

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Е.Г. Щеглова

ПРИМЕНЕНИЕ ЭВМ В ГЕОЛОГИИ: КУРС ЛЕКЦИЙ

Рекомендовано Ученым советом федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет» в качестве учебного пособия для студентов, обучающихся по программам высшего образования по направлению подготовки 21.05.02 Прикладная геология

Оренбург
2015

УДК 55:004.4:519.8(075.3)
ББК 26.3я73+32.973-018.2я73+22.18я73
Щ 33

Рецензент - кандидат технических наук, старший преподаватель кафедры экологии и природопользования ОГУ А.С. Степанов

Щеглова, Е.Г.

Щ 33 Применение ЭВМ в геологии: курс лекций : учебное пособие / Е.Г. Щеглова; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2015. – 110 с.

ISBN 978-5-7410-1315-1

Курс "Применение ЭВМ в геологии" разработан для последовательного освоения студентами основ применения ЭВМ в геологии, а также теоретических знаний по математическому моделированию.

В учебном пособии рассмотрены типы геологической информации и геологического моделирования, основные методы моделирования, а также сущность и условия применения одномерных, двумерных и многомерных статистических моделей. Описаны основные программы, используемые для компьютерного моделирования в геологии.

Учебное пособие предназначено для студентов очной, заочной, дистанционной форм обучения, обучающихся по направлению подготовки 21.05.02 "Прикладная геология"

УДК 55:004.4:519.8(075.3)
ББК 26.3я73+32.973-018.2я73+22.18я73

ISBN 978-5-7410-1315-1

© Щеглова Е.Г., 2015
© ОГУ, 2015

Содержание

Введение	4
1 Компьютерное моделирование в геологии	6
1.1 Характеристика геологической информации	6
1.2 Моделирование в геологии	8
1.3 Типы геолого-математических моделей	10
1.4 Принципы и методы геолого-математического моделирования	14
1.5 Методы изучения геологических объектов	16
2 Специфика геологических данных	20
2.1 Источники данных и их типы	20
2.2 Представление графической (пространственной) информации	26
2.3 Представление текстовой информации	38
3 Одномерный статистический анализ и его применение	44
3.1 Характеристика одномерного статистического анализа	44
3.2 Законы распределения случайных величин	52
4 Двумерный статистический анализ и его применение	74
4.1 Характеристика двумерного статистического анализа	74
4.2 Двумерное нормальное распределение. Эллипс рассеяния	80
4.3 Нелинейная регрессия. Метод квадратов	84
5 Многомерный статистический анализ и его применение	87
6 Основные программы при компьютерном моделировании в геологии	93
6.1 Программа Microsoft Word	93
6.2 Программа Microsoft Excel	95
6.3 Программа CorelDraw	101
6.4 Программа Microsoft Access	105
Список использованных источников	110

Введение

В настоящее время с развитием естественных наук в результате воздействия научно-технического прогресса происходят значительные изменения методов научных экспериментов, анализа, обобщения получаемых данных. Кроме этого, благодаря расширившимся возможностям фундаментальных наук, развитию электронно-вычислительной техники, а также комплексной автоматизации практически любой из видов человеческой деятельности.

В течение последних десятилетий отмечается более глубокое проникновение математических методов исследования в каждую отрасль естественных наук, что способствует значительному успеху в этих науках. Для более успешного и прогрессивного развития геологических наук также необходимы математические методы исследования и использование электронно-вычислительных машин.

Для современной геологии является недостаточным изучение только качественных сторон процессов и явлений. Помимо изучения качественных показателей, необходимо изучение количественных параметров, что позволит обеспечить наиболее высокий научный уровень исследований земных недр.

В последнее время острая необходимость использования математических методов обработки, анализа и обобщения геологических данных обнаруживается при проведении всех видов геологических исследований, а не только при прогнозных, поисковых, разведочных и оценочных работах.

Для примера, литологические, стратиграфические, палеонтологические, минералогические, петрографические и другие виды геологических исследований уже давно, кроме описательных методов исследования, требуют точных математических методов исследования с показателями числа и меры.

Каждый год различные геологические организации накапливают большое количество эмпирического материала, который представлен миллионами количественных определений химического состава по различным минералам и их агрега-

там, химическим и минеральным составом, физическими и горно-технологическими свойствами горных пород и полезных ископаемых, что требует применения электронно-вычислительных машин для обработки и обобщений, чтобы более полно извлекать содержащуюся в них информацию.

На данный момент острую необходимость математических методов при проведении геологоразведочных работ стали испытывать производственные геологические организации, так как возросли требования промышленности к конкретности и достоверности геологических данных.

По действующему положению на каждом этапе геологоразведочных работ количественные оценки прогнозных ресурсов полезных ископаемых должны быть обоснованы по данным геологических съемок с уточнением цифр прогнозных ресурсов (а далее запасов).

В результате резкого увеличения объемов количественной информации, получаемой при проведении геологической съемки, поисков и разведки полезных ископаемых, появилась необходимость в разработке новых способов ее хранения, поиска, обработки и анализа с помощью электронно-вычислительных машин.

Учитывая возрастающую роль математических методов обработки геологических данных и широкого применения электронно-вычислительных машин практически в каждой отрасли геологической науки, становится совершенно очевидно значение дисциплины "Применение ЭВМ в геологии" в образовании современного геолога, специалиста по геологической съемке, поискам и разведке месторождений полезных ископаемых, а также специалиста по геологии нефти и газа.

1 Компьютерное моделирование в геологии

1.1 Характеристика геологической информации

Обилие разнообразных геологических объектов и методов их изучения способствует получению разной по характеру информации: словесной (описательной), графической (картографической) и цифровой.

Геология в качестве теоретической науки долгое время развивалась в условиях почти полного отсутствия экспериментальных данных из-за недоступности геологических объектов для наблюдения и считалась описательной наукой.

Раньше геологическая информация в основном была представлена в виде словесного описания, численные показатели и показатели меры мало отражены и выполняли лишь иллюстративные функции. Теоретические выводы геологов часто носили субъективный характер, так как были основаны только на личном опыте и интуиции. В следствии этого на данный момент в геологии многие понятия и определения неоднозначны с большим количеством сравнений и аналогий. Например, понятия "горная порода", "минерал", "формация" имеют несколько десятков трактовок, а понятие "фация" даже больше 100 определений.

Достаточно распространенные формы объединения и обобщения знаний о свойствах геологических объектов - классификации и группировки. В их основу, в большинстве случаев, положены качественные признаки. При этом набор данных признаков, а также количество групп в классификациях могут быть разными. Например, существует несколько классификаций разделения изверженных пород по минеральному и химическому составу следующих авторов: Г. Розенбуш, С. Мишель-Леви, Ф.Ю. Левинсон-Лессинг, П. Ниггли, А.Н. Заварицкий.

При составлении карт, планов, разрезов (графические геологические документы) различными авторами неоднозначно используются определенные геологические объекты в качестве условных обозначений. Поэтому, представленная на таких картах геологическая информация также является неоднозначной и такие кар-

ты будут иметь существенные различия, хотя они могут быть составлены в том же масштабе и на одну и ту же территорию, но, например, разными исследователями и в разный период времени. Свои специфические особенности имеет количественная (цифровая) геологическая информация, объем которой значительно увеличился за последние десятилетия.

Из-за применения выборочного метода изучения, а также достаточной сложности геологических объектов количественная информация не достаточно точна и не в полной мере отражает их свойства. Кроме этого, на достоверность количественной информации негативно сказываются технические погрешности измерения. К тому же, некоторые свойства геологических объектов можно выразить разными числовыми характеристиками. В качестве примера может послужить изучение степени окатанности песчаных зерен и галек, которое позволяет судить о характере их транспортировки расстоянию до источника сноса. Хотя, в качестве оценки степени окатанности, можно использовать другие величины: частное от деления радиуса кривизны самого острого конца песчинки или гальки на ее средний радиус; отношение среднего радиуса максимальных окружностей, описывающих вершины всех углов границы в ее проекции на плоскости, к радиусу наибольшего круга, вписанного в эту проекцию и т.д.

При изучении полезных ископаемых могут анализироваться валовые содержания химических элементов, содержания их оксидов, сульфидов или других химических соединений, содержания минералов-носителей полезных компонентов или другие количественные показатели качества руд.

При изучении рудных месторождений чаще всего используются содержания химических элементов, при изучении россыпных месторождений – содержания полезных минералов, в некоторых случаях при изучении месторождений – содержания различных соединений металлов, которые обладают резко контрастными технологическими свойствами. Например, при переработке оловянных руд гораздо легче, по сравнению с сульфидами, извлекаются оксиды олова. Силикаты железа в металлургических процессах не выплавляются и уходят в шлаки. Поэтому важно

при выборе подходящего вида числовых измерений определить какая из количественных характеристик более полно выражает изменения нужного свойства.

1.2 Моделирование в геологии

Материальные системы подразделяются на хорошо организованные и плохо организованные.

Хорошо организованные системы - это системы, состоящие из ограниченного числа элементов. Эти элементы находятся в строго определенных и однозначных зависимостях. К хорошо организованным системам относят простейшие физические и химические процессы, механизмы, приборы и т.д. Свойства данных систем можно количественно описать с помощью законов физики и химии.

Плохо организованные системы - это системы, в которых можно установить только отдельные закономерности, и которые не поддаются строгому количественному выражению. Такими системами являются сложные природные объекты и явления, подверженные влиянию множества факторов различной природы. Кроме этого, к таким системам относятся живые организмы и их сообщества, объекты, изучаемые науками о Земле. Главный метод изучения таких систем - моделирование. При моделировании объект изучения заменяется упрощенным аналогом - моделью. По виду моделей выделяют знаковое (информационное) и предметное моделирование. Знаковое моделирование - это моделирование, при котором различные схемы, чертежи, формулы или мысли, записанные или высказанные на любом языке выступают в качестве моделей. Предметное моделирование - это моделирование, при котором исследование проводится на модели, воспроизводящей определенные геометрические, физические, динамические или функциональные характеристики объекта.

Выделяют модели структуры объекта и модели поведения (функционирования) объекта в зависимости от того, какие были выбраны его особенности для изу-

чения. Модели структуры объекта используют при исследовании статичных систем (свойств материальных предметов). Модели поведения (функционирования) объекта выбирают при изучении динамичных систем (процессов).

Охарактеризованные выше свойства геологических процессов и образований исключают вероятность достаточно широкого применения предметного моделирования в геологии, однако, не смотря на это, в последние десятилетия гораздо чаще стали предприниматься попытки воспроизведения отдельных элементов геологических процессов в условиях лабораторий. Возникли следующие научные направления: петрология, геохимия, экспериментальная геотектоника и т.д. Крупные достижения экспериментальной минералогии поспособствовали разработке технологических процессов получения синтетического кристаллосырья в промышленных масштабах. Предметное моделирование, при промышленной оценке месторождений, используется с целью исследования технологических свойств руд согласно лабораторным и полупромышленным пробам. При этом лабораторные установки, которые имитируют процесс переработки руды, являются действующими модификациями оснащения предстоящей обогатительной фабрики. Однако, основную роль в науках о нашей планете играют разные методы знакового (информационного) моделирования. По характеру информации их можно разбить на:

- словесные (вербальные);
- графические;
- математические (точные).

Словесные модели – это различные понятия, определения, а также многочисленные классификации, присутствующие во всех геологических дисциплинах.

Графические модели – это всевозможные графические геологические документы, а именно, планы, проекции, геологические разрезы, карты и т.д. Их отличительной особенностью является отражение свойств реальных объектов недр приблизительно и более упрощенно.

Математические модели в геологии – это числа и формулы, которые описывают взаимосвязи и закономерности изменения свойств геологических образований или параметров геологических процессов.

В последнее время, в связи с обширным введением в практику геологических исследований моделирования на ЭВМ с применением различной геологической информации, границы между этими видами моделей делаются в достаточной степени условными. При помощи номинальной шкалы измерений картографическая информация переводится в цифровую, а итоги замеров, полученные при геофизических и геохимических съемках, представляются с помощью графопостроителей или графических дисплеев в виде карт изолиний.

1.3 Типы геолого-математических моделей

Согласно принципу построения математической модели отличают статистическое и динамическое моделирование.

Статистическое моделирование – это моделирование, которое состоит в математическом описании свойств изучаемых объектов на основе результатов их исследования с помощью выборочного метода посредством индуктивного обобщения эмпирических данных.

Динамическое моделирование – это моделирование, которое применяет дедуктивный метод, основанный на том, что свойства конкретных объектов выводятся из общих представлений об их структуре и законах, характеризующих их свойства.

В настоящее время, в практике исследований в области геологии в основном используются статистические модели. Данный факт обуславливается сложностью и многообразием геологических объектов и трудностью описания геологических процессов даже в наиболее общих чертах.

Статистическое моделирование характеризуется следующими этапами:

1. Преобразование геологической информации в вид, который более удобен для анализа.
2. Выявление закономерностей свойств исследуемых объектов в массовых и случайных замерах.
3. Математическое описание выявленных закономерностей, а именно, составление математической модели.
4. Использование полученных количественных характеристик для решения конкретных геологических задач таких, как проверка геологических гипотез, выбор методов дальнейшего изучения объекта и т.д.
5. Оценка вероятности возможных ошибок в решении поставленной задачи посредством выборочного метода изучения объекта.

Совсем другой порядок решения геологических задач при помощи динамического моделирования. Отталкиваясь от единых суждений о генезисе изучаемого объекта, строится теоретическая математическая модель процесса его образования, которая учитывает ключевые факторы, влияющие на окончательный итог этого процесса, то есть на свойства объекта.

Данная модель, как правило, может быть предложена только в самом общем виде, так как параметры процесса неизвестны. Указанные параметры определяют посредством перебора разных вариантов и сопоставления теоретической реализации процесса с подлинными свойствами исследуемого объекта, которые были установлены эмпирическим путем.

Динамическое моделирование связано с огромным объемом достаточно трудных вычислений и возможно только на базе ЭВМ.

В зависимости от характера связи между параметрами и свойствами исследуемых объектов математические модели делят на статистические и детерминированные модели.

Статистические модели – это математические выражения, которые содержат хотя бы одну случайную компоненту, то есть переменную, значение которой для единичного наблюдения точно предсказать не представляется возможным. Та-

кие модели достаточно обширно применяют для целей математического моделирования, так как они хорошо учитывают случайные колебания экспериментальных данных.

Детерминированные модели – это модели, которые выражают функциональные связи между аргументом и зависимыми переменными. Их записывают в виде уравнений, где определенному значению аргумента будет соответствовать лишь одно значение переменной. Такие модели при моделировании геологических объектов применяют довольно редко. Данный факт можно объяснить тем, что детерминированные модели плохо согласуются с реальными явлениями, в которых функциональные связи сохраняются только в узких, достаточно ограниченных областях.

Многообразие геологических объектов исследования и задач спровоцировало потребность применения при геолого-математическом моделировании методов из различных разделов математики, а именно, теории множеств, теории вероятностей и математической статистики, теории групп, теории игр, теории информации, теории графов, дифференциальной геометрии, матричной и векторной алгебры и т.д. Однако, при этом, одну и ту же задачу можно решить различными методами, а в определенных случаях при решении одной задачи нужно будет использовать целый комплекс методов из различных разделов математики. Данное обстоятельство формирует определенные сложности при систематизации математических методов, используемых в геологии.

Наряду с этим, по тому какой тип решаемых задач, набор применяемых для этого методов и какие главные допущения относительно свойств геологических объектов все геолого-математические модели делятся на 2 категории.

К первой категории относят модели, которые используют в основном математический аппарат теории вероятностей и математической статистики. В таких моделях геологические объекты обычно являются внутренне однородными, изменения их свойств в пространстве носят случайный характер и не зависят от места

замера. Данные модели условно называют статистическими. От того, сколько одновременно рассматривается свойств их делят на:

- одномерные;
- двумерные;
- многомерные.

Статистические модели, как правило, применяют с целью:

1. Извлечения согласно выборочным данным более достоверных оценок свойств геологических объектов;
2. Подтверждения или опровержения геологических гипотез;
3. Раскрытия и отображения связей между свойствами геологических объектов;
4. Систематизации геологических объектов;
5. Установления объема выборочных данных, которые являются необходимыми для оценки свойств геологических объектов с установленной точностью.

Ко второй категории относят модели, которые рассматривают свойства геологических объектов как пространственные переменные. В таких моделях подразумевается то, что свойства геологических объектов находятся в зависимости от координат точки замера, а в изменении данных свойств в пространстве имеются определенные закономерности.

Однако, наравне с некоторыми вероятностными методами (например, дисперсионный анализ, случайные функции, временные ряды), используются кроме того приемы комбинаторики (полиномы), векторной алгебры, дифференциальной геометрии, гармонического анализа и иных разделов математики.

Для исследования пространственных геологических переменных применяют приемы обоих видов моделирования (статистическое, динамическое)

Модели пространственных геологических переменных применяют с целью решения задач, которые сопряжены с:

- проверкой гипотез о закономерностях размещения геологических объектов относительно друг друга;

- проверкой гипотез о характере процессов формирования геологических образований;
- выделением аномалий в геологических и геофизических полях;
- классификацией геологических объектов по особенностям их внутреннего строения;
- разработкой приемов интерполяции и экстраполяции при оконтуривании геологических объектов;
- выбором оптимальной густоты и формы сети наблюдений при изучении геологических объектов.

1.4 Принципы и методы геолого-математического моделирования

Использование математического моделирования в геологии связано с целым рядом проблем. Математическая модель, равно как и любая иная модель, - это упрощенный аналог изучаемого объекта.

В связи со сложностью геологических объектов, ни одна математическая модель не способна отобразить все без исключения их свойства. По этой причине, с целью описания разных свойств одного и того же объекта, зачастую требуется применять разные математические модели. При этом нужно удостовериться, что подобранная модель довольно подробно отображает непосредственно те свойства объекта, которые напрямую оказывают влияние на решение поставленной задачи.

Математические модели не имеют возможности подробно охарактеризовать исследуемые свойства. Они базируются на определенных допущениях о характере свойств объекта моделирования. По этой причине нужно, чтобы данные допущения не приводили к принципиальному искажению реальных свойств объекта в рамках поставленной задачи.

Так как, задачи, которые встречаются в практике геологических исследований, достаточно многообразны, то имеет место быть ситуация, при которой для

моделирования одного и того же свойства объекта нужно применение разных моделей.

Определенные трудности порой появляются еще по причине отсутствия четких пределов геологических совокупностей и рассмотренных выше особенностей их исследования.

Решение геологических задач на базе математического моделирования – это достаточно трудный процесс, состоящий из следующих этапов:

1. Формулирование геологической задачи.
2. Определение геологической совокупности, то есть установка пределов геологического объекта или же временного интервала геологического процесса.
3. Выявление ключевых свойств объекта или же параметров процесса в рамках установленной задачи.
4. Переход от геологической совокупности к опробуемой и выборочной с учетом особенностей методов исследования.
5. Выбор типа математической модели.
6. Формулирование математической задачи в рамках той математической модели, которая была выбрана.
7. Выбор метода для решения математической задачи.
8. Решение математической задачи на базе вычисления параметров математической модели объекта.
9. Интерпретация полученных результатов применительно к геологической задаче.
10. Оценка вероятности и величины возможной ошибки посредством неадекватности модели и объекта.
11. Подводя итог вышесказанному, можно сделать вывод о том, что этапу непосредственно математического моделирования предшествуют этапы по созданию геологической модели (опробуемой и выборочной геологической совокупности). По этой причине, модели, которые применяют для решения геологических задач математическими методами принято называть геолого-математическими.

Правильность итогового вывода при решении задач на базе геолого-математического моделирования находится в непосредственной зависимости от того, насколько верны решения, которые были приняты на каждом из этапов.

Несложно подметить тот факт, что решения на большинстве этапов принимают исходя из особенностей геологических задач и свойств геологических объектов, вследствие этого, они всецело находятся в компетенции геолога.

Консультант математик имеет возможность оказать значительную помощь геологу только при выборе метода решения геологической задачи. Как видно из многолетнего опыта, основная масса ошибок при применении математических методов в геологии допускается не из-за недостаточной подготовки геологов, а по причине того, что не учитывались особенности геологических объектов и задач.

Вследствие этого, при изложении последующего материала на эти нюансы геолого-математического моделирования обращено предельное внимание.

1.5 Методы изучения геологических объектов

Тот факт, что геологические образования и процессы являются недоступными для непосредственного наблюдения, то это определило достаточно широкое применение в практике геологических исследований выборочных методов исследования, при помощи естественных и искусственных обнажений, в пределах которых отбирают образцы и пробы для всевозможных анализов и исследований.

Локальные площади наблюдений и отбираемые пробы несопоставимо малы, если их сравнивать с площадями и объемами недр, на которые распространяются данные наблюдения. Поэтому возникают трудности с пространственным размещением пунктов локальных наблюдений, систематизацией выборочных данных и их распространением на прилегающие объемы недр.

Геологи обычно судят о свойствах всей геологической совокупности по ее какой-либо части. Эту часть геологической совокупности М.Розенфельд назвал опробуемой совокупностью.

