41. Отметим, что наблюдения располагаются в определенной последовательности, начиная с наблюдений, классифицированных как гиперстеновое габбро, и кончая кварцевыми сиенитами. В табл. 6.33 приведены также факторные значения исходных переменных (химических составляющих) на два фактора. Ясно, что интерпретация обоих факторов, характеризуемых большим весом и потому относительной избыточностью кремния и алюминия, соответствует традиционным представлениям о последовательности дифференциации изверженных пород.

Рис. 6.39. Графическое представление факторных нагрузок *Q*-метода на первые две факторные оси для данных по блокам, вычисленных по матрице сходства cos θ (см. табл. 6.28). Заметим, что вертикальная ось смещена относительно начала координат. Дуга окружности представляет общность, равную 1,00

Сделаем несколько заключительных замечаний относительно *Q*-метода факторного анализа. Его главная цель по существу – идентифицировать крайние члены – настоящие или гипотетические объекты, имеющие экстремальные свойства. Промежуточные объекты часто можно рассматривать как смеси двух крайних членов, как это сделано для изверженных пород, представленных на рис. 6.41, которые можно интерпретировать как члены последовательности градаций между сиалическими и фемическими породами. *Q*-метод факторного анализа также используется для классификации, объективно дающей в сущности тот же результат, что и кластерный анализ, да еще с большими затратами машинного времени. Если цель анализа–поиск групп или кластеров в совокупности проб, то

как кластерный анализ, так и *R*-метод факторного анализа дают более эффективные средства для этого. Если значимые факторы можно определить с помощью *R*-метода, то диаграмма рассеяния факторных меток или кластерных диаграмм обычно дает представление о соотношениях между пробами. В качестве примера на рис. 6.42 изображена дендрограмма, построенная на основании корреляционной матрицы для данных по изверженным породам, приведенным в табл. 6.30, методом усреднения взвешенных пар. Относительное расположение наблюдений почти в точности совпадает с расположением, полученным *Q*-методом.

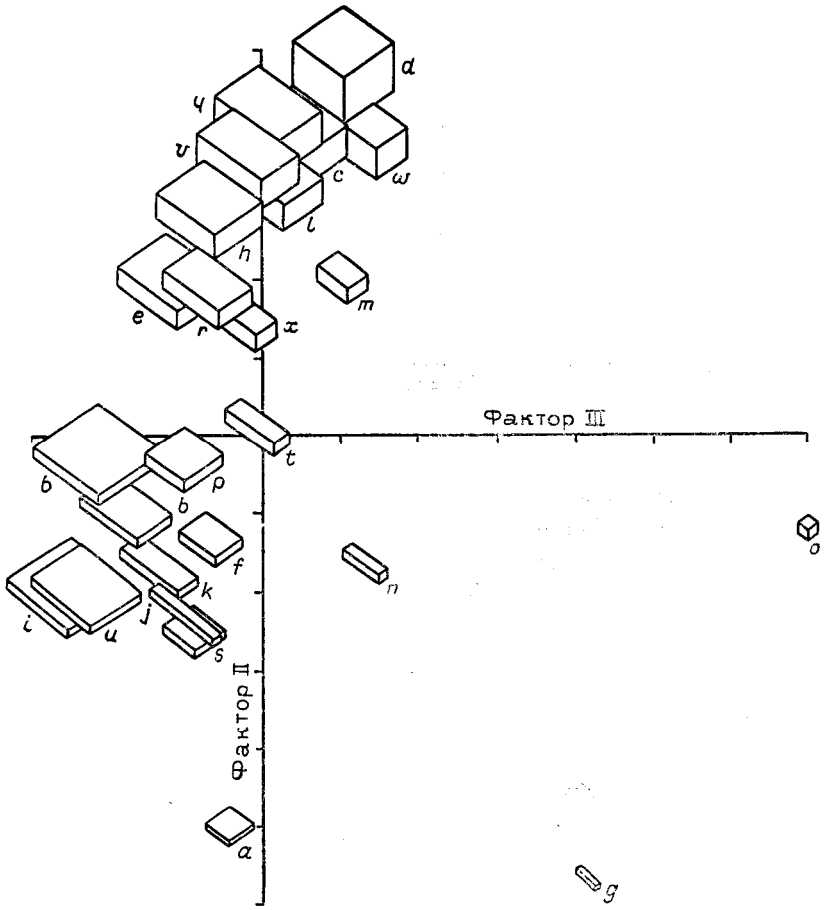


Рис. 6.40. Графическое представление факторных нагрузок *Q*-метода на факторные оси *II* и *III* для данных по блокам, вычисленных по матрице сходства cos θ (см. табл. 6.28)

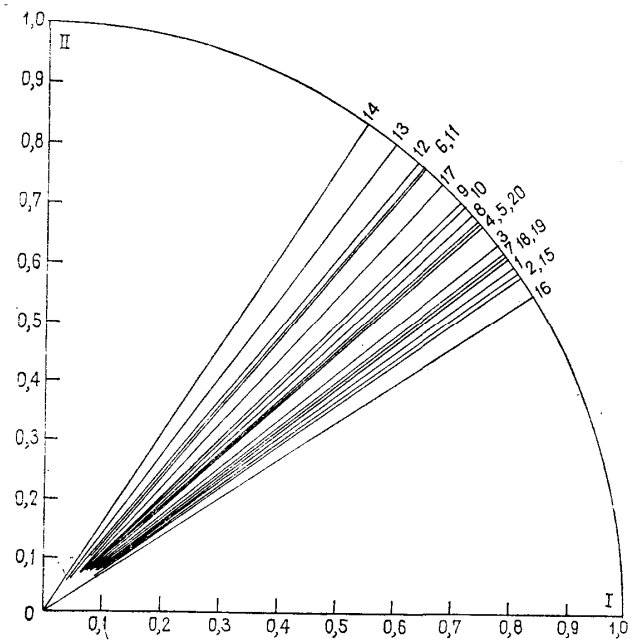


Рис. 6.41. Графическое изображение факторных нагрузок в Q-методе для 20 проб изверженных пород (см. рис. 6.42)

Таблица 6.30.Содержания основных породообразующих оксидов в 20 пробах изверженных пород

Название породы	Номер пробы	SiO_2	Al_2O_3	Fe_2O_3	FeO	MgO	CaO	Na_2O	K_2O
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Сиениты	1	61,7	15,1	2,0	2,3	3,7	4,6	4,4	4,5
Сиениты	2	58,3	17,9	3,2	1,7	1,5	3,7	5,9	5,3
Сиениты	3	51,2	17,6	3,5	4,3	3,2	4,5	5,7	4,4
Монцониты	4	54,4	14,3	3,3	4,1	6,1	7,7	3,4	4,2
Диориты	5	58,0	15,7	0,7	2,8	5,0	10,9	3,0	3,2
Диориты	6	46,9	15,9	2,9	10,0	7,0	9,6	2,7	0,7
Диориты	7	58,0	17,3	2,2	3,8	2,2	4,3	4,3	4,1
Кварцевые диориты	8	55,5	16,5	1,7	4,6	6,7	6,7	3,2	2,5
Габбро	9	55,4	15,3	2,7	5,5	5,8	9,9	2,9	1,5
Габбро	10	55,9	13,5	2,7	5,9	6,5	8,9	2,4	1,7
Нориты	11	47,2	14,5	1,6	13,8	5,2	8,1	3,1	1,2
Нориты	12	48,2	18,3	1,3	6,1	10,8	9,4	1,3	0,7
Гиперстеновое габбро	13	44,8	18,8	2,2	4,7	11,3	14,6	0,9	0,1
Гиперстеновое габбро	14	47,0	14,1	0,8	15,0	16,0	2,3	0,4	1,7
Сиениты	15	59,8	17,3	3,6	1,6	1,2	3,8	5,0	5,1
Кварцевые сиениты	16	66,2	16,2	2,0	0,2	0,8	1,3	6,5	5,8
Измененные сиениты*	17	50,0	9,9	3,5	5,0	11,9	8,3	2,4	5,0
Монцониты	18	57,4	18,5	3,7	2,1	1,7	6,8	4 3	3,7
Монцониты	19	59,8	15,8	3,8	3,3	2,2	3,9	3/J	4,4
Диабазы	20	52,2	18,2	3,3	4,4	4,7	6,5	4,6	1,9

* Содержат вторичные минералы, включая диопсид.

Таблица 6.32. Первые пять факторов, найденных по матрице cos θ, приведенной в табл. 6.31

Номер пробы	*I*	*II*	*III*	*IV*	*V*	Общность
1	0,9948	–0,0910	0,0242	0,0324	0,0069	0,9996
2	0,9918	–0,1223	0,0081	–0,0177	–0,0268	0,9997
3	0,9958	–0,0587	0,0085	–0,0457	–0,0344	0,9983
4	0,9989	–0,0126	–0,0070	0,0357	0,0178	0,9997
5	0,9963	–0,0191	–0,0596	0,0297	0,0353	0,9986
6	0,9904	0,1188	–0,0133	–0,0594	0,0309	0,9997
7	0,9959	–0,0838	0,0191	–0,0235	–0,0086	0,9998
8	0,9996	0,0010	–0,0017	0,0112	–0,0132	0,9996
9	0,9983	0,0204	–0,0336	0,0055	0,0391	0,9997
10	0,9978	0,0223	–0,0049	0,0291	0,0498	0,9994
11	0,9833	0,1202	0,0550	–0,0988	0,0746	0,9997
12	0,9890	0,1259	–0,0512	0,0008	–0,0538	0,9995
13	0,9721	0,1719	–0,1552	0,0066	–0,0365	0,9999
14	0,9561	0,2323	0,1691	0,0146	–0,0527	0,9997
15	0,9918	–0,1257	0,0102	–0,0084	–0,0137	0,9998
16	0,9844	–0,1665	0,0458	0,0203	–0,0113	0,9994
17	0,9866	0,0783	0,0214	0,1316	0,0259	0,9980
18	0,9950	–0,0870	–0,0367	–0,0275	–0,0089	0,9998
19	0,9945	–0,0946	0,0296	0,0035	0,0066	0,9989
20	0,9981	–0,0161	–0,0236	–0,0395	–0,0295	0,9995
Вклад в дисперсию (%)						
	98,124	1,148	0,349	0,204	0,116	
Кумулятивный вклад в дисперсию (%)						
	98,124	99,272	99,621	99,825	99,942	

Таблица 6.31. Значения cos θ для 20 проб изверженных пород

Номер пробы	Номер пробы																			
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
1	1,000																			
2	0,997	1,000																		
3	0,994	0,997	1,000																	
4	0,996	0,991	0,994	1,000																
5	0,993	0,988	0,989	0,997	1,000															
6	0,972	0,968	0,981	0,987	0,984	1,000														
7	0,998	0,999	0,998	0,995	0,992	0,977	1,000													
8	0,995	0,991	0,995	0,998	0,996	0,989	0,995	1,000												
9	0,991	0,986	0,990	0,998	0,998	0,992	0,991	0,998	1,000											
10	0,992	0,985	0,988	0,988	0,996	0,991	0,991	0,997	0,999	1,000										
11	0,966	0,961	0,975	0,978	0,974	0,996	0,972	0,981	0,984	0,984	1,000									
12	0,971	0,966	0,978	0,985	0,985	0,993	0,974	0,990	0,990	0,988	0,981	1,000								
13	0,948	0,943	0,958	0,970	0,973	0,984	0,951	0,973	0,978	0,973	0,965	0,993	1,000							
14	0,934	0,922	0,940	0,950	0,937	0,970	0,936	0,957	0,952	0,956	0,972	0,969	0,945	1,000						
15	0,998	1,000	0,996	0,992	0,989	0,967	0,999	0,991	0,987	0,986	0,960	0,965	0,941	0,921	1,000					
16	0,997	0,997	0,989	0,985	0,982	0,953	0,995	0,985	0,978	0,978	0,948	0,951	0,922	0,911	0,997	1,000				
17	0,979	0,967	0,973	0,990	0,984	0,980	0,973	0,987	0,987	0,990	0,970	0,982	0,969	0,965	0,968	0,961	1,000			
18	0,996	0,998	0,997	0,994	0,994	0,977	0,998	0,994	0,992	0,990	0,968	0,975	0,958	0,925	0,998	0,992	0,971	1,000		
19	0,999	0,998	0,995	0,995	0,991	0,973	0,999	0,994	0,992	0,990	0,968	0,970	0,946	0,934	0,999	0,996	0,975	0,997	1,000	
20	0,992	0,993	0,998	0,996	0,993	0,989	0,996	0,998	0,996	0,993	0,980	0,988	0,972	0,947	0,992	0,984	0,977	0,997	0,993	1,000

Таблица 6.33. Факторные нагрузки после вращения и факторные значения, вычисленные по методу варимакс

Номер пробы	I	II	Общность
1	0,7851	0,6177	0,9980
2	0,8044	0,5929	0,9986
3	0,7636	0,6418	0,9950
4	0,7342	0,6774	0,9980
5	0,7368	0,6709	0,9929
6	0,6377	0,7671	0,9950
7	0,7809	0,6236	0,9988
8	0,7254	0,6878	0,9993
9	0,7111	0,7009	0,9970
10	0,7094	0,7020	0,9960
11	0,6316	0,7632	0,9814
12	0,6319	0,7712	0,9940
13	0,5879	0,7930	0,9745
14	0,5348	0,8259	0,9681
15	0,8068	0,5904	0,9995
16	0,8295	0,5556	0,996b
17	0,6628	0,7350	0,9796
18	0,7825	0,6207	0,9976
19	0,7873	0,6148	0,9979
20	0,7360	0,6744	0,9965
Вклад в дисперсию (%)			
	52,311	46,962	
Кумулятивный вклад в дисперсию (%)			
	52,311	99,272	
Матрица факторных значений, найденная по методу варимакс			

Переменная	Фактор	
	I	*II*
X_1	70,2648	5,6766
X_2	14,5830	8,1431
X_3	4,5006	–1,0267
X_4	–16,2185	24,5371
X_5	–19,4934	28,8944
X_6	–6,5178	16,8625
X_7	11,4400	–6,9660
X_8	11,1204	–7,2130

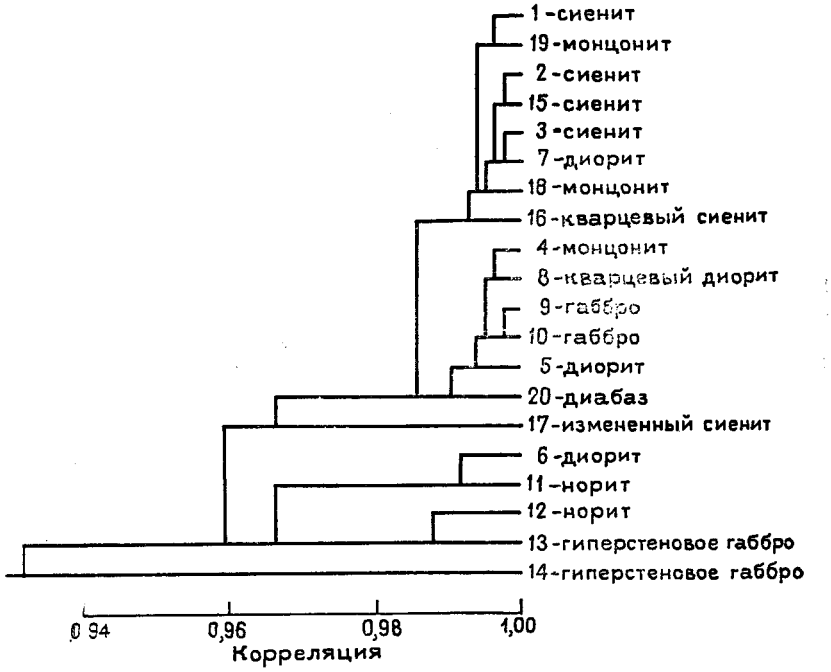


Рис. 6.42. Анализ групп для данных химического состава изверженных пород. Этот анализ дает в сущности тот же результат, что и *Q*-метод факторного анализа

АНАЛИЗ ГЛАВНЫХ КООРДИНАТ

Анализ главных координат – это широко распространенный *Q*-метод факторного анализа. Этот метод в основном применяется в количественной биологии и палеонтологии, хотя он используется также в петрологии. Метод главных координат популяризовался Говером [20] и подробно рассматривался другими авторами [56, 32]. Объективно с его помощью решают ту же задачу, что и *Q*-метод факторного анализа, т.е. определяют, является ли множество многомерных наблюдений выборкой из одной и той же совокупности или оно есть смесь представителей нескольких различных совокупностей.

В главных координатах первый шаг состоит в вычислении матрицы сходства порядка $n \times n$ или матрицы расстояний между п объектами множества данных. Можно использовать любое расстояние, например, евклидово

$$D_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \qquad (6.73)$$

Весьма широко используется также расстояние Говера

$$G_{ij} = \frac{1}{m} \sum_{k=1}^{m} \left(1 - \frac{|x_{ik} - x_{jk}|}{\text{размах } k} \right) \qquad (6.74)$$

Для вычисления расстояния Говера между i-м и j-м объектами необходимо вычислить абсолютное значение разности между ними для переменной с номером k и разделить на размах переменной k. Это дает некоторое число из интервала 0,0–1,0, причем близости объектов соответствует малое значение коэффициента сходства, а близким к единице числам – максимальное расхождение объектов. Для того чтобы получить меру сходства, ведущую себя по аналогии с коэффициентом корреляции, эта величина вычитается из 1,0. Вычисление повторяется для всех m переменных k, измеренных на объектах i и j, затем результаты суммируются и делятся на число переменных m, что в заключение приводит к величине G_{ij}.

При вычислении расстояния Говера не делается никаких допущений относительно природы данных; наблюдения могут быть номинальными или порядковыми, или более высокого ранга. Действительно, матрица данных может состоять из смеси чисел различного типа, таких как числа пла-

стинок в чашечках криноидей, длин их щупалец и отношений высот чашечек к их диаметрам. Меры сходства для всех возможных пар объектов представляются в виде матрицы ассоциаций [A] порядка $n \times n$-Эта матрица будет симметричной и будет иметь единичные значения на диагонали и числа, принадлежащие интервалу от нуля до единицы, в остальных местах.

Данные каждой строки матрицы [A] суммируются, полученная сумма делится на n; эта процедура дает среднее значение по строке. Данные каждого столбца матрицы [A] также суммируются и сумма делится на n, что дает среднее значение по столбцу. Обозначим эти средние соответственно через $\bar{a}_{j\bullet}$, и $\bar{a}_{\bullet k}$. Находится также общее среднее как строк, так и столбцов и обозначается $a_{\bullet\bullet}$. В результате этого элементы a_{jk} преобразуются, получается новая матрица [Q], элементы которой находятся по формуле

$$q_{jk} = a_{jk} + a_{\bullet\bullet} - (\bar{a}_{j\bullet} - \bar{a}_{\bullet k}) \qquad (6.75)$$

Рассмотрим n объектов, расположенных в m-мерном пространстве, определенном этими переменными. Преобразование (6.75) приводит к переносу начала координат m-мерного пространства в центроид множества точек. Эта операция приводит к замыканию множества данных, так как все строки и столбцы имеют теперь суммы элементов, равные нулю, поэтому одно из собственных значений матрицы [Q] обязано быть нулем. Это приводит к возрастанию относительной величины первых нескольких собственных значений.

Далее, находятся собственные значения и собственные векторы матрицы [Q]; это и есть главные координаты множества данных. Относительная важность каждой координаты может быть оценена простым вычислением процентного вклада каждого собственного значения в след матрицы [Q]. Обычно только первые несколько координат представляют интерес, так как нередко они учитывают большую часть различий между наблюдениями. В заключение индивидуальные нагрузки на главные координаты наносятся на график; это делается попарным изображением множества n собственных векторов, каждый из которых соответствует некоторому объекту.

Для иллюстрации анализа главных координат воспользуемся данными по искусственным блокам. Этот пример позволит нам сравнить результаты, полученные разными методами. В табл. 6.34 представлены коэффициенты сходства между индивидуальными блоками (матрица порядка 25×25), вычисленные с помощью расстояния Говера. В части, расположенной выше значений 1,000, представлены расстояния Говера, определенные для элементов матрицы [A] по формуле (6.74); в части, расположенной ниже этих значений, – меры сходства после указанного преобразования, состоящего в вычитании из каждого элемента среднего по строке и столбцу и последующего добавления общего среднего, как это указано в уравнении (6.75). Для этой матрицы [Q] находятся собственные векторы и собственные значения.

Таблица 6.35. Собственные значения, ассоциированные с первыми семью координатами, извлеченными из данных по блокам; графа 1 – последовательность собственных значении, графа 2 – процент от общей изменчивости, учитываемой для каждого собственного значения, графа 3 – кумулятивная изменчивость (%)

Координаты	1	2	3
I	13,3598	53,5758	53,5758
II	6,9122	27,7197	81,2954
III	4,2627	17,0943	98,3897
IV	0,3291	1,3200	99,7097
V	0,0682	0,2735	99,9832
VI	0,0042	0,0168	100,0000
VII	0,0000	0,0000	100,0000

Таблица 6.34. Матрица сходства между 25 случайными блоками; часть матрицы, расположенная над значениями 1,000, содержит расстояния Гувера; часть матрицы, расположенная под значениями 1,000, содержит расстояния Говера, преобразованные с помощью вычитания сумм строк и столбцов и прибавления общего среднего

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*
	1,0000	0,2262	-0,7412	-0,8077	-0,8564	0,9841	0,8343	-0,9455	0,3172	0,4510	0,4431	-0,8790	0,2417	0,7021	0,6164	0,4201	-0,1729	-0,9702	0,9750	0,0980	0,3836	-0,9852	-0,5220	-0,7157	-0,9656
a	1,0720	1,0000	-0,7339	-0,6850	0,1049	0,2151	-0,0963	-0,2045	0,8481	0,6541	0,8157	-0,6104	-0,8552	-0,0808	-0,5916	0,2330	0,7089	-0,0112	0,2568	0,1343	0,9295	-0,3299	-0,8934	-0,0419	-0,4429
b	0,2415	0,9586	1,0000	0,9852	0,5418	-0,7002	-0,6039	0,7751	-0,5441	-0,8599	-0,8943	0,9693	0,3308	-0,5915	-0,0947	-0,1612	-0,1661	0,5743	-0,6863	-0,4300	-0,6854	0,7814	0,9330	0,2685	0,8823
c	-0,7098	-0,7593	0,9906	1,0000	0,5617	0,7892	-0,6494	0,7944	-0,5669	-0,7646	-0,8084	0,9811	0,2700	-0,5855	-0,1766	-0,2993	-0,1815	0,6522	-0,7785	-0,3126	-0,6841	0,8549	0,9162	0,4077	0,9311
d	-0,7810	-0,7151	0,9711	0,9813	1,0000	-0,7909	-0,9036	0,9460	0,1543	-0,4760	-0,3594	0,6787	-0,5661	-0,9013	-0,7969	0,0458	0,6314	0,8881	-0,7372	-0,4208	0,0402	0,7629	0,2073	0,3848	0,7600
e	-0,8551	0,0495	0,5024	0,5177	0,9306	1,0000	0,7957	-0,8875	0,3480	0,3475	0,3649	-0,8397	0,2453	0,6321	0,6104	0,5261	-0,1112	-0,9484	0,9903	-0,0064	0,3933	-0,9799	-0,5069	-0,7764	-0,9469
f	1,0546	0,2289	-0,6704	-0,7640	-0,7910	1,0690	1,0000	-0,8896	-0,1708	0,4622	0,3298	-0,7092	0,5130	0,9431	0,8394	0,0199	-0,5568	-0,8426	0,7299	0,4235	-0,0696	-0,7789	-0,3278	-0,3912	-0,7870
g	0,9071	-0,0803	-0,5719	-0,6220	-0,9016	0,8669	1,0734	1,0000	-0,1405	-0,6391	-0,5781	0,8794	-0,2907	-0,8595	-0,6301	-0,1054	0,3669	0,9109	-0,8513	-0,3988	-0,2652	0,8967	0,5056	0,4622	0,9200
h	-0,9390	-0,2548	0,7408	0,7555	0,8817	-0,8826	-0,8825	0,9408	1,0000	0,2692	0,4780	-0,5221	-0,7642	-0,3110	-0,5023	0,6771	0,8581	-0,1720	0,4404	-0,3800	0,9755	-0,4453	-0,7232	-0,4756	-0,4603
i	0,3470	0,8211	-0,5550	-0,5825	0,1134	0,3762	-0,1403	-0,1763	0,9876	1,0000	0,9666	-0,7675	-0,3314	0,6108	-0,0437	-0,3316	-0,0378	-0,2985	0,3120	0,7650	0,4656	-0,4499	-0,7711	0,1556	-0,6032
j	0,4833	0,6296	-0,8683	-0,7777	-0,5144	0,3782	0,4952	-0,6724	0,2593	0,9925	1,0000	-0,7905	-0,4977	0,4512	-0,2039	-0,1683	0,1792	-0,2563	0,3482	0,6245	0,6569	-0,4680	-0,8647	0,1181	-0,6185
k	0,4697	0,7855	-0,9084	-0,8271	-0,4036	0,3900	0,3571	-0,6171	0,4623	0,9534	0,9811	1,0000	0,1706	-0,6479	-0,2683	-0,2786	-0,0794	0,7540	-0,8336	-0,3165	-0,6410	O,S060	0,8490	0,4593	0,9677
I	-0,8590	-0,6471	0,9487	0,9558	0,6280	-0,8211	-0,6885	0,8338	-0,5443	-0,7872	-0,8159	0,9680	1,0000	0,5012	0,8785	-0,1803	-0,8831	-0,4139	0,1718	0,0896	-0,7714	-0,1077	0,6221	-0,1589	-0,0154
m	0,3111	-0,8425	0,3595	0,2940	-0,5675	0,3131	0,5830	-0,2869	-0,7371	-0,3018	-0,4739	0,1880	1,0667	1,0000	0,7526	-0,2895	-0,6827	-0,6937	0,5425	0,6872	-0,1568	-0,6164	-0,2941	-0,0945	-0,6701
n	0,7659	-0,0741	-0,5688	-0,5674	-0,9086	0,6940	0,0072	-0,8617	-0,2898	0,6344	0,4691	-0,6365	0,5620	1,0548	1,0000	0,0295	-0,7698	-0,7373	0,5440	0,1419	-0,4812	-0,5195	0,2127	-0,4203	-0,4458
o	0,6957	-0,5691	-0,0561	-0,1427	-0,7883	0,6882	0,9193	-0,6164	-0,4653	-0,0041	-0,1701	-0,2410	0,9551	0,8232	1,0865	1,0000	0,5796	-0,4083	0,6087	-0,8063	0,5398	0,5221	-0,2617	-0,8887	-0,4085
p	0,4746	0,2308	-0,1474	-0,2901	0,0296	0,5791	0,0752	-0,1165	0,6894	-0,3168	-0,1592	-0,2761	-0,1285	-0,2437	0,0912	1,0370	1,0000	0,3030	-0,0128	-0,5204	0,7763	-0,0179	-0,4825	-0,2003	0,0061
q	-0,1584	0,6667	-0,1923	-0,2124	0,5752	0,0982	-0,5417	0,3158	0,8304	-0,0631	0,1482	-0,1169	-0,8713	-0,6769	-0,7481	0,5765	0,9569	1,0000	-0,9431	-0,0169	0,2108	0,9366	0,3190	0,7579	0,8824
r	-0,9751	-0,0728	0,5287	0,6020	0,8124	-0,9549	-0,8468	0,8404	-0,2192	-0,3431	-0,3066	0,6971	-0,4215	-0,7073	-0,7350	-0,4308	0,2406	0,9181	1,0000	-0,1066	0,4668	-0,9853	-0,5184	-0,8367	-0,0415
s	1,0461	0,2712	-0,6559	-0,7527	-0,7369	1,0598	0,8016	-0,8458	0,4692	0,3434	0,3738	-0,8145	0,2402	0,6050	0,6224	0,6623	0,0008	-0,9490	1,0701	1,0000	-0,1713	-0,0178	-0,2727	0,6199	-0,1728
t	0,1256	0,1052	-0,4431	-0,3302	-0,4639	0,0198	0,4518	-0,4368	-0,3946	0,7529	0,6067	-0,3409	0,1146	0,7062	0,1838	-0,7962	-0,5503	-0,0662	-0,0799	0,9833	1,0000	-0,4969	-0,8292	-0,3781	-0,5459
u	0,4113	0,9005	-0,6983	-0,7017	-0,0028	0,4195,	-0,0412	-0,3031	0,9610	0,4535	0,6391	-0,6653	-0,7464	-0,1378	-0,4463	0,5500	0,7465	-0,2601	0,4935	-0,1880	0,9834	1,0000	0,6137	0,7614	0,9819
v	-0,9803	-0,3818	0,7456	0,8145	0,6970	-0,9766	-0,7733	0,8360	-0,4826	-0,4848	-0,5086	0,8588	-0,1055	-0,6202	-0,5074	-0,5348	-0,0348	0,8646	-0,9813	-0,0573	-0,5364	0,9377	1,0000	0,2032	0,7235
w	-0,4782	-0,9064	0,9361	0,9146	0,1804	-0,4646	-0,2833	0,4838	-0,7217	-0,7670	-0,8664	0,8408	0,6632	-0,2589	0,2637	-0,2354	-0,4963	0,2858	-0,4756	-0,2733	-0,8297	0,5904	1,0155	1,0000	0,6345
x	-0,7142	-0,0972	0,2293	0,3639	0,3156	-0,7764	-0,3890	0,3961	-0,5164	0,1573	0,0741	0,4088	-0,1601	-0,1016	-0,4115	-0,9048	-0,2563	0,6825	-0,8361	0,5770	-0,4269	0,6958	0,1765	0,9310	1,0000
y	-0,9549	-0,4889	0,8523	0,8964	0,7001	-0,9377	-0,7756	0,8652	-0,4918	-0,6322	-0,6533	0,9264	-0,0074	-0,6680	-0,4279	-0,4153	-0,0407	0,8162	-0,9317	-0,2065	-0,5805	0,9254	0,7060	0,5747	0,9494

Таблица 6.36. Первые две главные координаты данных по блокам; каждый элемент соответствует конкретному блоку

Блок	Главные координаты	
	I	*II*
a	0,2685	–0,0847
b	0,1318	0,3113
c	–0,2405	–0,1235
d	–0,2503	–0,1110
e	–0,2071	0,2086
f	0,2606	–0,0745
g	0,2250	–0,2210
h	–0,2499	0,1019
i	0,1249	0,3108
j	0,1880	0,0683
k	0,1897	0,1463
l	–0,2609	–0,0727
m	0,0105	–0,3822
n	0,2005	–0,2333
o	0,1266	–0,3410
p	0,0978	0,1222
q	–0,0050	0,3500
r	–0,2304	0,1256
s	0,2573	–0,0431
t	0,0758	–0,1090
u	0,1532	0,3030
v	–0,2572	0,0085
w	–0,1942	–0,2370
x	–0,1491	–0,0076
y	–0,2657	–0,0159

В табл. 6.35 приведены первые семь собственных значений матрицы [*Q*]. Заметим, что седьмое и последующие собственные значения равны нулю. Действительно, первые два собственных значения дают вклад в общую изменчивость данных по блокам, равный 81%, а третье собственное значение дает еще дополнительный вклад, равный 17%, что составляет в сущности почти всю изменчивость. (Напомним, что данные были порождены только тремя независимыми переменными. Небольшая доля изменчивости, не учтенная первой, второй и третьей главными координатами, может быть объяснена ошибками округления в вычислениях.)

