Федеральное агентство по образованию

Государственное образовательное учреждение высшего профессионального образования
Санкт-Петербургский государственный горный институт им. Г.В.Плеханова
(технический университет)

Г.С.ПОРОТОВ

МАТЕМАТИЧЕСКИЕ МЕТОДЫ МОДЕЛИРОВАНИЯ В ГЕОЛОГИИ

Допущено
Министерством образования и науки Российской Федерации
в качестве учебника для студентов высших учебных заведений,
обучающихся по направлению подготовки бакалавров и магистров
«*Геология и разведка полезных ископаемых*» *и направлению*
подготовки дипломированных специалистов «*Прикладная геология*»

САНКТ-ПЕТЕРБУРГ
2006

УДК 550.8:519.2
ББК 26.386
П595

В учебнике рассмотрены геологические объекты и их свойства, принципы математического моделирования. Проанализированы одно-, двух- и трехмерные статистические модели, в том числе метод главных компонент, кластерный анализ, распознавание образов. Приведены примеры применения этих моделей к решению геологических задач. Охарактеризованы модели пространственных переменных, в том числе случайные функции, периодическая изменчивость, основы геостатистики, кригинг и их применение в геологии. Дано понятие о базах и банках данных при моделировании месторождений, рассмотрены некоторые приемы обработки банков данных и построения геологических границ на плане и в разрезах.

Учебник соответствует стандарту дисциплины и предназначен для студентов геологических специальностей вузов по направлению «Прикладная геология» и для геологов-производственников.

Рецензенты: кафедра геологии месторождений полезных ископаемых Санкт-Петербургского государственного университета; профессор В.И.Щеглов (Южно-Российский государственный технический университет)

Поротов Г.С.

П595. Математические методы моделирования в геологии: Учебник / Г.С.Поротов. Санкт-Петербургский государственный горный институт (технический университет). СПб, 2006. 223 с. + вклейка.

ISBN 5-94211-140-5

УДК 550.8:519.2
ББК 26.386

ISBN 5-94211-140-5

ВВЕДЕНИЕ

При геологических исследованиях быстрыми темпами накапливается большое количество геологической информации: результаты геологической документации буровых скважин, горных выработок и естественных обнажений, спектральных и химических анализов руд, горных пород и минералов, данные геофизических и геохимических измерений и др. Одно из важнейших направлений научно-технического прогресса в геологии состоит в широком внедрении автоматизированных методов накопления, хранения, обработки и передачи геологической информации с целью повышения эффективности геологических исследований.

Научно-техническая революция в области информатики и вычислительной техники обусловила широкое внедрение в геологическую отрасль компьютеров и современных методов обработки геологической информации. Успешное использование математических методов и компьютеров невозможно без повышения уровня математического образования. Предлагаемый учебник в какой-то мере восполняет этот пробел. Читатель сможет получить представление о принципах и особенностях математического моделирования геологических объектов и явлений, овладеть основными методами математической, преимущественно статистической, обработки геологической информации и научиться применять их для решения геологических задач.

К настоящему времени накоплен большой опыт использования математических методов в геологии. Первые упоминания о применении статистических методов в геологии относятся к началу XIX в. Так, Ч.Ляйель в 30-х годах XIX в. использовал статистическое соотношение распространенности раковин моллюсков для

стратиграфического расчленения разрезов. В начале XX в. Д.В.Наливкин применил статистику для описания изменчивости свойств ископаемых организмов.

В конце XIX – начале XX в. с помощью статистических методов изучали распространение химических элементов в земной коре, что нашло отражение в работах Ф.В.Кларка, В.И.Вернадского, А.Е.Ферсмана, А.П.Виноградова.

В начале XX в. С.Ю.Доборжинский, В.И.Бауман и П.К.Соболевский заложили основы горной геометрии для математического моделирования тел полезных ископаемых. В дальнейшем это направление получило развитие в работах П.А.Рыжова, Н.И.Ушакова, З.Д.Низгурецкого, В.А.Букринского и других исследователей.

В первой половине XX в. П.Н.Чирвинский, П.Ниггли, Ф.Ю.Левинсон-Лессинг, Г.Розенбуш, А.Н.Заварицкий и другие исследователи на основе статистической обработки минерального и химического состава разработали классификацию магматических горных пород.

Статистика была использована для изучения изменчивости оруденения (В.В.Котульский, Н.К.Разумовский, Л.И.Шаманский, Д.А.Родионов), для решения вопросов опробования (Н.В.Барышев, П.Жи), для обоснования плотности разведочной сети (В.Г.Соловьев, Д.А.Зенков, П.Л.Каллистов), для оценки точности подсчета запасов (А.М.Журавский, К.Л.Пожарицкий, Л.И.Шаманский, Д.А.Казаковский).

Большое значение имеют работы по изучению пространственных переменных на месторождениях полезных ископаемых. Они привели к созданию теории геостатистики, основы которой были заложены Д.П.Криге и Ж.Матероном и получили развитие в трудах А.Карлье, М.Давида, В.И.Щеглова и Ю.Е.Капутина.

Применение математических методов при построении структурных и фациальных карт отражено в работах У.Крамбейна, Ф.Грейбилла, Р.Миллера, Д.Кана, Н.Н.Боровко. Статистические методы обработки геологической информации освещены в исследованиях И.П.Шарапова, А.Б.Вистелиуса, Д.Н.Родионова, В.В.Бондаренко, Дж.С.Дэвиса и многих других.

При математической обработке геологической информации часто возникает необходимость формализации (однозначного определения) геологических понятий. Большой вклад в эту проблему внесли Ю.А.Воронин и Ю.А.Косыгин.

Д.А.Родионов, Р.И.Коган, В.А.Голубева и другие выпустили краткий справочник по математическим методам в геологии [15]. Имеются учебники А.Б.Каждана, О.И.Гуськова и А.А.Шиманского [8] и внутривузовские учебные пособия по математическим методам в геологии Г.С.Поротова и Ю.Г.Шестакова.

В применении математических методов в геологии можно условно выделить четыре периода. Первый охватывает отрезок времени с начала XIX в. до 30-х годов XX в. и характеризуется единичными работами отдельных исследователей.

Второй период протекал приблизительно в 1930-1965 гг. В это время началось широкое применение статистических и других математических методов в различных областях геологии.

Качественный скачок произошел после 1965 г. в связи с появлением ЭВМ. Большие возможности ЭВМ в обработке геологической информации способствовали резкому расширению круга математических методов и решаемых с их помощью задач.

С 1990 г. можно говорить о наступлении четвертого периода, вызванного широким распространением персональных компьютеров, которые стали доступны каждому геологу, позволяя ему оперативно обрабатывать геологическую информацию.

В настоящее время математические методы используют в геологии по следующим основным направлениям:

1) накопление, хранение и систематизация (сортировка, получение выборок и пр.) геологической информации с целью более полного и быстрого ее использования;

2) обработка геологической информации преимущественно на базе методов теории вероятностей и математической статистики для описания, сравнения, классификации геологических объектов и прогнозирования их свойств;

3) математическое моделирование геологических объектов и явлений для решения научных и прикладных задач;

4) автоматизация технологических операций, распространенных в геологии и горном деле, таких как построение геологических карт и разрезов, подсчет запасов и ресурсов, проектирование разведочных и эксплуатационных работ и др.

Разделы в учебнике расположены в порядке возрастания сложности, при этом особое внимание автор обращал на четкость и доступность изложения. При подготовке книги был учтен многолетний опыт преподавания дисциплины «Математические методы в геологии» студентам геологической специальности в Санкт-Петербургском государственном горном институте.

Учебник соответствует стандарту дисциплины «Математические методы моделирования в геологии» и использует опыт практических геологических работ. Для лучшего понимания математических операций в каждом разделе приведены подробные примеры вычислений.

Автор выражает благодарность проф. И.В.Булдакову, проф. В.И.Щеглову и доц. И.К.Котовой, которые своими замечаниями способствовали улучшению качества учебника.

Глава 1 ОБЩИЕ СВЕДЕНИЯ

1.1. ГЕОЛОГИЧЕСКИЕ ОБЪЕКТЫ И ИХ СВОЙСТВА

1.1.1. Понятие о геологических объектах

Геология – наука о Земле. Она занимается изучением как планеты в целом, так и ее составных частей различных порядков – от крупных геосфер до мельчайших атомов и молекул. Земля и ее составные части неоднородны, что выражается в плавном или скачкообразном изменении различных характеристик свойств. Изменчивость свойств позволяет проводить внутри Земли границы и тем самым разделять ее на множество геологических тел различных размеров, что обуславливает необходимость системно-структурного подхода к исследованию. Суть метода состоит в том, что в Земле выделяются геологические тела различных порядков и размеров, причем геологические тела n-го порядка являются составными частями геологических тел более низкого $(n-1)$-го порядка и сами, в свою очередь, состоят из множества геологических тел более высокого $(n+1)$-го порядка.

В строении Земли можно выделить геологические тела многих порядков, практически же количество порядков определяется задачами исследований. Например, при изучении литосферы объектами исследований могут быть:

○ литосфера (земная кора и верхняя часть мантии);

○ геотектонические области земной коры (платформы, складчатые области, океанические впадины и пр.);

○ геологические формации (закономерные сочетания горных пород);

○ горные породы (тела горных пород);

○ минералы (минеральные индивиды);

○ молекулы, ионы, атомы.

При изучении полезных ископаемых принято выделять геологические объекты следующих порядков:

◊ рудные провинции, районы и поля (группы месторождений);

◊ месторождения полезных ископаемых (группы рудных тел);

◊ рудные тела (множество природных типов руд);

◊ руды (минеральные агрегаты);

◊ минералы;

◊ компоненты (химические элементы, молекулы, ионы).

При разведке месторождений встречаются такие понятия, как подсчетный блок (при подсчете запасов), рудное сечение (в плоскости рудного тела), рудное пересечение разведочной выработкой (от точки входа до точки выходы из рудного тела), проба руды или минерала, состав проб. Подобные геологические тела различных порядков в настоящей работе называются *геологическими объектами*. Группа геологических тел одного порядка образует совокупность геологических объектов.

При изучении геологических объектов нередко приходится вводить понятия о конкретных и абстрактных объектах. Конкретный объект – это единичный отдельный объект множества, а абстрактный – это обобщенный усредненный типовой объект, отражающий свойства множества объектов. Например, объектом изучения может быть отдельное зерно минерала (конкретный объект) или минерал вообще, представленный множеством зерен (абстрактный объект).

Геологические объекты изучают в статике (без учета изменения во времени) и в динамике (с учетом изменения во времени). В последнем случае анализируют геологические процессы или события (явления), происходящие с геологическими объектами.

1.1.2. Свойства геологических объектов

Любой геологический объект обладает множеством разнообразных свойств. Например, минеральный индивид кварца имеет размеры, габитус, цвет, твердость, плотность и другие свойства. Слой горной породы характеризуется мощностью, элементами залегания, составом, строением и пр.

Свойства геологических объектов можно описать через качественные и количественные характеристики. Качественные характеристики выражаются логическими высказываниями. Например, для галита характерна совершенная спайность, пирит имеет желтый цвет, руда может иметь вкрапленную текстуру. Количественная мера свойства выражается числом: плотность алмаза 3,5 г/см^3, содержание меди в руде 1,58 %, азимут простирания рудного тела 56°. В геологической практике широко используют как качественные, так и количественные характеристики.

Для математической обработки характеристики качества переводят в числовую форму с помощью номинальной и порядковой шкал. Для количественных характеристик свойств используют интервальную и относительную шкалы.

Номинальная шкала имеет два значения: «да», которое кодируется единицей, и «нет», которое кодируется нулем. Если у объекта устанавливается изучаемое свойство, то присваивается значение «да», в противном случае «нет». Предварительно необходимо составить классификацию свойств и установить критерии различия между ними (формализовать свойства). Например, руда месторождения может иметь одну из следующих текстур: однородную (массивную), вкрапленную, полосчатую, пятнистую и брекчиевидную. Если известно, что какая-то проба руды имеет пятнистую текстуру, то ее нужно отнести в четвертый из названных классов, при этом в четвертом классе ставится единица, а в остальных классах – нули. Если имеется несколько проб руды, то результаты измерений текстур можно привести в табличной форме (см. таблицу).

Порядковая шкала применяется, когда значения свойства могут быть расположены в порядке возрастания или убывания. Например, если в рудах встречаются тонко-, мелко-, средне- и круп-

нозернистая структуры, то можно выделить классы по увеличению (уменьшению) зернистости и назначить им номера (баллы) с первого по четвертый. Тогда для кодирования структуры руды достаточно указать номер класса.

Характеристика текстуры руды

Номер пробы	Однородная	Вкрапленная	Полосчатая	Пятнистая	Брекчиевидная
1	0	0	0	1	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	0	0	1
5	0	1	0	0	0

Порядковые шкалы часто используются в геологии. Они удобны для записи последовательности напластования горных пород, очередности геологических событий и пр. Широко известна порядковая шкала твердости минералов, в которой самый мягкий минерал графит относится к первому классу, а самый твердый – алмаз – к десятому классу.

Интервальная шкала используется для количественных характеристик свойств с произвольной нулевой точкой отсчета. Примером являются топографические координаты пунктов измерений, которые исчисляются от условного репера, принимаемого за начало координат.

Относительная шкала имеет наибольшее распространение при измерении количественных характеристик. Особенность шкалы состоит в том, что ее начало имеет физический смысл. Так, при измерении содержаний компонентов в руде естественно взять за начало отсчета нулевое содержание.

Количественные характеристики свойств можно перевести в качественные. Например, можно условиться, что руды, содержащие 0,5-1 % меди, являются бедными; 1-2 % – рядовыми; 2-5 % – богатыми. Это позволяет закодировать состав руды с помощью номинальной или порядковой шкалы.

Таким образом, любые свойства геологических объектов можно записать в форме, удобной для математической обработки.

Если имеется n геологических объектов (или пунктов измерений) и у каждого объекта измерено k свойств, то результаты могут быть сведены в таблицу размером $n \times k$ клеток. Такая таблица называется матрицей и ее принято записывать в следующем виде:

$$\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{vmatrix}. \qquad (1.1)$$

Произвольный член матрицы можно обозначить x_{ij}, где i – номер объекта (или пункта измерений), j – номер свойства. При необходимости матрица (1.1) может быть дополнена координатами пунктов измерений x, y, z, а при изучении геологических процессов и временем t.

Матрица (1.1) является частным случаем таблиц более широкого типа, так называемых баз данных (БД). В базах данных строка матрицы называется записью, а столбец – полем записи. Базы данных можно записывать и хранить в компьютере с помощью программ Excel, Access, dBase, MS DOS, Word и других прикладных пакетов.

1.1.3. Выборочные методы изучения геологических объектов

Для того чтобы изучить какой-либо геологический объект, необходимо измерить характеристики его свойств. Так как геологические объекты имеют различные размеры – от тысячных и миллионных долей миллиметра до сотен и тысяч километров, измерения сопряжены с определенными трудностями. С одной стороны, количество объектов бывает столь велико (например, количество зерен минералов в руде), что каждый из них изучить невозможно и нецелесообразно. С другой стороны, крупные объекты (например, месторождения полезных ископаемых или интрузивные массивы) доступны для изучения в отдельных пунктах и об их свойствах приходится судить на основе данных по выбранной сети наблюдений.

Отмеченные особенности геологических объектов обуславливают необходимость применения выборочного метода измерения. В выборочном методе используются понятия – генеральный коллектив и выборка. *Генеральный коллектив* включает все множество однопорядковых геологических объектов, а *выборка* – часть объектов, выбираемых из генерального коллектива по определенным правилам. Исследованию подвергается выборка, а выводы распространяются на генеральный коллектив.

Различаются два варианта выборочного метода. Первый вариант применяется для изучения множества однопорядковых объектов, когда для исследования выбирается часть объектов. В частном случае в выборку может входить вся генеральная совокупность. Например, для характеристики какого-либо типа полезного ископаемого можно изучить несколько месторождений, а иногда и все месторождения данного типа.

Второй вариант используется для изучения крупных объектов (например, рудного тела). Геологический объект мысленно делится на элементарные объемы, совокупность которых рассматривается как генеральная совокупность. Изучению подвергается часть элементарных объемов, т.е. выборка, а выводы распространяются на объект в целом или на какую-то его часть.

На результаты выводов влияют два типа погрешностей. Первый тип появляется в процессе измерений и относится к *техническим погрешностям* или *погрешностям измерений*. Такие погрешности подразделяются на случайные и систематические. *Случайные погрешности* присутствуют во всех измерениях, они неустранимы и их стараются снизить до разумных пределов путем соответствующей организации работ. *Систематические погрешности* возникают в результате неправильной методики или технологии измерений, их значения направлены в одну сторону (завышения или занижения результатов измерений). Если такие погрешности появляются, необходимо устранять их либо изменением методики и технологии измерений, либо введением поправок.

Второй тип погрешностей возникает при распространении выборочных данных на генеральную совокупность. Их называют *погрешностями аналогии* или *погрешностями распространения*.

Значения их зависят от способа или методики распространения данных выборки на генеральную совокупность, поэтому иногда такие ошибки называют *методическими погрешностями*.

1.2. ПОНЯТИЕ О МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

1.2.1. Принцип и операции математического моделирования

Любые методы обработки экспериментальных данных содержат в своей основе явную или неявную модель изучаемого объекта или происходящего с ним явления (события).

Математическая модель – это совокупность представлений, предположений, гипотез и аксиом, отражающих существо изучаемого геологического объекта или явления.

Модель выражается в математической форме и позволяет описывать, анализировать и прогнозировать свойства геологических объектов или последствия явлений.

В основе математического моделирования лежит принцип *системного подхода*. Для исследования выделяются объект или группа однопорядковых объектов, которые рассматриваются как отдельная система, имеющая какие-то физические или условные границы и внутренние связи между частями или свойствами. Геологические объекты, расположенные за пределами системы, являются по отношению к ней окружающей средой.

Когда система определена, осуществляется ее исследование путем математического моделирования. Конечной целью моделирования может быть описание и классификация объектов, понимание геологической природы объектов и явлений, предсказание (прогнозирование) поведения или свойств системы, а в некоторых случаях и управление системой на основе контроля ее состояния. Например,

при разведке и эксплуатации месторождения необходимо понять его строение и происхождение, прогнозировать количество и качество минерального сырья, управлять процессом эксплуатации с целью рационального использования недр и решать много других практических задач.

Математическое моделирование геологических объектов можно разделить на несколько последовательных *операций*:

1. Определение системы, т.е. задание границ, перечня геологических объектов и их свойств, а иногда и характера взаимосвязей между свойствами.

2. Измерение характеристик свойств геологических объектов, входящих в систему. Иначе говоря, получение исходных данных для математической обработки. В некоторых задачах измерения могут отсутствовать, а изучению подвергаются предполагаемые значения, заданные исследователем.

3. Создание геологического представления (геологической модели) о существе изучаемой системы и формулировка геологической задачи, стоящей перед математическим моделированием. Часто можно выдвинуть несколько гипотез о системе, иногда взаимоисключающих друг друга. Тогда на основе последующего математического моделирования можно сделать заключение о том, какая из гипотез более соответствует действительности.

4. Выражение геологических представлений в математической форме, т.е. в виде формул, правил, уравнений и пр. Это и есть математическая постановка задачи. В процессе постановки часто приходится возвращаться к п.2 и 3 для уточнения недостающих сведений.

5. Исследование математической модели, которое чаще всего сводится к решению составленных в п.4 формул и уравнений и вычислению прогнозных значений свойств или параметров явлений, т.е. к получению ответа на геологическую задачу. Если исходные данные колеблются в некоторых пределах, то можно исследовать зависимость прогнозных значений от исходных данных. В некоторых случаях оценивается погрешность прогнозирования.

6. Проверка соответствия полученных результатов фактическим данным. В результате проверки можно определить, насколько правильно математическая модель описывает систему, насколько верны геологические представления, положенные в ее основу. Чаще всего оценивается степень совпадения или сходства фактических данных с теоретическими, вычисленными в ходе решения математической модели. Если имеется несколько геологических и соответствующих им математических моделей, то проверка может дать ответ на вопрос, какая из моделей лучше соответствует действительности. Следует отметить, что проверка не всегда возможна, особенно в тех случаях, когда получение фактических данных затруднено или невозможно.

В зависимости от постановки задачи в результате математического моделирования могут быть получены различные ответы. Во-первых, можно определить прогнозные значения тех свойств, которые трудно измерить или которые не поддаются непосредственному измерению. Во-вторых, можно оценить степень соответствия математической модели фактическим данным. В-третьих, можно установить, какая из математических и, соответственно, геологических моделей лучше соответствует действительности и тем самым выбрать для дальнейших исследований наилучшую модель.

Геологические системы являются весьма сложными структурами, находящимися под влиянием большого числа трудно учитываемых факторов, поэтому математическое моделирование не может дать их исчерпывающую характеристику. Следовательно, любая математическая модель является приближенным отражением реальных природных систем и для каждой природной системы можно построить несколько математических моделей различной степени сложности. Обычно по мере усложнения математической модели повышается достоверность прогнозирования и надежность выводов. Но существует оптимальная степень сложности математических моделей, такая, при которой дальнейшее усложнение не будет повышать достоверность прогнозирования и может даже ухудшить работоспособность модели. Нередко степень сложности математических моделей ограничивается техническими возможностями вычислительной техники.

1.2.2. Примеры математических моделей

Последовательность операций математического моделирования можно показать на нескольких примерах.

⏩ **Пример 1.1.** Рудное тело имеет длину по простиранию a = 500 м, по падению b = 200 м, видимую среднюю мощность на дневной поверхности m = 8 м, угол падения $\alpha = 65°$. Необходимо оценить объем рудного тела.

Из условия задачи понятно, что определена система (объект исследования) – рудное тело, измерены его параметры: размеры по простиранию и падению, мощность и угол падения, т.е. выполнены две операции моделирования.

Наиболее ответственна третья операция – создание геологической модели рудного тела. Возможно несколько альтернативных вариантов предположений о форме рудного тела:

а) рудное тело сохраняет протяженность и мощность на глубине, т.е. имеет форму параллелепипеда;

б) рудное тело выклинивается на глубине в линию, т.е. имеет форму клина;

в) рудное тело выклинивается на глубине в точку, т.е. имеет форму пирамиды.

Возможны и другие предположения о форме рудного тела на глубине. При существующем объеме геологической информации сделанные предположения о форме рудного тела равновероятны.

Четвертая операция – это выражение в виде математических формул геологических предположений о форме рудного тела. Предварительно необходимо уточнить, как ориентирована видимая мощность. Положим, что она горизонтальная, тогда истинная мощность рудного тела $m_{ист} = m\sin\alpha$. Запишем три формулы объема:

а) объем параллелепипеда $V = abm\sin\alpha$;

б) объем клина $V = 1/2\ abm\sin\alpha$;

в) объем пирамиды $V = 1/3\ abm\sin\alpha$.

Из сравнения формул видно, что объем рудного тела существенно зависит от предположения о его форме, различаясь по вариантам в 3 раза.

Пятая операция – вычисление (прогнозирование) объема рудного тела по приведенным формулам:

а) объем параллелепипеда $V = 725{,}0$ тыс.м^3;

б) объем клина $V = 362{,}5$ тыс.м^3;

в) объем пирамиды $V = 241{,}7$ тыс.м^3.

Шестая операция – проверка совпадения вычисленного и фактического объемов. Очевидно, что фактический объем рудного тела установить трудно. Это требует проведения дополнительных работ, например детального изучения рудного тела на глубине с помощью разведочных выработок или добычи руды. Предположим, что рудное тело добыто и его объем оказался 350 тыс.м^3, тогда можно заключить, что ближе всего к истине второй вариант (выклинивание рудного тела в линию). Погрешность прогнозирования объема рудного тела по второму варианту в абсолютном выражении $\delta = 362{,}5 \cdot 350 = 12{,}5$ тыс.м^3, в относительном $12{,}5/350 = 0{,}036 = 3{,}6$ %. ◀◀

▶▶ **Пример 1.2.** В рудах полиметаллического месторождения пробы проанализированы на цинк и кадмий. При построении графика обнаружено, что с возрастанием содержаний цинка растет содержание кадмия (рис.1.1). Требуется дать геологическое объяснение зависимости и построить математическую модель.

Исходными данными для построения модели являются содержания цинка x и кадмия y в пробах руды. Пока не будем рассматривать конкретные числовые данные и их обработку, а ограничимся логическими рассуждениями.

Зависимость между содержаниями цинка и кадмия вызвана тем, что оба компонента входят в состав одного минерала – сфалерита. При увеличении количества сфалерита растет содержание цинка и кадмия в руде. Это предполагаемая геологическая модель зависимости.

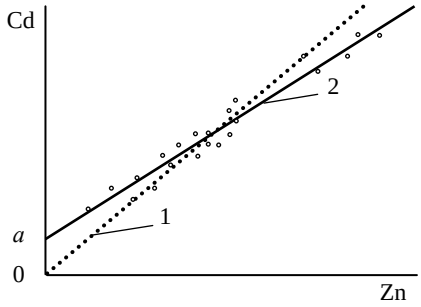

Рис.1.1. Пропорциональная (1) и линейная (2) зависимости между содержаниями цинка и кадмия

Математическая модель сводится к составлению уравнения зависимости между содер-

жаниями цинка и кадмия. Пренебрегая неизбежными колебаниями состава сфалерита, можно принять, что содержание в нем компонентов постоянное (это одно из допущений модели). Тогда между содержаниями цинка x и кадмия y должна существовать пропорциональная зависимость $y = bx$ (b – коэффициент пропорциональности). Прямая $y = bx$ должна проходить через две точки графика – начало координат и центр тяжести точек, соответствующий средним содержаниям цинка и кадмия в пробах.

Графический анализ зависимости (линия 1 на рис.1.1) показывает, что данная прямая не соответствует расположению точек. Это говорит о том, что геологическая и математические модели не соответствуют действительности. Прямая линия должна проходить вдоль удлинения облака точек, но тогда это будет не пропорциональная, а линейная зависимость, выражаемая уравнением $y = a + bx$ (a и b – коэффициенты). Линия уравнения не проходит через начало координат, а отсекает на оси ординат отрезок a, т.е. при нулевом содержании цинка и, следовательно, сфалерита содержание кадмия равно не нулю, а значению a. Геологическое объяснение данного факта состоит в том, что некоторая часть кадмия имеется в других минералах и нужно проверить их состав.

В данном случае математическое моделирование позволяет из двух моделей (пропорциональной и линейной) выбрать одну, более достоверную, и помогает более правильно объяснить наблюдаемую зависимость. ◀◀

▶▶ **Пример 1.3.** Известна плотность руды и содержание в ней полезного компонента. Необходимо построить математическую модель зависимости этих величин, что актуально для руд многих черных и цветных металлов.

Для упрощения модели с целью выделения ее главных особенностей примем, что руда состоит из двух минералов (рудного и нерудного), их массы m_1 и m_2, объемы V_1 и V_2, плотности ρ_1 и ρ_2, содержания в них компонента C_1 и C_2, причем положим $\rho_1 > \rho_2$ и $C_1 > C_2$. В качестве аргумента x будет служить содержание компонента в руде:

$$x = \frac{m_1 C_1 + m_2 C_2}{m_1 + m_2}. \qquad (1.2)$$

В качестве функции y будет плотность руды:

$$y = \frac{m_1 + m_2}{V_1 + V_2}. \qquad (1.3)$$

Требуется найти математическое выражение зависимости плотности y от содержания x.

Очевидно, что $V_1 = m_1/\rho_1$ и $V_2 = m_2/\rho_2$. Подставляя их в формулу (1.3), получим

$$y = \frac{(m_1 + m_2)\rho_1\rho_2}{m_1\rho_2 + m_2\rho_1}. \qquad (1.4)$$

Из формулы (1.3) найдем величину m_1:

$$m_1 = m_2 \frac{x - C_2}{C_1 - x}.$$

Подставим ее в выражение (1.4). После преобразований получим

$$y = \frac{\rho_1\rho_2(C_1 - C_2)}{C_1\rho_1 - C_2\rho_2} : \left(1 - x\frac{\rho_1 - \rho_2}{C_1\rho_1 - C_2\rho_2}\right).$$

Обозначим

$$\frac{\rho_1\rho_2(C_1 - C_2)}{C_1\rho_1 - C_2\rho_2} = a\,, \quad \frac{\rho_1 - \rho_2}{C_1\rho_1 - C_2\rho_2} = b\,.$$

В результате имеем гиперболическую зависимость плотности руды y от содержания в ней компонента x (рис.1.2):

$$y = a/(1 - bx)\,, \qquad (1.5)$$

где a и b – постоянные коэффициенты.

Формула (1.5) представляет собою математическую модель зависимости.

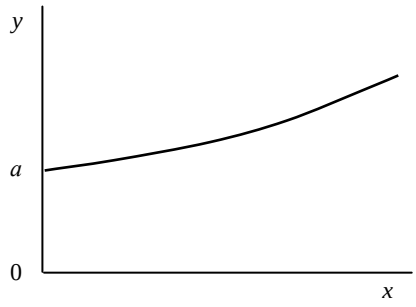

Рис.1.2. Гиперболическая зависимость плотности руды от ее состава

Подобная зависимость часто используется на практике. Ее характер принципиально не изменится, если руда состоит из нескольких минералов, но появится разброс исходных данных около гиперболической зависимости, что вызвано колебаниями количественных соотношений минералов в руде и их состава. ◀◀

1.2.3. Основные виды математических моделей, применяемых в геологии

Моделирование вообще и математическое моделирование в частности является эффективным средством изучения геологических систем, объектов и событий. Каждая модель служит некоторым их отражением и характеризует наиболее существенные особенности.

Модели можно разделить на материальные, аналоговые и символьные (рис.1.3).

Материальные модели представляют собой выполненные в определенном масштабе макеты геологических объектов. Например, существуют материальные модели кристаллических решеток минералов, модели идеальных кристаллов с различными наборами граней, морфологические модели рудных тел и др.

Аналоговые модели основаны на замене природных геологических процессов, явлений другими, воспроизводимыми в лаборатории, процессами, которые описываются одинаковыми математическими правилами и уравнениями. Например, движение подземных вод, процессы переноса в них вещества, явление диффузии и многие другие можно моделировать движением электрического тока в аналоговых устройствах.

Символьные модели, которые делятся на графические и математические, имеют особое значение при математическом моделировании. К графическим моделям относятся разнообразные геологические карты, разрезы, проекции, схемы и графики. Они позволяют наглядно изобразить геологические объекты и характеристики их свойств, а также дать интерпретацию многих операций математического моделирования.

Рис.1.3. Схема классификации моделей геологических объектов

Математические модели можно разделить на три группы. В первой группе анализируются характеристики в пределах однородных совокупностей свойств объектов вне связи их с пространственным размещением, это группа *статистических моделей*. Они бывают одномерные, двухмерные и многомерные.

Во второй группе учитываются пространственные координаты пунктов наблюдений, что позволяет изучать *пространственные геологические поля*. Модели делятся на детерминированные и вероятностные. В *детерминированных моделях* предполагается, что состояние системы однозначно определяется исходными или начальными данными и полностью предсказуемо в пространстве. *Вероятностные модели* характеризуются тем, что состояние системы и прогнозные значения свойств геологических объектов неоднозначно зависят от начальных или исходных данных и могут быть предсказаны с какой-то вероятностью в определенном диапазоне значений.

Третья группа охватывает случайные процессы, в которых учитывается фактор времени.

В настоящей книге будут рассмотрены модели первой и второй групп. Третья группа выходит за рамки дисциплины.

Глава 2 ОДНОМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ И ЕЕ ПРИМЕНЕНИЕ В ГЕОЛОГИИ

2.1. ОДНОМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ

2.1.1. Свойства геологических объектов как независимые случайные величины

Как указывалось выше, одномерная статистическая модель применяется для изучения одного свойства. Пусть имеется система, состоящая из множества однородных геологических объектов. Выборочным методом возьмем из множества n объектов и у каждого из них измерим характеристику свойства x. Результаты измерений обозначим x_1, x_2, ..., x_n и составим из них матрицу (1.1), в которой число строк равно n, а число столбцов $k = 1$.

В основе одномерной статистической модели лежат три гипотезы: а) измеренные значения x_1, x_2, ..., x_n носят случайный характер; б) они не зависят друг от друга; в) значения образуют однородную совокупность. Измеренные значения принято называть *реализациями* случайной величины x.

Гипотеза о случайном характере свойств обусловлена тем, что природные геологические системы и объекты являются весьма сложными, на каждое измеренное значение влияет множество разнонаправленных факторов. Кроме того, каждое измерение сопровождается случайной погрешностью. Данная гипотеза позволяет применять для математической обработки значений x_1, x_2, …, x_n аппарат (теоремы, формулы, уравнения, законы) теории вероятностей.

Вторая гипотеза о независимости измеренных значений менее очевидна. Она предполагает, что на результат каждого отдельного измерения не влияют результаты предыдущих или соседних измерений. Из этой гипотезы вытекает важное следствие, что для математической обработки не существенно пространственное размещение пунктов наблюдений, т.е. результаты измерений можно располагать в любом порядке, на выводы это не влияет. Эта гипотеза не всегда соответствует действительности: соседние измерения нередко зависят друг от друга, что можно проверить с помощью специального математического аппарата.

Статистическая обработка результатов измерений имеет смысл лишь только для однородных совокупностей, что лежит в основе третьей гипотезы. Если совокупность неоднородная, то ее необходимо разделить на однородные совокупности и каждую из них исследовать отдельно.

2.1.2. Статистические характеристики случайной величины

В основе большинства вычислений лежит расчет статистических характеристик случайной величины. К наиболее распространенным статистическим характеристикам одномерной случайной величины относятся размах, медиана, мода, среднее значение, дисперсия, среднеквадратичное отклонение, коэффициент вариации, асимметрия и эксцесс.

Пусть имеется n измерений свойства x. Необходимо найти статистические характеристики этого множества измерений.

Размах – это разность между максимальным $x_{\max}$ и минимальным $x_{\min}$ значениями свойства: $p = x_{\max} - x_{\min}$.

Медиана – средний член упорядоченного ряда значений. Для нахождения медианы нужно расположить все значения в порядке возрастания или убывания и найти средний по порядку член ряда. В случае n – четного числа в середине ряда окажутся два значения, тогда медиана будет равна их полусумме.

Мода – наиболее часто встречающееся значение случайной величины. Методику ее нахождения мы рассмотрим позднее.

Среднее значение – это среднеарифметическое из всех измеренных значений:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \qquad (2.1)$$

Существуют другие виды средних (среднее взвешенное, среднее геометрическое, среднее гармоническое и др.), которые вычисляются в особых случаях и здесь не рассматриваются.

Медиана, мода и среднее значение являются *характеристиками положения* – около них группируются измеренные значения случайной величины.

Дисперсия – это число, равное среднему квадрату отклонений значений случайной величины от ее среднего значения:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (2.2)$$

Среднеквадратичное отклонение – это число, равное квадратному корню из дисперсии:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (2.3)$$

Среднеквадратичное отклонение имеет размерность, совпадающую с размерностью случайной величины и среднего значения. Например, если значения случайной величины измерены в метрах, то и среднеквадратичное отклонение также будет выражаться в метрах.

Коэффициент вариации – это отношение среднеквадратичного отклонения к среднему значению:

$$V = \frac{\sigma}{\bar{x}}. \qquad (2.4)$$

Коэффициент вариации выражается в долях единицы или (после умножения на 100) в процентах. Вычисление коэффициента вариации имеет смысл для положительных случайных величин.

Дисперсия, среднеквадратичное отклонение и коэффициент вариации, а также размах являются *мерами рассеяния* значений случайной величины около среднего значения. Чем они больше, тем сильнее рассеяние.

Асимметрия – степень асимметричности распределения значений случайной величины относительно среднего значения,

$$A = \frac{1}{n\sigma^3} \sum_{i=1}^{n} (x_i - \bar{x})^3 . \qquad (2.5)$$

Эксцесс – степень остро- или плосковершинности распределения значений случайной величины относительно нормального закона распределения,

$$E = \frac{1}{n\sigma^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 - 3 . \qquad (2.6)$$

Асимметрия и эксцесс являются безразмерными величинами. Они отражают особенности группировки значений случайной величины около среднего значения.

Рассмотренные статистические характеристики относятся к множеству значений x_1, x_2, ..., x_n. Если множество представляет собой выборку из генеральной совокупности, то возникает задача оценки ее статистических характеристик по выборочным данным. Наибольшее значение имеют оценка математического ожидания и дисперсии генеральной совокупности.

Математическое ожидание случайной величины $M(x)$ – это ее среднее значение в генеральной совокупности. Оно, за редким исключением, бывает неизвестно, и приходится пользоваться его приближенной оценкой (точечной оценкой) – выборочным средним значением $\bar{x}$, определяемым по формуле (2.1). При увеличении числа наблюдений выборочное среднее стремится к пределу – к математическому ожиданию.

Дисперсия генеральной совокупности $D(x)$ – это число, равное среднему квадрату отклонений случайной величины от ее математического ожидания. Если математическое ожидание известно, то дисперсию находят по формуле

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - M(x)\right)^2. \qquad (2.7)$$

Если математическое ожидание неизвестно, то определяют *оценку дисперсии* по формуле

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (2.8)$$

Единица в знаменателе формулы (2.8) отражает одну использованную степень свободы: вместо математического ожидания в формулу подставлено выборочное среднее значение. При увеличении числа наблюдений n оценка дисперсии S^2 стремится к дисперсии генеральной совокупности $D(x)$.

Формулы (2.2) и (2.8) похожи друг на друга, но применяют их в разных случаях. Первая используется для характеристики выборки, а вторая – для характеристики генеральной совокупности.

В ряде задач возникает необходимость рассчитывать статистические характеристики суммы или разности случайных величин, а также произведения случайной величины на постоянный множитель.

Пусть имеется случайная величина x. Если умножить ее значения на постоянный множитель, то получим новую случайную величину $y = ax$. Статистические характеристики новой случайной величины преобразуются следующим образом:

среднее значение $y_{ср} = ax_{ср}$;

дисперсия $\sigma_y^2 = a^2\sigma_x^2$;

среднеквадратичное отклонение $\sigma_y = a\sigma_x$.

При этом коэффициент вариации, асимметрия и эксцесс не изменят своих значений. Очевидно, что деление значений случайной величины на постоянную величину a равносильно умножению на обратную величину $1/a$ и приведенные формулы сохраняют свою силу.

Если к случайной величине x прибавить постоянное слагаемое a, т.е. $y = x + a$, то изменится и среднее значение: $\bar{y} = \bar{x} + a$. Однако значения дисперсии, среднеквадратичного отклонения, асимметрии и эксцесса сохранятся. Вычитание постоянного слагаемого равносильно изменению знака слагаемого a на $-a$.

Наибольший интерес представляет ситуация, когда производится сложение (вычитание) двух и более случайных величин. Пусть имеются две независимые случайные величины x и y, их сумма (разность) образует третью случайную величину $z = x \pm y$. Статистические характеристики меняются следующим образом:

среднее значение $\bar{z} = \bar{x} \pm \bar{y}$;

дисперсия $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$.

Если имеется n независимых случайных величин x_1, x_2, ..., x_n и находится их сумма $z = x_1 + x_2 + ... + x_n$, то имеем соотношения:

средние значения $\bar{z} = \bar{x}_1 + \bar{x}_1 + \cdots + \bar{x}_n$;

дисперсии $\sigma_z^2 = \sigma_1^2 + \sigma_1^2 + \cdots + \sigma_x^2$. (2.9)

Особенно большое значение имеет последнее равенство, известное как теорема сложения дисперсий: дисперсия суммы независимых случайных величин равна сумме их дисперсий. Используя эту теорему, можно доказать, что дисперсия среднего значения $\bar{x}$ из n значений x_i в n раз меньше дисперсий исходных значений x_i:

$$\sigma_{\bar{x}}^2 = \sigma_x^2 / n. \quad (2.10)$$

Эта формула неоднократно будет использована в дальнейшем.

2.1.3. Моменты случайной величины, их связь со статистическими характеристиками

Вычисление статистических характеристик можно производить непосредственно по формулам (2.1)-(2.8), но на практике характеристики обычно находят с помощью моментов.

Моментом случайной величины k-го порядка относительно постоянного параметра a называется выражение

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k. \quad (2.11)$$

Порядок k может быть выражен любым целым числом, но интерес представляют первые четыре момента (порядка).

В зависимости от выбора параметра a различают начальные и центральные моменты. В первом случае a выбирается произвольно, что имеет смысл для ускорения вычислений. Часто полагают $a = 0$, и формула начальных моментов приобретает вид

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k. \qquad (2.12)$$

Во втором случае принимают $a = \bar{x}$ и получают центральные моменты

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k. \qquad (2.13)$$

От начальных моментов можно перейти к центральным:

$$\begin{aligned} \mu_1 &= 0; \\ \mu_2 &= m_2 - m_1^2; \\ \mu_3 &= m_3 - 3m_2m_1 + 2m_1^3; \\ \mu_4 &= m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4. \end{aligned} \qquad (2.14)$$

Зная моменты случайной величины, можно найти ее статистические характеристики по формулам

$$\bar{x} = m_1; \quad \sigma^2 = \mu_2; \quad \sigma = \sqrt{\mu_2};$$

$$V = \sigma / \bar{x}; \quad S^2 = \frac{1}{n-1}\mu_2; \qquad (2.15)$$

$$A = \mu_3 / \sigma^3; \quad E = \mu_4 / \sigma^4 - 3.$$

⏩ **Пример 2.1.** В 11 пробах руды определено содержание никеля (табл.2.1). Требуется рассчитать статистические характеристики.

Расчет статистических характеристик может быть выполнен двумя методами – через начальные (табл.2.2) или центральные

Таблица 2.1

Содержание никеля в руде *x*

Номер пробы	*x*, %	Номер пробы	*x*, %	Номер пробы	*x*, %
1	0,38	5	0,07	9	0,24
2	0,51	6	0,39	10	0,30
3	0,47	7	0,17	11	0,25
4	0,13	8	0,28		

(табл.2.3) моменты. Последняя строка табл.2.2 содержит начальные моменты $m_1 = 0{,}29$; $m_2 = 0{,}1015$; $m_3 = 0{,}039584$; $m_4 = 0{,}0166409$. По формулам (2.14) найдем центральные моменты:

$$\mu_2 = 0{,}1015 - 0{,}29^2 = 0{,}0174;$$

$$\mu_3 = 0{,}039584 - 3\cdot 0{,}10152\cdot 0{,}29 + 2\cdot 0{,}29^3 = 0{,}000139;$$

$$\mu_4 = 0{,}0166409 - 4\cdot 0{,}039584\cdot 0{,}29 + 6\cdot 0{,}1015\cdot 0{,}29^2 - 3\cdot 0{,}29^4 = 0{,}000617.$$

Таблица 2.2

Расчет начальных моментов случайной величины

№ п/п	Исходные данные *x*, %	Степень исходных данных		
		x^2	x^3	x^4
1	0,07	0,0049	0,000343	0,00002401
2	0,13	0,0169	0,002197	0,00028561
3	0,17	0,0289	0,004913	0,00083521
4	0,24	0,0576	0,013824	0,00331776
5	0,25	0,0625	0,015625	0,00390625
6	0,28	0,0784	0,021952	0,00614656
7	0,30	0,0900	0,027000	0,00810000
8	0,38	0,1444	0,054872	0,02085136
9	0,39	0,1521	0,059319	0,02313441
10	0,47	0,2209	0,103823	0,04879681
11	0,51	0,2601	0,132651	0,06765201
Сумма	3,19	1,1167	0,436519	0,18304999
Среднее	0,29	0,1015	0,039584	0,0166409
Моменты	m_1	m_2	m_3	m_4

Таблица 2.3

Расчет центральных моментов случайной величины

№ п/п	Исходные данные x, %	Степень отклонений исходных данных			
		$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
1	0,07	-0,22	0,0484	-0,010648	0,00234256
2	0,13	-0,16	0,0256	-0,004096	0,00065536
3	0,17	-0,12	0,0144	-0,001728	0,00020736
4	0,24	-0,05	0,0025	-0,000125	0,00000625
5	0,25	-0,04	0,0016	-0,000064	0,00000256
6	0,28	-0,01	0,0001	-0,000001	0,00000001
7	0,30	0,01	0,0001	0,000001	0,00000001
8	0,38	0,09	0,0081	0,000729	0,00006561
9	0,39	0,10	0,0100	0,001000	0,00010000
10	0,47	0,18	0,0324	0,005832	0,00104976
11	0,51	0,22	0,0484	0,010648	0,00234256
Сумма	3,19	0,00	0,1916	0,001548	0,00677204
Среднее	0,29	0,00	0,0174	0,000141	0,000616
Моменты	m_1	μ_1	μ_2	μ_3	μ_4

Эти же моменты другим способом вычислены в табл.2.3. Небольшие различия в значениях моментов, полученных разными способами, связаны с округлением промежуточных данных.

Зная центральные моменты, по формулам (2.15) найдем статистические характеристики:

$$\bar{x} = 0{,}29;\ \sigma^2 = 0{,}0174;\ \sigma = 0{,}132;$$

$$V = 0{,}132/0{,}29 = 0{,}455 = 45{,}5\ \%;$$

$$S^2 = 0{,}0174 \cdot 11/10 = 0{,}0191;$$

$$A = 0{,}000139/0{,}132^3 = 0{,}060;$$

$$E = 0{,}000617/0{,}132^4 - 3 = -0{,}968. \blacktriangleleft\blacktriangleleft$$

В примере 2.1 расчеты выполнены вручную, автоматизировать этом процесс позволяет прикладной пакет программ Excel.

2.1.4. Группировка исходных данных. Построение гистограммы

При большом числе исходных данных ($n > 50$) расчет статистических характеристик с помощью таблиц становится громоздким, поэтому применяется компактный метод расчета с предварительной группировкой данных. Для этого весь диапазон исходных значений от x_{min} до x_{max} разбивается на равные интервалы (классы), границы которых удобно брать округленными, хотя это не имеет принципиального значения. С округленными границами удобнее работать.

Число классов зависит от числа исходных данных. Обычно принимается от 6 до 20 классов, но можно использовать и больше. Для определения числа классов рекомендуется эмпирическая формула $N_{кл} = 16[0{,}4\ln(n) - 1]$. Далее подсчитывают число исходных значений, попавших в каждый класс, и результаты сводят в табл.2.4.

Таблица 2.4

Частота и частость содержания железа в руде

Класс содержаний, %	Число проб (частота)	Частость	
		в долях единицы	в процентах
30-32	2	0,014	1,4
32-34	6	0,041	4,1
34-36	9	0,061	6,1
36-38	14	0,095	9,5
38-40	20	0,136	13,6
40-42	25	0,170	17,0
42-44	21	0,143	14,3
44-46	17	0,116	11,6
46-48	13	0,088	8,8
48-50	10	0,068	6,8
50-52	5	0,034	3,4
52-54	3	0,020	2,0
54-56	2	0,014	1,4
Сумма	147	1,000	100,0

Некоторая трудность возникает в том случае, когда отдельные значения попадают на границу классов. Их можно относить в старший класс либо пытаться распределить примерно поровну между смежными классами.

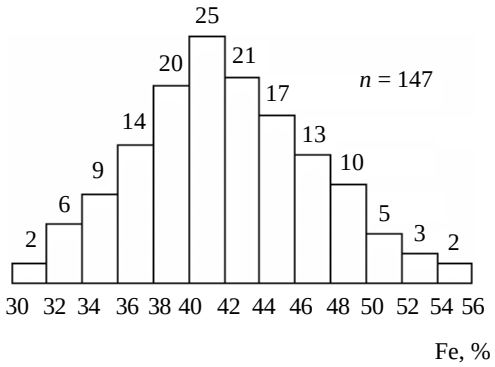

Рис.2.1. Гистограмма частот содержания железа в руде

Число значений в классе называется *частотой*. Если выразить частоту в относительных долях к общему числу значений, то получим *частость*. Ее можно выразить в процентах (табл.2.4).

Данные табл.2.4 позволяют построить гистограмму значений случайной величины (рис.2.1). По оси абсцисс откладывают классы, а по оси ординат – частоту или частость в виде ступенек. Для удобства обозрения над ступеньками выписана частота, а рядом с гистограммой указано суммарное значение *n*.

Гистограмма дает наглядное представление о поведении случайной величины. На ней видны размах и частота значений. Полезную информацию несет и форма гистограммы; она может быть симметричной и асимметричной, с одним, двумя и более максимумами частот.

Наличие нескольких максимумов свидетельствует о неоднородности изучаемой совокупности и позволяет ставить вопрос о выделении однородных совокупностей. В некоторых случаях отдельные частоты резко преобладают, это чаще всего связано с дефектами измерений. Например, при химическом анализе часто встречаются округленные значения и гораздо реже – промежуточные между ними. Чтобы устранить влияние подобных погрешностей, следует увеличить размер классов и построить гистограмму снова.

2.1.5. Расчет статистических характеристик по сгруппированным данным

Моменты и статистические характеристики можно рассчитать по сгруппированным данным, что сокращает объем вычислений при большом числе исходных данных.

⏩ **Пример 2.2.** Известны частота значений случайной величины в классах n_i, границы начального и конечного классов и размер (шаг) классов h. Требуется рассчитать статистические характеристики.

Расчет начинается с присвоения каждому классу условного порядкового номера x. Одному из классов присваивают нулевой номер, остальным – отрицательные и положительные номера (табл.2.5). Все классы располагают в порядке возрастания без пропусков. Нулевой класс выбирают произвольно, по возможности ближе к среднему значению, что уменьшает объем вычислений. Чаще всего за нулевой принимают класс с максимальной частотой. В табл.2.5 нулевой класс имеет пределы 40-42, его середина $x_0 = 41$, а шаг $h = 2$.

Вначале расчеты выполним в табл.2.5. С помощью условных номеров вычислим начальные моменты, но в условном масштабе, так как размер классов $h = 2$. Для этого найдем произведения $n_i x_i$, $n_i x_i^2$, $n_i x_i^3$, $n_i x_i^4$, суммируем их и определим среднее в каждой графе путем деления на общее число данных $n = 147$. Последняя строка таблицы содержит начальные моменты в условном масштабе $m_1 = 0{,}56$; $m_2 = 6{,}80$; $m_3 = 14{,}33$; $m_4 = 132{,}30$. От начальных моментов можно перейти к центральным моментам и далее к статистическим характеристикам.

Поскольку нулевой класс выбран произвольно и необходимо учесть размер классов, формулы вычисления среднего значения и центральных моментов выглядят следующим образом:

среднее значение

$$\bar{x} = x_0 + m_1 h; \qquad (2.16)$$

центральные моменты:

$$\mu_1 = 0;\ \mu_2 = (m_2 - m_1^2)h^2 - h/12;$$

$$\mu_3 = (m_3 - 3m_2m_1 + 2m_1^3)h^3 - m_1h^2; \qquad (2.17)$$

$$\mu_4 = (m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4)h^4 - \left(\frac{m_2 - m_1^3}{2}\right)h^2 + \frac{7}{240}h^4.$$

Таблица 2.5

Расчет статистических характеристик по сгруппированным данным (по данным гистограммы рис.2.1)

Класс x, %	Частота n_i	Номер класса x_i	Произведения				Сумма частот Σn_i
			n_ix_i	$n_ix_i^2$	$n_ix_i^3$	$n_ix_i^4$	
30-32	2	-5	-10	50	-250	1250	2
32-34	6	-4	-24	96	-376	1504	8
34-36	9	-3	-27	81	-243	729	17
36-38	14	-2	-28	56	-112	224	31
38-40	20	-1	-20	20	-20	20	51
40-42	25	0	0	0	0	0	76
42-44	21	1	21	21	21	21	97
44-46	17	2	34	68	136	272	114
46-48	13	3	39	117	351	1053	127
48-50	10	4	40	160	640	2560	137
50-52	5	5	25	125	625	3125	142
52-54	3	6	18	108	648	3888	145
54-56	2	7	14	98	686	4802	147
Сумма	147	–	82	1000	2106	19448	–
Среднее	–	–	0,56	6,80	14,33	132,30	–
Моменты	–	–	m_1	m_2	m_3	m_4	–

Между формулами (2.14) и (2.17) имеются различия. Так, в формулах (2.17) появляется размер классов h, играющий роль масштабного множителя, и поправки Шеппарда, которые возник-

ли из-за того, что внутри классов нивелированы различия между отдельными значениями. Поправка Шеппарда ко второму центральному моменту $-h/12$, к третьему $-m_1h^2$, к четвертому $-(m_2 - m_1^2)/2h^2 + 7/240h^4$.

