СПРАВОЧНИК
ПО МАТЕМАТИЧЕСКИМ МЕТОДАМ В ГЕОЛОГИИ

МОСКВА "НЕДРА" 1987

УДК 550.8.053 : 519.2 (031)

Справочник по математическим методам в геологии/Д. А. Родионов, Р. И. Коган, В. А. Голубева и др.— М.: Недра, 1987.— 335 с. ил.

Изложены основные понятия теории вероятностей и математической статистики, используемые при обработке геологоразведочной информации. Дана характеристика математических методов, применяемых при прогнозе, поисках и разведке месторождений полезных ископаемых. Приведены практические рекомендации по использованию этих методов в геологических исследованиях. Рассмотрены вопросы устойчивости и достоверности получаемых статистических выводов.

Для геологов различного профиля, связанных с использованием математических методов и ЭВМ.

Табл. 18, ил. 72, список лит.— 52 назв.

Рецензент: *Ю. В. Прохоров*, академик АН СССР (Математический институт им. В. А. Стеклова АН СССР)

Авторы: *Д. А. Родионов, Р. И. Коган, В. А. Голубева, Б. И. Смирнов, С. В. Сиротинская*

С $\frac{1904010000-240}{043(01)-87}$ 53—87

ПРЕДИСЛОВИЕ

Математические методы в геологических исследованиях интенсивно и с успехом применяются уже более четверти века. Число геологических работ с применением математических методов в настоящее время увеличивается. Весьма актуальными являются вопросы определения условий применимости разработанных методов, устойчивости используемых алгоритмов, оптимизации расчетов на ЭВМ, т. е. вопросы, требующие высокой математической культуры исследователя-геолога. Поэтому назрела настоятельная необходимость в издании справочника, в котором были бы представлены важнейшие математические методы обработки геологоразведочных данных, прошедших проверку на обширном экспериментальном материале.

В настоящем справочнике дается толкование математических терминов и понятий, полезных в практике геологических исследований. Они относятся к различным разделам математики: теории множеств, теории вероятностей, математической статистике, математической логике и к другим математическим дисциплинам. Однако главное внимание уделено изложению терминов теории вероятностей и математической статистики. Справочник содержит и конкретные математические методы решения практических задач, наиболее часто встречающихся в геологических исследованиях, как, например, прогнозирование геологических характеристик, классификация геологических объектов и проверка гипотез о параметрах, выделение информативных комбинаций признаков, разграничение геологических объектов по комплексу признаков, распознавание образов и т. д.

Степень детальности описания терминов различна. Одни освещены весьма подробно, для других приведены лишь определения и основные свойства, ряд терминов только упоминаются с соответствующей ссылкой на литературный источник. Среди этих терминов присутствуют как обобщающие понятия (например, вероятность, распознавание образов, математическая логика и т. п.), так и конкретные (методы, критерии, алгоритмы).

Следует отметить, что это первый опыт создания такого справочника, поэтому, естественно, он не свободен от ряда недостатков. Например, в нем представлены далеко не все методы современной

прикладной статистики, которые применялись уже при обработке геологоразведочных данных; недостаточно уделено внимания робастным (помехоустойчивым) процедурам получения оценок геологических параметров; нет теории нечетких (размытых) множеств и др.

В целом справочник будет полезен геологам, имеющим дело с обработкой геологической информации и использующим математические методы в своих исследованиях.

В написании ряда разделов принимали активное участие А. И. Ежов, Ю. П. Белов, С. Н. Тесаков. Авторы приносят благодарность Г. А. Зыбиной, О. А. Хрящевой, Е. А. Озерецковской, М. В. Родионовой, Г. М. Жарикову за помощь в подготовке рукописи.

ГЛАВА 1

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

Теория вероятностей — математическая дисциплина, основанная на интуитивном представлении вероятностей появления объекта или события. Анализ вероятностей позволяет оценить поведение изучаемого объекта или события в прошлом или будущем. Вероятность — количественное выражение возможности того, что данное событие произойдет. Ее принято выражать в долях единицы. При этом нуль соответствует невозможности появления события, а единица — полной уверенности в том, что оно произойдет.

Первоначально теория вероятностей развивалась для описания азартных игр. В настоящее время — это весьма богатая теория, в которой, как и во всякой теории, можно различать три стороны: а) формальное логическое содержание, б) интуитивные представления, в) приложения.

В настоящее время теория вероятностей строится на строгой аксиоматической основе, заложенной А. Н. Колмогоровым. В свою очередь, аксиомы теории вероятностей выведены из интуитивных представлений о вероятности события. Совершенно очевидно, что теория и приложения взаимодействуют друг с другом: прогресс в теории открывает новое поле для приложений, а каждое новое приложение создает новые теоретические проблемы и оказывает влияние на направление исследований.

В настоящее время число приложений неуклонно растет. Остановимся лишь на примерах геологических задач, в которых используется понятие вероятности. Вот простая задача: имеется керн значительной длины, взятый из скважины, проходящей через слой песчаника. Требуется оценить проницаемость песчаника.

Другой пример задачи, возникающий при поисково-разведочных работах, например при поисках месторождений нефти: следует оценить вероятность обнаружения месторождения с помощью анализа динамики открытия новых месторождений в достаточно длинном ряду испытаний (под испытанием понимается поисковое бурение). Соотношение успехов и неудач определяет коэффициент успешности бурения, т. е. вероятность успеха бурения.

Еще одна важная область применения теории вероятностей в геологических исследованиях: выработка стратегии эффективного поиска месторождения, для чего используются детальные сведения об относительных размерах, форме, вероятности и пространственном распределении месторождений, являющихся конечным объектом поискового бурения.

Известные вероятностные распределения, в свою очередь, используются в геологии для решения обширного класса задач: лог-

нормальное распределение — для характеристики изменений размеров частиц осадков и концентраций химических элементов, нормальное — при решении задач проверки гипотез о параметрах геологических объектов, а также, например, для описания распределений вероятностей величины запасов газовых месторождений и т. п.

Число примеров можно было бы увеличить. Однако это не входит в задачу справочника. Отметим только, что с появлением ЭВМ число приложений теории вероятностей в геологии растет, позволяя повышать надежность геологических выводов и оценивать экономический эффект той или иной стратегии исследований геологических объектов.

СОБЫТИЕ, ОПЕРАЦИИ НАД СОБЫТИЯМИ, ВЕРОЯТНОСТЬ СОБЫТИЯ

В е р о я т н о с т ь — число $P(A)$, поставленное в соответствие некоторому событию A и характеризующее возможность его появления в тех или иных заданных условиях, которые могут быть повторены неограниченное число раз. Оно удовлетворяет условию

$$0 \leqslant P(A) \leqslant 1.$$

С л у ч а й н о е с о б ы т и е. Событие, которое в результате проведения эксперимента может произойти, а может и не произойти, называется случайным. Если A — случайное событие, то вероятность $P(A)$ его появления больше нуля и меньше единицы т. е.

$$0 < P(A) < 1.$$

С о б ы т и е — это исход эксперимента. Последний называется активным, если исследователь может управлять хотя бы некоторыми его существенно важными параметрами, и пассивным, если не может. Для геологии типична ситуация пассивного эксперимента, в связи с чем понятие «событие» применительно к геологическим явлениям определяется как результат наблюдения (измерение мощности пласта, уровня радиоактивности, определение содержания химического элемента в пробе и т. п.). Предполагается, что комплекс условий, в которых проводится эксперимент или осуществляется наблюдение, может быть повторен неограниченное или, по крайней мере, достаточно большое число раз. В данном случае под комплексом условий понимается, с одной стороны, устойчивость основных характеристик, определяющих «лицо» и специфику исследуемого явления, а с другой — единообразие методологических основ и технических условий выполнения наблюдений.

В связи со случайным характером событий конкретный результат отдельного (единичного) набюдения не может быть известен до завершения измерения. Чаще всего исследователь может лишь перечислить возможные элементарные исходы (события), соответствующие данному комплексу условий. Элементарные события, обычно обозначаемые символом ω, обладают следующими свойствами: 1) они являются непересекающимися, т. е. взаимно исключают друг друга; 2) любой интересующий нас результат наблюдения (событие A), который в принципе возможен при реализации данного комплекса условий, может быть сопоставлен с некоторым множеством элементарных событий. В этом случае уместно отождествить событие A с соответствующей совокупностью элементарных исходов $\{\omega : \omega \in A\}$, $A \subseteq \Omega$, где Ω — пространство или множество элементарных событий. Заметим, что подмножество A может быть несчетным, счетным, конечным, содержать только один элемент и, наконец, быть пустым. Если событие A обязательно происходит в результате эксперимента, то оно называется достоверным; если никогда не происходит — невозможным; а если может произойти или не произойти — случайным. Для любого случайного события A по наступившему элементарному исходу ω_i можно судить о том, происходит ($\omega_i \in A$) или не происходит ($\omega_i \notin A$) это событие. Иллюстрацией к вышесказанному может служить следующий пример. При опробовании золотоносной россыпи некоторая конкретная проба может попасть в зону, «богатую» полезным ископаемым, либо в «бедную» зону (событие A). Условимся, что зона идентифицируется нами как «бедная» при числе золотинок в пробе менее m. Тогда совокупность возможных элементарных исходов, соответствующая событию A, будет составлена из следующих элементарных событий: $\omega_1 = 0$ (отсутствие золотинок), $\omega_2 = 1$, $\omega_3 = 2, \ldots,$ $\omega_m = m-1$ золотинок в пробе. Наступление любого из перечисленных исходов влечет за собой реализацию события A.

Представление событий в виде множеств элементарных исходов определяет возможность комбинирования событий с помощью таких операций, как объединение, пересечение и т. д., что позволяет глубже и полнее исследовать отношения между ними.

Д о с т о в е р н о е с о б ы т и е. Событие A, вероятность появления которого $P(A)$ равна 1, называется достоверным, т. е.

$$P(A) = 1.$$

Это значит, что такое событие будет обязательно происходить во всех экспериментах.

Н е в о з м о ж н о е с о б ы т и е. Событие A, вероятность появления которого $P(A)$ равна 0, называется невозможным, т. е.

$$P(A) = 0.$$

Это означает, что такое событие не произойдет ни в одном из экспериментов. Например, наличие кварца в пробе гранита является достоверным событием, а алмаза — практически невозможным.

О б ъ е д и н е н и е (с у м м а) с о б ы т и й. Событие B называется объединением событий $A_1, A_2, \ldots, A_i, \ldots, A_n$, если оно имеет вид: наступает A_1, или $A_2, \ldots,$ или A_n. Объединение событий обозначается знаком $\cup$, т. е.

$$B = A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i.$$

Под объединением событий в качестве примера можно понимать появление какого-либо граната: альмандина, андрадита, гроссуляра в скарне.

Совмещение (произведение) событий — это событие, состоящее в совместном появлении всех участвующих в этой операции событий. Произведение (иногда его называют пересечением) событий A_1 и A_2 означает реализацию и события A_1, и события A_2. Совмещение обозначают знаком $\cap$, т. е.:

$$A = A_1 \cap A_2.$$

Таким образом, событие A состоит из элементарных событий (исходов), принадлежащих одновременно и A_1 и A_2.

Более чем для двух событий

$$A = A_1 \cap A_2 \cap \ldots \cap A_n = \bigcap_{i=1}^{n} A_i.$$

В качестве примера совмещения событий можно рассматривать наличие натриевой и кальциевой составляющих в минералах группы плагиоклаза.

Вероятность условная — вероятность осуществления события A_1 при условии, что произошло событие A_2. Условная вероятность, обозначаемая как $P(A_1/A_2)$, определяется через отношение

$$P(A_1A_2)/P(A_2).$$

Условные вероятности обладают всеми свойствами обычных («безусловных») вероятностей. В частности,

$$0 \leqslant P(A_1/A_2) \leqslant 1.$$

Равенство $P(A_1/A_2)=0$ реализуется в том случае, если осуществление события A_2 исключает возможность появления A_1, т. е. $A_1 \cap A_2 = \phi$. Если же событие A_2 обеспечивает обязательное наступление события A_1, что соответствует ситуации $A_2 \subset A_1$, то $P(A_1/A_2) = 1$. Сравнивая условную вероятность $P(A_1/A_2)$ с «безусловной» $P(A_1)$, можно выяснить интенсивность и направление влияния на событие A_1 некоторого условия, фиксируемого исследователем в форме события A_2. Если наступление события A_1 никак не связано с событием A_2 (при $P(A_2) > 0$), то это найдет свое отражение в выполнении равенства

$$P(A_1/A_2) = P(A_1).$$

Любое нарушение этого равенства свидетельствует о зависимости события A_1 от события A_2. При этом, если $P(A_1/A_2) > P(A_1)$, то говорят, что событие A_2 благоприятствует осуществлению события A_1; если же $P(A_1/A_2) < P(A_1)$, то реализация события A_2 снижает шансы появления события A_1.

Идеи, лежащие в основе понятия условной вероятности, имеют важное методологическое значение. Естествоиспытатель, пытающийся раскрыть суть некоторого явления (событие A_1), стремится

по возможности фиксировать некоторые, вполне определенные и, на его взгляд, наиболее существенные условия, при которых это явление наблюдается (или не наблюдается). Фиксация условий, сопровождающих событие A_1, выполняется обычно путем регистрации событий A_2 и т. д., идентификация которых должна быть достаточно простой и вполне однозначной. Потребовав, например, обязательного осуществления события A_2, мы тем самым уточняем тот комплекс условий, в которых будет (или не будет) происходить событие A_1.

Таким образом, условные вероятности являются достаточно эффективным средством исследования взаимосвязи случайных событий. Если к тому же специально оговорить, что событие A_2 предшествует во времени событию A_1, то появляется возможность причинно-следственной интерпретации такой пары событий.

В качестве примера условной вероятности можно рассматривать вероятность появления алмазов при условии, что исследуется кимберлитовая порода.

Полная вероятность — вероятность осуществления события A в специальных, особым образом зафиксированных условиях, которые с вероятностной точки зрения, могут быть интерпретированы как совокупность случайных событий $H_1, H_2, \ldots, H_j, \ldots, H_n$. Специфичность $\{H_j\}$ в том, что они образуют так называемую полную группу несовместных событий, т. е. группу таких непересекающихся событий, из которых хотя бы одно событие обязательно происходит

$$\left(\sum_{j=1}^{n} P(H_j) = 1\right).$$

Если $A \subset \bigcup_{j=1}^{n} H_j$ (из осуществления хотя бы одного события из группы $\{H_j\}$ следует появление A), то полная вероятность имеет вид:

$$P(A) = \sum_{j=1}^{n} P(A/H_j)\, P(H_j),$$

где $P(A/H_j)$ — условная вероятность события A.

В геологии полная вероятность может найти применение при исследовании изменения вероятности появления некоторого признака (событие A) под влиянием меняющихся условий $\{H_j'\}$, $\{H_j''\}$ и т. д. Однако далеко не во всех случаях геолог располагает надежной информацией об условных вероятностях $P(A/H_j)$, что существенно ограничивает возможности этого подхода.

Формула Байеса позволяет произвести переоценку вероятностей любого из события $H_1, H_2, \ldots H_j, \ldots, H_n$ (полная группа несовместимых событий: $\bigcup_{j=1}^{n} H_j = \Omega$) после осуществления события A.

Предполагается, что вероятности $P(H_j)$ известны до опыта, в связи с чем вероятности такого рода называют априорными. Опыт

(наблюдение) планируют таким образом, чтобы одновременно регистрировалось и событие A, определяющее специфичность условий опыта, и одно из событий H_j, вероятность которого подлежит уточнению. Если условные вероятности $P(A/H_j)$ известны, то изменение вероятности H_j в связи с реализацией определенных условий (событие A) можно оценить с помощью формулы Байеса

$$P(H_j/A) = P(H_j)\,P(A/H_j)/\sum_{j=1}^{n} P(H_j)\,P(A/H_j).$$

Величина $P(H_j/A)$ определяется после того, как произведено испытание, в котором наблюдалось событие A. Поэтому вероятности, найденные таким путем, принято называть апостериорными.

Если события $\{H_j\}$ связать с определенными содержательными высказываниями, имеющими характер научных предположений (например, H_1 — наличие рудных месторождений, H_2 — их отсутствие), то $\{H_j\}$ вполне правомерно рассматривать как гипотезы. С этих позиций переоценка вероятностей по формуле Байеса, известной также под названием «теорема гипотез», есть не что иное, как вероятностная проверка степени реальности этих гипотез в некоторых фиксируемых исследователем условиях. Надежность получаемых при таком подходе выводов во многом зависит от обоснованности априорных вероятностей и статистической обеспеченности процедуры определения $P(A/H_j)$.

СЛУЧАЙНАЯ ВЕЛИЧИНА, ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Случайная величина. Величина ξ, которая в результате единичного эксперимента принимает то или иное заранее неизвестное значение, называется случайной величиной. Если множество значений $x_1, x_2, \ldots, x_m$ дискретно, то величина ξ называется дискретной случайной величиной.

В связи с тем что в геологии для большинства изучаемых характеристик нельзя предсказать точно, какое значение они примут в результате единичного эксперимента (например, взятия пробы породы и ее последующего химического анализа), случайная величина является весьма удобной математической моделью для формального представления геологических характеристик: содержаний и запасов рудных компонентов, линейных продуктивностей, геохимических ореолов, количества нефтегазоносных структур и т. п.

Функция распределения случайной величины $F(x)$ определяет вероятность того, что случайная величина ξ, примет значение, не превосходящее заданного значения x:

$$F(x) = P(\xi \leqslant x),$$

где x — произвольное действительное число.

Свойства $F(x)$:

1) $F(-\infty) = 0$, $F(+\infty) = 1$;

2) если $x_1 \leqslant x_2$, то $F(x_1) \leqslant F(x_2)$.

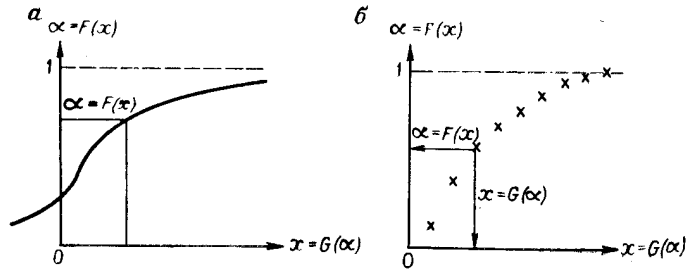


Рис. 1. Функции распределения $\alpha = F(x)$ и обратные функции распределения $x = G(\alpha)$ непрерывной (*а*) и дискретной (*б*) случайных величин

Если ξ — непрерывная случайная величина, то ее функция распределения есть неубывающая функция от x, непрерывная слева в точке x, если $F(x) = P(\xi \leqslant x)$. Так как функция распределения — монотонно возрастающая функция, то ее нередко называют кумулятивной функцией распределения (рис. 1).

Для дискретных случайных величин функция распределения является кусочно-постоянной (ступенчатой). Иногда функцию распределения называют интегральным законом распределения случайной величины.

Для любых x' и x'' $(x' < x'')$ справедливо равенство

$$P(x' \leqslant \xi \leqslant x'') = F(x'') - F(x').$$

Функция распределения полностью и единственным образом описывает распределение случайной величины. В естественнонаучных приложениях теории вероятностей обычно изучаются функции распределения, «восстановленные» по эмпирическим данным; при этом особую познавательную ценность имеют исследования тех теоретических функций распределения, которые выбраны на основе некоторых представлений о механизме образования (генезисе) данной случайной величины — аналога изучаемого геологического свойства.

Плотность распределения вероятностей — функция $f(x)$, определяющая вероятность того, что случайная величина ξ примет значение, принадлежащее интервалу $(x, x + \Delta x)$. Если функция распределения $F(x)$ непрерывна и дифференцируема, то

$$P[x < \xi < (x + \Delta x)] = F(x + \Delta x) - F(x)$$

и при $\Delta x \to 0$ имеем:

$$f(x) = \lim_{\Delta x \to 0} [F(x + \Delta x) - F(x)]/\Delta x$$

или

$$f(x) = dF(x)/dx.$$

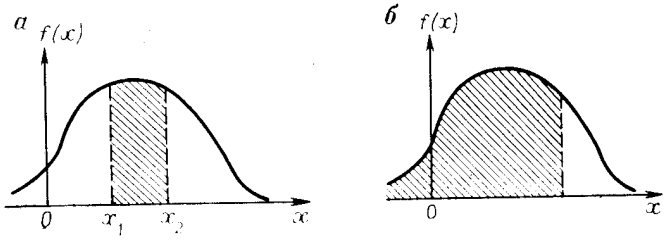


Рис. 2. Плотность вероятности $f(x)$. Заштрихованы площади:

$$а - [x_1 < x \leqslant x_2] = \int_{x_1}^{x_2} f(x)\,dx; \quad б - P(-\infty < x < x) = \int_{-\infty}^{x} f(x)\,dx$$

Связь между $f(x)$ и $F(x)$ можно выразить также следующей формулой:

$$F(x) = \int_{\infty}^{x} f(y)\,dy,$$

где $F(x) = P(\xi \leqslant x)$.

Свойства плотности распределения:

$$1) \quad f(x) \geqslant 0;$$

$$2) \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Графически $f(x)$ изображается в виде кривой, выборочным аналогом которой является гистограмма и полигон распределения (рис. 2). Плотность распределения позволяет компактно и предельно полно описать вероятностные свойства исследуемой случайной величины, что определяет широкое использование плотности распределения в естественнонаучных приложениях теории вероятностей (например, для свертывания многочисленных выборочных данных, характеризующих непрерывную случайную величину).

При вероятностном моделировании природных объектов подбор подходящей плотности распределения, аппроксимирующей наблюдаемое частотное распределение, занимает центральное место. Удачный выбор плотности распределения позволяет более точно оценить основные статистические характеристики исследуемого геологического признака, а также повысить надежность статистических выводов (за счет более обоснованного применения соответствующих тестов). Так, в условиях нормального распределения среднее арифметическое является эффективной оценкой математического ожидания случайной величины, однако эта оценка неэффективна в условиях логнормального распределения. В последнем случае оценка математического ожидания случайной величины находится на основе метода максимального правдоподобия. Точно так же различаются критерии проверки гипотез о равенстве математических ожиданий в условиях нормального и логнормального законов.

Известны попытки связать те или иные теоретические распределения с определенными моделями природных процессов, генерирующих исследуемый геологический признак. Существуют различные подходы к решению этой проблемы: чисто математический, концептуальный, механический, физический, физико-химический, термодинамический. В рамках математической модели (к ней обычно сводятся все остальные подходы) схема генерирования нормально распределенной случайной величины может быть записана так:

$$\xi = \sum_{i=1}^{n} \xi_i,$$

где ξ_i — случайные и независимые величины. Вклад их в общую сумму предполагается примерно одинаковым (см. раздел «Основные теоремы теории вероятностей»). Логнормальное распределение возникает в условиях модели:

$$\xi = \prod_{i=1}^{n} \xi_i,$$

при этом никаких ограничений на случайные величины ξ_i, кроме требования $\xi_i > 0$, не налагается.

Примерами концептуального подхода являются модели распределения химического элемента в горных породах, построенные А. Б. Вистелиусом, Д. А. Родионовым и др. С иными подходами к процедуре выбора теоретического распределения вероятностей (механическая, физическая и другие модели) можно познакомиться в работах Дж. Гриффитса, С. И. Романовского, М. И. Толстого. Общим условием использования теоретической плотности распределения для решения указанных выше задач является достаточно хорошая согласованность выбранной плотности распределения с эмпирически наблюдаемым частотным распределением. Следует, однако, помнить, что выборочные данные могут удовлетворять сразу нескольким плотностям распределения вероятностей, в связи с чем принятие того или иного конкретного закона распределения в качестве вероятностной модели изучаемого природного явления почти всегда гипотетично. С другой стороны, имеются определенные трудности в проверке согласованности теоретической плотности распределения, выбранной на основании вневероятностных соображений (например, при физико-химическом анализе явления), и эмпирического частотного распределения. Дело в том, что выборочные данные, как правило, отягощены различного рода методическими погрешностями. Последние могут существенно искажать исходное распределение. Еще одна сложность связана с возможной статистической неоднородностью исследуемого объекта, что может, как и в предыдущем случае, приводить к появлению смешанных распределений.

Распределение совокупности случайных величин — совместное распределение m случайных величин $\xi_1, \xi_2, \ldots, \xi_m$.

Функция распределения m-мерной случайной величины $(\xi_1, \xi_2, \ldots, \xi_m)$ определяет вероятность совместного выполнения m неравенств вида $\xi_j \leqslant x_j$:

$$F(x_1, x_2 \ldots, x_m) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \ldots, \xi_m \leqslant x_m).$$

Для случайных величин дискретного типа $F(x_1, x_2, \ldots, x_m)$ — ступенчатая функция. Для непрерывных величин плотность многомерного распределения определяется следующим образом:

$$f(x_1, x_2, \ldots, x_m) = \partial^m F(x_1, x_2, \ldots, x_m)/(\partial x_1\, \partial x_2, \ldots, \partial x_m).$$

Многомерным аналогом математического ожидания является вектор $(M\xi_1, M\xi_2, \ldots, M\xi_m)$, а дисперсии — ковариационная матрица:

$$\begin{pmatrix} D\xi_1 & \mathrm{Cov}(\xi_1, \xi_2) & \ldots & \mathrm{Cov}(\xi_1, \xi_m) \\ \mathrm{Cov}(\xi_2, \xi_1) & D\xi_2 & \ldots & \mathrm{Cov}(\xi_2, \xi_m) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{Cov}(\xi_m, \xi_1) & \mathrm{Cov}(\xi_m, \xi_2) & \ldots & D\xi_m \end{pmatrix}.$$

Если случайные величины $\xi_1, \xi_2, \ldots, \xi_m$ попарно независимы, то ковариационная матрица имеет вид диагональной матрицы (на главной диагонали — дисперсии, остальные элементы матрицы — нули).

Плотность распределения случайных величин $\xi_1, \xi_2, \ldots, \xi_m$, соответствующая условию $\xi_j = x_j$, $\xi_{j+1} = x_{j+1}, \ldots, \xi_m = x_m$, называется условной плотностью распределения. Она определяется следующим образом:

$$f(x_1, x_2, \ldots, x_{j-1} | x_j, \ldots, x_m) = \\ = f(x_1, x_2, \ldots, x_m)/f_M(x_j, \ldots, x_m,)$$

где $f_M(x_j, \ldots, x_m) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_m)\, dx_1 \ldots dx_{j-1}$.

Для более общего случая, включающего и дискретные случайные величины, удобнее воспользоваться понятием условной функции распределения

$$F(x_1, x_2, \ldots, x_{j-1} | x_j, \ldots, x_m).$$

Для независимых случайных величин $(\xi_1, \xi_2, \ldots, \xi_{j-1})$ и $(\xi_j, \ldots, \xi_m)$ справедливы равенства

$$F_y(x_1, x_2, \ldots, x_{j-1} | x_j, \ldots, x_m) = F_M(x_1, x_2, \ldots, x_{j-1}),$$

$$F(x_1, x_2, \ldots, x_m) = F_M(x_1, x_2, \ldots x_{j-1})\, F_M(x_j, \ldots, x_m).$$

Любые из этих равенств составляют необходимое и достаточное условие для независимости двух множеств случайных величин. Если многомерные случайные величины зависимы, то тесноту линейной связи между ними измеряют с помощью коэффициента корреляции.

Случайные величины $(\xi_1, \xi_2, \ldots, \xi_m)$ взаимно независимы, если выполняется равенство

$$F(x_1, x_2, \ldots, x_m) = F_1(x_1) F_2(x_2) \ldots F_m(x_m).$$

Это же условие в терминах плотностей вероятностей имеет вид

$$f(x_1, x_2, \ldots, x_m) = f_1(x_1) f_2(x_2) \ldots f_m(x_m).$$

Многомерные случайные величины все чаще используются при теоретико-вероятностном описании геологических объектов.

Широкое применение в геологии моделей, порождаемых многомерным пространством, обусловлено, с одной стороны, сложностью таких геологических задач, как расшифровка генезиса месторождений, их прогноз и т. п., а с другой стороны — низкой информативностью отдельно взятых геологических признаков (с точки зрения названных задач). Исследование совокупностей геологических характеристик, трактуемых с вероятностных позиций как m-мерные случайные величины, позволяет повысить надежность и объективность геологических описаний и выводов, получаемых на их основе.

Условное распределение — распределение случайной величины ξ при условии, что другая случайная величина η приняла некоторое определенное значение y. Если ξ и η — непрерывные случайные величины, имеющие совместную плотность распределения, то условные плотности распределения можно вычислить по формулам:

$$f_{\xi|\eta}(x/y) = f(x, y)/f_\eta(y),$$

$$f_{\eta|\xi}(y|x) = f(x, y)/f_\xi(x).$$

Условное математическое ожидание величины ξ при заданном η определяется (для непрерывного случая) следующим образом:

$$\mu_{\xi|\eta} = M|(\xi/\eta = y) = \frac{\int_{-\infty}^{\infty} x f(x, y)\, dx}{\int_{-\infty}^{\infty} f(x, y)\, dx} = \int_{-\infty}^{\infty} x F(x|y)\, dx,$$

где $F(x/y)$ — функция распределения, соответствующая условному распределению величины ξ при условии $\eta = y$.

Условная дисперсия вычисляется по формуле

$$\sigma^2_{\xi|\eta} = D(\xi|\eta = y) = \frac{\int_{-\infty}^{\infty} (x - \mu_{\xi|\eta})^2 f(x, y)\, dx}{\int_{-\infty}^{\infty} f(x, y)\, dx}.$$

Аналогично определяются условные математическое ожидание и дисперсия случайной величины η при заданном ξ.

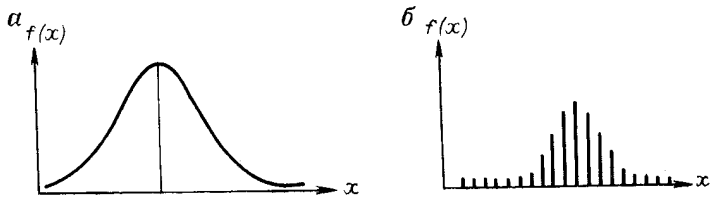


Рис. 3. Примеры непрерывного (*а*) и дискретного (*б*) распределений

Если ξ и η независимы, то из $f(x, y) = f_\xi(x) f_\eta(y)$ следует:

$$\mu_{\xi|\eta} = \mu_\xi \quad \text{и} \quad \mu_{\eta|\xi} = \mu_\eta,$$

т. е. условные математические ожидания совпадают с математическими ожиданиями частных распределений.

Дискретное распределение — распределение дискретной случайной величины ξ, принимающей конечное или счетное число значений $x_1, x_2, \ldots, x_j, \ldots$ с соответствующими вероятностями $p_1, p_2, \ldots, p_j, \ldots$, при этом $\sum_j p_j = 1$.

Таблица вида:

Значения ξ	$\ldots$	x_1	x_2	$\ldots$	x_j	$\ldots$
Вероятности	$\ldots$	p_1	p_2	$\ldots$	p_j	$\ldots$

называется таблицей распределения.

Наиболее часто в приложениях встречаются распределения дискретных случайных величин, принимающих лишь целочисленные значения: биномиальное, полиномиальное. Пуассона, гипергеометрическое, Паскаля и др. (рис. 3). Реализация тех или иных распределений регулируется степенью соответствия условий проведения случайных экспериментов комплексу требований, известному под названием «схема Бернулли». Независимость испытаний в сочетании с постоянством вероятностей наступления событий обеспечивает появление биномиального (если число взаимоисключающих исходов m не более двух) или полиномиального (если $m > 2$) распределений. Сохраняя те же условия, но приближая вероятность p одного из двух взаимоисключающих исходов к нулю и вводя дополнительное условие $np = \lambda$ (λ — некоторая постоянная, n — число наблюдений), получаем распределение Пуассона. Если p достаточно мало, то результаты эксперимента будут удовлетворительно описываться пуассоновским распределением даже в том случае, если требование постоянства вероятностей нарушено. Если же изменяющиеся от испытания к испытанию вероятности появления события имеют сравнительно большие значения, то ожидаемое распределение скорее всего будет близко к обобщенному биномиальному. При ослабленных требованиях независимости можно ожидать появления отрицательного биномиального распределения, а в случае сильной зависимости — распределения Пойа. Таким

образом, процедура подбора вероятностной модели исследуемого явления (если оно может быть описано в терминах дискретных распределений) должна включать в качестве обязательного этапа тщательное и всестороннее изучение условий проведения испытаний (наблюдений).

Математическое ожидание случайной величины ξ — момент первого порядка $M\xi$.

Математическое ожидание можно определить через функцию распределения $F(x)$:

$$M\xi = \int_{-\infty}^{\infty} x dF(x),$$

при этом, если $F(x)$ — ступенчатая функция (т. е. ξ — дискретная величина), то

$$M\xi = \sum_j x_j p(\xi = x_j).$$

Для функций распределения, имеющих плотность $f(x)$, можно записать

$$M\xi = \int_{-\infty}^{\infty} x f(x) dx.$$

Основные свойства математического ожидания:

1) если C — константа, то $MC = C$;

2) каково бы ни было постоянное C,

$$M(C\xi) = CM\xi, \qquad M(C + \xi) = C + M\xi;$$

3) если существуют математические ожидания случайных величин ξ_1 и ξ_2, то справедливо равенство

$$M(\xi_1 + \xi_2) = M\xi_1 + M\xi_2;$$

4) если ξ_1 и ξ_2 — независимые случайные величины, то

$$M(\xi_1 \xi_2) = M\xi_1 M\xi_2.$$

В геологических исследованиях, опирающихся на вероятностные модели, математическое ожидание является важнейшим показателем, характеризующим среднее значение исследуемой геологической характеристики.

Дисперсия — мера «разброса» или рассеивания значений случайной величины ξ относительно математического ожидания M_ξ. Дисперсия может быть определена как центральный момент второго порядка:

$$D\xi = \sigma^2 = M(\xi - M\xi)^2 = M\xi^2 - (M\xi)^2.$$

Величина $\sqrt{D\xi} = \sigma$ называется стандартным (средним квадратическим) отклонением.

Для дискретной случайной величины ξ величина $D\xi$ может быть найдена по формуле

$$D\xi = \sum_j (x_j - \mu)^2 P(\xi = x_j),$$

а для непрерывных случайных величин — по формуле

$$D\xi = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx,$$

где $\mu = M\xi$.

Основные свойства дисперсии:

1) для любой случайной величины

$$D\xi \geqslant 0;$$

2) если C — константа, то

$$DC = 0; \qquad DC\xi = C^2 D\xi; \qquad D(C+\xi) = D\xi;$$

3) если случайные величины ξ_1 и ξ_2 независимы, то

$$D(\xi_1+\xi_2) = D\xi_1 + D\xi_2;$$

4) для произвольных ξ_1 и ξ_2

$$D(\xi_1+\xi_2) = D\xi_1 + D\xi_2 + 2\operatorname{cov}(\xi_1, \xi_2),$$

где $\operatorname{cov}(\xi_1, \xi_2)$ — ковариация,

$$\operatorname{cov}(\xi_1, \xi_2) = M[(\xi_1 - M\xi_1)(\xi_2 - M\xi_2)].$$

В геологических исследованиях, опирающихся на вероятностные модели, дисперсия является основным показателем, характеризующим изменчивость измеряемых свойств природных объектов. Дисперсия вместе с математическим ожиданием служит не только средством более сжатого представления количественной геологической информации, но и имеет важное самостоятельное значение. Дисперсия широко используется, например, в исследованиях, направленных на выяснение некоторых существенных особенностей генезиса геологических объектов. Так, различие в условиях формирования приповерхностных (лавы) и глубинных изверженных пород ведет к резкому увеличению дисперсии содержаний химических элементов в абиссальных породах (главным образом за счет кристаллизационной дифференциации). Если породы к тому же комагматичны, то математические ожидания содержаний основных петрогенных элементов бывают близки; в таких случаях познавательная ценность дисперсии возрастает.

Момент k-го порядка центральный — математическое ожидание $M(\xi - M\xi)^k$ k-й степени центрированной случайной величины ξ.

Для дискретных распределений центральный момент может быть вычислен по формуле

$$M(\xi - M\xi)^k = \sum_i (x_i - \mu)^k p(\xi = x_i),$$

а для непрерывных случайных величин по формуле

$$M(\xi - M\xi)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,dx.$$

Отклонение ξ от ее математического ожидания также является случайной величиной (центрированной). Центральный момент первого порядка равен нулю, центральный момент второго порядка известен под названием дисперсии случайной величины ξ:

$$M(\xi - M\xi)^2 = D\xi.$$

Третий и четвертый центральные моменты служат для характеристики соответственно асимметричности («скошенности») и эксцесса («крутости») распределений случайной величины.

Реже употребляются абсолютные моменты — начальные $M|\xi|^\kappa$ и центральные — $M|\xi - M\xi|^\kappa$. Первый абсолютный центральный момент называют средним арифметическим отклонением.

Центральным моментом порядка k, S системы (ξ, η) называется математическое ожидание произведения соответствующих центрированных величин, взятых в k-й и S-й степени: $M((\xi - M\xi)^\kappa (\eta - M\eta)|^S)$. В случае дискретных величин центральный момент вычисляется по формуле

$$\sum_i \sum_j (x_i - \mu_\xi)^k (y_j - \mu_\eta)^S$$

для непрерывных величин —

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_\xi)^k (y-\mu_\eta)^S = f(x, y)\,dxdy.$$

Центральные моменты, соответствующие $k = 2$, $S = 0$ и $k = 0$, $S = 2$, являются дисперсиями случайных величин ξ и η. Особую роль играет второй смешанный центральный момент ($k = 1$, $S = 1$) — ковариация, широко используемая при исследовании связи между случайными величинами:

$$\operatorname{cov}(\xi, \eta) = M((\xi - M\xi)(\eta - M\eta)).$$

Медиана (Me) — значение случайной величины, для которой справедливы следующие соотношения:

$$P(\xi < \mathrm{Me}) = P(\xi > \mathrm{Me}) = 0{,}5.$$

Таким образом, в результате случайного эксперимента случайная величина ξ может с одинаковой вероятностью либо превысить медианное значение, либо оказаться ниже его. Медианное значение используется как характеристика положения случайной величины на числовой оси.

В случае нормального распределения медиана совпадает со средним значением M_ξ, а в случае логнормального

$$\mathrm{Me} = e^\mu,$$

где $\mu = M \ln \xi$.

Медиана играет важную роль в непараметрической статистике, а также при использовании помехоустойчивых, робастных процедур.

Мода (Mo) — наиболее вероятное значение случайной величины (в дискретных распределениях). Для непрерывных величин модой называется значение, при котором плотность вероятности

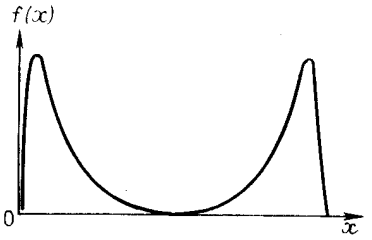


Рис. 4. Положительно и отрицательно асимметричные распределения $y = f(x)$

достигает максимума. Распределения, обладающие одним максимумом, называются одномодальными, если максимумов несколько — полимодальными. Мода, как и математическое ожидание и медиана, является одной из характеристик положения случайной величины на числовой оси (рис. 4). Для симметричных одномодальных распределений эти параметры совпадают. В естественнонаучных приложениях к модальным значениям нередко обращаются при работе с асимметричными распределениями. В геологии полимодальность распределения обычно трактуется как следствие неоднородности исследуемого объекта.

В случае нормального распределения мода совпадает с математическим ожиданием случайной величины, а в случае логнормального

$$Mo = e^{\mu-\sigma^2},$$

где $\mu = M \ln \xi$; $\sigma^2 = D \ln \xi$.

Показатель асимметрии — числовая характеристика степени асимметричности («скошенности») кривой плотности распределения вероятностей случайной величины. Показатель асимметрии определяется через центральный момент третьего порядка, нормированный кубом стандартного отклонения σ:

$$\gamma_1 = \frac{M(\xi - M\xi)^3}{\sigma^3}.$$

В симметричных распределениях моменты нечетного порядка равны нулю, следовательно, γ_1 для таких распределений также имеет нулевое значение. Если $\gamma_1 > 0$, то говорят, что распределение обладает положительной асимметрией, если $\gamma_1 < 0$, то отрицательной асимметрией. В первом случае длинная часть («хвост») кривой плотности расположена справа, а во втором случае — слева от моды. В геологических исследованиях, использующих вероятностные методы, положение асимметричных распределений исследуемых геологических признаков (например, содержаний химических элементов или минералов) нередко пытаются связать с особенностями генезиса природных объектов.

Эксцесс — числовая характеристика кривой плотности вероятностей, отражающая степень ее «крутости», т. е. островершинности или плосковершинности. Для распределений, обладающих чрезмерно (например, по сравнению с кривой плотности нормального закона) острой вершиной, характерна приуроченность подавляющего большинства значения случайной величины ξ к узкой области, примыкающей к моде. Плосковершинные распределения,

наоборот, характеризуются «размазанностью» случайной величины ξ по всему интервалу ее возможных значений. Показатель (коэффициент) эксцесса определяется как центральный момент четвертого порядка, нормированный стандартным отклонением, взятым в четвертой степени.

Эксцесс нормального распределения обычно рассматривается как эталон, с которым сравниваются эксцессы других теоретических распределений. Для нормального распределения величина $M(\xi - M\xi)^4/\sigma^4$ равна трем, поэтому формулу, определяющую значение показателя эксцесса, удобно представить в виде

$$\gamma_2 = \frac{M(\xi - M\xi)^4}{\sigma^4} - 3.$$

Таким образом, коэффициент эксцесса в условиях нормального распределения принимает нулевое значение.

РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Случайные величины могут описываться непрерывной или дискретной функцией распределения (рис. 5). В геологии дискретными законами распределения характеризуются частоты встречаемости акцессорных минералов изверженных горных пород (закон Пуассона); распределение запасов полезных ископаемых подчиняется другому дискретному (в частности, отрицательному биномиальному) распределению.

Распределения петрогенных и редких элементов в изверженных горных породах и минералах часто согласуются с нормальным, логнормальным и другими непрерывными законами.

Ниже приводится краткое описание наиболее распространенных распределений. Более подробное описание можно найти в руководствах [2, 6, 19].

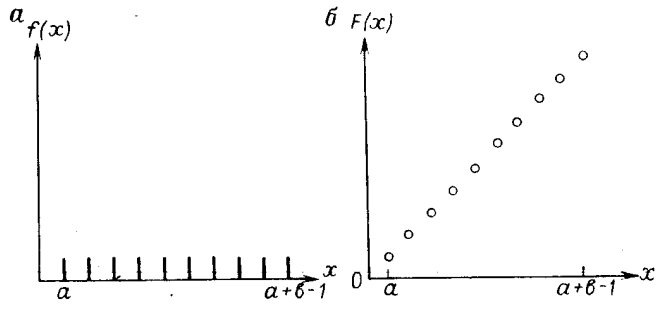


Рис. 5. Функции вероятности $f(x)$ и распределения $F(x)$ для дискретной равномерной случайной величины:
$а$ — $f(x) = 1/b$; $б$ — $F(x) = (x - a + 1)/b$

Дискретные распределения

Схема Бернулли — схема последовательно проводимых взаимно независимых испытаний (наблюдений), в каждом из которых с вероятностью p может появиться событие A. Независимость испытаний означает, что результат каждого из них (появление или непоявление некоторого события A) никак не зависит от результатов предыдущих или последующих наблюдений. Определяющим условием схемы Бернулли является также равновероятность появления события A в каждом из наблюдений: $p_j(A) = p$, где j — номер испытания. Принято называть осуществление события A «успехом» (часто обозначается единицей), а дополняющее его событие $(\overline{A})$ — «неудачей» (обозначается нулем).

Если вероятность p задана, а также зафиксировано число независимых испытаний n, то вероятность k появлений (в данной серии испытаний) события A может быть найдена по формуле Бернулли

$$p_{(k)} = C_n^k p^k (1-p)^{n-k},$$

где C_n^k — число сочетаний из n по k

$$C_n^k = \frac{n(n-1)\ldots(n-k+1)}{1\cdot 2\cdot \ldots \cdot k}.$$

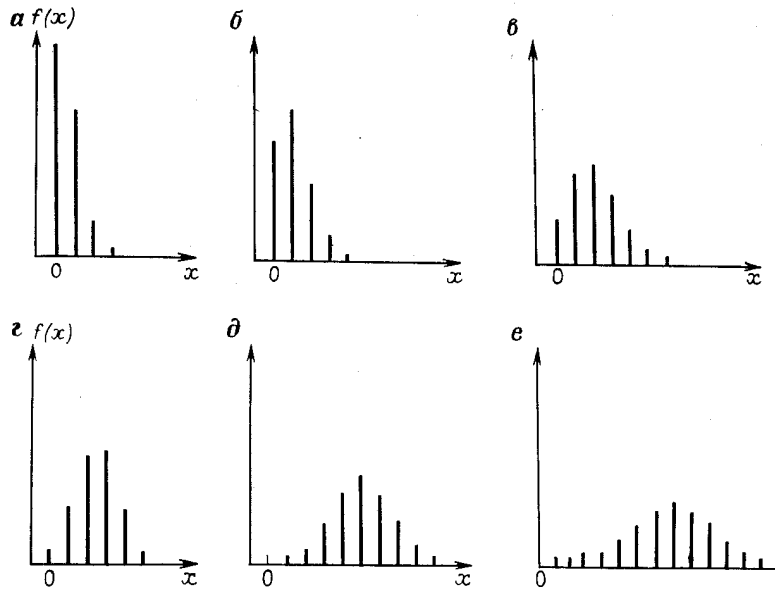


Рис. 6. Функция вероятности $f(x)$ для биномиальной случайной величины. Для распределения скарновых месторождений относительно контакта изверженных и осадочных пород, для распределения подсечения рудных тел заданных размеров и формы разведочными выработками:
$а$ — $n = 5$, $p = 0{,}1$; $б$ — $n = 10$, $p = 0{,}1$; $в$ — $n = 20$, $p = 0{,}1$; $г$ — $n = 5$, $p = 0{,}5$; $д$ — $n = 10$, $p = 0{,}5$; $е$ — $n = 20$, $p = 0{,}5$

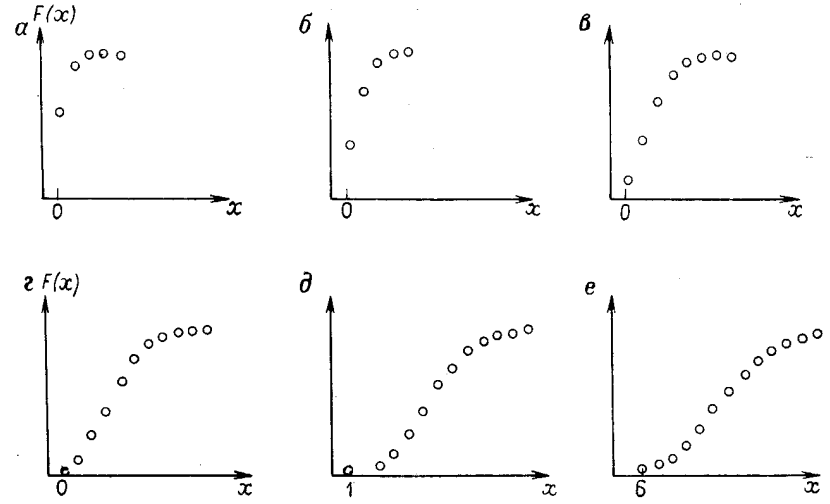


Рис. 7. Функции распределения биномиальной случайной величины

$$F(x) = \sum_{k=0}^{x} C_n^k p^k (1-p)^{n-k};$$

$а$ — $n = 5$, $p = 0{,}1$; $б$ — $n = 10$, $p = 0{,}1$; $в$ — $n = 20$, $p = 0{,}1$; $г$ — $n = 5$, $p = 0{,}5$; $д$ — $n = 10$, $p = 0{,}5$; $е$ — $n = 20$, $p = 0{,}5$

Если ξ — случайная величина, принимающая значения в пределах от 0 до n, где n — фиксированное число наблюдений, то распределение случайной величины ξ имеет следующий вид:

$$p(\xi \leqslant x) = F(x) = \begin{cases} 0; & x < 0 \\ \sum\limits^{x} C_n^k p^k (1-p)^{n-k}; & x \leqslant k \leqslant x+1. \\ 1; & x \geqslant n \end{cases}$$

Биномиальное распределение определяется двумя параметрами p и n, причем математическое ожидание и дисперсия равны соответственно (рис. 6, 7):

$$M\xi = np; \qquad D\xi = np(1-p).$$

Сложность вычислений по формуле Бернулли, а также труднодоступность достаточно полных таблиц биномиального распределения обусловливает обращение к асимптотическим формулам. Среди последних назовем формулу, обеспечивающую нормальное приближение:

$$F(x) \approx \Phi\left(x - 0{,}5 - np/\sqrt{np(1-p)}\right),$$

где Φ — функция стандартного нормального распределения.

Следующее приближение, называемое пуассоновским, рекомендуется применять, если $p < 0{,}1$. Точность приближения возрастает,

если при $n \to \infty$ вероятность $p \to 0$ таким образом, что $np \to \lambda$. Тогда для любого фиксированного k при $n \to \infty$

$$p(\xi = k) \to \frac{\lambda_k}{k!} e^{-\lambda}.$$

По мнению ряда исследователей, можно ожидать реализации биномиального распределения в таких геологических ситуациях, которые обеспечивают как постоянство вероятности появления изучаемого свойства, так и независимость во времени и пространстве результатов наблюдений. Так, биномиальному закону не противоречит распределение числа зерен тяжелых минералов в осадочных породах, если содержания этих минералов достаточно высокие. Удовлетворительно описываются биномиальным распределением частота встречаемости определенных типов контактов между минеральными зернами, а также частота определенных классов плотности и пористости осадочных пород.

Отрицательное биномиальное распределение порождается последовательностью испытаний Бернулли $\xi_1, \xi_2, \ldots$; оно описывает распределение случайной величины $\nu_{\bar{p}}(k)$ следующим образом:

$$\sum_{i=1}^{\nu_{\bar{p}}(k)-1} \xi_i = k-1; \qquad \sum_{i=1}^{\nu_{\bar{p}}(k)} \xi_i = k.$$

Случайная величина $\nu_{\bar{p}}(k)$ есть число испытаний в схеме Бернулли (с вероятностью p появления интересующего нас события в результате проведения одного испытания) до k-го появления ожидаемого события (включая последнее испытание).

Функция распределения случайной величины $\nu_{\bar{p}}(k)$ имеет вид (рис. 8)

$$P\{\nu_{\bar{p}}(k) = x\} = C_{x-1}^{x-k} p^k (1-p)^{x-k}, \qquad x = k, \quad k+1, \ \ldots \ .$$

Название закона объясняется совпадением правых частей последнего выражения с последовательными членами разложения бинома с отрицательным показателем $(pt)^k\,[1-(1-pt)]^{-k}$.

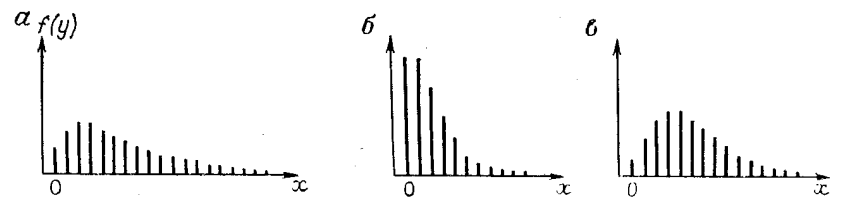


Рис. 8. Функции вероятности $f(x)$ для отрицательной биномиальной случайной величины. Для распределения количества кислых вулканических пород и залежей сульфидов вулканической природы, для распределения всех открытых и неоткрытых месторождений хорошо изученного региона:
$а$ — $x = 2$, $p = 0{,}25$; $б$ — $x = 2$, $p = 0{,}5$; $в$ — $x = 5$, $p = 0{,}5$

Основные параметры закона имеют вид:

среднее $E_{\nu_{\bar{p}}}(k) = k/p$;

дисперсия $D_{\nu_{\bar{p}}}(k) = k(1-p)/p^2$;

асимметрия $\beta_1 = (2-p)/\sqrt{k(1-p)}$;

эксцесс $\beta_2 = [1+4(1-p)+(1-p)^2]/[k(1-p)]+3$.

Известный исследователь Ф. Агтерберг использовал отрицательное биномиальное распределение для описания распределения месторождений и рудопроявлений определенного генезиса в конкретных регионах Канады.

Обобщенное биномиальное распределение — распределение дискретной случайной величины ξ, значения которой соответствуют числу появления k благоприятных исходов («успехов») в серии из n независимых испытаний, производимых в неодинаковых условиях. Последнее обстоятельство обусловливает изменение от испытания к испытанию вероятности «успеха» p_j. Случайная величина ξ имеет обобщенное биномиальное распределение с параметрами $n, p_1, p_2, \ldots, p_n$, если:

а) $k = 0$; $p(\xi = k) = \prod_{j=1}^{n} (1-p_j)$;

б) $k = 1, 2, \ldots, n-1$;

$$p(\xi = k) = p_1 p_2 \ldots p_k (1-p_{k+1}) \ldots (1-p_n) + \ldots + (1-p_1)(1-p_2) \ldots (1-p_{n-k}) p_{n-k+1} \ldots p_n;$$

в) $k = n$; $\quad p(\xi = k) = \prod_{j=1}^{n} p_j$.

Моменты обобщенного биномиального распределения:

$$M\xi = \sum_{k=1}^{n} p_k; \qquad D\xi = \sum_{k=1}^{n} p_k (1-p_k).$$

При переходе от биномиального распределения к обобщенному биномиальному распределению резко возрастают вычислительные трудности при нахождении $p(\xi = k)$. Если $k \neq 0$ и $k \neq n$, то искомая вероятность равна сумме всех возможных произведений, в которые сомножители «p» с разными индексами входят k раз, а сомножители «$1-p$» (также с разными индексами при p) входят $n-k$ раз. С целью упорядочения вычислительных процедур рекомендуется воспользоваться так называемой производящей функцией $\varphi(z)$:

$$\varphi(z) = (1-p_1+p_1 z)(1-p_2+p_2 z) \ldots (1-p_n+p_n z) = \prod_{j=1}^{n} (1-p_j+p_j z),$$

где z — произвольный параметр.

Разложение функции $\varphi(z)$ по степеням параметра z обеспечивает нахождение вероятности $p(\xi = k)$ как коэффициентов при k-й степени z:

$$\prod_{j=1}^{n}(1-p_j+p_jz)=\sum_{k=0}^{n}p(\xi=k)z^k.$$

В геологии нередки ситуации, в которых вероятности появления исследуемого признака (трактуемого как некоторое событие) неодинаковы для различных точек изучаемого объекта. Так, при исследовании распределения числа зерен определенного минерала в пробах, отобранных из различных слоев стратифицированного объекта, требование биномиальной модели о равенстве вероятностей будет скорее всего нарушено в связи с генетической неоднородностью объекта. В этом случае, если допущение о биномиальном (или пауссоновском) распределении числа зерен минерала в пределах отдельных слоев остается все еще в силе, уместно обратиться к обобщенному биномиальному распределению. Дж. Гриффитс [15] предлагает при обработке результатов наблюдений в рамках обобщенного биномиального распределения различать данные опробования, соответствующие отбору одного образца из каждого слоя (модель, или серия, Пуассона), и данные, относящиеся к многократному опробованию каждого слоя (модель или серия Лексиана). Полученная в последнем случае совокупность проб слагается из подсовокупностей, сформированных из образцов, взятых в пределах отдельных слоев.

Распределение Пуассона — это дискретное однопараметрическое распределение случайной величины ξ, определенное по формуле

$$p(\xi=r)=(\lambda^r/r!)\exp(-\lambda).$$

Параметр λ всегда положителен (рис. 9, 10).

Распределение Пуассона может быть получено преобразованием с помощью предельного перехода при $n \to \infty$ дискретного биномиального распределения.

Распределение Пуассона применяется в том случае, когда имеют дело с числом появлений некоторого события при большом числе наблюдений и при малой вероятности наступления этого события в каждом отдельном наблюдении.

Основные характеристики распределения: среднее $M(\xi) = \lambda$, дисперсия $D(\xi) = \lambda$.

Распределение Пуассона встречается при изучении размещения акцессорных минералов в изверженных горных породах.

Гипергеометрическое распределение — это распределение случайной величины $\nu_{MN}(m)$, равной числу объектов, обладающих определенным свойством среди m объектов, случайно извлеченных (без возвращения) из совокупности N объек-

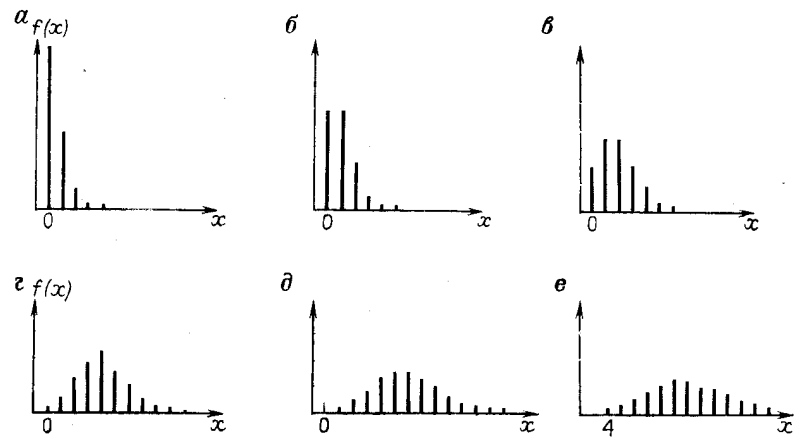


Рис. 9. Функции вероятности $f(x)$ для пуассоновской случайной величины

$$f(x)=\frac{\lambda^x e^{-\lambda}}{x!};$$

а — $\lambda = 1/2$; *б* — $\lambda = 1$; *в* — $\lambda = 2$; *г* — $\lambda = 4$; *д* — $\lambda = 5$; *е* — $\lambda = 10$

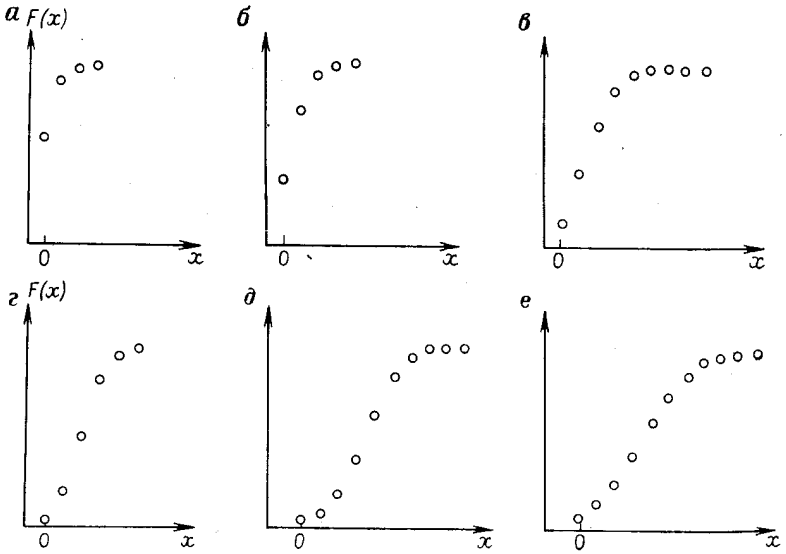


Рис. 10. Функции распределения $F(x)$ пуассоновской случайной величины

$$F(x)=\sum_{i=0}^{x}\frac{\lambda^i e^{-\lambda}}{i!}.$$

Для распределения числа нефтегазоносных ловушек среди множества изучаемых структур, для распределения появления 0, 1, 2, . . . ионов данного компонента или для распределения числа структурных элементов при исследовании шлифов:

а — $\lambda = 1/2$; *б* — $\lambda = 1$; *в* — $\lambda = 2$; *г* — $\lambda = 4$; *д* — $\lambda = 5$; *е* — $\lambda = 10$

тов, M из которых обладают этим свойством. Функция распределения $\nu_{MN}(m)$ имеет вид

$$P\{\nu_{MN}(m)=x\}=C_M^x C_{N-M}^{m-x}/C_N^m.$$

Основные параметры закона имеют вид:

$$\text{среднее}\quad E\nu_{MN}(m)=m(M/N);$$

$$\text{дисперсия}\quad D\nu_{MN}(m)=m(M/N-1)[(1-(M/N)][(1-(m/N)];$$

$$\text{асимметрия}\quad \beta_1=[1-2(M/N)]/\sqrt{[m(M/N)][1-(M/N)]}\times$$
$$\times[(N-2m)\sqrt{N-1}\,]/[(N-2)\sqrt{N-m}\,];$$

$$\text{эксцесс}\quad \beta_2=[C_1(N)-C_2(N)\,6(M/N)(1-M/N)]/m(M/N)\times$$
$$\times(1-M/N)+C_3(N)+C_4(N)],$$

где
$$C_1(N)=(N-1)N(N+1)/[(N-2)(N-3)(N-m)];$$
$$C_2(N)=(N-1)N^2/[(N-2)(N-3)(N-m)];$$
$$C_3(N)=3\{(N-1)N^2/[(N-2)(N-3)(N-m)]\}-1;$$
$$C_4(N)=\{18(N-1)/[(N-2)(N-3)]\}-$$
$$-\{6(N-1)/[(N-2)(N-3)m(M/N)(1-M/N)]\}-$$
$$-\{3(N-1)Nm/[(N-2)(N-3)(N-m)]\}.$$

При $N\to\infty$ распределение $\nu_{MN}(m)$ сводится к биномиальному закону.

Геологические области применимости гипергеометрического распределения близки к тем, где привлекаются модели Бернулли.

Распределение Пойа — распределение дискретной случайной величины ξ, значения которой могут быть получены в следующем эксперименте. В урне находится N шаров, из них Np белых ($0<p<1$) и $N(1-p)$ черных. Наугад вынимают шар и после определения цвета возвращают в урну вместе с S новыми шарами того же цвета. Число вынимания белого, например, шара в серии из n проб может рассматриваться как случайная величина ξ. Очевидно, случайный эксперимент, организованный подобным образом, выходит за рамки схемы Бернулли (нарушены требования независимости и постоянства вероятностей). Распределение вероятностей случайной величины ξ имеет вид

$$P(\xi=K)=C_n^k[b(b+S)]\ldots[b+(k-1)S]C(C+S)\ldots[C-$$
$$-(n-k-1)S]/[N(N+S)]\ldots[N+(n-1)S],$$
$$\text{где}\quad b=Np,\quad C=N(1-p).$$

Математическое ожидание и дисперсия определяются следующим образом:

$$M\xi=np;\qquad D\xi=np(1-p)(1+nS/N)/(1+S/N).$$

Если число испытаний n мало по сравнению с N, то допустимо биномиальное приближение:

$$P(\xi=k)\simeq C_n^k p^k(1-p)^{n-k}.$$

Распределение Пойа может оказаться подходящей моделью числа зерен минералов, характеризующихся крайне неравномерным распределением в исследуемом геологическом объекте (например, тяжелые минералы, образующие значительные локальные концентрации в результате естественного обогащения при намыве).

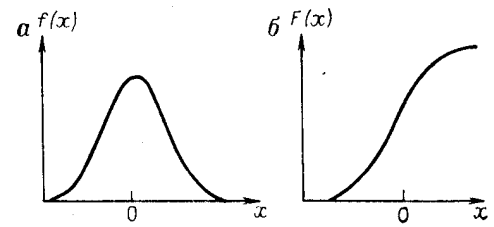


Рис. 11. Плотность вероятности $f(x)$ и функция распределения $F(x)$ стандартной нормальной случайной величины. Для распределения содержаний полезных компонентов в месторождениях калийных солей, бокситовых и железорудных месторождениях, для распределения коэффициента интенсивности как индикаторного отношения в сечении на поверхности, для распределения полей яркости при анализе аэроландшафтов:

$$а-f(x)=\frac{1}{\sqrt{2\pi}}\exp(-x^2/2);$$

$$б-F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp(-u^2/2)\,du$$

Непрерывные распределения

Нормальное распределение — распределение случайной величины ξ, характеризующееся плотностью вероятностей типа

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ и σ^2 — параметры распределения, соответствующие математическому ожиданию и дисперсии случайной величины ξ:

$$M\xi=\mu;\qquad D\xi=\sigma^2.$$

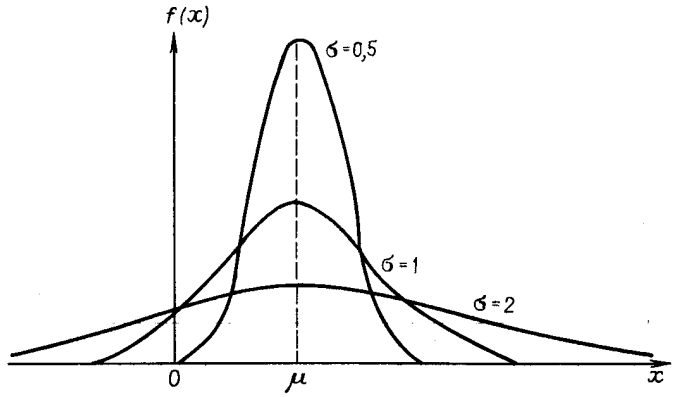


Рис. 12. Плотность вероятности $f(x)$ нормального распределения для различных значений σ.

Параметры μ и σ^2 однозначно характеризуют положение и форму кривой распределения (рис. 11, 12).

Функция нормального распределения определяется следующим образом:

$$P(\xi \leqslant x) = F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}}\, dv.$$

Заменив ξ ее центрированным и нормированным аналогом $\xi' = \frac{\xi - \mu}{\sigma}$, получим функцию стандартного нормального распределения Φ с параметрами $\mu' = 0$ и $\sigma' = 1$:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{\tau^2}{2}}\, d\tau,$$

где $t = \frac{x - \mu}{\sigma}$.

Значения этой функции, а также функции

$$\Phi^*(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-\frac{\tau^2}{2}}\, dt,$$

связанной с $\Phi(t)$ соотношением $\Phi(t) = \frac{1}{2} + \Phi^*(t)$, табулированы и легкодоступны.

Из симметричности нормального распределения с параметрами $\mu = 0$ и $\sigma = 1$ следует $\Phi(-t) = 1 - \Phi(t)$.

Приведем некоторые свойства нормального распределения.

1. Пусть $\eta = a + \xi$ — случайная величина, где a — константа, ξ — случайная величина, распределенная нормально с параметрами μ и σ^2. Тогда случайная величина η также распределена нормально с параметрами

$$\mu_\eta = \mu + a \quad \text{и} \quad D_\eta = \sigma^2.$$

2. Пусть $\eta = b\xi$, где b — константа, а случайная величина ξ, как и в предыдущем случае, распределена нормально с параметрами μ и σ^2. Тогда случайная величина η также распределена нормально с параметрами $M\eta = b\mu$ и $D\eta = b^2\sigma^2$.

3. Если $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины, распределенные нормально с параметрами $\mu_1, \mu_2, \ldots, \mu_n$ и $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, то распределение $\eta = \sum_{i=1}^{n} \xi_i$ нормально с параметрами:

$$M\eta = \sum_{i=1}^{n} \mu_i; \qquad D\eta = \sum_{i=1}^{n} \sigma_i^2.$$

4. Ксли $\xi_1, \xi_2, \ldots, \xi_n$ — зависимые нормально распределенные величины, то их сумма η также нормальна, а параметры ее распределения определяются следующим образом:

$$M\eta = \sum_{i=1}^{n} \mu_i; \qquad D\eta = \sum_{i=1}^{n} \sigma_i^2 + \sum_{i<j} \rho_{ij}\sigma_i\sigma_j; \qquad ij = 1, 2, \ldots, n,$$

где ρ_{ij} — коэффициент корреляции.

При достаточно широких предположениях распределение суммы случайных величин с ростом числа слагаемых очень быстро приближается к нормальному закону.

Важность нормального распределения в естественнонаучных приложениях определяется тем, что распределения значений многих (но не всех) количественно измеряемых свойств природных объектов вполне удовлетворительно аппроксимируются нормальным законом. В связи с этим нормальный закон часто принимается в качестве вероятностной модели исследуемого явления, что, вообще говоря, может привести к ошибочным выводам, так как согласие результатов наблюдений с тем или иным законом распределения отнюдь не доказывает единственность именно этой модели. С содержательных позиций, особенно если решаются задачи генетического плана, наиболее ценны такие вероятностные модели (распределения), которые выбираются с учетом теоретических предпосылок, характеризующих физическую природу изучаемого явления.

В геологии стало традицией проверять согласованность выборочных распределений с нормальным законом. Многочисленные исследования, проведенные в этом направлении, показали, что существенное уклонение от нормального закона встречается чаще, чем это предполагалось ранее. Наиболее вероятная причина анормальности распределения геологических характеристик заключается, скорее всего, в невыполнении требований центральной предельной теоремы — равномерной малости и независимости факторов, генерирующих исследуемую случайную величину.

Многомерное нормальное распределение с вектором математического ожидания $a = (a_1, \ldots, a_m)$ и ковариационной матрицей Σ — распределение многомерной случайной величины $\xi = (\xi_1, \ldots, \xi_m)$ с плотностью вида

$$p(x_1, \ldots, x_m) = (2\pi)^{-m/2} |\Sigma|^{1/2} \exp\{(X-a)\Sigma^{-1}(X-a)'\},$$

где Σ^{-1} — матрица, обратная ковариационной матрице Σ.

Примером аппроксимации многомерным нормальным распределением может служить распределение петрогенных компонент по данным силикатного анализа проб изверженных горных пород.

Логарифмически-нормальное (логнормальное) распределение — распределение случайной величины ξ, логарифм которой распределен по нормальному закону. Пусть η — случайная величина, распределенная нормально с параметрами (μ_η, σ^2) и $\eta = \ln \xi$. Тогда распределение величины η — логарифмически-нормально с плотностью вероятностей (рис. 13, 14)

$$p(x) = \begin{cases} 0; & x \leqslant 0, \\ \dfrac{1}{x\sigma_\eta\sqrt{2\pi}} e^{-(\ln x - \mu_\eta)^2/2\sigma_\eta^2}; & x > 0. \end{cases}$$

Здесь μ_η и σ_η^2 — параметры распределения, однако в отличие от нормального распределения μ_η и σ_η^2 в условиях логарифмически-

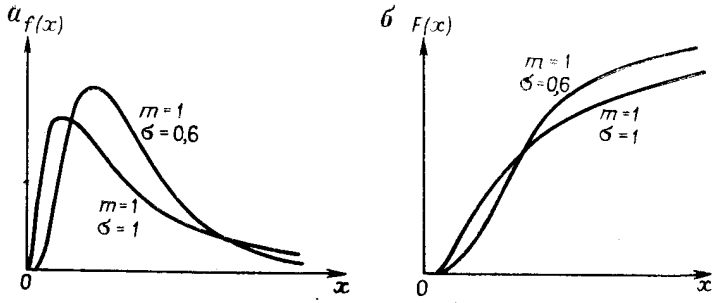


Рис. 13. Плотность вероятности $f(x)$ и функции распределения $F(x)$ логнормальной случайной величины. Для распределения частот размеров зерен в наблюдаемых образцах естественных осадков, для распределения содержаний полезных компонентов в рудах многих месторождений:

$$а - f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{\frac{-[\log(x-m)]^2}{2\sigma^2}\right\};$$

$$б - F(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_0^x \frac{1}{u}\exp\left\{\frac{-[\log(u-m)]^2}{2\sigma^2}\right\}du$$

нормального распределения не являются параметрами, характеризующими соответственно центр и масштаб кривой плотности вероятностей случайной величины ξ.

Математическое ожидание и дисперсия случайной величины ξ связаны с параметрами μ_η и σ_η^2 соотношениями

$$M\xi = e^{\mu_\eta + \sigma_\eta^2/2};$$

$$D\xi = e^{\sigma_\eta^2 + 2\mu_\eta}\left(e^{\sigma_\eta^2} - 1\right) = (M\xi)^2\left(e^{\sigma_\eta^2} - 1\right),$$

из которых следует, что в логарифмически-нормальном распределении математическое ожидание и дисперсия зависимы.

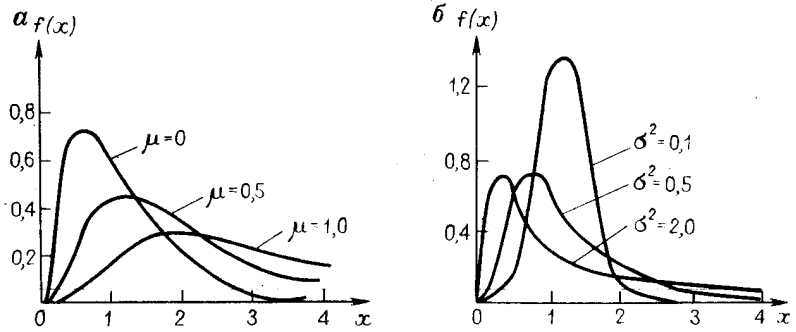


Рис. 14. Плотность вероятности $f(x)$ логнормального распределения для различных значений μ и σ:
$а$ — $\mu = 0$, $\mu = 0{,}5$, $\mu = 1{,}0$; $б$ — $\sigma^2 = 0{,}1$, $\sigma^2 = 0{,}5$, $\sigma^2 = 2$.

Структура вышеприведенных выражений не изменится, если воспользоваться не натуральными логарифмами (ln), а любыми другими.

Математическая модель, ведущая к возникновению логнормального распределения, опирается на центральную предельную теорему, распространенную на случай зависимых величин. Благодаря исследованиям Кептейна, С. Н. Бернштейна, Г. Крамера, Дж. Ачисона, Дж. Брауна и других, установлено, что логнормальный закон описывает распределения такой случайной величины ξ, которая может рассматриваться как предел последовательности случайных величин:

$$x_{i+1} = x_i + z_{i+1}h(x_i), \qquad i = 0,\ 1,\ 2,\ \ldots$$

Каждая из компонент x_{i+1} есть результат действия достаточно малых независимых импульсов $Z_1, Z_2, \ldots, Z_n$; при этом

$$Z_1 + Z_2 + \ldots + Z_n = \sum_{i=0}^{n-1}[(x_{i+1} - x_i)/h(x_i)] \simeq \int_{x_0}^{x}[dx/h(x)] = g(x).$$

Функция $h(x)$ подбирается таким образом, чтобы функция $g(x)$ была монотонной в заданном интервале изменения от $-\infty$ до ∞. Если $h(x) = \text{const} = C$, то $g(x) = (x - x_0)/C$, и в результате получаем нормальное распределение. Однако если $h(x) = x$, т. е. существует пропорциональность между действием импульса и достигнутым к этому времени значением случайной величины, то $g(x) = \ln x_n - \ln x_0$, и распределение случайной величины ξ логарифмически-нормально.

Логнормальное распределение широко применяется в самых различных областях естествознания. А. Н. Колмогоров показал, что логарифмически-нормальному распределению подчинены размеры частиц, образующихся при дроблении; на основании этой схемы предпринимались попытки привлечь логнормальный закон для описания гранулометрических характеристик кластических осадочных пород. Большой объем работ был также проделан геологами по проверке согласия с логнормальным законом выборочных распределений содержаний редких и малых элементов в породах различного генезиса. Хотя во многих случаях аппроксимация логнормальным распределением вполне удовлетворительна, надлежащее теоретическое обоснование, ведущее именно к логнормальной модели, почти всегда отсутствует. Это не позволяет исследователю дать генетическую интерпретацию полученных таким путем вероятностных моделей. Тем не менее если ограничиться задачей подбора достаточно приемлемой аппроксимации наблюдаемых в результате опыта асимметричных распределений (например, для более сжатого представления обширной выборочной информации или с целью более обоснованного применения некоторых статистических критериев и т. п.), то логнормальный закон нередко оказывается вполне подходящей моделью.

Д. А. Родионов расширил аппроксимирующие возможности логнормального распределения, предложив использовать при подборе кривой плотности вероятностей так называемое «трехпараметрическое распределение». Пусть η — положительная непрерывная случайная величина $\eta = a + \xi$, где a — константа, ξ — случайная величина, распределенная логнормально (обозначим это специальным символом Λ) с параметрами μ и σ^2. Тогда $p(\xi \leqslant x) = \Lambda(x; \mu, \sigma^2)$ и

$$p(\eta \leqslant y) = p[(\eta - a) \leqslant (y - a)] = p[\xi \leqslant (y - a)] = \Lambda(y - a; \mu, \sigma^2).$$

Плотность вероятностей случайной величины η имеет вид

$$f(y) = \frac{1}{(y-a)\,\sigma\sqrt{2\pi}} e^{-[\ln(y-a)-\mu]^2/2\sigma^2}; \qquad y > a;$$

$$M\eta = ae^{(\sigma^2+2\mu)/2} = a + M\xi; \qquad D\eta = e^{\sigma^2+2\mu}(e^{\sigma^2} - 1) = D\xi.$$

Распределение типа $\Lambda(g - a; \mu, \sigma^2)$ применялось при описании распределений таких химических элементов, содержания которых можно рассматривать как сумму константы и логнормальной компоненты.

Положив $\eta = a - \xi$, получаем распределение типа $1 - \lambda(a - y; \mu, \sigma^2)$. Кривая плотности вероятностей случайной величины $\eta = a - \xi$ является зеркальным отражением логнормальной кривой распределения случайной величины ξ от вертикали, проходящей через точку $y = a$. Распределение $1 - \lambda(a - y; \mu, \sigma^2)$ в отличие от классического логнормального распределения характеризуется отрицательной асимметрией.

Это распределение используется в геохимии при описании распределений тех химических элементов, концентрации которых можно представить в виде разности константы и логнормальной компоненты.

Распределение Стьюдента (t-распределение) — распределение случайной величины $\tau = \xi/\eta$, где

$$\eta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\xi_i^2},$$

случайные величины ξ и $\xi_1, \xi_2, \ldots, \xi_n$ независимы и нормально распределены с параметрами $\mu = \mu_1 = \mu_2 = \ldots = \mu_n = 0$; $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$.

Плотность распределения вероятностей случайной величины τ определяется формулой (рис. 15)

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \qquad -\infty < t < \infty,$$

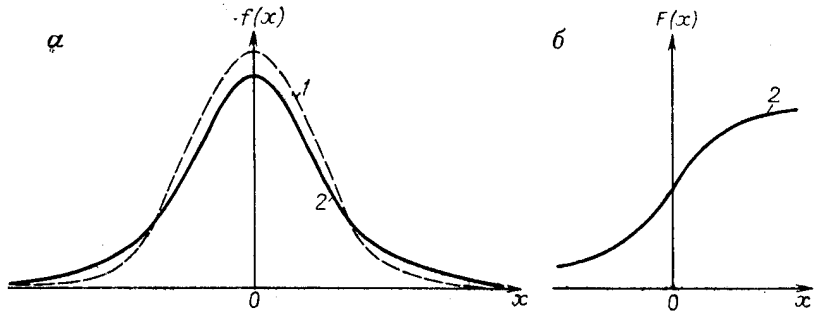


Рис. 15. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по закону Стьюдента. Для распределения химических элементов по диаметру атомов:
а — $f(x)$, б — $F(x)$. 1 — распределение Стьюдента; 2 — нормальное распределение

где n — число степеней свободы, а $\Gamma(x)$ — гамма-функция, равная

$$\int_0^\infty t^{x-1}e^{-t}dt.$$

В частности,

$$\Gamma(n+1) = n!, \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(x+1) = x\Gamma(x).$$

Распределение Стьюдента унимодально и симметрично относительно $x = 0$. Моменты τ-распределения:

$$M\tau^{2k-1} = 0; \qquad M\tau^{2k} = \frac{n^k\Gamma\left(\frac{n}{2} - k\right)\Gamma\left(k + \frac{1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}, \quad 2k < n;$$

$$D\tau = \begin{cases} \infty, & \text{если } n \leqslant 2 \\ n/(n-2), & \text{если } n > 2. \end{cases}$$

При $n = 1$ t-распределение есть не что иное, как распределение Коши. С ростом числа степеней свободы распределение Стьюдента приближается к стандартному нормальному распределению. Распределение Стьюдента табулировано в форме, наиболее удобной для статистических приложений.

Распределение Стьюдента — одно из наиболее важных специальных распределений, введенное в теорию вероятностей и статистику В. Госсетом, публиковавшим свои работы под псевдонимом Стьюдент.

Распределение Стьюдента широко используется в геологии при проверке гипотез о средних значениях геологических характеристик. Квантили распределения Стьюдента используются при построении доверительных интервалов геологоразведочных параметров.

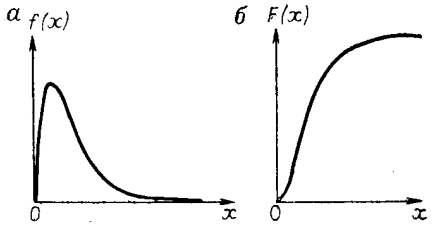


Рис. 16. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по закону Фишера. Для распределения показателей гранулометрического состава эоловых песков:
$а$ — $f(x)$; $б$ — $F(x)$

Распределение Фишера (F = распределение) — распределение случайной величины $\eta = \frac{\xi_1}{\xi_2}\frac{m_2}{m_1}$, где ξ_1 и ξ_2 — независимые случайные величины, имеющие χ^2-распределения соответственно с m_1 и m_2 степенями свободы.

Плотность вероятностей случайной величины η, имеющей распределение Фишера (рис. 16):

$$0; \quad x \leqslant 0;$$

$$f(x) = \frac{\Gamma\left(\frac{m_1+m_2}{2}\right) m_1^{m_1/2} m_2^{m_2/2} x^{m_1/2-1}}{\Gamma(m_1/2)\,\Gamma(m_2/2)\,(m_2+m_1x)^{(m_1+m_2)/2}}; \quad x>0,$$

где $\Gamma(\cdot)$ — гамма-функция.

Математическое ожидание и дисперсия случайной величины η:

$$M\eta = m_2/(m_2-2), \qquad m_2>2,$$

$$D\eta = 2m_2^2(m_1+m_2-2)/[m_1(m_2-2)^2(m_2-4)], \quad m_2>4.$$

Следует отметить, что отношение оценок дисперсии двух случайных величин, подчиняющихся нормальному закону, описывается F-распределением. Это обстоятельство определяет его широкое применение в дисперсионном анализе.

Если η — случайная величина, имеющая F-распределение с m_1 и m_2 степенями свободы, то $\xi = \frac{1}{2}\ln\eta$ имеет Z-распределение. Это распределение, называемое также распределением дисперсионного отношения, было предложено Фишером. Плотность z-распределения

$$f(Z) = 2m_1^{m_1/2} m_2^{m_2/2} \frac{\Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma(m_1/2)\,\Gamma(m_2/2)} \frac{e^{m_1Z}}{(m_2+m_1e^{2Z})^{(m_1+m_2)/2}};$$

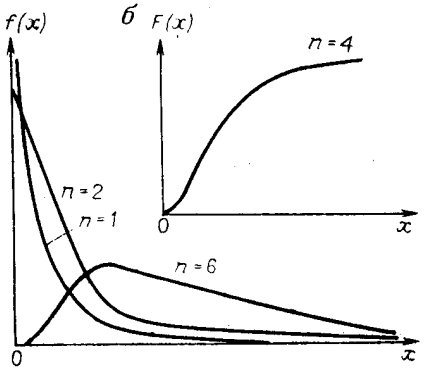
$$-\infty<Z<\infty.$$

Математическое ожидание и дисперсия случайной величины η, имеющей распределение Фишера, равны

$$M\eta = 0; \qquad D\eta = (m_1+m_2)/2(m_1m_2).$$

Распределение Фишера широко используется в геологии при проверке некоторых статистических гипотез, например при изуче-

Рис. 17. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по закону χ^2. Для распределения количества вторичного рудоносного цемента, развивающегося в пространстве между зернами в песчаниках или известняках:
$а$ — $f(x)$; $б$ — $F(x) = \{1/[2^{v/2}\Gamma(v/2)\}\int_0^x u^{(v-2)/2}\exp(-u/2)\,du$



нии степени рассеяния вокруг средних содержаний элементов и минералов в породах.

Распределение χ^2 (хи-квадрат) — распределение случайной величины χ^2, плотность вероятностей которого описывается формулой (рис. 17)

$$f(x) = \begin{cases} 0; & x \leqslant 0, \\ x^{(n-2)/2}\exp^{(-x/2)}/[2^{n/2}\Gamma(n/2)]; & 0<x<\infty, \end{cases}$$

где $\Gamma(\cdot)$ — гамма-функция; n — число степеней свободы (целое число).

Математическое ожидание и дисперсию случайной величины χ^2 можно записать:

$$M\chi^2 = n; \qquad D\chi^2 = 2n.$$

Широкое применение χ^2-распределения в теории вероятностей и математической статистике определяется тем обстоятельством, что случайную величину χ^2 можно представить как сумму квадратов независимых случайных величин, имеющих одно и то же стандартное нормальное распределение ($M\xi_i = 0$; $D\xi_i = 1$; $i = 1, \ldots, n$):

$$\chi^2 = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2 = \sum_{i=1}^{n}\xi_i^2.$$

Если $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины, имеющие нормальные распределения с параметрами $\mu_1 = M\xi_1$, $\mu_2 = M\xi_2, \ldots, \mu_n = M\xi_n$; $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$, то распределение случайной величины $\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\xi_n^2$ называется нецентральным χ^2-распределением с параметром нецентральности

$$m = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu_i^2.$$

С ростом числа степеней свободы n нецентральное χ^2-распределение приближается к нормальному распределению с параметрами

$$(n+m); \qquad \sqrt{2(n+2m)}\,.$$

Распределение χ^2 табулировано и легкодоступно.

Многочисленные статистические критерии проверки гипотез о таких параметрах, как средние содержания рудных компонент, степень изменчивости, коррелируемости изучаемых геологических свойств и др., сводятся к использованию центрального и нецентрального χ^2-распределений.

Распределение χ (хи-распределение) — распределение случайной величины

$$\chi=\sqrt{\xi_1^2+\xi_2^2+\ldots+\xi_n^2}\,,$$

где $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины, имеющие стандартное нормальное распределение (т. е. $M\xi_i=0$; $D\xi_i=1$; $i=1, 2, \ldots, n$).

Плотность вероятностей χ-распределения

$$f(x)=\begin{cases}0; & x\leqslant 0,\\ \dfrac{1}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)}\,x^{n-1}e^{-x^2/2}; & 0<x<\infty,\end{cases}$$

где $\Gamma(\cdot)$ — гамма-функция, n — число степеней свободы.

При $n=1$ распределение χ называется распределением Рэлея, при $n=2$ — распределением Максвелла (распределение скоростей молекул газа).

Математическое ожидание и дисперсия χ-распределения:

$$M\chi=\frac{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}; \qquad D\chi=n-2\left[\frac{\Gamma\left(\frac{n+1}{2}\right)^2}{\Gamma\left(\frac{n}{2}\right)}\right].$$

В геологических (инженерно-гидрогеологических) исследованиях χ-распределение находит ограниченное применение.

Распределение Рэлея — распределение случайной величины

$$\xi=\sqrt{\xi_1^2+\xi_2^2}\,,$$ где ξ_1, ξ_2 — независимые случайные величины, распределенные нормально с параметрами $\mu_1=\mu_2=0$ и $\sigma_1^2=\sigma_2^2=\sigma^2$.

Его плотность

$$f(x)=\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}; \qquad x\geqslant 0, \quad \sigma>0;$$

во всех остальных случаях математическое ожидание и дисперсия:

$$M\xi=\sigma\sqrt{\pi/2}\,; \qquad D\xi=0{,}429\sigma^2.$$

С. И. Романовский, исследуя кинематику водных потоков, вывел, что функция распределения размеров частиц донной популяции (т. е. частиц, которые при определенной скорости потока перемещаются волочением по дну либо находятся в покое) близка к плотности распределения Релея.

Распределение Лапласа, или двустороннее экспоненциальное распределение,— распределение, плотность вероятности которого имеет вид (рис. 18)

$$f(x)=\frac{1}{2}\lambda e^{-\lambda|x|};$$

$$-\infty<x<\infty.$$

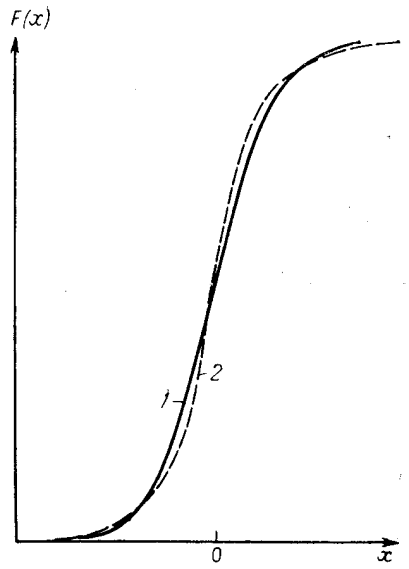


Рис. 18. Функции распределения Лапласа и нормальное:
1 — нормальное распределение; *2* — распределение Лапласа

Основные характеристики распределения Лапласа: среднее, мода, медиана, асимметрия — равны нулю, дисперсия $D\xi=2/\lambda^2$, эксцесс $\beta_2=3$.

Распределение Лапласа используется при обработке геохимических данных, при подсчете запасов месторождений полезных ископаемых.

Равномерное распределение — распределение случайной величины на конечном интервале (a, b) с плотностью $1/(b-a)$, равное нулю вне этого интервала. Функция плотности распределения такой случайной величины имеет вид (рис. 19):

$$f(x)=\begin{cases}0; & x<a,\\ 1/(b-a); & a\leqslant x\leqslant b,\\ 0; & x>b.\end{cases}$$

Равномерное распределение находит ограниченное применение при решении некоторых петрологических задач осадочного генезиса,

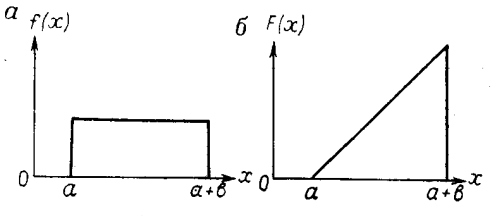


Рис. 19. Плотность вероятности $f(x)$ и функции распределения $F(x)$ для равномерной случайной величины. Для распределения количества кислорода в элементарной ячейке кристаллической решетки:
$а$ — $f(x)=1/b$; $a\leqslant x\leqslant(a+b-1)$;
$б$ — $F(x)=(x-a+1)\,b$; $a\leqslant x<(a+b-1)$

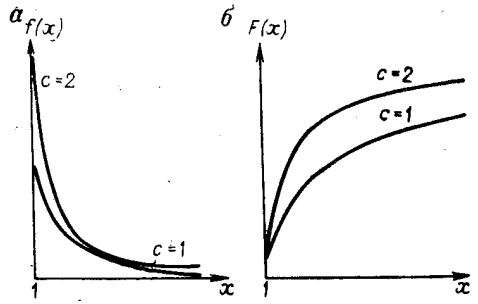


Рис. 20. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по закону Парето. Для распределения месторождений полезных ископаемых, имеющих годовую прибыль свыше 1 млн. руб. за определенный период времени:

$а - f(x) = cx^{-c-1}$; $б - F(x) = 1 - x^{-c}$; $c = 1,2$

а также при моделировании геологоразведочных свойств угольных, фосфоритовых и некоторых других типов месторождений.

Распределение Парето описывает неполную генеральную совокупность, усеченную, т. е. такую, из которой изъяты все элементы с признаком, превышающим некоторый заданный уровень c или не превышающим его.

Функция распределения случайной величины, распределенной по закону Парето, имеет вид (рис. 20)

$$F(x) = 1 - (c/x)^{\alpha},$$

а плотность вероятности

$$f_{\xi}(x) = (\alpha/c)\,(c/x)^{\alpha+1},$$

где $\alpha > 0$, $x > c$.

Числовые характеристики распределения Парето:

среднее $M\xi = \alpha c/(\alpha - 1)$ при $\alpha > 1$;

мода $x_{\text{mod}} = c$;

медиана $x_{\text{med}} = 2^{1/\alpha}c$;

дисперсия $D\xi = \alpha c^2/[(\alpha - 1)(\alpha - 2)]$ при $\alpha > 2$;

момент k-го порядка $M\xi^k = \alpha c^k/(\alpha - k)$ при $\alpha > k$.

Распределение Парето изредка используется при изучении распределения прогнозных минеральных ресурсов в рудных полях, узлах, регионах.

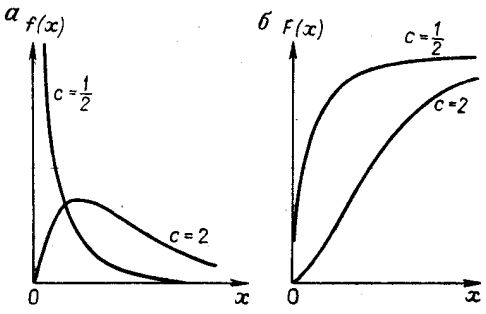


Рис. 21. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по гамма-закону.

Для распределения содержания кислых магматических пород:

$$а - f(x) = \frac{x^{c-1}\exp(-x)}{\Gamma(c)};$$

$$б - F(x) = \int_0^x \frac{u^{c-1}\exp(-u)}{\Gamma(u)}\,du$$

Гамма-распределение. Двухпараметрический закон распределения случайной величины $\gamma(b, a)$ описывается плотностью (рис. 21)

$$f(x) = \begin{cases} b^a x^{a-1} e^{-bx}/\Gamma(a); & 0 \leqslant x < \infty, \\ 0; & x = 0, \end{cases}$$

где $\Gamma(a) = \int_0^{\infty} x^{a-1}e^{-x}dx$ — гамма-функция; $a > 0$ — параметр формы; $b > 0$ — параметр масштаба.

Положив $a = m/2$, $b = 1/2$, получим x^2 (m)-распределение. Сумма независимых гамма-распределенных случайных величин (с равными параметрами масштаба b) также распределена по гамма-закону с параметрами $a_1 + a_2 + \ldots + ub$.

Важнейшие параметры гамма-распределения равны:

среднее $M\gamma(a, b) = a/b$;

мода $x_{\text{mod}} = \frac{a-1}{b}$, при $a \geqslant 1$;

дисперсия $D\gamma(a, b) = a/b^2$;

асимметрия $\beta_1 = 2/\sqrt{a}$;

эксцесс $\beta_2 = 6/a$.

Гамма-распределение нередко используется при описании зависимостей распределений содержаний рудных компонент от их запасов в геологоразведочном деле.

Бета-распределение — двухпараметрический закон распределения случайной величины $\beta(v, w)$, $0 < v < \infty$, $0 < w < \infty$, описываемый плотностью (рис. 22)

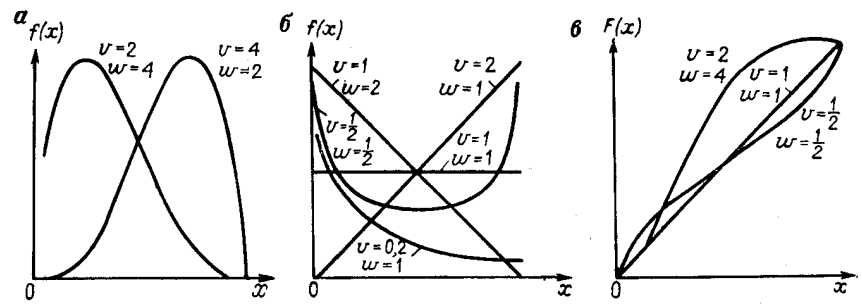


Рис. 22. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по бета-закону. Для распределения рассеянного химического элемента в кристаллической решетке изучаемого акцессорного минерала:

$$а, б - f(x) = \frac{x^{v-1}(1-x)^{w-1}}{B(v, w)}; \quad в - F(x) = \int_0^x \frac{u^{v-1}(1-u)^{w-1}}{B(v, w)}\,du$$

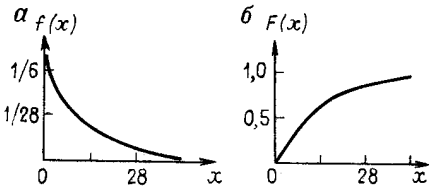


Рис. 23. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по экспоненциальному закону. Для распределения определенного числа лабораторных химических анализов по определению содержания компонента по времени, затраченному на их выполнение:
$а - f(x) = (1/b)\exp(-x/b) = \lambda\exp(-\lambda x)$;
$б - F(x) = 1 - \exp(-x/b)$

$$f(x) = \begin{cases} \Gamma(v+w)\,x^{v-1}(1-x)^{w-1}/[\Gamma(v)\,\Gamma(w)] & \text{для } x \leqslant 1, \\ 0 & \text{для остальных } x, \end{cases}$$

где $\Gamma(\cdot)$ — гамма-функция.

Если $\gamma(v, b)$ и $\gamma(w, b)$ — независимые гамма-распределенные случайные величины, то $\beta(v, w) = \gamma(v, b)/[\gamma(v,b)+\gamma(w, b)]$ имеет бета-распределение.

Экспоненциальное распределение (показательное распределение) — распределение случайной величины ξ, плотность вероятности которого описывается формулой (рис. 23)

$$f(x) = \begin{cases} 0, & x<0; \\ \beta e^{-\beta x}; & x \geqslant 0, \end{cases}$$

где β — параметр распределения.

Это распределение широко применяется в теории надежности при описании распределения времени безотказной работы системы.

Математическое ожидание и дисперсия экспоненциального распределения равны:

$$M\xi = 1/\beta; \qquad D\xi = 1/\beta^2.$$

С. И. Романовский показал, что в условиях, когда осаждение частиц происходит под действием главным образом силы тяжести, а подвижные факторы (течения, волнения и т. п.) не оказывают сколько-нибудь заметного влияния на седиментационный процесс, распределения максимального размера частиц и мощностей слоев будут близки к экспоненциальному закону. По его мнению, удовлетворительное согласие с экспоненциальным распределением будет наблюдаться для гранулометрических индексов тех осадочных образований, накопление которых происходило как бы из единой порции осадка, поступившего в зону аккумуляции.

Распределение Вейбулла порождается механизмом, характеризующим длительность жизни сложной системы. Пусть ξ — время жизни системы, $\lambda(t)$ — интенсивность отказа системы ко времени t. Если через $f_\xi(t)$ обозначить функцию плотности вероятности случайной величины ξ, а через $F_\xi(t)$ — ее функцию распределения, то будем иметь:

$$F_\xi(t) = -1 - e^{\int_0^t \lambda(\tau)\,d\tau}, \qquad \lambda(t) = -\frac{f_\xi(t)}{1 - F_\xi(t)}.$$

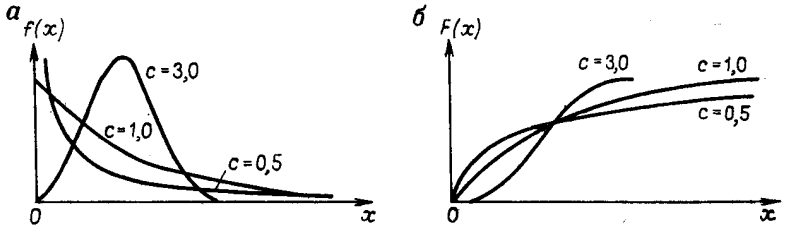


Рис. 24. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по закону Вейбулла. Для распределения размеров околожильных ореолов отдельных химических элементов:
$а - f(x) = cx^{c-1}\exp(-x^c)$; $б - F(x) = 1 - \exp(-x^c)$

Экспериментально установлено, что кривая $\lambda(t)$ обычно характеризуется следующим образом: сначала она сильно убывает, затем на некотором отрезке изменения аргумента остается постоянной и, наконец, начинает сильно возрастать. Первый отрезок изменения аргумента принято называть периодом приработки, второй — периодом нормальной эксплуатации, третий — периодом старения и износа. Такой характер поведения может быть описан функцией $\lambda(t) = \lambda_0 t^{\alpha-1}$, где $\lambda_0 > 0$ и $\alpha > 0$ — параметры, причем значения $\alpha < 1$, $\alpha = 1$ и $\alpha > 1$ соответствуют поведению функции интенсивности отказов в периоды приработки, нормальной эксплуатации и старения.

Функция распределения случайной величины ξ, подчиняющейся закону Вейбулла, имеет вид (рис. 24)

$$F_\xi(t) = 1 - e^{-\lambda_0 t^\alpha}, \qquad t \geqslant 0,$$

а плотность вероятности равна

$$f_\xi(t) = \lambda_0 \alpha t^{\alpha-1} e^{-\lambda_0 t^\alpha}, \qquad t \geqslant 0.$$

Основные числовые характеристики распределения:

$$\text{среднее } E\xi = \lambda_0^{-1/\alpha}\Gamma\left(1 + \frac{1}{\alpha}\right);$$

$$\text{мода } x_{\text{mod}} = \begin{cases} 0, & \text{если } \alpha > 1, \\ \lambda_0^{-1/\alpha}\left(1 - \frac{1}{\alpha}\right)^{1/\alpha}, & \text{если } \alpha > 1; \end{cases}$$

$$\text{дисперсия } D\xi = \lambda_0^{-2/\alpha}\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right];$$

$$\text{момент } k\text{-го порядка } m_k = E\xi^k = \lambda_0^{-k/\alpha}\Gamma\left(1 + \frac{k}{\alpha}\right).$$

Здесь $\Gamma(\cdot)$ — гамма-функция.

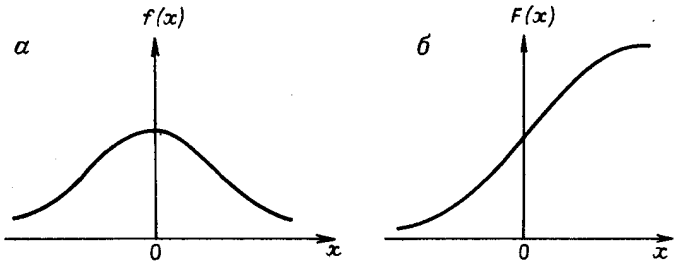


Рис. 25. Плотность вероятности $f(x)$ и функции распределения $F(x)$ случайной величины, распределенной по логистическому закону. Для распределения зональности ореолов, свойственной любым редкометалльным пегматитам:
$а$ — $f(x) = \exp(x)/[1 + \exp(x)]^2$; $б$ — $F(x) = 1/[1 + \exp(-x)]$

Логистическое распределение — распределение случайной величины ξ, задаваемое плотностью вероятностей вида (рис. 25)

$$f(x) = \frac{\pi e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\sigma\sqrt{3}\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)^2}\right]},$$

где μ и σ — параметры распределения, соответствующие математическому ожиданию и дисперсии случайной величины ξ:

$$M\xi = \mu; \qquad D\xi = \sigma^2.$$

Функция логистического распределения близка к нормальной функции.