Степень соответствия свойств опробуемой совокупности и изучаемой геологической совокупности находятся в непосредственной зависимости от таких показателей как: расположение, густота и общее количество точек наблюдения, размер, ориентировка, форма, объем отбираемых проб или способ измерения данного свойства.

Выделяют 3 главных системы расположения точек наблюдения:

1. Равномерное расположение точек наблюдения.
2. Случайное расположение точек наблюдения.
3. Многостадийное расположение точек наблюдения.

Более распространено равномерное опробование. При данном виде опробования, точки наблюдения в плоскости изучаемого объекта распределяются по правильной геометрической сети. Это опробование разрешает с одинаковой детальностью изучить все части изучаемого объекта и в связи с вышеуказанным фактом является основным при поисках и разведке месторождений полезных ископаемых.

Случайное опробование, как правило, используют, когда исследователей не интересуют закономерности изменения изучаемого свойства в пространстве или же точно известно отсутствие таких закономерностей. Кроме этого, случайное опробование применяют при невозможности или затруднительности создания сети равномерных наблюдений. Так, к примеру, в гористой местности берутся пробы главным образом из естественных обнажений, размещение которых в пределах исследуемой площади близко к случайному. Случайный способ может быть рекомендован и при отборе проб для проведения контрольных анализов.

Многостадийное опробование используется для изучения свойств сложных геологических объектов на различных масштабных уровнях их строения. С этой целью объект разделяют на участки, которые соответствуют элементам его неоднородности, а в них, в свою очередь, выделяют наиболее мелкие элементы неоднородности.

родности и т.д. В пределах каждого участка опробуется конкретная часть элементарных участков более высокого порядка. В результате этого общее количество наблюдений при данном опробывании значительно меньше по сравнению с равномерным. Многостадийное опробывание используют при составлении ландшафтных карт. В начале по итогам дешифрирования космоснимков масштаба 1:200000 и 1:500000 выполняется районирование территории по типам ландшафтов, далее в пределах каждого из этих типов выделяют ландшафты речных долин, водоразделов, склонов и т.д. С целью определения границ элементарных ландшафтов применяют аэрофотоснимки в масштабе 1:50000, а их основные характеристики, такие как состав и мощность рыхлых отложений, тип почвы и растительности оценивают методом изучения "ключевых участков" (сравнительно маленькие по площади участки, на которых выражены все особенности представленного ландшафта).

Каждой геологической совокупности можно поставить в соответствие набор числовых характеристик, которые были получены в результате измерения или анализа каких-либо свойств геологических объектов. Эти наборы числовых данных именуется выборочными (статистическими) совокупностями.

Для верного решения поставленных геологических задач принципиальное значение имеет однозначное и четкое определение соотношений геологической и выборочной совокупностей. Для определения конкретной геологической совокупности нужно сначала определить ее элементарные составляющие (исследуемые объекты), границы и виды последующих числовых измерений.

В зависимости от целей и задач исследований, геолог уточняет объекты (элементарные составляющие) и границы геологических совокупностей.

У. Крамбейн разделяет элементарные составляющие геологических совокупностей на две большие группы:

1. Те, которые образованы первичными индивидами (объектами).
2. Те, которые образованы наборами исходных объектов.

К совокупностям, которые образованы первичными индивидами (объектами) относят совокупности ископаемых организмов, минералов в шлихах или шлифах и др. По каждому из этих объектов измеряется одно свойство, несколько свойств или производится оценка средних значений свойств в группировках исследуемых объектов.

К совокупностям, которые образованы наборами исходных объектов относят совокупности образцов или проб, по которым определяются следующие показатели: физико-химические свойства, гранулометрический состав, содержание полезных или вредных компонентов и др.

В данных наборах свойства каждого исходного объекта не измеряют, а оценивают средние значения тех или иных свойств в объемах проб или образцов. Отличительная особенность этой группы совокупностей - это зависимость числовых характеристик свойств от размеров и объемов проб.

2 Специфика геологических данных

2.1 Источники данных и их типы

Пространственными данными (геоданные, геофизические данные, геопространственные данные) называются цифровые данные о пространственных объектах, которые содержат информацию об их местоположении и свойствах, пространственных и непространственных атрибутах.

Как правило, состоящие из 2-х взаимозависимых частей: позиционных и непозиционных данных, т.е. описания пространственного положения и тематического содержания данных, тополого-геометрических и атрибутивных данных.

Подробное описание пространственных данных формулируется из взаимозависимых описаний топологии, геометрии и атрибутики объектов.

Пространственные данные совместно с их смысловой средой составляют базу информационного обеспечения геоинформационных систем (ГИС).

При вводе в машинную среду применяются различные источники пространственных данных. Качество этих данных характеризуется такими показателями как: точность (безошибочность), достоверность, полнота, непротиворечивость.

Источниками пространственных данных являются цифровые или аналоговые данные, служащие основой информационного обеспечения ГИС. Основные типы источников пространственных данных следующие:

1. Картографические источники (карты, атласы, планы и другие картографические изображения).
2. Данные дистанционного зондирования.
3. Данные режимных наблюдений на гидрометеопостах, океанографических станциях и т.д.
4. Статистические данные ведомственной и государственной статистики и данные переписей.

Рассмотрим вышеуказанные типы пространственных геологических данных более подробно.

1. Картографические источники. К ним относят:

а) общегеографические карты: топографические (масштаб 1:200 000 и крупнее), обзорно-топографические (мельче 1:200 000 до 1 000 000 включительно) и обзорные (мельче 1:1 000 000) карты включают различные сведения о рельефе, гидрографии, почвенно-растительном покрове, населенных пунктах, хозяйственных объектах, путях сообщения, линиях коммуникации, границах. В геоинформатике такие карты работают для 2-х целей - получить информацию об указанных объектах местности и их привязку. К данной группе источников относят фотокарты и космофотокарты, которые составлены по результатам аэро- или космической съемки, с нанесенными на них горизонталями и иной картографической нагрузкой, являющейся обычной для общегеографических карт. Использование фотокарт в качестве источников данных дает возможность непосредственного применения для этой цели цифровых моделей местности, создаваемых в программах.

Карты природы - более различная по тематике группа карт, которая включает карты геологического строения и ресурсов недр, геофизические, рельефа земной поверхности и дна океанов, метеорологические и климатические, гидрологические и океанографические, почвенные, геоботанические, зоогеографические, медико-географические, ландшафтные и общие физико-географические, охраны природы.

Геофизические карты делят на карты магнитного поля (магнитные аномалии, магнитное склонение, вековой ход элементов геомагнитного поля), карты гравитационного поля (вертикальные движения земной коры, изменение силы тяжести, строение земной коры, гравиметрические карты), карты сейсмических явлений и вулканизма (сейсмическое районирование, землетрясения, цунами, вулканы). Группу карт рельефа составляют гипсометрические, морфометрические

(углы наклона местности, экспозиция склонов, горизонтальное и вертикальное расчленение рельефа и др.), геоморфологические карты.

Тематика климатических карт достаточно различна. К ним относят характеристики климатообразующих факторов, карты термического и ветрового режима, режима увлажнения, атмосферных явлений, прикладные синтетические и комплексные карты, например агроклиматические, биоклиматические и др. В меньшей степени традиционны карты оледенения, лавин, мерзлоты.

Из карт почв и земельных ресурсов базовой является типологическая почвенная карта, которая отображает генетическую характеристику почв, их механический состав и почвообразующие породы. В зависимости от того, какая является специфика территории создают карты эрозии, засоления почв и др. При описании растительного покрова отображают современный растительный покров (на фоне коренного), леса, естественные кормовые угодья.

Карты народонаселения. Эти карты отображают и характеризуют следующие показатели:

- размещение населения по территории и расселение (характеристика численности населения в пунктах и по районам, плотность сельского населения, равномерность размещения населенных пунктов, типы расселения и т.д.);

- этнографическую и антропологическую обстановку народонаселения (деление и описание населения по национальности, образу жизни, культуре и др.);

- демографическую обстановку (отображение половозрастной структуры, естественного и механического движения населения);

- социально-экономическую обстановку (отображение социально-классовой структуры общества и развернутое описание трудовых ресурсов).

Карты экономики. Этот вид карт в большей степени распространен и различен среди карт социально-экономической тематики. В данной тематике карты делят на карты промышленности с подразделением на добывающую и обрабатывающую или более подробно по каждой отрасли промышленности (нефтяная,

угольная, пищевая, кожевенная, текстильная, деревообрабатывающая, металлообрабатывающая, химическая и др.). Наиболее многочисленны карты сельского хозяйства. Достаточно широко применяется характеристика природных ресурсов, довольно часто с их хозяйственной оценкой и прежде всего земельных фондов, трудовых ресурсов, материально-технической базы сельского хозяйства и т.д.

Карты транспорта содержат данные о различных проявлениях деятельности всех разновидностей транспорта (автомобильного, железнодорожного и т.д.), помимо этого, они описывают их общую комплексную характеристику. На картах изображаются средства связи.

Карты науки, подготовки кадров, обслуживания населения.

Этот вид карт сопряжен равно как с картами народонаселения, так и экономики. В результате этого отдельные виды карт иногда характеризуются в 2х предыдущих разделах (карты торговли, связи и т.д.), а иногда их выделяют в качестве самостоятельных групп в пределах карт науки, подготовки кадров и обслуживания населения. В таком случае однозначной классификации карт не существует. Поэтому укажем лишь на один из возможных вариантов, когда выделяются карты образования, науки, культуры, здравоохранения, физкультуры и спорта, бытового и коммунального обслуживания, туризма и т.д. Отдельно выделяются политические, административные и исторические карты.

Следует отметить особую роль серий карт и комплексных атласов, где сведения приводятся в единообразной, систематизированной, взаимно согласованной форме: по проекции, масштабу, степени генерализации, современности, достоверности и другим параметрам. Такие наборы карт особенно удобны для создания тематических баз данных.

2. Данные дистанционного зондирования

Являются одним из базовых источников данных для ГИС. Они объединяют все типы данных, которые получают с носителей космического (пилотируемые орбитальные станции, автономные спутниковые съемочные системы и т.п.) и авиационного базирования (самолеты, вертолеты и микроавиационные радио-

управляемые аппараты), и составляют огромную часть дистанционных данных как антонима контактных (прежде всего наземных) видов съемок, способов получения данных измерительными системами в условиях физического контакта с объектом съемки.

К неконтактным (дистанционным) методам съемки кроме аэрокосмических относят различные измерительные системы морского (наводного) и наземного базирования, включая, например, фототеодолитную съемку, сейсмо-, электро- и магниторазведку и другие методы геофизического зондирования недр; гидроакустические съемки рельефа морского дна с помощью гидролокаторов бокового обзора, другие способы, базирующиеся на регистрации собственного или отраженного сигнала волновой природы.

Аэрофотосъемки регулярно проводятся в России с тридцатых годов и в настоящее время за более чем полувековой период собран фонд снимков, который полностью охватывает страну, а для многих районов с многократным перекрытием, что является наиболее важным при изучении динамики объектов.

Материалы аэрофотосъемки применяются зачастую для топографического картографирования страны, также довольно обширно используются в геологии, в лесном и сельском хозяйстве. Космические снимки стали поступать с шестидесятых годов, и в настоящий момент времени их количество достигает несколько десятков миллионов.

3. Данные режимных наблюдений

Достаточно обширны возможности применения стационарных измерительно-наблюдательных сетей с целью получения, в первую очередь, гидрологических и метеорологических данных, систематически сбор и обработка которых содержит определенную историю. Гидрологические материалы включают в себя информацию о прошлом, настоящем и для некоторых элементов будущем состоянии рек, озер и водохранилищ. Такие материалы собираются сетью опорных гидрометеорологических станций.

В массив гидрологических наблюдений включаются данные о ледовых явлениях на реках с устойчивым и неустойчивым ледоставом; о дождевом паводочном стоке; о средних месячных температурах воды по бассейнам; о средних, высших и низших уровнях воды; о средних месячных расходах воды; о максимальных расходах воды и слоях стока за половодье; о гранулометрическом составе взвешенных, влекомых и донных наносов; о расходах взвешенных наносов и мутности воды, и кроме вышеперечисленного ряд иных данных.

Наиболее многочисленны метеорологические наблюдения. Эти наблюдения содержат информацию по синоптическим показателям у поверхности земли, показатели термобарического поля в свободной атмосфере (средние месячные значения давления, геопотенциала и температуры воздуха для уровня моря и основных изобарических поверхностей); нормы и аномалии средней месячной температуры воздуха; месячные суммы осадков в процентах от нормы; нормы месячных сумм осадков; характеристики ветра в свободной атмосфере; а также разнообразные иные характеристики, количество которых достигает нескольких десятков.

Разнообразными являются работы, которые проводят исходя из потребностей океанологии. Сбор информации выполняется в глобальном масштабе с применением научно-исследовательских судов, плавмаяков, судов погоды, океанографических буйковых станций и т.д. Состав таких сведений содержит такие гидрометеорологические характеристики как: температура и соленость воды в приземном слое океана (моря) и на стандартных горизонтах (до глубины 500 м); высота и период ветровых волн и зыби; направление, скорость и направление ветра на установленной высоте; скорость и направление течений в поверхностном слое и на некоторых горизонтах; температуру воздуха; точки росы; температуру атмосферное давление; общую солнечную радиацию и т.д.

4. Статистические данные

Для того, чтобы упорядочить всю совокупность данных государственной службой предложены группы показателей по отраслям статистики. В качестве таких групп в России применялись отрасли статистики, такие как:

1. Промышленность.
2. Природные ресурсы и окружающая среда.
3. Технический прогресс.
4. Сельское хозяйства и заготовки.
5. Капитальное строительство.
6. Транспорт и связь.
7. Торговля.
8. Труд и заработная плата.
9. Население, здравоохранение и социальное обеспечение.
10. Народное образование, наука и культура.
11. Бюджет населения.
12. Жилищно-коммунальное хозяйство и бытовое обслуживание населения.
13. Материально-техническое снабжение и перепись.
14. Финансы.

2.2 Представление графической (пространственной) информации

Все без исключения форматы графических файлов возможно поделить на 2 типа: растровые и векторные. Они различаются принципом формирования изображения.

ГИС способен работать с векторными и растровыми форматами данных. В векторной модели сведения о точках, линиях и полигонах кодируется, а затем хранится в виде набора координат X,Y. Месторасположение точки (точечного объекта), к примеру, буровой скважины, описывается парой координат (X,Y). Линейные объекты, например, реки, дороги, трубопроводы, сохраняются равно как

наборы координат X,Y. Полигональные объекты, типа речных водосборов, земельных участков или же областей обслуживания, хранятся в виде замкнутого набора координат. Наиболее удобной векторная модель является при описании дискретных объектов и меньшей степени подходит для описания непрерывно меняющихся свойств, как например, типы почв или доступность объектов. Растровая модель является наиболее оптимальной при работе с непрерывными свойствами. Растровое изображение - это набор значений для отдельных элементарных составляющих (ячеек), оно аналогично отсканированной карте или картинке. Эти вышеуказанные 2 модели обладают своими достоинствами и недостатками.

2.1.1 Растровая модель графической информации

Картинка в растровой модели графической информации формируется аналогично мозаике, из отдельных точек (пикселей), любая из которых полностью определяется 2-мя базовыми параметрами, такими как: координаты расположения и цвет. В большей степени близким аналогом растрового изображения считается изображение на экране монитора компьютера (или обычного телевизора), что формирует электронный луч, который проскакивает поочередно по каждой строке создаваемого кадра изображения (растра). Почти все растровые форматы имеют возможность нести дополнительную информацию под которой подразумеваются разнообразные цветовые модели изображения, вектора, альфа-каналы (дополнительный канал, при помощи которого возможно сохранение выделенных или прозрачных областей изображения), слои различных типов, анимацию, возможности сжатия и многое иное. Плюсы растровых изображений заключаются в их возможности передачи тончайших нюансов изображения и, кроме того, в достаточно больших возможностях по его редактированию, что выражается в простом доступе к любому пикселю изображения, а также имеется возможность индиви-

дуального изменения любого из его параметров. Серьезным минусом является достаточно большие размеры полученного файла.

К наиболее распространенным форматам растровых изображений относят следующие форматы:

BMP (Windows Device Independent Bitmap). Данный формат является самым простым растровым форматом, этот формат также известный под именем DIB. Он является родным форматом Windows и поддерживается всеми графическими редакторами, которые работают под ее управлением. BMP данные о цвете хранятся только в модели RGB, поддерживаются как индексированные цвета (до 256 цветов), так и полноцветные изображения. Вся "мультиплатформенность" формата заключена только в поддержке Windows и почти полностью исчезнувшей OS/2.

Из-за примитивнейшего алгоритма записи изображения, при обработывании файлов формата BMP весьма в меньшей степени используются системные ресурсы, по этой причине данный формат зачастую применяется для того, чтобы хранить логотипы, экранные заставки, иконки и другие элементы графического оформления программ.

PCX (Soft Publisher's Paintbrush). Приблизительно аналогичными возможностями, как и предыдущий формат, обладает формат PCX, который был создан еще в начале компьютерной эпохи. Изображения в формате PCX возможно посмотреть большинством программ под DOS, в том числе и внутренним просмотрщиком Norton Commander. Цветовые возможности: 1, 2, 4, 8 или 24 - битовый цвет, поддерживается только лишь схема RGB, при этом целиком отсутствуют возможности по сохранению монохромного изображения в оттенках серого. Равно как и BMP, рассматриваемый формат в большей степени является устаревшим и поддерживается современными графическими программами только с целью совместимости с устаревшим программным обеспечением.

GIF (Graphics Interchange Format). Этот формат является одним из самых известных форматов на просторах сети Internet, был предложен фирмой CompuServe в 1987 году. Его отличает возможность использования режима индекс-

сированных цветов (не более 256), что ограничивает область применения формата изображениями, которые имеют резкие цветовые переходы. Формат GIF - это излюбленный формат веб-мастеров, которые используют его с целью сохранения многочисленных элементов оформления своих страничек. Маленькие размеры файлов изображений обусловлены использованием алгоритма сжатия без потерь качества LZW, из-за чего изображения в этом формате в большей степени удобные для пересылки по узким каналам связи глобальной сети. Самым заметным отличием является возможность применения режима постепенного проявления изображения (interleaved), в данном режиме строки изображения выводятся на экран не подряд, а в конкретном порядке - сначала каждая 8-я, затем каждая 4-я и т.д. Таким образом, всецело изображение показывается в четыре прохода. Данный факт дает возможность еще до полной загрузки изображения понять его суть, и, при необходимости, произвести прерывание его загрузки. В 1989 году формат обновили и дали ему название GIF 89A. От предшествующей версии он отличается тем, что у него есть дополнительный альфа-канал, который позволяет реализовать эффект прозрачности (не больше одной градации) и дает возможность хранения в одном файле нескольких картинок с указанием времени показа каждой (дает возможность создания анимированного изображения). Данное свойство формата GIF89A зачастую применялось при создании анимированных рекламных баннеров. На сегодняшний день GIF интенсивно вытесняет формат SWF (векторный формат Flash Player).

PNG (Portable Network Graphics). Этот вид формата сжимает графическую информацию без потерь качества. Данный формат использует алгоритм Deflate и в отличие от GIF или TIFF форматов он сжимает растровые изображения как по горизонтали, так и по вертикали. Этот факт дает в большей степени высокую степень сжатия и поддерживает цветные фотографические изображения вплоть до 48-битных включительно. В качестве недостатка данного формата можно указать тот факт, что здесь нельзя создавать анимационные ролики, хотя в настоящее время, при повальном переходе почти полностью всей анимации на технологию Flash, это

не является актуальным. Но однако, данный формат дает возможность создания изображения с 256 уровнями прозрачности за счет использования дополнительного альфа-канала с 256 градациями серого, и это, безоговорочно, дает возможность выделить его на фоне всех имеющихся в настоящее время форматов. В качестве иных особенностей отличающих этот формат от других форматов можно указать двумерную чересстрочную развертку (т.е. изображение проявляется постепенно не только по строкам, но и по столбцам) и встроенную гамма-коррекцию, которая позволяет сохранять изображения, яркость коих будет неизменна не только на любых машинах PC, но также на таких альтернативных платформах, как Mac, Sun или Silicon Graphics. В следствии того, что данный вид формата был создан для сети Internet, то в его заголовке не предназначено место для дополнительных параметров типа разрешения. Этот факт говорит о том, что данный формат плохо подходит для хранения изображений. С этой целью гораздо лучше подходит PSD или TIFF форматы. Однако, формат PNG замечательно подходит для публикации высококачественной растровой графики в Internet.