Первые две главных координаты, состоящие из элементов собственных векторов *I* и *II*, приведены в табл. 6.36. Каждый элемент соответствует индивидуальному наблюдению. Эти нагрузки изображены на рис. 6.43. Сравните результаты, полученные методом главных координат, с решением, полученным *Q*-методом факторного анализа (см. рис. 6.39). Заметим, что тот факт, что диагональные элементы матрицы [*Q*] могут быть не равными 1,00, означает, что представление общности в виде диаграммы невозможно осуществить на единичной окружности.

АНАЛИЗ СООТВЕТСТВИЯ

Факторный анализ предназначен для данных, представленных в интервальной форме или в шкале отношений, т.е для измерений, сделанных в непрерывной численной шкале. Он непригоден, например, для таких данных, как число ископаемых остатков различного типа в образцах. Такие номинальные или порядковые данные могут оказаться единственными доступными для исследования, и в некоторых случаях может оказаться полезным обработать их, используя методы теории собственных значений, аналогичные факторному анализу.

Задачи, в которых имеются данные-перечисления, обычно свойственны общественным наукам. В качестве примера можно назвать результаты анкетирования, которые подразделяются на категории. В силу этого большинство исследований, основанных на использовании методов теории собственных значений для анализа такого рода данных, были созданы социологами и статистиками, работающими над решением социологических проблем. Эти данные обычно представляются в виде условных таблиц; первая известная работа, в которой были применены такие таблицы, принадлежит Хиршфельду [27], см. также [17]. Совсем недавно Бензекри и другие исследователи [4] подробно изложили этот метод, и термин «анализ соответствия», введенный Бензекри, получил широкое распространение. Его работа стала основой для многих приложений в геологии [60, 61, 12]. В этих геологических приложениях, однако, методы Бензекри и его предшественников претерпели большие изменения. Хилл [26] излагает историю анализа соответствия и связи между работами различных авторов. Детальное изложение анализа соответствия и его обобщений содержится в монографии [41].

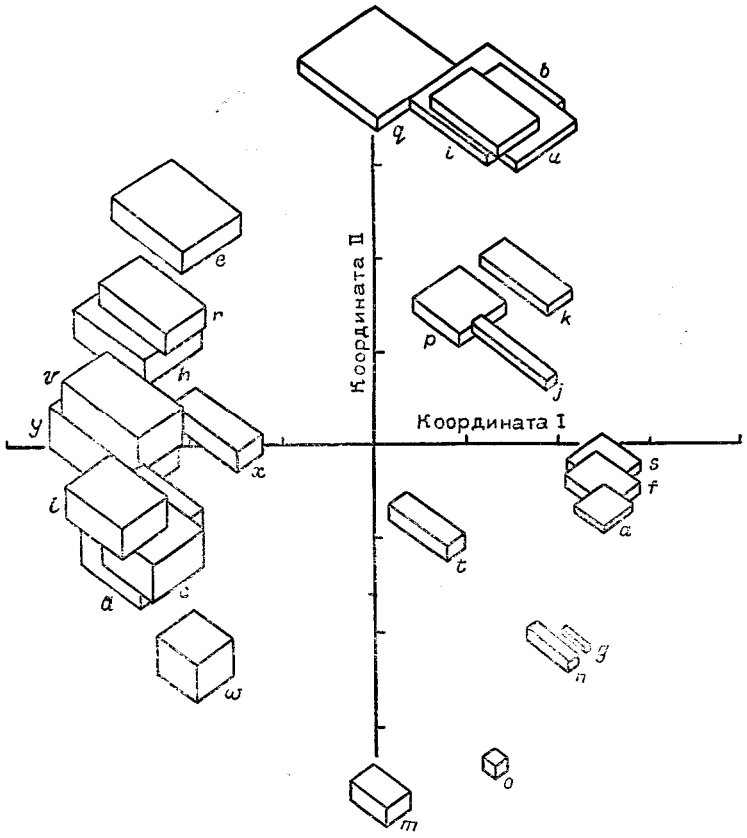


Рис. 6.43. Представление двух главных координат для данных по случайным блокам. Блоки изображены в положениях, соответствующих их нагрузкам на главные координаты

Анализ соответствия начинается с обработки матрицы, полученной из условной таблицы, которая преобразуется таким образом, чтобы ее элементы можно было рассматривать как условные вероятности. В силу природы этого преобразования (в действительности некоторая форма шкалирования) соотношения между строками и столбцами преобразованной таблицы такие же, как и в исходной матрице данных. Это означает, что теорема Эккарта–Юнга верна, и решения, полученные R- и Q-методами, эквивалентны.

Матрица необработанных данных $[X]$ имеет n строк, представляющих наблюдения, и m

столбцов переменных. Сами элементы рассматриваются как бирки. В задачах по палеонтологии, например, столбцы могут соответствовать видам останков микроорганизмов, строки могут представлять образцы, отобранные из различных стратиграфических интервалов в скважине, а элементы в таблице будут представлять собой результаты подсчета чисел образцов каждого вида останков микроорганизмов по выборкам. Общее число индивидуумов есть просто сумма всех элементов матрицы данных, или $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}$. Сумма элементов по строке $\sum_{i=1}^{n} x_{ij}$, есть число микроорганизмов всех типов, которые были обнаружены в каждой выборке, и сумма элементов по столбцу $\sum_{j=1}^{m} x_{ij}$, есть число микроорганизмов каждого вида, которые были обнаружены во всех выборках. Бирки можно обратить в проценты к общей сумме, а последние уже можно считать вероятностями

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}} \qquad (6.76)$$

Эти значения P_{ij} можно трактовать как совместные вероятности того, что конкретные виды остатков могут быть найдены в заданной выборке. Суммы строк, деленные на общую сумму, дают маргинальные вероятности

$$P_{i\bullet} = \frac{\sum_{j=1}^{m} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}} \qquad (6.77)$$

которые являются вероятностями того, что конкретные выборки будут содержать микроостанки, не взирая на их вид. Суммы столбцов, трактуемые аналогично, дают маргинальные вероятности

$$P_{\bullet j} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}} \qquad (6.78)$$

которые являются вероятностями того, что конкретные виды микроорганизмов имеются независимо от того, из какой выборки они были извлечены. Если объединенные вероятности разделить на соответствующие маргинальные вероятности, то в результате получим условные вероятности

$$P_{(j|i)} = P_{ij} \big/ P_{\bullet j} \quad ; \quad P_{(i|j)} = P_{ij} \big/ P_{i\bullet} \qquad (6.79)$$

Первая из этих условных вероятностей описывает ситуацию, когда, обнаружив микроорганизм вида j, мы хотим оценить вероятность того, что он появился в выборке с номером i. Вторая условная вероятность, основанная на суммах строк, дает вероятность того, что найденные микроорганизмы будут принадлежать к виду j, если известно, что этот образец был извлечен из i-й выборки.

В гл. 2 (см. кн.1) было показано, что таблица наблюдений может быть представлена через пропорции к общему числу наблюдений. Тогда, если строки и столбцы таблицы независимы, наблюдения должны быть приблизительно равными произведениям маргинальных вероятностей соответствующих им строк и столбцов. Если две переменные j и k тесно связаны, то все ожидаемые значения в j-м и k-м столбцах должны быть очень похожими. Это наводит на мысль, что степень сходства можно выразить с помощью вычисления попарного произведения, которое содержит наблюдаемые и ожидаемые вероятности для всех строк в двух сравниваемых столбцах. Такая мера используется в анализе соответствия и имеет вид коэффициента корреляции между двумя переменными [34]:

$$r_{ik} = \sum_{l=1}^{n} \left(\frac{P_{ij} - P_{i\bullet}P_{\bullet j}}{\sqrt{P_{i\bullet}P_{\bullet j}}} \right) \left(\frac{P_{i} - P_{i\bullet}P_{\bullet k}}{\sqrt{P_{i\bullet}P_{\bullet k}}} \right) \qquad (6.80)$$

где P_{ij}, – «наблюдаемая» вероятность в i-й строке и j-м столбце случайной таблицы; $P_{i\bullet}P_{\bullet j}$, – «ожи-

даемая» вероятность, вычисленная как произведение маргинальных вероятностей. Выражая r_{ik} через величины, введенные в гл. 2, получаем

$$r_{ik} = \sum_{l=1}^{n} \left(\frac{O_{ij} - E_j}{\sqrt{E_{ij}}} \right) \left(\frac{O_{ik} - E_{ik}}{\sqrt{E_{ik}}} \right) \qquad (6.81)$$

Связь между этим выражением и статистикой χ^2 в применении к случайной таблице становится более ясной, если возвести в квадрат один из членов:

$$\left(\frac{O_{ij} - E_j}{\sqrt{E_{ij}}} \right)^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Мы видим, что мера сходства, используемая в анализе соответствия, может рассматриваться как произведение двух значений χ^2. Это приводит к выражению «расстояние χ^2», которое иногда применяется к этой мере [41]. Если меры сходства r_{ik} вычислить для всех пар столбцов i и k, они образуют квадратную матрицу порядка $m \times m$. Из этой матрицы затем получаются собственные значения и собственные векторы. Это и есть главные оси анализа соответствия. Заметим, что так как все элементы случайной таблицы выражены как пропорции от общей суммы всех элементов, то сумма элементов столбца (и элементов строки) равна 1,00. Поэтому мы имеем дело с замкнутой таблицей данных, и одно собственное значение должно быть нулевым. Это означает, что размерность нашей задачи уменьшается от m до m–1,и, возможно, еще меньше. Вместо того чтобы прямо использовать уравнение (6.80), можно использовать другую формулу для вычисления коэффициента сходства, например следующую:

$$r_{ik} = \sum_{i=1}^{n} \frac{P_{ij} P_{ik}}{P_{i\bullet} \sqrt{P_{\bullet j} P_{\bullet k}}} \qquad (6.82)$$

Она дает то же множество собственных векторов.

Последнее собственное значение, как это вытекает из уравнения (6.80), тривиально и в точности равно нулю. Так как данные до выделения факторов не центрировались относительно нуля, то при использовании уравнения (6.82) факторное решение будет содержать исходное тривиальное собственное значение, которое тождественно равно 1,0. Вычисления, связанные с формулой (6.82), легче описать в матричной форме. Сначала обозначим исходную матрицу данных порядка $n \times m$ через $[X]$. Элементы $[X]$ преобразуются в объединенные вероятности с помощью деления каждого элемента матрицы на общую сумму, которая равна скаляру $\Sigma\Sigma x_{ij}$. В результате получаем матрицу $[B]$:

$$[B] = \frac{1}{\sum\sum x_{ij}} [X] \qquad (6.83)$$

Затем определим квадратную матрицу $[M]$ порядка $m \times m$, которая содержит суммы столбцов $[B]$, расположенных в порядке убывания по диагонали, и с нулями во всех внедиагональных позициях. Определим также другую квадратную матрицу $[N]$, которая имеет порядок $n \times n$ и содержит суммы строк $[B]$ по диагонали и нули в прочих местах. Эти две матрицы содержат столбец и строку маргинальных вероятностей и используются для преобразования матрицы $[B]$:

$$[W] = [N]^{-\frac{1}{2}} [B] [M]^{-\frac{1}{2}} \qquad (6.84)$$

(Так как мы имеем дело с диагональными матрицами, то операции $[N]^{-1/2}$ и $[M]^{-1/2}$ эквивалентны замене каждого диагонального элемента n_{ii} и m_{jj} на $1/\sqrt{n_{jj}}$ и $1/\sqrt{m_{jj}}$. Внедиагональные элементы каждой матрицы, конечно, равны нулю.)

Матрица $[W]$ будет иметь порядок $n \times m$, а ее элементы w_{ij}, – преобразования исходных элементов x_{ij}. Матрица попарных произведений столбцов есть просто

$$[R] = [W]'[W] \qquad (6.85)$$

Аналогично матрица попарных произведений строк есть

$$[Q] = [W][W]' \qquad (6.86)$$

Собственные значения матриц R и Q идентичны, только матрица Q будет иметь $n–m$ дополнительных собственных значений, равных нулю. Собственные векторы матрицы $[R]$ могут быть преобразованы в нагрузки анализа соответствия умножением каждого вектора на соответствующее сингулярное значение, которое равно квадратному корню из соответствующего собственного значения, т.е. нагрузки R-метода равны $\sqrt{\lambda}$, умноженному на собственный вектор R-метода. В матричном обозначечении, использовавшемся ранее, сингулярные значения $[R]$ можно представлять как диагональные элементы матрицы λ порядка $m \times m$, у которой все внедиагональные элементы равны нулю. Собственные векторы матрицы $[R]$ являются столбцами некоторой $m \times m$ матрицы $[U]$. Матричное уравнение, используемое для определения нагрузок R-метода, имеет тогда вид

$$[A^R] = [U][\Lambda] \qquad (6.87)$$

Нагрузки каждого из п наблюдений m факторов анализа соответствия есть просто

$$[S^R] = [W][A^R] \qquad (6.88)$$

Эти нагрузки могут быть нанесены вдоль осей, определенных Q-методом факторного анализа соответствия таким же образом, как главные компоненты или факторные метки.

Если вместо того, чтобы вычислять собственные значения матрицы $[R]$, мы будем вычислять их для матрицы $[Q]$, то сможем вычислить нагрузки и метки Q-метода анализа соответствия. Нагрузки находятся умножением элементов собственных векторов на квадратные корни из соответствующих собственных значений

$$[A^Q] = [V][\Lambda] \qquad (6.89)$$

где V – матрица порядка $n \times n$, столбцы которой содержат n собственных векторов матрицы Q. Метки Q-метода есть

$$[S^Q] = [W]'[A^Q] \qquad (6.90)$$

В силу теоремы Эккарта–Юнга и того, что шкалирование портит матрицу исходных данных как в столбцах, так и в строках, имеется прямая связь между решениями R- и Q-метода:

$$[A^Q] = [W][A^R][\Lambda]^{-1} = [S^R][\Lambda]^{-1} \qquad (6.91)$$

Другими словами, нагрузки Q-метода соответствия равны меткам R-метода анализа соответствия, деленным на соответствующие сингулярные значения. Таким образом, можно получить решение Q-методом, решая задачу R-метода, что дает огромное преимущество в вычислительном плане, так как матрица R обычно значительно меньше по размеру, чем матрица Q.

На один и тот же график можно нанести как наши наблюдения, так и переменные. Это можно сделать, преобразуя нагрузки R-метода и нагрузки Q-метода так, чтобы они были представлены в одной и той же метрике. Шкалирование нагрузок осуществляется по формулам

$$[\bar{A}^R] = [M]^{1/2}[A^R] \quad ; \quad [\bar{A}^Q] = [N]^{1/2}[A^Q] \qquad (6.92)$$

Теперь мы будем использовать геологические данные для проверки «классического» применения анализа соответствия, которое состоит в интерпретации данных, дающих перечисления. Табл. 6.37 содержит данные по числу конодонтов в 10-килограммовых пробах пород, собранных в Восточном Канзасе. Породы миссурийского возраста представляют четыре мегациклотемы или повторения литологических разновидностей, которые отражают циклические изменения условий осадконакопления. Каждая единица классифицировалась как часть идеализированной мегациклотемы; классификации указаны в таблице. Палеонтологи предсказывали, что конодонты, как и некоторые современные морские организмы, связаны с зонами конкретных глубин. Если бы и литология и сходство конодонтов были ответственны за изменения в уровне моря, то анализ соответствия дал бы возможность установить их распространенность, что было бы аналогично изменениям литологии.

Так как имеется 10 видов конодонтов и 20 стратиграфических единиц, то очень удобно построить матрицу сходства между переменными. Табл. 6.38 дает матрицу сходства χ^2, ее собственные значения и наиболее значащие собственные векторы. Также даны нагрузки R- и Q-методов на оси соответствия. Они нанесены на рис. 6.44. Указаны категории мегациклотем для каждой стратиграфической единицы. Как пробы (единицы пород), так и переменные (виды конодонтов) могут быть изображены в одном и том же пространстве.

На рис. 6.45 указаны порядки относительных глубин распространения конодонтов из Миссурийской стратиграфической последовательности. Заметим, что нагрузки R-метода, изображенные

на рис. 6.44, представляющие типы конодонтов, должны отражать глубину воды, на которой эти организмы жили. Конодонты с мелководья представлены на положительном конце фактора *I*, в то время как глубоководные появляются на положительном конце фактора *II*, конодонты с промежуточных (средних) глубин находятся вблизи начала координат. Исследуя положения нагрузок *Q*-метода факторного анализа на тон же диаграмме, можно установить глубины, на которых залегали различные единицы пород. Классификация стратиграфических единиц, представленная в табл. 6.37, заимствована из работы [25]. Заметим, что «внешние сланцы», например сланцы из Ви-лас, Боннер Спрингс, Ченьют, и известняк Фарли, который классифицируется как мелководная залежь, попадают в мелководную часть факторной диаграммы. Фосфатные черные сланцы из Евдора и Манси Грик попадают в глубоководную часть диаграммы, как и тонкозернистые известняки из Стоунер и Рэйтаун. Большая часть единиц породы, однако, расположена вблизи начала координат, включая так называемые «черные сланцы фантом», которые, похоже, не отличаются от большинства типов. горных пород.

Таблица 6.37 Результаты подсчета числа конодонтов, выделенных из 10-килограммовых проб. Столбцы соответствуют видам конодонтов, строки – стратиграфическим единицам разреза миссурийского возраста в Восточном Канзасе. Принята следующая классификация: *Q* – поверхностные сланцы, *S* – мелководный известняк, *V* – верхний известняк, *M* – средний известняк, *P* – «планктонный черный сланец», *B* – черный сланец

Номер	Класс	Порода	Число конодонтов										
			Adetognathus	*Ozarkodina*	*Aethotaxis*	*Idiognathodus delicatus*	*I. elegantulus*	*Magnilaterella*	*Hindeodella*	*Idioprioniodus*	*Gondolatla*	*Others*	Сумма
			A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	
1	M	*South Bend Ls.*	13	10	o	0	37	0	0	0	0	0	60
2	O	*Rock Lake Sh.*	0	o	o	0	11	0	0	0	0	0	11
3	U	*Stoner Ls.*	4	2	1	51	26	1	0	0	0	0	85
4	B	*Eudora Sh.*	0	7	1	207	350	0	0	34	14	3	606
5	M	*Captain Creek Ls.*	8	28	6	0	60	0	0	0	0	0	102
6	O	*Vilas Sh.*	145	20	5	0	10	0	0	0	0	0	180
7	U	*Spring Hill Ls.*	5	134	8	0	353	1	0	4	0	0	505
8	P	*Hickory Creek Ls.*	20	60	0	0	920	0	0	0	0	0	100
9	M	*Meriam Ls.*	115	255	10	0	1140	0	0	0	0	0	1520
10	S	*Bonner Springs Sh.*	1	0	0	0	3	0	0	0	0	0	4
11	S	*Parley Ls.*	31	21	7	0	4	1	0	0	0	0	61
12	–	*Island Creek Sh.*	100	5	0	0	5	0	0	0	0	0	110
13	U	*Argentine Ls.*	0	39	1	0	80	0	1	0	0	0	121
14	P	*Quindaro Sh.*	10	70	0	0	538	0	0	5	0	0	623
15	M	*Frisbee Ls.*	3	78	5	0	450	0	0	3	0	0	539
16	O	*Lane Sh.*	0	0	0	0	28	0	0	0	0	0	28
17	U	*Raytown Ls.*	38	20	3	100	267	3	0	25	0	0	456
18	S	*Muncie Creek Sh.*	15	8	0	243	515	0	10	85	55	13	946
19	M	*Paola Ls.*	10	130	10	200	900	0	0	50	0	0	1300
20	O	*Chanute Sh.*	117	20	0	63	57	0	0	7	0	0	264
Сумма			258	389	31	367	1929	4	0	82	32	7	3104

Таблица 6.38. Матрица сходства χ^2, собственные значения, первые два активных вектора по данным о распространении конодонта и факторные нагрузки, соответствующие первым двум факторам *R*- и *Q*-методов

Матрица сходства χ^2										
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
A	0,3843	0,0037	0,0257	–0,0273	–0,1136	0,0056	–0,0079	–0,0239	–0,0204	–0,0098
B	0,0037	0,0568	0,0196	–0,0645	0,0088	0,0015	–0,0076	–0,0292	–0,0268	–0,0129
C	0,0257	0,0196	0,0216	–0,0119	–0,0117	0,0063	–0,0026	–0,0066	–0,0070	–0,0034
D	–0,0273	–0,0645	–0,0119	0,1655	–0,0477	0,0090	0,0150	0,0620	0,0486	0,0233
E	–0,1136	0,0088	–0,0117	–0,0477	0,0592	–0,0066	–0,0052	–0,0159	–0,0136	0,0066
F	0,0056	0,0015	0,0063	0,0090	–0,0066	0,0075	–0,0010	0,0006	–0,0024	–0,0011
G	–0,0079	–0,0076	–0,0026	0,0150	–0,0052	–0,0010	0,0091	0,0129	0,0179	0,0088
H	–0,0239	–0,0292	–0,0066	0,0620	–0,0159	0,0006	0,0129	0,0365	0,0330	0,0160
I	–0,0204	–0,0268	–0,0070	0,0486	–0,0136	–0,0024	0,0179	0,0330	0,0430	0,0210
J	–0,0098	–0,0129	–0,0034	0,0233	–0,0066	–0,0011	0,0088	0,0160	0,0210	0,0102

Вектор	Собственное значение	Общее сходство (%)	Общее сходство (кумулятивные %)
1	0,4262	53,7003	53,7003
2	0,2634	33,1837	86,8841
3	0,0468	5,8936	92,7776
4	0,0385	4,8532	97,6308
5	0,0101	1,2691	98,8999
6	0,0044	0,5488	99,4487
7	0,0036	0,4523	99,9010
8	0,0008	0,0988	99,9997
9	0,0000	0,0003	100,0000
10	0,0000	0,0000	100,0000

	Собственный вектор	
Конодонт	*I*	*II*
A	0,9467	0,0653
B	0,0357	–0,3489
C	0,0756	–0,0647
D	–0,0945	0,7561
E	–0,2761	–0,2769
F	0,0166	0,0287
G	–0,0248	0,1011
H	–0,0735	0,3242
I	–0,0657	0,2923
J	–0,0316	0,1409

Нагрузка на оси в методе соответствия

R-метод			*Q*-метод		
Конодонты	*I*	*II*	Номер	*I*	*II*
A	2,2655	0,1229	1	0,5628	–0,3344
B	0,0715	–0,5492	2	–0,3362	–0,3372
C	0,6038	–0,4064	3	–0,0968	1,3119
D	–0,1938	1.2193	4	–0,3338	0,7964
E	–0,2195	–0,1730	5	0,1589	–0,5199
F	0,4087	0,5546	6	2,8147	0,0333
G	–0,4512	1,4456	7	–0,1593	–0,5114
H	–0,3037	1,0531	8	–0,2333	–0,3696
I	–0,4768	1,6680	9	0,0349	–0,4195
I	–0,4761	1,6703	10	0,6154	–0,1930
			11	1,8068	–0,3260
			12	3,1445	0,1537
			13	–0,1850	-0,5511
			14	–0,2260	0,3911
			15	–0,2395	–0,4309
			16	–0,3362	–0,3372
			17	0,0168	0,4110
			18	–0,3055	0,8712
			19	–0,2515	0,0998
			20	1,3905	0,5737

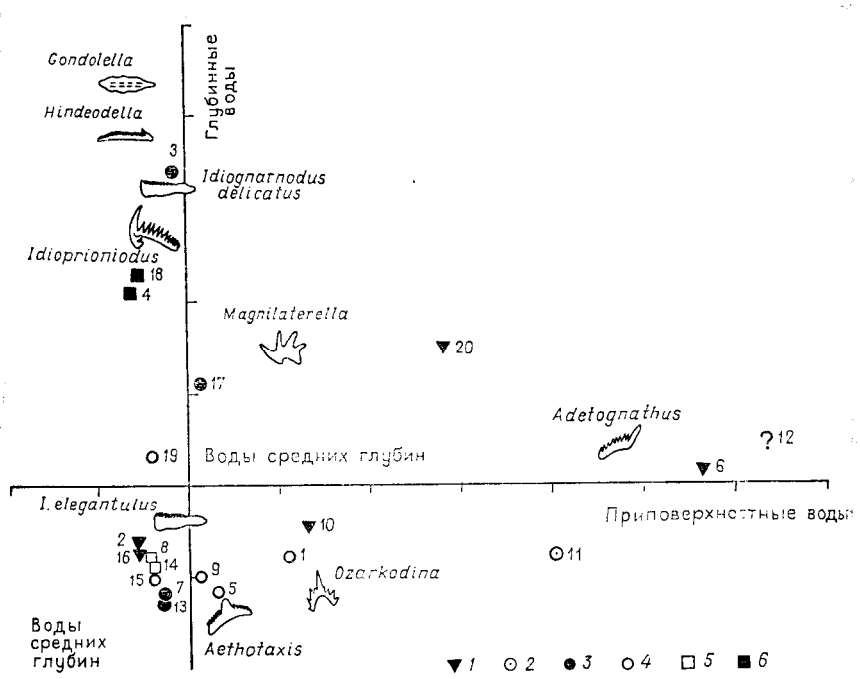


Рис. 6.44. Представление факторных нагрузок метода соответствия для данных по распространенности конодонтов, представленных в табл. 6.37. Указаны приблизительные интервалы значений глубин для различных типов конодонтов. Мегациклотемная классификация стратиграфических единиц включает: прибрежный сланец (1), мелководный известняк (2), верхний известняк (3), средний известняк (4), «черный сланец фантом» (5), черный сланец (6)

Анализ соответствия, проведенный по конодонтам, дает некоторое представление о природе циклического осадконакопления в этой последовательности частного вида. Экстремумы морских фаций, возможно, дают различные литологические разновидности, которые встречаются в характеристических позициях внутри мегациклотем. Большинство литологических типов, однако, не попадает в определенную схему. В частности, «черные сланцы фантом», которые по предположению имеют глубоководные характеристики, кажутся неотличимыми от поро; другого типа, которые имеют происхождение из промежуточных глубин.

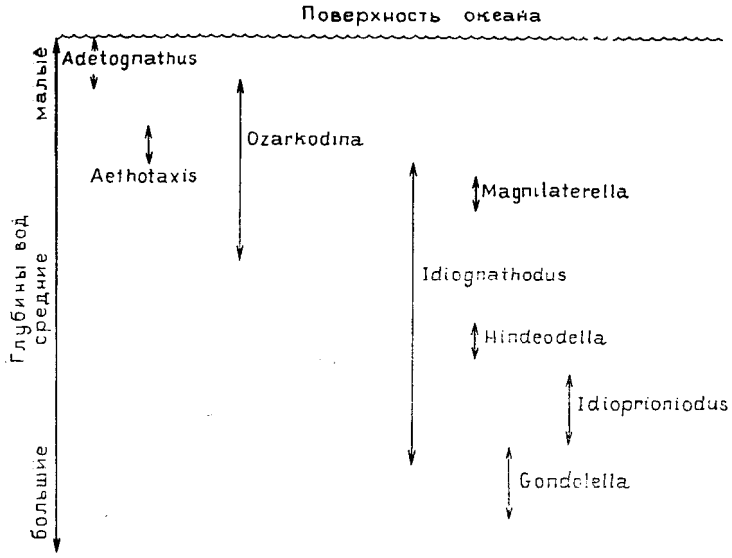


Рис. 6.45. Области значений относительных глубин для конодонтов из Миссурийской стратиграфической последовательности

Применение к непрерывным переменным

В геологии, как и в других областях, анализ соответствия применялся к непрерывным данным, а не к дискретным (см приложения в [60], [12]). Это ставит некоторые концептуальные проблемы, поскольку преобразованные переменные нельзя считать вероятностями, хотя некоторые авторы обращаются с ними таким образом [12]. Так как вообще полная сумма (так же как и суммы строк) состоит из смеси измерений различного типа. то процесс преобразования сильно зависит от

единиц измерения. В силу этого замечания теоретического характера анализ соответствия обычно применяется к множеству интервальных данных и шкал отношений. В этих приложениях процесс преобразования рассматривается как не более чем произвольная процедура, предназначенная для замыкания множества данных и для того, чтобы быть уверенными в том, что строки и столбцы матрицы данных шкалируются эквивалентным образом независимо от того, применяем ли мы *R*- или *Q*-метод. Используемая таким образом мера сходства часто называется «профильным» расстоянием; оно отражает относительную величину переменных, а не их абсолютные значения [66].