Распределение Коши — распределение случайной величины ξ, плотность вероятностей которой имеет вид (рис. 26)

$$f(x) = \frac{1}{\pi}\frac{\lambda}{\lambda^2 + (x-\mu)^2}; \qquad -\infty < x < \infty,$$

где λ и μ — параметры распределения ($\lambda > 0$); точка $x = \mu$ является модой и медианой распределения. Распределение Коши симметрично относительно μ, а моменты случайных величин — бесконечны. Отношение ξ_1/ξ_2 независимых случайных величин, каждая из которых распределена нормально с параметрами $M\xi_1 = M\xi_2 = 0$ и $D\xi_1 = D\xi_2 = 1$, имеет распределение вероятностей, совпадающее с распределением Коши. В этом случае параметр $\lambda = 1$ и плотность вероятностей приобретает вид

$$f(x) = 1/[\pi(x^2 + 1)].$$

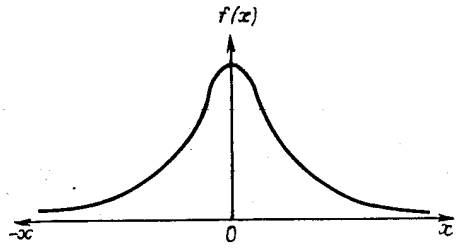


Рис. 26. Плотность вероятности распределения Коши $f(x) = 1/\pi(x^2 + 1)$. Для распределения отношений содержаний рудных и редких элементов в различных разрезах месторождения

Распределения, задаваемые кривыми Пирсона [9]. Плотность вероятностей $y = f(x)$, график которой принадлежит семейству кривых Пирсона, является решением дифференциального уравнения (рис. 27)

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x + c_1}{c_0 + c_1 x + c_2 x^2},$$

где началом отсчета для x, служит среднее значение. Вид решения зависит от постоянных величин c_0, c_1 и c_2, которые связаны простыми соотношениями с моментами соответствующего распределения вероятностей:

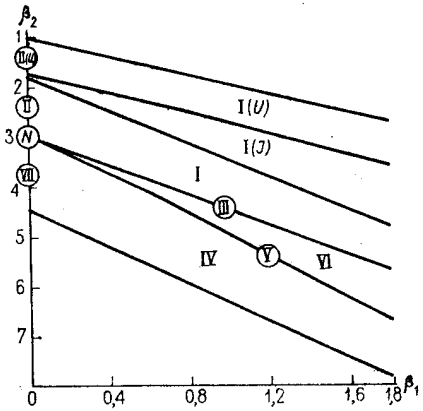


Рис. 27. Графики для определения типа кривой Пирсона в зависимости от β_1 и β_2

$$c_0 = \sigma^2(4\beta_2 - 3\beta_1)/[2(5\beta_2 - 6\beta_1 - 9)],$$

$$c_1 = \sigma\sqrt{\beta_1(\beta_2 + 3)}/[2(5\beta_2 - 6\beta_1 - 9)],$$

$$c_2 = (2\beta_2 - 3\beta_1 - 6)/[2(5\beta_2 - 6\beta_1 - 9)],$$

где $\sigma^2 = \mu^2$; $\beta_1 = \mu_3^2/\mu_2^3$; $\beta_2 = \mu_4/\mu_2^2$,

$$\mu_r = \int_{l_1}^{l_2} x^r f(x)\,dx, \qquad r = 2, 3, 4, \ldots (\mu_0 = 1; \quad \mu_1 = 0).$$

Величины l_1 и l_2 являются нижней и верхней границами естественной области определения плотности $f(x)$.

Таким образом, если вдоль осей прямоугольной системы координат условиться откладывать отрезки, отвечающие величинам β_2 и β_1, то в плоскости $\beta_2 0 \beta_1$ различным типам кривых Пирсона будут соответствовать области, кривые и точки. На рис. 27 указана такая разбивка плоскости $\beta_2 0 \beta_1$ для основных типов кривых Пирсона I—VII. Прямая линия с уравнением $\beta_2 - \beta_1 - 1 = 0$ представляет собой верхнюю границу для допустимых точек (β_2, β_1), так как не существует распределений, для которых $\beta_2 - \beta_1 - 1 < 0$. Кроме того, если кривая принадлежит семейству Пирсона, причем $8\beta_2 - 15\beta_1 - 36 \geqslant 0$, то $\mu_8 = \infty$. На рис. 27 прямая с уравнением $8\beta_2 - 15\beta_1 - 36 = 0$ служит нижней границей точек с координатами (β_2, β_1).

В табл. 1 приведены уравнения кривых, принадлежащих указанным семи типам семейства Пирсона.

Таблица 1

Уравнения кривых Пирсона

Тип	Уравнение	Начало отсчета для x	Область определения
I	$y = y_0\left(1+\frac{x}{a_2}\right)^{m_1}\left(1+\frac{x}{a_2}\right)^{m_2}$	Мода	$-a_1<x<a_2$
II	$y = y_0\left(1-\frac{x^2}{a_2}\right)^{m}$	Мода (среднее)	$-a<x<a$
III	$y = y_0 e^{-\gamma x}\left(1+\frac{x}{a}\right)^{\gamma a}$	Мода	$-a<x<\infty$
IV	$y = y_0 e^{-\gamma \operatorname{arctg}\frac{x}{a}}\left(1-\frac{x_2}{a_2}\right)$	Среднее $+\frac{\gamma a}{2m\cdot 2}$	$-\infty<x<\infty$
V	$y = y_0 e^{-\gamma/x}\, x^{-p}$	Начало кривой	$0<x<\infty$
VI	$y = y_0(x-a)^{q} - x^{-q_1}$	Точка, отстоящая на a^1 от начала кривой	$a<x<\infty$
VII	$y = y_0\left(1+\frac{x^2}{a_3}\right)^{-m}$	Среднее (мода)	$-\infty<x<\infty$

Распределения, задаваемые кривыми Пирсона, используются в геологических исследованиях при описании распределений химических элементов и минералов в горных породах, в частности при изучении химизма базальтов.

Таким образом, кривые третьего типа являются переходными между кривыми первого и шестого типов, кривые пятого типа — переходными между кривыми шестого и четвертого типов, кривые второго типа являются предельным классом кривых первого типа, а кривые седьмого типа — предельным классом кривых четвертого типа.

Практически тип кривой Пирсона может быть установлен также с помощью критерия $\varkappa$. Для этого вычисляют значение $\varkappa$:

$$\varkappa = -\frac{m_3^2(\Lambda+2)^2}{16(\Lambda+1)}, \qquad \Lambda = \frac{6(m_4-m_2^2-1)}{3m_3^2-2m_4+6},$$

где m_3 и m_4 — третий и четвертый центральные моменты распределений, и в соответствии с его величиной и знаком принимают решение о принадлежности кривой распределения к тому или иному типу кривых Пирсона:

$$\varkappa \begin{cases} <0 & \to \text{I};\\ =0, \quad m_4>3 & \to \text{II};\\ =0, \quad m_4\geqslant 3 & \to \text{VII};\\ 0<\varkappa<1 & \to \text{IV};\\ =1 & \to \text{V};\\ 1<\varkappa<\infty & \to \text{VI};\\ =\pm\infty & \to \text{III}. \end{cases}$$

ОСНОВНЫЕ ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Центральная предельная теорема — термин, объединяющий большое число теорем, суть которых сводится к следующему утверждению: при выполнении определенных условий (они конкретизируются в названных выше теоремах) функция распределения суммы случайных величин с ростом числа слагаемых сходится к нормальному закону. Группа центральных предельных теорем известна в различных формах, различающихся прежде всего требованиями, которые предъявляются к суммируемым случайным величинам и их распределениям.

Исторически первые формулировки идей центральной предельной теоремы связаны с именами Муавра (частный случай для вероятности ($P = 1/2$) и Лапласа (более общая форма $0 < P < 1$), показавших, что при $n \to \infty$ имеет место:

$$P\left(\frac{m-np}{\sqrt{np(p-1)}}<x\right)\to \Phi(x),$$

Где n — число испытаний, удовлетворяющих требованиям схемы Бернулли; m — число появления события A, P — вероятность наступления в одном испытании события A, $\Phi(x)$ — функция нормального распределения (функция Лапласа).

Теорема Муавра-Лапласа справедлива, если величина p не слишком близка к 0 или 1.

Для непрерывных случайных величин условия применимости центральной предельной теоремы сформулированы А. М. Ляпуновым. Пусть $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины, для каждой из которых существуют математические ожидания $M\xi_1, M\xi_2 \ldots, M\xi_n$, дисперсии $D\xi_1, D\xi_2, \ldots, D\xi_n$ и третий центральный абсолютный момент $M|\xi_1-M\xi_1|^3$, $M|\xi_2-M\xi_2|^3, \ldots,$ $M|\xi_n-M\xi_n|^3$. Положим

$$\mu_k = M\xi_k, \quad \sigma_k^2 = D\xi_k; \quad \alpha_k = M|\xi_k - M\xi_k|^3.$$

Если при $n \to \infty$ выполняется условие Ляпунова

$$\sum_{k=1}^{n} \alpha_k / \sum_{k=1}^{n} \sigma_k^3,$$

то $$P\left(\frac{\xi_1 + \xi_2 + \ldots + \xi_n - \sum_{k=1}^{n} \mu_k}{\sum_{k=1}^{n} \sigma_k} < x \right) \to \Phi(x).$$

Левая часть неравенства представляет собой так называемую нормированную величину $\xi_н$ с параметрами распределения $M\xi_н = 0$ и $D\xi_н = 1$.

Если суммируемые случайные величины имеют одинаковые распределения, т. е.

$$M\xi_1 = M\xi_2 = \ldots = M\xi_n,$$

$$D\xi_1 = D\xi_2 = \ldots = D\xi_n,$$

то для выполнения сходимости к нормальному закону достаточно потребовать взаимной независимости $\xi_1, \xi_2, \ldots, \xi_n$. Для разнораспределенных величин условие Ляпунова обеспечивает равномерно малый вклад каждой случайной величины в суммарную дисперсию. При нарушении условия «равномерной малости» закон распределения суммы $\xi_1, \xi_2, \ldots, \xi_n$ будет определяться распределением той случайной величины, изменчивость которой максимальна.

Наиболее общая форма записывается следующим образом:

$$p\left(A \leqslant \frac{\xi_1 + \xi_2 + \ldots + \xi_n - \sum_{k=1}^{n} \mu_k}{\sum_{k=1}^{n} \sigma_k} \leqslant B \right) \to \Phi(B) - \Phi(A).$$

Универсальность центральной предельной теоремы, т. е. ее применимость к суммированию случайных величин с любыми распределениями, а также устойчивость центральной предельной теоремы при нарушениях требования независимости (допускается «умеренная» зависимость $\xi_1, \xi_2, \ldots, \xi_n$) определяют относительно широкую распространенность нормального закона. В естественно-научных дисциплинах модель нормального распределения приемлема в тех случаях, когда есть основания предполагать, что формирование объекта происходило под влиянием достаточно большого числа более или менее равноценных по своей силе факторов. Однако в геологии не всегда можно ожидать «равномерной малости» вышеназванных факторов, что ведет к нарушению условий применимости центральной предельной теоремы и, как следствие, к заметному уклонению от нормального закона.

Удобной формой центральной предельной теоремы является правило Бэрри, используемое при описании геологических свойств рассматриваемых объектов прогнозирования, поисков и разведки месторождений полезных ископаемых. Оно представляет собой [39, с. 23]

$$\left| P\left\{ \frac{1}{B_n} \sum_{i=1}^{n} (\xi_i - \mu_i) < x \right\} - \Phi(x) \right| \leqslant C \frac{k}{B_n},$$

где $\Phi(x)$ — стандартная нормальная функция распределения; C — константа, например, $C = 7{,}5$; P — вероятность события, указанная в фигурных скобках.

Теорема Бернулли — теорема теории вероятностей, связывающая частоту m появления событий A в серии из n наблюдений с вероятностью этого события. В современном изложении теорема формулируется следующим образом: при неограниченном увеличении числа независимых наблюдений частость (m/n) события A сходится по вероятности к его вероятности, которая постоянна во всех испытаниях и равна P:

$$P\left(\left| \frac{m}{n} - p \right| > \varepsilon \right) \xrightarrow{p} 0 \text{ для любого } \varepsilon > 0, \text{ при } n \to \infty.$$

Теорема Бернулли — одна из основных предельных теорем, выражающая действие закона больших чисел в условиях схемы Бернулли, отражает весьма важное с познавательной и методологической точек зрения свойство случайных явлений, а именно — статистическую устойчивость частот.

Теорема Пуассона — обобщение теоремы Бернулли на случай, когда вероятность появления события A в серии последовательных независимых испытаний меняется от наблюдения к наблюдению: $p_1, p_2, \ldots, p_n$. Если символом p обозначить среднее арифметическое вероятностей наступления события в отдельных наблюдениях, то теорему Пуассона можно записать так:

$$p\left(\left| \frac{m}{n} - p \right| > \varepsilon \right) \xrightarrow{p} 0, \quad n \to \infty,$$

где m — число появлений события; n — число наблюдений; $\varepsilon > 0$; символ $\xrightarrow{p}$ — сходимость по вероятности.

Утверждение о сходимости по вероятности частоты появления события к его вероятности в достаточно длинном ряду независимых испытаний составляет методологическую основу практических приложений теории вероятностей в естественных науках вообще и в геологии в частности.

ГЛАВА 2

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Математическая статистика — раздел математики, объектом которого является получение надежных выводов из статистических данных и выработка методов, с помощью которых эти выводы могут быть получены. Основная задача математической статистики — на основании одной или нескольких выборок сделать вывод о всем содержимом той совокупности, из которой производится выбор. В геологических исследованиях наиболее широко используются следующие разделы математической статистики: оценка параметров распределений; проверка гипотез о некоторых хорошо изученных законах распределения, в первую очередь о нормальном законе построения доверительных интервалов для различных статистических характеристик; проверка статистических гипотез о равенстве или различии параметров распределения либо других статистических характеристик в двух или большем числе изучаемых совокупностей. При прогнозировании геологических характеристик широко используется аппарат корреляционного и регрессионного анализа.

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ, ВЫБОРКИ

Генеральная совокупность — математическая абстракция, используемая в математической статистике для описания совокупности объектов, которые подвергаются обследованиям с помощью случайного выбора ее представителей.

Рассмотрим случайный эксперимент. Дано некоторое множество, содержащее конечное число элементов. Эксперимент заключается в том, что мы наугад выбираем какой-нибудь элемент этого множества, регистрируем какую-либо его характеристику и затем возвращаем его назад. Предполагается, что вероятности быть выбранными равны для всех элементов. Заданное множество — генеральная совокупность, а процесс выбора, описанный выше, называется простым случайным выбором.

Если мы интересуемся значениями некоторой случайной величины ξ, то генеральная совокупность — это множество ее значений, а последовательность наблюденных значений $x_1, x_2, \ldots, x_n$ — случайная выборка из этой генеральной совокупности.

Генеральная совокупность может быть конечной или бесконечной.

Выборка — множество наблюдаемых значений одномерной или многомерной случайной величины с некоторой функцией распределения. Рассмотрим последовательность n наблюденных значений $x_1, \ldots, x_n$ одномерной случайной величины ξ с функцией распределения $F(x)$. Распределение выборки будет определяться

как дискретное распределение точек $x_1, \ldots, x_n$ на вещественной оси, в каждой из которых помещены массы, равные $1/n$. Соответствующая выборочная функция распределения $F^*(x)$ является ступенчатой функцией со скачками высоты $1/n$ в каждой точке x_i. Если обозначить через ν число выборочных значений, не превосходящих x, то будем иметь

$$F^*(x) = \nu/n,$$

т. е. $F^*(x)$ представляет собой частоту появления события $\xi < x$ в последовательности n наблюдений.

Для выборочного распределения $F^*(x)$ можно определить различные выборочные характеристики, например моменты выборки

$$M^\nu\xi = \frac{1}{n}\sum_{i=1}^{n} x_i^\nu.$$

Аналогично определяется выборка значений многомерной случайной величины $\xi = (\xi_1, \ldots, \xi_p)$. Пусть мы имеем n выборочных значений $(x_1^{(1)}, \ldots, x_p^{(1)}), \ldots, (x_1^{(n)}, \ldots, x_p^{(n)})$ случайной величины ξ. Эта выборка может быть представлена как множество точек в p-мерном пространстве. Распределение выборки получим, поместив в каждой из этих n точек массу, равную $1/n$. Для этого распределения можно вычислить моменты и другие выборочные характеристики по общим правилам.

Частоты распределения — выборочные характеристики распределения, являющиеся статистическим аналогом плотности вероятности распределения.

Статистические данные состоят обычно из измеряемых значений непрерывного или дискретного признака. В последовательности случайных экспериментов невозможно предсказать исход каждого результата измерения. В них обнаруживаются случайные колебания, не поддающиеся точному учету. Однако, если рассмотреть последовательность измерений в целом, оказывается, что несмотря на правильное поведение индивидуальных результатов, средние результаты достаточно длинной последовательности случайных экспериментов обнаруживают весьма заметную устойчивость.

Чтобы пояснить это явление, рассмотрим некоторый случайный эксперимент, который может быть многократно повторен при одинаковых условиях. Пусть S — множество всех возможных исходов эксперимента, а S_0 — некоторое его подмножество. Допустим, что в некотором эксперименте произошло событие $\xi \in S_0$, обозначим его E, т. е. E — событие, состоящее в том, что $\xi \in S_0$. Будем повторять эксперимент большое число раз, например n раз, Если в последовательности n испытаний событие E произойдет ν раз, то отношение ν/n будем называть частотой события E в последовательности n испытаний. Наблюдая частоты ν/n события E для возрастающих значений n, мы обнаружим, что эта частота стремится к постоянному значению.

В этом состоит явление статистической устойчивости, служащее основой математической статистики. Практически несомненно,

что частота события E в длинном ряду повторений эксперимента будет равна приблизительно вероятности его появления.

Гистограмма — график частот распределения в последовательности независимых случайных экспериментов.

Обозначим через S множество всех возможных исходов многократно повторяющегося эксперимента. Предположим, что множество S является объединением непересекающихся подмножеств S_i, т. е. $S = \bigcup_{i \in I} S_i$, где I — конечное или бесконечное множество индексов. Пусть в последовательности n испытаний, где n достаточно велико, обнаружилось, что событие $\xi_i \in S_i$ осуществилось ν_i раз, $i \in I$. Обозначим через ν_i/n частоту появления события из множества S_i. Очевидно, числа ν_i в сумме дают n, т. е. $\sum_{i \in I} \nu_i$.

Гистограмма — ступенчатая функция, график которой строится так: на оси абсцисс откладываются значения случайной величины, которые она принимает, а на оси ординат — частоты ν_i/n.

Выборочное среднее — первый момент выборочного распределения случайной величины. Для одномерных распределений — это среднее арифметическое значение наблюдаемых значений $x_1, \ldots, x_n$ случайной величины ξ:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Для многомерного распределения — это вектор средних арифметических значений компонент наблюдаемых векторов.

Если ξ — одномерная случайная величина с функцией распределения $F(x)$, математическим ожиданием μ_1 и дисперсией D, то ее наблюдаемые значения можно считать значениями одинаково распределенных случайных величин $\xi_1, \ldots, \xi_n$ с функцией распределения $F(x)$. Математическое ожидание случайной величины

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

равно

$$M\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} M\xi_i = \mu_1.$$

Оценкой для m служит среднее арифметическое x, а дисперсия $\bar{\xi}$ равна

$$D(\bar{\xi}) = \frac{1}{n^2} \sum_{i=1}^{n} D(\xi_i) = \frac{\mu_2}{n}.$$

Выборочная дисперсия — второй центральный момент выборочного распределения. Если $x_1, \ldots, x_n$ — наблюдаемые значения одномерной случайной величины ξ, то несмещенная выборочная дисперсия имеет вид

$$m_2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Таблица 2

Наилучшие линейные оценки S среднего квадратического отклонения

Объем выборки	Оценка среднего квадратического отклонения S
2	$0{,}8862\,(x_2 - x_1)$
3	$0{,}5908\,(x_3 - x_1)$
4	$0{,}4539\,(x_4 - x_1) + 0{,}1102\,(x_3 - x_2)$
5	$0{,}3724\,(x_5 - x_1) + 0{,}1352\,(x_4 - x_2)$
6	$0{,}3175\,(x_6 - x_1) + 0{,}1386\,(x_5 - x_2) + 0{,}0432\,(x_4 - x_3)$
7	$0{,}2778\,(x_7 - x_1) + 0{,}1351\,(x_6 - x_2) + 0{,}0625\,(x_5 - x_3)$
8	$0{,}2476\,(x_8 - x_1) + 0{,}1294\,(x_7 - x_2) + 0{,}0713\,(x_6 - x_3) + 0{,}0230\,(x_5 - x_4)$
9	$0{,}2237\,(x_9 - x_1) + 0{,}1233\,(x_8 - x_2) + 0{,}0751\,(x_7 - x_3) + 0{,}0360\,(x_6 - x_4)$
10	$0{,}2044\,(x_{10} - x_1) + 0{,}1172\,(x_9 - x_2) + 0{,}0763\,(x_8 - x_3) + 0{,}0436\,(x_7 - x_4) + 0{,}0142\,(x_6 - x_5)$

Примечание. $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ — вариационный ряд, т. е. наблюденные значения, расположенные в порядке возрастания членов.

Таблица 3

Оценка параметров в зависимости от степени загрязнения выборок. По Диксону, Масси, 1957 г.

Степень загрязнения	Примеры		Лучший уровень значимости	$N \leqslant 10$			$10 < N \leqslant 20$	
				μ	σ	μ	σ
Легкая $\gamma\lambda = 0{,}1$	0,1	1	0,05	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	По размаху: $\frac{x^{max} - x^{min}}{2}$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	$S = \sqrt{\frac{1}{N-1}\left[\Sigma x_i^2 - \frac{1}{N}(\Sigma x_i)^2\right]}$
	0,05	2					
	0,01	10					
Средняя $0{,}3 \leqslant \gamma\lambda \leqslant 0{,}45$	0,3	1	0,10	То же	То же	Ме	То же
	0,1	3					
	0,5	9					
Сильная $\gamma\lambda > 0{,}45$	—	—	0,20	Ме	»	То же	»

Примечание. Основная совокупность $N(\mu, \sigma^2)$ — нормальная совокупность с параметрами μ и σ^2; содержит $N-K$ членов; K — аномальных значений принадлежит совокупности $N(\mu+\lambda\sigma, \sigma^2)$; γ — доля аномальных значений в долях единицы; λ — величина сдвига аномального среднего в стандартных отклонениях основной совокупности; Ме — медиана.

Смещенная оценка дисперсии

$$m_2' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Положительное значение квадратного корня из m_2 называется выборочным средним квадратическим отклонением или выборочным стандартным отклонением. Предполагая, что математическое ожидание $M(m_2)$ наблюдений случайной величины ξ равно нулю, вычислим математическое ожидание выборочной дисперсии m_2':

$$M(m_2') = \mu_2 - \frac{\mu_2}{n} = \frac{n-1}{n}\mu_2.$$

При обработке геологических данных с числом наблюдений, не превышающим 10, в качестве наилучших оценок среднего используется обычное среднее арифметическое, а в качестве наилучшей оценки выборочного стандартного отклонения следует пользоваться рекомендациями Большева и Смирнова [8] (табл. 2).

В случае, если выборочные данные хорошо согласуются с логнормальным распределением, то максимально правдоподобными оценками среднего значения и дисперсии являются оценки по Ачисону и Брауну [39].

Пусть выборочные значения $x_1, \ldots, x_n$. Обозначим их логарифмы через $y_1, \ldots, y_n$, т. е.:

$$y_i = \lg x_i,\ i = 1, 2, \ldots, n;$$

$$\text{их среднее значение } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Максимально правдоподобная оценка среднего (эффективная и удовлетворяющая условию несмещенности) имеет вид

$$a = 10^{\bar{y}}\psi_n(t),$$

$$\text{где } \psi_n(t) \simeq e^t\left\{1 - \frac{t(t+1)}{n} + \frac{t^2(3t^2+22t+21)}{6n^2}\right\};$$

$t = 2{,}65\,S_y^2$ (S_y^2 — выборочная дисперсия y).

Максимально правдоподобная оценка дисперсии имеет вид

$$b^2 = 10^{2\bar{y}}\{\psi_n(t_2) - \psi_n(t_3)\},$$

$$\text{где } t_2 = 10{,}604 S_y^2; \qquad t_3 = \frac{(n-2)\,5{,}302}{n-1}S_y^2.$$

Для загрязненных выборок рекомендуются процедуры получения оценок (по Диксону и Месси) [19], показанные в табл. 3.

Число степеней свободы — ранг квадратичной формы, переменными в которой являются независимые нормально и одинаково распределенные случайные величины. Квадратичные формы от случайных переменных используются при построении распределений χ^2, t и распределения Фишера. Так, если $\xi_0, \xi_1, \ldots, \xi_{n-r+1}$ независимо распределенных нормальных случайных величин с математическим ожиданием 0 и дисперсией 1, то величина $\xi = \sum_{\nu=1}^{n}\xi_\nu^2$ имеет распределение χ^2 с n степенями свободы. Величина

$$\eta = \xi_0 \Big/ \sqrt{\frac{1}{n}\sum_{\nu=1}^{n}\xi_\nu^2}$$

имеет t-распределение с n степенями свободы.

Величина

$$\zeta = \sum_{\nu=1}^{n}\xi_\nu^2 \Big/ \sum_{\nu=n+1}^{n}\xi_\nu^2$$

имеет распределение Фишера с n и $n-m$ степенями свободы.

Матрица внутреннего рассеивания выборки — это $m \times m$-матрица $u = \|u_{ij}\|$, элемент u_{ij} которой равен скалярному произведению векторов $x_i - \bar{x}_i$ и $x_j - \bar{x}_j$, где x_i—i-я координата m-мерного наблюдаемого вектора $x = (x_1, \ldots, x_m)$, а $\bar{x}_i$ — среднее значение i-х выборочных координат по всем N наблюденным значениям:

$$u_{ij} = \sum_{j=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j), \qquad i, j = 1, 2, \ldots, m.$$

Матрица внутривыборочного рассеивания нескольких выборок — это сумма матриц рассеивания отдельных выборок. Пусть $x = (x_1, \ldots, x_m)$ — наблюдаемый m-мерный вектор. Рассмотрим k выборок наблюденных значений x:

$$\begin{matrix} x_{11}^{(\alpha)}, \ldots, x_{m1}^{(\alpha)}, \\ \\ x_{1n_\alpha}^{(\alpha)}, \ldots, x_{mn_\alpha}^{(\alpha)} \end{matrix} \qquad \alpha = 1, 2, \ldots, k.$$

Элементы матрицы внутривыборочного рассеивания имеют вид

$$u_{ij}^{W} = \sum_{\alpha=1}^{k} u_{ij}^{\alpha},$$

$$u_{ij}^{\alpha} = \sum_{g=1}^{n_\alpha}\left(x_{ig}^{(\alpha)} - \bar{x}_i^{(\alpha)}\right)\left(x_{jg}^{(\alpha)} - \bar{x}_j^{(\alpha)}\right),$$

где $(\bar{x}_1^{\alpha}, \ldots, \bar{x}_m^{\alpha})$ —вектор средних значений выборки с номером α. Обозначается эта матрица S_W. Матрица S_W используется в дискриминантном анализе.

Матрица межвыборочного рассеивания — разность между матрицей рассеивания объединенной выборки и матрицей внутривыборочного рассеивания тех же выборок.

Пусть $x = (x_1, \ldots, x_m)$ — наблюдаемый m-мерный вектор, и пусть

$$x_{11}^{(\alpha)}, \ldots, x_{m1}^{(\alpha)},$$
$$\alpha = 1, 2, \ldots, k,$$
$$x_{1n_\alpha}^{(\alpha)}, \ldots, x_{mn_\alpha}^{(\alpha)}$$

— наблюденные значения вектора x для k выборок. Элементы матрицы межвыборочного рассеивания имеют вид

$$u_{ij}^{b} = \sum_{\alpha=1}^{k} n_\alpha (\overline{X}_i^{(\alpha)} - \overline{X}_i)(\overline{X}_j^{(\alpha)} - \overline{X}_j),$$

где $i, j = 1, 2, \ldots m$, $X = (\overline{x}_1, \ldots, \overline{x}_m)$ — вектор-строка средних значений объединенной выборки, а

$$\overline{X}^{(\alpha)} = (\overline{x}_1^{(\alpha)}, \ldots, \overline{x}_j^{(\alpha)}, \ldots, \overline{x}_m^{(\alpha)})$$

вектор-строка средних значений выборки с номером α. Обозначается эта матрица S_B. Матрица S_B используется в дискриминантном анализе.

ТИПЫ ОЦЕНОК И МЕТОДЫ ОЦЕНИВАНИЯ

Для одного и того же неизвестного параметра могут существовать различные варианты оценок, и для того чтобы обоснованно подходить к выбору той или иной из них, необходимо рассмотреть их критерий качества.

Несмещенность. Пусть $x_1, x_2, \ldots, x_i, \ldots, x_n$ — выборка объема n, а θ — неизвестный оцениваемый параметр. Обозначим через $\hat{\theta}(x_1, x_2, \ldots, x_i, \ldots, x_n)$ оценку для θ. Напомним, что оценку $\hat{\theta}(x_1, x_2, \ldots, x_n)$ можно рассматривать как случайную величину.

Если при фиксированном n для оценки $\hat{\theta}(x_1, \ldots, x_n)$ неизвестного параметра θ выполнено условие

$$M\hat{\theta}(x_1, x_2, \ldots, x_n) = \theta,$$

то такая оценка называется несмещенной, т. е. не содержащей систематической ошибки.

Однако, если требование несмещенности не выполняется, этот недостаток обычно бывает легко устраним путем введения соответствующей поправки. Так, например, математическое ожидание оценки дисперсии

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

особенно при небольших n, будет несколько заниженным по сравнению с σ^2, что исправляется выражением

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

Эта оценка является несмещенной.

Состоятельность. Пусть $\hat{\theta}(x_1)$, $\hat{\theta}(x_1, x_2)$, $\hat{\theta}(x_1, x_2, x_3)$, $\ldots$, $\hat{\theta}(x_1, \ldots, x_n)$ — последовательность оценок, полученных по выборкам объема $k = 1, 2, 3, \ldots n$. Такую последовательность называют состоятельной, если

$$\lim_{k\to\infty} P\{|\hat{\theta}(x_1, x_2, \ldots, x_k) - \theta| < \varepsilon\} = 1,$$

где ε — сколь угодно малое заданное число. Иными словами, состоятельной называется такая последовательность оценок $\hat{\theta}(x_1, x_2, \ldots, x_k)$, для которой вероятность события, заключающегося в том, что $\hat{\theta}(x_1, x_2, \ldots, x_k)$ отличается от θ на величину, не превышающую сколь угодно малое заданное число ε, стремится к 1 при неограниченном возрастании k.

Эффективность. Оценка $\hat{\theta}(x_1, x_2, \ldots, x_n)$, обладающая минимальной дисперсией из всех возможных оценок, полученных по выборке объема n, называется эффективной:

$$D\hat{\theta}(x_1, x_2, \ldots, x_n) = \min.$$

Естественно, что такая оценка, при условии, что она не смещена, предпочтительнее любой другой, так как обеспечивает более тесную группировку результатов около истинного значения неизвестного оцениваемого параметра θ.

Достаточность. Примем, что $f(x_i, \theta)$ — плотность вероятности случайной величины в точке x_i. Тогда для выборки объема n функция правдоподобия будет определена выражением

$$\prod_{i=1}^{n} f(x_i, \theta).$$

Оценка $\hat{\theta}(x_1, x_2, \ldots, x_n)$ называется достаточной оценкой неизвестного параметра $\hat{\theta}$, если существует такая функция $h(x_1, x_2, \ldots, x_n)$, не зависящая от θ, для которой имеет место равенство

$$\prod_{i=1}^{n} f(x_i, \hat{\theta}) = q[\hat{\theta}(x_1, x_2, \ldots, x_n), \theta] h(x_1, \ldots, x_n).$$

Необходимо отметить, что достаточная оценка включает всю информацию, которую можно получить о неизвестном параметре по выборке объема n.

Метод моментов — метод получения точечных оценок параметров генеральной совокупности. Его сущность коротко можно охарактеризовать следующим образом. Предположим, что задача состоит в оценке k параметров $\theta_1, \ldots, \theta_k$ генеральной совокупности. Для получения этих оценок приравниваем k первых моментов генеральной совокупности к первым k выборочных моментов. Решение этих уравнений дает оценки параметров. Как правило, эти оценки состоятельны.

Метод максимального правдоподобия — статистический метод получения оценок параметров, основанный на принципе максимального правдоподобия, сформулированном Фишером и состоящий в том, что в качестве оценки параметра θ из области его допустимых значений выбирается то значение, для которого функция правдоподобия принимает наибольшее возможное значение.

Пусть $x_1, \ldots, x_n$ — выборка значений исследуемой случайной величины ξ, распределение которой задается плотностью $f(x, \theta)$, где θ — параметр распределения. Если распределение случайной величины ξ дискретно, то через $f(x, \theta)$ обозначается вероятность принятия значения x. Совместное распределение наблюденных значений $x_1, \ldots, x_n$, рассматриваемых как случайные величины, задается функцией параметра θ следующего вида:

$$L(x, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta).$$

$L(x, \theta)$ называется функцией правдоподобия.

Вид оценки максимального правдоподобия определяется следующим образом. Если функция правдоподобия дважды дифференцируема, то ее стационарные значения даются корнями уравнения

$$\frac{\partial L(x, \theta)}{\partial \theta} = 0.$$

Достаточным условием того, что корень этого уравнения будет локальным максимумом является равенство

$$\frac{\partial^2 L(x, \theta)}{\partial \theta^2} = 0.$$

Сначала находят локальные максимумы функции $L(x, \theta)$, а затем выбирают из них наибольший.

На практике иногда проще проводить вычисления не с функцией $L(x, \theta)$, а с ее логарифмом. Так как функция и ее логарифм имеют максимумы в одних и тех же точках, то условия $(L)'_\theta = 0$ и $(L)''_\theta < 0$ заменяются на условия $(\log L)'_\theta = 0$, $(\log L)''_\theta < 0$.

Например, найдем среднее значение θ нормального распределения с известной дисперсией. Имеем:

$$f(x, \theta) = (2\pi)^{-1/2} \sigma^{-1} \exp\left\{-\left(\sum_{i=1}^{n} x_i - \theta\right)^2 / 2\sigma^2\right\},$$

$$\frac{\partial \log f(x, \theta)}{\partial \theta} = \frac{n}{\sigma^2}(\bar{x} - \theta).$$

Приравнивая это выражение к нулю, находим $\hat{\theta} = \bar{x}$. Это несмещенная оценка для θ.

В теории проверки статистических гипотез часто используется отношение правдоподобия.

Предположим, что задача состоит в проверке гипотезы H_0 при конкурирующей гипотезе H_1 на основании выборки из n значений $x_1, \ldots, x_n$ случайной величины ξ с функцией распределения

$f(x, \theta)$, где θ — неизвестный параметр. Проверяемая гипотеза H_0 представляет собой равенство $\theta = \theta_0$, а конкурирующая гипотеза H_1 имеет вид $\theta = \theta_1$. Отношение правдоподобия, используемое при проверке гипотезы H_0 при конкурирующей гипотезе H_1, имеет вид

$$\frac{L(x_1, \theta_0)}{L(x_1, \theta_1)} = \frac{f(x_1, \theta_0) f(x_2, \theta_0) \ldots f(x_n, \theta_0)}{f(x_1, \theta_1) f(x_2, \theta_1) \ldots f(x_n, \theta_1)}.$$

Отношение правдоподобия широко используется в теории проверки статистических гипотез [29], например в последовательном анализе [11].

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Проверка гипотез, в результате которой можно подтвердить или опровергнуть какое-либо предположение, осуществляется с помощью некоторых случайных величин, называемых статистическими критериями. В связи с тем что критерий является случайной величиной, он полностью характеризуется соответствующей функцией распределения при условии, что проверяемая гипотеза верна.

Проиллюстрируем особенности применения статистических критериев на примере проверки гипотезы о соответствии модели нормального распределения эмпирическим данным для выборок большого объема. Последнее позволяет использовать нормальное приближение.

Метод проверки гипотезы о соответствии нормальной модели эмпирическим данным с помощью отношений оценок асимметрии и эксцесса к их стандартным отклонениям заключается в совместном выполнении двух предположений. Во-первых, в условиях близкого к нормальному распределения выборочных данных отношение $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ должно представлять собой значение случайной величины, распределенной асимптотически нормально с математическим ожиданием, равным 0, и дисперсией, равной 1. Во-вторых, в тех же условиях разность $\left(\bar{\nu}_4/\sigma_{\bar{\nu}_4}\right) - 3$ должна представлять собой значение аналогичной случайной величины, распределенной асимптотически нормально со средним 0 и дисперсией, равной 1. В качестве примера используем лишь часть общей гипотезы о нормальности — только предположение о равенстве нулю среднего для отношения $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$. Этому предположению соответствует нулевая гипотеза:

$$H_0 : M\left(\bar{\nu}_3/\sigma_{\bar{\nu}_3}\right) = 0. \tag{2.1}$$

Интересующий нас набор альтернатив:

$$H_1 : M\left(\bar{\nu}_3/\sigma_{\bar{\nu}_3}\right) \neq 0. \tag{2.2}$$

В результате проверки нулевой гипотезы может быть допущена ошибка, заключающаяся в принятии одной из альтернатив, хотя на самом деле верна нулевая гипотеза. Такая ошибка называется

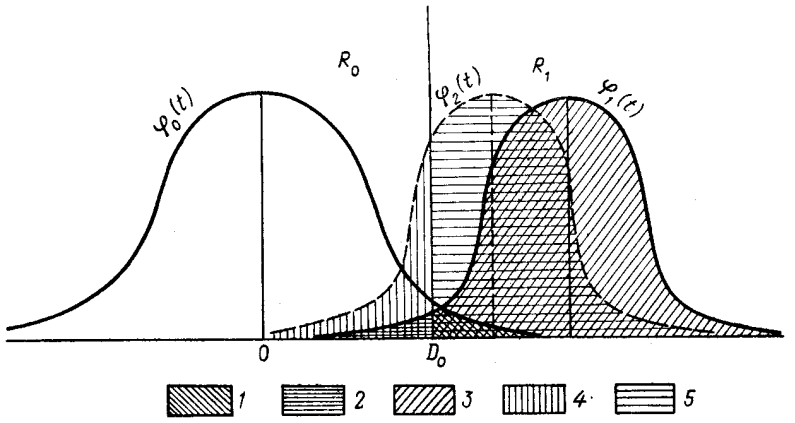


Рис. 28. Соотношения между вероятностями ошибок и мощностью критерия:
1 — площадь, соответствующая вероятности ошибок первого рода; *2* — то же, второго рода при первой альтернативе; *3* — площадь, соответствующая мощности критерия при первой альтернативе; *4* — площадь, соответствующая вероятности ошибки второго рода при второй альтернативе; *5* — площадь, соответствующая мощности критерия при второй альтернативе

ошибкой первого рода. Наоборот, ложное принятие нулевой гипотезы, хотя она неверна, называется ошибкой второго рода.

Прежде всего следует отметить, что набор альтернатив в том виде, в каком он приведен в неравенстве (2.2), называется двусторонним. Другими словами, двустороннее множество альтернатив охватывает все возможные предположения, отличные от нулевой гипотезы. Для простоты дальнейшее изложение будет основано на примере, в котором фигурирует одностороннее множество альтернатив вида

$$H_1 : M\left(\bar{\nu}_3/\sigma_{\bar{\nu}_3}\right) > 0.$$

Положим, что $\varphi_0(t)$ — функция плотности распределения $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ в условиях нулевой гипотезы (2.1). Функцию плотности распределения $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ в условиях одной из альтернатив одностороннего множества H_1^1 обозначим через $\varphi_1(t)$ (рис. 28). Задавая некоторое критическое значение D_0, восставим из этой точки перпендикуляр, делящий плоскость рисунка на две области; R_0 и R_1. Припишем областям R_0 и R_1 следующие свойства. Если при вычислении эмпирического значения отношения $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ мы получим величину, меньшую D_0, т. е. значение окажется в области R_0, то нулевую гипотезу следует принять. В противном случае, т. е. при попадании вычисленного значения $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ в область R_1, нулевая гипотеза должна быть отвергнута. Эти свойства областей R_0 и R_1 отражены в их названиях. Так, область R_0 называется областью принятия нулевой гипотезы, а область R_1 — областью отклонения нулевой гипотезы, или критической областью.

Как уже отмечалось, эти выводы, т. е. принятие или отклонение нулевой гипотезы, могут оказаться ошибочными. Какова же вероятность допустить ошибку при данном критическом значении D_0 и сформулированном выше множестве альтернатив? Вероятность ошибки первого рода при заданных D_0 и H_1^1 отвечает области, находящейся на чертеже (см. рис. 28) под кривой $\varphi_0(t)$ в области R_0. Обозначив вероятность ошибки первого рода, соответствующую критическому значению D_0, через α, выразим ее в рассматриваемом случае следующим образом:

$$\alpha = \int_{R_1} \varphi_0(t)\, dt..$$

Обозначим через β вероятность ошибки второго рода. При сформулированной выше альтернативе H_1^1 она равна

$$\beta = \int_{R_0} \varphi_1(t)\, dt.$$

Функция, заданная на множестве альтернатив

$$1-\beta = \int_{R_1} \varphi_1(t)\, dt,$$

называется функцией мощности критерия при заданном множестве альтернатив H_1^1.

Рассмотрим более подробно соотношение α, β, $1-\beta$. Нетрудно заметить, что α определяется значением D_0. С равным основанием можно сказать, что величина D_0 может быть определена через α. Дело в следующем. Зная функцию плотности распределения величины $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ в условиях нулевой гипотезы, можно указать такое значение $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$, чтобы появление значений, больших по величине, чем выбранное, происходило с заданной малой вероятностью α. Это и есть D_0. Вероятность α можно выбрать, в свою очередь, так, чтобы при единичном эксперименте (вычислении $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$) осуществление события $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ было бы практически невозможным. В этом случае вполне естественно считать, что данное значение $\bar{\nu}_3/\sigma_{\nu_3} > D_0$ практически нельзя считать принадлежащим совокупности, которая характеризуется функцией $\varphi_0(t)$, а следует отнести скорее к альтернативной совокупности, распределение которой подчиняется функции плотности $\varphi_1(t)$. Таким образом, значение D_0 устанавливается заранее и соответствует определенному заданному риску ошибочно отвергнуть верную нулевую гипотезу. Вероятность появления ошибки первого рода равна α и обычно называется уровнем значимости.

Уровень значимости и, следовательно, величина D_0 определяют вероятность ошибки второго рода, т. е. β. Последняя, кроме того,

зависит от альтернативы (см. рис. 28). Естественно, что для различных альтернатив ошибка второго рода и мощность критерия могут быть разными. Положим, что существует альтернатива H_1^2, в условиях которой $\bar{\nu}_3/\sigma_{\bar{\nu}_3}$ имеет распределение с функцией плотности $\varphi_2(t)$ (см. рис. 28). Нетрудно видеть, что при этой альтернативе значительно увеличилась вероятность ошибки второго рода. Ей соответствует площадь под кривой в области R_0, т. е.:

$$\beta' = \int_{R_0} \varphi_2(t)\,dt.$$

Это означает, что мы в большом числе случаев ошибочно будем принимать ложную гипотезу. Другими словами, в значительном числе экспериментов мы не сможем отличить значения величины $\bar{\nu}/\sigma_{\bar{\nu}_3}$, принадлежащие разным совокупностям: совокупности критерия в условиях нулевой гипотезы и в условиях альтернативы. Это можно назвать потерей чувствительности критерия к данной альтернативе.

Таким образом, падение мощности критерия влечет за собой уменьшение его чувствительности, т. е. снижает возможность различить действительно разные совокупности. Следует отметить, что для одних и тех же выборочных данных, применяя при проверке одной и той же нулевой гипотезы различные критерии, обладающие разной мощностью при заданной альтернативе, можно получить сильно отличающиеся результаты. Поэтому мощность критерия является показателем его качества.

Существует класс критериев, которые обладают наибольшей мощностью при проверке определенной нулевой гипотезы по отношению ко всему множеству возможных альтернатив. Этот класс критериев называется классом «равномерно наиболее мощных критериев» [29].

Из приведенного изложения следует, что выбор уровня значимости и принятие решения после проверки гипотезы определяются рядом причин: характером альтернативы, сравнительной ценностью потерь от совершения ошибок первого и второго рода, выполнением условий, накладываемых на критерий, и т. п. Все это следует учитывать в процессе применения статистических методов в геологии и при интерпретации полученных результатов.

Проверка гипотез о нормальном распределении

Проверка гипотезы об одномерном нормальном распределении — процедура выявления отклонения выборочного распределения от нормального.

Если случайная величина ξ подчиняется нормальному закону распределения $N(a, \sigma)$, с параметрами (a, σ), то известны следующие соотношения:

$$\delta = \frac{M|\xi - a|}{\sigma} = \sqrt{\frac{2}{\pi}} = 0{,}79788;$$

$$\gamma_1 = \frac{M(\xi - a)^3}{\sigma^{\frac{3}{2}}} = 0; \qquad \beta_2 = \frac{M(\xi - a)^4}{\sigma^4} - 3,$$

где δ — нормированное среднее абсолютное отклонение; γ — асимметрия; β_2 — 3 — эксцесс.

Таким образом, гипотезе о нормальном распределении равносильны нулевые гипотезы: $\delta = \sqrt{2/\pi}$, $\gamma_1 = 0$ и $\beta_2 = 0$, при альтернативах $\delta \neq \sqrt{2/\pi}$, $\gamma_1 \neq 0$, $\beta_2 \neq 0$.

Критические значения соответствующих статистик вычисляются в предположении о нормальности распределения.

Пусть $\xi_1, \ldots, \xi_n$ — выборочные значения случайной величины ξ, распределенной по нормальному закону с параметрами (a, σ). Для оценки величин δ, γ_1 и β_2 используются соответствующие выборочные характеристики:

$$d = \frac{1}{nS}\sum_{i=1}^{n}|\xi_i - \bar{\xi}|; \qquad g_1 = \frac{1}{nS^3}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^3;$$

$$b_2 = \frac{1}{nS^4}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^4,$$

где $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ — выборочное среднее,

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$$

— выборочная дисперсия; d — выборочное среднее абсолютное отклонение; g_1 — выборочный коэффициент асимметрии; $b_2 - 3$ — выборочный коэффициент эксцесса.

Статистики d, g_1, b_2 распределены асимптотически нормально, однако b_2 очень медленно приближается к нормальному распределение при $n \to \infty$. Параметры этих нормальных распределений следующие:

$$Md = \frac{2}{\sqrt{\pi(n-1)}}\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \sqrt{\frac{2}{\pi}}\left[1 + \frac{2}{8n-9} + O\left(\frac{1}{n^3}\right)\right];$$

$$Dd = \frac{1}{n}\left\{1 + \left[\frac{2}{\pi}\sqrt{n(n-2)} + \arcsin\frac{1}{n-1}\right]\right\} -$$

$$-\frac{n-1}{\pi}\left[\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right]^2 = \frac{1}{n}\left[\left(1 - \frac{3}{\pi}\right) - \frac{1}{4\pi n} + O\left(\frac{1}{n^2}\right)\right] =$$

$$= \frac{1}{n}\left[0{,}04507 - 0{,}0796\,\frac{1}{n} + O\left(\frac{1}{n^2}\right)\right];$$

$$Mg_1 = 0;$$

$$Dg_1 = \frac{6(n-2)}{(n+1)(n+3)} = \frac{6}{n}\left[1 - \frac{12}{2n+7} + O\left(\frac{1}{n^2}\right)\right];$$

$$Mb_2 = 3 - \frac{6}{n+1};$$

$$Db_2 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} = \frac{24}{n}\left[1 - \frac{225}{15n+124} + O\left(\frac{1}{n^3}\right)\right].$$

Решающее правило при проверке гипотезы об отклонении выборочного распределения от нормального следующее.

Выбираем n, $\alpha = .0{,}01$; $0{,}05$ и т. д. Вычисляем

$$d(n),\ Md(n),\ \sqrt{Dd(n)},\ \ldots$$

Если

$$\frac{1}{\sqrt{Dd(n)}}\,|d(n) - Md(n)| > n_\alpha,\ \text{где}\ n_\alpha - \alpha -$$

процентная точка нормального распределения, то с вероятностью 2α можно считать, что выборочное распределение отклоняется от нормального. Аналогично выглядит решающее правило для g_1. В связи с тем что выборочный эксцесс $b_2 - 3$ медленно стремится к нормальному распределению, процентные точки

$$(b_2 - Mb_2)/\sqrt{Db_2}$$

табулируются специально [8].

Таким образом, если

$$\left.\begin{aligned} t_1 &= \frac{\gamma_1}{\sqrt{D\gamma_1}} \leqslant t_{1-\alpha/2} \\ \text{и}\quad t_2 &= \frac{d - Md}{\sqrt{Dd}} \leqslant t_{1-\alpha/2} \end{aligned}\right\}\ \text{стремится к}\ N(0,1).$$

В противном случае при

$$\left.\begin{aligned} &t_1 > t_{1-\alpha/2} \\ &\text{и (или)}\ t_2 > t_{1-\alpha/2} \end{aligned}\right\}\ \text{не стремится к}\ N(0,1).$$

Критерий проверки соответствия одномерного выборочного распределения заданному (критерий Пирсона) позволяет установить степень соответствия выборочных данных, состоящих из n значений случайной величины, заданной функции распределения $f(x)$. Предполагается, что функция $F(x)$ или не содержит никаких неизвестных параметров, или же они оцениваются по выборке, и вероятность события $\xi < a$ может быть вычислена для любых вещественных значений a.

Критерий Пирсона χ^2 состоит в следующем [8]. Предположим, что пространство значений изучаемой случайной величины ξ разбито на конечное число r непересекающихся частей

$S_1, \ldots, S_r$. Обозначим через $p(S_i)$ вероятностную меру множества S_i, вычисленную с помощью функции распределения $F(x)$: $p(S_i) = p_F(x \in S_i)$. По n выборочным данным можно определить частоты $p_i = \frac{\nu_i}{n}$ попадания выборочных значений в множество S_i.

К. Пирсон показал, что статистика

$$M = \sum_{i=1}^{r} (\nu_i - np_i)^2/(np_i)$$

при $n \to \infty$ имеет асимптотическое распределение χ^2 с $r-1$ степенями свободы.

Таким образом, если при большом n статистика M превышает заданное α-процентное значение распределения χ^2 с $n-1$ степенями свободы, то нет оснований считать, что исследуемая выборка извлечена из распределения с функцией распределения $F(x)$. В противном случае можно считать, что выборочные данные находятся в соответствии с распределением, задаваемым функцией $F(x)$.

Большинство многомерных статистических критериев проверки гипотез применимо в предположении, что выборочные данные извлечены из многомерной нормальной совокупности. Так, асимптотические распределения статистик Хотелинга, Кульбака и других получены в предположении о нормальности распределений исходных случайных величин. Несмотря на то что эти критерии широко используются в настоящее время в геологических исследованиях, нормальность распределений обычно не проверяется. Естественно, при таком применении статистических критериев полученные результаты не всегда надежны.

Ниже приводятся два критерия проверки соответствия эмпирического распределения многомерному нормальному. Первый из них — критерий Б. Уэгла — очень прост и соответствует интуитивному желанию исследователя ограничиться после некоторых преобразований проверкой нормальности распределений соответствующих одномерных характеристик. Второй критерий Мардиа более сложен, но и более чувствителен к отклонению выборочного распределения от многомерного нормального. Используемые в нем статистики удобны при изучении влияния нарушения нормальности распределения на устойчивость известных критериев [7].

Критерий Уэгла. Пусть

$$X = \| x_{ij} \|_{\substack{1 \leqslant i \leqslant N \\ 1 \leqslant j \leqslant p}}$$

— выборка N наблюдений над p-мерным вектором $x = (x_1, \ldots, x_p)$.

Рассмотрим критерий проверки того, что X — выборка из многомерного нормального распределения с вектором средних значений a и ковариационной матрицей S (параметр a и матрица S неизвестны).

Пусть $\hat{S}$ — выборочная ковариационная матрица, построенная по матрице наблюдений X. Ее всегда можно записать в виде $\hat{S} = QQ'$. Действительно, существует такое линейное преобразо-

вание T, что $T'\hat{S}T$ — диагональная матрица Λ. Тогда $\hat{S} = (T')^{-1}\Lambda T^{-1}$, откуда вытекает требуемая запись

$$Q = (\Lambda^{1/2}T')^{-1/2}.$$

По матрице X строим совокупность векторов:

$$x_i = (x_{i1} - \bar{x}_{i1},\ x_{i2} - \bar{x}_{i2},\ \ldots,\ x_{ip} - \bar{x}_{ip}),\quad i = 1, 2, \ldots, N,$$

где $$\bar{x}_{ij} = \frac{1}{N-1}\sum_{t=1,\ t\neq i}^{N} x_{it},\quad j = 1, 2, \ldots, p.$$

По этой последовательности $x_1, \ldots, x_N$ строим последовательность

$$Z_i = Q^{-1}x_i, \qquad Q^{-1} = \Lambda^{-1/2}T'.$$

Критерий Уэгла состоит в проверке гипотезы о том, что X — выборка из p-мерной нормальной совокупности $N(a, S)$.

Пусть $\overline{X}$ — вектор выборочных средних и A — выборочная ковариационная матрица, вычисленная по N наблюдениям, взятым из p-мерной совокупности с нормальным законом распределения $N(0, E)$, где E — единичная матрица. Плотности распределения вектора $\overline{X}$ и элементов A равны соответственно

$$f(\overline{X})\,d\overline{X} = \text{const}\, e^{-\frac{N}{2}tr\overline{X}\overline{X}'}\,d\overline{X},$$

$$f(A)\,dA = \text{const}\,|A|^{\frac{N-p-1}{2}} e^{-\frac{1}{2}trA}\,dA.$$

Определим матрицу

$$\hat{X} = \overline{X} + \frac{N-1}{2}ZA^{1/2}.$$

Проверка того, что X — выборка из нормальной совокупности $N(a, S)$ сводится к проверке того, что $\overline{X}$ — выборка из нормальной совокупности $N(0, E)$. Проверка последнего условия сводится к проверке того, что p-мерная совокупность разбивается на p одномерных нормальных совокупностей.

Алгоритмы извлечения квадратного корня из матрицы имеются в работе [7].

Критерий Мардиа. Пусть $X = \|x_{ij}\|_{1\leqslant i\leqslant N,\ 1\leqslant j\leqslant p}$ — выборка объема N из p-мерной совокупности. Обозначим через $\overline{X} = (\bar{x}_1, \ldots, \bar{x}_p)$ вектор выборочных средних, а через

$$\hat{S} = \|S_{ij}\|_{1\leqslant i,\ j\leqslant p}$$

— выборочную ковариационную матрицу.

Для применения рассматриваемого критерия необходимо выполнение следующих условий.

1. Вторые моменты вектора $\overline{X}$ и вектора

$$\check{S} = (\hat{S}_{11},\ \hat{S}_{22},\ \ldots,\ \hat{S}_{pp},\ \hat{S}_{12},\ \ldots,\ \hat{S}_{1p},\ \hat{S}_{23},\ \ldots,\ \hat{S}_{2p},\ \ldots,\ \hat{S}_{p-1p}),$$

где S_{ij} — матричные элементы выборочной ковариационной матрицы, имеют порядок N^{-1}.

2. Моментами выше третьего порядка случайного вектора $x = (x_1, \ldots, x_p)$ можно пренебречь, т. е. они должны быть малыми по абсолютной величине.

Введем следующие суммы:

$$M^{rSi} = \frac{1}{N}\sum_{i=1}^{N}(x_{ir} - \bar{x}_i)(x_{is} - \bar{x}_S)(x_{it} - \bar{x}_t),$$

$$b_{1,\,p} = \sum_{r,\,S,\,t}\ \sum_{r',\,S',\,t'} \hat{S}^{rr'}\hat{S}^{SS'}\hat{S}^{tt'}M^{rSt}M^{r's't'},$$

где $\hat{S}^{ij}$ — элемент обратной матрицы $\hat{S}^{-1}$;

$$b_{2,\,p} = \frac{1}{N}\sum_{i=1}^{N}[(x_i - \overline{X})'\hat{S}^{-1}(x_i - \overline{X})]^2.$$

Критерий Мардиа для проверки гипотезы о том, что выборка X с вектором математического ожидания $\overline{X}$ и выборочной ковариационной матрицей $\bar{S}$ извлечена из многомерной нормальной совокупности, состоит в проверке двух условий:

$$b_{1,\,p} = 0, \qquad b_{2,\,p} = p(p+2).$$

Выполнение первого условия проверяется с помощью статистики $A = \frac{N}{6}b_{1,\,p}$, имеющей распределение χ^2 с $\frac{p(p+1)(p+2)}{2}$ степенями свободы.

Таким образом, если для заданного уровня значимости α значение статистики A, вычисленное по выборочным данным, превосходит значение χ^2_α то гипотеза о том, что $b_{1,\,p} = 0$, отвергается. Если $A \leqslant \chi^2_\alpha$, то гипотеза $b_{1,\,p} = 0$ принимается.

Для проверки такого условия $b_{2,\,p} = p(p+2)$ используется статистика

$$B = \left(b_{2,\,p} - \frac{p(p+2)(N-1)}{N+1}\right) \Big/ \frac{\sqrt{8p(p+2)}}{N},$$

распределенная в условиях проверяемой гипотезы по закону $N(0, 1)$, т. е. по нормальному закону с математическим ожиданием нуль и дисперсией единица. Задаваясь тем же уровнем значимости, что и выше, проверим, не противоречит ли вычисленное значение B гипотезе о нормальном распределении.

Приведем формулы для распределения коэффициентов $b_{1,\,p}$ и $b_{2,\,p}$ для случая $p = 2$. Пусть $\hat{S}_1^2$, $\hat{S}_2^2$ — выборочные дисперсии случайных величин x_1 и x_2, $\hat{r}$ — их выборочный коэффициент корре-

ляции. Если g_{rS} — выборочный центральный момент порядка r, S вектора (x_1, x_2), то через $\gamma_{r,s}$ обозначим отношение $g_{rS}/\hat{S}_1^r\hat{S}_2^r$. Тогда

$$\beta_{1,2} = (1-\hat{r}^2)^{-3}\{[\gamma_{30}^2+\gamma_{03}^2+3(1+2\hat{r}^2)][\gamma_{12}^2+\gamma_{21}^2]-$$
$$-2\hat{r}^3\gamma_{30}\gamma_{03}+6\hat{r}[\gamma_{30}(\hat{r}\gamma_{12}-\gamma_{21})+\gamma_{03}(\hat{r}\gamma_{21}-\gamma_{12})-(2+\hat{r}^2)\gamma_{12}\gamma_{21}]\};$$
$$\beta_{2,2} = \gamma_{40}+\gamma_{04}+2\gamma_{22}+4\hat{r}(\hat{r}\gamma_{22}-\gamma_{13}-\gamma_{31})/(1-\hat{r}^2)^2.$$

Последовательный анализ

Характерная черта метода статистического исследования, называемого последовательным анализом, заключается в том, что число наблюдений, необходимых в процессе испытания для получения статистических выводов, заранее не определено. Решение об окончании эксперимента зависит на каждой данной стадии эксперимента от результатов предыдущих наблюдений. Достоинство этого метода заключается в том, что он позволяет сконструировать такую методику проверки, которая требует в среднем меньшего числа наблюдений, чем равная ей по надежности проверка, основанная на заранее определенном количестве наблюдений.

В теории проверки статистических гипотез количество наблюдений, на которых основывается проверка, постоянно для каждой задачи. Существенной чертой последовательного анализа является то, что количество наблюдений, необходимых для принятия решения, зависит от исхода самих наблюдений и, следовательно, является случайной величиной.

Сущность последовательного анализа как метода проверки гипотез состоит в следующем: устанавливается некоторое правило, которым руководствуются на каждой стадии эксперимента (при m-м испытании, где m — целое число) при принятии одного из трех решений: принять гипотезу, отклонить гипотезу, продолжать эксперимент и провести дополнительное наблюдение. Если принимается первое или второе решение, то проверка на этом заканчивается. Если принимается третье решение, то производится следующее наблюдение. На основе этого наблюдения снова принимается одно из трех решений и т. д. Процесс заканчивается тогда, когда будет принято одно из первых двух решений.

Обозначим через M_m множество всех возможных выборок объема m $(x_1, x_2, \ldots, x_m)$. Правило, согласно которому принимается одно из трех решений, можно охарактеризовать с помощью трех областей пространства M_m: R_m^0, R_m^1, R_m (рис. 29—30). Если в результате наблюдения величины x_1 окажется, что она принадлежит области R_1^0, то принимается проверяемая гипотеза H_0. Если величина x_1 попадает в область R_1^1, то гипотеза H_0 отклоняется, т. е. принимается гипотеза H_1. Если величина x_1 попадает в область R_1, то производится наблюдение x_2, относительно которого проводятся те же процедуры с проверкой его принадлежности областям R_2^0,

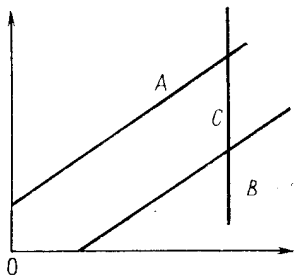


Рис. 29. Зоны принятия гипотезы (A), отклонения (B) и продолжение испытаний (C)

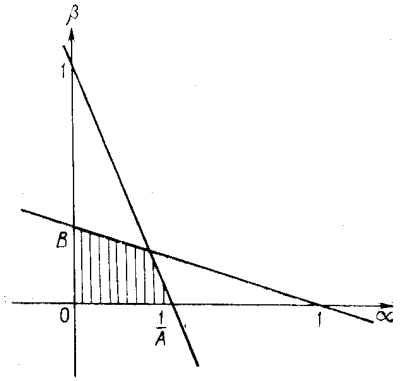


Рис. 30. Совокупность всех α и β, удовлетворяющих неравенствам $\frac{\alpha}{1-\beta} \leqslant \frac{1}{A}$, $\frac{\beta}{1-\alpha} \leqslant B$ при заданных A и B

R_2^1, R_2 и т. д., до остановки процесса, когда при некотором m принимается решение о принадлежности x_m области R_m^0 или R_m^1. Области R_m^0, R_m^1 и R_m не пересекаются друг с другом.

Предполагается, что вероятностное распределение исследуемой случайной величины ξ известно и описывается функцией плотности $f(x, \theta) = f(x, \theta_1, \ldots, \theta_k)$, где $\theta_1, \ldots, \theta_k$ — неизвестные параметры. Рассмотрим для простоты случай $k = 1$, когда параметр θ принимает только два значения, например θ_0 и θ_1. Предположим, что задача состоит в проверке гипотезы H_0, состоящей в том, что $\theta = \theta_0$, а конкурирующая гипотеза H_1 означает, что $\theta = \theta_1$. С последовательной проверкой гипотезы H_0 относительно H_1 связаны два числа α, β, заключенные между нулем и единицей. Если истинна гипотеза H_0, то вероятность того, что мы допустим ошибку первого рода (т. е. отклоним гипотезу H_0), будет равна α, а если истинна гипотеза H_1, то вероятность того, что мы совершим ошибку второго рода (т. е. примем гипотезу H_0), будет равна β. Пара чисел (α, β) называется силой критерия. Числа α, β задаются исследователем.

Для характеристики последовательного критерия проверки простой гипотезы H_0 против конкурирующей гипотезы H_1 рассмотрим вероятность получения выборки $x_1, \ldots x_m$ или функцию правдоподобия в предположении, что справедлива гипотеза H_0 (т. е. $\theta = \theta_0$):

$$L_{0m} = f(x_1, \theta_0) \ldots f(x_m, \theta_0),$$

и в предположении, что справедлива гипотеза H_1 (т. е. $\theta = \theta_1$):

$$L_{1m} = f(x_1, \theta_1) \ldots f(x_m, \theta_1),$$

Последовательный критерий отношения правдоподобия для проверки гипотезы H_0 относительно H_1 определяется таким обра-

зом. Выбираются два положительных числа A и B, $A > B$. На каждой стадии эксперимента, т. е. при $m = 1, 2, \ldots$ вычисляется отношение L_{1m}/L_{0m}. Если

$$B < (L_{1m}/L_{0m}) < A,$$

то эксперимент продолжается. Если $(L_{1m}/L_{0m}) > A$, то процесс оканчивается отклонением гипотезы H_0 (т. е. принятием гипотезы H_1). Если $(L_{1m}/L_{0m}) \leqslant B$, то процесс оканчивается принятием гипотезы H_0.

Постоянные A и B определяются так, чтобы критерий имел заданную силу (α, β). Обозначим эти числа $A(\alpha, \beta)$, $B(\alpha, \beta)$. Легко убедиться в том, что числа $A(\alpha, \beta)$ и $B(\alpha, \beta)$ должны удовлетворять следующим неравенствам:

$$A(\alpha, \beta) \leqslant (1-\beta)/\alpha, \qquad B(\alpha, \beta) \geqslant \beta/(1-\alpha).$$

Действительно, вероятность получения любой заданной выборки $(x_1, \ldots, x_n)$, удовлетворяющей условиям

$$B < \frac{L_{1m}}{L_{0m}} = \frac{f(x_1, \theta_1) \ldots f(x_m, \theta_1)}{f(x_1, \theta_0) \ldots f(x_m, \theta_0)} < A, \quad m = 1, 2, \ldots, n-1,$$

$$(L_{1n}/L_{0n}) < B,$$

по крайней мере в A раз больше при гипотезе H_1, чем при гипотезе H_0. Вероятность получения таких выборок является вероятностью того, что последовательный процесс окончится принятием гипотезы H_1 (отклонением H_0). Эта вероятность равна α, когда верна гипотеза H_0, и $1-\beta$, когда верна гипотеза H_1. Таким образом, получаем неравенство

$$1-\beta \geqslant A\alpha, \qquad \text{т. е.} \qquad A \leqslant (1-\beta)/\alpha.$$

Следовательно, $(1-\beta)/\alpha$ — верхняя грань для A.

Аналогично можно получить и нижнюю грань для B. Действительно, вероятность получения любой заданной выборки $(x_1, \ldots x_n)$, удовлетворяющей условиям

$$B \leqslant \frac{L_{1m}}{L_{0m}} = \frac{f(x_1, \theta_1) \ldots f(x_m, \theta_1)}{f(x_1, \theta_0) \ldots f(x_m, \theta_0)}, \qquad m = 1, 2, \ldots, n-1,$$

$$(L_{1n}/L_{0n}) \geqslant A,$$

по крайней мере в B раз больше при гипотезе H_1, чем при гипотезе H_0. Следовательно, и вероятность принятия H_0 при гипотезе H_1 по крайней мере в B раз больше, чем при H_0. Так как вероятность принятия H_0 равна $1-\alpha$, когда верна гипотеза H_0, и равна β, когда верна гипотеза H_1, то из этого следует неравенство

$$\beta \leqslant (1-\alpha)B,$$

которое можно записать в виде $B \geqslant \beta/(1-\alpha)$, т. е. $\beta/(1-\alpha)$ — нижняя граница для B.

Полученные неравенства можно переписать в виде $\alpha/(1-\beta) \leqslant 1/A$, $\beta/(1-\alpha) \leqslant B$.

Совокупность всех α и β, удовлетворяющих этим неравенствам при заданных A и B, можно изобразить графически (см. рис. 30, заштриховано).

Доказывается, что для достаточно большого класса совместных распределений $L_{0m}(x_1, \ldots, x_m)$ и $L_{1m}(x_1, \ldots, x_m)$ с вероятностью единица процесс рано или поздно заканчивается. Более того, этот вывод справедлив как для зависимых, так и для независимых наблюдений $x_1, \ldots, x_n$.

На практике числа α и β выбираются исходя из условий эксперимента. Числа A и B в первом приближении можно взять равными их верхней и соответственно нижней граням. При этом число наблюдений, необходимых для принятия решения, возрастает незначительно [11].

Проверка гипотез о параметрах распределения

Критерий Вилкоксона проверки гипотез о равенстве средних предназначен для проверки гипотезы H_0: $a_1 = a_2$ против набора альтернатив H_1: $a_1 \neq a_2$, где a_1 и a_2 — истинные средние для первого и второго объектов. Критерий Вилкоксона нечувствителен к нарушению условий нормальности распределения исходных геологических данных, к наличию аномальных значений и т. п. Предполагается, что элементы выборок взаимно независимы и подчиняются непрерывным распределениям.

Процедура применения критерия Вилкоксона следующая. Из двух выборок исходных данных $\{x_t\}$ и $\{y_t\}$ составляется общий вариационный ряд объемом $N = n_1 + n_2$ в порядке возрастания всех выборочных значений x и y. Далее нумеруют все члены этого ряда: $1, 2, \ldots, N$. Равным значениям (совпадающим членам) присваивают скорректированный средний ранг, представляющий собой среднее арифметическое рангов совпадающих (связанных) членов вариационного ряда.

Статистика W критерия Вилкоксона представляет собой сумму рангов r, относящихся к членам меньшей по объему выборки (сумму ранговых чисел):

$$W = \sum_{i=1}^{n_1} r_i, \quad n_1 \leqslant n_2.$$

Критические значения W_1 и W_2 определяются следующим образом в зависимости от объемов наблюдений n_1 и n_2 в выборках.

Ситуация 1. Объемы наблюдений в выборках не превышают 25. В работе [8, см. табл. 6.8] $n_1 = m$, $n_2 = n$, односторонний уровень значимости $\alpha/2 = Q$, удвоенное математическое ожидание статистики Вилкоксона — $2MW$.

По указанной таблице при заданных m, n, Q определяют нижнее критическое значение W_1, а также $2MW$, с помощью которого определяют верхнее критическое значение $W_2 = 2MW - W_1$.

В случае, когда имеются связанные ранги, но количество совпавших значений невелико, описанная процедура получения из таблицы критических значений W_1 и W_2 остается правомерной.

Ситуация 2. Объемы наблюдений в выборках превышают 25. Согласно [8] критические значения W_1 и W_2 определяют по следующим приближенным формулам:

$$W_1 = \{0{,}5\,[m(m+n+1)-1] - |t_{\alpha/2}|\sqrt{1/12mn(m+n+1)}\};$$

$$W_2 = m(m+n+1) - W_1,$$

где $m = n_1$, $n = n_2$, $n_1 \leqslant n_2$, $t_{\alpha/2}$ — квантиль гауссовского (нормального) распределения, причем двустороннему уровню значимости $\alpha = 0{,}05$ соответствует квантиль $t_{\alpha/2} = 1{,}96$.

При наличии совпадающих значений формула для W_1 может быть взята из [8, 19].

Для обеих ситуаций проверяемая гипотеза $H_0 : a_1 = a_2$ принимается как не противоречащая исходным данным, если вычисленная статистика W не выйдет за пределы, образованные критическими значениями W_1 и W_2, и отклоняется как неподтвердившаяся и тем самым принимаются альтернативы $H_1 : a_1 \neq a_2$, если статистика W окажется за допустимыми пределами W_1 и W_2; $H_0 : a_1 = a_2$, если $W_1 \leqslant W \leqslant W_2$; $H_1 : a_1 \neq a_2$, если $W < W_1$ или $W > W_2$.

Критерий Вэлча предназначен для проверки гипотезы о средних $H_0 : a_1 = a_2$ при наборе альтернатив $H_1 : a_1 \neq a_2$. Критерий использует предположения о нормальности распределений случайных величин — моделей изучаемых геологических признаков в сравниваемых объектах, об отсутствии аномальных наблюдений и некоторые другие.

Для целей проверки гипотезы о равенстве средних при не очень малых объемах наблюдений n_1 и n_2 в выборках следует воспользоваться статистикой Вэлча [19, 26]:

$$t = |\bar{x} - \bar{y}| / \sqrt{S_1^2/n_1 + S_2^2/n_2}, \quad \text{где } \bar{x} = \frac{1}{n_1}\sum_{t=1}^{n_1} x_t;$$

$$\bar{y} = \frac{1}{n_2}\sum_{t=1}^{n_2} y_t; \qquad S_1^2 = \frac{1}{n_1 - 1}\sum_{t=1}^{n_1}(x_t - \bar{x})^2;$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{t=1}^{n_2}(y_t - \bar{y})^2.$$

В условиях нулевой гипотезы $H_0 : a_1 = a_2$ величина t распределена асимптотически по закону Стьюдента с f степенями свободы

$$f = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\left(\frac{S_1^2}{n_1}\right)^2/(n_1+1)\right] + \left[\left(\frac{S_2^2}{n_2}\right)^2/(n_2+1)\right]} - 2\right],$$

где символ $[\cdot]$ (выражение в квадратных скобках) означает взятие целой части от числа.

Нулевая гипотеза $H_0 : a_1 = a_2$ принимается как подтвердившаяся, если вычисленная t-статистика Вэлча не превысит допустимый квантиль $t_{\alpha, f}$ распределения Стьюдента при заданном уровне значимости α и f степенях свободы, т. е. $t \leqslant t_{\alpha, f}$. Проверяемая гипотеза отклоняется и принимаются альтернативы о существенности различий в средних $H_1 : a_1 \neq a_2$, если $t > t_{\alpha, f}$.

Критерий Сиджела — Тьюки для проверки гипотез о равенстве дисперсий предназначен для проверки гипотезы: $H_0 : \sigma_1^2 = \sigma_2^2$ против набора альтернатив $H_1 : \sigma_1^2 \neq \sigma_2^2$, где σ_1^2 и σ_2^2 — истинные дисперсии для первого и второго объектов. Статистика Сиджела—Тьюки нечувствительна к нарушению условий нормальности распределения наблюдений, наличию аномальных значений и т. п. Она является полным аналогом статистики Вилкоксона, но проверка осуществляется в этом случае относительно параметра масштаба (дисперсии), а не параметра сдвига (среднего) [8, 26].

Учитывая это обстоятельство, можно для проверки нулевой гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ пользоваться теми же критическими значениями W_1 и W_2, что и в случае применения критерия Вилкоксона. Это, безусловно, удобно для практических расчетов при обработке геологических данных.

В некоторых крупных методических руководствах приводятся специальные таблицы критических значений R-критерия, которыми также можно пользоваться (они полностью совпадают с критическими значениями Вилкоксона).

Отличие критерия Сиджела—Тьюки от критерия Вилкоксона заключается в ином характере ранжирования выборочных данных. Номер (ранг) 1 приписывается наименьшему члену вариационного ряда, номер 2 — наибольшему, номер 3 — второму максимальному, номер 4 — второму наименьшему. Процедура ранжирования продолжается аналогичным способом. Если $n_1 + n_2$ нечетно, то медианный член устраняется.

Для применения R-критерия Сиджела—Тьюки следует убедиться в равенстве параметров сдвига (равенстве средних); если равенство средних не имеет места, то следует центрировать выборочные данные, например медианами.

Известны две схемы применения рангового критерия Сиджела—Тьюки.

Схема А.

1. С помощью критериев Вилкоксона или Вэлча убеждаемся в равенстве средних для двух сравниваемых объектов. При отсутствии сдвига можно пользоваться исходными данными, в противном случае — наблюдения (анализы проб) в обеих выборках центрируются своими медианами. Дальнейшие операции осуществляются тогда с центрированными данными.

2. Составляется общий вариационный ряд $N = n_1 + n_2$ в порядке возрастания всех исходных центрированных членов.

3. Вышеупомянутым специальным способом (ранг 1 — наименьшему члену, ранг 2 — наибольшему, ранг 3 — второму наибольшему, ранг 4 — второму наименьшему и т. д.) производится ран-

жирование всех членов общего вариационного ряда. Если число наблюдений нечетно, то среднее наблюдение (медиана) не получает никакого ранга, если четное — оно получает наивысший ранг.

4. Равным значениям (совпадающим членам) дается скорректированный средний ранг, представляющий собой среднее арифметическое рангов совпадающих членов вариационного ряда.

5. Статистика R-критерия Сиджела—Тьюки представляет собой сумму ранговых чисел, т. е. сумму рангов r_i, относящихся к членам меньшей по объему выборки:

$$R = \sum_{i=1}^{n_1} r_i, \qquad n_1 \leqslant n_2.$$

6. Аналогично процедуре применения критерия Вилкоксона определяют критические значения W_1 и W_2.

7. Проверяемая гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ принимается как не противоречащая выборочным данным, если вычисленная статистика R не выйдет за пределы, образованные критическими значениями W_1 и W_2 ($W_1 \leqslant R \leqslant W_2$), и отклоняется как неподтвердившаяся, если статистика R окажется за допустимыми пределами W_1 и W_2 ($R < W_1$ или $R > W_2$).

Схема Б.

Пункты 1—5 полностью совпадают с пунктами 1—5 схемы А.

6. Для не слишком малых выборок (n_1 и $n_2 > 9$) различия в дисперсиях ($H_1 : \sigma_1^2 \neq \sigma_2^2$) с достаточной точностью определяются с помощью стандартизованной нормальной переменной:

$$t = \frac{2R - n_1(n_1 + n_2 + 1) + \delta}{\sqrt{\frac{n_1 n_2}{3}(n_1 + n_2 + 1)}},$$

где $$\delta = \begin{cases} 1, & \text{если } 2R > n_1(n_1 + n_2 + 1), \\ -1, & \text{если } 2R \leqslant n_1(n_1 + n_2 + 1). \end{cases}$$

При сильно различающихся объемах выборок n_1 и n_2 следует пользоваться скорректированным выражением

$$t^* = 1 + \left(\frac{1}{10n_1} + \frac{1}{10n_2}\right)(t^3 - 3t).$$

Если пятая часть и более наблюдений связаны равенствами, то формула для t усложняется [19, 26].

7. Проверяемая гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ принимается как подтвердившаяся, если $|t| \leqslant t_{\alpha/2}$, и отклоняется и тем самым принимаются альтернативы $H_1 : \sigma_1^2 \neq \sigma_2^2$, если величина $|t|$ превысит допустимое $t_{\alpha/2}$ (при $\alpha = 0{,}05$ $t_{\alpha/2} = 1{,}96$).

Параметрические F-критерий Фишера и критерий Бартлета для проверки гипотез о равенстве дисперсий осуществляют проверку гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ при множестве альтернатив $H_1 : \sigma_1^2 \neq \sigma_2^2$ [9, 25]. Выбор этих двух параметрических

методов обусловлен необходимостью применения в задачах классификаций k ($k > 2$) геологических объектов критерия, критическое значение которого не зависело бы от объемов выборок. К сожалению, более простой в расчетной части и эффективный F-критерий Фишера этим свойством не обладает, и поэтому он рекомендуется к применению только при сопоставлении двух объектов. Для задач классификаций при $k > 2$ следует привлекать критерий Бартлета.

Применение F-критерия Фишера базируется на предположении о нормальности распределения случайных величин ξ и η — моделей геологических признаков.

Вычисляется F-статистика, представляющая собой отношение большей выборочной дисперсии к меньшей:

$$F = \frac{S_1^2}{S_2^2}, \quad \text{если } S_1^2 \geqslant S_2^2, \quad \text{и} \quad F = \frac{S_2^2}{S_1^2}, \quad \text{если } S_2^2 \geqslant S_1^2.$$

В условиях нулевой гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ величина F распределена по закону Фишера с $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$ степенями свободы.

Нулевая гипотеза считается подтвердившейся, т. е. не противоречащей эмпирическим данным, если рассчитанная величина F не превысит допустимого F_{α, f_1, f_2}, т. е. $F \leqslant F_{\alpha, f_1, f_2}$, соответствующего заданному двустороннему уровню значимости α при $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$ степенях свободы (для случая $S_1^2 \geqslant S_2^2$). Если же вычисление F превысит критическое, т. е. при $F > F_{\alpha, f_1, f_2}$, то нулевую гипотезу следует отклонить как противоречащую исходным данным и принять альтернативные гипотезы о существенности различий в истинных дисперсиях $H_1 : \sigma_1^2 \neq \sigma_2^2$. В этом случае более высокую истинную дисперсию (степень рассеяния) следует ожидать у объекта с более высокой выборочной дисперсией S^2.

Для критерия Бартлета условия применения те же, что и для критерия Фишера. Ниже описана упрощенная процедура применения критерия Бартлета, а более строгое ее изложение дано в работе Л. Н. Большева и Н. В. Смирнова [8].

Величина M для двух объектов определяется с помощью выражения

$$M = \frac{1 + \beta}{1 + 2\beta}\left[(n_1 - 1)\ln\frac{S^2}{S_1^2} + (n_2 - 1)\ln\frac{S^2}{S_2^2}\right] =$$

$$= \frac{2{,}3026}{C}\left[(N-2)\lg S^2 - (n_1 - 1)\lg S_1^2 - (n_2 - 1)\lg S_2^2\right],$$

где $$S^2 = \frac{1}{N-2}\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right],$$

$$C = 1 + \frac{1}{3}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{N-2}\right), \qquad N = n_1 + n_2.$$

В условиях нулевой гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ величина M распределена асимптотически по закону Пирсона χ^2 с 1-й степенью свободы. Нулевая гипотеза считается подтвердившейся, т. е. не противоречащей эмпирическим данным, если рассчитанная величина не превысит допустимого $\chi^2_{\alpha, f=1}$, т. е. $M \leqslant \chi^2_{\alpha, f}$, соответствующая заданному двустороннему уровню значимости α к $f = 1$ степени свободы. Нулевая гипотеза отклоняется как неподтвердившаяся, если вычисленное значение M превысит критическое χ^2, т. е. $M > \chi^2_{\alpha, f}$. В этом случае следует принять альтернативные гипотезы о существенности различий в истинных дисперсиях H_1 : $\sigma_1^2 \neq \sigma_2^2$ и полагать, что степень рассеяния выше у объекта, характеризующегося более высокой выборочной дисперсией S^2.

Критерий Пури—Сена—Тамуры применяется для проверки гипотез о равенстве многомерных средних в двух объектах. Этот ранговый критерий устойчив относительно нарушения условия предполагаемой нормальности (и даже унимодальности) распределения изучаемых случайных величин, а также относительно наличия в сопоставляемых выборках аномальных наблюдений.

Согласно Пури и Сену [26, 51], в качестве рабочих статистик критерия для проверки многомерных средних можно привлечь:

1) ранги $E_{tj} = R_{tj}$; тогда критерий Пури—Сена—Тамуры следует рассматривать как многомерный аналог одномерного критерия Вилкоксона:

2) нормальные метки, аналогичные участвующие в одномерном критерии Фишера—Йэйтса—Терри—Гефдинга [1, 26];

3) метки в виде обратных функций нормального распределения $E_{tj} = \Phi^{-1}(R_{tj}/(N+1))$, аналогичные тем, которые участвуют в одномерном критерии Ван дер Вардена [1, 26].

Процедура применения критерия Пури—Сена—Тамуры следующая.

1. По каждому геологическому признаку в отдельности сопоставляется общий вариационный ряд в порядке возрастания его членов, аналогично тому, как это производится при процедуре применения критерия Вилкоксона. Все члены нумеруются 1, 2, . . . , $(n_1 + n_2)$, т. е. определяются метки-ранги $E_{tj} = R_{tj}$; $t = 1, 2, \ldots, N$ $(N = n_1 + n_2)$; $j = 1, 2, \ldots, m$.

2. Так же аналогично критерию Вилкоксона равным значениям ставится в соответствие скорректированный ранг (метка) — среднее арифметическое из рангов R_{tj}.

Заметим, что уточненный средний ранг (среднюю метку) следует вводить лишь тогда, когда равные значения присутствуют в обеих выборках; а если они принадлежат одной выборке, то можно не вычислять скорректированный ранг (так как ранговая сумма для каждой выборки не будет изменяться в этом случае).

3. Определяются два m-мерных вектора средних меток-рангов T_1 и T_2:

$$T_1 = (T_{11}, T_{12}, \ldots, T_{1j}, \ldots, T_{1m}), \qquad j = 1, 2, \ldots, m,$$

$$T_2 = (T_{21}, T_{22}, \ldots, T_{2j}, \ldots, T_{2m}),$$

где $$T_{1j} = \frac{1}{n_1} \sum_{t=1}^{n_1} R_{tj}; \qquad T_{2j} = \frac{1}{n_2} \sum_{t=1}^{n_2} R_{tj}.$$

4. Определяется m-мерный вектор средних меток-рангов по всей объединенной выборке объема $N = n_1 + n_2$:

$$\bar{E} = (\bar{E}_1, \bar{E}_2, \ldots, \bar{E}_j, \ldots, \bar{E}_m), \qquad j = 1, 2, \ldots, m,$$

где $$E_j = \frac{1}{N} \sum_{t=1}^{N} R_{tj}, \qquad \bar{E}_1 = \bar{E}_2 = \ldots = \bar{E}_j = \ldots = \bar{E}_m.$$

5. Составляется ковариационная матрица меток V размерностью $m \times m$:

$$V = \{V_{ij}\}, \qquad i, j = 1, 2, \ldots, m,$$

где $$V_{ij} = \frac{1}{N} \sum_{t=1}^{N} (R_{ti} - \bar{E}_i)(R_{tj} - \bar{E}_j).$$

Заметим, что хотя $\bar{E}_i = \bar{E}_j$, в общем случае $R_{ti} \neq R_{tj}$, т. е. ранги R_{tj} первой (второй, . . . , N-й) пробы по признаку i не зависят от рангов R_{tj} первой (второй, . . . , N-й) пробы по признаку j.

6. Определяется обратная матрица $V^{-1} = \{V^{ij}\}$.

7. Вычисляется статистика Пури—Сена—Тамуры для проверки гипотезы о равенстве многомерных средних в двух объектах, представляющая собой квадратичную форму:

$$\Lambda = \sum_{u=1}^{2} n_u (T_u - \bar{E}) V^{-1} (T_u - \bar{E})',$$

где штрихом обозначается знак транспонирования разности вектора-строки $T_u - \bar{E}$.

8. В условиях нулевой гипотезы о равенстве многомерных средних в двух объектах статистика Λ распределена по закону Пирсона χ^2 с m степенями свободы.

Таким образом, если окажется $\Lambda \leqslant \chi^2_{\alpha, m}$, то для заданного уровня значимости α принимается нулевая гипотеза как подтвердившаяся. В противном случае, если $\Lambda > \chi^2_{\alpha, m}$, нулевая гипотеза отклоняется и принимаются альтернативы о существенности различий в многомерных средних сравниваемых двух объектов.

Критерий Джеймса—Сю для проверки гипотез о равенстве многомерных средних в двух объектах базируется на предположении о многомерном нормальном распределении случайных величин и отсутствии аномальных наблюдений, а также не предполагает равенства ковариационных матриц [26, 28].

Процедура применения критерия Джеймса—Сю следующая.

1. По двум исходным m-мерным выборочным данным объема n_1 и n_2 соответственно рассчитываются векторы средних арифме-

тических $\bar{x}^{(1)}$ и $\bar{x}^{(2)}$ и оценки ковариационных матриц $S^{(1)}$ и $S^{(2)}$ по каждой выборке:

$$\bar{X}^{(u)} = \{\bar{x}_j^{(u)}\}, \qquad \text{где } \bar{x}_j^{(u)} = \frac{1}{n_u}\sum_{i=1}^{n_u} x_{ij}^{(u)}, \qquad j = 1, 2, \ldots, m;$$

$$u = 1, 2; \qquad S^{(u)} = \{\hat{\sigma}_{ij}^{(u)}\}, \qquad u = 1, 2,$$

$$\text{где } \hat{\sigma}_{ij}^{(u)} = \frac{1}{n_u - 1}\left(x_{ti}^{(u)} - \bar{x}_i^{(u)}\right)\left(x_{tj}^{(u)} - \bar{x}_j^{(u)}\right),$$

$$i, j = 1, 2, \ldots, m; \qquad u = 1, 2.$$

2. Рассчитываются разности векторов средних арифметических:

$$\bar{X}^{(1)} - \bar{X}^{(2)} = \{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}\}, \qquad j = 1, 2, \ldots, m.$$

3. Рассчитывается оценка обобщенной ковариационной матрицы:

$$S = \{\hat{\sigma}_{ij}\} = \frac{S_1}{n_1} + \frac{S_2}{n_2} = \left\{\left(\frac{\hat{\sigma}_{ij}^{(1)}}{n_1} + \frac{\hat{\sigma}_{ij}^{(2)}}{n_2}\right)\right\},$$

$$i, j = 1, 2, \ldots, m.$$

4. Рассчитывается статистика Джеймса—Сю, представляющая собой квадратичную форму:

$$2I = \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)' S^{-1}\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right).$$

В условиях нулевой гипотезы о равенстве многомерных средних в двух объектах статистика $2I$ асимптотически распределена по закону Пирсона χ^2 с m степенями свободы. В работах [28, 39] даны более точные распределения статистики $2I$ критерия Джеймса—Сю в условиях нулевой гипотезы.

Поэтому, если окажется $2I \leqslant \chi^2_{\alpha, m}$, то для заданного уровня значимости α принимается нулевая гипотеза о равенстве многомерных средних как подтвердившаяся. В противном случае, если $2I > \chi^2_{\alpha, m}$, то нулевая гипотеза должна быть отклонена как противоречащая эмпирическим данным и приняты альтернативные гипотезы о существенности различий в многомерных средних сравниваемых двух объектов.

Критерий Пури—Сена проверки гипотез о равенстве ковариационных матриц в двух объектах [26, 51] базируется на предположении, что многомерные случайные величины (модели комплекса m геологических признаков в сопоставляемых объектах) имеют одинаковые медианы, поэтому, можно в качестве меток-статистик рассматриваемого критерия для проверки гипотезы о равенстве ковариационных матриц использовать:

1) модуль нормированных и центрированных рангов

$$\mathrm{E}_{tj} = \left|\frac{R_{tj}}{N+1} - 0{,}5\right|,$$

и тогда критерий Пури—Сена—Тамуры выступает как многомерный аналог одномерного критерия масштаба Ансари—Бредли [1, 26, 51];

2) квадрат нормированных и центрированных рангов

$$\mathrm{E}_{tj} = \left(\frac{R_{tj}}{N+1} - 0{,}5\right)^2,$$

и тогда критерий Пури—Сена—Тамуры следует рассматривать как многомерный аналог одномерного критерия масштаба Муда [1, 26, 51];

3) квадрат обратной функции нормального распределения

$$\mathrm{E}_{tj} = \left[\Phi^{-1}\left(\frac{R_{tj}}{N+1}\right)\right]^2,$$

и тогда критерий Пури—Сена—Тамуры выступает как многомерный аналог одномерного критерия Клотца [1, 26, 51];

4) квадрат j-й порядковый статистики выборки объема n из стандартного нормального распределения (1, 26, 51].

Второй вид метод (аналог критерия Муда), являющийся наиболее простым, описан ниже.

Обращаем внимание, что условие неравенства медиан означает необходимость центрирования исходных геологических данных медианами по каждому признаку аналогично тому, как рекомендовано при применении одномерного критерия масштаба Сиджела—Тьюки.

Процедура применения критерия Пури—Сена—Тамуры для проверки гипотезы о равенстве ковариационных матриц следующая.

1. По каждой выборке и каждому геологическому признаку в отдельности определяем медианы.

2. Центрируем исходные данные медианами.

3. По каждому геологическому признаку в отдельности по центрированным медианами данным составляется вариационный ряд в порядке возрастания его членов, аналогично тому, как это производится при процедуре применения критерия Сиджела—Тьюки. Все члены нумеруются 1, 2, ..., t, ..., N ($N = n_1 + n_2$), т. е. определяются ранги R_{tj}, $t > 1, 2, \ldots, N$; $j = 1, 2, \ldots, m$.

4. Аналогично статистике Муда для каждого ранга R_{tj} находим соответствующую ему метку E_{tj}: $\mathrm{E}_{tj} = [R_{tj}/(N+1) - 0{,}5]^2$.

5. В разных выборках (в одной выборке можно не исправлять) равным значениям центрированных медианами исходных данных ставится в соответствие скорректированная средняя метка — среднее арифметическое из меток для равных значений.

6—9. Процедура полностью аналогична пунктам 3—6 алгоритма вычисления статистики Пури—Сена—Тамуры для проверки гипотезы о равенстве многомерных средних.

Аналогично находим два вектора средних меток T_1 и T_2 и обратную матрицу V^{-1}.

10. Находим статистику Пури—Сена—Тамуры для проверки гипотезы о равенстве ковариационных матриц в двух объектах:

$$\Lambda_{\Sigma} = \sum_{u=1}^{2} n_u (T_u - \overline{E}) v^{-1} (T_u - \overline{E})', \qquad u = 1, 2.$$

11. В условиях нулевой гипотезы о равенстве ковариационных матриц в двух объектах статистика Λ_{Σ} распределена по закону Пирсона χ^2 с $f = m$ степенями свободы.

Поэтому, если окажется $\Lambda_{\Sigma} \leqslant \chi^2_{\alpha, f}$, то для заданного уровня значимости α нулевая гипотеза о равенстве ковариационных матриц в двух объектах принимается как подтвердившаяся. В противном случае, когда $\Lambda_{\Sigma} > \chi^2_{\alpha, f}$, нулевую гипотезу следует отклонить и принять альтернативные — о существенности различий в ковариационных матрицах сравниваемых объектов. Другими словами, меры рассеяния и зависимости геологических характеристик в сравниваемых геологических объектах статистики значимо различаются.

Критерий Кульбака для проверки гипотез о равенстве ковариационных матриц в двух объектах является своеобразным многомерным аналогом одномерного критерия Бартлета и учитывает не только дисперсии, но и ковариации признаков. Ограничения в применении критерия Кульбака полностью аналогичны вышеупомянутым при описании критерия Джеймса—Сю.

Процедура применения критерия Кульбака следующая [26, 28].

1. Полностью аналогично п. 1 процедуры вычисления статистики Джеймса—Сю рассчитываются оценки S_1 и S_2 ковариационных матриц по каждой выборке в отдельности.

2. Рассчитывается оценка S обобщенной ковариационной матрицы:

$$S = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1) S_1 + (n_2 - 1) S_2].$$

3. Рассчитываются определители $|S_1|$, $|S_2|$, $|S|$ трех выборочных ковариационных матриц S_1, S_2, S.

Вычисляется критерий Кульбака $2I_0$:

$$2I_0 = (n_1 - 1) \ln \frac{|S|}{|S_1|} + (n_2 - 1) \ln \frac{|S|}{|S_2|}.$$

В условиях нулевой гипотезы о равенстве ковариационных матриц в двух объектах $H_0 : \Sigma_1 = \Sigma_2$ статистика $2I_0$ распределена асимптотически по закону Пирсона χ^2 с $f = 0{,}5\, m\, (m + 1)$ степенями свободы. С более точным распределением статистики $2I_0$ в условиях нулевой гипотезы следует ознакомиться в работе [28].

Поэтому, если окажется $2I_0 \leqslant \chi^2_{\alpha, f}$, то для заданного уровня значимости α принимается как подтвердившаяся нулевая гипотеза о равенстве ковариационных матриц в двух объектах. Наоборот, если $2I_0 > \chi^2_{\alpha, f}$, то нулевую гипотезу следует отклонить и принять

альтернативные гипотезы о существенных отличиях ковариационных матриц в первом и втором объектах. Другими словами, в случае принятия альтернативы следует полагать, что характеристики рассеяния и зависимости между изучаемыми геологическими признаками в сопоставляемых объектах значимо различаются.

Статистические методы разграничения геологических объектов

Статистические методы разграничения геологических объектов — совокупность приемов статистической обработки многомерных данных, приводящая в итоге к разделению изучаемого набора наблюдений на некоторое число статистически однородных, отличающихся друг от друга групп.

Для геологии типична ситуация, когда относительно имеющегося набора многомерных наблюдений заранее неизвестно, является ли он однородным, т. е. состоит только из одной группы, или неоднородным, и тогда на какое число однородных групп его следует разделить, и какой состав этих групп. Такая задача возникает при самых разнообразных геологических исследованиях — геохимических, петрографических, биостратиграфических, палеонтологических и др. Следует особо подчеркнуть, что задача разграничения совокупности наблюдений на однородные группы принципиально отличается по своей постановке от дискриминантного анализа, в котором группы априори заданы, тогда как в задаче разграничения они неизвестны и их следует определить. Таким образом, задача разграничения должна предшествовать дискриминантному анализу.

Существует два типа задач разграничения. Первый соответствует ситуации, когда результаты наблюдения строго зафиксированы на одной линии, т. е. представляют собой линейно упорядоченную последовательность; второй — ситуации, когда наблюдения расположены на плоскости или в трехмерном объеме. Первый тип задачи характерен, например, для расчленения стратиграфических разрезов, тогда как примером второго типа может служить задача составления геохимической карты по комплексу химических элементов.

Тем не менее в условиях задач обоих типов исходные данные представляют собой матрицу, содержащую n строк и m столбцов, где n — число наблюдений, m — число признаков, т. е.

$$\mathbf{X} = \begin{Vmatrix} X_1 \\ X_2 \\ \cdots \\ X_t \\ \cdots \\ X_n \end{Vmatrix} = \begin{Vmatrix} x_{11} x_{12} \ldots x_{1j} \ldots x_{1m} \\ x_{21} x_{22} \ldots x_{2j} \ldots x_{2m} \\ \ldots\ldots\ldots\ldots \\ x_{t1} x_{t2} \ldots x_{tj} \ldots x_{tm} \\ \ldots\ldots\ldots\ldots \\ x_{n1} x_{n2} \ldots x_{nj} \ldots x_{nm} \end{Vmatrix}.$$

Одно многомерное наблюдение X_t в этой матрице представляет собой m-мерный вектор-строку, т. е.

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\},$$
$$t = 1, 2, \ldots, n; \qquad j = 1, 2, \ldots, m.$$

Каждое такое наблюдение можно рассматривать как одно значение некоторой m-мерной случайной величины, имеющей неизвестное m-мерное среднее μ_t, т. е.

$$\mu_t = \{\mu_{t1}, \mu_{t2}, \ldots, \mu_{tj}, \ldots, \mu_{tm}\}.$$

Формально задачу разграничения можно сформулировать как проверку гипотезы

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_t = \ldots = \mu_n = \mu_0$$

при альтернативе $H_1 : \mu_t \neq \mu_0$ хотя бы для одного $t = 1, 2, \ldots, n$, по результатам наблюдений X_t.

Если в результате проверки окажется, что следует принять H_0, то из этого следует, что изучаемый набор наблюдений разделять на группы нельзя, так как он является однородным. Если же будет принята альтернатива H_1, то это значит, что рассматриваемый набор объектов можно разделить на две или более однородные группы. Путем последовательной процедуры деления неоднородности набора объектов на две части достигается разделение на однородные отличающиеся одна от другой группы. Подробно эти вопросы приведены в работе Д. А. Родионова [39]. Ниже приведены два упомянутых алгоритма разграничения.

I. Алгоритм разграничения линейно упорядоченной последовательности наблюдений.

1. Для каждого из n—1 вариантов разделения последовательности наблюдений на две части из k и n—k наблюдений соответственно вычисляется n—1 значений статистического критерия:

$$v_k = \frac{n-1}{nk(n-k)} \sum_{j=1}^{m} \frac{\left[(n-k)\sum_{t=1}^{k} x_{tj} - k\sum_{t=k+1}^{n} x_{tj}\right]^2}{\sum_{t=1}^{n} x_{tj}^2 - \frac{1}{n}\left(\sum_{t=1}^{n} x_{tj}\right)^2},$$

$$k = 1, 2, \ldots, (n-1).$$

2. Если гипотеза об однородности верна, то v_k будут представлять собой значения случайных величин, распределенных как χ^2 с m степенями свободы. Поэтому гипотеза H_0 принимается, т. е. совокупность n m-мерных наблюдений рассматривается как однородная и разграничению не подлежит, если

$$\max_k v_k \leqslant \chi^2_{q,\,m},$$

где $\chi^2_{q,\,m}$ — значение χ^2, соответствующее уровню значимости q и m степеням свободы.

Если же

$$\max_k v_k > \chi^2_{q,\,m},$$

то гипотезу об однородности следует отклонить и перейти к поиску соответствующих границ в последовательности наблюдений.

3. Если $\max_k v_k > \chi^2_{q,\,m}$, то последовательность наблюдений делится на две части в точке, соответствующей $\max v_k$, и каждая часть анализируется отдельно описанным выше способом. Такая процедура деления неоднородных участков последовательности на две части продолжается до тех пор, пока вся совокупность наблюдений не будет разделена на статистические однородные группы.

4. В результате описанных выше операций последовательность наблюдений с множеством T значений индекса t будет разделена на смежные непересекающиеся подмножества

$$T_1, T_2, \ldots, T_h, \ldots, T_S, \quad \text{т. е.}$$

$$\bigcup T_h = T, \quad T_h \cap T_{h+1} = \varnothing \quad \text{(пустое множество)}$$
$$\text{для всех } h = 1, 2, \ldots, S.$$

Однако некоторые из установленных таким образом границ могут оказаться ложными, и заключительный этап работ состоит в выявлении и устранении таких ложных разграничений.

Для всех смежных пар T_h, T_{h+1}, начиная с $h = 1$, вычисляется значение критерия

$$v(T_h, T_{h+1}) = \frac{n_h + n_{h+1} - 1}{n_h n_{h+1}(n_h + n_{h+1})} \times$$
$$\times \sum_{j=1}^{m} \frac{\left(n_{h+1}\sum_{t\in T_h} x_{tj} - n_h \sum_{t\in T_{h+1}} x_{tj}\right)^2}{\sum_{t\in T_h \cup T_{h+1}} x_{tj}^2 - \frac{1}{(n_h + n_{h+1})}\left(\sum_{t\in T_h \cup T_{h+1}} x_{tj}\right)^2}.$$

Если $v(T_h, T_{h+1}) > \chi^2_{q,\,m}$, то граница сохраняется и производится аналогичная проверка для T_{h+1} и T_{h+2}. Если же $v(T_h, T_{h+1}) \leqslant \chi^2_{q,\,m}$, то граница рассматривается как ложная. Тогда T_h и T_{h+1} объединяются и проводится проверка для $T_h \cup T_{h+1}$ и T_{h+2}. В итоге будет получена новая последовательность $T'_1, T'_2, \ldots, T'_h, \ldots, T'_l$ при $l \leqslant S$, которая уже не будет содержать ложных разграничений.

II. Алгоритм разграничения набора m-мерных наблюдений, расположенных на плоскости или в трехмерном объеме.

А. Проверка гипотезы об однородности

1. Дана выборка, состоящая из n m-мерных наблюдений:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_t \\ \dots \\ X_n \end{pmatrix} = \begin{pmatrix} x_{11}x_{12} \dots x_{1j} \dots x_{1m} \\ x_{21}x_{22} \dots x_{2j} \dots x_{2m} \\ \dots \\ x_{t1}x_{t2} \dots x_{tj} \dots x_{tm} \\ \dots \\ x_{n1}x_{n2} \dots x_{nj} \dots x_{nm} \end{pmatrix},$$

где $X_t = \{x_{t1}, x_{t2}, \dots, x_{tj}, \dots, x_{tm}\}$ — m-мерный вектор-строка, $t = 1, 2, \dots, n$.

Вся выборка представляет собой матрицу порядка $n \times m$. Множество значений t будем обозначать, как и раньше, через T.

2. Рассматривается n вариантов разбивки совокупности n наблюдений на две части, причем одна из них содержит только одно наблюдение X_t, а другая — оставшиеся $n-1$ наблюдений. Для каждого из n вариантов такой разбивки на множества A^1 и A^{n-1} вычисляется значение критерия

$$v(A^1, A^{n-1}) = \frac{1}{n} \sum_{j=1}^{m} \frac{\left[(n-1)\, x_{t1} - \sum\limits_{t \in A^{n-1}} x_{tj}\right]^2}{\sum\limits_{t \in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t \in T} x_{tj}\right)^2}.$$

Из всех n значений критерия $v(A^1, A^{n-1})$ выбирается максимальное, чем определяется соответствующее этому максимуму наблюдение X_t.