JPEG (Joint Photographic Experts Group). Этот вид формата является наиболее популярным и общепринятым стандартом. Он замечательно подходит для хранения фотографических изображений JPEG (или JPG). JPEG имеет возможность хранения только 24-битовых полноцветных изображений. Одноименный с форматом, довольно трудный алгоритм сжатия базируется на особенностях человеческого зрения. И хоть формат JPEG превосходно делает сжатие фотографии, но данное сжатие осуществляется с потерями и портит качество, однако, несмотря на это, данный формат можно достаточно легко настроить на небольшие, почти полностью незаметные для человеческого глаза, потери. К тому же, усилить сжатие и уменьшить потери качества возможно, если предварительно размыть изображение (например, используя фильтр blur в Photoshop). Но не стоит применять формат JPEG для хранения изображений, которые подлежат дальнейшей обработке, так как при каждом сохранении документа в данном формате процесс ухудшения качества изображения носит лавинообразный характер. В большей степени целесооб-

разным будет корректирование изображения в каком-нибудь ином подходящем формате, например TIFF, и только по завершении всех работ заключительную версию возможно сохранить в JPEG. Таким образом, возможно сохранение достаточно приемлемого качества изображения при наименьшем размере конечного файла. Имеется разновидность формата JPEG, которая именуется "прогрессивный JPEG". Отличие прогрессивного JPEG от обычного это то, что при выводе его на экран изображение появляется практически сразу, но оно достаточно низкого качества, а по мере загрузки качество улучшается (процесс аналогичен постепенному показу GIF). Формат JPEG не поддерживает анимацию или прозрачный цвет, и является пригодным в подавляющем большинстве случаев только для публикации полноцветных изображений, таких как фотографические.

TIFF (Tag Image File Format). Этот вид формата является универсальным при хранении растровых изображений. Он в достаточной степени обширно применяется, прежде всего, в издательских системах, которые требуют более лучшего качества. Функция записи изображений в приведенном формате - это один из основных признаков, определяющих высокий класс современных цифровых фотокамер. Этот формат поддерживает большое количество алгоритмов сжатия (в их числе известные LZW, Deflate или JPEG), типов изображений от битового (1-, 2-, 4-, 8-, 24- и 32- битные изображения) и индексированных цветов до LAB, CMYK и RGB. Помимо этого, имеются 2 разновидности формата, соответственно для IBM PC и Macintosh, которые обусловлены разным порядком записи чисел, и который осуществляется на данных платформах. Со сжатием LZW файл TIFF занимает практически столько же места, сколько и GIF, однако, в отличие от GIF, формат TIFF поддерживает полноцветные изображения и хранит в своем теле полные данные об изображении, таких как: разрешение, тип принтера и иные детали, являющиеся необходимыми при профессиональной работе с изображениями. В формате TIFF поддерживаются следующие чисто профессиональные возможности: альфа-каналы, обтравочные контуры, имеется возможность сохранять несколько копий изображения с различным разрешением и даже включать в файл слои. Вследствие

его совместимости со множественными профессиональными программными обеспечениями, применяемыми для обработки изображений, данный формат является наиболее удобным при переносе изображений между компьютерами разнообразных типов (к примеру, с PC на Mac и обратно).

PSD (Adobe Photoshop). Данный формат - это стандартный формат пакета Adobe Photoshop и главное его отличие от многочисленных обычных растровых форматов наличие возможности хранения слоев (layers). Этот растровый формат включает в себя большое количество дополнительных переменных, не уступая формату TIFF по их количеству). Формат PSD сжимает изображения и использует при этом алгоритм сжатия без потерь, в редких случаях даже намного сильнее, чем PNG (лишь в тех случаях, когда размеры файла измеряют не в килобайтах, а в десятках или даже сотнях мегабайт). Это формат поддерживает глубины цвета, вплоть до 16 бит на канал (48-битные цветные и 16-битные черно-белые), и кроме этого, альфа-каналы, прозрачность, контуры, слои, векторные надписи и т.д. Замечательно подходит для переноса или хранения изображений, которые содержат специфические, свойственные лишь Adobe Photoshop, элементы. Файлы формата PSD беспрепятственно можно читать множеством известных просмотрщиков графики, но, однако, не нужно забывать тот факт, что, открывая данные файлы в некоторых графических редакторах третьих фирм, даже тех, которые декларируют поддержку формата PSD, имеется возможность потери большей части их специфических возможностей (особенно в части работы со слоями). Со всеми вышеуказанными форматами растровых изображений с большим успехом осуществляет работу уже названный выше графический пакет Adobe Photoshop.

2.2.2 Векторная модель графической информации

Векторное изображение, в отличие от растровых изображений, является совокупностью отрезков кривых линий, которые можно описать математическими

выражениями, а также совокупностью цветных заливок. Для того, чтобы компьютер мог нарисовать прямую линию необходимы координаты 2-х точек, соединяющихся по наиболее краткому пути. Для дуги задают координаты центра окружности, радиус и т.д.

Подводя итог вышесказанному, можно сказать, что векторная иллюстрация представляет собой набор геометрических примитивов (простейших объектов, к примеру, окружности, линии, многогранники и т.д.), которые применяют при создании наиболее сложных изображений. Поэтому главным достоинством векторных форматов являются компактность полученных файлов, и кроме того, высокое качество получаемых изображений, при этом на данную особенность не влияет разрешающая способность устройства отображения. Недостатком данной модели графической информации является характерная трудоемкость при создании и редактировании сложных элементов изображений, а также существуют определенные проблемы, которые возникают при распечатывании векторных изображений на некоторых принтерах.

К распространенным векторным форматам относят следующие форматы:

CDR (Corel DRAW Document). Формат CDR принадлежит фирме Corel. Данный формат представляет собой внутренний формат векторного редактора Corel DRAW. В этом формате используется компрессия (сжатие) для векторов и растра и могут быть внедрены шрифты. Файлы CDR могут иметь рабочее поле 45 на 45 метров, и именно поэтому его применяют при изготовлении наружной рекламы. Применять CDR чрезвычайно удобно и легко. Этот формат обладает неопровержимым лидерством на платформе PC. Множественные программы на PC (FreeHand, Illustrator, PageMaker) имеют возможность импортировать файлы CDR.

PDF (Portable Document Format). Формат PDF изначально проектировали в качестве компактного формата электронной документации. Данный формат представляет собой практически полностью платформонезависимый формат, в текстовой части которого имеется возможность использования большинства шрифтов (содержащихся напрямую в документе, поэтому документ всегда будет выгля-

деть так, как его задумывал автор, причем на любом компьютере, независимо от применяемого на нем программного обеспечения) и гипертекстовых ссылок, и кроме того, графические иллюстрации любого типа (векторные или растровые). Для достижения наиболее меньшего размера PDF-файла применяется компрессия, при этом каждый вид объектов сжимается по более выгодному для него алгоритму. Просматривать документы в формате PDF и распечатывать их на принтере возможно при помощи программы Acrobat Reader, которая распространяется компанией Adobe бесплатно.

WMF (Windows Metafile). Данный формат является родным форматом Windows, но на данный момент векторным. Распознается почти полностью всеми программами Windows, которые в той или иной мере связаны с векторной графикой. Хотя, несмотря на представляющую простоту и универсальность, применять этот формат целесообразно лишь в крайних случаях, так как формат WMF не имеет возможности сохранять некоторые параметры, что возможно присвоить объектам в разнообразных векторных редакторах. Кроме вышеуказанного, этот формат не будет восприниматься Macintoshами. И, наконец, главный недостаток данного формата - это то, что он может исказить цветовую схему изображения. WMF является аббревиатурой от "Windows Metafile Format" (Метафайл Windows). Применяется формат WMF при обмене графической информацией между приложениями Microsoft Windows. Поддерживает растровую и векторную графику. В метафайле записаны команды интерфейса графических устройств (GDI-команды), любая из которых способна описывать одну графическую функцию. Для того, чтобы отобразить метафайл, программа передает данные команды специальной функции, что осуществляет воспроизводство изображения. Метафайлы обеспечивают средства хранения и выборки графической информации, которые независят от устройства .

DXF (Drawing eXchange Format). Формат DXF является открытым форматом файлов, который используется с целью обмена двумерной графической информацией между приложениями Системы автоматизации проектных работ (САПР).

Этот вид формата создан фирмой Autodesk для системы AutoCAD. Поддерживается почти полностью всеми CAD-системами на платформе PC.

К векторным форматам, которые используются геоинформационными системами (ГИС) относят:

SHP (Shapefile - файл форм). Файлы этого вида формата поддерживаются системами ArcView и ARC/INFO. Файл формата SHP является двоичным файлом, который содержит в себе графические данные («формы»). Все формы одного SHP-файла имеют один и тот же геометрический тип, который называется типом формы. Он указывается в заголовочной части файла. Каждый SHP-файл имеет возможность быть связанным с индексным файлом формата SHX, который не применяется при импорте, и с файлом табличных данных формата DBF. При импорте может быть установлена связь между объектами, которые загружаются из SHP-файла, и соответствующими записями таблицы, которые загружаются из DBF-файла. Shapefiles имеют возможность включать в себя линии, точки и полигоны. Любой объект в shapefile - это один географический объект в совокупности с его атрибутами.

MIF-MID. Данный формат представляет собой собственный формат экспорта системы MapInfo и поддерживается множеством систем в режиме экспорта, гораздо реже - в режиме импорта. Данный факт связан с тем, что система MapInfo классифицируется в мире как ГИС конечного пользователя и не рассчитана использовать ее напрямую при создании картографических БД "на экспорт". Сам формат обмена включает в себя 2 текстовых файлов: файл MIF включает в себе координатное описание объектов карты, а файл MID - описание атрибутивных данных, которые связаны с объектами карты и записываются построчно в виде прямоугольной таблицы. Соответствие между данными 2-х файлов базируется на основе порядковых номеров объектов карты и строк таблицы данных.

Этот формат отличает относительная простота, что представляется его достоинством, но он все-таки значительно ориентирован на внутреннюю модель описания данных карты системы MapInfo. Эта особенность способствует ограни-

чению области его применения. Существует определенное количество версий формата MIF-MID, которые различаются, главным образом, объемом сопровождающих векторную графику описательных данных.

Существует несколько векторных моделей пространственных данных в геоинформационных системах, а именно:

1. Топологическая модель.
2. Нетопологическая модель.

Пространственные данные - это цифровые данные о пространственных объектах, которые включают в себя сведения об их местоположении и свойствах, пространственных и непространственных атрибутах.

Топологическая модель (линейно-узловое представление) - это векторная модель данных, которая включает в себя как описание топологических отношений между пространственными объектами, так и их геометрию.

Топология в ГИС - это часть пространственных данных, которая служит для описания взаимного положения геометрических объектов и их частей в векторно-топологическом представлении. В научно-популярной литературе топологию зачастую именуют «геометрией на резиновом листе», так как ее наглядно возможно представить себе как геометрию фигур, которые нарисованы на идеально упругих резиновых листах и что можно подвергнуть сжатию, растяжению, или же изгибанию без разрывов.

Базовые элементы топологического представления данных это (рисунок 2.1):

1. Точка – пара координат XY.
2. Отрезок – линия, соединяющая 2 точки.
3. Вершина (vertex) – начальная или конечная точка отрезка.
4. Дуга (arc) – упорядоченный набор связных отрезков.
5. Узел – начальная или конечная вершина дуги.
6. Висячий узел – узел, принадлежащий только 1 дуге, у которой начальная и конечная вершины не совпадают.
7. Висячая дуга – дуга, имеющая висячий узел.

8. Замкнутая дуга – совпадают начальная и конечная вершины.

9. Полигон – область, образованная замкнутой дугой или набором связанных дуг, который образует замкнутый контур.

Полигоны в топологии имеют ребра, которые определяют границы полигонов, узлы, где ребра пересекаются, и вершины, что определяют форму ребер.

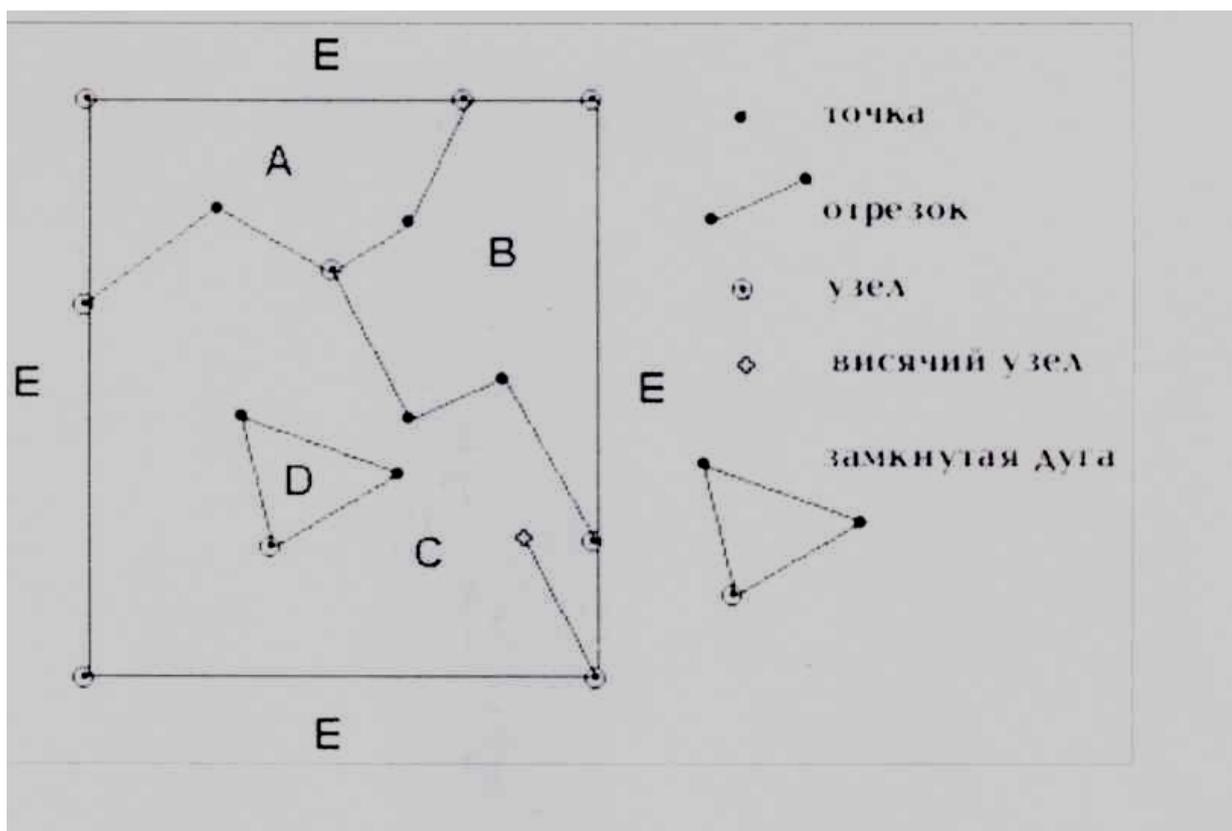


Рисунок 2.1 - Топологическое представление данных

Нетопологическая модель (модель "спагетти") - это представление линейных и полигональных пространственных объектов с описанием их геометрии в виде неупорядоченного набора дуг или же совокупности сегментов, которая не включает в себя описание топологических отношений между пространственными объектами. «Спагетти» - это название, которое в большей степени относится не к модели данных, а к её отсутствию, т.е. некорректно, хаотически и бес-

системно оцифрованным данным. В нетопологическом представлении любая линия, которая разделяет смежные объекты, дублируется.

2.3 Представление текстовой информации

2.3.1 Текстовые форматы

Рассмотрим текстовые форматы представления информации, которые в большей степени применяются :

ТХТ. Текстовый файл - это форма представления последовательности символов в компьютере. Любой символ из применяемого набора символов кодируется в виде одного байта, а в довольно редких случаях в виде последовательности подряд идущих двух, трёх и т.д. байтов.

Текстовые файлы разбиваются на несколько строк. На современных платформах разбиение на строки кодируется управляющим символом перевода строки, и реже последовательностью 2-х управляющих символов (на некоторых старых платформах разбиение на строки делалось по-другому). Текстовые файлы могут включать в себя печатные символы, а именно, цифры, буквы, а также знаки препинания и небольшое количество управляющих символов, таких как знаки табуляции и перевода строки. Языки программирования практически полностью применяют простой текст в качестве базы.

CSV (Comma Separated Values - это значения, которые разделены запятыми). CSV - это текстовый формат, который предназначен для представления табличных данных. Любая строка файла - это одна строка таблицы. Значения отдельных колонок разделяются разделительным символом (delimiter), к примеру, запятой (,), точкой с запятой (;), символом табуляции. Текстовые значения об-

рамляются символом двойные кавычки ("); если в значении встречаются кавычки - они представляются в файле в виде двух кавычек подряд.

Рассмотрим следующий пример:

Исходный текст:

1997, Ford, E350, "ac, abs, moon", 3000.00

1999, Chevy, "Venture ""Extended Edition""", , 4900.00

1996, Jeep, Grand Cherokee, "air, moon roof, loaded MUST SELL!", 4799.00

Результирующая таблица:

1997	Ford	E350	ac, abs, moon	3000.00
1999	Chevy	Venture "Extended Edition"		4900.00
1996	Jeep	Grand Cherokee	air, moon roof, loaded MUST SELL!	4799.00

Программами для редактирования файлов этого формата являются Microsoft Excel, Numbers, CSVed, Open Office.org Calc, KSpread.

XLS является форматом данных, который используется табличным процессором (редактором) Microsoft Office Excel. Электронные таблицы в формате XLS имеют возможность содержать в себе не только текстовые и числовые данные, но и, помимо этого, также мультимедийные и графические элементы, мощные математические функции, которые разрешают проводить достаточно сложные финансовые, статистические и другие расчеты.

DBF (Data Base Format). Этот вид формата осуществляет хранение данных и применяется как один из стандартных методов хранения и передачи информации системами управления базами данных (СУБД), электронными таблицами и т.п.

Файл вышеуказанного формата разделяется на заголовок, который хранит информацию о структуре базы (в современных версиях он хранит информацию и о

некоторых иных характеристиках, к примеру, применяемой кодовой таблице) и количестве записей, и непосредственно область данных, которая представляет собой последовательно организованную таблицу из записей определенной длины. В свою очередь записи разделяются на поля аналогично определенной длины, которые хранятся в файле собственно друг за другом, образуя запись. Первое поле DBF - формата - это пометка удаления. Длина поля - один символ. Если поле установлено в значение русское «х», то полагается, что запись отмечена на удаление.

Удалённые записи возможно восстановить, или физически удалить при проведении специальной операции, которая называется упаковкой.

2.3.2 Базы данных

Базы данных представляют собой информационные модели, которые содержат в себе информацию об объектах и их свойствах. Базы данных хранят информацию о группах объектов с аналогичными свойствами. Информация в базах данных хранится в определенном упорядоченном виде (к примеру, в записной книжке все записи упорядочены по алфавиту, в библиотечном каталоге записи упорядочены или по области знания, или также по алфавиту).

База данных - это информационная модель, которая позволяет упорядоченно хранить данные о группе объектов, что обладают аналогичным набором свойств.

Базой данных (БД) зачастую упрощённо или ошибочно именуют Системы Управления Базами Данных (СУБД). Нужно различать набор данных (собственно БД) и программное обеспечение, которое предполагается для организации и ведения баз данных (СУБД). На уровне физической модели электронная база данных является собой файл или набор таких файлов в формате Excel, TXT, DBF, CSV, XML, или же в специализированном формате определенной Системы Управления Базами Данных.

Современные СУБД, в их числе те, которые применяются в программном обеспечении ГИС, отличаются типами поддерживаемых в них моделей данных, среди коих выделяются иерархические, сетевые и реляционные модели и предназначенные для них программные средства СУБД. Последние из которых - реляционные СУБД – нашли определенно достаточно обширное применение при разработке программного обеспечения ГИС.

Иерархические базы данных. Иерархические базы данных графически возможно представить в качестве дерева, которое состоит из объектов разнообразных уровней. Верхний уровень занимает один объект, далее второй уровень занимает объекты второго уровня и т.д.

Между объектами присутствуют связи, каждый объект имеет возможность включать в себя несколько объектов, которые принадлежат более низкому уровню. Эти объекты находятся в отношении предка (объект более близкий к корню) к потомку (объект более низкого уровня), при этом существует возможность того, чтобы объект-предок не имел потомков или имел их несколько, тогда как у объекта-потомка обязательно только 1 предок. Объекты, которые имеют одного общего предка, именуют близнецами.