Воспользуемся искусственными данными по 25 случайно генерированным блокам для изучения свойств анализа соответствия применительно к измеряемым переменным. Исходные данные приведены в табл. 6.18; в табл. 6.39 представлена матрица сходства, вычисленная с помощью уравнения (6.80), приведены также собственные значения матрицы, собственные векторы, нагрузки *R*- и *Q*-методов на две первые оси, вычисленные методом соответствия. На рис. 6.46 нанесены первые *R*- и *Q*-нагрузки, представленные на одной и той же диаграмме. Из этого сравнительно легко усмотреть не только сходство между индивидуальными наблюдениями, но также и относительные вклады каждой исходной переменной в оси метода соответствия.

Хилл [26] рекомендует более подходящий механизм для обработки непрерывных данных методом соответствия. Если интервальную шкалу или шкалу отношений разделить на дискретные категории и подсчитать число измерений, попадающих в каждую категорию, то данные приведутся к ранговым. Действительно, каждая непрерывная переменная заменится некоторым числом дискретных переменных. Это уменьшает информацию, содержащуюся в множестве данных, но, учитывая неточность многих геологических наблюдений, эта потеря, по-видимому, не является значительной.

В качестве эксперимента выразим данные по блокам в порядковой шкале, разделив множество значений каждой переменной на подходящее число дискретных интервалов (такие, как «низкие», «средние», «высокие» значения) и затем определив, в какую из категорий попадает каждое наблюдение. Необработанная матрица данных поэтому состоит из набора единиц и нулей (табл. 6.40). Затем они преобразуются в совместные вероятности появления, и по этим данным вычисляются *R*- и *Q*-матрицы сходства. Заметим, что матрица $[Q]$ имеет порядок 25×25, а матрица $[R]$ – 21×21. Последовательные собственные значения меньше и уменьшаются медленнее, чем собственные значения матрицы сходства, вычисленной по метрическим данным. Первые два собственных значения, например, составляют лишь 43% следа матрицы сходства. Все собственные значения после четырнадцатого равны нулю. Даже несмотря на то, что несколько первых факторных осей соответствия не выглядят столь же эффективными, как оси, вычисленные из табл. 6.39, изображение нагрузок на оси соответствия показывает картину столь же значимую, как и та, которая была получена 'по метрическим данным. Рис. 6.47 показывает *R*- и *Q*-нагрузки на две первых оси. Сравните эти результаты с результатами, полученными по метрическим данным (рис. 6.46), имея в виду, что информация, содержащаяся в исходных данных, значительно меньше.

Таблица 6.39. Матрица сходства, собственные значения и два первых собственных вектора, вычисленных для данных по случайным блокам; перечислены факторные нагрузки *R*- и *Q*-метода соответствия на первые два фактора

Матрица сходства χ^2							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0,0094	0,0023	0,0089	0,0088	–0,0093	–0,0118	–0,0110
X_2	0,0023	0,0168	0,0142	0,0096	–0,0135	–0,0126	–0,0184
X_3	0,0089	0,0142	0,0376	0,0174	–0,0264	–0,0318	–0,0118
X_4	0,0088	0,0196	0,0174	0,0123	–0,0153	–0,0178	–0,0160
X_5	-0,0093	–0,0135	–0,0264	–0,0153	0,0218	0,0253	0,0148
X_6	-0,0118	–0,0126	–0,0318	–0,0178	0,0253	0,0307	0,0134
X_7	–0,0110	–0,0184	–0,0118	–0,0160	0,0148	0,0134	0,0428

Продолжение табл. 6.39. Матрица сходства, собственные значения и два первых собственных вектора, вычисленных для данных по случайным блокам; перечислены факторные нагрузки *R*- и *Q*-метода соответствия на первые два фактора

Вектор	Собственные значения	Общее сходство (%)	Общее сходство (кумулятивные %)
1	0,1213	70,8332	70,8332
2	0,0359	20,9437	91,7769
3	0,0110	6,4431	98,2201
4	0,0029	1,6891	99,9091
5	0,0001	0,0837	99,9928
6	0,0000	0,0072	100,0000
7	0,0000	0,0000	100,0000

Собственный вектор		
Переменная	*I*	*II*
X_1	0,1922	–0,0538
X_2	0,2783	–0,2148
X_3	0,4961	0,4085
X_4	0,3109	–0,0447
X_5	–0,4126	–0,1458
X_6	–0,4710	–0,2967
X_7	–0,3883	0,8203

Нагрузки, на оси в методе соответствия					
R-метод			*Q*-метод		
Переменная	*I*	*II*	Блок	*I*	*II*
X_1	0,1555	–0,0237	*a*	–0,5814	0,0088
X_2	0,2747	–0,1153	*b*	–0,1290	–0,1257
X_3	0,6980	0,3126	*c*	0,4855	0,0899
X_4	0,2219	–0,0173	*d*	0,6570	0,1459
X_5	–0,3809	–0,0732	*e*	0,1998	–0,0955
X_6	–0,3793	–0,1299	*f*	–0,1507	–0,0214
X_7	–0,5467	0,6280	*g*	–0,8040	0,5665
			h	0,3164	–0,0467
			i	–0,2322	–0,2435
			j	–0,3215	–0,0086
			k	–0,2154	–0,0698
			l	0,3782	0,0457
			m	0,1067	0,1565
			n	–0,2370	0,2185
			o	–0,2046	0,7453
			p	–0,0421	–0,0861
			q	–0,0520	–0,2084
			r	0,1873	–0,0313
			s	–0,2654	–0,0455
			t	–0,0217	0,0662
			u	–0,2393	–0,2097
			v	0,4181	0,0395
			w	0,5015	0,2009
			x	0,1477	0,0399
			y	0,5008	0,0539

Таблица 6.40. Данные по блокам: каждая исходная переменная представлена тремя группами (*L* – низкие, *M* – средние, *H* – высокие значения); в результате получается матрица исходных данных порядка 21×25

Блок	X_1			X_2			X_3			X_4			X_5			X_6			X_7		
	L	*M*	*N*	*L*	*M*	*N*	*L*	*M*	*N*	*L*	*M*	*N*	*L*	*M*	*N*	*L*	*M*	*N*	*L*	*M*	*N*
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0
b	0	0	1	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0
c	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	1	0	0
d	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
e	0	0	1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0
f	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0
g	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1
h	0	0	1	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
i	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0
j	0	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0
k	0	0	1	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0
l	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
m	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
n	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0
o	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0
p	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0
q	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0
r	0	0	1	0	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
s	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0
t	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
u	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0
v	0	0	1	0	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
w	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
x	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
y	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0

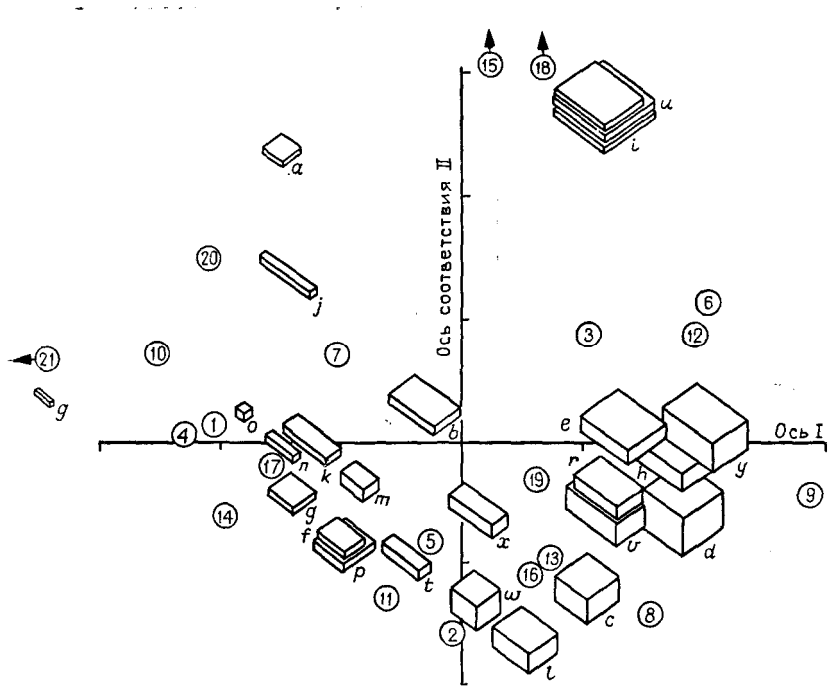


Рис. 6.47. Представление нагрузок *R*- и *Q*-методов на первые две оси соответствия для данных по блокам в порядке возрастания их объемов. Цифры в кружках – нагрузки *R*-метода (см. табл. 6.40. Значения 15, 18 и 21 – вне области диаграммы)

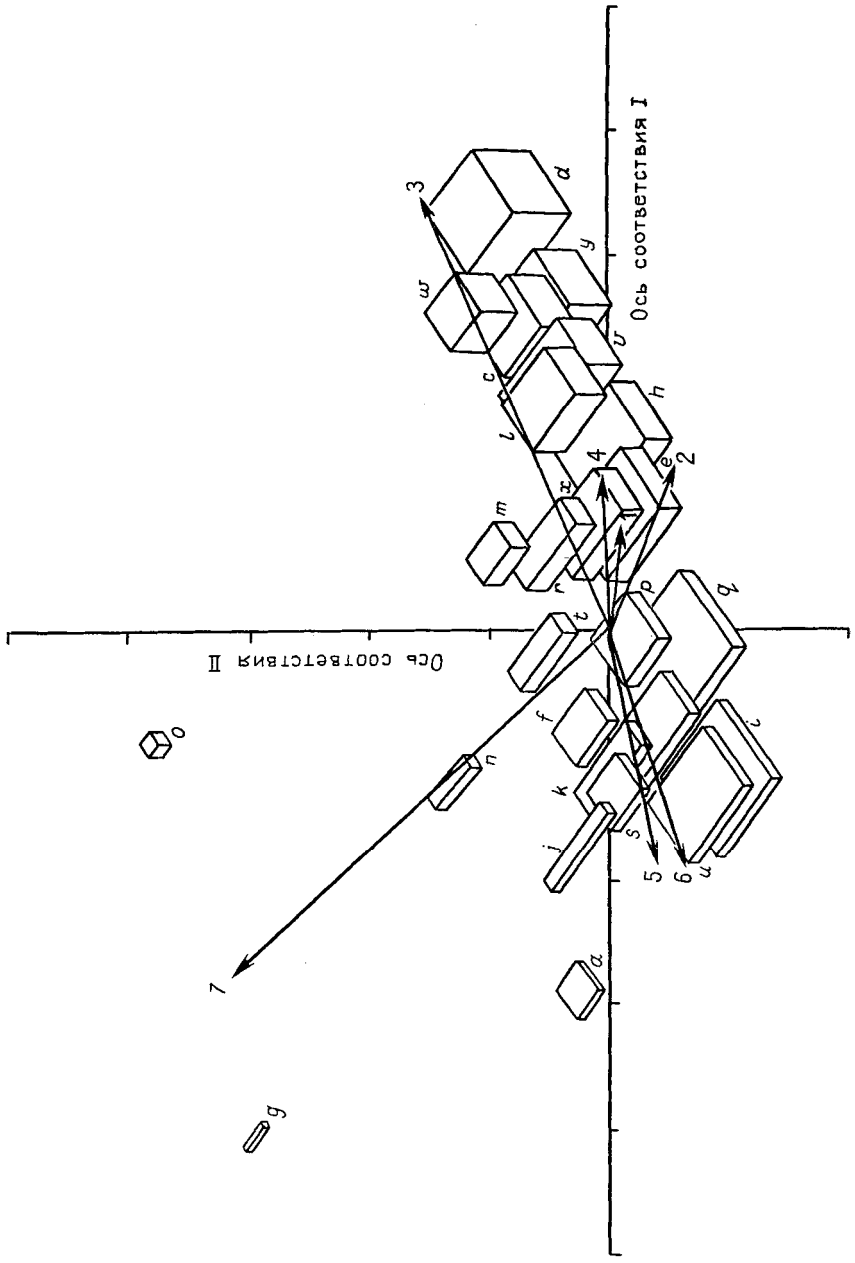


Рис. 6.46. Представление нагрузок *R*- и *Q*-методов анализа соответствия на первые две оси для данных по блокам

СОВМЕСТНЫЙ *R*- И *Q*-ФАКТОРНЫЙ АНАЛИЗ

Хотя теорема Эккарта–Юнга утверждает, что *R*- и *Q*-методы дают эквивалентные решения, на практике это не всегда так. Достаточно взглянуть на рис. 6.38 и 6.39, чтобы убедиться в том, что факторные метки *R*-метода выглядят иначе, чем факторные нагрузки *Q*-метода. Напомним, что решение *R*-метода получается из симметричной матрицы-произведения $[W]'[W]$, имеющей меньший порядок, в то время как решение *Q*-метода получается из произведения матриц $[W][W]'$, имеющего больший порядок. К сожалению, шкалирование, используемое для получения матрицы $[W]$ из матрицы необработанных данных $[X]$, не одинаково в этих двух методах. Например, глазные компоненты содержат преобразование каждого элемента матрицы $[X]$, состоящее в делении на стандартное

отклонение столбцов, что приводит к шкалированной матрице данных $[W]$. Q-метод факторного анализа использует стандартизацию, состоящую в делении каждого элемента $[X]$ на квадратный корень из суммы квадратов строк, что приводит к шкалированной матрице $[W]$. Однако матрица $[W]$ в анализе главных компонент не идентична матрице $[W]$, полученной на первом шаге Q-метода факторного анализа. Различие в шкалировании ухудшает решение в одном методе по отношению к другому.

Имеется несколько подходов к этой задаче. Очевидно, если не проводится никакого шкалирования, то собственные векторы и собственные значения для $[X]'[X]$ такие же, как и для $[X][X]'$, исключая лишь то различие, что одна из этих матриц имеет дополнительные нулевые собственные значения. Факторные метки R-метода будут пропорциональны нагрузкам Q-метода, и наоборот. Кроме того, R- и Q-факторные нагрузки располагаются в том же пространстве, что и факторные оси, поэтому их можно нанести на одну и ту же диаграмму, как это представлено на рис. 6.48.

К сожалению, использование необработанной матрицы непарных произведений имеет большие неудобства. Так как при этом не проводится никакого шкалирования, то результаты анализа очень чувствительны к выбору единиц измерения и могут просто отражать средние величины переменных, а не свойства дисперсий и ковариаций. Хотя, по-видимому, такой метод – наиболее простой способ получения факторов R- и Q-методов. он почти никогда не используется на практике.

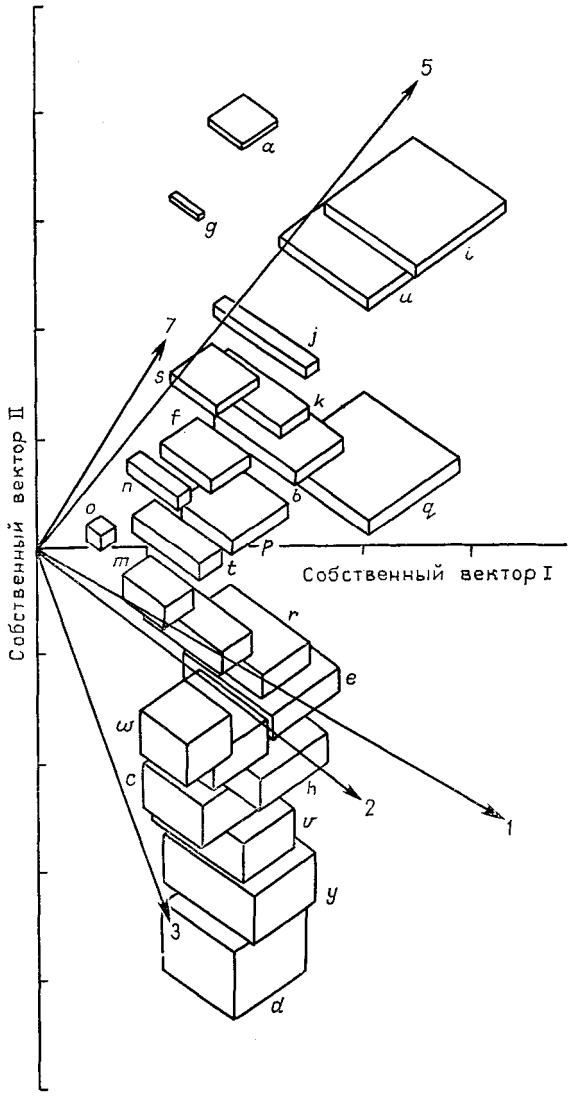


Рис. 6.48. Представление нагрузок *R*- и *Q*-методов на две первые факторныеоси для данных по блокам, вычисленных для матрицы необработанных сумм квадратов и попарных произведений [X]'[X]. Блоки представлены как объекты, характеризуемые нагрузками Q-метода; переменные – как векторы, определенные нагрузкам R-метода. Конец оси X_4 находится за пределами диаграммы в направлении оси X_2; конец оси X_6 – за пределами диаграммы в направлении оси X_5

Второй подход состоит в шкалировании матрицы $[X]$ так, чтобы строки и столбцы можно было трактовать одинаковым образом. В анализе соответствия это делается путем деления каждого элемента на произведение квадратных корней из сумм по строкам и столбцам. Хотя этот метод очень чувствителен, его недостатки менее заметны, если мы имеем дело с условными таблицами и если применяем его к результатам измерений.

Третий подход – это поиск способа такого шкалирования строк, который дал бы содержательную меру взаимных связей между строками матрицы $[W]$, и в то же время получалась бы содержательная мера взаимосвязей между столбцами. Эта задача оказывается не столь сложной, как кажется на первый взгляд, и это является базисом по меньшей мере двух практических методов одновременного нахождения факторов R- и Q-методов.

Элементы матрицы $[X]$ стандартизуются с помощью вычитания средних столбца (переменной) и деления на квадратный корень из п числа наблюдений, т.е.

$$w_{ij} = (x_{ij} - \bar{x}_j)\Big/\sqrt{n} \qquad (6.93)$$

Тогда меньшая матрица-произведение $[W]'[W]$ будет содержать дисперсии и ковариации переменных. В то же время большая матрица-произведение $[W][W]'$ эквивалентна матрице главных координат Q, когда сходство между объектами определяется евклидовым расстоянием, т.е.

$$q_{ij} = d_{ij} + d_{\bullet\bullet} - (\bar{d}_{i\bullet} + \bar{d}_{\bullet j}) \qquad (6.94),$$

и d_{ij} – элемент матрицы расстояний $[D]$:

$$d_{ij} = \sum_{i=1}^{m} (x_{ik} - x_{jk}) \Big/ \sqrt{n} \qquad (6.95)$$

С другой стороны, можно стандартизовать элементы матрицы: $[X]$, вычитая среднее по столбцу (переменной) и выполняя деление на произведение стандартных отклонений и квадратного, корня из n:

$$w_{ik} = \frac{x_{ik} - \bar{x}_{\bullet k}}{s_k \sqrt{n}} \qquad (6.96)$$

Меньшая матрица-произведение $[W]'[W]$ будет теперь содержать дисперсии и ковариации переменных в стандартизованной форме, которые есть не что иное как коэффициенты корреляции между переменными.

Снова большая матрица-произведение $[W][W]'$ эквивалентна одной из версий матрицы главных координат $[Q]$. Теперь, однако, матрица расстояний $[D]$ содержит евклидовы расстояния между наблюдениями, как определено стандартизованными переменными, или

$$d_{ij} = \sum_{k=1}^{m} \left(\frac{x_{ik} - x_{jk}}{s_k} \right) \Big/ \sqrt{n} \qquad (6.97)$$

Чтобы осуществить совместный R- и Q-факторный анализ, мы сначала вычислим меньшую матрицу-произведение $[W]'[W]$ после шкалирования по формулам (6.93) или (6.96). Затем вычислим собственные векторы и собственные значения, а также R-факторные нагрузки, умножая каждый элемент собственного вектора на соответствующее сингулярное значение или квадратный корень из его собственного значения. Если матрицу собственных векторов обозначить через $[U]$, то будем иметь

$$[A^R] = [U] \cdot [\Lambda] \qquad (6.98)$$

где снова Λ – диагональная матрица сингулярных значений $[W]'[W]$. Далее вычислим Q-факторные нагрузки, которые находятся как произведение шкалированной матрицы данных на матрицы собственных векторов:

$$[A^Q] = [W] \cdot [U] \qquad (6.99)$$

Конечно, можно также найти факторные метки, R-факторные метки находятся умножением шкалированной матрицы данных на матрицу R-факторных нагрузок:

$$[S^R] = [W] \cdot [A^R] \qquad (6.100)$$

Q-факторные нагрузки находятся по формуле

$$[S^Q] = [W]'[A^Q] = [W]'[W][U] \qquad (6.101)$$

Так как R-факторные нагрузки $[A^R]$ дают координаты переменных как точки факторного пространства и Q-факторные нагрузки $[A^Q]$ дают координаты объектов как точки того же факторного пространства, то обе совокупности факторных нагрузок можно нанести на одни и те же факторные диаграммы. Переменные, которые нанесены близко друг к другу, очень похожи. Теорема Экарта–Юнга дает соотношение между переменными и объектами. Уравнение (6.40) может быть переписано в виде

$$[A^Q][\Lambda]^{-\frac{1}{2}}[A^R]' = [W] \qquad (6.102)$$

Элемент матрицы $[W]$, таким образом, равен

$$\frac{1}{\sqrt{\lambda_k}} \sum_{k=1}^{m} a_{ik}^Q a_{jk}^R = w_{ij} \qquad (6.103)$$

Значение w_{ij}, которое является шкалированным значением переменной j, наблюдаемой на объекте i, можно рассматривать как произведение вектора нагрузок объекта $[a_i^Q]$ и вектора нагрузок перемен-

ной $[a_j^R]$, умноженного на $1/\sqrt{\lambda_k}$. Величина векторного произведения, как указывалось в гл. 5 (см. гл. 5), обратным образом связана с расстоянием между концами двух векторов. Таким образом, сила связи между объектом и переменной в факторной диаграмме прямо выражается как расстояние между точкой объекта и переменной точкой.

Хотя между факторными метками существуют эквивалентные соотношения, все же они не очень хорошо выражаются в терминах сходства. Наилучший способ изображения результатов одновременного *R*- и *Q*-факторного анализа – это представление двух множеств факторных нагрузок в проекции на факторные оси. Это сделано для данных по случайным блокам на рис. 6.49. Как для *R*-метода, так и для метода главных компонент, критические матрицы будут одинаковыми (см. табл. 6.21).

Двойственность между анализом главных компонент и анализом главных координат, основанном на использовании евклидова расстояния, была впервые указана Говером [20]. Однако эта двойственность до недавнего времени не использовалась, хотя множество программ факторного анализа использует сингулярное разложение матриц, которое является следствием теоремы Эккарта–Юнга.

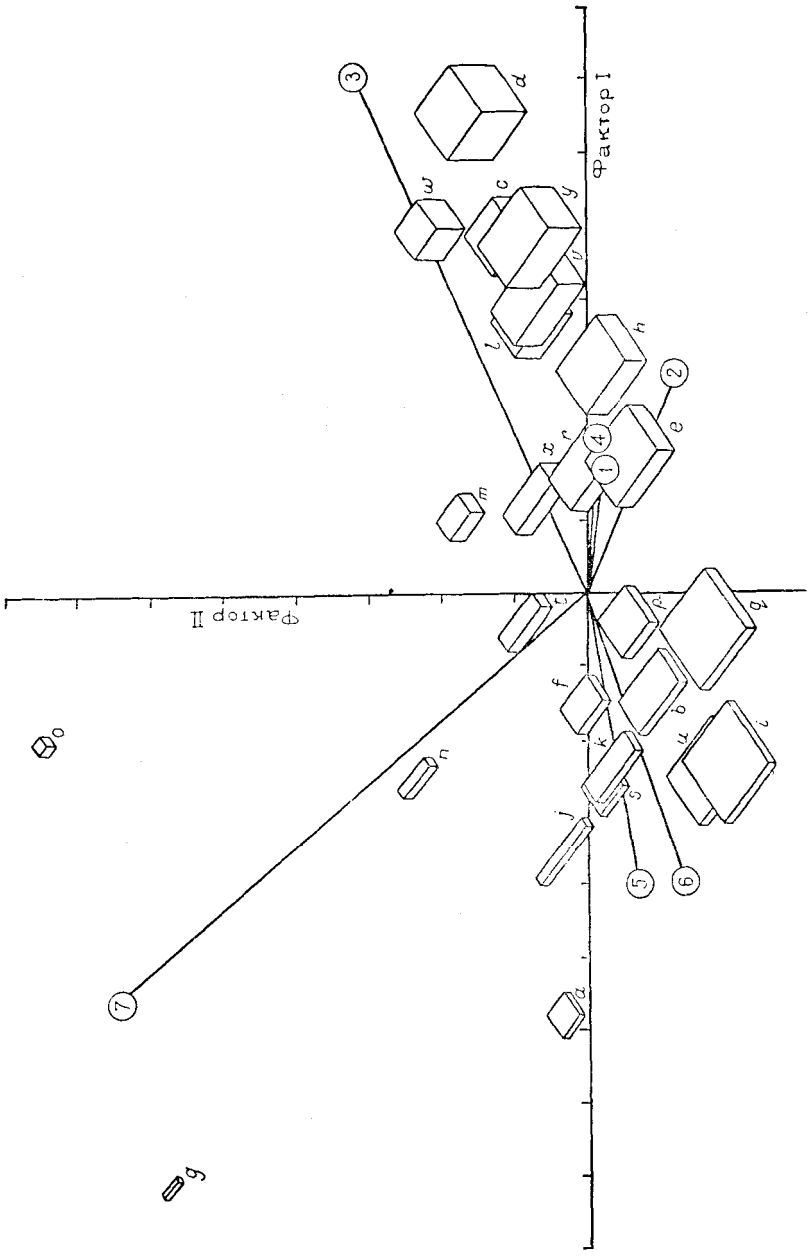


Рис. 6.49. Представление нагрузок *R*- и *Q*-методов на первые две факторные оси для данных по блокам, вычисленных из матрицы дисперсии и ковариации [*W*]'[*W*]. Блоки представлены как объекты, соответствующие нагрузкам *Q*-метода. Переменные представлены векторами, характеризующимися нагрузками *R*-метода

Математическое обоснование одновременного *R*- и *Q*-факторного анализа дано Зу, Чангом и Девисом [66], которые также приводят большое число примеров его использования в геологии.

В табл. 6.41 содержатся измерения, сделанные по профилю через небольшой высокорадиоактизный кварц-монцонитовый плутон, интрудированный в хлорит-актинолитовый сланец вблизи Береа, штат Виргиния. По этому профилю анализировались на содержание радиоактивных урана, тория и калия 22 образца из буровых скважин. В тех же местах вдоль профиля были сделаны воздушные радиометрические измерения. Цель исследования, первоначально проведенного Шерманом, Банкером и Вашем [57], состояла в установлении связи между содержаниями радиоактивных элементов и радиометрическими измерениями.

Данные были проанализированы Зу, Чангом и Девисом [66] с использованием шкалирования

по формуле (6.96). В табл. 6.42 приведены матрица коэффициентов корреляции, собственные значения, *R*- и *Q*-факторные нагрузки. На рис. 6.50 представлены *R*- и *Q*-факторные нагрузки на первые две оси в факторном пространстве. Очевидно различие между пробами, взятыми из плутона, и пробами, собранными в сланцах материнской породы и аллювиального покрытия.

Таблица 6.41. Содержания урана, тория и калия и интенсивность результатов воздушной радиометрической разведки (ВР), проведенной вдоль разреза интрузии кварцевых монцонитов вблизи Берея, штат Виргиния [57]

Номер	ВР*	U, г/т	Th, г/т	К, %
1	240	0,63	2,05	0,13
2	360	2,18	5,31	0,31
3	420	2,26	5,61	0,34
4	500	1,71	6,44	0,70
5	580	2,38	7,99	1,73
6	700	3,83	8,32	4,26
7	600	3,79	9,46	1,53
8	650	4,09	14,71	3,11
9	770	4,21	12,00	1,90
10	930	4,72	12,78	2,92
11	1020	6,24	16,31	2,29
12	1000	5,24	14,51	1,88
13	1000	4,73	15,79f	4,64
14	1040	4,67	10,30	4,17
15	1150	5,08	13,11	3,97
16	1000	5,27	13,40	4,36
17	960	5,61	10,31	2,05
18	420	2,33	6,83	0,47
19	370	2,64	9,88	0,58
20	400	2,29	6,02	0,34
21	480	2.32	6,14	0,32
22	730	5,94	12,86	1,35

* Воздушные радиометрические измерения в расчете на 1 с.