3. Рассматриваются все $n-1$ пары, образованные X_t и оставшимися $n-1$ наблюдениями, и соответствующие им $n-1$ вариантов разбивки пространства T на два подмножества A^2 и A^{n-2}. Для каждой такой разбивки вычисляются значения критерия, т. е.:

$$v(A^2, A^{n-2}) = \frac{n-1}{2(n-2)\, n} \sum^{m} \frac{\left[(n-2) \sum\limits_{t \in A^2} x_{tj} - 2 \sum\limits_{t \in A^{n-2}} x_{tj}\right]^2}{\sum\limits_{t \in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t \in T} x_{tj}\right)^2},$$

и определяется тот вариант из $n-1$ вариантов, которому соответствует $\max v(A^2, A^{n-2})$. Таким образом, устанавливается пара наблюдений X_{t_1}, X_{t_2}, включающая X_{t_1}, выявленное на предыдущем этапе.

4. Эта процедура продолжается до тех пор, пока не будет достигнута разбивка на $n/2$ наблюдений в случае четного n, и на $(n-1)/2$ и $(n+1)/2$ при нечетном n. Таким образом, для любого

$k \leqslant n/2$ при четном n и $k \leqslant (n-1)/2$ при нечетном n вычисляется значение критерия

$$v(A^k, A^{n-k}) = \frac{(n-1)}{k(n-k)\, n} \sum_{j=1}^{m} \frac{\left[(n-k) \sum\limits_{t \in A^k} x_{tj} - k \sum\limits_{t \in A^{n-k}} x_{tj}\right]^2}{\sum\limits_{t \in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t \in T} x_{tj}\right)^2}.$$

При этом множество A^k включает $n-1$ наблюдений, обеспечивающих максимальное значение критерия на $k-1$ предыдущей стадии вычислений.

5. В результате будет получена последовательность $n/2$ или $(n-1)/2$ максимальных значений критерия, полученных на $n/2$ или $(n-1)/2$ стадиях вычислений. Из всех этих значений выбирается максимальное, которому соответствует разбивка T на

$$A^{k^*} \text{ и } A^{n-k^*},$$

т. е. отыскивается значение

$$\max_{k} \max_{A^k \in A} v(A^k, A^{a \neq k}) = v(k^*),$$

где A^k — класс всех множеств, включающих выбранную на предыдущей стадии комбинацию $k-1$ наблюдений.

6. Если $v(k^*) \leqslant \chi^2_{q,\,m}$,

где $\chi^2_{q,\,m}$ — заданное значение χ^2, соответствующее уровню значимости q и m степеням свободы, то дальнейшие вычисления прекращаются, так как для данного набора наблюдений гипотеза об однородности не отклоняется, из чего следует, что любые разграничения этой совокупности не имеют смысла.

Если же $v(k^*) > \chi^2_{q,\,m}$,

то гипотеза об однородности набора наблюдений отклоняется, из чего следует, что изучаемую совокупность наблюдений нужно разделить не менее чем на две части. При этом выбирается тот вариант разбивки на две части, который соответствует $v(k^*)$.

Необходимо отметить, что в практической работе значительно удобнее пользоваться отношением

$$\tau = [v(k^*) - m]/\sqrt{2m}.$$

Вычислительные процедуры прекращаются и совокупность рассматривается как однородная, если $\tau \leqslant 3$, и гипотеза об однородности отклоняется, если $\tau > 3$.

Б. Поиск границ

7. Если гипотеза об однородности изучаемой совокупности наблюдений отклонена, то эта совокупность делится на две части в соответствии с $v(k^*)$.

8. Каждая из двух новых совокупностей анализируется отдельно по алгоритму, описанному в первой части, в результате чего принимается решение об однородности или неоднородности каждой из совокупностей. Если для какой-либо из этих совокупностей гипотеза об однородности принимается, то дальнейшие вычисления для нее прекращаются. Если же принимается альтернатива, то данная совокупность снова делится на две части, в соответствии с правилом, изложенным в п. 7, и анализ вновь полученных совокупностей продолжается.

9. Процедура такого дихотомического деления изучаемой совокупности продолжается до тех пор, пока во всех выделенных более дробных совокупностях не будет принята гипотеза об однородности. Однако некоторые из полученных разграничений могут оказаться ложными, и поэтому нужно перейти к третьей части алгоритма — устранению ложных границ.

В. Устранение ложных границ

10. В результате проведенных вычислений изучаемая выборка, объем которой n, будет разделена на h групп наблюдений. Обозначим через $T_1, T_2, \ldots, T_l, \ldots, T_h$ — непересекающиеся подмножества в T, которые соответствуют выделенным группам наблюдений.

11. Из упомянутых h групп наблюдений можно образовать, $h\,(n-1)/2$ пар и для каждой из них вычислять значение критерия

$$v(T_l,\ T_S) = \frac{n_l + n_S - 1}{n_l n_S\,(n_l + n_S)} \times$$

$$\times \sum_{j=1}^{m} \frac{\left(n_S \sum\limits_{t \in T_l} x_{tj} - n_l \sum\limits_{t \in T_S} x_{tj}\right)^2}{\sum\limits_{t \in T_l \cup T_S} x_{tj}^2 - \frac{1}{n_l + n_S}\left(\sum\limits_{t \in T_l \cup T_S} x_{tj}\right)^2}.$$

В результате будет получена треугольная матрица, содержащая $h\,(n-1)/2$ значений критерия:

$$\begin{Vmatrix} v(T_1, T_2 \ldots v(T_1, T_s) \ldots v(T_1, T_h) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ v(T_l,\ T_S) \ldots v(T_l,\ T_h) \\ \ldots\ldots\ldots\ldots \\ v(T_{h-1},\ T_h) \end{Vmatrix}.$$

12. Из всех этих значений выбирается минимальное, которое сравнивается с допустимым $\chi^2_{q,\,m}$ при заданном уровне значимости q и m степенях свободы.

$$\text{Если } \min_{l,\,S} v(T_l,\ T_S) > \chi^2_{q,\,m},$$

то дальнейшие вычисления прекращаются и все выделенные группы

наблюдений рассматриваются как существенно отличающиеся одна от другой. Если же

$$\min_{l,\,S} v(T_l,\ T_S) \leqslant \chi^2_{q,\,m},$$

то та пара групп T_l, T_S, на которой достигнуто это минимальное значение, объединяется в одну группу T_{l_1}.

13. В результате число групп будет $h-1$, и процедура проверки продолжается для данного уменьшенного набора групп. Для этого достаточно вычислить значения критерия для всех возможных пар, которые образует T_{l_1} с остальными $h-2$ группами. Значения критерия для тех пар, в которые не входит T_{l_1}, можно взять из матрицы, определенной в пункте 11. Из всех этих значений критерия опять выбирается минимальное, которое сравнивается с критическим $\chi^2_{q,\,m}$.

14. Такая последовательная процедура проверки, использующая парные объединения, продолжается до тех пор, пока минимальное значение не превысит допустимое $\chi^2_{q,\,m}$. Необходимо отметить, что на практике бывает удобно в качестве критерия использовать отношение:

$$\tau = \frac{\min\limits_{l,\,S} v(T_l,\ T_S) - m}{\sqrt{2m}}.$$

Процедура объединения прекращается как только будет достигнуто неравенство $\tau > 3$.

15. Полученные в результате группы наблюдений следует рассматривать как статистически однородные, отличающиеся одна от другой совокупности.

Многочисленные геологические примеры использования статистических методов разграничения геологических объектов приведены в работе [39].

ГЛАВА 3

ГЕОСТАТИСТИКА

Геостатистика — математическая теория разведки месторождений полезных ископаемых и оценки пространственных свойств их характеристик. Основы геостатистики заложены Ж. Матероном [35]; предпосылками для ее создания послужили работы Х. де Вийса, Д. Криге, Х. Зихеля и др., в которых были получены некоторые частные результаты, выходящие за рамки применения статистических методов. В последние два десятилетия геостатистика получила развитие в работах М. Давида [16], Ф. Формери, А. Жур-

неля и Ш. Хьюбре [50], А. Марешаля, Д. Серра, А. М. Марголина [30] и др. Она находит широкое применение при разведке месторождений промышленно развитыми странами и имеет особое значение для геологического изучения месторождений в процессе их разработки.

Математической моделью изучаемого в геостатистике объекта является геометрическое поле пространственной переменной $f(x)$, о которой известно, что в точках опробования x_i ($i = 1, \ldots, k$) с пространственными координатами (x_i), где $i = u, v, w$, она принимает фиксированные значения

$$f_i = f(x_i).$$

Пространственная переменная — геологический параметр, учитываемый при разведке месторождений полезных ископаемых (содержание полезного компонента в пробах и блоках, мощность рудного тела и т. д.). Пространственная переменная обладает следующими свойствами.

1. Каждая пространственная переменная всегда определена в конкретной области пространства — геометрическом поле. Обычно представляет интерес рассматривать не точечные, а средние значения пространственной переменной в пределах малой области — геометрической базы (пробы, блока). При изменении геометрической базы получается новая пространственная переменная. Геостатистика позволяет предсказать характеристики переменной на базе v в поле V по известным характеристикам точечной переменной в поле V', отличном от поля V.

2. Пространственная переменная характеризуется той или иной степенью непрерывности изменения в пространстве. Геометрические переменные (например, мощность) непрерывны в соответствии с определением Коши. Чаще наблюдается непрерывность в среднем: когда точка x стремится к x_0, лишь среднее значение $[f(x) - f(x_0)]^2$ стремится к нулю (изменение переменной нерегулярно и характеризуется точками разрыва). При отсутствии непрерывности в среднем переменная характеризуется крайней нерегулярностью. Тогда говорят об эффекте самородков, поскольку классическим примером нерегулярной изменчивости являются месторождения самородного золота. В качестве функции, характеризующей степень непрерывности пространственной переменной, в геостатистике используется вариограмма.

3. Пространственная переменная может быть изотропной (в этом случае вариограмма постоянна в любом направлении) или характеризоваться анизотропией (такой пространственной изменчивостью геологических параметров, когда вариограммы, построенные по различным направлениям в объеме тела полезного ископаемого, отличаются типом модели или значениями ее коэффициентов). Различаются анизотропии; геометрическая (которая может быть приведена к изотропности аффинным преобразованием оси координат аргумента вариограммы), зональная (характерная для расслоенных залежей) и функциональная (которая не может быть устранена

аффинными преобразованиями оси координат аргумента вариограммы).

4. В геометрическом поле пространственной переменной возможны разрывы. Они характеризуют явления перехода, которые можно проиллюстрировать следующими примерами: постоянная или почти постоянная в пределах каждого слоя переменная будет скачкообразно меняться при переходе от слоя к слою в этой вертикальной прерывистости, связанной с наличием безрудных пропластков; часто добавляется горизонтальная прерывистость, связанная с выклиниванием отдельных рудных линз; эффект самородков также представляет собой явление перехода, связанное с наличием микроструктур в геометрическом поле переменной: сеть разрывов в этом случае совпадает с границами раздела между рудными и безрудными зернами. На пространственную переменную накладывается единственное ограничение — предполагается, что функция $f(x)$ суммируема (интегрируема), т. е. что интегралы, $\int f(x)\,dx$, взятые по любой области поля V, конечны. Это условие отвечает реальной природе изучаемых объектов.

Геометрическое поле пространственной переменной — ее область определения. Оно представляет собой ограниченную область V, в которой пространственная переменная $f(x)$ принимает значения, отличные от нуля. Вне этого поля пространственная переменная равна нулю. С выходом за пределы геометрического поля связан граничный эффект, который проявляется тем сильнее, чем резче переход функции $f(x)$ к нулю. Геометрическое поле характеризуется геометрической переменной.

Геометрическая переменная — пространственная переменная, определяемая как

$$l(x) = \begin{cases} 1 & \text{при } x \in V \\ 0 & \text{при } x \notin V, \end{cases}$$

служащая геометрической характеристикой поля V (объема тела полезного ископаемого). Она характеризуется геометрической ковариограммой

$$L(h) = \frac{1}{V} \int\limits_V \int\limits_V \int l'(x)\, l'(x+h)\, dV,$$

где $l'(x) = l(X) - Ml$.

Ковариограмма $L(h)$ представляет собой меру (объем для $n = 3$, площадь поверхности для $n = 2$, длину для $n = 1$) пересечения исходного поля V и поля V', смещенного по отношению к исходному на вектор

$$h : L(h) = \mu V \cap V'_{-h} = \mu V \cap V'_h.$$

Граничный эффект — явление перехода пространственной переменной к нулевым значениям за пределами поля V (за контуром тела полезного ископаемого); проявляется как вклад погрешности определения границ (контуров) тел полезного ископаемого в дисперсию оценивания его запасов.

Задачи геостатистики сводятся к оцениванию функции F от пространственной переменной $f(x)$. В качестве F выступают среднее значение разведуемой характеристики в блоке объема V, ее значения в некоторой точке, не совпадающей с точкой опробования, запас в объеме блока.

Оценкой F является функция от совокупности наблюдаемых значений $\{f_i\}$: $\varphi = \varphi\{f_i\}$.

В силу дискретности системы опробования между оцениваемой величиной F и ее оценкой φ возникает случайное расхождение $\varepsilon = F - \varphi$. Дисперсия этого расхождения

$$D_\varepsilon = D\{A;\ B;\ C;\ F;\ \varphi\}$$

является функцией геометрических параметров разведочной сети (ее размеров, конфигурации и ориентировки) $\{A\}$, геометрических параметров проб объема v $\{B\}$, параметров пространственной изменчивости разведуемой переменной $f(x)$ — $\{C\}$, вида оцениваемой функции $\{F\}$, вида принятой оценки $\{\varphi\}$.

Предметом изучения в геостатистике служит исследование дисперсии результатов разведки (D_ε) как функции от перечисленных аргументов. Конечной целью этого изучения является, во-первых, отыскание таких оценок φ, которые при заданных значениях прочих аргументов минимизировали бы дисперсию D_ε и, во-вторых, нахождение таких параметров разведочных сетей $\{A\}$ и проб $\{B\}$, которые при известных параметрах изменчивости $\{C\}$, допустимой дисперсии D_ε и принятом виде оценки $\{\varphi\}$ минимизировали бы издержки на разведку месторождения.

Разработка математического аппарата геостатистики основана на двух подходах к исследованию пространственных переменных — транзитивной теории и теории случайных функций. Оптимальная в смысле минимизации дисперсии оценка месторождений базируется на разработанной в геостатистике процедуре крайгинга.

Развитие геостатистики характеризуется в настоящее время внедрением трехмерного моделирования по данным линейного опробования (метод «*turning bands*») и так называемой теории кондиционного моделирования, в рамках которой модельная случайная функция в пунктах опробования должна совпадать с фактическими значениями пространственной переменной [50]. Этот последний подход подобен предложенной А. М. Шурыгиным модели пространственных геологических переменных, рассматриваемых в качестве реализации условного случайного процесса. В стремлении к дальнейшему повышению эффективности геостатистических оценок ведутся разработки в области нелинейной геостатистики [50].

В зависимости от того, какая статистическая характеристика принята для описания пространственной изменчивости объектов разведки, геостатистические задачи и решения могут быть представлены в терминах корреляционного, структурного или спектрального анализов.

Транзитивная теория — один из подходов к решению задач геостатистики, в рамках которого пространственная переменная

рассматривается в качестве детерминированной функции пространственных координат. Источником случайности разведочных оценок является при этом только дискретность разведочной сети, не позволяющая составить исчерпывающего представления о непрерывной пространственной переменной. При построении этой теории не делается никаких вероятностных или других предположений о природе изучаемых явлений; она устанавливается в результате тщательного изучения имеющихся данных, а полученные результаты распространяются только на объекты того же типа. Этот подход близок кинематике, которая изучает общие законы движения, не вникая в причины и физический смысл. Основным инструментом транзитивной теории служит аналог корреляционной функции — транзитивная ковариограмма:

$$G(h) = \int f(x) f(x+h)\, dx$$

для векторного аргумента $h = \{h_1, \ldots, h_n\}$, где знак интеграла означает n-кратное интегрирование, dx — элемент объема dx_1, $dx_2, \ldots, dx_n$ n-мерного пространства. С учетом геометрического поля V функции $f(x)$ ковариограмма может быть представлена как:

$$G(h) = \frac{1}{V} \int\int\int_V [f(x) - Mf(x)]\,[f(x+h) - Mf(x)]\, dV,$$

где $Mf(x)$ — математическое ожидание функции $f(x)$. Ковариограмма характеризует изменчивость пространственной переменной $f(x)$ в геометрическом поле V.

Связь ковариограммы с геометрией поля V является ограничением транзитивной теории, так как не существует простого способа перехода от характеристик пространственной переменной, полученных в поле V, к ее характеристикам в другом поле V'. Кроме того, связь между ковариограммой и геометрическим полем часто выражает реально существующую зональность пространственной переменной. В подобных случаях невозможно разделить свойства пространственной переменной, обусловленные геометрией поля, и свойства, обусловленные ее пространственной изменчивостью. Приращения пространственной переменной в таких случаях являются нестационарными. Наоборот, в случаях, когда поведение пространственной переменной существенно не зависит от положения участка ее рассмотрения относительно границ геометрического поля, т. е. когда поле V можно рассматривать как произвольно взятое внутри обширной области однородной пространственной изменчивости, тогда имеются основания интерпретировать реальную пространственную переменную как реализацию случайной функции, произвольным образом включенную внутрь поля V. В этих условиях характеристики пространственной переменной, свободные от всякого влияния геометрической формы или размеров поля, могут быть выражены вариограммой [функцией $\gamma(h)$]. Связь между моделями случайных функций и транзитивными в подобных случаях выражается следующей формулой, устанавливающей зависимость математического ожидания $M[Gh)]$ транзитивной ковариограммы от

модели, основанной на рассмотрении случайных функций $\gamma(h)$:

$$M[G(h)] = [m^2 + \sigma^2 - \gamma(h)]\, L(h), \qquad (3.1)$$

где m — математическое ожидание стационарной случайной функции $f(x)$; σ^2 — ее дисперсия, $L(h)$ — геометрическая ковариограмма.

Сумма параметров m^2 и σ^2 может быть определена из следующего соотношения, справедливого для стационарных, в широком смысле, случайных функций:

$$m^2 + \sigma^2 = E[m^2(V) + \sigma^2(0 \mid V)], \qquad (3.2)$$

где $\sigma^2(0|V)$ — дисперсия распространения значений пространственной переменной, обладающей точечной геометрической базой.

С учетом (3.2) соотношение (3.1) может быть представлено в виде:

$$\gamma(h) = \sigma^2(0 \mid V) + E[m^2 - G(h)/L(h)].$$

Использование теории случайных функций в геостатистике основано на предположениях о том, что разведуемая пространственная переменная является реализацией стационарной случайной функции и что погрешности разведочных оценок могут быть обусловлены уклонением реализации случайной функции относительно ее математического ожидания. Этот подход возможен в частном случае, когда допустимо предположение, что изучаемый объект обладает пространственной однородностью, т. е. если закон распределения значений пространственной переменной в k произвольных точках пространства инвариантен относительно любого перемещения совокупности этих точек. Используемая в этой теории корреляционная функция $K(h)$ (ковариация значений случайной функции в двух точках как неслучайная функция расстояния между этими точками) представляет собой вероятностный эквивалент ковариограммы $G(h)$. Применение теории случайных функций возможно только тогда, когда изучаемая случайная функция обладает заведомо конечной дисперсией. Природные объекты обычно не удовлетворяют этому условию, но характеризуются конечной дисперсией приращений случайной функции, т. е. разностей значений функции в двух точках. Именно такие случайные функции изучает геостатистика. Роль транзитивной ковариограммы $G(h)$ или корреляционной функции $K(h)$ играет при этом вариограмма $\gamma(h)$. Применение теории случайных функций, как и транзитивной теории, также характеризуется определенными ограничениями.

Несоответствие свойств реальных геологических объектов стационарным (в широком смысле, случайным) функциям было установлено Д. Криге на золоторудных месторождениях Витватерсранда. Им была выявлена зависимость дисперсии содержаний золота в пробах от размеров разведуемых месторождений или их частей (блоков, участков, рудных полей):

$$\sigma^2 = \alpha \ln \frac{V}{v}.$$

ВАРИОГРАММА

В а р и о г р а м м а — функция, которая выражает зависимость половины квадрата разности значений геологического параметра в точках тела полезного ископаемого от расстояния между ними. В теории случайных функций известна как структурная функция:

$$\gamma(h) = \frac{1}{2V} \int_V \int [f(x+h) - f(x)]^2\, dV.$$

При этом $[f(x+h) - f(x)]$ может сохранять конечную дисперсию.

Основной характеристикой случайной функции $f(x)$ со стационарными, в широком смысле, приращениями является вариограмма, равная половине дисперсии приращения $f(x+h) - f(x)$:

$$\gamma(h) = \frac{1}{2} D[f(x+h) - f(x)] =$$

$$= \frac{1}{2} E\{[f(x+h) - f(x)]^2\}.$$

Одним из свойств вариограммы $\gamma(h)$ является уточнение смысла часто встречающегося в геологической практике, но недостаточно четко определенного понятия «зона влияния пробы или разведочной выработки». Поведение вариограммы при увеличении $|h|$ дает количественную характеристику уменьшения влияния данной пробы на различные участки месторождения по мере увеличения расстояния между пробой и участком.

При решении прикладных вопросов обычно стремятся получить возможность оперировать с вариограммой изотропной модели, т. е. с вариограммой, имеющей вид:

$$\gamma(h) = \gamma(r).$$

Возможность выявления анизотропии при изучении экспериментальных вариограмм представляет собой другое их важное свойство, так как дает в руки исследователя средство оценки основных структурных особенностей изучаемого объекта. На практике используют несколько простых моделей анизотропии. Простейшей является геометрическая анизотропия, при которой $\gamma(h)$ оказывается функцией $\gamma(Q)$ квадратической формы:

$$Q = \sum_{ij} a_i a_j h_i h_j$$

относительно координат вектора h. При этом достаточно провести линейное преобразование координат, чтобы геометрически анизотропную модель свести к изотропной, для которой $Q = r^2$. С этой целью по известным направлениям анизотропии параметры $h_1, h_2, \ldots, h_n$ заменяются на $\lambda_1 h_1$, $\lambda_2 h_2$, $\lambda_3 h_3$, . . . , $\lambda_n h_n$ с соответственно подобранными постоянными λ-модулями аффинного преобразования.

Возможна также зональная анизотропия, когда вариограмма $\gamma(h)$ зависит не от всех n компонент вектора h, а только от

одного или двух из них. Например, в случае трехмерного пространства может встретиться вариограмма $\gamma(h_3)$, зависящая только от вертикальной составляющей h_3 вектора h. При такой вариограмме приращение $f(x+h)-f(x)$ имеет нулевую дисперсию для всех горизонтальных векторов $h=(h_1, h_2, 0)$, т. е. пространственная переменная $f(x)$ остается постоянной в горизонтальной плоскости. Аналогично в случае потока, когда пространственная переменная $f(x)$ остается постоянной вдоль одной оси координат (например, h_1), вариограмма зависит от двух других координат: $\gamma(h)=\gamma(h_2, h_3)$. Практически пространственная переменная $f(x)$ никогда не остается строго постоянной в плоскости или вдоль профилей, параллельных некоторому направлению, а лишь проявляет значительно меньшую изменчивость в сравнении с изменчивостью в других направлениях. В таких случаях вариограмму можно представить суммой двух компонент, одна из которых будет характеризовать изотропную составляющую изменчивости, а другая — зональную (или вообще анизотропную):

$$\gamma(h)=\gamma_1(r)+\gamma_2(h_2, h_3).$$

или $$\gamma(h)=\gamma_1(r)+\gamma_2(h_3),$$

В большинстве задач и особенно при расчете дисперсий оценки эти две составляющие рассматриваются раздельно, а их эффект суммируется.

Следующим свойством вариограммы является то, что на ней в виде уступов и порогов выражаются явления перехода. Анализ явлений перехода часто позволяет выявить наложение нескольких структур различного масштаба.

Поведение вариограмм вблизи нуля хорошо отражает степень непрерывности и регулярности пространственной переменной. Ж. Матерон выделяет четыре типа вариограмм (рис. 31):

вариограммы первого типа с параболическим характером изменения вблизи нуля, дважды дифференцируемые в нуле, характеризуют непрерывные пространственные переменные, например мощность пластовых залежей полезных ископаемых;

вариограммы второго типа, тангенциальные, не являющиеся дважды дифференцируемыми в нуле, характеризуют пространственные переменные, непрерывные в среднем; к ним относятся содержания и линейные запасы многих компонентов в рудах;

вариограммы третьего типа, не проходящие через начало координат, характеризуют пространственные переменные, не обладающие непрерывностью даже в среднем, и отражают эффект самородков; резкое изменение вариограммы от нуля до некоторого уровня, начиная с которого она приобретает плавный характер, происходит в пределах очень узкой зоны, называемой носителем эффекта самородков; размер этой зоны близок к среднему диаметру скоплений полезных минералов;

вариограммы четвертого типа, параллельные оси абсцисс, характеризуют пространственные переменные, соответствующие независимым случайным величинам.

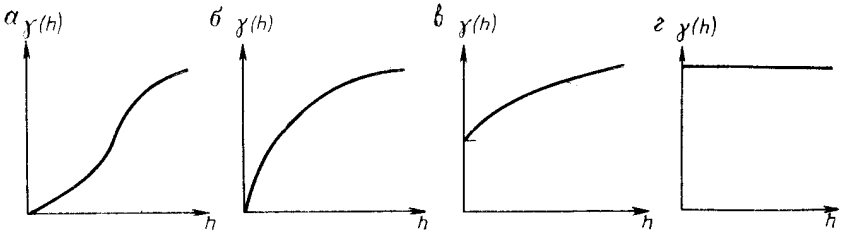


Рис. 31. Типы вариограмм по Ж. Матерсону:
а — непрерывный тип; *б* — линейный тип; *в* — тип эффекта самородков; *г* — случайный тип

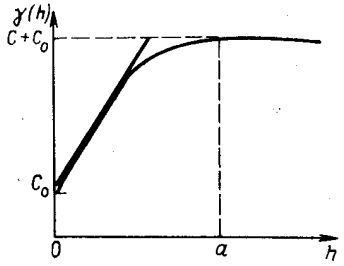


Рис. 32. Сферическая модель вариограммы:
а — интервал влияния; C_0 — эффект самородков; $C+C_0$ — порог

Наконец, когда $\gamma(r)$ является функцией радиус-вектора $r=|h|$, можно характеризовать ее поведение вблизи нуля ограниченным разложением вида:

$$\gamma(r)=\Sigma a_{2n}r^{2n}+\Sigma C_\lambda r^\lambda.$$

В таком разложении различают регулярную, содержащую только члены с r^{2n}, и нерегулярную части. Если нерегулярная часть в разложении отсутствует, то вариограмма $\gamma(h)$ бесконечно дифференцируема и, следовательно, представляет собой в высшей степени регулярную пространственную переменную $f(x)$. Таким образом, именно нерегулярная часть разложения, содержащая члены с r^λ (иногда также с $r^{2n}\ln r$), где λ — отличное от целого четного число, определяет степень регулярности пространственной переменной. При этом наиболее существенную роль играет член с наименьшим значением λ. Поэтому степень регулярности пространственной переменной удобно характеризовать наименьшим показателем степени λ в нерегулярной части разложения вариограммы.

Для аппроксимации вариограммы используются различные модели; к числу наиболее распространенных из них относятся следующие:

1) линейная модель:

$$\gamma(h)=Ah+B;$$

2) сферическая модель (рис. 32):

$$\gamma(h)=C\left(\frac{3h}{2a}-\frac{h^3}{2a^3}\right)+C_0, \quad \text{если } h\leqslant a;$$

$$\gamma(h)=C+C_0, \qquad \text{если } h>a;$$

$$\gamma(0)=0,$$

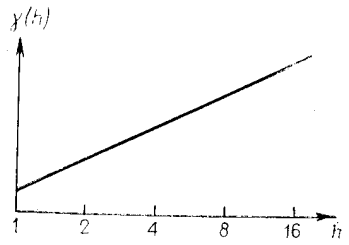


Рис. 33. Модель де Вийса

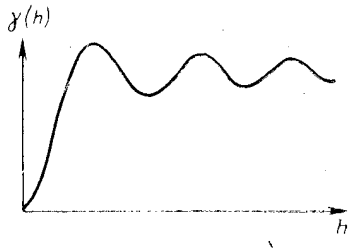


Рис. 34. Модель эффекта включений

где a — интервал влияния (расстояние, за пределами которого вариограмма приобретает характер, отражающий отсутствие зависимости между значениями геологических параметров); C_0 — эффект самородков; $C + C_0$ — порог (дисперсия геологического параметра, являющаяся предельным значением вариограммы);

3) модель Де Вийса (рис. 33):

$$\gamma(h) = A \ln h + B;$$

4) модель эффекта включений (рис. 34) — модель пространственной изменчивости геологических параметров крупных тел полезных ископаемых, характеризующихся квазипериодичностью, обусловленной наличием обогащенных или бедных участков; ее вариограмма имеет вид:

$$\gamma(h) = C\left(1 - \frac{\sin(ah)}{ah}\right),$$

где C — порог; a — интервал влияния;

5) степенна́я модель:

$$\gamma(h) = ah^{\lambda}.$$

Она может быть преобразована в линейную:

$$\ln\gamma(h) = \lambda \ln h + B;$$

6) экспоненциальная модель:

$$\gamma(h) = C_0 + C(1 - \exp(-|h|/a)),$$

где C_0 — эффект самородков; $C_0 + C$ — порог; a — интервал влияния.

В том случае, когда вариограммы не согласуются ни с одной теоретической моделью, имеют место скрытые структуры изменчивости геологических параметров. Часто такие вариограммы можно удовлетворительно представить в виде суммы двух теоретических функций: $\gamma(h) = \gamma_1(h) + \gamma_2(h)$.

ДИСПЕРСИЯ ОЦЕНКИ МЕСТОРОЖДЕНИЯ

В основе определения погрешностей разведки месторождения лежит понятие дисперсии распространения, которая представляет

собой меру рассеивания (так называемую ошибку аналогии) средних значений геологических параметров в неисследованном объеме тела полезного ископаемого по их известным оценкам в опробованном объеме.

В качестве дисперсии распространения рассматривается дисперсия

$$D(Y-Z) = \sigma_Y^2 + \sigma_Z^2 - 2\sigma_{YZ}.$$

Эта дисперсия служит для выражения ошибки, с которой среднее значение случайной функции в объеме V может быть принято в качестве оценки среднего значения в объеме V' (или наоборот). Дисперсия распространения имеет вид:

$$D(Y-Z) = \frac{1}{V^2}\int_V\int_V K(x-x')\,dxdx' + \frac{1}{(V')^2}\int_{V'}\int_{V'} K(x-x')\times$$
$$\times\, dxdx' - \frac{2}{VV'}\int_{V'}\int_V K(x-x')\,dxdx'. \qquad (3.3)$$

Если вместо объема V рассматривать дискретную совокупность, состоящую из N точек $x_1, \ldots, x_n$, то величина Z определяется выражением

$$Z = \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

Переменная Z представляет собой среднее значение случайной функции $f(x)$ в N точечных пробах. В этом случае дисперсия распространения $D(Y-Z)$ характеризует ошибку, с которой среднее значение Y случайной функции $f(x)$ в объеме V оценивается величиной Z, полученной по N точечным пробам. Эта дисперсия является дисперсией оценки. Используя (3.3), для дисперсии оценки получаем:

$$D(Y-Z) = \frac{1}{V^2}\int_V\int_V K(x-x')\,dxdx' + \frac{1}{N^2}\sum_{i,j=1}^{N} K(x_i-x_j) -$$
$$-\frac{2}{NV}\sum_{i=1}^{N}\int_V K(x-x_i)\,dx.$$

Если v и V — два объема, первый из которых содержится во втором, а Z и Y — средние стохастические значения функции $f(x)$ в этих объемах, то дисперсия распространения, характеризующая погрешность оценивания среднего значения Y в объеме V средним значением Z в объеме v, полностью определяется формулой (3.3). Для краткости эта дисперсия называется дисперсией распространения v на V.

Когда геометрическое поле V сложено из элементов одинакового объема v, дисперсия v в V определяется следующим выражением:

$$\sigma^2(v|V) = \frac{1}{V^2}\int_V\int_V \gamma(x-x')\,dxdx' - \frac{1}{v^2}\int_v\int_v \gamma(x-x')\,dxdx'.$$

Эта формула приводит к соотношению аддитивности (называемому также формулой Криге):

$$\sigma^2(v \mid V') = \sigma^2(v \mid V) + \sigma^2(V \mid V'), \qquad (3.4)$$

которое показывает, что дисперсия v в V' равна сумме дисперсий v в V и V в V'.

В сети опробования легко выделить плоскости и профили наибольшей плотности размещения проб. В случае трехмерного пространства, определив средние содержания по пробам, отобранным в профилях, принимают эти содержания в качестве оценок истинных содержаний по профилям. В свою очередь, среднее содержание по профилям, расположенным в одной плоскости, приписывают соответствующему сечению и, наконец, среднее содержание по всем сечениям берут в качестве среднего содержания для всего поля V в целом. При этом на всех стадиях распространения известных содержаний на неопробованные участки поля возникают некоторые ошибки распространения. Если эти ошибки независимы (т. е. ковариации ошибок равны нулю), то дисперсия результирующей ошибки, соответствующая (3.4), будет суммой трех дисперсий распространения:

$$\sigma_N^2 = \sigma_{N_1}^2 + \sigma_{N_2}^2 + \sigma_{N_3}^2,$$

где

$$\sigma_{N_1}^2,\ \sigma_{N_2}^2,\ \sigma_{N_3}^2$$

— дисперсии распространения соответственно на профиль, сечение и слой.

Одной из важнейших задач изучения пространственной переменной $f(x)$ является оценка количества металла $Q = \int f(x)\ dx$ по имеющимся экспериментальным данным, полученным в результате отбора проб по регулярной сети, ячейка которой имеет форму параллелепипеда с ребрами $a_1, a_2, \ldots, a_n$. В n-мерном пространстве задана прямоугольная система координат с осями, параллельными основным направлениям сети опробования, так что пробы отобраны в точках:

$$y + ka = (y_1 + k_1 a_1,\ \ldots,\ y_n + k_n a_n),$$

где k_i — целые числа, принимающие все возможные положительные и отрицательные значения; $y = (y_1, \ldots, y_n)$ — произвольные точки, принятые в качестве исходной при построении сети.

Обусловленная дискретностью опробования ошибка оценки запаса $Q'(y)$ может быть определена как среднее значение $[Q'(y) - Q]^2$ при условии, что точка y занимает все возможные положения внутри параллелепипеда p, образованного ребрами $a_1, \ldots, a_n$. Дисперсия оценки запаса имеет вид:

$$\sigma^2(a) = \frac{1}{|a|_p} \int [Q'(y) - Q]^2\, dy.$$

Используя кovариограмму пространственной переменной $f(x)$, получаем для дисперсии формулу

$$\sigma^2(a) = |a| \sum_k G(ka) - \int G(h)\, dh.$$

Дисперсия оценки, которая обеспечивает интервальное оценивание величины Q, представляет собой разность между приближенным и точным значениями интеграла $\int G(h)\, dh$, т. е. выражается как ошибка вычисления интеграла по численным значениям $G(h)$. Эта ошибка тем меньше, чем многочисленнее разведочные данные, а следовательно, плотнее сеть опробования и чем регулярнее пространственная изменчивость самой переменной $f(x)$, т. е. чем регулярнее кovариограмма $G(h)$.

Рассчитать одновременно оценку количества металла и дисперсию этой оценки по одним лишь разведочным данным невозможно. Эта трудность может быть преодолена, если заранее задаться определенным видом функции $G(h; \lambda, \mu, \ldots)$, параметры которой λ, $\mu, \ldots$ могут быть подобраны путем сопоставления с экспериментальными значениями $G'(ka)$. Ценность получаемых при этом результатов определяется степенью соответствия выбранной математической модели $G(h; \lambda, \mu, \ldots)$ физической реальности. Наиболее важен выбор поведения $G(h)$ при $|h| < a$ (где отсутствуют экспериментальные данные), т. е. поведения кovариограммы $G(h)$, определяемого ее нерегулярной частью (вблизи нуля).

Используя известную в теории численного интегрирования формулу Эйлера — Маклорена, можно получить разложение дисперсии оценки на два члена T_0 и T_z:

$$\sigma^2(a) = T_0(a) + T_z(a),$$

где T_0, называемый элементом распространения, зависит только от поведения $G(x)$ на отрезке $[0, a]$, т. е. в окрестности нуля; T_z — элемент флуктуаций — связан с поведением $G(x)$ при приближении к оси x в точке $x = b$ (где b — носитель кovариограммы).

Элемент флуктуации — фактор мешающий, который может вносить значительные, не поддающиеся оценке расхождения между истинным значением дисперсии $\sigma^2(a)$ и ее оценкой с помощью только элемента распространения $T_0(a)$. Однако этим элементом можно пренебречь при определении дисперсии оценки, так как среднее значение T_z по ε равно нулю, где

$$\varepsilon = \frac{b}{a} - (k + 1).$$

Замена элемента распространения дисперсии оценки ограниченным разложением $\sigma^2(a_1, \ldots, a_n)$, упорядоченным по возрастающим степеням одночленов $a_1^{\lambda_1}, a_2^{\lambda_2}, \ldots, a_n^{\lambda_n}$, приводит к принципу композиции дисперсии распространения, смысл которого легко пояснить в двумерном случае.

Если $G_2(r)$ — кovариограмма пространственной переменной в двумерном пространстве, $G_1(r)$ — кovариограмма полученная из

$G_2(r)$ посредством регуляризации, $\sigma_1^2(a_1)$ и $\sigma_2^2(a_2)$ — дисперсии оценки, полученные для одномерных случаев по ковариограммам соответственно $G_1(r)$ и $G_2(r)$, а $\sigma^2(a_1a_2)$ — дисперсия, рассчитанная для двумерного случая по ковариограмме $G_2(r)$, то принцип композиции дисперсии можно записать в следующем виде:

$$\sigma^2(a_1a_2)=\sigma_1^2(a_1)+a_1\sigma_2^2(a_2) \quad \text{при} \quad a_1>a_2.$$

Этот случай соответствует определению дисперсии оценки по известным дисперсиям распространения на разведочный профиль $(\sigma_1^2(a_1))$ и дисперсии распространения на ленту эксплуатационного блока $(\sigma_2^2(a_2))$.

Транзитивная теория полезна прежде всего для случаев, когда невозможны статистические выводы. Одним из таких случаев является оценка геометрической переменной, т. е. оценка объема или площади геометрического поля. Для такой оценки V пространственной переменной имеются те же данные, что и для оценки самой переменной. Например, если есть регулярная сеть опробования с началом в точке y и с шагом a, то по результатам опробования можно сказать, принадлежит ли каждая точка $y+pa=(y_1+p_1a_1, \ldots, y_n+p_na_n)$ полю V или не принадлежит. Другими словами, если $l(x)$ — геометрическая переменная, связанная с полем V, то нам известны численные значения (0 или 1) $l(y+pa)$ для всех точек опробования. Оцениваемая величина есть не что иное, как количество металла, связанное с геометрической переменной $l(x)$:

$$V=\int l(x)\,dx.$$

Таким образом, проблема оценки параметров геометрического поля является частным случаем общей проблемы оценки пространственных переменных. Следовательно, дисперсия оценки геометрической переменной может быть определена, если известна ее геометрическая ковариограмма $L(h)$. Однако обычно имеются не значения геометрической переменной в точках, а длины пересечения или площади сечений поля V параллельно некоторому направлению. Тогда необходимо оценивать не переменную l в n-мерном пространстве, а переменную l_{n-1} или l_{n-2}, полученную из $l_n(x)$ с помощью операции регуляризации первого и второго порядков. Ковариограммы этих переменных также могут быть получены посредством операции регуляризации $L(x)$.

Геометрическая ковариограмма обладает рядом важных особенностей, которые позволяют упростить расчет дисперсии оценки. Так как $L(h)$ представляет собой меру (объем для $h=3$) пересечения поля V и поля V' (смещенных на вектор h), то $L(0)=V$ и $\int L(h)\,dh=V^2$. Как значение $L(h)$ в нуле, так и ее интеграл либо известны, либо могут быть оценены. Рассмотрим меру $L(\delta h)$ пересечения полей V и $\tau_{\delta h}$, где вектор δh бесконечно мал по модулю. Разность мер $L(0)$ и $L(\delta h)$ (представляющая собой половину объема, описываемого вектором δh, когда его основание пробегает

всю поверхность, ограничивающую V) с точностью до члена второго порядка относительно (δh) равна:

$$L(0)-L(\delta h)=S(\alpha)\,|\delta h|.$$

Коэффициент пропорциональности $S(\alpha)$ называется диаметральной полувариацией в направлении α вектора δh. Если рассечь объем V плоскостями, перпендикулярными к направлению α и отнесенными к абсциссам z их пересечения с прямой, параллельной α, то диаметральная вариация $2S(\alpha)$ будет равна верхнему пределу всех возможных сумм вида:

$$|S(z_1)-S(z_2)|+\ \ldots\ +|S(z_{n-1})-S(z_n)|,$$

где $S(z_i)$ — площади соответствующих сечений.

Такой предел всегда существует для геометрических полей V, которые могут встретиться в природе. Следовательно, во всех случаях известны два первых члена разложения $L(h)$ в окрестности нуля:

$$L(h)=V-S(\alpha)\,|h|+\ \ldots,$$

а геометрическая ковариограмма в окрестности нуля линейна.

При условии, что ковариограмма $L(h)$ изотропна или может быть сведена к изотропной, можно получить основную часть дисперсии оценки V для плотных сетей опробования. В частности, если диаметральная полувариация $S(\alpha)$ практически постоянна при всех α и равна S, то для сети с размерами ячейки $a_1\geqslant a_2\geqslant a_3$ получим:

$$\sigma^2(a_1,\ a_2,\ a_3)=S\left[\frac{1}{6}a_1a_2a_3^2+0{,}0609a_1a_2^3+\frac{\pi}{90}a_1^4\right]+\ \ldots\ .$$

В случае, когда $a_3=0$, т. е. если измерены пересечения геометрического поля вдоль направления a_3, в выражении дисперсии оценки V сохраняются два последних члена.

Используя дисперсии оценок запаса Q и объема V, можно выразить относительную дисперсию среднего содержания:

$$\sigma_m^2/m^2=\sigma_Q^2/Q^2+\sigma_V^2/V^2-2\sigma_{QV}/QV,$$

где σ_{QV} — ковариация оценок Q и V.

Эта ковариация, в свою очередь, может быть рассчитана по смешанной ковариограмме:

$$C(h)=1/2\,[Q(h)+Q(-h)],$$

где $Q(h)=m(h)\,L(h)=\int f(x)\,l(x-h)\,dx$;
$Q(-h)=m(-h)\,L(h)=\int f(x)\,l(x+h)\,dx$.

$C(h)$ выражает среднее арифметическое количество металла, содержащегося в симметричных пересечениях V' и V'', которые сохраняются при переносах поля V соответственно на h и $-h$.

В частном случае, когда содержание $f(x)$ не зависит от своего поля V, т. е. когда характер изменения $f(x)$ остается одинаковым в пределах всего геометрического поля, $C(h)=mL(h)$ и, следовательно:

$$\sigma_m^2/m^2=\sigma_Q^2/Q^2-\sigma_V^2/V^2.$$

С использованием понятия линейного эквивалента пробы и с учетом модели Де Вийса выводятся выражения дисперсии оценки месторождений для различных систем их разведки.

Рассмотрим в качестве примера разведку жилы штреками и восстающими.

Прямоугольник lh принимается в качестве зоны влияния линейной пробы l, размещенной в его центре параллельно стороне l. Дисперсия, характеризующая точность, с которой истинное содержание Z в зоне влияния оценивается содержанием Y в линейной пробе l, имеет вид:

$$\sigma_E^2 = \sigma_Y^2 - 2\sigma_{YZ} + \sigma_Z^2 = 2\chi\left(\frac{h}{2}\right) - \gamma(0) - F(h), \qquad (3.5)$$

где, в частности, для $h < l$:

$$\frac{1}{3\alpha}\gamma(h) = \ln l - \frac{3}{2} + \pi\frac{h}{l} + \frac{h^2}{l^2}\ln\frac{h}{l} - \frac{3}{2}\frac{h^2}{l^2} - \frac{1}{12}\frac{h^4}{l^4} + \frac{1}{60}\frac{h^6}{l^6} + \ldots,$$

$$\frac{1}{3\alpha}\chi(h) = \ln l - \frac{3}{2} + \frac{\pi}{2}\frac{h}{l} + \frac{h^2}{3l^2}\ln\frac{h}{l} + \frac{11}{18}\frac{h^2}{l^2} - \frac{1}{60}\frac{h^4}{l^4} + \frac{1}{420}\frac{h^6}{l^6},$$

$$\frac{1}{3\alpha}F(h) = \ln l - \frac{3}{2} + \frac{\pi}{3}\frac{h}{l} + \frac{h^2}{6l^2}\ln\frac{h}{l} - \frac{25}{72}\frac{h^2}{l^2} - \frac{1}{180}\frac{h^4}{l^4} + \frac{1}{1680}\frac{h^6}{l^6}.$$

Приближенное выражение σ_E^2 в этом случае:

$$\sigma_E^2 = 3\alpha\left[\frac{\pi}{6}\frac{h}{l} - \left(\frac{\ln 2 - 1/4}{6}\right)\frac{h^2}{l^2} + \frac{1}{288}\frac{h^4}{l^4} - \frac{1}{1920}\frac{h^6}{l^6}\right]. \qquad (3.6)$$

Для малых величин h/l можно ограничиться линейным членом.

$$\text{Тогда } \sigma_E^2 = 3\alpha\frac{\pi}{6}\frac{h}{l}.$$

Если сечение рудного тела S разделить на n прямоугольных зон влияния $S_1, S_2, \ldots, S_n$, каждая из которых характеризуется линейной пробой, расположенной в центре, то результирующая дисперсия оценки среднего содержания по сечению примет следующий вид:

$$\sigma_n^2 = \frac{1}{S^2}\sum_{i=1}^{n} S_i^2\sigma_{E_i}^2. \qquad (3.7)$$

В частности, если все зоны влияния равны, то

$$\sigma_n^2 = \sigma_E^2/n.$$

Формулу (3.7) можно использовать для оценки точности определения среднего содержания в горной выработке, опробованной n равноотстоящими бороздовыми пробами. Она остается верной и при наличии зональной анизотропии, если пробы отобраны по всей мощности рудного тела. Эта формула может служить также для оценки точности определения среднего содержания по жиле, разведанной горизонтальными горными выработками. Если высота этажа h постоянна, то штрек длиной l характеризует два полуэтажа. Дисперсия распространения определяется согласно (3.5), а дисперсия оценки содержания по горизонтам принимает вид:

$$\sigma_n^2 = \left(\sum_{i=1}^{n} l_i^2\sigma_{E_i}^2\right)\Big/\left(\sum_{i=1}^{n} l_i\right)^2. \qquad (3.8)$$

При малом расстоянии h между горизонтами в сравнении с длиной выработки l, как и для (3.6), можно ограничиться первым членом разложения и считать, что

$$\sigma_{E_i}^2 = 3\alpha\frac{\pi}{6}\frac{h}{l_i}.$$

Тогда выражение дисперсии оценки (3.8) примет следующий простой вид:

$$\sigma_n^2 = \alpha\frac{\pi}{2}\frac{S}{L^2}, \qquad (3.9)$$

где $L = \sum_{i=1}^{n} l_i$ — суммарная длина выработок; $S = hL$ — общая разведанная площадь.

Если, кроме N горизонтальных выработок, по падению рудного тела пройдены N' восстающих со средним интервалом h', то можно получить оценку месторождения лучшую, чем только по штрекам. Обозначим символами t и t' средние взвешенные содержания соответственно по штрекам и восстающим, а σ_N^2 и $\sigma_{N'}^2$ — вычисленные по формулам (3.8) или (3.9) и характеризующие точность, с которой каждое из средних t и t' оценивает истинное содержание Z. С помощью крайгинга можно получить усредненную оценку содержания:

$$Y = \lambda t + (1-\lambda)t'.$$

Коэффициент λ следует определить таким образом, чтобы дисперсия оценки $D(Y-Z)$ истинного среднего содержания в рудном теле Z с помощью Y была минимальной. Тогда

$$D(Y-Z) = \lambda^2 D(t-Z) + (1+\lambda)^2 D(t'-Z) + 2\lambda(1-\lambda)M[(t-Z)(t'-Z)],$$

где $Y-Z = \lambda(t-Z) + (1-\lambda)(t'-Z)$ и
где дисперсии распространения $D(t-Z)$ и $D(t'-Z)$ идентичны σ_N^2 и $\sigma_{N'}^2$, а коэффициент ковариации $M[(t-Z)(t'-Z)]$ ничтожно мал.

В итоге дисперсия

$$D(Y-Z)=\lambda^2\sigma_N^2+(1-\lambda)^2\sigma_N^2, \qquad (3.10)$$

которая минимальна при

$$\lambda=(\sigma_{N'}^2)/(\sigma_N^2+\sigma_{N'}^2). \qquad (3.11)$$

Подставляя (3.11) в (3.10), получим минимальную дисперсию, или дисперсию крайгинга:

$$\sigma_k^2=(\sigma_N^2\sigma_{N'}^2)/(\sigma_N^2+\sigma_{N'}^2).$$

В частном случае, когда расстояния между штреками и восстающими меньше длин соответствующих выработок, дисперсии распространения рассчитываются по упрощенным формулам. Если обозначить символом L' суммарную длину восстающих выработок, то

$$\lambda=L^2/(L^2+L'^2)$$

и оценка Y определится выражением:

$$Y=(tL^2+t'L'^2)/(L^2+L'^2).$$

Другими словами, взвешивание оценок t и t' следует производить по квадратам суммарных длин соответствующих выработок (а не по самим длинам). Дисперсия оценки в этом случае

$$\sigma_K^2=\alpha\frac{\pi}{2}\frac{S}{L^2+L'^2}.$$

Как правило, истинные содержания в горных выработках неизвестны и оцениваются по результатам дискретного опробования. Например, оценка t получена по N_R бороздовым пробам, отобранным через равные интервалы a на полную мощность рудного тела p, предполагаемую постоянной. Среднее содержание в бороздовых пробах позволяет оценить истинное содержание с дисперсией, определяемой по формулам (3.8) или (3.9) с заменой h и l на a и p. Нельзя упускать из виду, что эти формулы применимы только тогда, когда мощность p мала по сравнению с величинами h и l.

Если мощность значительна, то необходимо обратиться к трехмерной модели Де Вийса и произвести регуляризацию первого и второго порядков (см. рис. 33).

Обозначив символом σ_R^2 дисперсию оценки содержания в горных выработках по бороздовым пробам, имеем:

$$\sigma_R^2=\frac{1}{N_R}\alpha\frac{\pi}{2}\frac{a}{p}=\frac{1}{N_R}\sigma_E^2.$$

Применяя принцип аддитивности дисперсий распространения, в данном случае на профиль и на слой, получаем общую дисперсию оценки:

$$\sigma_T^2=\sigma_N^2+\sigma_R^2.$$

Кроме того, необходимо добавить еще третий член, представляющий собой дисперсию случайных ошибок отбора и анализа проб, и проследить за тем, чтобы используемые методы отбора и анализа проб не вносили систематической ошибки.

Помимо варианта разведки жильного месторождения штреками и восстающими, дисперсия оценки месторождения может быть определена и для других систем: разведки тонкого пласта скважинами по квадратной сети, разведки рудного тела регулярной сетью (без ограничений на мощность и при различной конфигурации сети), разведки штокообразного тела горизонтальными сечениями и др.

КРАЙГИНГ

К р а й г и н г (метод Криге) — метод нахождения наилучшей оценки среднего содержания компонентов или мощности полезного ископаемого в блоке с использованием результатов опробования как внутри, так и вне оцениваемого блока; эти результаты учитываются с весами, обеспечивающими минимум дисперсии среднего значения. Крайгинг реализуется при различных системах разведки месторождений полезных ископаемых в следующих модификациях: дискретный, непрерывный, случайный крайгинг.

Дискретный крайгинг (точечный) применяется при интерполяции имеющихся разведочных данных в заданные точки тел полезных ископаемых. Целесообразность его применения возникает как на стадии разведки, например при выявлении участков промышленного оруденения в пластообразной залежи, разбуренной скважинами по регулярной сети, так и на стадии эксплуатации, когда, например, по данным опробования взрывных скважин решаются вопросы селективной отработки рудных тел. Если рудное тело разбурено по квадратной сети, то проблема крайгинга состоит в определении весов, которые должны быть приписаны значениям признака в центральной скважине A, в ближайших окружающих ее скважинах B_1, B_2, B_3, B_4 и в скважинах второй обрамляющей зоны C_1, C_2, C_3, C_4, для получения наилучшей оценки среднего значения признака в зоне влияния скважины A (рис. 35). Достаточно ограничиться учетом только двух ближайших к оцениваемому блоку (зона влияния скважины A) ореолов, поскольку в большинстве случаев использование данных по более удаленным скважинам не приносит заметного уточнения оценки.

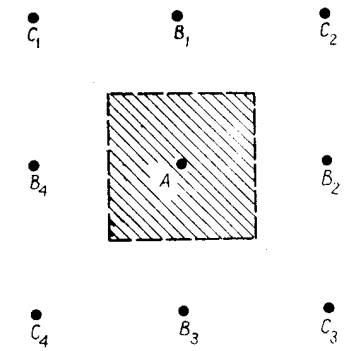


Рис. 35. Схема размещения разведочных скважин при дискретном крайгинге:

A — центральная скважина (заштрихована ее зона влияния, представляющая собой оцениваемый блок); B_1, B_2, B_3, B_4 — скважины ближайшей зоны, обрамляющей оцениваемый блок; C_1, C_2, C_3, C_4 — скважины второй зоны, обрамляющей оцениваемый блок

Если u — содержание компонента в центральной скважине A, v —

среднее содержание во всех имеющихся скважинах ореола B, w — среднее содержание во всех имеющихся скважинах ореола C, а $t = h/a$, где h — мощность рудного тела и a — шаг сети, то среднее содержание Z в оцениваемом блоке рассчитывается по следующей формуле крайгинга:

$$Z^* = (1-\lambda-\mu)u + \lambda v + \mu w.$$

Асимптотические формулы коэффициентов крайгинга λ и μ в условиях модели Де Вийса для случая, когда имеются центральная и все скважины обоих окружающих ее ореолов, при малых значениях t имеют вид:

$$\lambda = \frac{(0{,}4277 - \ln t)\left(0{,}5173 - \frac{1}{4}\ln t\right)}{0{,}9121 - 1{,}4739\ln t + 9/16\ln^2 t};$$

$$\mu = \frac{(0{,}4277 - \ln t)\,(0{,}0841 - 1/4\ln t)}{0{,}9121 - 1{,}4739\ln t + 9/16\ln^2 t};$$

а для больших t: $\lambda = 0{,}407$; $\mu = 0{,}017$.

Дисперсия крайгинга в первом случае:

$$\frac{1}{3\alpha}\sigma_K^2 = 0{,}1777 - \ln t - (\lambda-\mu)(0{,}4277 - \ln t)$$

и во втором:

$$\frac{1}{3\alpha}\sigma_K^2 = 0{,}311\,\frac{1}{t}.$$

Непрерывный крайгинг соответствует оценке месторождения при его разведке горными выработками, когда их объем делится на бесконечное число соприкасающихся элементарных объемов dV, каждый из которых отбирается в виде пробы и отдельно анализируется. Практически роль элементарных объемов dV могут играть отпалки или бороздовые пробы. Задача оценки ограниченного такими пробами блока представляет собой задачу оценки некоторой функции по известным значениям на границе заданной области. Если пространственную переменную представить как гармоническую функцию, то решение непрерывного крайгинга сведется к решению классической задачи Дирихле.

Обратимся к крайгингу участка, расположенного между двумя параллельными выработками D и D'. Будем считать эти выработки бесконечными прямыми и рассмотрим сначала крайгинг точки A, заключенной между этими прямыми. Примем точку A за начало координат, а расстояние между прямыми — за единицу длины h (т. е. $h = 1$) и обозначим через t и $1-t$ расстояния точки A от прямых D и D'. Выразим пространственную переменную в виде гармонической функции:

$$G = -\frac{1}{2}\ln\frac{e^{2\pi x} - 2e^{\pi x}\cos\pi y + 1}{e^{2\pi x} - 2e^{\pi x}\cos\pi(y-2t) + 1},$$

равной нулю при $y = t$ и $y = 1-t$, т. е. на прямых D и D', и такой, что $G + \frac{1}{2}\ln(x^2+y^2)$ будет гармонической функцией в точке A. Эта гармоническая функция представляет собой функцию Грина, которая позволяет оценивать область, ограниченную двумя бесконечными прямыми. С этой целью находят весовую функцию

$$f(s) = \frac{1}{2\pi}\frac{dG(M, A)}{dn}, \qquad (3.12)$$

где s — абсцисса контура.

Весовые функции для области, ограниченной двумя бесконечными прямыми, получим в соответствии с (3.12) дифференцированием по y:

$$\left.\begin{aligned} f_1(x) &= \frac{e^{\pi x}\sin\pi t}{e^{2\pi x} - 2e^{\pi x}\cos\pi t + 1} \\ f_2(x) &= \frac{e^{\pi x}\sin\pi t}{e^{2\pi x} + 2e^{\pi x}\cos\pi t + 1} \end{aligned}\right\}. \qquad (3.13)$$

Функции $f_1(x)$ и $f_2(x)$ представляют собой решение проблемы точечного крайгинга соответственно для D и D'. Эти функции удобно представить также в виде разложения в ряд Фурье:

$$\left.\begin{aligned} f_1(x) &= \sum_{p=1}^{\infty} e^{-p\pi x}\sin p\pi t, \\ f_2(x) &= \sum_{p=1}^{\infty} e^{-p\pi x}(-1)^{p+1}\sin p\pi t \end{aligned}\right\}. \qquad (3.14)$$

Если вместо точечного крайгинга рассматривать крайгинг произвольного участка S, то достаточно проинтегрировать полученные выражения по всем точкам A площади S, чтобы получить функции взвешивания для неточечного крайгинга. Кроме того, хотя выработки, расположенные по обе стороны оцениваемого участка, имеют конечную длину, их можно считать практически бесконечными, если они имеют длину, превышающую расстояние h, так как экспонента в формулах (3.13) и (3.14) обеспечивает очень быстрое убывание функций $f_1(x)$ и $f_2(x)$ при возрастании x.

Кроме случая оценки области, ограниченной двумя параллельными прямыми, задача непрерывного крайгинга решена для окружности, кольца и бесконечной прямой.

Случайный крайгинг — крайгинг при нерегулярном, но достаточно равномерном расположении разведочных выработок.

Рассмотрим случайный крайгинг на примере оценки блока небольших размеров. Требуется оценить истинное содержание Z в блоке P по небольшому числу проб, отобранных в пределах этого блока, со средним содержанием X и по единственному ореолу, образованному всеми остальными пробами со средним содержанием Y, отобранными по всему рудному телу. Так как число внеш-

них по отношению к блоку P проб велико в сравнении с числом проб, отобранных внутри блока, то с допустимой точностью можно считать, что величина Y представляет собой истинное содержание во всем рудном теле без блока P. Для упрощения расчетов будем считать, что оцениваемый блок занимает в рудном теле случайное положение. Кроме того, общее содержание m в месторождении определено с большей точностью, чем содержание Z в блоке P, и его можно отождествить с содержанием всех внешних проб, т. е. положим $Y = m$ и примем в качестве оценки содержания Z выражение:

$$Z^* = \lambda X + (1-\lambda)\, m.$$

Так как $Z^*-Z = \lambda\,(X-m) - (Z-m)$, то $D\,(Z^*-Z) = \sigma_Z^2 + \lambda^2\sigma_X^2 - 2\lambda\sigma_{ZX}$, где σ_Z^2 и σ_X^2 — дисперсий содержаний Z и X в пределах рудного тела, σ_{ZX} — их коэффициент ковариации.

Поскольку пробы в пределах блока P размещены случайно, и сам блок размещен в рудном теле случайно, то ковариация σ_{ZX} равна дисперсии σ_Z^2 и, следовательно:

$$D\,(Z^*-Z) = \sigma_Z^2 + \lambda^2\sigma_X^2 - 2\lambda\sigma_Z^2.$$

Оптимальное значение λ и соответствующая минимальная величина дисперсии крайгинга σ_K^2 равны:

$$\lambda = \sigma_Z^2/\sigma_X^2,$$

$$\sigma_K^2 = (1-\lambda)\,\sigma_Z^2 = (\sigma_X^2-\sigma_Z^2)\,\frac{\sigma_Z^2}{\sigma_X^2}.$$

При переходе границ, разделяющих разнородные части месторождения, возникают разрывы пространственной переменной, вызывающие эффект самородков в широком смысле (отсюда и название — явление перехода). Таким образом, эффект самородков может интерпретироваться как проявление неоднородности поля. Слагающие такое поле однородные зоны часто бывают слишком небольшими и многочисленными, чтобы их можно было изучать порознь.

При малых a явление перехода может быть описано с помощью транзитивной модели, определяемой функцией:

$$\gamma(r) = C(0) - C(r),$$

где

$$C(r) = \begin{cases} C_0 & \text{при } r = 0, \\ C(r) \leqslant C_0 & \text{при } 0 \leqslant r \leqslant a, \\ 0 & \text{при } r > a. \end{cases}$$

Для выражения функции $C\,(r)$ в транзитивной модели используется вариограмма

$$\gamma(r) = \begin{cases} C\left(\frac{3}{2}\frac{r}{a} - \frac{1}{2}\frac{r^2}{a^3}\right) & \text{при } r \leqslant a, \\ C & \text{при } r > a, \end{cases} \qquad (3.15)$$

которая представляет собой функцию, пропорциональную транзитивной ковариограмме круга или сферы радиуса r.

В условиях сферической модели (3.15) в одномерном случае при различном соотношении отрезка l и периметра a имеем:

для $l < a$

$$\frac{1}{C}\gamma(l) = \frac{3}{2}\frac{l}{a} - \frac{1}{2}\frac{l^3}{a^3},$$

$$\frac{1}{C}\chi(l) = \frac{3}{4}\frac{l}{a} - \frac{1}{8}\frac{l^3}{a^3},$$

$$\frac{1}{C}F(l) = \frac{1}{2}\frac{l}{a} - \frac{1}{20}\frac{l^3}{a^3},$$

для $l \geqslant a$

$$\frac{1}{C}\gamma(l) = 1,$$

$$\frac{1}{C}\chi(l) = 1 - \frac{3}{8}\frac{a}{l},$$

$$\frac{1}{C}F(l) = 1 - \frac{3}{4}\frac{a}{l} + \frac{1}{5}\frac{a^2}{l^2}.$$

Эти формулы позволяют методом последовательных приближений определить носитель a и амплитуду эффекта перехода C по экспериментальным значениям вариограммы и дисперсии, определенной по содержаниям в бороздовых пробах, отобранных вдоль выработки. Наклон вариограммы дает величину отношение l/a. Затем подбором находят такое значение a, при котором функция $F\,(h)$ становится равной экспериментальному значению дисперсии. Дисперсия распространения точечной (бороздовой) пробы в линейной зоне влияния длины l (отрезке выработки) дается общей формулой:

$$\sigma_E^2 = 2\chi\left(\frac{l}{2}\right) - F(l).$$

Эта функция от l используется при расчете дисперсии оценки среднего содержания на участке профиля (ошибки, вносимой при оценке выработки по отобранным в ней пробам).

В двумерном пространстве средние содержания в линейных пробах длиной h определяют новую пространственную переменную, модель изменчивости которой выводится из модели для точечных содержаний посредством обычных операций регуляризации. Алгоритм

$$\gamma(l,\ h) = \frac{2}{h^2}\int_0^h (h-x)\,\gamma\left(\sqrt{x^2+l^2}\right)\,dx$$

дает ковариацию двух отрезков длиной h, отстоящих одна от другого на расстоянии l. В зависимости от соотношения между a, l и

D, где $D=\sqrt{l^2+h^2}$, имеем три выражения для функции $\gamma(l, h)$:

$$\left.\begin{aligned}
&\frac{1}{C}\gamma(l,h)=\frac{3}{2}\frac{D}{a}+\frac{3}{2}\frac{l^2}{ah}\ln\frac{h+D}{l}-\frac{D^3-l^3}{ah^2}-\frac{1}{4}\frac{D^3}{a^3}-\\
&-\frac{3}{8}\frac{l^2D}{a^3}-\frac{3}{8}\frac{l^4}{a^3h}\ln\frac{h+D}{l}+\frac{1}{5}\frac{D^5-l^5}{a^3l^2}\quad\text{при}\quad a>0;\\
&\frac{1}{C}\gamma(l,h)=1+\frac{3}{2}\frac{l^2}{ah}\left(1-\frac{1}{4}\frac{l^2}{a^2}\right)\ln\left(\frac{a+\sqrt{a^2-b^2}}{l}\right)-\\
&-\frac{3}{4}\frac{\sqrt{a^2-b^2}}{h}\left(1+\frac{1}{2}\frac{l^2}{a^2}+\frac{1}{5}\frac{a^2}{h^2}\left(1-\frac{l^5}{a^5}\right)-\right.\\
&\left.-\frac{l^2}{h^2}\left(1-\frac{l}{a}\right)\right)\quad\text{при}\quad D>a>l,\\
&\frac{1}{C}\gamma(l,h)=1\qquad\text{при}\quad a>D.
\end{aligned}\right\}\qquad(3.16$$

Интегрируя (3.16) по l, можно получить вспомогательные функции F и χ. Но число возможных случаев существенно увеличивается: 4 для F и 5 для χ. С помощью функций F и χ может быть получено выражение дисперсии распространения содержания, определенного по горной выработке со средней длиной h, на зону ее влияния, представляемую прямоугольником lh. Это, в свою очередь, позволяет получить дисперсию оценки.

ЛИНЕЙНЫЙ ЭКВИВАЛЕНТ

Линейный эквивалент тела полезного ископаемого или пробы — геометрическая фигура объемом v, включая вырожденный объем, представляющий собой отрезок длиной l, содержание полезного компонента в котором обладает той же дисперсией внутри большого объема V, что и в пробе объемом v'.

В практических приложениях и особенно при расчете дисперсий оценок последовательным проведением операций регуляризации удобно свести задачу к одномерному случаю. В дополнение к вариограмме $\gamma(h)$ одномерной модели для этого используются две вспомогательные вариограммы $\chi(h)$ и $F(h)$, определяемые равенствами:

$$\chi(h)=\frac{1}{h}\int_0^h\gamma(x)\,dx;$$

$$F(h)=\frac{2}{h^2}\int_0^h x\chi(x)\,dx=\frac{2}{h^2}\int_0^h(h-x)\gamma(x)\,dx.$$

Функция $\chi(h)$ выражает среднее значение функции $\gamma(x)$, когда x является расстоянием между текущей точкой отрезка h и одним из его фиксированных концов. Функция $F(h)$ — это среднее значение функции $\gamma(x)$, когда x — расстояние между двумя точками,

занимающими независимо одна от другой все возможные положения на отрезке h. Функция $F(h)$ позволяет рассчитать дисперсию среднего значения, определенного по отрезку l, в пределах поля длины L, а именно:

$$\sigma^2(l\,|\,L)=F(L)-F(l).$$

Аналогично, функция $\chi(h)$ позволяет вычислять в пределах поля L ковариацию $\sigma(0, l\,|\,L)$ отрезка l и одного из его концов:

$$\sigma(0;\ l\,|\,L)=F(L)-\chi(l).$$

Дисперсия распространения среднего содержания, определенного по малой пробе l, на отрезок L, в центре которого расположена проба, может быть выражена равенством:

$$\sigma_E^2=2\chi\left(\frac{l}{2}\right)-F(L)-F(l).$$

При решении многочисленных практических задач возникает необходимость оценить среднее значение пространственной переменной в пределах поля V по результатам дискретного опробования. Одной из таких задач является оценка величины

$$Z=\int_V p(x)f(x)\,dx,$$

где $p(x)$ — весовая функция, равная нулю вне поля V и удовлетворяющая условию

$$\int_V p(x)\,dx=1.$$

Величина Z представляет собой среднее содержание, взвешенное на $p(x)$ в поле V. Лучшая оценка среднего значения Z функции $f(x)$ в объеме V по N точечным пробам может быть получена, если заменить простое среднее арифметическое некоторым средним взвешенным, приписывая каждому значению $f(x_i)$ в точке x_i вес, определяемый с учетом большей или меньшей корреляции $f(x_i)$ с искомой величиной Z и зависящий, следовательно, от положения точки x_i относительно объема V.

Рассмотрим оценку

$$Y=\sum_{i=1}^N\lambda_i f(x_i)+\lambda_0 m,$$

считая истинное среднее значение функции $m=M[f(x)]$ известным. Коэффициенты взвешивания λ определим из условий минимума дисперсии оценки $M[(Z-Y)^2]$. Для выполнения условий $M(Y)=M(Z)$ коэффициент λ_0 должен быть взят равным $1-\Sigma\lambda_i$, в результате чего выбранная оценка должна принять следующий вид:

$$Y=m+\sum_{i=1}^N\lambda_i[f(x_i)-m].\qquad(3.17)$$

Тогда дисперсия оценки разности $Z-Y$ будет равна:

$$D(Z-Y)=\sigma_Z^2-2\sum_{i=1}^N\lambda_i\sigma_{Zi}+\sum_{i,j=1}^N\lambda_i\lambda_j\sigma_{ij},$$

где σ_{ij} — коэффициент ковариации значений $f(x_i)$ и $f(x_j)$; σ_{Zi} — коэффициент ковариации $f(x_i)$ и Z.

Чтобы это выражение было минимальным, коэффициенты λ_i должны удовлетворять системе линейных уравнений, получаемых в результате приравнивания к нулю частных производных дисперсий $D(Z—Y)$ по λ_i, т. е. системе:

$$\sum_i \lambda_i \sigma_{ij} = \sigma_{Zi}; \qquad i = 1, \ldots, N. \tag{3.18}$$

Оценка Y реализует крайгинг величины Z, когда весовые коэффициенты λ_i удовлетворяют этой системе. Дисперсия крайгинга равна минимальному значению дисперсий $D(Z—Y)$. Умножив обе части выражения (3.18) на λ_i и просуммировав их по i, получим:

$$\sum_{ij} \lambda_i \lambda_j \sigma_{ij} = \sum_i \lambda_i \sigma_{Zi}.$$

Отсюда вытекает, что дисперсия крайгинга равна

$$\sigma_K^2 = \sigma_Z^2 - \sum_{i=1}^{N} \lambda_i \sigma_{Zi}.$$

В практике среднее значение m неизвестно и должно быть оценено по имеющимся данным. Оно может быть представлено суммой:

$$m = \sum_i \mu_i f(x_i), \qquad \text{где} \quad \Sigma \mu_i = 1.$$

Тогда оценка (3.17) сводится к следующему:

$$Y = \sum_i a_i f(x_i) \qquad \text{где} \quad \sum_i a_i = 1.$$

Условие нормировки $\Sigma a_i = 1$ необходимо для выполнения равенства:

$$M(Y) = M(Z).$$

В рамках модели Де Вийса дисперсия пробы v в поле V выражается формулой:

$$\sigma^2(v|V) = \frac{3\alpha}{V^2} \int_V \int_V \ln|x-x'|\,dxdx' - \frac{3\alpha}{v^2} \int_v \int_v \ln|x-x'|\,dxdx'. \tag{3.19}$$

Справедливая для гомотетичных форм v и V формула этой дисперсии неприменима, если v и V имеют произвольные формы. Она может быть обобщена введением понятия эквивалентности двух проб. Пробы v и v' считаются эквивалентными, если в произвольном поле V имеют одну и ту же дисперсию.

Рассмотрим функцию

$$F(v) = \frac{1}{v^2} \int_v \int_v \ln|x-x'|\,dxdx', \tag{3.20}$$

которая представляет собой среднее значение $\ln r$ в объеме v и зависит не только от размера, но и от формы объема v. С учетом (3.20) выражение (3.19) приобретает вид:

$$\sigma^2(v/V) = 3\alpha\,[F(V) - F(v)]. \tag{3.21}$$

Из (3.21) следует, что для эквивалентности проб v и v' необходимо и достаточно, чтобы $F(v) = F(v')$. Эта эквивалентность не зависит от параметра модели Де Вийса (коэффициента абсолютного рассеивания) и имеет, следовательно, чисто геометрический смысл. Совокупность всех проб, эквивалентных пробе v, характеризуется общим значением $F(v)$. В частности, среди этих проб имеется отрезок l, при котором

$$F(l) = \ln l - \frac{3}{2} = F(v).$$

Этот отрезок l называется линейным эквивалентом v. Совокупность эквивалентных проб характеризуется их общим линейным эквивалентом:

$$\ln l = F(v) + \frac{3}{2}.$$

Обозначив линейные эквиваленты пробы и ее геометрического поля V через l и L, получим:

$$\sigma^2(v|V) = 3\alpha \ln \frac{L}{l}.$$

Определение линейных эквивалентов различных геометрических фигур представляет собой сложную проблему. В качестве примера приведем первые члены разложения линейного эквивалента прямоугольного параллелепипеда со сторонами $a \geqslant b \geqslant c$:

$$\begin{aligned} L(a, b, c) = {} & \ln a + \frac{\pi}{3}\frac{b}{a} + \frac{1}{6}\frac{b^2}{a^2}\ln\frac{b}{a} - \frac{25}{72}\frac{b^2}{a^2} - \\ & - \frac{1}{180}\frac{b^4}{a^4} + \frac{1}{1680}\frac{b^6}{a^6} + \ldots + \frac{c^2}{b^2}\left[\frac{\pi}{6}\left(\ln 2 + \frac{1}{12}\right)\frac{b}{a} + \right. \\ & \left. + \frac{1}{6}\frac{b^2}{a^2}\ln\frac{b}{a} - \frac{5}{12}\frac{b^2}{a^2} - \frac{1}{216}\frac{b^4}{a^4} + \frac{1}{1800}\frac{b^6}{a^6} + \ldots\right] - \\ & - \frac{\pi}{6}\frac{c^2}{ab}\ln\frac{c}{b} + \frac{\pi}{15}\frac{c^3}{b^3}\left[\frac{b}{a} + \frac{b^2}{a^2}\right] + \frac{c^4}{b^4}\left[-\frac{\pi}{120}\frac{b}{a} - \right. \\ & \left. - \frac{127}{1800}\frac{b^2}{a^2} - \frac{1}{180}\frac{b^4}{a^4} + \frac{1}{1800}\frac{b^6}{a^6} + \ldots\right] + \\ & + \frac{1}{30}\frac{c^4}{a^2b^2}\ln\frac{c}{b} + \frac{c^6}{b^6}\left[\frac{\pi}{2688}\frac{b}{a} - \right. \\ & \left. - \frac{1}{1008}\frac{b^2}{a^2} + \frac{1}{1680}\frac{b^6}{a^6} + \ldots\right] + \ldots. \end{aligned} \tag{3.22}$$

Эта функция и подобные ей табулированы Ж. Матероном. Кроме того, им предложены простые приближенные формулы, которые позволяют с удовлетворительной точностью рассчитать линейные эквиваленты различных многоугольников и многогранников. В частности, приближенным для (3.22) является выражение:

$$L = a + b + c/2.$$

Понятием эквивалентности проб пользуются для сопоставления различных видов опробования, а также для быстрой оценки коэффициента абсолютного рассеивания по экспериментальному значению дисперсии в предположении, что изменчивость изучаемого признака может быть описана моделью Де Вийса. Кроме того, это понятие позволяет прогнозировать поведение экспериментальной вариограммы при больших расстояниях между пробами. Если l — линейный эквивалент используемых проб, то дисперсия разности $[f(x+d) - f(x)]$ содержаний в точках, отстоящих одна от другой на расстоянии d, равна:

$$\sigma_d^2 = 6\alpha\left[\ln\frac{d}{l} + \frac{3}{2}\right]$$

при условии, что d не менее чем в два-три раза больше l.

РЕГУЛЯРИЗАЦИЯ ПРОСТРАНСТВЕННОЙ ПЕРЕМЕННОЙ

Регуляризация пространственной переменной — ее взвешивание на геометрическую базу. В рамках транзитивной теории регуляризация пространственной переменной интерпретируется следующим образом.

Если $f(x)$ — пространственная переменная на точечной геометрической базе, а v — некоторые реальные объемы исследуемого геологического объекта, то новая пространственная переменная

$$t(x) = \frac{1}{v}\int_v f(x+h)\,dh$$

отличается от первой большей регулярностью. Функция $t(x)$ называется функцией $f(x)$, регуляризованной по объему v.

Если $p(h)$ — некоторая весовая функция, определяемая соотношением $p_0 = S\,p(h)\,dh$, то в общем виде

$$t(x) = \frac{1}{P_0}\int f(x+h)\,p(h)\,dh.$$

Если в качестве весовой функции используется геометрическая переменная $l(x)$ объема v, то

$$t(x) = \frac{1}{v}\int_v f(x+h)\,l(h)\,dh \qquad \text{при} \qquad v = \int_v l(h)\,dh.$$

В этом случае $l(x)$ является геометрической базой.

Ковариограмма переменной $t(x)$ получается путем регуляризации ковариограммы $G(h)$ точечной функции $f(x)$ посредством взвешивания на ковариограмму весовой функции, в качестве которой выступает $L(h)$:

$$G_t(h) = \frac{1}{v^2}\int G(h+u)\,L(u)\,du,$$

т. е. $G_t(h)$ равна среднему значению $G(h+u)$, когда концы вектора $(h+u)$ занимают все возможные положения внутри областей v_1 и v_2, равных v и смещенных относительно друг друга на вектор h.

Влияние регуляризации на аналитические свойства пространственной переменной заключается в следующем. Если, например, весовая функция $p(h)$ дифференцируема один раз (хотя бы в среднем квадратическом), то ее ковариограмма $p(h)$ дифференцируема уже дважды. Тогда, согласно правилу дифференцирования свертки функций, $G_t(h)$ дифференцируема дважды и, следовательно, $t(x)$ дифференцируема в среднем квадратическом, даже если $f(x)$ недифференцируема.

Последовательное взвешивание пространственной переменной на зону влияния меньшей размерности (точечной переменной на линейную зону влияния, линейной переменной на площадную зону влияния) и всякое последовательное сокращение размерности зоны влияния переменной в n-мерном пространстве, т. е. переход от $f_n(x)$ к $f_{n-1}(x)$, будем называть регуляризацией первого порядка вдоль оси (x_n), а переход от $f_n(x)$ к $f_{n-2}(x)$ вдоль плоскости (x_{n-1}, x_n) регуляризацией второго порядка и т. д. Операцию, обратную регуляризации, т. е. отыскание функции по значениям ее интегралов на всех прямых, плоскостях или гиперплоскостях пространства, будем называть сокращением регулярности.

В наибольшей степени изучены пространственные переменные $f_n(x)$, которые следуют изотропной транзитивной модели или могут быть сведены к ней. Соответствующая изотропной пространственной переменной ковариограмма $G_n(h)$ зависит не от направления вектора h, а только от его модуля $|h| = r$, где

$$r = \sqrt{h_1^2 + h_2^2 + \ldots + h_n^2}.$$

В случае изотропной функции $f(x)$ ковариограммы регуляризации первого и второго порядков имеют вид:

$$G_{n-1}(r) = 2\int_0^\infty G_n\left(\sqrt{r^2+x^2}\right)dx,$$

$$G_{n-2}(r) = 2\pi\int_0^\infty G_n\left(\sqrt{r^2+\rho^2}\right)\rho\,d\rho,$$

где ρ — полярные координаты в плоскости (x_n, x_{n-1}).

Операция регуляризации в транзитивной теории толкуется как предельный случай регуляризации пространственной переменной с помощью функции опробования $p(x) = l(x)$, соответствующей пробе v, когда геометрической базой пробы является бесконечная прямая. При рассмотрении случайных функций операция регуляризации может быть проведена вдоль отрезка прямой конечной длины. Возможны три варианта операции регуляризации вдоль отрезка конечной длины в зависимости от того, является ли этот отрезок (называемый мощностью) случайной, функционально изменяющейся, или постоянной величиной. Ж. Матерон дает решение операции регуляризации для простейшего случая — для постоянной мощности. В частности, когда регуляризация прямая (т. е. оцениваемая плоскость П перпендикулярна к направлению α, вдоль ко-

торого проводится операция регуляризации), а модель изотропна, алгоритм изотропной регуляризации первого порядка приобретает вид:

$$\gamma_{n-1}(r)=\frac{2}{l^2}\int_0^l (l-x)\,\gamma_n\left[\sqrt{r^2+x^2}\right]dx-\frac{2}{l^2}\int_0^l (l-x)\,\gamma_n(x)\,dx.$$

Определенная таким образом функция $\gamma_{n-1}(h)$ позволяет рассчитать все дисперсии и ковариации распространения, связанные с моделью $f_{n-1}(x)$ в плоскости П.

Алгоритм изотропной прямой регуляризации второго порядка имеет следующий вид:

$$\gamma_{n-2}(r)=\frac{4}{l^2l'^2}\int_0^l\int_0^{l'}(l-x)(l'-x')\,\gamma_n\left[\sqrt{r^2+x^2+x'^2}\right]dxdx'-$$

$$-\frac{4}{l^2l'^2}\int_0^l\int_0^{l'}(l-x)(l'-x')\,\gamma_n\left[\sqrt{x^2+x'^2}\right]dxdx'.$$

Этот алгоритм описывает характер изменчивости среднего содержания в прямоугольном сечении прямой призмы (все три вектора — h, α, α' — взаимно перпендикулярны), основанием которой служит прямоугольник со сторонами l и l'.

В рамках теории случайных функций регуляризации пространственной переменной интерпретируется следующим образом.

Регуляризованной функцией fp является функция, полученная путем взвешивания случайной функции f на весовую функцию p. Регуляризованная функция fp сама является случайной функцией, несколько более регулярной, чем исходная. Это следует из выражения ковариационной функции регуляризованной случайной функции fp:

$$K_p(h)=E[f_p(x)f_p(x+h)]=E[\int\int f(y+x)\,p(y)f(y'+x+h)\,p\times$$
$$\times(y'-x-h)\,dydy']=\int\int K(y-y'-h)\,p(y)\,p(y')\,dvdv'$$

Эта формула представляет собой свертку ковариационной функции K случайной функции f и транзитивной ковариограммы P весовой функции p. Функция взвешивания p передает функции f присущие ей (функции взвешивания) черты регулярности.

Если V и V' — два объема, а Y и Z — средние значения случайной функции в этих объемах, т. е. значения стохастических интегралов

$$Y=\frac{1}{V}\int f(x)\,dx \quad \text{и} \quad Z=\frac{1}{V'}\int f(x)\,dx,$$

то ковариация Y и Z определяется формулой

$$\sigma_{YZ}=\frac{1}{VV'}E\left[\int_V\int_V f(x)f(x')\,dxdx'\right]=$$

$$=\frac{1}{VV'}\int_V\int_{V'}K(x-x')\,dxdx'.$$

Эта ковариация равна среднему значению ковариационной функции $K(h)$, когда концы вектора $h=x-x'$ занимают независимо один от другого все возможные положения в объемах V и V'.

ЭФФЕКТ САМОРОДКОВ

Эффект самородков — резкие изменения содержаний полезных компонентов на очень небольших расстояниях, в тех случаях, когда минерализация представлена самородками, мелкими гнездами или прожилками; здесь вариограмма отличается от нуля в начале графика ($\gamma(h)=C_0$ при $h\to 0$, где $C_0\neq 0$). Если содержания полезных компонентов независимы при любом значении h, то имеет место чистый эффект самородков.

В рамках модели Де Вийса нельзя учесть эффект самородков. Разрыв или скачок вариограммы точечной переменной, наблюдаемый в начале координат, называется константой самородков C_0. Величина этой константы выражает интенсивность эффекта самородков. Если обозначить символом $E(r)$ единичную ступень, т. е. функцию, определенную равенством:

$$E(r)=\begin{cases}1, & \text{при } r>0,\\ 0, & \text{при } r=0,\end{cases}$$

то вариограмма может быть представлена в виде суммы двух составляющих:

$$\gamma(r)=C_0E(r)+\gamma_1(r),$$

где $C_0E(r)$ — эффект самородков в чистом виде; $\gamma_1(r)$ — непрерывная в нуле функция, характеризующая пространственную изменчивость.

Таким образом, при расчете всех необходимых дисперсий и ковариаций точечная пространственная переменная X может рассматриваться как сумма некоторой теоретической пространственной переменной X_Y, свободной от эффекта самородков и обладающей вариограммой первого или второго типа, и чисто случайной переменной ε с нулевым средним и дисперсией C_0, т. е. $X=X_Y+\varepsilon$, где X_Y и ε статистически независимы.

Если ограничиться изучением изменения точечного содержания x в области, непосредственно прилегающей к данной точке, т. е. если рассматривать только малые значения расстояний r, то изменение функции $\gamma(r)$ в этой области также будет весьма малым и вариограмму $\gamma(r)$ можно считать постоянной и равной константе эффекта самородков C_0. Локальное изменение содержаний практически полностью происходит за счет случайной компоненты ε, т. е. пространственная переменная ведет себя в локальной области как случайная величина. Реальному физическому явлению не может соответствовать вариограмма с разрывом в точке начала координат. В месте действительного разрыва в окрестностях точки $r=0$ существует зона быстрого перехода, т. е. резкого изменения вариограммы, что

характеризуется функцией $C(r)$, удовлетворяющей следующим условиям:

$$C(0)=0,$$
$$C(r)\leqslant C_0, \quad \text{при} \quad r\leqslant a, \qquad (3.23)$$
$$C(r)=C_0, \quad \text{при} \quad r>a,$$

где a — постоянная, называемая носителем эффекта самородков, которая указывает порядок величины зоны перехода, т. е. порядок размеров самородков.

С учетом (3.23) вариограмма точечного содержания может быть представлена как

$$\gamma(r)=C_0-C(r)+\gamma_1(r),$$

где C_0 — константа самородков; $\gamma(r)$ — вариограмма, непрерывная в нуле.

Точечное содержание x в этом случае представляется в виде суммы:

$$x=x_y+\varepsilon \qquad (3.24)$$

двух независимых пространственных переменных x_y и ε, имеющих в качестве вариограммы функции $\gamma_1(r)$ и $C_0-C(r)$ соответственно. В этом случае значения ε независимы одно от другого только на расстояниях, превышающих носитель a. На расстояниях, меньших a, их изменчивость характеризуется линейной вариограммой. Следовательно, эффект самородков будет отражаться и на пробах ненулевого объема. Если этот объем больше объема a^3 рудных зерен, то зона перехода оказывается включенной в объем v, в пределах которого интегрируется вариограмма, и эффект самородков порождает дополнительную дисперсию, зависящую от a^3/v, т. е. обратно пропорциональную числу зерен, составляющих пробу v. Если v и V — объемы пробы и залежи, то дисперсия σ_c^2, связанная с эффектом самородков, т. е. дисперсия средних значений ε в пробах v в пределах залежи V, может быть представлена в виде:

$$\sigma_C^2=\frac{1}{v^2}\int_v dv_1\int_v C(r)\,dv_2-\frac{1}{V^2}\int_V dV_1\int C(r)\,dV_2.$$

Общую дисперсию таких проб можно записать в виде суммы:

$$\sigma^2=\sigma_C^2+\sigma_y^2,$$

где σ_c^2 — дисперсия самородков; σ_y^2 — теоретическая дисперсия, рассчитываемая с помощью непрерывной составляющей $\gamma_1(r)$ вариограммы и интерпретируемая как дисперсия пространственной переменной x_y из (3.24).

Эффект самородков оказывает серьезное влияние на решение проблемы крайгинга. В тех случаях, когда дисперсия самородков преобладает, наилучшей оценкой содержания в любом блоке будет среднее арифметическое из содержаний всех проб независимо от места их взятия. В промежуточных случаях (при наличии эффекта

самородков) ослабевает влияние экрана, и обычные схемы крайгинга, в которых учитывают один или два ореола ближайших скважин, в значительной мере теряют свою эффективность. Так как учет большого числа ореолов практически невозможен из-за возникающих при этом вычислительных трудностей, то целесообразно объединять все внешние ореолы в один, приписывая ему содержание m, равное общему среднему содержанию в месторождении. Учет эффекта самородков приводит к случайному крайгингу.

ГЛАВА 4

ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ГЕОЛОГИЧЕСКИХ ПЕРЕМЕННЫХ

Интервальные оценки геологических переменных характеризуют их точность при заданной надежности. Понятие «доверительный интервал» введено Дж. Нейманом и Е. Пирсоном. Так называют вычисленный по выборочным значениям интервал, который с заданной вероятностью — надежностью $P=1-\alpha$ — накрывает истинное, неизвестное исследователю значение параметра. В отличие от точечных оценок в виде одного числа интервальные оценки характеризуют нижнюю и верхнюю доверительные границы при заранее заданной надежности $1-\alpha$ (α — риск, вероятность того, что истинное значение параметра не накрывает данный интервал).

Известны три основных типа интервального оценивания: доверительные интервалы по Нейману, основанные на частотной теории вероятностей; фидуциальные интервалы по Фишеру, использующие идеи, не охватываемые частотной теорией; доверительные интервалы по Байесу, основанные на теореме Байеса и на одной из форм постулата Байеса. Для применения последних интервальных процедур нужна априорная информация, чаще всего отсутствующая для реальных геологических ситуаций [8, 23, 25]. Наиболее важные для геологических исследований неймановские доверительные интервалы можно разделить на интервалы для единственной геологической переменной и для набора переменных. Первые включают: точное оценивание параметров на основе достаточной статистики и стьюдентизации; асимптотическое оценивание параметров на основе первой производной или на основе второй и более высоких производных; оценивание параметров по расслоенным выборкам (методом повторных оценок) для нормально, логнормально и полимодально распределенных совокупностей данных.

Доверительные интервалы для набора геологических переменных включают: точное и асимптотическое оценивание параметров на основе стьюдентизации, оптимизационной основе, а также оценивание параметров по расслоенным выборкам для функций в виде

произведений, отношений и сложных отношений геологических переменных [23, 25]. В качестве геологических переменных можно рассматривать: содержания основных и попутных компонент в руде, элементов-индикаторов в геохимических аномалиях, элементов-примесей в минералах; индикаторные отношения элементов, продуктивностей, прогнозных ресурсов по категориям P_3, P_2 и P_1, запасов полезных ископаемых по категориям C_2, C_1, B и A; разнообразные кондиционные показатели разведки и освоения месторождений, показатели в поисковой и технологической минералогии, показатели осевой, продольной и поперечной зональности первичных и вторичных геохимических ореолов и потоков рассеяния, показатели мультипликативных и аддитивных геохимических суммарных ореолов и т. п.

Ниже приведены перспективные неймановские процедуры интервального оценивания по расслоенным выборкам.

ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПРОСТЫХ ГЕОЛОГИЧЕСКИХ ПЕРЕМЕННЫХ

Для получения интервальных оценок необходимо найти ряд характеристик выборочных распределений, а именно: среднее, стандартное отклонение, границы доверительного интервала Ниже эти характеристики приведены для различных распределений.

1. Для одномерного нормального распределения. Геологическая переменная — содержание компонента, мощность рудного тела и т. п.— замерена (опробована) в n точках:

$$\{x_t\}, \quad t=1, 2, \ldots, n;$$

$$\bar{x}=\frac{1}{n}\sum_{t=1}^{n} x_t;$$

$$S^2=\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\bar{x}_n)^2; \qquad S=\sqrt{S^2};$$

$$\pm\lambda_{\bar{x}}=t_{1-\alpha/2}\frac{S}{\sqrt{n}}; \qquad \pm\lambda^0_{\bar{x}}=(\lambda_{\bar{x}}/\bar{x})\,100\,\%;$$

$$(\bar{x}-\lambda_{\bar{x}},\ \bar{x}+\lambda_{\bar{x}});$$

$$P\left[(\bar{x}-\lambda_{\bar{x}})\leqslant a\leqslant(\bar{x}+\lambda_{\bar{x}})\right]\simeq 1-\alpha,$$

где a — неизвестный параметр.

2. Для одномерного логнормального распределения:

$$\{x_t\}, \quad t=1, 2, \ldots, n;$$

$$\{\lg x_t\}, \quad t=1, 2, \ldots, n;$$

$$\overline{\lg x}=\frac{1}{n}\sum_{t=1}^{n}\lg x_t; \qquad S^2_{\lg}=\frac{1}{n-1}\sum_{t=1}^{n}(\lg x_t-\overline{\lg x})^2;$$

$$S_{\lg}=\sqrt{S^2_{\lg}}\,.$$

Максимально правдоподобная оценка среднего $\hat{a}$ по Ачисону и Брауну [39]:

$$\hat{a}=10^{\overline{\lg x}}\,\psi_n(t); \quad t=2{,}65S^2_{\lg};$$

$$\psi_n(t)=e^t\left\{1-\frac{t(t+1)}{n}+\frac{t^2(3t^2+22t+21)}{6n^2}\ldots\right\},$$

$$e=2{,}718\ldots;$$

$$\pm\lambda_{\hat{a}}=t_{1-\alpha/2}\frac{\hat{a}}{\sqrt{n}}\sqrt{S^2_{\lg}+0{,}5S^4_{\lg}}\,;$$

$$\pm\lambda^0_{\hat{a}}=(\lambda_{\hat{a}}/\hat{a})\,100\,\%;$$

$$(\hat{a}-\lambda_{\hat{a}},\ \hat{a}+\lambda_{\hat{a}});$$

$$P\left[(\hat{a}-\lambda_{\hat{a}})\leqslant a\leqslant(\hat{a}+\lambda_{\hat{a}})\right]\simeq 1-\alpha.$$

3. При наличии аномальных наблюдений принимают во внимание рекомендации Диксона и Масси по оцениванию среднего $\bar{x}_{дм}$ и стандартного отклонений $S_{дм}$ по загрязненным выборкам. В качестве $\bar{x}_{дм}$ и $S_{дм}$ в зависимости от степени загрязнения выборки [8, 25] используются обычное среднее арифметическое $\bar{x}$ и медиана Ме, обычное стандартное отклонение S и оценка по размаху:

$$\pm\lambda_{дм}=t_{1-\alpha/2}\frac{S_{дм}}{\sqrt{n}}; \qquad \pm\lambda^0_{дм}=(\lambda_{дм}/\bar{x}_{дм})\,100\,\%;$$

$$(\bar{x}_{дм}-\lambda_{дм},\ \bar{x}+\lambda_{дм});$$

$$P\left[(\bar{x}_{дм}-\eta_{дм})\leqslant a\leqslant(\bar{x}_{дм}+\lambda_{дм})\right]\simeq 1-\alpha.$$

4. По бимодальной выборке наблюдений. Методом С. В. Гольдина подтверждаем бимодальность распределения. Для каждой локальной совокупности данных приемами 1—3 строим доверительные интервалы.

5. По устойчивым винзоризованным выборкам [6]:

$$\{x_t\}, \ t=1, 2, \ldots, n;$$

$$x_1\leqslant x_2\leqslant\ldots\leqslant x_n.$$

Процедура g-винзоризации заключается в замене g первых членов на $(g+1)$-й член, а g последних членов — на $(n-g)$-й член:

$$Z_1=Z_2=\ldots=Z_g=x_{g+1}\leqslant x_{g+2}\leqslant\ldots\leqslant x_{n-g-1}\leqslant x_{n-g}=$$
$$=x_{n-g+1}=\ldots=Z_{n-1}=Z_n;$$

$$\bar{Z}=\frac{1}{n}\sum_{t=1}^{n}Z_t;$$

$$S^2(Z)=\frac{1}{n-1}\sum_{t=1}^{n}(Z_t-\bar{Z})^2; \qquad S(Z)=\sqrt{S^2(Z)};$$

$$\pm\lambda_{\bar{z}} = t_{(1-\alpha/2),\,(h-1)}\frac{(n-1)\,S\,(Z)}{(h-1)\sqrt{n}}; \qquad \pm\lambda^0_{\bar{z}} = \frac{\lambda_{\bar{z}}}{\bar{z}}\,100\,\%;$$

$$h = n - 2g; \qquad t_{(1-\alpha/2),\,(n-1)},$$

где $t_{(1-\alpha/2),\,(h-1)}$ — квантиль распределения Стьюдента при $h-1$ степенях свободы;

$$\left(\bar{z}-\lambda_{\bar{z}},\ z+\lambda_{\bar{z}}\right);$$

$$P\left[\left(\bar{z}-\lambda_{\bar{z}}\right) \leqslant a \leqslant \left(\bar{z}+\lambda_{\bar{z}}\right)\right] \simeq 1-\alpha.$$

6. Для выборочных данных с пропущенными наблюдениями. Для «восстановления» пропущенных наблюдений используют рекомендации А. Афифи и С. Эйзена [6], а затем строят доверительные интервалы, применяя приемы 1—5.

7. По угловым ориентированным наблюдениям [26, 31]:

$\{\nu^0_t\}$, $t = 1, 2, \ldots, n$, ν^0_t — угловые замеры в градусах;

$$\bar{\nu}^\circ = \operatorname{arctg}\left(\bar{S}/\bar{C}\right);$$

$$\bar{S} = \frac{1}{n}\sum_{t=1}^{n}\sin\nu^\circ_t; \qquad \bar{C} = \frac{1}{n}\sum_{t=1}^{n}\cos\nu^\circ_t;$$

$$\bar{R} = \sqrt{\bar{C}^2+\bar{S}^2};$$

$$\hat{k} \simeq \frac{1}{6}\bar{R}\left(12+6\bar{R}^2+5\bar{R}^4\right), \quad \text{при} \quad \bar{R} \geqslant 0;$$

$$\hat{k} \simeq 1/\left[2\left(1-\bar{R}\right)-\left(1-\bar{R}\right)^2-\left(1-\bar{R}\right)^3\right], \qquad \text{при } R \leqslant 1;$$

$$k' = n\bar{R}\hat{k};$$

$$\pm\lambda_{\bar{\nu}^\circ} = 57{,}296^\circ\frac{t_{1-\alpha/2}}{\sqrt{k'}}; \qquad \pm\lambda^\circ_{\bar{\nu}^\circ} = \left(\lambda_{\bar{\nu}^\circ}/\bar{\nu}\right)100\,\%;$$

$$\left(\bar{\nu}^\circ-\lambda_{\bar{\nu}^\circ},\ \bar{\nu}+\lambda_{\bar{\nu}^\circ}\right);$$

$$P\left[\left(\bar{\nu}^\circ-\lambda_{\bar{\nu}^\circ}\right) \leqslant a^\circ \leqslant \left(\bar{\nu}^\circ+\lambda_{\bar{\nu}^\circ}\right)\right] \simeq 1-\alpha,$$

где a° — неизвестный угловой параметр, градусы (например, азимут падения рудной зоны).

ИНТЕРВАЛЬНЫЕ ОЦЕНКИ СЛОЖНЫХ ГЕОЛОГИЧЕСКИХ ПЕРЕМЕННЫХ

1. Для прогнозных ресурсов изученного рудоносного поля. Расчеты минеральных ресурсов Q осуществляется по формуле

$$Q = \frac{1}{100}CMLhdK_H,$$

где C — содержание полезного компонента, %; M — мощность рудной залежи; м; L — протяженность рудной залежи по простиранию, м; h — протяженность рудной залежи по вертикали, м; d — объемная масса руды, т/м^3; K_H — коэффициент надежности.

По трем переменным C, M и L в отдельности находят интервальные оценки (см. выше).

Сочетание вариантов:

$$\left(\bar{C}-\lambda_{\bar{C}},\ \bar{C},\ \bar{C}+\lambda_{\bar{C}};\ \bar{M}-\lambda_{\bar{M}},\ \bar{M},\ \bar{M}+\lambda_{\bar{M}};\ \bar{L}-\lambda_{\bar{L}},\ \bar{L},\ \bar{L}+\lambda_{\bar{L}}\right),$$

приводит к $n = 3\times3\times3 = 27$ оценкам прогнозных ресурсов

$$\hat{Q}_t,\ t = 1, 2, \ldots 27.$$

$$\{\hat{Q}_t\},\ t = 1, 2, \ldots, 27;$$

$$\bar{Q} = \frac{1}{n}\sum_{t=1}^{n}\hat{Q}_t = \frac{1}{27}\sum_{t=1}^{27}\hat{Q}_t;$$

$$S^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(\hat{Q}_t-\bar{Q}\right)^2 = \frac{1}{26}\sum_{t=1}^{27}\left(\hat{Q}_t-\bar{Q}\right)^2;$$

$$S = \sqrt{S^2};$$

$$\pm\lambda_{\bar{Q}} = t_{1-\alpha/2}S \qquad (\text{при } \alpha = 0{,}05 \quad \pm\lambda_{\bar{Q}} = 1{,}96S);$$

$$\pm\lambda^\circ_{\bar{Q}} = \left(\lambda_{\bar{Q}}/\bar{Q}\right)100\,\% \qquad \left(\text{при } \alpha = 0{,}05\ \pm\lambda^0_{\bar{Q}} = \frac{196S}{\bar{Q}}\ \%\right);$$

$$\left(\bar{Q}-\lambda_{\bar{Q}},\ \bar{Q}+\lambda_{\bar{Q}}\right);$$

$$P\left[\left(\bar{Q}-\lambda_{\bar{Q}}\right) \leqslant Q \leqslant \left(\bar{Q}+\lambda_{\bar{Q}}\right)\right] \simeq 1-\alpha.$$

2. Для прогнозных ресурсов новых рудоносных полей. Расчеты минеральных ресурсов $Q_н$ новых рудоносных полей осуществляются по формуле

$$Q_н = S_н D_н D_р p_н k_н,$$

где $S_н$ и $D_н$ — соответственно площадь нового рудоносного поля и доля площади рудовмещающей толщи в ней; $D_р$ — доля рудной площади продуктивного горизонта от площади рудовмещающей толщи; $p_н$ — количество металла, приходящееся на единицу площади продуктивного горизонта (продуктивность) по имеющимся фактическим данным на новом рудоносном поле; $k_н$ — коэффициент надежности, устанавливаемый экспертным путем.