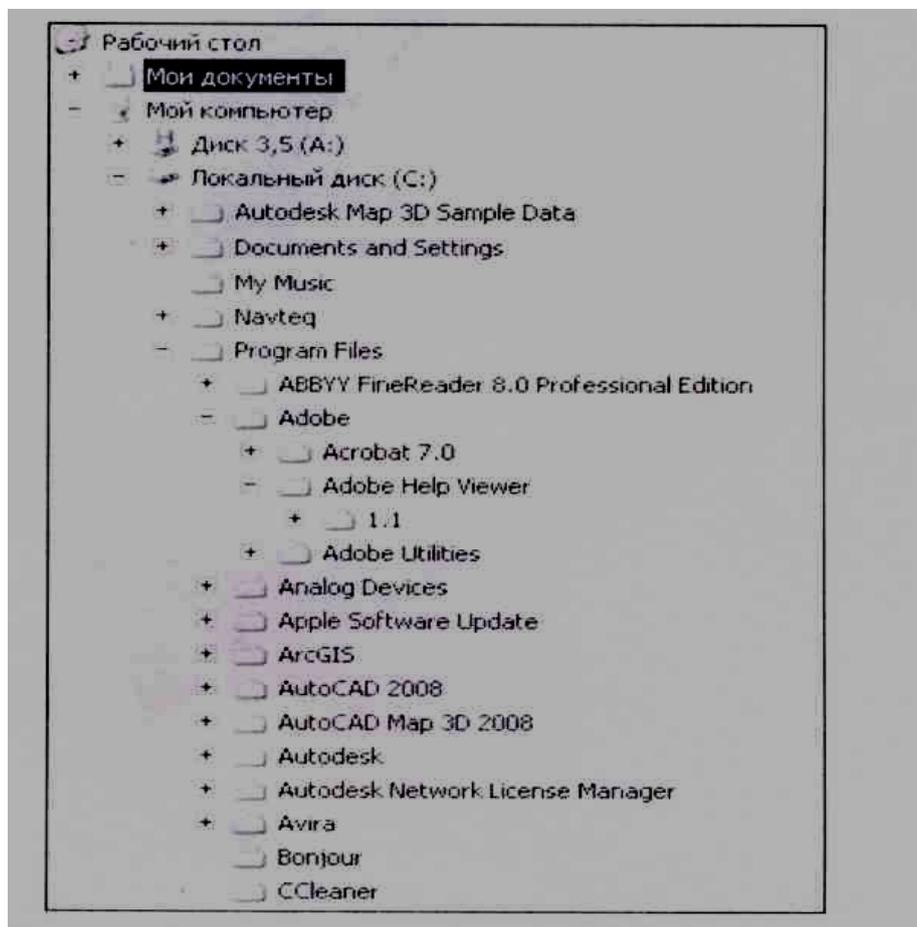


Рисунок 2.2 - Пример структуры иерархической БД

К примеру, иерархическая база данных - это Каталог папок Windows, с которым имеется возможность работать, запустив Проводник (рисунок 2.2).

Сетевые базы данных. Сетевая база данных образуется посредством обобщения иерархической базы данных за счет допущения объектов, которые имеют более 1 предка, т. е. любой элемент вышестоящего уровня может быть связан одновременно с любыми элементами следующего уровня. Вообще, на связи между объектами в сетевых моделях не накладывается никаких ограничений. Примером сетевой базы данных практически является Всемирная паутина глобальной компьютерной сети Интернет. Гиперссылки связывают между собой сотни миллионов документов в общую распределенную сетевую базу данных.

Реляционные базы данных. Реляционная база данных - база данных, которая базируется на реляционной модели. Происхождение слова «реляционный» от английского «relation», что означает отношение.

Теория реляционных баз данных разработана в 1970 году доктором Коддом, который являлся сотрудником компании IBM. В реляционных базах данных все данные отображены в виде простых таблиц, которые разбиты на строки и столбцы, и на пересечении этих строк и столбцов располагаются данные. Запросы к этим таблицам возвращают таблицы, которые имеют возможность становиться предметом для следующих запросов. Абсолютно любая база данных имеет возможность включать в себя несколько таблиц. Тезисно характерные особенности реляционной базы данных можно представить таким образом:

- данные хранятся в таблицах, которые состоят из строк («записей») и столбцов («атрибутов»);

- на пересечении любого столбца и строчки стоит в точности 1 значение. У любого столбца имеется собственное имя, что служит его названием, и все значения в одном столбце имеют 1 тип;

- запросы к базе данных возвращают конечный результат в виде таблиц, которые к тому же имеют возможность выступать в качестве объекта запросов;

- строки в реляционной базе данных неупорядочены. Упорядочивание реализуется в тот момент, когда происходит формирование ответа на запрос.

Вышеуказанные модели характеризуются простотой структуры данных, удобным для пользователя табличным представлением и возможностью применения формального аппарата алгебры отношений и реляционного исчисления для обработки данных.

Общепринятым стандартом языка работы с реляционными базами данных является язык SQL.

Для работы с реляционными БД используют Реляционные СУБД. Реляционная СУБД (РСУБД, по-другому Система управления реляционными базами данных, СУРБД) - СУБД, управляющая реляционными базами данных.

3 Одномерный статистический анализ и его применение

3.1 Характеристика одномерного статистического анализа

Одномерная статистическая модель используется при исследовании какого-либо одного свойства. К примеру, пусть имеется система, которая состоит из многочисленных однородных геологических объектов. Выборочным методом выберем из множества объектов и у каждого из них проведем измерение характеристики свойства x . Результаты измерений назовем x_1, x_2, \dots, x_n и далее составим из них матрицу, в которой число строк будет равно n , а число столбцов $k = 1$.

Одномерная статистическая модель базируется на 3-х основных гипотезах. Перечислим их:

1. У измеренных значений x_1, x_2, \dots, x_n случайный характер.
2. Значения x_1, x_2, \dots, x_n не зависят друг от друга.
3. Значения формируют однородную совокупность.

Измеренные значения обычно именуют как реализации случайной величины x .

Гипотеза о случайном характере свойств объясняется тем, что природные геологические системы и объекты представляются достаточно сложными. Каждое измеренное значение от многочисленных разнонаправленных факторов. Помимо вышеизложенного, любое из этих измерений сопровождается случайной погрешностью. Это суждение дает возможность использовать с целью точной математической обработки значений x_1, x_2, \dots, x_n следующий аппарат (теоремы, уравнения, формулы, законы) теории вероятностей.

Вторая суждение или гипотеза о независимости (самостоятельности) измеренных значений гораздо в меньшей степени является неоспоримой. Она подразумевает тот факт, что на конечный результат любого единичного измерения не оказывают влияния результаты предшествующих либо рядом располагающихся замеров. Из данной гипотезы следует важное следствие о том, что для математической обработки не значительно пространственное размещение пунктов наблюдений, т.е. имеется возможность располагать результаты измерений в любом порядке, на выводы этот факт влиять не будет. Вышеуказанная гипотеза не всегда отвечает реальности: смежные измерения зачастую находятся в зависимости друг от друга, что проверяется посредством специального математического аппарата.

Статистическая обработка результатов измерений содержит в себе смысл лишь для однородных совокупностей, на чем и базируется третья гипотеза. Когда совокупность является неоднородной, то в таком случае ее нужно поделить на однородные совокупности и каждую из них изучить в отдельности.

Базовой основой множества вычислений является расчет статистических характеристик случайной величины. В большей степени популярные статистические характеристики одномерной случайной величины это: *размах, медиана, мода, среднее значение, дисперсия, среднеквадратичное отклонение, коэффициент вариации, асимметрия, эксцесс.*

К примеру, 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 является средним членом упорядоченного ряда значений. Для того, чтобы найти медиану необходимо расположить все значения в порядке возрастания или убывания, затем определить средний по порядку член ряда. В

случае n - четного числа в середине ряда окажутся 2 значения, тогда медиана будет равна их полусумме.

Мода - это значение случайной величины, которое встречается больше всего.

Средним значением является среднеарифметическое из всех измеренных значений:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (3.1)$$

Медиана, мода и среднее значение относят к характеристикам положения. Около данных характеристик происходит группировка измеренных значений случайной величины.

Дисперсией называется число, которое равно среднему квадрату отклонений значений случайной величины от ее среднего значения (Дисперсия случайной величины является мерой разброса данной случайной величины, т.е. ее отклонением от математического ожидания):

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3.2)$$

Среднеквадратичным отклонением называется число, которое равно квадратному корню из дисперсии:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.3)$$

Среднеквадратичное отклонение имеет размерность, которая совпадает с размерностью случайной величины и среднего значения. К примеру, когда значения случайной величины были измерены в метрах, то и среднеквадратичное отклонение аналогично будет выражаться в метрах.

Коэффициентом вариации называется отношение среднеквадратичного отклонения к среднему значению:

$$V = \frac{\sigma}{\bar{x}} \quad (3.4)$$

Коэффициент вариации отображается в долях единицы или (при осуществлении умножения на 100) в процентах. Вычисление коэффициента вариации содержит в себе смысл для положительных случайных величин.

Дисперсия, среднеквадратичное отклонение и коэффициент вариации, и, кроме этого, размах - это *меры рассеяния* значений случайной величины вокруг среднего значения. Чем они будут больше, тем самым будет сильнее рассеяние.

Ассиметрией является степень асимметричности распределения значений случайной величины относительно среднего значения:

$$A = \frac{1}{n\sigma^3} \sum_{i=1}^n (X_i - \bar{X})^3 \quad (3.5)$$

Экцесс - это степень остро - или плосковершинности распределения значений случайной величины относительно нормального закона распределения:

$$E = \frac{1}{n\sigma^4} \sum_{i=1}^n (X_i - \bar{X})^4 - 3 \quad (3.6)$$

Асимметрия и эксцесс - это безразмерные величины. Они отображают особенности группировки значений случайной величины вокруг среднего значения.

Подводя итог вышесказанному, можно отметить, что медиана, мода и среднее значение являются характеристики положения; дисперсия, среднеквадратичное отклонение, коэффициент вариации, размах являются мерами рассеяния. Асимметрия и эксцесс отображают особенности группировки значений.

Рассмотренные статистические характеристики относят к множеству значений x_1, x_2, \dots, x_n . Когда множество являет собой выборку из генеральной совокупности, то тогда возникает задача оценки ее статистических характеристик по выборочным данным. Более значимыми являются оценка математического ожидания и дисперсии генеральной совокупности.

Математическим ожиданием случайной величины $M(x)$ называется ее среднее значение в генеральной совокупности. Данное значение, за редким исключением, бывает неизвестно, и приходится использовать его приближенную оценку (точечную оценку) - выборочное среднее значение x . При увеличении числа наблюдений выборочное среднее будет стремиться к пределу - к математическому ожиданию.

Дисперсией генеральной совокупности $D(x)$ называется число, которое равно среднему квадрату отклонений случайной величины от ее математического ожидания. В случае, когда математическое ожидание известно, то дисперсию определяют по указанной ниже формуле:

$$D(x) = \frac{1}{n} \sum_{i=1}^n [X_i - M(x)]^2 \quad (3.7)$$

В случае, когда математическое ожидание неизвестно, то тогда находят оценку дисперсии по следующей формуле:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3.8)$$

Единица в знаменателе формулы (3.8) отображает 1-ую использованную *степень свободы*: вместо математического ожидания в формулу подставлено выборочное среднее значение. При увеличении числа наблюдений n оценка дисперсии S^2 будет стремиться к дисперсии генеральной совокупности $D(x)$.

Формулы (3.2) и (3.8) схожи между собой, но используют их в различных случаях. Первая применяется при характеристике выборки, а вторая при характеристике генеральной совокупности.

Группировка исходных данных. Построение гистограммы.

При большом числе исходных данных ($n > 50$) расчет статистических характеристик при помощи таблиц представляется довольно громоздким. Вследствие этого, используется компактный метод расчета с предшествующей группировкой данных. С целью этого весь диапазон исходных значений от x_{\min} до x_{\max} делится на равные интервалы (классы), границы коих наиболее удобно брать округленными, хотя данный факт не является принципиальным. С округленными границами становится в большей степени комфортнее работать.

Число классов находится в зависимости от числа исходных данных. Зачастую принимается от 6 до 20 классов, но, однако, имеется возможность применения и большего их количества. С целью определения числа классов рекомендована эмпирическая следующая формула: $N_{кл}=16[0,41n(n)-1]$. Затем производят подсчет числа исходных значений, которые попали в каждый класс, и результаты сводят в таблицу 3.1.

Определенная сложность возникает в том случае, когда отдельные значения попадают на границу классов. В таком случае имеется возможность отнести их в старший класс, или же попытаться распределить поровну между смежными классами.

Таблица 3.1 - Частота и частость содержания железа в руде

Класс содержаний	Число проб (частота)	Частость	
		в долях	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

продолжение таблицы 3.1

Класс содержаний	Число проб (частота)	Частость	
		в долях	в процентах
52-54	3	0,020	2,0
54-56	2	0,014	1,4
Сумма	147	1,000	100,0

Число значений в классе именуется частотой. Если выразить частоту в относительных долях к общему числу значений, то получаем частость. Ее можно выразить в процентах (таблица 3.1).

Данные таблицы 3.1 позволяют построить гистограмму значений случайной величины (рисунок 3.1). 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 .

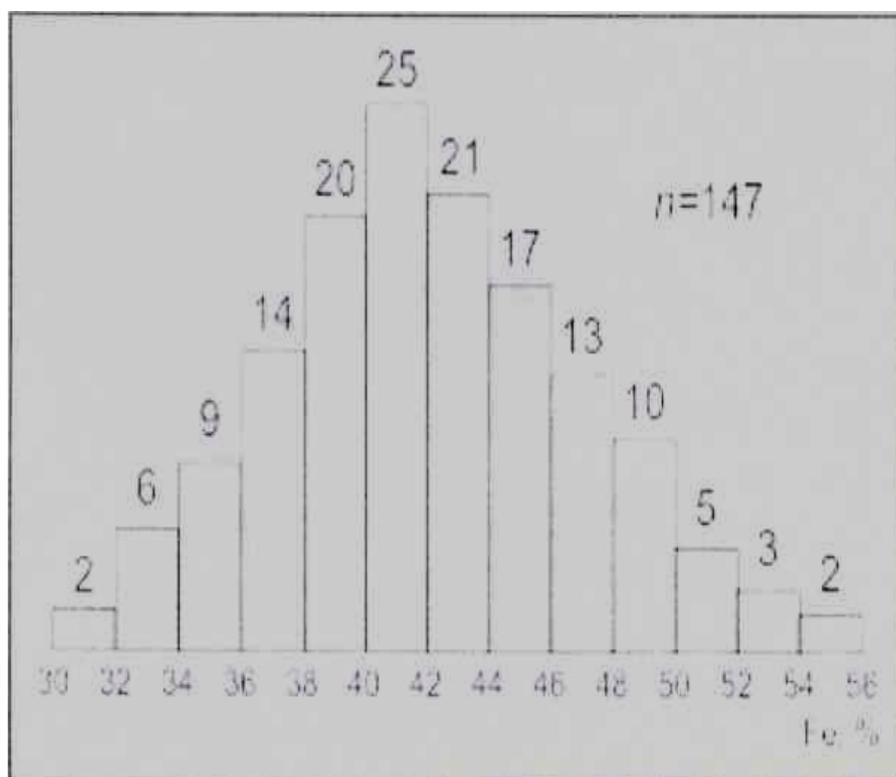


Рисунок 3.1 - Гистограмма частот содержания железа в руде

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. И 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.

3.2 Законы распределения случайных величин

Понятие о законах распределения. При увеличении числа наблюдений частоты будут стремиться к пределу, что характеризует вероятность появления случайной величины, а гистограмма частот будет стремиться к кривой, которая отражает *закон распределения вероятностей*. Вид кривой характеризуется сущностью изучаемого свойства. В редких случаях на вид кривой влияет методика измерений, к примеру, выбор размера проб. Число видов кривых и, соответственно, законов распределения бесконечно велико, но некоторые из них имеют теоретическое обоснование и встречаются чаще других. По крайней мере, реальные распределения приближаются к этим законам.

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

$$F(x) = \int_{-\infty}^x f(x)dx \quad (3.9)$$

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 частот при увеличении числа наблюдений.

Интеграл вероятности $F(x)$; при увеличении значения x монотонно растет от нуля до единицы (рисунок 3.2). Интеграл вероятности (рисунок 3.2 а) можно рассматривать как площадь (заштрихована на рисунке 3.2 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 .

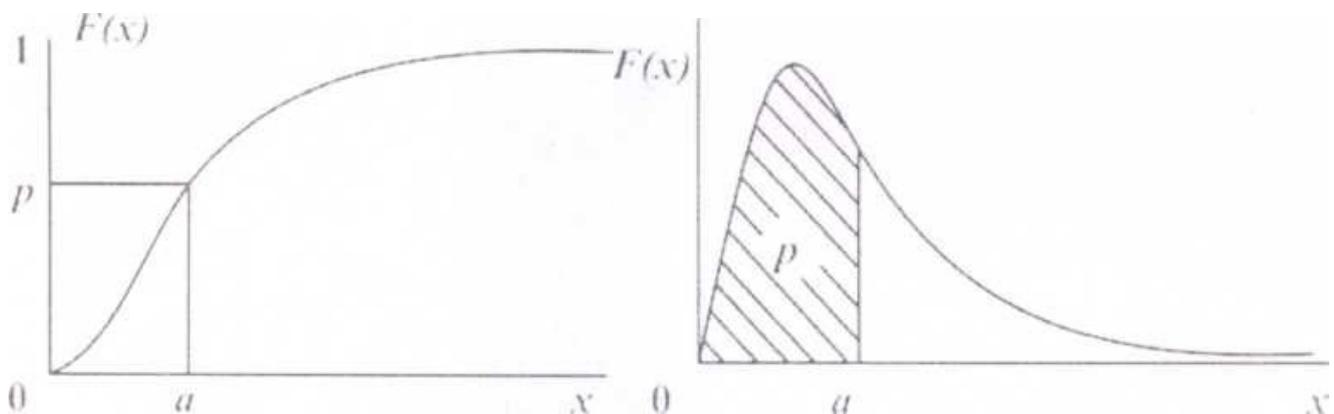


Рисунок 3.2 - Графики интеграла вероятности (а) и плотности вероятности (б)

Законы распределения случайных величин бывают дискретные и непрерывные. У дискретных законов график плотности вероятности имеет ступенчатый вид, как у гистограммы на рисунке 3.2, и случайная величина может принимать лишь прерывистые значения (например, число зерен минералов в пробе). К таким законам относятся биномиальный, Пуассона, гипергеометрический. Законы с непрерывным распределением имеют плавный график плотно-

сти вероятности, и случайная величина может принимать любые значения в области своего существования (например, содержание компонента в руде). Сюда относятся нормальный закон распределения случайных величин, лог-нормальный, Стьюдента, Фишера и др.

Рассмотрим более подробно законы распределения случайных величин, которые наиболее часто используются в геологии.

Нормальный закон распределения случайных величин.

Нормальный закон распределения случайных величин по отношению к другим законам распределения случайных величин, является самым применяемым, из-за что он носит предельный характер и при определенных условиях к нему приближаются многие другие законы.

Нормальный закон описывается интегралом вероятности:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx \quad (3.10)$$

Плотность вероятности имеет следующий вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \quad (3.11)$$

Кривая, выражаемая формулой (3.11), имеет симметричную форму относительно абсциссы \bar{x} (рисунок 3.3). 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.

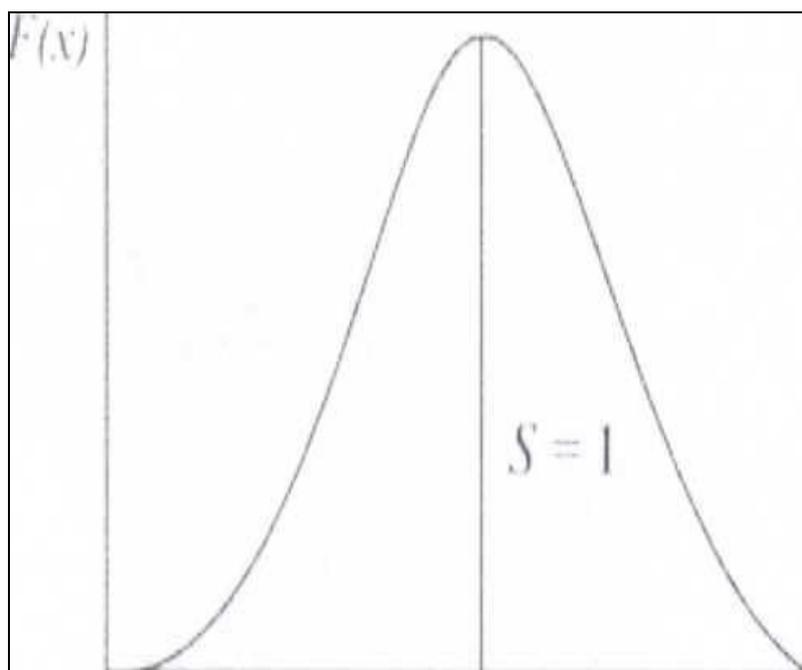


Рисунок 3.3 - График плотности вероятности нормального закона

Как следует из формул (3.10) и (3.11), 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.