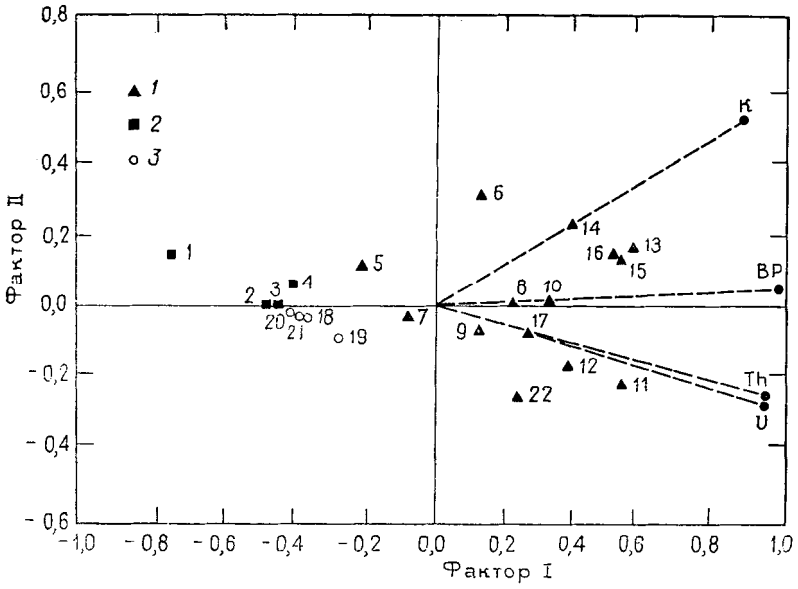


Рис. 6.50. Представление нагрузок *R*- и *Q*-методов на две первые факторные оси для данных Береа, вычисленных из ковариационной матрицы (см. табл. 6.42) [65]. Пунктирные линии – нагрузки *R*-метода; 1 – кварцевый монцонит; 2 – кристаллический сланец; 3 – песок и гравий

Таблица 6.42. Матрица коэффициентов корреляции, собственные значения, *R*- и *Q*-факторные нагрузки по данным состава и радиометрическим данным из Береа, штат Виргиния

Матрица коэффициентов корреляции порядка 4×4				
	BP	U	Th	K
BP	1,00			
U	0,89	1,00		
Th	0,82	0,89	1,00	
K	0,82	0,67	0,69	1,00
Собственные значения				
	Факторы			
	I	*II*	*III*	*IV*
λ	3,39	0,39	0,15	0,06
%	84,81	9,76	3,87	1,55
Σ, %	84,81	94,57	98,45	100,0
Нагрузки *Q*-метода				
	Факторы			
	I	*II*	*III*	*IV*
BP	0,96	0,05	–0,22	–0,16
U	0,94	–0,27	–0,13	0,17
Th	0,92	–0,25	0,28	–0,07
K	0,86	0,51	0,09	0,07
Нагрузки *Q*-метода				
	Факторы			
	I	*II*	*III*	*IV*
1	–0,75	0,12	–0,03	–0,01
2	–0,49	–0,02	–0,02	0,03
3	–0,45	–0,02	–0,04	0,00
4	–0,41	0,04	–0,00	–0,08
5	–0,22	0,09	0,03	–0,05
6	0,11	0,29	–0,00	0,12
7	–0,08	–0,05	0,00	0,04
8	0,21	0,00	0,23	0,03
9	0,12	–0,08	0,02	–0,03
10	0,31	0,00	–0,01	–0,03
11	0,51	–0,23	0,00	–0,01
12	0,35	–0,18	–0,03	–0,09
13	0,53	0,14	0,14	–0,04
14	0,36	0,21	–0,11	–0,00
15	0,50	0,11	–0,08	–0,07
16	0,49	0,13	0,01	0,03
17	0,26	–0,09	–0,19	0,04
18	–0,40	–0,04	0,01	–0,00
19	–0,31	–0,11	0,15	0,01
20	–0,44	—0,03	–0,01	0,0
21	–0,41	–0,03	–0,05	–0,03
22	0,21	–0,26	–0,03	0,12

МНОГОГРУППОВЫЕ ДИСКРИМИНАНТНЫЕ ФУНКЦИИ

Многогрупповой дискриминантный анализ вобрал в себя наиболее ценные черты дисперсионного анализа и связанные с ними вычислительные процедуры факторного анализа. Эта задача является обобщением уже рассмотренной процедуры дискриминантного анализа, связанной с разбиением на две группы. В качестве примера рассмотрим задачу из палеонтологии, связанную с изучением полового диморфизма гастропод по образцам, взятым из нескольких различных мест. По мягкой части их тел исследователь легко отличит самку от самца. Особи можно классифицировать как самцы из совокупности A, самки из совокупности A, самцы из совокупности B и т.д. Многомерные измерения, сделанные на раковинах, используются в дискриминантном анализе для нахождения комбинаций измерений, которые позволяют различить как их пол, так и совокупности, из которых они взяты. К счастью, различия пола оказываются более ярко выраженными, чем различия между совокупностями. Этот метод дает возможность выделить характеристики раковин гастропод и классифицировать ископаемые останки раковин в соответствии с их полом.

Аналогия с дисперсионным анализом состоит в том, как дисперсии и ковариации разделяются на категории или группы. Вспомним, как в гл. 2 в одномерном дисперсионном анализе общая сумма квадратов SS_T равнялась сумме квадратов в каждой группе SS_w плюс сумма квадратов между группами SS_B. Точно такая же структура удобна и в дискриминантном анализе.

Обозначения многогруппового дискриминантного анализа достаточно сложны, так как требуется рассмотреть не только объекты и переменные, но и группы, в которых лежат эти объекты. В силу этого нужно сначала использовать обозначения с тремя индексами, например, x_{ijk} обозначает t-ю переменную, измеренную на j-м объекте внутри группы k. Дело осложняется тем, что не все группы содержат одинаковое число объектов, поэтому необходимо ввести обозначение для числа объектов в k-й группе – n_k. Предположим, что наблюдения классифицируются в g различных групп. Если собрать все группы вместе, то мы найдем, что во всем множестве имеется всего $N = \sum_{k=1}^{g} n_k$ объектов, каждый из которых характеризуется некоторым набором переменных.

Среднее i-й переменной в k-й группе есть

$$\overline{X}_{i\cdot k} = \sum_{j=1}^{n_k} \frac{x_{ijk}}{n_k} \quad (6.104)$$

Общее среднее i-й переменной есть среднее всех наблюдений переменной i, несмотря на группы, в которых находятся наблюдения. Это общее среднее равно

$$\overline{\overline{X}}_{i\cdot\cdot} = \frac{1}{N} \sum_{k=1}^{g} \sum_{j=1}^{n_k} x_{ijk} \qquad (6.105)$$

Коэффициент ковариации между переменными i и l для всех объектов, несмотря на группы, равен

$$s_{ij} = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (x_{ijk} - \overline{\overline{X}}_{i\cdot\cdot})(x_{ijk} - \overline{\overline{X}}_{l\cdot\cdot}) \qquad (6.106)$$

Если вычислить эту меру для всех возможных пар переменных, то получим симметрическую матрицу $[S]$ порядка $p \times p$, которая будет называться матрицей общих сумм произведений. Вычислим также меру ковариации между переменными i и l внутри g групп по формуле

$$w_{il} = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (x_{ijk} - \overline{X}_{i\cdot k})(x_{ijk} - \overline{X}_{i\cdot k}) \qquad (6.107)$$

Для всех возможных пар переменных эти коэффициенты образуют матрицу $[W]$, которая представляет сумму произведений между группами. Эта матрица эквивалентна сумме матриц $[SPA]$ и $[SPB]$, используемых в простом дискриминантном анализе для двух групп. Наконец, выразим дисперсию между группами:

$$b_{ij} = \sum_{k=1}^{g} (\overline{X}_{i\cdot k} - \overline{\overline{X}}_{i\cdot\cdot})(\overline{X}_{l\cdot k} - \overline{\overline{X}}_{i\cdot\cdot}) \qquad (6.108)$$

Это также матрица $[B]$ порядка $p \times p$, которая содержит межгрупповые суммы произведений.

Как и в обычном дисперсионном анализе, внутригрупповые и межгрупповые суммы квадратов, сложенные вместе, дают общую сумму произведений

$$[S] = [B] + [W] \qquad (6.109)$$

Желательно, чтобы отношение $[B]/[W]$ было по возможности велико. Легко убедиться, что это отношение является многомерным аналогом отношения F, заданного по формуле $F = MS_B/MS_W$, используемого для проверки различия между группами в дисперсионном анализе. Если это отношение велико, то это значит, что группы широко разбросаны, в то время как наблюдения внутри групп плотно собраны вокруг своих средних.

Задача дискриминантного анализа состоит в нахождении множества линейных весов для переменных так, чтобы это отношение было максимальным. Если считать, что это множество весов образует вектор $[A_1]$, то дискриминантный анализ можно трактовать как задачу нахождения элементов этого вектора $[A_1]$ таким образом, чтобы отношение

$$\frac{[A_1]'[B][A_1]}{[A_1]'[W][A_1]}$$

достигало максимума. Конечно, на вектор $[A_1]$ необходимо наложить некоторые ограничения. В дискриминантном анализе обычно накладывается следующее ограничение: знаменатель этого выражения должен быть равен единице, т.е. $[A_1]'[W][A_1] = 1$.

При выполнении этого условия отношение будет достигать максимума тогда, когда $[A_1]$ – собственный вектор матрицы $[W]^{-1}[B]$ соответствует наибольшему собственному значению. Можно найти второе множество линейных весов $[A_2]$, которые являются элементами собственного вектора, соответствующего второму по величине собственному значению. Аналогично можно найти третье множество весов, четвертое и т.д. Таким образом, мы вычислим последовательность дискриминантных функций, которые дают разделение на заранее заданные группы настолько хорошо, насколько это возможно. В силу природы собственных векторов они ортогональны друг другу, и каждый следующий является вектором, дающим наилучшее разделение. Можно вычислить дискриминантную функцию для каждого положительного собственного значения. В общем случае число положительных собственных значений будет равно наименьшему из чисел $g–1$ или p. К сожалению, матрица, полученная по формуле $[W]^{-1}[B]$, не является симметричной, и поэтому ее собственные векторы находятся нелегко. В некоторых программах дискриминантного анализа собственные векторы находятся итерационными методами, основанными на процессе, называемом разложением на особые значения [7]. Другие программы сначала преобразовывают матрицу в симметричную форму, а затем находится множество собственных векторов, которое в свою очередь преобразуется в требуемое множество. Этот метод описан в [19], критические шаги пояснены в [47].

Наблюдения, используемые в вычислении дискриминантной функции, можно спроектировать на пространство, определенное дискриминантными осями. Это делается с помощью матричного умножения

$$Z = [A]'[X] \qquad (6.110)$$

где $[X]$ – исходная матрица данных порядка $N \times p$; $[A]$ – матрица порядка $p \times t$, столбцы которой состоят из t собственных векторов, соответствующих наибольшим собственным значениям, которые используются в дискриминантных функциях. Центроиды g групп можно спроектировать на дискриминантное пространство по формуле

$$[\overline{Z}] = [A]'[\overline{X}_k] \qquad (6.111)$$

где матрица $[X_k]$ имеет порядок $g \times p$ и состоит из средних всех переменных для каждой группы. Ограничим свое внимание на парах дискриминантных функций (обычно это первая и вторая) и нанесем наблюдения и центроиды групп на диаграмму рассеяния. Обычно предварительно данные шкалируются. В некоторых программах производится стандартизация, а из каждого наблюдения вычитается общее среднее и результат делится на стандартное отклонение, вычисленное по всему множеству данных. В других программах деление производится на объединенные внутригрупповые стандартные отклонения. Мараскилло и Левин [46] дают поучительное сравнение различных подходов.

Очевидно, наблюдение неизвестного происхождения можно спроектировать на дискриминантное пространство, просто умножая его слева на транспозицию матрицы $[A]$. Групповая принадлежность нового наблюдения становится очевидной из его положения на диаграмме рассеяния, од-

нако так же можно вычислить меру ее расстояния до центроида каждой группы. Новое наблюдение классифицируется как принадлежность к ближайшей группе.

Вычислить обобщенные расстояния от нового наблюдения до каждого из групповых центроидов можно, определив все разности $(x_i - \overline{X}_{i\bullet k})$, которые удобно расположить в виде матрицы U порядка $g \times p$. Тогда

$$[D^2] = [U]'[A][A]'[U] \qquad (6.112)$$

Это даст нам обобщенные расстояния от нового наблюдения до каждой из g групп, измеренные в t-мерном дискриминантном пространстве. С другой стороны, можно вычислить

$$[D^2] = [U]'[W]^{-1}[U] \quad (6.113)$$

что дает обобщенные расстояния от нового наблюдения до центроидов каждой группы, измеренные в исходном p-мерном пространстве. Объяснение этого и других используемых определений сходства между наблюдением и центроидами различных групп приводится в [19], где также дается метод построения доверительных областей относительно центроидов.

Дискриминантные функции – удобный метод определения групп, если несколько групп, предполагаемые различными, оказываются действительно различными. Рассмотрим одно приложение такого типа.

Когда осадочные породы формируются в морских условиях, в них остается морская вода. Химический состав этой реликтовой воды в последующем изменяется благодаря ионному обмену и другим реакциям, смешиванию с другими морскими водами и разбавлению за счет .просачивания поверхностных вод. Тем не менее, реликтовые воды, полученные при канатном бурении, могут иметь составные характеристики, которые дают ключ к происхождению или фациальным условиям формирования содержащих их пород.

Табл. 6.43 содержит анализы солености в виде характеристик (EPM), являющихся эквивалентом частям на миллион для вод нефтяных полей из трех карбонатных толщ в Техасе и Оклахоме. Пробы брались из разных нефтеносных подразделений. Для определения того, различны ли эти данные, к ним был применен дискриминантный анализ. Если они различны, то это наводит на мысль, что анализы солености дают информацию о природе их первоначальных фаций, так как все три исходные породы имеют приблизительно одни и те же литологические характеристики и, соответственно, сходные истории формирования.

Так как имеется три группы, то требуется вычислить только две дискриминантные функции и найти только два положительных собственных значения. Первое из них λ_1=13,29 составляет 93,6% межгрупповой дисперсии, второе λ_2=0,902 дает остальные 6,4%. На рис. 6.51 представлены две оси дискриминантных функций с метками дискриминанта и центроидами трех представленных групп. Метки могут быть стандартизованы так, чтобы их общее среднее равнялось нулю и стандартные отклонения комбинированных данных равнялись нулю.

Первая дискриминантная функция явно отделяет реликтовую воду из доломитов Грейбурга (группа 2) от воды, взятой из доломитов Элленбургера (группа 1) и известняков Виола (группа 3). Различия по второй дискриминантной функции устанавливаются менее явно из-за перекрытий групп 1 и 2 и перекрытий групп 2 и 3. Однако рассматриваемые вместе две дискриминантных функции полностью разделяют три группы. Это обнадеживающее заключение наводит на мысль, что пробы реликтовой воды из формаций, имеющих сходные литологические свойства, могут иметь одинаковые относительно однородные характеристики.

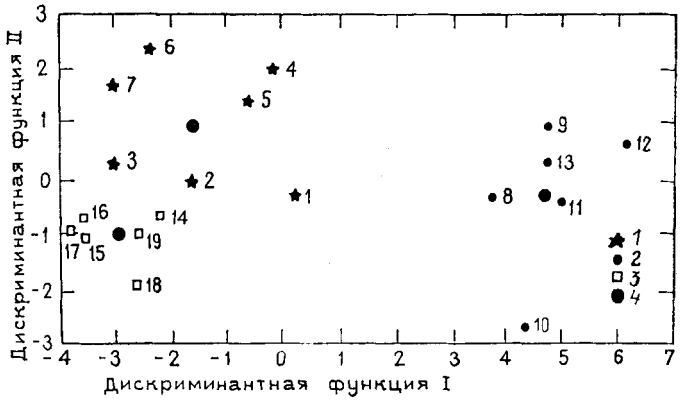


Рис. 6.51. Представление нагрузок на дискриминантные функции *I* и *II*: 1 – доломиты Эленбургера; 2 – доломиты Грейбурга; 3 – известняк Виола; 4 – центроиды групп. 1–19 – номера проб

Таблица 6.43 Химические анализы проб соленой воды, взятых из скважин в трех типах карбонатных отложении в Техасе и Оклахоме [54]

Номер пробы	HCO3	SO_4	Cl	Ca	Mg	Na
Группа 1 – доломиты Элленбургера						
1	10,4	30,0	967,1	95,9	53,7	857,7
2	6,2	29,6	1174,9	111,7	43,9	1054,7
3	2,1	11,4	2387,1	348,3	119,3	1932,4
4	8,5	22,5	2186,1	339,6	73,6	1803,4
5	6,7	32,8	2015,5	287,6	75,1	1691,8
6	3,8	18,9	2175,8	340,4	63,8	1793,9
7	1,5	16,5	2367,0	412,0	95,8	1872,5
$\overline{X}$	5,6	23,1	1896,2	276,5	75,0	1572,3
Группа 2 – доломиты Грейбурга						
8	25,6	0	134,7	12,7	7,1	134,7
9	12,0	104,6	3163,8	95,6	90,1	3093,9
10	9,0	104,6	1342,6	104,9	160,2	1190,1
11	13,7	103,3	2151,6	103,7	70,0	2054,6
12	16,6	92,3	905,1	91,5	50,9	871,4
13	14,1	80,1	554,8	118,9	62,3	472,4
$\overline{X}$	15,2	80,7	1375,4	87,9	73,4	1302,9
Группа 3 – известияки Виола						
14	1,3	10,4	3399,5	532,3	235,6	2642,5
15	3,6	5,2	974,5	147,5	69,0	768,1
16	0,8	9,8	1430,2	295,7	118,4	1027,1
17	1.8	25,6	183,2	35,4	13,5	161,5
18	8,8	3,4	289,9	32,8	22,4	225,2
19	6,3	16,7	360,9	41,9	24,0	318,1
$\overline{X}$	3,8	11,9	1106,4	180,9	80,5	857,1
$\overline{\overline{X}}$	8,0	37,7	1482,3	186,8	76,2	1261,4

КАНОНИЧЕСКАЯ КОРРЕЛЯЦИЯ

Мы снова обратимся к многомерным методам, которые имеют ту же вычислительную основу, что и факторный анализ, однако ало используемым 'понятиям и методам они тесно связаны с множественной регрессией. Вспомним, что множественная регрессия имеет дело с установлением связей между единственной зависимой переменной Y и множеством предсказывающих переменных X_1, X_2, ... Обобщение этого метода состоит в установлении связей между множеством переменных Y и другим множеством переменных X, измеренных на одних и тех же объектах. Эти соотношения определяются с помощью линейных комбинаций переменных X и Y, которые дают наивысшую корреляцию между этими двумя множествами.

Такие корреляции называются каноническими корреляциями, а линейные комбинации называются каноническими переменными. В самом деле, превратим множество переменных X в единственную новую переменную и множество переменных Y в другую единственную новую переменную и затем вычислим корреляцию между этими новыми переменными. Процесс этого превращения линеен, т.е. исходные переменные взвешиваются и складываются вместе, в результате получается каноническая переменная. Применение канонической корреляции может включать определение соот-

ношения между наборами геохимических и петрографических переменных или множеством петрографических откликов каротажа скважин и свойств формации, измеренных на образцах керна, взятых из тех же скважин.

Так как все переменные измерены в одних и тех же пробах, то наблюдения образуют матрицу, размерность которой $n\times(p+q)$, где p – число переменных Y и q – число переменных X. (Для удобства вычислений меньший из этих наборов переменных называем Y, $p\le q$.) Матрица дисперсий и ковариаций $[S]$ имеет порядок $(p+q)\times(p+q)$, и ее можно считать состоящей из четырех блоков: матрица S_{yy} порядка $p\times p$, содержащая дисперсии и ковариации переменных Y; матрица S_{xx} порядка $q\times q$, содержащая дисперсии и ковариации переменных X; матрица S_{xy} порядка $p\times q$ (и ее транспозиция S'_{xy}), которая содержит коэффициенты корреляции между переменными X и Y, т.е.

$$[S]=\left[\begin{array}{c|c} S_{xx} & S_{xy} \\ \hline S'_{xy} & S_{yy} \end{array}\right]$$

Хотя матрица $[S]$ расчленяется, она имеет вид нормальной ковариационной матрицы. Она симметрична относительно диагонали, на которой стоят дисперсии; внедиагональные элементы являются коэффициентами корреляции.

Обозначим подматрицу порядка $n\times p$ матрицы данных, содержащую переменные Y, через $[Y]$, а подматрицу порядка $n\times q$ матрицы данных, содержащую переменные X, через $[X]$. Матрицы $[Y]$ и $[X]$ можно преобразовать с помощью умножения на произвольные векторы, которые в новых координатах будут линейными комбинациями старых:

$$[A]'[Y]$$

$$[B]'[Y]$$

где $[A]$ – вектор порядка $1\times p$, $[B]$ – вектор порядка $1–q$. Дисперсии двух множеств преобразованных переменных будут

$$[A]'[S_{yy}][A]$$

$$[B]'[S_{xx}][B]$$

Ковариации между преобразованными переменными X и Y будут

$$[A]'[S_{xy}]'[B]$$

Цель анализа канонических корреляций состоит в выборе элементов двух векторов $[A]$ и $[B]$ так, чтобы коэффициенты корреляции достигали максимума при условии равенства дисперсий единице. Если с самого начала дисперсии стандартизируются так, чтобы они равнялись единице, то одновременно стандартизируются и коэффициенты ковариации, становясь коэффициентами корреляции между переменными. Используя методы теории собственных значений, можно найти векторы A и B, обладающие желаемыми свойствами. Имеется гарантия того, что каноническая корреляция будет больше, чем наибольшая корреляция между исходными переменными X и любой исходной переменной Y, т.е. больше, чем любой элемент матрицы $[S_{xy}]$. Это объясняется тем, что можно немедленно указать линейную комбинацию, которая будет иметь высокую корреляцию, считая все элементы векторов A и B равными нулю, исключая те, что соответствуют переменным с наивысшей корреляцией, которую надо принять равной единице.

Уравнение, которое требуется решить в этом случае, очень напоминает основное уравнение для нахождения собственных значений в анализе главных компонент:

$$\left|[\Lambda]-\lambda[I]\right|=[0] \qquad (6.114)$$

Здесь $[\Lambda]$ – матрица, получаемая умножением различных блоков расчлененной ковариационной матрицы $[S]$. Матричное умножение дает матрицу-произведение $[\Lambda]$, имеющую порядок $q\times q$ и представляющую объединенную матрицу дисперсий двух множеств переменных, т.е.

$$[\Lambda]=[S_{xx}]^{-1}[S_{xy}]'[S_{yy}]^{-1}[S_{xy}] \qquad (6.115)$$

К сожалению, матрица Λ асимметрична, поэтому для нахождения ее определителя и решения системы уравнений требуется привлечение трудоемких вычислительных процедур.

Собственное значение λ равно квадрату коэффициента корреляции между двумя каноническими переменными. Так как матрица Λ имеет порядок $q\times q$, то существует q различных собствен-

ных значений, каждое из которых представляет коэффициент корреляции между различными парами канонических переменных. Последовательные собственные значения будут уменьшаться по величине и каждая пара канонических переменных будет некоррелирована со всеми другими парами канонических переменных.

Вектор B, используемый для преобразования $[X]$ в канонические переменные, находится с помощью определения собственных векторов, соответствующих уже найденным собственным значениям:

$$([\Lambda]-\lambda[I])[B]=[0]$$

или

$$([S_{xx}]^{-1}[S_{xy}]'[S_{yy}]^{-1}[S_{xy}]-\lambda[I])[B]=[0]$$

Напомним, что собственные векторы вычисляют, подставляя собственные значения в систему совместных уравнений и затем решая ее. Как только вектор преобразования $[B]$ найден, находим эквивалентное каноническое преобразование Y по формуле

$$[A]=[S_{yy}]^{-1}[S_{xy}][B]/\sqrt{\lambda}$$

Конечно, для каждого λ имеется соответствующий вектор $[A]$ и вектор $[B]$. Каждая векторная пара будет преобразованием пары $[X]$ и $[Y]$ в канонические переменные; коэффициент корреляции между новыми переменными будет равен $r=\sqrt{\lambda}$.

Проиллюстрируем метод канонических корреляций, обратившись снова к данным по случайным блокам. Данные естественно попадают в два класса, так как переменные X_1, X_2 и X_3 являются основными измерениями блоков, в то время как переменные от X_4 до X_7 выражаются через них. Первые три переменные можно определить как образующие множество переменных Y, а последующие четыре – как образующие множество переменных X.

В табл. 6.44 приведены стандартная ковариационная (или корреляционная) матрица данных по блокам, расчлененная для канонической корреляции, матрица $[\Lambda]$, для которой требуется найти собственные значения, а также ее собственные значения, канонические корреляции, которые они представляют, и векторы $[A]$ и $[B]$, соответствующие наибольшей канонической корреляции. Мы можем использовать вектор $[A]$ для преобразования переменных Y в метки канонических переменных и вектор $[B]$ для преобразования переменных X в другое множество меток или канонических переменных. На рис. 6.52 указана диаграмма первой пары канонических переменных для данных по блокам. Совершенно очевидно, что имеется явно выраженная линейная связь между двумя множествами, представленными в канонической форме. Рис. 6.53 есть аналог той же диаграммы для второй пары канонических переменных. Диаграмма указывает на очень сильную корреляцию, даже несмотря на то, что преобразование совсем другое.

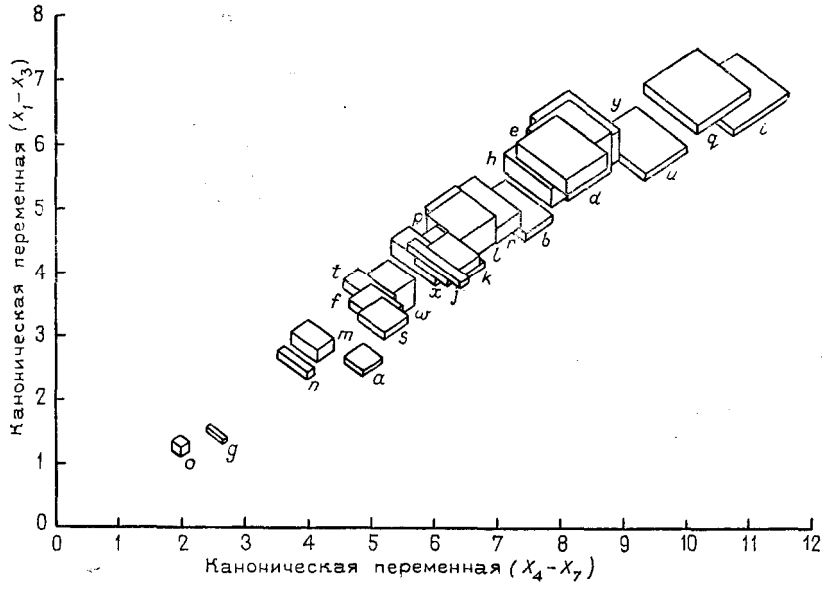


Рис. 6.52. Совместное представление первой пары канонических переменных для данных по случайным блокам. Канонические метки переменных X измерены вдоль горизонтальной оси, а переменных Y – вдоль вертикальной оси. Каноническая корреляция между первой канонической переменной X и первой канонической переменной Y равна R = 1,00

Таблица 6.44. Разбиение матрицы стандартных дисперсии и ковариации (коэффициентов корреляции) для данных по блокам (*а*); собственные значения и соответствующие канонические корреляции (*б*); преобразование векторов с целью превращения переменных по блокам в канонические переменные для трех ненулевых канонических корреляции (*в*); вектора $[A]$ используются для преобразования переменных X_1–X_3 к каноническому виду; векторы $[B]$ – для преобразования переменных X_4–X_7 к каноническому виду

(а)

$$\left[\begin{array}{ccc|cccc}
1{,}0000 & 0{,}5802 & -0{,}2011 & 0{,}9112 & 0{,}2833 & 0{,}2865 & -0{,}5331 \\
0{,}5802 & 1{,}0000 & -0{,}3637 & 0{,}8337 & 0{,}1658 & 0{,}2610 & -0{,}6087 \\
-0{,}2011 & -0{,}3637 & 1{,}0000 & 0{,}4385 & -0{,}7041 & -0{,}6805 & -0{,}6488 \\
\hline
0{,}9112 & 0{,}8337 & 0{,}4385 & 1{,}0000 & 0{,}1630 & 0{,}2022 & -0{,}6755 \\
0{,}2833 & 0{,}1658 & -0{,}7041 & 0{,}1630 & 1{,}0000 & 0{,}9902 & 0{,}4272 \\
0{,}2865 & 0{,}2610 & -0{,}6805 & 0{,}2022 & 0{,}9902 & 1{,}0000 & 0{,}3571 \\
-0{,}5331 & -0{,}6087 & -0{,}6488 & -0{,}6755 & 0{,}4272 & 0{,}3571 & 1{,}0000
\end{array}\right]$$

(*б*)

$$\lambda_1 = 0{,}9999 \quad r_1 = 1{,}00$$
$$\lambda_2 = 0{,}8485 \quad r_2 = 0{,}92$$
$$\lambda_3 = 0{,}6538 \quad r_3 = 0{,}81$$
$$\lambda_{14} = 0{,}0000 \quad r_4 = 0{,}00$$

(*г*)

Первая каноническая кореляция		Вторая каноническая кореляция		Третья каноническая кореляция	
Вектор $[A]$	Вектор $[B]$	Вектор $[A]$	Вектор $[B]$	Вектор $[A]$	Вектор $[B]$
$\begin{bmatrix} 0{,}3453 \\ 0{,}2945 \\ 0{,}1359 \end{bmatrix}$	$\begin{bmatrix} 0{,}6453 \\ -0{,}5731 \\ 0{,}5027 \\ 0{,}0494 \end{bmatrix}$	$\begin{bmatrix} -0{,}0594 \\ 0{,}1438 \\ -0{,}1559 \end{bmatrix}$	$\begin{bmatrix} -0{,}0415 \\ -0{,}6351 \\ 0{,}7712 \\ 0{,}0137 \end{bmatrix}$	$\begin{bmatrix} 0{,}0956 \\ -0{,}0723 \\ -0{,}0404 \end{bmatrix}$	$\begin{bmatrix} -0{,}0321 \\ -0{,}7468 \\ -0{,}6593 \\ -0{,}0826 \end{bmatrix}$

Комментарии, сделанные в разделе о главных компонентах относительно «чтения» нагрузок на компоненты, в равной степени справедливы и для канонического преобразования. Векторы $[A]$ и $[B]$ являются весами, используемыми для преобразования исходных переменных Y и X в канонические переменные. При некоторых обстоятельствах можно указать физические значения комбинации весов частного вида. Канонические переменные можно трактовать по аналогии с факторами. Однако нет уверенности в том, что набор весов будет иметь какую-либо осмысленную интерпретацию, и канонические корреляции могут просто отражать математические комбинации переменных двух множеств. Это в особенности справедливо, если канонические корреляции между X и Y являются слабыми. Мараскилло и Левин [46], дающие исключительно ясное изложение канонической корреляции, рассматривают также и другие виды интерпретации канонических весов.