В формуле переменной является по крайней мере $p_н$, т. е. определяется по исходным данным объема $n \geqslant 2$ в новом рудоносном поле. Подставляя их в формулу, имеем $\{\hat{Q}_н(t)\}$, $t = 1, 2, \ldots, n$ локальных определений прогнозных ресурсов $\hat{Q}_н(t)$:

$$\{\hat{Q}_н(t)\},\ t = 1, 2, \ldots, n;$$

$$\bar{Q}_н = \frac{1}{n}\sum_{t=1}^{n}\hat{Q}_н(t);$$

$$S^2 = \frac{1}{n-1}\sum_{t=1}^{n}[\hat{Q}_н(t) - \bar{Q}_н]^2; \qquad S = \sqrt{S^2};$$

$$\pm\lambda_{\bar{Q}_н} = t_{1-\alpha/2}\, S/\sqrt{n}; \qquad \pm\overset{\circ}{\alpha}_{\bar{Q}_н} = \left(\lambda_{\bar{Q}_н}/\bar{Q}_н\right) 100\ \%;$$

$$\left(\bar{Q}_н - \lambda_{\bar{Q}_н},\ \ \bar{Q}_н + \lambda_{\bar{Q}_н}\right);$$

$$p\left[\left(\bar{Q}_н - \lambda_{\bar{Q}_н}\right) \leqslant Q_н \leqslant \left(\bar{Q}_н + \lambda_{\bar{Q}_н}\right)\right] \simeq 1-\alpha.$$

3. Для прогнозных ресурсов по параметрам вторичных остаточных ореолов рассеяния элементов-индикаторов. Прогнозные ресурсы Q рассчитываются по формуле

$$Q = kHq = kHp/40,$$

где k — коэффициент пропорциональности, устанавливаемый в каждом районе специальными опытно-методическими работами; H — целесообразная, по геологическим данным, глубина подсчета; q — продуктивность, выраженная в тоннах металла для слоя мощностью 1 м, $q = 2{,}5 \cdot p/100 = p/40$; p — площадная продуктивность, 2,5 — усредненная объемная масса горных пород; 100 — коэффициент для перехода от весовых процентов к тоннам металла.

В формуле прогнозных ресурсов площадная продуктивность p выступает в качестве переменной, что обусловливает $\{\hat{Q}_t\}$, $t = 1, 2, \ldots, n$ $(n \geqslant 2)$ определений прогнозных ресурсов:

$$\{\bar{Q}_t\}, \quad t = 1, 2, \ldots, n;$$

$$\bar{Q} = \frac{1}{n}\sum_{t=1}^{n}\hat{Q}_t; \qquad S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(\hat{Q}_t - \bar{Q})^2;$$

$$S = \sqrt{S^2}; \qquad \pm\lambda_{\bar{Q}} = t_{1-\alpha/2}\, S/\sqrt{n};$$

$$\pm\lambda^0_{\bar{Q}} = \left(\lambda_{\bar{Q}}/\bar{Q}\right) 100\ \%;$$

$$\left(\bar{Q} - \lambda_{\bar{Q}},\ \bar{Q} + \lambda_{\bar{Q}}\right);$$

$$p\left[\left(\bar{Q} - \lambda_{\bar{Q}}\right) \leqslant Q \leqslant \left(\bar{Q} + \lambda_{\bar{Q}}\right)\right] \simeq 1-\alpha.$$

4. Для запасов полезного компонента единичного блока при аппроксимации нормальной моделью [25].

Группируем N полных пересечений в блоке произвольным образом в m групп по n пересечений в каждой группе: $N \simeq m \cdot n$, (можно $N = m \cdot 1$, т. е. $m = N$) при соблюдении условия: $m \gg n$ [25].

Для каждой серии $i = 1, 2, \ldots, m$ в отдельности находят средневзвешенное содержание рудного компонента $\bar{C}_i$ и среднее арифметическое значение мощности рудного тела $\bar{M}_i$:

$$\bar{C}_i = \sum_{t=1}^{n} C_t M_t \Big/ \sum_{t=1}^{n} M_t; \qquad \bar{M}_i = \frac{1}{n}\sum_{t=1}^{n} M_t;$$
$$i = 1, 2, \ldots, m,$$

где M_t — полные пересечения (мощности рудного тела).

Каждое полное пересечение состоит из k $(k \geqslant 1)$ секционных проб с содержанием рудного компонента C_u и длиной секции l_u, $u = 1, 2, \ldots, k$. Поэтому средневзвешенное содержание рудного компонента по каждому пересечению C_t и мощность по пересечению M_t, $t = 1, 2, \ldots, N$ определяются как

$$C_t = \sum_{u=1}^{k} C_u l_u \Big/ \sum_{u=1}^{k} l_u;$$

$$M_t = \sum_{u=1}^{k} l_u; \qquad u = 1, 2, \ldots, K,$$

где l_u — длина секционной пробы.

Для каждой серии $i = 1, 2, \ldots, m$ в отдельности по общепринятой формуле подсчета запасов определяют оценки запасов рудного компонента $\hat{Q}_i$:

$$\hat{Q}_i = \frac{1}{100}\bar{C}_i\bar{M}_i\Pi d k_p, \qquad i = 1, 2, \ldots, m,$$

где Π — площадь блока в плоскости блокировки запасов, м^2; d — объемная масса руды, т/м^3; k_p — коэффициент рудоносности (часто принимаемый за единицу); $\frac{1}{100}$ — коэффициент для перехода от весовых процентов к тоннам металла.

В дальнейшем процедура построения интервальных оценок аналогична вышерассмотренному

$$\{\hat{Q}_i\}, \qquad i = 1, 2, \ldots, m;$$

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i; \qquad S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2;$$

$$S = \sqrt{S^2}; \qquad \pm\lambda_{\bar{Q}} = t_{1-\alpha/2}\frac{S}{\sqrt{m}};$$

$$\pm\lambda^0_{\bar{Q}} = \left(\lambda_{\bar{Q}}/\bar{Q}\right) 100\ \%;$$

$$\left(\bar{Q} - \lambda_{\bar{Q}},\ \bar{Q} + \lambda_{\bar{Q}}\right);$$

$$p\left[\left(\bar{Q} - \lambda_{\bar{Q}}\right) \leqslant Q \leqslant \left(\bar{Q} + \lambda_{\bar{Q}}\right)\right] \simeq 1-\alpha.$$

5. Для запасов полезного компонента единичного блока при аппроксимации логнормальной моделью, с учетом наличия ураганных сечений. Процедуры интервального оценивания аналогичны в этом случае пунктам 2, 3 и 11 [25].

6. Для суммарных запасов полезного компонента группы блоков, рудных тел, участков месторождения [25]. По каждому V-му блоку ($V = 1, 2, \ldots, k$) имеется N_V полных пересечений, в каждом из которых оценивается среднее взвешенное содержание рудного компонента C_t, $t = 1, 2, \ldots, N_V$, производятся замеры мощности m_t, на геолого-маркшейдерских планах замеряются площади блоков Π_V, $V = 1, 2, \ldots, k$, определяются объемные массы руды d_V, $V = 1, 2, \ldots, k$ по всем блокам, а также в случае необходимости коэффициенты рудоносности k_p^V, $V = 1, 2, \ldots, k$.

По каждому из k блоков в отдельности находят оценки запасов $\overline{Q}_V$ и их дисперсии $S^2_{\overline{Q}_V}$ (п. 4):

$$\overline{Q}_1\ \overline{Q}_2 \ldots \overline{Q}_V \ldots \overline{Q}_k;$$

$$S^2_{\overline{Q}_1}\ S^2_{\overline{Q}_2} \ldots S^2_{\overline{Q}_V} \ldots S^2_{\overline{Q}_k}.$$

Затем определяются: оценка суммарных запасов $\hat{Q}_\Sigma$, ее дисперсия S^2_Σ, стандартное отклонение S_Σ, точность суммарных запасов при заданной надежности $\pm L_{\hat{Q}_\Sigma}$:

$$\hat{Q}_\Sigma = \sum_{V=1}^{k} \overline{Q}_V; \qquad S^2_\Sigma = \sum_{V=1}^{k} S^2_{\overline{Q}_V}; \qquad S_\Sigma = \sqrt{S^2_\Sigma};$$

$$\pm L_{\hat{Q}_\Sigma} = t_{1-\alpha/2} S_\Sigma.$$

Строятся интервальные оценки суммарных запасов:

$$\left(\hat{Q}_\Sigma - L_{\hat{Q}_\Sigma},\ \hat{Q}_\Sigma + L_{\hat{Q}_\Sigma}\right);$$

$$p\left[\left(\hat{Q}_\Sigma - L_{\hat{Q}_\Sigma}\right) \leqslant Q_\Sigma \leqslant \left(\hat{Q}_\Sigma + L_{\hat{Q}_\Sigma}\right)\right] \simeq 1 - \alpha.$$

7. Для суммарных запасов полезного компонента по сумме промышленных категорий $A + B + C_1$ [25]. Предположим, что по категории A разведано k_1, по категории B—k_2, а по категории C_1—k_3 блоков. По каждому V-му блоку находят оценку запасов $\hat{Q}$ и оценку дисперсии $\hat{\sigma}^2$ (п. 4):

$\left\{\hat{Q}_{V1},\ \hat{\sigma}^2_{V1}\right\}$, $V = 1, 2, \ldots, k_1$ (категория A);

$\left\{\hat{Q}_{V2},\ \hat{\sigma}^2_{V2}\right\}$, $V = 1, 2, \ldots, k_2$ (категория B);

$\left\{\hat{Q}_{V3},\ \hat{\sigma}^2_{V3}\right\}$, $V = 1, 2, \ldots, k_3$ (категория C_1).

Находят оценки суммарных запасов и их дисперсий по категории $A\left(\overline{Q}_{\Sigma_1},\ \hat{\sigma}^2_1\right)$, по категории $B\left(\overline{Q}_{\Sigma_2},\ \hat{\sigma}^2_2\right)$ и по категории

$C_1\left(\overline{Q}_{\Sigma_3},\ \hat{\sigma}^2_3\right)$:

$$\overline{Q}_{\Sigma_1} = \sum_{V=1}^{k_1} \hat{Q}_{V1}; \qquad \hat{\sigma}^2_1 = k_1 \sum_{V=1}^{k_1} \hat{\sigma}^2_{V1}\eta_{V1}, \qquad \text{где } \eta_{V1} = \hat{Q}_{V1}/\overline{Q}_{\Sigma_1};$$

$$\sum_{V=1}^{k_1} \eta_{V1} = 1{,}0;$$

$$\overline{Q}_{\Sigma_2} = \sum_{V=1}^{k_2} \hat{Q}_{V2}; \qquad \hat{\sigma}^2_2 = k_2 \sum_{V=1}^{k_2} \hat{\sigma}^2_{V2}\eta_{V2},$$

$$\text{где } \eta_{V2} = \hat{Q}_{V2}/\overline{Q}_{\Sigma_2}; \qquad \sum_{V=1}^{k_2} \eta_{V2} = 1{,}0;$$

$$\overline{Q}_{\Sigma_3} = \sum_{V=1}^{k_3} \hat{Q}_{V3}; \qquad \hat{\sigma}^2_3 = k_3 \sum_{V=1}^{k_3} \hat{\sigma}^2_{V3}\eta_{V3}, \quad \text{где}$$

$$\eta_{V3} = \hat{Q}_{V3}/\overline{Q}_{\Sigma_3}; \qquad \sum_{V=1}^{k_3} \eta_{V3} = 1{,}0.$$

Строят интервальные оценки суммарных запасов для каждой промышленной категории в отдельности:

$$\left(\overline{Q}_{\Sigma_1} - L_{\Sigma_1},\ \overline{Q}_{\Sigma_1} + L_{\Sigma_1}\right), \qquad \text{где } \pm L_{\Sigma_1} = t_{1-\alpha/2}\overline{\sigma}_1;$$

$$\hat{\sigma}_1 = \sqrt{\hat{\sigma}^2_1};$$

$$\left(\overline{Q}_{\Sigma_2} - L_{\Sigma_2},\ \overline{Q}_{\Sigma_2} + L_{\Sigma_2}\right), \qquad \text{где } \pm L_{\Sigma_2} = t_{1-\alpha/2}\hat{\sigma}_2;$$

$$\hat{\sigma}_2 = \sqrt{\hat{\sigma}^2_2};$$

$$\left(\overline{Q}_{\Sigma_3} - L_{\Sigma_3},\ \overline{Q}_{\Sigma_3} + L_{\Sigma_3}\right), \qquad \text{где } \pm L_{\Sigma_3} = t_{1-\alpha/2}\hat{\sigma}_3;$$

$$\hat{\sigma}_3 = \sqrt{\hat{\sigma}^2_3}.$$

Находят оценку общих разведанных запасов по сумме категорий $(A + B + C_1)$ и ее дисперсию:

$$\overline{Q}_{\Sigma\Sigma} = \overline{Q}_{\Sigma_1} + \overline{Q}_{\Sigma_2} + \overline{Q}_{\Sigma_3};$$

$$\hat{\sigma}^2_{\Sigma\Sigma} = 3\left(\hat{\sigma}^2_1\beta_1 + \hat{\sigma}^2_2\beta_2 + \hat{\sigma}^2_3\beta_3\right), \quad \text{где } \beta_1 = \overline{Q}_{\Sigma_1}/\overline{Q}_{\Sigma\Sigma}; \quad \beta_2 = \overline{Q}_{\Sigma_2}/\overline{Q}_{\Sigma\Sigma};$$

$$\beta_3 = \overline{Q}_{\Sigma_3}/\overline{Q}_{\Sigma\Sigma}; \qquad \beta_1 + \beta_2 + \beta_3 = 1{,}0.$$

Строят доверительный интервал для разведанных запасов по сумме категорий ($A + B + C_1$) при надежности $1-\alpha$ (например, при 95 % надежности):

$$(\bar{Q}_{\Sigma\Sigma} - L_{\Sigma\Sigma},\ \bar{Q}_{\Sigma\Sigma} + L_{\Sigma\Sigma}), \quad \text{где } \pm L_{\Sigma\Sigma} = t_{1-\alpha/2}\hat{\sigma}_{\Sigma\Sigma};$$

$$\hat{\sigma}_{\Sigma\Sigma} = \sqrt{\hat{\sigma}^2_{\Sigma\Sigma}};$$

$$(\bar{Q}_{\Sigma\Sigma} - 1{,}96\hat{\sigma}_{\Sigma\Sigma},\ \bar{Q}_{\Sigma\Sigma} + 1{,}96\hat{\sigma}_{\Sigma\Sigma}).$$

Величина $(\bar{Q}_{\Sigma\Sigma} - L_{\Sigma\Sigma})$ представляет собой минимально гарантированные запасы по сумме категорий ($A + B + C_1$).

8. Для индикаторных отношений элементов, продуктивностей, показателей осевой, продольной, поперечной зональности геохимических аномалий, включая мультипликативные геохимические ореолы.

Индикаторные отношения рассмотрены на примере отношения двух элементов-индикаторов: свинца и цинка. В n точках конкретного эрозионного среза первичного геохимического ореола опробованы концентрации свинца x_t и цинка — y_t, $t = 1, 2, \ldots, n$.

В каждой точке в отдельности находят индикаторные отношения:

$$I_t = x_t/y_t, \qquad t = 1, 2, \ldots, n.$$

Процедуры интервального оценивания рассматриваемого индикаторного отношения аналогичны пунктам 1, 2, 3, 4, 5, 6. В частности, при согласованности индикаторных отношений с нормальной моделью имеем:

$$\{I_t = x_t/y_t\}, \qquad t = 1, 2, \ldots, n,$$

или

$$I_1, I_2, \ldots, I_t, \ldots, I_n;$$

$$\bar{I} = \frac{1}{n}\sum_{t=1}^{n} I_t; \qquad S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(I_t - \bar{I})^2;$$

$$S = \sqrt{S^2}; \qquad \pm\lambda_{\bar{I}} = t_{1-\alpha/2}\frac{S}{\sqrt{n}}; \qquad \pm\lambda^0_{\bar{I}} = \left(\lambda_{\bar{I}}/\bar{I}\right)100\,\%;$$

$$\left(\bar{I} - \lambda_{\bar{I}},\ \bar{I} + \lambda_{\bar{I}}\right);$$

$$P\left[\left(\bar{I} - \lambda_{\bar{I}}\right) \leqslant I \leqslant \left(I + \lambda_{\bar{I}}\right)\right] \simeq 1 - \alpha,$$

где I — неизвестное истинное значение индикаторного отношения свинца к цинку.

Интервальные оценки параметров в корреляционном, регрессионном, дисперсионном анализе приведены в соответствующих разделах.

ГЛАВА 5

КЛАССИФИКАЦИЯ И КЛАСТЕРНЫЙ АНАЛИЗ

При проведении прогнозно-металлогенических исследований, поисково-съемочных работ, при разведке и геолого-экономической оценке месторождений полезных ископаемых геолог постоянно сталкивается с необходимостью классифицировать геологические объекты и процессы.

При решении прогнозных задач геолог группирует изученные геологические объекты, а затем уточняет геологические свойства каждой полученной однородной классификационной группы. Если исследователь получает данные по новому геологическому объекту, то он должен отнести изучаемый объект к одной из априорно известных однородных классификационных единиц либо построить по имеющимся данным новую классификацию.

Максимальная типичность и максимальная аномальность — важные принципы прогнозирования при отсутствии информации по эталонным месторождениям и недостатке сведений о благоприятных признаках [9].

Классификация — один из фундаментальных процессов в науке. Факты и явления должны быть упорядочены, прежде чем мы сможем их понять и разработать общие принципы, объясняющие как их появление, так и наблюдаемый среди них порядок [24].

Классификация — это упорядочение объектов по их сходству.

Под термином «классификация» обычно понимается распределение предметов по заданным классам (понятие «класс» см. ниже) согласно наиболее существенным признакам, присущим предметам данного типа и отличающим их от предметов других типов [1, 2].

Составление классификаций подчиняется следующим правилам:

в одной классификации применяется одно и то же основание;

объем классифицируемого класса равняется сумме объемов подклассов;

классы и подклассы не пересекаются;

подразделение на подклассы производится непрерывно.

Геолог обычно решает одну из двух задач классификации: выявление естественного расслоения исходных геологических наблюдений и объектов на четко выраженные группы (кластеры, таксоны см. ниже), лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные друг от друга части; типизация, при которой совокупность данных и объектов разбивают на сравнительно небольшое число областей группирования так, чтобы элементы одной области лежали друг от друга по возможности на небольшом расстоянии. Задача типизации всегда имеет решение, а кластеризация не всегда, т. е. может существовать один единственный кластер [1, 2]. Имеются три основных типа данных, исполь-

зуемых в кластерном анализе: многомерные, данные о близости и данные о кластерах.

Классификацию геологических объектов можно производить с помощью набора числовых, качественных или классификационных признаков, используя формальные математические методы для разбивки на классы. Альтернативой к такому формализованному подходу является экспертный метод, при котором разбивка объектов на классы производится геологами-петрологами, тектонистами, геохимиками, геофизиками и другими на основании профессиональных знаний, опыта, интуиции.

Функция расстояния и мера сходства (понятия см. ниже) определяют понятие однородности объектов, которое в кластерном анализе является наименее формализованным. Выбор расстояния или коэффициента сходства является узловым моментом исследования, от которого решающим образом зависит окончательный вариант разбивки объектов на классы при заданном алгоритме разбивки [1].

СХЕМЫ КЛАССИФИКАЦИИ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

Целесообразно различать три аспекта процедуры применения кластерного анализа [1, 26]:

выбор функций расстояний d или мер сходства r между любыми парами многомерных геологических наблюдений;

выбор функций расстояний d или мер сходства r между любыми геологическими объектами, каждый из которых охарактеризован наборами многомерных геологических наблюдений;

выбор функций расстояний d или мер сходства r между любыми парами групп объектов, в том числе между объектом и группой объектов.

С. А. Айвазян и др. [1] подразделяют задачи кластерного анализа на два типа: 1) классификация сравнительно небольших по объему совокупностей многомерных наблюдений, когда их несколько десятков; 2) классификация больших массивов многомерных наблюдений, когда их сотни и тысячи. Задачи классификации делят также по типу априорной информации на три типа: а) число классов априорно задано; б) число классов неизвестно и его следует определить; в) число классов неизвестно, но его определение не входит в условие задачи [1].

Две последние ситуации приводят к построению иерархических деревьев — дендрограмм. Существуют два типа иерархических деревьев — агломеративное и дивизимное (см. ниже) (рис. 36).

Выделяют три основные кластерные процедуры: 1) иерархические агломеративные и дивизимные; 2) параллельные, реализуемые с помощью итерационных алгоритмов; 3) последовательные, реализуемые с помощью итерационных алгоритмов, причем на каждом шаге итерации привлекается небольшая часть наблюдений.

Трудно установить точные правила кластерного анализа, применяемые во всех ситуациях, и построить объективный критерий для сравнения кластеров, полученных с помощью различных процедур. Можно воспользоваться оценочными индексами, введенными Дж. Меззихом [24, с. 132]. Речь идет о величинах внешнего и внутреннего критериев значимости и мере воспроизводимости (см. ниже).

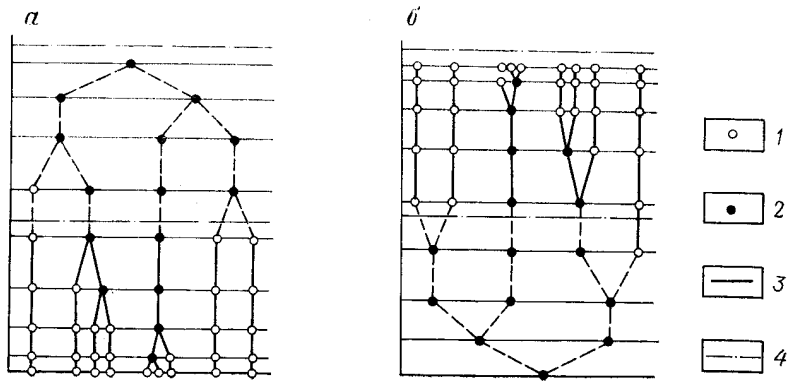


Рис. 36. Иерархическое дерево классификации объектов. На оси ординат откладывается абсолютное значение статистик:
а — агломеративное дерево; *б* — дивизимное дерево; *1* — объекты; *2* — однородные группы; *3* — этапы классификации; *4* — критические значения

Важный для прикладных геологических задач раздел кластерного анализа составляют статистические методы разграничения геологических объектов по комплексу признаков (см. гл. 2). Более подробно с процедурами кластерного анализа можно ознакомиться в работах [1, 18, 24].

Кластерный анализ — совокупность методов классификации и разбивка объектов и многомерных наблюдений на однородные группы [1].

Кластер (*cluster* — англ.) — скопление, пучок, группа элементов, характеризуемых каким-либо общим свойством; методы их нахождения — собственно кластерный анализ.

Таксон (*taxon* — англ.) — систематизированная группа любой категории; методы их нахождения — численная таксономия.

Метрическим пространством называется пара (X, d), состоящая из некоторого множества (пространства) X элементов x, $x \in X$ и расстояния d.

Функцией расстояния (метрикой) называется однозначная неотрицательная вещественная функция $d(x_u, x_S)$, определенная для любых $x_u \in X$ и $x_S \in X$, если соблюдаются следующие аксиомы:

1) $d(x_u, x_S) \geqslant 0$ для всех x_u и x_S из X;

2) аксиома максимальной близости объекта с самим собой: $d(x_u, x_S) = 0$ тогда и только тогда, когда $x_u = x_S$;

3) аксиома симметрии: $d(x_u, x_S) = d(x_S, x_u)$;

4) аксиома треугольника: $d(x_u, x_S) \leqslant d(x_u, x_Z) + d(x_Z, x_S)$.

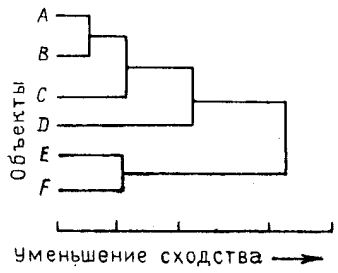


Рис. 37. Дендрограмма

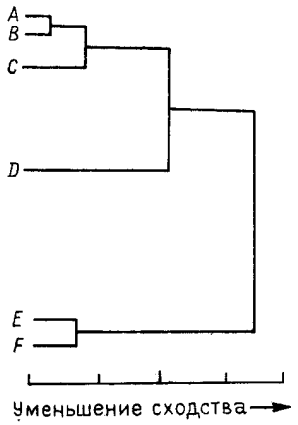


Рис. 38. Дендрограф

Мерой сходства (коэффициентом сходства, по В. Дюрану и П. Оделлу [17]) называется однозначная неотрицательная вещественная функция $r(x_u, x_S)$, определенная для любых x_u и x_S из X, $x_u \in X$, $x_S \in X$, если соблюдаются следующие аксиомы:

1) $0 \leqslant r(x_u, x_S) < 1$ для всех $x_u \neq x_S$;

2) аксиома максимального сходства объекта с самим собой: $r(x_u, x_S) = 1$ тогда и только тогда, когда $x_u = x_S$;

3) аксиома симметрии: $r(x_u, x_S) = r(x_S, x_u)$;

4) аксиома монотонного убывания коэффициентов сходства $r(x_u, x_S)$ по функции расстояния $d(x_u, x_S)$, т. е. из $d(x_Z, x_V) \geqslant d(x_u, x_S)$ должно следовать выполнение неравенства $r(x_Z, x_V) \leqslant r(x_u, x_S)$.

Термин «близость», по терминологии Р. Н. Шепарда [24], относится к сходству, различию, корреляции, мере пересечения или же к любой другой переменной, используемой в качестве меры сходства $r(x_u, x_S)$ или расстояния $d(x_u, x_S)$ между двумя объектами одного вида.

В иерархических схемах группировки объектов (кластерного анализа) наиболее распространенной формой графического изображения результатов является дендрограмма, а также ее двумерный аналог — дендрограф.

Дендрограмма представляет собой одномерный граф, напоминающий дерево, который используется для изображения взаимных связей между объектами из заданного их множества (рис. 37). Объекты располагаются по иерархическим уровням так, чтобы подчеркнуть их взаимное сходство на основе измеряемых свойств — геологических переменных. Объединение в группы объектов имеет смысл только в условиях высокой степени сходства. Компактное группирование свидетельствует о силе взаимных связей между объектами, а некомпактное — указывает на слабую зависимость. В дендрограммах объекты располагаются на равном расстоянии друг от друга, выбранном произвольно. Ветви дерева характеризуют иерархический порядок объектов, но при этом не отражаются иерархические зависимости между объектами. Их можно учесть, если расстояние между объектами сделать неравнозначным.

Дендрограф — это двумерная дендрограмма. Дендрограф описывает зависимости как внутри групп объектов, так и между группами (рис. 38). В результате имеем более наглядное графическое изображение связей между объектами.

Агломеративная кластерная процедура связана с вычислением функций расстояний и мер сходства между всеми парами объектов и объединением на каждом шагу той пары, для которой достигается минимум (максимум) функций расстояний и мер сходства. Кластеризация осуществляется путем объединений первоначально разобщенных k-объектов (см. рис. 36).

Дивизимная кластерная процедура связана с вычислением функций расстояний и мер сходства между всеми парами объектов и выделением на каждом шаге той пары (группы) объектов, для которой достигается их максимум (минимум). Кластеризация осуществляется путем разграничения первоначально единой группы, состоящей из k объектов.

Однородной будем называть такую совокупность, элементы которой формируются под воздействием общих основных причин и условий, а их законы распределения имеют простую структуру. Неоднородная совокупность такая, когда разные ее элементы формируются под влиянием разных причин и условий, либо если она может быть представлена в виде объединения некоторого числа однородных совокупностей с более простой структурой законов распределения элементов.

Оценочные индексы кластерного анализа, рассматриваемые Дж. Меззихом [24]:

а) величина внешнего критерия значимости — процент совпадения предсказаний экспертов и результатов процедуры кластерного анализа;

б) в качестве величины внутреннего критерия значимости предполагается кофенетический коэффициент корреляций (введен Р. Сокалом и Ф. Рольфом [24]);

в) мера воспроизводимости — в сущности специальный коэффициент корреляции.

Многомерное шкалирование — метод, полезный для анализа значений близости, чаще всего нижней или верхней полуматрицы близости. Многомерное шкалирование — это процедура описания матрицы близости в терминах расстояний между точками [24].

ТИПЫ РАССТОЯНИЙ И МЕРЫ СХОДСТВА

Коэффициенты сходства или различий между t-ми многомерными наблюдениями X_{ut} и X_{St}, по Р. Сокалу [24], подразделяются на три типа.

Первый тип — коэффициенты расстояния. Они имеют общий вид:

$$d_r(X_{ut},\ X_{St}) = \left(\frac{1}{m}\sum_{j=1}^{m} |x_{utj} - x_{Stj}|^r\right)^{1/r},$$
$$j = 1,\ 2,\ \ldots,\ m,$$

где u, S — индекс объектов; r — положительное целое число; m — число переменных.

Два случая особенно полезны: при $r = 1$ имеем дело с Манхэттенским расстоянием, при $r = 2$ получаем таксономическое расстояние (см. функцию расстояния).

Второй тип — коэффициенты ассоциативности. Они предназначены для оценивания сходства между парами многомерных наблюдений, описываемыми значениями признаков в виде двоичного кода (бинарными признаками). Общий вид коэффициентов ассоциативности представлен коэффициентом общего сходства Гауэра [24]:

$$r(X_{ut},\ X_{vt}) = \sum_{j=1}^{m} W_{uvj} S_{uvj} / \sum_{j=1}^{m} W_{uvj},$$
$$j = 1,\ 2,\ \ldots,\ m,$$

где $0 \leqslant S_{uvj} \leqslant 1$ — сходство между состояниями признака j для многомерных наблюдений X_{ut} и X_{vt}; W_{uvj} — вес, приписываемый этому признаку.

Третий тип — коэффициенты корреляции. Коэффициент корреляции есть отношение ковариации двух переменных к произведению их стандартных отклонений.

Расстояния и меры сходства между многомерными геологическими наблюдениями

Приведем наиболее употребительные функции для расстояний $d(X_{ut},\ X_{St})$ и коэффициентов сходства $r(X_{ut},\ X_{St})$ между парами многомерных наблюдений [1, 9, 18, 24, 26].

1. Обычное евклидово расстояние:

$$d(X_{ut},\ X_{St}) = \left[\sum_{j=1}^{m} (x_{utj} - x_{Stj})^2\right]^{1/2},$$

где x_{utj} (x_{Stj}) — значения j-го признака в t-й точке u (S)-го объекта.

2. Взвешенное евклидово расстояние [1].

3. Некоторые коэффициенты типа расстояния: L_1 — норма, супремум — норма, L_m — норма [18].

4. Обычное расстояние Махаланобиса:

$$d^2(X_{ut},\ X_{St}) = (X_{ut} - X_{St})' S^{-1} (X_{ut} - X_{St}),$$

где $(X_{ut} - X_{St})$ — разность векторов (m-мерных сравниваемых наблюдений); штрих означает операцию транспонирования разности векторов; символ — 1 — операция обращения матрицы S, т. е.

$$SS^{-1} = S^{-1}S = I,$$

где I — единичная матрица (размерностью $m \times m$); S — ковариационная матрица (размерностью $m \times m$) генеральной совокупности, из которой извлекаются многомерные наблюдения X_{ut} и X_{St}.

5. Взвешенное расстояние Махаланобиса [1].

6. Расстояние Минковского [24, 26].

7. Генетические расстояния Сангхви, Нея, расстояние Хеллингера для качественных данных [18, 24, 26].

8. Расстояние Минковского, классификационный индекс, квадратическая дифференциальная метрика Рао для количественных данных [18, 24, 26].

9. Хеммингово расстояние как мера различия наблюдений, задаваемых дихотомическими (0 и 1) признаками [9, 18, 24, 26]:

$$d(X_{ut},\ X_{St}) = \sum_{j=1}^{m} |x_{utj} - x_{Stj}| = n,$$

где m — число несовпадений значений соответствующих j-х признаков в сравниваемых наблюдениях X_{ut} и X_{St}.

10. Эвристические меры отдаленности, строго говоря, не являющиеся метриками (расстояниями d) из-за несоблюдения каких-либо вышеуказанных аксиом, определяющих функции расстояния d, но применяемые на практике. Среди таких мер следует упомянуть:

меру отдаленности Джеффриса—Матуситы [18],

коэффициент дивергенции [18],

информационный радиус Джардайна и Сибсона [18, 24].

11. Расстояния, задаваемые с помощью потенциальных функций [1, 2, 9].

12. Коэффициент сходства, основанный на взвешенном евклидовом расстоянии [1, 9, 24, 26].

13. Коэффициент подобия при разномасштабных признаках — как непрерывных, так и альтернативных (0 и 1) [9, 26].

14. Серия наиболее известных коэффициентов сходства для бинарных данных, т. е. булевых (0 и 1) векторов многомерных наблюдений [18, 24, 26]:

коэффициент сходства Сокала и Миченера,

коэффициент сходства Рао и Рассела,

коэффициент сходства Роджерса и Танимото,

коэффициент композиционного сходства (модификация коэффициента Роджерса и Танимото),

коэффициент сходства Джаккара, Танимото и Снита,

коэффициент сходства Дайса и Соренсона и др.

15. Аналогично функциям расстояний, задаваемых с помощью потенциальных функций, существуют разнообразные коэффициенты сходства на основе этих потенциальных функций [1, 9, 26].

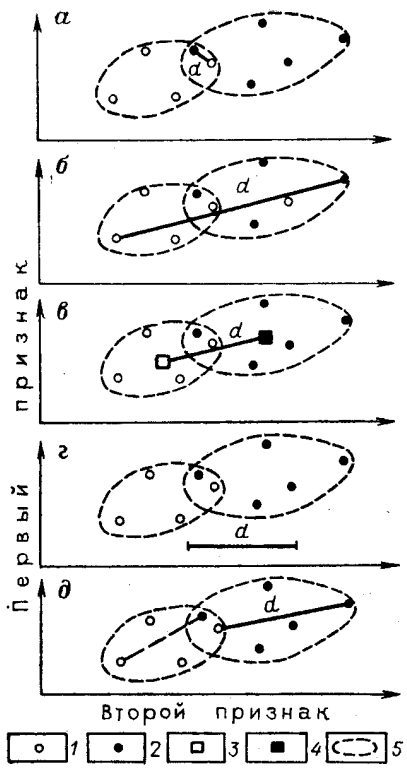


Рис. 39. Расстояния между парами объектов:

а — по принципу «ближайшего соседа»; *б* — по принципу «дальнего соседа»; *в* — между центрами тяжести; *г* — по принципу «средней связи»; *д* — хаусдорфово расстояние. *1* — наблюденные значения первого объекта; *2* — наблюденные значения второго объекта; *3* — «центр тяжести» наблюденных точек первого объекта; *4* — «центр тяжести» наблюденных точек второго объекта; *5* — выборочные совокупности

Расстояния и меры сходства между геологическими объектами

Приведем наиболее употребительные функции для расстояний $d(X_u, X_S)$ и коэффициентов сходства $r(X_u, X_S)$ между двумя объектами, т. е. матрицами наблюдений X_u и X_S (рис. 39) [1, 10, 17, 23, 25].

1. Минимальное локальное расстояние, измеряемое по принципу «ближайшего соседа» [1, 18, 24, 26]:

$$d(X_u, X_S) = \min d(X_{ut}, X_{St'}),$$
$$t = 1, 2, \ldots, n_u;$$
$$t' = 1, 2, \ldots, n_S,$$

где $d(X_{ut}, X_{St'})$ — расстояние типа Махаланобиса, евклидово расстояние, хеммингово расстояние и другие вышерассмотренные метрики. Геометрическое его представление для двумерных наблюдений ($m = 2$) приведено на рис. 39.

2. Максимальное локальное расстояние, измеряемое по принципу «дальнего соседа» (см. рис. 39) [1, 18, 24, 26]:

$$d(X_u, X_S) = \max d(X_{ut}, X_{St'}),$$
$$t = 1, 2, \ldots, n_u; \qquad t' = 1, 2, \ldots, n_S.$$

3. Расстояние центроидное (измеряемое по «центрам тяжести» кластеров): $d(X_u, X_S) = d(\overline{X}_{ut}, \overline{X}_{St'})$. Геометрическое представление центроидного расстояния см. на рис. 39.

4. Среднее расстояние (измеряемое по принципу «средней связи») [1, 18, 24, 26]:

$$d(X_u, X_S) = \frac{1}{n_u n_S} \sum_{t=1}^{n_u} \sum_{t'=1}^{n_S} d(X_{ut}, X_{St'}).$$

5. Расстояние медианное — разновидность расстояния, измеряемого по «центрам тяжести» (см. рис. 39) [1, 18, 24, 26]. При использовании медианного расстояния предполагается, что $n_u = n_S$.

6. Хаусдорфово расстояние — разновидность максиминной метрики, учитывающей минимальное и одновременно максимальное локальное расстояния [26]. Оно имеет вид

$$d(X_u, X_S) = \max \left\{ \max_t \min_{t'} d(X_{ut}, X_{St'}), \quad \max_{t'} \min_t d(X_{ut}, X_{St'}) \right\},$$
$$t = 1, 2, \ldots, n_u, \qquad t' = 1, 2, \ldots, n_S,$$

Хаусдорфово расстояние является максимальным значением из двух метрик, каждое из которых получено как максимальное из минимальных расстояний: минимальные расстояния $\min d$ определяются первоначально путем подстановки точки наблюдения S-го объекта (т. е. $X_S = \{X_{St'}\}$) и вычисления расстояния до точек наблюдения u-го объекта (т. е. $X_u = \{X_{ut}\}$), а затем, наоборот, между u-м и S-м объектами. Геометрическое представление хаусдорфова расстояния для двумерных наблюдений ($m = 2$) см. на рис. 39.

7. Мера близости (сходства), основанная на потенциальной функции.

8. Обобщенное по Колмогорову расстояние, основанное на понятии степенного среднего, включает многие из вышерассмотренных видов расстояний [1, 18, 26].

9. Расстояние Махаланобиса:

$$d^2(X_u, X_S) = \left(\overline{X}_{ut} - \overline{X}_{St'}\right)' S^{-1} \left(\overline{X}_{ut} - \overline{X}_{St'}\right),$$
$$t = 1, 2, \ldots, n_u; \qquad t' = 1, 2, \ldots, n_S,$$

где S — оценка обобщенной ковариционной матрицы размерности $m \times m$ по $(n_u + n_S)$ наблюдениям.

Рао [22, 23] предложил следующую формулу:

$$d^2(X_u, X_S) = \frac{1}{|R|} \sum_{i=1}^{m} \sum_{j=1}^{m} R_{ij} \left(\frac{\bar{x}_{ui} - \bar{x}_{si}}{\hat{\sigma}_i} \right) \left(\frac{\bar{x}_{uj} - \bar{x}_{sj}}{\hat{\sigma}_j} \right),$$

где предполагается, что $\sigma_{ui} = \sigma_{Si} = \sigma_i$ и $\sigma_{uj} = \sigma_{Sj} = \sigma_j$; $i, j = 1, 2, \ldots, m$; $|R|$ — определитель выборочной корреляционной матрицы $R = \{r_{ij}\}$, $i, j = 1, 2, \ldots, m$, r_{ij} — оценка парного коэффициента корреляции; R_{ij} — алгебраическое дополнение элемента, стоящего на пересечении i-й строки и j-го столбца корреляционной матрицы R.

10. Расстояние Свейна-Фу [18, 24].

11. Неиерархический метод средних [18, 24].

12. Неиерархический адаптивный метод «*Isodata*», являющийся своеобразным аналогом метода средних [18, 24].

Расстояния и меры сходства между группами геологических объектов

Приведем наиболее употребительные функции расстояния $d[X_{u1} \cup X_{u2} \cup \ldots (X_{S1} \cup X_{S2} \cup \ldots)$ и коэффициента сходства $r[X_{u1} \cup X_{u2} \cup \ldots (X_{S1} \cup X_{S2} \cup \ldots)$ между группами объектов,

т. е. между одной группой матриц $\{X_{u1}, X_{u2}, \ldots\}$ и другой $\{X_{S1}, X_{S2}, \ldots\}$ [1, 9, 18, 24, 26].

1. Большинство алгоритмов группирования (классифицирования) в иерархических процедурах можно получить из общей формулы Г. Ланса и В. Вильямса:

$$d[X_u(X_{v1} \cup X_{v2})] = \alpha d(X_u, X_{v1}) + \beta d(X_u, X_{v2}) + \gamma d(X_{v1}, X_{v2}) + \\ + \delta |d(X_u, X_{v1}) - d(X_u, X_{v2})|,$$

где α, β, γ, δ — числовые коэффициенты, значения которых определяют специфику кластерного анализа, $\cup$ — символ объединения (группа из двух матриц данных X_{v1} и X_{v2}). Задаваясь определенными значениями коэффициентов α, β, γ, δ, легко получить расстояние между группами объектов (или объектом и группой объектов) по принципу «ближайшего соседа», «дальнего соседа», средней группы, центроидной процедуры и т. п.

2. Обобщенное по Колмогорову расстояния для объединения групп объектов с использованием расстояния, основанного на понятии степенного среднего [1, 26].

3. Модификация Уишартом общей формулы Г. Ланса и В. Вильямса (метод Уорда) [1, 26].

4. Процедура Г. Ланса и В. Вильямса при ограничениях на параметры: $\alpha + \beta + \gamma = 1$, $\alpha = \beta = a$, $\gamma < 1$, $\delta = 0$ (наиболее целесообразно выбирать значения γ от — 0,25 до 0) [9, 26].

5. Процедура В. Н. Елкиной и Н. Г. Загоруйко [9, 26].

6. Весьма эффективная парагрупповая процедура Р. Мак-Кеммона и Г. Венингера представляет частный случай агломеративной иерархической процедуры кластерного анализа. В одну группу объединяются такие объекты X_u и X_V, на которых достигается минимум функционала

$$T_{uV} = (S_{uu} + S_{VV} - S_{uV}) / C^2_{n_u + n_V},$$

где S_{uu} и S_{VV} — величины, характеризующие расстояние внутри u-й и V-й групп; S_{uV} — величина, характеризующая межгрупповое расстояние; n_u и n_V — объемы групп; $C^2_{n_u+n_V}$ — число сочетаний из $(n_u + n_V)$ элементов по два. Парагрупповая процедура сохраняет пространство и монотонность меры сходства, а также учитывает внутригрупповое рассеяние.

7. Эффективная пороговая агломеративная иерархическая процедура кластерного анализа очень напоминает парагрупповой алгоритм Р. Мак-Кеммона и Г. Венингера, т. е. сохраняет все его достоинства [26]. Она включает задание расстояний типа Махаланобиса, Джеймса-Сю, т. е. статистик с известными типами распределений в условиях гипотез о близости объектов. Такие процедуры связаны с вычислением треугольных матриц примененных статистик между всеми парами объектов и объединением на каждом шаге той пары, где достигается минимум статистики, при условии, что эта величина меньше критического значения. Именно эта процедура рекомендуется для широкого применения при решении задач классификации геологических объектов [26].

ГЛАВА 6

ВЕРОЯТНОСТНОЕ РАСПОЗНАВАНИЕ ОБРАЗОВ И ДИСКРИМИНАНТНЫЙ АНАЛИЗ

ВЕРОЯТНОСТНЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Распознавание образов [42] — классификация некоторой группы объектов на основе заданных требований. Под образом понимается некоторая область, заданная в многомерном пространстве.

Требования, определяющие классификацию, могут быть различными, так как в различных ситуациях возникают свои типы классификаций, а именно: в зависимости от цели можно выбрать различные фиксированные множества признаков или всевозможные подмножества в этих множествах, в результате чего одна и та же пара объектов может быть отнесена как к одному и тому же, так и к различным образам.

В геометрической интерпретации под образом понимается область в m-мерном пространстве, вдоль координатных осей которого отложены значения признаков. Описание этой области называется эталоном, а отдельная точка в ней — реализацией образа.

Задачу распознавания образов можно понимать как сопоставление некоторой реализации, относительно которой неизвестно, к какому образу она относится, с эталонами. В этом случае реализация сравнивается с эталоном каждого образа и относится к тому или иному образу на основе заранее выбранного критерия соответствия или критерия подобия.

Постановку задачи распознавания образов будем называть детерминистской или вероятностной в зависимости от того, пересекаются образы между собой или нет.

Другими словами, ситуация будет детерминистской, если в любой точке выборочного пространства с ненулевой априорной вероятностью могут появляться реализации лишь одного образа, и вероятностной — в противном случае.

Необходимо особенно подчеркнуть, что отношение неизвестной (контрольной) реализации к тому или иному образу производится на основании априорной модели, при этом делаются предположения либо о характере распределения генеральной совокупности, либо о возможной структуре множества обучающих и контрольных реализаций, либо о типе допустимых правил принятия решений.

Процессу распознавания предшествует процесс обучения. Существует два различных метода обучения. Первый метод предполагает существование достаточно простых правил, настолько, что их можно четко описать, для того чтобы затем, сообразуясь с этими правилами, получать каждый раз требуемый результат.

Второй метод — метод показов предполагает, что учитель, сам верно классифицируя предъявленные объекты (реализации), не может сформулировать правило, по которому он действует.

Кроме обучения, иногда имеет место самообучение. Оно происходит в отсутствие учителя, когда не поступает информация о том, к какому образу относятся предъявляемые для самообучения эталонные реализации. В этом случае ученик сам определяет схожесть предъявленных реализаций.

Решающей функцией D называется некоторое правило (соответствие, функция, оператор, функционал и т. п.), которое относит каждую реализацию X к какому либо образу A.

В зависимости от того, какой из трех основных процедурных элементов X, D, A неизвестен, имеется три группы задач, связанных с распознаванием образов:

1) задан список образов A и указаны признаки, по которым эти образы следует отличать друг от друга. Требуется найти такое решающее правило D, чтобы распознавание произошло успешно;

2) задан список образов A и тип решающих правил D. Требуется выделить информативную комбинацию признаков, которая обеспечивала бы извлечение достаточного количества информации для распознавания;

3) задано множество реализаций или признаков и класс решающих функций D. Требуется разделить это множество на некоторое число (заданное или произвольное) однородных областей (классов) (задачи таксономии).

Введем следующие обозначения, принятые при описании различных методов распознавания образов:

$A_1, A_2, \ldots, A_S, \ldots, A_l$ — образы, l — их число, S, $r = 1, 2, \ldots, l$;

$x_1, x_2, \ldots, x_t, \ldots, x_m$ — признаки, m — их число, $i, j = 1, 2, \ldots, m$;

$$X_S = \{X_{S1}, \ldots, X_{Sn_s}\}$$

— множество эталонных реализаций S-го образа, $S = 1, 2, \ldots, l$;

x_{Sit} — значение i-го признака в t-й реализации для S-го образа;

x_{tu} — значение i-го признака в u-й реализации, $u = 1, 2, \ldots, N$;

X — реализация, подлежащая распознаванию.

Указанные три группы задач совпадают с основными группами задач, решаемых методами прикладной статистики и лишь слегка переформулированных. В прикладной статистике выделяют три основные задачи, для решения которых применим аппарат математической статистики: задачи классификации объектов, задачи выделения информативных комбинаций признаков и задачи оценивания зависимостей между случайными величинами. Таким образом, методы распознавания образов позволяют решать все основные задачи прикладной статистики, что свидетельствует о широкой применимости их при решении различных геологических задач, начиная от поисковой геологии вплоть до подсчета запасов месторождений полезных ископаемых. Методы распознавания образов

были эффективно применены для разделения нефтеносных и водоносных пластов по каротажным данным, определения нефтеносности структур по результатам химического анализа пластовых вод, уточнения связей геохимических показателей с нефтеносностью и битуминозностью, определения перспективных площадей и участков, прогнозной оценки геомагнитных аномалий, идентификации сейсмических волн, определения генетической принадлежности минералов, прогнозирования различных геологических характеристик. Ниже рассмотрены основные вероятностные методы распознавания образов, тогда как детерминированные методы («обобщенный портрет» Вапника и Червоненкиса, «Потенциальная функция» М. А. Айзермана, Э. М. Бравермана, Л. И. Розоноэра, «Кора-3» М. М. Бонгарда, «Тупиковые тесты и тесторы» А. М. Дмитриева, Ю. И. Журавлева, Ф. П. Кренделева и многие другие) в настоящем издании не приводятся. Подробнее с методами распознавания образов можно ознакомиться в работах [17, 38, 42, 47, 48]. Логические методы распознавания рассмотрены в гл. 16.

Правило Байеса. Рассмотрим случай многих образов и будем считать $x_1, x_2, \ldots, x_m$ случайными величинами, а именно результатами измерения признаков в условиях помех.

Пусть для каждого образа A_S, где $S = 1, 2, \ldots, l$, известна m-мерная функция плотности вероятности (или распределения) $p(X \mid A_S)$ вектора признаков X, т. е. функция плотности условной вероятности (или распределения) появления в X точек из A_S и вероятность $p(A_S)$ появления образа A_S, где $S = 1, 2, \ldots, l$. Тогда задача распознавания образов может быть сформулирована как определение решающей функции $D = D(X)$, где $D(X) = D_S$ означает, что принимается гипотеза $H_S : X \in A_S$.

Потери, когда принято решение D_S, т. е. $X \in A_S$ (хотя в действительности $X \in A_r$), обозначим $L(A_r, A_S)$. Тогда условные потери (или условный риск) для $X \in A_S$ равны: $r(A_S, D) = \int L(A_S, D)\, p(X/A_S)\, dx$, и для множества $p = \{p(A_S)\}$, $S = 1, 2, \ldots, l$, средние потери (средний риск) равны

$$Rp, D = \sum_{S=1}^{l} p(A_S)\, r(A_S, D) = \int p(x)\, r_X(p, D)\, dx,$$

где

$$r_X(p, D) = \left[\sum_{S=1}^{l} L(A_S, D)\, p(A_S)\, p(X/A_S)\right] / p(X)$$

— апостериорный условный средний риск решения D при фиксированном x.

Задача заключается в выборе такого решения D_S ($S = 1, 2, \ldots, l$), которое минимизирует средний риск $R(p, D)$ или максимум условного риска $r(A_S, D)$.

Оптимальное решающее правило минимизации среднего риска называется правилом Байеса.

Пусть D^* — оптимальное решение в смысле минимума среднего риска; тогда

$$r_X(p,\ D^*) \leqslant r_X(p,\ D),$$

т. е.

$$\sum_{S=1}^{l} L(A_S,\ D^*)\, p(A_S)\, p(X/A_S) \leqslant \sum_{S=1}^{l} L(A_S,\ D)\, p(A_S)\, p(X/A_S).$$

Пусть функция потерь

$$L(A_S,\ D_r) = \begin{cases} 0, & \text{если } S = r, \\ 1, & \text{если } S \neq r. \end{cases}$$

Тогда $D^* = D_S$, если

$$p(A_S)\, p(X/A_S) \geqslant p(A_r)\, p(X/A_r),$$

для всех $r = 1, 2, \ldots, l$.

Пусть λ — отношение правдоподобия для образов A_r и A_S:

$$\lambda = p(X/A_S)/p(X/A_r).$$

Тогда $D^* = D_s$, если

$$\lambda \geqslant p(A_r)/p(A_S),$$

для всех $r = 1, 2, \ldots, l$.

Разделяющей границей между A_S и A_r будет:

$$p(A_S)\, p(X/A_S) - p(A_r)\, p(X/A_r) = 0,$$

или

$$\log \frac{p(A_S)\, p(X/A_S)}{p(A_r)\, p(X/A_r)} = 0.$$

Пусть $p(X/A_S)$ — функция плотности многомерного нормального распределения с вектором средних M_S и ковариационной матрицей K_S:

$$p(X/A_S) = \frac{\exp\left[-\frac{1}{2}(X - M_S)^T K_S^{-1} (X - M_S)\right]}{(2\pi)^{m/2} |K_S|^{1/2}}.$$

Тогда разделяющей будет:

$$\log \frac{p(A_S)}{p(A_r)} + \log \frac{p(X/A_S)}{p(X/A_r)} = \log \frac{p(A_S)}{p(A_r)} - \frac{1}{2} \log \frac{|K_S|}{|K_r|} -$$

$$- \frac{1}{2}\left[(X - M_S)^T K_S^{-1} (X - M_S) - (X - M_r)^T K_r^{-1} (X - M_r)\right] = 0.$$

В случае равенства матриц ковариации $K_r = K_S = K$ разделяющая граница принимает вид гиперплоскости:

$$\log \frac{p(A_S)}{p(A_r)} + X^T K^{-1} (M_S - M_r) -$$

$$- \frac{1}{2}(M_S + M_r)^T K^{-1} (M_S - M_r) = 0.$$

Модификациями метода распознавания на основе правила Байеса являются следующие:

1) метод заданного превышения максимальной вероятности гипотезы по отношению к ближайшей к ней. $D^* = D_S$, если

$$p(A_S)\, p(X/A_S) \geqslant C p(A_r)\, p(X/A_r),$$

где $p(A_S)\, p(X/A_S)$ — максимальное значение для $p(A_r)\, p(X/A_r)$ по всем r; $p(A_r)\, p(X/A_r)$ — ближайшее к максимальному значению того же выражения; C — константа, зависящая от требуемой надежности распознавания;

2) метод превышения максимальной вероятности гипотезы над суммарной вероятностью всех остальных гипотез. $D^* = D_S$, если

$$p(A_S)\, p(X/A_S) \geqslant \sum_{r \neq S} p(A_r)\, p(X/A_r).$$

Критерий Вальда. Задача распознавания может быть решена не полным набором признаков, а некоторым его подмножеством. Рациональное соотношение между ложным распознаванием и числом использованных при этом признаков дает возможность добавлять признаки последовательно до тех пор, пока не будет достигнута требуемая точность распознавания.

В подобной процедуре становится существенным, в какой последовательности добавляются признаки. Понятно, что признаки следует расположить в такой последовательности, чтобы получить решение как можно скорее. Однако задача подобного упорядочивания признаков является самостоятельной задачей.

Пусть задано два класса A_1 и A_2. Рассмотрим последовательный процесс добавления признаков. На i-м шаге процесса, т. е. после измерения i-го признака, вычисляется последовательное отношение правдоподобия:

$$\lambda_i = \frac{p_i(X/A_1)}{p_i(X/A_2)},$$

где $p_i(X/A_S)$ — i-мерная функция условной плотности вероятности для образа A_S ($S = 1, 2$). После этого λ_i сравнивается с двумя останавливающими границами (порогами) B_1 и B_2, и если $\lambda_i \geqslant B_1$, то принимается решение $X \in A_1$, если $\lambda_i \leqslant B_2$, то принимается $X \in A_2$. В случае, если $B_2 < \lambda_i < B_1$, то добавляется следующий признак и производится $(i + 1)$-й шаг. В качестве B_1 и B_2 выбраны выражения

$$B_1 = \frac{1 - l_{21}}{l_{12}}, \qquad B_2 = \frac{l_{21}}{1 - l_{12}},$$

где l_{Sr} — вероятность принятия гипотезы $X \in A_S$, когда в действительности истинна гипотеза $X \in A_r$; $S,\ r = 1,\ 2$.

Можно доказать, что при заданных l_{12} и l_{21} не существует другой процедуры, которая обладала бы меньшими значениями вероятностей ошибок или среднего числа и давала бы выигрыш в среднем числе признаков по сравнению с последовательной процедурой классификации.

Пусть $x_1, x_2, \ldots$ — независимые измерения признаков реализации X с одномерной нормальной функцией плотности $p(x_j/A_j)$, $s=1, 2; j=1, 2, \ldots$, со средним значением m_1 и дисперсией σ^2. Тогда на первом шаге выбрано x_1 и вычисляется

$$\log \lambda_1 = \log \frac{p(x_1/A_1)}{p(x_1/A_2)} = \frac{1}{\sigma^2}\left[(m_1 - m_2)x_1 - \frac{1}{2}(m_1^2 - m_2^2)\right].$$

Значение $\log \lambda_1$ сравнивается с $\log B_1$ и $\log B_2$.

Если $$x_1 \geqslant \frac{\sigma^2}{m_1 - m_2}\log B_1 + \frac{1}{2}(m_1 + m_2),$$

то $X \in A_1$,

если

$$x_1 \leqslant \frac{\sigma^2}{m_1 - m_2}\log B_2 + \frac{1}{2}(m_1 + m_2),$$

то $X \in A_2$,

и если $$\left[\frac{\sigma^2}{m_1 - m_2}\log B_2 + \frac{1}{2}(m_1 + m_2)\right] < x_1 < \left[\frac{\sigma^2}{m_1 - m_2}\log B_1 + \frac{1}{2}(m_1 + m_2)\right],$$

то добавляется x_2 и т. д. На i-м шаге

$$\log \lambda_i = \sum_{j=1}^{i} \log \frac{p(x_j/A_1)}{p(x_j/A_2)} = \frac{m_1 - m_2}{\sigma^2}\sum_{j=1}^{i}\left[x_j - \frac{1}{2}(m_1 + m_2)\right].$$

Если

$$\sum_{j=1}^{i} x_j \geqslant \frac{\sigma^2}{m_1 - m_2}\log B_1 + \frac{i}{2}(m_1 + m_2),$$

то $X \in A_1$,

если

$$\sum_{j=1}^{i} x_j \leqslant \frac{\sigma^2}{m_1 - m_2}\log B_2 + \frac{i}{2}(m_1 + m_2),$$

то $X \in A_2$,

и если

$$\left[\frac{\sigma^2}{m_1 - m_2}\log B_2 + \frac{i}{2}(m_1 + m_2)\right] < \sum_{j=1}^{i} x_j < \left[\frac{\sigma^2}{m_1 - m_2}\log B_1 + \frac{i}{2}(m_1 + m_2)\right],$$

то берется x_{i+1} и т. д.

Обобщенный критерий Вальда. Применяется для случая, когда число образов превышает два.

В этом случае на каждом i-м шаге для каждого S-го образа вычисляется обобщенное последовательное отношение правдоподобия:

$$u_i(X/A_S) = \frac{p_i(X/A_S)}{\left[\prod_{r=1}^{l} p_i(X/A_r)\right]^{1/l}}, \qquad S = 1, 2, \ldots, l.$$

Затем $u_i(X/A_S)$ сравнивается с останавливающей границей для S-го образа $B(A_S)$, и если

$$u_i(X/A_S) < B(A_S), \quad S = 1, 2, \ldots, l,$$

то образ A_S из дальнейших операций исключается. После исключения S-го образа составляется новый набор последовательных отношений вероятностей.

Так продолжается до тех пор, пока не останется единственный образ, с которым и отождествляется X. В качестве останавливающих границ принимается выражение

$$B(A_S) = \frac{1 - l_{SS}}{\left[\prod_{r=1}^{l}(1 - l_{Sr})\right]^{1/l}}, \qquad S = 1, 2, \ldots, l.$$

Для случая двух образов обобщенный критерий Вальда эквивалентен последовательному критерию отношения правдоподобия Вальда и поэтому оптимален. Сохраняется ли оптимальность при $l > 2$, не доказано.

Усеченный критерий Вальда. При реализации последовательного критерия отношения правдоподобия или обобщенного критерия Вальда возможны два нежелательных случая:

1) алгоритмы могут потребовать слишком большого числа признаков;

2) среднее число признаков может стать очень большим, если величины l_{Sr} выбраны слишком малыми.

В этих случаях следует прервать последовательную процедуру на m-м шаге, и если до этих пор решение не было получено, то принимают решение $X' \in A_1$, если $\lambda_m \leqslant 1$, или решение $X' \in A_2$, если $\lambda_m > 1$.

Для обобщенного последовательного критерия отношения правдоподобия Вальда процедура усечения проводится аналогично.

Модифицированный критерий Вальда. В рассмотренных выше критериях Вальда и обобщенном последовательном критерии отношения правдоподобия вероятности появления ошибок l_{Sr} задавались заранее. При этом число признаков, необходимое для принятия решения, является случайной величиной, зависящей от l_{Sr}, и может принимать любое значение. Вместе с тем желательно использовать алгоритм, который за конечное и заданное исследователем число шагов получит окончательное решение. Такими алгоритмами являются рассмотренные алгоритмы с усечением. К сожале-

нию, алгоритмы с усечением являются алгоритмами с принудительной остановкой процесса, когда переход от продолжения к окончанию происходит скачкообразно. Чтобы сгладить скачок, может быть применена процедура с меняющимися от шага к шагу останавливающими границами.

Выберем невозрастающую функцию $g_1(x)$ и неубывающую функцию $g_2(x)$.

Метод заключается в последовательной проверке неравенства

$$e^{g_2(i)} < \lambda_i < e^{g_1(i)}, \qquad i = 1, 2, \ldots .$$

Если $\lambda_i \geqslant e^{g_1(i)}$, то принимается решение $X \in A_1$, если $\lambda_i \leqslant e^{g_2(i)}$, то $X \in A_2$.

Положим:

$$g_1(i) = a\left(1 - \frac{i}{m}\right)^{r_1}; \qquad g_2(i) = -b\left(1 - \frac{i}{m}\right)^{r_2},$$

где $0 < r_1,\ r_2 \leqslant 1,\ a > 0,\ b > 0$.

В этом случае процесс закончится после m-го шага, так как в этом случае $g_1(m) = g_2(m) = 0$, и неравенство $e^{g_2(i)} < e^{g_1(i)}$, являющееся необходимым условием продолжения процесса, не выполняется.

Пусть $x_1, x_2, \ldots$ — независимые признаки с одномерной нормальной функцией плотности $p(x_i/A_S)$, $S = 1, 2$; $i = 1, 2, \ldots$, со средними значениями m_i и дисперсией σ^2.

Тогда, вычислив

$$\log \lambda_i = \sum_{j=1}^{i} \log \frac{p(x_j/A_1)}{p(x_j/A_2)} = \frac{m_1 - m_2}{\sigma^2} \sum_{j=1}^{i} \left[x_j - \frac{1}{2}(m_1 + m_2)\right],$$

получим следующую процедуру.

Если $\sum_{j=1}^{i} x_j \geqslant \frac{\sigma^2}{m_1 - m_2} g_1(i) + \frac{i}{2}(m_1 + m_2)$, то $X \in A_1$,

если $\sum_{j=1}^{i} x_j \leqslant \frac{\sigma^2}{m_1 - m_2} g_2(i) + \frac{i}{2}(m_1 + m_2)$, то $X \in A_2$;

в противном случае, т. е. если

$$\left[\frac{\sigma^2}{m_1 - m_2} g_2(i) + \frac{i}{2}(m_1 + m_2)\right] < \sum_{j=1}^{i} x_j <$$

$$< \left[\frac{\sigma^2}{m_1 - m_2} g_1(i) + \frac{i}{2}(m_1 + m_2)\right],$$

добавляют $(i+1)$-й признак и делают $(i+1)$-й шаг вычислений. Если границы $g_1(i)$ и $g_2(i)$ заданы так же, как и раньше, то процесс закончится не позднее m-го шага.

Модифицированный обобщенный критерий Вальда. Как и раньше, для каждого S-го образа на i-м шаге вычисляется обобщенное последовательное отношение вероятностей:

$$u_i(X/A_S) = \frac{p_i(X/A_S)}{\left[\prod_{r=1}^{l} p_i(X/A_r)\right]^{1/l}}$$

и сравнивается с останавливающими границами $g_S(i)$, где индекс S означает принадлежность к S-му образу.

Если $u_i(X/A_S) < g_S(i)$, то образ A_S отбрасывается и в дальнейших вычислениях не участвует.

Такое отбрасывание производится до тех пор, пока не останется только один образ. Тогда к этому образу и относят исследуемую реализацию x.

В качестве порогов берут

$$g_S(i) = C\left(1 - \frac{i}{m}\right)^{r_S}, \quad S = 1, 2, \ldots, m; \quad C > 0,\ 0 < r_S \leqslant 1.$$

В описанных трех модифицированных алгоритмах можно выразить вероятности ошибок l_{ij} и среднее число измерений признаков, необходимое для принятия решения (это число равно числу шагов), через известные величины m, l_{Sr}, a, b, c, r_1, r_2.

Оптимальность выбора параметров a, b, c, r_1, r_2, r_s не исследовалась.

Последовательные ранги. Последовательным рангом значения признака x_i реализации $X = (x_1, x_2, \ldots, x_i, \ldots, x_m)$ называется число, равное S_i, если x_i является S_i-й наименьшей величиной в этом множестве, $i = 1, 2, \ldots, m$.

Например, последовательный ранг x_1 всегда равен 1, последовательный ранг x_2 равен 1 или 2, в зависимости от того, $x_2 < x_1$ или $x_2 > x_1$ и т. д.

Каждой реализации x будет соответствовать вектор последовательных рангов $S(m) = (S_1, S_2, \ldots, S_m)$.

Поскольку между порядком расположения признаков и векторами последовательных рангов существует взаимнооднозначное соответствие, то в качестве функции распределения последних может быть взята следующая функция:

$$P(x_{g_1} \leqslant x_{g_2} \leqslant \ldots \leqslant x_{g_m}) = \int \ldots \int \prod_{i=1}^{m} dP_{g_i}(x_{g_i}),$$

$$-\infty < x_{g_i} \leqslant \ldots \leqslant x_{g_m} < \infty,$$

где $x_{g_1}, x_{g_2}, \ldots, x_{g_m}$ — упорядоченные в неубывающем порядке значения признаков в реализации X.

Предполагается, что все эти значения независимы.

Пусть имеем две реализации $X_1 = (x_1^1, x_2^1, \ldots, x_m^1)$ и $X = (x_1, x_2, \ldots, x_m)$, причем относительно первой известно, что она принадлежит образу A_1. Тогда проверяем гипотезу H_0 о том, что

обе реализации имеют одинаковые распределения (т. е. X также относится к A_1) против альтернативной гипотезы H_1 о различии этих распределений. При этом предполагается, что $x_1^1, x_2^1, \ldots, x_m^1$ и $x_1, x_2, \ldots, x_m$ — независимые случайные величины с функциями распределения соответственно $P(X_1)$ и $f[P(X_1)]$.

Символически такую проверку гипотез можно записать так: проверяемая гипотеза $H_0 : P = P(X_1)$ при альтернативе $H_1 : P = f[P(X_1)]$.

Образуем из компонентов X_1 и X последовательность $x_1^1, x_1, x_2^1, x_2, \ldots, x_i^1, x_i, \ldots, x_m^1, x_m$ и обозначим $V_1 = x_1^1$, $V_2 = x_1, \ldots$ и т. д. Тогда получим $V(k) = (V_1, V_2, \ldots, V_k)$, $k = 1, 2, \ldots, 2m$, где k — номер шага процесса. И пусть $S(k)$ — вектор последовательных рангов для $V(k)$, а

$$\lambda_k = \frac{P_k[S(k)/H_1]}{P_k[S(k)/H_0]}$$

— последовательное отношение вероятностей (на каждом шаге).

Если H_0 истинна, то для любого S из $S(k)$

$$P_k[S(k) = S/H_0] = \frac{1}{k}.$$

Таким образом, найден знаменатель в выражении для λ_k, Теперь для нахождения числителя в том же выражении достаточно заметить, что

$$P_k[S(k) = S/H_1] = p(V_1 \leqslant V_2 \leqslant \ldots \leqslant V_k/H_1)$$

и

$$P(V_1 \leqslant V_2 \leqslant \ldots \leqslant V_k/H_1) = \int \ldots \int \prod_{u=1}^{k} df_u[P(V_u)],$$

где $$f_u[P(V_u)] = \begin{cases} P(V_u), & \text{если } V_u \text{ взята из } X_1, \\ f(V_u), & \text{если } V_u \text{ взята из } X. \end{cases}$$

Полученные на каждом шаге значения λ_k сравниваются с двумя останавливающими границами и в случае выхода из этих границ принимается соответствующая гипотеза H_0 или H_1. Если λ_k не выходит за границы на данном k-м шаге, то процесс продолжается, k увеличивается на единицу и вычисляется λ_{k+1}, и т. д.

Описанный метод применим, когда в эталоне имеются представители только одного образа, и распознавание заключается в том, чтобы вынести решение, относится ли предъявляемая для распознавания реализация к этому образу или нет.

Вопрос выбора вида функции f в альтернативе может быть решен различным способом, одним из которых являются альтернативы Лемана.

Альтернативы Лемана имеют вид $f[P(X)] = P^r(X)$, $r > 0$.

В случае допустимости альтернатив Лемана для последовательного отношения правдоподобия получим

$$\lambda_k = \frac{P_k[S(k)/H_1]}{P_k[S(k)/H_0]} = \frac{k!\, r^{\frac{k-\delta}{2}}}{\prod_{h=1}^{k}\left(\sum_{u=1}^{h} C_u\right)},$$

где $\delta = \begin{cases} 0, & \text{для четных } k, \\ 1, & \text{для нечетных } k, \end{cases}$ $\quad C_u = \begin{cases} 1, & \text{если } V_u \text{ из } X_1, \\ 0, & \text{если } V_u \text{ из } X. \end{cases}$

λ_k сравнивается с парой останавливающих границ, и как только происходит их пересечение, процесс останавливается и принимается соответствующее решение.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Почти в любом геологическом исследовании приходится сталкиваться с ситуацией, когда нужно сделать вывод о принадлежности одного объекта или их набора к одной из нескольких заранее заданных групп. Иногда эти выводы делаются по одному, а иногда по комплексу признаков, причем до последнего десятилетия все эти выводы были интуитивными, что приводило к существенному влиянию субъективных факторов.

Однако начиная с середины шестидесятых годов в геологии для решения упомянутой задачи различные исследователи начинают применять математические методы. Примерами могут служить работы Ш. А. Губермана и др., А. Н. Бугайца [9], А. А. Дорофеюка, Ю. А. Воронина и др. [39].

Решение задач классификационного отнесения изучаемых объектов к одной из заданных групп по комплексу признаков называется дискриминантным анализом. Методы дискриминантного анализа требуют только количественных данных, что несколько сужает области их применения. Однако в основе этого анализа лежит хорошо развитая математическая теория, что позволяет учитывать риск, связанный с принятием ошибочных решений.

Формально задача дискриминантного анализа сводится к следующему. Пусть $A_1, A_2, \ldots, A_e, \ldots, A_k$ — k множеств объектов. Не нарушая общности, мы ограничимся для простоты рассмотрением только двух множеств A_1 и A_2. Каждому из множеств A_1 и A_2 поставим в соответствие m-мерную случайную величину:

$$\Xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\};$$
$$[\mathrm{H}] = \{\eta_1, \eta_2, \ldots, \eta_j, \ldots, \eta_m\},$$

причем известно, что некоторые параметры θ_1 и θ_2, присущие Ξ и H, различны, т. е. $\theta_1 \neq \theta_2$. В большинстве случаев под θ_1 и θ_2 понимаются многомерные средние, но не исключено и рассмотрение ковариационных матриц или же многомерных средних и ковариационных матриц совместно.

Допустим, что из каждой совокупности A_1 и A_2 объектов a взята выборка объемом соответственно n_1 и n_2 и по выборочным данным $X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}$ и $Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}$ требуется построить решающее правило D, которое бы позволяло относить объекты из третьей совокупности A, представляющей собой смесь объектов из A_1 и A_2, к A_1 и к A_2. Обозначим результат m-мерного наблюдения из совокупности A, которая требует распознавания, через $z_t = \{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$. Таким образом, наше решающее правило должно заключаться в том, что рассматриваемое наблюдение z_t относится к совокупности A_1 (например, к перспективным в отношении рудоносности образованиям), если оно характеризуется определенным множеством значений $\{z_1, z_2, \ldots, z_j, \ldots, z_m\}$, а к совокупности A_2 (например, к бесперспективным в отношении рудоносности объектам) при других значениях $\{z_1, z_2, \ldots, z_j, \ldots, z_m\}$. Такое условие приводит к тому, что все m-мерное пространство, в котором результаты наблюдений представлены m-мерными точками, будет разделено на две области R_1 и R_2, причем если результат наблюдения попадет в R_1, мы принимаем решение о его принадлежности к группе A_1, а если он попадет в R_2, то мы относим его к совокупности A_2.

Естественно, что оба решения не исключают возможности появления ошибок, которые заключаются в следующем. Решение о принадлежности классифицируемого объекта $a \in A$ к A_1, т. е. $a \in A_1$, ошибочно и он в действительности принадлежит к A_2, т. е. $a \in A_2$. Вторая возможная ошибка заключается в том, что принимается решение $a \in A_2$, тогда как в действительности $a \in A_1$.

Каждой из этих ошибок можно приписать соответствующую цену, так как нередко их появление приводит к тем или иным потерям. Например, ошибочное отнесение объекта к перспективно рудоносным, тогда как в действительности он бесперспективен, приведет к потерям, связанным с безрезультативным проведением поисковых работ на этом объекте. Наоборот, ошибочное отнесение перспективного объекта к бесперспективным приведет к потере месторождения, которое стоит, как правило, дороже, чем затраты на поисковые работы. Обозначим стоимости этих потерь соответственно через $C(A_1/A_2)$ и $C(A_2/A_1)$. Ошибки и их стоимости сведены воедино в табл. 4.

Допустим, что в выборке, которую нужно подвергнуть разделению на объекты, принадлежащие совокупностям A_1 и A_2, эти объекты смешаны в определенных соотношениях, и доля объектов $a \in A_1$ равна q_1, а объектов $a \in A_2$ — q_2 и $q_1 + q_2 = 1$. Тогда величину q_1 можно рассматривать как вероятность события, заключающегося в том, что взятый наудачу из изучаемой смешанной совокупности объект будет принадлежать к A_1. Аналогично интерпретируется и вероятность q_2.

Кроме того, не нарушая общности, будем считать, что вероятностные свойства совокупностей A_1 и A_2 описываются плотностями вероятности:

$$f_1(X) = f_1(x_1, x_2, \ldots, x_j, \ldots, x_m),\ f_2 = (y_1, \ldots, y_m).$$

Таблица 4

Ошибки и их стоимости

Принимаемое решение	Действительное состояние	
	$a \in A_1$	$a \in A_2$
$a \in A_1$	Правильное решение $C(A_1/A_1) = 0$	$a \in A_1/a \in A_2$, $C(A_1/A_2) > 0$
$a \in A_2$	$a \in A_2/a \in A_1$, $C(A_2/A_1) > 0$	Правильное решение $C(A_2/A_2) = 0$

Таким образом, если, как мы это сделали выше, область R значений X и Y разделена на две непересекающиеся области R_1 и R_2, то вероятности появления ошибочных решений будут определены следующим образом:

$$P(A_1/A_2) = \int_{R_1} f_2(X)\,dx; \tag{6.1}$$

$$P(A_2/A_1) = \int_{R_2} f_1(X)\,dx. \tag{6.2}$$

Теперь мы можем охарактеризовать и потери, связанные с неправильной классификацией, которые, если известны значения q_1 и q_2, определяются следующей формулой:

$$W = C(A_1/A_2)\,P(A_1/A_2)\,q_2 + C(A_2/A_1)\,P(A_2/A_1)\,q_1. \tag{6.3}$$

Это выражение представляет собой математическое ожидание потерь классификации или, как его еще называют, средние потери. Таким образом, области принятия решений R_1 и R_2 нужно выбрать так, чтобы потери W были по возможности меньшими. Метод, который обеспечивает минимум W при заданных q_1 и q_2, называется методом Байеса. Подробно вопрос о риске, связанном с принятием решений в задачах классификации, рассмотрен в работах Т. Андерсона, а также Д. Блэкуела и М. Гиршика. Здесь же мы ограничимся только общим правилом, заключающимся в том, что R_1 и R_2 выбираются следующим образом:

$$R_1 : \frac{f_1(X)}{f_2(X)} \geqslant \frac{C(A_1/A_2)q_2}{C(A_2/A_1)\,q_1}; \tag{6.4}$$

$$R_2 : \frac{f_1(X)}{f_2(X)} < \frac{C(A_1/A_2)\,q_2}{C(A_2/A_1)\,q_1}. \tag{6.5}$$

Построение решающих правил для двух многомерных совокупностей. *Случай, когда $f_1(X)$ и $f_2(X)$ известны.*

Предположим, что $f_1(X)$ и $f_2(X)$ являются m-мерными нормальными плотностями с параметрами соответственно μ_1, Σ и μ_2, Σ. Та-

ким образом, мы заранее вводим условие, что ковариационные матрицы распределений f_1 и f_2 равны Σ.

В соответствии с выражением

$$R_1 : \frac{f_1(X)}{f_2(X)} \geqslant \frac{C(A_1/A_2)\, q_2}{C(A_2/A_1)\, q_1}$$

и приняв

$$\frac{C(A_1/A_2)\, q_2}{C(A_2/A_1)\, q_1} = k, \tag{6.6}$$

можно определить область R_1 с помощью неравенства:

$$R_1 : \frac{f_1(X)}{f_2(X)} = \frac{\exp\left[-\frac{1}{2}\{X-\mu_1\}\,\Sigma^{-1}\{X-\mu_1\}'\right]}{\exp\left[-\frac{1}{2}\{X-\mu_2\}\,\Sigma^{-1}\{X-\mu_2\}'\right]}. \tag{6.7}$$

Заметим, что μ_1 и μ_2 рассматриваются как m-мерные векторы-строки. После несложных преобразований предыдущее неравенство можно записать в виде:

$$R_1 : X\Sigma^{-1}\{\mu_1-\mu_2\}' - \frac{1}{2}\{\mu_1+\mu_2\}\,\Sigma^{-1}\{\mu_1-\mu_2\}' \geqslant \ln k. \tag{6.8}$$

В подавляющем большинстве геологических ситуаций нет никаких данных, позволяющих судить о вероятности q_1 и q_2, а нередко и цены потерь $C(A_2/A_1)$ и $C(A_1/A_2)$ невозможно бывает определить. В таких ситуациях ничего не остается делать, как допустиить, что $C(A_1/A_2) = C(A_2/A_1)$ и $q_1 = q_2$. Тогда $\ln k = 0$ и R_1 определится неравенством

$$R_1 : X\Sigma^{-1}\{\mu_1-\mu_2\}' \geqslant \frac{1}{2}\{\mu_1+\mu_2\}\,\Sigma^{-1}\{\mu_1-\mu_2\}'. \tag{6.9}$$

Область R_2 будет определена строгим неравенством, имеющим обратный знак:

$$R_2 : X\Sigma^{-1}\{\mu_1-\mu_2\}' < \frac{1}{2}\{\mu_1+\mu_2\}\,\Sigma^{-1}\{\mu_1-\mu_2\}'. \tag{6.10}$$

Выражение, стоящее в левой части этих неравенств, называется дискриминантной функцией, которая и служит критерием для отнесения рассматриваемого результата наблюдения z_t в выборке к совокупности A_1 или A_2. Подставив значение z_t на место X в формулу (6.7), получим в результате или неравенство (6.9), или (6.10), что даст возможность сделать вывод о принадлежности z_t. Заметим также, что правая часть неравенств (6.9) и (6.10), равная

$$\frac{1}{2}\{\mu_1+\mu_2\}\,\Sigma^{-1}\{\mu_1-\mu_2\}', \tag{6.11}$$

представляет собой константу, являющуюся пороговым значением в условиях сделанных ограничений.

Необходимо подчеркнуть, что описанный метод позволяет проводить классификационное отнесение только одного элемента выборки и не распространяется на выборку в целом. Более того, выводы, получаемые в результате применения такого метода относительно геологического объекта, могут оказаться неоднозначными.

Случай, когда μ_1, μ_2 и Σ оцениваются по выборке. Хотя ситуация, описанная в предыдущем разделе, и может встретиться в практике при геологических исследованиях, обычно параметры μ_1, μ_2 и Σ остаются неизвестными и оцениваются по выборке. Таким образом, в распоряжении исследователя имеются выборка объема n_1 наблюдений X_t ($t = 1, 2, \ldots, n_1$) над объектами совокупности A_1 и вторая выборка, объем которой n_2 наблюдений v_t ($t = 1, 2, \ldots, n_2$) над объектами совокупности A_2. По этим данным мы должны построить решающее правило, позволяющее относить некоторое наблюдение z_t из третьей выборки к A_1 или A_2.

Наилучшими оценками для μ_1 и μ_2 будут векторы средних арифметических:

$$\overline{X} = \frac{1}{n_1}\sum_{t=1}^{n_1} X_t = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_t, \ldots, \bar{x}_m\}, \tag{6.12}$$

$$\overline{Y} = \frac{1}{n_2}\sum_{t=1}^{n_2} Y_t = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_t, \ldots, \bar{y}_m\}, \tag{6.13}$$

а оценкой неизвестной ковариационной матрицы Σ будет матрица

$$S = \frac{1}{n_1+n_2+2}\left[\sum_{t=1}^{n_1}\{X_t-\overline{X}\}'\{X_t-\overline{X}\} + \sum_{t=1}^{n_2}\{Y_t-\overline{Y}\}^1\{Y_t-\overline{Y}\}\right]. \tag{6.14}$$

Если мы подставим эти оценки в выражение (6.9) вместо μ_1, μ_2 и Σ, то получим решающее правило, определяющее область R_1:

$$R_1 : ZS^{-1}\{\overline{X}-\overline{Y}\}' \geqslant \frac{1}{2}\{\overline{X}+\overline{Y}\}\,S^{-1}\{\overline{X}-\overline{Y}\}'. \tag{6.15}$$

Необходимо отметить, что левая часть неравенства, предложенная Р. Фишером, представляет собой линейную дискриминантную функцию, обладающую наибольшей дисперсией между выборками относительно дисперсии внутри выборок.

Таким образом, m-мерное наблюдение z_t подставляется на место z в выражение (6.15), чем определяется его принадлежность к A_1, если выполнено неравенство (6.15), и, наоборот, к A_2, если неравенство имеет обратный знак.

Случай, когда ковариационные матрицы известны и неравны. В предыдущих разделах мы рассматривали случай двух m-мерных нормальных распределений при условии, что соответствующие им ковариационные матрицы Σ_1 и Σ_2 равны. Однако это требование в большинстве геологических ситуаций может не выполняться, или же вообще могут отсутствовать какие-либо данные о соотношении

ковариационных матриц. В связи с этим мы сначала рассмотрим наиболее простой случай, когда μ_1 и μ_2, Σ_1 и Σ_2 известны и неравны, а затем перейдем к более сложной ситуации.

Как показано в работе Г. С. Лбова, в данной ситуации полезно применить квадратичное решающее правило, основанное на отношении правдоподобия, которое определит области R_1 и R_2:

$$R_1: \{X-\mu_2\}\,\Sigma_2^{-1}\,\{X-\mu_2\}' - \{X-\mu_1\}\,\Sigma_1^{-1}\,\{X-\mu_1\}' \geqslant \geqslant 2\ln\frac{|\Sigma_1^{1/2}|}{|\Sigma_2^{1/2}|}, \tag{6.16}$$

$$R_2: \{X-\mu_2\}\,\Sigma_2^{-1}\,\{X-\mu_2\}' - \{X-\mu_1\}\,\Sigma_1^{-1}\,\{X-\mu_1\}' < <2\ln\frac{\left|\Sigma_1^{\frac{1}{2}}\right|}{\left|\Sigma_2^{\frac{1}{2}}\right|}. \tag{6.17}$$

Таким образом, подлежащее классификации m-мерное наблюдение $z_t = \{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$ будет отнесено к первой группе, если в результате подстановки на место X в формулу (6.16) будет получено значение, принадлежащее R_1, и, наоборот, ко второй группе, если вычисленное значение дискриминантной функции окажется в области R_2.

Вопрос о риске, связанном с принятием решений в данной ситуации, подробно рассмотрен в работах А. Н. Бугайца, Г. С. Лбова и М. К. Камалова.

Случай, когда ковариационные матрицы неравны и μ_1, μ_2, Σ_1 *и* Σ_2 *оцениваются по выборке.* Наиболее часто наблюдаемая в геологии реальная ситуация заключается в том, что параметры многомерных распределений μ_1, μ_2, Σ_1 и Σ_2 остаются неизвестными и оцениваются по выборочным значениям:

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}, \quad t = 1, 2, \ldots, n_1, \tag{6.18}$$

$$Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}, \quad t = 1, 2, \ldots, n_2, \tag{6.19}$$

взятым из двух совокупностей, для которых нужно построить дискриминантную функцию. Оценки для μ_1, μ_2, Σ_1 и Σ_2 будут определены формулами:

$$\overline{X} = \frac{1}{n_1}\sum_{t=1}^{n_1} X_t = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_j, \ldots, \overline{x}_m\}, \tag{6.20}$$

$$\overline{Y} = \frac{1}{n_2}\sum_{t=1}^{n_2} Y_t = \{\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_j, \ldots, \overline{y}_m\}, \tag{6.21}$$

$$S_1 = \frac{1}{n_1 - 1}\sum_{t=1}^{n_1} \{X_t - \overline{X}\}'\{X_t - X\}, \tag{6.22}$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{t=1}^{n_2} \{Y_t - \overline{Y}\}'\{Y_t - \overline{Y}\}. \tag{6.23}$$

В данной ситуации возможны два решения. Первое из них — это построение линейной дискриминантной функции, которая будет определена выражением

$$z\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}\{\overline{X} - \overline{Y}\}'. \tag{6.24}$$

Критические области R_1 и R_2 для этой функции определяются неравенствами:

$$R_1: z\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}\{\overline{X} - \overline{Y}\}' \geqslant \frac{1}{2}\{\overline{X} + \overline{Y}\}\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}\{\overline{X} - \overline{Y}\}', \tag{6.25}$$

$$R_2: z\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}\{\overline{X} - \overline{Y}\}' < \frac{1}{2}\{\overline{X} + \overline{Y}\}\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}\{\overline{X} - \overline{Y}\}'. \tag{6.26}$$

Таким образом, результат наблюдения $z_t = \{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$ относится к первой совокупности, если после его подстановки в формулу (6.24) на место z окажется, что вычисленное значение попадет в область R_1. Наоборот, если это значение попадет в область R_2, то z_t следует отнести ко второй совокупности.

Квадратичная дискриминантная функция, построенная по этим данным, будет иметь вид:

$$\{z - \overline{Y}\}\,S_2^{-1}\,\{z - \overline{Y}\}' - \{z - \overline{X}\}\,S_1^{-1}\,\{z - \overline{X}\}'. \tag{6.27}$$

Критические области R_1 и R_2 на основе этой функции определяются следующим образом:

$$R_1: \{z - \overline{Y}\}\,S_2^{-1}\,\{z - \overline{Y}\}' - \{z - \overline{X}\}\,S_1^{-1}\,\{z - \overline{X}\}' \geqslant \geqslant 2\ln\frac{|S_1|^{1/2}}{|S_2|^{1/2}}, \tag{6.28}$$

$$R_2: \{z - \overline{Y}\}\,S_2^{-1}\,\{z - \overline{Y}\}' - \{z - \overline{X}\}\,S_1^{-1}\,\{z - X\}' < <2\ln\frac{|S_1|^{1/2}}{|S_2|^{1/2}}. \tag{6.29}$$

Таким образом, если в результате подстановки классифицируемого наблюдения $z_t = \{z_{t1}, z_{t2}, \ldots, z_{tm}\}$ в формулу (6.27) на место z будет выполнено неравенство (6.28), то z_t относится к первой совокупности, если же окажется, что вычисленное значение функции (6.27) принадлежит R_2, то z_t нужно отнести ко второй совокупности.

Об эффективности применения линейных и квадратичных решающих правил. Вопрос об эффективности применения того или иного решающего правила в различных реальных геологических ситуациях подробно изучался А. Н. Бугайцом [9], которым были получены следующие результаты.

Им установлено, что линейные решающие правила оптимальны или близки к оптимальным не только в условиях многомерных нормальных распределений изучаемых совокупностей, но и при весьма более широких условиях. Оказалось, что линейные решающие правила являются оптимальными, если изучаемые распределения являются унимодальными и их плотности равномерно убывают с удалением от средних значений. Этот результат в значительной степени расширяет область применения линейных решающих правил, которые отличаются значительной простотой построения.

Квадратичные решающие правила, как отмечает А. Н. Бугаец, казываются полезными в тех случаях, когда число классифицируемых совокупностей более двух.

ГЛАВА 7

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ — статистический метод исследования выборочных данных, проводимого с целью выявления и оценки степени влияния на изучаемую случайную величину различных одновременно действующих факторов. В основе дисперсионного анализа лежит такое разложение общей изменчивости выборочных данных, при котором удается отделить изменчивость, связанную с некоторыми фиксируемыми исследователем факторами, от изменчивости, обусловленной факторами, неконтролируемыми в данном эксперименте.

Надежная статистическая оценка вклада контролируемых факторов возможна лишь при условии, что эксперимент (наблюдение) некоторым образом организован. Это определяет тесную связь дисперсионного анализа с планированием эксперимента. В тех случаях, когда изменение хотя бы части контролируемых факторов может быть измерено количественно, пользуются комбинацией дисперсионного и регрессионного анализа (см. гл. 10).

Взаимоотношения между этими видами статистического анализа [46] ясно видны, если рассматривать выборочные значения исследуемой случайной величины как линейную комбинацию:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{ji}\beta_j + \ldots + x_{pi}\beta_p + \varepsilon_i,$$

$$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, p,$$

где y_i — результат наблюдения с номером i; $\{\beta_j\}$ — фиксированные в данном эксперименте факторы; $\{x_{ij}\}$ — некоторые постоянные коэффициенты; ε_i — случайная нормально распределенная величина с нулевым математическим ожиданием и дисперсией σ^2. Компонента ε_i обычно трактуется как «ошибка» измерений, ее величина отражает влияние на результат наблюдения некоторых неконтролируемых в данном эксперименте факторов [46]. Это множество уравнений в матричной записи имеет вид:

$$y = X'\beta + \varepsilon,$$

где
$$y = \begin{Bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_i \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{Bmatrix}, \quad \beta = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_i \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{Bmatrix}, \quad \varepsilon = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_i \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{Bmatrix},$$

$$X = \begin{Bmatrix} x_{11} & x_{12} & & x_{1i} & & x_{1n} \\ x_{21} & x_{22} & & x_{2i} & & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{j1} & x_{j2} & \ldots & x_{ji} & \ldots & x_{jn} \\ x_{p1} & x_{p2} & \ldots & x_{pi} & \ldots & x_{pn} \end{Bmatrix}.$$

Если элементы матрицы X суть 0 или 1, то x_{ji} называется «переменной-указателем»; значение, принимаемое x_{ji}, указывает на наличие ($x_{ji} = 1$) или отсутствие ($x_{ji} = 0$) влияния факторов $\beta_j \in \beta$ (в условиях данного наблюдения). В этом случае для получения выводов относительно $\{\varepsilon_i\}$ и некоторых $\{\beta_j\}$ пользуются методами дисперсионного анализа. Если элементы матрицы X пробегают непрерывные множества значений, то выводы относительно вышеуказанных величин делаются на основе регрессионного анализа. Наконец, если $\{x_{ji}\}$ — обоих видов, то используют ковариационный анализ. Общим условием применения дисперсионного анализа является выполнение следующих равенств:

$$M(\varepsilon) = 0, \quad M(\varepsilon\varepsilon') = \sigma^2 I,$$

где $M(\cdot)$ — математическое ожидание; 0 — нулевой вектор; I — единичная матрица.

Отсюда вытекает, что случайные величины ε_i: а) некоррелированы и б) имеют равные дисперсии. Обычно полагают также, что величины ε_i распределены нормально и тогда условие «а» формули-

руется как независимость. Другие предположения будут введены при обсуждении различных моделей дисперсионного анализа.

Модель I — модель с постоянными факторами, в которой все β_j, $j = 1, 2, \ldots, p$ могут рассматриваться как неизвестные постоянные. Величина β_j называется аддитивной постоянной (в прикладной статистике она именуется «генеральным средним», если $x_{jt} = 1$ при любом i.

Модель II — модель, в которой все параметры β_j случайны, за исключением, может быть, одного, являющегося постоянным. Такая модель называется также моделью со случайными факторами.

Если хотя бы один параметр β_j случаен и по крайней мере один неслучаен (но не является аддитивной постоянной), то модель называется смешанной моделью (модель III).

В геологических исследованиях ситуации, ведущие к основным моделям дисперсионного анализа, могут быть проиллюстрированы следующими примерами. При изучении некоторой определенной территории обнаружены выходы трех гранитных массивов, которые после специальных исследований были отнесены к трем последовательно сменяющим друг друга фазам магматизма. На образцах из этих массивов проведены измерения радиоактивности. Был поставлен вопрос, различаются ли средние значения радиоактивности опробованных гранитов. Ответ на него позволяет проверить влияние некоторого фактора, названного «приуроченностью к определенной фазе магматического процесса, проявляемого в данном регионе». В приведенном случае исследуемые гранитные массивы фиксированы и, естественно, неслучайны: все элементы конечной совокупности вовлечены в анализ. Вместе с тем результаты измерений радиоактивности, выполненных на этих массивах, обладают всеми свойствами случайных величин. Изменим теперь ситуацию таким образом, что исследователь получает доступ к достаточно большому множеству гранитных массивов, каждый из которых опять-таки может быть отнесен к одной из вышеназванных фаз. Если опробование организовано так, что радиоактивность измеряется лишь на некоторых массивах, то последние можно рассматривать как случайную выборку из множества всех возможных вариантов. Разделив последнюю на три группы, каждая из которых соответствует определенной фазе магматизма, вновь получаем возможность оценить влияние того же фактора, теперь уже по схеме модели II. Таким образом, по характеру постановки вопроса модели I и II идентичны, чего, однако, нельзя сказать о параметрах, вовлекаемых в дисперсионный анализ. В первом случае (модель I) математические ожидания выборочных средних значений радиоактивности трактуются как неизвестные постоянные, тогда как во втором случае (модель II) математические ожидания того же геологического признака должны рассматриваться как реализация случайных величин. Ниже будет показано, что эти различия существенным образом сказываются и на формулировании исследуемых методами дисперсионного анализа гипотез, и на выборе статистических критериев для их проверки.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Модель I. Результаты измерений некоторого геологического признака на p объектах запишем в виде матрицы Y:

$$Y = \begin{Bmatrix} y_{11} & y_{12} & \ldots & y_{1j} & \ldots & y_{1p} \\ y_{21} & y_{22} & \ldots & y_{2j} & \ldots & y_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ y_{n1} & y_{n2} & \ldots & y_{nj} & \ldots & y_{np} \end{Bmatrix}.$$

Такая запись матрицы Y означает, что на каждом объекте, соответствующем j-й градации проверяемого фактора, произведено одинаковое число наблюдений, равное n. В условиях модели I результаты наблюдений $\{y_{ij}\}$ следует рассматривать как выборочные значения случайных нормально распределенных величин ξ_1, $\xi_2, \ldots, \xi_j, \ldots, \xi_p$ с параметрами: $M(\xi_j) = \mu_j$, $D(\xi_j) = \sigma^2$ для $j = 1, 2, \ldots, p$ (равенство дисперсий).

Таким образом, для каждой градации фактора мы имеем вполне определенное (фиксированное) среднее, являющееся постоянным.

Основное уравнение однофакторного дисперсионного анализа в условиях модели I имеет вид

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

где μ — генеральное среднее, определяемое формулой

$$\mu = \frac{1}{p} \sum_{j=1}^{p} \mu;$$

α_j — эффект j-й градации исследуемого фактора, определяемый формулой $\alpha_j = \mu_j - \mu$; ε_{ij} — случайная независимая величина («ошибка» наблюдения для i-го измерения величины ξ_j), отражающая влияние на результаты эксперимента неконтролируемых в данном наблюдении факторов.

Так как μ_j — константа ($j = 1, 2, \ldots, p$), то постоянна и α_j. Поэтому модель I называют также моделью с постоянными эффектами.

Проверяемая статистическая гипотеза может быть сформулирована следующим образом:

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_j = \ldots = \alpha_p,$$

т. е. влияние исследуемого фактора на всех его уровнях (градациях) одинаково.

Другими словами, в условиях H_0 справедливо равенство:

$$\mu_1 = \mu_2 = \ldots = \mu_j = \mu_p = \mu.$$

Проверка нулевой гипотезы осуществляется по следующей схеме;

а) вычисляют выборочные средние $\bar{y}_j$, $\bar{y}$:

$$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij},$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}, \quad \text{где} \quad N = np;$$

б) находят суммы квадратов отклонений выборочных значений от соответствующих средних:

сумму, характеризующую изменчивость, обусловленную исследуемым фактором:

$$Q_1 = n\sum_{j=1}^{p}(\bar{y}_j - \bar{y})^2;$$

сумму, характеризующую изменчивость внутри каждой градации фактора (остаточная изменчивость):

$$Q_2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij} - \bar{y}_j)^2;$$

сумма, характеризующая общую изменчивость наблюдаемого признака:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij} - \bar{y})^2.$$

Если все названные выше предположения о величинах y_{ij}, α_j, ε_{ij} выполнены, то справедливо равенство

$$Q = Q_1 + Q_2.$$

Опираясь на Q, Q_1, Q_2, можно получить оценки соответствующих дисперсий:

$$S^2 = \frac{Q}{N-1}; \qquad S_1^2 = \frac{Q_1}{p-1}; \qquad S_2^2 = \frac{Q_2}{N-p}.$$

Критерий, используемый для проверки гипотезы H_0, имеет вид

$$F = \frac{Q_1(N-p)}{Q_2(p-1)} = \frac{S_1^2}{S_2^2}.$$

При условии, что гипотеза H_0 — верна, распределение критерия подчиняется закону Фишера (F-распределение). Гипотеза отклоняется, если $F > F_{\alpha, f_1, f_2}$, где F_{α, f_1, f_2} — табличное значение F-распределения, соответствующее уровню значимости α при степенях свободы $f_1 = p-1$, $f_2 = N-p$.

Модель II. Как и прежде, результаты наблюдений записываются в виде матрицы Y. Однако в отличие от модели I выбор объектов, соответствующих некоторым градациям исследуемого фактора, рандомизирован, что определяет случайный характер факторных эффектов.

Это приводит к изменению структуры основного уравнения однофакторного дисперсионного анализа; последнее в рамках модели II приобретает вид

$$y_{ij} = \mu + a_{ij} + \varepsilon_{ij},$$

где μ — генеральное среднее (аддитивная постоянная); a_j — значение случайной величины, представляющей отклонения среднего значения измеряемого признака на j-м объекте (m_j) от генерального среднего, т. е. $a_j = m_j - \mu$.

Как и ранее, $M(\varepsilon_{ij}) = 0$, $D(\varepsilon_{ij}) = \sigma^2$. Введем следующее предположение: распределение случайных эффектов a_j нормально с нулевым средним и дисперсией $\sigma_a^2 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_j^2 \ldots = \sigma_p^2$. Если $\{a_j\}$ и $\{\varepsilon_{ij}\}$ независимы в совокупности, то распределение y_{ij} подчиняется нормальному закону с математическим ожиданием и дисперсией, равной $\sigma_a^2 + \sigma_\varepsilon^2$. Параметры $\sigma_\alpha^2 + \sigma_\varepsilon^2$ называются компонентами дисперсии $D(y_{ij}) = \sigma_a^2 + \sigma_\varepsilon^2$; последняя может рассматриваться как мера неоднородности сравниваемых объектов.

В связи с этим нулевая гипотеза, соответствующая отсутствию влияния исследуемого фактора на переменную y_{ij}, может быть записана как: $H_0 : \sigma_a^2 = 0$.

Для проверки H_0 вполне применима описанная выше схема однофакторного дисперсионного анализа (модель I). Таким образом, модели I и II с точки зрения вычислительных процедур идентичны. Однако на этапе интерпретации должны учитываться различия в структурах моделей. Статистические выводы, полученные в рамках модели I, относятся в общем случае только к тем объектам, на базе которых выполняли дисперсионный анализ. В условиях модели II результаты дисперсионного анализа естественным образом распространяются на всю совокупность исследуемых объектов, из которой была получена выборка.

Однофакторный дисперсионный анализ с неравным числом наблюдений:

$$y = \begin{Vmatrix} y_{11} & y_{12} & \ldots & y_{1j} & \ldots & y_{1p} \\ y_{21} & y_{22} & \ldots & y_{2j} & \ldots & y_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ y_{i1} & y_{i2} & \ldots & y_{ij} & \ldots & y_{ip} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ y_{n_1 1} & y_{n_2 2} & \ldots & y_{n_j} & \ldots & y_{n_p} \end{Vmatrix}.$$

Общая схема анализа остается прежней, некоторые изменения вносятся лишь в формулы, по которым вычисляются оценки средних и суммы квадратов отклонений:

$$y_{*j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij};$$

$$y_{**}=\frac{1}{\sum\limits_{j=1}^{e} nj}\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij}=\frac{1}{N}\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ij};$$

$$Q_1=\sum_{j=1}^{p} nj\,(y_{*j}-y_{**})^2;$$

$$Q_2=\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij}-y_{*j})^2;$$

$$Q=\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij}-y_{**})^2.$$

Если величина F-критерия превысила критическое значение F_{α, f_1, f_2}, то нулевая гипотеза бракуется. В этом случае полагают, что существует по крайней мере одна пара средних, например, μ_k и μ_l ($k \neq l$), при которой $\mu_k \neq \mu_l$.

Возникает задача: отыскать те объекты (градации фактора), для которых различия средних значимы для принятого уровня α. Применение в этой ситуации двухвыборочного t-критерия малоэффективно, так как последний не позволяет выполнить сравнение некоторого интересующего нас среднего с совокупностью остальных средних. Более общее решение задачи дает метод Шеффе (S-метод). С его помощью удается построить доверительные интервалы для любой линейной комбинации средних:

$$\theta=C_1\mu_2+C_2\mu_2+\ \ldots\ +C_p\mu_p,$$

$$\text{где}\quad \sum_{j=1}^{p} C_j=0.$$

Функция θ, определенная таким образом, называется контрастом. Например, пусть $p=5$; требуется сопоставить μ_1 с остальными средними. Тогда $C_1=4$; $C_2=-1$, $C_3=-1$, $C_4=-1$. Выражение $4\mu_1-\mu_2-\mu_3-\mu_4-\mu_5$ представляет собой контраст. Выборочную оценку θ найдем путем замены μ_j величинами $\bar{y}_j$: $H=\sum\limits_{j=1}^{p} C_j\bar{y}_j$. Как и ранее, будем полагать, что $\bar{y}_1, \ldots, \bar{y}_p$ независимы и нормально распределены, а $D_{\bar{y}_j}=\sigma^2/n_j$. Отсюда следует, что

$$D(H)=\sigma^2\sum_{j=1}^{p}\frac{C_j^2}{n_j}.$$

Выборочной оценкой $D(H)$ является величина

$$\hat{D}(H)=S^2\sum_{j=1}^{p}\frac{C_j^2}{n_j},$$

где $S^2=Q/(N-1)$, $Q=\sum\limits_{j=1}^{p}\sum\limits_{i=1}^{n_j}(y_{ij}-\bar{y})^2$, $N=\sum\limits_{j=1}^{p} n_j$, n_i — объем выборки, соответствующей j-й градации исследуемого фактора. Доверительный интервал величины θ определяется следующим соотношением:

$$p\left\{H-S\sqrt{\hat{D}H}<\theta<H+S\sqrt{\hat{D}(H)}\right\}=1-\alpha,$$

где $S=[(p-1)F_{\alpha, p-1, N-1}]^{1/2}$.

Если построенный таким образом доверительный интервал не содержит нуля, то проверяемое среднее существенно (при заданном уровне значимости α) отличается от совокупности остальных средних.

Проведя $p(p-1)/2$ таких сравнений, можно выделить все «контрастирующие» значения μ_j, $j=1, 2, \ldots, p$ и тем самым обнаружить источник неоднородности средних.

ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

В геологии нередки ситуации, в которых удается контролировать два и более факторов, предположительно управляющих исследуемой случайной величиной. В таких случаях вскрытие возможных причин изменчивости изучаемого геологического признака может быть выполнено методами многофакторного дисперсионного анализа. Последний позволяет не только оценить влияние отдельных факторов, но и обнаружить (при определенных условиях) их взаимодействие. Рассматриваемый ниже случай с двумя факторами легко может быть обобщен на большее число факторов. В двухфакторном анализе обычно различают многостороннюю (перекрестную) классификацию, когда в таблице исходных данных $\{y_{ij}\}$ каждый j-й столбец (j-я градация фактора B) содержит одинаковое число групп, соответствующих градациям i-го фактора A, и иерархическую классификацию, если фактор B (второстепенный в рамках данной задачи) сгруппирован внутри главного фактора A. В иерархической классификации число градаций фактора B, фиксируемых внутри различных градаций фактора A, может быть неодинаковым.

Если же в каждой группе A_i имеется равное число подгрупп B_j, то такая иерархическая классификация обозначается специальным термином «гнездовая классификация». В случае иерархической классификации проблемы взаимодействия факторов не возникает. Это же замечание справедливо и для перекрестной классификации, если в каждой ячейке только одно наблюдение. В связи с этим, в дальнейшем при описании техники дисперсионного анализа во внимание будет приниматься не только характер исследуемых факторов (модели I, II и III), но и тип классификации, а также число наблюдений в ячейках (n_{ij} — наблюдения без повторений, $n_{ij}>1$ — наблюдения с повторениями).

Как и в однофакторном дисперсионном анализе, почти все вычислительные процедуры для моделей I, II и III идентичны, поэ-

Таблица 5

Исходные данные для двухфакторного дисперсионного анализа

A_1	B_1	...	B_j	...	B_p
	$y_{111}, y_{112}, \ldots, y_{11n}$	...	$y_{1j1}, y_{1j2}, \ldots, y_{1jn}$	...	$y_{1p1}, y_{1p2}, \ldots, y_{1pn}$
	...	...	...	...	...
A_i	$y_{i11}, y_{i12}, \ldots, y_{i1n}$	...	$y_{ij1}, y_{ij2}, \ldots, y_{ijn}$	...	$y_{ip1}, y_{ip2}, \ldots, y_{ipn}$
...	...	...	...	...	...
A_q	$y_{q11}, y_{q12}, \ldots, y_{q1n}$	...	$y_{qj1}, y_{qj2}, \ldots, y_{qjn}$	...	$y_{qp1}, y_{qp2}, \ldots, y_{qpn}$

тому указания на ту или иную модель будут делаться лишь при появлении существенных различий между ними.

Перекрестная классификация с n повторениями в ячейке. Результаты наблюдений удобно представить в виде табл. 5.

Для дальнейшего изложения удобно представить наблюдение в виде следующей общей модели:

$$y_{ijm} = \mu + \alpha_i + \beta_j + \gamma + \varepsilon_{ijm}, \quad i = 1, 2, \ldots, q;$$
$$j = 1, 2, \ldots, p; \quad m = 1, 2, \ldots, n.$$

Здесь μ — генеральное среднее; α_i — влияние, обусловленное i-й градацией фактора A; β_j — влияние, обусловленное j-й градацией фактора B; γ — эффект взаимодействия факторов; ε_{ijm} — вариация внутри отдельной ячейки. Предполагается, что α_i, β_j, γ, ε_{ijm} распределены нормально с параметрами соответственно $(0, \sigma_\alpha^2)$, $(0, \sigma_\beta^2)$, $(0, \sigma_\gamma^2)$, $(0, \sigma_\varepsilon^2)$. В условиях модели I факторы A, B, AB рассматриваются как фиксированные, применительно же к модели II — как случайные. Для смешанной модели одному из факторов приписывается систематическое влияние, другому — случайное. Для определенности будем полагать, что в модели III фактор B — фиксирован, а фактор A, так же как и их взаимодействие AB,— случаен.

Общая схема двухфакторного дисперсионного анализа с повторениями следующая:

а) вычисляются выборочные средние:

$$\bar{y}_{ij} = \frac{1}{n}\sum_{m=1}^{n} y_{ijm}; \quad \bar{y}_i = \frac{1}{np}\sum_{j=1}^{p}\sum_{m=1}^{n} y_{ijm}; \quad \bar{y}_j = \frac{1}{ng}\sum_{i=1}^{q}\sum_{m=1}^{n} y_{ijm};$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n} y_{ijm};$$

б) определяются суммы квадратов отклонений, соответствующие различным источникам изменчивости:

изменчивость, обусловленная влиянием фактора A:

$$Q_1 = np\sum_{i=1}^{q}(\bar{y}_i - \bar{y})^2;$$

изменчивость, обусловленная влиянием фактора B:

$$Q_2 = nq\sum_{j=1}^{p}(\bar{y}_j - \bar{y})^2;$$

изменчивость, обусловленная взаимодействием факторов AB:

$$Q_3 = n\sum_{i=1}^{q}\sum_{j=1}^{p}(\bar{y}_{ij} - \bar{y}_i + \bar{y})^2;$$

изменчивость, связанная с различиями внутри ячеек:

$$Q_4 = \sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n}(y_{ijm} - \bar{y}_{ij})^2;$$

общая изменчивость наблюдаемого признака:

$$Q = \sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n}(y_{ijm} - \bar{y})^2.$$

Справедливо равенство $Q = Q_1 + Q_2 + Q_3 + Q_4$. Величина Q_1 соответствует $(q-1)$ степеням свободы, $Q_2 - (p-1)$, $Q_3 - (q-1)\cdot(p-1)$, $Q_4 - (N-qp)$, $Q - (N-1)$.

Теперь нетрудно найти средние квадраты отклонений:

$$M_1 = Q_1/(q-1); \quad M_2 = Q_2/(p-1);$$
$$M_3 = Q_3/[(q-1)(p-1)]; \quad M_4 = Q_4/N - qp.$$

Проверка гипотез. Модель I. Обозначим символами $\mu_1, \ldots, \mu_i, \ldots, \mu_q$ математические ожидания величин, составляющих строчки в табл. 4, символами $\mu_1, \ldots, \mu_j, \ldots, \mu_{0p}$ — математические ожидания столбцов из той же таблицы. Математическое ожидание наблюдений, входящих в определенную ячейку, запишем как μ_{ij}. Тогда $\alpha_i = \mu_i - \mu$ будет соответствовать эффекту i-й градации фактора A, $\beta_{0j} = \mu_{0j} - \mu$ — эффекту j-й градации фактора B, $\gamma_{2j} = \mu_{ij} - \alpha_i - p_i - \mu$ — эффекту j-го уровня фактора B в условиях i-й градации фактора A (эффект взаимодействия).

Сформулируем нулевые гипотезы, утверждающие, что влияния фактора A и фактора B одинаковы, а взаимодействие между A и B отсутствует:

$$H_0': \alpha_1 = \alpha_2 = \ldots = \alpha_i = \ldots = \alpha_q;$$
$$H_0'': \beta_1 = \beta_2 = \ldots = \beta_j = \ldots = \beta_p;$$
$$H_0''': \gamma_{ij} = 0 \quad \text{для всех} \quad i, j.$$

Критерии, используемые для проверки этих гипотез, имеют вид:

$$F' = \frac{M_1}{M_4} = \frac{Q_1(N-qp)}{Q_4(q-1)}; \qquad F'' = \frac{M_2}{M_3} = \frac{Q_2(N-qp)}{Q_4(p-1)};$$

$$F''' = \frac{M_3}{M_4} = \frac{Q_3(N-qp)}{Q_4(p-1)(q-1)}.$$

Если гипотеза H_0 верна, то отношения M_1/M_4, M_2/M_4, M_3/M_4 подчиняются F-распределению с соответствующими степенями свободы. Действие факторов A, B, AB считается существенным (при уровне значимости α), если

$$F' \geqslant F_{\alpha,\, q-1,\, N-qp}; \quad F'' \geqslant F_{\alpha,\, p-1,\, N-qp}; \quad F''' \geqslant F_{\alpha,\, (p-1)(q-1),\, N-qp}.$$

Модель II. Проверяемые нулевые гипотезы запишем как:

$$H_0' : \sigma_\alpha^2 = 0; \quad H_0'' : \sigma_\beta^2 = 0; \quad H_0''' : \sigma_{\alpha\beta}^2 = 0.$$

Соответствующие критерии имеют вид

$$F' = \frac{M_1}{M_3} = \frac{Q_1(q-1)(p-1)}{Q_3(q-1)} = (p-1)\frac{Q_1}{Q_3};$$

$$F'' = \frac{M_2}{M_3} = \frac{Q_2(q-1)(p-1)}{Q_3(p-1)} = (q-1)\frac{Q_2}{Q_3};$$

$$F''' = \frac{M_3}{M_4} = \frac{Q_3(N-qp)}{Q_4(q-1)(p-1)}.$$

В условиях нулевой гипотезы M_1/M_3, M_2/M_3 и M_3/M_4 имеют F-распределение с соответствующими степенями свободы. Гипотезы H_0', H_0'', H_0''' отклоняются (при заданном уровне значимости α), если

$$F' \geqslant F_{\alpha,\, q-1,\, (q-1)(p-1)}; \quad F'' \geqslant F_{\alpha,\, p-1,\, (q-1)(p-1)};$$

$$F''' \geqslant F_{\alpha,\, (q-1)(p-1),\, N-qp}.$$

Модель III. Проверяемые нулевые гипотезы:

$$H_0' : \sigma_\alpha^2 = 0; \quad H_0'' : \beta_1 = \beta_2 = \ldots = \beta_j = \ldots = \beta_p;$$

$$H_0''' : \sigma_\gamma^2 = 0.$$

Критерии для проверки нулевых гипотез:

$$F' = \frac{M_1}{M_4} = \frac{Q_1(N-qp)}{Q_4(Q_1-1)}; \qquad F''' = \frac{M_3}{M_4} = \frac{Q_3(N-qp)}{Q_4(q-1)(p-1)};$$

$$F'' = \frac{M_2}{M_3} = (q-1)\frac{Q_2}{Q_3}.$$

Принятие решения осуществляется так же, как и в моделях I и II.

Перекрестная классификация с одним наблюдением в ячейке. Дисперсионный анализ данных такого рода проводят лишь в том случае, если известно, что

взаимодействие между столбцами и строчками исходной матрицы наблюдений отсутствует. В этом случае приемлемо следующее представление наблюдения y_{ijm}:

$$y_{ijm} = \mu + \alpha_i + \beta_j + \varepsilon_{ijm}$$

(так как $n = 1$, то индекс m может быть опущен).

Критерии для проверки главных эффектов (H_0' и H_0'') получают, разделив соответствующий средний квадрат (M_1 или M_2) на величину M_3, которую в этой ситуации уместно назвать «средним квадратом ошибки». Далее выполняется сравнение найденных отношений с F-распределением, при этом процедура выбора числа степеней свободы сохраняется прежней.

Перекрестная классификация с неравными числами наблюдений в ячейках. Неодинаковое количество наблюдений в ячейках приводит к нарушению ортогональности, что, в свою очередь, усложняет критерии проверки гипотез.

Гипотезу об отсутствии взаимодействия запишем следующим образом:

$$H_0''' : M(y_{ijm}) = \mu + \alpha_i + \beta_i, \quad i = 1, 2, \ldots, q;$$

$$j = 1, 2, \ldots, p; \quad m = 1, 2, \ldots, n_{ij},$$

где n_{ij} — число наблюдений в ячейке с номером i, j.

Введем обозначения:

$$N = \sum_{i=1}^{q}\sum_{j=1}^{p} n_{ij}; \quad Q_i = \sum_{j=1}^{p} n_{ij}; \quad H_j = \sum_{i=1}^{q} n_{ij};$$

$$q_i = \sum_{j=1}^{p}\sum_{m=1}^{n_{ij}} y_{ijm}; \quad h_j = \sum_{i=1}^{q}\sum_{m=1}^{n_{ij}} y_{ijm}.$$

Вычисление главных эффектов $\{\alpha_i\}$ и $\{\beta_j\}$ и взаимодействия сводится к решению системы линейных уравнений, получаемых при минимизации

$$\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m=1}^{n_{ij}} (y_{ijm} - \mu - \alpha_i - \beta_j)^2.$$

Приравнивая к нулю производные ∂/∂_μ, $\partial/\partial\alpha_i$, $\partial/\partial\beta_j$, получим:

$$N\hat{\mu} = \sum_{i=1}^{q} Q_i\hat{\alpha}_i + \sum_{j=1}^{p} H_j\beta_j = \sum_{i=1}^{q}\sum_{i=1}^{p}\sum_{m=1}^{n_{ij}} y_{ijm};$$

$$Q_i\hat{\mu} = Q_i\hat{\alpha}_i + \sum_{j=1} n_{ij}\beta_j = q_i; \quad i = 1, 2, \ldots, q;$$

$$H_j\hat{\mu} + H_j\hat{\beta}_j + \sum_{i=1}^{q} n_{ij}\hat{\alpha}_i = h_j; \quad j = 1, 2, \ldots, p.$$

Два последних выражения образуют систему линейных уравнений. Структура этой системы позволяет выполнить ряд упроще-

ний, в частности исключить $\{\hat{\beta}_j\}$, если $p > q$, или $\{\hat{\alpha}_i\}$, если $p < q$. Одновременно исключается и $\hat{\mu}$.

Новая система уравнений может быть записана как:

$$\sum_{i'=1}^{p} a_{ii'}\hat{\alpha}_i = \varphi_i, \quad i = 1, 2, \ldots, q,$$

где i' — фиксированное значение индекса j, организующее исключение $\{\hat{\beta}_j\}$ (при каждом заданном значении $i \in \{j\}$ индекс i пробегает все возможные значения от 1 до q);

$$a_{ii'} = \sigma_{ii'} Q_i = \sum_{j=1}^{p} \frac{n_{ij} n_{i'j}}{H_j},$$

$\sigma_{ii'}$ — символ Кронекера;

$$\varphi_i = q_i - \sum_{j=1}^{p} \frac{n_{ij} h_j}{H_j}.$$

Если исключаются $\{\hat{\alpha}_i\}_{i=1}^{i=q}$, то система имеет вид

$$\sum_{j'=1}^{p} b_{jj'}\hat{\beta}_{j'} = \varkappa_j, \quad j = 1, 2, \ldots, p,$$

где j' — фиксированное значение индекса i.

Критерий для проверки H_0''':

$$F''' = \frac{N-n}{n-q-p+1} \frac{Q_3 - Q_4}{Q_4},$$

где n — число непустых ячеек (если все ячейки заполнены, то $n = qp$);

$$Q_4 = \sum_{(i,j) \in D} \sum_{m=1}^{n_{ij}} (y_{ijm} - \bar{y}_{ij})^2;$$

$$\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{m=1}^{n_{ij}} y_{ijm}; \quad Q_3 = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{m=1}^{n_{ij}} y_{ijm}^2 - \sum_{i=1}^{q} \varphi_i \hat{\alpha}_i - \sum_{j=1}^{p} \frac{h_j^2}{H_j},$$

если исключались $\{\beta_j\}$, и

$$Q_3 = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{m=1}^{n_{ij}} y_{ijm}^2 - \sum_{j=1}^{p} \varkappa_i \hat{\beta}_i - \sum_{j=1}^{p} h_j^2/H_i,$$

если исключались $\{\hat{\alpha}_i\}$.

Статистика F''' имеет F-распределение с $(N-q-p+1)$ и $(N-n)$ степенями свободы.

Проверка гипотез H_0' и H_0'' в предположении, что H_0''' верна:

$$H_0': \quad \text{все} \quad \alpha_i = 0,$$

$$F' = \frac{N-q-p+1}{q-1} \frac{Q_1 - Q_3}{Q_3},$$

где

$$Q_1 = \sum_{i=1} \sum_{j=1}^{p} \sum_{m=1}^{n_{ij}} y_{ijm} - \frac{h_j}{H_j},$$

а Q_3 определено выше.

Статистика F' имеет F-распределение с $(q-1)$ и $(N-p-q+1)$ степенями свободы.

Аналогично проверяется гипотеза H_0''.

Обязательное условие проверки гипотез H_0' и H_0'' без предположения аддитивности — отсутствие пустых ячеек.

Найдем средние взвешенные:

$$\hat{A}_i = \sum_{j=1}^{p} \bar{y}_{ij} w_j, \quad \hat{B}_j = \sum_{i=1}^{q} \bar{y}_{ij} v_i,$$

где w_j, v_i — «веса» соответствующих столбцов и строк таблицы исходных данных, $w_j = H_j/N$; $v_i = Q_i/N$.

Определим суммы квадратов отклонений Q_A и Q_B:

$$Q_A = \sum_{i=1}^{q} W_i \hat{A}_i^2 - \left(\sum_{i=1}^{q} W_i\right)^{-1} \left(\sum_{i=1}^{q} W_i \hat{A}_i\right)^2;$$

$$Q_B = \sum_{j=1}^{p} V_j \hat{B}_j^2 - \left(\sum_{i=1}^{p} V_j\right)^{-1} \left(\sum_{j=1}^{p} V_j \hat{B}_j\right)^2,$$

где $W_i = \left(\sum_{j=1}^{p} \frac{w_j^2}{n_{ij}}\right)^{-1}$; $V_j = \left(\sum_{i=1}^{q} \frac{v_i^2}{n_{ij}}\right)^{-1}$.

Критерий для проверки гипотезы H_0':

$$F' = \frac{N-qp}{q-1} \frac{Q_A}{Q_4}$$

имеет F-распределение с $q-1$ и $N-qp$ степенями свободы.

Аналогично проверяется гипотеза H_0''.

Иерархическая классификация результатов наблюдений применяется в дисперсионном анализе в тех случаях, когда один фактор сгруппирован внутри другого, «главного» фактора. В геологии такая ситуация возникает, например, при разделении ошибок измерений и природной («внутренней») изменчивости исследуемого признака. Методические исследования такого

Таблица 6

Иерархическая классификация результатов наблюдений

A_1					...	A_i					...	A_q				
B_1	...	B_j	...	B_p	...	B_1	...	B_j	...	B_p	...	B_1	...	B_j	...	B_p
y_{111}		y_{1j1}		y_{ip1}	...	y_{i11}		y_{ij1}		y_{ip1}	...	y_{q11}		y_{qj1}		y_{qp1}
...		...		...		...		...		...		...		...		...
y_{11m}		y_{1jm}		y_{ipm}	...	y_{i1m}		y_{ijm}		y_{ipm}	...	y_{q1m}		y_{qjm}		y_{qpm}
...		...		...		...		...		...		...		...		...
y_{11n}		y_{1jn}		y_{ipn}	...	y_{i1n}		y_{ijn}		y_{ipn}	...	y_{q1n}		y_{qjn}		y_{qpn}

рода обычно планируются заранее, поэтому почти всегда удается обеспечить и равное число групп второстепенного фактора, и равное число наблюдений в этих подгруппах. Кроме того, случайный характер как ошибок измерений, так и значений изучаемого геологического признака определяет преимущественное появление ситуаций, соответствующих условиям модели II. В связи с этим при описании иерархического дисперсионного анализа ограничимся случаем гнездовой рандомизированной модели.

Исходные данные представлены в виде табл. 6.

Так как в рамках иерархической классификации взаимодействие отсутствует, то основное уравнение принятой модели дисперсионного анализа имеет вид

$$y_{ijm} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijm},$$

где μ — математическое ожидание всех результатов наблюдений; α_i — эффект i-й градации фактора A; $\beta_{j(i)}$ — эффект j-го уровня фактора B в пределах i-й градации фактора A; ε_{ijm} — эффект неконтролируемых факторов.

Дисперсию Y представим в виде суммы:

$$\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2.$$

Проверке подлежат гипотезы:

$$H_0': \sigma_\alpha^2 = 0; \quad H_0'': \sigma_\beta^2 = 0.$$

Определим суммы квадратов отклонений:

$$Q_A = np \sum_{i=1}^{q} (\bar{y}_i - \bar{y})^2; \quad Q_B = n \sum_{i=1}^{q} \sum_{j=1}^{p} (\bar{y}_{ij} - \bar{y})^2;$$

$$Q_0 = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{m=1}^{n} (y_{ijm} - \bar{y}_{ij})^2,$$

где $\bar{y} = \frac{1}{qpn} \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{m=1}^{n} y_{ijm}$; $\quad \bar{y}_{ij} = \frac{1}{n} \sum_{m=1}^{n} y_{ijm}$;

$$y_i = \frac{1}{pn} \sum_{j=1}^{p} \sum_{m=1}^{n} y_{ijm}.$$

Величина Q_A отражает влияние фактора A, Q_B — влияние фактора B, Q_0 характеризует изменчивость внутри ячеек. Средние квадраты отыскиваются по формулам

$$M_A = Q_A/(q-1); \quad M_B = Q_B/q(p-1);$$
$$M_0 = Q_0/qp(n-1).$$

Статистики $F' = M_A/M_B$ и $F'' = M_B/M_0$ имеют в условиях справедливости соответствующих нулевых гипотез F-распределение со степенями свободы f_1 и f_2; для F' — $f_1' = q-1, f_2' = q(p-1)$, и для F'' — $f_1'' = q(p-1)$ и $f_2'' = qp(n-1)$.

Рассмотрим вопрос о влиянии нарушений основных предположений дисперсионного анализа на статистические решения. В реальных ситуациях нередко наблюдается невыполнение требований нормальности ошибок, некоррелированности результатов наблюдений и равенства дисперсий. Если нарушения значительны, то статистические решения, принимаемые на основе дисперсионного анализа, могут оказаться ошибочными. В связи с этим перед проведением дисперсионного анализа рекомендуется проверить соответствие исходных данных указанным выше требованиям и при необходимости выполнить такое их преобразование (например, логарифмирование), которое устраняет ненормальность, а также стабилизирует дисперсии. Следует учитывать, что нарушение нормальности мало сказывается при работе с моделью I. Кроме того, опасность ошибочных выводов, возникающих из-за неравенства дисперсий, уменьшается, если дисперсионный анализ опирается на исходную матрицу с равными числами наблюдений в ячейках.

Труднее всего устраняется влияние стохастической зависимости наблюдений. В связи с этим при планировании экспериментов, подлежащих обработке дисперсионным анализом, особое внимание должно быть обращено на методы рандомизации. Если устранить нарушение основных предположений дисперсионного анализа не удается, то рекомендуется обращение к непараметрическому дисперсионному анализу.

МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Латинские квадраты. В геологии естественная классификация проводится по трем переменным, или по трем осям координат.

Одним примером трехфакторного дисперсионного анализа является классификация, с помощью которой сравнивают между собой различные методы измерения параметров минералов; другим примером — выявление осевой, продольной и поперечной зональности геохимических ореолов и связанных с ними зон концентрированной минерализации.

По мере возрастания сложности классификации (по четырем и более признакам) количество наблюдений растет очень быстро, так что экспериментатор вынужден игнорировать некоторые при-

чины изменчивости и использовать блоки с целью уменьшения количества наблюдений.

Изобретались сложные приемы, дающие возможность сократить размеры блоков путем использования неполных повторений и ликвидации некоторых эффектов, которые можно считать несущественными. Применяются различные планы эксперимента, включая такие, как метод рандомизированных полных сбалансированных блоков и сбалансированных блоков, латинский квадрат и прямоугольные решетки [1, 15, 22].

Проведение эксперимента латинским квадратом — одна из форм трехфакторного анализа, использующая объединение факторов для уменьшения количества наблюдений. Структура полной классификации по трем признакам:

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{ik} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl},$$

где μ — генеральное среднее; α_i — эффект i-й градации первого фактора; β_j — эффект j-й градации второго фактора; γ_k — эффект k-й градации третьего фактора; β_{ij}, γ_{ik}, γ_{jk} — эффекты парных взаимодействий; γ_{ijk} — эффект тройного взаимодействия; ε_{ijkl} — ошибка; α, β — коэффициенты.

Применение метода латинских квадратов (в частности, в сельском хозяйстве) показало, что результат взаимодействий можно не принимать во внимание. В геологической практике метод латинских квадратов еще не прошел тщательной проверки, и потому нельзя с уверенностью утверждать, что предположение об отсутствии эффектов взаимодействия правомерно.

Впервые в геологии метод латинских квадратов был применен для проверки изменчивости показаний операторов при точечных подсчетах оптических свойств минералов в шлифах осадочных пород. Использование латинского квадрата предшествует трехфакторному анализу. Обычно с его помощью делается попытка доказать однородность данных, т. е. определить барьер, отделяющий те изменения, которые можно считать относительно небольшими, от тех, которые следует рассматривать как значительные.

Пока при решении геологических задач схема латинского квадрата имеет ограниченное применение.

В последнее время сфера применения дисперсионного анализа в геологии расширяется, например, известны примеры его использования при оценке точности геохимического и разведочного опробования.

Схема латинского квадрата является неполным трехфакторным анализом, в котором все три фактора имеют одно и то же число уровней, а наблюдения проводятся только в m^2 из m^3 возможных совокупностей условий, которые выбираются по следующей схеме: в квадратной матрице порядка $m \times m$ каждое из чисел $1, 2, \ldots, m$ появляется один раз в каждой строке и в каждом столбце.

Основная модель латинских квадратов записывается в виде

$$y_{ijk} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + l_{ijk},$$

где y_{ijk} — случайные величины; l_{ijk} — независимые случайные величины, имеющие математическое ожидание нуль и дисперсию σ^2; для главных эффектов α_i^A, α_j^B и α_k^C выполнены условия:

$$\Sigma\alpha_i^A = \Sigma\alpha_j^B = \Sigma\alpha_k^C = 0.$$

Здесь эффект от действия всех неучтенных источников изменчивости определяет все вклады в дисперсию от трех взаимодействий первого порядка и взаимодействия второго порядка.

Проверяемые гипотезы:

$$H_A: \quad \text{все} \quad \alpha_i^A = 0,$$
$$H_B: \quad \text{все} \quad \alpha_j^B = 0,$$
$$H_C: \quad \text{все} \quad \alpha_k^C = 0.$$

Греко-латинские квадраты. (удобная модель изучения геологических объектов при желании устранить влияние трех мешающих факторов) образуются путем наложения одного латинского квадрата на другой таким образом, чтобы каждая комбинация из двух букв, первая из которых соответствует первому квадрату, а вторая — второму, встречалась бы ровно по одному разу. При выполнении этого условия исходные латинские квадраты называются ортогональными.

Латинские буквы одного квадрата заменяются на греческие. Например, если наложить латинский квадрат

$$\begin{matrix} A & B & C & D \\ C & D & A & B \\ D & C & B & A \\ B & A & D & C \end{matrix}$$

на латинский квадрат

$$\begin{matrix} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A, \end{matrix}$$

заменив в первом квадрате латинские буквы A, B, C, D соответственно на греческие α, β, γ, δ, получим греко-латинский квадрат

$$\begin{matrix} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma. \end{matrix}$$

При числе обработок $t = 2$ и $t = 6$ греко-латинских квадратов не существует, для всех остальных значений t греко-латинские квадраты всегда существуют [6]. Наличие примеров греко-латин-

ских квадратов для различного числа обработок можно найти в работе [6].

При дисперсионном анализе строки и столбцы греко-латинского квадрата идентифицируют первый и второй мешающие факторы, а греческие буквы — третий. Латинские буквы соответствуют обработкам, так что план устраняет влияние трех мешающих факторов.

Гипер-греко-латинские квадраты (удобная модель изучения геологических объектов при желании устранить влияние четырех мешающих факторов). Они образуются путем наложения на греко-латинский квадрат латинского квадрата таким образом, чтобы каждая комбинация из двух множеств символов встречалась бы ровно один раз. Три таких исходных латинских квадрата называются ортогональными в совокупности. Используя в качестве третьего множества символов числа, приводим пример гипер-греко-латинского квадрата:

$A\alpha 1$	$B\beta 2$	$C\gamma 3$	$D\delta 4$
$B\gamma 4$	$A\delta 3$	$D\alpha 2$	$C\beta 1$
$C\delta 2$	$D\gamma 1$	$A\beta 4$	$B\alpha 3$
$D\beta 3$	$C\alpha 4$	$B\delta 1$	$A\gamma 2$.

На этот квадрат уже нельзя наложить ни одного латинского квадрата с сохранением условия ортогональности. Существует не более t—1 взаимно ортогональных латинских квадратов порядка t [6]. Множество из t—1 таких квадратов называется полным множеством ортогональных латинских квадратов. Такие полные множества существуют для простого числа t или для степени простого числа [6]. О наличии примеров полных множеств для $t \leqslant 9$ сообщается в [6].

Важным условием практического использования схемы латинских квадратов является требование совпадения числа обработок с числом уровней каждого из мешающих факторов.

Схема латинских квадратов применима, например, при изучении коэффициентов зональности первичных геохимических ореолов, когда нужно устранить влияние двух географических координат или географической и временной координат.

НЕПАРАМЕТРИЧЕСКИЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ непараметрический — группа статистических методов, позволяющих оценить влияние различных одновременно действующих факторов на случайную величину с непрерывным распределением. Никаких других предположений относительно распределения исследуемой случайной величины не формулируется. В этом случае устойчивость процедур дисперсионного анализа обеспечивается переходом от значений случайной величины к их рангам (Краскла—Уэллиса метод) или соответствующим нормальным меткам (Пури и Сена метод). Другой подход опирается на так называемые медианные критерии (Брауна и Муда метод).

Однофакторный ранговый дисперсионный анализ Краскла-Уэллиса [19]. Исследуется влияние фактора A на случайную величину, выборочные значения которой в соответствии с градациями A образуют p выборок каждая объемом n_i:

$$Y = \{Y_1, Y_2, \ldots, Y_i, \ldots, Y_p\},$$

$$Y = \{y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{in_i}\}.$$

Проверяемая нулевая гипотеза утверждает, что выборочные средние, вычисленные по p выборкам, незначимо отличаются друг от друга. Если H_0-гипотеза отклоняется, то принимается решение о существенном влиянии фактора A на случайную величину ξ.

Процедура проверки гипотезы H_0 сводится к следующему. Все наблюдения $\{y_{ij} : i = 1, 2, \ldots, p : j = 1, 2, \ldots, n_i\}$ объединяются в одну общую выборку объемом $N = \sum_{i=1}^{p} n_i$. Производится ранжирование элементов общей выборки, при этом минимальному значению y_{ij} присваивается ранг 1, следующему по величине — ранг 2 и т. д. до исчерпания выборки. Максимальный элемент будет иметь, таким образом, ранг N.

Вычисляется статистика

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{p} \frac{R_i^2}{n_i} - 3(N+1),$$

где R_i — сумма рангов в i-й выборке (i-я градация фактора A).

В условиях H_0 величина H имеет χ^2-распределение с $(p-1)$ степенями свободы. Дополнительное условие: $n_i > 5$ при любом i.

Однофакторный дисперсионный анализ, опирающийся на медианный критерий Брауна и Муда [19]. Как и в методе Краскла-Уэллиса p выборок объединяются в одну. Отыскивается медианное значение y и строится табл. 7.

Таблица 7

Градации фактора и число наблюдений

Градации фактора A		1	2	...	i	...	p	Σ
Число наблюдений	$\geqslant \tilde{y}$	m_1	m_2	...	m_i	...	m_p	$N/2$
	$< \tilde{y}$	$n_1 - m_1$	$n_2 - m_2$	...	$n_i - m_i$	...	$n_p - m_p$	$N/2$
Общее число наблюдений в группах		n_1	n_2	...	n_i	...	n_p	N

Если влияние фактора A несущественно, то можно ожидать, что все p группы (выборок) будут иметь одну и ту же медиану, т. е. $m_i = n_i/2$ для каждого i.

Сформулированную таким образом гипотезу об однородности проверяют с помощью статистики

$$X^2 = \sum_{i=1}^{p} \frac{(m_i - n_i/2)^2}{n_i/4},$$

которая асимптотически распределена как χ^2 с $(p-1)$ степенями свободы. Действие фактора A считается несущественным (при уровне значимости α), если $\chi^2 < \chi^2_{\alpha,\, p-1}$.

ГЛАВА 8

СЛУЧАЙНЫЕ ПРОЦЕССЫ

Случайный процесс — функция $\xi(t)$ на конечном или бесконечном интервале изменения t, значения которой в каждой точке являются случайными величинами. Дать формальное определение случайного процесса, сочетающее в себе физическую сущность и математическую строгость, очень трудно, но интуитивное представление о случайном процессе связано с непредсказуемостью его мгновенных значений. Математически строгое определение требует введения понятия ансамбля, т. е. бесконечной совокупности реализаций. С физической точки зрения, вполне допустимо представление случайного процесса одной реализацией. С математической точки зрения, отдельная реализация — детерминированная функция времени, с помощью которой определить статистические свойства процесса можно лишь при выполнении определенных условий (рис. 40, 41, 42).

Свойства случайных процессов описываются вероятностными характеристиками, такими как распределение вероятностей, кор-

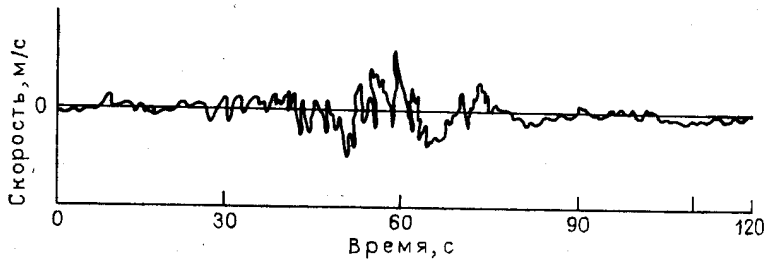


Рис. 40. Пример реализации нестандартного процесса

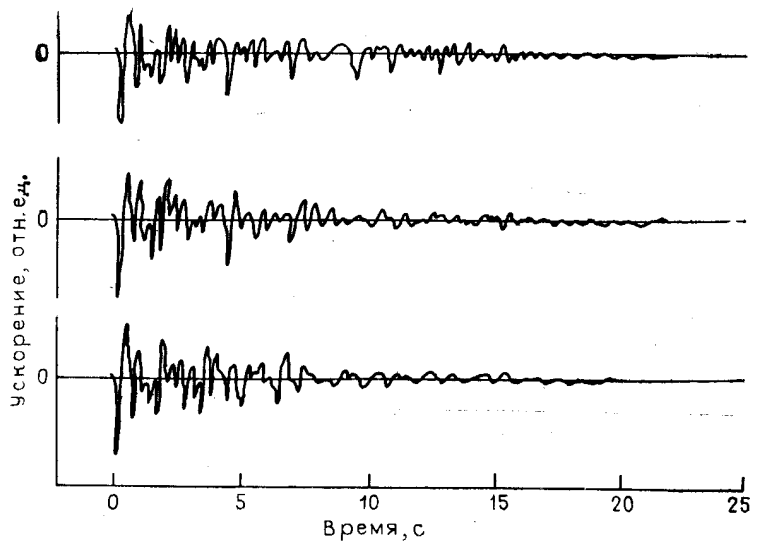


Рис. 41. Пример ансамбля реализаций переходного процесса

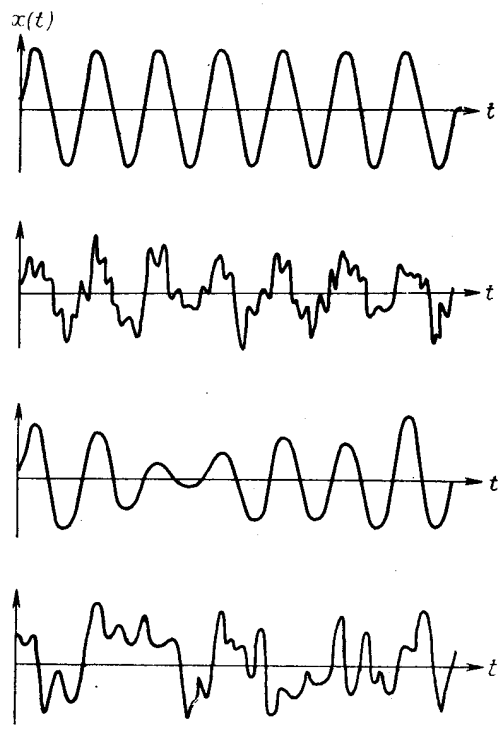


Рис. 42. Реализация случайных процессов

реляционная функция, спектральная функция, интервал корреляции, эффективная ширина спектра и т. д. (рис. 43, 44, 45).

Отнесение процесса к классу случайных и использование для описания его свойств вероятностных характеристик может быть обусловлено либо его физической природой, либо условиями изучения, приводящими к недостаточности данных.

К первой группе относятся процессы, являющиеся результатом суперпозиции большого числа элементарных процессов. Ко второй группе относятся процессы, природа которых позволяет описать их с любой степенью надежности. К классу случайных они относятся в том случае, если объем информации, которым располагает исследователь, недостаточен для определения всех его характеристик.

Наиболее распространено описание случайных процессов с помощью функций распределения вероятностей, корреляционных и спектральных функций. Описание случайного процесса на основе функций распределения состоит в следующем.

Случайный процесс $\xi(t)$ считается заданным, если для любого набора $t_1, t_2, \ldots, t_n$ на интервале изменения t при любом n задана функция распределения

$$F_n(x_1, \ldots, x_n) = p(\xi(t_1) \leqslant x_1, \ldots, \xi(t_n) \leqslant x_n),$$

где p — вероятность.

Функция распределения $F_n(x_1, \ldots, x_n)$ должна удовлетворять следующим условиям:

а) условию симметрии: для любой перестановки $i_1, \ldots, i_n$ чисел $1, 2, \ldots, n$ должно выполняться равенство

$$F_n(\xi(t_{i_1}), \xi(t_{i_2}), \ldots, \xi(t_{i_n})) = F_n(\xi(t_1), \xi(t_2), \ldots, \xi(t_n));$$

б) условию согласования: если $m < n$, то при любых $t_{m+1}, t_{m+2}, \ldots, t_n$

$$F_n(\xi(t_1), \ldots, \xi(t_m), \infty, \ldots, \infty) = F_m(\xi(t_1), \ldots, \xi(t_m)).$$

Случайные процессы, изучаемые в прикладных науках, бывают стационарными, нестационарными, эргодическими, неэргодическими, гауссовскими, винеровскими, со стационарными приращениями, марковскими и др.

Аппарат теории случайных процессов используется при описании геохимических полей, при прогнозировании геолого-геофизических характеристик, интерполяции геологических характеристик, в тренд-анализе, геостатистике, при подсчете запасов, в гидрогеологии, при исследовании тектонических явлений.

Примеры реализаций случайных функций в геологии: изучение зональности геохимических ореолов, выявление ореолов развития метасоматически измененных пород, пространственная изменчивость содержаний рудных компонент, линейных запасов в залежах, рудных участках месторождений полезных ископаемых, распределение прогнозных ресурсов в регионах.

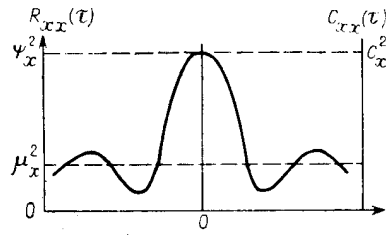


Рис. 43. Ковариационная и корреляционная функции

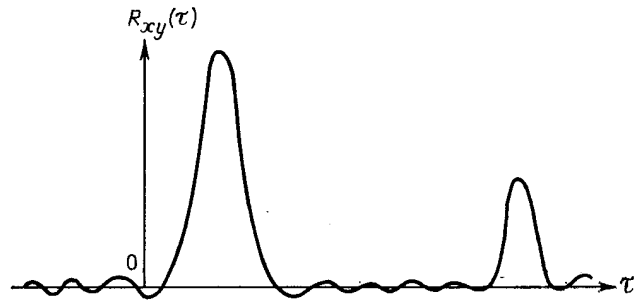


Рис. 44. Типичная ковариационная функция

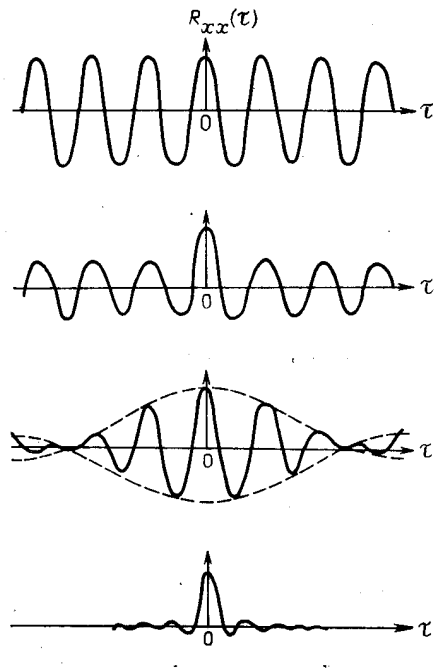


Рис. 45. Идеализированные ковариационные функции. Пунктиром показана огибающая сигнала

С помощью аппарата стационарных случайных процессов можно исследовать неоднородность внутреннего строения природных скоплений полезных ископаемых и проводить количественное описание наблюдаемой изменчивости их других важнейших свойств с выделением элементов неоднородности различных порядков в зависимости от геометрического расположения проб в разведочной сети. При этом очень важно выявление предельных расстояний зависимостей между изучаемыми параметрами (радиуса корреляции), так как только в его пределах правомерна интерполяция и экстраполяция геологоразведочных данных. Некоррелированные значения геологоразведочных параметров должны рассматриваться как совокупность независимых случайных величин.

Корреляционные функции, вычисленные по взаимно перпендикулярным разведочным сечениям, могут служить основой для расчета показателей анизотропии изучаемого геологического свойства, включая важнейшие кондиции.

Характеристики стационарных случайных функций можно использовать при интерполировании значений параметров в промежуточных интервалах между точками наблюдений, для выбора надежной геометрии проб и для решения других геологоразведочных задач.

Приведем примеры случайных процессов.

1. $\xi(t) = \xi f(t)$, где ξ — случайная величина; $f(t)$ — числовая функция аргумента t, $-\infty < t < \infty$.

2. $\xi(t) = \xi e^{i\lambda kt}$, где ξ — случайная величина; λ — вещественная постоянная $i = \sqrt{-1}$. Этот случайный процесс описывает периодическое колебание круговой частоты λ со случайной амплитудой и случайной фазой.

3. Можно рассмотреть также суперпозицию n случайных периодических колебаний с различными частотами:

$$\xi(t) = \sum_{k=1}^{n} \xi_k e^{i\lambda kt},$$

где $\xi_1, \ldots, \xi_n$ — случайные величины с нулевыми математическими ожиданиями.

Случайный процесс такого вида можно назвать процессом с дискретным спектром. Спектром этого процесса является совокупность чисел $\lambda_1, \lambda_2, \ldots, \lambda_n$.

ХАРАКТЕРИСТИКИ СЛУЧАЙНОГО ПРОЦЕССА

Моменты случайного процесса — вероятностные характеристики случайного процесса, представляющие собой интегралы от степеней переменных интегрирования с плотностью, определяемой соответствующей конечно-мерной функцией распределения.

Момент первого порядка — функция, которая является средним значением случайного процесса $\xi(t)$ и определяется по формуле

$$\mu_1(t) = M[\xi(t)] = \int_{-\infty}^{+\infty} x dF_t(x).$$

Второй момент — средний квадрат случайного процесса $\{\xi(t)\}$ — определяется по формуле:

$$\mu_2(t) = M[\xi^2(t)] = \int_{-\infty}^{+\infty} x^2 dF_t(x).$$

Второй и высшие моменты, вычисленные относительно среднего значения, называются центральными моментами. Так, второй центральный момент

$$\mu_2^c(t) = M[\xi(t) - M_1(t)]^2 = \int_{-\infty}^{+\infty} [x - \mu_1(t)]^2 dF_t(x)$$

называется дисперсией случайного процесса.

Корреляционная функция случайного процесса $\xi(t)$ — его момент второго порядка:

$$\mu_{11}(t, s) = M[\xi(t)\xi(s)] = \sigma(t, s).$$

Если $\xi(t)$ — стационарный случайный процесс, то $\mu_{11}(t, s) = \sigma(t-s)$. Раздел теории стационарных случайных процессов, рассматривающий лишь те его свойства, которые определяются моментами первых двух порядков, называется корреляционной теорией случайных процессов.

Многие свойства реализаций случайных процессов хорошо описываются первыми двумя моментами, характеризующими положение центра процесса и меру рассеяния его значений. Эти характеристики можно оценить путем усреднения реализаций случайного процесса.

ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНОГО ПРОЦЕССА

Выборочные характеристики стационарного случайного процесса — это статистические оценки его характеристик по наблюдаемым значениям.

Пусть $\xi_1, \ldots, \xi_T$ — последовательность наблюдений стационарного случайного процесса $\xi(t)$. Среднее значение и ковариационная последовательность этого случайного процесса:

$$M\xi(t) = \mu, \quad t = 0, 1, 2, \ldots,$$

$$\sigma(h) = \operatorname{cov}(\xi_t, \xi_{t+h}) = M[\xi(t) - \mu][\xi(t+h) - \mu],$$

$$h = 0, \pm 1, \pm 2, \ldots \quad t = 0, 1, 2, \ldots,$$

причем $\sigma(-h) = \sigma(h)$. Спектральное разложение $f(\lambda)$ связано с функцией $\sigma(h)$ следующим образом:

$$\sigma(h) = \int_{-\pi}^{\pi} e^{i\lambda h} f(\lambda)\, d\lambda, \quad h = 0, \pm 1, \pm 2, \ldots,$$

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \sigma^2(h) e^{i\lambda h},$$

где $f(\lambda)$ — спектральная функция [4].

Несмещенной оценкой для μ является

$$\bar{\xi} = \frac{1}{T} \sum_{t=1}^{T} \xi_t.$$

Если μ известно, то несмещенная оценка для $\sigma(h)$

$$C_h = C_{-h} = \frac{1}{T-h} \sum_{t=1}^{T-h} (\xi_t - \mu)(\xi_{t+\mu} - \mu),$$

$$h = 0, 1, \ldots, T-1.$$

Если μ неизвестно, то можно построить следующие оценки:

$$C_h = C_{-h} = \frac{1}{T-h} \sum_{t=1}^{T-h} (\xi_t - \bar{\xi})(\xi_{t+h} - \bar{\xi}), \qquad h = 0, 1, \ldots, T-1,$$

$$C_h = C_{-h} = \frac{1}{T-h} \sum_{t=1}^{T-h} (\xi_t - \bar{\xi}_h)(\xi_{t+h} - \bar{\xi}_h),$$

где $$\bar{\xi}_h = \frac{1}{T-h} \sum_{t=1}^{T-h} \xi_t, \qquad h = 0, 1, \ldots, T-2,$$

$$\bar{\xi} = \frac{1}{T-h} \sum_{t=1}^{T-h} \xi_{t+h}, \qquad h = 0, 1, \ldots, T-2.$$

Выборочная спектральная плотность равна

$$I(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} (\xi_t - \mu) e^{i\lambda t} \right|^2, \qquad -\pi \leqslant \lambda \leqslant \pi,$$

когда μ известно. Если μ неизвестно, то

$$I(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} (\xi_t - \bar{\xi}) e^{i\lambda t} \right|^2, \qquad -\pi \leqslant \lambda \leqslant \pi.$$

Дисперсия выборочного среднего $\bar{\xi}$:

$$D\xi = \frac{1}{T^2} \sum_{t,\, s=1}^{T} \sigma(t-s) =$$

$$= \int_{-\pi}^{\pi} \frac{1}{T} \left(\frac{\sin \frac{\lambda T}{2}}{\sin \frac{\lambda}{2}} \right)^2 f(\lambda)\, d\lambda.$$

Автокорреляция — мера взаимной зависимости между элементами последовательности наблюдений ξ_t, сдвинутых на j единиц (автокорреляция i-го порядка).

Пусть наблюдаемый ряд $\xi_1, \ldots, \xi_T$ является реализацией процесса, имеющего нулевое математическое ожидание. Если математическое ожидание $\xi_1, \ldots, \xi_T$ отлично от нуля, то путем центрирования можно перейти к новой последовательности, удовлетво-

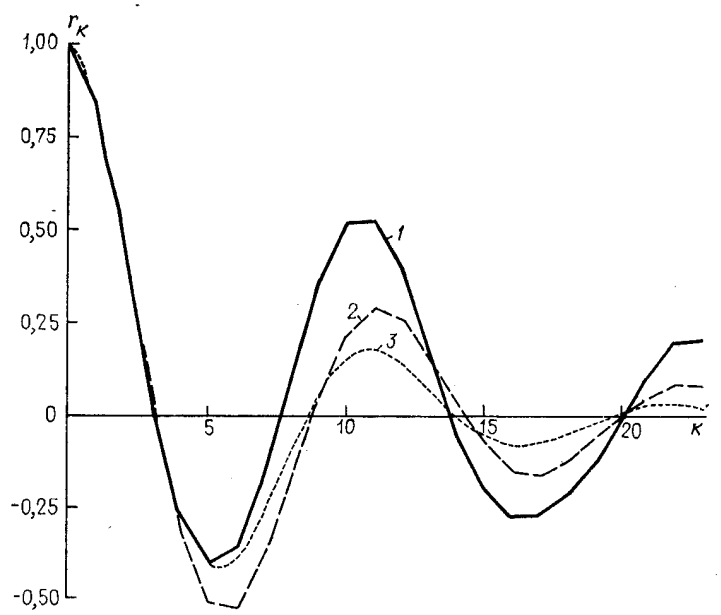


Рис. 46. Наблюденная коррелограмма (*1*) числа солнечных пятен (числа лет наблюдений) по сравнению с теоретической коррелограммой модели авторегрессии второго порядка для дискретного (*2*) и непрерывного (*3*) времени

ряющей этому условию. Тогда сериальная корреляция j-го порядка вычисляется по формуле

$$r_j = \frac{\sum_{t=j+1}^{T} \xi_t \xi_{t-j}}{\sum_{t=1}^{T} \xi_t^2}.$$

Реализация ξ_t заменяется отклонением от выборочного среднего также, если средние значения неизвестны.

Автокорреляция может использоваться для проверки нулевой гипотезы о независимости против альтернативы, утверждающей существование зависимости между наблюдениями, сдвинутыми на j единиц времени.

Набор коэффициентов $r_1, r_2, \ldots$, нанесенный на график с k в качестве абсциссы и r_k в качестве ординаты, называется коррелограммой. Коррелограмма — полезный инструмент анализа временны́х рядов, позволяющий различать их типы (рис. 46) [4].

ТИПЫ СЛУЧАЙНЫХ ПРОЦЕССОВ

Стационарный случайный процесс — случайный процесс $\xi(t)$, распределение вероятностей которого $F_{t_1, t_2, \ldots, t_n}(x_1, x_2, \ldots, x_n)$ для любого множества моментов времени

$t_1, \ldots, t_n$ зависит только от разностей $t_r - t_s$ и не зависит от абсолютных значений t_r. Другими словами, он инвариантен по отношению к сдвигу по оси t.

Из этого определения вытекает, что для стационарного случайного процесса все одномерные функции распределения не зависят от t, а все двумерные функции распределения зависят только от разности аргументов $t_1 - t_2$ и т. д.

Первый момент стационарного случайного процесса $\mu_1(t)$ постоянен, т. е. не зависит от t. Центральный момент стационарного случайного процесса порядка (1.1) — ковариационная функция

$$\mu_2(t_1, t_2) = M[\xi(t_1) - \mu_1][\xi(t_2) - \mu_1] = \sigma(t_1, t_2)$$

зависит от разности аргументов, т. е.

$$\sigma(t_1, t_2) = \sigma(t_1 - t_2).$$

Часто многие из свойств случайного процесса определяются его первыми и вторыми моментами, даже если процесс не является стационарным. Если при этом первые и вторые моменты инвариантны по отношению к сдвигу по времени t, то процесс называется стационарным в широком смысле. Аналогично можно определить стационарный процесс r-го порядка.

Эргодичность — свойство стационарных случайных процессов, позволяющее вычислять его момент первого порядка и корреляционную функцию только по одной реализации:

$$\mu_1 = M\xi(t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t)\, dt,$$

$$\sigma(\tau) = M[\xi(t+\tau)\xi(t)] = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t+\tau)\xi(t)\, dt,$$

где интегралы справа понимаются как пределы соответствующих интегральных сумм, а $(0, T)$ — интервал изменения параметра t.

Значение эргодической теоремы вытекает из того, что она позволяет найти μ и $\sigma(\tau)$ по одной реализации стационарного процесса. Корреляционная теория стационарных случайных процессов позволяет по μ_1 и $\sigma(\tau)$ получить довольно обширную информацию о случайном процессе.

Гауссовский случайный процесс — случайный процесс $\xi(t)$, у которого все совместные распределения для любой конечной совокупности значений $t_1, \ldots, t_k$ являются нормальными k-мерными распределениями. Пусть $\mu(t)$ — математическое ожидание случайного процесса $\xi(t)$, т. е. $\mu(t) = M(\xi(t))$. Пусть также $r(S, t)$ — его второй центральный момент:

$$M(x(s)\overline{x(t)} - \mu(s)\overline{\mu(t)}) = r(s, t).$$

Очевидно, $r(t, s) = r(s, t)$. Кроме того, для любого конечного набора $t_1, \ldots, t_N$ матрица $R = \|r(t_m, t_n)\|_{1\leqslant m,\, n\leqslant N}$ неотрицательно определена. Плотность N-мерного гауссовского распределения при заданных $t_1, \ldots, t_N$, функциях $\mu(t_1), \ldots, \mu(t_N)$ и матрице R определяется по формуле

$$\frac{\det\|a_{mn}\|}{\sqrt{(2\pi)^{\frac{N}{2}}}} \exp\left\{-\frac{1}{2}\sum_{m,\,n=1}^{N} a_{mn}(x_m - \mu(t_m))[(x_n - \mu(t_n)]\right\},$$

где $\|a_{mn}\|_{1\leqslant m,\, n\leqslant N}$ — матрица, обратная к матрице R, а $\det\|a_{mn}\|$ ее детерминант. Все условия согласования N-мерных распределений для гауссовского случайного процесса выполняются.

Винеровский случайный процесс — случайный процесс $\xi(t)$ с математическим ожиданием, равным нулю $[\xi(0) = 0]$, и независимыми приращениями, распределенными по нормальному закону с нулевым средним и дисперсиями, равными разности аргументов. Так, для винеровского процесса $\xi(t)$ вероятность выполнения неравенств $a < \xi(t) \leqslant b$ для некоторых заданных конечных значений a и b составляет

$$\frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2}}\, dx.$$

Вероятность выполнения неравенств

$$a_1 < \xi(t_1) \leqslant b_1,$$
$$a_2 < \xi(t_2) \leqslant b_2,$$
$$\ldots\ldots\ldots\ldots$$
$$a_n < \xi(t_n) \leqslant b_n$$

для $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$ и некоторых заданных конечных a_i, b_i, $i = 1, 2, \ldots, n$ (это множество называется n-мерным параллелепипедом) равна

$$\frac{1}{\sqrt{(2\pi)^n t_1 (t_2 - t_1)\ldots(t_n - t_{n-1})}} \int_{a_1}^{b_1} dz_1 \int_{a_2}^{b_2} dz_2 \ldots \int_{a_n}^{b_n} dz_n \times$$
$$\times e^{-\frac{z_1^2}{2t_1} - \frac{(z_2 - z_1)^2}{2(t_2 - t_1)} - \ldots - \frac{(z_n - z_{n-1})^2}{2(t_n - t_{n-1})}}.$$

Задание вероятностной меры на таких n-мерных параллелепипедах позволяет определить меру на любом измеримом множестве.

Случайный процесс $\xi(t)$ называется процессом со стационарными приращениями, если математическое ожидание приращения этого процесса за какой-либо промежуток времени пропорционально длине этого промежутка:

$$M[\xi(s) - \xi(t)] = a(s - t),$$

а

$$M[\xi(u) - \xi(t)]\overline{[\xi(v) - \xi(t)]} = f(u - t, v - t).$$

Обычно рассматривают процесс $\xi(t) - M\xi(t)$, так что можно считать $a = 0$, а f — функция, которая в вещественном случае становится функцией одной переменной.

Марковский случайный процесс — случайный процесс $\xi(t)$ (t принадлежит некоторому интервалу T, а $\xi(t)$ — некоторому пространству X), при котором для любого конечного набора $t_1 < \ldots < t_n$ значений параметра t условное распределение вероятностей величины $\xi(t_n)$ относительно $\xi(t_1), \ldots, \xi(t_{n-1})$ совпадает с условным распределением вероятностей величины $\xi(t_n)$ относительно величины $\xi(t_{n-1})$.

Многие природные процессы, которые рассматриваются как случайные, характеризуются тем, что в них вероятность находиться в данном состоянии в заданный момент времени определяется непосредственно предшествующим состоянием. Таковы, например, диффузионные процессы, процессы осадконакопления. Марковскую теорию для описания геологических процессов успешно применял А. Б. Вистелиус [12].

Примером дискретного марковского процесса является марковская цепь. В случае, когда X — совокупность чисел $1, 2, \ldots, k$, марковский процесс превращается в цепь Маркова с k состояниями.

Цепь Маркова — последовательность испытаний (в каждом из которых может осуществиться одно из k несовместимых событий $A_1, \ldots, A_k$), обладающая тем свойством, что условная вероятность в $(s+1)$-м испытании $s = 1, 2, 3, \ldots$ осуществиться событию A_i^{s+1} $(i = 1, 2, \ldots, k)$ после того, как в s-м испытании произошло известное нам событие, зависит от того, каким было событие, происшедшее в s-м испытании, и не изменяется от добавочных сведений о том, какие события происходили в более ранних испытаниях.

Цепь Маркова — одна из форм марковского процесса. В общем виде цепь Маркова — это серия переходов между различными состояниями, когда вероятности каждого перехода зависят только от непосредственно предшествующих состояний и не зависят от более ранних предшествующих состояний. Цепи Маркова, вероятности перехода которых зависят лишь от непосредственно предшествующего состояния, называются цепями первого порядка. Если вероятности перехода определяются двумя или более предшествующими состояниями, то соответствующие цепи Маркова называются цепями второго или n-го порядка (если учитывается n предшествующих состояний).

Цепи Маркова с успехом используются в стратиграфии при моделировании стратиграфических последовательностей, при изучении процессов диффузии, осадконакопления, при построении моделей процессов кристаллизации гранитоидов. Вычисление вероятностей перехода при изучении последовательностей геологических данных описано в работе [45].

Однородная цепь Маркова — это такая цепь, в которой условная вероятность появления события A_j^{s+1} в $(s+1)$-м испытании при условии, что в s-м испытании осуществилось событие $A_i^{(s)}$, не зависит от номера испытания. Эта вероятность называется вероятностью перехода из состояния с номером i в состояние с номером j

и обозначается p_{ij}. Полная вероятностная картина возможных изменений, осуществляющихся при переходе от одного испытания к следующему, задается матрицей

$$\pi = \begin{pmatrix} p_{11} & p_{12} & \ldots & p_{1k} \\ p_{21} & p_{22} & \ldots & p_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ p_{k1} & p_{k2} & \ldots & p_{kk} \end{pmatrix};$$

которая называется матрицей перехода.

Очевидны следующие свойства этой матрицы:

$$0 \leqslant p_{ij} \leqslant 1,$$

$$\sum_{j=1}^{k} p_{ij} = 1, \qquad i = 1, 2, \ldots, k.$$

Вероятность перехода из состояния $A_i^{(s)}$ в s-м испытании в состояние $A_j^{(s+n)}$ через n испытаний обозначается $p_{ij}(n)$:

$$p_{ij}(n) = \sum_{m=1}^{n} p_{ir}(m)\, p_{rj}(n-m).$$

Обозначим через π_n матрицу перехода через n испытаний:

$$\pi_n = \begin{pmatrix} p_{11}(n) & p_{12}(n) & \ldots & p_{1k}(n) \\ \ldots & \ldots & \ldots & \ldots \\ p_{k1}(n) & p_{k2}(n) & \ldots & p_{kk}(n) \end{pmatrix}.$$

Между матрицами π_n существует следующее соотношение:

$$\pi_n = \pi_m \pi_{n-m}, \qquad 0 < m < n.$$

Если при некотором начальном состоянии s все элементы матрицы π_s положительны, то существуют пределы:

$$\lim_{n\to\infty} p_{ij}(n) = p_j, \qquad j = 1, 2, \ldots, k.$$

Таким образом, на числа p_j можно смотреть, как на вероятность появления события $A_j^{(n)}$ при n-м испытании, когда n велико.

Смысл этого состоит в следующем: вероятность находиться системе в состоянии A_j практически не зависит от того, в каком состоянии она находилась в далеком прошлом.

Проверка марковского свойства производится, например, с помощью критерия Андерсона—Гудмана [45]. Пусть $\| p_{ij} \|_{1 \leqslant i,\, j \leqslant k}$ — матрица переходных вероятностей:

$$p_j = \frac{1}{\sum_{i,\, j=1}^{k} n_{ij}} \sum_{i=1}^{k} n_{ij},$$

где n_{ij} — частота переходов для i-й строки и j-го столбца; n — общее число состояний. Нулевая гипотеза состоит в том, что события,

образующие последовательность, независимы, а множество альтернатив — в том, что они зависимы. Например, этому может соответствовать марковская цепь первого порядка. Для проверки нулевой гипотезы вычисляется величина

$$\lambda = \sum_{i,\,j=1}^{k} \left(\frac{p_j}{p_{ij}}\right)^{n_{ij}}.$$

Если проверяемая гипотеза верна, то величина $2 \ln \lambda$ распределена асимптотически как χ^2 с $(k-1)^2$ степенями свободы.

Если для заданного уровня значимости α и заданного числа степеней свободы ν величина $-2 \ln \lambda > \chi^2_{\alpha\nu}$, то принимается нулевая гипотеза. В противном случае считается, что исходные данные не согласуются с предположением о независимости событий, и принимается альтернатива.

Если переходные вероятности p_{ij} зависят от t, т. е. фактически мы имеем дело с марковским случайным процессом, то он, как и всякий случайный процесс, может быть стационарным или нет.

Проверка стационарного свойства производится на основании следующего критерия Андерсона и Гудмана [45].

В стационарной марковской цепи переходная вероятность p_{ij} — это постоянная вероятность перехода из состояния i в момент времени $t-1$ в состояние j в момент времени t. В нестационарной марковской цепи $p_{ij}(t)$ — функция времени. Проверяется нулевая гипотеза:

$$H_0 : p_{ij}(t) = p_{ij} \qquad \text{для всех } t = 1, 2, \ldots, T.$$

Множество альтернатив состоит в утверждении:

$$H_1 : p_{ij}(t) \neq p_{ij}.$$

Критерий имеет вид:

$$\lambda = \prod_{t=1}^{T} \prod_{i,\,j=1}^{k} \left[\frac{p_{ij}}{p_{ij}(t)}\right]^{n_{ij}(t)},$$

где $n_{ij}(t)$ — наблюдаемая частота перехода из состояния i в состояние j. Если проверяемая гипотеза о стационарности верна, то величина $-2 \ln \lambda$ будет распределена как χ^2 с $(T-1)\,k\,(k-1)$ степенями свободы.

Если при заданном уровне значимости α и заданном числе степеней свободы ν величина $-2 \ln \lambda > \chi^2_{\alpha,\nu}$, то нулевая гипотеза принимается. В противном случае принимается альтернатива.

В геологии аппарат теории марковских цепей использовался А. Б. Вистелиусом [12], М. А. Романовой, В. Шварцахером, Х. Тиргертнером и другими для описания моделей процессов кристаллизации гранитоидов и процессов осадконакопления.

СПЕКТРАЛЬНОЕ РАЗЛОЖЕНИЕ СЛУЧАЙНОГО ПРОЦЕССА

Спектральное разложение стационарного случайного процесса — это представление стационарного процесса $\xi(t)$ в виде интеграла Фурье—Стилтьеса:

$$\xi(t) = \int_{-\infty}^{+\infty} e^{i\lambda t} dz(\lambda),$$

где $i = \sqrt{-1}$; $z(\lambda)^{-\infty}$ — случайная функция точки λ, удовлетворяющая условиям:

$$Mz(\lambda) = 0 \quad \text{для всех} \quad \lambda,$$

$$M\,[z(\lambda_1 + \Delta\lambda_1) - z(\lambda_1)]\,[z(\lambda_2 + \Delta\lambda_2) - z(\lambda_2)] = 0,$$

если интервалы $(\lambda_1, \lambda_1 + \Delta\lambda_1)$ и $(\lambda_2, \lambda_2 + \Delta\lambda_2)$ не пересекаются.

Интеграл в правой части этого равенства понимается как предел:

$$\lim_{T\to\infty} \left\{ \lim_{\max |\lambda_k - \lambda_{k-1}| \to 0} \sum_{k=1}^{n} e^{i\lambda_k t}\,[z(\lambda_k) - z(\lambda_{k-1})] \right\},$$

где $\lambda_{k-1} \leqslant \lambda_k' \leqslant \lambda_k$; $-T = \lambda_0 < \lambda_1 < \ldots < \lambda_n = T$.

Имеется формула обращения, позволяющая находить функцию $z(\lambda)$, если известен вид $\xi(t)$:

$$z(\lambda) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\lambda t} - 1}{-it} \xi(t)\, dt.$$

Если функция $z(\lambda)$ дифференцируема, то $dz(\lambda) = z'(\lambda)\, d\lambda$ и производная $z'(\lambda) = f(\lambda)$ называется спектральной плотностью случайного процесса.

Если случайный процесс дискретен, то его спектральное разложение имеет вид

$$x(t) = \sum_{k=1}^{\infty} x_k e^{i\lambda_k t},$$

где $Mx_k = 0$; $Mx_k\bar{x}_l = 0$, $k \neq l$. Числа $\lambda_1, \lambda_2, \ldots$ называются спектром этого дискретного случайного процесса.

Спектральное разложение корреляционной функции стационарного случайного процесса — представление корреляционной функции $\sigma(\tau)$ случайного процесса в виде интеграла:

$$\sigma(\tau) = \int_{-\infty}^{+\infty} e^{i\lambda\tau} dF(\lambda),$$

где $dF_{(\lambda)} = F(\lambda + \Delta\lambda) - F(\lambda) = M\,[z(\lambda + \Delta\lambda) - z(\lambda)]^2$,

$F(\lambda)$ — неубывающая функция [свойства функции $z(\lambda)$] [4].

Очевидно,

$$\sigma(0) = M\,|\xi(t)|^2 = \int_{-\infty}^{+\infty} dF(\lambda).$$

Функция $F(\lambda)$ называется спектральной функцией стационарного случайного процесса. Если функция $\sigma(\tau)$ такова, что

$$\int_{-\infty}^{\infty} |\sigma(\tau)|\,d\tau < \infty,$$

то существует функция $f(\lambda) = F'(\lambda)$, называемая спектральной плотностью процесса $x(t)$. Тогда имеют место формулы

$$\sigma(\tau) = \int_{-\infty}^{+\infty} e^{i\lambda\tau} f(\lambda)\,d\lambda,$$

$$f(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-i\lambda T}\sigma(\tau)\,d\tau,$$

т. е. функции $\sigma(\tau)$ и $f(\lambda)$ связаны преобразованием Фурье.

Спектральная плотность неотрицательна, т. е. $f(\lambda) \geqslant 0$.

Пример 1. Пусть $\sigma(\tau) = Ce^{-\alpha|\tau|}\cos\beta\tau$.

Тогда $f(\lambda) = \dfrac{C\alpha}{2\pi}\left[\dfrac{1}{\alpha^2+(\lambda-\beta)^2} + \dfrac{1}{\alpha^2+(\lambda+\beta)^2}\right]$.

Пример 2.

$\sigma(\tau) = Ca^{|\tau|}$, $C>0$, $|a|<1$, $\tau = 0, \pm 1, \pm 2, \ldots$

Имеем

$$f(\lambda) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{+\infty}\sigma(\tau)e^{-i\tau\lambda} = \frac{C}{2\pi}\left\{\sum_{k=-\infty}^{-1} a^{-k}e^{-ik\lambda} + \sum_{k=0}^{\infty} a^k e^{-ik\lambda}\right\} =$$

$$= \frac{C}{2\pi}\left\{\sum_{k=1}^{\infty} a^k e^{ik\lambda} + \sum_{k=0}^{\infty} a^k e^{-ik\lambda}\right\} = \frac{C}{2\pi}\left\{\frac{ae^{i\lambda}}{1-ae^{i\lambda}} + \frac{1}{1-ae^{-i\lambda}}\right\} =$$

$$= \frac{C}{2\pi}\left\{\frac{a}{e^{-i\lambda}-a} + \frac{e^{i\lambda}}{e^{i\lambda}-a}\right\} = \frac{C}{2\pi}\,\frac{1-a^2}{|e^{i\lambda}-a|^2}.$$

Можно показать, что последовательность $\xi(t)$ с корреляционной функцией $\sigma(\tau) = Ca^{|\tau|}$ может быть получена с помощью скользящего суммирования из последовательности некоррелированных случайных величин.

Пример 3. Пусть

$$\sigma(\tau) = \begin{cases} 1, & \tau = 0, \\ 0, & \tau \neq 0. \end{cases}$$

Это корреляционная функция стационарной последовательности некоррелированных случайных величин. В этом случае $f(\lambda) = 1/2\pi$.

ЭКСТРАПОЛЯЦИЯ И ФИЛЬТРАЦИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ

Экстраполяция стационарного случайного процесса — предсказание значения случайной величины $\xi(t)$ по известным значениям $\xi(t)$ в моменты времени $t-1$, $t-2$, . . . , $t-p$ (иногда по всему прошлому процесса, т. е. при $p \to \infty$).

Для простоты предположим, что математическое ожидание случайного процесса равно нулю, $M\xi(t) = 0$, а его ковариационная функция $M[\xi(t)\,\xi(s)]$ есть $\sigma(t-s)$. В общем виде задача предсказания случайного процесса сводится к отысканию такой функции g, при которой прогнозируемое значение может быть представлено в виде:

$$\hat{\xi}(t) = g[\xi(t-1),\ \xi(t-q),\ \ldots,\ \xi(t-p)].$$

Качество предсказания оценивается величиной среднего квадрата ошибки $\sigma_p^2 = M\varepsilon_p^2$, где

$$\varepsilon_p = \xi(t) - \hat{\xi}(t) = \xi(t) - g[\xi(t-1),\ \ldots,\ \xi(t-p)].$$

Наилучшим предсказанием называется такая функция g, для которой величина σ_p^2 принимает наименьшее значение. Простейший вид функции g — линейный.

Линейным предсказанием называется представление значения $\hat{\xi}(t)$ в виде $\sum_{s=1}^{p} C_s\xi(t-s)$ или при $p \to \infty$ — в виде ряда $\sum_{s=1}^{\infty} C_s\xi(t-s)$. В этом случае $M[\xi(t) - \hat{\xi}(t)]^2$ зависит только от функции $\sigma(h)$.

Пусть P — линейное пространство, натянутое на точки $\xi(t-1)$, $\xi(t-2)$, ... Наилучшее линейное предсказание для $\xi(t)$ сводится к нахождению точки линейного пространства P, наиболее близкой по норме к точке $\xi(t)$. Эта задача эквивалентна геометрической задаче об опускании перпендикуляра из точки $\xi(t)$ на линейное пространство P_p, натянутое на векторы $\xi(t-1)$, $\xi(t-2)$, . . . , $\xi(t-p)$.

Принимая в качестве скалярного произведения линейного пространства P $M[\xi(t)\,\eta(t)]$, приходим к следующему выводу. Так как перпендикуляр из точки $\xi(t)$ на подпространство P_p имеет минимальную длину, то предсказание $\hat{\xi}(t)$ получается из условия минимума среднего квадрата отклонения:

$$M\left[\xi(t) - \sum_{i=1}^{p} C_i\xi(t-i)\right]^2 = \sigma(0) + 2\sum_{j=1}^{p} C_j\sigma(j) + \sum_{i,\,j=1}^{p} C_iC_j\sigma(i-j).$$

Приравнивая к нулю частные производные этого выражения по $C_1, \ldots, C_p$, получим для определения коэффициентов $C_1, \ldots, C_p$ следующую систему линейных уравнений [4]:

$$\begin{pmatrix} \sigma_{(0)} & \sigma_{(1)} \ldots \sigma_{(p-1)} \\ \sigma_{(1)} & \sigma_{(0)} \ldots \sigma_{(p-2)} \\ \ldots\ldots\ldots\ldots \\ \sigma_{(p-1)} & \sigma_{(p-2)} \ldots \sigma_{(0)} \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ \ldots \\ C_p \end{pmatrix} = \begin{pmatrix} \sigma(1) \\ \sigma(2) \\ \ldots \\ \sigma(p) \end{pmatrix}.$$

Прогнозирование случайных процессов используется в геолого-геофизических исследованиях для построения карт различных характеристик в тех случаях, когда требуется получить сведения о значении картируемой величины в точке, недоступной для наблюдения или по разным причинам не подвергнутой наблюдению.

Линейная фильтрация стационарной случайной последовательности — нахождение с наибольшей возможной точностью по известным значениям $\zeta(t-1), \ldots, \zeta(t-n)$ стационарной последовательности $\zeta(t) = \xi(t) + \eta(t)$, где $\eta(t)$ — функция ошибок значений $\hat{\xi}(t+m)$, вычисляемых по формуле

$$\hat{\xi}(t+m) = a_1\zeta(t-1) + a_2\zeta(t-2) + \ldots + a_n\zeta(t-n),$$

дающей наилучшее приближение к $\zeta(t+m)$. При этом считается, что обе последовательности $\xi(t)$ и $\eta(t)$ также стационарны.

Будем предполагать, что нам известны корреляционные функции $\sigma_\zeta(\tau)$ и $\sigma_\eta(\tau)$ последовательностей $\zeta(t)$ и $\eta(t)$. На практике корреляционные функции $\sigma_\zeta(\tau)$ и $\sigma_\eta(\tau)$ подсчитываются соответственно по наблюденным значениям и точности измерительного прибора. Зная $\sigma_\zeta(\tau)$ и $\sigma_\eta(\tau)$, можно найти $\sigma_\xi(\tau)$ по формуле:

$$\sigma_\xi(\tau) = \sigma_\zeta(\tau) - \sigma_\eta(\tau).$$

Качество приближения

$$\hat{\xi}(t+m) = \sum_{i=1}^{n} a_i\zeta(t-i)$$

характеризуется средним квадратом ошибки:

$$\sigma^2_{nm} = M\left|\zeta(t+m) - \hat{\xi}(t+m)\right|^2.$$

Найдем те значения коэффициентов, при которых σ^2_{nm} достигает минимума.

Поставленная задача о линейной фильтрации сводится к задаче опускания перпендикуляра из точки $\xi(t+m)$ на линейное подпространство, натянутое на векторы $\zeta(t-1), \ldots, \zeta(t-m)$. Вектор

$$\xi(t+m) - \sum_{i=1}^{n} a_i\zeta(t-i)$$

и является таким перпендикуляром.

Искомые значения коэффициентов $a_1, a_2, \ldots, a_n$ находятся из равенств

$$\left(\xi(t+m) - \sum_{i=1}^{n} a_i\zeta(t-i),\ \zeta(t-k)\right) = 0,$$

где $(,)$ — скалярное произведение

$$(\xi(t), \xi(s)) = M(\xi(t), \xi(s)).$$

Используя определение $\sigma_\zeta(\tau)$, $\sigma_\eta(\tau)$ и некоррелированность последовательностей $\zeta(t)$ и $\eta(s)$, получаем систему линейных уравнений для определения коэффициентов $a_1, \ldots, a_n$:

$$\sigma_\xi(m+k) - \sum_{l=1}^{n} a_l\sigma_\zeta(k-l) = 0,\ k = 1, \ldots, n.$$

Средний квадрат ошибки фильтрации вычисляется по формуле

$$\sigma^2_{mn} = \sigma_\xi(0) - \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l \sigma_\zeta(k-l).$$

Теорию фильтрации случайных процессов для выделения полезной геологической и геофизической информации из экспериментальных данных использовали В. И. Аронов [5], И. Д. Савинский и др. [30, 45].

ГЛАВА 9

ТРЕНД-АНАЛИЗ

Тренд-анализ — математический метод, используемый для исследования закономерностей изменения геологического признака в пространстве и (или) во времени (анализ временных рядов). В более узком смысле под тренд-анализом понимается процедура аппроксимации эмпирических данных некоторыми вполне определенными функциями, аргументами которых являются координаты точек наблюдения. Предполагается, что любое из наблюдаемых значений z может быть представлено в виде суммы двух компонент, одна из которых (F) рассматривается как неслучайная функция от координат, а другая (φ) — как случайная:

$$z(x) = F(x) + \varphi(x),$$
$$z(x, y) = F(x, y) + \varphi(x, y),$$

где x, y — координаты точек наблюдения.

Детерминированная часть $F(\cdot)$ отражает закономерное изменение признака z в пределах исследуемой пространственной или временной области. Обычно такую систематическую составляющую

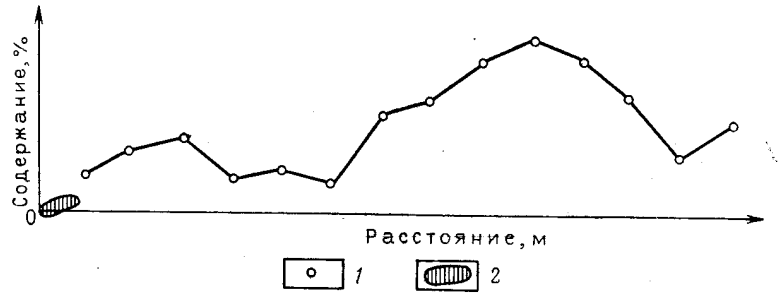


Рис. 47. Иллюстрация понятия линейного тренда при анализе участка рудного тела:
1 — точки наблюдения; *2* — место расположения рудного тела

связывают с действием регионального геологического фактора, сфера влияния которого заметно превышает размеры участка аппроксимации. Появление флуктуации $\varphi(\cdot)$ может быть вызвано следующими причинами: а) влиянием локально действующих геологических факторов; б) случайными ошибками измерений признака z.

В зависимости от смысла решаемой геологической задачи внимание исследователя может быть сосредоточено на следующих вопросах: 1) выявлении общей тенденции (тренда) в изменении признака z; 2) обособлении локальной составляющей (поиск положительных и отрицательных аномалий) (рис. 47).

ВЫДЕЛЕНИЕ РЕГИОНАЛЬНОЙ СОСТАВЛЯЮЩЕЙ

Задача решается вполне однозначно лишь в том случае, когда исследователю известны основные параметры процесса, формирующего переменную z. Чаще всего, однако, такая информация отсутствует, в связи с чем точное решение недостижимо. Неопределенность удается в какой-то мере уменьшить, если ввести некоторые ограничения на вид аппроксимирующей функции $F(\cdot)$ и функции $\varphi(\cdot)$. Такой подход оказывается особенно успешным тогда, когда соответствующее сужение класса функций $F(\cdot)$ и (или) $\varphi(\cdot)$ определяется особенностями исследуемого признака z.

А. Б. Вистелиус описал общую процедуру выделения региональной составляющей при следующих предположениях относительно $\varphi(x, y)$: распределение $\varphi(x, y)$ не зависит от координат точек наблюдения, $\varphi(x_i, y_j)$ и $\varphi(x_k, y_l)$ независимы для любых сочетаний (i, j) (k, l). Обычно также предполагают, что $\varphi(x, y)$ распределена нормально с параметрами (μ, σ^2). Функция $F(x, y)$ выбирается (из некоторого фиксированного класса) так, чтобы обеспечить получение остатков $\varphi(x, y) = z(x, y) - F(x, y)$, удовлетворяющих вышеуказанным условиям.

На практике чаще реализуется иной, более простой критерий, а именно: $\|z - F\|_H \leqslant \varepsilon$, где $\|\cdot\|$ — норма в функциональном пространстве H; $z \in H$, $F \in H$; ε — требуемая точность аппроксимации.

Как и в предыдущем случае, доброкачественность решения во многом зависит от удачного выбора класса функции H. Если в рамках H сразу несколько функций удовлетворяют условию $\|z - F\|_H \leqslant \varepsilon$, то обращаются к критерию наилучшего приближения: $\min \|z - F\|_H$.

В геологии чаще всего используются следующие приемы сглаживания: 1) методы, опирающиеся на скользящие средние; 2) аппроксимация алгебраическими полиномами; 3) приближение гармониками; 4) сплайн-аппроксимация.

Методы скользящего среднего

В основе этих методов лежит следующая общая процедура. Для первых m членов (m — нечетно) сглаживаемого ряда объемом n наблюдений ($m < n$) строится полином степени P ($P \leqslant m-1$), после

чего определяется его значение для точки $k = (m+1)/2$. Затем вновь берется m членов, начиная со второго, и все расчеты выполняются заново. Если учесть, что на каждом шаге производится сдвиг на одно наблюдение, то последней точкой, для которой вычисляется значение полинома, является точка с номером $n - (m-1)/2$.

В простейшем случае ($p = 1$) сглаживание выполняется обычным усреднением значений $z(\cdot)$:

$$z_k = \frac{1}{m} \sum_{t=k-(m-1/2)}^{t=k+(m-1)/2} z_t,$$

где z_k — сглаженное значение, относимое к точке k; m — число точек сглаживания; z_t — исходное значение аппроксимируемого признака.

Для $p > 1$ значения z_t вводятся в расчетную формулу с весами C_t, зависящими как от величины t, определяющей степень удаленности от «центральной точки», так и от степени полинома p:

$$\hat{z}_k = \frac{1}{S} \sum_{t=k-(m-1)/2}^{t=k+(m-1)/2} C_t z_t,$$

где S — алгебраическая сумма весовых коэффициентов (нормированный множитель). Заметим, что $C_{k-t} = C_{k+t}$.

Например, для $m = 5$ и $p = 3$ получаем так называемую первую формулу Шеппарда [22]:

$$\hat{z}_k = \frac{1}{35}[17z_k + 12(z_{k+1} + z_{k-1}) - 3(z_{k+2} + z_{k-2})].$$

Ниже приведены величины S (в круглых скобках) и весовые коэффициенты (в фигурных скобках), соответствующие определенному числу сглаживаемых наблюдений m (в квадратных скобках).

Так как случай нечетного p включает случай предыдущего четного p, то соответствующие полиномы объединены:

$p = 2,\ p = 3$

[5] $\left(\frac{1}{35}\right)$ {17, 12, −3},

[7] $\left(\frac{1}{21}\right)$ {7, 6, 3, −2},

[9] $\left(\frac{1}{231}\right)$ {59, 54, 39, 14, −21},

[11] $\left(\frac{1}{429}\right)$ {89, 84, 69, 44, 9, −36},

[13] $\left(\frac{1}{143}\right)$ {25, 24, 21, 16, 9, 0, −11},

[15] $\left(\frac{1}{1105}\right)$ {167, 162, 147, 122, 87, 42, — 13, — 78},

[17] $\left(\frac{1}{323}\right)$ {43, 42, 29, 34, 27, 18, 7, — 6, — 21},

[19] $\left(\frac{1}{2261}\right)$ {269, 264, 249, 224, 189, 144, 89, 24, — 51, — 136},

[21] $\left(\frac{1}{231}\right)$ {329, 324, 309, 284, 249, 204, 149, 84, 9, 276, — 171},

$P = 4,\ P = 5$

[7] $\left(\frac{1}{231}\right)$ {131, 75, — 30, 3},

[9] $\left(\frac{1}{429}\right)$ {179, 135, 30, — 55, 15},

[11] $\left(\frac{1}{429}\right)$ {143, 120, 60, — 10, — 45, 18},

[13] $\left(\frac{1}{2431}\right)$ {677, 600, 390, 110, — 135, — 198, 110},

[15] $\left(\frac{1}{46189}\right)$ {11063, 10125, 7500, 3755, — 165, — 2937, — 2860, 2145},

[17] $\left(\frac{1}{4199}\right)$ {883, 825, 660, 415, 135, — 177, — 260, — 195, 195},

[19] $\left(\frac{1}{7429}\right)$ {1393, 1320, 1110, 750, 405, 18, — 290, — 420, — 225, 340},

[21] $\left(\frac{1}{260015}\right)$ {44003, 42120, 36660, 28190, 17655, 6378, 2940, 11220, — 13005, — 6460, 11628}.

Выбор той или иной формулы определяется обычно опытным путем. Аппроксимация скользящим средним, подавляя высокочастотную компоненту, сохраняет в то же время общую конфигурацию крупных пиков, соответствующих региональной составляющей. Вместе с тем следует отметить довольно часто наблюдаемое несовпадение местоположения пиков до и после сглаживания. Еще одним недостатком этого метода является отсутствие сглаженных значений на краях аппроксимируемых рядов.

В двумерном случае получаем (двумерный тренд-анализ) (рис. 48, 49):

$$\hat{z}(x_i,\ y_i) = \alpha_{ij} \sum_{(x_r,\ y_s) \in \Pi_{ij}} C(x_r,\ y_s)\, z(x_r,\ y_s),$$

где x_i, y_j — координаты центра площадки («окна») сглаживания Π_{ij}; x_r, y_s — координаты точек наблюдения, принадлежащих площадке Π; $C(x_r, y_s)$ — весовая функция; α_{ij} — нормировочный множитель.

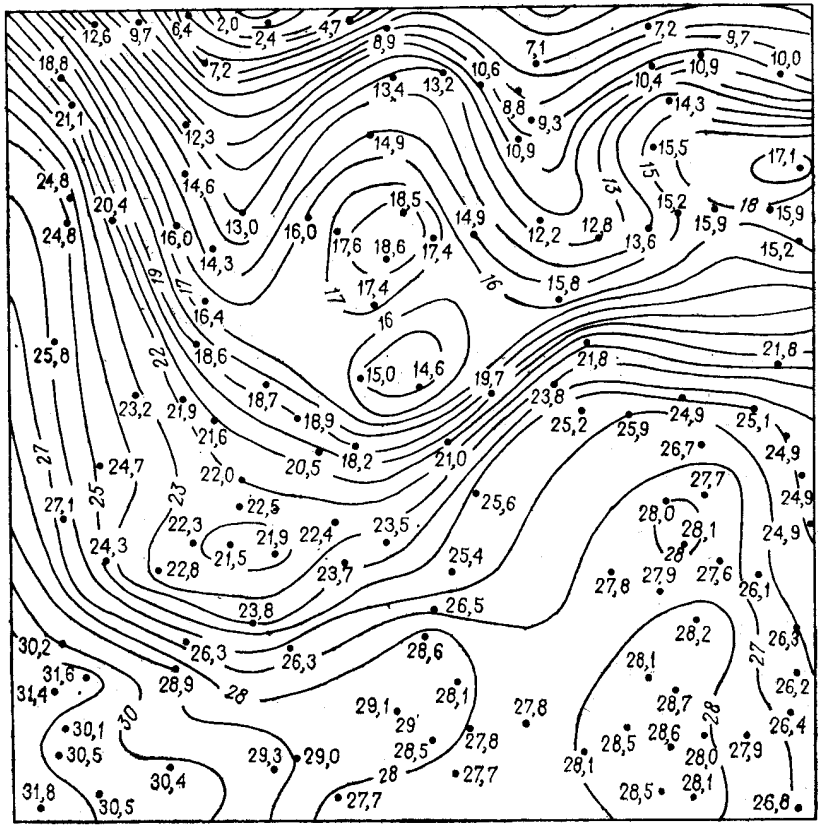


Рис. 48. Пример карты, построенной с применением тренд-анализа вручную

Многочисленные модификации метода «скользящего окна», используемые в геологических и геофизических исследованиях, отличаются друг от друга формой и размерами площадок трансформации, весовыми функциями, требованиями к расположению и количеству точек, охватываемых сглаживающим окном и т. п.

Простейший вариант «скользяшего окна» — сглаживание невзвешенным осреднением:

$$\frac{1}{z}(x_i,\ y_j) = \frac{1}{m} \sum_{x_r,\ y_s \in \Pi_{ij}} z(x_r,\ y_s),$$

где m — число наблюдений в пределах площадки Π_{ij}.

Этот способ применяется в тех случаях, когда есть основания полагать, что в границах площадки сглаживания $F(x, y) = \text{const}$, а $\varphi(x, y)$ однородна, распределена нормально и ее значения в соседних точках взаимно независимы.

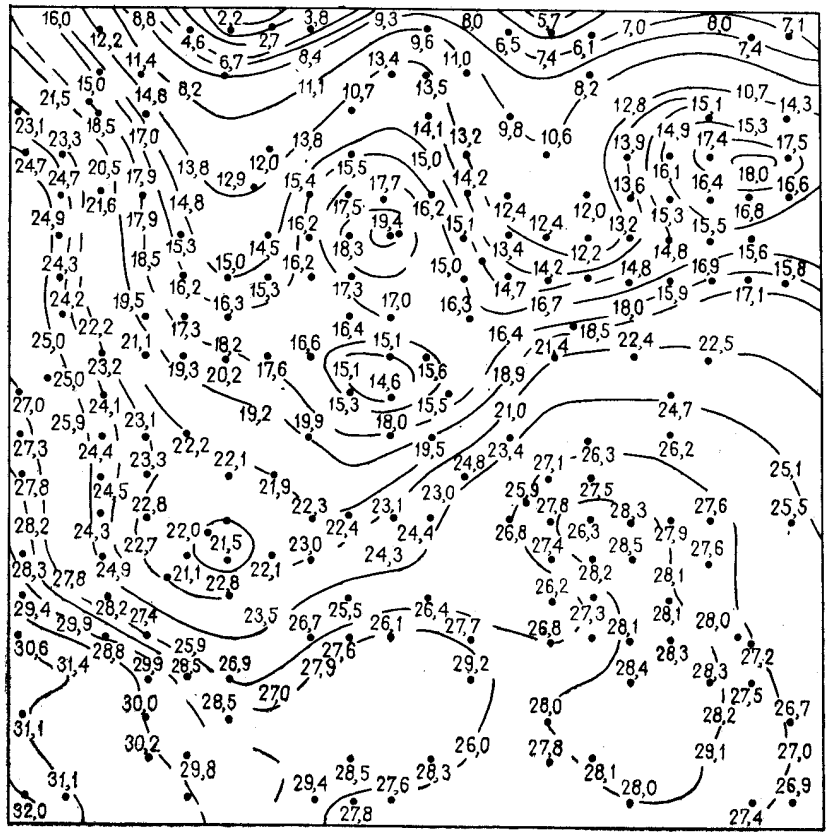


Рис. 49. Пример карты, построенной с применением тренд-анализа на ЭВМ

Роль весовых функций проиллюстрируем на примере метода, который условно может быть назван методом «ближайших точек». Эта модификация скользящего усреднения отличается следующими особенностями: а) размер и форма площадки заранее не устанавливаются, б) исходные точки могут располагаться на карте неравномерно и не обязательно в узлах регулярной сети, в) число «ближайших точек» m, участвующих в сглаживании, постоянно.

Чтобы построить аппроксимирующую поверхность методом ближайших точек, вся исследуемая территория покрывается прямоугольной (квадратной) координатной сеткой. Далее вычисляются значения $\hat{z}$, соответствующие узлам этой сетки. Для этого отыскивается m точек, ближайших к узлу с координатами (x_i, y_j), а затем рассчитываются расстояния (D_{rs}) между узлом сетки и каждой из этих точек:

$$D_{rs} = [(x_i - x_r)^2 + (y_j - y_s)^2]^{1/2}.$$

Величина $1/D_{rS}$ играет роль весовой функции. Нормирующий множитель определяется так:

$$\alpha_{ij} = 1/\sum_{r,\,s=1}^{m} (1/D_{rs}).$$

Подставляя значение α_{ij} в формулу двумерного тренд-анализа, получаем:

$$\hat{z}(x_i,\ y_j) = \left[\sum_{1}^{m} z(x_r,\ \ y_s)\, D_{rs}\right] / \sum_{1}^{m} (1/D_{rs}).$$

Таким образом, исходные точки учитываются с весами, обратно пропорциональными их расстояниям до узла координатной сетки. Обратим внимание, что при $x_i = x_r$, $y_j = y_s$ аппроксимированное значение $\hat{z}(x_i, y_j)$ совпадает с наблюдаемым значением $z(x_r, y_s)$.

Аппроксимация алгебраическими полиномами

Выделение региональной составляющей $F(\cdot)$ осуществляется с помощью полиномиального приближения всей наблюдаемой совокупности эмпирических данных. Таким образом, искомая функция $F(\cdot)$ заменяется полиномами P_l степени l:

$$P_l(x) = \sum_{r} a_r^{(l)} x_r; \qquad r = 0,\ 1,\ \ldots,\ l;$$

$$P_l(x,\ y) = \sum_{r,\,s} a_{rs}^{(l)} x_r,\ y,\ s;$$

$$r = 0,\ 1,\ \ldots,\ l; \qquad s = 0,\ 1,\ \ldots,\ l; \qquad r + s = l,$$

коэффициенты которых $A = \{a^{(l)}\}$ определяются методом наименьших квадратов из условия:

$$\min \sqrt{\sum_{i=1}^{n} [z(x_i) - P_l(x_i)]^2} = \min \sqrt{\sum_{i=1}^{n} l_i^2}$$

(аналогично для двумерного случая).

Полином $P_l(x, y)$ в развернутой форме имеет следующий вид (ограничимся первыми тремя степенями):

$$P_1(x,\ y) = a_{00}^{(1)} + a_{10}^{(1)} x + a_{01}^{(1)} y;$$

$$P_2(x,\ y) = a_{00}^{(2)} + a_{10}^{(2)} x + a_{01}^{(2)} y + a_{20}^{(2)} x^{(2)} + a_{11}^{(2)} xy + a_{02}^{(2)} y^2;$$

$$P_3(x,\ y) = a_{00}^{(3)} + a_{10}^{(3)} x + a_{01}^{(3)} y + a_{20}^{(3)} x^2 + a_{11}^{(3)} xy + a_{02}^{(3)} y^2 + a_{20}^{(3)} x^3 +$$

$$+ a_{12}^{(3)} xy^2 + a_{21}^{(3)} x^2 y + a_{03}^{(3)} y^3.$$

Линейность отыскиваемых полиномиальных функций относительно $\{a^{(l)}\}$ облегчает составление системы уравнений, решение

которой дает оценки этих неизвестных коэффициентов. В матричной форме систему линейных уравнений можно записать так:

$$BA = z + E,$$

где $B = (1, x, y, x^2, \ldots, x^s, y^s, \ldots, y^l)$; z — вектор-столбец исходных данных; E — вектор-столбец случайных ошибок [5], относительно которых предполагается, что они распределены нормально с нулевым математическим ожиданием и постоянной дисперсией.

Минимизируя $E'E$, определяем оценку матрицы A:

$$\hat{A} = (B'B)^{-1}B'z.$$

Выбор степени полинома, как и при сглаживании скользящим средним, является столь же сложной проблемой. Ее однозначное решение требует использования такой дополнительной информации о процессах, формирующих исследуемый признак z, которой геолог в подавляющем большинстве случаев не располагает. Лишь для отдельных классов геологических задач удается сформулировать самые общие требования к форме аппроксимирующих поверхностей. Например, если $F(\cdot)$ рассматривается как модель регионального геохимического фона, то аппроксимирующая поверхность должна иметь достаточно простой, плавно изменяющийся рельеф, задаваемый полиномами невысоких степеней.

Если же информация о возможной форме моделируемой поверхности отсутствует, то рекомендуется строить несколько карт, последовательно повышая степень аппроксимирующего полинома и проводя на каждом шаге содержательный анализ получаемых результатов. При этом не следует упускать из виду следующие обстоятельства: а) с повышением степени полинома на картах тренда все больший вес начинают получать эффекты, связанные с действием локальных факторов, б) при небольшом числе неравномерно расположенных точек наблюдения и высоких (выше 4—5) степенях полинома возможны неконтролируемые отклонения аппроксимирующей поверхности от моделируемой геологической поверхности, что связано с появлением плохо обусловленных матриц $B'B$. Один из первых признаков наличия таких искажений — так называемый «краевой эффект», выражающийся в появлении на краях карты чрезмерно высоких или низких значений $\hat{z}$. В какой-то мере искажающее влияние этого эффекта удается уменьшить, если прибегнуть к масштабированию исходных данных. Предлагаемое А. Б. Вистелиусом логарифмирование исходных данных не только уменьшает краевой эффект (за счет стабилизации дисперсии z), но и исключает появление отрицательных значений $\hat{z}$. Последнее особенно важно в тех случаях, когда величины $z < 0$ не имеют геологического смысла (например, в задачах сглаживания содержаний химических элементов). С целью уменьшения ошибок округления (они являются одной из причин потери устойчивости при обращении матрицы $B'B$) рекомендуется совмещать точку ($x = 0$, $y = 0$) с центром исследуемого района.

Аппроксимация гармониками

Сглаживание эмпирических данных с помощью рядов Фурье (гармонический анализ) уместно в тех случаях, когда переменная z периодически изменяется в пространстве и (или) во времени. Такую цикличность можно ожидать в ситуациях, характеризующихся более или менее регулярной повторяемостью в пределах изучаемого участка земной коры определенного комплекса геологических условий, что в свою очередь обеспечивает периодическое изменение значений признака z, фиксируемого исследователем.

О д н о м е р н ы й с л у ч а й. Периодическую составляющую запишем в виде следующего ряда Фурье:

$$\hat{z}(x) = \sum_{k=0}^{\infty}\left(\alpha_k \cos\frac{2k\pi x}{\lambda} + \beta_k \sin\frac{2k\pi x}{\lambda}\right),$$

где k — номер гармоники (гармоническое число); α_k, β_k — коэффициенты, зависящие от гармонического числа; λ — длина основной волны.

Выбор величины λ произволен, обычно задают $\lambda = L$, где L равна или превышает длину исследуемого ряда наблюдений. Таким образом, периодическая составляющая величины z представляется суммой синусоидальных и косинусоидальных функций, частоты которых кратны основной частоте $1/\lambda$. Обычно ограничиваются конечным и к тому же небольшим числом гармоник: $k = 0, 1, 2, \ldots, m$.

Нахождение неизвестных коэффициентов α_k и β_k осуществляется методом наименьших квадратов. Формальная запись системы уравнений, необходимых для оценивания α_k и β_k, существенно упрощается, если воспользоваться следующими обозначениями:

$$C_k^i = \cos\frac{2k\pi x_i}{\lambda} \quad \text{и} \quad S_k^i = \sin\frac{2k\pi x_i}{\lambda}.$$

В матричной форме система линейных уравнений имеет вид:

$$\begin{Vmatrix} C_0C_0' & C_0S_0' & C_0C_1' & \ldots & C_0S_k' \\ S_0C_0' & S_0S_0' & S_0C_1' & \ldots & S_0S_k' \\ C_1C_0' & C_1S_0' & C_1C_1' & \ldots & C_1S_k' \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ S_kC_0' & S_kS_0' & S_kC_1' & \ldots & S_kS_k' \end{Vmatrix} \times$$

$$\times\begin{Vmatrix} \alpha_0 \\ \beta_0 \\ \alpha_1 \\ \dots \\ \beta_k \end{Vmatrix} = \begin{Vmatrix} zC_0' \\ zS_0' \\ zC_1' \\ \dots \\ zS_k' \end{Vmatrix},$$

где C_0, C_1, . . . , S_k — векторы, элементы которых соответствуют различным значениям x_i; точно также $z = \{z(x_i)\}$, $i = 1, 2, \dots, n$.

Для $k = 0$ (гармоника нулевого порядка) $C_0^i = 1$ при любом x_i, так как $\cos 0 = 1$, а $S_0^{(i)} = 0$ при любом x_i, так как $\sin 0 = 0$. В связи с этим в вышеприведенной матрице все элементы, содержащие S_0, обращаются в нуль, что приводит к устранению второго столбца и второй строки. Элементы первого столбца и первой строки также упрощаются:

$$C_0C_0' = n,\ C_0C_1' = \sum_{i=1}^{n} C_1^{(i)},\ C_0S_1' = \sum_{i=1}^{n} S_1^{(i)} \qquad \text{и т. д.}$$

При большем числе гармоник система уравнений оказывается весьма громоздкой, что серьезно осложняет ее решение даже на современных ЭВМ. Однако если наблюдения z выполнены через равные интервалы, то все элементы матрицы, расположенные выше и ниже ее главной диагонали, становятся равными нулю. В этом случае коэффициенты α и β определяются так:

$$\alpha_n = \frac{2}{n}\sum_{i=1}^{n} z_i \cos\frac{2n\pi x_i}{\lambda};$$

$$\beta_n = \frac{2}{n}\sum_{i=1}^{n} z_i \sin\frac{2n\pi x_i}{\lambda}.$$

Заметим, что для конечного ряда, состоящего из n наблюдений, можно вычислить $k = n/2$ гармоник.

Двумерный случай. В двумерном случае получаем:

$$\hat{z}(x, y) = \sum_{k=0}^{kC}\sum_{l=0}^{LS} CC_{kl}\cos(2\pi kx/Mx)\cos(2\pi ly/My) +$$
$$+ \sum_{k=0}^{kC}\sum_{l=1}^{LS} CS_{kl}\cos(2\pi kx/Mx)\sin(2\pi ly/My) +$$
$$+ \sum_{k=1}^{kS}\sum_{l=0}^{LS} SC_{kl}\sin(2\pi kx/Mx)\cos(2\pi ly/My) +$$
$$+ \sum_{k=1}^{kS}\sum_{l=1}^{LS} SS_{kl}\sin(2\pi kx/Mx)\sin(2\pi ly/My),$$

где Mx, Mg — оценки длин основных волн в направлении осей соответственно x и y, kC — максимальное число косинусоидальных

гармоник по оси x; LS — то же, по оси y; kS — максимальное число синусоидальных гармоник по оси x; LS — то же по оси y; CC_{kl}, CS_{kl}, SC_{kl}, SS_{kl} — коэффициенты двойного ряда Фурье.

Названные выше величины определяются следующим образом:

$$Mx = \max x + 1; \qquad My = \max y + 1;$$
$$KC = Mx/2, \qquad kS = (Mx-2)/2 \quad \text{если } Mx \text{ четно};$$
$$LC = My/2,\ LS = (My-2)/2, \text{ если } My \text{ нечетно};$$
$$kC = kS = (Mx-1)/2, \text{ если } Mx \text{ нечетно};$$
$$LC = LS = (My-1)/2, \quad \text{если } My \text{ нечетно}.$$

Чтобы найти неизвестные коэффициенты CC_{kl}, CS_{kl}, SC_{kl}, SS_{kl}, необходимо составить систему линейных уравнений. Матричная запись этой системы будет менее громоздкой, если воспользоваться следующими обозначениями:

$$A_k = \cos(2\pi kx/Mx); \qquad B_k = \sin(2\pi kx/Mx);$$
$$C_l = \cos(2\pi ly/My); \qquad D_l = \sin(2\pi ly/My).$$

Тогда система уравнений приобретает вид

$$\begin{Vmatrix} \sum_{x,y}(A_0C_0)^2 & \sum_{x,y} A_1C_0A_0C_0 & \dots & \sum_{x,y} B_{kS}D_{LS}A_0C_0 \\ \sum_{x,y} A_0C_0A_1C_0 & \sum_{x,y}(A_1C_0^2) & \dots & \sum_{x,y} B_{kS}D_{LS}A_1C_0 \\ \dots & \dots & \dots & \dots \\ \sum_{x,y} A_0C_0A_3D_1 & \sum_{x,y} A_1C_0A_3D_1 & \dots & \sum_{x,y} B_{kS}D_{LS}A_3D_1 \\ \dots & \dots & \dots & \dots \\ \sum_{x,y} A_0C_0B_{kS}D_{LS} & \sum_{x,y} A_1C_0B_{kS}D_{LS} & \dots & \sum_{x,y} B_{kS}D_{LS}^2 \end{Vmatrix} \times$$
$$\times\begin{Vmatrix} CC_{00} \\ CC_{10} \\ \dots \\ CS_{31} \\ \dots \\ SS_{kSLS} \end{Vmatrix} = \begin{Vmatrix} \sum_{x,y} z(x, y)A_0C_0 \\ \sum_{x,y} z(x, y)A_1C_0 \\ \dots \\ \sum_{x,y} z(x, y)A_3D_1 \\ \dots \\ \sum_{x,y} z(x, y)B_{kS}D_{LS} \end{Vmatrix}.$$

Трудоемкость вычисления коэффициентов $CC_{00} \dots SS_{kSLS}$ существенно снижается, если точки наблюдений расположены в узлах регулярной сети. В этом случае недиагональные элементы матрицы обращаются в нуль, что заметно упрощает решение системы уравнений.