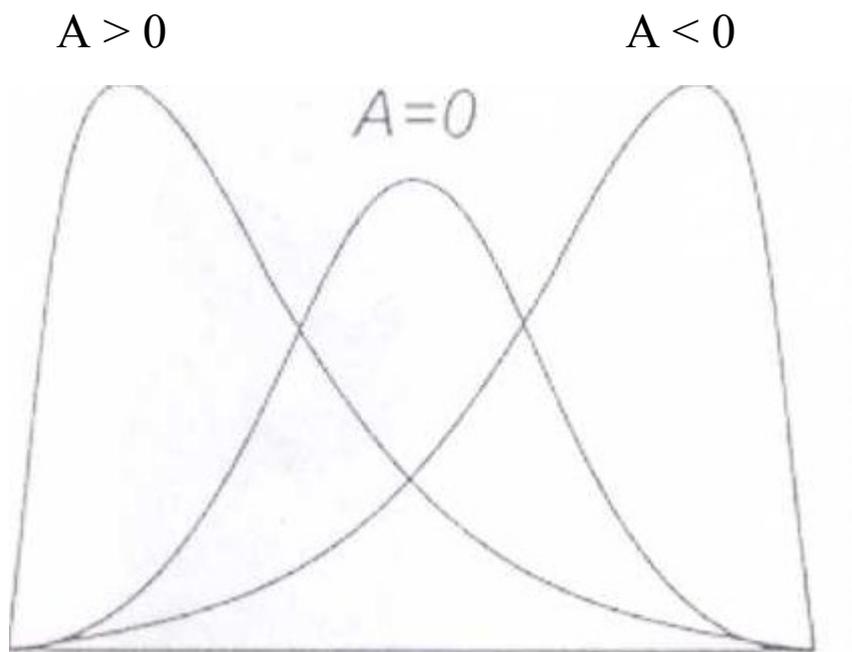


Рисунок 3.4 - Графики плотности вероятности с различной асимметрией

Иногда распределения бывают асимметричными (рисунок 3.4). Отклонение эксцесса от нуля в ту или иную сторону связано с остро- или плосковершинностью кривой распределения по отношению к нормальному распределению (рисунок 3.5). 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.

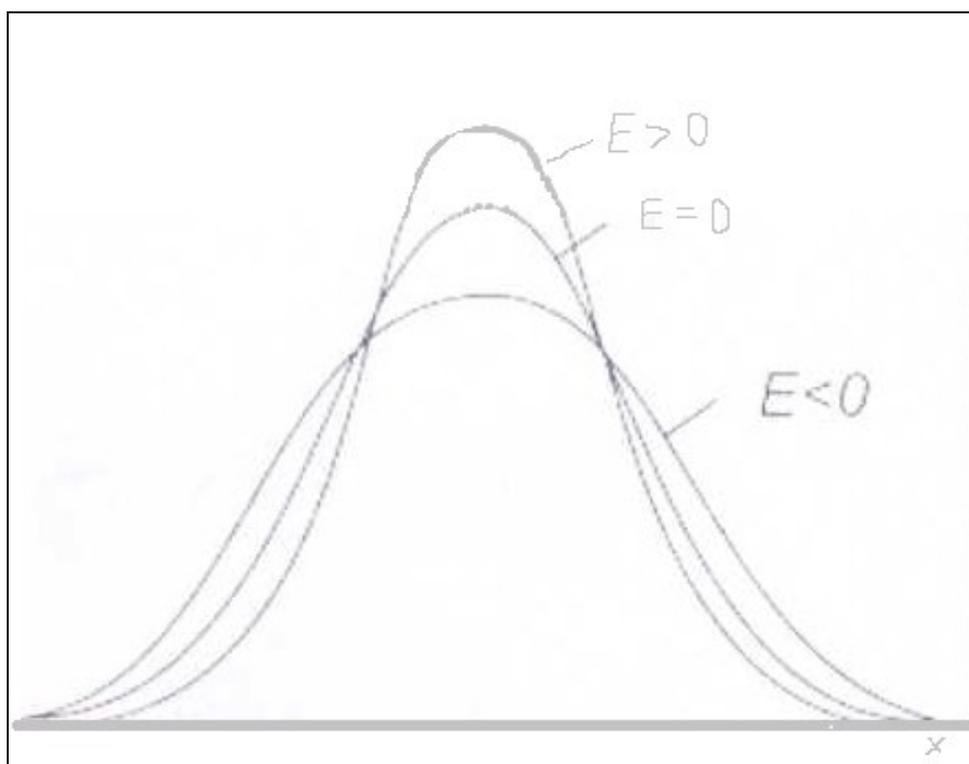


Рисунок 3.5 - Графики плотности вероятности с различными эксцессами

Вероятность p того, что значение случайной величины не превысит заданное значение a (заштрихованная площадь на рисунке 3.2) определяется интегралом, т.е. $p = F(a)$. Наоборот, вероятность a того, что значение случайной величины больше заданного значения a (незаштрихованная площадь на рисунке 3.2), равна $1 - p$. Часто приходится оценивать вероятность q попадания случайной величины в заданный интервал от a до b , которому соответствует заштрихованная площадь на рисунке 3.6.

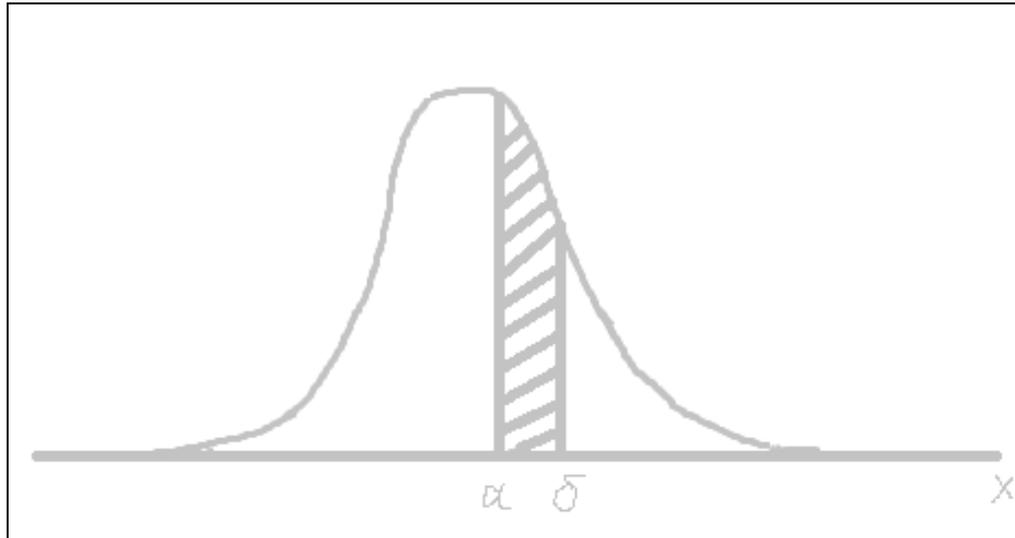


Рисунок 3.6 - График плотности вероятности. Заштрихованная площадь соответствует вероятности q попадания в интервал от a до b .

Вычисление вероятностей сводится к нахождению определенных интегралов (формула 3.10) или (формула 3.11). Интеграл вероятности не интегрируется в алгебраических выражениях, поэтому для нахождения вероятности принято пользоваться специальными таблицами. Поскольку среднее значение x и среднеквадратичное отклонение 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 ($t = 0$) и единичное среднеквадратичное отклонение ($\sigma = 1$). Чтобы пользоваться такими таблицами, нужно предварительно нормировать исходные значения случайной величины x по формуле:

$$t = \frac{x - \bar{x}}{\sigma} \quad (3.12)$$

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

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-t^2/2} dt \quad ; \quad f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad (3.13)$$

Таблицы значений $F(t)$ и $f(t)$ приведены во всех справочниках и пособиях по теории вероятностей. Значения $f(t)$ можно вычислять непосредственно по формулам (3.11) или (3.13).

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

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^t e^{-t^2/2} dt \quad (3.14)$$

Она характеризует вероятность q попадания случайной величины в симметричный интервал от $-t$ до $+t$ (рисунок 3.7) и связана с интегралом вероятности соотношением $\Phi(t) = 2F(t) - 1$.

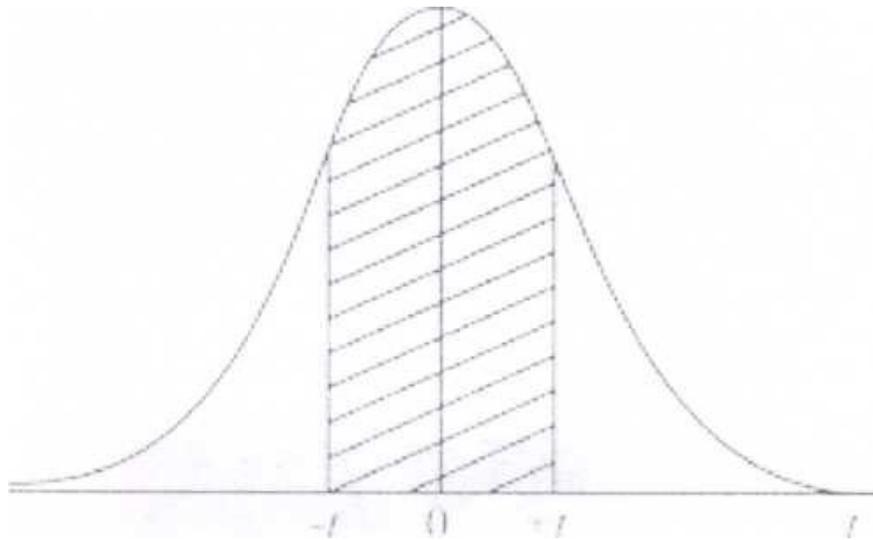


Рисунок 3.7 - График плотности вероятности. Заштрихованная площадь соответствует вероятности попадания значений случайной величины от $-t$ до $+t$

Вероятность попадания случайной величины в интервал от нуля до $+t$ называется *функцией Лапласа*. Из-за симметричности интеграл (формула 3.14) можно представить как удвоенную функцию Лапласа:

$$\Phi(t) = \sqrt{\frac{2}{\pi}} \int_0^t e^{-t^2/2} dt \quad (3.15)$$

Значения функций $F(t)$, $f(t)$ и $\Phi(t)$ в пределах от $t = 0$ до $t = 3,1$ с шагом аргумента приведены в таблице 3.2. Поскольку функции $f(t)$ и $\Phi(t)$ симметричны относительно $t = 0$, их значения при отрицательном значении t находят из таблицы 3.2 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 таблицы 3.2 имеем $f(t) = 0,0940$; $\Phi(t) = 0,9109$; $F(t) = 1 - 0,9554 = 0,0446$.

Таблица 3.2 - Функции нормального распределения

t	F(t)	f(t)	Φ(t)
0,0	0,5000	0,3989	0,0000
0,1	0,5398	0,3970	0,0797
0,2	0,5793	0,3910	0,1585
0,3	0,6179	0,3814	0,2358
0,4	0,6554	0,3683	0,3108
0,5	0,6915	0,3521	0,3829
0,6	0,7257	0,3332	0,4515
0,7	0,7580	0,3123	0,5161
0,8	0,7881	0,2897	0,5763
0,9	0,8159	0,2661	0,6319
1,0	0,8413	0,2420	0,6827
1,1	0,8643	0,2179	0,7287

Продолжение таблицы 3.2

t	F(t)	f(t)	$\Phi(t)$
1,2	0,8849	0,1942	0,7699
1,3	0,9032	0,1714	0,8064
1,4	0,9192	0,1497	0,8385
1,5	0,9332	0,1296	0,8664
1,6	0,9452	0,1109	0,8904
1,7	0,9554	0,0940	0,9109
1,8	0,9641	0,0790	0,9281
1,9	0,9713	0,0656	0,9426
2,0	0,9772	0,0540	0,9545
2,1	0,9821	0,0440	0,9643
2,2	0,9861	0,0355	0,9722
2,3	0,9893	0,0283	0,9786
2,4	0,9918	0,0224	0,9836
2,5	0,9938	0,0175	0,9876
2,6	0,9953	0,0136	0,9907
2,7	0,9965	0,0104	0,9931
2,8	0,9974	0,0079	0,9949
2,9	0,9981	0,0060	0,9963
3,0	0,9987	0,0044	0,9973
3,1	0,9990	0,0033	0,9981

Рассмотрим, как определяется вероятность с помощью таблицы 3.2. Пусть имеется интервал от $a = 2,72$ до $b = 2,96$; известны также характеристики $\bar{x} = 2,2$ и $\sigma = 0,40$. По формуле (3.12) вычислим нормированные значения $t_1 = (2.72 - 2.2)/0.40 = 1.31$; $t_2 = (2,96 - 2,2)/0,40 = 1,90$. В таблице 3.2 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 (3.13) и (3.14), играющие большую роль, можно вычислять без применения таблиц путем численного интегрирования на компьютере или с помощью пакета "Statistica".

Обратная задача - находить интервал возможных значений случайной величины t при заданных вероятностях p , a , q , b . Если задана вероятность $p = F(t)$, то соответствующее ей значение t называется *квантилью* распределения. Она является функцией, обратной интегралу вероятности ($F^{-1}(p)$), и обозначается $t = F^{-1}(p)$. Квантиль можно найти интерполяцией (от лат. "interpolatio" - изменение, переделка) данных таблицы 3.2. 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 *квартлями*. 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 таблицы 3.2, но лучше иметь специальную таблицу 3.3 зависимости t от $\Phi(t)$.

Например, задана вероятность $q = 0,96$, тогда соответствующий ей коэффициент вероятности $t = 2,054$. Таблица 3.3 может быть использована и для нахождения квантилей. По заданной вероятности p вычисляется вероятность $q = 2p-1$ и по таблице 3.3 определяется квантиль. Например, дана

вероятность $p = 0,9$. Вычисляем $q = 2 \times 0,9 - 1 = 0,8$, соответствующая ей квантиль $t = 1.282$.

Таблица 3.3 - 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

Логарифмически-нормальный закон распределения случайных величин.

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 ($2,3026 = \ln 10$).

Случайная величина в логнормальном законе, в отличие от нормального, имеет область существования от нуля до $+\infty$. Если присутствуют нулевые значения (или следы), что нередко бывает при спектральном и химическом анализах, то это вызывает трудности, так как логарифм нуля равен $-\infty$. Обычно нулевые содержания заменяют какими-то минимальными значениями, например пределом чувствительности 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 (3.16)

$$f(z) = \frac{1}{\sigma_z \sqrt{2\pi}} e^{-\frac{(z-\bar{z})^2}{2\sigma_z^2}} \quad (3.16)$$

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

Плотность вероятности исходных значений x выражается формулой логнормального закона:

$$f(x) = \frac{1}{x\sigma_x\sqrt{2\pi}} e^{-\frac{(\ln x - \overline{\ln x})^2}{2\sigma_z^2}} \quad (3.17)$$

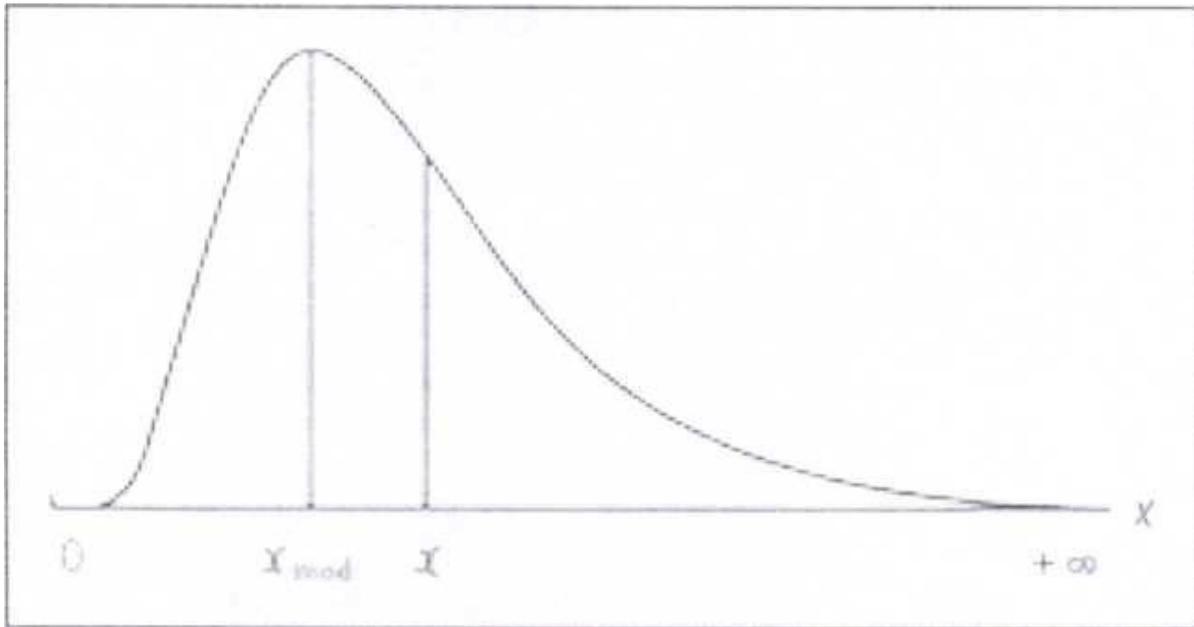


Рисунок 3.8 - График плотности вероятности логнормального закона

График функции $f(x)$ асимметричен (рисунок 3.8), среднее значение, мода и медиана не совпадают между собой. Они связаны с величинами \bar{z} и σ_z^2 следующими соотношениями:

$$\bar{x} = e^{\bar{z} + \sigma_z^2/2}; \quad X_{mod} = e^{\bar{z} - \sigma_z^2}; \quad X_{med} = e^{\bar{z}} \quad (3.18)$$

Дисперсия исходных данных также определяется соотношением:

$$\sigma^2 = e^{2\bar{x} + \sigma_x^2} (e^{\sigma_x^2} - 1) \quad (3.19)$$

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

Распределение Стьюдента

Распределение Стьюдента, которое также именуется t-распределением, имеет огромное значение: с его помощью проверяют гипотезы о равенстве средних значений случайных величин. Функция распределения Стьюдента можно выразить интегралом:

$$F_k(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma[\frac{k+1}{2}]}{\Gamma[\frac{k}{2}]} \int_{-\infty}^t [1 + \frac{x^2}{k}]^{\frac{k+1}{2}} dx \quad (3.20)$$

Плотность вероятности имеет вид:

$$f_k(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma[\frac{k+1}{2}]}{\Gamma[\frac{k}{2}]} \int_{-\infty}^t [1 + \frac{t^2}{k}]^{\frac{k+1}{2}} \quad (3.21)$$

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

$$\Gamma(x) = \int_0^{\infty} y^{x-1} e^{-y} dy \quad (3.22)$$

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 n₁ и n₂, то k = n₁ + n₂ - 2. Могут быть и другие варианты гипотез.

Из функции (3.21) следует, что случайная величина t может принимать любые значения в пределах от - ∞ до + ∞.

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 (3.20) обычно приводится в табличном виде. В таблице 3.4 приведены лишь некоторые значения, по которым видно, как с увеличением числа k функция распределения Стьюдента приближается к нормальному закону (последняя графа таблицы 3.4).

Таблица 3.4 - 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				

Плотность вероятности (3.21) имеет симметричный график, похожий на кривую нормального закона, но более вытянутый по горизонтальной оси (рисунок 3.9). При увеличении значения k график приближается к кривой нормального закона.

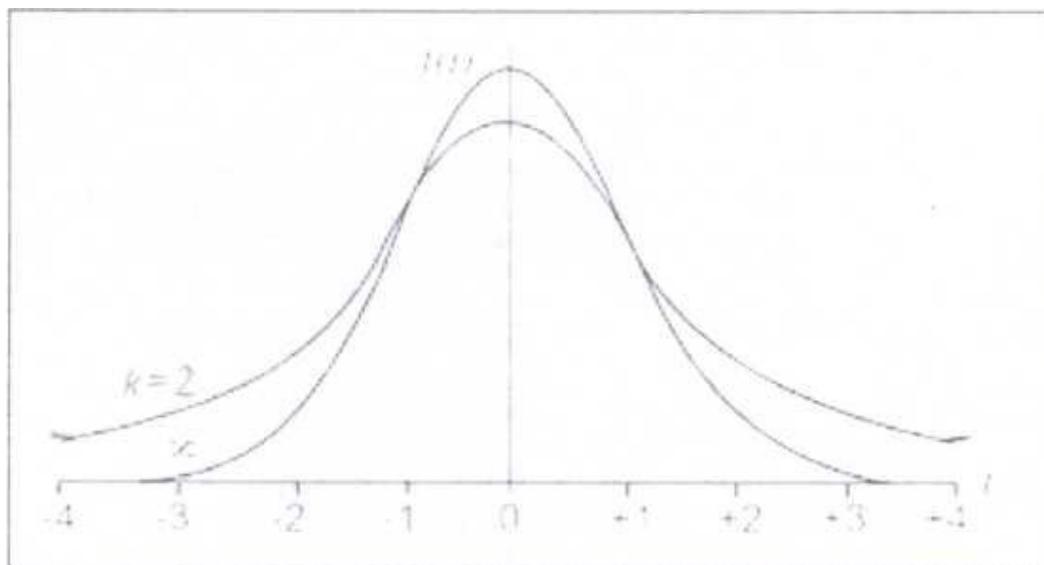


Рисунок 3.9 - График плотности распределения Стьюдента. При $k = \infty$ распределение совпадает с нормальным

Значения плотности вероятности обычно находят по таблицам, но их не сложно вычислить, преобразовав формулу (3.21) к следующему виду:

$$f_k(t) = \frac{A_k}{\sqrt{\pi k}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}; \quad A_k = \Gamma\left(\frac{k+1}{2}\right) \Gamma\left(\frac{k}{2}\right) \quad (3.23)$$

Для нахождения A_k достаточно знать одно значение, например $A_1 = 1/\sqrt{\pi}$, и далее пользоваться рекуррентной формулой (формула, выражающая каждый член последовательности через предыдущих членов) $A_{k+1} = k/2 A_k$. Так, $A_2 = \sqrt{\pi}/2$, $A_3 = 2/\sqrt{\pi}$ и т.д. Некоторые значения плотности вероятности приведены в таблице 3.5. Последняя графа соответствует нормальному распределению.