Существует два часто используемых статистических критерия проверки гипотез о канонической корреляции. Один из них основан на проверке наличия любой значимой канонической корреляции среди канонических соотношений. Этот критерий будет значимым, если одна или более пар канонических переменных коррелированы. Второй критерий использует лишь значимость наибольшей канонической корреляции.

Критерий всеобщей значимости был определен Бартлеттом [3] и имеет вид

$$L = (1-\lambda_1)(1-\lambda_2)\ldots(1-\lambda_q) \quad (6.116)$$

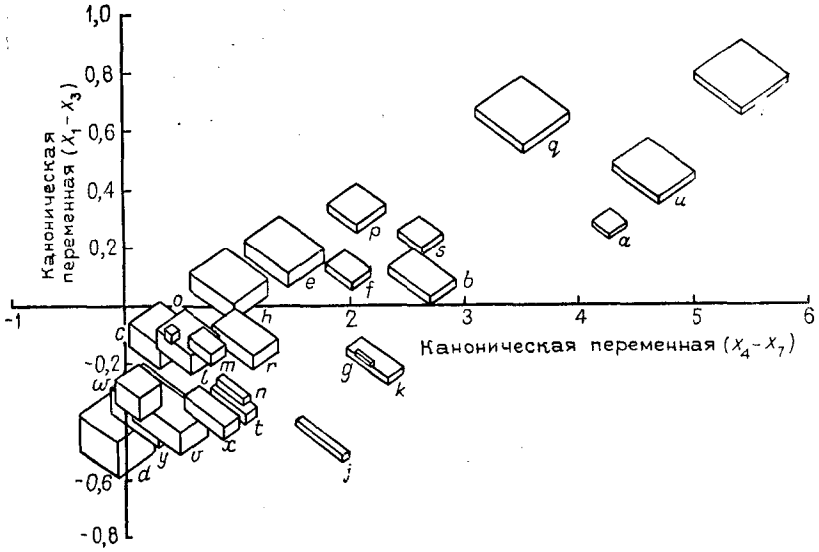


Рис. 6.53. Совместное представление второй пары канонических переменных для данных по случайным блокам.
Канонические метки переменных X и Y измеряются соответственно вдоль горизонтальной и вертикальной осей. Каноническая корреляция между вторым X и вторым Y равна $R = 0{,}85$

Величина

$$\chi^2 = \left\{-(n-1)+\frac{1}{2}(p+q+1)\right\}\ln L \qquad (6.117)$$

распределена по закону χ^2 с pq степенями свободы. Нулевая гипотеза состоит в том, что все канонические корреляции равны нулю. Если проверяемая статистика попадает в критическую область, то по меньшей мере один коэффициент канонической корреляции значимо отличается от нуля.

Критерий проверки только наибольшей канонической корреляции может быть получен из критерия Бартлетта в виде

$$\chi^2 = \left\{-(n-1)+\frac{1}{2}(p+q+1)\right\}\ln(1-\lambda_1) \qquad (6.118)$$

где λ_1 – наибольшее собственное значение. Этот критерий имеет число степеней свободы, приблизительно равное

$$df \approx p+q+1+\frac{1}{2}\{(p-1)(q-1)\}^{\frac{2}{3}} \qquad (6.119)$$

Число степеней свободы, вычисленное по формуле (6.119), округляется до ближайшего целого числа, не превосходящего вычисленное.

В качестве геологического примера рассмотрим данные каротажа скважин, приведенные в табл. 6.45. Проводились измерения интенсивности гамма-лучей, скорости распространения звука и электрического сопротивления в доступном интервале. Эти измеряемые свойства отражают как характеристики пород, так и свойства жидкостей в поровом пространстве. В таблице также представлены лабораторные измерения, сделанные на кернах, взятых из некоторого интервала. Они включают проницаемость, пористость, насыщенность нефтью и водой.

В анализе каротажа скважин измеренные аппаратурой характеристики преобразуются, в результате чего получаются оценки пористости, насыщенности нефтью и водой. Эти оценки основаны на различных комбинациях откликов каротажа в зависимости от стандартов. Это наводит на мысль, что можно найти значимую каноническую корреляцию между откликами каротажа и измерениями на керне.

В табл. 6.46 представлена стандартизованная ковариационная матрица семи переменных, приведенных в табл. 6.45. Указано разбиение матрицы на блоки $[S_{yy}]$, $[S_{xx}]$ и $[S_{xy}]$ и $[S_{yx}]$. Указана также матрица $[\Lambda]$, получающаяся из объединения дисперсий и ковариаций в двух множествах в соответствии с уравнением (6.115), так же, как и собственные значения матрицы $[\Lambda]$ и соответствующие канонические корреляции. Преобразованные векторы $[A]$ и $[B]$, связанные с тремя ненулевыми собственными значениями, также приведенными в табл. 6.46, в графической форме представлены на рис. 6.54.

Первая каноническая корреляция отражает по существу связь между интенсивностью гамма-лучей и временем прохождения звука, с одной стороны, и пористостью – с другой. Знаки весов наводят на мысль о том, что высокие значения гамма-излучения и низкие значения скорости звука могут указывать на низкую пористость. Сланец дает относительно высокий уровень отклика гамма-лучей из-за наличия радиоактивного калия-40 в минерале. Скорость звуковых волн значительно ниже в жидкости, которая заполняет поры, чем в твердом теле. Поэтому можно утверждать, что комбинация гамма-лучей и звукового каротажа дает хороший индикатор сланцев и других непористых пород и, следовательно, коррелированна с пористостью керна.

Вторая каноническая корреляция еще более интересна, так как она связана с комбинацией каротажа и проницаемостью керна, – замечательное свойство, которое трудно предсказать. Веса указывают на то, что низкий отклик гамма-излучения и низкое микросопротивление почти равносильны низкой проницаемости и (или) высокой пористости и насыщенности флюидами. К сожалению, корреляция не является статистически значимой и может оказаться просто случайным эффектом опробования. Третья каноническая корреляция, которая очень мала и статистически незначима, связывает микросопротивление и отклик звукового каротажа с относительной насыщенностью нефтью и водой.

Таблица 6.45. Результаты каротажа и изучения керна скважин в известняках Пенсильвании в Северо-Западном Канзасе

γ	Δ_t	R_t	K	Ф	S_0	S_w
3,1	64,0	28,8	0,1	3,9	28,2	53,8
3,4	69,0	25,1	0,4	7,0	17,2	55,6
3,4	65,0	38,0	0,1	6,1	24,6	54,2
2,8	62,0	15,1	0,4	6,2	19,3	63,0
2,5	56,0	58,9	0,1	5,9	15,3	73,0
2,3	56,0	61,7	0,3	4,7	14,9	61,6
2,3	54,0	129,0	0,2	6,2	29,0	37,1
2,6	60,0	110,0	1,6	12,7	26,7	34,6
6,0	97,0	5,2	0,0	3,0	0,0	96,6
5,2	67,0	18,2	18,0	18,9	26,4	32,3
3,9	82,0	26,9	6,5	18,4	19,0	48,3
4,7	80,0	12,9	2,5	17,9	16,8	48,0
5,1	77,0	12,0	0,1	12,3	11,7	72,6
5,0	79,0	11,0	0,0	10,4	0,0	91,4
6,1	81,0	70,8	0,0	5,2	0,0	97,8

γ – интенсивность γ-излучения, в гамма-единицах; Δ_t – время перехода, мкс/фут; R_t – микросопротивление запаздывания, Ом·м; K – проницаемость, 10^{-6} мкм^2; Ф – пористость, %; S_0 – нефтяное насыщение, % от пористости, S_w – насыщение водой % от пористости.

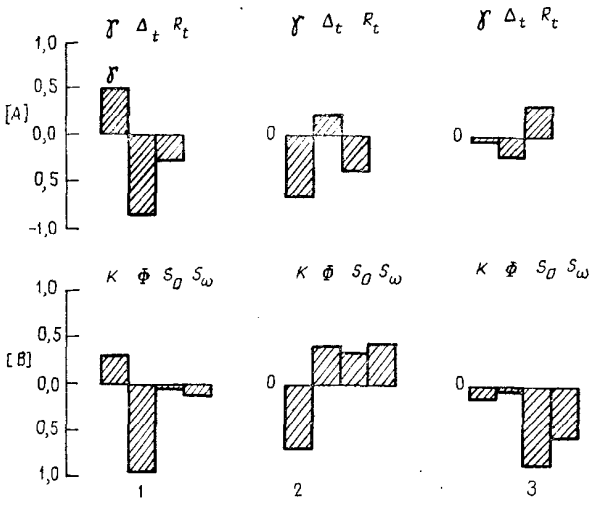


Рис. 6.54. Каноническое преобразование векторов [A] **и** [B]**, вычисленное по данным каротажа скважин и измерениям керна в скважине Северо-Западного Канзаса**. Элементы векторов преобразования представлены как веса. Переменные те же, что в табл. 6.45

Таблица 6.46. Разбиение матрицы стандартизованных дисперсий и коэффициентов корреляции для данных по скважине в табл. 6.45(*а*); объединенная матрица дисперсии $[A]$ (*б*); собственные значения матрицы $[A]$ (*в*) и соответствующие канонические корреляции (*г*); $[A]$ – преобразование вектора с целью приведения значении каротажа к каноническому виду, $[B]$ – с целью приведения измерений к каноническому виду

(а)

$$\left[\begin{array}{ccc|cccc} 1{,}0000 & 0{,}6459 & -0{,}2320 & 0{,}3031 & 0{,}1027 & 0{,}0881 & -0{,}1102 \\ 0{,}6459 & 1{,}0000 & -0{,}0663 & 0{,}3063 & 0{,}4753 & 0{,}1899 & -0{,}2996 \\ -0{,}2320 & -0{,}0663 & 1{,}0000 & 0{,}2161 & 0{,}3118 & -0{,}0862 & -0{,}0789 \\ \hline 0{,}3031 & 0{,}3063 & 0{,}2161 & 1{,}0000 & 0{,}6334 & 0{,}0251 & -0{,}2150 \\ 0{,}1027 & 0{,}4753 & 0{,}3118 & 0{,}6334 & 1{,}0000 & 0{,}2099 & -0{,}5134 \\ 0{,}0881 & 0{,}1899 & -0{,}0862 & 0{,}0251 & 0{,}2099 & 1{,}0000 & -0{,}8639 \\ -0{,}1102 & -0{,}2996 & -0{,}0789 & -0{,}2150 & -0{,}5134 & -0{,}8639 & 1{,}0000 \end{array}\right]$$

(*б*)

$$\begin{bmatrix} 0{,}1237 & -0{,}0891 & -0{,}0337 & 0{,}0406 \\ 0{,}0769 & 0{,}4445 & 0{,}1144 & -0{,}2278 \\ -0{,}0750 & 0{,}0006 & 0{,}0762 & -0{,}0372 \\ -0{,}0962 & 0{,}0167 & 0{,}0620 & -0{,}0332 \end{bmatrix}$$

(*в*)

$$\begin{aligned} \lambda_1 &= 0{,}4010 \quad r_1 = 0{,}63 \\ \lambda_2 &= 0{,}1697 \quad r_2 = 0{,}41 \\ \lambda_3 &= 0{,}0405 \quad r_3 = 0{,}20 \\ \lambda_{14} &= 0{,}0000 \quad r_4 = 0{,}00 \end{aligned}$$

(*г*)

Первая каноническая корреляция		Вторая каноническая корреляция		Третья каноническая корреляция	
Вектор $[A]$	Вектор $[B]$	Вектор $[A]$	Вектор $[B]$	Вектор $[A]$	Вектор $[B]$
$\begin{bmatrix} 0{,}4960 \\ -0{,}8727 \\ -0{,}2869 \end{bmatrix}$	$\begin{bmatrix} 0{,}2953 \\ -0{,}9473 \\ -0{,}0573 \\ -0{,}1100 \end{bmatrix}$	$\begin{bmatrix} -0{,}6627 \\ 0{,}2195 \\ -0{,}3380 \end{bmatrix}$	$\begin{bmatrix} -0{,}6816 \\ 0{,}4287 \\ 0{,}3627 \\ 0{,}4692 \end{bmatrix}$	$\begin{bmatrix} -0{,}0589 \\ -0{,}2128 \\ 0{,}3301 \end{bmatrix}$	$\begin{bmatrix} -0{,}1225 \\ -0{,}0510 \\ -0{,}8266 \\ -0{,}5469 \end{bmatrix}$

СПИСОК ЛИТЕРАТУРЫ К ГЛАВЕ 6 «АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ

1. *Anderberg M. R.* Cluster analysis for applications. Academic press., Inc, New York, 1973, 359 p.
   Отличное изложение неиерархических кластерных процедур. Приложены программы на ФОРТРАНе.
2. *Anderson T. W.* An introduction to multivariate statistical analysis. John Wiley and Sons, Inc., New York, 1958, 374 p.
   Имеется русский перевод: Т. Андерсон. Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. Современное изложение многомерных статистических методов, T^2-статистика рассматривается в гл. 5, а дискриминантный анализ излагается в гл. 6, где он называется класси-

фикацией. Метод главных компонент рассматривается в гл. 11 и коротко как статистический метод – на стр. 443–444.

3. *Bartlett M. S.* Further aspects of the theory of multiple regression. Proc. of the Cambridge Philosophical Soc., 34, 1938, 33–40.
4. *Benzecri J.-P.* a. oth. L'Analyse des donnees, v. 2, L'Analyse des correspondences. Dunod, Paris, 1980, 628 p.
5. *Bijnen E. J.* Cluster analysis – Survey and evaluation of techniques. Tilburg Univ. Press. Groningen. The Netherlands, 1973, 112 p.
   Компактный хорошо написанный обзор всех типов кластерных процедур.
6. *Burt C.* Correlations between persons. British Journal of Psychology, General Section, 28, 1937, 59–96.
   В этой ранней работе обсуждаются связи между R- и Q-методами факторного анализа. Гипотетический пример, используемый для иллюстрации теоремы Эккарта–Юнга, взят из этой книги.
7. *Businger P. A.. Golub G. H.* Algorithm 358, singilar value decomposition of a complex matrix. Communications of the Assoc. for Computing Machinery, 12, 1969, p. 564–565.
8. *Child D.* The essentials of factor analysis. Holt, Rinehart and Winstin Ltd., London, 1970, 107 p.
   Компактный обзор факторного анализа и близких методов, изданный в бумажном переплете. Изложение нематематическое и очень понятное.
9. *Clark S. P. Jr.* (ed.) Handbook of Physical Constants. Geol. Soc. America, Mem., 97, 1966, 587 p.
10. *Clifford H. T., Sfefenson W.* An introduction to numerical classification. Academic Press, Inc., New York, 1973, 229 p.
    Точка зрения биолога на кластерный анализ и другие процедуры классификации. Глава 8 (о матрицах сходства) и глава 13 (многомерный анализ) представляют особый интерес.
11. *Cooley W. W., Lohnes P. R.* Multivariate data analysis. John Wiley and Sons, Inc., New York, 1971, 364 p.
    Наряду со многими другими вопросами излагаются теории множественной регрессии, дискриминантных функций и факторный анализ. В книге приводятся также тексты подпрограмм на ФОРТРАНе и вычислительные схемы разнообразных процедур.
12. *David M., Dagbert M., Beauchemin Y.* Statistical analysis in geology. Correspondence analysis method. Quart, Colorado Sch. Mines 72, № 1, 1977, 60 p.
    Авторы предлагают применить анализ соответствий к непрерывным переменным и дают обоснование метода, основанное на теории вероятностей. Приводится программа на ФОРТРАНе.
13. *Davis I. C.* Information contained in sediment-size analysis. Jour. Int'l. Assoc. Mathematical Geology, 2, № 2, 1970, p. 105–112.
    Экспериментальные данные по осадкам бассейна Баратария, рассмотренные в настоящей главе, взяты из этой книги.
14. *Draper N. R., Smith H.* Applied regression analysis, 2nd ed. John Wiley and Sons, Inc., New York, 1981, 709 p.
    Имеется русский перевод: Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М., Финансы и статистика, 1987. – Т. 1 и 2.
15. *Eckart C., Young B.* The approximation of one matrix by another of lower rank. Psychometrika, 1, № 3, 1936, p. 211–218.
    Фундаментальная работа о связи *R*- и *Q*-методов факторного анализа.
16. *Everitt B. S.* Graphical techniques for multivariate data North Holland. New York, 1978, 117 p.
17. *Fisher R. A.* The precision of discriminant functions. Annals of Eugenics, 10, 1940, p. 422–429.
18. *Folk R. L.* Petrology of sedimentary rocks, 4th ed. Hemphill's, Austin, Tex, 1980, 184 p.
19. *Ganadesikan R.* Methods for statistical data analysis of multivariate observations. John Wiley and Sons, Inc., New York, 1977, 311 p.
    В тексте представлены графические методы, разработанные в Телефонных лабораториях Белла. Процедуры кластерного анализа и другие процедуры классификации рассмотрены в гл. 4.
20. *Gower I. C.* Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, № 3, 1966, p. 325–338.
    Рассматривается двойственность между анализом соответствия и методом главных компонент.
21. *Griffiths J. C., Ondrick C.* W. Modelling the petrology of detrital sediments. In: Merriam D. F. (ed.). Computer applications in the earth sciences. Plenum Press, New York, 1969, p. 73–97.
    Статья посвящена моделированию процесса отклика при исследовании происхождения кластических осадков. Факторы, представляющие осадки, обсуждаются на стр. 76–84. Эксперимент, свя-

занный с применением МГК в этой главе, описан на стр. 86–88.
22. *Harman H. H*. Modern factor analysis, 2nd ed. Univ. Chicago, Press, 1967, 474p.
Имеется русский перевод: Харман Г. Современный факторный анализ. М., Статистика, 1972. Несмотря на то что большинство примеров этой книги взято из психометрики, основной упор сделан на вычислительные аспекты излагаемых теорий и поэтому изложение относительно свободно от психологической терминологии.
23. *Harris R. J*. A primer of multivariate statistics. Academic Press, New York, 1975, 332 p.
Рассмотрены главные компоненты (гл. 6) и факторный анализ (гл. 7). Хотя книга написана в расчете на общественные науки, изложенные в ней методы в равной мере применимы и к геологическим данным. Рассмотрены также преобразование данных и вращение факторов.
24. *Harfigan J. A*. Clustering algorithms, John Wiley and Sons. Inc. New York, 1975, 351 p.
Кластерный анализ изложен на строгом алгоритмическом уровне. Книга содержит многочисленные подпрограммы на ФОРТРАНе и тестовые данные, взятые из различных дисциплин.
25. *Heckel P. H., Baesemann J. F*. Environmental enterpretation of conodont distribution in Upper Pennsilvanian (Missourian) megacyclothems in eastern Kansas. Bull. American Assoc. Petroleum Geologists, 59, № 3, 1975, p. 486– 509.
26. *Hill M*. O. Correspondence analysis. A neglected multivariate method. Jour. Royal Stat. Soc., Ser. C, Appl. Stat., 23, № 3, 1974, p. 340–354.
27. *Hirschfeld H. O*. A connection between correlation and contingency. Proc. of the Cambridge Philosophical Soc., 31, 1935, p. 520–524.
28. *Howarth R. I*. (ed.) Statistics and data analysis in geochemical prospecting. Elsevier Publ. Co., Amsterdam, 1983, 437 p.
В гл. 6 рассмотрено применение многомерных методов при геохимической разведке и приведен перечень источников вычислительных программ, в гл. 9– применение дискриминантного анализа при прогнозировании олова, в главах 10 и 11 дан обзор публикаций о применении многомерных методов.
29. *Imbrie I., Purdy E. G*. Classification of modern Bahamian carbonate sediments, in Classification of Carbonate Rocks, a Symposium American Assoc. Petroleum Geologists, Mem, 1, 1962, p. 243–272.
Основная ссылка в геологической литературе на метод главных компонент.
30. *Sardine N., Sibson R*. Mathematical taxonomy, John Wiley and Sons, Ltd., London, 1971,286p.
Методы кластерного анализа и классификации рассматриваются с использованием обозначений. В книге имеется ценный словарь терминов численной таксономии.
31. *Johnson R. M*. On the theorem stated by Eckart and Young. Psychomet-rika, 28, N 3, 1963, p. 259–263.
Математическое доказательство теоремы Эккарта–Юнга.
32. *Joreskog K. G*. Factor analysis by least-squares and maximum-likehood methods, in Enslein K. and others (eds.) Statistical methods for digital computers, v. 3, John Wiley and Sons, Inc., New York, 1977, p. 125–153.
Полное изложение, доведенное до алгоритма «истинного» *R*-метода факторного анализа, описанного в этой главе.
33. *Joreskog K., G., Klovan J. E., Reyment R. A*. Geological factor analysis. Elsevier Publ. Co., Amsterdam, 1976, 178 p.
Рекомендуется тем, кто интересуется применением факторного анализа к решению геологических задач. Автор не затрудняет себя высказыванием мнений об относительной ценности тех или иных методов.
34. *Kendall M. G., Stuart A*. Advanced theory of statistics, 2nd ed., v. 2, Charles Griffin and Co., Ltd, London, 1967, 690 p.
35. *Klouan J. E*. The use of factor analysis in determining depositional environments from grain-size distributions, Jour. Sedimentary Petrology, 36, 1966, p. 115–125.
Дан анализ осадков бассейна Баратария *Q*-методом факторного анализа.
36. *Koch G. S. Je; Link R. F*. Statistical analysis of geological data. Dover Publications, Inc., New York, 1980, 850 p.
37. *Krumbein W. C., Aberdeen E*. The sediments of Barataria Bay. Jour. Sedimentary Petrology, 7, 1937, p. 3–17.
Данные табл. 6.22 представляют собой адаптацию данных табл. 1 этой работы.
38. *Krumbein W. C., Shreve R. L*. Some statistical properties of dendritic channel networks. Office of Naval Research. Tech. Rept., 13, ONR Task No. 389–150, 1970, 117 p. (available from Documents Clearing-

house, Arlington, VA, as document AD 705625).
Результаты исследования дренирующих сетей на площадях в восточной части штата Кентукки, проведенных студентами-геологами и географами. Классы были достаточно большими, что позволило собрать обширные данные по операционным ошибкам, а также большое количество данных по самому предмету. Табл. 6.1 составлена по результатам этих занятий.

39. *Lawley D. N.* The estimation of factor loading by the method of maximum likehood. Proc. Royal Soc. Edinbourgh, Ser. A60, 1940, p. 64–82.
40. *Lawley D. N., Maxwell A. E.* Factor analysis as a statistical method. 2nd ed., Butterworth and Co., Ltd, London, 1971, 153 p.
Имеется русский перевод: Лоули Д., Максвелл А. Факторный анализ как статистический метод. М., Мир, 1960. В этой короткой монографии факторный анализ изложен независимо от его тесной связи с психологией. Особенно интересны подробно разобранные примеры.
41. *Lebart L. Л., Morineau A., Warwick K. M.* Multivariate descriptive statistical analysis. John Wiley and Sons, Inc., New York, 1984, 231 p.
Перевод французского издания, в котором изложен анализ соответствия, развитый Бензекри.
42. *Le Maitre R. W.* Numerical Petrology. Elsevier Publ. Co., Amsterdam, 1982, 281 p.
Этот весьма специальный текст представляет собой ценность из-за того, что в нем обсуждается эффект замкнутости в анализе многомерных данных, а также применения этого метода к геохимическим и петрологическим данным. Рекомендуется тем, кто анализирует данные с постоянными суммами.
43. *Li C.* C. Introduction to experimental statistics. McGraw-Hill, Inc., New York, 1964, 460 p.
Глава 30 «Измерения в многомерном пространстве» содержит очень хорошее и краткое изложение дискриминантного анализа.
44. *Longley J. W.* An appraisal of least squares program of the electronic computer from the point of view of the user, Jour. American Statistical Assoc., 62, № 319, 1967, p. 819–841.
Обсуждение грубых ошибок, которые вкрадываются в программы множественной регрессии из-за плохого программирования и сохранения недостаточного числа знаков.
45. *Lunneborg C. E., Abbott R. D.* Elementary Multivariaty analysis lor the behavioral sciences. North Holland, New York, 1983, 522 p.
46. *Marascuilo L. A., Levin J. R.* Multivariate statistics in the social science. A researcher's guide. Brooks/Cole Publ. Co., Monterey, Calif., 1983, 530 p.
Дает подробный анализ библиотеки программ широко используемых многомерных статистик и сравнение получаемых результатов.
47. *Maron M. J.* Numerical Analysis. A practial approach. McMillan Publ. Co., Inc., New York, 1982, 471 p.
Содержит хорошее описание математических операций, используемых во всех математических процедурах, включающих вычисление собственных значений.
48. *Marriot F. H. C.* The interpretation of multiple observations. Academic Press, Inc., Ltd., London, 1974, 117 p.
Очень компактный том, содержащий хорошо написанный обзор многомерного статистического анализа.
49. *Mafalas N. C., Reiner B. J.* Some comments on the use of factor analysis, Water Resources Research, 3, 1967, p. 213–223.
50. *McQueen J.* Some methods for classification and analysis of multivariate- observations 5th Berkeley Symposium on Mathematics. Statistics and Probability, 1, 1967, p. 281–298.
51. *Morrison D. F.* Multivariate statistical methods, 2nd ed. McGraw-Hill, Inc., New York, 1976, 415 p.
Одно из самых хороших руководств по многомерному статистическому анализу, содержащее особенно полезный раздел о связи между наиболее известными одномерными методами и их многомерными обобщениями. Гл. 4, 6–9 охватывают темы, изложенные в этом разделе. Изложение факторного анализа в гл. 9 особенно просто.
52. *Morrison D. F.* Applied statistical methods. Prentice-Hall, Inc, Engle- wood Cliffs, N.J., 1983, 562 p.
Содержит отличное изложение дискриминантного анализа для двух групп.
53. *Ondrick C. W., Srivastava G. S.* COREAN – FORTRAN IV computer program for correlations, factor analysis (*R*- and *Q*-mode) and varimax rotation. Kansas Geological Survey Computer Contribution, 42, 1970, 92 p.
54. *Osfroff A. G.* Comparison of some formation water classification systems. Bull. American Assoc. Petro-

leum Geologists, 51, N 3, 1967, p. 403–416.
55. *Potter P. E., Skimp N. F.. Witters I.* Trace elements in marine and fresh-water agrillaceous sediments. Geochimica et Cosmochimica. Acta, 27, 1963, 669–694.
56. *Reyinent R. A., Blackith R. E., Campbell N. A*. Multivariate morphometrics, 2nd ed. Academic Press, Inc., Lrd., London, 1984, 233 p.
Полный обзор применения статистических методов к изучению морфологии животного мира, в особенности ископаемых остатков. Содержит большое количество цитат из текущей литературы и многочисленные примеры исследования.
57. *Sherman K. N.. Bunker C. M.. Buck C. A*. Correlation of uranium, thorium and potassium with aeroradioactivity in the Berea area, Virginia Economic Geology, 66, 1971, p. 302–308.
58. *Sneath P. H. A., Sokal R. R*. Numerical Taxonomy. W. H. Freeman and Co., San Francisco, 1973, 573 p.
Одна из основополагающих работ по численной таксономии. К сожалению, терминология может смутить геолога, а длинные рассуждения по поводу численных методов могут показаться не относящимися к делу. Однако это наиболее легкое введение в биологическую литературу по численной классификации.
59. *Switzer P*. Numerical classification in Merriam D. F. (ed.) Geostatistics, .a Colloquium. Plenum Press. New York, 1970, p. 31–43.
60. *Teil H*. Correspondence factor analysis. An outline of its method. Jour. ilnt'1 Assoc. Mathematical Geology, 7, N 1, 1975, 3–12.
61. *Teil H., Cheminee J. L*. Application of correspondence factor analysis -to the study of major and trace elements in the Erta Ale Chain (Altar, Ethiopia). .Jour. Int'l Assoc. Mathematical Geology, 7, № 1, 1975, p. 13–30.
62. *Temple J. T*. The use of factor analysis in geology. Jour. Int'1 Assoc. Mathematical Geology, 10, 1978, № 4, p. 379–387.
Серьезная критика факторного анализа, в особенности его применения в геологии и связанных с нею областей.
63. *Thurstone L L*. Multiple factor analysis. Univ. Chicago Press, Chicago, 1974. 535 p.
Классическая работа по факторному анализу в его первоначальном виде, приспособленном к задачам психометрии. Поскольку она написана до появления вычислительных машин, многие детали сейчас кажутся устаревшими. Однако изложение факторного анализа в ней фундаментально.
64. *Tryon R. C., Bailey D. E*. Cluster analysis McGraw-Hill. Inc., New York, 1970, 347 p.
Методы классификации представлены с точки зрения психологов и социологов. Вычислительные процедуры, приведенные Трайоном и Бэли, отличаются от представленных в этой главе; наиболее близки к ним методы *Q*-анализа.
65. *Wanke H*. a. oth. Major and trace elements and cosmic-ray producted radioisotopes in lunar samples. Science, 167, № 3918, 1970, p. 523–525.
66. *Zhou D., Chang T., Davis J. C*. Dual extraction of *R*-mode and *Q*-mode factor solutions. Jour. Int'l Assoc. Mathematical Geology, 15, № 5, 1983, p. 581– 606.