По данным табл.2.5 вычисляем статистические характеристики:

$$\bar{x} = 41 + 0{,}56\cdot 2 = 42{,}12;\ \ \mu_2 = (6{,}80 - 0{,}56^2 - 1/12)2^2 = 25{,}6;$$

$$\mu_3 = (14{,}33 - 3\cdot 6{,}80\cdot 0{,}56 + 2\cdot 0{,}56^3)2^3 - 0{,}56\cdot 2^2 = 23{,}82;$$

$$\mu_4 = (132{,}3 - 4\cdot 14{,}33\cdot 0{,}56 +$$

$$+ 6\cdot 6{,}80\cdot 0{,}56^2 - 3\cdot 0{,}56^4)2^4 - (6{,}80 - 0{,}56)/2\cdot 2^2 +$$

$$+ 7/240\cdot 2^4 = 1790{,}7;$$

$$\sigma^2 = 25{,}6;\ \ \sigma = 5{,}06;\ \ \sigma^3 = 129{,}5;\ \ \sigma^4 = 655{,}36;$$

$$V = 5{,}06/42{,}12 = 0{,}120 = 12{,}0\ \%;\ A = 23{,}82/129{,}5 = 0{,}184;$$

$$E = 1790{,}7/655{,}36 - 3 = -0{,}268.$$

Медиану в сгруппированных данных находят линейной интерполяцией в том классе, где нарастающая сумма частот (последняя графа табл.2.5) переходит через половину общего числа значений n. В рассматриваемом примере из 147 значений средний член имеет порядковый номер (147 + 1)/2 = 74. Следовательно, медиана заключена в классе 40-42, где находятся порядковые номера с 52 по 76. Обозначим начало класса $x_н = 40$, число значений в классе $n_i = 25$. Порядковый номер медианы в классе найдем как разность $n_т = 74 - 51 = 23$. Тогда медиана

$$x_{\text{med}} = x_н + \frac{n_т}{n} h. \qquad (2.18)$$

Подставляя данные, получим $x_{\text{med}} = 40 + 23/25\cdot 2 = 41{,}84$.

Группировка значений случайной величины в классы позволяет найти моду, которой на гистограмме (см. рис.2.1) соответствует максимум частот. Один из приемов нахождения моды основан на параболической интерполяции частот по трем соседним классам,

включая класс с максимальной частотой. В рассматриваемом примере это будут классы 38-40, 40-42, 42-44 с частотами соответственно 20, 25, 21. Обозначим частоты этих классов n_1, n_2, n_3. Тогда мода

$$x_{\text{mod}} = x_0 + \frac{h}{2}\frac{n_1 - n_3}{n_1 - 2n_2 + n_3}, \qquad (2.19)$$

где x_0 – середина класса с максимальной частотой.

Подставляя численные значения, найдем

$$x_{\text{mod}} = 41 + \frac{2}{2}\frac{20 - 21}{20 - 2 \cdot 25 + 21} = 41{,}11.$$

Подведем итог расчета статистических характеристик: среднее значение $\bar{x} = 42{,}12$; медиана $x_{\text{med}} = 41{,}84$; мода $x_{\text{mod}} = 41{,}11$; дисперсия $\sigma^2 = 25{,}6$; среднеквадратичное отклонение $\sigma = 5{,}06$; коэффициент вариации $V = 12{,}0\ \%$; асимметрия $A = 0{,}184$; эксцесс $E = -0{,}268$. ◀◀

Освоив расчет статистических характеристик, можно переходить к рассмотрению законов распределения случайных величин.

2.2. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

2.2.1. Понятие о законах распределения

При увеличении числа наблюдений частости стремятся к пределу, который характеризует вероятность появления случайной величины, а гистограмма частостей стремится к кривой, отражающей *закон распределения вероятностей*. Вид кривой определяется сущностью изучаемого свойства. Иногда на вид кривой влияет методика измерений, например выбор размера проб. Число видов кривых и, соответственно, законов распределения бесконечно велико, но некоторые из них имеют теоретическое обоснование и встреча-

ются чаще других. По крайней мере, реальные распределения приближаются к этим законам.

Закон распределения случайной величины наиболее часто выражается в виде интеграла вероятности:

$$F(x) = \int_{-\infty}^{x} f(x)dx, \qquad (2.20)$$

где $F(x)$ – вероятность p того, что значение случайной величины не превысит значения x, т.е. $p = F(x)$; функция под интегралом $f(x)$ – плотность вероятности случайной величины; к кривой, описываемой функцией $f(x)$, стремится гистограмма частостей при увеличении числа наблюдений.

Интеграл вероятности $F(x)$ при увеличении значения x монотонно растет от нуля до единицы (рис.2.2). Интеграл вероятности (2.20) можно рассматривать как площадь (заштрихована на рис.2.2, *б*), ограниченную осью абсцисс, кривой $f(x)$ и отрезком перпендикуляра, проведенного из точки a. Вся площадь под кривой $f(x)$ равна единице, поэтому заштрихованная площадь меньше единицы и соответствует вероятности p.

Законы распределения случайных величин бывают дискретные и непрерывные. У дискретных законов график плотности вероятности имеет ступенчатый вид, как у гистограммы на рис.2.1, и случайная величина может принимать лишь прерывистые значения (например, число зерен минералов в пробе). К таким законам

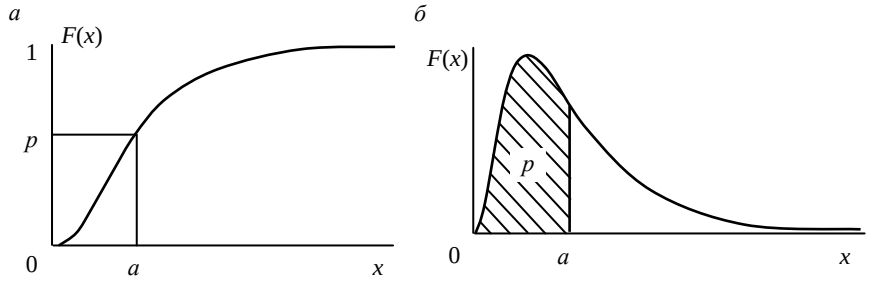

Рис.2.2. Графики интеграла вероятности (*а*) и плотности вероятности (*б*)

относятся биномиальный, Пуассона, гипергеометрический. Законы с непрерывным распределением имеют плавный график плотности вероятности, и случайная величина может принимать любые значения в области своего существования (например, содержание компонента в руде). Сюда относятся законы нормальный, логнормальный, Стьюдента, χ^2, Фишера и некоторые другие.

Рассмотрим наиболее часто употребляемые в геологической практике законы распределения.

2.2.2. Нормальный закон распределения

Среди всех законов распределения чаще других используют нормальный, потому что он носит предельный характер и при определенных условиях к нему приближаются многие другие законы. Нормальный закон описывается интегралом вероятности

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx, \qquad (2.21)$$

плотность вероятности имеет следующий вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}. \qquad (2.22)$$

Кривая, выражаемая формулой (2.22), имеет симметричную форму относительно абсциссы $\bar{x}$ (рис.2.3). Площадь между кривой и осью абсцисс равна единице. Ветви кривой не ограничены и уходят в плюс и минус бесконечность, сливаясь в удалении от величины $\bar{x}$ с осью абсцисс.

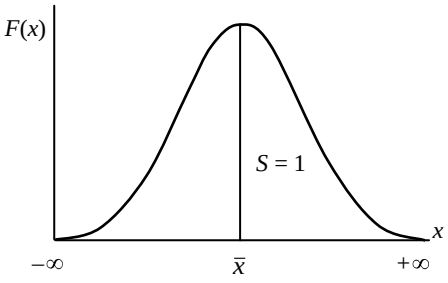

Рис.2.3. График плотности вероятности нормального закона

Как следует из формул (2.21) и (2.22), нормальный закон полностью определяется двумя статистическими характеристиками: средним значением $\bar{x}$ и дисперсией σ^2. Среднее значение определяет положение графика на оси абсцисс, а дисперсия - крутизну ветвей. Кривая плотности вероятности симметричная, асимметрия и эксцесс равны нулю. Вследствие симметричности среднее, медианное и модальное значения совпадают.

Иногда распределения бывают асимметричными (рис.2.4). Отклонение эксцесса от нуля в ту или иную сторону связано с остро- или плосковершинностью кривой распределения по отношению к нормальному распределению (рис.2.5). В частности, кривые с плоской вершиной или с несколькими максимумами имеют отрицательный эксцесс.

Наиболее важное применение нормального закона распределения, как и других законов, состоит в решении задач двух типов: 1) определение вероятности появления случайной величины в заданном интервале; 2) определение интервала возможных значений случайной величины при заданной вероятности.

Вероятность p того, что значение случайной величины не превысит заданное значение a (заштрихованная площадь на рис.2.2) определяется интегралом (2.20), т.е. $p = F(a)$. Наоборот, вероятность α того, что значение случайной величины больше заданного значения a (незаштрихованная площадь на рис.2.2), равна $1 - p$. Часто

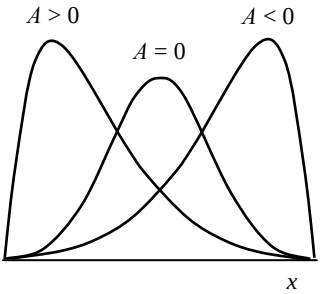

Рис.2.4. Графики плотности вероятности с различной асимметрией

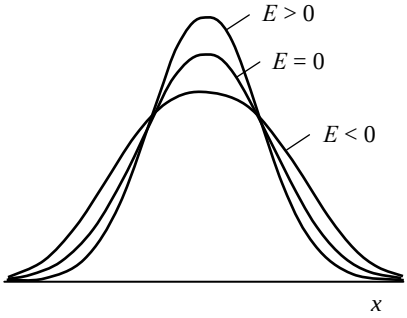

Рис.2.5. Графики плотности вероятности с различными эксцессами

приходится оценивать вероятность q попадания случайной величины в заданный интервал от a до b, ее находят как интеграл:

$$q = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx = F(b) - F(a), \qquad (2.23)$$

которому соответствует заштрихованная площадь на рис.2.6.

Наконец, иногда используется вероятность β того, что случайная величина находится за пределами интервала от a до b, тогда $\beta = 1 - q$.

Особый интерес представляет ситуация, когда размер интервалов берется равным среднеквадратичному отклонению σ. В этом случае практически вся площадь под кривой плотности вероятности (точнее, 99,7 % площади) охватывается интервалом в шесть среднеквадратичных отклонений, т.е. от среднего значения вправо и влево по 3σ (рис.2.7). За пределами этого интервала остается незначительная часть площади, и ею часто пренебрегают.

Вычисление вероятностей сводится к нахождению определенных интегралов (2.21) или (2.23). Интеграл вероятности не интегрируется в алгебраических выражениях, поэтому для нахождения вероятности принято пользоваться специальными таблицами. По-

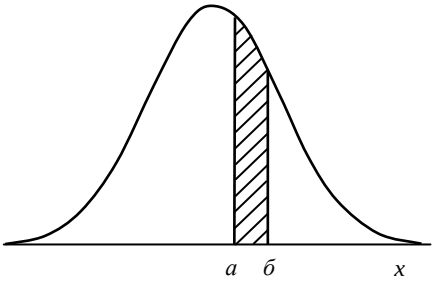

Рис.2.6. График плотности вероятности. Заштрихованная площадь соответствует вероятности q попадания в интервал от a до b

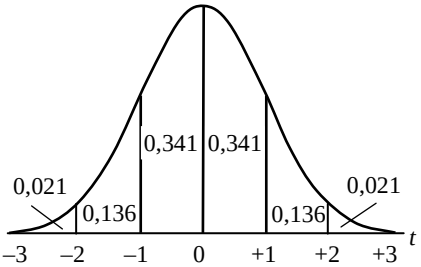

Рис.2.7. График плотности вероятности. Площадь под кривой разделена на шесть частей. В каждой части указан размер площади и, соответственно, вероятность попадания значений случайной величины в этот интервал. Оцифровка оси абсцисс – нормированные среднеквадратичные отклонения

скольку среднее значение $\bar{x}$ и среднеквадратичное отклонение σ могут принимать любые значения, в таблицах трудно учесть все возможные варианты. В связи с этим таблицы составляют в одном варианте для стандартного нормального закона – для нормированных значений случайной величины *t*, которая имеет нулевое математическое ожидание $(\bar{t} = 0)$ и единичное среднеквадратичное отклонение (σ = 1). Чтобы пользоваться такими таблицами, нужно предварительно нормировать исходные значения случайной величины *x* по формуле

$$t = \frac{x - \bar{x}}{\sigma}. \qquad (2.24)$$

Интеграл вероятности $F(t)$ и плотность вероятности $f(t)$ стандартного нормального закона имеют вид

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \mathrm{e}^{-t^2/2} dt; \quad f(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}. \qquad (2.25)$$

Таблицы значений $F(t)$ и $f(t)$ приведены во всех справочниках и пособиях по теории вероятностей, самые распространенные из которых «Таблицы математической статистики» [4] и «Справочник по математике для инженеров и учащихся втузов» [6]. Значения $f(t)$ можно вычислять непосредственно по формулам (2.22) или (2.25).

Большое значение имеет функция $\Phi(t)$, выражаемая интегралом:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} \mathrm{e}^{-t^2/2} dt. \qquad (2.26)$$

Она характеризует вероятность *q* попадания случайной величины в симметричный интервал от $-t$ до $+t$ (рис.2.8) и связана с интегралом вероятности соотношением $\Phi(t) = 2F(t) - 1$.

Отметим, что вероятность попадания случайной величины в интервал от нуля до $+t$ называется функцией Лапласа. Из-за симметричности интеграл (2.26) можно представить как удвоенную функцию Лапласа:

$$\Phi(t) = \sqrt{\frac{2}{\pi}} \int_{0}^{t} \mathrm{e}^{-t^2/2} dt. \qquad (2.27)$$

Значения функций $F(t)$, $f(t)$ и $\Phi(t)$ в пределах от $t = 0$ до $t = 3{,}1$ с шагом аргумента 0,1 приведены в табл.2.6. Для более детального определения значений функций рекомендуются вышеупомянутые справочники [4, 6]. Поскольку функции $f(t)$ и $\Phi(t)$ симметричны относительно $t = 0$, их значения при отрицательном значении t находят из табл.2.6 без учета знака. Для нахождения функции $F(t)$ при отрицательных значениях t нужно использовать соотношение $F(-t) = 1 - F(t)$. Например, при $t = -1{,}7$ из табл.2.6 имеем $f(t) = 0{,}0940$; $\Phi(t) = 0{,}9109$; $F(t) = 1 - 0{,}9554 = 0{,}0446$.

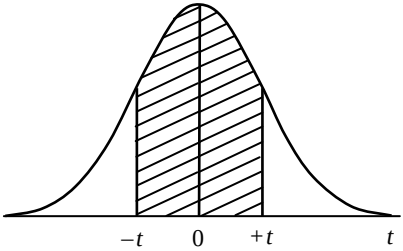

Рис.2.8. График плотности вероятности. Заштрихованная площадь соответствует вероятности попадания значений случайной величины в интервал от $-t$ до $+t$

Таблица 2.6

Функции нормального распределения

t	$F(t)$	$f(t)$	$\Phi(t)$	t	$F(t)$	$f(t)$	$\Phi(t)$
0,0	0,5000	0,3989	0,0000	1,6	0,9452	0,1109	0,8904
0,1	0,5398	0,3970	0,0797	1,7	0,9554	0,0940	0,9109
0,2	0,5793	0,3910	0,1585	1,8	0,9641	0,0790	0,9281
0,3	0,6179	0,3814	0,2358	1,9	0,9713	0,0656	0,9426
0,4	0,6554	0,3683	0,3108	2,0	0,9772	0,0540	0,9545
0,5	0,6915	0,3521	0,3829	2,1	0,9821	0,0440	0,9643
0,6	0,7257	0,3332	0,4515	2,2	0,9861	0,0355	0,9722
0,7	0,7580	0,3123	0,5161	2,3	0,9893	0,0283	0,9786
0,8	0,7881	0,2897	0,5763	2,4	0,9918	0,0224	0,9836
0,9	0,8159	0,2661	0,6319	2,5	0,9938	0,0175	0,9876
1,0	0,8413	0,2420	0,6827	2,6	0,9953	0,0136	0,9907
1,1	0,8643	0,2179	0,7287	2,7	0,9965	0,0104	0,9931
1,2	0,8849	0,1942	0,7699	2,8	0,9974	0,0079	0,9949
1,3	0,9032	0,1714	0,8064	2,9	0,9981	0,0060	0,9963
1,4	0,9192	0,1497	0,8385	3,0	0,9987	0,0044	0,9973
1,5	0,9332	0,1296	0,8664	3,1	0,9990	0,0033	0,9981

Рассмотрим, как определяется вероятность с помощью табл.2.6. Пусть имеется интервал от $a = 2{,}72$ до $b = 2{,}96$; известны также характеристики $\bar{x} = 2{,}2$ и $\sigma = 0{,}40$. По формуле (2.24) вычислим нормированные значения $t_1 = (2{,}72 - 2{,}2)/0{,}40 = 1{,}31$; $t_2 = (2{,}96 - 2{,}2)/0{,}40 = 1{,}90$. В табл.2.6 найдем вероятности $F(t_1) = 0{,}9032$; $F(t_2) = 0{,}9713$. Отсюда имеем вероятность попадания случайной величины в заданный интервал $q = F(t_2) - F(t_1) = 0{,}0681$.

Задача нахождения вероятностей упрощается, если a и b симметричны относительно $\bar{x}$. Тогда достаточно найти $t = t_2$ и вероятность $q = \Phi(t)$. Например, $a = 1{,}94$; $b = 2{,}26$; $\bar{x} = 2{,}10$; $\sigma = 0{,}32$. Имеем $t = (2{,}26 - 2{,}10)/0{,}32 = 0{,}50$ и вероятность $q = \Phi(t) = 0{,}3829$.

Интегралы вероятности (2.25) и (2.26), играющие большую роль, можно вычислять и без применения таблиц путем численного интегрирования на компьютере или с помощью пакета «Stat».

Часто приходится решать обратную задачу – находить интервал возможных значений случайной величины t при заданных вероятностях p, α, q или β. Если задана вероятность $p = F(t)$, то соответствующее ей значение t называется *квантилью* распределения. Она является функцией, обратной интегралу вероятности (2.20), и обозначается $t = F^{-1}(p)$. Квантиль можно найти интерполяцией данных табл.2.6. Например, задана вероятность $p = 0{,}9$. В таблице имеются значения $p = 0{,}8849$ (при $t = 1{,}2$) и $p = 0{,}9032$ (при $t = 1{,}3$). Интерполируя эти значения, найдем, что при $p = 0{,}9$ квантиль $t = 1{,}28$. Квантили, соответствующие вероятностям 1/4; 2/4; 3/4, называются *квартилями*. Вторая квартиль, соответствующая вероятности $p = 0{,}5$, называется *медианой распределения*.

Наиболее часто используют значения t, соответствующие заданной вероятности $q = \Phi(t)$, они называются *коэффициентами вероятности* и служат критериями принятия разнообразных решений. Для нахождения коэффициента вероятности можно воспользоваться интерполяцией данных табл.2.6, но лучше иметь специальную табл.2.7 зависимости t от $\Phi(t)$. Например, задана вероятность $q = 0{,}96$, тогда соответствующий ей коэффициент вероятности $t = 2{,}054$. Табл.2.7 может быть использована и для нахождения квантилей. По заданной вероятности p вычисляется вероятность $q = 2p - 1$ и по табл.2.7 определяется квантиль. Например, дана вероятность

p = 0,9. Вычисляем $q = 2 \cdot 0{,}9 - 1 = 0{,}8$, соответствующая ей квантиль t = 1,282.

Таблица 2.7

Коэффициенты вероятности t при заданной вероятности $q = \Phi(t)$

q	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,013	0,025	0,038	0,050	0,063	0,075	0,088	0,100	0,130
0,1	0,126	0,138	0,151	0,164	0,176	0,189	0,202	0,215	0,228	0,240
0,2	0,253	0,266	0,279	0,292	0,305	0,319	0,332	0,345	0,358	0,372
0,3	0,385	0,399	0,412	0,426	0,440	0,454	0,468	0,482	0,496	0,510
0,4	0,524	0,539	0,553	0,568	0,583	0,598	0,613	0,628	0,643	0,659
0,5	0,674	0,690	0,706	0,722	0,739	0,755	0,772	0,789	0,806	0,824
0,6	0,842	0,860	0,878	0,896	0,915	0,935	0,954	0,974	0,994	1,015
0,7	1,036	1,058	1,080	1,103	1,126	1,150	1,175	1,200	1,227	1,254
0,8	1,282	1,311	1,341	1,372	1,405	1,440	1,476	1,514	1,555	1,598
0,9	1,645	1,695	1,751	1,812	1,881	1,960	2,054	2,170	2,326	2,576
0,99	2,576	2,612	2,652	2,697	2,748	2,807	2,878	2,968	3,090	3,291

На практике наиболее часто используются значения вероятностей q = 0,5 и q = 0,9 (β = 0,5 и β = 0,01). Им соответствуют коэффициенты вероятности t = 1,960 и t = 2,576. С другой стороны, часто задаются значения $t = 2$ и $t = 3$, им соответствуют вероятности q = 0,9545 и q = 0,9973 (см. табл.2.6).

2.2.3. Логарифмически-нормальный закон распределения

В тесной связи с нормальным находится логарифмически-нормальный (сокращенно логнормальный) закон распределения, широко применяемый в геохимии. Замечено, что этим законом удовлетворительно описывается частота появления низких содержаний химических элементов. Академик А.Н.Колмогоров теоретически обосновал логнормальное распределение частиц при дроблении, что используется при гранулометрическом анализе обломочных пород.

Логнормальный закон описывает ситуацию, когда нормальному распределению подчиняются логарифмы значений случайной величины. При расчетах вначале находят натуральные или десятичные логарифмы значений случайной величины. Далее вся работа ведется с логарифмами: вычисляют их среднее значение, дисперсию, среднеквадратичное отклонение, асимметрию, эксцесс, а по таблицам нормального закона определяют вероятности. Какие логарифмы – натуральные или десятичные – использовать для расчетов, не играет роли, потому что они связаны постоянным множителем: натуральные логарифмы в 2,3026 раз больше десятичных (2,3026 = ln10).

Случайная величина в логнормальном законе, в отличие от нормального, имеет область существования от нуля до +∞. Если присутствуют нулевые значения (или следы), что нередко бывает при спектральном и химическом анализах, то это вызывает трудности, так как логарифм нуля равен –∞. Обычно нулевые содержания заменяют какими-то минимальными значениями, например пределом чуствительности анализа. Существуют также способы обработки усеченных распределений, позволяющие получать статистические характеристики при отбрасывании крайних исходных значений.

Обозначим логарифм случайной величины: $z = \ln x$. Плотность вероятности логарифмов описывается формулой нормального закона (2.22)

$$f(z) = \frac{1}{\sigma_z \sqrt{2\pi}} \mathrm{e}^{-\frac{(z-\bar{z})^2}{2\sigma_z^2}}, \qquad (2.28)$$

где $\bar{z}$ – среднее значение логарифмов; σ – среднеквадратичное отклонение логарифмов.

Плотность вероятности исходных значений x выражается формулой логнормального закона

$$f(x) = \frac{1}{x\sigma_x \sqrt{2\pi}} \mathrm{e}^{-\frac{(\ln x - \overline{\ln x})^2}{2\sigma_x^2}}. \qquad (2.29)$$

График функции $f(x)$ асимметричен (рис.2.9), среднее значение, мода и медиана не совпадают между собой. Они связаны с величинами $\bar{z}$ и σ_z^2 следующими соотношениями:

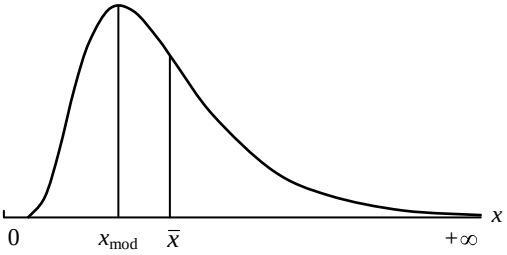

Рис.2.9. График плотности вероятности логнормального закона

$$\bar{x} = e^{\bar{z}+\sigma_z^2/2};\ x_{\text{mod}} = e^{\bar{z}-\sigma_z^2};\ \ x_{\text{med}} = e^{\bar{z}}. \qquad (2.30)$$

Дисперсия исходных данных также определяется соотношением

$$\sigma^2 = e^{2\bar{z}+\sigma_z^2}(e^{\sigma_z^2}-1). \qquad (2.31)$$

При малой дисперсии кривые плотности вероятности логнормального и нормального законов близки между собой и в пределе, при стремлении дисперсии к нулю, совпадают.

2.2.4. Распределение Стьюдента

Распределение Стьюдента, называемое также *t*-распределением, играет большую роль – с его помощью проверяют гипотезы о равенстве средних значений случайных величин. Функция распределения Стьюдента выражается интегралом:

$$F_k(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \int_{-\infty}^{t} \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}} dx, \qquad (2.32)$$

соответственно, плотность вероятности имеет вид

$$f_k(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \qquad (2.33)$$

где k – число степеней свободы, Γ – гамма-функция, выражаемая интегралом:

$$\Gamma(x) = \int_0^\infty y^{x-1} \mathrm{e}^{-y}\, dx\,. \qquad (2.34)$$

Число степеней свободы зависит от числа измерений n случайной величины и от существа поставленной задачи. Если проверяется гипотеза о равенстве вычисленного среднего значения какому-то заранее заданному числу, то $k = n - 1$. Если сравниваются два средних значения из двух совокупностей с числом измерений n_1 и n_2, то $k = n_1 + n_2 - 2$. Могут быть и другие варианты гипотез.

Из функции (2.33) следует, что случайная величина t может принимать любые значения в пределах от $-\infty$ до $+\infty$.

Особенность распределения Стьюдента состоит в том, что его функции зависят от числа степеней свободы, а они, в свою очередь, – от числа измерений. При увеличении значения k распределение приближается к нормальному и в пределе (при $k = \infty$) совпадает с ним. Практически уже при $k = 20$ можно пользоваться таблицами нормального распределения.

Функция распределения (2.32) обычно приводится в табличном виде. Подробные сведения даны в справочнике Л.Н.Большева, Н.В.Смирнова [4]. Здесь приведены лишь некоторые значения (табл.2.8), из них видно, как с увеличением числа k функция распределения Стьюдента приближается к нормальному закону (последняя графа табл.2.8).

Плотность вероятности (2.33) имеет симметричный график, похожий на кривую нормального закона, но более вытянутый по горизонтальной оси (рис.2.10). При увеличении значения k график приближается к кривой нормального закона.

Таблица 2.8

Функция распределения Стьюдента $F_k(t)$ в зависимости от числа степеней свободы k

t	Число степеней свободы k							
	1	2	5	10	20	50	100	∞
0,0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	0,6467	0,6667	0,6808	0,6860	0,6887	0,6903	0,6909	0,6915
1,0	0,7500	0,7887	0,8184	0,8284	0,8354	0,8388	0,8400	0,8413
1,5	0,8128	0,8638	0,9030	0,9178	0,9254	0,9299	0,9314	0,9332
2,0	0,8554	0,9082	0,9490	0,9633	0,9704	0,9744	0,9757	0,9772
2,5	0,8789	0,9352	0,9728	0,9843	0,9884	0,9921	0,9929	0,9938
3,0	0,8976	0,9523	0,9850	0,9933	0,9965	0,9979	0,9983	0,9987
3,5	0,9114	0,9636	0,9914	0,9971	0,9989	0,9995	0,9996	0,9998
4,0	0,9220	0,9714	0,9948	0,9987	0,9996	0,9999	0,9999	
4,5	0,9304	0,9770	0,9968	0,9994	0,9999			
5,0	0,9372	0,9811	0,9980	0,9997				

Значения плотности вероятности обычно находят по таблицам, но их несложно вычислить, преобразовав формулу (2.33) к следующему виду:

$$f_k(t) = \frac{A_k}{\sqrt{\pi k}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}; \quad A_k = \Gamma\left(\frac{k+1}{2}\right) : \Gamma\left(\frac{k}{2}\right).$$

Для нахождения A_k достаточно знать одно значение, например $A_1 = 1/\sqrt{\pi}$, и далее пользоваться рекуррентной формулой $A_{k+1} = k/2A_k$. Так, $A_2 = \sqrt{\pi}/2$, $A_3 = 2/\sqrt{\pi}$ и т.д. Некоторые значения плотности вероятности приведены в табл.2.9. Последняя

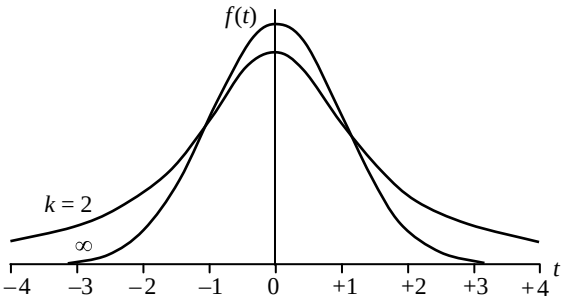

Рис.2.10. График плотности распределения Стьюдента. При $k = \infty$ распределение совпадает с нормальным

графа соответствует нормальному распределению.

Таблица 2.9

Плотность вероятности распределения Стьюдента $f_k(t)$ в зависимости от числа степеней свободы k

t	Число степеней свободы k							
	1	2	5	10	20	50	100	∞
0,0	0,3183	0,3536	0,3796	0,3891	0,3940	0,3970	0,3979	0,3989
0,5	0,2546	0,2962	0,3279	0,3397	0,3458	0,3495	0,3508	0,3521
1,0	0,1592	0,1925	0,219	0,2304	0,2360	0,2396	0,2408	0,2420
1,5	0,0979	0,1141	0,1245	0,1274	0,1286	0,1292	0,1294	0,1295
2,0	0,0637	0,0680	0,0651	0,0611	0,0581	0,0558	0,0549	0,0540
2,5	0,0439	0,0422	0,0333	0,0269	0,0227	0,0197	0,0186	0,0175
3,0	0,0318	0,0274	0,0173	0,0114	0,0080	0,0058	0,0051	0,0044
3,5	0,0240	0,0186	0,0092	0,0048	0,0026	0,0015	0,0012	0,0009
4,0	0,0187	0,0131	0,0051	0,0020	0,0008	0,0003	0,0002	0,0001
4,5	0,0150	0,0095	0,0029	0,0009	0,0003	0,0001	0,0000	0,0000
5,0	0,0122	0,0071	0,0018	0,0004	0,0001	0,0000	0,0000	0,0000

Асимметрия распределения Стьюдента равна нулю, а эксцесс отрицательный. Как и в случае нормального закона, для распределения Стьюдента может быть вычислена функция $\Phi_k(t) = 2F_k(t) - 1$, которая характеризует вероятность q попадания случайной величины в симметричный интервал от $-t$ до $+t$.

На практике для принятия решений чаще используется противоположный показатель $\beta = 1 - q$. Можно составить таблицу зависимости β от t при различных степенях свободы k. Но гораздо важнее знать обратную функцию: чему равно значение коэффициента t при заданной вероятности β и известной степени свободы k (табл.2.10), так как коэффициент t часто используют в качестве критерия принятия решений. Именно такие таблицы с различными вариациями и приводятся в разнообразных справочниках. Например, задана вероятность $\beta = 0{,}05 = 5\ \%$ и число степеней свободы $k = 15$. Из табл.2.10 имеем $t = 2{,}131$.

Таблица 2.10

Коэффициенты вероятности *t* распределения Стьюдента при заданной вероятности β и степени свободы *k*

k	Вероятность β						
	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	2,353	3,182	4,541	5,841	7,453	10,214	12,924
4	2,132	2,776	3,747	4,604	5,597	7,173	8,610
5	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	1,763	2,131	2,602	2,947	3,286	3,733	4,073
16	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	1,740	2,110	2,567	2,898	3,222	3,645	3,985
18	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	1,729	2,093	2,540	2,861	3,174	3,579	3,883
20	1,725	22,086	2,528	2,845	3,153	3,552	3,849
22	1,717	2,074	2,508	2,819	3,119	3,505	3,792
24	1,711	2,064	2,492	2,797	3,091	3,467	3,745
26	1,706	2,056	2,479	2,779	3,067	3,435	3,707
28	1,701	2,048	2,467	2,763	3,047	3,408	3,674
30	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	1,676	2,009	2,403	2,678	2,937	3,261	3,496
100	1,660	1,984	2,364	2,626	2,871	3,174	3,390
β	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Если число степеней свободы велико (несколько десятков и более), то можно пользоваться таблицами нормального закона (см. табл.2.7). Так, при $\beta = 0{,}01$ имеем $q = 1 - \beta = 1 - 0{,}01 = 0{,}99$ и по табл.2.7 находим $t = 2{,}576$.

Таким образом, распределения нормальное и Стьюдента близки между собой. При малом числе измерений (и, соответственно, степеней свободы) более надежные выводы могут быть сделаны по таблицам распределения Стьюдента, а при большом числе наблюдений следует пользоваться таблицами нормального закона.

2.2.5. Распределение χ^2

Распределение χ^2 служит преимущественно для проверки гипотез о соответствии наблюдаемых частот теоретическим законам распределения. Плотность вероятности распределения описывается формулой

$$f(\chi^2) = f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}}, \qquad (2.35)$$

где k – число степеней свободы, зависящее от числа классов гистограммы n_k (обычно $k = n_k - 3$).

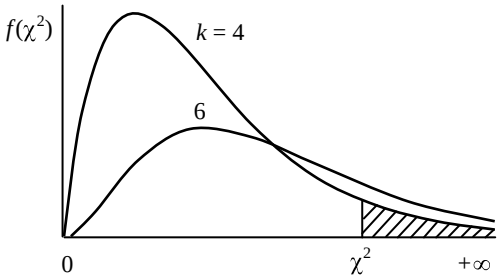

Рис.2.11. График распределения χ^2

Случайная величина χ^2 имеет область существования от нуля до $+\infty$. График плотности вероятности асимметричен (рис.2.11), модальное значение $\chi^2_{\text{mod}} = k - 2$.

При увеличении числа степеней свободы распределение χ^2 прибли-

жается к нормальному с математическим ожиданием k и дисперсией $2k$. Практически при числе степеней свободы $k > 30$ можно переходить к таблицам нормального распределения, заменив величину χ^2 нормированной случайной величиной t:

$$t = (\chi^2 - k)/\sqrt{2k}. \qquad (2.36)$$

Для практических целей требуется иметь таблицу коэффициентов вероятности, играющих роль критериев. В зависимости от вероятности α при заданной степени свободы k (табл.2.11) критерию χ^2 соответствует заштрихованная площадь на рис.2.11.

Таблица 2.11

Коэффициенты вероятности распределения χ^2 при заданных вероятности α и числе степеней свободы k

k	Вероятность α					k	Вероятность α				
	0,10	0,05	0,025	0,01	0,005		0,10	0,05	0,025	0,01	0,005
1	2,71	3,84	5,02	6,64	7,88	16	23,54	26,30	28,84	32,00	34,27
2	4,60	5,99	7,38	9,21	10,66	17	24,77	27,59	30,19	33,41	35,72
3	6,25	7,82	9,35	11,34	12,54	18	25,99	28,87	31,53	34,80	37,16
4	7,78	9,49	11,14	13,28	14,86	19	27,20	30,14	32,85	36,19	38,58
5	9,24	11,07	12,83	15,09	16,75	20	28,41	31,41	34,17	37,57	40,00
6	10,64	12,59	14,45	16,81	18,55	21	29,62	32,67	35,48	38,93	41,40
7	12,02	14,07	16,01	18,48	20,28	22	30,81	33,92	36,78	40,29	42,80
8	13,36	15,51	17,54	20,09	21,96	23	32,01	35,16	38,08	41,64	44,18
9	14,68	16,92	19,02	21,67	23,59	24	33,20	36,42	39,36	42,98	45,56
10	15,99	18,31	20,48	23,21	25,19	25	34,38	37,65	40,65	44,31	46,93
11	17,28	19,68	21,92	24,72	26,76	26	35,56	38,88	41,92	45,64	46,29
12	18,55	21,03	23,34	26,22	28,30	27	36,74	40,11	43,19	46,96	49,64
13	19,81	22,36	24,74	27,69	29,82	28	37,92	41,34	44,46	48,28	50,99
14	21,06	23,68	26,12	29,14	31,32	29	39,09	42,56	45,72	49,59	52,34
15	22,31	25,00	27,49	30,58	32,80	30	40,26	43,77	46,98	50,89	53,67

⏩ Пример 2.3. Выбрана вероятность $\alpha = 0{,}05$, число степеней свободы $k = 15$. Необходимо найти χ^2.

Из табл.2.11 получаем $\chi^2 = 25{,}0$. Если число степеней свободы большое (например, $k = 50$), а вероятность та же, то воспользуемся табл.2.6 нормального закона, для чего найдем $\Phi(t) = 1 - 2\alpha = 0{,}90$, ей соответствует найденное интерполяцией значение $t = 1{,}645$. Из формулы (2.35) следует, что $\chi^2 = 66{,}45$. Для сравнения точное значение $\chi^2 = 67{,}50$. ⏪

2.2.6. Распределение Фишера

Распределение Фишера, называемое также *F*-распределением, используется для проверки гипотезы о равенстве дисперсий случайных величин. В качестве критерия служит отношение несмещенных оценок дисперсий $F = S_1^2 / S_2^2$, причем в числитель отношения всегда помещают бóльшую дисперсию, т.е. $S_1^2 > S_2^2$. Плотность вероятности распределения величины *F* выражается формулой

$$f(t) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} k_1^{k_1/2} k_2^{k_2/2} (k_2 + k_1 t)^{-\frac{k_1+k_2}{2}} t^{(k_1/2-1)}, \quad (2.37)$$

где k_1 и k_2 – количество степеней свободы, зависящее от числа измерений случайных величин n_1 и n_2, т.е. $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$.

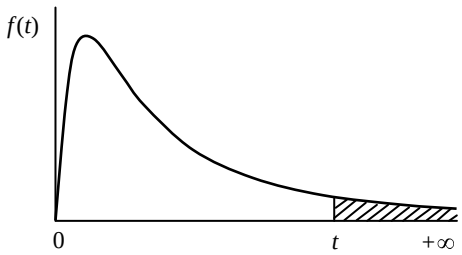

Рис.2.12. График плотности вероятности *F*-распределения

График плотности вероятности асимметричен (рис.2.12) и имеет максимум (моду)

$$t_{\text{mod}} = \frac{k_2(k_1 - 2)}{k_1(k_2 + 2)}. \quad (2.38)$$

Практическое значение имеет зависимость коэффициента t (критерия) от вероятности α (ей соответствует заштрихованная площадь на рис.2.12) при заданных степенях свободы k_1 и k_2. Оценивается вероятность того, что отношение S_1^2 / S_2^2 превысит некоторое критическое значение t. Если отношение S_1^2 / S_2^2 больше t, то дисперсии различаются между собой с вероятностью $p = 1 - \alpha$.

Таблица 2.12

Коэффициенты вероятности *F*-распределения при $\alpha = 0{,}05 = 5$ %

k_2	Число степеней свободы k_1														
	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
3	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,59	8,57	8,55	8,53
4	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,72	5,69	5,66	5,63
5	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,46	4,43	4,40	4,36
6	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,77	3,74	3,70	3,67
7	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,34	3,30	3,27	3,23
8	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,04	3,01	2,97	2,93
9	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,83	2,79	2,75	2,71
10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,66	2,62	2,58	2,54
15	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,20	2,16	2,11	2,07
20	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,99	1,95	1,90	1,84
30	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,79	1,74	1,68	1,62
40	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,69	164	1,58	1,51
60	2,76	2,63	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,59	1,53	1,47	1,39
120	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,75	1,66	1,55	1,50	1,43	1,35	1,25
∞	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,39	1,32	1,22	1,00

Коэффициент t зависит от трех величин: вероятности α, степеней свободы k_1 и k_2, что трудно отобразить в одной таблице, поэтому применяется серия таблиц с различными значениями вероятности α. В качестве примера приведена табл.2.12 значений коэффициента t при наиболее часто используемой вероятности $\alpha = 0{,}05$.

⏩ **Пример 2.4.** Необходимо сравнить две дисперсии: S_1^2 = = 5,22 и S_2^2 = 1,86. Число измерений соответственно $n_1 = 16$, $n_2 = 31$.

Находим отношение $F = 5{,}22/1{,}86 = 2{,}81$, степени свободы $k_1 = 16 - 1 = 15$, $k_2 = 31 - 1 = 30$. Из табл.2.12 имеем $t = 2{,}01$. Поскольку значение $F = 2{,}81$ больше критерия $t = 2{,}01$, дисперсии S_1^2 и S_2^2 различаются между собой с вероятностью более $p = 1 - 0{,}05 = 0{,}95$. ⏪

2.2.7. Построение графика плотности вероятности, проверка гипотезы о законе распределения

Во многих случаях желательно построить график кривой плотности вероятности того или иного закона распределения и совместить его с гистограммой, что позволяет наглядно оценить степень их сходства. В процессе расчета точек кривой можно получить количественные меры соответствия фактической гистограммы теоретическому закону распределения случайной величины.

Рассмотрим построение кривой плотности вероятности нормального закона на примере данных табл.2.4 и рис.2.1. Чтобы построить кривую, необходимо рассчитать значения функции $f(t)$ по формуле (2.22) и нанести их на график. Имеет смысл рассчитать лишь те значения, которые соответствуют серединам классов гистограммы. Расчет выполнен по форме табл.2.13, в которой первые семь граф заимствованы из табл.2.5. Они нужны для расчета статистических характеристик. Для построения кривой достаточно иметь две характеристики: среднее значение и дисперсию. В качестве аргумента можно брать номера классов y, что значительно упрощает расчеты. В результате вычислений получены $\bar{y} = 0{,}56$; $\sigma_y^2 = 6{,}40$; $\sigma_y = 2{,}53$. Далее по формуле (2.24) нужно перейти от условной величины y к нормированной величине t:

$$t = \frac{y - \bar{y}}{\sigma_y} = \frac{y - 0{,}56}{2{,}53}.$$

Зная t, можно рассчитать плотность вероятности $f(t)$ либо по табл.2.6, либо по формуле (2.22). Все необходимые расчеты выполнены в табл.2.13.

Таблица 2.13

Расчет графика плотности вероятности нормального закона

Класс содержаний x, %	n	y	Произведения				t	$f(t)$	$n_т$	Значения $\frac{(n-n_т)^2}{n_т}$
			ny	ny^2	ny^3	ny^4				
30-32	2	-5	-10	50	-250	1250	-2,20	0,0355	2,1	
32-34	6	-4	-24	96	-376	1504	-1,80	0,0790	4,6	0,252
34-36	9	-3	-27	81	-243	729	-1,41	0,1476	8,6	0,019
36-38	14	-2	-28	56	-112	224	-1,01	0,2396	13,9	0,001
38-40	20	-1	-20	20	-20	20	-0,62	0,3292	19,1	0,042
40-42	25	0	0	0	0	0	-0,22	0,3894	22,6	0,255
42-44	21	1	21	21	21	21	0,17	0,3932	22,8	0,142
44-46	17	2	34	68	136	272	0,57	0,3391	19,7	0,370
46-48	13	3	39	117	351	1053	0,96	0,2516	14,6	0,175
48-50	10	4	40	160	640	2560	1,38	0,1582	9,2	0,070
50-52	5	5	25	125	625	3125	1,75	0,0863	5,0	0,000
52-54	3	6	18	108	648	3880	2,15	0,0396	2,3	1,012
54-56	2	7	14	98	686	4802	2,55	0,0154	0,9	
Сумма	147	–	82	1000	2106	19448	–	2,5037	145,4	$\chi^2 = 2{,}338$
Среднее	–	–	0,56	6,80	14,33	1323,3	–	–	–	–

Примечание. n – частота фактическая; y – условный номер класса; t – нормированные значения; $f(t)$ – плотность вероятности; $n_т$ – частота теоретическая.

Сумма плотностей вероятности, равная 2,5037, близка к среднеквадратичному отклонению $\sigma_y = 2{,}53$. Если продолжить кривую плотности вероятности за пределы графика (рис.2.13), то сумма будет точно равна σ_y.

Чтобы перейти от плотности вероятности к теоретической частоте $n_т$, применяется формула

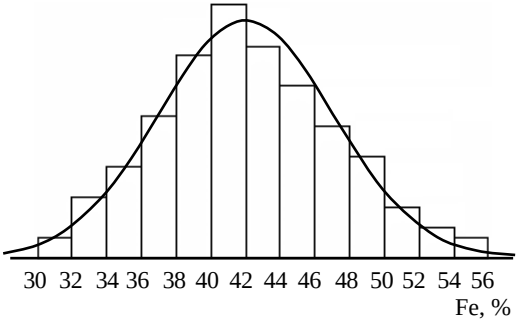

Рис.2.13. Совмещение гистограммы и соответствующей ей кривой плотности вероятности нормального закона распределения

$$n_т = \frac{Nh}{\sigma} f(t). \quad (2.36)$$

Здесь $N = 147$ (суммарное число измерений); h – размер (шаг) класса.

Поскольку в качестве аргумента взят номер класса y, шаг $h = 1$, $\sigma_y = 2{,}53$. Следовательно,

$$n_т = \frac{147 \cdot 1}{2{,}53} f(t) = 58{,}1 f(t).$$

Такое же соотношение получится, если в качестве аргумента взять случайную величину x. Из табл.2.5 имеем $h = 2$, $\sigma = 5{,}06$, и отношение h/σ не изменится. Сумма теоретических частот 145,4 близка к общему числу значений $N = 147$. Продолжив кривую за пределы графика, можно убедиться, что суммы теоретических и фактических частот совпадут.

Сравнение фактических и теоретических частот в табл.2.13 показывает их сходство. Можно совместить частоты на одном графике, построив гистограмму по фактическим частотам n, а кривую плотности вероятности – по теоретическим частотам $n_т$ (рис.2.13).

Последняя графа табл.2.13 позволяет рассчитать критерий χ^2:

$$\chi^2 = \sum_1^{N_k} \frac{(n - n_т)^2}{n_т}, \quad (2.40)$$

где N_k – число классов сравнения частот.

Чтобы избежать грубых случайных расхождений, рекомендуется объединять соседние классы с малым числом фактических частот. В табл.2.13 объединены два первых и два последних класса. В результате число классов $N_k = 11$. Вычисленное значение $\chi^2 = 2{,}338$. Приняв вероятность $\alpha = 0{,}05$ и зная число степеней свободы $k = 11 - 3 = 8$, из табл.2.11 найдем предельное значение $\chi^2_{пр} = 15{,}51$. Так как $\chi^2 < \chi^2_{пр}$, то следует признать, что с вероятно-

стью $p > (1 - \alpha) = 0{,}95$ распределение фактических частот не противоречит нормальному закону. Такая же проверка соответствия фактических частот теоретическим может быть выполнена для других законов распределения.

Возможен и другой способ проверки соответствия, но только нормальному закону. Он основан на том, что асимметрия и эксцесс нормального закона равны нулю. Оценивая степень отклонения фактических значений асимметрии и эксцесса от нуля с помощью какого-либо критерия, можно сделать заключение о соответствии или несоответствии распределения случайной величины нормальному закону. Обычно используется критерий распределения Стьюдента (см. табл.2.10), а при большом числе исходных данных – критерий нормального закона (см. табл.2.7). Проверяют две гипотезы: при числе степеней свободы $k = n - 1$ асимметрия $A = 0$ и эксцесс $E = 0$. Согласно формуле (2.24) имеем

$$t_A = \frac{|A - 0|}{\sigma_A}; \quad t_E = \frac{|E - 0|}{\sigma_E} \tag{2.41}$$

или проще $t_A = |A| / \sigma_A$, $t_E = |E| / \sigma_E$, где $\sigma_A = \sqrt{6/N}$, $\sigma_E = \sqrt{24/N}$. Если t_A и t_E будут меньше предельного значения $t_{\text{пр}}$, то распределение случайной величины не противоречит нормальному закону. Если t_A или t_E больше предельного значения $t_{\text{пр}}$, то распределение противоречит нормальному закону. В качестве предельного значения можно брать $t_{\text{пр}} = 3$, что соответствует вероятности $q = 0{,}997$ (см. табл.2.6). Иногда вероятность принятия решения задается (например, $q = 0{,}95$), тогда $t_{\text{пр}}$ определяют по табл.2.7 ($t_{\text{пр}} = 1{,}96$).

В рассматриваемом примере асимметрия и эксцесс рассчитаны по табл.2.5: $A = 0{,}166$; $E = -0{,}269$ (см. пример 2.2). Вычислим $\sigma_A = \sqrt{6/147} = 0{,}202$; $\sigma_E = \sqrt{24/147} = 0{,}404$, тогда $t_A = 0{,}166/0{,}202 = 0{,}82$; $t_E = 0{,}269/0{,}404 = 0{,}67$. При вероятности $q = 0{,}95$ имеем $t_{\text{пр}} = 1{,}96$. Так как t_A и t_E меньше $t_{\text{пр}}$, то еще раз получаем подтверждение того, что распределение содержаний не противоречит нормальному закону.

Аналогичным способом можно проверить соответствие распределения случайной величины логнормальному распределению, оперируя с логарифмами значений случайной величины.

Критерии t_A и t_E не требуют построения гистограммы, их удобно использовать при любом числе значений случайной величины, но область их применения ограничена нормальным и логнормальным законами. Критерий χ^2 более универсален, его можно использовать для сравнения гистограмм с любыми законами распределения.

Возникает вопрос, с какой целью производится проверка гипотез о законах распределения. Ответ заключается в том, что при статистической обработке значений случайной величины нужно знать вероятность (т.е. надежность) принятия решений, а вероятность можно определить лишь тогда, когда известен закон распределения случайной величины.

2.2.8. Преобразование случайной величины

Большинство решений, принимаемых на базе статистических закономерностей, основано на нормальном законе распределения, играющем универсальную роль. Как было отмечено, при определенных условиях к нему приближаются логнормальный закон, распределение Стьюдента, распределение χ^2 и многие другие. Однако реальное распределение свойств геологических объектов часто отличается от нормального, что вызывает затруднения в принятии решений и в оценке достоверности получаемых выводов. Поэтому принятию решений обычно предшествует проверка соответствия распределения случайной величины нормальному закону, и, если соответствия нет, то можно попытаться преобразовать случайную величину, приведя ее распределение к нормальному. Подобное преобразование применялось выше, когда вместо случайной величины x вводилась новая случайная величина $z = \ln x$. В результате асимметричное логнормальное распределение преобразовывалось в симметричное нормальное.

Представляют интерес такие преобразования, которые превращают произвольно распределенную случайную величину x в слу-

чайную величину z, распределение которой близко к нормальному. Задача заключается в подборе наилучшей функции преобразования.

Преобразование обычно меняет область существования случайной величины. Например, если случайная величина x меняется в пределах от нуля до $+\infty$, то преобразованная случайная величина $z = \ln x$ имеет область существования от $-\infty$ до $+\infty$. Поэтому учет области существования случайной величины может помочь в выборе наилучшего преобразования.

Если случайная величина x имеет область существования от a до b, то преобразование

$$z = \ln \frac{x - a}{b - x} \tag{2.42}$$

меняет пределы ее существования от $-\infty$ до $+\infty$, что во многих случаях эффективно. Частным случаем является ситуация, когда $a = 0$, $b = 1 = 100\ \%$ (например, содержание химических элементов не может быть меньше нуля и больше 100 %), и формула преобразования имеет вид

$$z = \ln \frac{x}{1 - x}. \tag{2.43}$$

Если значения случайной величины x очень малы, то ею в знаменателе можно пренебречь, и получается формула $z = \ln x$, лежащая в основе логнормального распределения. Наоборот, если значения x близки к единице, то получается формула преобразования в правоасимметричное логнормальное распределение $z = -\ln(1 - x)$.

Если случайная величина колеблется в пределах от -1 до $+1$ (например, коэффициент корреляции или многие тригонометрические функции), то эффективным является преобразование

$$z = \ln \frac{1 + x}{1 - x} \tag{2.44}$$

или преобразование, предложенное Фишером

$$z = -\frac{1}{2} \ln \frac{1 + x}{1 - x}. \tag{2.45}$$

Для преобразования могут быть использованы также степенные функции вида $z = x^a$ или $z = x^{-a}$, где a может принимать значения от 1/2 до 3 [2].

⏩ **Пример 2.5.** Рассмотрим подбор функции преобразования для случайной величины x с асимметричным распределением (табл.2.14).