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 $+t$.

Таблица 3.5 - 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	0.397	0.3979	0.3989
0,5	0.254	0.2962	0.3279	0.3397	0.3458	0.3495	0.3508	0.3521
1,0	0.159	0.1925	0.219	0.2304	0.2360	0.239	0.2408	0.2420
1,5	0.097	0.1141	0.1245	0.1274	0.1286	0.129	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	0.013	0.0051	0.0020	0.0008	0.0003	0.0002	0.0001
4,5	0.0150	0.009	0.0029	0.0009	0.0003	0.000	0.0000	0.0000
5,0	0.012	0.007	0.0018	0.0004	0.0001	0.000	0.0000	0.0000

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 (таблица 3.6), 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 таблицы 3.6 имеем $t = 2,131$.

Если число степеней свободы велико (несколько десятков и более), то можно пользоваться таблицами нормального закона (таблица 3.3). Так, при $\beta = 0,01$ имеем $q = 1 - \beta = 1 - 0,01 = 0,99$ и по таблица 3.3 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 распределения случайных величин.

Таблица 3.6 - 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

продолжение таблицы 3.6

k	Вероятность β						
	0,10	0,05	0,02	0,01	0,005	0,002	0,001
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	2,086	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
P	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Распределение Фишера

Распределение Фишера, которое называется также F-распределением, применяется для проверки гипотезы о равенстве дисперсий случайных величин. Критерием является отношение несмещенных оценок дисперсий $F = S_1^2/S_2^2$, причем в числитель отношения практически всегда помещают большую дисперсию, т.е. $S_1^2 > S_2^2$. Плотность вероятности распределения величины F можно выразить следующей формулой:

$$f(t) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} k_1^{\frac{k_1}{2}} k_2^{\frac{k_2}{2}} (k_2 + k_1 t)^{-\frac{k_1+k_2}{2}} t^{(k_1/2-1)} \quad (3.24)$$

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 $k_2 = n_2 - 1$.

График плотности вероятности асимметричен (рисунок 3.10) и имеет максимум (моду):

$$t_{\text{mod}} = \frac{k_2(k_1 - 2)}{k_1(k_2 + 2)} \quad (3.25)$$

Практическое значение имеет зависимость коэффициента t (критерия) от вероятности α (ей соответствует заштрихованная площадь на рисунке 3.10 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 $p = 1 - \alpha$.

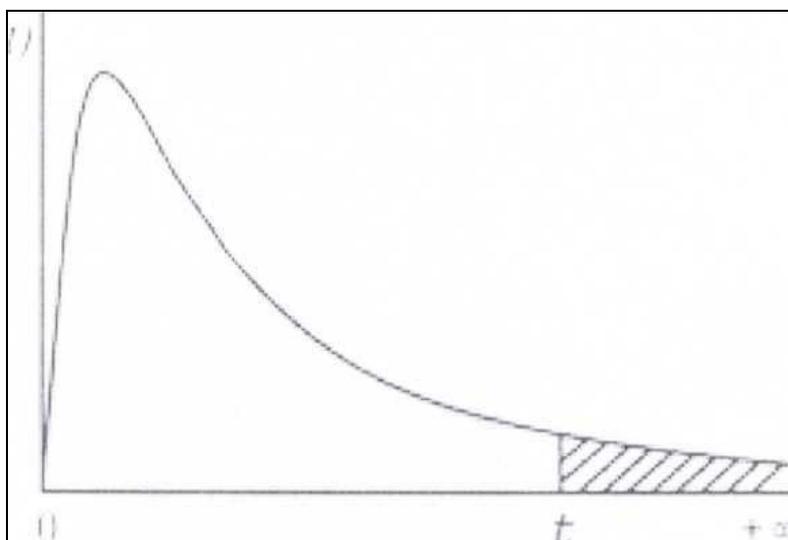


Рисунок 3.10 - График плотности вероятности F-распределения

Таблица 3.7 - 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	1,64	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

4 Двумерный статистический анализ и его применение

Во многочисленных задачах геологии исследуют 2 связанных между собой свойства многообразия геологических объектов. Данный анализ реализуется на базе двумерной статистической модели.

4.1 Характеристика двумерного статистического анализа

Например, пусть существует система из n однородных геологических объектов, у любой из них измерены характеристики 2-х свойств. Результаты измерений одного свойства назовем x_1, x_2, \dots, x_n , второго свойства y_1, y_2, \dots, y_n .

Двухмерная модель базируется на аналогичных гипотезах, которые были перечислены при рассмотрении одномерной статистической модели, а именно:

1. Значения $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n$ имеют случайный характер.
2. Значения первого свойства x_1, x_2, \dots, x_n не находятся в зависимости друг от друга, значения второго свойства y_1, y_2, \dots, y_n аналогично не зависят друг от друга (но, однако, имеется возможность наличия зависимости между свойствами x и y).
3. Совокупность измеренных свойств представляется однородной.

Система значений $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n$ именуется системой 2-х случайных величин, двумерной случайной величиной или случайным вектором.

Результаты измерений двумерной случайной величины обычно изображают на графике, где по оси абсцисс откладывают характеристику одного свойства, а по оси ординат - другого. Любой геологический на вышеупомянутом графике изображают точкой, а большое количество объектов - облаком точек. Расположение точек на графике дает возможность делать предварительные выводы о характере зависимости между свойствами.

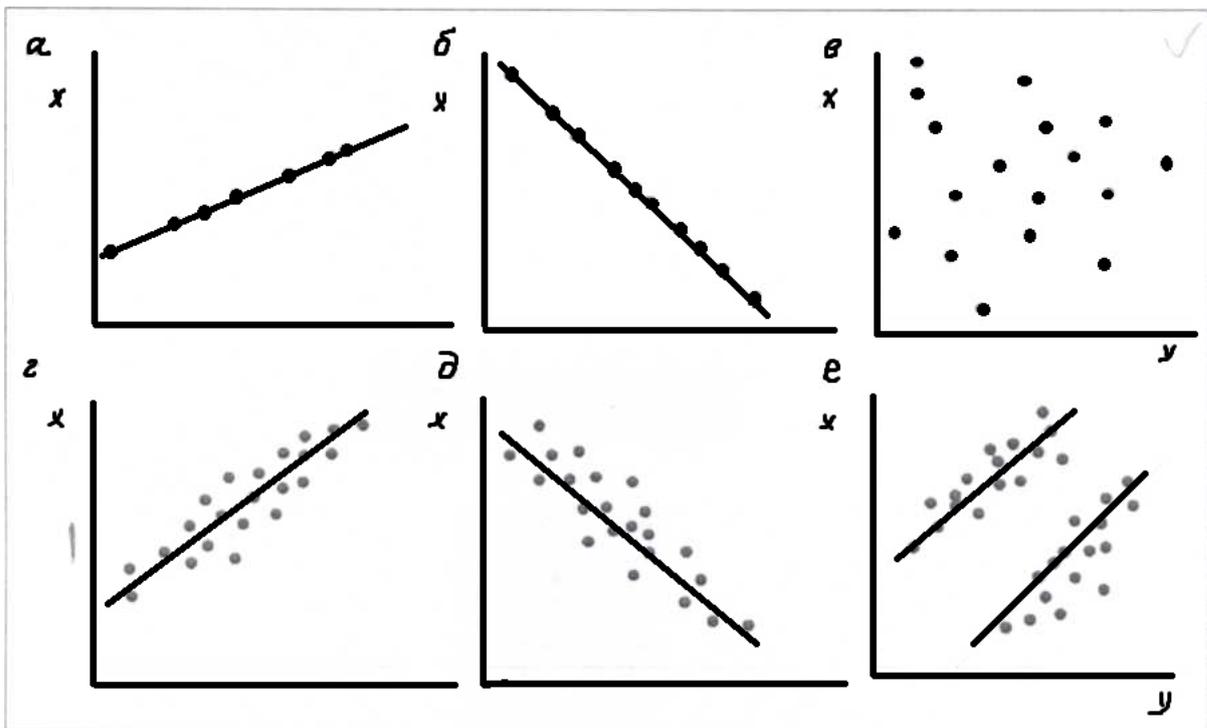


Рисунок 4.1 - Виды зависимостей между характеристиками свойств x и y

В том случае, когда точки располагаются вдоль линии (рисунок 4.1 а, б), то тогда между характеристиками свойств существует функциональная зависимость. Такая зависимость может быть линейной и нелинейной. Когда точки располагаются хаотично (рисунок 4.1 в), тогда такой зависимости между характеристиками свойств не наблюдается. В большинстве случаев точки располагаются в виде облака, которое группируется вдоль какой-то линии (рисунок 4.1 г, д), в данном случае отмечается нестрогая статистическая зависимость между свойствами. Эта зависимость аналогично может быть линейной и нелинейной, Функциональные и статистические зависимости имеют возможность быть положительными, когда с возрастанием характеристики одного свойства увеличивается и другая (рисунок 4.1 а, г), но также могут быть и отрицательными, когда характеристика одного свойства растет, а другого убывает (рисунок 4.1 б, д). Гораздо реже точки могут образовывать 2 и более изолированных или частично перекрывающихся облака (рисунок 4.1 в), что говорит о 2-х и более однородных совокупностях, что рекомендуется исследовать отдельно.

Система двух случайных величин содержит 5 базовых статистических характеристик: средние значения \bar{x} и \bar{y} , дисперсии δ_x^2 и δ_y^2 и корреляционный момент (или ковариацию) K_{xy} , которые можно вычислить по следующим формулам:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad ; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (4.1)$$

$$\sigma_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \quad ; \quad \sigma_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4.2)$$

$$K_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (4.3)$$

Первые 4 формулы встречались гораздо раньше. Особый интерес представляет 5-я формула, которая отображает взаимосвязь между случайными величинами x и y . Поскольку корреляционный момент имеет размерность, его преобразуют в безразмерную величину по нижеприведенной формуле:

$$r = \frac{K_{xy}}{\sigma_x \sigma_y} \quad (4.4)$$

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 (рисунок 4.1 в). При $r = 1$ связь функциональная положительная (рисунок 4.1). При $r = -1$ связь функциональная отрицательная (рисунок 4.1 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 свойствами; 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_B заметно влияет случайная погрешность измерений исходных данных, уменьшая истинное значение коэффициента корреляции r :

$$r_B = r \div \sqrt{\left(1 + \frac{\sigma_1^2}{\sigma_x^2}\right)\left(1 + \frac{\sigma_2^2}{\sigma_y^2}\right)} \quad (4.5)$$

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

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 $k = n - 2$:

$$t = \frac{|r|}{S_r} \text{ при } S_r = \sqrt{\frac{1 - r^2}{n - 2}} \quad (4.6)$$

где S_r - оценка среднеквадратичного отклонения коэффициента корреляции.

Если критерий t будет больше допустимого $t_{доп}$ при заданной вероятности β , 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:

$$t = \frac{|r|}{\sigma_r} \text{ при } \sigma_r = \sqrt{\frac{1 - r^2}{n}} \quad (4.7)$$

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

Еще один критерий предложен Фишером:

$$t = \frac{|z|}{\sigma_z} \text{ при } \sigma_r = \sqrt{\frac{1-r^2}{n}} \quad (4.8)$$

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

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} = \operatorname{arcth}(r) \quad (4.9)$$

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

Из соотношения (4.1.6) выводится формула значимого коэффициента корреляции:

$$r_{\text{зн}} = \frac{t_{\text{доп}}}{\sqrt{t_{\text{доп}}^2 + n + 2}} \quad (4.10)$$

Так как $t_{\text{доп}}$ 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 (4.7), формула (4.10) упрощается:

$$r_{\text{зн}} = \frac{t_{\text{доп}}}{\sqrt{t_{\text{доп}}^2 + n}} \quad (4.11)$$

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

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 :

$$y = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad (4.12)$$

$$x = \bar{x} + r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \quad (4.13)$$

Эти уравнения пересекаются в точке средних значений \bar{x} и \bar{y} . Уравнение регрессии (4.12) позволяет определить еще одну остаточную дисперсию σ_δ , которая характеризует рассеяние значений случайной величины около линии регрессии:

$$\sigma_\delta^2 = \frac{1}{n} \sum_{i=1}^n \delta_i^2 \quad (4.14)$$

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

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

$$\sigma_\delta^2 = \sigma_y^2 (1 - r^2) \quad (4.15)$$

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:

$$\sigma_y^2 = \sigma_{\text{тр}}^2 + \sigma_{\delta}^2 \quad (4.16)$$

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 (4.12) будет сопровождаться погрешностью $t\sigma_{\delta}$ или, учитывая формулу (4.15), погрешностью $t\sigma_y\sqrt{1-r^2}$, где t - коэффициент вероятности. Чем больше коэффициент корреляции по абсолютной величине, чем меньше погрешность прогнозирования. Для надежного прогнозирования необходимо использовать лишь такие зависимости, у которых коэффициент корреляции больше 0,87.

4.2 Двумерное нормальное распределение. Эллипс рассеяния

Облако точек на рисунке 4.1, как и во многих других случаях, в первом приближении имеет эллипсовидную форму. В ряде задач нужно знать параметры эллипса, охватывающего облако, и построить эллипс по 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:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \left[\frac{(x-\bar{x})^2}{\sigma_x^2} - \frac{2r(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} \right] \quad (4.17)$$

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 (рисунок 4.2):

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}} \quad ; \quad f(y) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\bar{y})^2}{2\sigma_y^2}} \quad (4.18)$$

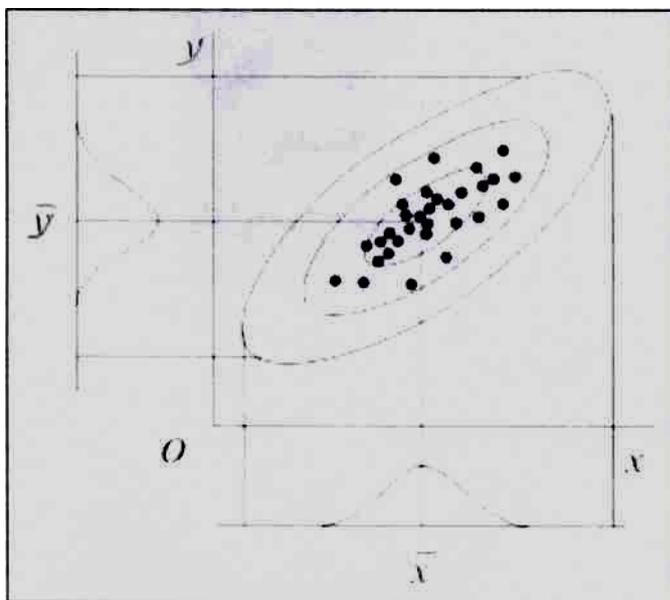


Рисунок 4.2 - Схема эллипса рассеяния двухмерного нормального распределения и проекции вероятности на оси Ox и Oy .

Облако точек заключено внутри эллипса, выраженного уравнением:

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - \frac{2r(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = t^2 \quad (4.19)$$

где t - коэффициент вероятности.

Если t будет принимать другие значения, будут построены подобные эллипсы иного размера (рисунок 4.3).

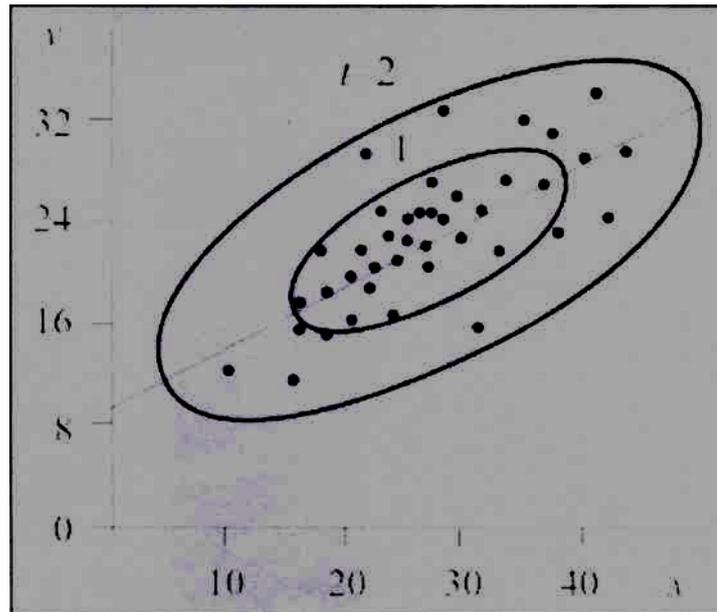


Рисунок 4.3 - Подобные эллипсы рассеяния

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

$$p = 1 - e^{-\frac{t^2}{\sqrt{1-r^2}}} \quad (4.20)$$

Для того, чтобы построить эллипс нужно обладать знаниями о положении его центра, о размерах осей (полуосей) и об их ориентировке по отношению к осям координат.

Центр эллипса имеет координаты x и y . Эллипс можно охарактеризовать следующими показателями: размер, форма и ориентировка осей на плоскости.

Размер эллипса возрастает при увеличении рассеяния точек, т.е. при возрастании дисперсий σ_x и σ_y . Форма эллипса находится в зависимости в основном от коэффициента корреляции r . Чем ближе он будет по модулю к единице, тем гораздо более узким и вытянутым представляется 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 представляется возможным найти из уравнения:

$$\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \quad (4.21)$$

Его решение дает 2 угла α_1 и α_2 , которые отличаются друг от друга на 90° . Для того, чтобы найти полуоси эллипса, начало координат переносят в центр эллипса, в точку (\bar{x}, \bar{y}) и поворачивают координатные оси на угол α_1 и α_2 . Далее обозначают новые оси координат u и v , тогда такое уравнение эллипса (4.19) будет иметь канонический вид:

$$\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2} = t^2 \quad (4.22)$$

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 взаимосвязаны с дисперсиями σ_x^2 и σ_y^2 следующими соотношениями:

$$\sigma_u^2 = \sigma_x^2 \cos^2 \alpha + r\sigma_x\sigma_y \sin 2\alpha + \sigma_y^2 \sin^2 \alpha \quad (4.23)$$

$$\sigma_v^2 = \sigma_x^2 \sin^2 \alpha - r\sigma_x\sigma_y \sin 2\alpha + \sigma_y^2 \cos^2 \alpha \quad (4.24)$$

Сумма дисперсий при переносе и повороте координат не изменяется. Она находится в зависимости от взаимного расположения точек в облаке и является инвариантом:

$$\sigma_v^2 + \sigma_u^2 = \sigma_x^2 + \sigma_y^2 = \text{const} \quad (4.25)$$

Исходя из вышесказанного можно сделать вывод о том, что при необходимости построения эллипса рассеяния, будет достаточно знать координаты его центра (\bar{x}, \bar{y}) , угол поворота осей α_1 и α_2 и длину полуосей $t\sigma_u$ $t\sigma_v$.