ПРИЛОЖЕНИЕ

Тексты программ СТАТ

Настоящее приложение написано на основе приложения к книге Дж. С. Дэвиса «Как пользоваться программой СТАТ», написанного Стивом Йи, в котором приведены рекомендации об использовании дискеты. На дискете помещены основные статистические процедуры, изложенные в книге. Дискета предназначена для персонального компьютера типа IBM-PC, использующего операционную систему PC-DOS или MS-DOS (версия 2.10) с памятью 128 Кб. В настоящем приложении приведены тексты соответствующих программ, написанные на языке ФОРТРАН. СТАТ – комплекс программ на языке ФОРТРАН, содержащих ряд статистических процедур, описанных в этой книге. В него вошли программы ввода данных, элементарной статистики и основных операций над матрицами. Поиск необходимого пользователю модуля происходит путем указания соответствующего режима.

Ввод данных

Вводимые данные имеют структуру матрицы. Ниже приводится таблица, характеризующая эту структуру в различных случаях:

Данные	Строки	Столбцы
Одномерные данные	Наблюдения	1
Многомерные данные	Наблюдения	Значения координат
Одномерный дисперсионный анализ	Копии	Пробы
Двумерный дисперсионный анализ	Уровни	Пробы

Элементарная статистика

Если пользователь выбрал режим элементарной статистики, то можно воспользоваться следующими пятью программами: 1 – одномерные статистики и одновыборочный критерий; 2 – двумерные статистики; 3 – линейная регрессия; 4 – одномерный дисперсионный анализ; 5 – двумерный дисперсионный анализ.

В следующей таблице приведены соответствующие ограничения на размеры матрицы данных:

Метод	Число строк		Число столбцов	
	Максимум	Минимум	Максимум	Минимум
Одномерный	Нет	2	1	1
Двумерный	Нет	2	2	2
Линейная регрессия	100	2	2	2
Одномерный дисперсионный анализ	Нет	2	100	2
Двумерный дисперсионный анализ	Нет	2	100	2

Примечание. Максимальные размеры матрицы ограничены только размерами оперативной памяти используемого компьютера.

Результаты вычислений по каждой программе по желанию заносятся в файл, который может

быть использован в дальнейшем. Ниже приводится перечень различных статистик, которые вычисляются в программе:

Метод	Статистики
Одномерные статистики	Число наблюдений, сумма, сумма квадратов, дисперсия, стандартное отклонение, среднее, одновыборочная *t*-статистика
Двумерные статистики	Число наблюденных пар, сумма попарных произведений, коэффициент ковариации, коэффициент корреляции, между X_1 и X_2; сумма, сумма квадратов, среднее, дисперсия, стандартное отклонение
Линейная регрессия	Члены нормального уравнения (т. е. число наблюденных пар; сумма квадратов *X*-ов и *Y*-ов), сумма Y-ов, сумма попарных произведений *X* и *Y*, оцененные параметры уравнения регрессии, остатки, общая сумма квадратов, качество аппроксимации и коэффициент корреляции
Одномерный ANOVA	Таблица одномерного ANOVA
Двумерный ANOVA	Таблица двумерного ANOVA

Матричные операции

В разделе матричные операции содержатся следующие модули:

1. напечатать матрицу;
2. сложить две матрицы;
3. перемножить две матрицы;
4. умножить матрицу на постоянную;
5. обратить матрицу;
6. транспонировать матрицу;
7. вычислить собственные векторы и собственные значения симметрической матрицы.

Для всех программ максимальный размер матрицы 50x50.

Программа 7 записывает вектор собственных значений в убывающем порядке в некоторый файл.

Апробацию программ СТАТ рекомендуется проводить на примерах данных, помещенных в тексте. Там же можно найти и результаты вычислений и их интерпретацию.

Дополнительные программы

«Террастат» – обширная библиотека программ почти всех процедур, описанных в этой книге. Они реализованы на дискете для персонального компьютера типа IBM–PC и совместимых с ним, использующих операционные системы PC–DOS или MS–DOS. Информацию о библиотеке программ «Террастат», включая цены и совместимость компьютера, можно получить по адресу:

TERRASCIENCE, INC.
755 West Tenth Avenue
Lakewood, Colorado USA 80215.

Желающие получить копию дискеты СТАТ к книге Дж. С. Дэвиса с инструкцией и комментариями на русском языке могут обратиться по адресу:

ВИЭМС, 123853 Москва, 3-я Магистральная ул., 38.