Таблица 2.14

Подбор функции преобразования случайной величины x (содержание TiO_2 в магнетите)

Номер пробы n	Исходная случайная величина x	Преобразованные случайные величины		Вероятность p	Квантиль t
		$\sqrt{x}$	$\ln x$		
1	0,04	0,2000	-3,219	0,025	-1,960
2	0,06	0,2449	-2,813	0,075	-1,440
3	0,07	0,2646	-2,659	0,125	-1,150
4	0,08	0,2828	-2,526	0,175	-0,935
5	0,09	0,3000	-2,408	0,225	-0,755
6	0,09	0,3000	-2,408	0,275	-0,598
7	0,10	0,3162	-2,303	0,325	-0,454
8	0,12	0,3464	-2,120	0,375	-0,319
9	0,13	0,3606	-2,040	0,425	-0,189
10	0,14	0,3742	-1,966	0,475	-0,063
11	0,16	0,4000	-1,833	0,525	0,063
12	0,17	0,4123	-1,772	0,575	0,189
13	0,18	0,4243	-1,715	0,625	0,319
14	0,21	0,4583	-1,561	0,675	0,454
15	0,23	0,4796	-1,470	0,725	0,598
16	0,23	0,4796	-1,470	0,775	0,755
17	0,28	0,5292	-1,273	0,825	0,935
18	0,30	0,5447	-1,205	0,875	1,150
19	0,34	0,5831	-1,079	0,925	1,440
20	0,45	0,6708	-0,799	0,975	1,960
Среднее	0,1735	0,3987	-1,932	–	–
Дисперсия	0,0107	0,0145	0,379	–	–
Асимметрия	0,979	0,430	-0,145	–	–

Окончание табл.2.14

Номер пробы n	Исходная случайная величина x	Преобразованные случайные величины		Вероятность p	Квантиль t
		$\sqrt{x}$	$\ln x$		
Эксцесс	0,394	-0,537	-0,688	–	–
Критерий t_A	1,79	0,79	0,26	–	–
Критерий t_E	0,36	0,49	0,63	–	–

Применим формулы преобразования: $z = \sqrt{x}$ и $z = \ln x$. Они дали положительные результаты – распределение преобразованной величины по критериям асимметрии и эксцесса не противоречит нормальному закону. Лучший результат дало преобразование $z = \ln x$, особенно по критерию асимметрии. ◀◀

Подбор функции преобразования удобно контролировать с помощью графика пробит-функции [12]. По оси абсцисс откладывают значения случайной величины x или z, а по оси ординат – квантили нормального распределения t (рис.2.14), которые соответствуют вероятностям $p = (2n - 1)/2N$, где n – порядковый номер случайной величины в упорядоченном ряду наблюдений (табл.2.14), $N = 20$ – число наблюдений. На пересечении абсциссы и ординаты ставят точки. Если точки на графике расположены вдоль прямой линии, то

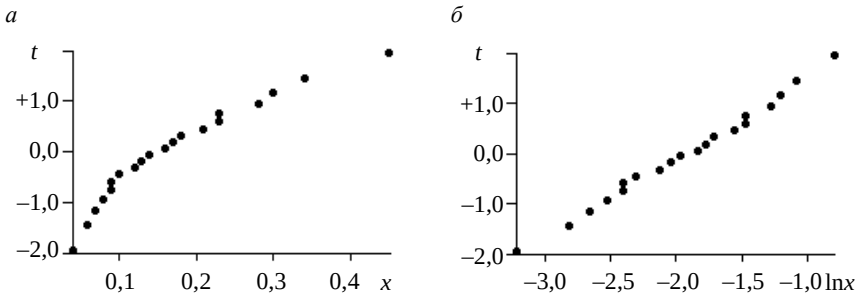

Рис.2.14. Графики пробит-функции для исходных (*а*) и преобразованных (*б*) данных

распределение случайной величины близко к нормальному. В противном случае распределение отличается от нормального. В рассматриваемом примере график пробит-функции для исходной случайной величины явно нелинейный, а для преобразованной случайной величины $z = \ln x$ – близкий к линейному, что подтверждает эффективность выбранного преобразования.

Для построения графика пробит-функции рассчитаны значения вероятностей p, а соответствующие им квантили t получены интерполяцией данных табл.2.7.

2.3. ГЕОЛОГИЧЕСКИЕ ПРИЛОЖЕНИЯ ОДНОМЕРНОЙ СТАТИСТИЧЕСКОЙ МОДЕЛИ

2.3.1. Точечная оценка погрешности среднего значения

Среднее значение $\bar{x}$ из n независимых значений случайной величины x также является случайной величиной. Если случайная величина x имеет дисперсию σ^2, то среднее значение $\bar{x}$, как показано в подразделе 2.1.2, имеет дисперсию δ^2 в n раз меньше:

$$\delta^2 = \sigma^2/n \text{ или } \delta = \sigma / \sqrt{n}\,. \qquad (2.46)$$

Величину δ можно рассматривать как *абсолютную* среднеквадратичную случайную погрешность среднего значения $\bar{x}$.

Если разделить обе части равенства (2.46) на среднее значение $\bar{x}$, то получим *относительную* погрешность

$$\tau = \frac{\delta}{\bar{x}} = \frac{V}{\sqrt{n}}\,, \qquad (2.47)$$

где V – коэффициент вариации. Относительная погрешность может быть выражена в долях единицы или в процентах.

Формулы (2.46) и (2.47) играют большую роль: они показывают, что погрешность среднего значения прямо пропорциональна

изменчивости случайной величины и обратно пропорциональна корню квадратному из числа измерений. Это позволяет решать две задачи: 1) оценивать абсолютную δ или относительную τ погрешность среднего значения при известном числе наблюдений n; 2) находить необходимое число измерений n для достижения заданной погрешности среднего значения.

⏩ **Пример 2.6.** В результате анализа 16 проб гранита рассчитано среднее содержание кремнезема $\bar{x}$ = 70,35 % и среднеквадратичное отклонение σ^2 = 3,20 %. Определить, чему равна среднеквадратичная погрешность среднего содержания и сколько дополнительно нужно взять проб, чтобы снизить относительную погрешность до 1 %.

Абсолютная среднеквадратичная случайная погрешность δ = $= 3{,}20 / \sqrt{16} = 0{,}80$ %; относительная случайная погрешность τ = $= 0{,}80 / 70{,}35 = 1{,}14$ %.

Продолжим задачу. Если τ = 1 % = 0,01, то из формулы (2.47) получим $\delta = \tau\bar{x} = 0{,}01 \cdot 70{,}35 = 0{,}70$. Из формулы (2.46) имеем $n = \sigma^2 / \delta^2 = 3{,}20^2 / 0{,}70^2 = 21$. Следовательно, дополнительно нужно взять и проанализировать 21 – 16 = 5 проб. ⏪

2.3.2. Интервальная оценка математического ожидания случайной величины

Обычно среднее значение случайной величины $\bar{x}$ находят по выборке из генеральной совокупности. Математическое ожидание случайной величины в генеральной совокупности $M(x)$ обычно неизвестно. Его можно приближенно оценить с помощью выборочного среднего значения $\bar{x}$, которое является случайной величиной и имеет дисперсию δ^2. Чаще всего с достаточным основанием предполагается, что случайная величина $\bar{x}$, как представляющая собою сумму многих случайных величин, имеет распределение, близкое к нормальному. Размах значений нормально распределенной вели-

чины составляет приближенно ±3δ (ширина кривой нормального распределения на рис.2.7). Где-то в этом интервале и заключено математическое ожидание $M(x)$. Наиболее вероятно, что оно совпадает со средним значением $\bar{x}$, которое является *точечной оценкой* математического ожидания. Менее вероятно, что математическое ожидание смещено в ту или иную сторону от среднего значения. Интервал возможных значений математического ожидания зависит от вероятности $q = \Phi(t)$ и выражается через коэффициент вероятности t соотношением

$$\bar{x} - t\delta < M(x) < \bar{x} + t\delta. \qquad (2.48)$$

Данный интервал называется *доверительным интервалом* или *интервальной оценкой* математического ожидания. Каждому значению вероятности q соответствует определенный коэффициент вероятности t (табл.2.6 и 2.7) и размер доверительного интервала:

Вероятность $q = \Phi(t)$	Коэффициент вероятности t	Доверительный интервал
0,683	1	$\bar{x} - \delta < M(x) < \bar{x} + \delta$
0,954	2	$\bar{x} - 2\delta < M(x) < \bar{x} + 2\delta$
0,997	3	$\bar{x} - 3\delta < M(x) < \bar{x} + 3\delta$

Используя данные примера 2.6, в котором известно среднее содержание кремнезема в граните $\bar{x} = 70{,}35$ %, и $\delta = 0{,}80$ %, получаем доверительные интервалы:

Вероятность q	Доверительный интервал
0,683	$69{,}65 < M(x) < 71{,}15$
0,954	$68{,}75 < M(x) < 71{,}95$
0,997	$67{,}95 < M(x) < 72{,}75$

Какую из вероятностей q принять за основу, нельзя решить математическим путем, так как ответ лежит в области принятия решений и должен опираться на какое-то логическое или экономическое обоснование. Практически в менее ответственных случаях принимают $t = 2$ и $q = 0{,}954$, в более ответственных случаях $t = 3$

и q = 0,997. При наличии достаточного обоснования могут приниматься и дробные значения t.

Если среднее значение $\bar{x}$ или другая оцениваемая величина подчиняются не нормальному, а другому закону распределения, то, естественно, вероятность q будет иная.

2.3.3. Выделение аномальных значений

Статистические характеристики и получаемые на их основе выводы имеют смысл лишь для однородных совокупностей. При объединении двух и более однородных совокупностей с различными статистическими характеристиками расчеты по объединенной совокупности обычно не имеют смысла. Искажение статистических характеристик происходит и в том случае, когда в однородную совокупность попадают единичные значения, значительно отличающиеся от среднего, называемые аномальными или ураганными. Поэтому актуальной является задача о разделении неоднородной совокупности на однородные, о выделении из неоднородных совокупностей аномальных значений. Данная задача имеет несколько способов решения при условии, что известен или задан закон распределения случайной величины.

Распространенный способ выделения аномальных значений называется правилом «*трех сигм*» и основан на том, что случайная величина при нормальном законе распределения практически полностью (на 99,7 %) заключена в пределах от $\bar{x} - 3\sigma$ до $\bar{x} + 3\sigma$ (см. рис.2.7). Если значение случайной величины отличается от среднего значения $\bar{x}$ больше чем на 3σ, то оно является аномальным. Естественно, что испытуемое значение не должно участвовать в расчете среднего значения и среднеквадратичного отклонения. Для удобства расчетов можно нормировать случайную величину по формуле (2.24). Тогда правило «трех сигм» преобразуется: если нормированное значение $|t| > 3$, то оно является аномальным.

⏩ **Пример 2.7.** Средняя зольность угля $\bar{x} = 6{,}5$ %, среднеквадратичное отклонение $\sigma = 2{,}1$ %. Определить, не является ли аномальной проба угля с зольностью 15 %.

Найдем нормированное значение $t = (15 - 6{,}5)/2{,}1 = 4{,}05$. Поскольку $t > 3$, проба является аномальной и относится к другой совокупности.

На основе приведенных данных можно определить, какие вообще значения зольности являются аномальными. Так как $\bar{x} - 3\sigma = 6{,}5 - 3 \cdot 2{,}1 = 0{,}2$ %; $\bar{x} - 3\sigma = 6{,}5 + 3 \cdot 2{,}1 = 12{,}8$ %, то аномальными являются значения зольности менее 0,2 и более 12,8 %. ⏪

Если распределение случайной величины логнормальное, то правило «трех сигм» применяется к логарифмам значений, что используется при геохимическом методе поисков месторождений для выделения геохимических аномалий.

⏩ **Пример 2.8.** Среднее (фоновое) содержание меди $\bar{x} = 0{,}018$, дисперсия натуральных логарифмов $\sigma_z^2 = 0{,}22$. Определить, какие содержания меди надо считать аномальными.

Используя формулы подраздела 2.2.3, найдем $\sigma_z = \sqrt{0{,}22} = 0{,}47$; $\bar{z} = \ln \bar{x} - \sigma_z^2/2 = \ln 0{,}018 - 0{,}22/2 = -4{,}13$. Нижний предел логарифмов $z_1 = \bar{z} - 3\sigma_z = -4{,}13 - 3 \cdot 0{,}47 = -5{,}54$. Верхний предел логарифмов $z_2 = \ln \bar{x} + \sigma_z^2/2 = -4{,}13 + 3 \cdot 0{,}47 = -2{,}72$. Так как $z = \ln x$, то $x = e^z$ и получаем нижний предел содержаний $x_1 = e^{-5{,}54} = 0{,}004$ %, верхний предел $x_2 = e^{-2{,}72} = 0{,}066$ %. Следовательно, аномальными являются содержания меди менее 0,004 и более 0,066 %. На практике нижним пределом обычно пренебрегают, полагая его равным нулю. ⏪

Наряду с правилом «трех сигм» существуют и другие правила выявления аномальных значений. Более общее правило состоит в том, что задается либо вероятность q, либо соответствующая ей предельная величина критерия t. Если нормированное значение пре-

вышает предельное значение t, то значение случайной величины является аномальным.

Следует учесть, что при исключении аномальных значений происходит искажение (смещение) статистических характеристик оставшейся совокупности. Так, если из нормально распределенной совокупности исключить одно или несколько максимальных значений, то уменьшатся среднее значение и дисперсия – возникает усеченное нормальной распределение. Это обстоятельство рекомендуется учитывать при выделении аномальных значений.

Обозначим смещенные характеристики усеченного распределения: среднее значение $\bar{x}_{\text{смещ}}$ и дисперсия $\sigma^2_{\text{смещ}}$, тогда их связь с несмещенными характеристиками выражается формулами

$$\bar{x}_{\text{смещ}} = \bar{x} - \sigma_y; \tag{2.49}$$

$$\sigma^2_{\text{смещ}} = \sigma^2_y (1 - ty - y^2); \tag{2.50}$$

$$y = \frac{N}{N-n} f(t), \tag{2.51}$$

где y – нормированное смещение среднего; n – число исключенных значений; N – общее число значений случайной величины; $f(t)$ – функция плотности вероятности (2.25); t – квантиль нормального распределения, соответствующая вероятности $p = 1 - n/N$, т.е. $t = F^{-1}(1 - n/N)$.

Поскольку статистические характеристики изменяются, происходит и смещение критерия t:

$$t_{\text{смещ}} = \frac{t + y}{\sqrt{1 - ty - y^2}}. \tag{2.52}$$

Из приведенных формул следует, что величины t, $f(t)$, y, $t_{\text{смещ}}$ зависят только от отношения n/N.

⏩ **Пример 2.9.** Необходимо проверить аномальность максимальных значений табл.2.15.

Таблица 2.15

Пример выявления аномальных значений

Номер пробы n	Значения x	Квантиль t	Номер пробы n	Значения x	Квантиль t
1	0,06	-2,07	14	0,49	0,05
2	0,15	-1,57	15	0,50	0,15
3	0,21	-1,30	16	0,52	0,24
4	0,25	-1,10	17	0,53	0,34
5	0,28	-0,94	18	0,57	0,45
6	0,29	-0,80	19	0,60	0,56
7	0,32	-0,67	20	0,64	0,67
8	0,35	-0,56	21	0,67	0,80
9	0,38	-0,45	22	0,73	0,94
10	0,39	-0,34	23	0,75	1,10
11	0,42	-0,24	24	0,80	130
12	0,45	-0,15	25	1,14	1,57
13	0,47	-0,05	26	1,19	2,07

Вначале найдем среднее и дисперсию из всех 26 значений: $\bar{x} = 0{,}502$; $\sigma^2 = 0{,}06478$; $\sigma = 0{,}2545$. Далее вычислим среднее и дисперсию из 24 значений, исключив максимальные значения. Получим смещенные оценки $\bar{x}_{\text{смещ}} = 0{,}451$; $\sigma^2_{\text{смещ}} = 0{,}03577$; $\sigma = 0{,}1891$. Вычислим нормированные значения исключенных значений: $t_{25} = (1{,}14 - 0{,}451)/0{,}1891 = 3{,}64$; $t_{26} = (1{,}19 - 0{,}451)/0{,}1891 = 3{,}91$. Поскольку нормированные значения $t_{25} > 3$ и $t_{26} > 3$, по правилу «трех сигм» оба исключенных значения являются аномальными. Однако полученный вывод является некорректным, так как он построен на смещенных оценках.

Оценим размер смещения, обусловленный исключением двух максимальных значений. Имеем $p = 1 - n/N = 0{,}923$. Вероятности p соответствует квантиль $t = F^{-1}(p) = 1{,}426$. По формуле (2.25) найдем $f(t) = 0{,}1443$, по формуле (2.51) определим нормированное смещение $y = 26/24 \cdot 0{,}1443 = 0{,}1563$. Из формулы (2.50) следует $\sigma^2 = \sigma^2_{\text{смещ}} / (1 - ty - t^2) = 0{,}03577 / (1 - 1{,}426 \cdot 0{,}1563 - 0{,}1563^2) =$

= 0,04752; $\sigma = 0{,}218$. Из формулы (2.49) получаем $\bar{x} = x_{\text{смещ}} + \sigma_y = 0{,}451 + 0{,}218 \cdot 0{,}1563 = 0{,}485$. Полученные оценки приведены в табл.2.16.

Таблица 2.16

Результат вычисления истинных характеристик

Параметр	Исходная совокупность с учетом аномальных значений	Смещенные характеристики после исключения аномальных значений	Несмещенные «истинные» характеристики
Среднее значение	0,502	0,451	0,485
Дисперсия	0,06478	0,03577	0,4752
Среднеквадратичное отклонение	0,2545	0,1891	0,2180

По формуле (2.52) найдем смещенный критерий:

$$t_{\text{смещ}} = (3 + 0{,}1563) / \sqrt{1 - 1{,}426 \cdot 0{,}1563 - 0{,}1563^2} = 3{,}638.$$

Отсюда следует, что, вместо $t = 3$ для проверки аномальности значений нужно пользоваться $t_{\text{смещ}} = 3{,}638$, что довольно существенно. Но и с учетом смещенного критерия исключенные значения являются аномальными. ◀◀

Поскольку смещение критерия $t_{\text{смещ}}$ зависит только от отношения n/N, на основе формул (2.51) и (2.52) могут быть составлены таблицы $t_{\text{смещ}}$ для различных значений t. Для примера приведена табл.2.17, соответствующая $t = 3$, т.е. правилу «трех сигм».

Аномальные значения можно выявить и на графике пробит-функции (рис.2.15), построенном

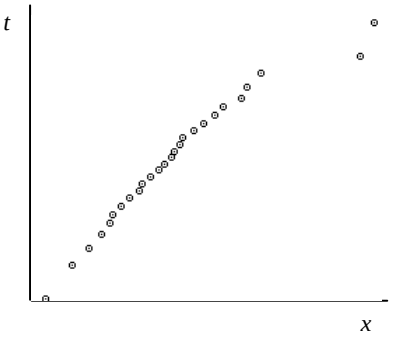

Рис.2.15. График пробит-функции с аномальными значениями

по данным табл.2.15 по методике, описанной в подразделе 2.3.2. Большинство точек укладывается в одну линию, но две точки заметно отклоняются от нее, что свидетельствует об аномальности соответствующих им значений.

Таблица 2.17

Значения смещенного критерия $t_{смещ}$ при заданном t = 3

n/N	0	1	2	3	4	5	6	7	8	9
0,00	3,000	3,019	3,034	3,048	3,061	3,073	3,085	3,096	3,107	3,117
0,01	3,128	3,138	3,148	3,157	3,167	3,176	3,186	3,195	3,204	3,213
0,02	3,221	3,230	3,239	3,247	2,256	3,264	3,272	3,280	3,289	3,297
0,03	3,305	3,313	3,320	3,328	3,336	3,344	3,351	3,359	3,367	3,374
0,04	3,382	3,389	3,397	3,404	3,411	3,419	3,426	3,433	3,440	3,448
0,05	3,455	3,462	3,469	3,476	3,483	3,490	3,497	3,504	3,511	3,518
0,06	3,525	3,532	3,538	3,545	3,552	3,559	3,566	3,572	3,579	3,586
0,07	3,592	3,599	3,606	3,612	3,619	3,626	3,632	3,639	3,645	3,652
0,08	3,658	3,665	3,671	3,678	3,684	3,691	3,697	3,704	3,710	3,717
0,09	3,723	3,729	3,736	3,742	3,748	3,755	3,761	3,767	3,774	3,780
0,10	3,786	3,793	3,799	3,805	3,811	3,818	3,824	3,830	3,836	3,842
0,11	3,849	3,855	3,861	3,867	3,873	3,879	3,886	3,892	3,898	3,904
0,12	3,910	3,916	3,922	3,928	3,935	3,941	3,947	3,953	3,959	3,965
0,13	3,971	3,977	3,983	3,989	3,995	4,001	4,007	4,013	4,019	4,025
0,14	4,031	4,037	4,043	4,049	4,055	4,061	4,067	4,073	4,079	4,085
0,15	4,091	4,097	4,103	4,109	4,115	4,121	4,127	4,133	4,139	4,145
0,16	4,151	4,157	4,163	4,169	4,175	4,181	4,187	4,193	4,198	4,204
0,17	4,210	4,216	4,222	4,228	4,234	4,240	4,246	4,252	4,258	4,264
0,18	4,269	4,275	4,281	4,287	4,293	4,299	4,305	4,311	4,317	4,323
0,19	4,328	4,334	4,340	4,346	4,352	4,358	4,364	4,370	4,376	4,381
0,20	4,387	4,393	4,399	4,405	4,411	4,417	4,423	4,429	4,434	4,440

Еще один способ выявления аномальных значений основан на применении критерия Титьена – Мура [14]. Если из нормально распределенной совокупности, содержащей N значений, исключить n максимальных или минимальных значений, то дисперсия умень-

шится, и по степени ее уменьшения можно судить об аномальности исключенных значений. Вначале вычисляется величина

$$L = \frac{N-n}{N}\frac{\sigma^2_{N-n}}{\sigma^2_N}, \qquad (2.53)$$

где σ^2_N – дисперсия исходной совокупности; σ^2_{N-n} – дисперсия после исключения n предполагаемых аномальных значений.

Если значение L окажется меньше критерия $L_{доп}$ при заданной вероятности α, то исключенные значения являются аномальными. Для примера приведена табл.2.18 с вероятностью α = 0,05 [14].

Таблица 2.18

Критерий Титьена – Мура при α = 0,05

N	Количество исключенных значений n									
	1	2	3	4	5	6	7	8	9	10
3	0,003									
4	0,051	0,001								
5	0,125	0,018								
6	0,203	0,055	0,010							
7	0,273	0,106	0,032							
8	0,326	0,146	0,064	0,022						
9	0,372	0,194	0,099	0,045						
10	0,418	0,233	0,129	0,070	0,034					
11	0,454	0,270	0,162	0,098	0,054					
12	0,489	0,305	0,196	0,125	0,076	0,042				
13	0,517	0,337	0,224	0,150	0,098	0,060				
14	0,540	0,363	0,250	0,174	0,122	0,079	0,050			
15	0,556	0,387	0,276	0,197	0,140	0,097	0,066			
16	0,575	0,410	0,300	0,219	0,159	0,115	0,082	0,055		
17	0,594	0,427	0,322	0,240	0,181	0,136	0,100	0,072		
18	0,608	0,447	0,337	0,259	0,200	0,154	0,116	0,086	0,062	
19	0,624	0,462	0,354	0,277	0,209	0,168	0,130	0,099	0,074	
20	0,639	0,484	0,377	0,299	0,238	0,188	0,150	0,115	0,088	0,066
25	0,696	0,550	0,450	0,374	0,312	0,262	0,222	0,184	0,154	0,126

Окончание табл.2.18

N	Количество исключенных значений n									
	1	2	3	4	5	6	7	8	9	10
30	0,730	0,599	0,506	0,434	0,376	0,327	0,283	0,245	0,212	0,183
35	0,762	0,642	0,554	0,482	0,424	0,376	0,334	0,297	0,264	0,235
40	0,784	0,672	0,588	0,523	0,468	0,421	0,378	0,342	0,310	0,280
45	0,802	0,696	0,618	0,556	0,502	0,456	0,417	0,382	0,350	0,320
50	0,820	0,722	0,646	0,588	0,535	0,490	0,450	0,414	0,383	0,355

Примерим критерий Титьена – Мура к данным табл.2.15. Дисперсия исходной совокупности $\sigma_N^2 = 0{,}06478$; дисперсия после исключения двух значений $\sigma_{N-n}^2 = 0{,}03577$. Следовательно,

$$L = \frac{26-2}{26}\,\frac{0{,}03577}{0{,}06478} = 0{,}510.$$

Из табл.2.18 интерполяцией находим критерий $L_{\text{доп}} = 0{,}560$. Так как $L < L_{\text{доп}}$, то исключенные значения являются аномальными.

На графике пробит-функции при большом количестве данных можно выявить и другие особенности поведения случайной величины. На рис.2.16 показаны фактические данные по содержанию меди на колчеданном месторождении. Стрелками выделены две точки – нижняя и верхняя. В верхней точке проходит граница аномальных проб (более 16 %), в нижней точке – естественная природная граница кондиционных руд (около 0,5 %). Средняя часть графика близка к прямой линии, что соответствует нормальному закону распределения.

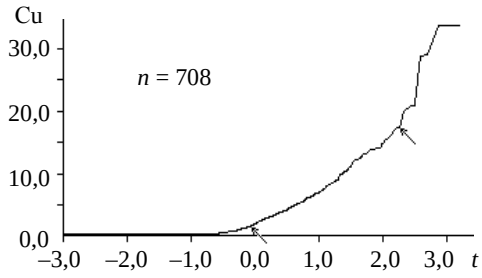

Рис.2.16. Пробит-график содержаний меди в руде

2.3.4. Выделение однородных совокупностей

Одна из сложных проблем при обработке статистических данных – это разделение неоднородной совокупности на однородные. Заключение о неоднородности совокупности лучше всего делать по гистограмме частот. Например, на рис.2.17 явно выделяются два максимума частот, соответствующие двум однородным совокупностям. Одна совокупность имеет моду при 27 % содержания железа, другая – при 55 %. Геологическая причина появления двух совокупностей заключается в том, что бедные руды возникли путем замещения алюмосиликатных пород, а богатые – карбонатных пород.

Для статистического исследования рекомендуется разделить данные опробования на две однородные совокупности. Это можно сделать двумя способами: 1) раздельным изучением руд, образованных по алюмосиликатным и карбонатным породам (геологический способ изучения); 2) аналитическим способом, что требует применения сложных расчетов при условии, что задан или известен закон распределения каждой совокупности

Возможна и обратная ситуация: наличие неоднородной совокупности на гистограмме позволяет сделать определенные геологические выводы.

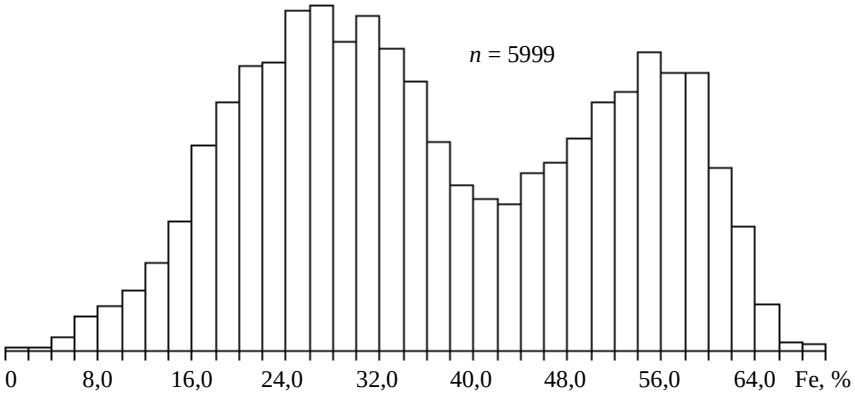

Рис.2.17. Гистограмма содержаний железа в рудах Качарского месторождения

Так, на рис.2.18 показано распределение стронция в апатите в логарифмическом масштабе. На гистограмме выделяются три однородные совокупности. Первая совокупность соответствует содержанию SrO 0,01-0,05 %, вторая 0,05-1 %, третья 1-13 %. Следовательно, имеется три разновидности апатита с различным содержанием стронция. Анализ адресов проб показывает, что они относятся к различным типам месторождений и горных пород. Наиболее чистыми по содержанию стронция являются апатиты из гранитоидов, ультрабазитов и метаморфических пород. Средние по содержанию стронция – это апатиты скарновых месторождений и некоторых массивов щелочных пород. Наиболее высокие содержания стронция наблюдаются в апатитах Хибинской группы месторождений.

Однородные совокупности, входящие в смешанную совокупность, различаются средними значениями $\bar{x}$, $\bar{y}$ и дисперсиями σ_x^2, σ_y^2. Важным показателем, определяющим возможность аналитического разделения смешанных совокупностей при условии нормального их распределения, является *раздвиг* распределений:

$$d = \frac{|\bar{x} - \bar{y}|}{\sqrt{\sigma_x^2 + \sigma_y^2}}, \qquad (2.54)$$

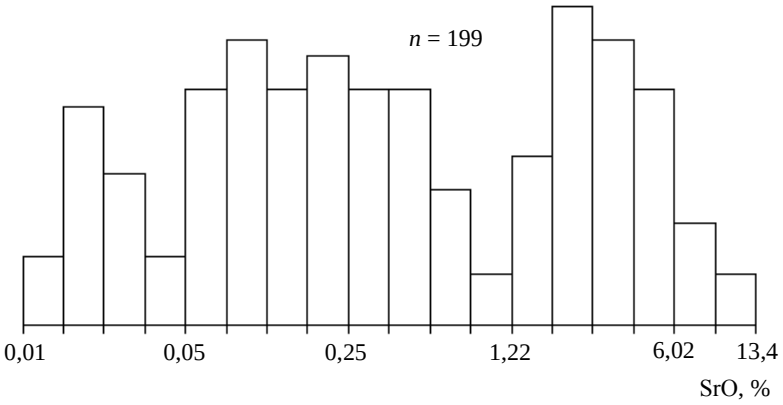

Рис.2.18. Гистограмма содержания стронция в апатитах различных природных образований. Масштаб по горизонтальной оси логарифмический

который по смыслу близок к критерию t. Чем больше раздвиг, тем легче разделить неоднородную совокупность на однородные и определить их характеристики. Можно выделить несколько вариантов разделения:

1. Раздвиг очень большой ($d > 4$), гистограмма распадается на две самостоятельные гистограммы, не перекрывающие друг друга (рис.2.19, *а*).

2. Раздвиг большой ($d = 2 \div 4$), гистограмма является бимодальной, совокупности частично перекрываются (рис.2.19, *б* и рис.2.17). Однородные совокупности можно разделить либо аналитическим путем, либо используя геологическую информацию.

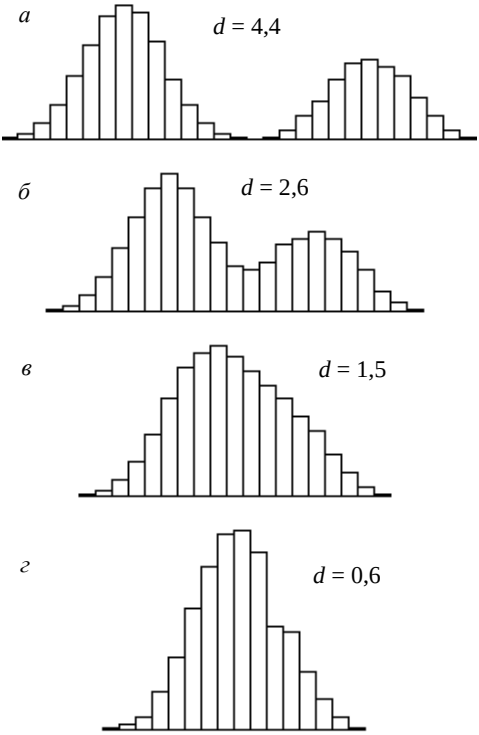

Рис. 2.19. Характер гистограмм при различном раздвиге

3. Раздвиг малый ($d = 0,7 \div 2$), гистограмма одномодальная, но имеет искаженную асимметричную форму (рис.2.19, *в*). Аналитическое разделение ее на однородные совокупности все же возможно.

4. Раздвиг незначительный ($d < 0,7$), гистограмма одномодальная (рис.2.19, *г*), разделить ее на однородные совокупности практически невозможно.

Таким образом, перед статистической обработкой данных необходимо стараться разделить неоднородную совокупность на однородные и удалить из расчетов аномальные значения.

Глава 3 ДВУХМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ И ЕЕ ПРИМЕНЕНИЕ В ГЕОЛОГИИ

3.1. ДВУХМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ

3.1.1. Система двух случайных величин и ее графическое изображение

Во многих геологических задачах изучают два взаимосвязанных свойства множества геологических объектов. Такой анализ проводится на основе двухмерной статистической модели.

Пусть имеется система из n однородных геологических объектов, у каждого из них измерены характеристики двух свойств. Результаты измерений одного свойства обозначим $x_1, x_2, \ldots, x_n$, второго свойства $y_1, y_2, \ldots, y_n$. Их можно записать в виде таблицы-матрицы (1.1), в которой число строк равно n, а число столбцов $k = 2$.

В основе двухмерной модели лежат те же гипотезы, что и в основе одномерной: а) значения $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$ носят случайный характер; б) значения первого свойства $x_1, x_2, \ldots, x_n$ не зависят между собой, значения второго свойства $y_1, y_2, \ldots, y_n$ также не зависят между собой (но могут существовать зависимости между свойствами x и y); в) совокупность измеренных свойств является однородной. Система значений $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$ называется системой двух случайных величин, двухмерной случайной величиной или случайным вектором.

Результаты измерений двухмерной случайной величины принято изображать на графике, где по оси абсцисс откладывают характеристику одного свойства, а по оси ординат – другого. Каж-

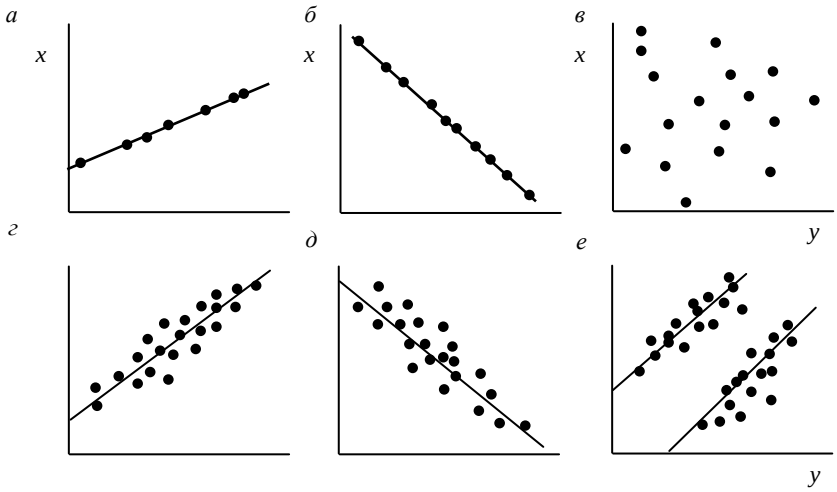

Рис.3.1. Виды зависимостей между характеристиками свойств *x* и *y*

дый геологический объект на таком графике изображают точкой, а множество объектов – облаком точек (рис.3.1). Расположение точек на графике позволяет сделать предварительные выводы о характере зависимости между свойствами. Если точки расположены вдоль линии (рис.3.1, *а, б*), то между характеристиками свойств имеется функциональная зависимость. Она может быть линейной и нелинейной. Если же точки расположены беспорядочно (рис.3.1, *в*), то зависимости между характеристиками свойств нет. Чаще всего точки располагаются в виде облака, группирующегося вдоль какой-то линии (рис.3.1, *г, д*), в этом случае наблюдается нестрогая статистическая зависимость между свойствами. Она также может быть линейной и нелинейной. Функциональные и статистические зависимости могут быть положительными, когда с возрастанием характеристики одного свойства увеличивается и другая (рис.3.1, *а, г*), но могут быть и отрицательными, когда характеристика одного свойства растет, а другого убывает (рис.3.1, *б, д*). Иногда точки могут образовать два и более изолированных или частично перекрывающихся облака (рис.3.1, *е*), что свидетельствует о двух и более однородных совокупностях, которые следует изучать раздельно.

3.1.2. Статистические характеристики системы двух случайных величин. Коэффициент корреляции

Система двух случайных величин имеет пять основных статистических характеристик: средние значения $\bar{x}$ и $\bar{y}$, дисперсии σ_x^2 и σ_y^2 и корреляционный момент (или ковариацию) K_{xy}, которые вычисляют по формулам:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \tag{3.1}$$

$$\sigma_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2; \quad \sigma_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2; \tag{3.2}$$

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{3.3}$$

Первые четыре формулы встречались ранее. Особый интерес представляет пятая формула, которая отражает взаимосвязь между случайными величинами *x* и *y*. Поскольку корреляционный момент имеет размерность, его преобразуют в безразмерную величину по формуле

$$r = \frac{K_{xy}}{\sigma_x \sigma_y}. \tag{3.4}$$

Величина *r* играет чрезвычайно большую роль в статистических исследованиях и называется *коэффициентом корреляции*. Его значения заключены в интервале между +1 и –1. Если коэффициент корреляции равен нулю, то линейная связь между случайными величинами отсутствует (рис.3.1, *в*). При *r* = 1 связь функциональная положительная (см. рис.3.1, *а*). При *r* = –1 связь функциональная отрицательная (см. рис.3.1, *б*). В реальных условиях коэффициент корреляции не бывает равен единице (или минус единице) и характеризует степень статистической связи между свойствами *x* и *y*. Чем ближе по абсолютной величине *r* к единице, тем сильнее связь между свой-

ствами; она может быть положительной ($r > 0$) и отрицательной ($r < 0$). Таким образом, коэффициент корреляции является мерой линейной зависимости между двумя величинами. Для оценки нелинейных зависимостей он непригоден.

На вычисленную величину $r_в$ заметно влияет случайная погрешность измерений исходных данных, уменьшая истинное значение коэффициента корреляции r:

$$r_в = r : \sqrt{\left(1+\frac{\sigma_1^2}{\sigma_x^2}\right)\left(1+\frac{\sigma_2^2}{\sigma_y^2}\right)}, \qquad (3.5)$$

где σ_1^2 и σ_2^2 – дисперсии случайной погрешности измерений величин x и y соответственно.

Влияние погрешности может оказаться настолько значительным, что зависимость между случайными величинами не будет выявлена.

Статистическая линейная связь между характеристиками двух свойств считается доказанной, если критерий t будет больше предельного $t_{доп}$. Коэффициент корреляции, при котором связь считается доказанной, называется *значимым коэффициентом корреляции*. Для установления значимости используется критерий t, основанный на распределении Стьюдента с числом степеней свободы $k = n - 2$:

$$t = \frac{|r|}{S_r} \text{ при } S_r = \sqrt{\frac{1-r^2}{n-2}}, \qquad (3.6)$$

где S_r – оценка среднеквадратичного отклонения коэффициента корреляции.

Если критерий t будет больше допустимого $t_{доп}$ при заданной вероятности β (см. табл.2.10), то связь считается доказанной. Имеет смысл принять вероятность $\beta = 0{,}0027$, что соответствует правилу «трех сигм».

При большом значении n можно пользоваться более простым критерием, основанным на нормальном законе распределения:

$$t = \frac{|r|}{\sigma_r} \text{ при } \sigma_r = \sqrt{\frac{1-r^2}{n}}. \quad (3.7)$$

Если $t > 3$ (что соответствует вероятности $\beta = 0{,}0027$), то связь считается доказанной.

Еще один критерий предложен Фишером:

$$t = \frac{|z|}{\sigma_z} \text{ при } \sigma_z = \frac{1}{\sqrt{n-3}}, \quad (3.8)$$

где z – новая переменная, полученная преобразованием коэффициента корреляции через гиперболический арктангенс,

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{arcth}(r). \quad (3.9)$$

И здесь для доказательства связи необходимо выполнение условия $t > 3$.

Из соотношения (3.6) выводится формула значимого коэффициента корреляции

$$r_{зн} = \frac{t_{доп}}{\sqrt{t_{доп}^2 + n + 2}}. \quad (3.10)$$

Так как $t_{доп}$ зависит от числа наблюдений (точнее, от числа степеней свободы $k = n - 2$), то и значимый коэффициент корреляции зависит от числа наблюдений. При увеличении числа наблюдений, как следует из соотношения (3.7), формула (3.10) упрощается:

$$r_{зн} = \frac{t_{доп}}{\sqrt{t_{доп}^2 + n}}. \quad (3.11)$$

Обычно принимается значение $t_{доп} = 3$.

⏩ **Пример 3.1.** Известны содержания общего и магнетитового железа в руде. Требуется рассчитать коэффициент корреляции между этими величинами (табл.3.1).

Таблица 3.1

Расчет коэффициента корреляции

Номер пробы n	Содержание железа, %		Отклонения и их произведения				
	общего x	магнетитового y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	52,0	45,7	14,9	16,6	222,01	275,56	247,34
2	49,4	45,4	12,3	16,3	151,29	265,69	200,49
3	34,5	28,4	-2,6	-0,7	6,76	0,49	1,82
4	41,5	36,6	4,8	7,5	232,04	56,25	36,00
5	36,5	22,1	-0,6	-7,0	0,36	49,00	4,20
6	22,7	10,9	-14,4	-18,2	207,36	331,24	282,08
7	42,3	27,5	5,2	-1,6	27,04	2,56	-8,32
8	20,0	10,3	-17,1	-18,8	292,41	353,44	321,48
9	23,9	17,3	-13,2	-11,8	174,24	139,24	155,76
10	23,8	16,0	-13,3	-13,1	176,89	171,61	174,23
11	33,2	23,8	-0,9	-5,3	15,21	28,09	20,67
12	61,8	55,8	24,7	26,7	610,09	712,89	659,49
13	63,7	57,3	26,6	28,2	707,56	795,24	750,12
14	22,1	15,2	-15,0	-13,9	225,00	193,21	208,50
15	50,0	45,7	12,9	16,6	166,41	275,56	214,14
16	43,4	35,4	6,3	6,3	39,69	39,69	39,69
17	37,0	29,6	-0,1	0,5	0,01	0,25	-0,05
18	28,6	20,7	-8,5	-8,4	72,25	70,56	71,40
19	23,5	13,4	-13,6	-15,7	184,96	246,49	213,52
20	32,0	24,7	-5,1	-4,4	26,01	19,36	22,44
Сумма	742,3	581,8	0,3	-0,2	3328,59	4026,42	3595,00
Среднее	37,1	29,1	-	-	166,43	201,32	179,75
Характеристики	$\bar{x}$	$\bar{y}$	-	-	σ_x^2	σ_y^2	K_{xy}

По данным таблицы 3.1 имеем: $\bar{x} = 37{,}1$; $\bar{y} = 29{,}1$; $\sigma_x^2 = 166{,}43$; $\sigma_y^2 = 201{,}32$; $\sigma_x = 12{,}90$; $\sigma_y = 14{,}19$; $K_{xy} = 179{,}75$; $r = 179{,}75/(12{,}90 \cdot 14{,}19) = 0{,}982$. Вычисленный коэффициент корреляции $r = 0{,}982$ близок к единице, следовательно, связь между свой-

ствами сильная и положительная. Чтобы убедиться в реальности связи, вычислим критерий Стьюдента по формулам (3.6):

$$S_r = \sqrt{\frac{1-0{,}982^2}{20-2}} = 0{,}0445; \quad t = \frac{0{,}982}{0{,}00445} = 22{,}1.$$

Тот же критерий на основе нормального закона распределения:

$$\sigma_r = \sqrt{\frac{1-0{,}982^2}{20}} = 0{,}0422; \quad t = \frac{0{,}982}{0{,}0422} = 23{,}3.$$

В обоих случаях критерий t значительно больше трех, поэтому линейная связь между содержаниями железа общего и магнетитового доказана надежно. ◀◀

3.1.3. Уравнение линейной регрессии

Если между величинами x и y установлена линейная статистическая зависимость, то представляет интерес найти ее выражение в виде уравнения прямой линии $y = ax + b$ (где a и b – коэффициенты). Такое уравнение называется *уравнением регрессии*. Если величина x неслучайная, то существует одно уравнение регрессии. Если обе величины (x и y) случайные, то имеется два уравнения регрессии и можно вычислять зависимости как y от x, так и x от y. Расчет уравнения сводится к определению наиболее вероятного значения y, когда известно значение x. Опуская вывод, запишем уравнение линейной зависимости через статистические характеристики:

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}). \qquad (3.12)$$

Аналогичный вид имеет второе уравнение зависимости x от y:

$$x = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y}). \qquad (3.13)$$

Эти уравнения пересекаются в точке средних значений $\bar{x}$ и $\bar{y}$. В уравнения входят пять статистических характеристик, рассмотренных в предыдущем подразделе.

Как указывалось, дисперсия случайной величины является характеристикой ее рассеяния около математического ожидания или среднего значения. Уравнение регрессии (3.12) позволяет определить еще одну *остаточную дисперсию* σ_δ, которая характеризует рассеяние значений случайной величины около линии регрессии:

$$\sigma_\delta^2 = \frac{1}{n}\sum_{i=1}^{n}\delta_i^2, \qquad (3.14)$$

где δ_i – отклонения значений случайной величины y от линии регрессии.

Дисперсии σ_δ^2 и σ_y^2 связаны между собой соотношением

$$\sigma_\delta^2 = \sigma_y^2(1-r^2). \qquad (3.15)$$

Разность между ними также является дисперсией, учтенной (поглощенной) уравнением регрессии. Она называется *дисперсией тренда* $\sigma_{\text{тр}}^2$. В некоторых публикациях ее называют *дисперсией закономерной изменчивости*, противопоставляя случайной остаточной дисперсии. Между тремя дисперсиями существует соотношение

$$\sigma_y^2 = \sigma_{\text{тр}}^2 + \sigma_\delta^2, \qquad (3.16)$$

которое можно рассматривать как разложение дисперсии σ_y^2 на две составляющие – закономерную и случайную. Если принять дисперсию σ_y^2 за 100 %, то дисперсии тренда и остаточную можно выразить в процентах от нее.

Уравнение линейной регрессии позволяет решать несколько практических задач. Первое назначение уравнения описательное, потому что часто важен сам факт линейной зависимости и ее аналитическое выражение. Но наибольшая эффективность уравнения заключается в возможности прогнозирования значения одной случайной величины, если известно значение другой. Поскольку зависимость носит статистический характер, прогнозирование по уравнению (3.12) будет сопровождаться погрешностью $t\sigma_\delta$ или, учитывая формулу (3.15), погрешностью $t\sigma_y\sqrt{1-r^2}$, где t – коэффициент ве-

роятности. Чем больше коэффициент корреляции по абсолютной величине, тем меньше погрешность прогнозирования. Для надежного прогнозирования необходимо использовать лишь такие зависимости, у которых коэффициент корреляции больше 0,87.

⏩ **Пример 3.2.** По условиям примера 3.1 необходимо рассчитать уравнение зависимости содержания железа магнетитового y от содержания железа общего x в руде.

По данным табл.3.1

$$y = 29{,}1 + 0{,}982\frac{14{,}19}{12{,}90}(x - 37{,}1)$$

или после раскрытия скобок $y = 1{,}080x - 11{,}0$. При $t = 2$ погрешность прогнозирования по уравнению $2 \cdot 14{,}19\sqrt{1 - 0{,}982^2} = 5{,}4$. Поэтому можно записать $y = 1{,}080x - 11{,}0 \pm 5{,}4$.

Из табл.3.1 имеем дисперсию $\sigma_y^2 = 201{,}32$; остаточную дисперсию $\sigma_\delta^2 = 201{,}32(1 - 0{,}982^2) = 7{,}18$; дисперсию тренда $\sigma_{тр}^2 = 201{,}32 - 7{,}18 = 194{,}14$. Приняв σ_y^2 за 100 %, найдем, что дисперсия тренда составит 96,4 %, а остаточная дисперсия отклонений равна 3,6 % от общей дисперсии.

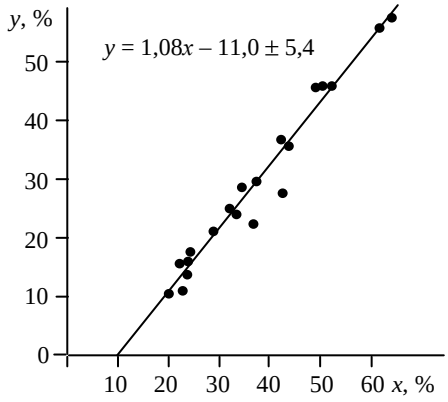

Рис.3.2. Зависимость содержания магнетитового железа x от содержания общего железа y

Линию полученного уравнения можно нанести на график (рис.3.2). Она пересечет ось абсцисс при значении $x = 11{,}0/1{,}080 = 10{,}2$ %, что указывает на вероятное среднее содержание железа в немагнитных минералах руды. В качестве второй точки для проведения линии регрессии можно использовать средние значения $\bar{x} = 37{,}1$ и $\bar{y} = 29{,}1$.

Отметим, что существует и второе уравнение зависимости x от y, оно имеет вид

$$x = 37{,}1 + 0{,}982\frac{12{,}90}{14{,}19}(y - 29{,}1)$$

или $x = 0{,}893y + 11{,}1$, его погрешность 4,9. Линии обоих уравнений пересекаются в точке средних значений $\bar{x}$ и $\bar{y}$. ◀◀

3.1.4. Двухмерное нормальное распределение. Эллипс рассеяния

Облако точек на рис.3.1, как и во многих других случаях, в первом приближении имеет эллипсовидную форму. В ряде задач нужно знать параметры эллипса, охватывающего облако, и построить эллипс на чертеже.

Идеальный эллипс возникает в том случае, когда система двух случайных величин и каждая из них в отдельности подчиняются нормальному закону распределения. Но и при заметных отклонениях от него конфигурация облака может быть охарактеризована эллипсом рассеяния.

Двухмерное нормальное распределение системы двух случайных величин описывается формулой плотности вероятности

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} - \frac{2r(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2}\right]. \quad (3.17)$$

В формулу входит пять статистических характеристик, рассмотренных выше. Если спроектировать облако точек на оси Ox и Oy и построить гистограммы частот величин x и y, то каждая из них подчиняется нормальному закону (рис.3.3):

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}}; \quad f(y) = \frac{1}{\sigma_y\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(y-\bar{y})^2}{2\sigma_y^2}}.$$

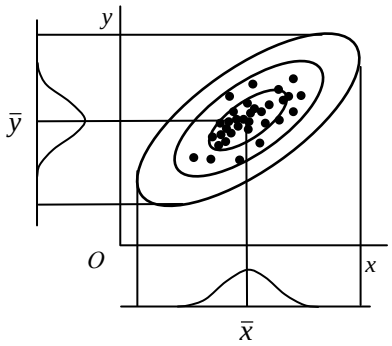

Рис.3.3.Схема эллипса рассеяния двухмерного нормального распределения и проекции плотности вероятности на оси *Ox* и *Oy*

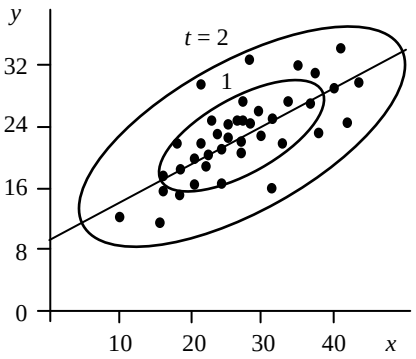

Рис.3.4. Подобные эллипсы рассеяния

Облако точек заключено внутри эллипса, выраженного уравнением

$$\frac{(x-\bar{x})^2}{\sigma_x^2} - \frac{2r(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} = t^2, \qquad (3.18)$$

где t – коэффициент вероятности.

Если t будет принимать другие значения, будут построены подобные эллипсы иного размера (рис.3.4).

В центре эллипса точки расположены гуще, к краям их плотность убывает. Вероятность попадания точек в эллипс при нормальном распределении с параметром (квантилью) t описывается формулой

$$p = 1 - \mathrm{e}^{-\frac{t}{\sqrt{1(1-r^2)}}}. \qquad (3.19)$$

Для построения эллипса необходимо знать положение его центра, размеры осей (полуосей) и их ориентировку по отношению к осям координат.

Центр эллипса имеет координаты $\bar{x}$ и $\bar{y}$. Эллипс характеризуется размером, формой и ориентировкой осей на плоскости. Размер эллипса возрастает при увеличении рассеяния точек, т.е. при

возрастании дисперсий σ_x^2 и σ_y^2. Форма эллипса зависит в основном от коэффициента корреляции r. Чем ближе он по модулю к единице, тем более узким и вытянутым оказывается эллипс. В пределе, при $r = 1$, эллипс вырождается в отрезок прямой линии. Ориентировка эллипса характеризуется углом поворота его осей по отношению к системе координат. Угол можно найти из уравнения

$$\operatorname{tg} 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{3.20}$$

Его решение дает два угла α_1 и α_2, отличающихся друг от друга на 90°. Чтобы найти полуоси эллипса, начало координат переносят в центр эллипса, в точку $(\bar{x}, \bar{y})$ и поворачивают координатные оси на угол α_1 или α_2. Обозначим новые оси координат u и v, тогда уравнение эллипса (3.18) приобретает канонический вид:

$$\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2} = t^2, \tag{3.21}$$

откуда следует, что полуоси эллипса равны $t\sigma_u$ и $t\sigma_v$.