4.3 Нелинейная регрессия. Метод квадратов

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

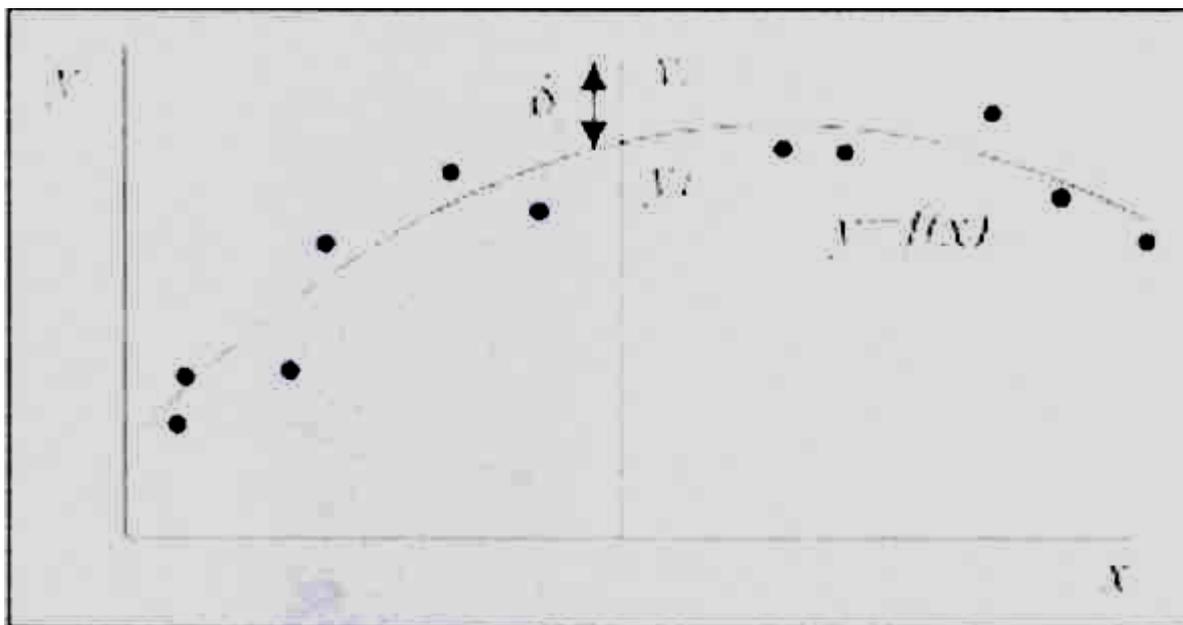


Рисунок 4.4 - Схема, иллюстрирующая отклонения точек от заданной кривой $y = f(x)$

Пусть значения величины y нелинейно зависят от значений величины x (точки на рисунок 4.4). Нужно подобрать такую функцию $f(x)$, в которой

отклонения между фактическими y_i и расчетными (теоретическими) $y_T = f(x)$ значениями будут наименьшими. Отклонения $\delta_i = y_i - y_T$ могут быть положительными и отрицательными. Главный принцип метода заключается в требовании, чтобы сумма квадратов всех отклонений от линии зависимости была минимальной:

$$\sum_{i=1}^n \delta_i^2 \rightarrow \min \quad (4.26)$$

Вид аппроксимирующей функции $f(x)$ должен быть задан либо на основании теоретических соображений (например, гиперболическая зависимость плотности руды от ее состава), либо путем эмпирического подбора. Например, в качестве функции $f(x)$ могут быть использованы полином порядка p : $f(x) = -a_0 + a_1x + a_2x^2 + \dots + a_px^p$; 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 (рисунок 4.4). Следовательно, и сумма квадратов отклонений также зависит от значений коэффициентов, т.е. является их функцией:

$$\sum_{i=1}^n \delta_i^2 = \Phi(a, b, c \dots) \quad (4.27)$$

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

$$\left| \frac{\partial \Phi}{\partial a} = 0; \frac{\partial \Phi}{\partial b} = 0; \frac{\partial \Phi}{\partial c} = 0 \right. \quad (4.28)$$

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 $y_T = f(x)$ для каждого x_i и сравнить их с фактическими y_i , т.е. найти отклонения $\delta_i = y_i - y_T$. Далее вычисляют дисперсии отклонений:

$$\sigma_{\delta}^2 = \frac{1}{n} \sum_{i=1}^n \delta_i^2 \quad (4.29)$$

и, наконец, определяют *корреляционное отношение*

$$\eta = \sqrt{1 - \sigma_{\delta}^2 / \sigma_y^2} \quad (4.30)$$

которое заключено в интервале от нуля до единицы ($0 \leq \eta \leq 1$) 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 $\sigma_{\text{зак}}^2 = \sigma_y^2 - \sigma_{\delta}^2$, которая характеризует изменчивость расчетных значений y_T и может быть названа закономерной. Приняв общую дисперсию за 100 %, можно найти соотношение между σ_{δ}^2 и $\sigma_{\text{зак}}^2$ в процентах.

Рассмотренная схема обработки данных применима к исследованию линейных и нелинейных, однофакторных и многофакторных зависимостей. В частном случае простой линейной зависимости $y = ax + b$ использование метода наименьших квадратов дает уравнение регрессии (4.12), а корреляционное отношение по абсолютной величине совпадает с коэффициентом корреляции.

5 Многомерный статистический анализ и его применение

□ Система множества случайных величин и ее статистические характеристики

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

$$\begin{matrix} X_{11}, X_{12}, \dots, X_{1k} \\ X_{21}, X_{22}, \dots, X_{2k} \\ \dots \\ X_{n1}, X_{n2}, \dots, X_{nk} \end{matrix} \quad (5.1)$$

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 временными 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 (5.1), - облаком точек.

Многомерная статистическая модель имеет различные статистические характеристики, наиболее употребительными из которых являются средние значения случайных величин $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$, их дисперсии $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$ и среднеквадратичные отклонения $\sigma_1, \sigma_2, \dots, \sigma_k$.

Кроме того, часто используются матрицы ковариации и коэффициентов корреляции случайных величин. Например, ковариация K_{ij} - это корреляционный смешанный момент двух случайных величин i и j . Матрица ковариации имеет симметричный вид:

$$\begin{matrix} \sigma^2_1, & K_{12}, & \dots, & K_{1k} \\ K_{21}, & \sigma^2_2, & \dots, & K_{2k} \\ K_{k1}, & K_{k2}, & \dots, & \sigma^2_k \end{matrix} \quad (5.2)$$

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:

$$\begin{matrix} 1, & r_{12}, & \dots, & r_{1k} \\ r_{21}, & 1, & \dots, & r_{2k} \\ r_{k1}, & r_{k2}, & \dots, & 1 \end{matrix} \quad (5.3)$$

В матрице по диагонали находятся единицы, а в остальных полях - собственно коэффициенты корреляции. Методика расчета корреляционных моментов и коэффициентов корреляции такая же, как в двумерной статистической модели. Данные матрицы коэффициентов корреляции могут быть представлены в виде графа связей (рисунок 5.1). Для построения графа использованы результаты силикатного анализа горных пород. Чем больше коэффициент корреляции между компонентами, тем толще соединяющая их линия.

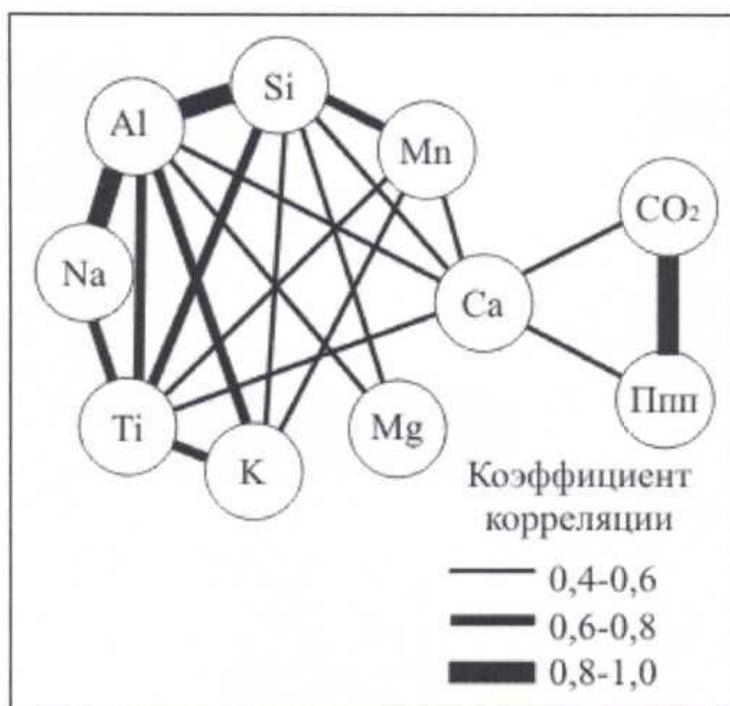


Рисунок 5.1 - Граф связей

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:

$$y = a_1x_1 + a_2x_2 + \dots + a_kx_k + b \quad (5.4)$$

где x_1, x_2, \dots, x_k - свойства; a_1, a_2, \dots, 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, \dots, x_k нормируют по формуле (3.12), т.е. заменяют величинами:

$$t_1 = \frac{x_1 - \bar{x}_1}{\sigma_1}, \quad t_2 = \frac{x_2 - \bar{x}_2}{\sigma_2}, \quad t_k = \frac{x_k - \bar{x}_k}{\sigma_k}, \quad t_y = \frac{y - \bar{y}}{\sigma_y} \quad (5.5)$$

В результате замены уравнение (5.4) приобретет следующий вид:

$$t_y = A_1 t_1 + A_2 t_2 + \dots + A_k t_k \quad (5.6)$$

где величины A_1, A_2, \dots, A_k - нормированные коэффициенты регрессии.

Метод главных компонент

Вышеуказанный метод является одним из самых популярных и эффективных способов по обработке многомерных статистических данных. Его суть заключается в линейном преобразование исходных данных в независимые величины, которые несут смысловую геологическую информацию.

Многомерные случайные величины изображают в многомерном признаковом пространстве облаком точек.

Предполагается, что облако имеет форму, близкую к многомерному эллипсоиду. Преобразование исходных данных обычно 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 этому 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 геологическую 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 (3.12). Единицами нормирования свойств служат среднеквадратичные отклонения.

Метод главных компонент позволяет выделить однородные совокупности и аномальные значения, а также дать геологическую интерпретацию причин изменения свойств объектов по значениям факторных нагрузок.

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. Геометрическая аналогия матрицы - облако точек в многомерном 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 од-

народным группам объектов. Эти группы называются кластерами, что и дало название методу.

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.

6 Основные программы при компьютерном моделировании в геологии

6.1 Программа Microsoft Word

Программа Microsoft Word - это многофункциональная система обработки различных текстов.

Управлять всеми пунктами меню и командами можно при помощи мыши, а также при помощи горячих клавиш, с целью экономии времени пользователя на выполнение определенного набора операций. Данная программа дает возможность просмотреть на экране монитора заверченный документ не затрачивая время и листы бумаги для распечатывания пробных страниц. Кроме этого, возможно форматирование текста, например, печатание курсивом или жирным шрифтом.

Редактор Microsoft Word, аналогично Microsoft Windows были разработаны одной корпорацией. Этот факт дает возможность гарантировать наибольшую согласованность данных систем друг с другом, а так же с другими программами, которые являются совместимыми с Microsoft Windows. Тексты и иллюстрации многочисленных форматов могут быть импортированы в Microsoft Word непосредственно из других программ и встроены в текст документа. Кроме этого, существует ряд положительно зарекомендовавших себя функций, например, распечатывание некоторых страниц, не прерывая непосредственно работу с данным документом. В таком случае программа управления печатью с буферизацией берет на себя контроль за печатью на заднем плане. Письма и факсы, записки и отчеты, публикации и web-страницы - это далеко не весь перечень документов с которыми пользователь может вести работу, используя текстовый процессор Microsoft Word.

В Microsoft Word присутствует широкий выбор средств автоматизации, которые упрощают выполнение типичных задач. Перечислим некоторые из них:

а) перемещение по документу выполняются при помощи мыши, а также при помощи клавиатуры (сочетания клавиш: Ctrl + - на слово вправо, Ctrl + - на слово влево, Ctrl + - к следующему абзацу, Ctrl + - к предыдущему абзацу, Ctrl + PgDown - на экран вниз, Ctrl + PgUp - в начало следующей страницы, Ctrl + Home - в начало документа, Ctrl + End - в конец документа, Alt + Ctrl + PgUp - в начало экрана, Alt + Ctrl + PgDown - в конец экрана);

б) копирование текста осуществляют при помощи команд копировать, вырезать, вставить (меню Правка). Кроме этого, при помощи клавиатуры с использованием клавиш Ctrl + Shift (копирование), Shift + Insert (вставить);

в) сохранение документа возможно при помощи кнопки Сохранить на панели инструментов или в окне редактора файл-сохранение. Редактор дает возможность сохранения документов в форматах: dot (шаблоны), rtf (форматированный текст), txt (текстовый без форматирования), HTML (текст с гиперссылками), в формате MS-DOS, doc (предыдущей версии Word 6.0/95), xls (электронные таблицы) и т.д.;

г) автоматическое форматирование текста при вводе, например, автоматическое форматирование порядковых номеров, дробей, автоматическое оформление сетевых путей и адресов Интернета как гиперссылок и т.д.;

д) создание таблиц при помощи мыши или при помощи команды Вставка - Таблица. Далее осуществляют выравнивание содержимого ячеек по ширине, по верхнему или нижнему краю ячейки, производят объединение ячеек, производят вертикальную ориентацию текста, например, поворот текста на 90 градусов и т.д.);

е) для оформления документов имеется более 150 различных типов границ, в том числе объемные границы и границы, которые состоят из нескольких линий, являющиеся наиболее популярными в профессиональных документах, например, границы страниц, границы текста.

Таким образом, данная программа является основной (базовой) программой для работы с текстовыми документами.

6.2 Программа Microsoft Excel

Microsoft Excel для Windows является удобным и мощным программным средством при работе с таблицами данных, позволяющим упорядочивать, анализировать и графически представлять различные виды данных.

Файл Microsoft Excel представляет собой книгу, состоящую из одного или нескольких листов. При открытии файла загружаются все листы книги, а один из них выводится на экран. Лист в Microsoft Excel имеет столбцы и строки, которые, пересекаясь, образуют ячейки данных. Столбцы озаглавлены латинскими буквами, строки - арабскими цифрами. В каждом листе имеется 256 столбцов и 16 384 строки. Имя ячейки образуется из буквы, обозначающей столбец, и номера строки (A1, B2 и т. д.). В ячейках хранятся числовые и текстовые данные, а также математические формулы, которые могут использовать данные, хранящиеся в других ячейках книги.

Рабочее окно Microsoft Excel имеет стандартный интерфейс. Под строкой заголовка с именем файла расположена строка меню, ниже - панели инструментов стандартная и форматирования. Под панелями слева находится поле имени, в котором обозначены текущие или выделенные ячейки, строки или столбцы. Правее находится строка формул, в которой показывается вводимый в ячейку текст или формула.

Проще всего запустить Microsoft Excel, щелкнув по ярлычку Microsoft Excel на рабочем столе или из меню Программы.

Вводить данные можно только в текущую ячейку. Текущую ячейку можно отличить по рамке вокруг нее. Сделать ячейку текущей можно, если выделить ее мышкой или с помощью клавиатуры клавишами управления курсором. Чтобы выделить другую ячейку, нужно щелкнуть на ней мышью. Введенные данные появляются в текущей ячейке и в строке формул.

Введенные в ячейку данные или формулы можно легко распространить (скопировать) на соседние ячейки. Для этого надо сделать ячейку с копируемыми данными текущей и потянуть за маркер автозаполнения, который расположен в правом нижнем углу рамки, распространяя данные на нужные ячейки.

Ввод последовательности данных используется для автоматического ввода последовательности или прогрессии числовых и текстовых данных. Достаточно ввести значения (например, 1 и 2) в две первые ячейки и перетащить мышью маркер автозаполнения, чтобы получить числовой ряд: 1, 2, 3, 4, 5, и т. д.

Достаточно часто стандартные размеры ячеек не позволяют увидеть введенные данные. Чтобы избежать просмотра данных по частям, можно изменять размеры ячеек. Изменение ширины столбца (высоты строки) производится путем установки указателя мыши на разделительную линию (он примет вид отрезка со стрелками), нажатия левой кнопки и передвижения мыши в нужное положение.

Форматирование данных в программе Microsoft Excel мало отличается от форматирования в редакторе MS-Word и других программах пакета MS-Office. В MS-Excel имеются возможности для автоматического форматирования данных. Для их использования выделяется ячейка или блок ячеек, подлежащих форматированию, в меню Формат выбирается командная строка Автоформат и осуществляется подбор из предложенных вариантов.

Для форматирования также используются кнопки на панели инструментов форматирования (рисунок 6.1). Большинство кнопок этой панели стандартны для всех программ пакета MS-Office.

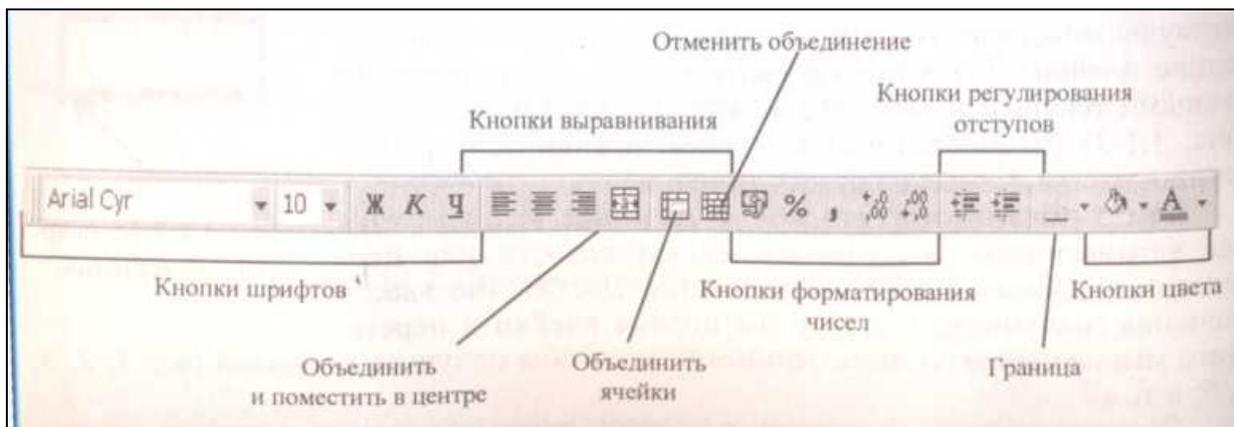


Рисунок 6.1 - Панель инструментов "Форматирование"

При форматировании рекомендовано активнее пользоваться контекстным меню (щелчок правой кнопки мыши по редактируемым ячейкам). Особенно большие возможности предоставляет опция контекстного меню **Формат ячеек**.

Проведение расчетов осуществляется следующими способами: автосуммирование, составление формул расчета, составление функций с помощью Мастера функций.

Формулы для вычисления записываются в ячейку, в которую будет записан результат вычислений. Ввод формул в ячейках необходимо начинать с арифметического оператора, обычно знака равенства «=», после которого указываются адреса ячеек, содержимое которых используется в формулах. Между именами ячеек вписываются математические операторы. Клавишей **Enter** выводятся результаты вычислений.

Формула может иметь вид не только суммы. Применяются операторы «-» (вычесть), «*» (умножить), «/» (разделить), «%» (процент от числа), «^» (степень). Порядок действий определяется скобками. Если задействованы несколько чисел, то применяются операторы связи: A1:A10 - все числа в столбце A от 1-й строки; A1; A3; A5—1-е, 3-е и 5-е число столбца A и т. д.

В вычислениях приходится использовать формулы, содержащие числа и функции. MS-Excel имеет сотни функций, которые помогают легко выполнять специальные вычисления. Любая функция представляет собой отдельную встроенную формулу, выполняющую операции над указанными данными.

Для ввода формулы необходимо выделить ячейку, в которую будет записан результат вычислений, а затем нажать на кнопку Мастер функций.

В диалоговых окнах Мастера необходимо выбрать функцию и выделить диапазоны ячеек с данными.

Если в строке или столбце несколько раз повторяется одна и та же формула, то для ускорения расчетов ее можно скопировать и применить к другим данным. Копирование формул в диапазон ячеек осуществляется так же, как и копирование данных:

Необходимо выделить ячейку, содержащую формулу, и потянуть за маркер автозаполнения, распространяя формулу на нужные ячейки.

MS-Excel - это прекрасный инструмент для построения различных типов графиков и диаграмм.

Построение диаграмм осуществляется с помощью Мастера диаграмм, который вызывается одноименной кнопкой на стандартной панели инструментов или опцией Диаграмма в меню Вставка.

Построение диаграмм необходимо начинать с выделения данных, которые будут в ней использоваться, а затем обращаться к Мастеру диаграмм. Диалоговые окна Мастера диаграмм позволяют выбрать тип и вид диаграммы, ввести ее название, легенду и заголовки осей. Тип создаваемой диаграммы зависит от решаемой задачи и характера выделенных данных. Можно выбрать как один ряд данных (отдельную строку или отдельный столбец), так и несколько.

Диаграмму можно форматировать, т.е. перемещать по рабочему столу, изменять ее размеры и пропорции, добавлять, удалять и переупорядочивать

ряды данных, изменять тип диаграммы, удалять и добавлять любые элементы, изменять форматы (оформление) ее элементов.

Форматирование диаграммы производится путем редактирования ее элементов. Наиболее часто форматируемые элементы диаграмм приведены на рисунке 6.2.