Тексты программ STAT на языке FORTRAN

```
C
C   MAIN PROGRAM FOR 8TAT STATISTICAL PACKAGE
C
    PROGRAM STAT
        COMMON /IOLUN/ I5,I6,J5,J6,J4
    COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
        COMMON /MATQ/ AM(50,50),BM(50,50),CM(50,50)
    1 NRA,NCA,NRB,NCB,NRC,NCC,CONST
    COMMON /NAMEF/ FNAME
    CHARACTER*16 FNAME
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
C
C... INITIALIZE CONSOLE I/O UNITS
C
    I5 = 0
    I6 = 0
C
C... INITIALIZE FILE I/O UNITS
C
    J4 = 0
    J5 = 0
    J6 = 0
C
C... SET MAX MATRIX DIMENSIONS
C
    MAXR = 50
    MAXC = 50
    CALL INITTS
    CALL INOUT(0,IOP,KOP,MAXR,MAXC)
    CALL CLEAR
        WR1TE(16,200)
    200 FORMAT(//' STAT - Version 1.5',
        1//' (C) Copyright by Terrasciences Inc. 1987',
        1//' This program presents a sample of the statistical,
        1//' procedures described in "Statistics and Data Analysis',
        1//' in Geology" by John C. Davis')
C
    111 CALL DATE
    101 CALL CLEAR
        WRITE(I6,204)
    204 FORMAT(///,15X,'* * * Main Menu of Options * * *',
        1//' Option Description',
        1//,4X,'1 -- Data Entry',
        1//,4X,'2 -- Elementary Statistics',
        1//,4X,'3 -- Matrix Operations',
        1//,4X;'0 -- Exit Slat and Return to DOS',
        1//,' Enter Option Number: ')
    108 CALL IREAD(1,KOP,IO5,IO5,IO5)
        IF(LZ.LT.0) CALL CLEAR
        IF(LZ.LT.0) GOTO 111
        IF(KOP.EQ.0) STOP
        IF(KOP.EQ.1) CALL DATENT
        IF(KOP EQ.1.AND.LZ.LT.0) CALL CLSALL
        IF(KOP.EQ.1) GOTO 101
        IF(KOP.EQ.2) GOTO 100
        IF(KOP.EQ.3) GOTO 102
        CALL INVLID(1)
        GOTO 108
    100 CALL CLEAR
        WRITE(I6,201)
    201 FORMAT(/,15X,'* Elementary Statistics Menu of Options
        1//' Option Description',
        1//,4X,'1 -- Univanate Statistics and One-Sample t Test'
        1//,4X,'2 -- Bivariate Statistics',
```

```
        1//,4X,'3 -- Linear Regression',
        1//,4X,'4 -- One Way Analysis of Variance'
        1//,4X,'5 -- Two Way Analysis of Variance'
        1//,4X,'0 -- Return to Main Menu of Options'
        1//,' Enter Option Number: ')
    109 CALL IREAD(1,IOP,IO5,IO5,IO5)
        IF(LZ.LT.0) GOTO 101
        IF(IOP.LT.0.OR.IOP.GT.5) CALL INVLID(1)
        IF(IOP.LT.0.OR.IOP.GT.5) GOTO 109
        IF(IOP.EQ.0) GOTO 101
        GOTO 103
    102 CALL CLEAR
        WRITE(I6,202)
    202 FORMAT(/,15X,'* Matrix Operations Menu of Options *',
        1//' Option Description',
        1//,4X,'1 -- Print Out a Matrix',
        1//,4X,'2 -- Add Two Matrices',
        1//,4X,'3 -- Multiply Two Matrices',
        1//,4X,'4 -- Multiply a Matrix by a Constant',
        1//,4X,'5 -- Invert a Matrix',
        1//,4X,'6 -- Transpose a Matrix',
        1//,4X,'7 -- Compute Eigenvalues and Vectors of Symmetric Matrix',
        1//,4X,'0 -- Return to Main Menu of Options',
        1//' Enter Option Number: ')
    110 CALL IREAD(1,IOP, IO5,IO5,IO5)
        IF(LZ.LT.0) GOTO 101
        IF(IOP.LT.0.OR.IOP.GT.7) CALL INVLID(1)
        IF(IOP.LT.0.OR.lOP.GT.7) GOTO 110
        IF(IOP.EQ.0) GOTO 101
C
C... DETERMINE JOP TYPE OF INPUT/OUTPUT
C
    103 IF(KOP.EQ.2) JOP = 1
        IF(KOP.EQ.3.AND.IOP.EQ.1) JOP = 1
        IF(KOP.EQ.3.AND.(IOP.GE.2.AND.IOP.LE.3)) JOP = 2
        IF(KOP.EQ.3.AND.IOP.GE.4) JOP = 3
        CALL INOUT(JOP,IOP,KOP,MAXR,MAXC)
        IF(LZ.GE.0) GOTO 112
    300 CALL CLSALL
        IF(KOP.EQ.2) GOTO 100
        GOTO 102
    112 IF(JOP.EQ.-I) GOTO 95
        IF(KOP.EQ.3) GOTO 104
        CALL CLEAR
        GOTO (1,10,15,20,25),IOP
    104 GOTO (30,35,40,45,50,55.60),IOP
    1 CALL VARIAN
        GOTO 99
    10 CALL BICOR
        GOTO 99
    15 CALL UNFIT
        GOTO 99
    20 CALL ONEOV
        GOTO 99
    25 CALL TWOVA
        GOTO 99
    30 CALL PRINTM(1,CM,NRC,NCC,MAXR,MAXC)
        CALL PAUSE(2)
        GOTO 99
    35 CALL ADDM(AM,BM,CM,NRA,NCA,MAXR,MAXC)
        GOTO 97
    40 CALL MMULT(AM,BM,CM,NRA,NCA,NCB,MAXR,MAXC,MAXR,MAXC,MAXR,MAXC)
        GOTO 97
    45 CALL CMULT(AM,CONST,CM,NRA,NCA,MAXR,MAXC)
        GOTO 97
    50 CALL MINV(AM,CM,NRA,MAXR,DET)
        GOTO 97
    55 CALL MTRANS(AM,CM,NRA,NCA,MAXR,MAXC)
        GOTO 97
    60 CALL EIGENJ(AM,CM,NRA,MAXR)
```

```
C
C... CLOSE I/O FILES
C
    97 IF(LZ.GE.0) GOTO 113
       GOTO 300
   113 IF(KORF0.EQ.'K') GOTO 94
       REWIND J4
       CALL ICLOSE(J4,1)
    94 IF(JOP.EQ.3) GOTO 98
    99 IF(L2.QE.0) GOTO 114
       GOTO 300
   114 IF(KORF.EQ.'K') GOTO 98
       REWIND J5
       CALL ICLOSE(J5,1)
    98 IF(KOP.EQ.3.AND.IOP.EQ.5.AND.DET.EQ.0.0) GOTO 102
       IF(KOP.EQ.3.AND.IOP.GE.2) GOTO 96
       IF(lPRT.EQ.'N') GOTO 95
       REWIND J6
       CALL ICLOSE(J6,1)
    95 IF(KOP.EQ.2) GOTO 100
       GOTO 102
C
C... WRITE OUT SOLUTION MATRICES
C
    96 WRITE(I6,205)
   205 FORMAT(/' Enter name of file to store matrix C: ')
       CALL FNRDIN(FNAME)
       IF(LZ.LT.0) GOTO 300
       CALL FILOPN(1,20,FNAME)
       WRITE(20,*) NRC,NCC
       DO 105 I = 1,NRC
       WRITE(20,*) (CM(1,J),J=1,NCC)
   105 CONTINUE
       REWIND 20
       CALL ICLOSE(20,1)
       WRITE(I6,207) FNAME
   207 FORMAT(/' Matrix C is now stored in file ',A16)
       IF(IOP.LE.6) CALL PAUSE(2)
       IF(IOP.LE.6) GOTO 102
C
C... STORE EIGENVALUES
C
       FNAME = 'EIGENVAL.DAT'
       CALL FILOPN(1,20,FNAME)
       WRITE(20,*) NRA,1
       DO 107 I = 1,NRA
     WRITE(20,*) AM(I,I)
   107 CONTINUE
       REWIND 20
       CALL ICLOSE(20,1)
       WRITE(I6,208) FNAME
   208 FORMAT(/' The eigenvalues are now stored as a column vector',
       1' in file ',A16)
       CALL PAUSE(2)
       GOTO 102
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   DETERMINE I/O DATA OPTIONS
C
C   JOP = 0:INITIALIZE I/O OPTIONS
C         1 I/O OPTIONS FOR ELEMENTARY STAT AND PRINTM
C         2 INPUT TWO MATRICES
C         3 INPUT ONE MATRIX
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE INOUT(JOP,IOP,KOP,MAXR,MAXC)
       COMMON /IOLUN/ I5,I6,J5,J6,J4
    COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
```

```
      COMMON /MATO/ AM(50,50),BM(50,50),CM(50,50),
    1 NRA,NCA,NRB,NCB,NRC,NCC,CONST
      COMMON /NAMEF/ FNAME
    CHARACTER*16 FNAME
      COMMON /DATIM/ IDATE(8),ITIME(5)
    CHARACTER*1 IDATE,ITIME
      COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZ2,NSP,DPC,DFT,UPA
      COMMON /ASC/ IALPHA(96)
    CHARACTER*1 IALPHA
C
   IF(JOP.EQ.0) GOTO 99
    GOTO (102,400,400), JOP
C
C... INITIALIZE I/O OPTIONS
C
    99 KORF = 'F'
       KORF0 = 'F'
       MON = 'Y'
       IPRT = 'N'
    RETURN
C
C... ASK USER FOR NATURE OF OUTPUT
C
    102 CALL CLEAR
      WRITE(I6,225)
    225 FORMAT(/' SELECT OUTPUT OPTIONS',//)
    100 WRITE(I6,200)
    200 FORMAT(/' Output results to monitor Y or N: ')
      CALL AREAD(2,JO5,KO5,IO5,IO5)
      IF(LZ.LT.0) RETURN
      MON = IALPHA(JO5)
      IF(MON.EQ.Y) MON = 'Y'
      IF(MON.EQ.'n') MON = 'N'
      IF(MON.NE.'Y'.AND.MON.NE.'N') CALL INVLID(1)
      IF(MON.NE.'Y'.AND.MON.NE.'N') GOTO 100
    101 WRITE(I6,201)
    201 FORMAT(/' Output results to file - Y or N: ')
      CALL AREAD(2,JO5,KO5,IO5,IO5)
      IF(LZ.LT.0) RETURN
      IPRT = IALPHA(JO5)
      IF(IPRT.EQ.y) IPRT = 'Y'
      IF(IPRT.EQ.'n') IPRT = 'N'
      IF(IPRT.NE.'Y'.AND.IPRT.NE.'N') CALL INVL1D(1)
      IF(IPRT.NE.T.AND.IPRT.NE.'N') GOTO 101
C
C... OPEN I/O FILES IF THOSE OPTIONS ARE REQUESTED
C
    CALL CLEAR
      WRITE(I6,226)
    226 FORMAT(/' SELECT INPUT/OUTPUT F1LES,//)
      IF(KORF.EQ.'F') GOTO 301
      J5 = 0
      GOTO 350
C
C... OPEN INPUT DATA PILE
C
    301 J5 = 2
      WRITE(I6,202)
    202 FORMAT(/' Enter name of input data file: ')
      CALL FNRDIN(FNAME)
      IF(LZ.LT.0) RETURN
      CALL FILOPN(0,J5,FNAME)
      IF(LZ.LT.0) RETURN
C
C... OPEN OUTPUT FILE
C
    350 IF(IPRT.EQ.'N') RETURN
      J6 = 3
      WRITE(I6,205)
    205 FORMAT(/' Enter name of output file: ')
```

```
        CALL FNRDIN(FNAME)
        IF(LZ.LT.0) RETURN
        CALL FILOPN(1,J6,FNAME)
C
C   WRITE OUT DATE AND TIME
C
        WRITE(J6,224) IDATE ITIME
    224 FORMAT(//, STAT Program Output'
        1//, Date   ,8A1/   Time   5A1,///)
    RETURN
C
C   INPUT ONE OR TWO MATRICES
C
    400 IOPM1 = IOP - 1
        CALL CLEAR
        GOTO (41,42,43,44,45,46),IOPM1
    41 WRITE(16,206)
    206 FORMAT(/' MATRIX ADDITION: A + B = C',//)
        GOTO  50
    42 WRITE(I6,207)
    207 FORMAT(/' MATRIX MULTIPLICATION: A * B = C',//)
        GOTO 50
    43 WRlTE(16,208)
    208 FORMAT(/' MULTIPLY MATRIX BY CONSTANT: CONST * A = C',//)
        WRITE(I6,204)
    204 FORMAT(/' Enter value for constant CONST: ')
        CALL RREAD(1,CONST,AO5,AO5,AO5)
        IF(LZ.LT.0) RETURN
        GOTO 50
    44 WRITE(I6,209)
    209 FORMAT(/' MATRIX INVERSION: Inverse of A = C',/,')
        GOTO 50
    45 WRITE(16,210)
    210 FORMAT(/' MATRIX TRANSPOSITION: Transpose of A = C',//) f
        GOTO 50
    46 WRITE(I6,211)
    211 FORMAT(/ EIGENVALUES AND EIGENVECTORS',//,
        1//' Compute eigenvalues and vectors of symmetric matrix A',
        1//' Eigenvalues (in descending order) are stored in file',
        1//' EIGENVAL.DAT',
        1//' Corresponding eigenvectors are represented by columns of,
        1//' matrix C')
C
C... READ IN MATRIX A
C
    50 KORF0 = 'F'
        IF(KCRF0.EQ.'K') J4 = 0
        IF(KORF0.EQ.'K') GOTO 51
        J4 = 1
        WRITE(16,214)
    214 FORMAT (/' Enter name of file containing matrix A: ')
        CALL FNRDIN(FNAME)
        IF(LZ.LT.0) RETURN
        CALL FILOPN(0,J4,FNAME)
        IP(LZ.LT.0) RETURN
    51 IF(KORF0.EQ.'K') WRITE(I6,215)
    215 FORMAT(/' Enter number of rows and columns for matrix A: ') '
        READ(J4,*,END=500) NRA,NCA
        IF(NRA.LE.MAXR.AND.NCA.LE.MAXC) GOTO 52
    57 WRITE(I6,216) MAXR,MAXC
    216 FORMAT (/' Number of rows, columns exceed max of ',13,' by ',13)
        IF(J4.EQ.1) REWIND J4
        IF(J4.EQ.1) CALL ICLOSE(J4,1)
        IF(J5.EQ.2) REWIND J5
        IF(J5.EQ.2) CALL ICLOSE(J5,1)
        JOP = -1
        CALL PAUSE(2)
    RETURN
    52 DO 53 1=1,NRA
        IF(KORF0.EQ.'K') WRITE(I6,217) I
```

```
    217 FORMAT(' Enter row ',13,' of matrix A: ')
        READ(J4,*,END=500) (AM(I,J),J=1,NCA)
    53 CONTINUE
        IF(KORF0.EQ.'F') REWIND J4
        IF(JOP.EQ.3) GOTO 54
C
C... READ IN MATRIX B
C
        KORF = 'F'
        IF(KORF.EQ.'K') J5 = 0
        IF(KORF.EQ.'K') GOTO 56
        J5 = 2
        WRITE(I6,219)
    219 FORMAT(/' Enter name of file containing matrix B: ')
        CALL FNRDIN(FNAME)
        IF(LZ.LT.0) RETURN
        CALL FILOPN(0,J5,FNAME)
        IF(LZ.LT.0) RETURN
    56 IF(KORF.EQ.'K') WRITE(I6,220)
    220 FORMAT(/" Enter number of rows and columns for matrix B: ')
       READ(J5,*,END=500) NRB,NCB
       IF(NRB.GT.MAXR.OR.NCB.GT.MAXC) GOTO 57
C
C... CHECK CONSISTENCY OF MATRIX DIMENSIONS
C
    54 GOTO (71,72,80,73,80,73), IOPM1
    71 IF(NRA.EQ.NRB.AND.NCA.EQ.NCB) GOTO 80
        CALL TONE
        WRITE(I6,221)
    221 FORMAT(/' Error - Matrices A and B do not have compatible',
        1' dimensions.')
        LZ = -9
        CALL PAUSE(2)
    RETURN
    72 IF(NCA.EQ.NRB) GOTO 80
        CALL TONE
        WRITE(I6,221)
        LZ = -9
        CALL PAUSE(2)
    RETURN
    73 IF(NRA.EQ.NCA) GOTO 80
        CALL TONE
        WRITE(I6,222)
    222 FORMAT(/' Error - Matrix A is not a square matrix.')
        LZ = -9
        CALL PAUSE(2)
    RETURN
    80 IF(JOP.EQ.3) GOTO 59
C
C... DIMENSIONS OF A AND B ARE COMPATIBLE. NOW READ IN MATRIX B.
C
        DO 58 I=1,NRB
        IF(KORF.EQ.'K') WRITE(I6,223) I
    223 FORMAT(' Enter row ',13,' of matrix B: ')
        READ(J5,',END=500) (BM(I,J),J=1,NCB)
    58 CONTINUE
        IF(KORF.EQ.'F') REWIND J5
C
C... DETERMINE DIMENSIONS OF C: THE SOLUTION MATRIX
C
    59 GOTO (81,82,81,81,83,81), IOPM
    81 NRC = NRA
       NCC = MCA
   RETURN
   82 NRC = NRA
      NCC = NCB
   RETURN
   83 NRC = NCA
      NCC = NRA
   RETURN
```

```
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   PAUSE AFTER PRINTOUT OF RESULTS ON MONITOR
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE PAUSE(IOP)
        COMMON /IOLUN/ l5,l6,J5,J6,J4
        COMMON  /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
    CHARACTER*1 IPAUSE
C
        IF(IOP.EQ.3) GOTO 100
        IF(IOP.EQ.1) WRITE(16,201)
    201 FORMAT(/' Press RETURN to go to the main menu of options')
        IF(IOP.EQ.2) WRITE(I6,202)
    202 FORMAT(/'  Press RETURN to return to the menu of options')
        IF(IOP.EQ.0) WRITE(I6,200)
    200 FORMAT(/' Press RETURN to continue')
    101 READ(I5,299) IPAUSE
    299 FORMAT(A1)
    RETURN
C
    100 IF(MON.EQ.'Y') WRITE(I6,202)
        IF(MON.EQ.'N'.AND.IPRT.EQ.'N') WRITE(I6,202)
        IF(MON.EQ.'N'.AND.IPRT.EQ.'Y') WRITE(I6,203)
    203 FORMAT(/' Output file created')
        IF(MON.EQ.'N'.AND.IPRT.EQ.'Y') WRITE(I6,202)
        GOTO 101
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   CHECKS IF FILE IS NEW OR OLD, THEN OPENS IT
C
C   IUNK = 0: FILE SHOULD EXIST
C        = 1: UNKNOWN FILE
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE FILOPN(IUNK,IFIL,FNAME)
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
    CHARACTER*16 FNAME
    LOGICAL*2 FILEOK
C
        IF(IUNK.EQ.0) GOTO 100
        INQUIRE(FILE=FNAME,EXIST=PILEOK)
        IF(FILEOK) OPEN(IFIL,FILE=FNAME,STATUS='OLD',FORM='FORMATTED')
        IF(NOT.FILEOK) OPEN(IFIL,FILE=FNAME,STATUS='NEW',FORM='FORMATTED')
        CALL FILADD(IFIL)
    RETURN
    100 INOUIRE(FILE=FNAME,EXIST=FILEOK)
        IF( FILEOK) GOTO 102
       WRITE(16,200) FNAME
    200 FORMAT(/' File ',A16,' does not exist --',
        1//' Enter correct file name or press RETURN to abort: ')
        CALL FNRDIN(FNAME)
        IF(LZ.LT.0) RETURN
        GOTO 100
    102 OPEN(IFIL,FILE=FNAME,STATUS='OLD',FORM='FORMATTED')
        CALL FILADD(IFIL)
    RETURN
    END
```

```
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   KEEPS TPACK OF ALL LUN OF CURRENTLY OPENED FILES
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE FILADD(IFIL)
        COMMON /FLINDX/ IFLUN(20)
C
        DO 100 1=1,20
        IF(IFLUN(I).EQ.0) GOTO 101
    100 CONTINUE
    101 J = l
        IFLUN(J) = IFIL
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   READS IN FILE NAME AND CHECKS FOR LENGTH AND
C   NONSTANDARD CHARACTERS
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE FNRDIN(FNAME)
        COMMON /IOLUN/ I5,I6,J5,JG,J4
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1 IALPHA
    CHARACTER*16 FNAME
    CHARACTER*1 ANAME(16),CHOL
C
C   EQUIVALENCE (ANAME(1),FNAME)
C
    105 READ(I5,200) ANAME
    200 FORMAT(16A1)
        WRITE(FNAME,'(16A1)') ANAME
        IF(FNAME.NE.IALPHA(I)) GOTO 100
        LZ = -9
        CALL CLSALL
    RETURN
C
C... CHECK FOR TOTAL LENGTH
C
    100 DO 101 I = 1,16
        IF(ANAME(I).EQ.IALPHA(1)) GOTO 102
    101 CONTINUE
        NT = 16
        GOTO 103
    102 NT = I - 1
C
C... CHECK FOR PERIOD = IALPHA(15)
C
    103 IF(NT.LE.O) GOTO 300
        DO 104 1=1,NT
        IF(ANAME(I).EQ.IALPHA(15)) GOTO 106
    104 CONTINUE
        IP = 0
        NL = NT
        NR = 0
        GOTO 107
    106 IP = I
        NL = IP - 1
        NR = NT - IP
    107 IF(NL.LE.0.OR.NL.GT.8) GOTO 300
        IF(NR.GT.3) GOTO 300
C
C... CHECK FOR APPROPRIATE CHARACTERS. CAN BE DIGIT, LOWER OR UPPER
C... CASE LETTER.
C
        DO 108 1=1,NT
        IF(I.EQ.IP) GOTO 108
        CHOL = ANAME(I)
```

```
        IC = ICHAR(CHOL)
C
C... 48 TO 57 - ASCII RANGE FOR DIGITS
 C
        IF(IC.GE.48.AND.IC.LE.57) GOTO 108
C
C... 65 TO 90 - ASCII RANGE FOR UPEER CASE LETTERS
C
        IF(IC.GE.65.AND.IC.LE.90) GOTO 108
C
C... 97 TO 122 - ASCII RANGE FOR LOWER CASE LETTERS
C
        IF(IC.GE.97.AND.IC.LE.122) GOTO 108
        GOTO 300
    108 CONTINUE
    RETURN
    300 CALL INVLID(2)
        GOTO 105
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   CLOSE ALL CURRENTLY OPENEO FILES
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE CLSALL
        COMMON /FLINDX/ IFLUN(20)
        DO 100 1=1,20
        IF(IFLUN(I).EQ.0) GOTO 100
        JFIL = IFLUN(I)
        CALL ICLOSE(JFIL,1)
    100 CONTINUE
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   CLOSE FILE I, IOP = 0 IS DELTE, IOP = 1 IS KEEP
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE ICLOSE(I,IOP)
        COMMON /FLINDX/ IFLUN(20)
        IF(IOP.EQ.0) CLOSE(I,STATUS='DELETE')
        IF(IOP.EQ.1) CLOSE(I,STATUS='KEEP')
        DO 100 J=1,20
        IF(IFLUN(J).EQ.I) IFLUN(J) = 0
    100 CONTINUE
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   DATA ENTRY SUBROUTINE
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE DATENT
        COMMON /TLC/ SV(20), LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /DATQ/ A(100,2),B(2,2),C(2),D(100,4)
        COMMON /NAMEF/ FNAME
    CHARACTER*16 FNAME
C
        NCMAX = 100
        CALL CLEAR
        WRITE(I6,205)
    205 FORMAT(/' DATA ENTRY MODULE',//)
        WRITE(I6,200)
    200 FORMAT(/' Enter name of file to store data. ')
        CALL FNRDIN(FNAME)
        IF(LZ.LT.0) RETURN
        CALL FILOPN(1,20,FNAME)
    104 WRITE(I6,202)
    202 FORMAT(/' Enter number of rows and columns of data matrix: ')
```

```
    102 CALL IREAD(2,NROW,NCOL,IO5,IO5)
        IF(LZ.LT.0) RETURN
        IF(NROW.GE.1.AND.NCOLGE.1) GOTO 101
        CALL INVLID(I)
        GOTO 102
    101 IF(NCOLLE.NCMAX) GOTO 105
        WRITE(I6,206) NCMAX
    206 FORMAT(/' Number of columns exceed max of ',14,' - Try again')
        GOTO 104
    105 WRITE(20,*) NROW,NCOL
        DO 100I = 1,NROW
        DO 103J = 1,NCOL
        WRITE(I6,203) I,J
    203 FORMAT('  Enter value for row ',I4,', column ',14,
        1' of data matrix: ')
        CALL RREAD(1,ZA,AO5,AO5,AO5)
        IF(LZ.LT.0) RETURN
        A(J,1) = ZA
    103 CONTINUE
        WRITE(20,*) (A(J,1),J=1,NCOL)
    100 CONTINUE
        REWIND 20
        CALL ICLOSE(20,1)
        WRITE(I6,204) FNAME
    204 FORMAT(/' Data now stored in file ',A16)
        CALL PAUSE(2)
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C  subroutine CLEAR clears the monitor screen
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
    SUBROUTING CLEAR
    CHARACTER*1 ESC
    CHARACTER*3 SEQ
    CHARACTER*4 CLS
        SEQ = (2J)
        ESC = CHAR(027)
        WRITE (CLS,'(A1,A3)') ESC,SEQ
        WRITE (*,*) CLS
   RETURN
   END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   ASKS USER FOR DATE AND TIME TO BE WRITTEN ONTO ALL
C   PRINTOUT FILES
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE DATE
        COMMON /IOLUN/ l5,l6,J5,J6,J4
        COMMON /ASC/ IALPHA(96)
        COMMON /DATIM/ IDATE(8),ITIME(5)
    CHARACTER*1 lALPHA,IDATE,ITIME
    DIMENSION IN(6)
C
    105 WRITE(I6,200)
    200 FORMAT(//,' Enter current date (MM-DD-YY):
        READ(I5,201) IDATE
    201 FORMAT(8A1)
C
C... CHECK FOR Q, U OR H
C
        CALL QUH(IDATE(1),IQUIT)
        IF(1QUIT.EQ.1)STOP '   '
C
        IF(1DATE(3).NE.1ALPHA(14)) GOTO 100
        IF(IDATE(6).NE.IALPHA(14)) GOTO 100
C
```

```
        DO 102 I = 1,6
        IN(I) = -1000
    102 CONTINUE
C
        DO 101 I=1,96
        IF(IDATE(1).EQ.1ALPHA(I)) IN(1) = -17
        IF(IDATE(2).EQ.IALPHA(I)) IN(2) = -17
        IF(IDATE(4).EQ.IALPHA(I)) 1N(3) = -17
        IF(IDATE(5).EQ.IALPHA(I)) IN(4) = -17
        IF(IDATE(7).EQ.IALPHA(l)) IN(5) = -17
        IF(IDATE(8).EO.IALPHA(I)) IN(6) = -17
    101 CONTINUE
C
        DO 103 I=1,6
        IF(IN(I).EQ.-1000) GOTO 100
    103 CONTINUE
C
        IF(IN(1).NE.-16.AND.IN(1).NE.O.AND.1N(1).NE.1) GOTO 100
        IF(lN(2).LT.0.OR.IN(2).GT.9) GOTO 100
        IF(IN(1).LE.0.AND.IN(2).EQ.0) GOTO 100
        IF(IN(1).EQ.1.AND.IN(2).GT.2) GOTO 100
        IF(IN(3).GT.-16.AND.IN(3).LT.0) GOTO 100
        IF(IN(3).GT.3) GOTO 100
        IF(IN(4).LT.0.OR.IN(4).GT.9) GOTO 100
        IF(IN(3).LE.0.AND.IN(4).EQ.0) GOTO 100
        IF(IN(3).EQ.3.AND.IN(4).GT.1) GOTO 100
        GOTO 104
C
    100 CALL INVLID(1)
        GOTO 105
C
    104 WRITE(I6,203)
    203 FORMAT (/' Enter current time (HH:MM) ')
        READ(I5,204) ITIME
    204 FORMAT(5A1)
C
C... CHECK FOR Q, U OR H
C
        CALL QUH(ITIME(1),IQUIT)
        IF(IQUIT.EQ.I) GOTO 105
C
        IF(ITIME(3).NE.lALPHA(27)) GOTO 106
C
        DO 107 I = 1,4
        IN(I) = 1000
    107 CONTINUE
C
        DO 108 I = 1,96
        IF(ITIME(1).EQ.IALPHA(1)) IN(1) = I - 17
        IF(ITIME(2).EQ.IALPHA(I)) IN(2) = I - 17
        IF(ITIME(4).EQ.IALPHA(I)) IN(3) = I - 17
        IF(ITlME(5).EQ.IALPHA(I)) IN'4) = I - 17
    108 CONTINUE
C
        DO 109 I = 1,4
        IF(IN(I).EQ.-1000) GOTO 106
    109 CONTINUE
C
        IF(1N(1).GT 16.AND.IN(1).LT.O) GOTO 106
        IF(1N(1).GT.2) GOTO 106
        IF(IN(2).LT.0.OR.IN(2).GT.9) GOTO 106
        IF(1N(1).EQ.2.AND.IN(2).GT.4) GOTO 106
        IF(IN(3).LT.0.OR.IN(3).GT.5) GOTO 106
        IF(IN(4).LT.0.OR.IN(4).GT.9) GOTO 106
        CALL PAUSE(1)
    RETURN
C
    106 CALL INVLID(1
        GOTO 104
    END
```

```
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   CHECKS FOR Q, U OR H
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE QUH(ICH,IQUIT)
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1 lALPHA,ICH
C
        IQUIT = 0
        IF(ICH.EQ.IALPHA(50).OR.ICH.EQ.IALPHA(82)) IQUIT = 1
        1F(ICH.EQ.IALPHA(54).OR.ICH.EQ.IALPHA(86)) IQUIT = 1
        IF(ICH.EQ.IALPHA(41).OR.ICH.EQ.IALPHA(73)) IQUIT = 1
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   WRITES INVALID ENTRY MESSAGE
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE INVLID(IOP)
        COMMON /IOLUN/ I5,I6,J5,J6,J4
C
        CALL TONE
        IF(IOP.EQ.I) WRITE(16,200)
    200 FORMAT(/' Invalid entry, Try again ')
        IF(IOP.EQ.2) WRITE(I6,201)
    201 FORMAT(/' Invalid file name, Try again or press RETURN 10 abort '\
        IF(IOP.EQ.3) WRITE(16,202)
    202 FORMAT(/' Error - End of file encountered.',
        1/,9X,'You must recreate your data file.')
        IF(IOP.EQ.4) WRITE(I6,203)
    203 FORMAT(/' Error - The dimensions of the data matrix are',
        1/,9X, 'invalid for this module.')
        IF(IOP.EQ.5) WRITE(I6,204)
    204 FORMAT(/' An internal error has occurred in the program.',
        1/' Please check if your data is appropriate for this option.')
    RETURN
    END
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   INITIALIZE CHARACTER ARRAYS
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE INITTS
        COMMON /FLINDX/ IFLUN(2O)
        COMMON /CHAR/ ISOUT(180),LPT
        COMMON /TLC/ SAVE(20),LZ,LU,Z1,Z2,D1,D2,NZ,NSTEP,DEPINC,DFLT,UPA
        COMMON /FLS/ K5,K6,K7,K8,K9,K10,K11,K12,K13,IPCH,ILP,IPLOT,NCHAN
        COMMON /ASC/ IALPHA(96)
    DIMENSION JALPHA(96)
        COMMON /HEADLN/ IYN(2)
   CHARACTER*1 IALPHA,JALPHA,IYN
C
            DATA JALPHA '/',',',',','I','"','#','$','%','&','"','(',')',
        2               '*','+',',',',','-','.','/','0','1','2','3',
        3               '4','5','6','7','8','9',':',';','<','=',
        4               '>','?','@','A','B','C','D','E','F','G',
        5               'H','T','J','K','L','M','N','0','P','Q',
        6               'R','S','T','U','V','W','X','Y','Z','[',
        7               '\',']','^','_','`','a','b','c','d','e',
        8               'f','g','h','i','j','k','l','m','n','o',
        9               'p','q','r','s','t','u','v','w','x','y',
        1               'z','{','|','}',' ','~','~'/
C
        DO 100 I = 1,96
        IALPHA(I) = JALPHA(I)
    100 CONTINUE
C
C... IYN(1) IS N(0), IYN(2) IS Y(ES)
```

```
C
        IYN(1) = IALPHA(47)
        IYN(2) = IALPHA(58)
C
        DO 101 I = 1,20
        IFLUN(I) = 0
    101 CONTINUE
C
        K5 = 0
        K6 = 0
        K7 = 0
        K8 = 0
        K9 = 0
        K10 = 0
        K11 = 0
        K12 = 0
        K13 = 0
C
        LZ = 0
C
        LPT = 0
        NCHAN = 60
        DFLT = -999
        UPA = 0
        ILP = 0
        IPLOT = 0
        IPCH = -13264
    RETURN
    END




C
CP32     PROGRAM 3-2
C
C   PROGRAM TO COMPUTE THE VARIANCE OF 'N' SAMPLES
C
    SUBROUTINE VARIAN
C
        COMMON /TLC/ SV(20),L2,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1 IALPHA
        COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0,ITEST
        WRITE(I6,200)
    200 FORMAT(/' UNIVARIATE STATISTICS MODULE',//)
C
C   SET SUMS TO ZERO
C
        SUMX = 0.0
        SUMX2 = 0.0
C
C   READ NUMBER  OF OBSERVATIONS TO BE USED
C
        IF(KORF.EQ.'K') WRITE(I6,201)
    201 FORMAT(/' Enter number of observations: ')
        READ (J5,*,END=500) NS,NCOL
        IF(NS.LE.1.OR.NCOL.NE.1) GOTO 501
        DO 100 I=1,NS
C
C   READ AN OBSERVATION  AND ADD TO SUM
C
        IF(KORF.EQ.'K') WRITE(I6,202) I
    202 FORMAT(' Enter observation ',14,': ')
        READ (J5,*,END=500) X
        SUMX = SUMX + X
        SUMX2 = SUMX2 + X * X
```

```
    100 CONTINUE
C
C... READ IN t-TEST INFORMATION
C
    110 WRITE(I6,205)
    205 FORMAT(/' Do you want to compute one-sample t statistic',
        1' - Y(es) or N(o): ')
        CALL AREAD(2,JO5,KO5,IO5,IO5)
        IF(LZ.LT.0) RETURN
        ITEST = IALPHA(JO5)
        IF(lTEST.EQ.'y') JTEST = 'Y'
        IF(ITEST.EQ.'n') ITEST = 'N'
        IF(ITEST.NE.'Y'.AND.ITEST.NE.'N') CALL INVLID(1)
        IF(ITEST.NE.'Y'.AND.ITEST.NE.'N') GOTO 110
        IF(ITEST.EQ.'N') GOTO 111
        WRITE(I6,207)
    207 FORMAT(/' Enter hypothetical mean of population: ')
        CALL RREAD(1,HYPM,AO5,AO5,AO5)
        IF(LZ,LT,0) RETURN
C
C   COMPUTE THE VARIANCE
C
    111 VAR=(FLOAT(NS) * SUMX2 - SUMX * SUMX) / FLOAT(NS * (NS - 1))
C
C   PRINT RESULTS
C
        IF(MON.EQ.'Y') WRITE (I6,2000) NS,SUMX,SUMX2,VAR
        IF(IPRT.EQ.'Y') WRITE(J6,2000) NS,SUMX,SUMX2,VAR
C
C   COMPUTE THE STANDARD DEVIATION AND MEAN
C
        STDEV = SQRT(VAR)
        AMEAN = SUMX / FLOAT(NS)
C
C   PRINT STANDARD DEVIATION AND MEAN
C
        IF(MON.EQ.'Y') WRITE (I6,2001) STDEV,AMEAN
        IF(IPRT.EQ.'Y') WRITE(J6,2001) STDEV,AMEAN
C
C... COMPUTE t STATISTIC
C
        IF(lTEST.EQ.'N') GOTO 112
        FLPT = FLOAT(NS)
        IF(STDEV.EQ.0.0) GOTO 502
        TSTATC = (SQRT(FLPT) * (AMEAN - HYPM)) / STDEV
C
C... PRINT t STATISTIC
C
        IF(MON.EQ.'Y') WRITE (I6,2002) TSTATC
        IF(IPRT.EQ.'Y') WRITE(J6,2002) TSTATC
    112 CALL PAUSE(3)
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 L2 = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
C
C... INVALID DATA MATRIX DIMENSIONS
C
    501 LZ = -9
        CALL CLSALL
        CALL INVLID(4)
        CALL PAUSE(2)
    RETURN
C
C... DIVIDE BY ZERO
```

```
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
        CALL PAUSE(2)
    RETURN
    2000 FORMAT (///,9X,25HNumber of Observations = ,10X,I10,//,
         112 X,22HSum of Observations = ,F20.5,//,
         234H Sum  of Squares of Observations = ,F20.5,//,
         37X,27HVariance of Observations = ,F20.5)
    2001 FORMAT (/.13X.21 HStandard Deviation = ,F20.5,//,
         111X,23HMean of Observations = ,F20.5)
    2002 FORMAT (/,9X,25HOne-Sample t Statistic = ,F20.5)
    END




C
C
CP34     PROGRAM 3-4
C
C   PR0GRAM TO COMPUTE
C     A. VARIANCE OF EACH VARIABLE
C     B. MEAN OF EACH VARIABLE
C     C. COVARIANCE BETWEEN VARIABLES
C     D. CORRELATION BETWEEN VARIABLES
C
    SUBROUTINE BICOR
C
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
C
        WRITE(I6,200)
    200 FORMAT(/' BIVARIATE STATISTICS MODULE',//)
C
C   SET SUMS TO ZERO
C
        SUMX1 = 0.0
        SUMX2 = 0.0
        SX1SQ = 0.0
        SX2SQ = 0.0
        SX1X2 = 0.0
C
C   READ NUMBER  OF OBSERVED PAIRS TO BE USED
C
        IF(KORF.EQ.'K') WRITE(I6,201)
    201 FORMAT(/' Enler number of observed pairs ')
        READ (JS,*,END=500) NS,NCOL
        IF(NS.LE.1.OR.NCOL.NE.2) GOTO 501
        DO 100 I=1,NS
C
C   READ AND OBSERVED PAIR AND ADD TO SUMS
C
        IF(KORF.EQ.'K') WRITE(I6,202) I
    202 FORMAT(' Enter X1, X2 of observed pair ,I4,
         READ(J5,*,END=500) X1,X2
         SUMX1 = SUMX1 + X1
         SUMX2 = SUMX2 + X2
         SX1SQ = SX1SQ + X1 * X1
         SX2SQ = SX2SQ + X2 * X2
         SX1X2 = SX1X2 + X1 * X2
    100 CONTINUE
C
C   PRINT SUMS
C
        IF(MON.EQ.'Y') WRITE(l6,2000) NS,SUMX1,SX1SO,SUMX2,SX2SQ,SX1X2
```

```
        IF(IPnT.EQ.'Y')WRITE(J6,2000) NS,SUMX1,SX1SQ,SUMX2,SX2SQ,SX1X2
        IF(MON.EQ.T) CALL PAUSE(O)
C
C   COMPUTE AND PRINT MEAN, VARIANCE AND STANDARD DEVIATION
C   OF VARIABLE X1
C
        AMEAN = SUMX1 / PLOAT(NS)
        VAR = (FLOAT(NS) * SX1SQ - SUMX1 * SUMX1) / FLOAT(NS * (NS - 1))
        STDEV1 = SQRT(VAR)
        IF(MON.EQ.'V') WRITE(I6,2001) AMEAN,VAR,STDEV1
        IF(IPRT.EQ.'Y') WRITE(J6,2001) AMEAN,VAR,STDEV1
C
C   COMPUTE AND PRINT MEAN, VARIANCE AND STANDARD DEVIATION
C   OF VARIABLE X2
C
        AMEAN = SUMX2 / FLOAT(NS)
        VAR=(FLOAT(NS) * SX2SQ - SUMX2 * SUMX2) / FLOAT(NS * (NS - 1))
        STDEV2 = SQRT(VAR)
        IF(MON.EQ.'Y') WRITE(16,2002) AMEAN,VAR,STDEV2
        IF(IPRT.EQ.'Y')WRITE (J6.2002) AMEAN,VAR,STDEV2
C
C   COMPUTE AND PRINT COVARIANCE BETWEEN X1 AND X2
C
        COV = (FLOAT(NS) * SX1X2 - SUMX1 * SUMX2) / FLOAT(NS * (NS - l))
        IF(MON.EQ.'Y') WRITE(16,2003) COV
        IF(IPRT.EQ.'Y')WRITE(J6,2003) COV
C
C   COMPUTE AND PRINT CORRELATION BETWEEN X1 AND X2
C
        DBY0 = STDEV1 * STDEV2
        IF(DBY0.EQ.0.0) GOTO 502
        COR = COV / (STDEV1 * STDEV2)
        IF(MON.EQ.'Y') WRITE(I6,2004) COR
        IF(IPRT.EQ.'Y') WRITE(J6,2004) COR
        CALL PAUSE(3)
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
C
C... INVALID DATA MATRIX DIMENSIONS
C
    501 LZ = -9
        CALL CLSALL
        CALL INVLID(4)
        CALL PAUSE(2)
    RETURN
C
C... DIVIDE BY ZERO
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
        CALL PAUSE(2)
    RETURN
    2000 FORMAT (//,.5X,28H Number of Observed Pairs = ,10X,I10,//,