Дисперсии разброса точек σ_u^2 и σ_v^2 в новой системе координат связаны с дисперсиями σ_x^2 и σ_y^2 соотношениями:

$$\sigma_u^2 = \sigma_x^2 \cos^2\alpha + r\sigma_x\sigma_y \sin 2\alpha + \sigma_y^2 \sin^2\alpha;$$

$$\sigma_v^2 = \sigma_x^2 \sin^2\alpha - r\sigma_x\sigma_y \sin 2\alpha + \sigma_y^2 \cos^2\alpha.$$

Сумма дисперсий при переносе и повороте координат не меняется. Она зависит от взаимного расположения точек в облаке и является инвариантом:

$$\sigma_u^2 + \sigma_v^2 = \sigma_x^2 + \sigma_y^2 = \text{const}. \tag{3.22}$$

Таким образом, чтобы построить эллипс рассеяния, достаточно знать координаты его центра $(\bar{x}, \bar{y})$, угол поворота осей α_1 или α_2 и длину полуосей $t\sigma_u$ и $t\sigma_v$.

3.1.5. Нелинейная регрессия. Метод наименьших квадратов

Зависимости между свойствами могут быть не только линейными, но и более сложными – нелинейными и многофакторными. Для обработки любых зависимостей существует эффективный *метод наименьших квадратов*. Суть метода состоит в том, что изучаемая зависимость аппроксимируется таким алгебраическим выражением (трендом), который дает наименьшее расхождение с наблюдаемыми значениями.

Пусть значения величины y нелинейно зависят от значений величины x (точки на рис.3.5). Нужно подобрать такую функцию $f(x)$, в которой отклонения между фактическими y_i и расчетными (теоретическими) $y_т = f(x)$ значениями будут наименьшими. Отклонения $\delta_i = y_i - y_т$ могут быть положительными и отрицательными. Главный принцип метода заключается в требовании, чтобы сумма квадратов всех отклонений от линии зависимости была минимальной:

$$\sum_{i=1}^{n} \delta_i^2 \to \min. \qquad (3.23)$$

Вид аппроксимирующей функции $f(x)$ должен быть задан либо на основании теоретических соображений (например, гиперболическая зависимость плотности руды от ее состава в примере 1.3), либо путем эмпирического подбора. Например, в качестве функции $f(x)$ могут быть использованы полином порядка p: $f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_px^p$; синусоида $f(x) = a\sin(bx + c)$; показательная функция $f(x) = ae^{bx}$ и др. В каждой функции присутствуют постоянные коэффициенты a, b, c (их число зависит от вида функции), значения которых заранее не известны и которые определяют положение кривой на графике (рис.3.5). Следовательно, и сумма квадратов отклонений также зависит от значений коэффициентов, т.е. является их функцией:

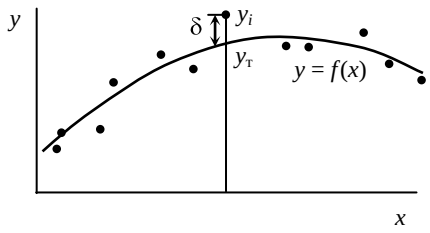

Рис.3.5 Схема, иллюстрирующая отклонения точек от заданной кривой $y = f(x)$

$$\sum_{i=1}^{n} \delta^2 = \phi(a, b, c \ldots).$$

Чтобы найти минимум этой функции, нужно взять частные производные по неизвестным коэффициентам и приравнять их нулю:

$$\frac{\partial \phi}{\partial a} = 0; \quad \frac{\partial \phi}{\partial b} = 0; \quad \frac{\partial \phi}{\partial c} = 0. \tag{3.24}$$

В результате будет получена система уравнений, в которой число уравнений равно числу неизвестных. Решая эту систему, найдем искомые коэффициенты *a, b, c* …

Когда коэффициенты в функции $f(x)$ определены, можно найти расчетные значения $y_т = f(x)$ для каждого x_i и сравнить их с фактическими y_i, т.е. найти отклонения $\delta_i = y_i - y_т$. Далее вычисляют дисперсии отклонений:

$$\sigma_\delta^2 = \frac{1}{n} \sum_{i=1}^{n} \delta_i^2 \tag{3.25}$$

и, наконец, определяют *корреляционное отношение*:

$$\eta = \sqrt{1 - \sigma_\delta^2 / \sigma_y^2}, \tag{3.26}$$

которое заключено в интервале от нуля до единицы ($0 \le \eta \le 1$) и характеризует степень нелинейной зависимости между величинами *x* и *y*. Чем ближе η к единице, тем сильнее зависимость. При $\eta = 0$ связь отсутствует.

Зная дисперсию исходных данных σ_y^2 и дисперсию случайных отклонений σ_δ^2, можно по их разности найти еще одну дисперсию $\sigma_{зак}^2 = \sigma_y^2 - \sigma_\delta^2$, которая характеризует изменчивость расчетных значений $y_т$ и может быть названа *закономерной*. Приняв общую дисперсию за 100 %, можно найти соотношение между σ_δ^2 и $\sigma_{зак}^2$ в процентах.

Рассмотренная схема обработки данных применима к исследованию линейных и нелинейных, однофакторных и многофактор-

ных зависимостей. В частном случае простой линейной зависимости $y = ax + b$ использование метода наименьших квадратов дает уравнение регрессии (3.12), а корреляционное отношение по абсолютной величине совпадет с коэффициентом корреляции.

3.1.6. Применение метода наименьших квадратов к параболической зависимости

Имеется нелинейная зависимость (рис.3.6). Требуется рассчитать нелинейную параболическую зависимость по методу наименьших квадратов. Уравнение параболы имеет вид

$$y = ax^2 + bx + c. \qquad (3.27)$$

Следовательно, для каждой точки графика справедливо соотношение (см. рис.3.5)

$$y_i = ax_i^2 + bx_i + c + \delta_i \cdot$$

Из этого выражения найдем отклонения δ_i и сумму квадратов отклонений, которая является функцией ψ от неизвестных коэффициентов a, b, c:

$$\psi(a, b, c) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} [y_i - (ax_i^2 + bx + c)]^2.$$

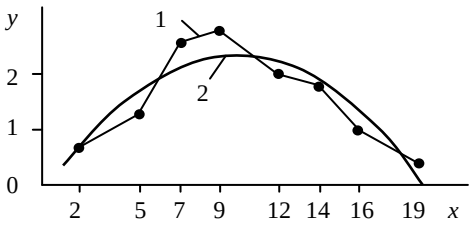

Рис.3.6. Аппроксимация изменения мощности рудного тела параболической зависимостью
1 – фактические данные; 2 – аппроксимирующая парабола $y = -0{,}027x^2 + 0{,}532x - 0{,}243$

Чтобы отыскать минимум функции $\psi(a, b, c)$, необходимо найти частные производные от функции по неизвестным a, b, c и приравнять производные нулю:

$$\frac{\partial \varphi}{\partial a} =$$

$$= -2\sum_{i=2}^{n} [y_i - (ax_i^2 + bx + c)]x_i^2 = 0;$$

$$\frac{\partial \varphi}{\partial b} = -2\sum_{i=2}^{n}[y_i - (ax_i^2 + bx + c)]x_i = 0;$$

$$\frac{\partial \varphi}{\partial c} = -2\sum_{i=2}^{n}[y_i - (ax_i^2 + bx + c)] = 0.$$

После раскрытия скобок и преобразования получим систему трех уравнений с тремя неизвестными

$$\begin{aligned} a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 &= \sum x_i^2 y_i; \\ a\sum x_i^3 + b\sum x_i^2 + c\sum x_i &= \sum x_i y_i; \\ a\sum x_i^2 + b\sum x_i + c\sum 1 &= \sum y_i. \end{aligned} \qquad (3.28)$$

Заметим, что $\sum_{i=1}^{n} 1 = n.$ Для удобства последующей записи введем смешанные начальные моменты:

$$m_{kl} = \frac{1}{n}\sum_{i=1}^{n} x_i^k y_i^l.$$

Разделим левые и правые части всех уравнений системы (3.28) на n и запишем систему через смешанные начальные моменты:

$$\begin{aligned} am_{40} + bm_{30} + cm_{20} &= m_{21}; \\ am_{30} + bm_{20} + cm_{10} &= m_{21}; \\ am_{20} + bm_{10} + c &= m_{01}. \end{aligned} \qquad (3.29)$$

Для того чтобы найти коэффициенты a, b, c в уравнении параболы (3.27), нужно вычислить все моменты, входящие в систему (3.29), и решить ее. Система уравнений (3.29) линейна относительно неизвестных a, b, c, что существенно облегчает расчеты. Нередко встречаются такие зависимости (например, гиперболические), которые приводят к сложной нелинейной системе, которую нельзя решить алгебраическим путем. Подобные системы решают методом последовательных приближений.

Пример 3.3. По простиранию рудного тела от произвольной точки отсчета на расстоянии x_i от нее измерена мощность y_i (рис.3.6, табл.3.2). Требуется рассчитать параболическую зависимость мощности линзообразного рудного тела.

Порядок расчета начальных моментов приведен в табл.3.2, последняя строка которой содержит данные, необходимые для составления системы уравнений (3.29):

$$33076a + 2079b + 139{,}5c = 178{,}8;$$

$$2079a + 139{,}5b + 10{,}5c = 15{,}61;$$

$$139{,}5a + 10{,}5c + c = 1{,}575.$$

Решая систему, найдем коэффициенты $a = -0{,}0270$; $b = 0{,}532$; $c = -0{,}242$. Следовательно, уравнение аппроксимирующей параболы имеет вид

$$y_т = -0{,}0270x^2 + 0{,}532x - 0{,}242.$$

Таблица 3.2

Расчет параболической зависимости, аппроксимирующей изменение мощности рудного тела

№ п/п	Исходные данные, м		Произведения					
	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$	y_i^2
1	2	0,7	4	8	16	1,4	2,8	0,49
2	6	1,3	25	125	625	6,5	32,5	1,69
3	7	2,6	49	343	2401	18,2	127,4	6,76
4	9	2,8	81	729	6561	25,2	226,8	7,84
5	12	2,0	144	1728	20738	24,0	288,0	4,00
6	14	1,8	196	2744	38416	25,2	352,8	3,24
7	16	1,0	256	4096	65536	16,0	256,0	1,00
8	19	0,4	381	6859	130321	7,6	144,4	0,16
Сумма	84	12,6	1116	18632	204612	124,1	1430,7	25,18
Среднее	10,5	1,575	139,5	2079	33076	15,51	178,8	3,148

Сравнение фактических y_i и теоретических $y_т$ мощностей, рассчитанных по уравнению параболы, свидетельствует об удовле-

творительном их совпадении (табл.3.3). Расхождения δ между фактическими и теоретическими значениями позволяют найти дисперсию случайных отклонений $\sigma_\delta^2 = 0{,}104$.

Таблица 3.3

Сравнение фактической и расчетной (теоретической) мощности

№ п/п	Исходные данные, м		Расчетные величины		
	x_i	y_i	$y_т$, м	δ_i	δ_i^2
1	2	0,7	0,7	0,0	0,00
2	6	1,3	1,7	–0,4	0,16
3	7	2,6	2,2	0,4	0,16
4	9	2,8	2,4	0,4	0,16
5	12	2,0	2,3	–0,3	0,09
6	14	1,8	1,9	–0,1	0,01
7	16	1,0	1,4	–0,4	0,16
8	19	0,4	0,1	0,3	0,0,9
Сумма	84	12,6	12,7	–	0,83
Среднее	10,5	1,575	1,6	–	0,104

По формуле перехода от начальных моментов к центральным (2.14) найдем $m_{40} = 33076$; $m_{30} = 2079$; $m_{20} = 139{,}5$; $m_{10} = 10{,}5$; $m_{21} = 178{,}8$; $m_{11} = 15{,}51$; $m_{01} = 1{,}575$; $m_{02} = 3{,}148$; $n = 8$.

Далее вычислим дисперсию исходных значений

$$\sigma_y^2 = 3{,}148 - 1{,}575^2 = 0{,}667,$$

откуда получим дисперсию, учтенную параболической зависимостью $\sigma_{зак}^2 = 0{,}667 - 0{,}104 = 0{,}563$, и по формуле (3.26) определим корреляционное отношение:

$$\eta = \sqrt{1 - 0{,}104 / 0{,}667} = 0{,}919.$$

Корреляционное отношение близко к единице, следовательно, параболическая зависимость хорошо аппроксимирует эмпирические данные. ◀◀

3.1.7. Выбор порядка полинома при аппроксимации нелинейной зависимости

Многие нелинейные зависимости могут быть аппроксимированы полиномом:

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m, \tag{3.30}$$

где m – порядок полинома; a_0, a_1, a_2, …, a_m – коэффициенты полинома.

Задача вычислений состоит в определении коэффициентов полинома с использованием метода наименьших квадратов. Чем выше порядок полинома, тем сложнее график, но при этом усиливается влияние случайных колебаний свойства, что отрицательно сказывается на надежности аппроксимации. Поэтому существует некоторый оптимальный порядок полинома, который наилучшим образом отражает исследуемую зависимость.

Критерием выбора наилучшего порядка полинома, как и любой другой аппроксимирующей функции, является дисперсия σ_k^2 случайных отклонений фактических значений от теоретических с учетом степеней свободы k,

$$\sigma_k^2 = \frac{n}{n-k}\sigma_\delta^2 . \tag{3.31}$$

Количество степеней свободы k равно количеству постоянных коэффициентов в аппроксимирующей функции, в которой n – число наблюдений. Так, в квадратной параболе (3.27) три постоянных коэффициента, в параболе пятого порядка шесть коэффициентов, в синусоиде три коэффициента и т.д. При исследовании полинома повышают его порядок, начиная с $m = 0$, и анализируют дис-

Таблица 3.4

Дисперсии отклонений

n	k	σ_δ^2	σ_k^2
0	1	0,08346	0,9527
1	2	0,07835	0,1044
2	3	0,01745	0,0279
3	4	0,01569	0,0314
4	5	0,01039	0,0277
5	6	0,00949	0,0380

персию отклонений с учетом использованных степеней свободы $k = m + 1$. Как только остаточная дисперсия отклонений достигнет минимума, оптимальный порядок полинома получен, дальнейшее его повышение приведет к увеличению данной дисперсии. Для условий примера 3.3 наилучшая функция, аппроксимирующая исходные данные, – это полином четвертой степени ($n = 4$), что подтверждается данными табл.3.4.

3.1.8. Приведение нелинейных зависимостей к линейному виду

Система уравнений (3.28), возникающая в результате применения метода наименьших квадратов к нелинейным зависимостям, лишь в редких случаях может быть решена алгебраическим путем. Простое решение системы возникает в случае полиномиальной зависимости. Система уравнений (3.28) для полиномов всегда является линейной. Поэтому по возможности стараются привести сложные для расчета зависимости к линейному или полиномиальному виду.

Например, показательная функция $y = ae^{bx}$ может быть приведена к линейному виду путем логарифмирования $\ln y = \ln a + bx$ и замены переменной $z = \ln y$, что приведет к линейному уравнению регрессии $z = \ln a + bx$. Здесь неизвестными являются коэффициенты $\ln a$ и b. Существенно то, что отклонения δ рассчитываются не от исходных значений y, а от их логарифмов, что не одно и то же.

Аналогично приводится к линейной логарифмическая функция $y = a + b\ln x$ путем замены переменной $z = \ln x$, что дает уравнение $y = a + bz$.

Гиперболическая функция $y = a/(1 + bx)$ приводится к общему знаменателю $y + bxy = a$, а потом делается замена $z = xy$. Получим линейную зависимость $y + bz = a$, обработка которой позволяет найти коэффициенты a и b. Подобные примеры можно продолжить и далее.

3.2. ГЕОЛОГИЧЕСКИЕ ПРИЛОЖЕНИЯ ДВУХМЕРНОЙ СТАТИСТИЧЕСКОЙ МОДЕЛИ

3.2.1. Прогнозирование свойств по уравнению регрессии

Выше отмечалось, что уравнение линейной регрессии позволяет прогнозировать одно свойство по другому, что имеет значение, если прямое измерение характеристики прогнозируемого свойства затруднено или связано с дополнительными затратами. Например, на одном из полиметаллических месторождений установлена линейная зависимость содержания кадмия от содержания цинка в руде (рис.3.7). Коэффициент корреляции между содержаниями 0,937, т.е. очень высокий. Разброс точек на рисунке обусловлен, во-первых, колебаниями состава сфалерита, во-вторых, значительной случайной погрешностью определения содержания кадмия.

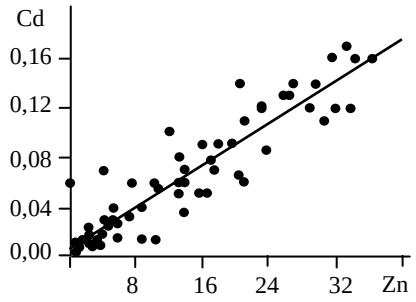

Рис.3.7. Зависимость содержания кадмия от содержания цинка, выраженная уравнением Cd = 0,0043Zn + 0,0056

Погрешность уравнения регрессии составляет 0,034 %, что ниже среднего содержания кадмия 0,058 %. Возможно, погрешность уравнения завышена из-за неизбежной случайной погрешности химического анализа (или опробования).

3.2.2. Выявление аномальных значений и однородных совокупностей

При построении графиков регрессии отдельные точки нередко далеко отходят от линии регрессии (рис.3.8). Без каких-либо расчетов можно считать, что удаленная точка соответствует аномаль-

ному значению. Если же точка аномального значения находится вблизи линии регрессии, то необходим специальный расчет. Вначале рассчитывается линия регрессии без предполагаемого аномального значения, далее находят отклонения δ точки от линии регрессии и с помощью различных критериев, рассмотренных в подразделе 2.3.3, решается вопрос об аномальности исследуемого значения. Следует отметить, что отклонения δ от линии регрессии обычно подчиняются нормальному закону, хотя исходные данные могут существенно отличаться от него. Возможен случай, когда на графике наблюдаются два облака точек, которым соответствуют различные линии регрессии (см. рис.3.1, *е*), что свидетельствует о неоднородности совокупности значений. Их нужно разделить на две самостоятельные совокупности и обрабатывать раздельно. Выделение однородных совокупностей решается геологическими методами, так как математические методы весьма сложны и в данном учебнике не рассматриваются.

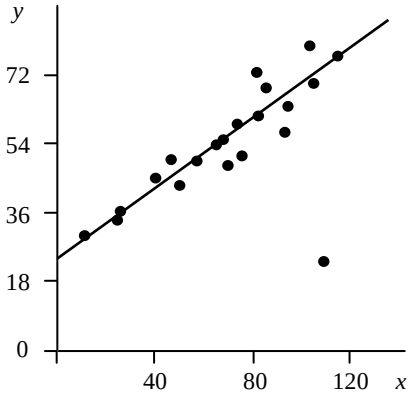

Рис.3.8. Построение линии регрессии при аномальном значении

3.2.3. Внутренний контроль химических анализов

Одним из возможных способов применения двухмерной статистической модели является *внутренний контроль* химических анализов. Однако подобная методика может быть использована также для контроля опробования, минералогического, спектрального анализа и пр.

В основе внутреннего контроля лежит условие *равноточности* основных и повторных анализов. Пробы делят на две партии и анализируют в одной и той же лаборатории, в одно и то же время и по одинаковой технологии. Первую партию называют основными пробами, вторую – контрольными. Контрольные пробы зашифрованы, так что их нельзя отличить от основных. Сравнение результатов анализов основных и контрольных проб позволяет оценить случайную погрешность анализов (ошибку воспроизводимости анализов). Вначале находят абсолютную случайную погрешность:

$$\delta_{сл} = \sqrt{\frac{\sum (x-y)^2}{2n}}, \qquad (3.32)$$

где x и y – соответственно основные и контрольные анализы; n – число контрольных проб.

Далее определяют относительную случайную погрешность, которую обычно выражают в процентах:

$$\tau_{сл} = \frac{2\delta_{сл}}{\bar{x}+\bar{y}} 100. \qquad (3.33)$$

Для относительных случайных погрешностей существуют допустимые значения, которые приводят в инструкциях по подсчету запасов для каждого вида минерального сырья. Если относительная случайная погрешность окажется больше допустимой, то подсчет запасов будет ненадежным.

В табл.3.5 приведен пример обработки данных внутреннего контроля анализов. При расчетах следует обращать внимание на грубые (аномальные) различия между основными и контрольными измерениями, которые могут быть вызваны неслучайными причинами и классифицируются как промахи. Их присутствие может существенно исказить (увеличить) случайную погрешность.

Таблица 3.5

Расчет случайной погрешности химического анализа

Номер пробы n_i	Содержание меди, %		Разность $x_i - y_i$	Квадрат разности $(x_i - y_i)^2$
	Основные пробы x_i	Контрольные пробы y_i		
1	2,74	2,70	0,04	0,0016
2	2,14	2,44	−0,30	0,0900
3	2,33	2,19	0,14	0,0196
4	2,57	2,54	0,03	0,0009
5	2,16	2,24	−0,08	0,0064
6	1,27	1,21	0,06	0,0036
7	1,00	1,23	−0,23	0,0529
8	0,95	0,59	0,36	0,1296
9	1,72	1,28	0,44	0,1936
10	2,06	1,76	0,30	0,0090
11	1,06	1,43	−0,37	0,1369
12	1,83	1,83	0,00	0,0000
13	2,13	1,81	0,32	0,1024
14	3,04	3,16	−0,12	0,0144
15	1,52	1,34	0,18	0,0324
16	1,48	1,63	−0,15	0,0225
17	0,78	0,82	−0,04	0,0016
18	0,92	0,60	0,32	0,1024
19	2,17	2,62	−0,45	0,2025
20	2,96	2,56	0,40	0,1600
21	1,45	1,79	−0,34	0,1156
22	1,82	1,83	−0,01	0,0001
23	2,51	2,29	0,22	0,0484
24	1,70	2,18	−0,48	0,2304
Сумма	44,31	44,07	0,24	1,7578
Среднее	1,85	1,84	–	–

Абсолютная случайная погрешность 0,191

Относительная случайная погрешность 10,4 %

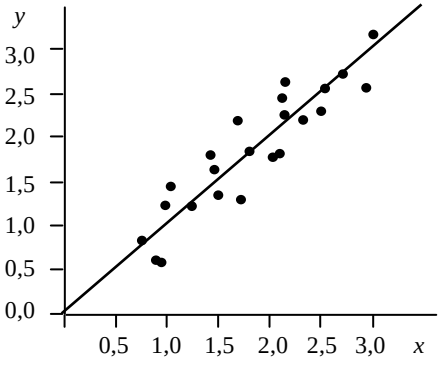

Рис.3.9. График случайных погрешностей химического анализа

Для выявления случайной погрешности рекомендуется проводить графический анализ (рис.3.9) – облако точек должно группироваться около биссектрисы угла *xOy*. Если какие-то точки сильно удалены от биссектрисы, то соответствующие им пробы классифицируют как промахи, они должны быть исключены из расчета.

3.2.4. Внешний контроль химических анализов

В геологической практике принято регулярно оценивать систематическую погрешность измерений. Наиболее часто определяется погрешность опробования или ее составная часть – погрешность химического анализа, для чего выполняется *внешний контроль* анализов. Главное требование при изучении систематических погрешностей (или систематических расхождений) – *неравноточность* основных и контрольных измерений. Для выполнения этого условия основные пробы посылают в одну лабораторию, а контрольные – в другую, где анализ выполняют, как правило, по более совершенной методике. Сравнение анализов основных и контрольных проб позволяет оценить систематическую погрешность анализов.

Обозначим x_i – данные основных проб, y_i – данные контрольных проб. Для выявления систематической погрешности применяются графический и аналитический методы. При графическом анализе проверяется расположение точек графика. При отсутствии систематической погрешности они должны располагаться вдоль биссектрисы $y = x$. Из-за наличия неизбежных случайных погрешно-

стей точки рассеиваются около биссектрисы, образуя облако. Если облако точек смещено относительно биссектрисы, то можно предполагать, что в основных (а иногда и в контрольных) пробах имеется систематическая погрешность (рис.3.10).

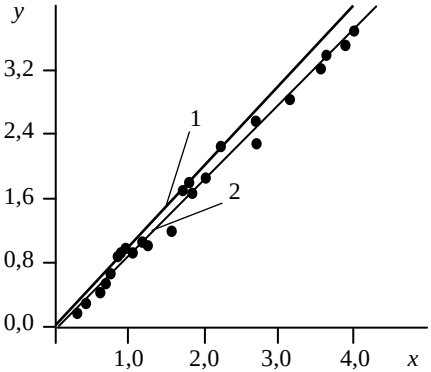

Рис.3.10. График систематической погрешности анализов
1 – биссектриса; 2 – уравнение регрессии

Для более точного доказательства систематической погрешности применяются аналитические методы. Наиболее распространенный прием основан на сравнении средних значений $\bar{x}$ и $\bar{y}$ с помощью критерия

$$t = \frac{|\bar{x} - \bar{y}|}{\sigma_{xy}} \quad \text{при} \quad \sigma_{xy} = \sqrt{\frac{\sigma_x^2 - 2r\sigma_x\sigma_y + \sigma_y^2}{n-1}}, \qquad (3.34)$$

где n – число контрольных проб.

Если критерий t окажется больше допустимого $t_{доп}$, то систематическая погрешность доказана. Допустимое значение критерия $t_{доп}$ находят на основе распределения Стьюдента при вероятности $\beta = 0{,}05$ и числе степеней свободы $k = n - 1$ (см. табл.2.10). При увеличении числа контрольных проб распределение Стьюдента приближается к нормальному и в пределе $t_{доп} = 1{,}960$.

Метод выявления систематической погрешности, основанный на сравнении средних значений $\bar{x}$ и $\bar{y}$, обладает существенным недостатком: равенство средних еще не гарантирует наличия систематических расхождений при низких и высоких содержаниях. Последние могут быть направлены в разные стороны и в среднем компенсируют друг друга.

Более рациональной является методика, основанная на оценке коэффициентов уравнения регрессии $y = ax + b$. При отсутствии систематической погрешности должны выполняться условия $a = 1$ и $b = 0$.

Таблица 3.6

Расчет систематической погрешности анализов серы

Номер пробы n	Содержание, %		
	Основные пробы x	Контрольные пробы y	Исправленные пробы $x_{испр}$
1	0,41	0,35	0,36
2	1,56	1,21	1,43
3	0,27	0,20	0,23
4	2,70	2,31	2,48
5	0,71	0,66	0,64
6	0,61	0,44	0,55
7	3,90	3,55	3,59
8	4,03	3,68	3,71
9	0,88	0,88	0,80
10	0,96	0,94	0,87
11	2,01	1,90	1,84
12	2,71	2,59	2,49
13	3,65	3,38	3,36
14	1,73	1,70	1,58
15	1,05	0,95	0,95
16	1,24	1,03	1,13
17	2,25	2,25	2,06
18	1,16	1,08	1,06
19	1,79	1,76	1,64
20	1,85	1,68	1,69
21	3,65	3,40	3,36
22	3,15	2,86	2,90
23	3,58	3,24	3,29
24	0,67	0,57	0,60
Сумма	46,42	42,60	42,61
Среднее	1,934	1,775	–

Поэтому вычисляют уравнение регрессии между содержаниями в основных и контрольных пробах и находят соответственно два критерия:

$$t_a = \frac{|a-1|}{\sigma_a} \text{ при } \sigma_a = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n-2}}$$

$$\text{и } t_b = \frac{|b|}{\sigma_b} \text{ при } \sigma_b = \sigma_a\sqrt{\sigma_x^2 + (\bar{x})^2}\,.$$

Если хотя бы один из критериев больше допустимого $t_{доп}$, то систематическая погрешность установлена. Значения $t_{доп}$ берут из табл.2.10 при вероятности $\beta = 0{,}05$ и числе степеней свободы $k = n - 2$.

Если систематическая погрешность установлена, то в основные данные могут быть введены поправки по уравнению регрессии $y = ax + b$. Подставляя в него содержания в основных пробах x_i, можно получить исправленные значения y_i, не содержащие систематической погрешности. В особо ответственных случаях контроль проводят несколько раз, чтобы убедиться в обоснованности введения поправок на систематическую погрешность.

▸▸ Пример 3.4. В 24 пробах выполнены основные и контрольные анализы на серу (табл.3.6,

рис.3.10). Требуется определить, имеется ли систематическая погрешность в основных анализах.

В результате статистической обработки рассчитаны характеристики: $\bar{x} = 1{,}934$; $\bar{y} = 1{,}775$; $\sigma_x = 1{,}185$; $\sigma_y = 1{,}100$; $r = 0{,}997$. Из них получено уравнение регрессии $y = 0{,}9253x - 0{,}0181$. На рис.3.10 видно, что линия регрессии несколько смещена относительно биссектрисы, что свидетельствует о возможной систематической погрешности.

Вычислим необходимые величины:

$$\sigma_{xy} = \sqrt{\frac{1{,}185^2 - 2 \cdot 0{,}997 \cdot 1{,}185 \cdot 1{,}100 + 1{,}100^2}{24 - 1}} = 0{,}0256;$$

$$t = \frac{|1{,}934 - 1{,}775|}{0{,}0256} = 6{,}21.$$

Из табл.2.10 при вероятности $\beta = 0{,}05$ и числе степеней свободы $k = 24 - 1 = 23$ найдем допустимую величину критерия $t_{\text{доп}} = 2{,}069$. Так как $t > t_{\text{доп}}$, то систематическая погрешность доказана.

Проверим наличие систематической погрешности по величине коэффициентов регрессии:

$$\sigma_a = \frac{1{,}775}{1{,}934}\sqrt{\frac{1 - 0{,}997^2}{24 - 2}} = 0{,}01515;$$

$$\sigma_b = 0{,}01515\sqrt{1{,}185^2 + 1{,}934^2} = 0{,}03436;$$

$$t_a = \frac{|0{,}9253 - 1|}{0{,}01515} = 4{,}93; \quad t_b = \frac{|-0{,}0181|}{0{,}03436} = 0{,}53.$$

Из табл.2.10 при вероятности $\beta = 0{,}05$ и числе степеней свободы $k = 24 - 2 = 22$ допустимое значение критерия $t_{\text{доп}} = 2{,}074$. Так как $t_a > t_{\text{доп}}$, то систематическая погрешность доказана.

Поправка на систематическую погрешность, выполненная по уравнению регрессии, дает исправленные значения содержания в основных пробах (табл.3.6). ◀◀

3.2.5. Оценка различия между геологическими объектами

Оценку сходства или различия между геологическими объектами можно производить по характеристикам как каждого отдельного свойства, так и множества свойств. Ограничимся оценкой различия по одному свойству.

Пусть имеются два геологических объекта, в каждом из которых имеется несколько измерений характеристик одного свойства. Средние значения $\bar{x}$ и $\bar{y}$, дисперсии σ_1^2 и σ_2^2, число измерений n_1 и n_2. Решение о различии объектов принимается с помощью критерия t с использованием распределения Стьюдента при вероятности $\beta = 0{,}05$ и числе степеней свободы $k = n_1 + n_2 - 2$:

$$t = \left(|\bar{x} - \bar{y}| \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \sqrt{\frac{n_1 + n_2 - 2}{n_1 + n_2}}. \qquad (3.35)$$

Если критерий t будет больше допустимого $t_{доп}$ при заданной вероятности (см. табл.2.10), то имеются существенные различия между геологическими объектами.

При увеличении числа наблюдений распределение Стьюдента стремится к нормальному и критерий t стремится к пределу, выражаемому формулой

$$t = |\bar{x} - \bar{y}| \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \qquad (3.36)$$

Эта формула обычно и применяется на практике.

▶▶ **Пример 3.5.** Проанализированы две серии проб базальтов из различных потоков вулкана. Средние содержания кремнезема составляют соответственно 50,35 и 49,76 %, дисперсии содержаний

4,16 и 3,32, число проб 20 и 25. Нужно установить, различаются ли базальты по содержанию кремнезема.

Число степеней свободы $k = 20 + 25 - 2 = 43$. Вычислим критерий Стьюдента:

$$t = \left(|50{,}35 - 49{,}76| \Big/ \sqrt{\frac{4{,}16}{20} + \frac{3{,}32}{25}} \right) \sqrt{\frac{20 + 25 - 2}{20 + 25}} = 0{,}988.$$

Из табл.2.10 при вероятности $\beta = 0{,}05$ и числе степеней свободы 43 найдем допустимое значение критерия $t_{доп} = 2{,}01$. Так как $t < t_{доп}$, то базальты по содержанию кремнезема не различаются. ◀◀

Различия между совокупностями измерений можно оценивать не только по средним значениям, но и по другим статистическим характеристикам: по дисперсиям, асимметриям и эксцессам. Сравнение дисперсий основано на *F*-распределении (см. подраздел 2.2.6). Сравнение асимметрий и эксцессов проводится по критериям Стьюдента или нормального закона:

$$t_A = |A_1 - A_2| \Big/ \sqrt{6\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}; \quad t_E = |E_1 - E_2| \Big/ \sqrt{24\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \quad (3.37)$$

3.2.6. Оценка постоянной радиоактивного распада

Как известно, радиоактивный распад атомов происходит по экспоненциальному закону:

$$y = y_0 \, e^{-\lambda t}, \quad (3.38)$$

где y_0 – число распадов в произвольный начальный момент времени $t = 0$; λ – постоянная распада.

Если имеются измерения радиоактивности y в различные моменты времени t, то можно рассчитать постоянную λ, которая связана с периодом полураспада T соотношением $T = \ln 2/\lambda$. Период

Таблица 3.7

Результаты опытов по измерению радиоактивности препарата

№ п/п	Время, мин	Среднее время t, мин	Число распадов в минуту	
			фактическое y	расчетное $y_{расч}$
1	0-2	1	66,1	60,1
2	2-6	4	64,9	59,3
3	6-10	8	55,4	58,3
4	10-16	13	57,4	57,0
5	16-28	22	59,2	54,8
6	28-48	38	41,2	51,0
7	48-58	53	49,9	47,8
8	58-89	74	33,9	43,6
9	89-95	92	41,8	40,2
10	95-101	98	46,8	39,2
11	101-115	108	28,8	37,5
12	115-121	118	43,3	35,8
13	121-127	124	37,1	34,9

полураспада – это время, в течение которого распадается половина атомов. Эта величина является постоянной для каждого изотопа и позволяет идентифицировать его.

Пример экспериментальных данных описан в научно-популярном журнале. В начале XX в. немецкий физик Отто Ган, измеряя импульсы от препарата урана, облученного нейтронами, установил тенденцию уменьшения числа распадов с течением времени (табл.3.7, рис.3.11).

Взяв за аргумент среднее время в интервале измерения, а за функцию – число распадов атомов в минуту, можно определить постоянную распада λ и период полураспада T. Уравнение (3.38) предварительно прологарифмируем и приведем к линейному виду:

$$\ln y = \ln y_0 - \lambda t. \quad (3.39)$$

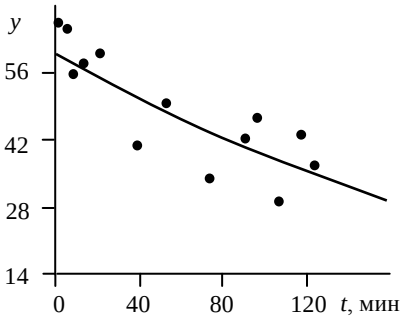

Рис.3.11. График распада атомов урана

Найдем линейную зависимость $\ln y$ от t. В результате вычислений по данным табл.3.7 получим уравнение линейной регрессии $\ln y = 4{,}1006 - 0{,}00442t$. Потенцируя его, найдем искомую зави-

симость $y = 60{,}38e^{-0{,}0044}$. Значения $y_{расч}$, вычисленные по этой формуле, приведены в табл.3.7 и по ним построена плавная кривая на рис.3.11. Заметный разброс точек относительно линии обусловлен тем, что распад атомов происходит неравномерно во времени.

Таким образом, постоянная распада $\lambda = 0{,}00442$. Период полураспада $T = \ln(2)/0{,}00442 = 157$ мин = 2 ч 37 мин. Эти данные позволили Отто Гану установить, что импульсы создаются не ураном или радием, периоды полураспада которых были уже известны, а каким-то другим химическим элементом (изотопом). Дальнейшими исследованиями было установлено, что атом урана распадается на два осколка – на атомы бария и криптона. Так эксперимент перерос в крупное научное открытие явления искусственного распада атомного ядра.

3.2.7. Зависимость плотности руды от ее состава

Плотность многих видов полезных ископаемых зависит от их состава. Так, плотность железной руды зависит от содержания в ней железа, сульфидной руды – от содержания серы и т.д. Как показано в подразделе 1.2.2, данная зависимость является гиперболической, хотя часто приближается к линейной.

Рассматриваемая зависимость позволяет решать две задачи: 1) определять плотность руды при известном ее составе, что используется для подсчета запасов; 2) определять состав руды по ее плотности, что применяется при геофизическом опробовании руд.

⏩ **Пример 3.6.** Известны плотность железной руды y и содержание в ней железа x (табл.3.8). Необходимо рассчитать зависимость между этими величинами.

Теоретическое уравнение зависимости выражается гиперболой (1.5). Преобразуем уравнение к линейному виду $a + bxy = y$, как в подразделе 3.1.8. Вычисление уравнения регрессии дает коэффициенты $a = 2{,}722$, $b = 0{,}00655$. Следовательно, уравнение зависимости имеет вид

$$y = \frac{2{,}722}{1 - 0{,}00655x}.$$

Вычисленные по уравнению расчетные значения плотности $y_{\text{расч}}$ близки к фактическим. Дисперсия отклонений фактических значений от расчетных $\sigma_{\delta}^2 = 0{,}006556$, дисперсия фактических значений $\sigma_y^2 = 0{,}2541$. Отсюда получены дисперсия тренда $\sigma_{\text{тр}}^2 = 0{,}2475$ и корреляционное отношение $\eta = 0{,}987$.

Таблица 3.8

Содержание железа в руде и плотность руды

Номер пробы n	Содержание железа x, %	Плотность, т/м^3		Номер пробы n	Содержание железа x, %	Плотность, т/м^3	
		фактическая y	расчетная $y_{\text{расч}}$			фактическая y	расчетная $y_{\text{расч}}$
1	41,33	3,77	3,73	17	34,00	3,57	3,50
2	64,00	4,71	4,69	18	40,05	3,66	3,69
3	64,78	4,79	4,73	19	62,46	4,63	4,61
4	57,58	4,25	4,37	20	20,18	3,16	3,14
5	27,63	3,34	3,32	21	44,28	3,82	3,83
6	32,14	3,59	3,45	22	26,77	3,22	3,30
7	54,38	4,07	4,23	23	39,34	3,63	3,67
8	49,66	4,14	4,03	24	43,19	3,65	3,80
9	25,13	3,36	3,26	25	50,16	3,96	4,05
10	46,06	3,84	3,90	26	26,21	3,16	3,29
11	52,83	4,11	4,16	27	38,10	3,64	3,63
12	49,98	4,20	4,05	28	27,96	3,34	3,33
13	62,36	4,65	4,60	29	62,78	4,68	4,62
14	30,58	3,49	3,40	30	20,78	3,11	3,15
15	61,82	4,54	4,57	31	24,05	3,25	3,23
16	34,34	3,58	3,51	32	38,98	3,58	3,63

Из уравнения зависимости можно извлечь дополнительную геологическую информацию. При отсутствии рудного минерала (магнетита) содержание железа в нерудных минералах близко к 10 %. Подставив это значение в формулу зависимости, получим плотность суммы нерудных минералов 2,91 т/м^3. Если же взять чистый магне-

тит, в котором за счет примесей содержание железа несколько ниже теоретического и близко к 71,5 %, то плотность магнетита составит 5,12 т/м^3. ◀◀

3.2.8. Вычисление параметров усеченного нормального распределения

В ряде случаев гистограмма искусственно ограничена (обычно слева) пределом точности анализа или кондициями. Необходимо восстановить параметры распределения по усеченной гистограмме. Такая задача может быть решена, если известен или предполагается закон распределения случайной величины. Для решения применяется разновидность метода наименьших квадратов со взвешиванием наблюдений. Суть метода состоит в том, что классы гистограммы имеют различный вес, пропорциональный частоте появления свойства в данном классе. Запись метода наименьших квадратов в данном случае имеет вид

$$\sum_{i=1}^{n} p_i \delta_i^2 \Rightarrow \min, \tag{3.40}$$

где p_i – весовые коэффициенты.

В случае нормального закона распределения характеристики усеченного распределения находят следующим образом. Уравнение кривой, аппроксимирующей гистограмму, имеет вид

$$n = \frac{Nh}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\bar{x})^2}{2\sigma^2}}, \tag{3.41}$$

где n – теоретические частоты; N – общее количество наблюдений, которое неизвестно (гистограмма обрезана); h – размер классов.

Необходимо определить характеристики полного распределения: среднее значение $\bar{x}$, дисперсию σ^2 и число наблюдений N, чтобы оценить, какая часть гистограммы отсутствует. Перед применением метода наименьших квадратов функцию (3.41) логарифмируют:

$$\ln n = \ln N + \ln\left(\frac{h}{\sqrt{2\pi}}\right) - \ln\sigma - \frac{(x-\bar{x})^2}{2\sigma^2}.$$

Отклонения для такой функции имеют вид

$$\delta_i = \ln N + \ln\left(\frac{h}{\sqrt{2\pi}}\right) - \ln\sigma - \frac{(x-\bar{x})^2}{2\sigma^2} - \ln n. \qquad (3.42)$$

В качестве весов наблюдений берется частота свойств в классах гистограммы. Чем больше n, тем весомее роль класса. Следовательно, сумма квадратов отклонений является функцией неизвестных величин $\bar{x}$, σ и N и должна быть минимальной:

$$\sum_{i=1}^{k} n_i \delta_i^2 = \psi(\bar{x}, \sigma, N) \Rightarrow \min, \qquad (3.43)$$

где k – число имеющихся классов гистограммы.

Чтобы найти минимум суммы квадратов, нужно взять частные производные от выражения (3.43) по каждой из неизвестных величин и приравнять их нулю, в результате получим линейную систему уравнений относительно неизвестных $\bar{x}$, σ и N. Опуская громоздкие промежуточные выкладки, приведем порядок вычислений. Вначале находят восемь сумм: Σn, Σnx, Σnx^2, Σnx^3, Σnx^4, $\Sigma n\ln n$, $\Sigma nx\ln n$, $\Sigma nx^2\ln n$ (индекс i опущен). Разделив каждую сумму на Σn, получим семь начальных моментов: m_{10}, m_{20}, m_{30}, m_{40}, m_{01}, m_{11} и m_{21} соответственно. Далее вычислим вспомогательные величины:

$$a_1 = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2}; \quad a_2 = \frac{m_{21} - m_{20}m_{01}}{m_{30} - m_{20}m_{10}};$$

$$b_1 = \frac{m_{40} - m_{20}^2}{m_{30} - m_{20}m_{10}}; \quad b_2 = \frac{m_{30} - m_{20}m_{10}}{m_{20} - m_{10}^2}.$$

И, наконец, искомые величины:

$$\sigma^2 = \frac{b_1 - b_2}{2(a_1 - a_2)}; \quad \bar{x} = a_1\sigma^2 + b^2/2;$$

$$\ln N = \ln \sigma\sqrt{2\pi} + m_{01} + \frac{m_{20} - 2\bar{x}m_{10} + \bar{x}^2}{2\sigma^2}.$$

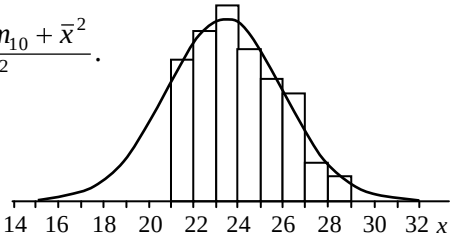

Рис.3.12. Построение кривой плотности нормального закона распределения по усеченной гистограмме

В заключение нужно учесть размер класса, умножив σ^2 на h^2, а $\bar{x}$ на h, чтобы иметь статистические характеристики в истинном масштабе. Для достаточно надежного определения статистических характеристик нужно иметь не менее 2/3 классов гистограммы, включая классы с максимальной частотой.

⏩ **Пример 3.7.** Имеется неполная гистограмма (рис.3.12, табл.3.9). Размер классов $h = 1$. Необходимо рассчитать недостающие частоты и достроить кривую нормального распределения.

В ходе расчетов последовательно получены суммы: $\Sigma n = 195$; $\Sigma nx = 696$; $\Sigma nx^2 = 3154$; $\Sigma nx^3 = 16554$; $\Sigma nx^4 = 95590$; $\Sigma n \ln n = 649{,}255$; $\Sigma nx \ln n = 2206{,}209$; $\Sigma nx^2 \ln n = 9416{,}719$.

Таблица 3.9

Расчет частот *n*

Класс *x*, %	Частоты			Класс *x*, %	Частоты		
	фактические	расчетные	округленные		фактические	расчетные	округленные
14-15	–	0,1	–	24-25	31	34,2	34
15-16	–	0,3	–	25-26	25	26,6	27
16-17	–	1,0	1	26-27	22	17,8	18
17-18	–	2,7	3	27-28	8	10,2	10
18-19	–	6,2	6	28-29	5	5,0	5
19-20	–	12,0	12	29-30	–	2,1	2
20-21	–	20,1	20	30-31	–	0,8	1
21-22	29	28,9	29	31-32	–	0,2	–
22-23	35	35,6	36	32-33	–	0,1	–
23-24	40	37,6	38				

Сумма	195	241,5	242

После деления сумм на Σn найдем моменты: $m_{10} = 3{,}5692307$; $m_{20} = 16{,}17436$; $m_{30} = 84{,}89231$; $m_{40} = 490{,}20512$; $m_{01} = 3{,}329513$; $m_{11} = 11{,}31389$; $m_{21} = 48{,}29086$. Далее вычисляем вспомогательные величины: $a_1 = -0{,}1659145$; $a_2 = -0{,}2047645$; $b_1 = 8{,}415905$; $b_2 = 7{,}907624$. Зная их, определим характеристики распределения: $\sigma_x^2 = 6{,}5416$; $\sigma_x = 2{,}5577$; $\bar{x} = 2{,}8685$; $\ln N = 5{,}4876$; $N = e^{5{,}4876} = 241{,}68$. Промежуточные расчеты полезно выполнять без округления, иначе при нахождении многочисленных разностей, имеющихся в формулах, точность расчетов заметно снизится.

Так как округленно $N = 242$, то полная гистограмма должна содержать 242 значения, фактически же имеется 195, следовательно, не хватает 47 значений. Чтобы узнать, в каких классах они должны быть, нужно найти расчетные частоты по формуле (3.41) и сравнить их с фактическими (табл.3.9). По расчетным значениям построена кривая частот нормального закона (рис.3.12), которая хорошо аппроксимирует гистограмму.

Глава 4 МНОГОМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ И ЕЕ ПРИМЕНЕНИЕ В ГЕОЛОГИИ

4.1. МНОГОМЕРНАЯ СТАТИСТИЧЕСКАЯ МОДЕЛЬ

4.1.1. Система множества случайных величин и ее статистические характеристики

Дальнейшим развитием двухмерной статистической модели служит многомерная статистическая модель, которая состоит из совокупности множества сопряженных случайных величин (называемых многомерными случайными векторами) и выражается матрицей свойств размером $k \times n$:

$$\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{vmatrix}, \qquad (4.1)$$

где n – число наблюдений; k – число свойств.

В основе многомерной статистической модели лежит гипотеза о том, что измеренные значения являются независимыми случайными величинами (векторами), т.е. строки матрицы можно располагать в любом порядке. Однако между столбцами матрицы связь

может присутствовать. В ряде задач некоторые из измерений могут быть неслучайными величинами, например заранее заданными пространственными или временны́ми координатами, что не является препятствием для статистической обработки.

Для изображения множества случайных величин используется многомерное признаковое пространство, имеющее k осей. Каждое отдельное измерение (строка матрицы) изображается в таком пространстве точкой, а их совокупность, т.е. матрица (4.1), – облаком точек.

Многомерная статистическая модель имеет различные статистические характеристики, наиболее употребительными из которых являются средние значения случайных величин $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_k$, их дисперсии σ_1^2, σ_2^2, …, σ_k^2 и среднеквадратичные отклонения σ_1, σ_2, …, σ_k. Кроме того, часто используются матрицы ковариации и коэффициентов корреляции случайных величин. Напомним, что ковариация K_{ij} – это корреляционный смешанный момент двух случайных величин i и j. Матрица ковариации имеет симметричный вид:

$$\begin{vmatrix} \sigma_1^2 & K_{12} & \cdots & K_{1k} \\ K_{21} & \sigma_2^2 & \cdots & K_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ K_{k1} & K_{k2} & \cdots & \sigma_k^2 \end{vmatrix}. \quad (4.2)$$

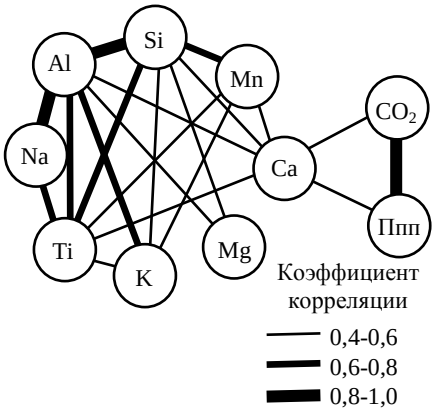

Рис.4.1. Граф связей

В ней по диагонали расположены дисперсии случайных величин, а в остальных полях – корреляционные моменты. Матрица коэффициентов корреляции между свойствами (их называют парными коэффициентами корреляции) также имеет симметричный вид:

$$\begin{vmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{vmatrix}. \qquad (4.3)$$

В матрице по диагонали находятся единицы, а в остальных полях – собственно коэффициенты корреляции. Методика расчета корреляционных моментов и коэффициентов корреляции такая же, как в двухмерной статистической модели. Данные матрицы коэффициентов корреляции могут быть представлены в виде графа связей (рис. 4.1). Для построения графа использованы результаты силикатного анализа горных пород. Чем больше коэффициент корреляции между компонентами, тем толще соединяющая их линия.

4.1.2. Множественная линейная регрессия. Коэффициент множественной корреляции

Во многих случаях возникает необходимость изучить зависимость одной случайной величины от множества других случайных величин. Многофакторная зависимость обычно выражается уравнением множественной линейной регрессии

$$y = a_1x_1 + a_2x_2 + \cdots + a_kx_k + b, \qquad (4.4)$$

где $x_1, x_2, \ldots, x_k$ – свойства; $a_1, a_2, \ldots, a_k, b$ – постоянные коэффициенты.

Коэффициенты находят методом наименьших квадратов или через значения статистических характеристик. Результат не зависит от способа вычислений. По второму способу переменные $x_1, x_2, \ldots, x_k$ нормируют по формуле (2.24), т.е. заменяют величинами:

$$t_1 = \frac{x_1 - \bar{x}_1}{\sigma_1};\ t_2 = \frac{x_2 - \bar{x}_2}{\sigma_2};\ t_k = \frac{x_k - \bar{x}_k}{\sigma_k};\ t_y = \frac{y - \bar{y}}{\sigma_y}. \qquad (4.5)$$

В результате замены уравнение (4.4) приобретет следующий вид:

$$t_y = A_1t_1 + A_2t_2 + \cdots + A_kt_k, \qquad (4.6)$$

где величины $A_1, A_2, \ldots, A_k$ – нормированные коэффициенты регрессии.

Если в формулу (4.6) подставить нормированные значения (4.5), получим выражение множественного уравнения регрессии еще в одной форме:

$$y = \bar{y} + A_1 \frac{\sigma_y}{\sigma_1}(x_1 - \bar{x}_1) + A_2 \frac{\sigma_y}{\sigma_2}(x_2 - \bar{x}_2) + \cdots + A_k \frac{\sigma_y}{\sigma_k}(x_k - \bar{x}_2). \quad (4.7)$$

Заметно сходство уравнений (3.12) и (4.7). В уравнении (4.7) находится несколько однотипных слагаемых, а вместо коэффициента корреляции r присутствуют нормированные коэффициенты регрессии $A_1, A_2, \ldots, A_k$. Значения $A_1, A_2, \ldots, A_k$ находят путем решения системы линейных уравнений, составленной из коэффициентов корреляции:

$$\begin{aligned} &A_1 + r_{12}A_2 + \cdots + r_{1k}A_k = r_{1y};\\ &r_{21}A_1 + A_2 + \cdots + r_{2k}A_k = r_{2y};\\ &\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots\\ &r_{1k}A_1 + r_{2k}A_2 + \cdots + A_k = r_{ky}. \end{aligned} \quad (4.8)$$

Сравнение фактических y и расчетных $y_{\text{рас}}$ значений по уравнению (4.7) дает отклонения δ. Рассчитав дисперсию отклонений σ_δ^2 и дисперсию исходных данных σ_y^2, можно найти *коэффициент множественной корреляции* R, который характеризует степень зависимости свойства y от множества других случайных величин $x_1, x_2, \ldots, x_k$:

$$R = \sqrt{1 - \sigma_\delta^2 / \sigma_y^2}\ . \quad (4.9)$$

Значения R колеблются от нуля до единицы. Чем ближе R к единице, тем более сильная зависимость величины y от множества величин $x_1, x_2, \ldots, x_k$. Кроме того, дисперсия отклонений позволяет рассчитать погрешность уравнения множественной регрессии (при вероятности $q = 0{,}95$ и коэффициенте вероятности $t = 2$), которая равна $t\sigma_\delta$.