Аппроксимация сплайн-функциями

Приближенное описание геологических поверхностей сплайн-функциями позволяет устранить ряд недостатков, присущих полиномиальной аппроксимации, а именно: а) снизить трудоемкость вычислительных процедур при моделировании сложных поверхностей полиномами высоких степеней; б) избежать искажения типа «краевых эффектов» в зонах, удаленных от центра карты и слабо обеспеченных наблюдениями. В то же время сглаживание сплайнами, являясь кусочно-полиномиальной аппроксимацией, сохраняет все преимущества приближения исследуемых геологических полей многочленами низких степеней. Возможность описания сложных поверхностей с помощью полиномов невысоких степеней определяется тем, что в сплайн-методе вся картируемая территория разбивается на относительно небольшие непересекающиеся участки — прямоугольники или треугольники, в вершинах которых размещены точки наблюдений. Аппроксимация полиномами осуществляется раздельно для каждого типа такого многоугольника. Обычно используют полином третьей степени — кубический сплайн.

Так как в каждом многоугольнике подбирается свой полином, то возникает задача по обеспечению непрерывности функций в точках сочленения. Сформулируем общие условия, необходимые для выполнения гладкого «склеивания» аппроксимирующих поверхностей.

Среди множества точек наблюдений выделим два связанных подмножества M_1 и M_2, в точках которых измерены значения $z(x, y)$ с некоторой случайной ошибкой $\varphi(x, y)$. Два полиномиальных (кубических) приближения $P_1\{(x, y): x \in M_1, y \in M_1\}$ и $P_2\{(x, y): x \in M_2\ y \in M\}_2$ будут составлять единую непрерывную и гладкую аппроксимирующую поверхность, если в любой точке (x_0, y_0), расположенной в области пересечения M_1 и M_2, выполняются равенства

$$P_1(x_0, y_0) = P_2(x_0, y_0); \qquad \frac{\partial^2 P_1(x_0, y_0)}{\partial x^2} = \frac{\partial^2 P_2(x_0, y_0)}{\partial x^2};$$

$$\frac{\partial P_1(x_0\ y_0)}{\partial x} = \frac{\partial P_2(x_0, y_0)}{\partial x}; \qquad \frac{\partial^2 P_1(x_0, y_0)}{\partial x \partial y} = \frac{\partial^2 P_2(x_0, y_0)}{\partial x \partial y};$$

$$\frac{\partial P_1(x_0, y_0)}{\partial y} = \frac{\partial P_2(x_0, y_0)}{\partial y}; \qquad \frac{\partial^2 P_1(x_0, y_0)}{\partial y^2} = \frac{\partial^2 P_2(x_0, y_0)}{\partial y^2}.$$

Эти равенства вводят необходимые ограничения на искомые коэффициенты полиномов P_1 и P_2, при этом первое из них обеспечивает непрерывность функций в точке склеивания (x_0, y_0), а остальные — непрерывность первых и вторых частных производных, т. е. гладкость склеивания полиномов.

Таким образом, в самом общем виде задача сплайн-аппроксимации сводится к отысканию коэффициентов полиномов с учетом вышеуказанных условий. Алгоритмы нахождения коэффициентов в узлах правильной прямоугольной сети (классическая ситуация)

описаны в работе [13]. Если же данные расположены хаотически, то задача не имеет однозначного решения — можно построить бесконечное множество сплайнов, согласующихся с наблюдаемыми значениями z. В этом случае для сведения задачи к классической схеме предлагается либо дополнительно вычислить значения картируемой переменной в необеспеченных узлах сети (например, с помощью интерполяционных методов), либо обратиться к некоторой внестатистической информации, позволяющей судить об общем поведении аппроксимируемой поверхности в заданной области. Так, если подбирается сплайн-функция, описывающая пространственное распределение значений регионального фона химических элементов, то самым общим требованием к соответствующей поверхности будет условие ее максимальной плавности (минимальность средней кривизны).

Пусть S — множество бикубических сплайнов, аппроксимирующих наблюдаемые значения $z(x, y)$; при этом для любого $S^* \in S$ выполняется условие $|S^*(x_i y_j| - z(x_i y_j| \leqslant \varepsilon$, где ε — допустимая ошибка аппроксимации. Требуется выбрать такой сплайн S_0, который формирует поверхность максимальной гладкости. В. А. Волков [13] в качестве соответствующих критериев предлагает воспользоваться функционалами:

$$\Phi'(S) = \frac{1}{\mathcal{G}} \iint\limits_{-x,\ y \in \mathcal{F}} \left[\left(\frac{\partial S(x, y)}{\partial x}\right)^2 + \left(\frac{\partial S(x, y)}{\partial y}\right)^2\right] dxdy;$$

$$\Phi''(S) = \frac{1}{\mathcal{G}} \iint\limits_{-x,\ y \in \mathcal{F}} \left[\left(\frac{\partial^2 S(x, y)}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 S(x, y)^2}{\partial x \partial y} + \frac{\partial^2 S(x, y)}{\partial y^2}\right)^2\right] dxdy.$$

Минимизация функционала $\Phi'(S)$ определяет сплайн-поверхность, обладающую минимальной площадью (критерий минимального «среднего угла наклона»), а функционала $\Phi''(S)$ — поверхность с минимальной средней скоростью изменения углов наклона (критерий минимума «средней кривизны»).

Сглаживание сплайн-функциями особенно удобно при моделировании поверхностей, осложненных разрывными нарушениями. В этом случае рекомендуется обращаться к сплайн-аппроксимации на триангулированных областях.

ОБОСОБЛЕНИЕ ЛОКАЛЬНОЙ СОСТАВЛЯЮЩЕЙ (ВЫДЕЛЕНИЕ АНОМАЛИЙ)

С задачей выделения аномалий геолог постоянно сталкивается при геохимических поисках. В нефтяной геологии аналогичная проблема возникает, например, при морфологическом анализе поверхности пластов осадочных пород, предпринимаемом с целью обнаружения структурных ловушек (выделение локальных поднятий).

В рамках данной задачи основная модель тренд-анализа приобретает вид:

$$z(x, y) = F(x, y) + L(x, y) + \varepsilon(x, y),$$

где $L(x, y)$ — локальная составляющая; $\varepsilon(x, y)$ — случайные флуктуации признака z.

Будем рассматривать искомые аномалии как полезный сигнал, а компоненты $F(x, y)$ и $\varepsilon(x, y)$ соответственно как низкочастотный и высокочастотный шумы. Тогда задача выделения аномалий сводится к подавлению шумов $F(x, y)$ и $\varepsilon(x, y)$. Обычно частота $L(x, y)$ заметно отличается от частоты региональной составляющей, но близка к частотам $\varepsilon(x, y)$. Последнее обстоятельство не позволяет полностью отфильтровать $\varepsilon(x, y)$, что приводит к выделению ложных аномалий. С другой стороны, существующие методы конструирования аппроксимирующих поверхностей $F(x, y)$ таковы, что вполне вероятно некоторое поглощение ими компоненты $L(x, y)$, что приводит к пропуску искомых аномалий.

Положение может быть улучшено, если исследователь располагает хотя бы приближенными спектральными характеристиками $F(x, y)$ и $\varepsilon(x, y)$. В этом случае удается построить полосный фильтр, задерживающий мешающие частоты $F(x, y)$ и $\varepsilon(x, y)$.

В условиях отсутствия достаточно полной информации о собственных функциях процессов, формирующих фон и аномалии, можно рекомендовать следующие приемы обособления локальной составляющей:

1) сглаживание исходных значений z скользящим окном постоянного размера. Аномалии определяются по разности $z-\hat{z}$. Для получения устойчивых результатов в качестве аномалий рассматриваются группы расположенных рядом точек с одинаковым знаком разности;

2) сглаживание наблюдаемых значений z площадками (окнами) разных размеров, аномалии выделяются по разности $\hat{z}''-\hat{z}'$, где $\hat{z}'$ — результаты сглаживания прямоугольным (квадратным) окном с диагональю R', а $\hat{z}''$ — с диагональю R'', $R' > R''$;

3) аппроксимация исходных данных полиномами низких ($l = 1, 2, 3$) степеней с последующим нахождением разностей $z-P_l$. Аномалии могут быть ранжированы по степени их устойчивости ко все возрастающей степени полинома: сильно выраженные аномалии сохраняются на всех картах отклонений, наиболее слабые проявляются лишь на карте отклонений от полинома первой степени;

4) аномалии определяются по разности $p_{l+1}-P_l$ (аналог п. 2);

5) сглаживание исходных данных скользящим окном небольшого размера (подавление $\varepsilon(x, y)$ с последующей аппроксимацией полученных $\hat{z}$ полиномами низких степеней). Аномалии выделяются по картам отклонений $\hat{z}-P_l(\hat{z})$, сглаженных, в свою очередь, методом «ближайших точек».

ГЛАВА 10

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляционный анализ из всех статистических методов исследования геологических объектов нашел наиболее широкое применение. Вычисление коэффициентов корреляции и сопряженности с последующей оценкой взаимозависимости случайных величин (моделей геологических, геофизических, геохимических и других характеристик) присутствует практически в любой работе, где применяются математические методы в геологических исследованиях. Так, например, это имеет место при парагенетическом анализе химических элементов (Б. И. Смирнов), установлении связей направлений трещин с элементами тектонических структур (Л. Д. Кноринг), изучении взаимозависимостей геохронологических и тектонических подразделений, реконструкции тектонических структур (А. В. Долицкий), выявлении петрофизических связей в промысловой геофизике (М. М. Элланский), изучении связей оруденения с магматизмом (В. В. Ляхович), геологической интерпретации гравитационных и магнитных аномалий (Г. И. Каратаев), прогнозирования геологических характеристик по набору или комплексу других геологических (геофизических, геохимических) характеристик.

Корреляция — это удобная модель, позволяющая исследовать комплексные свойства геологических объектов, и является необходимым элементом любой процедуры многомерного статистического анализа. Ниже описаны коэффициенты корреляции парные, частные, множественные, ранговые и коэффициенты сопряженности. Подробнее различные вопросы, связанные с оцениванием указанных величин, рассмотрены в работах [2, 6, 8, 19, 23, 28, 29, 40, 46].

Корреляционный анализ — статистическое исследование стохастической зависимости между случайными величинами $\{Y_j, j = 1, 2, \ldots, k\}$ и $\{X_j; j = 1, 2, \ldots, l\}$ [2, 9, 18, 22]. Задачами корреляционного анализа являются: 1) оценка по выборочным данным величины коэффициента парной корреляции ($k = l = 1$), множественной корреляции ($k = 1$; $l > 1$), канонической корреляции ($k > 1$; $l > 1$); 2) проверка значимости выборочных коэффициентов корреляции; 3) оценка степени близости выявленной связи к линейной. Если зависимость между X_j и Y_j (в дальнейшем в ситуациях $k = l = 1$ нижний индекс у коррелируемых переменных будем опускать) имеет линейный характер, то удается охарактеризовать не только тесноту связи, но и ее направление. Связь называется прямой (положительной), если при увеличении (уменьшении) значений одной из переменных другая обладает устойчивой тенденцией к увеличению (уменьшению) своих значений. В этом случае коэффициент парной корреляции положителен. При обратном со-

отношении между переменными X и Y имеем дело с отрицательной корреляцией, что находит свое выражение и в знаке коэффициента парной корреляции.

При конкретном выборе той или иной меры связи следует учитывать степень детальности описания коррелируемых геологических признаков.

ПАРНАЯ КОРРЕЛЯЦИЯ

1. Корреляция дихотомических признаков [22]. Подобная ситуация типична для тех случаев, когда исследователь фиксирует либо наличие, либо отсутствие некоторого определенного свойства.

Пусть случайные величины X и Y принимают значения $\{x, \bar{x}\}$ и $\{y, \bar{y}\}$, где x, y означает, что данные признаки фиксируются; тогда символами $\bar{x}$, $\bar{y}$ обозначим отсутствие этих признаков. В более общем случае x и $\bar{x}$ соответствуют регистрации двух различающихся значений переменной X (это же справедливо и для Y).

В результате единичного наблюдения можно ожидать появления таких сочетаний: x и y, x и $\bar{y}$, $\bar{x}$ и y, $\bar{x}$ и $\bar{y}$. Выполнив N наблюдений, получим соответствующие частоты $n_{11} = n(x, y)$, $n_{12} = n(x, \bar{y})$, $n_{21} = n(\bar{x}, y)$, $n_{22} = (\bar{x}, \bar{y})$.

Для данных такого типа величина выборочного коэффициента связи, являющегося оценкой ρ (коэффициент связи в генеральной совокупности), отыскивается по формуле

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{(n_1 n_2 n_3 n_4)^{1/2}}, \qquad (10.1)$$

где $n_1 = n_{11} + n_{12}$; $n_2 = n_{21} + n_{22}$; $n_3 = n_{11} + n_{21}$; $n_4 = n_{12} + n_{22}$. Коэффициент r изменяется от -1 до 1, достигая крайних пределов в таких случаях:

а) $n_{12} = n_{21} = 0$, тогда $r = 1$;

б) $n_{11} = n_{22} = 0$, тогда $r = -1$.

Так как величина r, найденная по выборочным данным, испытывает случайные флуктуации, то вывод о зависимости X и Y не может быть сделан лишь на основании выполнения неравенства $r \neq 0$. Суждение о связи между X и Y в генеральной совокупности будет обоснованным только после проверки гипотезы $H_0 : \rho = 0$; при альтернативе $H_1 : \rho \neq 0$.

Критерий, позволяющий выбрать одну из гипотез, имеет вид: $k = Nr^2$.

Распределение k в условиях нулевой гипотезы удовлетворительно описывается χ^2-распределением с одной степенью свободы. Таким образом, для уровня значимости α можно записать следующие соотношения.

Если $k < \chi^2_{\alpha, 1}$, то принимается H_0, а если $k \geqslant \chi^2_{\alpha, 1}$, то H_0 отклоняется и принимается H_1. Выполнение последнего соотношения свидетельствует о зависимости случайных величин X и Y; теснота этой зависимости оценивается значением коэффициента r, а направление связи (прямая или обратная) — знаком при r.

Кроме коэффициента, вычисляемого по формуле (10.1), в литературе описан ряд других мер связи, пригодных для работы с дихотомическими признаками. Наиболее употребительными являются коэффициент связи Юла Q и коэффициент коллигации Y, также введенный в статистическую практику Д. Юлом [23]:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}; \qquad Y = \frac{1 - (n_{22}/n_{11})^{1/2}}{1 + (n_{22}/n_{11})^{1/2}}.$$

Оба эти коэффициента, как и коэффициент связи r, меняются от -1 до $+1$. Однако предельные значения могут быть достигнуты коэффициентами Q и Y при обращении в нуль хотя бы одной из частот n_{12}, n_{21} ($Q = Y = 1$) или n_{11}, n_{22} ($Q = Y = -1$), что отличает эти меры связи от коэффициента r.

2. Корреляция для порядковых геологических данных. Следующий, более высокий уровень описания свойств геологических объектов соответствует измерениям, выполняемым с помощью порядковых шкал. Последние не только обеспечивают отнесение того или иного наблюдения к определенной категории (классу), но и позволяют упорядочить эти категории, т. е. расположить x_1, x_2, x_3, ... по возрастанию или убыванию степени проявленности, выраженности измеряемого признака. Характерной особенностью порядковой шкалы является отсутствие сведений о величине различия между ее градациями. В лучшем случае разницу между классами удается упорядочить по величине; такую шкалу обычно называют порядково-метрической.

В геологических исследованиях измерениям по порядковой шкале соответствуют, например, полуколичественные и приближенно-количественные спектральные анализы. Первые лишь весьма приблизительно оценивают содержания химических элементов, что проявляется, в частности, в форме записи результатов анализа («очень много», «много», «мало», «следы», «не обнаружен» и т. п.). Данные, полученные приближенно-количественным методом, более детальны, однако и в этом случае «расстояния» между соседними градациями точно не определимы и в связи с этим в значительной мере условны. Тем не менее различия в степени детальности измерения содержаний химических элементов, выполняемых на основе полуколичественного и приближенно-количественного анализов, достаточно существенны. Это дает основание рассматривать раздельно меры связи для полуколичественных данных (назовем их категоризованными упорядоченными данными) и для результатов приближенно-количественных анализов. Последние легко поддаются ранжированию (эта процедура более подробно будет описана ниже), поэтому данные такого рода назовем ранговыми.

А. Категоризованные (упорядоченные) данные [23]. Предположим, что значения случайных величин ξ и η принимают в результате испытания значения, соответственно $\{A_i : i = 1, 2, \ldots, r\}$ и $\{B_j : j = 1, 2, \ldots, S\}$. В общем случае число классов (категорий) r и S не совпадает.

Законы распределения этих величин можно записать в следующем виде:

$$\begin{array}{cccccc} A_1 & A_2 & \ldots & A_i & \ldots & A_r; \\ p(A_1) & p(A_2) & \ldots & p(A_i) & \ldots & p(A_r); \end{array} \qquad \sum_{i=1}^{r} p(A_i) = 1;$$

$$\begin{array}{cccccc} B_1 & B_2 & \ldots & B_j & \ldots & B_S; \\ p(B_1) & p(B_2) & \ldots & p(B_j) & \ldots & p(B_S); \end{array} \qquad \sum_{j=1}^{S} p(B_j) = 1.$$

Их совместные (двумерные) распределения отражает матрица:

$$\begin{pmatrix} p(A_1B_1) & p(A_1B_2) & \ldots & p(A_1B_j) & \ldots & p(A_1B_S) \\ p(A_2B_1) & p(A_2B_2) & \ldots & p(A_2B_j) & \ldots & p(A_2B_S) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ p(A_iB_1) & p(A_iB_2) & \ldots & p(A_iB_j) & \ldots & p(A_iB_S) \\ p(A_rB_1) & p(A_rB_2) & \ldots & p(A_rB_j) & \ldots & p(A_rB_S) \end{pmatrix}.$$

Элементами этой матрицы (таблицы сопряженности) являются вероятности совместного появления определенных значений случайных величин ξ и η. Полезно отметить следующие соотношения:

$$\sum_{i=1}^{r} p(A_iB_j) = p(B_j); \qquad \sum_{j=1}^{S} p(A_iB_j) = p(A_i);$$

$$\sum_{i=1}^{r}\sum_{j=1}^{S} p(A_iB_j) = 1.$$

Необходимым и достаточным условием независимости случайных величин ξ и η является выполнение равенства:

$$p(A_iB_j) = p(A_i)\,p(B_j).$$

В качестве меры зависимости Г. Крамером предложен коэффициент сопряженности, имеющий вид [22]:

$$\varphi^2 = 1/(q-1),$$

где φ^2 — показатель взаимной сопряженности, введенный К. Пирсоном,

$$\varphi^2 = \sum_{i=1}^{r}\sum_{j=1}^{S} \frac{[p(A_iB_j) - p(A_i)\,p(B_j)]^2}{p(A_i)\,p(B_j)}; \qquad q = \min(r,\ S).$$

Свойства коэффициента сопряженности Г. Крамера:

1) $0 \leqslant \varphi^2 \leqslant 1$;

2) $\varphi^2 = 0$ тогда и только тогда, когда случайные величины ξ и η независимы;

3) $\varphi^2 = 1$, если случайные величины ξ и η связаны однозначной функциональной зависимостью.

Если число градаций, т. е. число отличающихся возможных значений ξ и η равно двум ($r = 2$; $S = 2$), то коэффициент сопряженности Крамера, так же как и показатель взаимной сопряженности Пирсона, совпадает с квадратом коэффициента корреляции для дихотомических признаков (см. ниже).

В геологии коэффициент сопряженности применяется при исследовании зависимостей между такими свойствами геологических объектов, которые не могут быть получены в результате наблюдений (так называемые категоризованные данные с относительно небольшим числом градаций). Коэффициент сопряженности незаменим в тех случаях, когда качественные признаки не поддаются упорядочению по самой природе явления. На практике коэффициент сопряженности применяют и для исследования связи между непрерывными случайными величинами, если отсутствуют сведения как о законе их распределения, так и о форме ожидаемой связи между ними. Предположим теперь, что переменная X характеризуется r классами, а переменная Y—S классами:

$$X = \{x_1, x_2, \ldots, x_i, \ldots, x_r\};$$
$$Y = \{y_1, y_2, \ldots, y_j, \ldots, y_S\}.$$

Выборочные данные, используемые для оценки силы и направления связи между величинами X и Y, удобно представить в виде таблицы сопряженности (табл. 8).

Таблица 8

Сопряженность случайных величин X и Y

X	Y					Σ
	y_1	...	y_j	...	y_S	
x_1	n_{11}	...	n_{1j}	...	n_{1S}	$n_{1\cdot}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{iS}	$n_{i\cdot}$
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rj}	...	n_{rS}	$n_{r\cdot}$
Σ	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot S}$	N

В клетки этой таблицы вписаны частоты совместного появления значений случайных величин X и Y, принадлежащих определенным классам. Так, n_{ij} обозначает число наблюдений, в которых зарегистрированы одновременно i-й класс переменной X и j-й класс переменной Y.

Последняя строчка и крайний правый столбец отведены для суммарных частот:

$$n_i = \sum_{j=1}^{S} n_{ij}; \qquad n_j = \sum_{i=1}^{r} n_{ij};$$

$$N = \sum_{i=1}^{r}\sum_{j=1}^{S} n_{ij} = \sum_{j=1}^{S} n_j = \sum_{i=1}^{r} n_i.$$

Выборочную меру связи вычислим по формуле:

$$r_{\text{гк}} = \frac{\sum_{i>i'}^{r}\sum_{j>j'}^{S}(n_{ij}n_{i'j'} - n_{ij'}n_{i'j})}{N^2 - \sum_{i=1}^{r} n_i^2 - \sum_{j=1}^{S} n_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{S} n_{ij}}.$$

Введение дополнительных индексов $i' = 1, 2, \ldots, r-1$; $j' = 1, 2, \ldots, S-1$ и условия $i > i'$, $j > j'$ позволяет обособлять в таблице сопряженности блоки, соответствующие событиям типа «не x_i», «не y_j».

Коэффициент связи $r_{\text{гк}}$, введенный в практику статистических исследований Гудмэном и Красклом, изменяется в интервале от -1 до $+1$ [23]. Если величины X и Y связаны обратной зависимостью, коэффициент $r_{\text{гк}}$ отрицателен, в противном случае — положителен. Чем слабее связь, тем ближе величина $r_{\text{гк}}$ к нулю.

Оценку значимости коэффициента связи $r_{\text{гк}}$, т. е. проверку гипотезы $H_0 : \rho = 0$ выполним путем сравнения выборочного коэффициента $r_{\text{гк}}$ с величиной его стандартного отклонения. Верхняя граница дисперсии коэффициента $r_{\text{гк}}$ может быть найдена из следующего соотношения:

$$D(r_{\text{гк}}) \leqslant \frac{2N(1 - r_{\text{гк}}^2)}{N^2 - \sum_{i=1}^{r} n_{i\cdot}^2 - \sum_{j=1}^{S} n_{\cdot j}^2 + \sum_{i=1}^{r}\sum_{j=1}^{S} n_{ij}^2}.$$

Так как распределение $\sqrt{N}(\rho - r_{\text{гк}})$ асимптотически нормально, то нулевая гипотеза может быть проверена с помощью доверительных интервалов, вычисляемых для $r_{\text{гк}}$ при определенном уровне значимости α. Если принять, например, $\alpha = 0{,}05$, то $\Phi(\alpha) = 1{,}96$, а доверительные границы составляют, соответственно,

$$r_{\text{гк}} - 1{,}96\sqrt{D(r_{\text{гк}})} \quad \text{и} \quad r_{\text{гк}} + 1{,}96\sqrt{D(r_{\text{гк}})}.$$

Гипотеза H_0 отвергается при уровне значимости α, если $\rho = 0$ не принадлежит интервалу

$$r_{\text{гк}} \pm \Phi(\alpha)\sqrt{D(r_{\text{гк}})}.$$

Еще одной выборочной мерой связи, употребляемой при работе с категоризованными упорядоченными данными, является коэффициент t_c [22], рассчитываемый по формуле:

$$t_c = \frac{m \cdot 2 \sum_{i>i'}^{r}\sum_{j>j'}^{S}(n_{ij}n_{i'j'} - n_{ij'}n_{i'j})}{N^2(m-1)},$$

где $m = \min(r, S)$.

Коэффициент связи t_c вычисляется несколько проще, чем $r_{\text{гк}}$ (в этом его преимущество); однако для достижения предельных значений $+1$, соответствующих строго линейной зависимости, необходимо, чтобы N было кратно m. Оценка значимости выборочного коэффициента t_c, так же как и для $r_{\text{гк}}$, выполняется путем построения доверительных интервалов. Верхняя граница дисперсии t_c определяется из соотношения:

$$D(t_c) \leqslant \frac{2}{N}\left[\left(\frac{m}{m-1}\right)^2 - t_c^2\right].$$

Следует заметить (это касается и коэффициента $r_{\text{гк}}$), что, оперируя при проверке $H_0 : \rho = 0$ верхней границей дисперсии, мы тем самым несколько расширяем область принятия нулевой гипотезы — мера вынужденная, но необходимая, так как точное распределение обсуждаемых статистик связи неизвестно.

Если число градаций (классов) признака достаточно велико, то пользоваться таблицей сопряженности и соответствующими коэффициентами связи неудобно из-за громоздкости первых и трудоемкости вычисления вторых. Поэтому при работе с данными, измеренными по порядковой шкале с высокой степенью детальности, оценивание тесноты и направления линейной связи выполняется с помощью так называемых ранговых коэффициентов.

Б. Ранжируемые данные. Опишем вкратце процедуру ранжирования. Расположим значения переменной X в порядке их возрастания:

$$x_1 < x_2 < \ldots < x_i < \ldots < x_N$$

и определим ранг величины x_i как ее номер в этом упорядоченном ряду, т. е. ранг $(x_i) = i$. Для переменной Y ранжирование выполняется аналогичным образом. При повторяющихся значениях для последних вычисляется усредненный ранг. Так, если в упорядоченном ряду значения $x_i, x_{i+1}, \ldots, x_{i+m}$ оказались одинаковыми (что связано с недостаточной тонкостью измерений и с неизбежными огрублениями результатов измерений), то всем этим значениям приписывается один общий ранг, равный $[i + (i + \overline{m})]/2$. В дальнейшем множество рангов, соответствующих переменной X, будем обозначать R_X, а соответствующих переменной Y обозначим R_Y.

Ранговую меру связи определим так [23]:

$$r_c = \frac{\operatorname{cov}(R_X, R_Y)}{S_X S_Y} = \frac{\frac{1}{N}\sum_{i=1}^{N} R_{iX}R_{iY} - \overline{R}_X\overline{R}_Y}{S_X S_Y},$$

где $\overline{R}_X$, $\overline{R}_Y$ — средние арифметические рангов, приписанных соответствующим значениям переменных X и Y; S_X, S_Y — средние квадратические отклонения этих рангов.

Так как $\{R_{iX}\}$ и $\{R_{iY}\}$ — перестановки первых N натуральных чисел, то $\overline{R}_X = \overline{R}_Y = (N+1)/2$, а $S_X^2 = S_Y^2 = (N^2-1)/12$. После некоторых преобразований получаем упрощенный вариант формулы:

$$r_c = 1 - \frac{6\sum_{i=1}^{N}(R_{iX} - R_{iY})^2}{N^3 - N}. \qquad (10.2)$$

Показатель связи r_c был введен в статистику К. Спирменом, отсюда его название — ранговый коэффициент корреляции Спирмена.

Оценка существенности этого наиболее употребительного рангового коэффициента производится обычно путем сравнения выборочного значения r_c с предельно допустимым значением r_α, величина которого регулируется как числом наблюдений N, так и уровнем значимости α. В литературе имеются таблицы точного распределения коэффициента Спирмена, однако они не совсем удобны для пользования. Для практических приложений полезно иметь таблицы допустимых значений r_α, рассчитанных для общепринятых уровней значимости. Существует несколько методов нахождения таких значений, различие между ними определяется типом аппроксимирующей функции.

Процедура проверки гипотезы $H_0 : \rho(R_X, R_Y) = 0$ при множестве альтернатив $H_1 : \rho(R_X, R_Y) \neq 0$ сводится к сопоставлению r_c и r_α. При $r_c < r_\alpha$ H_0 принимается, а при $r_c \geqslant r_\alpha$ H_0 отклоняется.

Если среди выборочных данных, используемых при вычислении коэффициента корреляции Спирмена, имеется много совпадающих значений, то в формулу (10.2) необходимо ввести поправки T_X и T_Y. Для переменной X, например, величина корреляции определяется следующим образом [23]:

$$T_X = \sum_{j=1}^{m} \frac{t_j^3 - t_j}{12},$$

где m = число групп, охарактеризованных единым (усредненным) рангом; t_j — объем j-й группы. Аналогично вычисляется поправка T_Y.

Теперь можно записать уточненную формулу для нахождения r_c [23]:

$$r_c = \frac{(N^3 - N)/6 - \sum_{i=1}^{N}(R_{iX} - R_{iY})^2 T_X - T_Y}{[(N^3 - N)/6 - 2T_X]^{1/2}[(N^3 - N)/6 - 2T_Y]^{1/2}}.$$

Как и большинство коэффициентов связи, коэффициент корреляции Спирмена изменяется в интервале от -1 до $+1$, достигая крайних значений в случаях либо полной согласованности обоих рядов — R_X и R_Y, либо их полной несогласованности. С этих позиций ранговые коэффициенты являются, по сути дела, коэффициентами, выражающими степень неупорядоченности значений одной величины относительно другой. Эта особенность ранговых мер связи особенно четко выражена в коэффициенте Кендалла, вычисляемого по формуле [23]:

$$r_к = 1 - \frac{2Q}{0{,}5N(N-1)},$$

где Q — число инверсий между рангами R_{iY}, характеризующее степень беспорядка рангов величины Y при условии, что ранги переменной X расположены в натуральном порядке: 1, 2, . . . , N.

Число таких инверсий, т. е. нарушений в фиксированном ряду натурального порядка, может меняться от 0 до $0{,}5N(N-1)$. Если $Q = 0$, то ранги R_X и R_Y полностью согласованы и $r_к = 1$; если же $Q = 0{,}5N(N-1)$ (что соответствует обратному относительно R_X порядку Y рангов), то $r_к = -1$. Заметим, что коэффициенты Кендалла и Спирмена практически эквивалентны в условиях H_0. Значения $r_к$ и r_c совпадают лишь в ситуациях полной согласованности или несогласованности последовательностей R_{iX} и R_{iY}. В остальных случаях $|r_c|$ обычно превышает $|r_к|$, что связано с различиями в процедуре измерения расхождений между R_X и R_Y. При подсчете коэффициента r_c инверсиям более удаленных друг от друга ранжируемых значений X (или Y) приписываются большие веса.

Оценку значимости выборочного коэффициента корреляции Кендалла нетрудно выполнить, если учесть, что распределение $r_к/\sqrt{D(r_к)}$ в условиях нулевой гипотезы с увеличением N стремится к стандартному нормальному распределению. Дисперсия коэффициента $r_к$ зависит только от объема выборки:

$$D(r_к) = 2(2N+5)/9N(N-1).$$

Уже при $N \geqslant 10$ нормальное приближение при проверке гипотезы о некоррелированности рангов дает вполне удовлетворительные результаты.

Коэффициент корреляции Кендалла как оценка тесноты связи употребляется гораздо реже, чем коэффициент Спирмена, что связано как с трудоемкостью его вычисления, так и с нелинейностью самой статистики $r_к$. Кроме того, как отмечает Ван дер Варден,

коэффициент Спирмена при фиксированном уровне значимости имеет несколько большую мощность.

В тех случаях, когда есть основания предполагать нормальность распределения случайных величин X и Y, эффективность ранговых критериев может быть повышена, если заменить ранги R_X и R_Y их математическими ожиданиями в выборке объема N из совокупности со стандартным нормальным распределением. Получаемые таким путем условные числа называют нормальными метками. Вслед за Я. Гаеком и Э. Шидаком определим метку i-го значения величины X (последние упорядочены, например, по убыванию):

$$a_N(i, f_x) \approx \varphi(M_{ix}, f_x),$$

где f_x — плотность распределения порядковых статистик $\{x_i\}$; M_{ix} — математическое ожидание i-й порядковой статистики в условиях распределения f_x.

Если f непрерывна, то локально наиболее мощный критерий проверки гипотезы независимости может быть построен на основе суммы $\sum_{i=1}^{N} a_N(R_{ix}, f_x)\, a_N(R_{iy}, f_y)$. Дополнительно условившись, что выборка объемом N извлечена из нормальной совокупности с нулевым средним значением и стандартным отклонением, равным единице, получим коэффициент корреляции r_Φ, известный в литературе как коэффициент Фишера — Иэйтса [23]:

$$r_\Phi = \frac{\sum_{i=1}^{N} M(R_{ix}, N)\, M(R_{iy}, N)}{\sum_{i=1}^{N} [M(i, N)]^2},$$

где $M(R_{ix}, N)$ — математическое ожидание i-го ранга случайной величины X; $M(R_{iy}, N)$ — то же i-го ранга случайной величины Y; (i, N) — упорядоченная последовательность целых чисел $i = 1, 2, \ldots, N$.

Значения математических ожиданий порядковых статистик табулированы. К сожалению, эти таблицы малодоступны, поэтому при оценке направления и тесноты связи с помощью нормальных меток удобно воспользоваться коэффициентом корреляции Ван дер Вардена [23]:

$$r_B = \frac{\sum_{i=1}^{N} \psi\left(\frac{R_{ix}}{N+1}\right)\psi\left(\frac{R_{iy}}{N+1}\right)}{\sum_{i=1}^{N}\left[\psi\left(\frac{i}{N+1}\right)\right]^2}.$$

Здесь точные значения нормальных меток заменены их приближенными значениями, для нахождения которых достаточно обратиться к широко распространенным таблицам обратной функции стандартного нормального распределения. Иногда эти таблицы сопровождаются значениями $\sum_{i=1}^{N}\left[\psi\left(\frac{i}{N+1}\right)\right]^2$, что заметно облегчает вычисление r_B.

Коэффициенты Фишера—Иэйтса и Ван дер Вардена практически эквивалентны. Оценка их значимости может быть выполнена без большой потери точности с помощью стьюдентовского распределения.

3. Количественная геологическая информация. Современные геологические исследования в большинстве своем опираются на измерения, выполненные по интегральной шкале, что позволяет количественно оценить величину отличия одной степени проявления признака от другой. Если к тому же удается указать на этой шкале нулевую точку, то интегральная шкала преобразуется в шкалу отношений. Измерения, выполненные по шкале отношений (ее еще называют пропорциональной шкалой), относятся к наиболее высокому уровню измерений. Таковы, например, количественные химические и спектральные анализы горных пород, руд, минералов и других геологических объектов. Данные, получаемые в результате таких анализов, допускают, в отличие от качественных и порядковых измерений, использование при их обработке любых арифметических действий. Это обстоятельство позволяет при вычислении выборочных коэффициентов связи опираться не на частоты или ранги, а непосредственно на значения коррелируемых случайных величин. Тем самым обеспечивается полнота извлечения из результатов наблюдений необходимой, с точки зрения решаемой проблемы, информации.

С вероятностных позиций количественно измеренные геохимические признаки можно рассматривать как непрерывные случайные величины. Параметрический коэффициент парной корреляции — числовая характеристика силы линейной связи между случайными величинами. Коэффициент парной корреляции определяется как ковариация, нормированная стандартными отклонениями σ_ξ и σ_η, случайных величин ξ и η [2, 8, 19, 23]:

$$\rho(\xi, \eta) = \operatorname{cov}(\xi, \eta)/\sigma_\xi\sigma_\eta.$$

Другие формы выражения коэффициента парной корреляции:

$$\rho(\xi, \eta) = \frac{M(\xi\eta) - M\xi M\eta}{\sqrt{D\xi D\eta}};$$

$$\rho(\xi, \eta) = \frac{M[(\xi - M\xi)(\eta - M\eta)]}{\sqrt{D\xi D\eta}}.$$

Свойства коэффициента парной корреляции:

1) для любых случайных величин ξ и η, у которых $\sigma_\xi \neq 0$ и $\sigma_\eta \neq 0$, выполнено неравенство $-1 \leqslant \rho(\xi_1\eta) \leqslant 1$. Если $\rho(\xi, \eta) > 0$, то случайные величины ξ и η называются положительно коррелированными, если $\rho(\xi, \eta) < 0$ — отрицательно коррелированными. Таким образом, коэффициент парной корреляции характеризует не только силу, но и направление связи;

2) $\rho(\xi, \eta) = \rho(\eta, \xi)$;

3) если $\xi = a + b\eta$, где a и b — константы, то $\rho(\xi, \eta) = 1$;

4) если ξ и η — независимые случайные величины, то $\mathrm{cov}(\xi, \eta) = 0$ и, следовательно, $\rho(\xi, \eta) = 0$.

Равенство $\rho(\xi, \eta) = 0$ является необходимым и достаточным условием независимости ξ и η лишь в том случае, если двумерная случайная величина (ξ, η) нормально распределена. Если вид распределения (ξ, η) неизвестен, то при выполнении равенства $\rho(\xi, \eta) = 0$ говорят о некоррелированности случайных величин ξ и η.

Выборочный коэффициент корреляции r — оценка коэффициента корреляции случайных величин ξ и η — моделей геологических характеристик по выборочным данным следующего вида:

$$r = \frac{\hat{\mathrm{cov}}(\xi, \eta)}{\hat{\sigma}_\xi \hat{\sigma}_\eta} = \frac{\sum_{i=1}^{n} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{\sqrt{\sum_{i=1}^{n} (x_{ti} - \bar{x}_i)^2 \sum_{i=1}^{n} (x_{tj} - \bar{x}_j)^2}}.$$

Последнее выражение можно переписать в виде

$$r = \frac{\sum_{i=1}^{n} x_{ti}x_{tj} - \frac{1}{n}\left(\sum_{i=1}^{n} x_{ti}\right)\left(\sum_{i=1}^{n} x_{tj}\right)}{\sqrt{\left[\sum_{i=1}^{n} x_{ti}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{ti}\right)^2\right]\left[\sum_{i=1}^{n} x_{tj}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{tj}\right)^2\right]}},$$

позволяющем резко уменьшить число операций при вычислении r и тем самым уменьшить опасность потери точности оценки, полученной вручную или с помощью микрокалькулятора.

При малых объемах наблюдений n значение r получается заниженным по сравнению с истинным значением коэффициента корреляции. Поэтому при $n < 10$ для r следует использовать оценку r^*:

$$r^* = r\left[1 + \frac{1 - r^2}{2(n - 3)}\right].$$

Проверка значимости r, т. е. проверка гипотезы о том, что в генеральной совокупности истинная корреляция равна нулю $[H_0 : \rho(\xi, \eta) = 0]$, осуществляется с помощью специальных таблиц процентных точек выборочного коэффициента корреляции $r_{\alpha, \nu}$, вычисленных при условии что ξ и η распределены по двумерному нормальному закону $\rho(\xi, \eta) = 0$ [8, 19]. Эти таблицы имеют два входа — число степеней свободы ν равно числу наблюдений n, уменьшенному на 2, т. е. $\nu = n-2$. Уровень значимости α принимается обычно 5 %. Задав уровень значимости и определив число степеней свободы, находят соответствующее критическое значение $r_{\alpha, \nu}$. Нулевая гипотеза $H_0 : \rho = 0$ отклоняется, если $r > r_{\alpha, \nu}$, и принимается как подтвердившаяся, если $r \leqslant r_{\alpha, \nu}$.

Значимость выборочного коэффициента корреляции может быть оценена также с помощью таблиц квантилей распределения Стью-

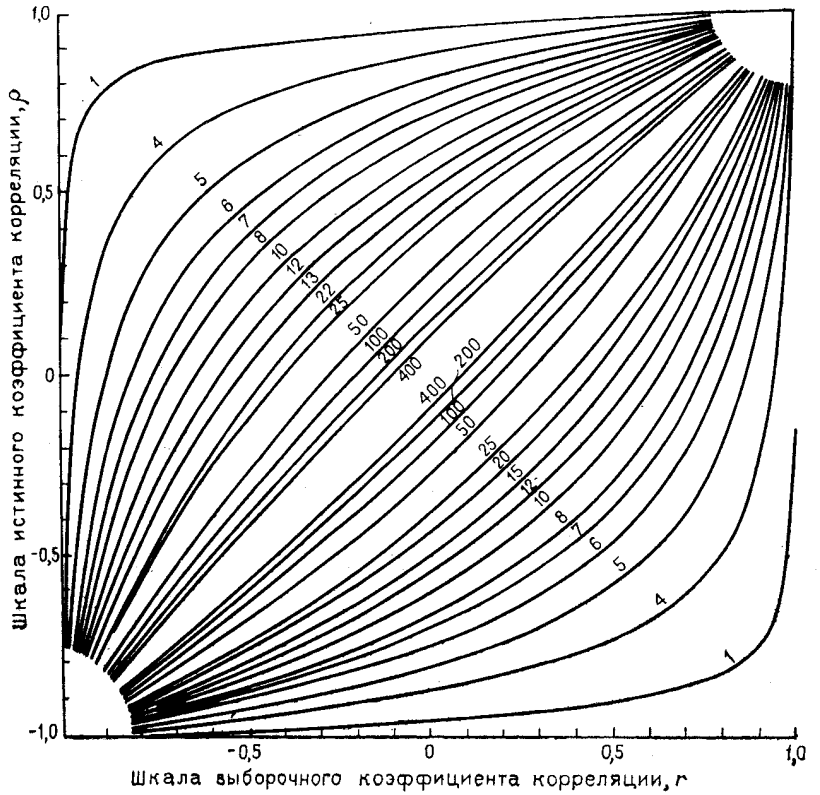


Рис. 50. 95 %-ные доверительные пределы для коэффициента корреляции. Цифры на кривых указывают объемы выборок

дента и F-распределения, а именно статистика $t = r\sqrt{n-2}/\sqrt{1-r^2}$ в условиях нулевой гипотезы $H_0 : \rho = 0$ распределена по закону Стьюдента с $\nu = n-2$ степенями свободы. Поэтому если $t > t_{\alpha, \nu}$, то нулевая гипотеза отклоняется, а если $t \leqslant t_{\alpha, \nu}$, то принимается как подтвердившаяся. Здесь α — уровень значимости, $t_{\alpha, \nu}$ — соответствующее значение квантиля распределения Стьюдента.

Аналогично, статистики $F = r^2(n-2)/(1-r^2)$ или $F' = (1 + r)/(1-r)$ в условиях нулевой гипотезы $H_0 : \rho = 0$ имеют F-распределения со степенями свободы: первая — 1 и $n-2$, вторая соответственно $n-2$ и $n-2$.

Доверительный интервал для коэффициента корреляции может быть определен с помощью графика 95 %-ных доверительных границ (рис. 50, 51) [8]. Для этого по оси абсцисс откладывается значение r выборочного коэффициента корреляции (см. рис. 50). Из этой точки восставляется перпендикуляр, пересекающий совокупность кривых линий. Числа на кривых означают объемы выборки. Находят точки пересечения перпендикуляра с двумя кривыми, со-

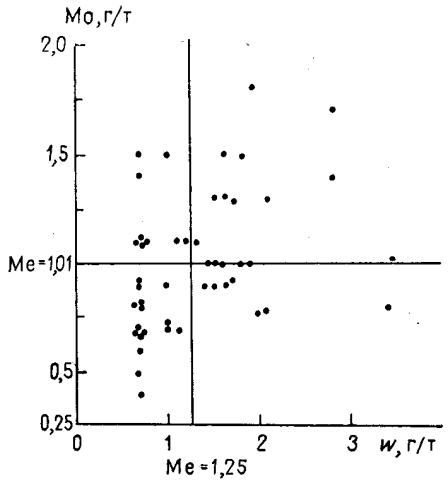


Рис. 51. Корреляционное поле и вычисление тетрахорического коэффициента корреляции Бломквиста

ответствующими заданному объему выборки. Ординаты этих точек пересечения принимаются за значения доверительных границ.

С помощью графика 95 %-ных доверительных границ нетрудно проверить значимость выборочного коэффициента корреляции. Если полученные с помощью рис. 50 доверительные границы не содержат нуля, то гипотеза $H_0 : \rho = 0$ должна быть отклонена, а если содержат нуль, то принята как подтвердившаяся.

Р. А. Фишер предложил следующее преобразование случайной величины r:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Распределение z при $n \geqslant 20$ близко к нормальному с моментами:

$$Mz = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-3)} \left[1 - \frac{3-\rho^2}{4(n-3)} + \cdots \right];$$

$$Dz = \frac{1}{n-3} \left[1 - \frac{\rho^2}{2(n-3)} - \frac{2-6\rho^2+3\rho^4}{6(n-3)^2} + \cdots \right].$$

Случайная величина $(z - Mz)/\sqrt{Dz}$ имеет приближенно нормальное распределение с параметрами (0, 1).

При значении r, близком к единице ($0{,}97 < r < 1$), формулу для z следует изменить на следующую [8]:

$$z = -\frac{1}{2} \left(\frac{1-r}{2} + \ln \frac{1-r}{2} \right).$$

Преобразование Фишера z позволяет построить доверительный интервал для коэффициента корреляции ρ с коэффициентом доверия $1-2\alpha$:

$$\frac{e^{2z_1}-1}{e^{2z_1}+1} < \rho < \frac{e^{2z_2}-1}{e^{2z_2}+1},$$

где

$$z_1 = \frac{1}{2} \ln \frac{1+r}{1-r} - \frac{\psi(1-\alpha)}{\sqrt{n-3}},$$

$$z_2 = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{\psi(1-\alpha)}{\sqrt{n-3}},$$

$\psi(p)$ есть p-квантиль нормального распределения [8].

Проиллюстрируем процедуру вычисления выборочного коэффициента корреляции, оценки его статистической значимости и нахождения 95 %-ного доверительного интервала для коэффициента корреляции по содержаниям рудных, редких и петрогенных элементов на поверхностном эрозионном срезе гранитного массива Эльджурту (Северный Кавказ). В работе [26] приводятся данные по содержаниям 19 элементов, из которых мы отбираем два столбца, соответствующие содержаниям вольфрама и молибдена:

1) находим средние по этим элементам:

$$\bar{x}_W = 1{,}462, \qquad \bar{x}_{Mo} = 1{,}048;$$

2) оцениваем соответствующие дисперсии:

$$S^2_W = 0{,}969, \qquad S^2_{Mo} = 0{,}121;$$

3) вычисляем стандартные отклонения:

$$S_W = 0{,}984, \qquad S_{Mo} = 0{,}348;$$

4) оцениваем ковариацию W и Mo:

$$\operatorname{cov}(W, Mo) = 0{,}193;$$

5) определяем значение выборочного коэффициента корреляции r:

$$r = \operatorname{cov}(W, Mo)/S_W S_{Mo} = 0{,}563;$$

6) оцениваем значимость r. В работе [8] из таблицы, задав уровень значимости $\alpha = 5\ \%$ и определив число степеней свободы $\nu = n-2 = 48$, находим критическое значение $r_{\alpha,\nu} = 0{,}273$. Поскольку $r = 0{,}563 > r_{\alpha,\nu} = 0{,}273$, то нулевая гипотеза $H_0 : \rho = 0$ отклоняется в пользу набора альтернатив $H_1 : \rho \neq 0$, т. е. найденное значение коэффициента корреляции статистически значимо при 5 %-ом уровне;

7) определяем 95 %-ный доверительный интервал для ρ (W, Mo). Для этого на графике 95 %-ных доверительных границ отложим на оси абсцисс значение $r = 0{,}563$ и восставим в этой точке перпендикуляр. Находим точки пересечения перпендикуляра с кривыми, соответствующими объему выборки $n = 50$. Определяем ординаты точек пересечения; они равны 0,35 и 0,7. Следовательно, с доверительной вероятностью 95 %-ный интервал (0,35; 0,7) накрывает истинное значение коэффициента корреляции ρ.

Поскольку доверительный интервал не содержит нулевого значения, то еще раз убеждаемся, что найденное значение r статистически значимо при 5 %-ном уровне, т. е. истинный коэффициент корреляции $\rho \neq 0$.

4. Нелинейная корреляция. В тех случаях когда есть основания предполагать, что связь между исследуемыми переменными нелинейна, оценку тесноты связи следует выполнить с помощью корреляционного отношения или же коэффициента сопряженности [19, 23].

Оценка корреляционного отношения вычисляется следующим образом. Пусть исследуется зависимость переменной I от X. Ра-

зобьем все множество наблюдаемых значений X на k классов (интервалов) и подсчитаем для каждого j-го класса среднее

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij},$$

где n_j — число точек, попавших в j-й интервал, $\sum_{j=1}^{k} n_j = N$; y_{ij} — значения переменной Y, принадлежащие j-му интервалу группирования. Оценка корреляционного отношения $r_{Y/X}$ находится по формуле

$$r_{Y/X} = \left[\frac{\sum_{j=1}^{k} (\bar{Y}_j - \bar{Y})^2 n_j}{(N-1) S_Y^2} \right]^{1/2},$$

где $\bar{Y}$ — оценка среднего; S_Y^2 — оценка дисперсии переменной Y.

Величина $r_{Y/X}$ меняется в пределах от 0 до 1. Заметим, что в общем случае $r_{Y/X} \neq r_{X/Y}$.

Проверка гипотезы $H_0 : \rho_{Y/X} = 0$ осуществляется на основе критерия

$$F = \frac{(N-k) r_{Y/X}^2}{(k-1)\left(1 - r_{Y/X}^2\right)},$$

имеющего в условиях H_0 F-распределение с k—1 и N—k степенями свободы. Если $F > F_{\alpha, k-1, N-k}$, то нулевая гипотеза отвергается при уровне значимости α.

Разность $r_{Y/X}^2 - r^2$, где r — коэффициент линейной корреляции, может служить мерой линейности связи. Чтобы сделать статистически обоснованный вывод о существенной нелинейности исследуемой зависимости, необходимо вычислить критерий

$$V_2 = \frac{N-k}{k-2} \cdot \frac{r_{Y/X}^2 - r^2}{1 - r_{Y/X}^2}$$

и сравнить V_2 с $F_{\alpha, k-2, N-k}$. Если V_2 ниже допустимого при данном значении α F-распределения, то нет никаких оснований отказываться от линейной модели.

Понятие корреляционного отношения может быть обобщено на многомерный случай.

Если значения переменных Y и X сведены в таблицу сопряженности, описанную ранее, то проверку гипотезы о независимости можно выполнить с помощью критерия [19, 23]

$$k_{rXS} = N\left(\sum_{i=1}^{r} \sum_{j=1}^{S} \frac{n_{ij}}{n_{i\cdot} n_{\cdot j}} - 1 \right).$$

В условиях гипотезы о независимости величина k_{rXS} удовлетворительно аппроксимируется χ^2-распределением с $(r-1)$ $(S-1)$ степенями свободы. Из

$$k_{rXS} > \chi^2_{\alpha, (r-1)(S-1)}$$

следует принятие гипотезы о зависимости переменных Y и X.

Оценка тесноты связи вычисляется по формуле [19, 23]:

$$C = \left[\frac{k_{rXS}}{N(q-1)} \right]^{1/2},$$

где C — коэффициент сопряженности (связанности), введенный в статистику Г. Крамером; $q = r$, если $r \leqslant S$, и $q = S$, если $r > S$. Коэффициент C меняется в пределах от 0 до 1.

ЧАСТНАЯ, МНОЖЕСТВЕННАЯ И КАНОНИЧЕСКАЯ КОРРЕЛЯЦИЯ

Частный коэффициент корреляции. При исследовании взаимосвязи случайных величин ξ_k и ξ_l, входящих в систему $\{\xi_j : j = 1, \ldots, m\}$, часто возникает подозрение, что величина парного коэффициента корреляции ρ_{kl} определяется не столько степенью взаимозависимости величин ξ_k и ξ_l, сколько согласованным воздействием на них остальных образующих систему величин. Метод частной корреляции позволяет произвести «очистку» коэффициента корреляции ρ_{kl} от влияния остальных величин, входящих в систему. Числовой характеристикой такой «очищенной» связи является частный коэффициент корреляции, измеряющий тесноту и направление связи между ξ_k и ξ_l при фиксированных значениях величин $\{\xi_j : j = l, \ldots, m;\ j \neq k;\ j \neq l\}$.

Если задана матрица парных коэффициентов корреляции

$$R = \{\rho_{kl} : k = 1, \ldots, m;\ \ l = 1, \ldots, m\},$$

то частный коэффициент корреляции между случайными величинами ξ_k и ξ_l при фиксированных значениях всех остальных переменных определяется по формуле [23]:

$$\rho_{kl}^{*} = \frac{A_{kl}}{(A_{kk} A_{ll})^{1/2}},$$

где A_{kl} — алгебраическое дополнение для элемента ρ_{kl} матрицы R:

$$R = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1j} & \ldots & \rho_{1m} \\ \rho_{21} & 1 & \ldots & \rho_{2j} & \ldots & \rho_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{j1} & \rho_{j2} & \ldots & 1 & \ldots & \rho_{jm} \\ \rho_{m1} & \rho_{m2} & \ldots & \rho_{mj} & \ldots & 1 \end{pmatrix}.$$

Величина частного коэффициента корреляции, как и любого линейного коэффициента корреляции, меняется в пределах от —1 до 1.

Если $m = 3$, то вычисление коэффициента корреляции частного значительно упрощается:

$$\rho^{*}_{kl\cdot S} = \frac{\rho_{kl} - \rho_{kS}\rho_{lS}}{\sqrt{(1-\rho^2_{kS})(1-\rho^2_{lS})}}.$$

В геологии, возможности которой в области активного эксперимента весьма ограничены, частная корреляция является одним из эффективных методов исследования взаимоотношения между компонентами сложных природных систем и параметрами внешней среды.

Выборочный частный коэффициент корреляции есть оценка $r_{ij\cdot q}$ по выборочным данным коэффициента частной корреляции случайных величин ξ_i и ξ_j при фиксированных m—2 величинах $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_m$ (если $i < j$) следующего вида:

$$r_{ij\cdot q} = -\frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}},$$

где B_{ij} — алгебраическое дополнение выборочной корреляционной матрицы $\{r_{ij}\}$ случайных величин $\xi_1, \ldots, \xi_m$, соответствующее элементу r_{ij}; r_{ij} — выборочный коэффициент корреляции случайных величин ξ_i и ξ_j, $i, j = 1, 2, \ldots, m$, q — набор индексов 1, 2, . . . , m без i и j.

Распределение $r_{ij\cdot q}$, построенного по n наблюдениям, совпадает с распределением выборочного коэффициента парной корреляции r_{ij} с заменой числа степеней свободы n—2 на n—m.

Для проверки значимости выборочного коэффициента частной корреляции применимы все способы проверки значимости выборочного коэффициента парной корреляции с единственным изменением — уменьшением числа степеней свободы на m—2. Для построения доверительных интервалов для коэффициента частной корреляции применимы все способы, описанные для коэффициента парной корреляции, с заменой числа степеней свободы n—2 на n—m.

Проиллюстрируем процедуру вычисления выборочного коэффициента частной корреляции по содержаниям рудных редких и петрогенных элементов на поверхностном эрозионном срезе гранитного массива Эльджурту (Северный Кавказ) [26]. Вычислим выборочный коэффициент частной корреляции для содержаний вольфрама и молибдена, очищенных от влияния содержаний железа. Выборочные коэффициенты парной корреляции равны: $r_{12} = 0{,}563$; $r_{13} = -0{,}125$; $r_{23} = -0{,}204$.

Выборочная корреляционная матрица имеет вид

$$\begin{bmatrix} 1 & 0{,}563 & -0{,}125 \\ 0{,}563 & 1 & -0{,}204 \\ -0{,}125 & -0{,}204 & 1 \end{bmatrix}.$$

Алгебраические дополнения этой матрицы равны

$$B_{12} = -\begin{vmatrix} 0{,}563 & -0{,}204 \\ -0{,}125 & 1 \end{vmatrix} = -(0{,}563 - 0{,}125\cdot 0{,}204)$$

$$B_{11} = \begin{vmatrix} 1 & -0{,}204 \\ -0{,}204 & 1 \end{vmatrix} = 1 - (0{,}204)^2$$

$$B_{22} = \begin{vmatrix} 1 & -0{,}125 \\ -0{,}125 & 1 \end{vmatrix} = 1 - (-0{,}125)^2.$$

Так что

$$r_{12\cdot 3} = \frac{0{,}563 - 0{,}125\cdot 0{,}204}{\sqrt{[1-(-0{,}204)^2]\,[1-(-0{,}125)^2}}.$$

Процедура проверки значимости найденного коэффициента и построение доверительного интервала идентична аналогичной процедуре, описанной в разделе выборочного парного коэффициента корреляции.

Множественная корреляция. Коэффициент множественной корреляции — мера линейной зависимости случайной величины ξ_k от совокупности случайных величин: $\{\xi_l : l = 1, 2, \ldots, m;\ l \neq k\}$.

Коэффициент множественной корреляции определяется формулой

$$\rho^2_{k\,\{l=1,\ \ldots\ m;\ l\neq k\}} = 1 - \frac{|R|}{A_{kk}},$$

где $|R|$ — определитель корреляционной матрицы R, имеющий размерность $m \times m$ (см. частный коэффициент корреляции); A_{kk} — алгебраическое дополнение для k-го элемента той же матрицы.

Свойства коэффициента множественной корреляции:

1) $0 \leqslant \rho^2_{k\,\{\cdot\}} \leqslant 1$; 2) $\rho^2_{k\,\{\cdot\}} = 0$

тогда и только тогда, когда $\rho_{kl} = 0$ для любого l.

3) $\rho^2_{\{\cdot\}} = 1$, если ξ_k является строго линейной функцией от $\{\xi_l : l = 1, \ldots, m;\ l \neq k\}$.

Так как $\rho^2_{k\,\{\cdot\}} \geqslant \rho^2_{kl}$, то равенство коэффициента множественной корреляции единице выполняется всегда, когда абсолютное значение хотя бы одного из парных коэффициентов корреляции с первым индексом k равно 1.

Множественная корреляция широко применяется в геологических исследованиях, в геологических задачах количественного прогнозирования таких геологических признаков, прямые измерения которых либо затруднительны по техническим причинам, либо невыгодны по экономическим соображениям.

Выборочным коэффициентом множественной корреляции

$$R_{i\cdot 1,\ \ldots,\ i-1,\ i+1,\ \ldots,\ m}$$

между случайной величиной ξ_i и набором $\xi_1, \xi_2, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$ называется величина

$$R_{i\cdot 1,\ \ldots,\ i-1,\ i+1,\ \ldots,\ m} = \sqrt{1 - \frac{1}{\hat{C}^{ii}}},$$

где $\hat{C}^{ii}$ — диагональный элемент матрицы $\{\hat{C}^{-1}\}$, обратной матрице выборочных коэффициентов корреляции.

Для проверки статистической гипотезы H_0 о равенстве нулю коэффициента множественной корреляции

$$H_0 : \rho^2_{i \cdot 1, \ldots, i-1, i+1, \ldots, m} = 0$$

при множестве альтернатив

$$H_1 : \rho^2_{i \cdot 1, \ldots, i-1, i+1, \ldots, m} \neq 0$$

вычисляется величина

$$F = \frac{(n-m) R^2_{i \cdot 1, \ldots, i-1, i+1, \ldots, m}}{(m-1)\left(1 - R^2_{i \cdot 1, \ldots, i-1, i+1, \ldots, m}\right)},$$

имеющая в условиях нулевой гипотезы F-распределение с $m-1$ и $n-m$ степенями свободы.

При уровне значимости α по таблицам квантилей F-распределения находят $F_{\alpha, m-1, n-m}$ критическое значение F-распределения с $m-1$ и $n-m$ степенями свободы. Тогда, если $F > F_{\alpha, m-1, n-m}$, то гипотеза H_0 отклоняется, в противном случае она принимается как подтвердившаяся.

Каноническая корреляция [23] измеряет силу связи между множествами случайных величин. Пусть $X^{(1)} = \{X_i,\ i = 1, 2, \ldots, k\}$,

$$X^{(2)} = \{X_j,\ j = k+1,\ k+2, \ldots, l\},$$

$$X^{(1)} \cap X^{(2)} \neq \varnothing, \qquad X = X^{(1)} \cup X^{(2)}.$$

Положим $p_1 = k$, $p_2 = l-k$, $p = k + l$ и условимся, что $p_1 \leqslant p_2$.

Корреляционную матрицу R размерностью $p \times p$ разобьем на блоки: R_{11} — матрица $p_1 \times p_1$ парных коэффициентов корреляции между элементами подмножества $X^{(1)}$; R_{22} — аналогичная матрица $p_2 \times p_2$, относящаяся к подмножеству $X^{(2)}$; $R_{12} = R_{21}^T$ — матрицы размерностью $p_1 \times p_2$ и $p_2 \times p_1$;

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

Нулевую гипотезу, предполагающую отсутствие линейной связи между подмножествами случайных величин $X^{(1)}$ и $X^{(2)}$, запишем следующим образом:

$$H_0 : R = \begin{pmatrix} R_{11} & 0 \\ 0 & R_{22} \end{pmatrix}.$$

Нулевая гипотеза противопоставляется гипотезе H_1, утверждающей, что подмножества случайных величин $X^{(1)}$ и $X^{(2)}$ не являются независимыми.

Выбор между гипотезами H_0 и H_1 осуществляется на основе коэффициентов канонической корреляции, оценки которых (v_1, $v_2, \ldots, v_{p-1} \in V$) определяются как ненулевые корни уравнения:

$$R_{12} R_{22}^{-1} R_{21} - v^2 R_{11} = 0,$$

где R_{11}, R_{22}, R_{21}, R_{12} — блоки выборочной корреляционной матрицы R.

Суть канонической корреляции заключается в отыскании таких линейных комбинаций величин, составляющих подмножества $X^{(1)}$ и $X^{(2)}$, которые дают максимальную корреляцию U_1. Затем в каждом подмножестве находим новые линейные комбинации, опять же удовлетворяющие условию максимальной корреляции U_2. При этом $U_1 \geqslant U_2$, а линейные комбинации, полученные при нахождении U_1 и U_2, ортогональны, т. е. некоррелированы. Используя терминологию факторного анализа, можно было бы сказать, что нулевая линейная комбинация соответствует наиболее мощному фактору, общему для обоих подмножеств, тогда как вторая и последующие комбинации (всего их p_1, если $p_1 \leqslant p_2$) учитывают все более слабеющие попарно некоррелированные факторы. В результате получаем следующий ряд коэффициентов канонической корреляции: $v_1 \geqslant v_2 \geqslant \ldots \geqslant v_S \geqslant \ldots \geqslant v_{p_1}$.

Принятие решения относительно гипотезы H_0 опирается на критерий, имеющий вид:

$$I = (N - p_2 - 1) \sum_{S=1}^{p_1} \frac{v_S^2}{1 - v_S^2},$$

где N — объем p-мерной выборки, на основе которой формировалась матрица R.

При условии, что нулевая гипотеза верна, величина критерия I имеет χ^2-распределение. Нуль-гипотеза отвергается при уровне значимости α, если вычисленное значение I превысит предельно допустимое значение $\chi^2_{\alpha, f}$, выбираемое из соответствующих таблиц. Число степеней свободы f регулируется объемами подмножеств $X^{(1)}$ и $X^{(2)}$ и составляет $p_1 \times p_2$. Итак, при выполнении неравенства $I > \chi^2_{\alpha, f}$ коррелируемые подмножества случайных величин считаются зависимыми.

ГЛАВА 11

РЕГРЕССИОННЫЙ И КОВАРИАЦИОННЫЙ АНАЛИЗ

Регрессионный анализ — совокупность статистических методов, ориентированных на исследование стохастической зависимости одной переменной Y от набора других переменных $\{X_j\} = (X_1, X_2, \ldots, X_p)$ [18, 26, 38]. Его основными задачами являются: 1) установление формы зависимости Y от $\{X_j\}$; 2) определение вида уравнения регрессии; 3) прогнозирование значений результирующей переменной Y, носящей название отклика по известным значениям переменных $X_1, X_2, \ldots, X_p$, которые нередко называют рег-

рессорами. Они могут либо определяться экспериментатором, т. е. быть контролируемыми (модель I), либо меняться случайным образом независимо от исследователя (модель II). В обоих случаях переменная Y рассматривается как случайная величина. Вычислительные процедуры для моделей I и II идентичны, однако содержательная интерпретация выводов, полученных в условиях этих моделей, имеет свои особенности (см. гл. 7) [40].

Регрессионный анализ нашел широкое применение при решении большого круга геологических задач: при комплексной интерпретации промыслово-геофизических данных в задачах нефтегазовой геологии (М. М. Элланский), при геологической интерпретации гравитационных и магнитных аномалий (Г. И. Каратаев), при математическом моделировании земной коры по геологической и геофизической информации (Ю. П. Белов, Б. Е. Большаков), при картировании геологических характеристик (В. И. Аронов, М. А. Романова, В. М. Омелин и др.), при геохимических поисках по первичным ореолам (А. А. Беус, А. П. Соловов, С. В. Григорян), п и поисках и оценке пегматитов по геохимическим данным (А. Н. Бугаец), при анализе пространственных геологических закономерностей (И. И. Боровко), при фациально-формационном анализе магматических комплексов (И. И. Абрамович, В. В. Груза), при оценке запасов минерального сырья (А. М. Марголин).

С различными аспектами проблемы регрессионного оценивания можно ознакомиться в работах [2, 6, 19, 23, 28, 40].

ЛИНЕЙНАЯ РЕГРЕССИЯ

Основное уравнение регрессионного анализа в матричной форме имеет следующий вид: $Y = X'\beta + \varepsilon$, где

$$Y = \begin{Bmatrix} y_1 \\ y_2 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{Bmatrix}; \qquad \beta = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_j \\ \cdots \\ \beta_p \end{Bmatrix}; \qquad \varepsilon = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_i \\ \cdots \\ \varepsilon_n \end{Bmatrix};$$

$$X = \begin{Bmatrix} x_{11} & x_{12} & \ldots & x_{1i} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2i} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{j1} & x_{j2} & \ldots & x_{ji} & \ldots & x_{jn} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{p1} & x_{p2} & \ldots & x_{pi} & \ldots & x_{pn} \end{Bmatrix}.$$

Здесь Y—n-мерный вектор значений зависимой переменной; X — матрица значений аргументов; β—p-мерный вектор неизвестных коэффициентов регрессии, оценки которых (b) отыскиваются по выборочным данным; ε — n-мерный вектор случайных отклонений, появление которых чаще всего связывают с действием факторов, не учтенных матрицей X.

Традиционный регрессионный анализ опирается на следующие допущения [22, 29, 38, 44]: 1) $M\varepsilon_j = 0$, $D\varepsilon_j = \sigma^2 < \infty$ для всех j; $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$; 2) ранг матрицы X равен p (т. е. среди регрессоров отсутствуют линейно связанные переменные); 3) значения результирующей переменной в достаточной мере однородны (в идеальном случае являются выборкой из генеральной совокупности с нормальным распределением); 4) измерения Y и $X_1, X_2, \ldots, X_p$ выполнены без ошибок (по крайней мере без сколько-нибудь существенных ошибок). В настоящее время активно разрабатываются методы регрессионного анализа, позволяющие ослабить или снять эти ограничения. Так, в случае разнородных наблюдений, «загрязненных» аномально высокими или аномально низкими значениями случайной величины Y, предлагаются различные методы робастного (устойчивого к «ураганным» значениям) оценивания параметров регрессии. Измерения, отягощенные ошибками, обрабатываются по схеме конфлюентного анализа (в этой ситуации связь между результирующей и объясняющими переменными называется структурной). Существуют также методы, позволяющие эффективно оценивать параметры в условиях сильной сопряженности регрессоров.

Различают линейную и нелинейную регрессию (рис. 52); при этом полезно выделить следующие классы: 1) регрессии, линейные и по X, и по β:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j + \ldots + \beta_p X_p + \varepsilon;$$

2) регрессии, линейные по β и нелинейные по X, например:

$$Y = \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_j X_1^j + \ldots + \beta_p X_1^p + \varepsilon; \qquad m,\ n > 1;$$

3) регрессии, нелинейные по β, например имеющие вид показательной функции:

$$Y = \beta_1 \beta_2^x.$$

Для регрессий, линейных по X и β или только по β, вычисление оценок b неизвестных коэффициентов β производится методом наименьших квадратов [23, 29, 40, 46], в основе которого лежит требование минимизации суммы квадратов отклонений эмпирических значений Y от значений $\hat{Y}$, вычисляемых по уравнению регрессии:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \min,$$

или в матричной записи: $(Y - \hat{Y})'(Y - \hat{Y}) = \varepsilon'\varepsilon = \min$.

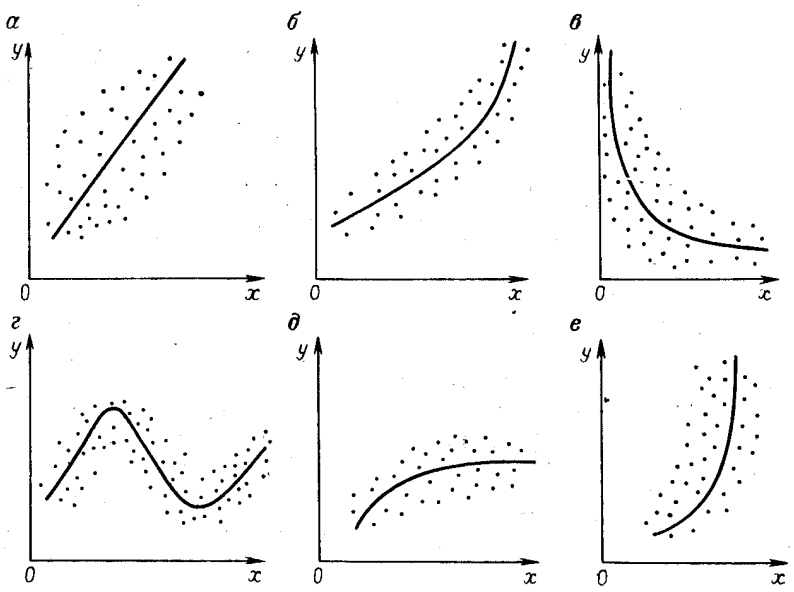


Рис. 52. Простейшие виды регрессионных зависимостей:

$а - y = ax + b$; $б - y = a + bx + cx^2$; $в - y = a + \frac{b}{x}$; $г - y = a + bx + cx^2 + dx^3$;

$д - y = a + b \log x$; $е - y = a + b + c^x$

Этот же метод может быть использован и для регрессий, нелинейных по β, если удастся подобрать подходящее линейное преобразование. Так, в вышеприведенном примере можно рекомендовать следующее преобразование:

$$\log Y = \log \beta_1 + X \log \beta_2.$$

Положив $\log Y = A$, $\log \beta_1 = B$, $\log \beta_2 = C$, получаем линейную зависимость, допускающую применение метода наименьших квадратов: $A = B + CX$.

В матричной форме вектор оценок b определяется следующим образом: $b = (X'X)^{-1}X'Y$.

В матрицу X обычно вводят фиктивную переменную $X_0 = \{1, 1, \ldots, 1\}$ размерности n, что позволяет вместе с коэффициентами $b_1, b_2, \ldots, b_P$ вычислять и b_0 — постоянную регрессии, выполняющую в данном случае функцию выравнивания: она сдвигает линию (если $P = 1$) или гиперповерхность (если $P > 1$) регрессии в область скопления точек $\{y_i, x_{ji}, j = 1, 2, \ldots, P; i = 1, 2, \ldots, n\}$.

Отыскав $(b_0, b_1, b_2, \ldots, b_j, \ldots, b_P)$, можно составить уравнение линейной регрессии:

$$\hat{Y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_j x_{ji} + \ldots + b_P x_{Pi},$$

где x_{ji} — значения j-го регрессора в i-ом наблюдении; $\hat{y}_i$ — расчетное значение прогнозируемой переменной.

Коэффициенты регрессии b_j, выраженные в натуральном масштабе, можно представить в стандартизированной форме, что более удобно при их сравнении. Вычисление стандартизированных коэффициентов b_j' выполняется либо предварительной стандартизацией исходных данных $[y_i' = (y_i - \bar{y})/S_y;\ x_{ji}' = (x_{ji} - \bar{x}_j)/S_{xj}]$, либо по формуле $b_j' = (S_{xj}/S_y)b_j$.

Величина и знак b_j' позволяют оценить интенсивность и направление влияния регрессоров на результирующую переменную. Значение $|b_j'|$ показывает, на какую долю стандартного отклонения изменится среднее значение переменной Y при условии, что X_j возросло (уменьшилось) на величину S_{xj}, а остальные объясняющие переменные остались бы на прежнем уровне. Если регрессоры $\{X_j\}$ можно отождествить с некоторыми природными факторами, то такого рода анализ регрессионной модели может оказаться эффективным средством решения генетических задач геологии. Однако прежде чем приступить к содержательной интерпретации коэффициентов регрессии, необходимо убедиться в статистической значимости последних. Качество уравнения регрессии в целом можно оценить следующим образом. Сформулируем нулевую гипотезу: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ при альтернативе $H_1 : \beta_j \neq 0$ хотя бы для одного $j = 1, 2, \ldots, p$.

Для проверки этой гипотезы используется критерий F:

$$F = \frac{n - p - 1}{p} \cdot \frac{R^2}{1 - R^2},$$

где R вычисляется по формуле

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}.$$

В условиях H_0 критерий F имеет F-распределение со степенями свободы p и $n-p-1$. Если $F > F_\alpha$, p, $n-p-1$, то нулевая гипотеза отвергается и принимается решение об удовлетворительном качестве соответствия регрессии эмпирическим данным. Если уравнение регрессии служит для прогнозирования Y по $\{X_j\}$, то для повышения надежности рекомендуется добиться путем подбора соответствующего уравнения выполнения соотношения $F > 4F_\alpha$ $p,n-p-1$ [6, 40].

Существует мнение, что интерпретация величины R в условиях модели I и модели II регрессионного анализа должна быть различной. Если регрессоры $\{X_j\}$ — случайные величины, то вполне допустима трактовка R^2 как индикатора адекватности регрессионной модели. В этом случае R^2 оценивает ту долю изменчивости Y,

которая «объясняется» регрессией. Если же регрессоры $\{X_j\}$ детерминированы (модель II), то R^2 следует рассматривать как показатель, величина которого указывает, насколько регрессионная модель лучше описывает исходные данные, чем модель при условии $\hat{Y}_i = \bar{Y}$, $i = 1, 2, \ldots, n$.

Для небольших n предлагается специальная коррекция R^2, устраняющая его смещение. Нахождение исправленного значения R_n^2 выполняется следующим образом:

$$R_n^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Отклонение гипотезы $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ не означает, что среди всего набора регрессоров нет переменных, вклад которых в объяснение результирующей Y близок или равен нулю. Поэтому следующей задачей является проверка гипотез о равенстве нулю каждого из p коэффициентов регрессии: $H_0 : \beta_j = 0$ при альтернативе $H_1 : \beta_j \neq 0$.

Для проверки H_0 используется критерий $|t| = b_j/S_{bj}$, где

$$S_{bj} = S_\varepsilon \sqrt{(X'X)^{-1}},$$

$$S_\varepsilon = \left[\frac{1}{n-p-1}\varepsilon'\varepsilon\right]^{1/2},$$

а $(X'X)^{-1}$ — j-й элемент главной диагонали матрицы, обратной матрице $X'X$.

Если $t > t_{\alpha, n-p-1}$, где $t_{\alpha, n-p-1}$ выбирается из таблиц распределения Стьюдента, то нулевая гипотеза отклоняется с уровнем значимости α, т. е. можно считать, что имеет место существенное уклонение от 0 коэффициента β_j.

Выполнив такого рода проверку всех коэффициентов β_j, исследователь получает возможность сосредоточить особое внимание на содержательном анализе тех β_j, для которых H_0 была отвергнута. Для оценки их точности полезно построить доверительный интервал:

$$P\{b_j - t_{\alpha, n-p-1}S_{bj} \leqslant \beta_j + t_{\alpha, n-p-1}S_{bj}\} = 1 - \alpha,$$

накрывающий с надежностью $(1-\alpha)\cdot 100$ % истинный коэффициент регрессии β_j. Чем уже ширина такого интервала, тем «лучше» выборочная оценка b_j, а значит и более надежна генетическая или иная интерпретация соответствующего регрессора.

При использовании уравнений регрессии в прогнозных целях полезно построить доверительные интервалы для предсказываемой переменной Y. Доверительный интервал для отдельных значений y_i имеет следующие границы: $\hat{Y}_i \pm t_{\alpha, n-p-1}S_{l_i}$, где t_α — критическое значение распределения Стьюдента при заданном уровне значимости α; S_{l_i} — оценка стандартной ошибки прогноза в точке X_i:

$$S_{l_i} = \{S_\varepsilon[1 + X_i'(X'X)^{-1}X_i]\}^{1/2}$$

где $X_i' = (x_{1i}, x_{2i}, \ldots, x_{ji}, \ldots, x_{pi})$.

Если точность предсказания Y по данному набору регрессоров невысока, то обычно пытаются либо сменить вид функции (например, от линейной функции перейти к степенной, полиномиальной и т. п.), либо произвести ревизию $\{X_j\}$. В последнем случае используются так называемые пошаговые процедуры, в основе которых лежат операции удаления или включения тех или иных регрессоров. Общее правило, на основании которого принимается решение о включении или невключении переменной в множество $\{X_j, j = 0, 1, \ldots, k\}$, сводится к выяснению вопроса, улучшается или нет предсказание Y по набору $\{X_j, j = 0, 1, \ldots, k+1\}$ в сравнении с набором $\{X_j, j = 0, 1, \ldots, k\}$. Эта задача может быть сформулирована как проверка гипотезы

$$H_0 : \rho_{YX_{k+1}}\{X_j,\ j \neq k+1\} = 0$$

при альтернативе:

$$H_1 : \rho_{YX_{k+1}}\{X_j,\ j \neq k+1\} \neq 0,$$

где

$$\rho_{YX_{k+1}}\{X_j,\ j \neq k+1\}$$

— частный коэффициент корреляции.

Проверку H_0 можно провести с помощью следующего критерия:

$$F = (n-j-2)\frac{r^2_{YX_{k+1}}\{X_j,\ j \neq k+1\}}{1 - r^2_{YX_{k+1}}\{X_j,\ j \neq k+1\}},$$

где

$$r_{YX_{k+1}}\{X_j,\ j \neq k+1\}$$

— оценка частного коэффициента корреляции.

Если $F > F_{\alpha, 1, n-j-2}$, то принимается альтернатива $\rho_{YX_{k+1}}\{X_j, j \neq k+1\} \neq 0$, вклад переменной X_{k+1} считается существенным и она присоединяется к набору $\{X_j, j = 0, 1, \ldots, k\}$. Процедура повторяется для всех изучаемых переменных, при этом в качестве наилучшего регрессора выбирается такой X_j, для которого

$$F_{YX_j} = r^2_{YX_j}(n-2)/\left(1 - r^2_{YX_j}\right) = \max,$$

т. е. является максимальным для всего набора изучаемых переменных. Далее последовательно рассматриваются все оставшиеся аргументы и после вычисления и сравнения между собой значений критерия $F_Y X_j$, один из них, обладающий максимальным F, присоединяется к регрессору X_j. Следует учитывать, что на каждом шаге этой процедуры число степеней свободы для критического значения F_{α, f_1, f_2} меняется только для f_2 : n—2, n—3, n—4 и т. д., тогда как f_1 постоянно равно 1. Формирование набора аргументов считается завершенным, когда все вычисляемые значения F-критерия станут меньше заданного критического значения. Опираясь на тот же F-критерий, можно построить иную пошаговую процедуру,

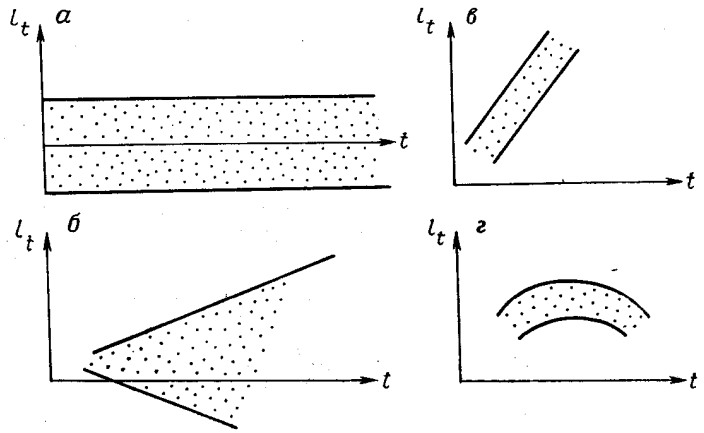


Рис. 53. Зависимость величин остатков регрессии от регрессоров и отклика:
а — адекватность линейной модели исходным данным: остатки заполняют горизонтальную полосу с центром на оси абсцисс; *б* — неадекватность модели исходным данным: остатки заполняют полосу, которая расширяется при возрастании аргумента (дисперсия σ^2 непостоянна); *в* — неадекватность модели исходным данным: остатки заполняют наклонную полосу, указывающую на наличие линейного тренда; *г* — неадекватность модели исходным данным: остатки заполняют график сложного вида, указывающий на то, что пропущен линейный член

которую уместно было бы назвать методом исключения. В этом случае сначала рассматривается максимально полный набор регрессоров, а затем производится их последовательное удаление, начиная с тех, которые не обладают способностями предсказания.

При построении линейной регрессионной модели были сделаны некоторые предположения относительно ошибок e_t. Проверка этих предположений может быть проведена графически, построением графиков остатков, выражающих зависимость величин остатков $e_t = y_t - \hat{y}_t$ от x_t или от $\hat{y}_t$, $t = 1, 2, \ldots, n$ (рис. 53). Если остатки заполняют горизонтальную полосу с центром на оси абсцисс, то принимается решение об адекватности линейной модели. В противном случае модель неадекватна, причем расположение остатков на графике указывает источник неадекватности. Если полоса расширяется при возрастании аргумента, то это свидетельствует о том, что дисперсия σ^2 непостоянна и следует применить преобразование переменной y. Наклонная полоса указывает на наличие линейного тренда, и в этом случае в модель следует добавить дополнительную переменную. График сложного вида указывает, что пропущен линейный член. Также следует проверить гипотезу о нормальном распределении ошибок e_t, $t = 1, 2, \ldots, n$, например, путем построения гистограмм.

РЕГРЕССИЯ НАИМЕНЬШИХ АБСОЛЮТНЫХ ОТКЛОНЕНИЙ

При аппроксимации зависимой переменной y линейной комбинацией $\hat{y} = \alpha_1 x_1 + \ldots + \alpha_m x_m$ независимых переменных $x_1 \ldots, x_m$ с помощью уравнения регрессии параметры оцениваются из условия обращения в минимум средней суммы квадратов отклонений:

$$\sigma^2_{\text{ост}} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2.$$

Однако имеются соображения в пользу другого критерия при построении регрессионной модели — критерия минимизации среднего абсолютного отклонения:

$$\Delta = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{\hat{y}}_t\right|.$$

Эти соображения следующие.

1. В ряде задач средняя погрешность Δ является естественной мерой точности, не искажая величину отклонений, в то время как среднеквадратическая погрешность $\sigma^2_{\text{ост}}$ увеличивает роль больших отклонений и преуменьшает роль малых. Мера точности $\sigma^2_{\text{ост}}$ естественна в задачах, где цена отклонения пропорциональна квадрату его значения.

2. Регрессия $\hat{\hat{y}}$ более устойчива, нежели $\hat{y}$, так как при больших отклонениях она менее сдвинута в сторону точек с большими отклонениями.

3. Метод нахождения $\hat{\hat{y}}$ прост и легко реализуем, в отдельных случаях решение находится меньшим числом вычислений, нежели нахождение $\hat{y}$.

4. Метод наименьших квадратов естествен в случае нормального распределения. Критерий наименьших абсолютных отклонений естествен в случае закона Лапласа (двустороннего экспоненциального) с плотностью

$$f(x) = \frac{\lambda}{2}e^{-\lambda(x-a)}.$$

Законы Гаусса и Лапласа близки, однако плотность закона Лапласа обладает большей островершинностью и весомостью хвостов.

Случай $m = 1$. Для построения уравнения регрессии

$$\hat{\hat{y}} = a + bx = \frac{x_{t1}y_{t2} - x_{t2}y_{t1}}{x_{t1} - x_{t2}} + \frac{y_{t1} - y_{t2}}{x_{t1} - x_{t2}}$$

необходимо найти две оптимальные точки t_1 и t_2. Для этого, задавшись какой-либо точкой (например, $t_1 = n$), определяют наилучшую точку t_2 в пару t_1. Для этого находят a: $a = y_n - bx_n$, а b — из условия обращения в минимум кусочно-линейной функции:

$$\Delta(b) = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - y_n - b(x_t - x_n)\right| = \frac{1}{n}\sum_{t=1}^{n}\left|\tilde{y}_t - b\tilde{x}_t\right|,$$

который достигается в одной из абсцисс ее излома:

$$b_t = \frac{y_t - y_n}{x_t - x_n} = \frac{\tilde{y}_t}{\tilde{x}_t} = \frac{\begin{vmatrix} y_t & 1 \\ y_n & 1 \end{vmatrix}}{\begin{vmatrix} x_t & 1 \\ x_n & 1 \end{vmatrix}}.$$

Упорядочим b_t по возрастанию. Пусть порядок b_t после упорядочения не изменился (в противном случае меняем индексацию по t в соответствии с новым порядком). Найдем

$$\min_r \left| \sum_{t=1}^{r} |\tilde{x}_t| - \sum_{t=r+1}^{n} |\tilde{x}_t| \right| = \left| \sum_{t=1}^{r_0} |\tilde{x}_t| - \sum_{t=r_0+1}^{n} |\tilde{x}_t| \right|.$$

Если выражение $\sum_{t=1}^{r_0} |\tilde{x}_t| - \sum_{t=r_0+1}^{n} |\tilde{x}_t|$ положительно, то $r = r_0$, если отрицательно, то $r = r_0 + 1$, а если равно нулю — то задача допускает множество решений: $b_{r_0} \leqslant b_r \leqslant b_{r_0+1}$.

Определив точку $t = r$, наилучшую для $t = n$, отбрасываем последнюю и повторяем всю процедуру поиска наилучшей точки уже для $t = r$.

Процедура заканчивается, когда очередная наилучшая точка совпадает с отбрасываемой точкой.

Общий случай m независимых переменных. Из общего числа n точек выбирается каким-либо образом $m + 1$ точек $k_1, k_2, \ldots, k_{m+1}$. Затем поочередно одна из этих точек отбрасывается, а из остальных $n-m$ точек находится наилучшая.

Для этого вычисляется величина

$$b_t^* = \frac{\Delta\tilde{y}_t}{\Delta\tilde{x}_t} = \begin{vmatrix} y_{k_1} & x_{k_1}^{(1)} & \ldots & x_{k_1}^{(m-1)} & 1 \\ y_{k_2} & x_{k_2}^{(1)} & \ldots & x_{k_2}^{(m-1)} & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ y_{k_{(m+1)}} & x_{k_{(m+1)}}^{(1)} & \ldots & x_{k_{(m+1)}}^{(m-1)} & 1 \end{vmatrix} =$$

$$= \begin{vmatrix} x_{k_1}^{(m)} & x_{k_1}^{(1)} & \ldots & x_{k_1}^{(m-1)} & 1 \\ x_{k_2}^{(m)} & x_{k_2}^{(1)} & \ldots & x_{k_2}^{(m-1)} & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{k_{(m+1)}}^{(m)} & x_{k_{(m+1)}}^{(1)} & \ldots & x_{k_{(m+1)}}^{(m-1)} & 1 \end{vmatrix}$$

Заменяя в процедуре нахождения r для случая $m = 1$ b_t на b_t^*, а $\tilde{x}_t$ на $\tilde{y}_t$, сведем задачу к известному алгоритму нахождения регрессии.

РЕГРЕССИЯ НА ОРТОГОНАЛЬНЫХ ПЕРЕМЕННЫХ

Оценки, полученные на основе классической линейной модели, обладают тем недостатком, что с добавлением в модель новой независимой переменной все полученные ранее оценки приходится пересчитывать.

От такого недостатка свободна модель, в которой матрица X имеет ортогональные столбцы.

В общем случае если система векторов $x_1, x_2, \ldots, x_m$ линейно независима, то к ней можно применить процесс ортогонализации, в результате чего получим новую систему попарно ортогональных векторов:

$$z_0 = x_0 \equiv 1,$$

$$z_1 = x_1 - \frac{(x_1, z_0)}{(z_0, z_0)} z_0,$$

$$\ldots\ldots\ldots\ldots$$

$$z_m = x_m - \frac{(x_m, z_0)}{(z_0, z_0)} z_0 - \ldots - \frac{(x_m, z_{m-1})}{(z_{m-1}, z_{m-1})} z_{m-1}.$$

Тогда в новых переменных общая регрессионная модель будет иметь вид:

$$y = \alpha_0 + \alpha_1 z_1 + \ldots + \alpha_m z_m,$$

так что $$z = \begin{pmatrix} 1 & z_{11} & \ldots & z_{1m} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & z_{n1} & \ldots & z_{nm} \end{pmatrix},$$

$$z'z = \begin{pmatrix} (z_0, z_0) & & & 0 \\ & (z_1, z_1) & & \\ & & \ddots & \\ 0 & & & (z_m, z_m) \end{pmatrix},$$

$$\hat{\alpha}_i = \frac{(y, z_i)}{(z_i z_i)}, \quad i = 0, 1, \ldots, m,$$

$$v(\hat{\alpha}_i) = \frac{\sigma^2}{(z_i z_i)}.$$

Сумма квадратов $SS(z_t)$, обусловленная вкладом одной переменной z_t, равна:

$$SS(z_t) = \frac{(y, z_i)^2}{(z_t, z_i)}.$$

Следовательно,

$$SSR = SS(z_1) + \ldots + SS(z_m),$$

$$SST = y'y - n\bar{y}^2,$$

и $$SSE = SST - SSR$$

$$S^2 = \hat{\sigma}^2 = \frac{SSE}{n-m-1}$$

$$\hat{y} = z_{\hat{\alpha}}.$$

В предположении нормальной распределенности ошибок строятся, как и в случае общей модели, соответствующие доверительные интервалы и доверительные области.

Существенным достоинством перехода к ортогональным переменным является возможность провести регрессионный анализ в случае, когда столбцы матрицы X линейно зависимы. В этом случае матрица $X'X$ вырождена и общая модель регрессии неприменима.

В процессе ортогонализации векторов матрицы X с вырожденной матрицей $X'X$ на некотором шаге получим $(z_j, z_j) = 0$. Это будет означать, что столбцы $x_0, x_1, \ldots, x_{j-1}$ линейно независимы, а $x_0, x_1, \ldots, x_{j-1}, x_j$ — линейно зависимы. Исключив вектор x_j из рассмотрения, мы сможем продолжить процесс ортогонализации дальше. В результате ряд столбцов матрицы X будет исключен и останется матрица с линейно независимыми столбцами.

Обозначим через $u_i = (z_0, z_1, \ldots, z_{i-1})$, $i = 1, \ldots, m$, $u_0 = z_0$ матрицу из уже преобразованных столбцов на $(i-1)$-ом шаге.

Тогда процесс ортогонализации может быть записан в виде: $z_i = x_i - u_i (u_i' u_i)^{-1} u_i x_i$.

Коэффициенты регрессии β_i и α_i связаны линейными соотношениями и могут быть получены приравниванием коэффициентов при соответствующих переменных, для чего следует в модели с переменными z_i заменить их на выражения через x_i. Рассмотрим исходную задачу с точки зрения функционального анализа.

В процессе ортогонализации можно дополнительно нормировать полученные переменные z_i. Обозначим их через e_i.

Рассмотрим бесконечномерное пространство, для которого определено скалярное произведение элементов (такие пространства называются Гильбертовыми), и выберем в нем подпространство L, порожденное бесконечномерной ортонормальной системой e_i, $i = 0, 1, \ldots, m, \ldots$

Пусть $y \in L$. Тогда будет выполнено соотношение:

$$\left\| y - \sum_{i=0}^{m} \hat{\alpha}_i e_i \right\|^2 = \| y \|^2 - \sum_{i=0}^{m} | e_i |^2 + \sum_{i=0}^{m} | \hat{\alpha}_i - C_i |^2,$$

где $C_i = (y, e_i)$ называются коэффициентами Фурье элемента относительно ортонормальной системы $\{e_i\}$.

Норма, стоящая в левой части равенства, принимает наименьшее значение, когда $\hat{\alpha}_i = C_i$.

Тогда

$$\left\| y - \sum_{i=0}^{m} C_i e_i \right\|^2 = \| y \|^2 - \sum_{i=0}^{m} | C_i |^2.$$

Доказывается, что ряд $\sum_{i=0}^{m} | C_i |^2$ сходится, причем $\sum_{i=0}^{\infty} | C_i |^2 = \| y \|^2$.

Если теперь y — произвольный элемент из H, то справедливо неравенство Бесселя: $\sum_{i=0}^{\infty} | C_i |^2 \leqslant \| y \|^2$.

(Равенство достигается, если $y \in L$).

Ортонормальная система $\{e_i\}$ называется полной, если не существует элемента $x \in H$, отличного от нулевого и ортогонального всем элементам из $\{e_i\}$.

Ортонормальная система $\{e_i\}$ называется замкнутой, если подпространство L, порожденное этой системой, совпадает с H.

Ряд Фурье по замкнутой системе для любого элемента x из H сходится к x и для любого $x \in H$ справедливо равенство Парсеваля—Стеклова:

$$\sum_{i=0}^{\infty} C_i = \| x \|^2.$$

Если ортонормальная система полная, то она и замкнутая.

К недостаткам использования ортогональных переменных следует отнести необходимость пересчета всех коэффициентов при добавлении или исключении отдельных наблюдений.

КОВАРИАЦИОННЫЙ АНАЛИЗ

Ковариационный анализ — статистический метод оценки влияния на случайную величину различных одновременно действующих факторов, одни из которых заданы качественно (см. гл. 7), а другие могут быть измерены количественно (ситуация, соответствующая регрессионному анализу). Таким образом, ковариационный анализ может рассматриваться как комбинация дисперсионного и регрессионного анализов.

Линейная модель, обобщающая дисперсионный и регрессионный подходы, имеет вид: $Y = X'\beta + z'\gamma + \varepsilon$, где Y, X, β, ε имеют тот же смысл, что и в модели дисперсионного анализа, а $z'\gamma$ определяет вклад факторов, поддающихся количественному исследованию, при этом z — значения факторов (регрессоров), γ — коэффициенты регрессии Y на z.

В дальнейшем будем полагать, что коэффициенты регрессии не зависят от градаций качественного фактора, задающего разбивку исходных данных на p групп: $\gamma_1 = \gamma_2 = \ldots = \gamma_i = \ldots = \gamma_p = \gamma$.

Основные предположения ковариационного анализа [46]:

1) Y имеет нормальное распределение с параметрами $(X'\beta, \sigma^2 I)$;

2) Y имеет нормальное распределение с параметрами $(X'\beta + z'\gamma, \sigma^2 I)$.

Как и в дисперсионном и регрессионном анализах, распределение ε также предполагается нормальным с параметрами $(0, \sigma^2)$.

Таблица 9

Исходные данные для ковариационного анализа

Градации фактора A	
1	$(y_{11}, z_{11})\ (y_{12}, z_{12}) \ldots (y_{1j}, z_{1j}) \ldots (y_{1ni}, z_{1ni})$
. . .	. . . . .
i	$(y_{i1}, z_{i1})\ (y_{i2}, z_{i2}) \ldots (y_{ij}, z_{ij}) \ldots (y_{ini}, z_{ini})$
. . .	. . . . .
p	$(y_{p1}, z_{p1})\ (y_{p2}, z_{p2}) \ldots (y_{pj}, z_{pj}) \ldots (y_{pnp}, z_{pnp})$

Предположение (1) соответствует нулевой гипотезе $H_\gamma : \gamma = 0$, а (2) — гипотезе $H_\beta : \beta_1 = \beta_2 = \ldots = \beta_i = \ldots = \beta_p$. Если H_γ выполняется, то проверка H_β сводится к общему дисперсионному анализу. Если же H_γ отклоняется, то перед проверкой требуется внести определенные коррективы, исключающие эффект регрессии.

Принципиальную схему ковариационного анализа удобно рассмотреть на примере однофакторного анализа с одним независимым переменным (регрессором): $y_{ij} = \beta_i + \gamma z_{ij} + \varepsilon_{ij}$, где β_i — эффект i-й градации фактора A; γz_{ij} — эффект, обусловленный действием переменной z; γ — коэффициент регрессии; ε_{ij} — эффект неконтролируемых факторов, $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, n_i$ (табл. 9).

Проверка гипотезы

$$H_\gamma : \gamma = 0.$$

Определим суммы квадратов и произведений отклонений, отражающих изменчивость Y и z.

А. Внутри групп (градаций):

$$a_1 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - y_{i*})^2; \qquad b_1 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (z_{ij} - z_{i*});$$

$$c_1 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - y_{i*})(z_{ij} - z_{i*}),$$

где $y_{i*} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}; \qquad z_{i*} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}.$

Б. Между группами: $a_2 = \sum_{i=1}^{p} n_i (y_{i*} - y_{**})^2;$

$$b_2 = \sum_{i=1}^{p} n_i (z_{i*} - z_{**})^2;$$

$$c_2 = \sum_{i=1}^{p} n_i (y_{i*} - y_{**})(z_{i*} - z_{**}),$$

где $y_{**} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}; \qquad z_{**} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{n_i} z_{ij};$

$$N = \sum_{i=1}^{p} n_i.$$

Если H_γ верна, то статистика $\frac{c_1^2}{a_1} : \left(b_1 - \frac{c_1^2}{a_1}\right)$ имеет F-распределение с $f_1 = 1$ и $f_2 = N - p - 1$ степенями свободы. Гипотеза о равенстве нулю коэффициента регрессии γ отклоняется, если вычисленное значение критерия превысит табличное F_{α, f_1, f_2}.

Проверка гипотезы H_β в условиях $\gamma \neq 0$:

$$H_\beta : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_i = \ldots = \beta_p.$$

Суммы квадратов «между группами» и «внутри групп» должны быть скорректированы так, чтобы влияние независимой переменной z было исключено:

$$a = a_1 + a_2; \qquad b = b_1 + b_2; \qquad c = c_1 + c_2;$$

$$S = b - c^2/a, \qquad S_1 = b_1 - C_1^2/a_1, \qquad S_2 = b_2 - C_2^2/a_2.$$

Статистика S_2/S_1 в условиях гипотезы H_β имеет F-распределение с $f_1 = p - 1$, $f_2 = N - p - 1$ степенями свободы.

Рассмотренную схему можно обобщить на случае, когда классификация наблюдений выполнена по двум и более факторам.

В геологии ковариационный анализ применяется пока реже дисперсионного и регрессионного анализов, хотя информация, привлекаемая геологом для решения генетических задач, большей частью носит комбинированный характер. Так, при выяснении условий локализации месторождений полезных ископаемых исследователь опирается как на качественную (например, структурно-тектонические и литолого-фациальные сведения), так и на количественную информацию (например, содержание химических элементов, мощность осадочных и метаморфических образований и т. п.). В этой ситуации наиболее надежные выводы о факторах, управляющих изучаемым геологическим объектом, могут быть получены методами ковариационного анализа.

ГЛАВА 12

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Главными компонентами случайного P-мерного вектора x называются такие ортогональные линейные комбинации v_j, $j = 1, 2, \ldots, r$, $r \leqslant P$ составляющих этого вектора $x_1, \ldots, x_P$, что если

упорядочить v_j по их дисперсиям так, чтобы были выполнены неравенства $S^2(v_1) \geqslant S^2(v_2) \geqslant \ldots \geqslant S^2(v_r)$, то дисперсия первой линейной комбинации $S^2(v_1)$ должна достигать максимального значения среди всех линейных комбинаций вектора x, дисперсия второй линейной комбинации $S^2(v_2)$ должна быть максимальной среди всех линейных комбинаций вектора x, ортогональных первой линейной комбинации, и т. д. до линейной комбинации v_r, $r \leqslant P$, ортогональной всем предыдущим линейным комбинациям.

Метод главных компонент — статистический метод сжатия информации, основанный на нахождении собственных векторов и собственных значений ковариационной матрицы P-мерного случайного вектора, распределенного по многомерному нормальному закону.

Основной задачей, в которой метод главных компонент играет важную самостоятельную роль, является задача выяснения сущности геологических процессов по данным изучения современного облика природных объектов. Она сводится к выяснению и оценке роли факторов в становлении наблюдаемых явлений и существующих закономерностей размещения полезных ископаемых в земных недрах. С ней связаны задачи построения корреляционных моделей в предположении действия определенной совокупности природных процессов, определения особенностей изменения по площади и разрезу составляющих, обязанных действию как отдельно взятых факторов, так и любых их сочетаний. В последнее время появились работы по факторному крайгингу, в которых факторный анализ эффективно используется для выделения систематических и случайных составляющих изменчивости комплекса геологических характеристик.

Метод главных компонент нашел применение при поисках локальных структур и залежей нефти и газа в пределах бортовых зон, при изучении вопросов становления состава магматических образований, парагенетических ассоциаций и при решении ряда других задач.

Метод главных компонент при решении некоторых задач выполняет также вспомогательные функции в комплексе с другими методами прикладного статистического анализа.

Такова его роль в задачах классификации, где он позволяет уменьшить число геологических признаков, в задачах прогнозирования на основе построения регрессионной модели. Метод главных компонент с успехом используется при картировании геолого-геофизических характеристик, при сравнительном изучении природных систем и выделении эволюционирующих составляющих.

СТАТИСТИЧЕСКИЙ МЕТОД ХОТЕЛЛИНГА

Рассмотрим вычислительные аспекты метода главных компонент. Пусть $x = (x_1, \ldots, x_P)$ — P-мерный случайный вектор, имеющий многомерное нормальное распределение с математическим ожиданием нуль и ковариационной матрицей S:

$$M(x) = 0; \qquad M(x, x') = S.$$

Тогда можно найти ортогональное преобразование $v = Ax$ такое, что ковариационная матрица случайного вектора v будет диагональной, т. е. $M(v, v') = \Lambda$, где $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_P)$, причем $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \ldots \geqslant \lambda_n$ — корни уравнения $|S - \lambda E| = 0$, а j-й столбец матрицы A удовлетворяет уравнению $(S - \lambda_j E)\, a_j = 0$ (этот вектор можно нормировать, так что $a_j a_j = 1$) и j-я компонента $v_j = a_j x$ вектора v имеет наибольшую дисперсию среди всех нормированных линейных комбинаций, некоррелированных с предыдущими компонентами $v_1, v_2, \ldots, v_{j-1}$.