         121X,12HSum  of X1 = ,F20.5,//.
         210X,23HSum of Squares of X1 = ,F20.5,//,
         321X,12HSum  of X2 = ,F20.5,//,
         410X,23HSum of Squares of X2 = ,F20.5,//,
         509X, 241HSum of Cross Products = ,F20.5)
    2001 FORMAT (/,20X,13HMean of X1 = ,F20.5,//,
         116X,17HVariance of X1 = ,F20.5,//,
         216X,27HStandard Deviation of X1 = ,F20.5)
    2002 FORMAT (/,20X,13HMean of X2 = ,F20.5,//,
```

```
     116X,17HVariance of X2 = ,F20.5,//,
     216X,27HStandard Deviation of X2 = ,F20.5)
2003 FORMAT (/,2X,31 HCovariance Between X1 and X2 = ,F20.5)
2004 FORMAT (/,1X,32HCorrelation Between X1 and X2 = ,F20.5)
    END




C
CP35     PROGRAM 3-5
C
C   PROGRAM FOR ONE WAY ANALYSIS OF VARIANCE
C
    SUBROUTINE ONEOVA
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,N22,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /IONAT/ KORF,MONJPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
        COMMON /DATQ/ P1SSA(100,2),BJNK(2,2),CJNK(2),X(100,4)
C
        WRITE(I6,200)
    200 FORMAT(/' ONE WAY ANALYSIS OF VARIANCE MODULE',//)
C
        NCMAX = 100
C
C   READ NUMBER OFREPUCATIONS PER SAMPLE AND NUMBER OF SAMPLES
C
        IF(KORF.EQ.'K') WRITE(I6,201)
    201 FORMAT(/' Enter number of replications per sample',
        1' and number of samples: ')
        READ(J5,*,END=500) NR,NS
        IF(NR.LE.1.OR.NS.LE.1) GOTO 501
        IF(NS.GT.NCMAX) GOTO 501
C
C   SET SUMS TO ZERO
C
        P1SST = 0.0
        P2SST = 0.0
        SSA = 0.0
        DO 100 J = 1,NS
        P1SSA(J,1) = 0.0
    100 CONTINUE
        DO 101 = 1,NR
C
C   READ A REPLICATE AND ADD TO SUMS
C
        IF(KORF.EQ.'K') WRITE(I6,202) I
    202 FORMAT(/' Enter results of replicate ,14,' for each sample
        READ(J5,*,END=500) (X(J,1),J=1,NS)
        DO 102 J = 1,NS
        P1SST = P1SST + X(J,1) * X(J,1)
        P2SST = P2SST + X(J,1)
        P1SSA(J,1) = P1SSA(J,1) + X(J 1)
    102 CONTINUE
    101 CONTINUE
C
C   CALCULATE SST
C
        SST = P1SST - P2SST * P2SST / FLOAT(NS * NR)
        NRSI = NS * NR - 1
C
C   CALCULATE SSA
C
        DO 103 J = 1,NS
        SSA = SSA + P1SSA(J,1) * P1SSA(J,1 ) / FLOAT(NR)
    103 CONTINUE
        SSA = SSA - P2SST * P2SST / PLOAT(NR * NS)
        NS1 = NS - 1
```

```
C
C   CALCULATE SSW
C
        SSW = SST - SSA
        NNS = NR * NS - NS
C
C   CALCULATE MEAN SQUARES
C
        AMSA = SSA / FLOAT(NS1)
        AMSW = SSW / FLOAT(NNS)
C
C   CALCULATE F-RATIO
C
        IF(AMSW.EQ.0.0) GOTO 502
        F = AMSA / AMSW
C
C   PRINT RESULTS
C
        IF(MON.EQ.'N') GOTO 104
        WRITE(I6,2000)
        IF(SSA.LT 1.0E7) WRITE(I6,2001) SSA,NS1,AMSA,F
        IF(SSA.GE.1.0E7) WRITE(I6,2004) SSA,NS1,AMSA,F
        IF(SSW.LT.1.0E7) WRITE(I6,2002) SSW,NNS,AMSW
        IF(SSW.GE.1.0E7) WRITE(I6,2005) SSW,NNS,AMSW
        IF(SST.LT.1.0E7) WRITE(I6,2003) SST,NRS1
        IF(SST.GE.1.0E7) WRITE(I6,2006) SST,NRS1
    104 IP(lPRT.EQ.'N') GOTO 150
        WRITE(J6,2000)
        IF(SSA.LT.1.0E7) WR1TE(J6,2001) SSA,NS1,AMSA,F
        IF(SSA.GE.1.0E7) WRITE(J6,2004) SSA,NS1,AMSA,F
        IF(SSW.LT.1.0E7) WRITE(J6,2002) SSW,NNS,AMSW
        IF(SSW.GE.1.0E7) WRITE(J6,2005) SSW,NNS,AMSW
        IF(SST.LT.1.0E7) WRITE(J6,2003) SST,NRS1
        IF(SST.GE.1.0E7) WRITE(J6,2006) SST,NRS1
    150 CALL PAUSE(3)
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
C
C   INVALID DATA MATRIX DIMENSIONS
C
    501 LZ = -9
        CALL CLSALL
        CALL INVLID(4)
        CALL PAUSE(2)
    RETURN
C
C... DIVIDE BY ZERO
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
        CALL PAUSE(2)
    RETURN
    2000 FORMAT(///; Source of,14X,'Sum of Degrees of Mean',/
         1' Variation',14X,'Squares Freedom Squares F-Test',/,
         21X,60(1H-))
    2001 FORMAT(14H Among Samples,7X,F10.2,I10,F10.2,/,
         141X,F20.4)
    2004 FORMAT(14H Among Samples,7X,E10.3,I10,E10.3,/,
         141X,F20.4)
    2002 FORMAT(21H Within Replications ,F10.2,I10,F10.2)
    2005 FORMAT(21H Within Replications ,E10.3,I10,E10.3)
    2003 FORMAT(/,16H Total Variation,5X,F10.2,I10)
```

```
     2006 FORMAT(/,16H Total Variation,5X,E10.3,I10)
     END




C
CP35     PROGRAM 3-6
C
C    PROGRAM FOR TWO WAY ANALYSIS OF VARIANCE
C
      SUBROUTINE TWOVA
C
          COMMON /TLC/ SV(20),L2,LU,Z11,Z22,D11,D22,NZ2,NSP,DPC,DFT,UPA
          COMMON /IOLUN/ I5,I6,J5,J6,J4
          COMMON /IONAT/ KORF,MON,IPRT,KORF0
     CHARACTER*1 KORF,MON,IPRT,KORF0
          COMMON /DATQ/ P1SSA(100,2),BJNK(2,2),CJNK(2),X(100,4)
C
          WRITE(I6,200)
     200 FORMAT(/' TWO WAY ANALYSIS OF VARIANCE MODULE',//)
C
          NCMAX = 100
C
C    READ NUMBER  OF TREATMENTS AND NUMBER OF SAMPLES
C
         IF(KORF.EQ.'K') WRITE(I6,201)
     201 FORMAT(/' Enter number of tieatments and number of samples ,'/)
         READ(J5,*,END=500) NT,NS
         IF(NT.LE.1.OR.NS.LE.1) GOTO 501
         IF(NS.GT.NCMAX) GOTO 501
C
C    SET SUMS TO ZERO
C
         P1SST = 0.0
         P2SST = 0.0
         SSA = 0.0
         SSB = 0.0
         DO 100 J = 1,NS
         P1SSA(J,1) = 0.0
     100 CONTINUE
         DO 101 I=1,NT
C
C    READ A TREATMENT AND ADD TO SUMS
C
         IF(KORF EQ 'K") WRITE(I6,202) I
     202 FORMAT(' Enter results of treatment ',14, for each sample ,'/)
         READ (J5,*,END=500) (X(J,1),J=1,NS)
         TSSB = 0.0
         DO 102 J = 1,NS
         P1SST = P1SST + X(J,1) * X(J,1)
         P2SST = P2SST + X(J,1)
         P1SSA(J,1) = P1SSA(J,1) + X(J,1)
         TSSB = TSSB + X(J,1)
     102 CONTINUE
         SSB = SSB + TSSB * TSSB / FLOAT(NS)
     101 CONTINUE
C
C    CALCULATE SST
C
         SST = P1SST + P2SST * P2SST / FLOAT(NS * NT)
         NTS1 = NS * NT - 1
C
C    CALCULATE SSA
C
         DO 103 J = 1,NS
         SSA = SSA + P1SSA(J,1) * P1SSA(J,1) / FLOAT(NT)
     103 CONTINUE
         SSA = SSA - P2SST * P2SST / FLOAT(NT * NS)
```

```
        NS1 = NS - 1
C
C   CALCULATE SSB
C
        SSB = SSB - P2SST * P2SST / FLOAT(NS * NT)
        NT1 = NT - 1
C
C   CALCULATE SSE
C
        SSE = SST - (SSA + SSB)
        NT1S1 = (NS - 1) * (NT - 1)
C
C   CALCULATE MEAN SQUARES
C
        AMSA = SSA / FLOAT(NS1)
        AMSB = SSB / FLOAT(NT1)
        AMSE = SSE / FLOAT(NT1S1)
C
C   CALCULATE F-TESTS
C
        IF(AMSE.EQ.0.0) GOTO 502
        F1 = AMSA / AMSE
        F2 = AMSB / AMSE
C
C   PRINT RESULTS
C
        IF(MON.EQ.'N') GOTO 104
        WRITE (I6.2000)
        IF(SSA.LT.1.0E7) WRITE(16,2001) SSA,NS1,AMSA,F1
        IF(SSA.GE.1.0E7) WRITE(16,2005) SSA,NS1,AMSA,F1
        IF(SSB.LT.1.0E7) WRITE(16,2002) SSB,NT1,AMSB,F2
        IF(SSB.QE.1.0E7) WRITE(I6,20G6) SSB,NT1,AMSB,F2
        IF(SSE.LT.1.0E7) WRITE(16,2003) SSE,NT1S1,AMSE
        IF(SSE.GE.1.0E7) WRITE(16,2007) SSE,NT1S1,AMSE
        IF(SST.LT 1.0E7) WRITE(16,2004) SST,NTS1
        IF(SST.GE 1.0E7) WRITE(16,2008) SST,NTS1
    l4 IF(IPRT.EQ.'N') GOTO 150
        WRITE (J6.2000)
        IF(SSA.LT.1.0E7) WRITE(J6,2001) SSA,NS1,AMSA,F1
        IF(SSA.GE.1.0E7) WRITE(J6,2005) SSA,NS1,AMSA,F1
        IF(SSB.LT.1.0E7) WRITE(J6,2002) SSB,NT1,AMSB,F2
        IF(SSB.GE.1.0E7) WRITE(J6,2006) SSB,NT1,AMSB,F2
        IP(SSE.LT.1.0E7) WRITE(J6,2003) SSE,NT1S1,AMSE
        IF(SSE.GE.1.0E7) WRITE(J6,2007) SSE,NT1S1,AMSE
        IF(SST.LT.1.0E7) WRITE(J6,2Q04) SST,NTS1
        IF(SST.GE.1.0E7) WRITE(J6,2008) SST,NTS1
    150 CALL PAUSE(3)
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
C
C... INVALID DATA MATRIX DIMENSIONS
C
    501 LZ = -9
        CALL CLSALL
        CALL INVUD(4)
        CALL PAUSE(2)
    RETURN
C
C... DIVIDE BY ZERO
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
```

```
        CALL PAUSE(2)
    RETURN
    2000 FORMAT (///,' Source of',14X,'Sum of Degrees of Mean',/
        1' Variation',13X,'Squares Freedom Squares F-Tests',/,
        21X,60(1H-))
    2001 FORMAT (14H Among Samples,7X,F10.2,18,2X.F10.3,',
        141X,F20.4)
    2005 FORMAT (14H Among Samples,7X,E10.3,I8,2X,E10.3,/,
        141X,F20.4)
    2002 FORMAT (17H Among Treatments,4X,F10.2,I8,2X,F10.3,/,41X,F20.4)
    2006 FORMAT (17H Among Treatments,4X,E10.3,18,2X,E10.3,/,41X,F20.4)
    2003 FORMAT (6H Error,15X,F10.2,l8,2X,F10.3)
    2007 FORMAT (6H Error,15X.E10.3,I3,2X,E10.3)
    2004 FORMAT (/,16H Total Variation,5X,F10.2,l8)
    2008 FORMAT (/,16H Total Variation,5X,E10.3,l8)
    END




C
CLINFIT LINFIT - PROGRAM 5-3
C
C   ROUTINE LINFIT
C
C   PROGRAM TO FIT A UNEAR REGRESSION.
C
C   ARRAY A CONTAINS X AND Y DATA THAT IS READ IN.
C   ARRAY B CONTAINS THE COEFFICIENTS OF THE UNKNOWN B'S IN THE
C   NORMAL EQUATIONS 5.7 AND 5.8.
C   ARRAY C ORIGINALLY IS A VECTOR THAT CONTAINS THE SUM OF THE Y'S
C   AND THE SUMS OF THE CROSSPRODUCTS OF X AND Y IN EQUATION 5.11
C   AFTER THE NORMAL EQUATIONS ARE SOLVED, ARRAY C CONTAINS THE
C   COEFFICIENTS OF THE REGRESSION EQUATION.
C   SUMS OF THE CROSSPRODUCTS OF X AND Y IN EQUATION 5.11.
C   ARRAY D CONTAINS X, Y, Y-CALCULATED. AND DEVIATIONS FOR ALL
C   DATA POINTS.
C
C   THE MAXIMUM NUMBER OF OBSERVATIONS IS 100.
C
C   SUBROUTINES NEEDED ARE READM, PRINTM, AND SLE.
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
    SUBROUTINE UNFIT
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,N22,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ l5,l6,J5,J6,J4
        COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTER*1 KORF,MON,IPRT,KORF0
        COMMON /DATQ/ A(100,2),B(2,2),C(2),D(100,4)
C
        NMAX = 100
        WRITE(I6,200)
    200 FORMAT(/' LINEAR REGRESSION MODULE'//)
C
C   READ X-Y DATA
C
        IF(KORF.EQ.'K') WRITE(16,201)
    201 FORMAT(/' Enter number of observed pairs: ')
        READ(JS,*,.END=500  N,NCOL
        IF(N.LE.1.OR.NCOL.NE.2) GOTO 501
        IF(N.LE.NMAX) GOTO 105
        WRITE(I6,202) NMAX
    202 FORMAT(/' Number of observed pairs exceed max of ',14)
        LZ = -9
        CALL PAUSE(2)
    RETURN
    105 DO 106 I = 1,N
        IF(KORF.EQ.'K') WRITE(I6,203) I
```

```
    203 FORMAT(/' Enter X, Y of observed pair ',14,': ')
        READ(J5,*,END=500) (A(I,J),J=1,2)
    106 CONTINUE
C
C... CALCULATE SUMS FOR LEAST SQUARES SOLUTION
C
        DO  100 I = 1,2
        C(I) = 0.0
        DO 101 J = 1,2
        B(I,J) = 0.0
    101 CONTINUE
    100 CONTINUE
        DO 102 I = 1,N
        B(1,1) = B(1,1) + 1.0
        B(1,2) = B(1,2) + A(I,1)
        B(2,2) = B(2,2) + A(I,1) * A(I.1)
        C(1) = C(1) + A(I,2)
        C(2) = C(2) + A(I,1) * A(I,2)
    102 CONTINUE
        B(2,1) = B(1,2)
C
C   SOLVE THE SIMULTANEOUS LINEAR EQUATIONS WHICH ARE OF THE FORM
C   OF 5.7 AND 5.8 IN THE TEXT.
C
        IF(MON.EQ.'Y') WRITE(I6,1001)
        IF(IPRT.EQ.T) WRITE(J6,1001)
        CALL PRINTM(0,B,2,2,2,2)
        IF(MON.EQ.'Y') WRITE(I6,1002)
        IF(IPRT.EQ.'Y') WRITE(J6,1002)
        CALL PRINTM(0,C,1,2,1,2)
C
        CALL SLE(B,C,2,2,1.0E-05)
        IF(LZ.LT.0) CALL PAUSE(2)
        IF(LZ.LT.0) RETURN
C
        SLOPE = C(2)
        IF(MON.EQ.'Y') WRITE(I6,1003)
        IF(IPRT.EQ.T) WRI7E(J6,1003)
        CALL PRINTM(0,C,1,2,1.2)
C
        DO 103 I = 1,N
        D(I,1) = A(I,1)
        D(I,2) = A(I,2)
        D(I,3) = C(1) + C(2) * D(I,1)
        D(I,4) = D(I,2) - D(I,3)
    103 CONTINUE
        IF(MON.EQ.'Y') WRITE(I6,1004)
        IF(IPRT.EQ.'Y') WRITE(J6,1004)
        CALL PRINTM(0,D,N,4,100,4)
C
C... CALCULATE ERROR MEASURES
C
        SY = 0.0
        SY2 = 0.0
        SYC = 0.0
        SYC2 = 0.0
        DO 104 I = 1,N
        SY = SY + D(I,2)
        SY2 = SY2 + D(I,2) * D(I,2)
        SYC = SYC + D(I,3)
        SYC2 = SYC2 + D(I,3) * D(I,3)
    104 CONTINUE
        SST = SY2 - SY * SY / FLOAT(N)
        SSR = SYC2 - SYC * SYC / FLOAT(N)
        SSD = SST - SSR
        IF(SST.EQ.0.0) GOTO 502
        R2 = SSR / SST
        IF(SLOPE.GE.0.0) R = SQRT(R2)
        IF(SLOPE.LT.0.0) R = -1.0 * SQRT(R2)
        IF(MON.EQ.'N') GOTO 110
```

```
        WRITE(16,2000) N
        WRITE(16,2001 )SST
        WRITE(16,2002) SSR
        WRITE(16,2003) SSD
        WRITE(16,2004) R2
        WRITE(16,2005) R
    140 IF(IPRT.EQ.'N') GOTO 150
        WRITE(J6,2000) N
        WRITE(J6,2001) SST
        WRITE(J6,2002) SSR
        WRITE(J6,2003) SSD
        WRITE(J6,2004) R2
        WRITE(J6,2005) R
    150 CALL PAUSE(3)
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
        CALL PAUSE(2)
    RETURN
C
C... INVALID DATA MATRIX DIMENSIONS
C
    501 LZ = -9
        CALL CLSALL
        CALL INVLID(4)
        CALL PAUSE(2)
    RETURN
C
C... DIVIDE BY ZERO
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
        CALL PAUSE(2)
    RETURN
    1001 FORMAT(///,' Coefficient Matrix of Unknown Parameters',
        1' in Normal Equations')
    1002 FORMAT(/' Vector of Sum of Y"s and Sum of Crossproducts',
        1' of X and Y')
    1003 FORMAT(/' The Estimated Parameters of the Regression Equation')
    1004 FORMAT(/' Column 1 = X Variable' /' Column 2 = Y Variable
        1' Column 3 = Y Value Based On Regression Equation'
        2' Column 4 = Column 2 - Column 3')
    2000 FORMAT(//,28H Number of Observed Pairs = ,15)
    2001 FORMAT(25H To al Sums of Squares = ,F20.5)
    2002 FORMAT(37H Sums of Squares Due To Regression = ,F20 5)
    2003 FORMAT(36H Sums of Squares Due To Deviation = ,F20 5)
    2004 FORMAT(19H Goodness of Fit = ,F15.6)
    2005 FORMAT(27H Correlation Coefficient = ,F15.6)
    END




C
CPRINTM    PRINTM - PROGRAM 4-2
C
C   SUBROUTINE TO PRINT A MATRIX
C   HAVING N ROWS AND M COLUMNS
C
    SUBROUTINE PRINTM(KFIL,A,N,M,N1,M1)
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /IONAT/ KORF,MON,IPRT,KORF0
    CHARACTERS KORF,MON,IPRT,KORF0
```

```
    DIMENSION A(N1,M1)
C
        IF(KFIL.EQ.1) CALL CLEAR
        IF(KFIL.EQ.1) WRITE(16,203)
    203 FORMAT (/' PRINT OUT MATRIX',//)
C
C... PRINT MATRIX OUT IN STRIPS OF 5 COLUMNS
C
        NROW = 0
        IF(KFIL.EQ.0) GOTO 102
C
C... READ MATRIX FROM INPUT DATA FILE
C
        IF(KORF.EQ.'K') WRlTE(l6,200)
    200 FORMAT(/' Enter number of rows and columns for matrix: ')
        READ(J5,*,END=500) N,M
        IF(N.LE.N1.AND.M.LE.M1) GOTO 103
        WRITE(I6,201) N1,M1
    201 FORMAT(/' Number of rows, columns exceed max of ',14,' by ',14)
        LZ = -9
    RETURN
    103 DO 104 I = 1 ,N
        IF(KORF.EQ.'K') WRITE(I6,202) I
    202 FORMAT(' Enter row ',14,' of matrix: ')
        READ(J5,*,END=500) (A(I,J),J=1,M)
   104 CONTINUE
C
C... NOW PRINT OUT MATRIX
C
    102 DO 100 1B=1,M,5
        IE = IB + 4
        IF(IE - M) 2,2,1
    1 IE = M
C
C... PRINT HEADING
C
    2 IF(MON.EQ.'Y') WRITE(I6,2000) (I,MB,IE)
        IF(IPRT.EQ.'Y') WRITE(J6,2000) (I,I=IB,IE)
        DO 101 J = 1,N
C
C... PRINT ROW OF MATRIX
C
        IF(MON.EQ.'Y') WRITE(I6,2001) J,(A(J,K),K=IB,IE)
        IF(IPRT.EQ.'Y')WRITE(J6,2001) J,(A(J,K),K=IB,IE)
        NROW = NROW + 1
        IF(MON.EQ.'Y'.AND.MOD(NROW,10).EQ.0) CALL PAUSE(0)
    101 CONTINUE
    100 CONTINUE
        IF(MON.EQ.'Y'.AND.MOD(NROW,10).NE.0.AND.KFIL.EQ.0) CALL PAUSE(0)
        IF(MON.EQ.'N'.AND.IPRT.EQ.'Y'.AND.KFIL.EQ.1) WRITE(16,204)
    204 FORMAT(/' Output file created')
    RETURN
C
C... END OF FILE ENCOUNTERED
C
    500 LZ = -9
        CALL CLSALL
        CALL INVLID(3)
    RETURN
    2000 FORMAT(/,5X,5(8X,14,3X),/)
    2001 FORMAT(I5,5F15.5)
    END
```

```
C
CADDM    ADDM - PROGRAM 4-3
C
C   SUBROUTINE TO ADD TWO MATRICES
C   A AND B TO FORM C. ALL HAVE N ROWS AND M COLUMNS
C
    SUBROUTINE ADDM(A,B,C,N,M,N1,M1)
    DIMENSION A(N1,M1),B(N1,M1),C(N1,M1)
        DO 100 I = 1,N
        DO 101 J = 1,M
        C(I,J) = A(I,J) + B(I,J)
    101 CONTINUE
    100 CONTINUE
    RETURN
    END
```

```
C
CCMULT     CMULT - PROGRAM 4-5
C
C   SUBROUTINE TO MULTIPLY EACH ELEMENT OF A MATRIX BY A
C   CONSTANT B TO FORM THE MATRIX C.
C   EACH MATRIX HAS N ROWS AND M COLUMNS.
C
C   C = B * MAT(A)
C
    SUBROUTINE CMULT(A,B,C,N,M,N1,M1)
    DIMENSION A(N1,M1),C(N1,M1)
        DO 100 I = 1,N
        DO 101 J = 1,M
        C(I,J) = B * A(I,J)
    101 CONTINUE
    100 CONTINUE
    RETURN
    END
```

```
C
CMMULT    MMULT - PROGRAM 4-7
C
C   SUBROUTINE FOR MULTIPLICATION OF MATRIX A BY MATRIX B
C   TO GIVE MATRIX C. A IS L ROWS BY N COLUMNS.
C   B IS N ROWS BY M COLUMNS, AND C WILL BE L ROWS BY M COLUMNS
C
    SUBROUTINE MMULT(A,B,C,L,N,M,NA,MA,NB,MB,NC,MC)
    DIMENSION A(NA,MA),B(NB,MB),C(NC,MC)
        DO 100 I = 1,L
        DO 101 J = 1,M
        C(I,J) = 0.0
        DO 102 K = 1,N
        C(I,J) = C(I,J) + A(I,K) * B(K,J)
    102 CONTINUE
    101 CONTINUE
    100 CONTINUE
    RETURN
    END
```

```
C
CMINV    MINV - PROGRAM 4-8
C
C   SUBROUTINE TO FIND INVERSE OF MATRIX A. B IS THE INVERSE
C   OF A. A IS REDUCED TO THE IDENTITY MATRIX.
C   A AND B ARE N X N. DET IS THE DETERMINANT OF A.
C
    SUBROUTINE MINV(A,B,N,N1,DET)
    DIMENSION A(N1,N1),B(N1,N1)
        COMMON /IOLUN/ I5,I6,J5,J6,J4
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
C
        ZERO = 1.0E-05
C
C   SET B TO IDENTITY MATRIX AND SAVE THE ORIGINAL A MATRIX
C
        DO 100 I = 1,N
        DO 101 J = 1,N
        B(I,J) = 0.0
    101 CONTINUE
        B(I,I) = 1.0
    100 CONTINUE
        DET = 1.0
C
C   CALCULATE INVERSE
C
        DO 102 I = 1,N
C
C   DIVIDE ITH ROW OF A AND B BY A(l,l)
C
        DlV = A(I,I)
        IF(ABS(DIV) - ZERO) 99,99,2
    2 DET = DET * DIV
        D0 103 J = 1,N
        A(I,J) = A(I,J) / DIV
        B(I,J) = B(I,J) / DIV
    103 CONTINUE
C
C   REDUCE THE ITH COLUMN OF A TO ZERO
C
        DO 104 J = 1,N
        IF(I-J) 1,104,1
    1 RATIO = A(J,I)
        D0 105 K = 1,N
        A(J,K) = A(J,K) - RATIO * A(I,K)
        B(J,K) = B(J,K) - RATIO * B(I,K)
    105 CONTINUE
    104 CONTINUE
    102 CONTINUE
        WRITE(I6,201) DET
    201 FORMAT(/' Determinant of Matrix A F155)
    RETURN
C
C   MATRIX CANNOT BE INVERTED A DIAGONAL ELEMENT IS APPROXIMATEL ZERO
C
    99 DET = 0.0
        WRITE(I6,202)
    202 FORMAT(/' Matrix Cannot Be Inverted')
        LZ = -9
        CALL PAUSE(2)
        RETURN
        END
```

```
C
CSLE    SLE PROGRAM 4 9
C
C   SUBROUTINE FOR SOLUTION OF N SIMULTANEOUS EQUATIONS
C   MATRIX A IS N X N AND B IS A COLUMN VECTOR OF N ELEMENTS
C   A IS CONVERTED TO THE IDENTITY MATRIX
C   B CONTAINS SOLUTION
C
    SUBROUTINE SLE(A,B,N,N1,ZERO)
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D11,D22,NZZ,NSP,DPC,DFT,UPA
        COMMON /IOLUN/ I5,I6,J5,JP,J4
    DIMENSION A(N1,N1),B(N1)
        DO 100 I = 1,N
        DIV = A(I,I)
        IF(ABS(DIV) - ZERO) 99,99,1
    1 DO 101 J=1,N
        A(I,J) = A(1,J) / DIV
    101 CONTINUE
        B(I) = B(I) / DIV
        D0 102 J = 1,N
        IF(I-J) 2,102,2
    2 RATIO=A(J,I)
        DO 103 K = 1,N
        A(J,K) = A(J,K) - RATIO * A(I,K)
    103 CONTINUE
        B(J) = B(J) - RATIO * B(I)
    102 CONTINUE
    100 CONTINUE
    RETURN
    99 WRITE(16,200)
    200 FORMAT(/' Matrix Cannot Be Inverted. System of,
        1/' Simultaneous Linear Equations Cannot Be Solved.')
        LZ = -9
    RETURN
    END
```

```
C
CEIGENJ   EIQTNJ - PROGRAM 4-10
C
C   SUBROUTINE TO CALCULATE. THE EIGENVALUES AND EIGENVECTORS
C   OF AN NXN SYMMETRIC MATRIX.
C
C   UPON  COMPLETION THE EIGENVALUES ARE STORED IN THE DIAGONAL
C   ELEMENTS OF MATRIX A (IN DESCENDING ORDER). THE EIGENVECTORS
C   ARE STORED BY COLUMNS  IN MATRIX B.
C
C   EIGENVALUE A(1,1) CORRESPONDS TO EIGENVECTOR (B(J,I),J=1,N)
C
    SUBROUTINE EIGENJ(A,B,N,N1)
    DIMENSION A(N1,N1),B(N1,N1)
        COMMON /TLC/ SV(20),LZ,LU,Z11,Z22,D1I,D22,NZZ,NSP,DPC,DFT,UPA
C
C..  CALCULATE INITIAL AND FINAL NORMS
C    SET B TO IDENTITY MATRIX
C
        IF(N.LE.0) GOTO 502
        ANORM = 0.0
        DO 100 I = 1,N
        DO 101 J = 1,N
        IF(I-J) 2,1,2
    1 B(I,J) = 1.0
        GOTO  101
    2 B(I,J) = 0.0
        ANORM = ANORM + A(I,J) * A(I,J)
    101 CONTINUE
    100 CONTINUE
```

```
        ANORM = SQRT(ANORM)
        FNORM = ANORM * 1.0E-09 / FLOAT(N)
C
C... INITIALIZE INDICATORS AND COMPUTE THRESOLD
C
        THR = ANORM
    23 THR = THR / FLOAT(N)
    3 IND = 0
C
C... SCAN DOWN COLUMNS FOR OFF-DIAGONAL ELEMENTS
C    GREATER THAN OR EQUAL TO THRESOLD
C
        DO 102 I = 2,N
        I1 = I-1
        DO 103 J = 1,I1
        IF(ABS(A(J,I))-THR) 103,4,4
C
C... COMPUTE SIN AND COS
C
    4 IND = 1
        AL = -A(J,I)
        AM = (A(J,J) - A(I,I)) / 2.0
        DBY0 = SQRT(AL * AL + AM * AM)
        IF(DBY0.EQ.0.0) GOTO 502
        AO = AL / SQRT(AL * AL + AM * AM)
        IF(AM) 5,6,6
    5 AO = -AO
    6 SINX = AO / SORT(2.0(1.0 + SQRT(1.0 - AO * AO)))
        SINX2 = SINX * SINX
        COSX = SQRT(1.0 - SINX2)
        COSX2 = COSX * COSX
C
C... ROTATE COLUMNS I AND J
C
        DO 104 K = 1,N
        IF (K-J) 7,10,7
    7 IF (K-l) 8,10,8
    8 AT = A(K,J)
        A(K,J) = AT * COSX - A(K,I) * SINX
        A(K,I) = AT * SINX + A(K,1) * COSX
    10 BT = B(K,J)
        B(K,J) = BT * COSX - B(K,I) * SINX
        B(K,I) = BT * SINX + B(K,I) * COSX
    104 CONTINUE
        XT = 2.0 * A(J,I) * SINX * COSX
        AT = A(J,J)
        BT = A(I,I)
        A(J,J) = AT * COSX2 + BT * SINX2 - XT
        A(I,I) = AT * SINX2 + BT * COSX2 + XT
        A(J,I) = (AT - BT) * SINX * COSX + A(J,I) * (COSX2 - SINX2)
        A(I,J) = A(J,I)
        D0 105 K = 1,N
        A(J,K) = A(K,J)
        A(I,K) = A(K,I)
    105 CONTINUE
    103 CONTINUE
    102 CONTINUE
        IF(IND) 20,20,3
    20 IF(THR-FNORM) 25,25,23
C
C... SORT EIGENVALUES AND EIGENVECTORS
C
    25 DO 110 I = 2,N
    29 IF (A(J-1,J-1)-A(J,J)) 30,110,110
    30 AT = A(J-1,J-1)
        A(J-1,J-1) = A(J,J)
        A(J,J) = AT
        DO 111 K = 1,N
        AT = B(K,J-1)
        B(K,J-1) = B(K,J)
```

```
        B(K,J) = AT
    111 CONTINUE
        J = J - 1
        IF(J-1) 110,110,29
    110 CONTINUE
    RETURN
C
C... DIVIDE BY ZERO
C
    502 LZ = -9
        CALL CLSALL
        CALL INVLID(5)
        CALL PAUSE(2)
    RETURN
    END




C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   SUBROUTINE MTRANS COMPUTES THE TRANSPOSE MATRIX C OF MATRiX A
C
C * * * * * * * * * * * * * * * * * * * * * * * *  * * * * * * * * * * * * * *
    SUBROUTINE MTRANS(A,C,NRA,NCA,MAXR,MAXC)
    DIMENSION A(MAXR,MAXC),C(MAXR,MAXC)
        DO 100 I = 1,NRA
        DO 101 J = 1,NCA
        C(J,I) = A(I,J)
    101 CONTINUE
    100 CONTINUE
    RETURN
    END




FUNCTION IJEOF(I)
    ,JEOP=0
    RETURN
    END
C./ ADD NAME=RREAD
C
C   TO READ N REAL.NUMBERSFROMKEYBOARD,CHECKING FOR
C   "QUIT", "UP" OP, "HELP"
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
    SUBROUTINE RREAD(N,A,B,C,D)
        COMMON /TLC/ SAVE(20),LZ,LU,Z1,Z2,D1,D2,NZ,NSTEP,DEPlNC,DFLT,UPA
        COMMON /FLS/ I5,I6,I7,I8,I9,I10,I11,I12,I13,lPCH,lLP,IPLOT,NCHAN
        COMMON /CHAR/ MN(720),LPT
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1 IALPHA,MN
        A = 0.0
        B = 0.0
        C = 0.0
        D = 0.0
        NUM = 0.0
C   WRITE(I6,9)
    9 FORMAT(1X)
    1 READ(I5,10) (MN(M),M=1,50)
    10 FORMAT(50A1)
        DUM = IJEOF(I5)
        IF(112.EQ.-99) WRITE(113,10) (MN(M),M=1,50)
C       IF(16.EQ.I13) WRITE(I6,11) (MN(M),M=1,50)
    11 FORMAT(1X,50A1)
        I1 = 1
```

```
        DO 20 I = 1,50
        I2 = 51 - I
        IF(MN(12).EQ.1ALPHA(1)) GOTO 20
        GOTO 30
    20 CONTINUE
        Z = 0.0
        GOTO 40
    30 CALLNUMB(I1,I2,IERR,Z)
        IF(LZ.LT.0) GOT0 90
        IF(IERR.EQ.0) GOTO 40
C
C... TONE FOR IBM PC ONLY
C
        CALL TONE
C
        WRITE(I6,110)
        GOTO 1
    40 NUM=NUM+1
        IF(NUM.EQ.1) A = Z
        IF(NUM.EQ.2) B = Z
        IF(NUM.EQ.3) C = Z
        IF(NUM.EQ.4) D = Z
        IF(I1.LE.I2)GOTO 30
        IF(NUM.LT.N) GOTO 1
    90 RETURN
    110 FORMAT(1X,13HInvalid entry,12H, Try again ')
    END
C./ADDNAME=IREAD
C
C   TO READ N INTEGERS FROM KEYBOARD
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    SUBROUTINE IREAD(N,I,J,K,L)
        COMMON /TLC/ SAVE(20),LZ,LU,Z1,22,D1,D2,NZ,NSTEP,DEPINC,DFLT,UPA
        CALL RREAD(N,A,B,C,D)
        IF(LZ.LT.0) RETURN
        I = A
        J = B
        K = C
        L = D
    RETURN
    END
C
C./ADD NAME=AREAD
C
C   TO READ N A1 CHARACTERS FROM KEYBOARD AS INTEGER VARIABLES
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *  * * * * * * * *
C
    SUBROUTINE AREAD(N,I,J,K,L)
        COMMON /TLC/ SAVE(20),LZ,LU,Z1,Z2,D1,D2,NZ,NSTEP,DEPINC,DFLT,UPA
        COMMON /FLS/ I5,I6,I7,I8,I9,I10,I11,I12,I13,IPCH,ILP,IPLOT,NCHAN
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1 IALPHA,IBL,M1,M2,M3,M4
C
C   DATA IBL/' '/
C
        IBL = IALPHA(1)
        IF(LZ.EQ.-5.AND.N.EQ.1) RETURN
        M1 = IBL
        M2 = IBL
        M3 = IBL
        M4 = IBL
C
C   WRITE(I6,9)
C
    9 FORMAT(1X)
        READ(I5,10) M1,M2,M3,M4
    10 FORMAT(4A1)
        DUM = IJEOF(I5)
```

```
          IF(112.EQ.-99) WRITE(I13,10) M1,M2,M3,M4
C         IF(I6.EQ.I13) WRITE(I6,110) M1,M2,M3,M4
      110 FORMAT(1X,4A1)
          LZ = 0
          CALL CHKALL(M1,M2,M3,M4)
          IF(LZ.LT.0) RETURN
          I = 1
          J = 1
          K = 1
          L = 1
          DO 120 KK = 1,96
          IF(I.EQ.1.AND.M1.EQ.IALPHA(KK)) I = KK
          IF(J.EQ.1.AND.M2.EQ.IALPHA(KK)) J = KK
          IF(K.EQ.1.AND.M3.EQ.IALPHA(KK)) K = KK
          IF(L.EQ.1.AND.M4.EQ.IALPHA(KK)) L = KK
      120 CONTINUE
      RETURN
      END
C
C./ ADD NAME=NUMB
C
C   CONVERTS  AN M STRING OF CHARACTERS  FROM MN(I1) TO MN(I2
C   TO A REAL NUMBER
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

C
      SUBROUTINE NUMB(I1,I2,IERR,A)
          COMMON /TLC/ SAVE(20),LZ,LU,Z1,Z2,D1,D2,NZ,NSTEP,DEPINC,DFLT,UPA
          COMMON /FLS/ E5,I6,I7,I8,I9,I10,I11,I12,I13, IPCH,ILP,IPLOT,NCHAN
          COMMON /ASC/ IALPHA(96)
          COMMON /CHAR/ MN(720),LPT
      DIMENSION NM(132)
      CHARACTERS IALPHA,MN,M1,M2,M3,M4
          A = DFLT
          LZ = 0
          M1 = MN(I1)
          M2 = MN(I1+1)
          M3 = MN(I1+2)
          M4 = MN(I1+3)
          CALL CHKALL(M1,M2,M3,M4)
          IF(LZ.LT.0) RETURN
          NDIG = 0
          NEG = 0
          IDEC = 999
          DO 100 I = 11,12
          IE=I
          IF(MN(I).NE.IALPHA(1)) GOTO 10
          IF(NEG.EQ.0) GOTO 100
          GOTO 200
      10 DO 20 K = 10,26
          IF(MN(1).NE.IALPHA(K)) GOTO 20
          IF(K.EQ.13.OR.K.EQ.16) GOTO 200
          IF(NEG.NE.0) GOTO 30
          IB = I
          IF(K.EQ.12) NEG = 1
          IF(K.EQ.14) NEG = -1
          IF(NEG.NE.0) GOTO 40
          NEG = 1
          IB = I - 1
          GOTO 30
      20 CONTINUE
          GOTO 900
C
C   AFTER FIRST DIGIT
C
      30 IF(K.LE.14.0R.K.EQ.1S)GOT0200
          NDIG = NDIG + 1
      40 IF(K.E0.15) IDEC = NDIG
          NM(I) = K - 17
```

```
    100 CONTINUE
C
C   END OF NUMBER
C
    200 A = 0.0
        IF(NDIG.GT.0) GOTO 210
        I1 = I1 + 1
    RETURN
    210 IF(NDIG*IDEC.EQ.1) GOTO 900
        DO 300 J = 1,NDIG
        IF(J.LT.IDEC) A=10.0 * A + FLOAT(NM(IB + J)
        IF(J.EQ.IDEC) GOTO 300
        IF(J.GT.IDEC) A = A + FLOAT(NM(IB + J)) * 10.0 * FLOAT(IDEC - J)
    300 CONTINUE
        A = A * FLOAT(NEG)
        IERR = 0
        I1 = IE + 1
    RETURN
C
C   ERFCR
C
    900 WRITE(I6,910)
    910 FORMAT(1X,2iHError in number field)
        IERR = -1
    RETURN
    END
C
C./ ADD NAME=CHKALL
C
C   CHECKS  FOR CHARACTERS  SPELLING "QUIT", "UP", "HELP" ETC
C
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
    SUBROUTINE CHKALL(M1,M2,M3,M4)
    CHARACTER*1 M1,M2,M3,M4
C
        CALL CHECK(M1,M2,M3,M4,52,53,48,49,-9)
        CALL CHECK(M1,M2,M3,M4,38,57,42,53.-9)
        CALL CHECK(M1,M2,M3,M4,34,35,48,51,-9)
        CALL CHECK(M1,M2,M1,M1.54,49,54.54,-1)
        CALL CHECK(M1,M1,M1,M1,50,5Q,50,50,-9)
        CALL CHECK(M1,M2,M1,M1,54,1,54,54,-1)
        CALL CHECK(M1,M2,M3,M4,41,38,45,49,-10)
        CALL CHECK(M1,M1,M1,M1,32,32,32,32,-10)
        CALL CHECK(M1,M2,M3,M4,36,48,49,58,-5)
        CALL CHECK(M1,M1,M1,M1,54,54,54,54,-1)
        CALL CHECK(M1,M1,M1,M1,41,41,41,41,-10)
        CALL CHECK(M1,M1,M1,M1,82,82,82,82,-9)
        CALL CHECK(M1,M1,M1,M1,85,86,86,86,-1)
        CALL CHECK(M1,M1,M1,M1,73,73,73,73,-10)
    RETURN
    END
C
C./ ADD NAME=CHECK
C
C   COMPARES STRING M1,M2,M3,M4 TO ASCIl CHARACTERS K1,K2,K3,K4
C * * * * * * * * * * * * * * * * * * * * * * *  * * * * * * * * * * * * * * *
C
    SUBROUTINE CHECK(M1,M2,M3,M4,K1,K2,K3,K4,L)
        COMMON /TLC/ SAVE(20),LZ,LU,Z1,Z2,D1.D2,NZ,NSTEP,DEPINC,DFLT,UPA
        COMMON /ASC/ IALPHA(96)
    CHARACTER*1  IALPHA,M1,M2,M3,M4
        IF(M1.NE.IALPHA(K1)) RETURN
        IF(M2.NE.IALPHA(K2)) RETURN
        IF(M3.NE.IALPHA(K3)) RETURN
        IF(M4.NE.IALPHA(K4)) RETURN
        LZ = L
    RETURN
    END
```

```
    This routine will create the audio tone (subroutine TONE)


     assume  cs:codes


    codes segment 'code'
    public tone
    tone proc far
        pusn bp             ;save frame pointer
        mov  bp,sp          ;address parameters

        mov  ax,0           ;numerator (low word)
        mov  dx,12h         (high word)
        mov  bx,1250        divisor
        div  bx             ;divide
        mov  bx.ax          ;save quotient in BX

    set 1/pitch into timer than turn on tone
        mov  a1,10110110b   ;put the number
        out  43h,al         ;into timer
        mov  ax,bx          1/pitch into AX
        out  42h,a!         ;LSB into timer2
        mov  al,ah          ;MSB  into timer2
        out  42h,al
        in   al,61h         ;read port B into AL
        or   al,3           ;turn on bits 0,1
        out  61h,al         lo turn on speaker

    sound the tone for a while, then turn it off
        mov  cx,02FFFh     ;set up for delay
wait    loop  wait          ;delay
       in    al,61h        ;read portB into AL
       and   al,1111100b   ;mask  lower two bits
       out   61h,al        to turn off speaker

       pop   bp            restore frame pointer
       ret                 return
tone   endp
codes  ends
       end
```