⏩ **Пример 4.1.** В табл.4.1 приведены содержания меди, цинка и золота. Необходимо выполнить статистические расчеты.

Имея начальные моменты и формулы перехода от них к центральным моментам (2.14) и далее к статистическим характеристикам (2.15), найдем

$$\bar{x}_1 = m_1 = 1{,}81; \quad \bar{x}_2 = m_2 = 3{,}19; \quad \bar{y} = m_y = 18{,}2;$$

$$\sigma_1^2 = m_{11} - m_1^2 = 3{,}907 - 1{,}81^2 = 0{,}6309; \ \sigma_1 = \sqrt{0{,}6309} = 0{,}7943;$$
$$\sigma_2^2 = m_{22} - m_2^2 = 13{,}283 - 3{,}19^2 = 3{,}107; \ \sigma_2 = \sqrt{3{,}107} = 1{,}763;$$

$$\sigma_y^2 = m_{yy} - m_y^2 = 355{,}4 - 18{,}2^2 = 24{,}16; \ \sigma_y = \sqrt{24{,}16} = 4{,}915;$$

$$K_{12} = m_{12} - m_1 m_2 = 6{,}922 - 1{,}81 \cdot 3{,}19 = 1{,}148;$$
$$r_{12} = 1{,}148/(0{,}7943 \cdot 1{,}763) = 0{,}820;$$

$$K_{1y} = m_{1y} - m_1 m_y = 36{,}57 - 1{,}81 \cdot 18{,}2 = 3{,}628;$$
$$r_{1y} = 3{,}628/(0{,}7943 \cdot 4{,}915) = 0{,}929;$$

$$K_{2y} = m_{2y} - m_2 m_y = 64{,}1 - 3{,}19 \cdot 18{,}2 = 6{,}042;$$
$$r_{2y} = 6{,}042/(1{,}763 \cdot 4{,}915) = 0{,}697.$$

Таблица 4.1

Расчет статистических характеристик многофакторной зависимости

Номер пробы n_i	Содержание			Произведения					
	меди x_1, %	цинка x_2, %	золота y, г/т	x_1^2	x_2^2	y^2	x_1x_2	x_1y	x_2y
1	2,5	4,4	23	6,25	19,36	529	11,00	57,5	101,2
2	1,6	4,8	18	2,56	23,04	324	7,68	28,8	86,4
3	0,8	2,2	9	0,64	4,84	81	1,76	7,2	19,8
4	1,5	3,1	15	2,25	9,61	225	4,65	22,5	46,5
5	1,4	1,6	17	1,96	2,56	289	2,24	23,8	27,2
6	2,8	5,4	21	7,84	29,16	441	15,12	58,8	113,4
7	3,4	6,0	28	11,56	36,00	784	20,40	95,2	168,0
8	0,9	1,5	15	0,81	2,25	225	1,35	13,5	22,5
9	1,4	0,5	16	1,96	0,25	256	0,70	22,4	8,0
10	1,8	2,4	20	3,24	5,76	400	4,32	36,0	48,0

Сумма	18,1	31,9	182	39,07	132,83	3554	69,22	365,7	641,0
Среднее	1,81	3,19	18,2	3,907	13,283	355,4	6,922	36,57	64,1
Момент	m_1	m_2	m_y	m_{11}	m_{22}	m_{yy}	m_{12}	m_{1y}	m_{2y}

Систематизируем статистические характеристики и запишем систему уравнений (4.8):

$$A_1 + 0{,}820A_2 = 0{,}929;$$

$$0{,}820A_1 + A_2 = 0{,}697.$$

Решая ее, найдем нормированные коэффициенты регрессии:

$$A_1 = 1{,}091;\ A_2 = -0{,}1973.$$

Теперь можно составить уравнение множественной регрессии (4.7):

$$y = 18{,}2 + 1{,}091\frac{4{,}915}{0{,}7943}(x_1 - 1{,}81) - 0{,}19$$

После раскрытия скобок уравнение приобретет вид

$$y = 6{,}75x_1 - 0{,}55x_2 + 7{,}74.$$

Из табл.4.1 известны значения x_1 и x_2. Подставив их в полученное уравнение, найдем расчетные теоретические значения $y_{рас}$. Сравнивая их с фактическими значениями y, получим отклонения δ и дисперсию σ_δ^2 (табл.4.2). Далее найдем дисперсию значений y по формуле (2.14): $\sigma_y^2 = 355{,}4 - 18{,}2^2 =$

Таблица 4.2

Сравнение фактических и расчетных содержаний золота

Номер пробы n_i	Содержание золота, г/т		Отклонение δ	Квадрат отклонения δ^2
	фактическое y	теоретическое $y_{рас}$		
1	23	22,2	0,8	0,64
2	18	15,9	2,1	4,41
3	9	11,9	–2,9	8,41
4	15	16,2	–1,2	1,44
5	17	16,3	0,7	0,49
6	21	23,7	–2,7	7,29
7	28	27,4	0,6	0,36
8	15	13,0	2,0	4,00
9	16	16,9	–0,9	0,81
10	10	8,6	1,4	1,96
Сумма	172	172,1	–0,1	29,81
Среднее	17,2	17,21	0,0	2,98

=24,16. Это позволит рассчитать коэффициент множественной корреляции:

$$R = \sqrt{1 - \frac{2{,}98}{24{,}16}} = 0{,}936.$$

Полезно проанализировать рассчитанное уравнение регрессии. Коэффициент перед содержанием меди – положительный, а перед содержанием цинка – отрицательный. Первый коэффициент на порядок больше второго, следовательно, содержание меди оказывает более сильное влияние на содержание золота, чем на содержание цинка. Можно также рассчитать погрешность уравнения регрессии: $2\sigma_\delta$ = 3,4 г/т. ◀◀

4.1.3. Отбор информативных свойств в уравнении множественной линейной регрессии

Главное назначение уравнения множественной регрессии – прогнозирование значений одной случайной величины по множеству значений других случайных величин. Однако, как показано в примере 4.1, роль последних бывает различной, поэтому возникает необходимость выявить в уравнении информативные свойства, а неинформативные свойства исключить из расчета.

Отбор информативных факторов основан на анализе дисперсии отклонений σ_δ^2 с учетом степеней свободы $m = k + 1$, где k – количество свойств в уравнении множественной регрессии. Для этого вычисляется дисперсия с учетом степеней свободы:

$$\sigma_k^2 = \sigma_\delta^2 \frac{n}{n-k-1}. \qquad (4.10)$$

При увеличении числа учитываемых случайных величин дисперсия σ_k^2 вначале убывает, потом достигает минимума и далее начинает расти. Когда дисперсия достигнет минимума, информативные свойства определены. Дальнейшее увеличение числа слу-

чайных величин приведет к росту дисперсии и внесет искусственный «шум» в результаты прогнозирования по уравнению регрессии.

Информативные свойства определяют путем перебора сочетаний случайных величин. Вначале выбирают одну величину, которая имеет самый высокий парный коэффициент с прогнозируемой величиной y. Далее находят сочетание этой величины попарно со всеми остальными величинами, и каждый раз вычисляют дисперсию с учетом степеней свободы. Лучшим будет такое сочетание случайных величин, при котором дисперсия σ_k^2 минимальна. Потом к двум найденным величинам добавляют третью, четвертую и т.д. до тех пор, пока дисперсия σ_k^2 продолжает убывать. Когда дисперсия σ_k^2 начнет возрастать, процесс отыскания информативных свойств прекращается.

⏩ **Пример 4.2.** Имеется 20 проб полиметаллической руды, проанализированных на пять компонентов (табл.4.3). Требуется изучить влияние первых четырех компонентов на содержание серебра, выступающего в роли функции y, и выбрать среди них наиболее информативные.

По исходным данным табл.4.3 вычислим статистические характеристики (табл.4.3 и 4.4). Дисперсия содержаний серебра $\sigma_y^2 = 2{,}782^2 = 7{,}740$. Содержания серебра имеют самый высокий коэффициент корреляции с содержанием свинца ($r = 0{,}811$), которое, очевидно, является наиболее информативным признаком. Дисперсия отклонений для содержаний серебра $\sigma_\delta^2 = 7{,}740(1 - 0{,}811^2) = 2{,}649$, с учетом степеней свободы дисперсия $\sigma_k^2 = 2{,}649 \cdot 20/18 = 2{,}943$.

Далее к ведущему фактору – содержанию свинца – поочередно присоединим содержания других компонентов и рассчитаем уравнения регрессии, а потом дисперсии отклонений:

содержания Pb и Cu	$\sigma_\delta^2 = 1{,}170;$	$\sigma_k^2 = 1{,}376;$
содержания Pb и Zn	$\sigma_\delta^2 = 2{,}554;$	$\sigma_k^2 = 3{,}005;$
содержания Pb и S	$\sigma_\delta^2 = 2{,}269;$	$\sigma_k^2 = 2{,}669.$

Наименьшая дисперсия σ_k^2 имеет место для содержаний Pb и Cu, следовательно, медь является вторым по силе влияния фактором.

Таким же образом изучим тройные сочетания компонентов:

содержания Pb, Cu и Zn $\sigma_\delta^2 = 1{,}167$; $\sigma_k^2 = 1{,}459$;

содержания Pb, Cu и S $\sigma_\delta^2 = 1{,}127$; $\sigma_k^2 = 1{,}409$;

Третьим по силе влияния является содержание серы.

Таблица 4,3

Химические анализы проб руды

Номер пробы	Cu, %	Zn, %	S, %	Pb, %	Ag, г/т
1	0,25	3,94	19,1	4,04	6,9
2	0,62	5,10	15,0	2,61	5,3
3	0,38	5,11	21,0	3,58	4,7
4	1,86	3,06	41,0	3,02	5,3
5	1,23	3,58	24,2	1,71	0,1
6	2,11	2,04	28,8	2,53	3,7
7	2,75	1,89	11,5	2,76	5,4
8	2,73	3,81	41,5	2,29	4,6
9	1,40	2,70	44,2	2,86	3,9
10	1,04	0,88	18,7	1,72	0,7
11	2,66	2,47	39,8	3,34	8,0
12	3,99	4,37	40,2	5,12	10,9
13	4,29	4,28	30,9	4,85	8,9
14	1,80	4,35	40,2	2,20	3,0
15	1,43	6,80	30,0	3,92	5,8
16	2,42	4,65	44,5	2,68	5,5
17	3,20	3,89	44,3	3,30	8,5
18	2,02	0,81	28,7	1,35	1,8
19	3,82	1,17	17,6	0,80	4,3
20	1,20	1,70	12,8	2,12	0,7
Среднее	2,06	3,33	29,7	2,89	4,9
σ	1,142	1,561	11,3	1,089	2,782

Таблица 4.4

Матрица коэффициентов корреляции

Компонент	Cu	Zn	S	Pb	Ag
Cu	1,000	−0,130	0,370	0,132	0,541
Zn	−0,130	1,000	0,276	0,659	0,451
S	0,370	0,276	1,000	0,217	0,394
Pb	0,132	0,659	0,217	1,000	0,811
Ag	0,541	0,451	0,394	0,811	1,000

Рассчитаем последний вариант – совместное влияние четырех компонентов:

содержания Pb, Cu, S и Zn $\sigma_{\delta}^2 = 1{,}127$; $\sigma_k^2 = 1{,}503$.

Теперь можно сопоставить итоговые данные:

содержание Pb $\sigma_k^2 = 2{,}943$;

содержания Pb и Cu $\sigma_k^2 = 1{,}376$;

содержания Pb,Cu и S $\sigma_k^2 = 1{,}409$;

содержания Pb,Cu,S и Zn $\sigma_k^2 = 1{,}503$.

Минимальная дисперсия отклонений с учетом степеней свободы достигается при учете содержаний Pb и Cu, которые и являются информативными. Содержания S и Zn являются неинформативными, их учитывать не следует. Окончательное уравнение множественной линейной регрессии имеет вид

$$Ag = 1{,}923Pb + 1{,}074Cu + 2{,}871 \pm 2{,}2.$$

Коэффициент множественной корреляции $R = 0{,}921$.

Установленная зависимость объясняется тем, что в природе серебро связано в основном со свинцом и медью, а не с цинком. ◀◀

4.2. ПРИМЕНЕНИЕ МНОГОМЕРНОЙ СТАТИСТИЧЕСКОЙ МОДЕЛИ В ГЕОЛОГИИ

4.2.1. Анализ матрицы коэффициентов корреляции

Выше был приведен пример построения графа связей по значениям коэффициента корреляции (рис.4.1), который наглядно иллюстрирует характер взаимосвязей между свойствами и несет определенную геологическую информацию. Матрица коэффициентов корреляции может быть непосредственно использована для выделения групп взаимосвязанных свойств.

⏩ **Пример 4.3.** Имеется матрица коэффициентов корреляции между свойствами, рассчитанная по нескольким сотням групповых проб, взятых из железных руд и проанализированных на 19 компонентов (табл.4.5 на вклейке). Требуется выделить геохимические группы компонентов.

При беглом взгляде на таблицу трудно выявить какие-либо закономерности. Но если переставить компоненты местами, сгруппировав вместе элементы со значимыми положительными коэффициентами корреляции, то в матрице выявляется несколько взаимосвязанных групп компонентов, имеющих геологический смысл (табл.4.6 на вклейке). Внутри групп связи положительные, а между группами связи либо отсутствуют, либо отрицательные.

Первую группу образуют железо, кобальт, сера, медь и никель. Это рудные компоненты одного (главного) этапа рудообразования. Вторая группа включает цинк, свинец и серебро. Они относятся ко второму наложенному этапу минерализации и ведут себя независимо от компонентов первой группы. Третья группа объединяет углекислоту, потери при прокаливании (ппп) и кальций. Они входят в состав известняков, которые замещены рудами. Четвертую группу слагают компоненты алюмосиликатных горных пород: кальций, марганец, кремний, алюминий, натрий, титан, калий, магний и фосфор, так как руда частично заместила силикатные породы. Кальций входит в обе группы, поскольку он присутствует и в карбонатах, и в силикатах. Таким образом, анализ матрицы коэффициентов корреляции позволяет выделить геохимические группы компо-

нентов и содержит информацию о типах горных пород, замещенных железными рудами. Геохимические группы компонентов можно изобразить в виде графа связей, как на рис.4.1.

Подобная группировка компонентов в корреляционной матрице может быть сделана во многих случаях и позволяет получать полезные геологические выводы. Нередко группы различных свойств частично перекрывают друг друга, что свидетельствует о сложности и многостадийности геологических процессов. ◀◀

4.2.2. Метод главных компонент

Одним из распространенных и эффективных способов обработки многомерных статистических данных является метод главных компонент*, суть которого заключается в линейном преобразовании исходных данных в независимые величины, несущие смысловую геологическую информацию.

Как отмечалось в подразделе 4.1.1, многомерные случайные величины изображают в многомерном признаковом пространстве облаком точек. Предполагается, что облако имеет форму, близкую к многомерному эллипсоиду. Преобразование исходных данных сводится к переносу и вращению системы координат в признаковом пространстве. Начало координат переносится в центр тяжести облака, а поворот осуществляется таким образом, чтобы оси многомерного эллипсоида совпали с осями координат. Оси эллипсоида ранжируются по длине, и та координатная ось, которая совпадает с наиболее длинной осью эллипсоида, называется первой, следующая по длине – второй и т.д. Новые координаты точек облака после переноса и вращения системы координат называются главными компонентами, которые и дали название методу.

В процессе вращения сумма дисперсий остается постоянной, т.е. является инвариантом (она зависит только от взаимного расположения точек в облаке), но происходит перераспределение дисперсий. Максимальная дисперсия оказывается сосредоточенной в первых главных компонентах, которые и несут основную геологи-

* *Компонента* – новая координата точки в признаковом пространстве после переноса и вращения системы координат.

ческую информацию. Минимальной дисперсией обладают последние компоненты. Они несут малую информацию, и ими можно пренебречь. Происходит как бы сворачивание информации в первых главных компонентах. Направляющие косинусы между осями старой и новой систем координат называются факторными нагрузками и часто имеют геологическое содержание.

Поскольку свойства могут иметь различную физическую природу, возникает необходимость приведения значений случайных величин к одному масштабу, что существенно влияет на результаты вычислений. Обычно по осям координат откладывают нормированные случайные величины, вычисленные по формуле (2.24). Единицами нормирования свойств служат среднеквадратичные отклонения.

Метод главных компонент широко распространен, но слабо освещен в литературе, поэтому подробно рассмотрим последовательность обработки исходных данных и геологическую интерпретацию результатов.

⏩ **Пример 4.4.** Имеется 20 проб магнетита, проанализированных на семь компонентов (табл.4.7). Требуется обработать данные по методу главных компонент.

По исходным данным вычислим средние значения, среднеквадратичные отклонения и составим матрицу коэффициентов корреляции между компонентами магнетита (табл.4.7 и 4.8).

Следующая операция – отыскание собственных чисел и собственных векторов матрицы коэффициентов корреляции. Решение состоит в нахождении корней алгебраического уравнения степени k (k – число свойств) путем последовательных приближений. Собственные числа – это дисперсии главных компонент.

Порядок вычисления первого собственного числа матрицы коэффициентов корреляции приведен в табл.4.9. Вначале запишем матрицу коэффициентов корреляции и найдем суммы коэффициентов по строкам. Суммы составляют вектор W, записанный справа от матрицы. Среди сумм найдем максимальную, она равна 3,236, все суммы разделим на нее и определим начальный вектор V (0,951; 0,760 …), который запишем в виде строки под матрицей.

Далее найдем новые суммы путем построчного умножения членов матрицы на начальный вектор и суммирования произведений

по строкам, что даст уточненный вектор *W* (первая итерация), записанный справа (2,627; 2,059 …). Снова отыщем максимальную сумму, равную 3,086, все суммы разделим на нее и получим второй уточненный вектор *V* (0,851; 0,667 …). Повторяя перечисленные операции, достигнем стабилизации вектора *V*. Операция закончена. Окончательные значения векторов приведены в последнем столбце и в последней строке таблицы. Максимальная сумма в векторе *W* дает первое собственное число $\lambda_1 = 2{,}934$. Это дисперсия первой главной компоненты. Естественно, что подобные расчеты выполняют на компьютере.

Таблица 4.7

Состав магнетита, %

Номер пробы	TiO_2	MnO	V_2O_5	SiO_2	Al_2O_3	MgO	CaO
1	0,25	0,17	0,31	0,23	0,30	0,06	0,06
2	0,32	0,15	0,26	0,25	0,65	0,12	0,12
3	0,30	0,13	0,25	0,12	0,52	0,06	0,02
4	0,28	0,10	0,29	0,10	0,46	0,06	0,03
5	0,33	0,08	0,27	0,24	0,42	0,10	0,04
6	0,12	0,08	0,23	0,16	0,29	0,10	0,02
7	0,33	0,22	0,25	0,20	0,65	0,12	0,02
8	0,43	0,15	0,39	0,46	0,40	0,29	0,15
9	0,36	0,17	0,28	0,50	0,70	0,06	0,04
10	0,46	0,12	0,20	0,12	0,70	0,13	0,02
11	0,39	0,12	0,27	0,06	0,67	0,12	0,02
12	0,24	0,12	0,27	0,16	0,45	0,05	0,02
13	0,45	0,07	0,22	0,16	0,63	0,04	0,02
14	0,54	0,19	0,30	0,25	0,40	0,17	0,17
15	0,35	0,23	0,22	0,17	0,25	0,06	0,02
16	0,35	0,07	0,24	0,12	0,43	0,12	0,03
17	0,10	0,06	0,19	0,16	0,36	0,09	0,02
18	0,07	0,08	0,29	0,11	0,36	0,06	0,05
19	0,32	0,15	0,25	0,15	0,60	0,10	0,04
20	0,43	0,27	0,35	0,07	0,52	0,08	0,02
$\bar{x}$	0,321	0,136	0,266	0,190	0,483	0,099	0,046
σ	0,119	0,057	0,047	0,111	0,138	0,055	0,044

Далее необходимо умножить вектор *V* или *W* на такой множитель, чтобы сумма квадратов членов вектора была равна собственному числу λ. Нетрудно определить, что к вектору *V* нужно применить множитель $a_v = \sqrt{\lambda / \sum v^2}$, а к вектору *W* множитель $a_w = \sqrt{\lambda / \sum w^2}$. В рассматриваемом примере $\sum v^2 = 4{,}207$, $a_v = 0{,}835$. Умножая все члены вектора *V* на множитель a_v, получим первый собственный вектор Φ_1 (первую факторную нагрузку),

Таблица 4.5

Матрица коэффициентов корреляции ($r_{знач}$ = 0,197)

Компонент	Si	Ti	Al	Fe	Mn	Mg	Ca	Na	K	P	Ппп	CO_2	S	Cu	Zn	Pb	Co	Ni	Ag
Si	1,000	0,779	0,902	−0,900	0,604	0,417	0,596	0,781	0,582	0,300	−0,158	−0,136	−0,612	−0,299	0,013	−0,116	−0,715	−0,209	−0,045
Ti	0,779	1,000	0,765	−0,676	0,583	0,317	0,417	0,625	0,487	0,228	−0,250	−0,226	−0,497	−0,145	0,080	−0,050	−0,576	−0,250	0,021
Al	0,902	0,765	1,000	−0,818	0,588	0,433	0,447	0,808	0,613	0,270	−0,255	−0,255	−0,662	−0,245	0,014	−0,108	−0,679	−0,255	−0,084
Fe	−0,900	−0,676	−0,818	1,000	−0,411	−0,421	−0,798	−0,692	−0,556	−0,266	−0,181	−0,196	0,584	0,418	0,082	0,146	0,710	0,242	0,036
Mn	0,604	0,583	0,588	−0,411	1,000	0,354	0,565	0,387	0,440	0,222	−0,062	−0,244	−0,419	−0,201	0,136	0,057	−0,475	−0,290	0,034
Mg	0,417	0,317	0,433	−0,421	0,354	1,000	0,151	0,142	0,242	0,150	−0,112	−0,244	−0,371	−0,185	0,030	0,013	−0,318	−0,022	−0,039
Ca	0,596	0,417	0,447	−0,798	0,565	0,151	1,000	0,356	0,344	0,174	0,488	0,544	−0,377	−0,415	−0,104	−0,127	−0,517	−0,241	−0,009
Na	0,781	0,625	0,808	−0,692	0,387	0,142	0,356	1,000	0,319	0,204	−0,175	−0,123	−0,542	−0,195	−0,027	−0,121	−0,586	−0,268	0,015
K	0,582	0,487	0,613	−0,556	0,440	0,242	0,344	0,319	1,000	0,272	−0,082	−0,078	−0,398	−0,189	−0,059	−0,091	−0,454	−0,141	−0,020
P	0,300	0,228	0,270	−0,266	0,222	0,150	0,174	0,204	0,272	1,000	−0,110	−0,117	−0,108	0,014	−0,057	−0,002	−0,123	0,072	−0,071
Ппп	−0,158	−0,250	−0,255	−0,181	−0,062	−0,112	0,488	−0,175	−0,082	−0,110	1,000	0,951	0,126	−0,267	−0,275	−0,091	0,006	−0,015	0,049
CO_2	−0,136	−0,226	−0,255	−0,196	−0,244	−0,244	0,544	−0,123	−0,078	−0,117	0,951	1,000	0,084	−0,321	−0,256	−0,099	−0,017	−0,079	0,058
S	−0,612	−0,497	−0,662	0,584	−0,419	−0,371	−0,377	−0,542	−0,398	−0,108	0,126	0,084	1,000	0,549	0,012	0,041	0,832	0,349	0,077
Cu	−0,299	−0,145	−0,245	0,418	−0,201	−0,185	−0,415	−0,195	−0,189	0,014	−0,267	−0,321	0,549	1,000	0,116	0,079	0,482	0,213	0,105
Zn	0,013	0,080	0,014	0,082	0,136	0,030	−0,104	−0,027	−0,059	−0,057	−0,257	−0,256	0,012	0,116	1,000	0,604	0,072	0,006	0,576
Pb	−0,116	−0,050	−0,108	0,146	0,057	0,013	−0,127	−0,121	−0,091	−0,002	−0,091	−0,099	0,041	0,079	0,604	1,000	0,128	0,114	0,369
Co	−0,715	−0,576	−0,679	0,710	−0,475	−0,318	−0,517	−0,586	−0,454	−0,123	0,006	−0,017	0,832	0,482	0,072	0,128	1,000	0,365	0,105
Ni	−0,209	−0,250	−0,255	0,242	−0,290	−0,022	−0,241	−0,268	−0,141	0,072	−0,015	−0,079	0,349	0,213	0,006	0,114	0,365	1,000	−0,081
Ag	−0,045	0,021	−0,084	0,036	0,034	−0,039	−0,009	0,015	−0,020	−0,071	0,049	0,058	0,077	0,105	0,576	0,369	0,105	−0,081	1,000

Таблица 4.6

Преобразованная матрица коэффициентов корреляции

Компонент	Fe	Co	S	Cu	Ni	Zn	Pb	Ag	CO_2	Ппп	Ca	Mn	Si	Al	Na	Ti	K	Mg	P
Fe	1,000	0,710	0,584	0,418	0,242	0,082	0,146	0,036	−0,196	−0,181	−0,798	−0,411	−0,900	−0,818	−0,692	−0,676	−0556	−0,421	−0,266
Co	0,710	1,000	0,832	0,482	0,365	0,072	0,128	0,105	−0,017	0,006	−0,517	−0,475	−0,715	−0,679	−0,586	−0,576	−0,454	−0,318	−0,123
S	0,584	0,832	1,000	0,549	0,349	0,012	0,041	0,077	0,084	0,126	−0,377	−0,419	−0,612	−0,662	−0,542	−0,497	−0,398	−0,371	−0,108
Cu	0,418	0,482	0,549	1,000	0,213	0,116	0,079	0105	−0,321	−0,267	−0,415	−0,201	−0,299	−0,245	−0,195	−0,145	−0,189	−0,185	0,014
Ni	0,242	0,365	0,349	0,213	1,000	0,006	0,114	−0,081	−0,079	−0,015	−0,241	−0,290	−0,209	−0,255	−0,268	−0,250	−0,141	−0,022	0,072
Zn	0,082	0,072	0,012	0,116	0,006	1,000	0,604	0576	−0,256	−0,275	−0,104	0,136	0,013	0,014	−0,027	0,080	−0,059	0,030	−0,057
Pb	0,146	0,128	0,041	0,079	0,114	0,604	1,000	0,369	−0,099	−0,091	−0,127	0,057	−0,116	−0,108	−0,121	−0,050	−0,091	0,013	−0,002
Ag	0,036	0,105	0,077	0,105	−0,081	0,576	0,369	1,000	0,058	0,049	−0,009	0,034	−0,045	−0,084	0,015	0,021	−0,020	−0,039	−0,071
CO_2	−0,196	−0,017	0,084	−0,321	−0,079	−0,256	−0,099	0,058	1,000	0,951	0,544	−0,244	−0,136	−0,255	−0,123	−0,226	−0,078	−0,224	−0,117
Ппп	−0,181	0,006	0,126	−0,267	−0,015	−0,275	−0,091	0,049	0,951	1,000	0,488	−0,062	−0,158	−0,255	−0,175	−0,250	−0,082	−0,112	−0,110
Ca	−0,798	−0,517	−0,377	−0,415	−0,241	−0,104	−0,127	−0,009	0,544	0,488	1,000	0,565	0,596	0,447	0,356	0,417	0,344	0,151	0,174
Mn	−0,411	−0,475	−0,419	−0,201	−0,290	0,136	0,057	0,034	−0,244	−0,062	0,565	1,000	0,604	0,588	0,387	0,583	0,440	0,354	0,222
Si	−0,900	−0,715	−0,612	−0,299	−0,209	0,013	−0,116	−0,045	−0,136	−0,158	0,596	0,604	1,000	0,902	0,781	0,779	0,582	0,417	0,300
Al	−0,818	−0,679	−0,662	−0,245	−0,255	0,014	−0,108	−0,084	−0,255	−0,255	0,447	0,588	0,902	1,000	0,808	0,765	0,613	0,433	0,270
Na	−0,692	−0,586	−0,542	−0,195	−0,268	−0,027	−0,121	0,015	−0,123	−0,175	0,356	0,387	0,781	0,808	1,000	0,625	0,319	0,142	0,204
Ti	−0,676	−0,576	−0,497	−0,145	−0,250	0,080	−0,050	0,021	−0,226	−0,250	0,417	0,583	0,779	0,765	0,625	1,000	0,487	0,317	0,228
K	−0,556	−0,454	−0,398	−0,189	−0,141	−0,059	−0,091	−0,020	−0,078	−0,082	0,344	0,440	0,582	0,613	0,319	0,487	1,000	0,242	0,272
Mg	−0,421	−0,318	−0,371	−0,185	−0,022	0,030	0,013	−0,039	−0,244	−0,112	0,151	0,354	0,417	0,433	0,142	0,317	0,242	1,000	0,150
P	−0,266	−0,123	−0,108	0,014	0,072	−0,057	−0,002	−0,071	−0,117	−0,110	0,174	0,222	0,300	0,270	0,204	0,228	0,272	0,150	1,000

значения которой записаны в боковике и в головке табл.4.10. Произведения данных боковика и головки табл.4.10 дают матрицу коэффициентов первого фактора в этой таблице. Так, умножив 0,620 на 0,490, получим член матрицы 0,304. Таким же способом получены остальные члены матрицы (табл.4.10).

Таблица 4.8

Матрица коэффициентов корреляции

Компонент	TiO_2	MnO	V_2O_5	SiO_2	Al_2O_3	MgO	CaO
TiO_2	1,000	0,449	0,247	0,206	0,426	0,396	0,355
MnO	0,449	1,000	0,408	0,170	0,140	0,105	0,188
V_2O_5	0,247	0,408	1,000	0,387	–0,096	0,448	0,550
SiO_2	0,206	0,170	0,387	1,000	0,063	0,439	0,553
Al_2O_3	0,426	0,140	–0,096	0,063	1,000	0,039	–0,105
MgO	0,396	0,105	0,448	0,439	0,039	1,000	0,695
CaO	0,355	0,188	0,550	0,553	–0,105	0,696	1,000

Таблица 4.9

Расчет первого собственного числа

Коэффициент корреляции							Собственное число *W*			
1,000	0,449	0,247	0,206	0,426	0,396	0,355	3,079	2,627	…	2,179
0,449	1,000	0,408	0,170	0,140	0,105	0,188	2,460	2,059	…	1,722
0,247	0,408	1,000	0,387	–0,096	0,448	0,550	2,944	2,781	…	2,563
0,206	0,170	0,387	1,000	0,063	0,439	0,553	2,818	2,553	…	2,371
0,426	0,140	–0,096	0,063	1,000	0,039	–0,105	1,467	0,865	…	0,450
0,396	0,105	0,448	0,439	0,039	1,000	0,695	3,122	2,924	…	2,736
0,355	0,188	0,550	0,553	–0,105	0,696	1,000	3,236	3,086	…	2,934
Вектор *V*										
0,951	0,760	0,910	0,871	0,453	0,965	1,000				
0,851	0,667	0,885	0,627	0,280	0,948	1,000	$\lambda_1 = 2{,}934$			
…	…	…	…	…	…	…				
0,743	0,587	0,874	0,808	0,153	0,933	1,000				

Таблица 4.10

Расчет матрицы коэффициентов первого фактора

Φ_1	0,620	0,490	0,730	0,675	0,128	0,779	0,835
0,620	0,384	0,304	0,453	0,418	0,079	0,483	0,518
0,490	0,304	0,240	0,358	0,331	0,063	0,382	0,409
0,730	0,453	0,358	0,533	0,493	0,093	0,569	0,610
0,675	0,418	0,331	0,493	0,456	0,086	0,526	0,564
0,128	0,079	0,063	0,093	0,086	0,016	0,100	0,167
0,779	0,483	0,382	0,569	0,526	0,100	0,607	0,650
0,835	0,518	0,409	0,610	0,564	0,107	0,650	0,697

Если из матрицы коэффициентов корреляции (см. табл.4.8) вычесть почленно матрицу (табл.4.10), то получим первую остаточную матрицу коэффициентов корреляции после исключения первого фактора (табл.4.11).

Таблица 4.11

Первая остаточная матрица коэффициентов корреляции

0,616	0,145	−0,206	−0,212	0,347	−0,087	−0,163
0,145	0,760	0,050	−0,161	0,077	−0,277	−0,221
−0,206	0,050	0,467	−0,106	−0,189	−0,121	−0,060
−0,212	−0,161	−0,106	0,544	−0,023	−0,087	−0,011
0,347	0,077	−0,189	−0,023	0,984	−0,061	−0,212
−0,087	−0,277	−0,121	−0,087	−0,061	0,393	0,045
−0,163	−0,221	−0,060	−0,011	−0,212	0,045	0,303

С первой остаточной матрицей повторяется итерационный процесс по образцу табл.4.9, что дает второе собственное число λ_2 и второй вектор Φ_2. Исключив влияние второго фактора по образцу табл.4.10 и 4.11, получим вторую остаточную матрицу. Процесс нахождения собственных чисел λ и векторов факторных нагрузок Φ повторяется k раз. При этом все члены последней остаточной матрицы будут равны нулю, что служит одним из способов проверки правильности вычислений.

Результаты расчета систематизируем в табл.4.12. Факторы расположим в порядке убывания собственных чисел λ. Для контроля можно убедиться, что сумма собственных чисел равна числу свойств. Чем больше собственное число, тем больше роль соответствующего фактора. Роль факторов выражают в процентах от суммы факторов (последняя строка табл.4.12).

Таблица 4.12

Факторные нагрузки

Компонент	Фактор						
	1	2	3	4	5	6	7
TiO_2	0,620	0,605	0,068	−0,267	−0,230	−0,301	−0,134
MnO	0,490	0,438	−0,676	0,105	−0,175	0,257	0,056
V_2O_5	0,730	−0,200	−0,397	0,030	0,483	−0,172	−0,081
SiO_2	0,675	−0,199	0,228	0,632	−0,189	−-0,045	−0,125
Al_2O_3	0,128	0,811	0,404	0,198	0,315	0,085	0,131
MgO	0,779	−0,189	0,334	−0,316	0,053	0,329	−0,187
CaO	0,835	−0,320	0,135	−0,121	−0,093	−0,061	0,393
λ	2,934	1,434	0,964	0,647	0,464	0,308	0,250
λ, %	41,9	20,5	13,8	9,2	6,6	4,4	3,6

Из данных табл.4.12 следует, что на долю первых трех факторов приходится 76,2 % нагрузки (информации), остальными факторами можно пренебречь. В первом по значению факторе (41,9 % нагрузки) все компоненты вектора положительные. Этот факт указывает на то, что есть какая-то геологическая причина, которая вызывает одновременное возрастание содержаний всех компонентов магнетита. Среди компонентов есть такие, которые входят в состав магнетита как изоморфные примеси (титан, ванадий, марганец, часть алюминия и магния) и как механические примеси других минералов в пробах (кремний, алюминий, магний и кальций). Одновременное вхождение в состав проб изоморфных и механических примесей можно объяснить высокой скоростью кристаллизации магнетита, в результате чего руды становятся тонкозернистыми.

В тонкозернистых рудах магнетит обычно обогащен изоморфными примесями, а при отборе проб неизбежно будут захвачены посторонние минералы.

Второй фактор (20,5 % нагрузки) имеет положительные максимальные коэффициенты у алюминия, титана и марганца, остальные коэффициенты отрицательные. Фактор, который способствует накоплению в пробах алюминия, титана и марганца, вероятнее всего, – состав замещаемых магнетитом пород. При образовании магнетита по алюмосиликатным породам в пробах обычно повышено содержание алюминия и титана, а при образовании по известнякам – содержание кальция, что и подтверждается соответствующими факторными нагрузками.

В третьем факторе (13,8 % нагрузки) большие отрицательные коэффициенты у марганца и ванадия и положительные – у остальных компонентов. Третий фактор можно объяснить либо чистотой отбора проб, либо тем, что эти пробы взяты из висячего бока рудных тел, где обычно накапливаются подвижные компоненты – ванадий и марганец.

Таким образом, можно дать следующее объяснение результатам факторного анализа: состав проб магнетита зависит от скорости кристаллизации руд, типа замещаемых пород и чистоты отбора проб. Но не исключено, что могут быть предложены и другие объяснения закономерностей изменения состава магнетита. Обычно хорошо интерпретируются те факторы, у которых собственные числа больше единицы. Это два-три ведущих фактора.

В рассматриваемом методе, как указывалось, происходит перенос и вращение системы координат. Имеет смысл определить новые координаты точек (т.е. главные компоненты), что осуществляется с помощью векторов Ф, которые представляют собою направляющие косинусы углов между осями старой и новой систем координат. Прежде чем рассчитывать новые координаты, необходимо исходные данные нормировать по формуле (2.24) (табл.4.13).

Новые координаты (главные компоненты) вычисляют по формуле

$$z = \frac{1}{\sqrt{\gamma}} \sum_{i=1}^{k} \Phi_i t_i. \qquad (4.11)$$

Таблица 4.13

Нормированный состав магнетита

Номер пробы	TiO_2 (t_1)	MnO (t_2)	V_2O_5 (t_3)	SiO_2 (t_4)	Al_2O_3 (t_5)	MgO (t_6)	CaO (t_7)
1	−0,597	0,596	0,936	0,369	−1,326	−0,709	0,318
2	−0,008	0,246	−0,128	0,550	1,210	0,382	1,682
3	−0,176	−0,105	−0,340	−0,622	0,268	−0,709	−0,591
4	−0,345	−0,632	0,511	−0,802	−0,167	−0,709	−0,364
5	0,076	−0,982	0,085	0,459	−0,457	0,018	−0,136
6	−1,689	−0,982	−0,766	−0,261	−0,399	0,018	−0,591
7	0,076	1,474	−0,340	0,090	1,210	0,382	−0,591
8	0,916	0,246	2,638	2,441	−0,601	3,473	2,364
9	0,328	0,596	0,298	2,802	1,572	−0,709	−0,136
10	1,168	−0,281	−0,404	−0,622	1,572	0,564	−0,591
11	0,580	−0,281	0,085	−1,162	1,355	0,382	−0,591
12	−0,681	−0,281	0,085	−0,261	−0,239	−0,891	−0,591
13	1,084	−1,158	−0,979	−0,261	0,341	−1,073	−0,591
14	1,840	0,947	0,723	0,550	−0,601	1,291	2,818
15	0,244	1,649	−0,979	−0,171	−1,638	−0,709	−0,591
16	0,244	−1,158	−0,553	−0,622	−0,384	0,382	−0,364
17	−1,857	−1,333	−1,617	−0,261	−0,891	−0,164	−0,591
18	−2,109	−0,982	0,511	−0,712	−0,891	−0,709	0,091
19	−0,008	0,246	−0,340	−0,351	0,848	0,018	−0,136
20	0,916	2,351	1,787	−1,072	0,268	−0,345	−0,591

Их вычисляют для каждого фактора и пробы отдельно. Например, для первого фактора имеем

$$z = \frac{1}{\sqrt{2{,}934}}(0{,}620t_1 + 0{,}490t_2 + 0{,}730t_3 + 0{,}675t_4 + 0{,}128t_5 + \\ + 0{,}779t_6 + 00{,}835t_7) = 0{,}210.$$

Совокупность значений z для всех проб магнетита, рассчитанная по формуле (4.11), приведена в табл.4.14. Дисперсия значений z по каждому фактору равна собственному числу λ, что служит еще одной проверкой правильности всех предыдущих вычислений.

Таблица 4.14

Главные компоненты z

Номер пробы	Новые координатные оси (факторы)						
	1	2	3	4	5	6	7
1	0,210	–1,169	–1,483	0,522	–0,102	–0,387	0,147
2	1,290	0,331	0,873	0,357	0,054	0,338	1,395
3	–1,093	0,487	–0,156	–0,016	0,163	–0,119	0,119
4	–0,935	–0,256	–0,341	–0,275	0,778	–0,611	0,075
5	–0,145	–0,682	0,555	0,111	–0,036	–0,606	–0,495
6	–1,724	–1,832	0,160	–0,021	–0,215	0,581	–0,303
7	0,302	1,535	–0,290	0,465	–0,006	1,210	–0,118
8	5,161	–1,875	0,665	–0,156	0,494	0,320	–0,882
9	1,229	1,085	0,534	2,860	–0,143	–0,394	–0,204
10	–0,429	1,964	1,461	–0,741	–0,303	0,359	–0,226
11	–0,319	1,389	0,540	–0,881	0,991	0,119	–0,164
12	–1,124	–0,278	–0,430	0,387	0,333	–0,273	0,011
13	–1,227	0,897	0,900	–0,182	–0,544	–1,328	–0,168
14	3,355	–0,291	–0,103	–1,157	–1,068	–0,468	0,910
15	–0,682	0,046	–1,770	–0,091	–1,911	0,322	–0,319
16	–0,771	–0,322	0,817	–0,937	–0,109	–0,240	–0,480
17	–2,293	–1,530	0,879	0,158	–0,453	0,744	0,043
18	–1,474	–1,905	–0,438	0,134	1,045	0,019	0,654
19	–0,229	0,812	0,228	–-0,031	0,204	0,403	0,278
20	0,900	1,597	–2,603	–0,505	0,830	0,013	–0,273
σ_z^2	2,934	1,434	0,964	0,647	0,464	0,308	0,250

Таблица 4.15

Средний состав групп магнетита, %

Группа	Число проб	TiO_2	MnO	V_2O_5	SiO_2	Al_2O_3	MgO	CaO
1	15	0,34	0,14	0,26	0,18	0,52	0,08	0,03
2	3	0,10	0,07	0,24	0,14	0,34	0,08	0,03
3	1	0,43	0,15	0,39	0,46	0,40	0,29	0,15
4	1	0,54	0,19	0,30	0,25	0,40	0,17	0,17

Главные компоненты (табл.4.14) позволяют изобразить проекцию облака точек на любую плоскость в новых координатах. Обычно используется проекция точек на первые две оси (рис.4.2), но можно выбрать и другие пары координатных осей. Подобные графики также несут геологическую информацию. Расположение точек на проекции позволяет выявить однородные совокупности или отдельные аномальные значения, т.е. может быть использовано для классификации объектов. На рис.4.2 видно, что магнетиты различаются по составу, соответствующие точки группируются в два облака. Кроме того, две точки удалены от облаков, что свидетельствует об аномальности двух проб магнетита. Следовательно, имеются две однородные группы магнетита и две аномальные пробы магнетита (группы состоят из отдельных проб). Чтобы определить, какие точки соответствуют различным номерам проб, следует обратиться к табл.4.14 (первые два фактора), где приведены координаты точек в новой системе координат. Аномальными являются пробы 8 и 14. Малое облако точек соответствует пробам 6, 17 и 18. В большом облаке находятся точки остальных проб. Имеет смысл рассчитать средний состав групп и сравнить их между собой (табл.4.15). ◀◀

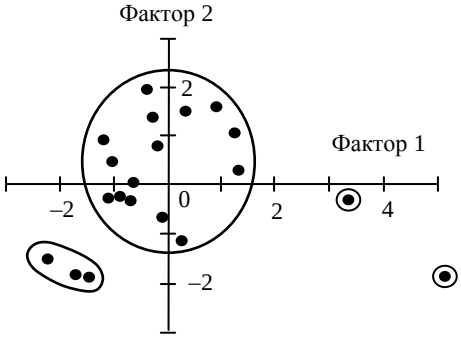

Рис.4.2. Проекция облака точек на плоскость первых двух факторов

Иногда на рисунках главных компонент имеется только одна однородная совокупность. В некоторых случаях вместо облаков точек наблюдаются ряды точек, отражающие особенности эволюции свойств объектов в пространстве или во времени (рис.4.3). На рисунки часто выносят значения главных факторных нагрузок, которые позволяют наглядно видеть направленность свойств в признаковом пространстве.

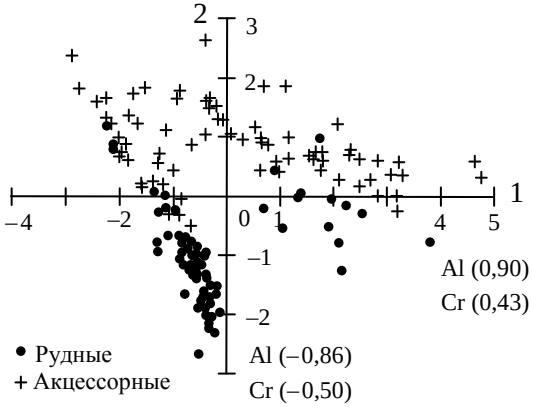

Рис.4.3. График первых двух факторов состава хромитов Кемпирсайского месторождения

На рис.4.3 отчетливо видны два эволюционных ряда. В одном ряду преобладают акцессорные хромиты, в другом – рудные хромиты. Вероятно, они различаются способом образования. Чтобы определить состав хромитов двух ветвей, нужно рассчитать состав хромита в каждой ветви. Эта процедура выполняется с помощью значений главных компонент, таких же, как в табл.4.14. Выполненные расчеты по образцу табл.4.15 показали, что верхний эволюционный ряд, где расположены преимущественно акцессорные магнетиты, направлен в сторону алюмохромитов, а нижний эволюционный ряд – в сторону магнохромитов.

Итак, метод главных компонент позволяет выделить однородные совокупности и аномальные значения, а также дать геологическую интерпретацию причин изменения свойств объектов по значениям факторных нагрузок.

4.2.3. Кластерный анализ. Дендрограмма

На основе многомерной статистической модели разработан еще один способ классификации объектов по множеству свойств – *кластерный анализ*. Существо его заключается в выделении однородных групп объектов и в установлении количественной меры сходства (различия) между объектами и группами объектов.

Пусть имеется совокупность геологических объектов, обладающих множеством свойств. Сведения о свойствах образуют матрицу (4.1). Геометрическая аналогия матрицы – облако точек в мно-

гомерном признаковом пространстве, в котором отдельные точки соответствуют единичным объектам. При кластерном анализе исследуется взаимное расположение точек. Чем ближе расположены точки, тем более сходны между собой соответствующие объекты. Задача состоит в том, чтобы объединить скопления близлежащих точек, соответствующие однородным группам объектов. Эти группы называются *кластерами*, что и дало название методу. Поставленная задача имеет много вариантов решения.

Вначале необходимо выбрать масштаб по осям координат. Если величины имеют одинаковую размерность и приблизительно один порядок, то применяют натуральный масштаб – по координатным осям откладывают исходные свойства.

Если величины различаются размерностью или порядком значений, то необходима нормализация свойств. Один из способов нормализации основан на использовании размаха значений $x_{\max} - x_{\min}$ и осуществляется по формуле

$$t = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{4.12}$$

где x – исходные; t – преобразованные (нормализованные) свойства, нормализованные значения меняются от нуля до +1.

Второй способ нормализации производится по формуле

$$t = 1 - 2\frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{4.13}$$

Здесь нормализованные значения колеблются от –1 до +1.

Третий способ нормализации выполняется по формуле (2.24), в которой в качестве масштаба используется среднеквадратичное отклонение σ. При нормальном распределении свойства нормализованные значения заключены в основном в пределах от –3 до +3, при иных законах могут выходить за эти пределы.

Когда масштаб по координатным осям задан, можно приступить к определению мер сходства (различия) между объектами по множеству свойств. Наиболее распространенная мера сходства между объектом i и объектом j – это взвешенное евклидово расстояние между точками в многомерном признаковом пространстве:

$$\rho_{ij} = \sqrt{\frac{1}{k}\sum_{l=1}^{k}(t_{il} - t_{jl})^2}\ . \qquad (4.14)$$

Напомним, что k – число свойств. Чем меньше ρ_{ij}, тем ближе расположены точки в признаковом пространстве, тем больше сходство между соответствующими объектами.

В качестве меры сходства можно применять среднеарифметическое значение абсолютных значений свойств:

$$\rho_{ij} = \frac{1}{k}\sum_{l=1}^{k}\left|t_{il} - t_{jl}\right|. \qquad (4.15)$$

Иной характер имеет угловая мера сходства, основанная на корреляционной связи между объектами:

$$\rho_{ij} = \frac{\sum_{l=1}^{k} t_{il}t_{jl}}{\sqrt{\sum_{l=1}^{k} t_{il}^2}\sqrt{\sum_{l=1}^{k} t_{jl}^2}}. \qquad (4.16)$$

Она характеризует косинус угла между двумя многомерными векторами, соединяющими начало координат с точкой i и с точкой j. Эта мера заключена в пределах от –1 до +1. Чем она ближе к +1, тем больше сходство между объектами. Чем она ближе к –1, тем больше различие между объектами. Применение данной меры оправданно, если точки находятся приблизительно на одном удалении от начала координат, так как расстояние между точками не учитывается.

Существует много других мер сходства (различия) между объектами, их обзор дан в справочнике [15].

Если имеется совокупность из n объектов, то совокупность мер сходства между всеми парами объектов составляет симметричную матрицу размером $n \times n$. Если используется формула (4.14) или (4.15), то матрица сходства имеет вид

$$\begin{vmatrix} 0 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 0 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & 0 \end{vmatrix}. \qquad (4.17)$$

По диагонали матрицы расположены нули, которые можно заменить прочерками. Чем меньше мера сходства, тем больше объекты сходны между собой.

Если используется мера сходства (4.16), то матрица сходства имеет другой вид:

$$\begin{vmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{vmatrix}. \qquad (4.18)$$

В этой матрице, чем ближе мера сходства к +1, тем объекты более сходны между собой.

В ходе кластерного анализа близкие между собой объекты объединяются в группы (кластеры). Вначале находят два объекта, наиболее близких между собой. Их свойства усредняют, и далее они выступают как один объект (кластер). Находят меры сходства полученного кластера со всеми остальными объектами. Данная операция объединения объектов продолжается, пока все они не объединятся в один объект. В итоге получается последовательность объединения объектов и мера сходства на каждом шаге объединения. Эти данные изображают на графике, который называется *дендрограммой*. По оси абсцисс откладывают номера объектов в порядке объединения, а по оси ординат – соответствующие меры сходства.

⏩ **Пример 4.5.** Имеется 14 анализов циркона на пять компонентов (табл.4.16). Необходимо провести кластерный анализ.

Таблица 4.16

Состав циркона, %

Номер пробы	SiO_2	ZrO_2	HfO_2	Fe_2O_3	TR_2O_3
1	32,74	65,27	1,29	0,12	0,23
2	32,74	64,92	1,74	0,04	0,23
3	33,03	65,30	0,50	0,18	0,23
4	32,07	66,45	1,92	0,18	0,02
5	33,65	63,65	1,63	0,15	0,23

Окончание табл.4.16

Номер пробы	SiO_2	ZrO_2	HfO_2	Fe_2O_3	TR_2O_3
6	31,34	66,57	1,52	0,18	0,17
7	31,03	67,33	0,51	0,09	0,49
8	31,08	68,36	0,49	0,18	0,05
9	30,96	67,84	0,49	0,09	0,20
10	34,53	63,74	1,59	0,30	0,10
11	34,00	63,58	1,45	0,40	0,23
12	34,40	63,58	1,61	0,16	0,27
13	32,81	63,64	1,50	0,18	0,62
14	31,34	66,57	1,52	0,27	0,17

Предварительно все свойства нормализуем по формуле (2.14). После вычисления матрицы сходства (4.17) стало ясно, что наименьшая мера сходства 0,302 имеется между пробами 5 и 12. Объединим их в кластер и усредним. Далее они выступают как одна проба. Следующая наиболее близкая пара объектов с мерой сходства 0,451 – это объекты 6 и 14. Продолжая объединение проб и кластеров далее, получим следующую последовательность объединения проб:

Мера сходства	Номера объектов														
0,302							5	12							
0,451							6	14							
0,572							1	2							
0,648							8	9							
0,665							4	6	14						
0,672							10	11							
0,730						5	12	1	2						
1,034						3	5	12	1	2					
1,057				4	6	14	3	5	12	1	2				
1,118							7	8	9						
1,206		4	6	14	3	5	12	1	2	7	8	9			
1,401		13	4	6	14	3	5	12	1	2	7	8	9		
1,415	10	11	13	4	6	14	3	5	12	1	2	7	8	9	

В последней строке все пробы объединились в один кластер. На дендрограмме, построенной по этим данным (рис.4.4), видна последовательность объединения проб. Кроме того, на графике выделяются, по крайней мере, четыре группы проб (четыре типа цирконов по составу) и три пробы (13, 3 и 7), отличающиеся от других проб по составу. ◀◀

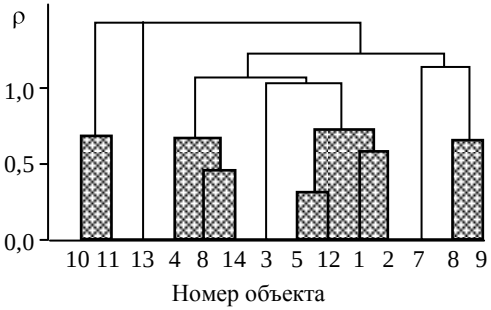

Рис.4.4. Дендрограмма проб циркона

Таким образом, кластерный анализ позволяет выполнить группировку объектов по степени их сходства по множеству свойств. Группировка при кластерном анализе может несколько отличаться от группировки по методу главных компонент, так как в основу этих методов заложены различные критерии. Можно сказать, что кластерный анализ дополняет метод главных компонент и помогает принять более правильное решение при выделении однородных совокупностей. Кроме того, представляет интерес и сама дендрограмма, наглядно характеризующая типизацию объектов по их свойствам.