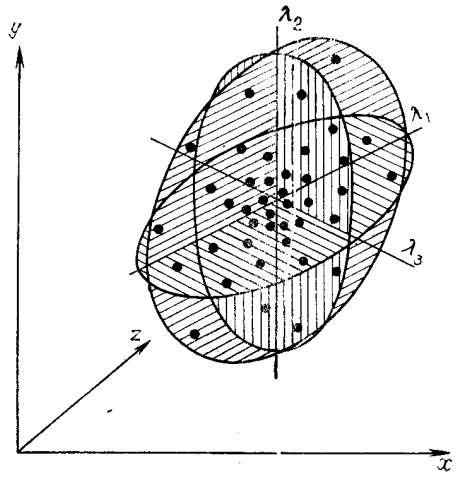


Рис. 54. Трехмерное распределение точек с соответствующими главными осями

Обычно в практических исследованиях истинная ковариационная матрица S неизвестна. Ее приходится заменять выборочной ковариационной матрицей

$$\hat{S} = \|\hat{S}_{ij}\|_{1 \leqslant i,\ j \leqslant P},$$

$$\text{где } \hat{S}_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j),$$

x_{ki}, $k = 1, 2, \ldots, N$ — наблюденные значения компоненты x_i вектора x, $\bar{x}_i = \frac{1}{N} \sum_{k=1}^{N} x_{ki}$.

Для нахождения значений главных компонент $v_1, \ldots, v_r$, $r \leqslant P$, P-мерного случайного вектора x вычисляют собственные значения $\lambda_1, \ldots, \lambda_P$ и собственные векторы $\hat{a}_1, \ldots, \hat{a}_P$ матрицы $\hat{S}$, причем собственные векторы нормируют, т. е. подбирают так, чтобы выполнялось условие $\hat{a}_i' \hat{a}_i = 1$ (рис. 54). Далее находят проекции векторов $x_{k1}, \ldots, x_{kP}$ ($k = 1, 2, \ldots, N$) на направления главных компонент $\hat{a}_1, \ldots, \hat{a}_P$.

Тогда $\hat{v}_{kl} = (x_k, \hat{a}_l)$, $k = 1, 2, \ldots, N$; $l = 1, 2, \ldots, p$,

или $\hat{v}_{kl} = \sum_{S=1}^{P} x_{kS} a_{Sl}$ (рис. 55, 56).

Математическое ожидание компоненты вектора $\hat{v} = (v_1, \ldots, v_P)$ равно нулю:

$$M(v_l) = \frac{1}{N} \sum_{k=1}^{N} \hat{v}_{kl} = 0.$$

Дисперсия $S^2(v_l)$ компоненты v_l вектора $v = (v_1, \ldots, v_P)$ равна $\hat{\lambda}_l$.

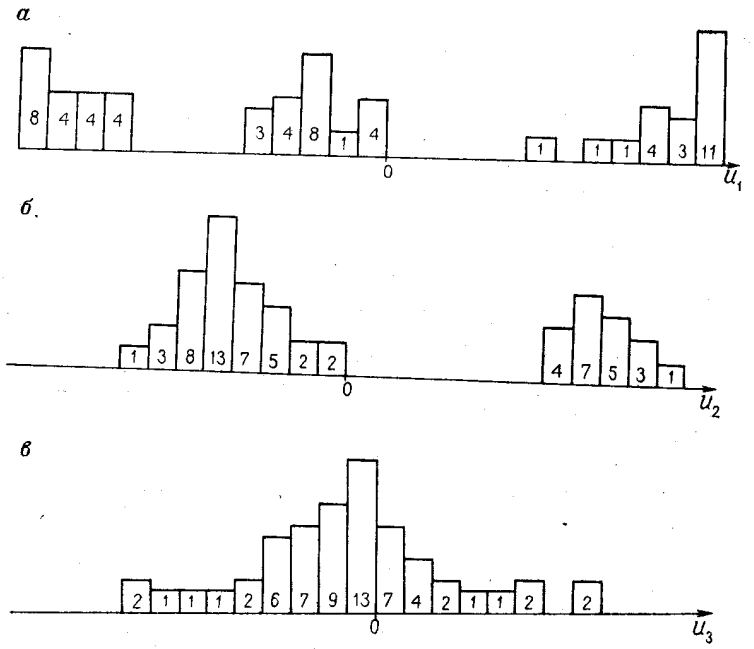


Рис. 55. Гистограммы проекций наблюдений на первую (*а*), вторую (*б*) и третью компоненты (*в*). В колонках — число наблюдений; общее число наблюдений 61

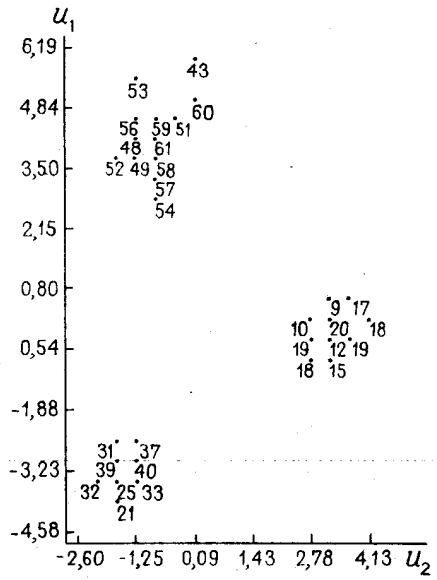


Рис. 56. Расположение наблюдений на плоскости первых двух компонент из трех различных совокупностей

Применение метода главных компонент для исследования ковариационной матрицы предполагает отбрасывание последних главных компонент, оценка дисперсии которых незначимо отличается от нуля. Критерии проверки соответствующих гипотез относительно собственных значений выборочной ковариационной матрицы основаны на использовании закона распределения собственных значений этой матрицы для случая, когда выборка извлечена из совокупности с нормальным законом распределения. Известен критерий М. С. Бартлетта проверки гипотезы о том, что собственные значения ковариационной матрицы равны друг другу. Однако на практике для отбрасывания незначимых главных компонент используется обычно эмпирический критерий, состоящий в том, что вклад в суммарную дисперсию отбрасываемых компонент должен быть меньше 25 или 10 %.

Остановимся на некоторых критериях проверки гипотез о собственных значениях ковариационной матрицы. Предполагая, что случайный вектор $x=(x_1, \ldots, x_P)$ имеет многомерное нормальное распределение с математическим ожиданием нуль и ковариационной матрицей S, можно указать некоторые свойства распределения собственных значений матрицы S.

Укажем доверительные интервалы для оценок собственных значений матрицы S. Обозначим через $\hat{\lambda}_i$ собственное значение выборочной ковариационной матрицы. Тогда $\sqrt{\frac{N}{2}}\frac{\lambda_i-\lambda_i}{2}$, где N — число наблюдений, является случайной величиной, распределенной нормально с математическим ожиданием нуль и дисперсией единица. Доверительный интервал для собственного значения λ_i при (100—α) %-ном уровне значимости находится по формуле

$$\frac{\hat{\lambda}_i}{1+F_{\alpha/2}\sqrt{\frac{N}{2}}} \leqslant \lambda_i \leqslant \frac{\hat{\lambda}_i}{1-F_{\alpha/2}\sqrt{\frac{N}{2}}},$$

где $F_{\alpha/2}$ — $\alpha/2$ %-ная точка стандартного нормального распределения.

Относительно собственных значений выборочной ковариационной матрицы $\hat{S}$ также можно сделать некоторые утверждения. Имеется критерий проверки равенства наименьших $P-k$ собственных значений матрицы $\hat{S}$. Если эти собственные значения равны, то достаточно рассматривать лишь k главных компонент.

Гипотеза о равенстве между собой $P-k$ наименьших собственных значений матрицы $\hat{S}$ проверяется с помощью критерия

$$\Lambda = N'\left[-\ln|S|+\ln\hat{\lambda}_1 \ldots \hat{\lambda}_k + q\ln\lambda\right],$$

где $q=P-k$; $\quad \lambda=\frac{1}{q}(tr\hat{S}-\hat{\lambda}_1-\ldots-\lambda_k)$;

$$N'=N-k-\frac{1}{6}\left(2q+1+\frac{2}{q}\right).$$

С целью получения более хорошего приближения к χ^2-распределению Д. Лоули и А. Максвелл добавляют величину

$$\lambda^2 \cdot \sum_{i=1}^{k} \frac{1}{(\lambda_i - \lambda)^2}.$$

Проверяемая статистика сравнивается со значением χ^2 с

$$\frac{1}{2}(q+2)(q-1)$$

степенями свободы для уровня значимости α. Гипотеза о том, что $\lambda_{q+1} = \ldots = \lambda_P$, отвергается, если полученное значение критерия χ^2 превосходит табличное значение для заданного α и числа степеней свободы $1/2\,(q+2)\,(q-1)$.

Схема применения метода главных компонент в прикладных исследованиях такова:

1) проверка нормальности закона распределения;

2) нахождение выборочной ковариационной матрицы и ее собственных векторов и собственных значений;

3) отбрасывание незначимых собственных векторов;

4) проектирование точек выборочного ореола на направления главных компонент;

5) построение гистограмм проекций на отдельные оси и пары главных осей;

6) интерпретация полученных результатов.

ГЕОМЕТРИЧЕСКИЙ МЕТОД ПИРСОНА

Выше приведена схема статистического варианта метода главных компонент. В том случае когда нет оснований считать исходные данные случайной выборкой из совокупности, положения изучаемых объектов строго фиксированы и изучению подлежит их конфигурация, можно использовать детерминистский вариант метода главных компонент — геометрический метод Пирсона, в основе которого лежит уже изложенная выше схема исследования выборочных данных, примененная к матрице рассеяния данной системы точек, т. е. к матрице $\Sigma = \|\sigma_{ij}\|_1 \leqslant i, j \leqslant P$, где

$$\sigma_{ij} = \sum_{k=1}^{N} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j).$$

Нахождение пространства главных компонент сводится к проектированию заданного множества точек на пространство меньшей размерности. В качестве критерия наилучшего проектирования выберем условие минимизации суммы квадратов перпендикуляров, опущенных из заданных точек на подпространство, которое определено центром тяжести системы точек и k ортогональными векторами $L_1, \ldots, L_q$, т. е. условием минимизации выражения

$$\sum_{i=1}^{P}\sum_{r=1}^{N} (x_{ir} - \bar{x}_i)^2 - \sum_{i=1}^{k} L_i \sum L_i'.$$

Нахождение минимума этого выражения равносильно нахождению максимума выражения $\sum_{i=1}^{k} L_i \sum L_i'$. Этот максимум достигается, когда L_i, $i = 1, \ldots k$ — собственные векторы матрицы Σ.

Необходимо отметить, что вращение выделенных главных компонент так, как это делается в факторном анализе, производить нельзя. Обоснование процедур вращения заданного пространства в связи с задачей снижения размерности пространства дано В. М. Бухштабером и В. К. Масловым [10].

КОМПЛЕКСИРОВАНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ С ДРУГИМИ СТАТИСТИЧЕСКИМИ МЕТОДАМИ

Метод главных компонент обычно используется в комплексе с другими статистическими методами: регрессионным, дискриминантным анализом, методами геостатистики, дисперсионным анализом, кластерным анализом.

Опишем вкратце некоторые схемы комплексирования.

Регрессия на главных компонентах. Пусть имеется два случайных вектора y и x размерности соответственно q и P. Предположим, что $M(x) = M(y) = 0$. Совместная дисперсионная матрица векторов y и x имеет вид:

$$\begin{pmatrix} \Gamma & \theta \\ \theta' & \Sigma \end{pmatrix},$$

где $\Sigma = M\{x, x'\}$; $\theta' = M\{y, x'\}$, $\Gamma = M\{y, y'\}$ или $\Sigma = X'X$; $\theta = Y'X$, $\Gamma = Y'Y$.

Здесь Σ — матрица порядка $p \times p$; θ — матрица порядка $q \times p$; Γ — матрица порядка $q \times q$. Задача состоит в нахождении такого линейного преобразования T вектора x в вектор z ($z = T'x$), что вектор z наилучшим образом предсказывает значение y в каждой точке наблюдения. Пусть T — матрица преобразования порядка $P \times q$. Тогда совместная дисперсионная матрица вектора (y, z) есть

$$\begin{pmatrix} \Gamma & \theta T \\ T'\theta' & T'\Sigma T \end{pmatrix}.$$

Она имеет порядок $2q \times 2q$. Остаточная дисперсионная матрица вектора y—z

$$D_{y-z} = \Gamma - \theta T (T'\Sigma T)^{-1} T'\theta'.$$

Качество предсказания y через z оценивается через след матрицы D_{y-z}. Оказывается, что $\min_T tr(D_{y-z})$ достигается для матрицы $T = u_0$; ее столбцы являются собственными векторами, соответствующими паре матриц $\theta'\theta$ и Σ, найденными из уравнения $(\theta'\theta - \lambda^\circ\Sigma)\, u_0 = 0$.

В частном случае, когда вектор y совпадает с вектором x, получаются обычные главные компоненты, так как $\Gamma = \theta = \Sigma$, и урав-

нение для определения собственных векторов принимает вид $(\Sigma - \lambda E)\, u = 0$. В этом случае собственные векторы матрицы Σ минимизируют не только след матрицы $D_{y-z} = \Sigma - \Sigma u\,(u'\Sigma u)^{-1} u^1 \Sigma'$, но и евклидову норму этой матрицы, т. е. корень квадратный из суммы квадратов ее элементов.

Описанная процедура позволяет записать уравнение линейной регрессии вектора y через главные компоненты вектора x. Таким образом достигаются сокращение размерности пространства и сжатие информации.

Дискриминантный анализ и главные компоненты. Рассмотрим метод построения линейной дискриминантной функции, основанной на расстоянии Махалонобиса. Пусть даны две выборки

$$\| X_{ij}^{(1)} \|_{1 \leqslant i \leqslant N_1,\ 1 \leqslant j \leqslant P};$$
$$\| x_{ij}^{2} \|_{1 \leqslant i \leqslant N_2,\ 1 \leqslant j \leqslant P}$$

объема N_1 и N_2, извлеченные из двух P-мерных совокупностей. Обозначим через $\overline{X}^{(1)} = (\bar{x}_1^{(1)}, \ldots, \bar{x}_P^{(1)})$ и $\overline{X}_2^{2} = (\bar{x}_1^{(2)}, \ldots, \bar{x}_P^{(2)})$ — векторы выборочных средних двух выборок, а через W ковариационную матрицу объединенной выборки:

$$W = \| W_{ij} \|_{1 \leqslant i,\ j \leqslant P},$$

где
$$W_{ij} = \frac{1}{N_1 + N_2 - 2} \left[\sum_{t=1}^{N_1} (x_{ti}^{(1)} - \bar{x}_i^{(1)})(x_{tj}^{(1)} - \bar{x}_j^{(1)}) + \right.$$
$$\left. + \sum_{t=1}^{N_2} (x_{ti}^{(2)} - \bar{x}_i^{2})(x_{tj}^{(2)} - \bar{x}_j^{(2)}). \right]$$

Определим расстояние Махалонобиса:

$$D^2 = (\bar{x}^{(1)} - \bar{x}^{(2)})\, W^{-1}\, (\bar{x}^{(1)} - \bar{x}^{(2)}).$$

Гипотеза об отсутствии различий в векторах средних для двух выборок проверяется на основании статистики:

$$\frac{N_1 N_2 (N_1 + N_2 - p - 1)}{P\,(N_1 + N_2)\,(N_1 + N_2 - 2)} D^2,$$

распределенной по закону $F_{P, N_1 + N_2 - P - 1}$.

С. Р. Рао устанавливает тесную связь расстояний D^2 с линейной дискриминантной функцией Р. Фишера. Задача, по Р. Фишеру, состоит в нахождении такой линейной комбинации переменных $x_1, \ldots, x_P$, что расстояние между выборками, вычисленное исходя из этой линейной комбинации, является максимальным. Пусть требуемая линейная комбинация есть

$$f = \sum_{i=1}^{P} l_i d_i = (l,\ d_x).$$

Максимизируемая функция — это отношение

$$\frac{N_1 N_2}{N_1 + N_2} \cdot \frac{l_1 (\bar{x}^{(1)} - x_1^{(2)}) + \ldots + l_P (\bar{x}_P^{(1)} - \bar{x}_P^{(2)})}{\sum_{i,j=1}^{P} l_i l_j W_{ij}}.$$

Вид этой функции показывает, что искомые коэффициенты линейной комбинации $l_1, \ldots, l_P$ определяются лишь с точностью до постоянного множителя. Поэтому условия для нахождения экстремума указанной функции можно записать в виде $Wl = d_x$, где $d_x = (d_1, \ldots, d_P) = (x_1^{(1)} - \bar{x}_1^{(2)}, \ldots, x_P^{(1)} - \bar{x}_P^{(2)})$—вектор разности выборочных средних двух выборок. Решая эту систему, получим $l = W^{-1} d_x$. Умножая уравнения системы соответственно на $l_1, \ldots, l_P$ и складывая полученные результаты, получаем

$$\sum_{i,j} l_i l_j W_{ij} = \sum_{i} l_i d_i = \sum_{i,j} W^{ij} d_i d_j = D^2 \left(\| W^{ij} \| = W^{-1} \right).$$

Поэтому оптимальное значение максимизируемой функции будет

$$\frac{N_1 N_2}{N_1 + N_2} \sum_{i,j} W^{ij} d_i d_j = \frac{N_1 N_2}{N_1 + N_2} D^2.$$

Установим связь вектора $l = (l_1, \ldots, l_P)$ с главными компонентами матрицы W. С этой целью в P-мерном пространстве переменных x сделаем ортогональное преобразование $y = U_x$; так что матрица $UWU' = \Lambda$ будет диагональной. Тогда вектор $y = Ux$ будет собственным вектором матрицы W. Введем обозначение $d_y = Ud_x$. На основании уравнения $Wl = d_x$ имеем $U'\Lambda U l = d_x$, откуда, учитывая, что $U^{-1} = U'$, получаем $\Lambda U l = d_y$ или $Ul = \Lambda^{-1} d_y$. Представляя линейную комбинацию $f = \sum_{i=1}^{P} l_i d_i$ в виде скалярного произведения $f = (l,\ d_x)$ векторов l и d_x и используя свойство инвариантности скалярного произведения относительно ортогонального преобразования пространства, получаем

$$f = (l,\ d_x) = (Ul,\ Ud_x) = (\Lambda^{-1} d_y,\ d_y).$$

С другой стороны, имеем формулы для расстояния Махалонобиса

$$D^2 = (W^{-1} d_x,\ d_x) = (W^{-1} U^{-1} d_y,\ U^{-1} d_y) =$$
$$= (UW^{-1} U^{-1} d_y,\ d_y) = (\Lambda^{-1} d_y,\ d_y), \text{ т. е. } f = D^2.$$

Таким образом, коэффициенты функции f, записанной в переменных y или в базисе, состоящем из собственных векторов матрицы W, выражаются в виде $\Lambda^{-1} d_y$, где Λ — матрица собственных значений матрицы W.

Из приведенной формулы вытекает следующая процедура построения линейной дискриминантной функции. Если $\lambda_1, \lambda_2, \ldots, \lambda_q$ — q максимальных собственных значений матрицы W, состав-

ляющих в сумме не менее 75 % от их общей суммы, то коэффициенты линейной дискриминантной функции f в переменных y принимаются равными

$$(1/\lambda_1)\, d_{y_1},\ \ldots,\ (1/\lambda_q)\, d_{y_q}.$$

Остальные $P-q$ коэффициентов заменяются нулями.

Метод, сочетающий основные черты дисперсионного и факторного анализа, был создан Г. Ф. Голлобом [7].

ГЛАВА 13

ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ — статистический метод описания природных и социальных явлений с помощью некоторого числа основополагающих внутренних параметров — общих факторов, т. е. в виде

$$z_i = F_i(f_1, f_2, \ldots, f_k) + e_i, \qquad i = 1, 2, \ldots, n,$$

где $z = (z_1, \ldots, z_n)'$ — n-мерный вектор-столбец наблюдаемых переменных; F_i — некоторые многочлены переменных — факторов $f_1, \ldots, f_k$; $l = (l_1, \ldots, l_n)$ — n-мерный вектор-столбец специфических факторов, влияющих только на данную переменную. Предполагается, что они не коррелированы как между собой, так и с общими факторами f. Факторы $f_1, \ldots, f_k$ обычно предполагаются некоррелированными между собой. Все они имеют определенную интерпретацию в рамках решаемой задачи.

Рассмотрим простейшую линейную модель факторного анализа, когда функции F_i являются линейными функциями переменных — факторов $f_1, f_2, \ldots, f_k$ [7]:

$$z_i = \sum_{r=1}^{k} l_{ir} f_r + e_i, \qquad i = 1, 2, \ldots, n.$$

Коэффициент l_{ir} называют нагрузкой i-й переменной на r-й фактор. Обозначая через

$$L = \| l_{ir} \|_{1 \leqslant i \leqslant n,\ 1 \leqslant k \leqslant r}$$

матрицу факторных нагрузок, запишем основное уравнение факторного анализа в матричной форме: $z = Lf + e$, $f = (f_1, \ldots, f_k)'$.

Неизвестными параметрами линейной модели являются факторные нагрузки и дисперсии специфических факторов. Число неизвестных в этой системе, равное $nk + n$, значительно превышает число уравнений. Поэтому для их оценки прибегают обычно к информации, содержащейся в корреляционной матрице. Из уравнения $z = Lf + e$ легко получить соотношение $R = L\Phi L' + v$, где R — корреляционная матрица наблюдаемых переменных; Φ — корреляционная матрица общих факторов, которая в предположении некоррелированности факторов становится единичной матрицей; v — ковариационная матрица специфических факторов, являющаяся диагональной. Таким образом, в случае некоррелированных факторов, т. е. при $\Phi = E$, получаем $R = LL' + V$.

К числу исходных предпосылок, удовлетворение которых позволяет использовать линейную модель факторного анализа с достаточно высокой надежностью, можно назвать следующие.

1. Исходный набор наблюдаемых переменных равноправен с точки зрения причинно-следственных связей, т. е. изменения переменных обусловлены влиянием ряда общих и специфических факторов.
2. Исследуемый набор наблюдаемых переменных подчиняется многомерному нормальному закону распределения.
3. Специфические факторы l_i не коррелированы ни между собой, ни с общими факторами.
4. Число общих факторов, определяющих изучаемое явление, должно быть значительно меньше числа анализируемых переменных.
5. Корреляционная матрица наблюдаемых переменных должна быть устойчивой по отношению к изменению выборки.
6. В исходных наблюдениях отсутствует автокорреляция.
7. Выборка исходных данных должна быть представительной.

На практике линейная модель часто используется даже в тех случаях, когда какое-либо из условий не выполняется. Исследуя полученные результаты, можно так модифицировать модель, чтобы она максимально приближалась к реальной ситуации.

Наиболее обоснованные с математической точки зрения модели факторного анализа — это метод минимальных остатков Хармана и метод максимального правдоподобия Лоули и Максвелла. Охарактеризуем их вкратце.

МЕТОД МИНИМАЛЬНЫХ ОСТАТКОВ ХАРМАНА

Факторные нагрузки в методе Хармана определяются из условия минимизации по методу наименьших квадратов суммы квадратов внедиагональных элементов остаточной корреляционной матрицы. Пусть

$S = \| S_{ik} \|_{1 \leqslant j,\ k \leqslant n}$ — выборочная корреляционная матрица,

$L = \| l_{ir} \|_{1 \leqslant i \leqslant n,\ 1 \leqslant r \leqslant k}$ — искомая матрица факторных нагрузок.

Метод Хармана соответствует минимизации нормы матрицы:

$$\begin{pmatrix} 0 & S_{12} - \sum_{r=1}^{k} l_{1r} l_{2r} & \ldots & S_{1n} - \sum_{r=1}^{k} l_{1r} l_{nr} \\ S_{21} - \sum_{r=1}^{k} l_{2r} l_{1r} & 0 & \ldots & S_{2n} - \sum_{r=1}^{k} l_{2r} l_{nr} \\ \ldots & \ldots & \ldots & \ldots \\ S_{n1} - \sum_{r=1}^{k} l_{nr} l_{1r} & S_{n2} - \sum_{r=1}^{k} l_{nr} l_{2r} & \ldots & 0 \end{pmatrix}$$

Минимизируемая функция имеет вид:

$$f(L)=\sum_{k=j+1}^{n}\sum_{j=1}^{n-1}\left(S_{jk}-\sum_{r=1}^{k}l_{jr}l_{kr}\right)^2.$$

Цель метода минимальных остатков состоит в том, чтобы, меняя значения факторных нагрузок при фиксированном k, минимизировать функцию $f(L)$ при условии

$$\sum_{r=1}^{k}l_{jr}^2\leqslant 1, \qquad j=1,\ 2,\ \ldots,\ n.$$

Указанное условие вытекает из соотношений для элементов матрицы:

$$l_{j1}^2+l_{j2}^2+\ \ldots\ +l_{jk}^2+d_j^2=1, \qquad j=1,\ 2,\ \ldots,\ n,$$

где члены d_j^2 соответствуют второй компоненте факторного отображения, d_j^2 — дисперсия j-го специфического фактора.

Задача нахождения минимума функции $f(L)$ решается методом последовательных приближений на основе стандартных процедур оптимизации функций.

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ЛОУЛИ И МАКСВЕЛЛА

Задача ставится так: используя выборочную корреляционную матрицу

$$\hat{S}=\|\hat{S}_{ij}\|_{1\leqslant i,\ j\leqslant n}$$

n-мерной случайной величины $z=(z_1,\ \ldots,\ z_n)$ и предполагая число факторов k заданным, дать эффективные оценки параметров l_{ir} и элементов v_i диагональной матрицы V.

Для решения этой задачи строим функцию правдоподобия:

$$L=-\frac{N}{2}\ln|R|-\frac{N}{2}\sum_{i,\ j=1}^{N}\hat{S}_{ij}r^{ij},$$

где N — объем выборки. Максимум этой функции реализуется при выполнении следующих условий:

$$v_i=S_{ii}-\sum_{r=1}^{k}l_{ir}^2; \qquad L'-L'R^{-1}S=0,$$

где $R=LL'+V$.

Полученная система уравнений решается методом последовательных приближений. Задаваясь начальными факторными нагрузками $L_{ir}^{(1)}$, выбранными произвольно, из первого уравнения находим первое приближение $v^{(1)}$ к матрице v. Затем по матрицам $L^{(1)}$ и $v^{(1)}$ вычисляют первое приближение $R^{(1)}$ к матрице $R=LL'+v$. Второе уравнение позволяет определить второе приближение к матрице факторных нагрузок L и т. д.

Доказательство сходимости алгоритма в общем случае отсутствует. В связи с этим результат часто оказывается зависящим от

выбора начального приближения. В геологии этот метод не применялся.

Оценка числа факторов. В рассмотренных методах факторного анализа предполагается заранее заданным число факторов k. Приведем критерий оценки числа факторов, принадлежащий Д. Риппу. Статистика

$$U_k=(N-1)\ln\frac{LL'+v}{|S|}$$

в предположении нормального распределения исходных параметров распределена по закону

$$\chi_\nu^2 \text{ с } \nu=\frac{1}{2}[(n-k)^2+(n+k)^2]$$

степенями свободы. Если статистика U_k превышает значение χ^2 при некотором уровне значимости, то гипотеза о том, что число факторов равно k, отклоняется. В противном случае гипотеза принимается. При отклонении гипотезы можно предположить, что число факторов больше k.

Вращение в пространстве факторов. В основе процедуры вращения в пространстве факторов лежит наглядная интерпретация факторов и факторных нагрузок. Если представить себе факторные нагрузки как координаты точки в k-мерном пространстве факторов, а сами факторы считать ортогональными осями в этом пространстве, то преобразование факторного решения есть по существу вращение этих осей вокруг начала координат. Ясно, что это вращение можно выбрать бесконечным числом способов.

Существуют различные методы реализации вращения факторов: графические и аналитические. В основе графического метода лежит следующий принцип: изображая на плоскости точки с координатами, равными факторным нагрузкам, мы получаем ореол точек; новые оси выбираются так, чтобы вблизи них лежало по возможности больше точек. Угол поворота в этом случае определяется лишь приближенно, так что графический метод вращения факторов является в сущности эвристическим.

При большем числе факторов рассматриваются всевозможные двумерные проекции многомерной картины, в каждой из двумерных плоскостей производится вращение. Матрица вращений в пространстве факторов будет произведением матриц вращений в каждой плоскости.

Аналитические методы вращения факторов основаны на следующей идее: применение ортогонального преобразования в пространстве факторов ведет к ортогональному преобразованию матрицы факторных нагрузок. Матрица L преобразуется в матрицу M так, что выполняются следующие условия:

$$\sum_{r=1}^{k}m_{ir}^2=\sum_{r=1}^{k}l_{ir}^2=\text{const}, \qquad i=1,\ 2,\ \ldots,\ n.$$

Так как теоретический верхний предел суммы $Q=\sum_{i=1}^{n}\sum_{r=1}^{k} m_{ir}^4$ достигается в том идеальном случае, когда каждая переменная зависит лишь от одного фактора, то в качестве критерия качества преобразования Г. А. Фергюсон выбирает величину Q. Наилучшее вращение соответствует ее максимуму.

ГЛАВА 14

ИНФОРМАТИВНЫЕ КОМБИНАЦИИ ПРИЗНАКОВ

ВЫБОР ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ПРИЗНАКОВ ОТНОСИТЕЛЬНО МНОГОМЕРНЫХ СРЕДНИХ

Выбор полной и наилучшей комбинаций признаков, информативных относительно многомерных средних, на основе рангового критерия Пури—Сена—Тамуры [26]. (рис. 57).

1. Из двух m-мерных выборок объема n_1 и n_2 по каждому признаку j в отдельности составляются вариационные ряды для объединенной выборки объема $n = n_1 + n_2$ и рассчитываются m одномерных модификаций критерия Пури—Сена—Тамуры, т. е. Λ_j ($j = 1, 2, \ldots, m$), которые имеют χ^2-распределение с одной степенью свободы. Выбираем то значение $j = j_1$, для которого Λ_j максимальна, т. е. $\Lambda_{j_1} = \max \Lambda_j$.

2. Вычисляется m—1 величин ранговой статистики Λ критерия Пури—Сена—Тамуры для двумерных величин, образованных j_1 признаком и одним из оставшихся $j \neq j_1$. Выбирается то значение j_2, для которого максимальна статистика

$$\Lambda_{j_1 j_2}, \text{ т. е. } \Lambda_{j_1 j_2} = \max_{j} \Lambda_{j_1, j}.$$

3. Процедура повторяется для всех $j = 2, 3, \ldots, m$. Так, для $j = j_q \leqslant m$ вычисляется $(m-g)$ значений ранговой статистики Λ критерия Пури—Сена—Тамуры для g-мерных величин, образованных $j_1 j_2, \ldots, j_{(q-1)}$ признаками на предыдущем шаге и одним из оставшихся $j \neq j_1, j_2, \ldots, j_{(q-1)}$. Вновь выбирается то значение j_q, для которого максимальна статистика

$$\Lambda_{j_1, j_2, \ldots, j_g}, \text{ т. е. } \Lambda_{j_1, j_2, \ldots, j_g} = \max_{j} \Lambda_{j_1, j_2, \ldots, j_{(g-1)}, j}.$$

При $g = m$ полученная последовательность номеров признаков $j_1, j_2, \ldots, j_m$ будет соответствовать их расположению от наилучшего к наихудшему.

4. Ранжируя с помощью приведенного критерия имеющейся в распоряжении геолога ряд из m признаков, получим набор статистик:

$$\max_j \Lambda_j^{(1)},\ \max_j \Lambda_{j_1, j}^{(2)},\ \max_j \Lambda_{j_1, j_2, j}^{(3)} \ldots \max_j \Lambda_{j_1, j_2, \ldots, j_{(g-1)}, j}^{(g)}, \ldots, \Lambda^{(m)},$$

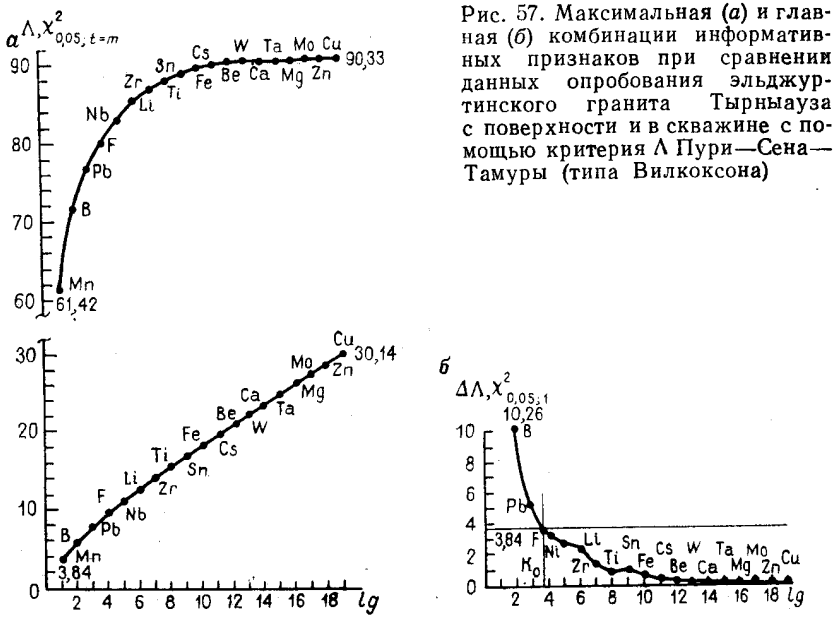


Рис. 57. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия Λ Пури—Сена—Тамуры (типа Вилкоксона)

поставленный в соответствие с упорядоченным рядом самих признаков $j_1, j_2, \ldots, j_g, \ldots, j_m$. В обозначении

$$\Lambda_{j_1, j_2, \ldots, j_{(g-1)}, j}^{(g)}$$

индекс сверху в скобках показывает число признаков, участвующих в вычислении статистики, а внизу перечислены номера признаков.

Обозначим для большей краткости указанные статистики следующим образом:

$$\Lambda_g = \max_j \Lambda_{j_1, j_2, \ldots, j_{(g-1)}, j}^{(g)}.$$

5. Статистика λ_g критерия Пури—Сена—Тамуры имеет асимптотическое χ^2-распределение с g степенями свободы.

При заданном уровне значимости α ряд из $(1-\alpha)$ квантилей χ^2-распределения с 1, 2, . . . , m степенями свободы (т. е. 3,84; 5,99; 7,81; 9,49 и т. д.) является рядом критических точек, превышения значений которых соответствующими статистиками λ_1, λ_2, λ_3, . . . , λ_m означает в конечном счете информативность той комбинации признаков, для которой вычислена соответствующая статистика λ_g.

6. Комбинация из g первых признаков $(j_1, j_2, \ldots, j_g)$ упорядоченного ряда признаков называется полной информативной комбинацией, если для комбинации из $i = m-g$ признаков, где $i > g$, принимается нулевая гипотеза о равенстве многомерных средних, т. е. если

$$\Lambda_i > \chi^2_{\alpha, f=i} \text{ для } i = 1, 2, \ldots, g \quad (H_1: M\xi \neq M\eta),$$

$$\Lambda_i \leqslant \chi^2_{\alpha,\, f=i} \text{ для } i>g \qquad (H_0: M\xi = M\eta).$$

7. Если полная информативная комбинация признаков состоит более чем из одного признака ($g \geqslant 2$), то с помощью ряда из приращений значений статистик

$$\Delta\Lambda_i = \Lambda_i - \Lambda_{i-1} \qquad (i = 2,\ 3,\ \ldots,\ g)$$

можно сделать дополнительные выводы о той доле различающей информации, которую вносит каждый добавляемый признак.

8. Убывающая информативность признаков в ранжированном ряду является причиной того, что величины $\Delta\Lambda_i$ статистики уменьшаются с увеличением номера i. Первые g_0 признаков будут наилучшей информативной комбинацией, если

$$\Delta\Lambda_i > \chi^2_{\alpha,\, f=i} \text{ для } i = 2,\ 3,\ \ldots,\ g_0 \qquad (H_1: \Delta\beta_i \neq 0),$$

$$\Delta\Lambda_i \leqslant \chi^2_{\alpha,\, f=i} \text{ для } i > g_0 \qquad (H_0: \Delta\beta_i = 0).$$

9. Поскольку в полной информативной комбинации признаков всегда $\Lambda_1 > \chi^2_{\alpha,\, f=1}$, то при незначимости всех приращений $\Delta\Lambda_i$ для $i = 2,\ 3,\ \ldots,\ g$ наилучшую информативную комбинацию g_0 будет составлять один наиболее информативный признак, а именно первый j_1 в ранжированном ряду.

10. Комбинацию признаков $i = m-g$, для которой принимается нулевая гипотеза о равенстве многомерных средних

$$\Lambda_i \leqslant \chi^2_{\alpha,\, f=i} \text{ для } i > g,$$

следует полагать неинформативной или малоинформативной, т. е. обеспечивающей, скорее всего, черты сходства сопоставляемых объектов.

Выбор полной и наилучшей комбинации признаков, информативных относительно многомерных средних, на основе параметрического критерия Джеймса — Сю (рис. 58). Процедура ранжирования геологических признаков (свойств) от наиболее информативного к наименее информативному, приемы поиска полной I_g и наилучшей I_{g0} информативной, а также неинформативной I_{m-g} комбинаций признаков на основе параметрического критерия Джеймса—Сю полностью идентичны тем, которые приведены при описании выбора полной и наилучшей комбинаций признаков на основе рангового критерия Пури—Сена—Тамуры. Вновь будем иметь 10 этапов от поиска наиболее информативного признака (этап 1) до определения неинформативной комбинации признаков (этап 10). Естественно, что применяться должны не λ статистика критерия Пури—Сена—Тамуры, а статистика $2I$ критерия Джеймса—Сю. Так, для ранжирования g-го признака, т. е. $j = j_g \leqslant m$, вычисляется $(m-g)$ значений параметрической статистики $2I$ критерия Джеймса—Сю для g-мерных величин, образованных $(j_1, j_2, \ldots, i_{(g-1)})$ признаками

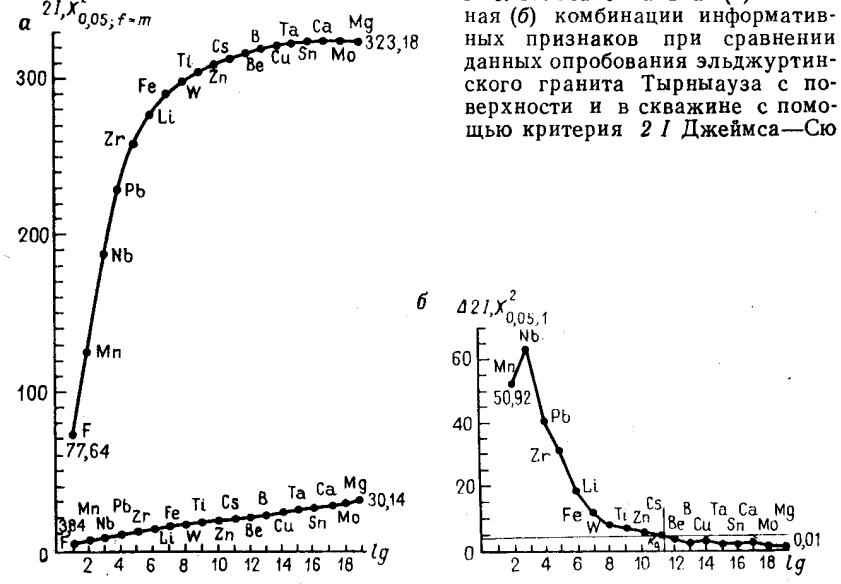


Рис. 58. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия *2 I* Джеймса—Сю

на предыдущем шаге и одним из оставшихся $j \neq (j_1, j_2, \ldots, j_{(g-1)})$. Выбирается то значение j_g, для которого максимальна статистика

$$2I = \max 2I_{j_1,\, j_2,\, \ldots,\, i_{(g-1)}}.$$

ВЫБОР ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ПРИЗНАКОВ ОТНОСИТЕЛЬНО КОВАРИАЦИОННЫХ МАТРИЦ

Выбор полной и наилучшей комбинаций признаков, информативных относительно ковариационных матриц, на основе рангового критерия Пури — Сена — Тамуры (рис. 59). По-прежнему отмечаем, что процедуры ранжирования геологических признаков от наиболее к наименее информативным, поиск полной I_g и наилучшей I_{g_0} информативной, а также неинформативной $I_{(m-g)}$ комбинаций признаков на основе рангового критерия Пури—Сена—Тамуры полностью идентичны тем, которые описаны выше. Следует пользоваться рабочей статистикой Λ_Σ критерия Пури—Сена—Тамуры.

Выбор полной и наилучшей комбинаций признаков, информативных относительно ковариационных матриц, на основе параметрического критерия Кульбака (рис. 60). Процедура ранжирования признаков, поиска полной и наилучшей информативной и неинформативной комбинаций признаков идентична вышеописанной.

Специфические отличия — это то, что привлекается рабочая статистика $2I_0$ критерия Кульбака и при выборе наилучшей информативной комбинации признаков критические значения χ^2 пред-

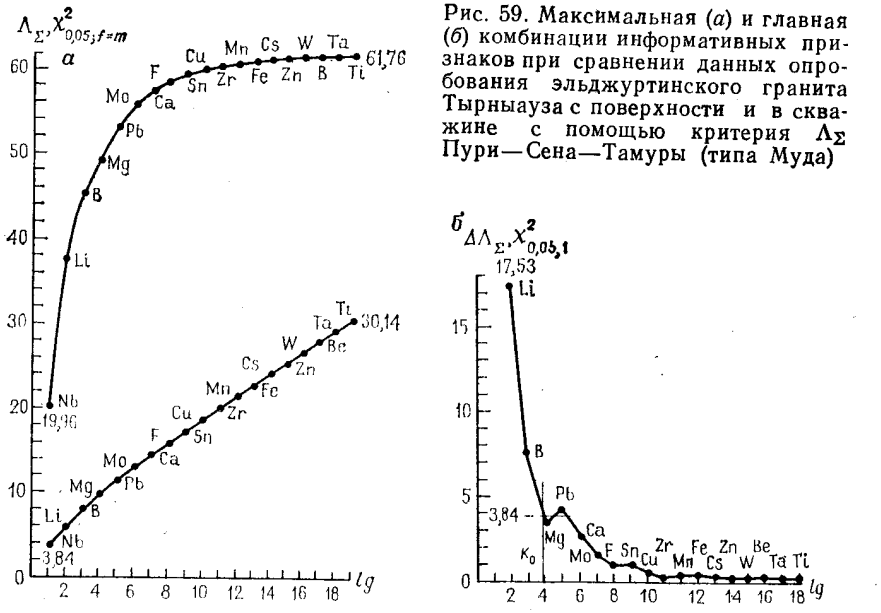


Рис. 59. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия Λ_Σ Пури—Сена—Тамуры (типа Муда)

ставляют собой монотонно возрастающий ряд: $\chi^2_{j=1} = 3{,}84$; $\chi^2_{j=2} = 5{,}99$; $\chi^2_{j=3} = 7{,}81$ и т. д., а не константу 3,84 (как при использовании статистики Λ_Σ Пури—Сена—Тамуры).

Выбор информативных комбинаций геологических характеристик для одного геологического объекта. Наряду с задачей выявлений информативных, контрастных, геологических характеристик при сопоставлении двух геологических объектов, не меньшее практическое значение имеет задача определения наиболее важных информативных геологических характеристик для одного геологического объекта.

При определении формы зависимости геологических характеристик центральной является проблема выбора из всего набора регрессоров — геологических характеристик такого их подмножества, которое позволит статистически значимо описать зависимую переменную с помощью уравнения регрессии. Другими словами, все множество регрессоров нужно разделить на две группы, одна из которых должна содержать переменные, которые позволяют построить значимую регрессию, а другая — переменные, влиянием которых (дополнительно к влиянию информативной комбинации) на зависимую переменную можно пренебречь.

Можно предложить три процедуры поиска.

1. Жесткая процедура. На каждом шаге число регрессоров, входящих в информативную комбинацию, увеличивается на единицу, причем комбинация регрессоров, полученная на предыдущем шаге, обязательно включается в новую комбинацию. Дополнительный регрессор определяется из условия обращения в максимум (или минимум) определенной статистики, соответствующей выбранному критерию. Жесткая процедура может быть реализована в четырех модификациях, а именно поиск очередного регрессора может проводиться как из обращения в максимум выбранного критерия, так и из условия обращения этого критерия в минимум. Выделенное подмножество регрессоров в одном случае удобно называть информативной комбинацией регрессоров, а в другом случае — неинформативной комбинацией регрессоров. Кроме того, процедура может быть начата с анализа либо единственного признака, либо полного набора регрессоров.

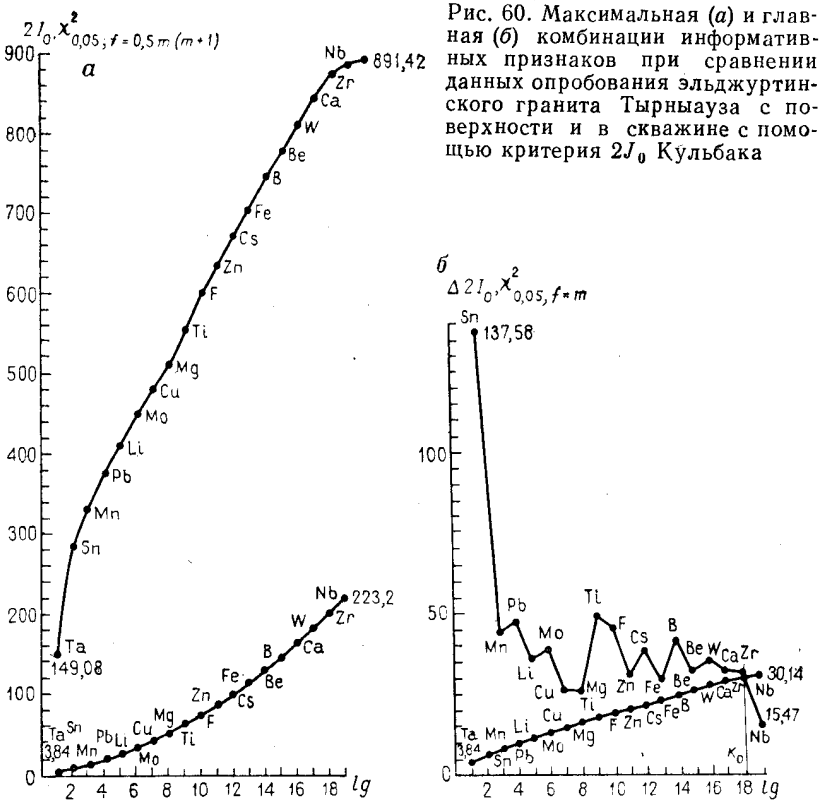


Рис. 60. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия $2J_0$ Кульбака

2. Процедура полного перебора. В этой процедуре выделенные на предыдущих шагах наилучшие подмножества регрессоров не включаются автоматически в последующую комбинацию, а находятся из условия обращения в максимум или минимум определенной статистики, соответствующей принятому критерию, последовательным полным перебором всех возможных вариантов сочетаний признаков.

Как и для жесткой процедуры, для процедуры полного перебора возможны те же четыре модификации.

3. Процедура с нахождением ядра информативности. Поскольку жесткая процедура экономична в вычислительном плане, а процедура полного перебора приводит к точному решению задачи поиска, нам представляется разумным рекомендовать компромиссную стратегию поиска наилучшего из подмножества регрессоров. Анализ практических результатов, полученный на ЭВМ с использованием жесткой процедуры и процедуры полного перебора, показал, что поиск небольшого числа наиболее важных регрессоров следует производить максимально точно, остальные же регрессоры могут быть выделены более экономичным путем, хотя бы и с определенной потерей точности.

Это будет процедура, которая в начале работы использует полный перебор регрессоров или шаговый метод для определения так называемого ядра информативности, а затем экономичную жесткую процедуру. Использование шагового метода вместо полного перебора позволит сделать процедуру более экономичной.

ГЛАВА 15

ЭКСТРЕМАЛЬНЫЕ ЗНАЧЕНИЯ

Теория экстремальных значений обобщает статистические сведения об экстремальных значениях и позволяет предсказать их при последующих экспериментах. При этом предполагают, что экстремальные значения имеют случайный характер. Задачи, которые позволяет решать теория экстремальных значений, следующие.

1. Не выпадает ли за разумно ожидаемые границы значение некоторой случайной величины — модели геологической характеристики в выборке, взятой из совокупности с известным распределением?

2. Наблюдается ли какая-либо закономерность в ряду экстремальных значений?

Теория экстремальных значений позволяет адекватно описывать такие геологические процессы, как паводки, землетрясения, горные обвалы, сели, различные атмосферные явления и т. п.

Большой вклад в развитие теории экстремальных значений внес Б. В. Гнеденко.

Теория экстремальных значений оперирует с некоторыми специальными понятиями, такими как характеристическое n-е наибольшее значение, функция интенсивности (рис. 61), период повторяемости и т. п. Характеристическое n-е наибольшее значение u_n для $n \geqslant 2$ определяется из уравнения

$$F(u_n) = 1 - 1/n,$$

функция интенсивности

$$\mu(x) = \frac{f(x)}{1 - F(x)}.$$

Функция распределения $F(x)$ может быть получена исходя из функции интенсивности:

$$1 - F(x) = [1 - F(x_0)] \exp\left[-\int_{x_0}^{x} \mu(z)\, dz\right],$$

где x_0 — произвольное значение x.

Период повторяемости определяется выражением

$$T(x) = \frac{1}{1 - F(x)}.$$

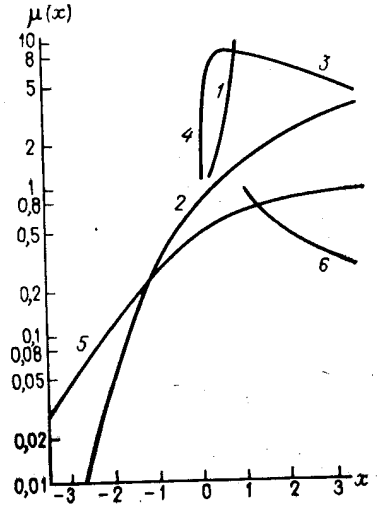


Рис. 61. Функции интенсивности $\mu(x)$ для различных типов распределений:
1 — равномерное; *2* — нормальное; *3* — логнормальное; *4* — экспоненциальное; *5* — логистическое; *6* — Парето

Теория экстремальных значений частично перекрывается с теорией порядковых статистик, так что методы последней применимы для решения экстремальных проблем. Точные распределения экстремальных значений известны для исходного нормального распределения [8, 19]. Теория экстремальных значений базируется на предположении об экспоненциальном характере исходного распределения и позволяет изучать соотношения между математическим ожиданием, медианой, модой, характеристическим наибольшим значением и экстремальными интенсивностями.

Эти соотношения могут быть получены в предположении экспоненциальности распределения случайной величины — модели геологической характеристики и будут справедливы для распределений специального (экспоненциального) типа.

Распределения экспоненциального типа — это распределения случайной величины ξ, для которых асимптотически выполнены равенства:

$$\frac{f(x)}{1 - F(x)} = \frac{-f'(x)}{f(x)}, \quad \frac{f(x)}{F(x)} = \frac{f'(x)}{f(x)},$$

где $f(x)$ и $F(x)$ — соответственно плотность вероятности и функция распределения случайной величины ξ.

Распределения экспоненциального типа разбиваются на три класса в зависимости от того, чему равняется для больших положительных x значение критического отношения $Q(x)$:

$$Q(x) = \frac{-f^2(x)}{f'(x)[1 - F(x)]},$$

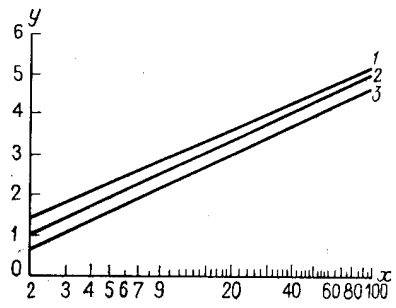


Рис. 62. Средние наибольших значений экспоненциального распределения:
1 — среднее; *2* — медиана; *3* — мода (характеристическое наибольшее значение)

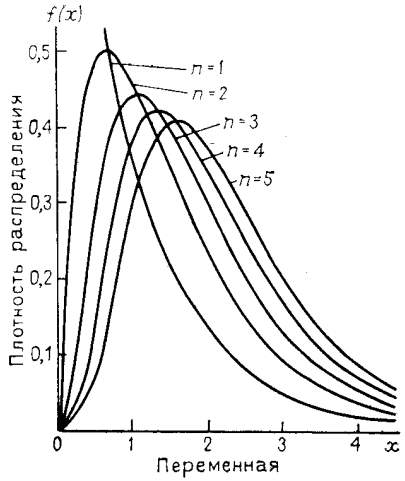


Рис. 63. Плотности распределения $f(x)$ наибольших значений экспоненциального распределения для различных объемов выборки n

1 класс: $Q(x) = 1 + |\varepsilon(x)|$; 2 класс: $Q(x) = 1$; 3 класс: $Q(x) = 1 - |\varepsilon(x)|$, где

$$\lim_{x\to\infty} |\varepsilon(x)| = 0.$$

К первому классу принадлежат такие распределения, как логистическое, гамма, нормальное и др., к третьему классу — логарифмически нормальное.

Для экспоненциального распределения справедливо

$$F(x) = 1 - e^{-x}; \quad f(x) = -f'(x) = 1 - F(x) = e^{-x};$$

$$\mu(x) = 1; \qquad T(x) = e^{x}, \qquad x \geqslant 0; \qquad u_n = \tilde{x}\tilde{x}_n = \ln n,$$

где x_n — мода наибольшего значения; $\tilde{x}_n \approx \ln n - \ln \ln 2$, где $\tilde{x}_n$ — медиана наибольшего значения.

На рис. 62, 63 представлены среднее, медиана, мода и характеристическое наибольшее значение для различного числа наблюдений.

Функция распределения наибольшего значения

$$\Phi_n(x) = (1 - e^{-x})^n,$$

а плотность $\varphi_n(x) = n(1 - e^{-x})^{n-1} e^{-x}$.

Равенство наиболее вероятного и характеристического значений свидетельствует о том, что для увеличения вдвое наиболее вероятного наибольшего значения необходимо квадратичное увеличение числа наблюдений.

Для распределений экспоненциального типа равенство наиболее вероятного экстремума и характеристического экстремума выпол-

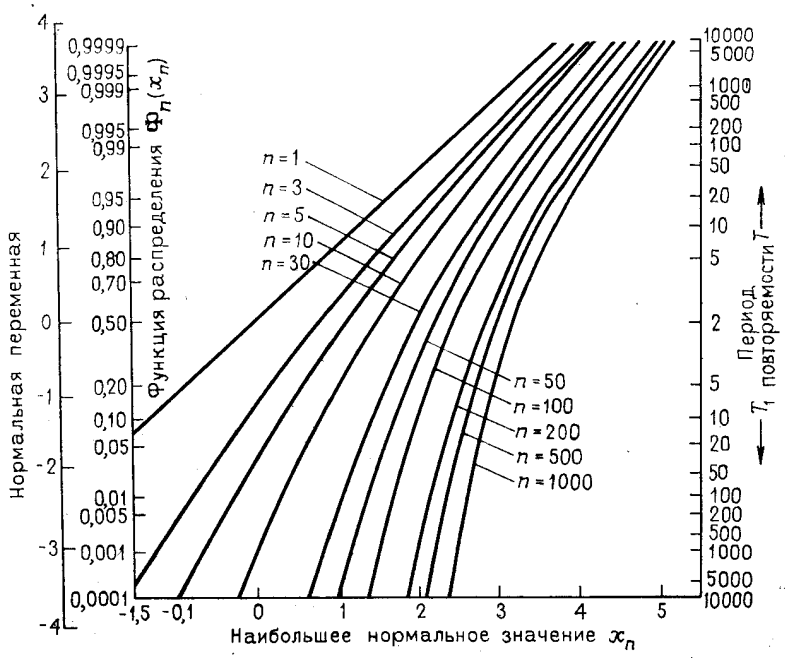


Рис. 64. Функции распределения нормальных экстремальных значений

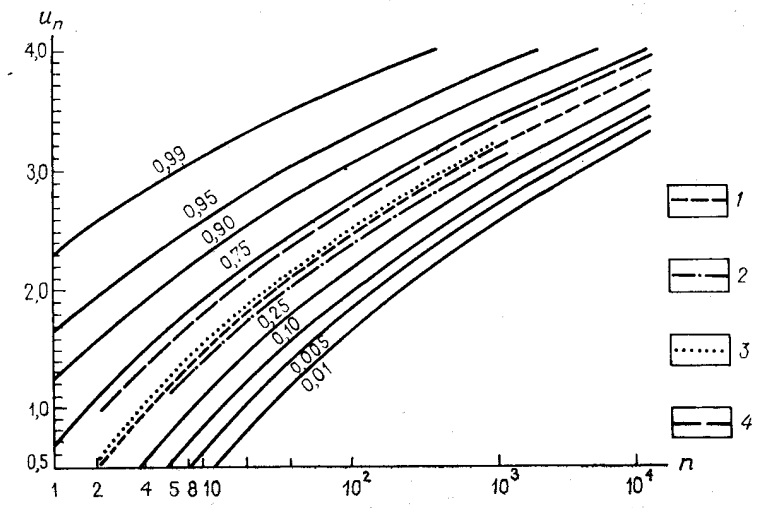


Рис. 65. Квантили нормальных экстремальных значений:
1 — медиана; *2* — мода; *3* — среднее; *4* — медиана наибольшего отклонения

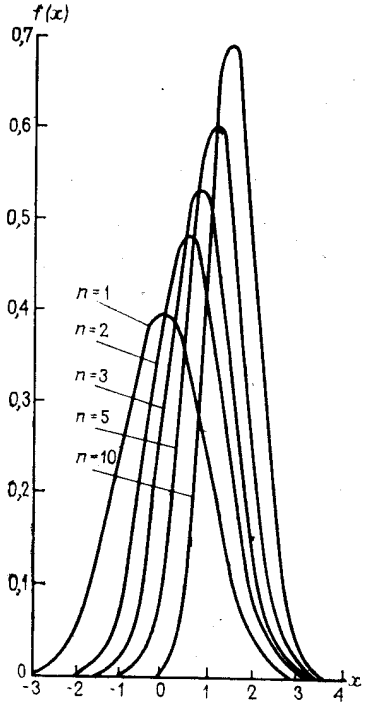


Рис. 66. Плотности распределения нормальных экстремальных значений для различных объемов выборок n

няется асимптотически. Число наблюдений, практически достаточное для такого предположения, зависит от исходного распределения и от требуемой точности вычисления.

Также асимптотически выполняется условие, что медианы наибольшего значения больше, чем моды, и другие важные условия, выполняющиеся точно для экспоненциального распределения.

Для нормального распределения приводим основные графики для экстремальных значений: функции распределения (рис. 64), квантили (рис. 65), плотности распределения (рис. 66), средние (рис. 67), различные средние квадратичные отклонения (рис. 68).

Разность между наблюдаемым экстремумом и выборочным средним называется экстремальным отклонением и используется в качестве критерия для возможного отбрасывания крайних выборочных значений.

Наряду с классом распределений экспоненциального типа в теории экстремальных значений широко используются распределения, являющиеся обобщением распределения Парето. Эти распределения образуют класс распределений типа Коши, для которых справедливы отношения:

$$\lim_{z\to\infty}[1-F(z)]z^k=A, \qquad A>0,$$

$$\lim_{z\to-\infty}F(z)(-z)^{k_1}=A_1,$$

где

$$k=\lim_{n\to\infty}\mu_n v_n;$$

v_n — характеристическое наибольшее значение; μ_n — значение функции интенсивности в точке $x=v_n$, k_1 определяется аналогично k.

Распределения типа Коши разбиваются на три класса в зависимости от величины

$$d^2v_n/dn^2,$$

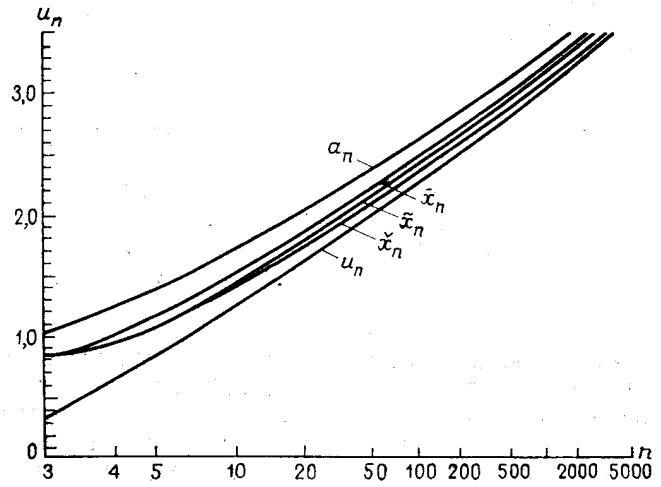


Рис. 67. Средние нормальных экстремальных значений в натуральном масштабе:

a_n — экстремальная интенсивность; $\bar{x}_n$ — среднее; $\tilde{x}_n$ — медиана; $\hat{\lambda}_n$ — мода; u_n — характеристическое наибольшее значение

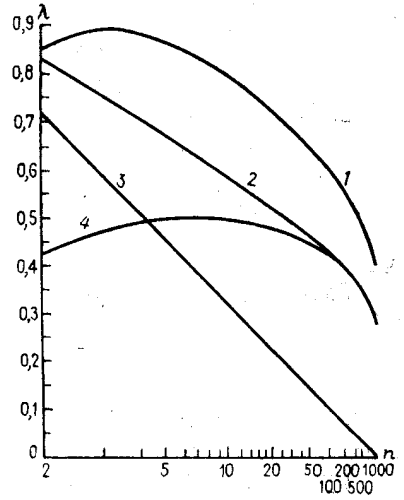


Рис. 68. Средние квадратические отклонения для нормальных экстремальных значений, отклонений и размахов:

1 — размахов; *2* — наибольших значений; *3* — среднего; *4* — экстремального отклонения

где

$$\frac{d^2v_n}{dn^2}=\frac{1}{k}\frac{v_n}{n^2}\left(\frac{1}{k}-1\right).$$

Если d^2v_n/dn^2 отрицательно (при $k>1$), то имеет место третий класс, если положительно (при $k<1$), то первый, а если равна нулю (при $k=1$), то второй.

Кроме точных распределений экстремальных значений (в зависимости от объема выборки) известны асимптотические (предельные) распределения.

Первое предельное распределение имеет место для исходных распределений экспоненциального типа, второе — для распределений типа Коши, а третье — для ограниченных распределений.

Все три предельных распределения получены при условии выполнения специального постулата устойчивости, который утверждает, что предельное распределение должно быть таким, чтобы наибольшее значение в выборке объема n, взятой из этого распределения, имело то же самое предельное распределение.

Первое предельное распределение имеет вид

$$F(x) = \exp\left[-e^{-\alpha(x-u)}\right],$$

где α, u — положительные параметры, не зависящие от n.

Первое исходное распределение, устойчивое по отношению к наибольшему значению, совпадает с этим распределением.

Функция распределения наибольшего значения для первого исходного распределения образуется путем сдвига последнего по оси абсцисс на величину $\ln n/\alpha$.

Второе предельное распределение имеет вид:

$$F(x) = \exp\left[-(v/x)^k\right],$$
$$x \geqslant 0, \quad v > 0, \quad k > 0.$$

Второе исходное распределение, устойчивое по отношению к наибольшему значению, совпадает с этим распределением.

Функция распределения наибольшего значения для второго исходного распределения образуется из этого распределения путем изменения масштаба случайной величины (умножением на величину $n^{1/k}$).

Третье предельное распределение имеет вид

$$F(x) = \exp\left[-(x/v)^k\right],$$
$$x \leqslant 0, \quad v < 0, \quad k > 0.$$

Третье исходное распределение, устойчивое по отношению к наибольшему значению, совпадает с этим распределением.

Функция распределения наибольшего значения для третьего исходного распределения получается из этого распределения путем изменения масштаба случайной величины (умножения на величину $n^{-1/k}$).

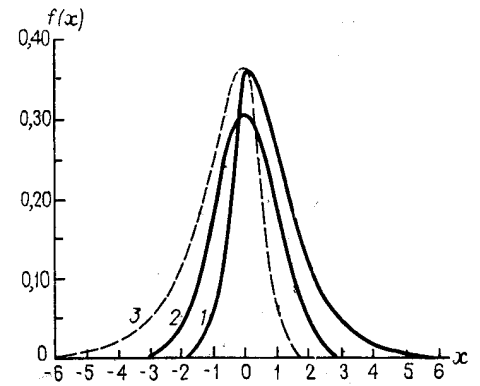
Проблема учета взаимозависимости наблюдений частично решается переходом от точных распределений экстремальных значений к предельным.

Первое исходное распределение (функцию распределения и плотность), устойчивое по отношению к наибольшему значению, можно записать так:

$$\Phi(x) = \exp\left(-e^{-y}\right), \qquad \varphi(x) = \alpha_n \exp\left(-y - e^{-y}\right),$$

где $y = \alpha_n (x - u_n)$, u_n — характеристическое наибольшее значение. Это распределение носит название двойного экспоненциального

Рис. 69. Плотности распределений наибольшего значения (*1*), нормального (*2*) и наименьшего (*3*)



распределения. Полагая $\lambda_n = e^{\alpha_n u_n}$, получают асимптотическую функцию распределения и плотность вероятности:

$$\Phi(x) = \exp\left(-\lambda_n e^{-\alpha_n x}\right),$$
$$\varphi(x) = \alpha_n \lambda_n \exp\left(-\alpha_n x - \lambda_n e^{-\alpha_n x}\right).$$

Двойное экспоненциальное распределение содержит два параметра: u_n и $1/\alpha_n$. Они зависят от исходного распределения и от объема выборки. Первый из них, u_n является характеристическим наибольшим значением и возрастает как логарифм n. Другой параметр, есть функция интенсивности.

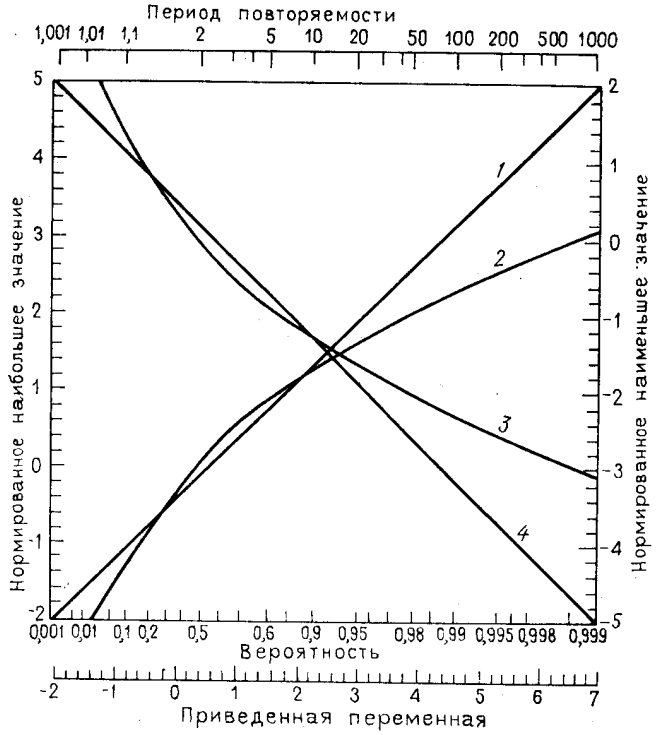


Рис. 70. Функции распределений наибольшего значения (*1*), нормального (*2, 3*) и наименьшего (*4*)

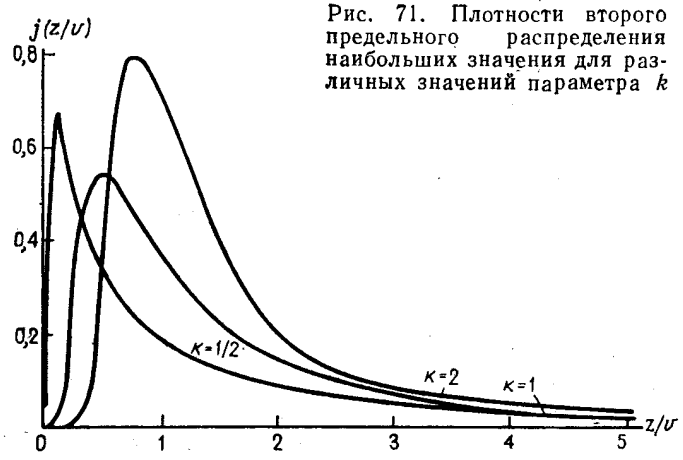


Рис. 71. Плотности второго предельного распределения наибольших значения для различных значений параметра k

На рис. 69, 70 представлены графики соответственно плотностей и функций распределения двух двойных экспоненциальных распределений и нормального распределения.

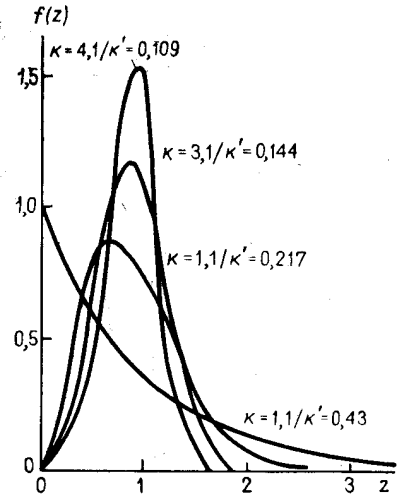


Рис. 72. Плотность третьего предельного распределения наименьших значений для различных значений параметра k

Стандартные отклонения наибольшего и наименьшего значений:

$$\sigma_n = \frac{\pi}{\sqrt{6}\,\alpha_n},$$

$$\sigma_1 = \frac{\pi}{\sqrt{6}\,\alpha_1},$$

а произведение параметров αu является функцией коэффициента вариации v, где γ — среднее:

$$\alpha_u = \frac{\pi}{\sqrt{6}\,v} - \gamma.$$

Плотности распределений второго предельного распределения наибольших значений приведены на рис. 71, а третьего предельного распределения наименьших значений — на рис. 72.

ГЛАВА 16

МАТЕМАТИЧЕСКАЯ ЛОГИКА

ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ ЛОГИКИ

Математическая логика — логика, развиваемая математическими методами. Характерным для математической логики является использование формальных языков с точным синтаксисом и четкой семантикой, однозначно определяющими понимание формул (А. А. Марков).

Основное различие между математической логикой и другими областями математики заключается в том, что она имеет дело не с количественными, а с качественными формами. Некоторые математики определяют математическую логику как науку, которая рассматривает главным образом «неколичественные отношения». Изучение их не позволяет непосредственно применить в математической логике принципы и приемы, разработанные в других областях математики, однако некоторые из этих приемов удается приспособить и к исследованию качественных форм. Так, еще в XIX в. английский математик и логик Джордж Буль, заметив аналогию между логическими и алгебраическими операциями, положил начало одному из главных разделов математической логики — алгебре логики, или булевой алгебре, в которой в логических преобразованиях используются методы алгебры. Помимо алгебры логики, другими важными разделами математической логики являются алгебра высказываний и логика предикатов. Основное практическое применение в настоящее время получила алгебра логики, которая составляет математический аппарат кибернетики и используется при проектировании и работе ЭВМ, различных автоматических устройств, в теории программирования и т. д.

В геологии, в математических, в частности логических, методах также главным образом используется аппарат алгебры логики и только в новом классе этих методов, методах анализа логических зависимостей, кроме того, применяется алгебра высказываний и логика предикатов.

Проникновение математической логики в геологические исследования связано с потребностью получать обоснованные выводы при анализе больших массивов качественной информации, представленной описаниями наблюдаемых геологических объектов (минерализованных тел, стратиграфических разрезов, разрывных нарушений и т. д.), а также различными картами и схемами. Наиболее распространенными геологическими задачами, при решении которых возникает необходимость в применении методов математической логики для обработки фактических, в том числе и качественных данных, являются следующие:

а) задача группировки геологических образований, например, выделение осадочных, рудных формаций, магматических комплексов, минеральных типов;

б) задача классификационного отнесения геологических образований к одной из заданных групп. Примером такой задачи является оценка возможных масштабов изучаемых рудопроявлений путем определения их принадлежности к классу промышленных или классу бесперспективных объектов;

в) задача установления связей между свойствами различных групп геологических образований с последующим прогнозированием этих свойств. Например, между зональностью оруденения и изменением свойств вмещающей среды, между размещением оруденения и геологическим строением региона, между соотношением полезных компонентов в рудах и особенностями геологических образований, слагающих участок с оруденением.

В случае количественных данных задачи этих трех типов решаются на основе методов математической статистики, в частности первая из них методами статистического разграничения геологических объектов [39], вторая — с помощью дискриминантного анализа [3], третья — корреляционного и регрессионного анализа [3]. Для качественной информации также разработан ряд методов решения этих задач, в частности эвристические методы автоматической классификации (методы таксономии) для задач первого типа и эвристические методы распознавания образов для задач второго типа. Иногда делались попытки применить методы распознавания образов и для решения задач третьего типа, например, для изучения связей между соотношением полезных компонентов в рудах и особенностями геологической обстановки, между минеральным составом руд и глубинностью их формирования, между вещественным составом рудных формаций и геологической обстановкой их локализации и т. д.

Аппарат математической логики получил распространение в основном при решении задач классификационного отнесения, а также задач выявления связей между свойствами геологических образований. Причем для решения последних создан принципиально новый класс методов, названных методами анализа логических зависимостей.

Что касается задачи группировки геологических образований, то в настоящее время при их решении аппарат математической логики удалось использовать только в качестве языка для формального описания и постановки этих задач.

АЛГЕБРА ВЫСКАЗЫВАНИЙ

Определение 1. Под *высказыванием* понимается имеющее смысл языковое выражение, относительно которого можно утверждать только то, что оно либо истинно, либо ложно. Поэтому каждому высказыванию можно приписать истинное значение И (истина) или Л (ложь).

Основная задача алгебры высказываний состоит в описании преобразований над высказываниями на основе определенных логических законов. Используя частицу «не», а также союзы «и», «или», «если, . . . то», «тогда и только тогда, когда» и т. п., можно из одних высказываний строить другие, новые высказывания. При этом исходные высказывания принято называть *простыми*, а вновь образованные — *сложными*. Эти названия не носят абсолютного характера, т. е. высказывание, которое в одной ситуации считается простым, в другой может рассматриваться как сложное, и наоборот [14, 27, 36].

Определение 2. Построение из простых высказываний сложного высказывания называется *логической операцией*.

Логические операции по существу выражают упомянутые выше связки, употребительные в обычной речи. В алгебре высказываний исследуется вопрос об истинности сложного высказывания, построенного с помощью заданных логических операций, в зависимости от истинности входящих в него простых высказываний. Основные операции алгебры высказываний выражаются таблицами.

Определение 3. *Конъюнкция* — логическая операция, соединяющая два или более высказываний при помощи союза «и» в сложное высказывание, которое истинно тогда и только тогда, когда каждое из простых высказываний истинно и ложно, когда по крайней мере одно из исходных высказываний ложно. Операция конъюнкции записывается знаками «$\wedge$», «$\cdot$» или он вообще опускается, т. е. пишут AB. Эта операция полностью определяется таблицей операции конъюнкции (табл. 10).

Определение 4. *Дизъюнкция* — операция, соответствующая неальтернативному (неисключающему) союзу «или». Эта операция записывается символом «$\vee$» и задается таблицей операции дизъюнкции (табл. 11). Операция дизъюнкции определяется следующим образом: высказывание $A \vee B$ истинно тогда и только тогда, когда истинно хотя бы одно из первоначальных высказываний A или B.

Определение 5. *Отрицание* $\overline{A}$ высказывания A задается табл. 12.

Таблица 10

Операция конъюнкции

A	B	$A \wedge B$
И	И	И
И	Л	Л
Л	И	Л
Л	Л	Л

Таблица 11

Операция дизъюнкции

A	B	$A \vee B$
И	И	И
И	Л	И
Л	И	И
Л	Л	Л

Таблица 12

Операция отрицания

A	$\overline{A}$
И	Л
Л	И

Таблица 13

Операция импликации

A	B	$A \to B$
И	И	И
И	Л	Л
Л	И	И
Л	Л	И

Таблица 14

Операция эквивалентности

A	B	$A \sim B$
И	И	И
Л	Л	Л
Л	И	Л
Л	Л	И

Эта операция, в отличие от конъюнкции и дизъюнкции, одноместна, т. е. новое высказывание $\overline{A}$ строится не из двух, а из одного простого высказывания.

Определение 6. *Импликация* — операция, соответствующая обороту «если . . . , то». Обозначается символом «→» и задается табл. 13. Сложное высказывание $A \to B$, полученное с помощью этой операции, считается ложным тогда и только тогда, когда A истинно, а B ложно. Высказывание A называется посылкой, а B заключением.

Определение 7. *Эквивалентность* — операция, соответствующая обороту типа «тогда и только тогда, когда». Обозначается символом «∼» и задается табл. 14. Операция эквивалентность определяется следующим образом: сложное высказывание $A \sim B$ истинно тогда и только тогда, когда A и B оба истинны или оба ложны.

Определение 8. Логическая операция называется *тождественно истинной* (*ложной*), если при любых истинных значениях простых высказываний сложное высказывание истинно (ложно).

Определение 9. Всякое сложное высказывание, составленное из некоторых исходных высказываний посредством применения логических операций 1—5, называется формулой алгебры высказываний.

Простые высказывания при этом могут принимать одно постоянное значение И (Л) или не иметь определенного значения И или Л. В первом случае простые высказывания называются *постоянными*, во втором — *переменными*. Переменные высказывания обозначаются большими латинскими буквами (например, $Y_1, Y_2, \ldots, Y_n$). Если задать значения всех переменных высказываний, составляющих произвольную формулу, то она примет определенное значение. Таким образом, каждой формулой реализуется некоторая функция, аргументами которой являются переменные простые высказывания. Как оказалось, любую сложную логическую операцию можно построить, комбинируя указанные выше пять основных операций. Более того, для этого достаточно только трех из них — дизъюнкции, конъюнкции и отрицания. Поэтому произвольная, не тождественно ложная функция алгебры высказываний $F(Y_1, \ldots, Y_n)$ может быть представлена формулой, содержащей только эти три операции.

Определение 10. Представление функции $F(Y_1, \ldots, Y_n)$ в виде формулы, являющейся дизъюнкцией (логической суммой) различных конъюнкций (логических произведений) $Y_1' \wedge \overline{Y}_i \wedge \wedge \ldots \wedge Y_i' \wedge \ldots \wedge Y_k'$, где Y_i' обозначает Y_i или $\overline{Y}_i$, называется дизъюнктивной нормальной формой.

АЛГЕБРА ЛОГИКИ

Алгебра логики [14, 48] представляет собой такой раздел математической логики, в котором логические операции над высказываниями интерпретируются как операции, действующие на множестве, состоящем из двух элементов, в частности на множестве {0, 1}. При этом с каждой из основных логических операций сопоставляется определенная двузначная функция.

Определение 1. Функция $f(z_1, \ldots, z_n)$, у которой аргументы принимают значения из множества {0, 1} и которая на любом наборе значений аргументов принимает значение из того же множества {0, 1}, называется функцией алгебры, логики, или булевой функцией. Из определения функции $f(z_1, \ldots, z_n)$ следует, что для ее задания достаточно указать, какие значения функции соответствуют каждому из наборов значений аргументов (табл. 15).

При числе аргументов n число различных наборов их значений равно 2^n. Множество таких наборов обозначают E^n. Эти наборы в таблицах для булевых функций записываются стандартным образом, т. е. каждый набор рассматривается как запись некоторого числа в двоичном исчислении и в таблице они располагаются в порядке возрастания этих чисел 0, 1, . . . , 2^{n-1}. При фиксированных переменных можно построить 2^{2^n} таблиц, различающихся правыми частями, т. е. число функций алгебры логики, зависящих от n переменных, равно 2^{2^n}.

Определение 2. Функция алгебры логики называется не всюду определенной или частичной, если

$$f(z_1, \ldots, z_n) = \begin{cases} 1 \text{ на } M_1^f \\ 0 \text{ на } M_0^f \\ \text{не определена на } E^n/M^f, \end{cases}$$

Таблица 15

Представление функции алгебры логики таблицей

$z_1\ z_2\ \ldots\ z_{n-1}\ z_n$	$f(z_1, z_2, \ldots, z_{n-1}, z_n)$
0 0 . . . 0 0	f (0, 0, . . . , 0, 0)
0 0 . . . 0 1	f (0, 0, . . . , 0, 1)
0 0 . . . 1 0	f (0, 0, . . . , 1, 0)
. . . . .	. . . . .
1 1 . . . 1 0	f (1, 1, . . . , 1, 0)
1 1 . . . 1 1	f (1, 1, . . . , 1, 1)

где M^f — множество, на котором задана функция $f(z_1, \ldots, z_n)$, $M^f \subseteq E^n$; M_1^f, M_0^f — подмножества множества M^f, на котором функция $f(z_1, \ldots z_n)$ принимает значения, соответственно, 1 или 0; (E^n/M^f) — множество наборов, являющееся дополнением к множеству M^f до E^n.

Определение 3. Булева функция $f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n)$ существенно зависит от аргумента z_i, если имеются такие значения $\alpha_i, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n$ переменных $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$, что $f(\alpha_1, \ldots, \alpha_{i-1}, 0, \alpha_{i+1}, \ldots, \alpha_n) \neq f(\alpha_1, \ldots, \alpha_{i-1}, 1, \alpha_{i+1}, \ldots, \alpha_n)$.

В этом случае переменная z_i называется существенной. Если z_i не является существенной переменной, то она называется фиктивной. Функции $f(z_1, \ldots, z_n)$ называются равными ($f_1 = f_2$), если функцию f_2 можно получить из f_1 путем добавления и изъятия фиктивных переменных. В алгебре логики функции рассматриваются с точностью до фиктивных переменных, т. е. считается, что если задана функция f_1, то задана и любая равная ей функция f_2.

Среди булевых функций выделяются так называемые элементарные функции, которые тесно связаны с основными логическими операциями в алгебре высказываний и играют такую же роль в алгебре логики, как, например, $\sin z$, z^n в математическом анализе.

Определение 4. Элементарными функциями в алгебре логики считаются следующие:

1) $f_1(z) = 0$ — константа 0;
2) $f_2(z) = 1$ — константа 1;
3) $f_3(z) = z$ — тождественная функция;
4) $f_4(z) = \bar{z}$ — отрицание z ($\bar{z}$ читается «не z»);
5) $f_5(z_1, z_2) = (z_1 \wedge z_2)$ — конъюнкция z_1 и z_2 (читается «z_1» и «z_2»). Вместо знака $\wedge$ употребляется знак «$\cdot$» или вообще знак отсутствует, т. е. пишут (z_1, z_2). Эту функцию часто называют логическим умножением;
6) $f_6(z_1, z_2) = (z_1 \vee z_2)$ — дизъюнкция z_1 и z_2 (читается «z_1 или z_2»). Эту функцию часто называют логическим сложением;
7) $f_7(z_1, z_2) = (z_1 \to z_2)$ — импликация z_1 и z_2 (читается «из z_1 следует z_2»). Эту функцию часто называют логическим следованием;
8) $f_8(z_1, z_2) = (z_1 + z_2)$ — сложение z_1 и z_2 по mod 2;
9) $f_9(z_1, z_2) = (z_1/z_2)$ — функция Шеффера.

Значения этих функций приведены в табл. 16 и 17.

Таблица 16

Значения одноместных элементарных булевых функций

z	0	1	z	$\bar{z}$
0	0	1	0	1
1	0	1	1	0

Как и в алгебре высказываний, в алгебре логики, исходя из элементарных функций, можно строить формулы, описывающие сложные булевые функции, в том числе и записать в виде формулы любую произвольную булеву функцию, заданную таблицей. В свою очередь каждой формуле алгебры логики соответствует некоторая функция, причем различным формулам могут соответствовать равные функции.

Таблица 17

Значения двуместных элементарных булевых функций

z_1	z_2	$(z_1 \wedge z_2)$	$(z_1 \vee z_2)$	$(z_1 \to z_2)$	$(z_1 + z_2)$	(z_1/z_2)
0	0	0	0	1	0	1
0	1	0	1	1	1	1
1	0	0	1	0	1	1
1	1	1	1	1	0	0

Определение 5. Формулы $\mathfrak{A}$ и $\mathfrak{B}$ называются эквивалентными, если соответствующие им функции равны.

На этом понятии основано осуществление эквивалентных преобразований алгебры логики, в частности преобразований, позволяющих упрощать формулы, реализующие некоторые булевы функции.

Для элементарных функций (главным образом множества $\{0, 1, z, (z_1 \wedge z_2), (z_1 \vee z_2)\}$) основные эквивалентности следующие:

а) функция $(z_1 \bigcirc z_2)$ обладает свойством ассоциативности $((z_1 \bigcirc z_2) \bigcirc z_3) = (z_1 \bigcirc (z_2 \bigcirc z_3))$,
где $(z_1 \bigcirc z_2)$ — любая из функций $(z_1 \wedge z_2)$, $(z_1 \vee z_2)$, $(z_1 + z_2)$;

б) функция $(z_1 \bigcirc z_2)$ обладает свойством коммутативности $(z_1 \bigcirc z_2) = (z_2 \bigcirc z_1)$;

в) для конъюнкции и дизъюнкности выполняются дистрибутивные законы: $((z_1 \vee z_2) \wedge z_3) = ((z_1 \wedge z_3) \vee (z_2 \wedge z_3))$; $((z_1 \wedge z_2) \vee z_3) = ((z_1 \vee z_3) \wedge (z_2 \vee z_3))$;

г) имеют место следующие соотношения между отрицанием, конъюнкцией и дизъюнкцией:
$\bar{\bar{z}} = z$; $\overline{(z_1 \wedge z_2)} = (\bar{z}_1 \vee \bar{z}_2)$; $\overline{(z_1 \vee z_2)} = (\bar{z}_1 \wedge \bar{z}_2)$.
Последние два тождества иногда называют правилами Моргана;

д) выполняются следующие свойства конъюнкции и дизъюнкции:

$(z \wedge z) = z$, $\quad (z \vee z) = z$,
$(z \wedge \bar{z}) = 0$, $\quad (z \vee \bar{z}) = 1$,
$(z \wedge 0) = 0$ $\quad (z \vee 0) = z$,
$(z \wedge 1) = z$ $\quad (z \vee 1) = 1$.

Определение 6. Пусть задан набор переменных $z_1, \ldots, z_n$. Выражение

$$K = z'_{i_1} \wedge \ldots \wedge z'_{i_\nu} \& \ldots \wedge z'_{i_r},$$

где $z'_{i\nu}$ обозначает z_{i_ν} или $\bar{z}_{i_\nu}$, $i_\nu \neq i_\mu$ при $\nu \neq \mu$ называется элементарной конъюнкцией, а число r — рангом элементарной конъюнкции.

Определение 7. Выражение

$$\mathfrak{N} = \bigvee_{j=1}^{s} K_j, \ (K_j \neq K_t \text{ при } j \neq t),$$

где K_j ($j = 1, \ldots, s$) — элементарная конъюнкция ранга r_j, называется дизъюнктивной нормальной формой (Д. Н. Ф.).

Одним из вопросов, который решается в алгебре логики, является следующий: если выделить некоторое подмножество элементарных функций, то всякая ли функция алгебры логики, заданная таблицей, может быть выражена в виде формулы. Ответ на этот вопрос будет положительным: любая функция алгебры логики может быть записана в виде формулы через отрицание, конъюнкцию и дизъюнкцию, в частности в виде Д. Н. Ф. Помимо дизъюнкции, конъюнкции и отрицания, такого рода свойством обладают и некоторые другие системы элементарных функций.

Так как произвольная булева функция представима в виде Д. Н. Ф. не единственным образом, то возникает вопрос о выборе более предпочтительной ее реализации. Для этого вводится индекс простоты $L(\mathfrak{N})$, характеризующий «сложность» Д. Н. Ф. $\mathfrak{N}$. Этот индекс может выражать сложность Д. Н. Ф. в разных аспектах и, следовательно, быть различным. В частности, таким:

а) $L_б(\mathfrak{N})$ — число букв переменных, встречающихся в записи Д. Н. Ф.;

б) $L_к(\mathfrak{N})$ — число элементарных конъюнкций, входящих в $\mathfrak{N}$;

в) $L_о(\mathfrak{N})$ — число символов отрицания, встречающихся в записи Д. Н. Ф. $\mathfrak{N}$.

Определение. 8. Д. Н. Ф. $\mathfrak{N}$, реализующая функцию $f(z_1, \ldots, z_n)$ и имеющая минимальный индекс $L(\mathfrak{N})$, называется минимальной относительно L.

Обычно собственно *минимальной* Д. Н. Ф. называют Д. Н. Ф., минимальную относительно индекса $L_б(\mathfrak{N})$, а *кратчайшей* — Д. Н. Ф., минимальную относительно индекса $L_к$.

Вопрос о том, как для произвольной функции алгебры логики построить минимальную Д. Н. Ф. относительно заданного индекса простоты, называется *проблемой минимизации булевых функций*. Способы нахождения таких Д. Н. Ф. называются алгоритмами построения минимальных Д. Н. Ф. Иногда также используется термин — алгоритмы минимизации булевых функций.

ЛОГИКА ПРЕДИКАТОВ

Логика предикатов [36] представляет собой дальнейшее развитие алгебры высказываний, при котором исследуются операции с высказываниями, отнесенные к предметам.

Пусть $\mathfrak{M}$ — некоторое множество предметов (объектов), а x — произвольный предмет из этого множества. Через $P(x)$ обозначим некоторое высказывание о предметах множества $\mathfrak{M}$. Например, если $\mathfrak{M}$ — натуральный ряд чисел, то в качестве $P(x)$ может рассматриваться высказывание «x есть четное число». Это неопределенное высказывание становится определенным, если *предметную переменную* x заменить *индивидуальным предметом* из множества $\mathfrak{M}$; например, «4 есть четное число», «7 есть четное число». Каждое такое определенное высказывание является истинным или ложным,

т. е. каждому предикату из $\mathfrak{M}$ ставится в соответствие один из двух символов И или Л. Таким образом, $P(x)$ представляет собой функцию, определенную на множестве $\mathfrak{M}$ и принимающую только два значения — И или Л. Аналогично этому неопределенные высказывания о двух и более предметах $H(x, y)$, $G(x, y, z)$ и т. д. представляют собой функцию двух, трех и т. д. переменных.

Эти неопределенные высказывания, или функции одной или нескольких переменных, называются логическими функциями, или *предикатами*. Предикат с одной переменной соответствует понятию предиката в классической логике Аристотеля и выражает свойство предметов. Так, например, «x — кварц», если в качестве $\mathfrak{M}$ рассматривается множество минералов гранитоидов. Предикатом с несколькими переменными выражается *отношение* между предметами; например, $S(x, y)$ может обозначать «x моложе y», если $\mathfrak{M}$ — множество разновозрастных пород, слагающих участок с оруденением. Множество $\mathfrak{M}$ с определенными на нем предикатами называется предметной областью. Среди предикатов различают постоянные или индивидуальные и переменные. Под *индивидуальным* предикатом понимается некоторое конкретное высказывание о предметах области $\mathfrak{M}$, как, например, в приведенных выше примерах. В качестве *переменного* предиката рассматривается произвольный предикат из некоторого (подходящего) множества постоянных предикатов.

ЛОГИЧЕСКИЕ МЕТОДЫ

Среди обширного круга математических методов, разработанных в настоящее время для обработки качественной геологической информации, находится набор методов, объединенных общим термином «логические» (логико-информационные, логико-дискретные, логико-комбинаторные, логико-математические, просто логические и т. д.). Если исключить из этого набора те методы, для которых термин «логические» использован как синоним термина «эвристические», то остальные методы по степени использования в них аппарата математической логики можно разделить на следующие три типа.

1. Эвристические методы, в которых понятия математической логики применяются только в качестве языка для формальной постановки задачи. В процедурах, из которых состоят эти методы, аппарат математической логики никак не используется.

Очевидно, что этот тип методов можно относить к логическим лишь с большой долей условности, так как при описании постановки задачи на ином формальном языке, они ничем не будут отличаться от других нелогических методов обработки качественной информации. Поэтому в данной главе эти методы рассматриваться не будут.

2. Эвристические методы, включающие некоторые математические приемы, для выполнения которых используется аппарат математической логики. При этом ряд других приемов, а также общие принципы построения методов не связаны с указанной областью математики.

3. Методы, которые основаны на строгих теоретических разработках математической логики.

К первому и второму типам относятся все логические методы решения задач классификационного отнесения, а к первому типу, кроме того,— один из разработанных в рудоформационном анализе методов группировки рудных объектов.

Так как логические методы второго из перечисленных типов входят в класс эвристических методов распознавания образов, которыми решаются задачи классификационного отнесения в случае качественных данных, то указанные методы называют также *логическими методами распознавания*.

К третьему типу принадлежат *методы анализа логических зависимостей*, разработанные для выявления связей между свойствами геологических образований, с последующим прогнозированием этих свойств.

Логические методы первых двух типов, в отличие от методов анализа логических зависимостей, называются иногда логическими методами сравнения. Введение этого термина обусловлено принципиальным различием в общем подходе к обработке качественных данных между указанными типами методов. Так, все логические (а также и нелогические) методы решения задач группировки и классификационного отнесения обладают одной общей особенностью — они обязательно включают процедуру сравнения отдельных геологических объектов по заданному набору описывающих их характеристик (признаков); в отличие от них, методы анализа логических зависимостей основаны на сопоставлении геологических признаков по многим объектам.

Ниже рассматриваются характерные особенности логических методов распознавания и методов анализа логических зависимостей, заключающиеся в определенной форме представления фактического материала, принципах выбора информативных комбинаций признаков и использовании понятия избыточности качественной информации.

Представление фактических данных для применения логических методов. При применении логических методов к решению геологических задач обязательной формой представления фактического материала является бинарная таблица, т. е. таблица, в которой содержатся два значения — 1 и 0. Исключение составляют лишь те логические методы, в частности метод непрерывных дизъюнктивных форм, в которых хотя и используется свойственный всем логическим методам распознавания критерий неизбыточности информации, однако оценка информативности признаков по этому критерию проводится без участия аппарата двузначной алгебры логики, оперирующей только значениями 1 и 0.

При представлении фактического геологического материала в виде бинарной таблицы объектам исследования сопоставляются строки таблицы, а признакам, описывающим эти объекты, ее столбцы. Наличие тех или иных признаков на данном объекте отмечается в соответствующих столбцах символом 1, их отсутствие — символом 0.

В зависимости от решаемой задачи формирование системы признаков может осуществляться с привлечением литературных данных, описаний объектов, содержащихся в геологических отчетах, а также геологических карт. Причем эта система может включать самые разнообразные сведения об изучаемых объектах (например, карбонатизация, дайки кислого состава и т. д. при характеристике участков с оруденением). Аналогично системам признаков от постановки геологических задач зависит и выбор объектов исследования, в качестве которых часто рассматриваются отдельные рудные тела, участки с оруденением, различные геологические структуры, площади карты заданных размеров и формы, включающие точки минерализации и т. д.

Система признаков, составленная для характеристики выделенных элементарных площадей карты, помимо ее легенды, может включать и некоторые признаки, сформированные на основании содержания самой карты.

Иногда при применении логических методов возникает вопрос об учете не только качественной, но и количественной информации. В таких случаях количественные значения рассматриваемого параметра (например, протяженность разломов, их простирание, густота трещиноватости и т. д.) тем или иным способом разбиваются на отдельные градации, и каждая градация затем рассматривается как самостоятельный качественный признак.

Выбор информативных комбинаций признаков. Одной из основных процедур, которая выполняется при решении логическими методами задач классификационного отнесения и задач выявления связей между свойствами геологических образований, является выбор информативных комбинаций признаков. Для этих двух задач, одна из которых решается логическими методами распознавания, а другая методами анализа логических зависимостей, информативность признаков, точнее комбинаций признаков, рассматривается в разных аспектах. Так, в логических методах распознавания обязательным критерием информативности является эффективность данной комбинации признаков для разделения объектов из разных классов эталонной выборки. Причем мера, по которой следует оценивать эту эффективность, задается исследователем произвольно. Другим обязательным критерием информативности в логических методах распознавания, в отличие от нелогических методов этого типа, является критерий неизбыточности данной комбинации признаков относительно ее свойства разделять разные классы эталонной выборки (см. «Метод тупиковых тестов»). Кроме указанных двух критериев в логических методах распознавания при выборе информативных комбинаций признаков могут использоваться и другие критерии, однако их применение не является характерной чертой этих методов.

В методах анализа логических зависимостей информативность комбинаций признаков также определяется по нескольким крите-

риям. Одним из них является существование между признаками рассматриваемой комбинации заданной логической зависимости. Причем задание этой зависимости выполняется на основе логико-математического моделирования геологических понятий.

Другим критерием служит минимальность данной комбинации признаков относительно связывающей последние логической зависимости. Очевидно, что этот критерий является более жесткой модификацией используемого в логических методах распознавания критерия неизбыточности информации (см. «Устранение избыточности в качественной информации»).

Как в логических методах распознавания, так и в методах анализа логических зависимостей выбор информативных комбинаций признаков, иногда с последующим переходом к оценке информативности отдельных признаков, осуществляется на основе специальных алгоритмов анализа бинарных таблиц с фактическим материалом (см. «Логические методы распознавания» и «Методы анализа логических зависимостей»).

Устранение избыточности в качественной информации. В логических методах исследования геологических объектов широко используется общий прием алгебры логики, позволяющий устранять избыточность в качественной информации. Этот прием применяется при выборе информативных комбинаций признаков, где в зависимости от общего подхода избыточность информации может рассматриваться в двух аспектах: во-первых, относительно эффективности данного сочетания признаков для разделения объектов разных классов эталонной выборки (в логических методах распознавания) и, во-вторых, относительно выявленной логической связи между признаками (в методах анализа логических зависимостей). Независимо от аспекта, в котором рассматривается избыточность информации, для ее устранения существует общий прием алгебры логики [48], первоначально разработанный для построения неизбыточного набора входных сигналов при диагностике неисправностей в вычислительных устройствах. Этот прием применим во всех случаях, когда имеется m объектов (в том числе и математических), описанных n характеристиками, каждому из которых можно сопоставить дизъюнкцию каких-либо из n характеристик, а всем m объектам — конъюнкцию этих m дизъюнкций, т. е. построить выражение $\Pi\Sigma$ (произведение логических сумм):

$$\Pi\Sigma = \bigwedge_{i=1}^{m} (e_{i1} \vee \ldots \vee e_{ik_i}),$$

где k_i — число выделенных для i-го объекта характеристик. При этом в зависимости от исследуемых характеристик изменяются только элементы в скобках. Например, в логических методах распознавания ими являются номера признаков, которыми геологический объект одного класса отличается от объекта других классов; в методах анализа логических зависимостей — номера признаков, с которыми исследуемое свойство имеет заданную логическую

зависимость и которые для выделенного подмножества объектов принимают какое-то одно из двух значений, 1 или 0.

Устранение избыточной информации достигается путем эквивалентных преобразований (см. «Алгебра логики», определение 5) произведения логических слагаемых $\Pi\Sigma$ в сумму логических произведений $\Sigma\Pi$ с поглощением одинаковых и дублирующих членов по формулам: $e \vee e = e$; $e \wedge e = e$; $e \wedge \tilde{e} \vee e = e$.

Это преобразование часто обозначается символом $\wedge\vee\to\vee\wedge$ и используется для построения минимальных Д. Н. Ф. функций алгебры логики (см. «Алгебра логики, определения 7, 8»).

В полученном с помощью преобразования $\wedge\vee\to\vee\wedge$ выражении $\Sigma\Pi$ каждая конъюнкция соответствует неизбыточному набору исследуемых в данной задаче характеристик: например, в логических методах распознавания набору признаков, удовлетворяющих понятию тупикового теста или какой-либо из его модификаций (см. «Метод тупиковых тестов»), в методах анализа логических зависимостей,— комбинациям существенных аргументов функций алгебры логики (см. «Алгебра логики», определение 3).

Так как среди последних находятся и комбинации аргументов минимальной длины, то на основе этого преобразования в методах анализа логических зависимостей не только устраняется избыточная информация, но также минимизируется оставшаяся, т. е. выбирается оптимальное решение.

ЛОГИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ

Разработка логических методов распознавания тесно связана с развитием теории распознавания образов, в сущности, представляющей собой определенный подход к решению задач классификационного отнесения при изучении технических и природных объектов. В наиболее общем виде этот подход, получивший название «распознавание с учителем», состоит в следующем.

В рассматриваемом множестве объектов выделено подмножество эталонных объектов, составляющих один или более классов по исследуемому свойству («целевому признаку») и описанных какими-либо количественными или качественными характеристиками (признаками). Требуется для любого другого объекта множества на основании сходства его описания с описаниями эталонной выборки решить вопрос о принадлежности к одному из заданных классов.

С момента своего возникновения разработка методов распознавания образов шла по двум направлениям: с одной стороны, различающихся типом используемых математических моделей изучаемых объектов — вероятностных и детерминированных, с другой — характером обрабатываемых данных — количественных и, в основном, описательных, качественных. Вероятностный подход обычно применялся для решения задачи распознавания в случае описания геологических объектов количественными характеристиками. На его основе для данных, отвечающих определенным требованиям, разработаны методы дискриминантного анализа [3], по-

зволяющие получать результат с минимальным риском принятия ошибочных решений. В случае описательной информации при малом числе эталонных объектов, а также количественных данных, которые не отвечают требованиям, предъявляемым методами математической статистики, решение задач распознавания проводилось так называемыми детерминированными методами. В настоящее время разработано много этих методов, их подробный обзор имеется в соответствующей литературе [38].

Появление такого разнообразия детерминированных методов распознавания связано главным образом с неопределенностью понятия различия рассматриваемых объектов, что позволяет в качестве меры различия описаний объектов из разных классов эталонной выборки выдвигать любую, в той или иной степени основанную на специфике рассматриваемой задачи.

Кроме того, решающее правило для отнесения объектов с неизвестной принадлежностью к одному из эталонных классов (правило принятия решения) построенное с использованием заданной меры различия описаний, также выбирается произвольно. Эти два обстоятельства и делают детерминированные методы распознавания эвристическими. Обоснованием эффективности выбранного критерия различия описаний объектов и правила принятия решения обычно служит та или иная степень совпадения группировки эталонных и контрольных объектов, полученной на основе процедур распознавания, с заданной разбивкой этих объектов на классы по целевому признаку. Однако в условиях незначительного числа объектов, для которых, собственно, и разработано большинство детерминированных методов, подобная проверка не всегда приводит к обоснованным выводам.