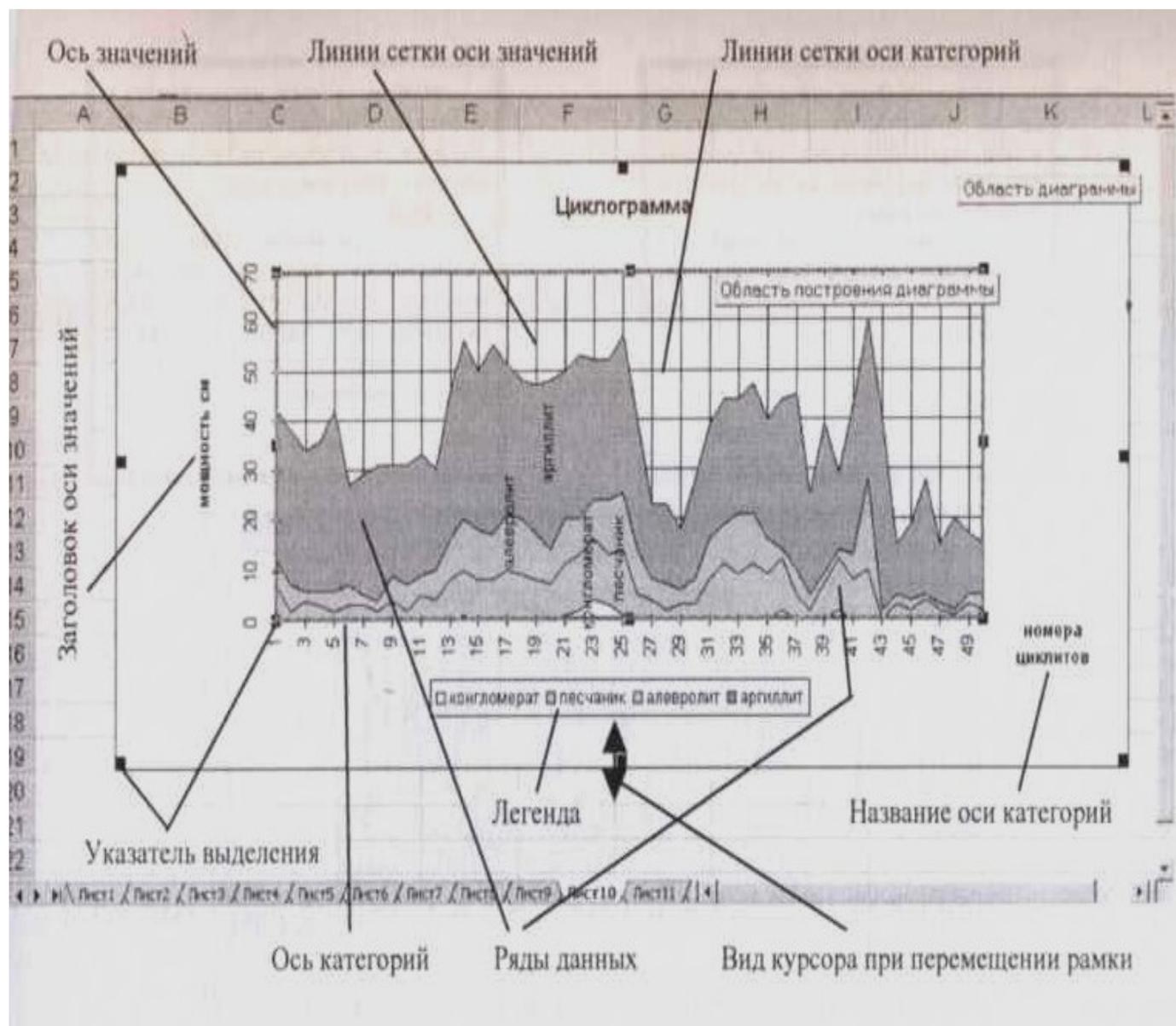


Рисунок 6.2 - Элементы диаграммы

При форматировании диаграммы удобно пользоваться кнопками панели инструментов Диаграмма (рисунок 6.3).



Рисунок 6.3 - Панель инструментов "Диаграммы"

Если при работе над отчетом нужно выделить на диаграмме какие-либо данные, то можно добавить стрелки, линии или текст. Они строятся при помощи кнопок на панели Рисование (рисунок 6.4).

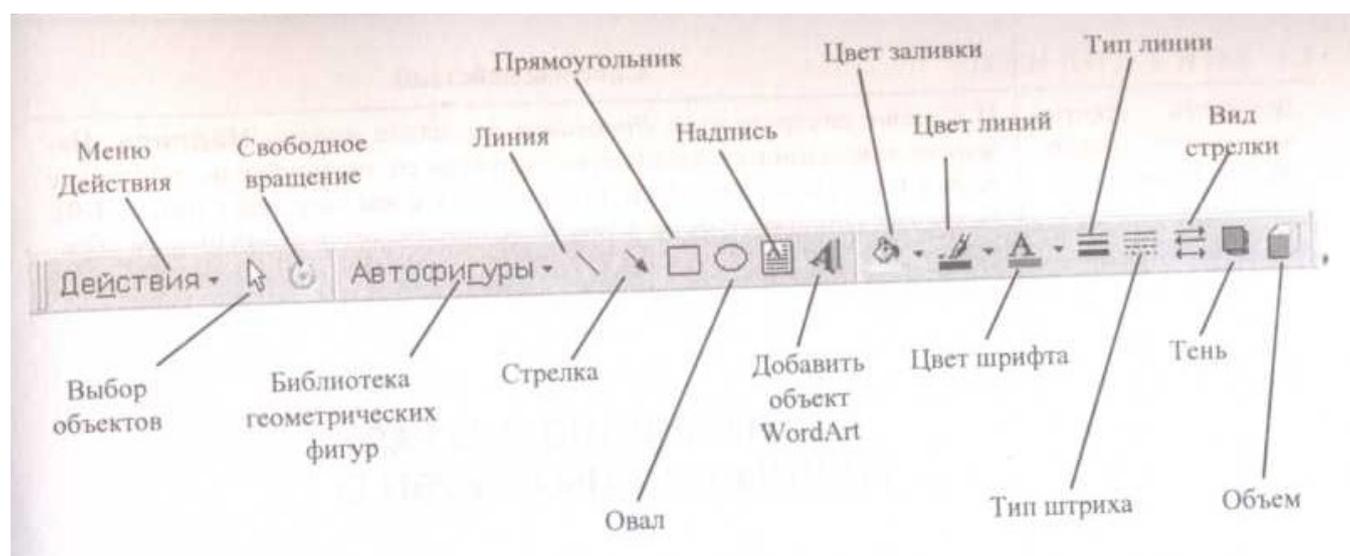


Рисунок 6.4 - Панель инструментов "Рисование"

6.3 Программа CorelDraw

CorelDRAW представляет собой объектно-ориентированный пакет программ для работы с векторной графикой.

Термин «объектно-ориентированный» следует понимать в том смысле, что все операции, выполняющиеся в процессе создания и изменения изображений, пользователь проводит не с изображением в целом и не с его мельчайшими частицами (пикселями точечного изображения), а с объектами — элементами изображения. Из стандартных объектов (линий, кругов, прямоугольников и т. д.), пользователь может строить составные объекты и манипулировать ими как единым целым.

Каждому стандартному классу объектов CorelDRAW ставится в соответствие уникальная совокупность управляющих параметров или атрибутов класса. Например, если имеется объект «прямоугольник», то для него зафиксированы определенные значения управляющих параметров: высота, ширина, цвет контура и заливки и др.

Для каждого стандартного класса объектов в CorelDRAW определен перечень стандартных операций. Например, прямоугольник можно развернуть, масштабировать, закруглить ему углы, преобразовать его в объект другого класса - замкнутую кривую.

Объектная ориентация CorelDRAW дает пользователю почти неограниченную гибкость в работе. Можно выделять отдельные объекты изображения и модифицировать их на любом этапе работы, что невозможно ни для точечных изображений, ни при использовании традиционных инструментов художника — бумаги, кисти, пера, красок, карандашей.

До логического завершения концепцию объектной ориентированности пакета доводит введение в его состав объектно-ориентированного языка программирования VBA. Программные модули на этом языке позволяют автоматизировать вы-

полнение часто повторяющихся действий и даже строить на основе CorelDRAW специализированные графические системы, определяя новые классы объектов и операции над ними. В программный модуль на языке VBA можно преобразовать последовательность действий, причем их запись ведется автоматически (для возможной отмены последних действий).

В CorelDRAW можно просто рисовать с помощью соответствующих инструментов, однако CorelDRAW больше используют как конструктор сложных графических изображений из множества составляющих частей — объектов, которые редактируются независимо друг от друга, поэтому основными рабочими инструментами являются кривые линии и геометрические фигуры (примитивы), которые можно модифицировать различным образом. Из объектов можно создавать группы для дальнейшего редактирования группы как единого объекта.

Вне зависимости от внешнего вида, любой векторный объект CorelDRAW имеет некоторое количество точек или узлов, соединенных прямыми или кривыми линиями — *сегментами*. Координаты узлов и параметры сегментов определяют внешний вид объекта. Изменение этих составляющих частей объекта приводит в результате к созданию требуемого изображения в редакторе векторной графики.

Область внутри объекта можно закрасить или залить одним цветом, смесью цветов или узором. Заливка в CorelDRAW может быть не только стандартной, но и с эффектами протекания цветов, размытия, а также состоять из сложной узорчатой текстуры. Сегменты объекта образуют контур, имеющий свой цвет и толщину, которые тоже можно изменять. Различают замкнутые и разомкнутые контуры.

Созданные в CorelDRAW векторные объекты можно копировать, дублировать и клонировать, накладывать друг на друга (используя различные формы объединения), а также изменять их атрибуты. К ним можно применять операции манипулирования (размещение, поворот, масштабирование, зеркальное отражение), изменять их внешний вид (используя точки или узлы модификации). К объектам можно применять различные оригинальные эффекты, используя для этого интерактивные средства (перетекание, контур, искажение, оболочку, прозрачность), линзы

и др., а также высококачественную иллюзию объема (путем применения соответствующего эффекта, усиленного применением цветовых эффектов и освещения).

В CorelDRAW имеется библиотека профессионально нарисованных символов, охватывающих различные темы (бизнес, окружающая среда, наука, транспорт и пр.), и готовые рисунки, которые можно использовать при создании собственных иллюстраций. После вставки в документ их можно редактировать подобно любому другому объекту (менять форму, добавлять и удалять заливку и др.).

В CorelDRAW можно использовать растровые изображения, вставляя их в графически документ. При этом каждый растровый рисунок является отдельным объектом, и его можно редактировать независимо от других объектов. Хотя CorelDRAW предназначен для работы с векторной графикой, у него имеются некоторые средства для работы с растровыми рисунками.

Шрифты в CorelDRAW (как и любой другой объект) также можно модифицировать: вытягивать и сжимать, наклонять и переворачивать, раскрашивать и покрывать текстурой, снабжать тенью и объемом, делать прозрачным или размытым и т.п. В CorelDRAW существует возможность работы с двумя разновидностями текстовых объектов: с фигурным и обычным текстом.

Фигурный текст представляет собой графический объект, с которым можно работать как с любым другим объектом. Фигурный текст можно разместить по любой заданной прямой или кривой линии, прямоугольнику, эллипсу, многоугольнику, другой букве или текстовой строке.

Обычный текст представляет собой массив текста в рамке, вставленный в рисунок. Можно менять границы рамки обычного текста или придавать ей замысловатую форму, но внутри текст будет располагаться точно так же, как и в любом текстовом редакторе, например, в Word. CorelDRAW имеет многие возможности форматирования текста (изменение типа, размера и начертания шрифта, выравнивание и т.п.).

Изображения, созданные в CorelDRAW, можно сохранять как в векторной форме, так и формате растровой графики.

CorelDRAW поддерживает слоистую структуру рисунков. Объекты, из которых состоят сложные рисунки, могут быть разделены на тематические группы и распределены по разным слоям (уровням). Слои CorelDRAW позволяют выбирать объекты, принадлежащие только к определенному слою или группе слоев, делая их доступными для просмотра, редактирования и вывода на печать.

Электронные слои напоминают кальки с нарисованными на них объектами карты: горизонталями рельефа, речной и овражной сетью, дорогами и населенными пунктами. Можно просмотреть одну кальку или наложить их друг на друга.

Слои рисунка отображаются в окне Диспетчера объектов меню Инструменты. В окне Диспетчера объектов приведены все имеющиеся слои. Линии сетки находятся в своем слое, направляющие - в своем. Выделив любой слой и вызвав контекстное меню → опция Свойства, можно редактировать параметры слоя. Списки слоев, расположенные в верхней части окна Диспетчера объектов, используются при работе с многостраничными документами.

При создании нового рисунка CorelDRAW предлагает разместить объекты рисунка в слое 1. По мере необходимости можно создавать новые слои, а их названия - редактировать. Создавать и редактировать объекты можно только в рабочем (активном) слое, который выделен в списке слоев красным цветом.

Управление слоями осуществляется с помощью пиктограмм в окне Диспетчера объектов. Один щелчок мыши по соответствующей пиктограмме переводит слой в состояние "видимый/невидимый", "печатаемый/непечатаемый", "редактируемый/нередатируемый".

6.4 Программа Microsoft Access

Microsoft Access - это интерактивная реляционная СУБД для Windows. Данная программа дает возможность хранения и извлечения данных в зависимости от отношений, которые были установлены пользователем. Работа в данной программе облегчена благодаря использованию манипулятора мыши. Графические возможности оболочки производят большое впечатление при изготовлении высококачественных отчетов и распечаток.

Access дает возможность пользователю работать с базами данных различных форматов, например, базы данных dBASE, Paradox или Vtrieve без конвертирования их в формат, используемый Access. В составе пакета Access имеется язык Access Basic (встроенный диалект языка Visual Basic), который дает возможность формирования специализированных систем управления базами данных.

Microsoft Access for Windows является удобной и мощной СУБД - системой управления базами данных, которая позволяет: создавать базы данных, осуществлять ввод, просмотр и обновление данных, обращаться к ним при помощи запросов для извлечения и обработки информации. MS-Access относится к реляционным базам данных. Термин «реляционная» означает, что вся информация об объектах может быть записана в нескольких связанных между собой таблицах, а каждая запись в такой базе данных содержит информацию только об одном объекте.

Файл Microsoft Access представляет собой базу данных, которая включает таблицы, запросы, формы, отчеты и другие типы объектов MS-Access, доступ к которым осуществляется из диалога База данных.

База данных представляет собой набор данных, который связан с определенной темой или назначением. Когда Вы создаете базу данных в Microsoft Access, Вы

создаете один файл, который содержит все таблицы, запросы, формуляры, отчеты и другие объекты, которые помогают Вам использовать информацию в базе данных.

Для создания этих объектов в базе данных используется специальный режим Конструктора объектов (конструктор таблицы, конструктор запроса и т.д.). Ввод, просмотр и редактирование данных осуществляется в режиме соответствующего типа объекта (режим таблицы, режим запроса и т.д.)

Таблицы предназначены для хранения данных. Вертикальные графы таблицы называются полями (столбцами), горизонтальные - записями (строками). Каждая таблица содержит информацию об определенных объектах. В столбцах таблицы хранятся различные сведения об этих объектах (атрибуты объекта). В строках приведены описания значений всех атрибутов каждого объекта. Например, объектами таблицы с данными геохимического опробования горных пород являются пробы (точки опробования), их атрибутами - оксиды, значения атрибутов - содержания оксидов в конкретной пробе.

Структура таблиц задается в режиме конструктора. В этом режиме каждая строка в верхней части панели окна таблицы соответствует одному из полей создаваемой или редактируемой таблицы. В столбце Имя поля окна таблицы вводятся заголовки полей таблицы, в столбце тип данных - указывается тип данных этого поля (текстовой, числовой и т. д.), который выбирается из предложенного списка. Столбец Описание предназначен для описания полей таблицы, примечаний.

Запросы позволяют извлекать нужные данные из одной или нескольких таблиц. С помощью запросов можно выбирать, обновлять, добавлять или удалять данные, выполнять различные вычисления, а также создавать новые таблицы, используя для этого данные уже существующих таблиц. Запросы создаются и редактируются в режиме конструктора запросов. Окно конструктора запросов разделено на две панели. В верхней панели помещены списки полей тех таблиц или запросов, к которым будет обращен создаваемый

запрос. Нижняя панель представляет собой бланк, заполняющийся при работе над запросом. Каждый столбец бланка отвечает полю таблицы (или запроса), по которым формируется запрос.

Запрос можно представить себе как точку зрения на данные, включенные в таблицу. Запросы служат для селекции и фильтрации набора данных. Они позволяют выбрать из базы только необходимую информацию, т.е. ту, которая соответствует определенному критерию (условию) и нужна для решения конкретной задачи. Например, Вам может понадобиться информация о поставщиках и поставляемых ими товарах. Вы можете создать запрос типа «Какие товары поставляют московские поставщики» или «Кто из петербургских поставщиков сигарет продал за последний квартал наибольшую партию». Результат обработки программой Access такого запроса представляет собой таблицу называемую Dynaset. В эту таблицу включены выбранные из основной таблицы (или нескольких таблиц) блоки данных, которые удовлетворяют критериям запроса. Dynaset - динамический, временный набор данных, поэтому при каждом выполнении запроса он строится вновь на основе «свежих» табличных данных.

Microsoft Access дает большие возможности в гибкости проектирования запросов.

Например:

1. Выбор поля. Нет необходимости включать все поля таблицы в запрос.
2. Выбор записи. Дает возможность специфицировать запрос таким образом, чтобы получить данные определенного вида. Например, цены на товары за определенный месяц.
3. Отсортировка записи. Дает возможность просматривать записи в определенном порядке. Например, можно вывести имена клиентов на экран в алфавитном порядке.
4. Запрашивать данные из нескольких таблиц, данные из нескольких таблиц могут обрабатываться вместе и это дает возможность просмотреть совмещенные данные. Также можно запрашивать данные из других баз данных.

5. Выполнять вычисления. Это дает возможность создавать новые поля которые будут содержать результаты вычислений.

6. Использовать запрос в качестве источника данных для формуляров, отчетов и других запросов. Каждый раз когда вы будете открывать формуляр или печатать отчет, ваш запрос будет извлекать новейшую информацию из таблиц. Вы также сможете модифицировать существующие данные прямо в формуляре основанном на запросе.

7. Изменять данные в таблицах. Вы можете обновлять, удалять, добавлять группы записей все сразу. Вы также можете создать новую таблицу на базе существующей таблицы или группы таблиц.

Формы предназначены для ввода, редактирования и визуализации данных на экране; кроме того, они позволяют создавать пользовательский интерфейс для управления работой базы данных. Формы создаются и редактируются в режиме Конструктор. Ввод, редактирование и просмотр данных осуществляется собственно в режиме Форма. Стандартное окно формы, открывающееся по умолчанию в режиме конструктора, состоит из области данных, в которой размещают информацию из базовой таблицы или запроса. В верхнюю часть окна может быть добавлена область заголовка, в нижнюю часть - область примечания. В форме для просмотра данных из базовой таблицы используются специальные поля, которые называются присоединенными или связанными, так как они связаны с полями в источнике данных. Их также называют элементами управления данными. Элементы управления располагаются в области данных окна формы. В стандартном виде элементы управления состоят из двух частей: справа размещается сам элемент управления (окошко поля, в котором выводятся данные для просмотра), слева располагается связанная с ним подпись (название этого поля).

Перед началом работы в режиме конструктора необходимо проверить, какие панели инструментов имеет интерфейс MS-Access на рабочем компьютере. Для этого в меню Вид окна режима конструктора формы нужно

воспользоваться командой Панели инструментов и просмотреть список панелей. Панели Конструктор форм, Панель элементов и Формат должны быть отмечены флажками. Если они не установлены, то необходимо их установить, последовательно выбрав из списка.

Отчеты. Отчет содержит в себе информацию, которая была оформлена пользователем в соответствии с его спецификациями. Отчет дает возможность извлечения и представления данных как значимую информацию, которую далее пользователь имеет возможность использовать и распространять. Отчетами являются почтовые адреса, накладные, суммы продаж или списки телефонов. При помощи Microsoft Access можно спроектировать отчет, который будет содержать информацию в том виде, в котором нужно пользователю. Существует возможность использования большого количества разнообразных элементов проектирования, например, текст, данные, рисунки, линии, поля и графики для создания нужного отчета. От пользователя зависит, как и какие элементы он будет использовать для построения отчета.

Отчеты аналогично являются эффективным средством для распечатки информации регулярного использования. Пользователь может создать проект отчета и сохранить его для дальнейшего использования. Проект отчета останется неизменным, но каждый раз пользователь будет получать распечатку текущих данных.

Список использованных источников

1. Васильев, А.А. Компьютерное моделирование / А.А. Васильев - Режим доступа: http://www.ctc.msiu.ru/materials/Book2/ch_09_model/01_model/01_comp_model/index.html
2. Виноградов, Е.Б. Современная геология: научное издание /Е.Б. Виноградов - Екатеринбург : Изд. - во ЕПД, 2007. - 81 с.
3. Дьяконов, В.В. Компьютерные методы обработки геологической информации: учебное пособие / В.В. Дьяконов, Н.В. Жорж. – М.: РУДН, 2008. – 266 с.
4. Мартянова А.Е. Математические методы моделирования в геологии: учебное пособие. Часть I / А.Е. Мартянова. – Астрахань: АГТУ, 2008. – 218 с.
5. Приемы ввода и обработки первичной геологической документации: учебное методическое пособие по компьютерным технологиям. СПб: Изд-во СПб карт-фабрики ВСЕГЕИ, 2001. - 204 с.
6. Исаев, В.Ю. Моделирование в геологии / В.Ю. Исаев - Режим доступа: http://www.drtisygeology.ru/materials/Book2/ch_09_model/_model423/index

Учебное пособие

Елена Григорьевна Щеглова

**ПРИМЕНЕНИЕ ЭВМ В ГЕОЛОГИИ:
КУРС ЛЕКЦИЙ**

ISBN 978-5-7410-1315-1



9 785741 013151