4.2.4. Распознавание образов

Постановка задачи о распознавании образов. В геологической практике часто необходимо определить принадлежность объекта по множеству свойств к заданной совокупности однородных объектов. Сюда относятся задачи выделения перспективных территорий, оценки рудопроявлений, отнесения руд к какому-либо типу и многие другие. Решение подобных задач основано на приемах распознавания образов.

Образ – это совокупность однородных эталонных объектов с набором свойств. Образ выражается матрицей свойств (4.1), в при-

знаковом пространстве ему соответствует облако точек. В задачах распознавания обычно участвует не менее двух образов, а также испытуемые объекты. Задача распознавания заключается в определении принадлежности каждого испытуемого объекта к тому или иному образу. Распознавание образов наиболее часто используется при поисках месторождений, хотя область его применения гораздо шире.

Пусть имеется n перспективных территорий с заведомо промышленным оруденением определенного типа. Совокупность таких территорий, обладающих множеством свойств, создает образ рудных объектов. Им соответствует матрица свойств

$$X_{\text{рудн}} = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{vmatrix}. \qquad (4.19)$$

Далее имеется m заведомо неперспективных территорий, где оруденение отсутствует, – это образ безрудных объектов:

$$Y_{\text{безр}} = \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m1} & y_{m2} & \cdots & y_{mk} \end{vmatrix}. \qquad (4.20)$$

Наконец, имеются группы из l испытуемых территорий, перспективность которых неизвестна, им соответствует матрица

$$Z_{\text{исп}} = \begin{vmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ z_{l1} & z_{l2} & \cdots & z_{lk} \end{vmatrix}. \qquad (4.21)$$

У всех объектов должны быть измерены одни и те же величины (например, длина, мощность и т.д.). Цель исследования состоит в определении, к какому образу – рудному или безрудному – относится каждая из испытуемых территорий.

Может быть задано несколько типов рудных объектов и, соответственно, рудных образов. Испытуемые объекты могут быть объединены с безрудными. Иногда рудные объекты делят на две части. Одна часть используется «для обучения», другая – «для экзамена», т.е. для проверки эффективности методики распознавания.

Существует много способов решения задачи распознавания. Последовательность операций по решению задачи распознавания называется решающим правилом или *алгоритмом распознавания*.

Определение информативных свойств. Распознаванию образов предшествует отбор информативных свойств, чтобы исключить из рассмотрения неинформативные свойства и сократить объем вычислений. Более того, избыток неинформативных свойств ухудшает результаты распознавания образов.

Один из методов определения информативности свойств основан на оценке частот сочетаний качественных свойств рудных объектов, поскольку сочетания свойств более информативны, чем сумма информативностей отдельных свойств. Например, оруденение может появляться при благоприятном сочетании нескольких свойств (при пересечении разрывных нарушений, на контакте интрузивных пород с определенными литологическими типами горных пород и т.д.). В простейшем случае информативность J отдельного свойства i определяется по формуле

$$J = \frac{1}{n}\sqrt{\frac{1}{k}\sum_{i=1}^{k} N_{ij}^2}\ , \qquad (4.22)$$

где n – число объектов; k – число свойств; N_{ij} – частота совместного появления свойства i и свойства j.

Расположив свойства в порядке информативности, можно найти суммарную информативность m свойств:

$$J_m = \sqrt{\sum_{j=1}^{m} J_j^2}\ . \qquad (4.23)$$

Известны приложения рассмотренного приема к определению информативности руководящей фауны.

В алгоритме распознавания образов «Кора-3», предложенного М.М.Бонгардом [15], используются не только двойные, но тройные сочетания свойств.

⏩ **Пример 4.6.** Имеется шесть рудных объектов, у которых измерено пять качественных свойств:

$$x_{ij}^{\text{руд}} = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{vmatrix}.$$

Требуется определить информативность свойств.

По формуле (4.22) найдем информативность первого свойства:

$$J_1 = \frac{1}{6}\sqrt{\frac{1}{5}(3^2 + 2^2 + 1^2 + 1^2 + 2^2)} = 0{,}32.$$

Здесь числа в скобках – частота сочетания первого свойства со всеми свойствами (включая первое) в квадрате. Аналогично найдем информативность остальных свойств. В результате получим набор (вектор) информативности всех свойств:

$$\{J\} = (0{,}32 \quad 0{,}57 \quad 0{,}28 \quad 0{,}39 \quad 0{,}53).$$

Наиболее информативным оказалось второе свойство, далее идут пятое, четвертое, первое и третье свойства. По формуле (4.23) определим суммарную информативность, последовательно добавляя свойства в порядке убывания их информативности:

$$\{J_m\} = (0{,}57 \quad 0{,}78 \quad 0{,}87 \quad 0{,}93 \quad 0{,}97).$$

Если принять информативность суммы всех свойств за 100 %, то на долю двух важнейших свойств приходится 81 %, а при добавлении третьего свойства она возрастет до 90 %. Для решения задачи распознавания образов можно ограничиться тремя первыми свойствами. ⏪

Второй метод определения информативности качественных свойств основан на энтропии свойства j:

$$J_j = \sum_{i=1}^{k} (p_{1i} - p_{2i}) \ln \frac{p_{1i}}{p_{2i}}, \qquad (4.24)$$

где p_{1i} и p_{2i} – частота появления свойства j соответственно на рудных и безрудных объектах.

Чем больше частоты p_{1i} и p_{2i} отличаются друг от друга, тем более информативны соответствующие свойства. Пример применения формулы (4.24) приведен в работе [1], где оценена информативность свойств в районе развития медно-никелевых месторождений и выделены территории, наиболее перспективные на поиски этого оруденения (рис.4.5).

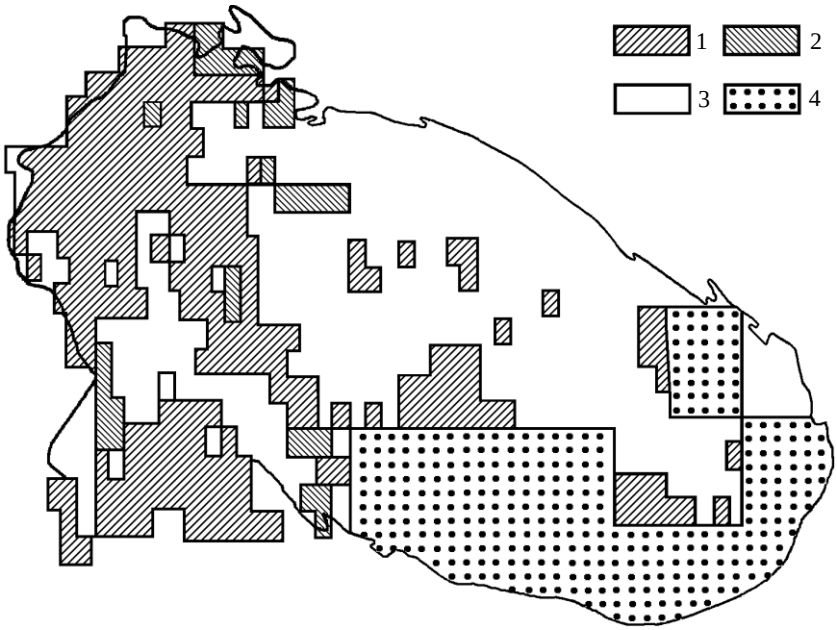

Рис.4.5. Схема перспективности на никель Кольского полуострова (по листам геологической карты масштаба 1:50 000) [1]
1 – перспективные территории; 2 – неперспективные территории с массивами ультраосновных пород; 3 – территории без ультраосновных пород; 4 – территории, не охваченные прогнозной оценкой

Информативность количественных свойств оценивают путем анализа расстояний между облаками точек рудных и безрудных объектов в признаковом пространстве. Информативность свойства j характеризуется квадратом нормированного расстояния между проекциями центров облаков на ось j признакового пространства:

$$J_j = \rho^2 = \frac{(\bar{x}_j - \bar{y}_j)^2}{\sigma_j^2}, \qquad (4.25)$$

где $\bar{x}_j$ и $\bar{y}_j$ – средние значения свойства j рудных и безрудных объектов.

Дисперсию определяют по формуле

$$\sigma_j^2 = \frac{\sigma_{1j}^2}{n_1} + \frac{\sigma_{2j}^2}{n_2}, \qquad (4.26)$$

где n_1 и n_2 – количество соответственно рудных и безрудных объектов; σ_{1j}^2 и σ_{2j}^2 – дисперсии свойства j тех же объектов.

Чем больше значение J_j, тем более информативным является данное свойство.

Учет зависимости между свойствами может изменить порядок информативности свойств. На основе матриц исходных данных составляют матрицы ковариации для рудных $\{K\}_{\text{рудн}}$ и безрудных $\{K\}_{\text{безр}}$ объектов. Элементы матриц находят из выражения

$$K_{ij} = \frac{1}{n}\sum_{m=1}^{k}(x_{im} - \bar{x}_{jm})(x_{jm} - \bar{x}_{jm}). \qquad (4.27)$$

Далее определяют средневзвешенную матрицу ковариации по формуле, аналогичной (4.26). Элементы средневзвешенной матрицы

$$\overline{K}_{ij} = \frac{K_{ij\text{рудн}}}{n_{\text{рудн}}} + \frac{K_{ij\text{безр}}}{n_{\text{безр}}}. \qquad (4.28)$$

Из элементов матрицы составляют систему из k уравнений с неизвестными коэффициентами a_1, a_2, …, a_k, которые находят путем решения этой системы:

$$\begin{vmatrix} K_{11} & K_{12} & \cdots & K_{1m} & \bar{x}_1 - \bar{y}_1 \\ K_{21} & K_{22} & \cdots & K_{2m} & \bar{x}_2 - \bar{y}_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ K_{m1} & K_{m2} & \cdots & K_{mm} & \bar{x}_m - \bar{y}_m \end{vmatrix}, \qquad (4.29)$$

где K_{ij} – средневзвешенные корреляционные моменты между свойствами объектов (раздельно для рудных и безрудных объектов).

Информативность k количественных свойств выражается квадратом обобщенного расстояния между облаками в признаковом пространстве:

$$J_k = \rho_k^2 = \sum_{i=1}^{k} a_i (\bar{x}_i - \bar{y}_i). \qquad (4.30)$$

Отсюда вытекает порядок определения информативности свойств. Вначале по формуле (4.25) отыскивают наиболее информативное свойство. К нему поочередно добавляют каждое из оставшихся свойств, находя наиболее информативную комбинацию из двух свойств по формуле (4.27). Далее добавляют следующее свойство и снова находят наиболее информативное сочетание свойств. Повторяя эти операции, определяют все последующие свойства в ряду информативности. Отбор информативных свойств прекращается, когда дальнейшее их увеличение не приводит к заметному повышению совместной информативности.

⏩ **Пример 4.7.** Имеется 10 рудных и 14 безрудных объектов, у которых изучено пять свойств:

$$\{x\} = \begin{vmatrix} 0 & 2{,}15 & 12 & 0{,}46 & 2 \\ 1 & 1{,}46 & 9 & -0{,}23 & 1 \\ 1 & 4{,}16 & 15 & -0{,}21 & 2 \\ 0 & 0{,}44 & 9 & 0{,}25 & 3 \\ 1 & 3{,}17 & 11 & 0{,}71 & 1 \\ 0 & 5{,}10 & 12 & -0{,}32 & 2 \\ 0 & 2{,}44 & 16 & 0{,}05 & 2 \\ 1 & 1{,}38 & 10 & 0{,}21 & 1 \\ 1 & 3{,}51 & 15 & -0{,}53 & 3 \\ 0 & 2{,}12 & 8 & -0{,}06 & 2 \end{vmatrix}; \quad \{y\} = \begin{vmatrix} 1 & 2{,}16 & 25 & 0{,}12 & 1 \\ 1 & 3{,}14 & 31 & -0{,}40 & 3 \\ 0 & 0{,}35 & 15 & 0{,}18 & 3 \\ 0 & 1{,}76 & 44 & 0{,}55 & 3 \\ 1 & 0{,}10 & 11 & 0{,}36 & 2 \\ 0 & -0{,}05 & 3 & 0{,}37 & 1 \\ 1 & 1{,}20 & 8 & 0{,}00 & 2 \\ 0 & 2{,}22 & 16 & 0{,}48 & 2 \\ 1 & 0{,}48 & 6 & 0{,}12 & 1 \\ 0 & 1{,}15 & 14 & -0{,}15 & 3 \\ 0 & -0{,}11 & 5 & 0{,}44 & 1 \\ 1 & 1{,}65 & 18 & 0{,}35 & 2 \\ 1 & 3{,}00 & 35 & 0{,}20 & 3 \\ 1 & 2{,}20 & 26 & 0{,}06 & 2 \end{vmatrix}.$$

Требуется определить информативность свойств.

Найдем средние значения свойств:

$$\{\bar{x}\} = \left| 0{,}500 \quad 2{,}58 \quad 11{,}70 \quad 0{,}034 \quad 1{,}900 \right|;$$

$$\{\bar{y}\} = \left| 0{,}571 \quad 1{,}38 \quad 18{,}36 \quad 0{,}191 \quad 2{,}071 \right|.$$

Далее вычислим дисперсии значений свойств:

$$\sigma_x^2 = \left| 0{,}250 \quad 1{,}800 \quad 7{,}21 \quad 0{,}1296 \quad 0{,}490 \right|;$$

$$\sigma_y^2 = \left| 0{,}245 \quad 1{,}127 \quad 140{,}1 \quad 0{,}0633 \quad 0{,}638 \right|.$$

По формуле (4.26) рассчитаем дисперсии разности средних значений свойств:

$$\sigma^2 = \left| 0{,}0425 \quad 0{,}261 \quad 10{,}73 \quad 0{,}01748 \quad 0{,}0946 \right|.$$

По формуле (4.25) найдем квадраты расстояний:

$$\rho^2 = \left| 0{,}12 \quad 5{,}52 \quad 4{,}13 \quad 1{,}41 \quad 0{,}31 \right|.$$

Из полученных данных видно, что наибольшей информативностью обладает второе свойство, далее идут третье, четвертое, пятое и первое.

Рассмотрим, как изменится информативность свойств, если учитывать зависимости между ними. Вначале рассчитаем матрицы ковариации свойств рудных и безрудных объектов по формуле (4.27):

$$\{K\}_{\text{рудн}} = \begin{vmatrix} 0{,}260 & 0{,}067 & 0{,}150 & -0{,}0220 & -0{,}150 \\ 0{,}067 & 1{,}800 & 2{,}060 & -0{,}1945 & 0{,}062 \\ 0{,}150 & 2{,}060 & 7{,}210 & -0{,}2910 & 0{,}570 \\ -0{,}022 & -0{,}1945 & -0{,}291 & 0{,}1296 & -0{,}0936 \\ -0{,}150 & 0{,}062 & 0{,}570 & -0{,}0936 & 0{,}490 \end{vmatrix};$$

$$\{K\}_{\text{безр}} = \begin{vmatrix} 0{,}245 & 0{,}209 & 0{,}939 & -0{,}0518 & -0{,}041 \\ 0{,}209 & 1{,}127 & 9{,}790 & -0{,}1006 & 0{,}396 \\ 0{,}939 & 9{,}790 & 140{,}1 & -0{,}2280 & 5{,}830 \\ -0{,}0518 & -0{,}1006 & -0{,}228 & 0{,}0633 & -0{,}0615 \\ -0{,}041 & 0{,}396 & 5{,}830 & -0{,}0615 & 0{,}638 \end{vmatrix}.$$

Далее найдем средневзвешенную матрицу по формуле (4.28):

$$\{\overline{K}\} = \begin{vmatrix} 0{,}0425 & 0{,}0216 & 0{,}0821 & -0{,}0059 & -0{,}0179 \\ 0{,}0216 & 0{,}2605 & 0{,}9050 & -0{,}0266 & 0{,}0345 \\ 0{,}0821 & 0{,}9050 & 10{,}73 & -0{,}0454 & 0{,}4730 \\ -0{,}0059 & -0{,}0266 & -0{,}0454 & 0{,}0175 & -0{,}0138 \\ -0{,}0179 & 0{,}0345 & 0{,}4730 & -0{,}0138 & 0{,}0946 \end{vmatrix}.$$

Как отмечалось, наиболее информативным является второе свойство, для него $\rho^2 = 5{,}52$. Будем поочередно добавлять ко второму свойству остальные и вычислять совместную информатив-

ность двух свойств. Так, найдем совместную информативность первого и второго свойств. Матрица их ковариации, согласно (4.28), имеет вид

$$\{\overline{K}\} = \begin{vmatrix} 0{,}0425 & 0{,}0216 \\ 0{,}0216 & 0{,}2605 \end{vmatrix}.$$

Составим систему уравнений (4.29):

$$0{,}0425a_1 + 0{,}0216a_2 = 0{,}500 - 0{,}571;$$

$$0{,}0216a_1 + 0{,}2605a_2 = 2{,}58 - 1{,}38.$$

Решая систему, найдем $a_1 = -4{,}19$; $a_2 = 4{,}96$. Совместная информативность первого и второго свойств по формуле (4.30) $\rho_{12}^2 = 6{,}25$, что больше, чем их сумма $\rho_1^2 + \rho_2^2 = 5{,}64$, т.е. информативность двух свойств с учетом зависимости больше, чем без ее учета.

Так же оценивается совместная информативность второго свойства с третьим, четвертым и пятым свойствами. Имеем $\rho_{23}^2 = 20{,}99$; $\rho_{24}^2 = 5{,}60$; $\rho_{25}^2 = 6{,}73$; следовательно, наиболее информативным является сочетание второго и третьего свойств. К этим двум свойствам поочередно добавляют оставшиеся свойства (первое, четвертое и пятое) и по той же методике находят наиболее информативную комбинацию трех, четырех и пяти свойств. В рассматриваемом примере порядок наиболее информативных свойств следующий: второе, третье, первое, пятое и четвертое, т.е. иной, чем без учета зависимости между свойствами. Совместная информативность перечисленных свойств в нарастающем порядке выражается строкой матрицы

$$\rho_k = \{5{,}52 \quad 20{,}99 \quad 21{,}45 \quad 21{,}75 \quad 21{,}89\},$$

откуда видно, что для распознавания образа достаточно ограничиться двумя свойствами (вторым и третьим), остальные свойства вносят малый вклад в информативность и могут быть отброшены. ◀◀

Методы распознавания образов. После отбора информативных свойств можно приступать к распознаванию образов. Методы распознавания образов весьма многочисленны. Их можно разделить на три группы:

- методы, основанные на анализе частоты появления свойств;
- методы, основанные на анализе расстояний между точками объектов в многомерном признаковом пространстве;
- методы, основанные на разделении многомерного признакового пространства на области различной формы, в каждой из которых преобладают объекты одного вида.

В настоящей работе не ставится задача дать обзор различных методов распознавания образов [15], ограничимся лишь примерами алгоритмов в каждой из групп. Следует отметить, что не существует методов, дающих стопроцентный результат распознавания. Хорошим результатом считается уверенное распознавание испытуемых объектов на уровне 75-85 %.

При частотном анализе исходные свойства выражаются нулями и единицами, т.е. одни свойства у объекта есть, другие отсутствуют. В алгоритме «Кора-3» перебираются все двойные и тройные комбинации свойств, которые характерны для одного образа и отсутствуют у другого. Найденные комбинации свойств называются *сложными признаками*. Для каждого испытуемого объекта определяется количество сложных признаков каждого образа. Испытуемый объект относят к образу рудных или безрудных объектов по большинству «голосов».

⏩ **Пример 4.8.** Имеются шесть рудных, шесть безрудных и три испытуемых объекта, у которых изучены четыре свойства:

$$\{X\}_{\text{рудн}} = \begin{vmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{vmatrix}; \quad \{Y\}_{\text{безр}} = \begin{vmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix};$$

$$\{Z\}_{\text{исп}} = \begin{vmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix}.$$

Требуется выявить сложные признаки и определить принадлежность каждого испытуемого объекта к образу рудных или безрудных объектов.

Нетрудно установить, что существует 11 сочетаний свойств по два и по три, которые имеются у рудных объектов и отсутствуют у безрудных, и, наоборот, 13 сочетаний свойств типичны только для безрудных объектов. Всего имеем 11 + 13 = 24 сложных признака (табл. 4.17).

Таблица 4.17

Сложные признаки

Номер признака	Свойство				Номер признака	Свойство			
	1	2	3	4		1	2	3	4
Рудные объекты					Безрудные объекты				
1	0	0	–	–	1	1	1	–	–
2	–	0	–	1	2	1	–	–	0
3	0	0	1	–	3	–	–	0	0
4	0	0	–	0	4	1	1	0	–
5	0	0	–	1	5	1	1	1	–
6	1	0	–	1	6	1	0	–	0
7	0	–	0	1	7	1	1	–	0
8	0	–	1	0	8	0	–	0	0
9	1	–	1	1	9	1	–	0	0
10	–	0	0	1	10	1	–	1	0
11	–	0	1	1	11	–	0	0	0
					12	–	1	0	0
					13	–	1	1	1

Таблица 4.18

Результаты распознавания объектов

Номер испытуемого объекта	Количество признаков объектов		Вывод о принадлежности испытуемых объектов
	рудных	безрудных	
1	5	0	Рудный
2	0	6	Безрудный
3	1	3	"–"

Используя сведения по испытуемым объектам и данные табл.4.17, можно подсчитать, что для первого испытуемого объекта характерны признаки только рудных объектов, для второго – только безрудных, у третьего испытуемого объекта есть те и другие признаки, но больше «голосов» в пользу безрудных объектов (табл.4.18). ◀◀

Другой алгоритм распознавания основан на анализе образов рудных и безрудных объектов, которые в признаковом пространстве слагают два соответствующих облака, которые могут частично перекрывать друг друга. Принадлежность испытуемого объекта к образу рудных или безрудных объектов можно оценить по расстоянию точки испытуемого объекта от облаков. К какому облаку ближе точка, к тому образу и следует отнести испытуемый объект.

Существуют различные способы оценки расстояний между точкой и облаком. Можно рекомендовать следующий способ: вначале найти расстояния до всех точек облака, а потом взять среднее из них. Для оценки расстояния между точкой испытуемого объекта $x_{\text{исп}}$ и точкой облака x_i применима формула

$$\rho_i = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i - x_{\text{исп}}}{x_{\max} - x_{\min}}\right)^2}, \qquad (4.31)$$

где $x_{\max}$ и $x_{\min}$ – максимальные и минимальные значения различных свойств i; $x_{\max} - x_{\min}$ – размах свойств.

Далее вычислим среднее расстояние точки до облака:

$$\overline{\rho} = \frac{1}{n}\sum_{i=1}^{n}\rho_i . \qquad (4.32)$$

⏩ **Пример 4.9.** Даны пять рудных, шесть безрудных и два испытуемых объекта, у которых изучены три свойства:

$$\{X\}_{\text{рудн}} = \begin{vmatrix} 2 & 0 & 4{,}4 \\ 1 & 1 & 0{,}6 \\ 3 & 1 & 2{,}1 \\ 2 & 0 & -0{,}3 \\ 1 & 1 & 1{,}8 \end{vmatrix}; \{Y\}_{\text{безр}} = \begin{vmatrix} 1 & 1 & 2{,}3 \\ 1 & 0 & 1{,}6 \\ 2 & 1 & -2{,}5 \\ 1 & 0 & 0{,}2 \\ 3 & 0 & 1{,}0 \\ 1 & 1 & -0{,}5 \end{vmatrix}; \{Z\}_{\text{исп}} = \begin{vmatrix} 2 & 1 & 3{,}3 \\ 1 & 0 & 1{,}2 \end{vmatrix}.$$

Требуется определить принадлежность испытуемых объектов к рудным или к безрудным.

Вначале найдем максимальные и минимальные значения свойств:

$$\{X\}_{\max} = \begin{vmatrix} 3 & 1 & 4{,}4 \end{vmatrix}; \ \{X\}_{\min} = \begin{vmatrix} 1 & 0 & -2{,}5 \end{vmatrix}.$$

По формуле (4.31) определим расстояние точки первого испытуемого объекта от одной из точек облака рудных объектов:

$$\rho_i = \sqrt{\left(\frac{2-2}{3-1}\right)^2 + \left(\frac{0-1}{1-0}\right)^2 + \left(\frac{4{,}4-3{,}3}{4{,}4+2{,}5}\right)^2} = 1{,}01.$$

Таким же способом найдем расстояние для всех остальных точек облаков и рассчитаем среднее расстояние до каждого облака по формуле (4.32). Для первого испытуемого объекта получим $\overline{\rho}_{\text{рудн}} = 0{,}76$, $\overline{\rho}_{\text{безр}} = 0{,}94$, следовательно, он ближе к рудным объектам. Второй испытуемый объект имеет $\overline{\rho}_{\text{рудн}} = 0{,}93$, $\overline{\rho}_{\text{безр}} = 0{,}75$, он ближе к безрудным. ⏪

Следующая группа методов основана на разделении признакового пространства на области, в которых преобладают объекты одного типа. Чаще области выделяются путем проведения плоскостей (гиперплоскостей) между облаками рудных и безрудных объектов. Такой способ называется *дискриминантным анализом*.

Уравнение плоскости имеет вид

$$\sum_{i=1}^{k} a_i x_i - b = 0, \tag{4.33}$$

где a_i – коэффициенты, определяющие ориентировку плоскости по отношению к осям координат признакового пространства и вычисляемые путем решения системы уравнений (4.29); b – коэффициент, влияющий на параллельное перемещение плоскости в пространстве.

Если подставить координаты x_i любой точки облака в уравнение (4.33), то получим дискриминант этой точки:

$$D = \sum_{i=1}^{k} a_i x_i - b. \tag{4.34}$$

По одну сторону от плоскости знак дискриминанта положительный, по другую – отрицательный. Следовательно, по знаку дискриминанта можно судить о том, в какую область пространства попадает точка испытуемого объекта, т.е. к какому образу следует отнести объект.

Если облака рудных и безрудных объектов разобщены между собой, то плоскость можно провести в любом месте между ними. Если облака частично перекрываются, то нужно провести разделяющую плоскость так, чтобы наименьшая часть точек попадала в чужую область, что достигается нахождением наилучшего значения коэффициента b. Подбор коэффициента b можно осуществлять вручную (а также и коэффициентов a_i), но можно вычислить его следующим способом. Вначале находят коэффициент p для каждой точки облаков:

$$p = \sum_{i=1}^{k} a_i x_i . \tag{4.35}$$

Далее вычисляют средние значения и дисперсии значений p для рудных и безрудных объектов, т.е. $\overline{p}_{\text{рудн}}$, $\overline{p}_{\text{безр}}$, $\sigma^2_{\text{рудн}}$ и $\sigma^2_{\text{безр}}$. Обычно предполагают, что значения коэффициентов p подчиняются или близки к нормальному закону распределения, тогда определение коэффициентов b сводится к решению квадратного уравнения

$$\left(\frac{b-\bar{p}_{\text{рудн}}}{\sigma_{\text{рудн}}}\right)^2 - 2\ln\frac{n_{\text{рудн}}}{\sigma_{\text{рудн}}} = \left(\frac{b-\bar{p}_{\text{безр}}}{\sigma_{\text{безр}}}\right)^2 - 2\ln\frac{n_{\text{безр}}}{\sigma_{\text{безр}}}. \qquad (4.36)$$

⏩ **Пример 4.10.** Используя данные примера 4.7, необходимо найти уравнение плоскости, разделяющей образы рудных и безрудных объектов, и дискриминанты объектов.

Выпишем значения информативных свойств (второго и третьего):

$$\{X\}_{\text{рудн}} = \begin{vmatrix} 2{,}15 & 12 \\ 1{,}46 & 9 \\ 4{,}16 & 15 \\ 0{,}44 & 9 \\ 3{,}17 & 11 \\ 5{,}10 & 12 \\ 2{,}44 & 16 \\ 1{,}28 & 10 \\ 3{,}51 & 15 \\ 2{,}12 & 8 \end{vmatrix}; \quad \{Y\}_{\text{безр}} = \begin{vmatrix} 2{,}16 & 25 \\ 3{,}24 & 31 \\ 0{,}35 & 15 \\ 1{,}76 & 44 \\ 0{,}10 & 11 \\ -0{,}05 & 3 \\ 1{,}20 & 8 \\ 2{,}22 & 16 \\ 0{,}48 & 6 \\ 1{,}15 & 14 \\ -0{,}11 & 5 \\ 1{,}65 & 18 \\ 3{,}00 & 35 \\ 2{,}20 & 26 \end{vmatrix}.$$

Средние значения свойств:

$$\bar{X} = \begin{vmatrix} 2{,}58 & 11{,}70 \end{vmatrix}; \quad \bar{Y} = \begin{vmatrix} 1{,}38 & 18{,}36 \end{vmatrix}.$$

Матрица ковариации свойств: $\{K\} = \begin{vmatrix} 0{,}2605 & 0{,}905 \\ 0{,}905 & 10{,}73 \end{vmatrix}$.

Составим систему уравнений (4.29):

$$0{,}2605a_1 + 0{,}905a_2 = 2{,}58 - 1{,}38;$$
$$0{,}905a_1 + 10{,}73a_2 = 11{,}70 - 18{,}36.$$

Решение системы дает коэффициенты $a_1 = 9{,}57,\ a_2 = -1{,}427.$

Вычислим значения p для рудных и безрудных объектов по формуле (4.35):

$$\{p\}_{\text{рудн}} = \begin{vmatrix} 3{,}45 \\ 1{,}13 \\ 18{,}41 \\ -8{,}63 \\ 14{,}69 \\ 31{,}68 \\ 0{,}52 \\ -2{,}02 \\ 12{,}19 \\ 8{,}87 \end{vmatrix};\ \{p\}_{\text{безр}} = \begin{vmatrix} -15{,}00 \\ -14{,}19 \\ -18{,}06 \\ -45{,}94 \\ -14{,}74 \\ -4{,}76 \\ 0{,}07 \\ -1{,}59 \\ -3{,}97 \\ -8{,}97 \\ -8{,}19 \\ -9{,}90 \\ -21{,}24 \\ -16{,}05 \end{vmatrix}.$$

Средние значения $\overline{p}_{\text{рудн}} = 8{,}02,\ \overline{p}_{\text{безр}} = -13{,}04,$ дисперсии $\sigma^2_{\text{рудн}} = 123{,}2.$ Подставив полученные значения в уравнение (4.36) и решая его, найдем один из корней b = 0,56, следовательно, уравнение плоскости (в данном случае – это прямая линия, рис.4.6) имеет вид

$$9{,}57x_1 - 1{,}427x_2 + 0{,}56 = 0,$$

а дискриминант выражается формулой

$$D = 9{,}57x_1 - 1{,}427x_2 + 0{,}56.$$

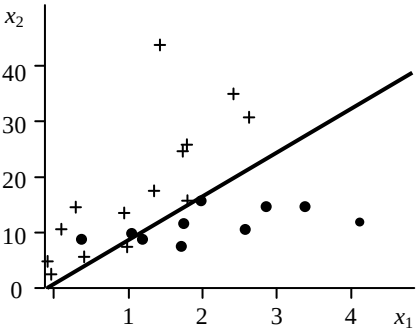

Рис.4.6. Разделение рудных (точки) и безрудных (крестики) объектов дискриминантной плоскостью (линией)

Значения дискриминанта отличаются от значений p лишь коэффициентом $b = 0{,}56$. Для рудных объектов значения дискриминанта, как правило, положительные, для безрудных – отрицательные. Только в трех случаях из 24 знак дискриминанта не отвечает принадлежности объектов, т.е. ошибка распознавания составляет $3/24 = 0{,}125 =$ $= 12{,}5\ \%$. ◀◀

Следует отметить, что на практике (например, при обогащении руд) может быть использована комбинация дискриминантного

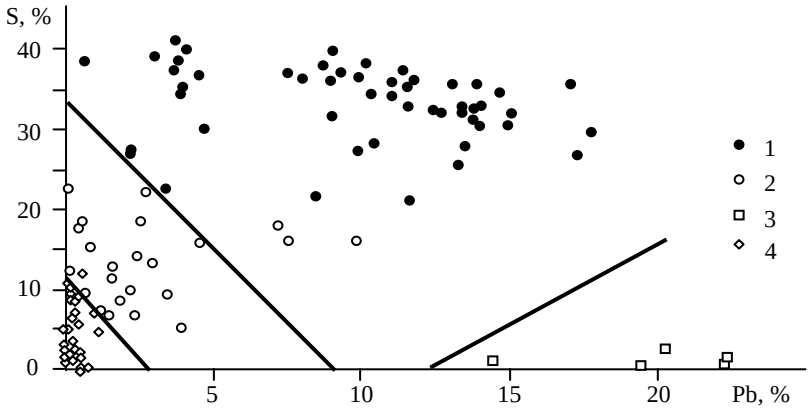

Рис.4.7. Разделение полиметаллических руд с помощью дискриминантного анализа
1 – сплошные полиметаллические руды; 2 – вкрапленные полиметаллические руды; 3 – охристые окисленные руды; 4 – оруденелые метасоматиты

анализа с методом главных компонент (подраздел 4.2.2). Если на графике главных компонент рудные и безрудные объекты дают два различных облака, то между ними можно провести дискриминантную плоскость, как описано в данном подразделе (рис.4.7).

Глава 5 МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННЫХ ГЕОЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

5.1. СВОЙСТВА ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ КАК ПРОСТРАНСТВЕННЫЕ ПЕРЕМЕННЫЕ

В данной главе свойства будут изучены как функции координат пространства – *пространственные переменные*. В роли пространственных переменных могут выступать мощность пластообразных тел, абсолютные отметки кровли и почвы пласта, содержание компонентов в рудном теле и многие другие величины.

Пространство, в котором существует изучаемая величина, называется *геологическим полем* пространственной переменной. В каждой точке или области геологического поля с координатами x, y, z свойство принимает конкретное значение $\varphi(x, y, z)$. В общем случае в каждой точке или области геологического поля могут быть измерены несколько величин.

Значения пространственной переменной измеряют в пределах геологических объектов конечных размеров по какой-то сети в ограниченных областях геологического поля. В связи с этим рассмотрим параметры геометрии сети наблюдений и области измерений.

Геометрия сети наблюдений характеризуется формой, расположением и плотностью сети. Сеть бывает одномерная – вдоль линии, двухмерная – по площади и трехмерная – в объеме геологического тела. Измерения делят на непрерывные и дискретные (прерывистые). Для математической обработки непрерывные измерения обычно преобразуют в дискретные.

Наблюдения могут размещаться по равномерной, кратной или неравномерной сети. Равномерная сеть характеризуется постоянным шагом h – равным расстоянием между пунктами наблюдений. В двухмерной сети имеется два постоянных шага h_1 и h_2, которые образуют ячейку сети площадью $s = h_1h_2$. Если $h_1 = h_2$, то сеть квадратная. Трехмерная сеть имеет три постоянных шага h_1, h_2, h_3, образующих ячейку объемом $v = h_1h_2h_3$. В реальных условиях не всегда удается соблюдать строго равномерную сеть наблюдений. Небольшими отклонениями от равномерной сети часто можно пренебречь. У кратной сети расстояния между пунктами наблюдений непостоянные, но кратные шагу h.

Плотность сети – количество наблюдений на единицу длины, площади или объема геологического объекта. Если сеть наблюдений отличается от равномерной, то можно говорить о средней плотности сети.

Обычно пространственная переменная предполагается непрерывной величиной, плавно меняющей свое значение в геологическом поле. Но возможны сравнительно резкие и даже скачкообразные изменения пространственной переменной, отражающие дискретность строения геологических объектов и позволяющие проводить внутри них геологические границы.

Геометрия области измерения характеризуется формой, размером и ориентировкой. Форма области может быть точечной, линейной, сферической, цилиндрической и пр. Во многих случаях формой области можно пренебречь, полагая ее точечной.

Размер области наблюдения влияет на некоторые характеристики. Например, при увеличении размера области уменьшается дисперсия величин. Размером области измерения также часто пренебрегают, считая измерения точечными, тем более что размеры области обычно на порядок ниже шага сети наблюдений.

При уменьшении размеров области некоторые измерения стремятся к предельному значению величины в данной точке, что характерно для непрерывных величин, например для мощности пластообразных геологических тел. В других случаях такой предел от-

сутствует из-за дискретности строения геологических объектов (например, руда состоит из зерен рудных и нерудных минералов). В этих случаях принято говорить о средних значениях пространственной переменной в некоторой малой области геологического поля.

Ориентировка области измерений имеет значение для анизотропных геологических тел, а их большинство. При различной ориентировке линейных или цилиндрических областей в анизотропной среде можно получить разные результаты. Обычно стараются линейные пробы располагать по направлению наибольшей изменчивости свойств, т.е. по мощности рудных тел или пластов горных пород.

Результаты измерений пространственных переменных сводят в матрицу, в которой присутствуют значения величины φ и координаты пунктов наблюдений (центров областей измерений) x, y, z:

$$\begin{vmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1k} & x_1 & y_1 & z_1 \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2k} & x_2 & y_2 & z_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nk} & x_n & y_n & z_n \end{vmatrix}. \qquad (5.1)$$

Напомним, что n – количество объектов, k – количество величин. В зависимости от количества учитываемых координат пространственные переменные и геологические поля делятся на одно-, двух- и трехмерные.

5.2. ВИДЫ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ И ГЕОЛОГИЧЕСКИХ ПОЛЕЙ

Математическое моделирование геологического поля ставит своей целью описание поведения пространственной переменной по имеющимся результатам наблюдений, а также прогнозирование ее значений в заданных точках или областях геологического поля. По-

путно могут быть решены задачи оценки погрешности прогнозирования, рациональной плотности сети измерений и другие.

Реальные геологические поля пространственных переменных обладают большой сложностью. Математические модели геологических полей не позволяют дать их исчерпывающую характеристику, а отражают лишь наиболее существенные черты. Для каждого геологического поля можно построить много математических моделей, различающихся характером предположений о поведении величин в пространстве.

Математические модели геологических полей делятся на детерминированные и вероятностные. В *детерминированных* моделях предполагается, что пространственная переменная является неслучайной функцией координат и однозначно зависит от местоположения пунктов измерений. В тех пунктах, где проводились измерения, значения пространственной переменной принимают фактическими, а в промежутках между ними находят путем интерполяции. Способ интерполяции определяет вид математической модели. Среди детерминированных моделей можно выделить модели линейные, полиномиальные, обратных расстояний и сплайн-модели.

В *вероятностных* моделях предполагается, что значения пространственной переменной (в том числе и в пунктах измерений) содержат элементы случайности. Различают две группы математических моделей: случайные функции и геостатистические модели. В разных группах по-разному объясняется появление случайной составляющей.

Случайные функции основаны на предположении о том, что значения пространственной переменной $\varphi(x)$ испытывают случайные колебания $\delta(x)$ около неслучайной составляющей, называемой математическим ожиданием $m(x)$:

$$\varphi(x) = m(x) + \delta(x). \qquad (5.2)$$

В геологической литературе математическое ожидание называют также регулярной, координированной, закономерной составляющей или трендом. Математическое ожидание иногда делят на регулярную $f(x)$ и периодическую $\omega(x)$ составляющие:

$$m(x) = f(x) + \omega(x). \qquad (5.3)$$

Может быть несколько периодических составляющих, различающихся амплитудой и длиной волны.

Геостатистические модели содержат предположение о том, что случайный результат измерений вызван случайным расположением пунктов наблюдений. Любое перемещение сети наблюдений приводит к получению новых результатов, но при этом остается неизменным средний квадрат разности между результатами измерений в пунктах, отстоящих друг от друга на шаг *h*. Полусумма среднего квадрата разностей называется *вариограммой* $\gamma(h)$

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n} [\varphi(x+h) - \varphi(x)]^2 . \qquad (5.4)$$

Геостатистические модели различаются способом аппроксимации эмпирической вариограммы теоретической вариограммой и последующей интерполяцией результатов наблюдений.

5.3. ДЕТЕРМИНИРОВАННЫЕ МОДЕЛИ ГЕОЛОГИЧЕСКИХ ПОЛЕЙ

5.3.1. Линейная интерполяционная модель

В основе модели лежит предположение о том, что между пунктами измерений значения пространственной переменной меняются по закону прямой линии. При густой сети измерений и слабой изменчивости величин предположение может быть близким к действительности и не повлечет за собой существенных погрешностей при прогнозировании значений между пунктами измерений. Может быть противоположная ситуация, когда величина настолько изменчива, что какое-нибудь разумное предположение о ее поведении между пунктами измерений сделать невозможно. В этом случае линейная модель выбирается из соображений максимальной простоты, что обеспечивает высокую достоверность прогнозирования. Подобные соображения используют, например, для оконтуривания рудных тел при подсчете запасов, когда сложные контуры тел заменяют многоугольниками, состоящими из прямолинейных отрезков.

Если в пункте с координатой x_1 измерено значение пространственной переменной φ_1, в пункте x_2 – значение φ_2, то при линейной интерполяции в любом пункте x между x_1 и x_2 интерполированное (прогнозное) значение

$$\varphi = \frac{x - x_1}{x_2 - x_1}(\varphi_2 - \varphi_1) + \varphi_1 . \qquad (5.5)$$

Линейную интерполяцию можно представить графически в виде отрезков ломаной линии, опирающейся на измеренные значения φ_1, φ_2, …, φ_n (табл.5.1, рис.5.1).

Таблица 5.1

Результаты измерения мощности рудного тела

Номер пункта	Расстояние от начальной точки, м	Мощность, м
1	0,0	2,2
2	7,0	1,9
3	20,1	2,5
4	28,6	4,2
5	34,8	3,8
6	44,0	3,1
7	54,9	3,9
8	62,1	2,6

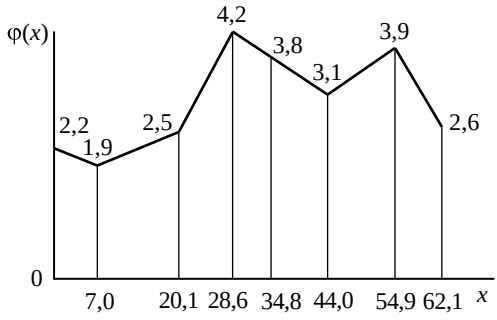

Рис.5.1. График линейной модели, построенный по данным табл.5.1

Линейная интерполяция может быть выполнена и в двухмерном пространстве внутри треугольника, образованного тремя пунктами наблюдений, не лежащими на одной прямой. По данным в вершинах треугольника можно найти уравнение плоскости:

$$\varphi = ax + by + c. \qquad (5.6)$$

Уравнение позволяет вычислять интерполированное значение φ в любой точке с заданными координатами x и y внутри треугольника. Если имеется много пунктов наблюдений, то охваченная ими площадь разбивается на

несколько треугольников и в каждом из них рассчитывается свое интерполяционное уравнение (5.6).

Таблица 5.2

Данные по скважинам

Номер скважины n	Координаты скважины, м		Абсолютная отметка пласта z, м
	x	y	
1	355	142	125,6
2	210	163	148,3
3	224	281	105,2

⏩ **Пример 5.1.** Имеются три вертикальные скважины, в которых определены абсолютные отметки кровли пласта (табл.5.2). Необходимо рассчитать абсолютную отметку кровли пласта с координатами x = 240 м, y = 200 м.

Составим систему уравнений:

$$355a + 142b + c = 125{,}6;$$

$$210a + 163b + c = 148{,}3;$$

$$224a + 201b + c = 105{,}2.$$

Решение системы дает коэффициенты a = –0,206; b = –0,341; c = 247,1. Следовательно, интерполяционное уравнение (5.6) имеет вид

$$z = -0{,}206x - 0{,}341y + 247{,}1.$$

Подставляя в него заданные координаты, найдем абсолютную отметку кровли в точке x = 240 м, y = 200 м:

$$z = -0{,}206 \cdot 240 - 0{,}341 \cdot 200 + 247{,}1 = 129{,}5 \text{ м}.$$

Зная коэффициенты уравнения (5.6), из примера можно извлечь дополнительную геологическую информацию об элементах залегания кровли пласта. Азимут простирания $\alpha = \operatorname{arctg}(-b/a)$, а угол падения $\gamma = \operatorname{arctg}\sqrt{a^2 + b^2}$. В данном примере имеем $\alpha = \operatorname{arctg}(-0{,}341/0{,}206) = -59° = 301°$, $\gamma = \operatorname{arctg}\sqrt{0{,}206^2 + 0{,}341^2} = 22°$. ⏪

Можно распространить линейную интерполяцию и на трехмерное пространство, которое разделяется на совокупность тетраэдров. Каждый тетраэдр образован четырьмя пунктами измерений, не

лежащими на одной плоскости. Внутри тетраэдра интерполяция осуществляется с помощью уравнения (гиперплоскости)

$$\varphi = ax + by + cz + d,$$

которое опирается на вершины тетраэдра и рассчитывается аналогично приведенному в примере 5.1.

Следует отметить, что линейная интерполяционная модель, как и другие детерминированные модели, не позволяет оценить погрешность интерполяции без привлечения дополнительных данных. Чтобы решить эту задачу, надо выполнить дополнительные измерения пространственной переменной внутри интервалов интерполяции и сравнить интерполированные и измеренные данные. Статистическая обработка таких материалов позволяет выявить, как погрешность интерполяции зависит от расстояния между пунктами наблюдений.

5.3.2. Полиномиальная модель

В основе модели лежит предположение о том, что поведение пространственной переменной нелинейное и может быть описано полиномиальной функцией, значения которой совпадают с фактическими данными в пунктах измерений. Полиномиальная функция может быть одно-, двух- и трехмерной. Простейшая одномерная полиномиальная функция имеет вид

$$\varphi(x) = a_0 + a_1x + a_2x^2 + \ldots + a_kx^k. \qquad (5.7)$$

Порядок полинома k на единицу меньше числа измерений n.

Полиномиальная модель редко применяется на практике, так как по мере увеличения числа измерений степень полинома растет, а при высоких степенях полинома интерполированные значения вычисляются с большой погрешностью и испытывают столь сильные изменения, что часто переходят границы реальности.

⏩ **Пример 5.2.** По данным табл.5.1 необходимо рассчитать полиномиальную модель.

Поскольку таблица содержит восемь пунктов измерений, необходимо иметь полином седьмого порядка. Подобный полином содержит восемь неизвестных коэффициентов, для нахождения которых составим систему из восьми уравнений:

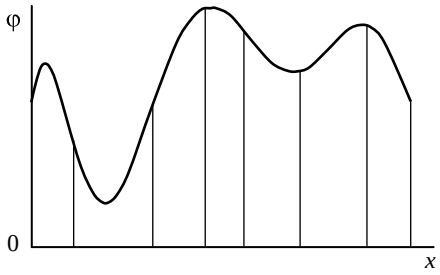

Рис.5.2. График полинома, построенный по данным табл.5.1

$$a_0 + 0{,}0a_1 + 0{,}0^2a_2 + 0{,}0^3a_3 + \\ + 0{,}0^4a_4 + 0{,}0^5a_5 + 0{,}0^6a_6 + \\ + 0{,}0^7a_7 = 2{,}2;$$

$$a_0 + 7{,}0a_1 + 7{,}0^2a_2 + 7{,}0^3a_3 + \\ + 7{,}0^4a_4 + 7{,}0^5a_5 + 7{,}0^6a_6 + \\ + 7{,}0^7a_7 = 1{,}9;$$

$$a_0 + 20{,}1a_1 + 20{,}1^2a_2 + 20{,}1^3a_3 + 20{,}1^4a_4 + 20{,}1^5a_5 + 20{,}1^6a_6 + 20{,}1^7a_7 = 2{,}5;$$

$$a_0 + 28{,}6a_1 + 28{,}6^2a_2 + 28{,}6^3a_3 + 28{,}6^4a_4 + 28{,}6^5a_5 + 28{,}6^6a_6 + 28{,}6^7a_7 = 4{,}2;$$

$$a_0 + 34{,}8a_1 + 34{,}8^2a_2 + 34{,}8^3a_3 + 34{,}8^4a_4 + 34{,}8^5a_5 + 34{,}8^6a_6 + 34{,}8^7a_7 = 3{,}8;$$

$$a_0 + 44{,}0a_1 + 44{,}0^2a_2 + 44{,}0^3a_3 + 44{,}0^4a_4 + 44{,}0^5a_5 + 44{,}0^6a_6 + 44{,}0^7a_7 = 3{,}1;$$

$$a_0 + 54{,}9a_1 + 54{,}9^2a_2 + 54{,}9^3a_3 + 54{,}9^4a_4 + 54{,}9^5a_5 + 54{,}9^6a_6 + 54{,}9^7a_7 = 3{,}9;$$

$$a_0 + 62{,}1a_1 + 62{,}1^2a_2 + 62{,}1^3a_3 + 62{,}1^4a_4 + 62{,}1^5a_5 + 62{,}1^6a_6 + 62{,}1^7a_7 = 2{,}6.$$

Решение системы дает коэффициенты a_0, a_1, a_2, …, a_7, которые позволяют составить уравнение полинома (5.7) и построить график (рис.5.2). Численные значения по осям графика такие же, как на рис.5.1. ◀◀

5.3.3. Модель обратных расстояний

В основу модели положена идея о том, что влияние измерений убывает обратно пропорционально квадрату расстояния r от пункта измерения (как в законе всемирного тяготения или в электрическом поле заряженных частиц), поэтому модель часто называют потенциальной. Интерполированное значение φ в каждой точке находят как средневзвешенное из измеренных значений в соседних пунктах n:

$$\varphi = \sum_{i=1}^{n} \frac{\varphi_i}{r^2} : \sum_{i=1}^{n} \frac{1}{r^2}. \qquad (5.8)$$

Если расстояние r равно нулю, то в данном пункте принимается измеренное фактическое значение. Для прогнозирования берут три-пять ближайших пунктов измерений или ограничиваются каким-то произвольным радиусом R. В расчет принимают все пункты измерений в пределах этого радиуса. За пределами радиуса влияние измеренных значений не учитывается.

⏩ **Пример 5.3.** Используя данные табл.5.1, построим интерполяционную модель обратных расстояний (рис.5.3).

Все расчеты проведем по четырем соседним пунктам измерений. Вначале берем первые четыре пункта и по формуле (5.8) находим значения кривой между первым и третьим пунктами. Далее, чтобы найти значения кривой между третьим и четвертым пунктами, необходимо брать в расчет слева и справа по два пункта (второй, третий, четвертый и пятый). Передвигаясь далее на один пункт, повторим расчеты по следующим четырем пунктам (с третьего по шестой). В результате получим кривую между четвертым и пятым пунктами. Подобные операции повторяют до конечного пункта. Такой прием называется «расчет в скользящем окне размером в четыре наблюдения». Полученные результаты представлены

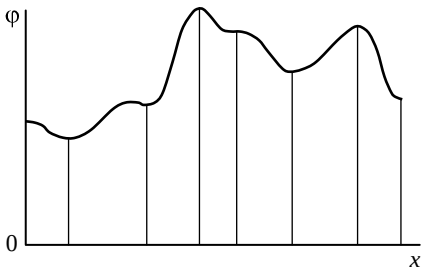

Рис.5.3. График обратных расстояний, построенный по данным табл.5.1

на рис.5.3. Численные значения по осям графика такие же, как на рис.5.1.

На графике виден недостаток метода обратных расстояний. В пунктах измерения касательная к кривой всегда горизонтальная, что не соответствует действительности. Тем не менее, метод широко применяется на практике, когда другие методы не работают. ◀◀

5.3.4. Сплайн-модель

Сплайн – это кусочно-непрерывная гладкая функция, состоящая из множества полиномиальных функций третьего порядка, плавно переходящих друг в друга. Сплайн позволяет построить плавный график пространственной переменной, хорошо согласующийся с геологическими представлениями, и поэтому весьма популярен. Его можно уподобить гибкой упругой линейке, опирающейся на ординаты фактических значений пространственной переменной. Концы линейки могут быть свободными или закрепленными с заданным углом наклона. Сплайн бывает одно-, двух- и трехмерным.

Рассмотрим методику расчета одномерного сплайна. Пусть имеется n пунктов измерений с координатами $x_1, x_2, \ldots, x_n$. В каждом пункте измерено значение пространственной переменной $\varphi_1, \varphi_2, \ldots, \varphi_n$. Между пунктами измерений имеется $n - 1$ отрезок. Каждый отрезок представлен полиномом третьего порядка:

$$\varphi = a + bx + cx^2 + dx^3. \qquad (5.9)$$

В каждом полиноме четыре неизвестных коэффициента a, b, c, d. Следовательно, всего имеем $4(n - 1)$ неизвестный коэффициент. Их нужно подобрать такими, чтобы выполнялись следующие условия:

- в пунктах $x_1, x_2, \ldots, x_n$ значения полиномов должны совпадать с измеренными значениями $\varphi_1, \varphi_2, \ldots, \varphi_n$;
- в пунктах стыковки соседних полиномов $x_2, x_3, \ldots, x_{n-1}$ не должно быть изломов, т.е. наклоны линий должны быть одинаковыми, что сводится к равенству первых производных от полиномов (5.9);
- в тех же пунктах стыковки не должно быть скачка кривизны, что соответствует равенству вторых производных от полиномов (5.9);
- в начальном и конечном пунктах должны быть заданы граничные условия.

Нахождение коэффициентов a, b, c, d сводится к решению системы $4(n-1)$ уравнения первой степени, которые получаются следующим образом.

Каждое значение полинома должно совпадать с левым измеренным значением каждого отрезка пространственной переменной, что дает $n-1$ уравнение:

$$\begin{cases} a_1 + b_1 x_1 + c_1 x_1^2 + d_1 x_1^3 = \varphi_1; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n-1} + b_{n-1} x_{n-1} + c_{n-1} x_{n-1}^2 + d_{n-1} x_{n-1}^3 = \varphi_{n-1}. \end{cases}$$

Каждое значение полинома должно совпадать с правым измеренным значением каждого отрезка пространственной переменной, что дает еще $n-1$ уравнение:

$$\begin{cases} a_1 + b_1 x_2 + c_1 x_2^2 + d_1 x_2^3 = \varphi_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n-1} + b_{n-1} x_n + c_{n-1} x_n^2 + d_{n-1} x_n^3 = \varphi_n. \end{cases}$$

В $n-2$ пунктах стыковки должны быть равны первые производные соседних полиномов ($\varphi' = b + 2cx + 3dx^2$) – равенство касательных:

$$\begin{cases} b_1 + 2c_1 x_2 + 3d_1 x_2^2 = b_2 + 2c_2 x_2 + 3d_2 x_2^2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_{n-2} + 2c_{n-2} x_{n-1} + 3d_{n-2} x_{n-1}^2 = b_{n-1} + 2c_{n-1} x_{n-1} + 3d_{n-1} x_{n-1}^2. \end{cases}$$

В тех же $n-2$ пунктах стыковки должны быть равны вторые производные соседних полиномов ($\varphi'' = 2c + 6dx$) – равенство кривизны:

$$\begin{cases} 2c_1 + 6d_1 x_2 = 2c_2 + 6d_2 x_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 2c_{n-2} + 6d_{n-2} x_{n-1} = 2c_{n-1} + 6d_{n-1} x_{n-1}. \end{cases}$$

Нужны еще два уравнения, учитывающие граничные условия в конечных пунктах x_1 и x_n. Чаще других применяется условие, что в конечных пунктах кривизна нулевая (концы упругой линейки не закреплены), т.е. вторые производные должны быть нулевыми:

$$\begin{cases} 2c_1 + 6d_1 x_1 = 0; \\ 2c_{n-1} 6d_{n-1} x_n = 0. \end{cases}$$

Могут быть заданы и другие условия, например наклон касательной (т.е. первой производной) в крайних пунктах.

В результате получится необходимое число уравнений $4(n-1)$. Решение системы уравнений дает все необходимые коэффициенты полиномов, что позволяет рассчитывать (прогнозировать) значения пространственной переменной между пунктами измерений.

При большом числе пунктов измерений получается довольно громоздкая система уравнений первой степени, и, хотя ее решение на компьютере не составляет труда, объем вычислений довольно значителен.

Чтобы избежать большого объема вычислений, рекомендуется применять *скользящий сплайн*. Для скользящего сплайна берут четыре или шесть смежных пунктов измерений, по ним рассчитывают уравнения сплайна, а для построения и прогнозирования выбирают один полином, занимающий среднее положение. Потом осуществляется перемещение на один пункт измерения (скольжение) и расчет повторяется. Эти операции доводят до конечного пункта измерений. Исключение делается для первого отрезка, для которого используют первое уравнение, и для последнего отрезка, для которого используют последнее уравнение.

⏩ **Пример 5.4**. По данным табл.5.1 необходимо рассчитать скользящий сплайн.

Возьмем первые шесть пунктов измерений и составим необходимую систему уравнений. Между шестью пунктами имеются пять отрезков, следовательно, нужно рассчитать пять кубических полиномов, содержащих 20 коэффициентов.

Первое условие – все полиномы должны проходить через левую точку отрезков:

$$a_{11} + a_{12} + a_{13} + a_{14} = 2{,}2;$$
$$a_{21} + 7{,}0a_{22} + 7{,}0^2a_{23} + 7{,}0^3a_{24} = 1{,}9;$$
$$a_{31} + 20{,}1a_{32} + 20{,}1^2a_{33} + 20{,}1^3a_{34} = 2{,}5;$$
$$a_{41} + 28{,}6a_{42} + 28{,}6^2a_{43} + 28{,}6^3a_{44} = 4{,}2;$$
$$a_{51} + 34{,}8a_{52} + 34{,}8^2a_{53} + 34{,}8^3a_{54} = 3{,}8.$$

Все полиномы должны проходить через правую точку отрезков:

$$a_{11} + 7{,}0a_{12} + 7{,}0^2a_{13} + 7{,}0^3a_{14} = 1{,}9;$$
$$a_{21} + 20{,}1a_{22} + 20{,}1^2a_{23} + 20{,}1^3a_{24} = 2{,}5;$$
$$a_{31} + 28{,}6a_{32} + 28{,}6^2a_{33} + 28{,}6^3a_{34} = 4{,}2;$$
$$a_{41} + 34{,}8a_{42} + 34{,}8^2a_{43} + 34{,}8^3a_{44} = 3{,}8;$$
$$a_{51} + 44{,}0a_{52} + 44{,}0^2a_{53} + 44{,}0^3a_{54} = 3{,}1.$$

Второе условие – на стыках соседних отрезков должны быть равны первые производные:

$$a_{12} + 2\cdot7{,}0a_{13} + 3\cdot7{,}0^2a_{14} = a_{22} + 2\cdot7{,}0a_{23} + 3\cdot7{,}0^2a_{24};$$
$$a_{22} + 2\cdot28{,}6a_{23} + 3\cdot28{,}6^2a_{24} = a_{32} + 2\cdot28{,}6a_{33} + 3\cdot28{,}6^2a_{34};$$
$$a_{32} + 2\cdot34{,}8a_{33} + 3\cdot34{,}8^2a_{34} = a_{42} + 2\cdot34{,}8a_{43} + 3\cdot34{,}8^2a_{44};$$
$$a_{42} + 2\cdot44{,}0a_{43} + 3\cdot44{,}0^2a_{44} = a_{52} + 2\cdot44{,}0a_{53} + 3\cdot44{,}0^2a_{54}.$$

Третье условие – на стыках соседних отрезков должны быть равны вторые производные:

$$2a_{13} + 6\cdot7{,}0a_{14} = 2a_{23} + 6\cdot7{,}0a_{24};$$
$$2a_{23} + 6\cdot28{,}6a_{24} = 2a_{33} + 6\cdot28{,}6a_{34};$$
$$2a_{33} + 6\cdot34{,}8a_{34} = 2a_{43} + 6\cdot34{,}8a_{44};$$
$$2a_{43} + 6\cdot44{,}0a_{44} = 2a_{53} + 6\cdot44{,}0a_{54}.$$

Четвертое условие – на свободных концах первого и последнего отрезков примем граничные условия – вторые производные равны нулю:

$$2a_{13} + 6\cdot 0{,}0a_{14} = 0;$$

$$2a_{53} + 6\cdot 44{,}0a_{54} = 0.$$

Всего получено 20 уравнений первой степени с 20 неизвестными коэффициентами. Решение системы дает значения коэффициентов:

$a_{11} = 2{,}200000$; $a_{12} = -0{,}069706$; $a_{13} = 0{,}005753$; $a_{14} = 0{,}000274$;

$a_{21} = 1{,}955657$; $a_{22} = 0{,}035012$; $a_{23} = -0{,}009206$; $a_{24} = 0{,}000438$;

$a_{31} = 21{,}556675$; $a_{32} = -2{,}890512$; $a_{33} = 0{,}136342$; $a_{34} = -0{,}001975$;

$a_{41} = -73{,}310478$; $a_{42} = 7{,}060588$; $a_{43} = -0{,}211599$; $a_{44} = 0{,}002080$;

$a_{51} = 22{,}811948$; $a_{52} = 1{,}225829$; $a_{53} = 0{,}026517$; $a_{54} = -0{,}000201$.

Далее происходит перемещение исходных измерений на один шаг, все расчеты повторяют. Перемещение доводится до конца исходных данных. Конечный результат виден на рис.5.4. Очевидно преимущество полученного графика перед графиками рис.5.2 и 5.3. ◀◀

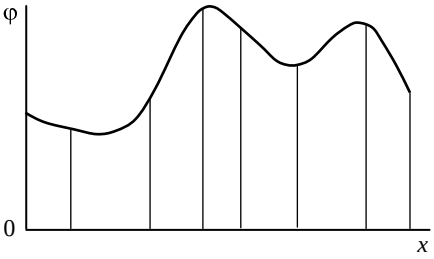

Рис.5.4. График скользящего сплайна, построенный по данным табл.5.1

Более сложным является расчет двухмерного сплайна. В литературе не удалось найти полное решение двухмерного сплайна для произвольно расположенных пунктов измерений. Имеющиеся публикации рассчитаны только на прямоугольную сеть, причем вычисляются сплайны в двух перпендикулярных направлениях, а в промежутках между ними используется интерполяция.

Теоретически можно построить и трехмерный сплайн, но на практике его, по-видимому, не применяют.

5.4. ВЕРОЯТНОСТНЫЕ МОДЕЛИ ГЕОЛОГИЧЕСКИХ ПОЛЕЙ

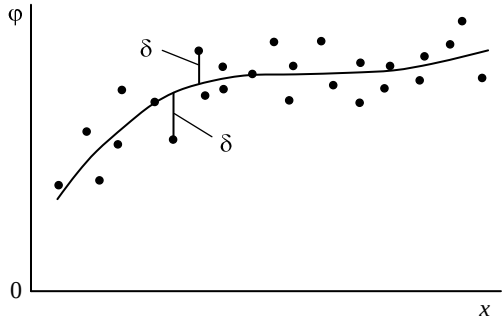

Рис.5.5. График случайной функции

5.4.1. Модель на основе случайной функции

Основой случайной функции служит предположение, что измеренные значения являются случайными функциями координат и содержат две составляющие: математическое ожидание $m(x)$ (закономерная изменчивость, или тренд) и случайные колебания $\delta(x)$ относительно его [см. формулу (5.2)].

Если математическое ожидание – величина постоянная, то случайная функция называется *стационарной*, в противном случае – *нестационарной*. Математическое ожидание позволяет прогнозировать значения пространственной переменной между пунктами измерений, тогда как случайные колебания служат для оценки погрешности прогнозирования.

Стационарная случайная функция может обладать еще одним свойством. Если на любом ее отрезке характеристики одинаковые, то функция *эргодичная*, что редко используется в геологической практике.

Измеренные значения в отдельных точках называются *реализациями* случайной функции. Случайную функцию можно изобразить на графике, на котором точки пунктов измерений имеют случайные отклонения δ от плавной линии математического ожидания $m(x)$ (рис.5.5). Отклонения бывают положительные, отрицательные и нулевые.

Случайная функция имеет три главные характеристики: математическое ожидание, дисперсию случайных колебаний и автокорреляционную функцию.

Математическое ожидание может рассматриваться как тренд, заданный на основе теоретических соображений (зависимость плотности от состава руды, кривая радиоактивного распада) или эм-

пирическим способом, чаще всего в виде полинома. Эмпирический полином является приближенной оценкой математического ожидания. Вычисление тренда осуществляется по методу наименьших квадратов (см. подраздел 3.1.5), а наилучший порядок полинома находят согласно подразделу 4.1.3. Возможен еще один метод оценки математического ожидания путем сглаживания исходных данных способом скользящего окна.

Математическое ожидание стационарной случайной функции, как отмечалось, величина постоянная и равная среднеарифметическому из всех измеренных значений. Если из нестационарной случайной функции вычесть математическое ожидание, то, согласно формуле (5.2), она превратится в стационарную с нулевым математическим ожиданием. Во многих случаях математическое ожидание (закономерная изменчивость, или тренд) слабо проявлено, тогда им пренебрегают, полагая случайную функцию стационарной.

Дисперсия случайной функции равна дисперсии отклонений $\delta(x)$:

$$D = \frac{1}{n}\sum_{i=1}^{n}\delta^2(x_i)\,. \qquad (5.10)$$

Если из дисперсии извлечь квадратный корень, то можно получить среднеквадратичное отклонение σ_δ.

Автоковариационная функция

$$K(h) = \frac{1}{m}\sum_{i=1}^{m}\delta(x_i+h)\delta(x_i)\,, \qquad (5.11)$$

где m – количество слагаемых под знаком суммы; h – шаг измерений.

Особый интерес представляет *автокорреляционная функция*

$$r(h) = K(h)/D\,. \qquad (5.12)$$

Она является аналогом коэффициента корреляции случайных величин, колеблется в пределах от –1 до +1 и характеризует зависи-

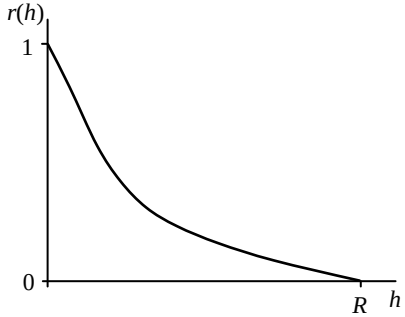

Рис.5.6. Идеальная форма автокорреляционной функции

мость между отклонениями δ на расстоянии *h*.

Автокорреляционная функция зависит от шага измерений *h*. При нулевом шаге она равна единице, при увеличении шага убывает, приближаясь к нулю. В идеальном виде функция показана на рис.5.6. Шаг, при котором автокорреляционная функция неотличима от нуля, называется *радиусом автокорреляции R*. Он является важной величиной, характеризующей радиус влияния отдельного измерения.

Следует отметить, что шаг и радиус автокорреляции являются векторными величинами. В изотропной среде радиус влияния одинаков по всем направлениям. В анизотропной среде, а геологические объекты – большей частью анизотропные тела, радиус автокорреляции зависит от направления. Чем сильнее проявлена изменчивость в каком-либо направлении, тем меньше радиус автокорреляции.

Таблица 5.3

Результаты измерения мощности рудного тела

Номер пункта	*x*, м	φ, м	Номер пункта	*x*, м	φ, м
1	0	1,8	16	30	2,1
2	2	1,5	17	32	1,9
3	4	1,6	18	34	1,6
4	6	1,8	19	36	1,4
5	8	1,9	20	38	1,3
6	10	2,2	21	40	1,7
7	12	2.5	22	42	1,9
8	14	2,4	23	44	2,1
9	16	2,4	24	46	2,2
10	18	2,6	25	48	1,5
11	20	2,4	26	50	1,8
12	22	2,2	27	52	2,0
13	24	1,8	28	54	2,3
14	26	1,7	29	56	1,8
15	28	1,9	30	58	1,5

Идеальная форма автокорреляционной функции искажается по нескольким причинам. На ее форму сильнее всего влияет периодическая изменчивость, придавая кривой волнистый характер. Если используется модель стационарной случайной функции, заметно сказывается монотонный тренд. В практических расчетах значения автокорреляционной функции находят по дискретным данным, а ее график имеет вид ломаной линии, состоящей из отдельных отрезков. В этих условиях в качестве радиуса автокорреляции принимается первое пересечение линии автокорреляции с осью абсцисс.

⏩ Пример 5.5. Имеются измерения мощности рудного тела, произведенные через 2 м (табл.5.3). Требуется рассчитать характеристики стационарной случайной функции.

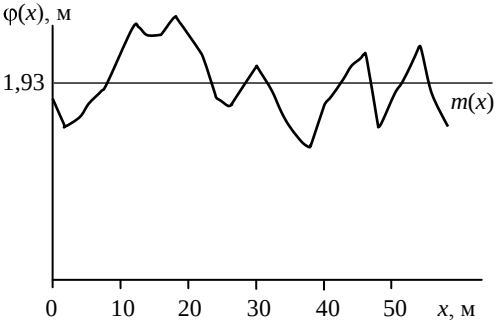

Рис.5.7. График изменения мощности рудного тела

Графический анализ (рис.5.7) показывает, что в поведении мощности отсутствует заметный тренд, поэтому можно принять модель стационарной случайной функции. Математическое ожидание стационарной случайной функции постоянно и равно среднему значению мощности $m(x) = = 1{,}93$ м. Дисперсия отклонений равна обычной дисперсии $D(x) = 0{,}117$. График автокорреляционной функции (рис.5.8) показывает, что имеются небольшие периодические колебания в исходных измерениях с длиной волны порядка 11 м, а радиус автокорреляции $R = 10{,}43$ м. ⏪

Изучение нестационарной случайной функции начинается с выделения математического ожидания. Обычно с помощью задан-

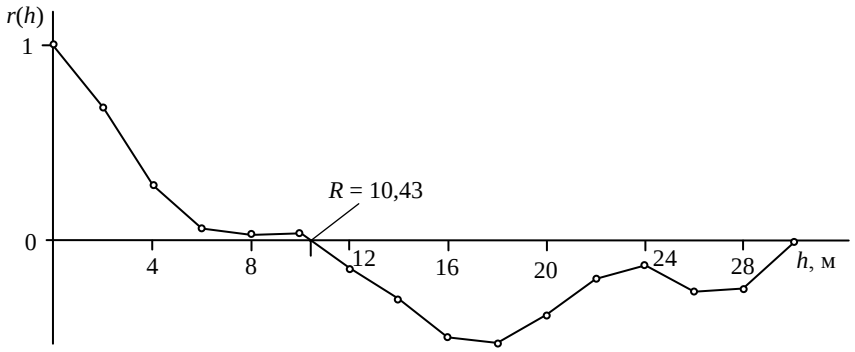

Рис.5.8. Эмпирический график автокорреляционной функции

Таблица 5.4

Значения пространственной переменной

Номер пункта	x, м	$f(x)$, м	Номер пункта	x, м	$f(x)$, м
1	8,6	9,5	11	50,5	33,4
2	9,8	11,6	12	55,2	27,9
3	11,5	16,3	13	61,5	32,1
4	13,8	14,8	14	62,5	24,0
5	20,0	25,0	15	63,9	26,0
6	24,0	19,8	16	69,2	31,0
7	25,5	24,0	17	73,1	23,2
8	29,2	28,5	18	79,6	19,5
9	35.8	28,0	19	82,2	22,1
10	43,8	31,4	20	87,7	15,8

ной функции (тренда) или путем сглаживания исходных данных с использованием скользящего окна удается выявить только оценку математического ожидания. Чаще всего задается аппроксимирующий полином, коэффициенты и порядок которого определяются, как было показано в подразделах 3.1.5 и 3.1.7. После вычитания из исходных данных тренда получается остаток, т.е. случайные отклонения, у которых определяют дисперсию отклонений. Важно отметить, что, в отличие от предыдущего расчета, для вычисления тренда не обязательна равномерная сеть наблюдений, но чтобы построить график исходных измерений, значения координаты x должны быть расположены в порядке возрастания.

⏩ **Пример 5.6.** Имеются результаты измерения переменной (табл.5.4). Требуется рассчитать наилучший полиномиальный тренд.

Для оценки математического ожидания выберем полином, рассчитанный по методу наименьших квадратов. Оптимальный порядок полинома определим по методике, описанной в подразделе 3.1.7. Наилучшим оказался полином третьего порядка. Его уравнение имеет вид

$$m(x) = 0{,}05178 + 1{,}41332x - 0{,}018784x^2 + 0{,}0000055x^3.$$

После вычитания из исходных данных полинома получим остаток – случайные отклонения, которые колеблются около оси абсцисс (рис.5.9). Дисперсия случайных отклонений с учетом использованных степеней свободы $D = 0{,}447026$. ◀◀

На практике часто используют двухмерный (площадной) тренд, который аппроксимируют двухмерным полиномом невысокого порядка (не более третьего), хотя теоретически можно использовать методику отыскания наилучшего порядка полинома. Тренды высокого порядка требуют большого объема вычислений и часто дают нереальные значения между пунктами измерений.

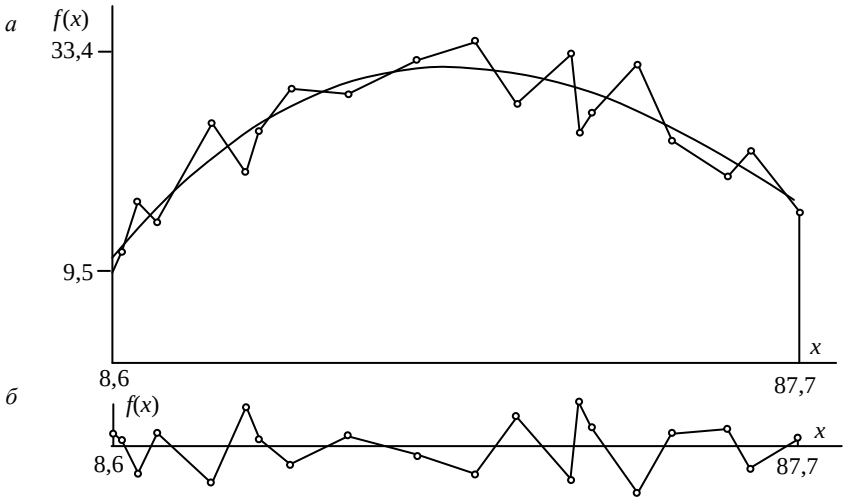

Рис.5.9. Аппроксимация исходных данных (ломаная линия) трендом (*а*) и остаток от тренда (*б*)

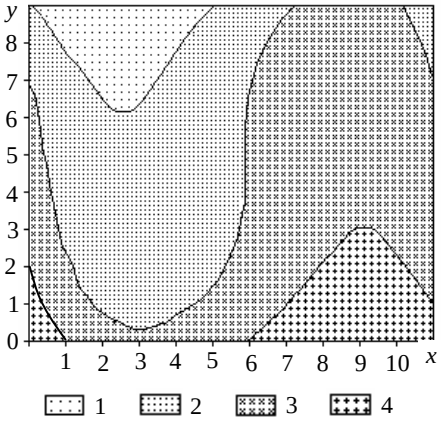

Рис.5.10. Тренд третьего порядка содержания молибдена

1 – 0,1-0,2 %; 2 – 0,2-0,3 %; 3 – 0,3-0,4 %; 4 – 0,4-0,5 %

▸▸ Пример 5.7. Имеются данные по опробованию штокверка молибденового месторождения на одном из горизонтов (табл.5.5). Требуется построить тренд содержания молибдена.

Таблица 5.5

Содержание молибдена, %

Номер пункта	0	1	2	3	4	5	6	7	8	9	10	11
0	0,12	0,04	0,10	0,13	0,11	0,17	0,14	0,32	0,66	0,28	0,25	0,33
1	0,15	0,51	0,19	0,09	0,09	0,20	0,21	0,31	0,35	0,32	0,26	0,35
2	0,07	0,54	0,27	0,17	0,11	0,16	0,42	0,67	0,23	0,35	0,29	0,28
3	0,22	0,34	0,24	0,25	0,07	0,20	0,44	0,46	0,24	0,36	0,24	0,27
4	0,21	0,28	0,37	0,18	0,13	0,22	0,57	0,48	0,20	0,28	0,25	0,25
5	0,10	0,15	0,17	0,20	0,30	0,15	0,50	0,43	0,41	0,66	0,41	0,33
6	0,15	0,08	0,13	0,25	0,64	0,21	0,16	0,19	0,34	0,62	0,44	0,36
7	0,78	0,17	0,14	0,19	0,25	0,40	0,27	0,21	0,23	0,58	0,54	0,33
8	0,58	0,57	0,17	0,20	0,28	0,64	0,77	0,23	0,23	0,33	0,43	0,51
9	0,27	0,22	0,19	0,21	0,25	0,26	0,42	0,42	0,24	0,34	0,37	0,49

Обозначим ось абсцисс x, ось ординат y. Для тренда применим двухмерный полином третьего порядка:

$$m(x,y) = a_0 + a_1x + a_2y +$$
$$+ a_3x^2 + a_4xy + a_5y^2 + a_6x^3 +$$
$$+ a_7x^2y + a_8xy^2 + a_9y^3.$$

Вычислим тренд методом наименьших квадратов. В результате расчета найдем коэффициенты тренда:

$a_0 = 0{,}504237;$

$a_1 = -0{,}130796;$

$a_2 = -0{,}0669908;$

$a_3 = 0{,}0279553;$

$a_4 = 0{,}00381132;$ $a_5 = 0{,}0108503;$

$a_6 = -0{,}0014979;$ $a_7 = -0{,}000349273;$

$a_8 = 0{,}0000063219;$ $a_9 = -0{,}000772421.$

Используя коэффициенты, построим график, характеризующий тренд содержания молибдена на горизонте (рис.5.10). ◀◀

Еще один способ приближенной оценки математического ожидания основан на методе сглаживания исходных данных с помощью скользящего окна. Его часто называют сглаживающим фильтром и используют для выделения полезного сигнала на фоне случайных помех. По существу, сглаженные данные характеризуют не математическое ожидание, а тенденцию изменения пространственной переменной. Сглаживание – простая операция, не требующая больших вычислений. Существует много способов сглаживания. Наиболее часто сглаживание осуществляется скользящим окном, содержащим три соседних наблюдения. По этим трем наблюдениям находят среднеарифметическое значение, которое сопоставляют с серединой окна. Потом окно передвигают на одно наблюдение, расчет повторяют и так поступают до конца ряда измерений.

Могут быть использованы окна с различным нечетным числом измерений. Роль измерений в окне также может быть различной – центральным значениям чаще придают бóльший вес. В случае значительного разброса исходных данных хороший результат дает медианное сглаживание, когда в окне в качестве среднего значения используют медиану.

▶▶ **Пример 5.8.** Имеются исходные данные из 20 измерений (табл.5.6). Требуется выполнить сглаживание данных.

Таблица 5.6

Сглаживание исходных данных

Номер пункта	Исходные данные		Первое сглаживание		Второе сглаживание		Третье сглаживание	
	x	$f(x)$	f_1	δ_1	f_2	δ_2	f_3	δ_3
1	8,6	9,5	10,2	0,7	11,0	1,5	11,4	1,9
2	9,8	11,6	12,5	0,9	12,3	0,7	12,8	1,2
3	11,5	16,3	14,2	–2,1	15,1	–1,2	15,0	–1,3
4	13,8	14,8	18,7	3,9	17,6	2,8	17,7	2,9
5	20,0	25,0	19,9	–5,1	20,5	–4,5	20,1	–4,9
6	24,0	19,8	22,9	3,1	22,3	2,5	22,5	2,7
7	25,5	24,0	24,1	0,1	24,6	0,6	24,5	0,5
8	29,2	28,5	26,8	–1,7	26,7	–1,8	26,8	–1,7
9	35,8	28,0	29,3	1,3	29,0	1,0	28,7	0,7
10	43,8	31,4	30,9	–0,5	30,4	–1,0	30,1	–1,3
11	50,5	33,4	30,9	–2,5	31,0	–2,4	30,5	–2,9
12	55,2	27,9	31,1	3,2	30,0	2,1	29,9	2,0
13	61,5	32,1	28,0	–4,1	28,8	–3,3	28,8	–3,3
14	62,5	24,0	27,4	3,4	27,5	3,5	27,8	3,8
15	63,9	26,0	27,0	1,0	27,0	1,0	26,9	0,9
16	69,2	31,0	26,7	–4,3	26,1	–4,9	26,1	–4,9
17	73,1	23,2	24,6	1,4	24,3	1,1	24,1	0,9
18	79,6	19,5	21,6	2,1	21,8	2,3	21,9	2,4
19	82,2	22,1	19,1	–3,0	19,5	–2,6	19,9	–2,2
20	87,7	15,8	17,9	2,1	18,3	2,5	18,7	2,9
Дисперсия		45,9	-	7,28	-	6,10	-	6,71

Для сглаживания выберем окно размером в три измерения. Возьмем первые три значения $f(x)$ и вычислим среднее значение:

$$f_1 = (9{,}5 + 11{,}6 + 16{,}3) : 3 = 12{,}5,$$

получим сглаженное значение во второй строке таблицы. Переместим окно на одно измерение и вычислим следующее значение:

$$f_1 = (11{,}6 + 16{,}3 + 14{,}8) : 3 = 14{,}2.$$

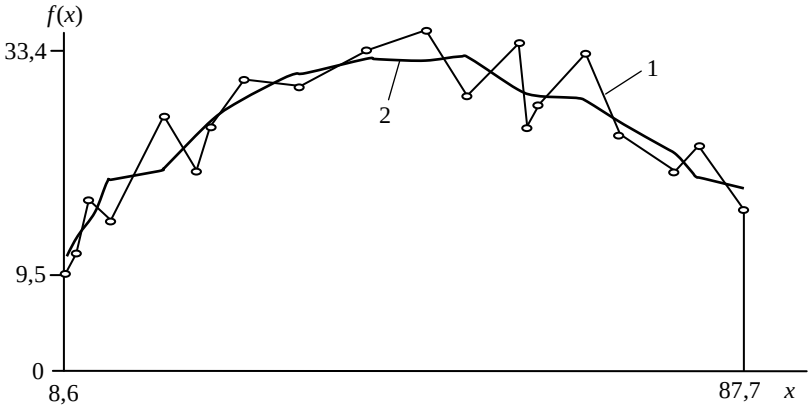

Рис.5.11. График исходных данных (линия 1)
и сглаженных значений (линия 2)

Продолжая перемещать окно, аналогично найдем остальные сглаженные значения. Исключение составляют крайние сглаженные значения, при расчете которых крайние значения используются дважды.

Когда первое сглаживание закончено, рассчитаны отклонения сглаженных значений от исходных значений δ_1 и найдена дисперсия отклонений (7,28), можно выполнить второе сглаживание и снова найти отклонения сглаженных значений от исходных значений $f(x)$, а также дисперсию отклонений (6,10). Сглаживание повторяется и оканчивается после достижения минимума дисперсии отклонений. В данном примере минимум дисперсии достигнут после второго сглаживания (табл.5.5). На рис.5.11 совмещены исходные и сглаженные значения. ◀◀

5.4.2. Гармонический анализ

Как упоминалось выше, в составе закономерной изменчивости часто присутствует периодическая составляющая. Ее можно выделить и вычесть из исходных данных с помощью гармонического или периодограммного анализа. Гармонический анализ позволяет ряд исходных данных представить как сумму синусоид. Но здесь, в

отличие от классического гармонического анализа, необходимо выделить наиболее существенные синусоиды, которые вносят основной вклад в изменчивость пространственной переменной.

Периодические явления широко распространены в природе, они находят свое выражение в ритмичном напластовании горных пород, в развитии систем упорядоченных трещин, в формировании рудных столбов. Чтобы выявить и охарактеризовать периодичность, необходимо применить специальные математические приемы.

Предлагаемая методика характеризуется тем, что в ряду наблюдений последовательно находят синусоиды в порядке убывания их значимости и вычитают из исходных данных. Можно извлечь несколько (иногда одну) важнейших синусоид и этим ограничиться. Но можно извлекать синусоиды до тех пор, пока дисперсия отклонений суммы синусоид от исходных данных с учетом использованных степеней свободы не будет минимальной.

Каждая синусоида имеет три характеристики: *амплитуду A*, *длину волны L* (или обратную ей величину – *частоту*) и *начальную фазу* ψ, поэтому каждая синусоида использует три степени свободы. Кроме того, часто находят *дисперсию* синусоиды, которая пропорциональна квадрату амплитуды. Ряд значений пространственной переменной $f(x)$ представляют в виде суммы N гармоник (синусоид) $\omega_k(x)$:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{N} \omega_x(x) , \qquad (5.13)$$

где $a_0 / 2$ – свободный член, равный среднеарифметическому значению пространственной переменной.

В качестве первой длины волны L чаще всего принимается длина ряда наблюдений. Каждая гармоника $\omega_k(x)$ выражается формулой

$$\omega_k(x) = a_k \cos(2\pi k \frac{x}{L}) + b_k \sin(2\pi k \frac{x}{L}) \qquad (5.14)$$

или в другой форме записи:

$$\omega_k(x) = A_k \sin(2\pi k \frac{x}{L} + \psi_k) . \qquad (5.15)$$

Здесь a_k, b_k – коэффициенты; k – номер гармоники (синусоиды); L/k – длина волны; A_k – амплитуда; ψ_k – начальная фаза гармоники; $k = 1, 2, \ldots, N$.

Амплитуда и начальная фаза связаны с коэффициентами a_k, b_k соотношениями

$$A_k = \sqrt{a_k^2 + b_k^2}; \qquad \psi_k = \operatorname{arctg}(a_k / b_k). \tag{5.16}$$

Количество гармоник N может быть большим, но на практике принимается конечным. Количество гармоник не должно превышать $n/2$ (где n – количество измерений), но наиболее правильно ограничивать их количество по минимальной дисперсии отклонений с учетом степеней свободы, как в подразделе 3.1.7 или 4.1.3. Каждая синусоида использует три степени свободы, поэтому k синусоид поглощают $3k$ степеней свободы. Впрочем, для геологических целей обычно выбирают одну или две важнейшие синусоиды, пренебрегая остальными. За счет ограничения количества синусоид их сумма в формуле (5.13) приближенно равна пространственной переменной $f(x)$.

Объединение формул (5.13) и (5.14) дает ряд Фурье (гармонический ряд):

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{N}\left[a_k \cos(2\pi k \frac{x}{L}) + b_k \sin(2\pi k \frac{x}{L}) \right]. \tag{5.17}$$

Коэффициенты a_k, b_k находят по формулам

$$a_k = \frac{2}{n}\sum_{i=1}^{n} f(x_i)\cos(2\pi k \frac{x_i}{L}); \tag{5.18}$$

$$b_k = \frac{2}{n}\sum_{i=1}^{n} f(x_i)\sin(2\pi k \frac{x_i}{L}). \tag{5.19}$$

Когда коэффициенты a_k и b_k определены, вычисляют амплитуды гармоник по формуле (5.16). Совокупность всех гармоник образует спектр амплитуд. Чем больше амплитуда, тем бóльшую роль играет соответствующая гармоника. Для количественной оценки роли гармоник используют их дисперсии D_k, совокупность которых составляет спектр дисперсий. Сумма дисперсий всех гармоник равна дисперсии пространственной переменной, что позволяет оценить роль каждой гармоники в абсолютных или относительных единицах.

Вычисление дисперсии любой гармоники лучше всего осуществлять путем сравнения остаточных дисперсий до вычитания и после вычитания гармоники из значений пространственной переменной. Уменьшение остаточной дисперсии характеризует дисперсию D_k, поглощенную данной гармоникой.

⏩ **Пример 5.9.** Имеются результаты 25 измерений пространственной переменной $f(x)$ – содержаний цинка в скважине по сети с шагом 1 м (табл.5.7). Очевидно, что длина отрезка (начальная длина волны) $L = 25 - 1 = 24$ м. Необходимо разложить пространственную переменную в ряд Фурье и оценить роль различных гармоник.

Задачу решим путем последовательного вычисления и вычитания гармоник из значений пространственной переменной (кстати, наиболее рациональный метод при неравномерной сети наблюдений). Вначале из значений пространственной переменной вычтем свободный член ряда (5.17), что даст отклонения $\delta_0(x) = f(x) - a_0/2$. Далее из отклонений $\delta_0(x)$ вычтем первую гармонику $\omega_1(x)$, что даст остаток $\delta_1(x)$. Из него вычтем вторую гармонику и этот процесс продолжим до заданного числа гармоник $N = 12$. Начало вычислений, включая первые две гармоники, показано в табл.5.7. Коэффициенты a_k и b_k каждой очередной гармоники найдем через отклонения $\delta(x)$ по преобразованным формулам (5.18) и (5.19):

$$a_k = \frac{2}{n}\sum_{i=1}^{n}\delta_{k-1}(x_i)\cos(kt_i)\,; \qquad (5.20)$$

$$b_k = \frac{2}{n}\sum_{i=1}^{n}\delta_{k-1}(x_i)\sin(kt_i)\,, \qquad (5.21)$$

где $t_i = 2\pi[(x_i - x_1)/(x_n - x_1)]$ – вспомогательный аргумент, вводимый для сокращения вычислений.

Таблица 5.7

Результаты измерений пространственной переменной и некоторые вычисленные данные

Номер	Исходные	Среднее	Откло-	Первая	Откло-	Вторая	Откло-

пункта n	измерения $f(x)$	значение $a_0/2$	нение $\delta_0(x)$	гармоника $\omega_1(x)$	нение $\delta_1(x)$	гармоника $\omega_2(x)$	нение $\delta_2(x)$
1	2,36	3,14	−0,78	0,34	–1,12	−0,38	−0,74
2	2,08	3,14	−1,06	0,31	–1,37	−0,30	–1,07
3	1,95	3,14	–1,19	0,26	–1,45	−0,14	–1,31
4	7,68	3,14	4,54	0,20	4,34	0,05	4,29
5	5,72	3,14	2,58	0,12	2,46	0,23	2,23
6	3,08	3,14	−0,06	0,03	−0,09	0,35	−0,44
7	2,20	3,14	−0,94	−0,06	−0,88	0,38	−1,26
8	2,76	3,14	−0,38	−0,14	−0,24	0,30	−0,54
9	1,32	3,14	–1,82	−0,22	–1,60	0,14	–1,74
10	1,02	3,14	–2,12	−0,28	–1,84	−0,05	–1,79
11	3,24	3,14	0,10	−0,32	0,42	−0,23	0,65
12	4,90	3,14	1,76	−0,34	2,10	−0,35	2,45
13	4,48	3,14	1,34	−0,34	1,68	−0,38	2,05
14	2,30	3,14	−0,84	−0,31	−0,53	−0,30	−0,23
15	1,97	3,14	–1,17	−0,26	−0,91	−0,14	−0,77
16	2,27	3,14	−0,87	−0,20	−0,67	0,05	−0,73
17	2,08	3,14	–1,06	−0,12	−0,94	0,23	–1,18
18	1,45	3,14	–1,69	−0,03	–1,66	0,35	–2,01
19	3,85	3,14	0,71	0,06	0,65	0,38	0,28
20	5,18	3,14	2,04	0,14	1,90	0,30	1,60
21	6,34	3,14	3,20	0,22	2,98	0,14	2,84
22	3,93	3,14	0,79	0,28	0,51	−0,05	0,57
23	2,33	3,14	−0,81	0,32	–1,13	−0,23	−0,89
24	2,11	3,14	–1,03	0,34	–1,37	−0,35	–1,02
25	1,90	3,14	–1,24	0,34	–1,58	−0,38	–1,20
Дисперсия			2,790	-	2,734	-	2,666

Отклонения $\delta(x)$ дают возможность вычислить остаточные дисперсии:

$$D(\mathrm{o},k)=\frac{1}{n}\sum_{i=1}^{n}\delta_k^2(x_i)\,, \qquad (5.22)$$

которые образуют убывающий ряд.

Разность соседних остаточных дисперсий дает дисперсию текущей гармоники k:

$$D(k) = D(\text{o}, k-1) - D(\text{o}, k). \qquad (5.23)$$

Ее можно выразить в процентах от дисперсии пространственной переменной D = 2,790, принимаемой за 100 %.

Для вычисления гармоник, часть которых приведена в табл.5.7, применим преобразованную формулу (5.14):

$$\omega_k(x) = a_k \cos(kt) + b_k \sin(kt). \qquad (5.24)$$

Таблица 5.8

Результаты расчета всех гармоник

Номер гармоники k	Длина волны L/k	Коэффициент		Амплитуда A	Дисперсия отклонений $D(\text{o},k)$	Дисперсия гармоник	
		a_k	b_k			$D(k)$	$D(k)$,%
0	–	3,140	0,000	0,000	2,790	0,000	0,0
1	24,0	0,335	–0,058	0,340	2,734	0,056	2,0
2	12,0	–0,376	0,054	0,380	2,665	0,070	2,5
3	8,0	–1,828	0,535	1,905	0,912	1,753	62,8
4	5,0	–0,154	–0,070	0,169	0,898	0,014	0,5
5	4,8	0,084	0,040	0,093	0,894	0,004	0,2
6	4,0	0,597	–0,883	1,066	0,317	0,577	20,7
7	3,4	0,469	0,042	0,471	0,210	0,107	3,8
8	3,0	0,241	0,178	0,300	0,166	0,044	1,6
9	2,7	–0,041	0,128	0,135	0,156	0,009	0,3
10	2,4	–0,146	0,263	0,301	0,110	0,046	1,7
11	2,2	–0,243	0,326	0,406	0,027	0,084	3,0
12	2,0	–0,088	0,000	0,088	0,027	0,000	0,0
Сумма	–	–	–	–	–	2,764	99,1

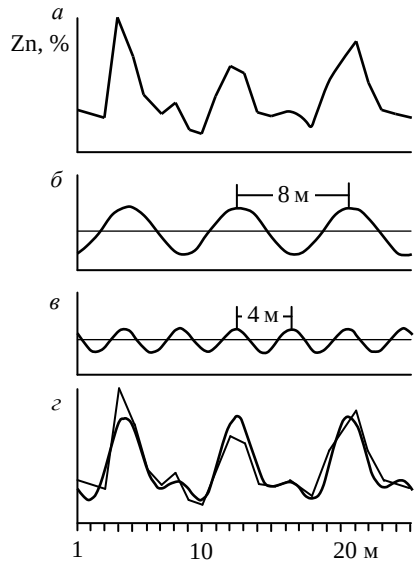

Рис.5.12. Изменчивость в распределении содержания цинка по скважине: *а* – исходные данные; *б*, *в* – периодическая изменчивость; *г* – сравнение исходной и периодической изменчивости

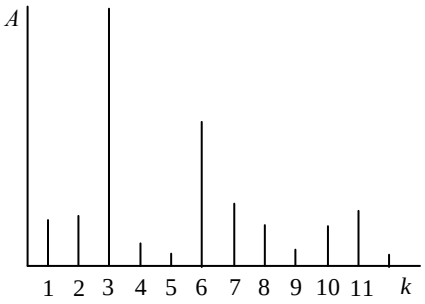

Рис.5.13. Спектр амплитуд

Коэффициенты a_k и b_k позволяют рассчитать амплитуды всех гармоник по формуле (5.16), а отклонения – их дисперсии по формуле (5.22).

Результаты расчетов по всем 12 гармоникам суммированы в табл.5.8. Отметим, что отношение L/k – это длина волны гармоники k. В таблице приведены длина волны, коэффициенты a_k и b_k, амплитуда A, дисперсии отклонений $D(\text{о}, k)$, их разности – дисперсии гармоник $D(k)$ и они же, выраженные в процентах от дисперсии пространственной переменной.

Вычисления показали, что главное значение имеют третья гармоника с длиной волны 8 м (рис.5.12, *б*), ее доля составляет 62,8 %, и шестая гармоника с длиной волны 4 м (рис.5.12, *в*), ее доля 20,7 %. Остальные гармоники играют малую роль.

Следовательно, можно ограничиться двумя важнейшими гармониками, которые в сумме дают 83,5 % от дисперсии пространственной переменной (рис.5.12, *г*).

Данные табл.5.8 позволяют построить спектр амплитуд (рис.5.13). ◀◀

5.4.3. Периодограммный анализ

Периодическая изменчивость не обязательно должна быть синусоидальной, она может иметь любую форму. Можно выявить периодическую изменчивость любой формы путем последовательного перебора длин волн, кратных шагу наблюдений. Такая методика, по-видимому, не описана в литературе. Предлагается назвать ее периодограммным анализом.

Если имеется ряд значений пространственной переменной $f(x)$, измеренных с шагом h, имеющих дисперсию D, то, задавая длину волны, можно выявить периодическую изменчивость $\omega(x)$ при данной длине волны и вычесть ее из исходных данных, что дает отклонение $\delta(x) = f(x) - \omega(x)$. Разность дисперсии исходных данных и дисперсии отклонений дает дисперсию, поглощенную периодической изменчивостью. Чем больше поглощенная дисперсия, тем сильнее проявлена периодическая изменчивость. Задавая различную длину волны, можно найти наибольшую дисперсию, что и лежит в основе периодограммного анализа.

Отыскание наилучшей длины волны начинается с наименьшей длины волны $L = h$. Затем длина волны последовательно увеличивается до достижения половины ряда наблюдений $N = n/2$.

Таблица 5.9

Расчет периодической изменчивости и отклонений при длине волны L = 5 м

Номер столбца				
1	2	3	4	5
Исходные данные				
2,36	2,08	1,95	7,68	5,72
3,08	2,20	2,76	1,32	1,02
3,24	4,90	4,48	2,30	1,97
2,27	2,08	1,45	3,85	5,18
6,34	3,93	2,33	2,11	1,90
Средние по столбцам (периодическая изменчивость)				
3,46	3,04	2,59	3,45	3,16
Отклонения от средних				
–1,10	–0,96	–0,64	4,23	2,56
–0,38	–0,84	0,17	–2,13	–2,14
–0,22	1,86	1,89	–1,15	–1,19
–1,19	–0,96	–1,14	0,40	2,02
2,88	0,89	–0,26	–1,34	–1,26

Дисперсия исходных данных 2,790
Дисперсия отклонений 2,689
Дисперсия периодической изменчивости 0,101

Таблица 5.10

Дисперсия различных длин волн

Длина волны, м	Дисперсия отклонений	Дисперсия периодической изменчивости	
		D	*D*, %
0	2,790	0,000	0,0
2	2,783	0,007	0,2
3	2,759	0,031	1,1
4	2,279	0,511	18,3
5	2,689	0,101	3,6
6	2,702	0,088	3,2
7	1,724	1,066	38,2
8	0,445	2,345	85,1
9	1,436	1,354	48,5
10	2,367	0,423	15,2
11	2,148	0,642	23,0
12	2,073	0,717	25,7

▸▸ Пример 5.10. Измерения пространственной переменной приведены в табл.5.7. Необходимо рассчитать периодическую изменчивость ряда значений путем перебора различных длин волн.

Покажем порядок расчета при длине волны $L = 5$ м. Исходные данные «разрежем» на части по пять измерений и запишем один под другим в пять столбцов (табл.5.9). Средние по столбцам – это периодическая изменчивость при длине волны $L = 5$ м. После вычитания из исходных данных периодической изменчивости получим отклонения. Далее рассчитаем их дисперсию и вычтем из дисперсии исходных данных.

Подобные расчеты выполнены для всех длин волн от $L = 2$ до $L = 12$ (табл.5.10). Наибольшая дисперсия периодической изменчивости 85,1 % установлена для длины волны 8 м.

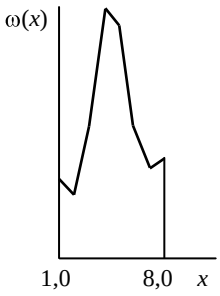

Рис.5.14. Одна волна периодической изменчивости

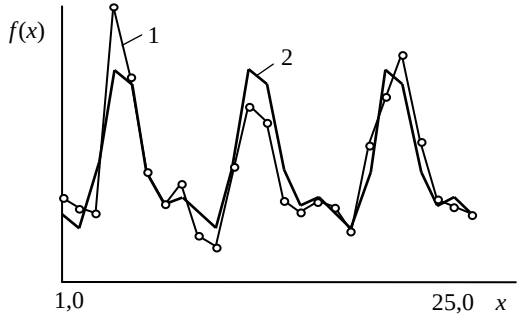

Рис.5.15. Совмещение исходных данных (линия 1) и периодической изменчивости (линия 2)

Наиболее сильно выраженная волна периодической изменчивости, показанная на рис.5.14, имеет следующий вид:

{1,91 1,52 3,01 5,92 5,51 3,10 2,17 2,38}.

Можно совместить график исходных данных с графиком периодической изменчивости и убедиться в удовлетворительном их совпадении (рис.5.15). Следует отметить, что дисперсия периодической изменчивости в данном случае несколько больше, чем при гармоническом анализе с использованием двух важнейших гармоник. ◀◀

5.5. ОСНОВЫ ГЕОСТАТИСТИКИ

5.5.1. Вариограмма и ее аппроксимации

В основе геостатистической группы математических моделей лежит гипотеза о том, что случайный результат измерений обусловлен случайным расположением сети наблюдений. При перемещении сети наблюдений результаты измерений будут другие, но сохраняется одна характеристика – средний квадрат разности между результатами измерений на расстоянии h. Это возможно при эргодичном характере пространственной переменной. На основе гипотезы введена *вариограмма* $\gamma(h)$ – главная характеристика в геостатистике. Она равна полусумме среднего квадрата разности между результатами измерений при шаге h и выражается формулой (5.4).

Следует отметить, что вариограмма тесно связана со случайными функциями. Сумма вариограммы и ковариации (автокорреляционной функции) равна дисперсии исходных данных:

$$\gamma(h) + K(h) = D. \qquad (5.25)$$

График вариограммы зависит от характера дискретности пространственной переменной. Для непрерывных пространственных переменных (например, мощности пласта) вариограмма начинается с нулевой отметки и возрастает до дисперсии исходных данных (рис.5.16). Для дискретных пространственных переменных вариограмма начинается с некоторой величины C (рис.5.17), называемой

эффектом самородков, потому что он был вначале установлен на месторождениях золота.

Вариограмма имеет радиус влияния *R*, который идентичен радиусу автокорреляции в случайной функции. За пределами радиуса влияния вариограмма постоянная и равна дисперсии *D*, т.е. ее влияние отсутствует. Как упоминалось, радиус *R* является векторной величиной. В изотропных геологических телах по всем направлениям радиус влияния описывает окружность, за пределами которой влияние вариограммы отсутствует. В анизотропных геологических телах значения радиуса *R* по разным направлениям столь различны, что напоминают восьмерку.

В пункте с координатой *R* вариограмма может иметь горизонтальную касательную и плавно переходить в линию дисперсии.

Эмпирическая вариограмма, получаемая на основе дискретных измерений, строится по отдельным точкам и имеет вид ломаной линии (рис.5.18). Чтобы использовать ее для дальнейших вычислений, необходимо выполнить аппроксимацию вариограммы какой-либо теоретической кривой. Вид аппроксимирующей функции определяет вид геостатистической модели. Существует много видов аппроксимирующих функций. Наибольшее распространение получили четыре функции и, соответственно, четыре геостатистические модели.

Первая модель линейная. В ней вариограмма аппроксимируется прямой линией (рис.5.19), что нередко близко к действительности. Во второй модели вариограмма аппроксимируется кубическим полиномом, который не имеет общей касательной с линией дисперсии (рис.5.20). Эта модель лучше всего соответствует действительности. Третья модель (сферическая) также аппроксимируется кубическим полиномом, но он имеет горизонтальную касательную при соприкосновении с линией дисперсии (см. рис.5.16 или рис.5.17). Четвертая модель (модель Де-Вийса) характеризуется тем, что шаг вариограммы откладывается в логарифмическом масштабе, а сама вариограмма представляется в виде прямой линии (рис.5.21).

Выбор варианта аппроксимации носит произвольный характер. Наилучшим вариантом можно считать тот, который дает наименьшую дисперсию отклонений эмпирических значений от теоретических.

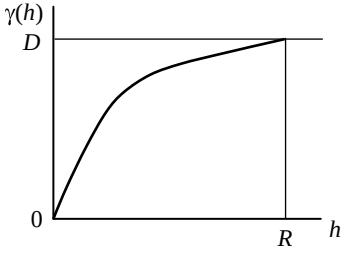

Рис.5.16. Теоретический вид вариограммы

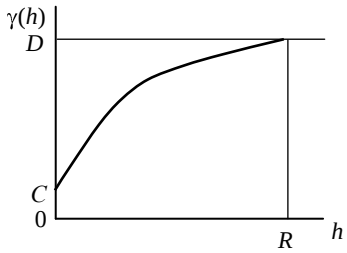

Рис.5.17. Теоретический вид вариограммы с эффектом самородков

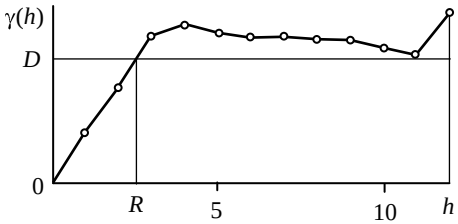

Рис.5.18. Эмпирическая вариограмма

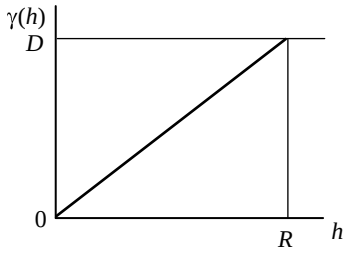

Рис.5.19. Линейная модель вариограммы

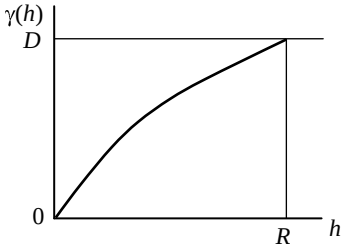

Рис.5.20. Полиномиальная модель вариограммы

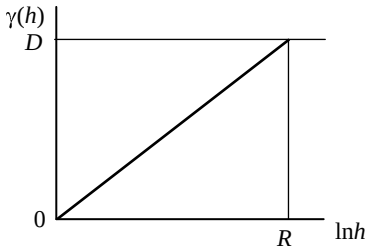

Рис.5.21. Модель Де-Вийса

На вид вариограммы отрицательно влияет периодическая изменчивость свойств пространственной переменной. Вариограмма колеблется около линии дисперсии, создавая так называемый эф-

фект включений. При сильно выраженной периодической изменчивости радиус влияния определяется неточно.

Таблица 5.11

Содержание золота в пробах, г/т

Пункт измерений x, м	Содержание золота в пробах $f(x)$, г/т	Пункт измерений x, м	Содержание золота в пробах $f(x)$, г/т
1	5,2	14	4,6
2	3,5	15	4,7
3	4,6	16	6,7
4	5,2	17	6,8
5	6,8	18	4,9
6	6,1	19	3,0
7	7,0	20	3,5
8	7,4	21	3,8
9	7,8	22	5,4
10	7,6	23	4,1
11	6,2	24	4,0
12	5,2	25	5,5
13	2,6	26	5,1

⏩ Пример 5.11. В горной выработке определено содержание золота через 1 м (табл.5.11, рис.5.22). Требуется построить вариограмму содержаний и аппроксимировать ее какой-либо моделью.

Эмпирическая вариограмма, вычисленная по формуле (5.4), показана на рис.5.18. Она содержит 13 точек, т.е. рассчитана до 12-го шага (половина длины ряда наблюдений).

Начальная часть вариограммы (первые четыре точки) расположена практически на одной линии, поэтому для построения теоретической вариограммы используем линейную модель, как на рис.5.19. Уравнение теоретической вариограммы, рассчитанное по четырем точкам, $\gamma(h) = 0{,}8h$. Радиус автокорреляции $R = 2{,}5$ м получен при пересечении линии теоретической вариограммы с линией дисперсии исходных данных $D = 2{,}03$.

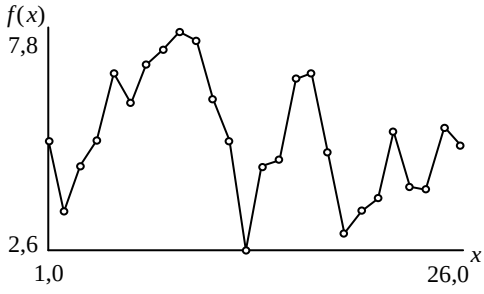

Рис.5.22. График исходных данных

5.5.2. Влияние на вариограмму геометрической базы измерений

Многие характеристики пространственной переменной зависят от геометрической базы измерений, т.е. от формы, размеров, а в анизотропных телах и от ориентировки области измерений. К таким характеристикам, в первую очередь, относятся дисперсия и вариограмма.

Если область, в которой производится измерение, настолько мала, что ее размерами можно пренебречь, то измерение рассматривается как точечное. Характеристики, полученные из таких измерений, рассматриваются как измерения на точечной геометрической базе.

Пусть в геологическом поле имеется множество точек, в каждой из которых пространственная переменная имеет значение $f(x)$. Если геологическое поле разделить на области объемом v и в каждом из них найти среднее из N точечных значений, то получим новую пространственную переменную $f_v(x)$:

$$f_v(x) = \frac{1}{p_0} \sum_{m=1}^{N} h(m) f(x+m) \quad \text{при} \quad p_0 = \sum_{m=1}^{N} p(m), \qquad (5.26)$$

где $p(m)$ – весовая функция, зависящая от взаимного расположения точек в объеме v.

При увеличении размеров геометрической базы происходит усреднение значений пространственной переменной, соответственно, будут меняться некоторые статистические характеристики. При переходе к бесконечному множеству точек в объеме v выражение (5.26) преобразуется в интегральную форму:

$$f_v(x) = \frac{1}{p_0} \int_v p(m) f(x+m) dm \quad \text{при} \quad p_0 = \int_v p(m) dm. \qquad (5.27)$$

Нахождение средних значений $f_v(x)$ по точечным значениям $f(x)$ называется *регуляризацией* (сглаживанием) пространственной переменной в объеме v. Меняя объем v, будем получать различные регуляризованные пространственные переменные, соответственно, будут меняться и их характеристики.

Изменение дисперсии по мере увеличения объема v описывается формулой

$$D_v = D - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma(i-j), \qquad (5.28)$$

где D – дисперсия пространственной переменной на точечной базе; N – число точек в объеме v.

Вычитаемое в формуле (5.28) представляет собою среднее значение вариограммы, получаемое при всех возможных положениях точек i и j в объеме v. При переходе к бесконечному множеству точек формула (5.28) приобретает интегральный вид:

$$D_v = D - \frac{1}{v^2}\int_v dx\int_v \gamma(x-y)dy. \qquad (5.29)$$

Из формул (5.28) и (5.29) следует, что по мере увеличения объема v дисперсия пространственной переменной уменьшается.

Подобные формулы существуют и для вариограммы:

$$\gamma_v(h) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma(i-j+h) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma(i-j) \qquad (5.30)$$

или в интегральной форме

$$\gamma_v(h) = \frac{1}{v^2}\int_v dx\int_v \gamma(x-y+h)dy - \frac{1}{v^2}\int_v dx\int_v \gamma(x-y)dy. \qquad (5.31)$$

Первое слагаемое в этих формулах характеризует среднюю ковариограмму, которая получается как средняя из вариограмм между всеми точками в объеме v, а вычитаемое – в таком же объеме, смещенном на расстояние h.

Формулы (5.28)-(5.31) выведены в предположении, что объем v мал по сравнению с объемом изучаемого геологического объекта, иначе сказывается граничный эффект, который можно учесть введением специальной весовой функции – геометрической ковариограммы $p(h)$:

$$p(h) = \frac{1}{v^2}\int_v k(x)k(x+h)dx, \qquad (5.32)$$

где $k(x) = 1$, когда точка x находится внутри геологического объекта, $k(x) = 0$, когда точка x находится за его пределами.

Геометрический смысл ковариограммы заключается в степени перекрытия двух объемов, смещенных относительно друг друга на расстояние h. С учетом функции $p(h)$ формулы (5.29) и (5.31) приобретают следующий вид:

$$D_v = D - \int_v p(x)\gamma(x)dx; \tag{5.33}$$

$$\gamma_v(h) = \int_v p(x)\gamma(x+h)dx - \int_v p(x)\gamma(x)dx. \tag{5.34}$$

Если вариограмма аппроксимирована каким-либо алгебраическим выражением, т.е. задан вид математической модели, то интегралы во многих случаях могут быть вычислены и заменены соответствующими алгебраическими выражениями.

⏩ **Пример 5.12.** Имеются исходные данные по содержанию золота в горной выработке (табл.5.11). Требуется показать изменение дисперсии и вариограммы при увеличении размера проб.

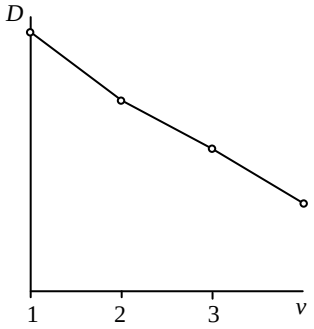

Рис.5.23. Уменьшение дисперсии D при увеличении геометрической базы измерений v

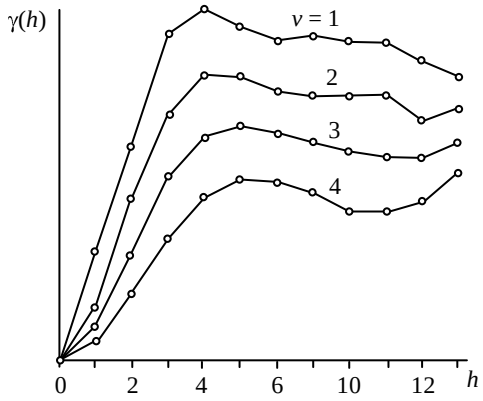

Рис.5.24. Изменение вариограммы $\gamma(h)$ при увеличении геометрической базы измерений v

Будем объединять и усреднять исходные данные по два, три и четыре соседних измерения путем сглаживания методом скользящего окна, что аналогично увеличению размеров проб исходных данных *v*. Количество исходных данных будет уменьшаться. Рассчитаем дисперсию *D* усредненных данных по формуле (2.2) и вариограмму $\gamma(h)$ по формуле (5.4) по каждому варианту усреднения. Результаты расчета изображены на рис.5.23 и 5.24. На рисунках видно, что при увеличении размера проб дисперсия уменьшается, по сравнению с рис.5.18 убывает также и вариограмма. ◀◀

5.5.3. Понятие о кригинге

На основе геостатистических моделей создан новый метод интерполяции результатов между пунктами измерений, который получил название *кригинг* (*крайгинг*) в честь автора метода Д.П.Крига [7]. Как показано Д.Матероном [12], кригинг дает минимальную сумму квадратов отклонений прогнозных значений от фактических значений пространственной переменной. Существует несколько видов кригинга. Вначале рассмотрим простейший – *точечный кригинг.*

Пусть имеется область *v*, в которой произведены измерения пространственной переменной по дискретной сети (рис.5.25). Каждое измерение считаем точечным, т.е. пренебрегаем геометрической базой измерений.

Необходимо рассчитать прогнозное значение пространственной переменной в точке *Z*. Предположим, что геологический объект изотропный, тогда вокруг точки *Z* можно провести окружность с радиусом влияния *R*. Все измерения внутри окружности влияют на значение пространственной переменной в точке *Z*. За пределами окружности результаты измерений в прогнозировании не участвуют.

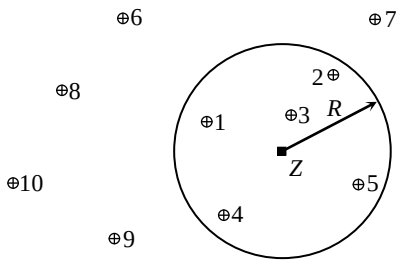

Рис.5.25. Схема точечного кригинга

Прогнозное значение находится в результате выполнения ряда последовательных операций:

1. Рассчитывается эмпирическая вариограмма, как на рис.5.18.

2. Осуществляется аппроксимация эмпирической вариограммы теоретической вариограммой (см. рис.5.19-5.21), т.е. выбирается геостатистическая модель.

3. В области с радиусом R рассчитываются все расстояния между пунктами измерений, а также между пунктами измерений и пунктом прогноза.

4. Составляется симметричная система уравнений кригинга следующего вида:

$$\begin{vmatrix} D & D-\gamma_{12} & D-\gamma_{13} & \dots & D-\gamma_{1m} & D-\gamma_{1z} \\ D-\gamma_{21} & D & D-\gamma_{231} & \dots & D-\gamma_{2m} & D-\gamma_{2z} \\ D-\gamma_{31} & D-\gamma_{32} & D & \dots & D-\gamma_{3m} & D-\gamma_{3z} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ D-\gamma_{m1} & D-\gamma_{m2} & D-\gamma_{m3} & \dots & D & D-\gamma_{mz} \end{vmatrix}. \qquad (5.35)$$

Здесь γ_{12} – значение теоретической вариограммы между пунктами 1 и 2 (учитывается расстояние между этими пунктами); m – количество пунктов, участвующих в расчете (в данном случае $m = 5$).

5. Решение системы дает весовые коэффициенты p_1, p_2, p_3, …, p_m.

6. Сумма весовых коэффициентов должна быть равна единице, для чего все коэффициенты делятся на их сумму.

7. С помощью весовых коэффициентов вычисляется прогнозное значение свойства в пункте Z:

$$Z = \sum_{i=1}^{m} p_i z_i \,, \qquad (5.36)$$

где z_i – значения свойства в пунктах с 1-го по m-й.

Если понадобится вычислить прогнозное значение в следующем пункте, то операции 3-6 придется выполнить снова. Если в пределах радиуса влияния не окажется ни одного пункта измерения, то кригинг нельзя применять. Приходится обращаться к другим

методам интерполяции данных, чаще всего используется метод обратных расстояний.

Иногда при составлении системы уравнений (5.35) в матрице коэффициентов оказывается много нулей (расстояния между точками больше радиуса автокорреляции). В этом случае система уравнений является несовместимой, т.е. не имеет решения (или имеет много решений). Тогда весовые коэффициенты лучше всего находить путем деления значений $D - \gamma_{1z}$, $D - \gamma_{2z}$, $D - \gamma_{3z}$, …, $D - \gamma_{mz}$ на дисперсию D и потом приводить сумму коэффициентов к единице.

У точечного кригинга имеется один недостаток – в окрестностях пунктов измерений касательная к прогнозным значениям ориентирована горизонтально, как в полиномиальной модели (см. рис.5.2). Чтобы устранить этот недостаток, применяется *универсальный кригинг* – комбинация тренда с кригингом. Вначале вычисляется тренд $f(x)$, потом остаток от тренда и по остатку осуществляется кригинг. Прогнозное значение находится по формуле

$$Z = f(x) + \sum_{i=1}^{m} p_i z_i. \qquad (5.37)$$

В качестве тренда используется аппроксимирующий полином не выше третьего порядка. Использование тренда частично снимает явление анизотропии, что позволяет применять изотропную схему кригинга.

Точечный кригинг можно распространить на объем v. Во всех точках объема с помощью кригинга рассчитывают прогнозное значение свойства, а потом находят среднее из всех значений. Если пункты измерений не точечные, а имеют какой-то объем, то при прогнозировании учитывают все точки в этом объеме. Практически все задачи с объемными геометрическими базами решаются путем вычисления определенных интегралов по этим объемам численными методами.

⏩ **Пример 5.13.** Нужно провести точечный кригинг, используя схему рис.5.25 и дополнив недостающие данные.

Примем изотропную линейную модель, радиус влияния R = 14 м, дисперсию исходных данных D = 0,42. Исходные данные

приведены в табл.5.12. Вариограмма для такой модели имеет следующий вид: до 14 м и после 14 м соответственно

$$\gamma(h) = D/R = 0{,}03h;$$
$$\gamma(h) = D = 0{,}420. \qquad (5.3)$$

Иначе говоря, до 14 м вариограмма растет по линейному закону, после 14 м имеет постоянное значение.

Проведем вокруг прогнозного пункта Z окружность радиусом 14 м. Только те пункты и соответствующие значения z, которые попадают в окружность, влияют на результат прогноза. Это пять первых пунктов (см. рис.5.25).

Таблица 5.12

Данные для кригинга, м

Номер пункта	x	y	z
1	26	20	3,3
2	41	26	3,1
3	36	21	3,0
4	28	8	2,6
5	44	12	2,1
6	16	33	3,0
7	46	33	2,8
8	9	24	2,5
9	15	5	2,6
10	3	12	2,0
Z	35	16	?

Далее нужно рассчитать расстояния между теми пунктами, которые будут участвовать в прогнозе, по координатам, которые имеются в табл.5.12, и по расстояниям найти значения вариограммы (табл.5.13).

Таблица 5.13

Значения вариограммы $\gamma(h)$

Номер пункта	Исходные пункты					Прогнозный пункт z
	1	2	3	4	5	
1	0	0,4200	0,3015	0,3651	0,4200	0,2955
2	0,4200	0	0,2121	0,4200	0,4200	0,3498
3	0,3015	0,2121	0	0,4200	0,3612	0,1530
4	0,3651	0,4200	0,4200	0	0,4200	0,3189
5	0,4200	0,4200	0,3612	0,4200	0	0,2955

Все готово для составления системы уравнений кригинга (5.35):

$$\left|\begin{array}{ccccc|c} 0{,}4200 & 0 & 0{,}1185 & 0{,}0549 & 0 & 0{,}1245 \\ 0 & 0{,}4200 & 0{,}2079 & 0 & 0 & 0{,}0702 \\ 0{,}1185 & 0{,}2079 & 0{,}4200 & 0 & 0{,}0588 & 0{,}2670 \\ 0{,}0549 & 0 & 0 & 0{,}4200 & 0 & 0{,}1011 \\ 0 & 0 & 0{,}0588 & 0 & 0{,}4200 & 0{,}1245 \end{array}\right| .$$

Решение системы уравнений после приведения суммы корней к единице даст пять весовых коэффициентов:

$$\{0{,}0752;\ -0{,}1659;\ 0{,}6647;\ 0{,}2268;\ 0{,}1993\}.$$

Подставляя их в формулу (5.36), получим

$$Z = 0{,}0752 \cdot 3{,}3 - 0{,}1659 \cdot 3{,}1 + 0{,}6647 \cdot 3{,}0 +$$
$$+ 0{,}2268 \cdot 2{,}6 + 0{,}1993 \cdot 2{,}1 = 2{,}73 \text{ м}.$$

Обращает на себя внимание, что максимальный весовой коэффициент имеет ближайший пункт 3, который «затеняет» пункт 2, у которого весовой коэффициент отрицательный. Остальные коэффициенты уменьшаются по мере удаления от точки Z. Таким образом, на результат кригинга влияет как удаленность пунктов измерений, так и взаимное их расположение. ◀◀

Глава 6 ОСНОВЫ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ МЕСТОРОЖДЕНИЙ

6.1. ЗАДАЧИ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ МЕСТОРОЖДЕНИЙ

При разведке месторождений накапливается большое количество информации: геологическая документация разведочных выработок, данные опробования, результаты геофизических, геохимических исследований и др. В дальнейшем информация перерабатывается с целью построения геологических карт, разрезов, погоризонтных планов, проекций рудных тел, подсчета запасов и решения других вопросов. До появления ЭВМ информацию обрабатывали вручную, что приводило к значительным затратам труда и времени. После появления ЭВМ, особенно персональных компьютеров, мощных серверов и сетей, накопление и обработка геологической информации значительно ускорилась, но и сейчас еще ряд технологических операций в разведке месторождений осуществляется вручную.

Одно из первых назначений компьютера при разведке месторождений состоит в *накоплении*, *систематизации*, *обработке и передаче* геологической информации. Но главное направление при разведке месторождений заключается в *математическом моделировании месторождений*, что позволяет решать вопросы, касающиеся подсчета запасов, определения качества минерального сырья, геолого-экономической оценки месторождений. На базе математического моделирования месторождений можно проектировать горнодобывающие предприятия, планировать и управлять добычей минерального сырья и решать многие другие прикладные задачи.

Существует, по крайней мере, три направления моделирования месторождений: геоинформационное, аналитическое и блочное. Все они имеют между собой много общего.

Геоинформационное моделирование предназначено в основном для моделирования и построения карт любого назначения, в том числе геологических карт земной поверхности и, как частный случай, построения геологических карт месторождений. Существуют специальные пакеты программ для построения карт, такие как ArcInfo, Arcwiev и др. Пакеты позволяют редактировать и преобразовывать полученную информацию, получать «слои» с различной информацией и совмещать их на одном чертеже.

Аналитическое моделирование предназначено для построения геологических карт и разрезов по данным геологической документации разведочных выработок. Для границ всех горных пород и руд путем ряда преобразований рассчитывают координаты в разведочных выработках. Различные способы интерполяции, рассмотренные в предыдущей главе, позволяют строить геологические границы на планах и в разрезах, т.е. делать графическую модель месторождения (или отдельного рудного тела).

Блочное моделирование основано на разделении пространства, в котором находятся рудные тела, на блоки (ячейки) квадратной (на маломощных рудных телах) или кубической (на мощных рудных телах) формы одинакового размера. Разбиение на блоки осуществляется созданием сети параллельных линий и плоскостей. По данным разведочных выработок путем различных способов интерполяции в каждом блоке рассчитывают параметры оруденения (качество руды, ее свойства, достоверность сведений и другие данные). Задав шкалу значений параметров, можно раскрасить блоки различным цветом и получить цветную модель месторождения в виде множества кубиков. Модель можно поворачивать в пространстве, изучая форму рудного тела и качество руд в трех измерениях. Для работы с подобными моделями месторождений создана серия пакетов прикладных программ: Datamine, Micromine и др. С помощью пакетов можно рассчитывать запасы минерального сырья, осуществлять геолого-экономическую оценку месторождений и проектировать рудники.

Во всех направлениях моделирования предусмотрены сбор данных, их систематизация и обработка (моделирование геологических объектов), хранение данных и представление итоговой информации в графическом или табличном виде.

Исходная информация чаще всего накапливается в бумажном виде. Ее необходимо перенести на машинные носители вручную или с помощью различных технических средств. Часть информации можно получить сразу на машинных носителях, например, путем сканирования изображений или текста, с цифровых фотоаппаратов, с аэро- и космических фотоаппаратов, в результате геофизических измерений физических полей и др. Как при ручном наборе информации, так и с применением технических средств исходные данные обычно преобразуются в цифровую или символьную форму и хранятся в виде файлов баз данных в Excel, dBASE, Access, FoxPRO, Paradox, Word, MS DOS и др. База состоит из отдельных банков однородных данных.

Каждая модель или пакет преобразует банки данных в удобную для обработки форму с помощью специальных программ – конвертеров.

6.2. БАНКИ ИСХОДНЫХ ДАННЫХ ПРИ РАЗВЕДКЕ МЕСТОРОЖДЕНИЙ

6.2.1. Банк координат устьев разведочных выработок

Разведка месторождений ведется по дискретной сети наблюдений с помощью разведочных выработок (скважин, горных выработок) при вспомогательной роли геофизических работ. Все разведочные выработки подвергаются геологической документации и опробованию. В рудных пересечениях и их ближайших окрестностях, как правило, берут пробы для химического, минералогического или технического анализа, чтобы установить границы рудных тел и определить качество полезного ископаемого. Вмещающие породы нередко подвергают геохимическому опробованию, чтобы изучить геохимические ореолы вокруг рудных тел и попутно выявить при-

знаки не вскрытых разведочными выработками новых рудных тел. В разведочных выработках проводятся и другие исследования инженерно-геологического и гидрогеологического характера.

В процессе разведки создают банк координат устьев (начала) разведочных выработок, банк искривлений скважин или маркшейдерских замеров, чтобы определить положение разведочных выработок в недрах земной коры, банк геологической документации разведочных выработок и банк опробования (может быть несколько банков опробования – раздельно рядовые, групповые, минералогические, технические и технологические пробы). Все эти банки данных вначале создают на бумажных носителях, а потом переносят на машинные носители – в компьютер.

Поскольку в большинстве случаев разведка месторождений осуществляется скважинами, приведем образцы банков данных именно по скважинам. Банки данных по горным выработкам принципиально от них не отличаются.

Таблица 6.1

Реестр разведочных скважин

№ п/п	Номер скважины	Координаты устья, м			Глубина скважины, м
		X	*Y*	*Z*	
1	22	1911,2	863,0	290,6	613,0
2	23	1577,6	794,2	279,4	383,7
3	25	1752,3	713,4	282,8	421,4

Банк координат устьев разведочных скважин (и поисковых скважин), который часто называется *реестром* скважин, рекомендуется оформлять в виде табл.6.1.

Полезно добавить в этот реестр год окончания бурения скважин, что позволяет в дальнейшем получать интересные данные по динамике разведки месторождения и другие полезные сведения.

6.2.2. Банк искривлений скважин

Банк искривлений скважин содержит данные о глубинах замеров искривлений, зенитных и азимутальных углах в пунктах замеров (табл.6.2). Этот банк данных позволяет определить положение ствола в пространстве. Замеры искривлений в скважинах проводят через 20-50 м. По опытным данным, среднее отклонение скважин от

вертикали на глубине 1000 м составляет около 100 м, а на глубине 1500 м – около 200 м, хотя скважина первоначально задана вертикальной. Пока зенитный угол небольшой (до 1°) гироскопическим методом азимутальный угол измеряется весьма ненадежно. Чтобы точнее определять зенитный угол, часто задают начальный зенитный угол скважины порядка 3-5°.

Таблица 6.2

Замеры искривлений скважин

№ п/п	Номер скважины	Глубина замера, м	Азимутальный угол, град.	Зенитный угол, град.
1	25	0	216	0,00
2	25	50	79	0,50
3	25	100	86	1,25
4	25	150	95	1,50
5	25	200	120	2,50
6	25	250	128	3,50

6.2.3. Банк геологической документации

Банк геологической документации включает номера скважин, интервалы документации и индексацию типов горных пород и руд (табл.6.3). Индексация позволяет сокращать и формализовать геологическую документацию и облегчает обработку данных.

Таблица 6.3

Банк данных геологической документации

№ п/п	Номер скважины	Интервал, м			Индекс	Примечание
		От	До	Длина		
1	25	0,0	5,6	5,6	пг	Песчано-глинистые отложения
2	25	5,6	14,2	8,6	изв	Известняк серый массивный
3	25	14,2	19,5	5,3	ск	Скарн пироксен-гранатовый
4	25	19,5	32,9	13,4	руда	Магнетитовая руда
5	25	32,9	37,6	4,7	ск	Скарн пироксен-гранатовый
6	25	37,6	37,9	0,3	рн	Разрывное нарушение
7	25	37,6	45,0	11,4	изв	Известняк белый массивный
8	25	45,0	50,1	5,1	гр	Гранит серый

Графа «Примечание» для расчетов не нужна, она поясняет индекс и в некоторых случаях может помочь идентифицировать горные породы в геологическом разрезе, когда увязка горных пород и руд неоднозначная.

6.2.4. Банк опробования

Последний банк исходных данных содержит результаты опробования (табл.6.4) или результаты измерения качества полезного ископаемого.

Таблица 6.4

Банк опробования рядовых проб

№ п/п	Номер скважины	Интервал, м			Индекс (сорт)	Состав руды, %		
		От	До	Длина		Cu	Zn	S
1	26	19,5	22,0	2,5	Ц	0,36	3,48	36,55
2	26	22,0	23,8	1,8	МЦ	1,34	2,16	38,43
3	26	23,8	26,0	2,2	МЦ	1,45	3,11	35,17
4	26	26,0	29,3	3,3	М	2,14	0,88	28,66
5	26	29,3	32,9	4,5	М	2,09	0,65	19,78
6	27	36,1	39,3	3,2	Ц	0,45	5,16	35,87
7	27	39,3	42,2	2,9	МЦ	1,54	4,14	40,21
8	27	42,4	44,6	2,4	МЦ	2,08	3,33	37,32

Таблицы 6.3 и 6.4 иногда дополняют графой «Выход керна», используемой для оценки достоверности геологических границ. Банков опробования может быть несколько (рядовые, групповые, минералогические и другие пробы). В банки опробования часто добавляют еще одну графу – плотность руды, необходимую для расчета средних содержаний в пределах интервалов однотипных руд. Добавление этой величины имеет смысл в тех случаях, когда плотность зависит от состава руды.

При составлении банков геологической документации и опробования большую роль играет однозначная формализованная запись индексов, что важно при увязке горных пород и руд в геологических разрезах и на карте. Увязка часто бывает неоднозначной и, как правило, делается в интерактивном режиме.

6.3. ВТОРИЧНЫЕ (РАСЧЕТНЫЕ) БАНКИ ДАННЫХ

6.3.1. Банк координат пунктов измерения искривлений

Имея банк координат устьев скважин (см. табл.6.1) и замеры искривлений в скважинах (см. табл.6.2), можно рассчитать координаты всех пунктов, где произведены измерения искривлений. Расчет ведется от устья скважины. Существует несколько вариантов расчета. Берется первый отрезок, длина его (расстояние между соседними замерами d) известна, имеются также замеры зенитных (γ) и азимутальных (α) углов на концах отрезка. Вначале находят вертикальную dz и горизонтальную dxy проекции отрезка:

$$dz = d\cos[(\gamma_1 + \gamma_2)/2]; \quad dxy = d\sin[(\gamma_1 + \gamma_2)/2], \qquad (6.1)$$

где $(\gamma_1 + \gamma_2)/2$ – полусумма зенитных углов на концах отрезка.

Далее по полусумме азимутов на концах отрезка вычисляют горизонтальные проекции:

$$dx = dxy\sin[(\alpha_1 + \alpha_2)/2]; \quad dy = dxy\cos[(\alpha_1 + \alpha_2)/2]. \qquad (6.2)$$

Проекции отрезков суммируют с координатами устья скважин, получают координаты конца отрезка, т.е. координаты пункта искривления.

Подобные операции повторяют для каждого отрезка, в результате определяют координаты всех пунктов измерений искривлений последовательно, начиная с устья скважин:

Таблица 6.5

Координаты пунктов искривлений

№ п/п	Номер скважины	Глубина, м	Координаты, м		
			X	*Y*	*Z*
1	25	0,0	1543,7	894,2	245,1
2	25	50,0	1544,3	895,1	195,2
3	25	100,0	1545,6	896,4	145,4
4	25	150,0	1547,2	898,2	96,5
5	25	200,0	1549,9	900,7	48,6
6	25	250,0	1551,3	903,3	–00,8
7	25	300,0	1551,2	905,0	–49,3
8	25	350,0	1550,4	907,2	–98,5

$$X_{i+1} = X_i + dx;\ \ Y_{i+1} = Y_i + dy;\ \ Z_{i+1} = Z_i - dz. \qquad (6.3)$$

Обращает на себя внимание, что приращение *dz* берется со знаком минус, так как отсчет координаты *Z* ведется снизу вверх от уровня Мирового океана (в России от Кронштадского футштока). В некоторых странах (например, в Канаде) отсчет координаты *Z* осуществляется в обратном порядке – сверху вниз (например, в пакете Micromine). Некоторая сложность в расчетах возникает, если азимуты на концах отрезка переходят через 360°, что можно учесть введением отрицательного азимута, если он меньше 360°. Если замеры искривлений не доведены до забоя скважины, то оставшаяся часть скважины принимается прямолинейной с последними измеренными зенитным и азимутальным углами.

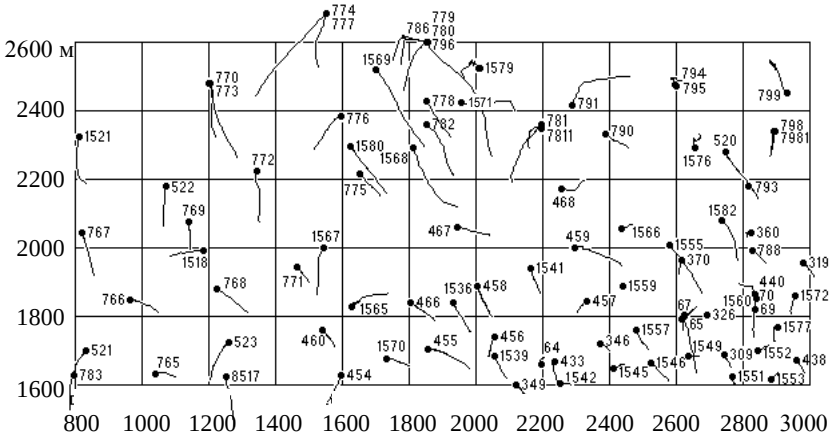

Рис.6.1. Проекции искривленных скважин на горизонтальную плоскость в северной части апатитового месторождения Коашва (Хибины)

В результате расчетов получается банк координат искривленных скважин (табл.6.5), который позволяет показать положение скважин в пространстве на горизонтальной и на любой другой проекции (рис.6.1).

6.3.2. Банк рудных пересечений

В банке опробования содержатся рядовые пробы (см. табл.6.4), взятые по отдельным типам руд. Непрерывная совокупность рядовых проб дает рудное пересечение.

Рудное пересечение – это отрезок от точки входа до точки выхода из рудного тела.

Иногда внутри рудного пересечения располагаются руды различных промышленных сортов: окисленные и первичные, медные и цинковые, гематитовые и магнетитовые и пр. Тогда внутри рудного пересечения выделяются отдельные пересечения сортов руд. В каждом рудном пересечении или в пересечении промышленного сорта руды рассчитывают средний состав по формуле

$$C_{ср} = \frac{\sum Cm\rho}{\sum m\rho} \text{ или } C_{ср} = \frac{\sum Cm}{\sum m}, \qquad (6.4)$$

где m – длина проб; C – состав проб; ρ – плотность руды.

Когда плотность руды зависит от ее состава, применяется первая формула (6.4), в других случаях – вторая.

Таблица 6.6

Банк рудных пересечений

№ п/п	Номер скважины	Интервал, м			Сорт	Содержание, %		
		От	До	Длина		Cu	Zn	S
1	23	38,4	40,2	1,8	Ц	0,22	6,43	32,15
2	23	40,2	43,6	3,4	МЦ	1,87	4,35	33,48
3	23	43,6	45,8	2,2	М	1,96	0,87	32,16
4	23	45.8	48,4	3,6	МВ	1,25	0,23	18,14
Рудное пересечение		38,4	48,4	10,0	Р	1,56	2,91	30,78

В результате подобных расчетов создается банк рудных пересечений (табл.6.6). Фактически в табл.6.6 два банка: банк сортов руд, а последняя строчка – составная часть банка рудных пересечений. Одна скважина может пересечь несколько рудных пересечений, так же, как и несколько однотипных горных пород.

6.3.3. Банк координат геологических границ

Имея банк данных геологической документации (см. табл.6.3), банк рудных пересечений (см. табл.6.6) и банк координат пунктов искривлений скважин (см. табл.6.5), можно рассчитать координаты всех геологических границ в скважинах. Координаты любой геологической границы в скважине находят путем линейной интерполяции между координатами соседних пунктов искривлений.

Пусть имеются координаты скважины на глубине 150 м (X_1 = 1254,2 м, Y_1 = 754,6 м, Z_1 = 247,4 м) и на глубине 200 м (X_2 = 1256,4 м, Y_2 = 752,8 м, Z_2 = 198,6 м). Требуется определить координаты геологической границы на глубине 183 м.

Ответ получаем по формуле линейной интерполяции (путем решения пропорции):

$$X = 1254{,}2 + (1256{,}4 - 1254{,}2)(183 - 150)/(200 - 150) = 1255{,}4 \text{ м};$$

$$Y = 754{,}6 + (754{,}6 - 752{,}8)(183 - 150)/(200 - 150) = 753{,}4 \text{ м};$$

$$Z = 247{,}4 - (247{,}4 - 198{,}6)(183 - 150)/(200 - 150) = 215{,}8 \text{ м}.$$

Нужно обратить внимание, что приращение координаты Z идет со знаком минус. Подобным образом находят координаты всех геологических границ. В результате получают банк всех геологических границ (табл.6.7).

Таблица 6.7

Банк геологических границ

№ п/п	Номер скважины	Координаты, м			Индекс
		X	*Y*	*Z*	
1	22	2154,3	1457,8	236,9	ПГ
2	22	2156,2	1458,0	232,7	ИЗ
3	22	2163,5	1462,1	200,6	АРГ
4	22	2165,7	1463,5	197,6	Р
5	22	2168,8	1471,4	148,9	ГР
6	23а	1547,8	2175,5	233,3	ПГ
7	23а	1547,2	2176,1	230,8	ПОРФ
8	23а	1545,7	2187,3	208,1	ИЗ

Расчетом банка геологических границ завершаются практически все направления математического моделирования геологических объектов. Далее пути моделирования расходятся. В геоинформатике, где главная задача состоит в построении геологических карт, основное внима-

ние уделяется построению контуров геологических границ, их векторизации, введению условных обозначений для площадных и точечных объектов, созданию слоев с различной информацией, к преобразованию масштабов карт, к совмещению на одном чертеже нескольких слоев информации и т.д.

В данной книге мы не будем касаться геоинформатики, так как она входит в специальных курс «Математическая картография» и достаточно подробно изложена в работах В.Я.Цветкова [17, 18].

Рассмотрим второе направление – аналитическое – использование банка геологических границ для построения геологических карт и разрезов как одной из основ подсчета запасов и геолого-экономической оценки месторождений.

В конце книги коснемся методики построения и работы с блочными моделями месторождений.

6.4. О МОДЕЛИРОВАНИИ МЕСТОРОЖДЕНИЙ

6.4.1. Аналитические модели месторождений

В понятие аналитические модели месторождений вкладывается несколько иной смысл, чем в монографии И.И.Шаталова и В.И.Щеглова [19], где математическое моделирование ведется для создания искусственных учебных моделей месторождений. Здесь ставится задача построения графических изображений реальных месторождений на основе координат разведочных выработок. Эта задача довольно сложная и не всегда поддается исчерпывающему решению.

Можно выделить несколько уровней сложности месторождений применительно к математическому моделированию.

1. Наиболее простые месторождения разведаны короткими вертикальными разведочными выработками, единичные рудные тела однозначно увязаны между собой, разрывные нарушения, крутые складки и размывы отсутствуют.

2. Более сложные месторождения характеризуются наличием нескольких рудных тел, увязка которых неоднозначная и возможна

лишь в диалоговом режиме. Могут присутствовать разрывные и складчатые нарушения и размывы. Разведочные выработки прямолинейные.

3. Месторождения следующей группы сложности отличаются тем, что скважины искривлены, но увязка рудных тел однозначная. Проблема заключается в основном в том, что скважины не лежат в одной плоскости, разведочные выработки часто находятся далеко друг от друга и интерполяция границ носит неоднозначный характер.

4. Наиболее сложная ситуация возникает, когда в разведочных скважинах много рудных пересечений, увязка их неоднозначная, скважины искривлены, не находятся в одной плоскости (рис.6.1) и трудно построить плоские разрезы. В данном случае кроме диалогового режима приходится принимать специальные меры по построению плоских разрезов.

Для месторождений последних двух групп рекомендуется сделать дополнительную опорную сеть вертикальных псевдоскважин, ориентированных по заданным профилям. Так, для скважин, изображенных на рис.6.1, построена опорная сеть псевдоскважин, расположенных по линиям с азимутом СЗ 235° в соответствии с генеральным направлением линий, принятых на месторождении (рис.6.2). Опорная сеть позволяет строить разрезы по линиям опорных скважин.

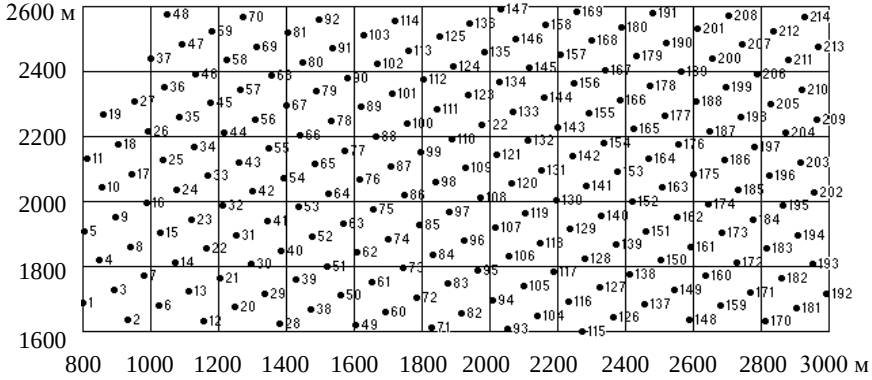

Рис.6.2. Сеть опорных скважин 100 × 100 м

Идея заключается в том, чтобы, используя геологические границы в разведочных скважинах, подобные табл.6.7, путем интерполяции рассчитать координаты геологических границ во всех псевдоскважинах. Далее по координатам геологических границ, имеющихся в псевдоскважинах, проводят геологические границы в плоскостях вертикальных геологических разрезов. Создание сети псевдоскважин удобно во многих случаях математического моделирования месторождений.

Интерполяция координат геологических границ между разведочными скважинами может быть осуществлена различными способами. Наиболее приемлемый способ основан на комбинации тренда и кригинга. Вначале рассчитывается двухмерный тренд, потом остаток от тренда, по остаткам находится вариограмма и применяется кригинг по остаткам. Полезно отметить, что тренд частично снимает явление анизотропии. Наиболее сложная задача состоит в том, что перед расчетом тренда нужно идентифицировать рудные тела в диалоговом режиме. Иногда приходится рассчитывать несколько вариантов трендов, добиваясь наименьших отклонений поверхности тренда от координат рудных тел в разведочных скважинах.

Когда рудные тела индентифицированы, рассчитаны тренд, вариограмма остатков и определен радиус автокорреляции, можно прогнозировать значения координат геологических границ в опорных псевдоскважинах. При этом возникает несколько вариантов.

Если в пределы радиуса автокорреляции попадает более трех разведочных скважин, то кригинг осуществляется обычным способом по формуле (5.37). Если в пределах радиуса автокорреляции имеется одна или две разведочные скважины, то нужно найти вес координат геологической границы по формуле $p_i = D - \gamma(h)$, где h – расстояние от опорной псевдоскважины до разведочной скважины, и далее воспользоваться формулой (5.37). Если псевдоскважина находится за пределами радиуса влияния разведочных выработок, то применяют либо линейную интерполяцию, либо интерполяцию методом обратных расстояний.

Еще одна проблема возникает при расщеплении или слиянии рудных тел или при выклинивании геологических тел. При слиянии рудных тел между ними проводится условная геологическая гра-

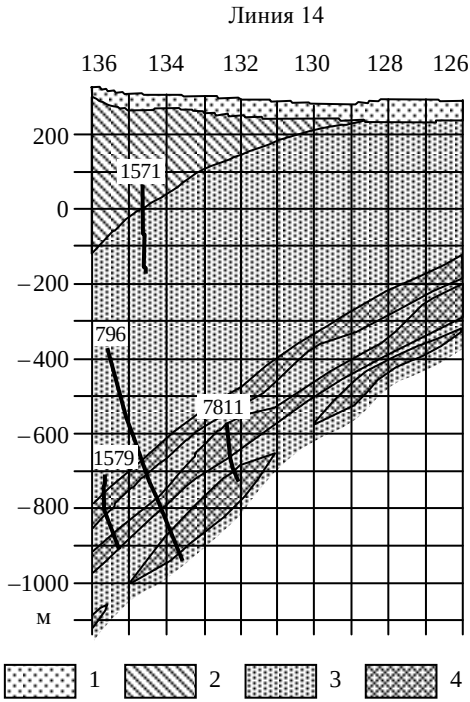

Рис.6.3. Геологический разрез по линии псевдоскважин месторождения Коашва
1 – покровные отложения; 2 – сиениты; 3 – ийолит-уртиты; 4 – апатитовые руды

ница либо по геологическим данным, либо с помощью тренда. При графическом изображении эту границу, естественно, не показывают. При выклинивании геологических тел для каждой расщепленной части строится своя граница, которая совпадает с границей нижележащих пород или руд, т.е. мощность выклинившегося геологического тела считается нулевой.

Имея координаты геологических границ в каждой псевдоскважине, можно по ним построить вертикальный геологический разрез. Наилучший вариант получается, если в качестве геологической границы в разрезе используют сплайн (скользящий сплайн). Каждое геологическое тело имеет верхнюю и нижнюю границу. Их строят одновременно, но условные обозначения горной породы или руды (цветные или штриховые) задают верхней границей (рис.6.3).

6.4.2. Блочные модели месторождений

Как указывалось выше, месторождение с помощью сетки разбивается на элементарные одинаковые по размеру блоки. На маломощных рудных телах блоки имеют квадратную форму, а на мощных рудных телах – кубическую (рис.6.4 и 6.5).

Размер блоков может быть любой, но предпочтительнее делать блоки меньше эксплуатационных. На экране монитора размер блоков можно довести до пикселя. Обычно стараются выбрать блоки такого размера, чтобы в него попадало не менее трех разведочных выработок (в пределах радиуса автокорреляции). Чем больше в блоке разведочных выработок, тем достовернее будут определены в нем параметры.

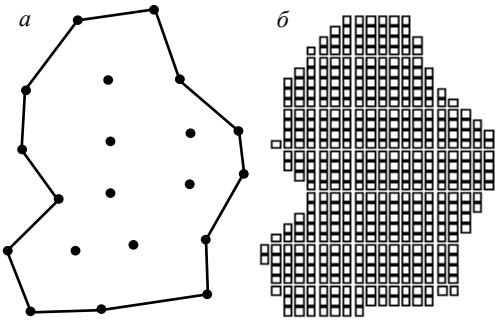

Рис.6.4. Схема выделения ячеек на проекции рудного тела железорудного месторождения во Вьетнаме: *а* – план расположения скважин; *б* – разделение рудного тела на ячейки

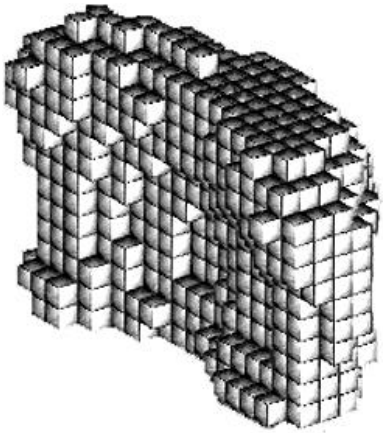

Рис.6.5. Блочная модель месторождения, по данным Горного института КНЦ АН СССР

Разбиению на блоки предшествует оконтуривание рудных тел («натягивание каркаса на рудное тело»). Блоки выделяют в пределах установленного каркаса. Используя различные методы интерполяции, описанные в главе 5, в каждом блоке по данным разведочных выработок определяют параметры оруденения (состав и плотность руды, запасы руд и металла, достоверность и категорийность запасов и др.).

Для построения блочных моделей месторождений разработано большое количество пакетов прикладных программ, такие как Datamine, Micromine, Minescape, Vulkan, Geostat и др.

Особенность пакетов состоит в специальной обработке банка опробования. Рудные пересечения делят на равные интервалы (например, по 3 м) без учета типов руд. В каждом интервале по данным рядовых проб без учета плотности руды рассчитывают среднее содержание металлов. Далее определяют координаты центров рудных интервалов (методику см. в подразделах 6.3.1 и 6.3.3), которые и используют в дальнейших расчетах. Попутно проверяют арифметические ошибки, которые могут появиться при ручном вводе исходных данных. Совокупность блоков дает пространственное представление о размещении оруденения. Для наглядности можно поворачивать полученное изображение в трехмерном пространстве. Блоки можно раскрашивать по содержанию металлов, по сортам руд, по достоверности запасов и т.д.

Суммируя содержимое блоков в заданных границах, можно рассчитать запасы руды, среднее содержание металлов как в отдельных блоках, так и по месторождению в целом.

Дополнительные программы, введенные в пакеты прикладных программ, позволяют проводить экономическую оценку месторождений (для этого потребуется дополнительная экономическая информация), обосновывать кондиции методом вариантов. Наконец, в пакетах имеются программы, предназначенные для проектирования рудников. Многообразие задач, решаемых с помощью моделей, обусловило их широкое применение при математическом моделировании и оценке месторождений.

ЗАКЛЮЧЕНИЕ

Рассмотренные математические методы не исчерпывают всего их многообразия. Приведены лишь наиболее важные и распространенные в геологической практике, по мнению автора, методы. Так, исключены из рассмотрения дисперсионный анализ, частные коэффициенты корреляции, некоторые редкие законы распределения случайных величин. Некоторые методы, например основы геостатистики, кригинг даны в сокращенном виде. Подробные сведения студенты могут найти в справочнике [15].

Вместе с тем в учебник включены новые методы, как созданные автором, так и заимствованные из других источников: нахождение информативных факторов, определение оптимального порядка полинома, обработка усеченного закона распределения, выделение отдельных синусоид из гармонического ряда и др. Данные методы проверены на практике и могут быть полезны при математической обработке геологических данных.

В геологии появились также такие новые математические методы, как фрактальный анализ, явления самоорганизации вещества и др. Они находятся в стадии становления, поэтому в учебник пока не вошли.

Нужно обратить внимание, что в работе мало говорится о применении компьютеров, хотя многие из рассмотренных методов содержатся в таких прикладных пакетах, как Статистика, Сёрфер, Excel и пр. В пакетах отсутствует прозрачность в идеологии вычислений, часто отсутствуют формулы, по которым производятся те или иные вычисления, дается лишь готовый результат. Прикладные пакеты удобны для подготовки презентаций, построения диаграмм различного типа, научных докладов, но по ним трудно изучать математические методы. Предполагается, что пользователь пакетов хорошо знаком с математическими методами и использует программы для иллюстративных целей.

Можно надеяться, что учебник будет полезен не только студентам, но и геологам-производственникам, а также другим специалистам, занимающимся математической обработкой данных. Математические методы большей частью универсальны и применимы в самых различных областях знаний.

РЕКОМЕНДАТЕЛЬНЫЙ БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Абрамович И.И.* Математическая геология и математический прогноз / И.И.Абрамович, Л.Н.Дуденко, Ю.И.Михайлова // Тр. ВСЕГЕИ. 1972. Т.178. С.103-122.

2. *Айвазян С.А.* Прикладная статистика. Исследование зависимостей: Справочное пособие / С.А.Айвазян, И.Г.Енюков, Л.Д.Мешалкин. М.: Финансы и статистика, 1985. 187 с.

3. *Афифи А.* Статистический анализ. Подход с использованием ЭВМ / А.Афифи, С.М.Эйзен. М.: Мир, 1982. 488 с.

4. *Большев Л.Н.* Таблицы математической статистики / Л.Н.Большев, Н.В.Смирнов. М.: Наука, 1983. 416 с.

5. *Боровиков В.* Statistica. Искусство анализа данных на компьютере. 2-е изд. СПб: Питер, 2003. 688 с.

6. *Бронштейн И.Н.* Справочник по математике для инженеров и учащихся втузов / И.Н.Бронштейн, К.А.Семендяев. 13-е изд. М.: Наука, 1986. 544 с.

7. *Давид М.* Геостатистические методы при оценке запасов. Л.: Недра, 1980. 360 с.

8. *Каждан А.Б.* Математическое моделирование в геологии и разведке полезных ископаемых / А.Б.Каждан, О.И.Гуськов, А.А.Шиманский. М.: Недра, 1979. 168 с.

9. *Каждан А.Б.* Математические методы в геологии: Учебник для вузов / А.Б.Каждан, О.И.Гуськов. А.А.Шиманский. М.: Недра, 1990. 251 с.

10. *Капутин Ю.Е.* Геостатистическое исследование месторождений полезных ископаемых: Методические рекомендации. Петрозаводск: Изд-во КарФАН СССР, 1988. 190 с.

11. Математическая энциклопедия: В 5 томах. М.: Советская энциклопедия, 1977-1985.

12. *Матерон Ж.* Основы прикладной геостатистики. М.: Мир, 1968. 408 с.

13. *Поротов Г.С.* Основы статистической обработки материалов разведки месторождений: Учебное пособие / Ленинградский горный институт. Л., 1985. 97 с.

14. *Смоляк С.А.* Устойчивые методы оценивания / С.А.Смоляк, Б.П.Титаренко. М.: Статистика, 1980. 208 с.

15. Справочник по математическим методам в геологии / А.А.Родионов, Р.И.Коган, В.А.Голубев и др. М.: Недра, 1987. 334 с.

16. *Тьюки Дж.* Анализ результатов наблюдений. М.: Мир, 1981. 302 с.

17. *Цветков В.Я.* Геоинформационные системы и технологии. М.: Финансы и статистика, 1998. 288 с.

18. *Цветков В.Я.* Геоинформационные системы и технологии: Учебное пособие / Московский институт геодезии и картографии. М., 1996. 112 с.

19. *Шаталов И.И.* Моделирование месторождений и рудных полей на ЭВМ: Учебное пособие / И.И.Шаталов, В.И.Щеглов. М.: Недра, 1989. 150 с.

20. *Шестаков Ю.Г.* Математические методы в геологии: Учебное пособие. Красноярск: Изд-во Красноярского ун-та, 1988. 208 с.

ОГЛАВЛЕНИЕ