Логические методы распознавания возникли как ответвление детерминированных методов распознавания образов. Поэтому общие принципы построения последних используются и в логических методах распознавания. Эти принципы, определяющие выполнение отдельных этапов решения задач классификационного отнесения, следующие.

На первом этапе составляется таблица описаний эталонных объектов, относящихся к одному или чаще нескольким классам по исследуемому свойству. Строки этой таблицы содержат описания m объектов в заданной системе из k геологических характеристик (признаков). Столбцы состоят из значений таких признаков. В одном типе таблиц, так называемых бинарных, эти значения во всех столбцах таблицы представлены единицей или нулем, обозначающими соответственно факт наличия или отсутствия признака на конкретном объекте, в другом типе они могут являться количественными или полуколичественными (балльными) характеристиками. Кроме значений признаков, описывающих эталонные объекты, иногда в таблицу включается $n + 1$ «целевой» признак, характеризующий исследуемое свойство.

Второй этап решения задач классификационного отнесения обычно состоит в оценке эффективности («информативности») отдельных признаков или сочетаний признаков с точки зрения возможности различить по ним эталонные объекты, относящиеся по целевому признаку к разным классам. Причем для такой оценки в разных методах используются различные критерии и разные меры, хотя и построенные по одному общему принципу, т. е. информативны те признаки (сочетания признаков), по которым в заданном смысле и с заданной точностью можно различить эталонные объекты из разных классов.

Третий этап заключается в построении решающего правила для определения принадлежности объектов с неизвестным значением целевого признака к одному из классов эталонных объектов. Для этого обычно используются три вида решающих правил: основанных на введении линейной разделяющей функции; использующих процедуру голосования; основанных на вычислении меры сходства распознаваемых объектов с обобщенными характеристиками заданных классов.

В последнее время известную популярность приобрели методы распознавания, использующие принцип «переобучения» или адаптации. Особенностью этих методов, названных методами вычисления оценок, является слияние второго и третьего этапов решения задачи распознавания за счет использования итерационных процедур.

Многие методы распознавания, помимо изложенных трех этапов, включают этап нахождения информативной системы признаков, заключающийся в удалении из исходной системы тех признаков, учет которых не улучшает или даже снижает качество распознавания. В большинстве методов этот этап практически неотделим от этапа принятия решения и обычно имеет место в тех методах, в которых отнесение распознаваемых объектов к определенному классу выполняется на основе вычисления значений некоторой разделяющей функции.

В методах, в которых оценка информативности признаков основана не только на критерии различия объектов из разных классов эталонной выборки, но и на комбинации нескольких критериев, что, в частности, свойственно логическим методам распознавания, этап выбора информативной системы может осуществляться непосредственно после составления таблицы описаний эталонных объектов.

Логические методы распознавания отличаются от остальных детерминированных методов этого класса только используемыми критериями информативности признаков и, следовательно, выполнением второго этапа решения задач классификационного отнесения. Эти критерии базируются на понятии тупикового теста, впервые введенного в практику геологических исследований А. Н. Дмитриевым, Ю. И. Журавлевым, Ф. П. Кренделевым в 1966 г. В качестве тупикового теста (см. «Метод тупиковых тестов») рассматривается неизбыточное сочетание признаков, которое позволяет различить объекты из разных классов эталонной выборки. С помощью понятия тупикового теста для оценки информативности признаков

вводятся два критерия. Один из них — критерий различия значений признаков (в данном случае составных признаков, образованных их сочетаниями) для объектов разных классов эталонной выборки является общим для всех детерминированных методов распознавания образов, и мера информативности признаков с точки зрения этого критерия сходна с мерой, используемой в известных методах «Кора-2», «Кора-3», которые не являются логическими. Другой критерий — неизбыточность, «тупиковость» сочетаний признаков относительно возможности различить по ним объекты разных классов эталонной выборки является исключительно особенностью методов, использующих понятие тупикового теста.

С этим критерием связано и основное отличие логических методов распознавания от других детерминированных методов распознавания образов, которое заключается в том, что для выбора информативных сочетаний признаков, удовлетворяющих понятию тупикового теста, можно использовать аппарат алгебры логики. Однако последнее не является обязательным, и в большинстве алгоритмов поиска тупиковых тестов эти сочетания устанавливаются обычным перебором столбцов таблицы исходных данных и сравнением строк в выделенных сочетаниях столбцов.

Другой причиной включения рассматриваемых методов в логические, видимо, послужили обстоятельства возникновения самого понятия тупикового теста. Это понятие было введено в технической кибернетике в связи с решением задачи построения оптимального диагностического теста, т. е. такого неизбыточного списка входных наборов, с помощью которого можно было бы различить все неисправности заданного класса для некоторого типа вычислительных устройств. Задача построения этого, оптимального в смысле неизбыточности, диагностического теста ставилась как задача минимизации некоторых функций алгебры логики (см. «Алгебра логика», определения 7, 8) и для ее решения использовался соответствующий математический аппарат (см. «Устранение избыточности качественной информации»).

В процессе усовершенствования первого, основанного на понятии тупикового теста логического метода распознавания, были созданы различные модификации этого метода, а также практически новые методы.

С созданием указанных модификаций и методов связано возникновение термина тестовый подход. Этим термином обозначается определенный способ оценки информативности признаков в методах распознавания образов, в основе которого лежит понятие тупикового теста, вернее, критерий неизбыточности информации, входящий в понятие тупикового теста. Так как этот критерий применяется к сочетаниям столбцов (сочетаниям признаков) таблицы исходных данных, обладающих каким-либо заранее заданным исследователем свойством, то в зависимости от определения этого свойства могут быть формально построены самые различные модификации понятия тупиковых тестов: Q-тесты, H-тесты, тупиковые пакеты и др.

Например, в собственно тупиковых тестах это свойство задается требованием, чтобы для объектов из разных классов эталонной выборки в рассматриваемом сочетании столбцов все строки были различны; в Q-тестах, чтобы в пределах одного класса они были одинаковы; в тупиковых пакетах, чтобы в пределах одного класса в анализируемых сочетаниях столбцов для каждого объекта содержалось хотя бы одно значение «1» и т. д.

При использовании различных модификаций понятия тупикового теста возникали различные модификации метода тупиковых тестов. Помимо таких модификаций, были разработаны также и другие, полученные видоизменением некоторых процедур, свойственных всем методам распознавания образов, при неизмененном понятии тупикового теста. При дополнительном введении новых формальных критериев оценки информативности признаков возникали новые методы, в частности метод вариационных рядов, метод Т-свойств, метод непрерывных дизъюнктивных форм.

Таким образом, во всех методах распознавания, применяющих тестовый подход, при обязательном использовании критерия неизбыточности все другие критерии информативности признаков, а также способы выполнения отдельных процедур методов распознавания образов могут быть самыми различными.

Перечисленные три новых метода отражают определенный путь развития логических методов распознавания, которые усовершенствовались как за счет введения новых формальных процедур и критериев, так и за счет более полного учета специфики геологических объектов (см. «Метод вариационных рядов»).

В настоящее время с помощью логических методов распознавания решается довольно широкий круг геологических задач. Большинство результатов получено для задач оценки возможных масштабов проявлений различных полезных ископаемых и оценки перспективных площадей или каких-либо геологических структур (тектонических узлов, купольных структур, локальных поднятий и т. д.) на обнаружение новых объектов с изучаемым видом полезного ископаемого.

Значительно реже логические методы распознавания образов применялись при решении задач, основной целью которых являлось не классификационное отнесение, а выявление геологических условий, влиявших на возникновение или изменение исследуемого свойства геологических объектов (особенностей геологической обстановки, определявших соотношение в рудах изучаемых полезных компонентов, условий формирования различных гранитоидных комплексов, критериев глубинности формирования руд). Результаты решения указанных двух типов задач для различных геологических объектов подробно освещены в литературе.

Ниже приводятся четыре логических метода распознавания: метод тупиковых тестов; метод вариационных рядов; метод Т-свойств; метод непрерывных дизъюнктивных нормальных форм (Н. Д. Н. Ф.).

Метод тупиковых тестов [17]. Термином «метод тупиковых тестов» обозначается первый разработанный в 1966 г. А. Н. Дмитриевым, Ю. И. Журавлевым, Ф. П. Кренделевым логический метод распознавания, использующий для оценки информативности признаков понятие тупикового теста. Иногда этот же термин применяется и к последующим модификациям указанного метода.

Понятие «тупиковый тест» введено для бинарных таблиц, в виде которых представляются массивы качественных данных. Оно определяется следующим образом.

Пусть задана эталонная выборка из m объектов, разбитая на l непересекающихся классов $k_1, \ldots, k_l$, и пусть эта выборка описывается бинарной таблицей Т, m строк которой представляют собой характеристики объектов со значениями n признаков. Тестом таблицы называется сочетание столбцов $t_{j_1}, \ldots, t_{j_S}$, образующее такую подтаблицу, в которой любая пара строк, описывающая объекты из разных классов, различна. Тупиковым тестом называется такой тест, при удалении из которого любого столбца, описания по крайней мере одной пары объектов из разных классов становятся неразличимы. Иногда термин «тупиковый тест» используется только для случая, когда $l = m$; в случае же когда классы эталонных объектов представлены не одним, а несколькими объектами, употребляется термин «тупиковый тестор». Например, в приведенной табл. 18 с параметрами $l = 2$, $m = 5$, $n = 4$ понятию тупикового теста соответствуют два сочетания столбцов — одно с номерами 1, 2, 3, другое с номерами 1, 3, 4. Все четыре столбца таблицы составляют тест, а любое сочетание этих столбцов по два не является ни тупиковым тестом, ни просто тестом.

Каждый столбец бинарных таблиц, для которых определяется понятие тупикового теста, описывает некоторый геологический признак. Поэтому сочетанию столбцов, являющемуся тупиковым тестом, соответствует неизбыточное сочетание геологических признаков, которыми различаются анализируемые классы эталонных объектов.

Таблица 18

Пример тупиковых тестов

Классы	Объекты	Номера признаков			
		1	2	3	4
I	1	1	0	1	1
	2	1	1	0	0
II	3	0	0	1	1
	4	1	1	1	0
	5	1	1	1	0

Первый вариант метода тупиковых тестов предназначался для обработки эталонной выборки, представленной только одним классом (например, классов крупных месторождений). Он состоял в том, что для таблицы исходных данных Т, описывающей эту выборку, находились тупиковые тесты и вычислялись оценки информативности отдельных признаков по формуле $p_j = n_j/N$, где n_j — число тупиковых тестов, в которые вошел столбец значений j-го признака, N — общее число тупиковых тестов. По этим оценкам строилась линейная функция. В качестве значений этой функции для каждого эталонного объекта принималась сумма оценок (весов) тех признаков, которые имели значение «1» в строке, описывающей данный объект. Аналогично вычислялись значения этой функции для распознаваемых объектов с неизвестным значением целевого признака (например, масштабами оруденения). Найденные значения функции (веса объектов), рассматривались в качестве меры принадлежности объекта к заданному классу крупных месторождений. В настоящее время разработан ряд модификаций этого метода. Так, Ю. И. Журавлевым было предложено заменить процедуру принятия решения на основе построения линейной функции процедурой голосования. В результате возникли две модификации метода: «голосование по тупиковым тестам» и «голосование по тупиковым тесторам». Как в первой, так и во второй модификации в качестве меры сходства распознаваемого объекта с объектами некоторого класса эталонной выборки рассматривалась функция.

$$k_t = \max_i \{k_t\},\ i = 1, \ldots, m,$$

или

$$\delta = \sum_{i=1}^{m} k_t/k,$$

где k_t — число тупиковых тестов (тесторов) таблицы Т, для которых описание распознаваемого объекта совпадает с описанием i-го эталона; k — общее число тупиковых тестов (тесторов) таблицы Т. При этом применение модификации «голосование по тупиковым тесторам» было более предпочтительным, чем использование модификации «голосование по тупиковым тестам», так как оно предполагало задание нескольких классов эталонных объектов, и выбор информативных сочетаний признаков проводился относительно различия описаний объектов из разных, а не из одного и того же класса.

Несколько иная модификация метода тестов была предложена Е. А. Смертиным. Она основывалась на понятии Q-тестов, т. е. таких неизбыточных сочетаний признаков, наборы значений которых попарно неразличимы для объектов данного класса. При этом критерием неразличимости наборов считалось их совпадение хотя бы по одному признаку. Поиск Q-тестов осуществлялся на основе тех же алгоритмов, что и поиск тупиковых тестов. Оценка информативности признаков производилась, как и в первом варианте метода

тестов, в виде отношения числа тупиковых Q-тестов, в которые входит j-й признак к общему числу их в таблице, и по полученным оценкам аналогично строилась линейная разделяющая функция. Применение Q-тестов в случае задания одного класса эталонных объектов было более обоснованно, чем использование первоначального метода тупиковых тестов.

Другой попыткой усовершенствовать первый вариант метода тупиковых тестов явилось предложение А. П. Мацака проводить процедуру голосования не по всем тупиковым тестам (тесторам) таблицы, а только тем из них, которые имеют небольшую длину (меньше некоторого заданного порога). Такие тупиковые тесты (тесторы) с большим основанием можно было считать связанными с реальными свойствами эталонных объектов, а не случайными, появившимися в таблице из-за статистической непредставительности эталонной выборки.

На основе этого предложения А. Н. Дмитриевым в оценку информативности отдельных признаков были введены числовые характеристики, учитывающие длину тупиковых тестов (тесторов), в которые входит данный признак.

Помимо охарактеризованных модификаций метода тупиковых тестов, имеется и ряд других, менее известных. Эти модификации достаточно полно освещены в литературе. Основное применение метод тупиковых тестов и его модификации нашли при оценке масштабов нефтяных и рудных проявлений.

Метод вариационных рядов [41]. В основе метода вариационных рядов лежит введение некоторых критериев и, соответственно, мер информативности признаков, учитывающих специфику геологических объектов. Необходимость в таком учете обусловлена тем, что принципы построения первых, использующих тестовый подход методов, были заимствованы из методов распознавания образов, предназначенных для диагностики технических объектов, которые существенно отличаются от геологических.

Так, в технической кибернетике, с развитием которой связано создание методов распознавания образов, процедуре распознавания подлежат сами объекты, а не их свойства, как при геологических исследованиях (например, определяется принадлежность принятого сигнала к одной из двух групп эталонных сигналов с известными источниками). Поэтому в качестве признаков анализируются сведения, заведомо характеризующие объект распознавания (например, частота и длительность сигналов), относительно которых стоит вопрос только о том, насколько важно их все учитывать при распознавании или некоторые из них можно удалить как избыточные.

При геологических исследованиях, как правило, неясно, как «выглядит» объект распознавания, какими характеристиками следует его описывать. Поэтому в качестве описаний эталонных объектов приводится не характеристика распознаваемого свойства (например, масштаба оруденения), а некоторого сложного объекта, обладающего этим свойством (участка с оруденением). При использовании не только логических, но и большинства других детерминированных методов распознавания образов всегда возникает сомнение в том, что если в целом по всем признакам эталонные объекты сходны (например, сходно геологическое строение участков с оруденением), то они будут сходны и по распознаваемому свойству (масштабу оруденения).

Однако, как показали многочисленные опыты по решению геологических задач, в самом обычном случае 70—80 % информации, содержащейся в описаниях геологических объектов, не имеет отношения к исследуемому свойству, и ее удаление существенно улучшает качество распознавания. Поэтому при исследовании геологических объектов методами распознавания образов возникает самостоятельная задача селекции информационных шумов, которая хотя и сводится к задаче выбора информативного набора признаков, однако не может быть эффективно решена на основе оценки информативности признаков относительно критерия различия описаний объектов из разных классов эталонной выборки. Это связано с тем, что при выборке комбинаций информативных признаков по такому критерию в большинстве методов распознавания значение меры информативности отдельных признаков вычисляется с учетом остальных признаков. Исключение составляют лишь меры информативности, основанные на вычислении частости появления значений отдельных признаков; однако их применение обоснованно только в случае большого объема эталонной выборки, т. е. когда при определенных условиях целесообразнее вместо детерминированных методов распознавания использовать строгие статистические методы.

Помимо рассмотренной особенности геологических объектов, другой их отличительной чертой является то, что для них сведения об отсутствии некоторых признаков могут нести не меньше важной информации, чем факты их наличия.

Эти две особенности геологических объектов были учтены при разработке метода вариационных рядов, который, хотя и является логическим методом распознавания, но уже содержит в себе некоторые элементы методов анализа логических зависимостей. Так, в методе вариационных рядов в качестве основного критерия информативности выдвигается не критерий различия описаний эталонных объектов, а критерий связи анализируемых геологических признаков с исследуемым свойством («целевым признаком»). При этом понятие связи вводится на основе представления о том, что если направленному изменению некоторого геологического явления сопутствует такое же направленное изменение каких-либо иных геологических характеристик, то это свидетельствует о взаимосвязи изучаемого явления с данными характеристиками. Такое представление, например, использовалось Р. М. Константиновым в рудно-информационном анализе в качестве способа для нахождения взаимосвязи между особенностями минерального состава рассматриваемой группы месторождений и геологическими условиями их нахождения.

На основе введенного понятия связи устанавливаются и границы применимости метода вариационных рядов, который разработан для случая, когда известна упорядоченность классов эталонных объектов по интенсивности проявления исследуемого свойства (например, по окраске минералов, масштабов оруденения), т. е. по возрастанию (убыванию) значений целевого признака при неизвестной величине этих значений. При выборке информативной комбинации признаков на основе критерия их связи с целевым признаком из объектов эталонной выборки составляется по крайней мере два вариационных ряда, представляющих собой таблицы, в которых объекты упорядочены в соответствии с возрастанием (убыванием) значений целевого признака и каждый класс представлен одним объектом. Оценка информативности признаков, значения которых (единица или нуль) содержатся в столбцах вариационных рядов, проводится путем сравнения последних. Мерой информативности признаков служит некоторая функция от биноминальных коэффициентов, составляющих смещенный треугольник Паскаля, значения которой находятся во взаимно однозначном соответствии со степенью упорядоченности (монотонности) наборов значений признаков. Указанная функция принимает максимальные значения тогда, когда для любой пары значений j-го признака из соседних строк с номерами i, $i+1$ выполняется одно и то же неравенство:

$$\alpha_{ij} \leqslant \alpha_{i+1,\,j} (\alpha_{ij} \geqslant \alpha_{i+1,\,j}).$$

Введение такой меры информативности признаков позволяет учесть и другую особенность геологических объектов, ценность сведений об отсутствии на объекте некоторых признаков, чем также устраняется и зависимость конечного результата от способа кодирования исходного материала. Это достигается путем вычисления двух оценок степени упорядоченности признаков как относительно «1» (P^* (1)), так и относительно «0» (P^* (0)). Соотношение величин двух оценок показывает, какое из двух значений данного признака сопутствует возрастанию значений целевого признака, т. е. отражает положительную или отрицательную связь между этими признаками. То же соотношение используется при выборе информативных признаков. При этом признак считается информативным, если во всех сравниваемых вариационных рядах знак неравенства

$$\overline{P}_j^*(1) > \overline{P}_j^*(0) \quad \text{или} \quad \overline{P}_j^*(1) < \overline{P}_j^*(0)$$

не меняется.

После выбора информативных признаков вариационные ряды объединяются в одну таблицу, в которой каждый класс представлен уже несколькими эталонными объектами, охарактеризованными только информативными признаками. Для этой таблицы находятся все тупиковые тесты и с их помощью по той же формуле, что и в методе тупиковых тестов, оценивается информативность признаков (p_j), $j = 1, \ldots, n$, относительно критерия различия описаний эталонных объектов из разных классов. Кроме того, для каждого признака вычисляются в процентах обобщенные оценки $P_j^*(1)$, $P_j^*(0)$

по модифицированному смещенному треугольнику Паскаля, которые обладают тем свойством, что в сумме всегда составляют 100 %. На основании этих двух характеристик далее вычисляются комбинированные оценки информативности признаков:

$$\hat{P}_j(1) = \overline{P}_j^*(1) \cdot p_j, \qquad \hat{P}_j(0) = \overline{P}_j^*(0) \cdot p_j,$$

которые преобразовываются к виду:

$$R_j = \left| \hat{P}_j(1) - \hat{P}_j(0) \right|.$$

В зависимости от того, какое из двух неравенств, $\hat{P}_j(1) > \hat{P}_j(0)$ или $\hat{P}_j(1) < \hat{P}_j(0)$, выполняется в качестве оценки информативности значений «1» и «0» j-го признака, в первом случае рассматриваются величины соответственно $R_j(1) = R_j$ или $R_j(0) = 0$; во втором оценками значений «1» и «0» служат величины $R_j(1) = 0$ и $R_j(0) = R_j$. Значения разделяющей функции для эталонной выборки определяются по формуле:

$$I_i = \sum_{j=1}^{n} R_j(1,\ 0),\ i = 1, \ldots, M,$$

где i — номер эталонного объекта; M — число эталонных объектов в объединенной таблице; R_j (1, 0), либо R_j (1), либо R_j (0) — в зависимости от того, какое значение (1 или 0) принимает j-й признак для i-го объекта.

Решение вопроса о принадлежности распознаваемого объекта к одному из классов эталонной выборки, так же как и во многих других методах распознавания образов, производится на основании сравнения значения линейной разделяющей функции, вычисленной для распознаваемого объекта со значениями этой функции для эталонных объектов. Распознаваемый объект относится к тому же классу, к какому принадлежит эталонный объект, для которого разность между сравниваемыми значениями разделяющей функции минимальна.

Как видно из принципов построения метода вариационных рядов, одним из условий его применения является одинаковое число объектов (не менее двух) в заданных классах эталонной выборки. Это условие иногда создает определенные осложнения, так как не позволяет включить в эталонную выборку часть хорошо изученных объектов, если они принадлежат одному классу.

Чтобы учесть всю имеющуюся информацию по эталонным объектам, И. А. Чижовой разработана модификация метода вариационных рядов. Она предусматривает перебор всех эталонных выборок с одинаковым числом объектов в классах, построенных по заданной эталонной выборке с различным их числом. При этом для каждой просматриваемой системы эталонных объектов по смещенному треугольнику Паскаля с помощью той же процедуры, что и в основном методе вариационных рядов, производится выбор информативной комбинации признаков. Для разных систем эталонных объектов эти комбинации могут незначительно различаться, и поэтому про-

водится дополнительная оценка выявленных информативных признаков с точки зрения их «устойчивости». Мерой устойчивости служит отношение:

$$\mu_j = Q_j/Q, \qquad j = 1, \ldots, q,$$

где q — общее число информативных признаков; Q_j — число систем эталонных объектов, в которых j-й признак является информативным; Q — общее число просматриваемых систем эталонных объектов.

На основе вычисленных значений меры устойчивости оценивается качество перебираемых систем. Оптимальной считается такая система эталонных объектов, для которой среднее арифметическое из значений меры устойчивости по выявленным для нее информативным признакам максимально. Эта система затем анализируется по методу вариационных рядов, и относительно нее определяется принадлежность распознаваемых объектов к заданным классам.

Процедура вычисления мер устойчивости, помимо повышения надежности распознавания, позволяет сделать некоторые дополнительные содержательные выводы о связи рассматриваемых признаков с исследуемым свойством.

Метод вариационных рядов, а также его модификация использовались для решения следующих типов задач:

а) оценки возможных масштабов рудопроявлений;

б) исследования особенностей геологической обстановки, влиявших на соотношение в рудах полезных компонентов;

в) нахождения оптимального технологического режима для синтеза минералов с заданными свойствами.

Метод Т-свойств [34] является одним из логических методов распознавания, которые в значительно большей степени, чем все другие методы этого типа, построен на аппарате алгебры логики и неразрывно с ним связан. В методе Т-свойств, разработанном В. О. Красавчиковым, информативность признаков оценивается по трем критериям, два из которых базируются на понятиях алгебры логики. Один из них — критерий общности объектов данного класса эталонной выборки — вводится с помощью определения табличного свойства (Т-свойства), т. е. булевой функции заданного типа (обычно дизъюнкции; см. «Алгебра логики», определение 3), не являющейся константой и принимающей значение «1» для каждой строки рассматриваемого сочетания столбцов таблицы. Другой критерий — критерий неизбыточности, общий для всех логических методов распознавания, основанных на тестовом подходе. По третьему, «нелогическому» критерию, как и во всех методах распознавания образов, оценивается информативность сочетаний признаков с точки зрения возможности различать по ним объекты разных классов эталонной выборки или объекты, не относящиеся к данному классу, если задан только один класс.

Таким образом, в методе Т-свойств представление о тупиковых тестах как неизбыточных сочетаниях признаков, позволяющих

различать все объекты одного класса от всех объектов другого класса, трансформируется в представление о неизбыточных сочетаниях признаков, обладающих заданным табличным свойством и различающих распознаваемый объект S от всех объектов данного класса V или (в случае задания двух классов V и Q) q объектов класса Q ($q \geqslant 1$) от всех объектов класса V, и наоборот.

В качестве меры сходства распознаваемого объекта S с одним из классов V или Q эталонной выборки рассматриваются отношения, полученные на основе процедуры голосования:

$$B_S^1 = \sum_{z \in \omega(V/Q)} z(S)^1 / |\omega(V/Q)|;$$

$$B_S^2 = \sum_{z \in \omega(Q/V)} z(S)^2 / |\omega(Q/V)|,$$

где $\omega(V/Q)$, $\omega(Q/V)$ — множества неизбыточных сочетаний признаков с табличным свойством, отличающих класс V от Q и наоборот; $\omega(V/Q)$, $\omega(Q/V)$ — мощности этих множеств; $z(S)^1$, $z(S)^2$ — неизбыточные сочетания признаков, обладающие заданным табличным свойством, по которым описание распознаваемого объекта S не отличается от описаний эталонных объектов классов, соответственно V и Q.

Результат распознавания определяется сравнением $B^1_{(S)}$ и $B^2_{(S)}$ при некоторых заданных порогах. Кроме того, так же как и в методе тупиковых тестов, вычисляются оценки информативности отдельных признаков по каждому классу эталонных объектов. Особенностью метода Т-свойств является возможность его использования тогда, когда задан только один класс эталонных объектов, а также несколько классов эталонной выборки, слабо сопоставимых по целевому признаку.

Основное применение метод Т-свойств получил в нефтяной геологии, где с его помощью определялась продуктивность различных геологических структур. Кроме того, имеется опыт использования этого метода в рентгено-структурном анализе для диагностики минералов сложного состава.

Метод непрерывных дизъюнктивных нормальных форм (Н. Д. Н. Ф.) [34], разработанный В. О. Красавчиковым, предназначен для обработки таблиц, столбцы которых представлены значениями количественных признаков, принадлежащих отрезку [0,1] или являющихся целыми положительными числами.

В методе Н. Д. Н. Ф., как и в методе тупиковых тестов, выбор информативных сочетаний признаков осуществляется на основе двух критериев: критерия различия (или сходства) описаний объектов из разных классов эталонной выборки и критерия неизбыточности рассматриваемых сочетаний признаков относительно введенной меры сходства. Последний критерий и определяет включение метода Н. Д. Н. Ф. в логические методы распознавания. При выполнении этого метода просмотр сочетаний таблицы Т для оценки их информативности производится путем перебора, начиная с со-

четаний, состоящих из одного признака ($\lambda = 1$), кончая сочетаниями длины $\lambda = \lambda^*$, где λ^* — некоторый заранее заданный порог. Мера сходства описаний объекта из класса Q с описаниями объектов класса P эталонной выборки по рассматриваемому сочетанию t задается следующим образом. Для каждого признака x в сочетании t длины λ_t вычисляется функция $\mu(x)$:

$$\mu(x) = 1 - |\alpha_i - \alpha_p|,$$

где $|\alpha_i - \alpha_p|$ — абсолютная величина нормированной разности значений признака x для сравниваемой пары объектов из классов Q и P с номерами i и p.

Из полученных λ_t величин, которые убывают с увеличением различия по данному признаку для сравниваемых объектов, выбирается минимальная. Эта процедура производится для всех пар сравниваемых объектов, один из которых (S_i) принадлежит, например, классу Q, а остальные классу P. В качестве меры сходства объекта S_i со всеми объектами класса P из полученного на предыдущем шаге набора минимальных величин выбирается максимальная ($D_t(S_i)$), т. е. находится так называемая Н. Д. Н. Ф.

Оценка информативности сочетания признаков t для класса P относительно введенной меры сходства производится по формуле

$$\Delta_t = \sum_{i=1}^{m_2} D_t(S_i),$$

где m_2 — число объектов в классе Q.

Оценка информативности сочетания t, с точки зрения критерия неизбыточности, основывается на неравенстве:

$$\sum_{i=1}^{\lceil m_2 \rceil} D_{t'}(S_i) - \sum_{i=1}^{\lceil m_2 \rceil} D_t(S_i) \geqslant 0,$$

где t' — произвольный набор признаков из сочетания t.

Согласно этому неравенству неизбыточным считается такое сочетание t, при удалении из которого хотя бы одного признака сходство объектов класса Q с объектами класса P по введенной мере возрастает.

На основе этих двух оценок отбираются информативные сочетания. Так, сочетание t считается информативным, если для него $\Delta_t \leqslant \Delta$, $\psi_t \geqslant \psi$, где

$$\psi_t = \max_i \left\{ \min_{t' \subset t} \left[D_{t'}(S_i) - D_t(S_i) \right] \right\};$$

Δ и ψ — некоторые заданные пороги. Выбор информативных сочетаний признаков осуществляется для классов P и Q раздельно.

Для определения принадлежности рассматриваемого объекта S^* к классу P или Q вычисляются значения следующих разделяющих функций:

$$B^P(S^*) = \sum_{t \in r(P)} D_t(S^*) / |\tau(P)|;$$

$$B^Q(S^*) = \sum_{t \in r(Q)} D_t(S^*) / |\tau(Q)|,$$

где $r(P)$, $r(Q)$ — множество информативных сочетаний признаков в классах соответственно P и Q; $|\tau(P)|$ и $|\tau(Q)|$ — число информативных сочетаний в классах P и Q. Распознаваемый объект S^* относится к классу P или Q или ни к одному из них (отказ от распознавания) на основе сравнения по определенным правилам величин $B^P(S^*)$ и $B^Q(S^*)$.

Метод Н. Д. Н. Ф. применялся в нефтяной геологии для установления границ развития пород-коллекторов, а также для оценки продуктивности некоторых нефтеносных структур.

Методы анализа логических зависимостей

Данным термином [41] обозначается комплекс методов, предназначенный для решения задач выявления связей между свойствами различных геологических образований на основе качественной информации с последующим прогнозированием одного из этих свойств.

В основе разработки методов анализа логических зависимостей, базирующихся на трех разделах математической логики — алгебры высказываний, алгебры логики и логики предикатов, лежит определенный подход к анализу качественной геологической информации, резко их отличающий как от логических, так и нелогических методов распознавания образов. Этот подход при прогнозировании оруденения включает:

а) запись общих представлений о типах связей оруденения с геологической обстановкой в виде уравнения алгебры логики;

б) выбор среди анализируемых признаков геологической обстановки их комбинаций, удовлетворяющих составленному уравнению;

в) выбор оптимальной комбинации признаков и подстановку в уравнение соответствующих логических переменных;

г) прогнозирование по полученной формуле исследуемого свойства оруденения на новом объекте.

Единый подход при разработке отдельных методов анализа логических зависимостей обусловил использование в каждом из них ряда общих элементов. Такими общими элементами, на основе которых строится любой из методов анализа зависимостей, являются следующие:

а) логико-математическое моделирование общих представлений о типах связей между изучаемыми геологическими явлениями;

б) выявление логических зависимостей в массивах с фактическими данными;

в) выбор оптимального решения из множества допустимых.

При использовании методов анализа логических зависимостей фактический материал, так же как и в логических методах распознавания, представляется в виде бинарной таблицы T. Однако в отличие от последних в методах анализа логических зависимостей символам 1 и 0, характеризующим в таблице Т наличие или отсутствие геологических признаков, придается несколько иной смысл. Так, относительно этих символов считается, что они заменяют слова

«истина» и «ложь» и обозначают соответственно истинность или ложность высказывания о том, что рассматриваемый геологический признак установлен на данном объекте (при данном наблюдении). Например, при изучении некоторого множества месторождений M высказывание «на месторождении A из M установлены дайки диабазов» может оказаться либо справедливым, либо ложным для месторождения A. В зависимости от этого в столбце таблицы Т, соответствующем признаку «дайки диабазов», в строке, описывающей месторождение A, отмечается либо 1, либо 0.

При такой интерпретации исходных данных можно считать, что бинарная таблица Т содержит значения логических функций — предикатов (см. «Логика предикатов»). Тогда выявление по ней связей между свойствами геологических образований, в частности проявлениями оруденения и геологическим строением площадей, с позиций математической логики можно рассматривать как задачу выделения комбинаций одноместных предикатов, зависимость между которыми описывается какой-либо из функций алгебры логики.

Определение вида этой функции, точнее — вида формулы, которой она задается составляет первый элемент методов анализа логических зависимостей — логико-математическое моделирование общих понятий о связях геологических явлений, лежащих в основе решаемой прогнозной задачи. Это моделирование осуществляется с помощью аппарата алгебры высказываний и алгебры логики. В результате найденная формула алгебры логики далее задается в качестве исходного условия решаемой задачи и используется при выполнении второго элемента рассматриваемых методов — выявления логических зависимостей в бинарной таблице с фактическими данными. Этот элемент методов основан на понятии «логические зависимости», которое вводится следующим образом.

Пусть задана бинарная таблица Т, состоящая из m строк, описывающих множество объектов M и $(n+1)$-го столбцов, содержащих значения $n+1$ предикатов $P^*(x)$, $P_1(x), \ldots, P_2(x)$. Причем столбец значений предиката $P^*(x)$ характеризует прогнозируемое свойство объектов (например, масштаб оруденения, тип минерализации и т. д.).

Очевидно, что в таблице Т можно выделить множество подтаблиц, образованных разными сочетаниями столбцов, но обязательно содержащих столбец значений предиката $P^*(x)$. Среди выделенных подтаблиц могут находиться такие, которыми задаются булевы функции алгебры логики, в том числе и частично определенные функции (см. «Алгебра логики», определение 2). Если при этом в столбце, характеризующем предикат $P^*(x)$, содержатся значения этой функции, а в остальных k столбцах, описывающих другие k предикатов $\hat{P}_1(x), \ldots, \hat{P}_k(x)$ из набора $P_1(x), \ldots, P_n(x)$ — значения логических переменных, то говорят, что предикат $P^*(x)$ логически зависит от предикатов $\hat{P}_1(x), \ldots, \hat{P}_k(x)$.

Так как любую булеву функцию алгебры логики всегда можно

выразить через отрицание, конъюнкцию и дизъюнкцию (в виде Д. Н. Ф.), то в методах анализа логических зависимостей специальными алгоритмами выявляются только зависимости типа конъюнкции и дизъюнкции (выявление зависимостей типа отрицания тривиально). При этом считается, что между предикатом $P^*(x)$ и предикатами $\hat{P}_1(x), \ldots, \hat{P}_k(x)$ существует зависимость типа конъюнкции, если $\beta_i = \alpha_{i1} \wedge \ldots \wedge \alpha_{ik}$, $i=1, \ldots m$, и типа дизъюнкции, если $\beta_i = \alpha_{i1} \vee \ldots \vee \alpha_{ik}$, где β_i — значение в i-й строке предиката $P^*(x)$; $\alpha_{i1}, \ldots, \alpha_{ik}$ — значения в i-й строке предикатов $\hat{P}_1(x), \ldots, \hat{P}_k(x)$.

В результате выполнения второго элемента рассматриваемых методов, выявления логических зависимостей в бинарной таблице Т находится множество комбинаций предикатов, зависимость между которыми описывается заданной функцией алгебры логики. Так как каждому предикату соответствует некоторый геологический признак, то любой выделенной комбинации предикатов соответствует комбинация геологических признаков, имеющих между собой связь определенного типа.

Третьим элементом разработанных логических методов является выбор оптимального решения из множества допустимых. Введение этого элемента обусловлено тем, что при выявлении логических зависимостей одним и тем же типом связи могут обладать несколько комбинаций геологических признаков. Введение критерия оптимальности основывается на практических требованиях к качеству результатов решения геологических задач. На основе этих требований оптимальным решением считается такая из выявленных комбинаций, которая содержит минимальное или неизбыточное число последних. В математическом плане принцип выбора оптимального решения реализуется путем минимизации булевых функций (см. «Алгебра логики», определения 7, 8), описывающих связи в анализируемом массиве эмпирических данных.

В настоящее время разработаны следующие методы анализа логических зависимостей: метод прогнозирования размещения оруденения на основе анализа геологических карт; метод выявления связи оруденения с направлениями глубинных разломов; метод прогнозирования размещения оруденения по комплексу карт в изолиниях; метод прогнозной оценки рудопроявлений. Разработанный комплекс методов анализа логических зависимостей применялся для решения задач прогнозирования, возникших в процессе изучения оруденения ряда регионов Советского Союза и зарубежных стран.

Метод прогнозирования размещения оруденения на основе анализа геологических карт. Рассматриваемый метод предназначен для анализа геологической карты (комплекса карт), на которой отражены различные особенности геологического строения региона и отмечены месторождения и рудопроявления, относящиеся к исследуемому типу оруденения. Требуется с помощью разработанного ме-

тода из легенды карты выбрать комбинации ее элементов (условных обозначений), которыми определяется размещение изучаемого типа оруденения, и затем по ним выделить на карте область развития этого оруденения, в том числе и площади, перспективные на поиски новых рудных объектов.

Формальная постановка этой задачи заключается в следующем. Пусть легенда геологической карты включает n условных обозначений (признаков), которые можно разделить на k групп ($A_1, \ldots, A_\nu, \ldots, A_k$), соответствующих k различным геологическим явлениям (например, магматизму, дизъюнктивной тектонике и т. д.), и пусть имеется m объектов, представляющих собой некоторые участки карты определенного размера, каждый из которых включает точку с оруденением. Кроме того, пусть задано уравнение алгебры логики, описывающей тип связи оруденения с другими геологическими явлениями. Требуется выявить такие комбинации рассматриваемых признаков, связь которых с оруденением и характеризуется заданным уравнением.

Вид уравнения, о котором идет речь в условии задачи, заранее неизвестен, и для его определения проводится логико-математическое моделирование представлений о связи оруденения с другими геологическими явлениями., Они состоят в следующем.

1. Формирование оруденения происходит при взаимодействии нескольких геологических явлений. Это представление можно записать в виде следующей формулы алгебры высказываний:

$$G^* \sim G_1 \wedge \ldots \wedge G_\nu \wedge \ldots \wedge G_k, \qquad (16.1)$$

где G^* — высказывание о том, что установлено оруденение; G_ν — высказывание о том, что установлено ν-е геологическое явление.

2. Всегда, когда устанавливается оруденение, оно представлено изучаемым его типом, и наоборот. Это понятие справедливо только в рамках конкретной задачи, когда все m выделенных на карте точек минерализации относятся к одному ее типу (например, медному). Оно описывается следующей формулой алгебры высказываний:

$$G^* \sim Y^*, \qquad (16.2)$$

где $\wedge^*$ — высказывание о наличии оруденения; Y^* — высказывание о том, что установлен исследуемый тип оруденения.

3. Если обнаружено некоторое геологическое явление, то оно будет выражаться хотя бы одним признаком из характеризующей его группы признаков, и, наоборот, если установлен хотя бы один из таких признаков, то будет установлено и соответствующее явление. Это понятие также справедливо только для конкретной геологической карты, на которой каждое ν-е геологическое явление описывается вполне определенной группой признаков A_ν. Сформулированное понятие выражается формулой:

$$G_\nu \sim Y_1^\nu \vee \ldots \vee Y_l^\nu \vee \ldots \vee Y_{l_\nu}^\nu, \qquad (16.3)$$

где G_ν — высказывание о том, что обнаружено ν-е геологическое явление; $Y_1^\nu, \ldots, Y_{l_\nu}^\nu$ —высказывания о том, что установлены признаки из группы A_ν, характеризующей это явление.

Построенная логико-математическая модель после применения к ней ряда логических преобразований позволяет определить вид формулы алгебры логики, описывающей связь исследуемого оруденения с геологическими признаками из группы $A_1, \ldots, A_k$:

$$F = (u_1^1 \wedge \ldots \wedge u_1^k) \vee \ldots \vee (u_{r_1}^1 \wedge \ldots \wedge u_{r_k}^k). \qquad (16.4)$$

Для решения задачи по полученному уравнению и бинарной таблице T с фактическими данными необходимо определить, какие из n геологических признаков легенды карты имеют заданный этим уравнением тип связи с оруденением.

Бинарная таблица T, на основании которой решается данная задача, включает m строк, характеризующих m участков карты с точками минерализации, для каждого из которых значениями 1 и 0 отмечено наличие или отсутствие геологических признаков из легенды карты. Таблица T разбита на k подтаблиц ($T_1, \ldots, T_\nu, \ldots, T_k$), соответствующих k группам легенды. Принадлежность всех m объектов к одному типу оруденения выражается единичным столбцом таблицы. Этот столбец рассматривается как столбец значений функции F, а остальные столбцы таблицы считаются столбцами значений α_{ij} ($\alpha_{ij} \in \{0, 1\}$, $i = 1, \ldots, m$) некоторых логических переменных $z_1, \ldots, z_j, \ldots, z_n$, разбитых на k подмножеств $\{Z_1\}, \ldots, \{Z_\nu\}, \ldots, \{Z_k\}$. Предполагается, что среди этого набора переменных могут находиться и такие, которые удовлетворяют уравнению (16.4), т. е. являются аргументами $u_1^1 \ldots, u_k^k$ функции F.

Обработка таблицы T по рассматриваемому методу включает два этапа. На первом этапе осуществляется выбор аргументов функции F. С этой целью на основании формулы (16.3) по таблице T составляется m формул, описывающих связь оруденения с ν-й группой геологических признаков для каждого i-го объекта ($i = 1, \ldots, m$):

$$f_i\left(\tilde{z}_{i1}^\nu, \ldots, \tilde{z}_{it_\nu(i)}^\nu\right) = \tilde{z}_{i1}^\nu \vee \ldots \vee \tilde{z}_{is}^\nu \vee \ldots \vee \tilde{z}_{it_\nu(i)}^\nu, \qquad (16.5)$$

где $\tilde{z}_{is}^\nu (s_\nu = 1, \ldots, t_\nu(i))$ — переменная из подмножества $\{Z_\nu\}$, для которой $\tilde{\alpha}_{is}^\nu = 1$.

Формула, описывающая связь оруденения с ν-й группой признаков одновременно для всей выборки из m объектов, имеет вид:

$$\tilde{f} = \tilde{f}_1^\nu \wedge \ldots \wedge \tilde{f}_i^\nu \wedge \ldots \wedge \tilde{f}_m^\nu. \qquad (16.6)$$

После подстановки в формулу (16.6) правых частей формул (16.5) в результате преобразования $\wedge\vee \to \vee\wedge$ находится выражение, в котором конъюнкции образованы наборами переменных, удовлетворяющих уравнению (16.4). Из подмножества $\{X_\nu\}$ среди полученных наборов выбирается один набор $И_\nu$, который содержит

минимальное число переменных. Весь набор аргументов функции F находится в виде объединения $И_k^*$ наборов $И_\nu$:

$$U^* = \bigcup_{\nu=1}^{k} U_\nu. \qquad (16.7)$$

На втором этапе обработки таблицы Т для каждого i-го объекта строится формула вида (16.4):

$$f_i\left(\hat{x}_{i1}^1, \ldots, \hat{x}_{ir_k(i)}^k\right) = (\hat{x}_{i1}^1 \wedge \ldots \wedge \hat{x}_{i1}^\nu \wedge \ldots \wedge \hat{x}_{i1}^k) \vee \ldots$$
$$\ldots \vee (\hat{x}_{it}^1 \wedge \ldots \wedge \hat{x}_{is}^\nu \wedge \ldots \wedge \hat{x}_{il}^k) \vee \ldots$$
$$\ldots \vee \left(\hat{x}_{ir_1(i)}^1 \wedge \ldots \wedge \hat{x}_{ir_\nu(i)}^\nu \wedge \ldots \wedge \hat{x}_{ir_k(i)}^k\right), \qquad (16.8)$$

где $\hat{x}_{is}^\nu$ — такая переменная, что $\hat{x}_{is}^\nu \in U_\nu$; $\hat{\alpha}_{is}^\nu = 1$, $s = 1, \ldots, r_\nu(i)$; $\nu = 1, \ldots, k$. Обозначая логические выражения в скобках (конъюнкции) через $y_1, \ldots, y_\omega$, где ω — общее число различных конъюнкций в формулах (16.8), можно любую формулу (16.8) представить в виде (16.5) и затем к рассматриваемой системе формул применить преобразование $\wedge\vee \to \vee\wedge$.

В результате применения на первом и втором этапах обработки таблицы Т преобразования $\wedge\vee \to \vee\wedge$ строится такая формула вида (16.4), которая при минимальном числе переменных содержит минимальное число конъюнкций. Этим конъюнкциям соответствуют определенные сочетания геологических признаков, для которых общие представления о связи оруденения с комплексом геологических явлений согласуются с фактическими данными. Повторением преобразования $\wedge\vee \to \vee\wedge$ достигается двойная оптимизация решения — сначала по числу переменных, а затем по числу конъюнкций, образованных этими переменными.

Так как анализируемые признаки составляют легенду карты, то выделенному набору их сочетаний (типовым геологическим ситуациям размещения оруденения) на карте соответствуют определенные области. Эти области охватывают все рассматриваемые рудные точки и занимают минимальную площадь по сравнению с любыми другими вариантами построения подобных областей. Области с типовыми геологическими ситуациями, на которых в настоящее время оруденение не установлено, рассматриваются как перспективные на обнаружение новых рудных объектов.

Метод выявления связи оруденения с направлениями глубинных разломов. Этот метод является частным случаем метода прогнозирования размещения оруденения на основе анализа геологических карт, когда исследуется связь оруденения только с одним геологическим явлением— дизъюнктивной тектоникой и соответственно одной группой признаков ($k = 1$) — направлениями глубинных разломов. Ввиду отсутствия принципиальных различий в построении указанных методов метод выявления связи оруденения с направлениями глубинных разломов здесь не приводится.

Метод прогнозирования размещения оруденения по комплексу карт в изолиниях. В геологической информации особое место занимают данные о признаках, которые имеют площадное распространение и в различных точках территории характеризуются числовыми значениями (например, интенсивность трещиноватости, содержание химических элементов, геофизические характеристики). Эти данные по каждому из признаков обычно представляются в виде отдельной карты изолиний, а для совокупности признаков строится комплекс карт в изолиниях. Иногда на таких картах также показано расположение рудных объектов. Тогда возникает вопрос о том, связано ли и как связано оруденение с отображенными на картах признаками и можно ли по этим картам прогнозировать размещение оруденения. Для решения этого вопроса разработан метод анализа логических зависимостей, который применим для исследования территорий, на которых установлено m различных рудных объектов, относящихся к r ($r \geqslant 2$) группам, и построено n ($n \geqslant 2$) карт в изолиниях. В этот метод также входит способ графического представления выявленных связей с выделением на их основе перспективных площадей. Построению прогнозных карт предшествуют следующие основные шаги метода.

1. Представление исходного картографического материала в виде бинарных таблиц Т. Для этого:

а) на картах в изолиниях выделяются m элементарных площадей заданной формы и размера, каждая из которых включает рудный объект;

б) строится числовая таблица C, в m строках которой для каждой элементарной площади отмечается минимальное и максимальное значения признаков по всем n картам в изолиниях (значения a_{ij}, d_{ij} и соответственно $i = 1, \ldots, m$, $j = 1, \ldots, n$);

в) числовая таблица C преобразуется в r бинарных таблиц Т. При построении этих таблиц в числовой таблице C для каждой ν-й, $\nu = 1, \ldots, r$ группы с числом объектов m_ν по каждому j-му признаку фиксируется максимальное

$$c_j^\nu = \max_s a_{sj}^\nu, \qquad s = 1, \ldots, m_\nu, \qquad j = 1, \ldots, n$$

и минимальное $d_j^\nu = \min_s b_{sj}^\nu$ значения.

Полученным $r \cdot n$ интервалам ставятся в соответствие столбцы r новых таблиц Т. В каждой такой таблице T_ν, составленной по интервалам значений признаков для ν-й группы объектов, значения признаков для объектов из остальных $r-1$ групп могут попадать или не попадать в выделенные интервалы. В зависимости от этого в строках таблицы T_ν отмечаются значения 1 или 0. При таком построении любая таблица T_ν в свою очередь разбивается на r подтаблиц. Одна из этих подтаблиц, M^*, включает m_ν единичных строк, являющихся описанием объектов ν-й группы, а другие подтаблицы,

M_k, $k = 1, \ldots, r-1$, $k \neq \nu$ состоят из m_k характеристик объектов $r-1$ остальных групп.

2. Построение логико-математической модели связи оруденения с геологическими признаками, отображенными на картах в изолиниях. Данный этап метода разработан для случая, когда выделенные интервалы значений признаков по отдельным группам объектов перекрываются. Это свидетельствует о связи оруденения не с отдельными признаками, а с их сочетаниями, точнее сочетаниями интервалов значений признаков. Такой тип связи характеризуется уравнением алгебры логики следующего вида:

$$F = u_1 \wedge \ldots \wedge z_t.$$

3. Выявление логических зависимостей в бинарных таблицах T_ν, $\nu = 1, \ldots, r$. Этот этап метода выполняется отдельно для каждой таблицы T_ν и состоит в выявлении в ней сочетаний столбцов (признаков), соотношение между которыми в соответствии с заданным уравнением описывается булевой функцией — конъюнкцией. При этом в каждую таблицу T_ν дополнительно вводится столбец, характеризующий принадлежность объектов к ν-й группе (1, если данный объект относится к ν-й группе, 0 — в ином случае). Вновь введенный столбец рассматривается в качестве столбца значений функций F, а среди остальных столбцов таблицы T_ν выявляются столбцы со значениями аргументов этой функции.

4. Выбор оптимального решения. При выполнении этого шага ставится условие, чтобы среди сочетаний столбцов, связанных логической зависимостью типа конъюнкции, учитывались только сочетания минимальной длины. Выбор сочетаний с указанными свойствами проводится на основе логических преобразований, применяемых при построении минимальных дизъюнктивных нормальных форм булевых функций.

В результате применение метода для каждой таблицы T_ν находится одно информативное сочетание признаков Z_ν, которое характеризуется следующими свойствами: а) на любом объекте ν-й группы значения признаков выделенного сочетания не выходят за пределы анализируемых в данной таблице интервалов; б) на всех объектах других групп значения хотя бы одного признака из выделенного сочетания находятся вне рассматриваемых интервалов*; в) информативное сочетание содержит минимальное число признаков по сравнению с любыми другими сочетаниями, также обладающими свойствами (а) и (б).

С найденными информативными сочетаниями признаков и интервалов значений этих признаков связано размещение изучаемых групп рудных объектов. Такие сочетания можно использовать в целях прогнозирования. Для этого разработаны способы их графической интерпретации с построением схемы металлогенического районирования территории. Последняя представляет собой оптимальный вариант районирования, так как строится на основе только таких сочетаний признаков и интервалов их значений, с которыми связано размещение исследуемых групп рудных объектов, причем с учетом минимального числа признаков. Построенную схему можно использовать и для составления прогнозных карт, заключающегося в выделении в пределах областей развития тех или иных групп рудных объектов отдельных площадей, перспективных на поиски новых проявлений оруденения.

Метод прогнозной оценки рудопроявлений. Рассматриваемый метод предназначен для анализа качественной информации о строении месторождений полезных ископаемых, содержащейся в геологических отчетах, с выявлением на основе этого анализа геологических факторов, контролирующих масштаб изучаемой минерализации (факторов рудообразования) и последующей прогнозной оценки по ним масштабов рудопроявлений. При этом под факторами рудообразования понимаются особенности геологической обстановки, имевшие место до или во время рудообразования и влиявшие на масштаб рудных объектов. Эти факторы рассматриваются как некоторые неизвестные параметры. Предполагается, что с ними, помимо масштаба оруденения, связаны и некоторые геологические признаки строения рудных объектов, наблюдающиеся в настоящее время. На основе такого предположения выявление указанных факторов сводится к задаче нахождения этих геологических признаков с последующей их разбивкой на группы, каждая из которых характеризует результат действия отдельного фактора.

Формальная постановка рассматриваемой задачи состоит в следующем. Пусть эталонная выборка представлена m_1 крупными и m_2 мелкими месторождениями, строение которых охарактеризовано n геологическими признаками, и пусть задано уравнение алгебры логики, описывающей связь масштаба оруденения с геологическими признаками строения месторождений. Требуется выявить такие комбинации признаков, удовлетворяющие уравнению заданного вида, чтобы каждая из них была связана с отдельным фактором рудообразования.

Для определения уравнения, задание которого входит в условие задачи, проводится логико-математическое моделирование общих понятий о связи оруденения с геологической обстановкой его формирования. Эти понятия следующие.

1. Формирование крупных месторождений происходит при взаимодействии нескольких факторов рудообразования. Набор этих факторов одинаков для месторождений со сходным типом оруденения и близкими масштабами. Это понятие описывается формулой алгебры высказываний:

$$Y^* \sim G_1 \wedge \ldots \wedge G_\nu \wedge \ldots \wedge G_k, \qquad (16.9)$$

где Y^* — высказывание о наличии крупного месторождения; G_ν — высказывание о наличии ν-го фактора рудообразования.

* Это свойство не распространяется на группы объектов, которые вообще нельзя отличить от исследуемой, т. е. на такие группы, для которых в подтаблицах M_K таблицы T_ν содержатся единичные строки.

2. На территориях с различным геологическим строением результат действия одного и того же фактора рудообразования может выражаться разными признаками. Формально это представление записывается следующим образом:

$$G_\nu \sim Y_1^\nu \vee \ldots \vee Y_j^\nu \vee \ldots \vee Y_{t_\nu}^\nu, \quad \nu = 1, \ldots, k, \quad (16.10)$$

где Y_j^ν — высказывание о наличии j-го признака, относящегося к ν-му фактору рудообразования.

3. Присутствие признаков, отражающих действие факторов рудообразования, является или только необходимым, или только достаточным условием выявления крупного месторождения, т. е.

или $$Y^* \to Y_j^\nu, \quad j = 1, \ldots, t_\nu, \quad \nu = 1, \ldots, k, \quad (16.11)$$

или $$Y_j^\nu \to Y^*, \quad j = 1, \ldots, t_\nu, \quad \nu = 1, \ldots, k. \quad (16.12)$$

Формулы (16.9) и (16.10) позволяют найти уравнение алгебры логики (уравнение (16.13)), описывающее модель связи масштаба оруденения с факторами рудообразования и характеризующими эти факторы геологическими признаками,

$$F = (u_1^1 \vee \ldots \vee u_{t_1}^1) \wedge \ldots \wedge (u_1^\nu \vee \ldots \vee u_{t_\nu}^\nu) \wedge \ldots \wedge (u_1^k \vee \ldots \vee u_{t_k}^k), \quad (16.13)$$

а из формул (16.11), (16.12) следует способ нахождения решения полученного уравнения.

В частности доказано, что если выполняется условие (16.11), то из истинности высказываний (16.9) и (16.10) следует истинность высказывания

$$Y^* \sim Y_{j_1}^1 \wedge \ldots \wedge Y_{j_\nu}^\nu \wedge \ldots \wedge Y_{j_k}^k.$$

При этом для любых $\nu, \mu = 1, \ldots, k$, $\nu \neq \mu$, выполняются соотношения $Y_{j_\nu}^\nu \in \{Y^\nu\}$, $Y_{j_\mu}^\mu \in \{Y^\mu\}$, где $\{Y^\nu\}$, $\{Y^\mu\}$ — множества простых высказываний, составляющих правые части ν-го и μ-го выражений (16.10).

Если выполняется условие (16.12), то из истинности высказываний (16.9) и (16.10) следует

$$Y^* \sim Y_1^\nu \vee \ldots \vee Y_j^\nu \vee \ldots \vee Y_{t_\nu}^\nu, \quad \nu = 1, \ldots, k.$$

$$Y_j^\nu \in \{Y_\nu\}, \quad j = 1, \ldots, t_\nu.$$

Это означает, что решение уравнения (16.13) можно получить в виде формул алгебры логики, обладающих следующими свойствами:

а) они описывают элементарные функции алгебры логики — конъюнкцию и дизъюнкцию.

б) любая пара аргументов булевой функции конъюнкции в формулах первого типа входит в разные скобки уравнения (16.13), а все аргументы булевой функции — дизъюнкции в формулах второго типа относятся к одной скобке уравнения (16.13).

Построение формул алгебры логики с указанными свойствами проводится на основе бинарной таблицы Т с фактическими данными, в которой масштаб оруденения описывается столбцом, содержащим значение I для m_1 крупных месторождений и значение 0 для m_2

мелких. Этот столбец рассматривается в качестве столбца значений булевой функции F в уравнении (16.13). Остальные столбцы таблицы, соответствующие признакам строения месторождений, считаются столбцами значений логических переменных $z_1, \ldots, \ldots, z_n$, среди которых находятся и неизвестные аргументы функции F.

Функция F может иметь много различных представлений в виде конъюнкций и дизъюнкций каких-либо переменных из набора $z_1, \ldots, z_n$, из которых не каждое обладает свойством (б).

Однако существует возможность из всего множества построенных формул выделить формулы, имеющие это свойство. В частности, как доказано, построенная формула будет обладать свойством (б) только в том случае, если все входящие в нее переменные будут существенными аргументами функции F (см. «Алгебра логики», определение 3).

На основе полученных теоретических результатов на этапе выявления логических зависимостей по таблице Т выбираются такие два набора логических переменных, что функция F выражается как конъюнкция переменных одного из них и дизъюнкция другого.

На этапе выбора оптимального решения для каждого из выявленных наборов осуществляется преобразование $\wedge\vee \to \vee\wedge$. В результате его выполнения строятся две системы формул, любая из которых содержит только существенные аргументы функции F. Системы формул с этим свойством являются единственным решением данной задачи.

После построения двух систем формул набор переменных, входящих в эти формулы, с помощью специальных алгоритмов разбивается на группы, каждая из которых образована переменными, являющимися элементами одной из скобок уравнения (16.13). При этом число элементов в скобке уравнения (16.13) определяется числом переменных в группе, а число таких скобок — числом групп. Так как группе переменных, составляющих любую из скобок уравнения (16.13), соответствует комбинация признаков, характеризующих влияние отдельного фактора рудообразования, то, основываясь на геологической интерпретации полученных групп, можно судить о том, что представляют собой эти факторы. После подстановки в уравнение (16.13) выявленного набора переменных найденная формула используется для прогнозной оценки рудопроявлений с изучаемой минерализацией.

Разработанный метод прогнозной оценки масштабов рудопроявлений применим не только для случая, когда объекты эталонной выборки разбиты на два класса — крупные и мелкие месторождения. При большем числа классов (например, если эталонная выборка разбита на классы крупных, средних и мелких месторождений) сначала находится формула алгебры логики, позволяющая отделить классы крупных и средних месторождений от класса мелких; затем строится формула, по которой можно различить класс крупных месторождений от класса средних и мелких. Соответственно и оценка масштабов оруденения на изучаемых рудопроявлениях проводится на основе двух формул.

ГЛАВА 17

ИНФОРМАТИКА

Научная дисциплина, изучающая структуру и общие свойства научной информации, а также закономерности всех процессов научной коммуникации — получения, передачи, поиска и использования научной информации, обозначается в современной науке термином «информатика». Научная и производственная деятельность во всех областях осуществляется на основе обработки научной информации и данных о ходе и результатах природных и социальных процессов. Прикладные задачи информатики заключаются в разработке более эффективных методов и средств осуществления информационных процессов, в определении способов оптимальной коммуникации с широким применением современных технических средств в конкретных отраслях науки и производства. В информатике применяются методы, средства и аппарат ряда других наук — математики, физики, кибернетики, лингвистики.

Одной из основных проблем информатики является создание теории проектирования и практического применения систем поиска информации или систем управления данными, которые, в свою очередь, являются ядром любой системы обработки информации (данных). Функциональный комплекс средств и методов, применяемых для накопления, систематизации и хранения данных, а также для осуществления операций выделения из имеющегося множества данных таких подмножеств, которые соответствуют требованию (запросу) на их извлечение, обозначается в информатике термином «информационно-поисковая система», или «банк данных». Приведенные термины в определенном смысле являются синонимами. Семантической основой термина «информационно-поисковая система» являются понятия «информация» и «поиск», а термина «банк данных» — понятия «данные», «управление» и «банк» («хранилище»). Терминология, основанная на понятиях «данные» и «управление», предпочтительнее терминологии, основанной на понятиях «информация» и «поиск», поскольку понятия «данные» и «управление» имеют более четкие определения.

Понятия, методы и средства информатики широко используются в практике функционирования государственной системы научно-технической информации и при реализации программ комплексной автоматизации системы управления народным хозяйством и создании автоматизированных систем обработки данных в науке и на производстве.

В геологии проблема создания систематизированной, качественной и достоверной информационной основы для задач прогноза, поисков и разведки полезных ископаемых, а также оптимального управления геологоразведочным производством решается с широким использованием современных представлений и средств информатики. В данной главе приведены наиболее важные и широко употребительные в геологии понятия информатики. При этом определение понятия дано в большинстве случаев в прикладном их значении, а именно так, как эти понятия применяются в практике анализа геологической информации и создания геологических баз данных. Для удобства пользования справочником термины сгруппированы в три подраздела: общенаучные термины, банки данных и информационно-поисковые системы.

Более подробно с понятиями информатики можно ознакомиться в специальной литературе [20, 32, 44, 47].

ОБЩЕНАУЧНЫЕ ТЕРМИНЫ

А в т о м а т и з а ц и я — применение научно обоснованной методологии и технических средств для регулирования некоторого процесса передачи методик, энергии или информации, в результате которого полностью или частично устраняется участие человека.

В прикладном значении под автоматизацией в большинстве случаев понимается использование средств вычислительной техники и экономико-математического аппарата для совершенствования (оптимизации) управления конкретными процессами (предприятием, отраслью, технологической операцией, процессом решения какой-либо задачи и т. д.). Процессы и регулирующие ход процессов организационные и технические системы, в которых применяются средства автоматизации, обычно определяются как автоматизированные. Например, автоматизированная система управления геологоразведочной отраслью народного хозяйства — АСУ-Геология; автоматизированная система подсчета запасов нефти и газа; автоматизированный банк данных по скважинам глубокого разведочного бурения; автоматизированная информационно-поисковая система по геологии рудных полезных ископаемых.

И н ф о р м а ц и я. Понятие «информация» является одной из общенаучных, философских категорий и из-за многообразия его толкований и приложений не имеет общепринятого определения. До середины XX в. этот термин имел смысл сообщения, сведения о чем-либо, передаваемом людьми. В настоящее время отдельные аспекты понятия «информация» рассматриваются в таких научных дисциплинах, как теория информации, теория связи, кибернетика, информационная теория управления, лингвистика, социология, общая теория систем, генетика и др.

В прикладных отраслях науки и практики, таких как создание вычислительных систем, автоматизированных систем управления, систем информационного обслуживания и т. д., понятие «информация» в большинстве случаев совпадает с понятием «данные».

В геологии широко используется термин «геологическая информация», понимаемый как совокупность данных о строении, свойствах и закономерностях образования и развития геологических объектов.

С и с т е м а — совокупность элементов, находящихся в отношениях и связях друг с другом, которая образует определенную целостность, единство. С середины XX в. понятие «система» становится одним из ключевых философско-методологических и специально-научных понятий. В современном научном и техническом значении разработка проблематики, связанной с исследованием и конструированием систем разного рода, проводится в рамках общей теории систем, различных специальных теорий систем, в кибернетике, системо-технике, системном анализе и т. д.

Д а н н ы е — факты и идеи, представленные в формализованном виде, позволяющем передавать или обрабатывать эти факты и идеи при помощи некоторого процесса (и соответствующих технических средств). Данные всегда зафиксированы на каких-либо материальных носителях и характеризуют строение и свойства объектов, ход и результаты процессов.

Наблюденные, измеренные и зафиксированные факты о строении и свойствах геологических объектов, о ходе и результатах природных процессов или производственно-технических процессов геологического изучения недр образуют множество геологических данных.

К л а с с и ф и к а ц и я: 1) правило (или совокупность правил) отнесения объектов или понятий к группам (разделам, классам), характеризующимся некоторыми общими свойствами; 2) система соподчиненных понятий (классов объектов); 3) процедура отнесения понятий или объектов к классам (группам); 4) перечень понятий (объектов), входящих в классы (группы).

Классификация служит мощным научным методом исследования, который позволяет систематизировать результаты предшествующего развития данной отрасли познания, представить в обобщенном виде картину состояния науки, а также делать обоснованные прогнозы относительно неизвестных еще факторов или закономерностей.

Создание автоматизированных систем обработки данных невозможно без использования естественных и искусственных классификаций объектов, свойств объектов, значений свойств объектов. Систематизированные в форме классификаций значения о строении и свойствах объектов служат основой для разработки моделей организации данных в базах данных.

Для проектирования и создания баз данных в геологии необходимым условием является наличие четких классификаций, природных, технологических и организационно-экономических объектов.

Примерами классификаций в геологии могут служить: виды полезных ископаемых, типы месторождений полезных ископаемых, горные породы, типы горных выработок, типы буровых установок и т. д.

П о и с к и н ф о р м а ц и и — последовательность операций, выполняемых с целью выделения из имеющегося множества данных таких подмножеств, которые соответствуют требованию (запросу)

на их извлечение. Поиск информации осуществляется путем сравнения содержания каждого из имеющихся в множестве элементов с содержанием запроса. Решение о соответствии (или несоответствии) элемента множества запросу принимается в зависимости некоторого содержательного или формального критерия. Такой критерий называют критерием смыслового соответствия. Поиск информации осуществляется, например, при получении очередного номера реферативного журнала «Геология». Специалист прочитывает каждый реферат, сравнивает его содержание со своими научными потребностями и в отношении каждого реферата решает, в какой степени реферат (вернее, научная статья или отчет, по которым составлен реферат) отвечает его интересам.

В связи с ростом объема научной и производственной информации ее поиск становится все более трудоемким. Поэтому в информатике интенсивно разрабатываются методы и средства для формализации, механизации и автоматизации операций поиска информации на основе применения современных быстродействующих электронных вычислительных машин (ЭВМ). В ЭВМ с большой скоростью (миллионы операций в секунду) осуществляются простейшие операции сравнения двоичных символов и их последовательностей. Если обеспечить представление информации и данных в форме последовательностей двоичных символов (эта операция обозначается термином «кодирование»), то поиск информации по четким однозначным правилам (алгоритм) можно поручить ЭВМ. Автоматизированный поиск информации осуществляется при помощи информационно-поисковых систем и систем управления базами данных.

У п р а в л е н и е д а н н ы м и — комплекс операций, выполняемых с целью организации множеств данных (баз данных) и использование данных для обработки при решении задач. Является более широким понятием, чем поиск информации. Управление данными осуществляется в любой системе обработки данных.

О б ъ е к т — философское общенаучное понятие для отображения представлений субъекта об организации объективной реальности, о строении, свойствах и взаимосвязях множества предметов, тел, процессов. В информатике понятие «объект» используется для определения множества объектов материального мира, информация о строении и свойствах которых обрабатывается в информационных системах. При создании информационных систем или банков данных всегда требуется определить, какие именно объекты и процессы (или их типы) будут являться объектами описания в информационных массивах или базах данных. Процесс определения объектов описания и связей между ними и составляет существо работы, которую в специальной литературе называют проектированием концептуальной модели предметной области.

Многообразие объектов и процессов в природе и обществе делает проблему создания всеобъемлющей классификации явлений, процессов и объектов практически и теоретически трудноразрешимой. Поэтому на практике применяются частные, ограниченные класси-

фикации некоторых подмножеств объектов, процессов и явлений. Задача разработки прагматической классификации объектов предметной области, данные о строении и свойствах которых используются в геологической науке и геологоразведочном производстве, имеет важное значение для создания автоматизированных банков данных и автоматизированных систем обработки данных в отрасли. В основу содержательной структуризации геологических данных могут быть положены классификации наук о Земле, классификации методов получения данных о геологическом строении недр, классификации природных геологических объектов, организационно-функциональная структура геологоразведочного производства и другие подходы. В большинстве предложенных классификаций геологических данных используется следующая схема выделения объектов описания.

1. Геологические (природные) объекты.

1.1. Точки геологических тел — нульмерные геологические объекты (образцы, пробы и т. д.).

1.2. Векторы (линии геологических тел) — объекты одномерного пространства (геологические разрезы скважин, керна и т. д.).

1.3. Плоскости сечения геологических тел — объекты двумерного пространства (геологические разрезы, обнажения, проекции геологических тел в изолиниях).

1.4. Объемы геологических тел — объекты трехмерного пространства (залежи, рудные тела, литолого-стратиграфические комплексы отложений и др.).

2. Технологические объекты.

2.1. Технологические точки (точки заложения скважин, пункты взрыва, точки наблюдения, точки опробования и др.).

2.2. Технологические линии (ствол скважины, профиль, маршрут).

2.3. Технологические площади (поисково-разведочные площади, территории геологосъемочных работ разного масштаба).

2.4. Технологические объемы, т. е. совокупность технологических объектов на территории проведения геологоразведочных работ (страны, области, региона, района).

3. Организационно-экономические объекты.

3.1. Организационные точки — организационные объекты, имеющие самостоятельное финансирование на проведение геологоразведочных работ (отряды, партии, экспедиции).

3.2. Организационные линии — организационные объекты (геологические организации), выполняющие законченную последовательность видов геологоразведочных работ определенного целевого назначения.

3.3. Организационные площади — организационные объекты (геологические организации), выполняющие весь комплекс работ по изучению и разведке недр определенной территории.

3.4. Организационные объемы — совокупность организационных объектов определенного уровня структуры управления отраслью.

Все указанные типы объектов описания используются при производстве и управлении геологоразведочными работами. Геологические организации выполняют определенные виды и методы работ на технологических объектах с целью получения первичных геологических данных. Затем на основе первичных данных строятся с той или иной детальностью и достоверностью модели природных (геологических) объектов. На основе анализа моделей геологических объектов, с одной стороны, и потребностей общества в ресурсах полезных ископаемых — с другой, вновь планируется работа организаций. Таким образом, уже на содержательном этапе структуризации геологических и технико-экономических данных отчетливо видна целесообразность организации их в виде комплекса баз данных. В каждом таком комплексе баз данных используются различные типы объектов описания. Кроме того, указываются и задаются логические связи между объектами описания и показателями (атрибутами) в комплексе баз данных. По конкретной территории (например, по стране, части страны, нефтегазоносной провинции, месторождению, конкретной организации и т. д.) имеется реальная возможность и необходимость создания и эксплуатации следующих баз данных:

— характеристика территории работ;
— паспорта организационно-экономических объектов;
— технологические объекты геологоразведочных работ;
— геологические (природные) объекты и модели;
— нормативно-справочные данные.

Система баз данных (БД) по той или иной территории формируется в соответствии с комплексом видов и методов производственных геологоразведочных работ (наполнение БД «Технологические объекты»), специализацией работ по видам полезных ископаемых (наполнение БД «Геологические (природные) объекты и модели»), организационной структурой геологической службы (наполнение БД «Паспорта организационно-экономических объектов»). Объемные характеристики системы баз данных определяются объемами выполненных геологоразведочных работ. Объекты разных типов и масштаба описываются совокупностями показателей, состав которых определяется исходя из информационных потребностей задач обработки данных и управления производством.

БАНКИ ДАННЫХ

Банк данных — основная компонента в современных автоматизированных системах обработки данных разного назначения, обеспечивающая выполнение всех операций по созданию, ведению и использованию данных. Банк данных понимается как совокупность баз (или набора баз) данных, программных средств (называемых системой управления базами данных), технических средств и коллектива специалистов, ответственных за функционирование банка данных (администрация банка данных). Банк данных обеспечивает интеграцию функций информационного обслу-

живания пользователей и задач обработки данных в системе. Главные функции банка данных: ввод данных в базы данных; поддержание баз данных в актуальном состоянии; защита данных от несанкционированного доступа; поиск и предоставление пользователям и для задач требуемых данных.

Общая цель разработки и эксплуатации банков данных в геологической отрасли заключается в создании систематизированной, качественной и достоверной информационной основы для функционирования автоматизированных систем обработки информации автоматизированных систем обработки данных разного целевого назначения. Создание банка данных заключается в формировании массивов (баз) данных о строении и свойствах природных, технологических и организационно-экономических объектов и обеспечении средств для эффективного автоматизированного поиска и извлечения данных для обработки.

При проектировании банков данных решаются следующие комплексы задач:

— создается концептуальная информационная модель предметной области или информационных потребностей абонентов;

— разрабатывается логическая модель баз данных;

— выбираются (или создаются) программные средства, с помощью которых будет создаваться и эксплуатироваться банк данных;

— разрабатывается физическая модель баз данных, которая отображается на физические средства вычислительной техники, применяемой для эксплуатации банка данных;

— разрабатываются формы, способы и средства предоставления данных потребителям;

— разрабатываются формы и способы поступления данных на вход банка данных;

— разрабатывается технология эксплуатации банка данных и комплекс эксплуатационной документации для специалистов, обеспечивающих эксплуатацию банка данных.

В зависимости от области применения и режима эксплуатации банки данных имеют специфическую целевую направленность, которая отражается в собственном названии банка данных. Эксплуатируются или проектируются, например, такие банки данных: «Скважины глубокого разведочного бурения на нефть и газ», «Банк данных технико-экономических показателей автоматизированной системы плановых расчетов», «Ресурсы полезных ископаемых земного шара», «Физико-химические свойства горных пород», «Кадастр месторождений и залежей руд черных металлов» и др.

В создании и эксплуатации банков данных участвуют работники всех основных служб геологической отрасли: производственных геологосъемочных, поисковых и разведочных организаций, специализированных подразделений по созданию АСУ-Геология, научно-исследовательских организаций, отраслевой сети геологических фондов, отраслевой системы научно-технической информации. Основой для координации деятельности этих служб при создании

распределенной отраслевой сети банков данных и обеспечения совместимости формируемых баз данных служит комплекс отраслевых соглашений (стандартов) по принципиальным вопросам проектирования и эксплуатации банков данных.

Б а з а д а н н ы х — упорядоченное множество данных о строении и свойствах объектов некоторой предметной области, организованное в соответствии с принятой моделью предметной области и реализованное на материальных носителях в форме, позволяющей манипулировать данными с помощью технических средств вычислительной техники в процессе целенаправленной обработки данных. База данных является наиболее существенной и важной составной частью банка данных.

М о д е л ь б а з ы д а н н ы х. Понятие «модель базы данных» введено для отображения множества объектов, их свойств и взаимосвязей, для задания определенной структуры, в соответствии с которой организовано размещение данных в базе данных (БД). Понятие «модель БД», с одной стороны, служит для организации процесса разработки базы данных, с другой — оно означает конкретное строение и состав, структурный план, в соответствии с которым построена база данных. Часто наряду с термином «модель БД» применяется термин «схема БД». Различают общую, генеральную БД (схему БД) и прикладную (подсхему) модели; последняя отражает представление о структуре базы данных с точки зрения отдельного пользователя. В процессе проектирования базы данных последовательно разрабатывают: 1) понятийную модель предметной области; 2) понятийную модель базы данных, 3) логическую модель БД; 4) физическую модель организации БД.

Для разработки моделей БД применяют специальные языки описания данных. Логическая модель БД может быть построена с использованием разных принципов и правил отражений связи (отношений) меджду объектами описания и их свойствами (атрибутами). Наиболее известными являются следующие типы моделей: файловая, иерархическая, сетевая, реляционная и соответствующие этим типам моделей языки описания данных. Во всех языках описания данных вводятся понятия «объект», «атрибут», «тип связи между объектами и атрибутами», а также задаются средства и правила описания объектов, атрибутов и связей (отношений). После построения логической модели БД переходят к построению физической модели организации БД, т. е. к планированию размещения данных на физических носителях информации в памяти ЭВМ.

Модель базы данных логическая — модель данных для некоторой части предметной области, в которой общая модель отображается в структуре данных определением организационных единиц структуры данных и спецификацией их свойств и отношений между ними. Логическим средством структуризации данных посвящены предложения КОДАСИЛ (Ассоциация по языкам систем обработки данных).

Модель базы данных реляционная — модель базы данных, предложенная в 1970 г. американским ученым Е. Ф. Коддом. Реляцион-

ная модель основана на представлении данных в виде отношений между ними, при этом представление отношений подвергается нормализации и пошаговому процессу приведения их к двумерной табличной форме, причем информация о них сохраняется полностью. Имеется несколько подходов построения такого рода моделей, основанных на реляционной алгебре и реляционном исчислении.

Модель базы данных физическая — модель базы данных, отображающая логическую модель в выбранной структуре хранения с учетом свойств конкретной вычислительной и программной обстановке.

Модель базы данных сетевая представлена структурной диаграммой в виде произвольного ориентированного графа. Каждая вершина графа соответствует записи, и он отличается от дерева тем, что некоторые порожденные записи могут иметь несколько исходных. Сетевая модель упрощает проблемы, связанные с хранением данных, обладает большей по сравнению с иерархической моделью симметрией, но меньшей наглядностью.

Система управления базами данных (СУБД) — комплекс программных средств, специально предназначенных для реализации на ЭВМ всех процедур создания и ведения баз данных и обеспечивающих извлечение данных по требованию пользователей или задач обработки данных. В зависимости от представляемых возможностей по организации баз данных в соответствии с различными моделями баз данных (иерархической, сетевой, реляционной и др.), а также в зависимости от вида обрабатываемой информации (документальной, фактографической, графической и др.) разработаны специальные СУБД: для обработки текстов на естественном языке, на формализованных информационно-поисковых языках, для обработки баз данных, организованных в соответствии с различными типами моделей БД. Выбор конкретных комплексов программных средств для создания и эксплуатации банка данных осуществляется на этапе технического проектирования и является одним из наиболее ответственных проектных решений.

Администратор базы данных — это специалист, имеющий представление о прикладных задачах, решаемых пользователями с помощью базы данных, работающий в тесном контакте с пользователями и администраторами других баз данных и отвечающий за определение, загрузку, защиту и эффективность баз данных в банке данных.

Администратор данных в базе данных — лицо или группа лиц, ответственные за функционирование базы данных и развитие ее схемы данных. Администратор данных отвечает за сохранность данных всего учреждения или той их части, с которой связана его система. Он осуществляет контроль за всей структурой данных.

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

Информационно-поисковая система представляет собой совокупность языковых, программных и технических средств, предназначенных для хранения, поиска и выдачи искомой информации из имеющегося множества информационных единиц (информационного массива).

При выполнении исследований и производственной деятельности всегда требуется осуществлять операции поиска информации. Например, поиск статей и отчетов, в которых рассматривается проблема рациональной методики разведки залежей нефти и газа на больших глубинах; поиск геологических разрезов скважин, в которых мощность толщи песчаников в верхнемеловых отложениях больше 10 м; поиск результатов измерения значений плотности пород в определенном районе территории города и т. п.

Поиск информации возможен тогда, когда имеется множество информационных единиц — статей, скважин, значений плотности пород и др. и информационный запрос, в котором конкретизирована потребность в информации, т. е. указано, какими свойствами (или значениями свойств) должны обладать искомые информационные единицы, а также правило установления соответствия свойств информационных единиц предъявленному запросу. В случае когда множество информационных единиц заранее никак не упорядочено, операция поиска информационных единиц осуществляется последовательным сравнением каждой из имеющихся информационных единиц с запросом. На практике множества информационных единиц часто столь велики, что последовательное сравнение каждой информационной единицы с запросом требует больших затрат труда и времени. Например, практически невозможно для отыскания статей по методике разведки глубокозалегающих залежей нефти и газа последовательно прочитать все имеющиеся в библиотеке книги и статьи. Для того чтобы облегчить и ускорить выполнение операций поиска информации, используется метод упорядочения, систематизации и унификации представления информационных единиц в множестве (в информационном массиве). Средством для этого служит искусственный формализованный язык описания информационных единиц и запросов.

В отличие от естественного языка, искусственные формализованные языки призваны обеспечить однозначность описания информационных единиц. Такие искусственные языки обозначаются в информатике термином «информационно-поисковые языки (ИПЯ)». Информационно-поисковая система обеспечивает автоматизированный поиск информации. Автоматизация поиска информации становится возможной благодаря тому, что описание информационных единиц в информационном массиве и информационных запросов выполнено на искусственном информационно-поисковом языке, а операция сравнения описаний информационных единиц с запросами является точно определенной, подчиняющейся формально-логическим правилам. Чтобы был возможен автоматизированный

поиск информации, необходимо предварительно перевести описания информационных единиц и запросов с естественного языка на информационно-поисковый язык. Процедура перевода обозначается термином «индексирование». Правило сравнения описаний информационных единиц с описанием информационного запроса, выраженное средствами информационно-поискового языка, обозначается термином «критерий смыслового соответствия».

Таким образом, в абстрактном понимании информационно-поисковая система представляет собой совокупность информационно-поискового языка, правил индексирования и критерия смыслового соответствия. Практическая реализация информационно-поисковой системы означает:

1) создание информационного массива, т. е. множества описаний информационных единиц на информационно-поисковом языке, зафиксированных на каком-либо материальном носителе (перфокарте, фотопленке, магнитной ленте и др.);

2) применение некоторого технического информационно-поискового устройства (например, устройств для сортировки библиографических карт и перфокарт, устройств счетно-перфорационной техники, электронно-вычислительных машин ЭВМ);

3) разработку технологии поиска информации с применением технических средств в виде набора инструкций по эксплуатации информационно-поисковой системы;

4) создание коллектива специалистов для эксплуатации ИПС.

Следовательно, в практическом смысле информационно-поисковая система представляет собой совокупность информационного массива на материальных носителях информации, информационно-поискового устройства, эксплуатационных инструкций для всех операций обработки информации и коллектива специалистов, обеспечивающих эксплуатацию ИПС.

ГЛАВА 18

МНОЖЕСТВА

За последнее десятилетие в связи с ростом научного уровня в области применения математических методов в геологии все чаще стали использоваться различные понятия теории множеств и соответствующие им теоретико-множественные операции. Дело в том, что конечным результатом решения поставленной геологической задачи с формальных позиций является нахождение некоторого множества объектов или точек. Так, например, задачу обычного оконтуривания рудного тела можно представить как нахождение верхней грани (см. ниже) всех множеств точек — таких, чтобы соответствующие им содержания полезного компонента были больше или равны заданному значению.

Аналогично можно сформулировать некоторые задачи разграничения, выбора информативных признаков, изучения зональности, прогнозирования и др.

А л г е б р а — не пустой класс множеств, замкнутый относительно всех конечных операций. Если Ω — пространство, а $\varnothing$ — пустое множество, то очевидно, что каждая алгебра содержит $\varnothing$ и Ω. Иногда алгебру называют полем.

σ-а л г е б р а — непустой класс множеств, замкнутый относительно всех счетных операций. Каждая σ-алгебра является алгеброй. Иногда σ-алгебру называют σ-полем.

В к л ю ч е н и е. Говорят, что A есть подмножество B или включается в B, или содержится в B, если все точки A являются точками B. Это записывают как $A \subset B$ или $B \supset A$. Иными словами, если точка $t \in A$ влечет за собой $t \in B$, то $A \subset B$, и обратно. Отношение включения рефлексивно и транзитивно, т. е. $A \subset A$; $A \subset B$ и $B \subset C$ влекут за собой $A \subset C$.

В е р х н я я г р а н ь — это объединение всех множеств класса, представляющее собой множество точек, которые принадлежат хотя бы одному A_t. Верхняя грань обозначается как $\bigcup_{t \in T} A_t$ или $\sup_{t \in T} A$, где T — множество индексов t. Обозначение sup читается как супремум.

Д о п о л н е н и е. Разность $A—B$ множеств A, B представляет собой множество всех точек в A, не принадлежащих B, называется дополнением B до A и обычно обозначается B.

З а м к н у т о с т ь. Класс A множеств A является замкнутым относительно теоретико-множественной операции, если множества, полученные из множеств класса A с помощью этой операции, также входят в A. В частности, полный класс множеств $S(\Omega)$, порожденных пространством Ω, замкнут относительно любой операции.

К л а с с м н о ж е с т в. Множество множеств называется классом. Обычно классы обозначают $A, B, \ldots$, в отличие от множеств $A, B, \ldots$ с индексами и без них. Если Ω — фиксированное, непустое множество, называемое пространством, то класс всех множеств из Ω называется пространством множеств в Ω и обычно обозначается $S(\Omega)$; все теоретико-множественные понятия и операции применимы к классам, рассматриваемым как множества в соответствующем пространстве множеств.

М н о ж е с т в о. Совокупность произвольных объектов называется множеством. Множество, не содержащее элементов, называют пустым и обычно обозначают $\varnothing$. В общем случае каждое множество будет состоять из элементов ω некоторого фиксированного множества, Ω, которое называется пространством. Элементы ω пространства Ω обычно называют точками.

Если точка ω принадлежит множеству A, то это записывается $\omega \in A$, если же точка ω не принадлежит A, то обычно пишут $\omega \notin A$ или $\omega \overline{\in} A$.

Нижняя грань — это пересечение, или множество, всех точек, принадлежащих каждому множеству A_t множеств класса $\{A_t\}$. Нижняя грань обычно обозначается

$$\inf_{t\in T} A_t = \bigcap_{t\in T} A_t,$$

где inf читается «инфимум».

Объединение множеств — см. «Верхняя грань».

Пересечение множеств — см. «Нижняя грань».

Пространство. Фиксированное непустое множество (обозначим его Ω) называется пространством. Элементы Ω пространства называются точками.

Последовательность множеств. Упорядоченный счетный бесконечный класс множеств $A_1, A_2, \ldots$ называется последовательностью A_n, где каждому значению $n = 1, 2, 3, \ldots$ ставится в соответствие множество A_n. Множества A_n, различные или совпадающие, отличаются одно от другого только индексом.

Последовательность множеств монотонная. Последовательность A называется монотонной, если она неубывающая $(A_n\uparrow)$, т. е. $A_1 \subset A_2 \subset A_3 \subset \ldots$, или невозрастающая $(A_n\downarrow)$, т. е. $A_1 \supset A_2 \supset A_3 \supset \ldots$.

Предел монотонной последовательности множеств. Каждая монотонная последовательность сходится, причем $\lim A_n\uparrow = \bigcup A_n$; $\lim A_n\downarrow = \bigcap A_n$.

Пределы же для произвольной последовательности B можно определить следующими формулами:

$$\liminf B_n = \lim_n (\inf_{k\geqslant n} B_k);$$

$$\limsup B_n = \lim_n (\sup_{k\geqslant n} B_k).$$

Поле — см. «Алгебра».

σ-поле — см. «σ-алгебра».

Сумма множеств. Если каждые два множества из класса $\{A_t\}$ не пересекаются, то $\{A_t\}$ называют классом без пересечений, а объединение множеств из такого класса называется суммой и обозначается ΣA_t.

ГЛАВА 19

МАТРИЦЫ

Широкое применение при решении геологических задач с помощью многомерных математических методов, позволяющих обрабатывать совместно данные по комплексу признаков, приводит к необходимости использовать такие понятия, как «матрица», «вектор-столбец», вектор-строка» и к действиям над ними.

Особенно широко эти понятия используются в многомерных статистических методах, а также при решении систем уравнений, при обработке данных по угловым измерениям и др.

В связи с этим ниже приведены основные сведения о матрицах, векторах и действиях над ними.

ОПЕРАЦИИ НАД МАТРИЦАМИ

Матрицей называется прямоугольная таблица чисел:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{s1} & a_{s2} & \ldots & a_{sn} \end{pmatrix}.$$

Число a_{ij}, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, n+1$ называется элементом матрицы A, индексы i и j указывают соответственно номер строки и столбца, на пересечении которых находится a_{ij} n-мерный вектор $a_i = (a_{i1}, a_{i2}, \ldots, a_{in})$ называется i-й строкой матрицы A, а s-мерный вектор

$$a^j = \begin{pmatrix} a_{1j} \\ \ldots \\ a_{sj} \end{pmatrix}$$

ее j-м столбцом.

Если $s = n$, то матрица A называется квадратной, а число n — ее порядком.

Упорядоченная совокупность элементов $a_{11}, a_{22}, \ldots, a_{nn}$ квадратной матрицы A называется главной диагональю этой матрицы.

Сумма $a_{11} + a_{22} + \ldots + a_{nn}$ элементов главной диагонали матрицы называется следом этой матрицы.

Квадратная матрица называется вырожденной (или особенной), если ее определитель равен нулю, и невырожденной (или неособенной), если он отличен от нуля.

Квадратная матрица A называется диагональной, если все ее элементы равны нулю, кроме элементов $a_{11}, a_{22}, \ldots, a_{nn}$.

Прямоугольная матрица, содержащая s строк и n столбцов, называется диагональной, если все ее элементы равны нулю, кроме элементов $a_{11}, a_{22}, \ldots, a_{kk}$, где $0 \leqslant k \leqslant \min(s, n)$, равных единице. Ранг этой матрицы равен k.

Единичной матрицей называется матрица E:

$$E=\begin{pmatrix} 10 & \dots & 0 \\ 01 & \dots & 0 \\ \dots & \dots & \dots \\ 00 & \dots & 1 \end{pmatrix}.$$

Обратной матрицей для данной матрицы A называется матрица A^{-1}, удовлетворяющая условию

$$AA^{-1}=A^{-1}A=E.$$

Для невырожденной матрицы A существует единственная матрица A^{-1}, удовлетворяющая условию $AA^{-1}=A^{-1}A=E$.

Для нахождения элементов обратной матрицы следует разделить соответствующие элементы присоединенной матрицы на значение определителя матрицы A (для невырожденной матрицы A):

$$A^{-1}=\begin{pmatrix} \frac{A_{11}}{d} & \frac{A_{21}}{d} & \dots & \frac{A_{n1}}{d} \\ \frac{A_{12}}{d} & \frac{A_{22}}{d} & \dots & \frac{A_{n2}}{d} \\ \frac{A_{1n}}{d} & \frac{A_{2n}}{d} & \dots & \frac{A_{nn}}{d} \end{pmatrix}.$$

Определитель обратной матрицы A^{-1} связан с определителем исходной матрицы A следующим соотношением:

$$|A^{-1}|=1/|A|.$$

Ортогональной матрицей Q называется матрица, для которой транспонированная матрица Q' совпадает с обратной Q^{-1}. Для того чтобы квадратная матрица Q была ортогональной, необходимо и достаточно, чтобы сумма квадратов всех элементов любой ее строки равнялась единице, а сумма произведений соответственных элементов любых двух ее различных строк равнялась нулю. Аналогичное утверждение можно сформулировать и для столбцов матрицы. Определитель ортогональной матрицы равняется $+1$ или -1.

Матрица, обратная ортогональной, сама является ортогональной. Матрица, получающаяся в результате умножения двух ортогональных матриц, также ортогональна.

Присоединенной (или взаимной) матрицей к матрице A называется матрица

$$A^{*}=\begin{pmatrix} A_{11} & A_{12} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix},$$

элементы которой A_{ij} являются алгебраическими дополнениями к элементам a_{ji} матрицы A.

Если исходная матрица A невырожденная, то и присоединенная к ней матрица A^{*} тоже невырожденная, а определитель присоединенной матрицы равен $(n-1)$-й степени определителя исходной матрицы A порядка n.

Квадратная матрица порядка n называется симметрической, если ее элементы, симметричные относительно главной диагонали, равны:

$$a_{ij}=a_{ji}.$$

Матрица A тогда и только тогда будет симметрической, когда она совпадает со своей транспонированной матрицей, т. е. если $A'=A$.

Все характеристические корни симметрической матрицы действительны.

Для любой симметрической матрицы A найдется ортогональная матрица Q, приводящая матрицу A к диагональному виду. Другими словами, матрица $Q^{-1}AQ$, полученная трансформированием матрицы A матрицей Q, будет диагональной.

При этом на главной диагонали полученной диагональной матрицы будут расположены характеристические корни матрицы A, взятые с их кратностями.

Сопряженной матрицей к матрице $A=\{a_{ij},\ i=\overline{1,\ m},\ j=\overline{1,\ n}\}$ называется матрица $A^{*}=\{a_{ji}^{*}\}$ с элементами $a_{ji}^{*}=\bar{a}_{ij}$, где $\bar{a}_{ij}$ — величина, комплексно-сопряженная к a_{ij}. Если $\gamma=\alpha+i\beta$ — некоторая комплексная величина, то комплексно-сопряженной к ней называется величина $\bar{\gamma}=\alpha-i\beta$ (α, β — действительные). Для действительных γ справедливо $\bar{\gamma}=\gamma$. Таким образом, сопряженная матрица A^{*} получается из исходной матрицы A путем транспонирования последней и перехода к комплексно-сопряженным величинам.

Эрмитовой матрицей называется квадратная матрица H, совпадающая со своей сопряженной, т. е. $H=H^{*}$.

Кососимметрической матрицей называется квадратная матрица k, отличающаяся множителем (-1) от своей транспонированной, т. е. $k'=-k$. Диагональные элементы кососимметрической матрицы равны нулю, а симметричные относительно главной диагонали — отличаются множителем (-1).

Транспонированная матрица. Пусть

$$A=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{s1} & a_{s2} & \dots & a_{sn} \end{pmatrix}.$$

Транспонированной матрицей A' называется матрица

$$A' = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{s1} \\ a_{12} & a_{22} & \dots & a_{s2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{sn} \end{pmatrix},$$

строки которой совпадают с соответствующими столбцами матрицы A.

Справедливы следующие соотношения:

$$(A+B)' = A' + B',$$
$$(\alpha A)' = \alpha A',$$
$$(AB)' = B'A'.$$

Для неособенной матрицы A

$$(A^{-1})' = (A')^{-1}.$$

Квадратная матрица называется верхней (нижней) треугольной матрицей, если все ее элементы, расположенные ниже (выше) главной диагонали, равны нулю:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \dots & \dots \\ & 0 & & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} a_{11} & & & 0 \\ a_{21} & a_{22} & & \\ \dots & \dots & \dots & \dots \\ a_{s1} & a_{s2} & \dots & a_{sn} \end{pmatrix}.$$

Характеристической матрицей матрицы A называется матрица $A-\lambda E$:

$$A-\lambda E = \begin{pmatrix} \alpha_{11}-\lambda & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22}-\lambda & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn}-\lambda \end{pmatrix},$$

где $A = (\alpha_{ij})$ — квадратная матрица порядка n с действительными элементами; λ — некоторое неизвестное; E — единичная матрица n-го порядка.

Матрицы A и B называются подобными, если найдется такая невырожденная матрица Q, что будет выполнено соотношение

$$B = Q^{-1}AQ.$$

Подобные матрицы имеют одинаковые характеристические многочлены и характеристические корни.

Скелетное разложение матрицы $A = \{a_{ij}\}$ ранга r и размеров $m \times n$ называется представление A в виде произведения двух матриц B и C с размерами соответственно $m \times r$ и $r \times n$:

$$\underset{[m\times n]}{A} = \underset{[m\times r]}{B}\,\underset{[r\times n]}{C}.$$

Ранги B и C также равны r.

Для получения ранга A достаточно в качестве столбца матрицы B взять любые r линейно независимых столбцов матрицы A.

Сложение матриц и умножение на число. Суммой $A + B$ двух матриц $A = (a_{ij})$ и $B = (b_{ij})$, $i = 1, 2, \dots, s$; $j = 1, 2, \dots, n$, имеющих одинаковые размеры, называется матрица $C = (C_{ij})$ того же размера, каждый элемент которой равен сумме соответственных элементов матриц A и B:

$$C_{ij} = a_{ij} + b_{ij}.$$

Произведением матрицы A на число k называется матрица kA, получающаяся умножением на k всех элементов матрицы A:

$$kA = (ka_{ij}).$$

Справедливы соотношения:

1) $A + B = B + A$;
2) $(A + B) + C = A + (B + C)$;
3) $k(A + B) = kA + kB$;
4) $(k_1 + k_2)A = k_1A + k_2A$;
5) $(k_1k_2)A = k_1(k_2A)$.

Сложение матриц и их умножение связаны законами дистрибутивности:

$$(A + B)C = AC + BC; \qquad C(A + B) = CA + CB.$$

Умножение матрицы на число и умножение матриц связаны соотношениями:

$$k(AB) = (kA)B = A(kB).$$

Умножение матриц. Пусть даны две матрицы:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{s1} & a_{s2} & \dots & a_{sn} \end{pmatrix};$$

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix}.$$

Тогда матрица $C = \{C_{ij}\}$, составленная из элементов

$$C_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \qquad i = 1, 2, \ldots, s; \qquad j = 1, 2, \ldots, m,$$

называется произведением матрицы A на матрицу B, а соответствующая операция называется операцией умножения матриц: $C = AB$.

Умножение матриц некоммутативно, т. е. результат зависит от порядка сомножителей.

Умножение матриц ассоциативно $(AB)\,C = A\,(BC)$.

Определитель произведения нескольких квадратных матриц одного порядка равен произведению определителей этих матриц.

Ранг произведения матриц не превышает ранга каждого из сомножителей.

Ранг произведения любой матрицы A справа или слева на невырожденную квадратную матрицу Q равен рангу матрицы A.

Минор. Пусть дан определитель d порядка n. Берем целое число k, удовлетворяющее условию $1 \leqslant k \leqslant n-1$, и в определителе d выбираем произвольные k строк и k столбцов. Элементы, стоящие на пересечении этих строк и столбцов, составляют матрицу порядка k. Определитель этой матрицы называется минором k-го порядка определителя d.

Другими словами, минор k-го порядка есть определитель, получающийся после вычеркивания в определителе d $n-k$ строк и $n-k$ столбцов.

Обобщим на случай прямоугольных матриц понятие минора. Выбираем в матрице A произвольные k строк и k столбцов, $k \leqslant \min(s, n)$. Элементы, стоящие на пересечении этих строк и столбцов, составляют квадратную матрицу k-го порядка, определитель которой называется минором k-го порядка матрицы A.

Наивысший порядок отличных от нуля миноров матрицы A равен рангу этой матрицы.

Дополнительный минор. Пусть в определителе d n-го порядка взят минор M k-го порядка. Если мы вычеркнем те строки и столбцы, на пересечении которых стоит этот минор, то останется минор M' $(n-k)$-го порядка, который называется дополнительным минором для минора M.

Определитель. Определителем n-го порядка, соответствующим матрице

$$A = \begin{Vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{Vmatrix},$$

называется алгебраическая сумма n' членов, составленная следующим образом: членами служат всевозможные произведения n элементов матрицы, взятых по одному в каждой строке и в каждом столбце, причем член берется со знаком плюс, если его индексы составляют четную подстановку, и со знаком минус,— если нечетную.

Для записи определителя n-го порядка, соответствующего матрице A, употребляются символы:

$$d,\ \det A \text{ или } \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

Выражение $d = a_{i1}A_{i1} + \ldots + a_{in}A_{in}$, где A_{ij} — алгебраическое дополнение элемента a_{ji}, называется разложением определителя d по i-ой строке $i = 1, 2, \ldots, n$, т. е. определитель d равен сумме произведений всех элементов произвольной его строки на их алгебраические дополнения.

Определитель характеризуется следующими свойствами.

1. Определитель не меняется при транспонировании.
2. От перестановки двух строк определитель лишь меняет знак.
3. Если все элементы некоторой строки определителя умножить на некоторое число k, то сам определитель умножится на k.
4. Если все элементы i-й строки определителя n-го порядка представлены в виде суммы двух слагаемых

$$a_{ij} = b_j + C_j, \qquad j = 1, 2, \ldots, n,$$

то определитель равен сумме двух определителей, у которых все строки, кроме i-й, такие же, как и в исходном определителе, а i-я строка в одном из слагаемых состоит из элементов b_j, в другом — из элементов C_j.

5. Если одна из строк определителя есть линейная комбинация его других строк, то определитель равен нулю.
6. Если одна из строк определителя состоит из нулей, то определитель равен нулю.
7. Определитель не меняется, если к элементам одной из его строк прибавляется любая линейная комбинация других строк.

Рангом матрицы A называется максимальное число линейно независимых столбцов (или строк) матрицы A.

Наивысший порядок отличных от нуля миноров матрицы A равен рангу этой матрицы.

При вычислении ранга матрицы следует переходить от миноров меньших порядков k минорам больших порядков. При этом если уже определен минор k-го порядка B, отличный от нуля, то следует вычислить лишь миноры $(k+1)$-го порядка, окаймляющие минор B. Если все они равны нулю, то ранг матрицы равен k.

Максимальное число линейно независимых строк любой матрицы равно максимальному числу ее линейно независимых столбцов.

Элементарные преобразования не меняют ранга матрицы.

Для нахождения ранга матрицы следует элементарными преобразованиями привести ее к диагональной форме и подсчитать число единиц, стоящих на главной диагонали.

Ортогональной базой евклидова пространства E_n называется база этого пространства, которая одновременно является ортогональной системой.

Получить ортогональную базу можно путем применения процесса ортогонализации к произвольной базе евклидова пространства.

Ортогональным преобразованием неизвестных называется линейное преобразование

$$x_i = \sum_{j=1}^{n} q_{ij} y_j, \qquad i = 1, 2, \ldots, n,$$

которое оставляет неизменной сумму квадратов неизвестных. Матрица ортогонального преобразования

$$Q = \{q_{ij}\}$$

является ортогональной матрицей.

Ортогональными называются векторы a и b, скалярное произведение которых равно нулю:

$$(a, b) = 0.$$

Система векторов с попарно ортогональными между собой векторами называется ортогональной системой.

Всякая ортогональная система ненулевых векторов линейно независима.

Ортонормированной называется база l_1, l_2, l_n евклидова пространства E_n с попарно ортогональными и нормированными векторами:

$$(l_i, l_j) = 0, \qquad i \neq j;$$
$$(l_i, l_j) = 1, \qquad i = 1, 2, \ldots, n.$$

ОБОБЩЕННЫЕ ОБРАТНЫЕ МАТРИЦЫ

Обобщенная обратная матрица для комплексной прямоугольной $m \times n$ матрицы A определяется как единственная матрица A^+, удовлетворяющая четырем условиям:
$AA^+A = A$; $A^+AA^+ = A^+$; $AA^+ = (AA^+)^*$; $A^+A = (A^+A)^*$.

Здесь «*» означает переход к комплексно-сопряженной транспонированной матрице.

Если A — квадратная невырожденная матрица, то $A^+ = A^{-1}$.

Известно, что для любой прямоугольной матрицы A существует так называемое сингулярное разложение

$$A = v \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} u^*,$$

где v и u — унитарные матрицы; D — диагональная матрица.

Столбцы матрицы u — собственные векторы матрицы A^+A, столбцы матрицы v — собственные векторы матрицы AA^+. Если известно сингулярное разложение матрицы A, то

$$A^+ = u \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} v^*.$$

Обобщенная обратная матрица и каждое из условий 1, 2, 3 и 4 имеют простую геометрическую интерпретацию. Пусть A — матрица линейного оператора из n-мерного унитарного пространства N в m-мерное унитарное пространство M по отношению к фиксированным ортонормальным базисам. Пусть P с N — ядро этого оператора (т. е. совокупность элементов N, удовлетворяющих условию $A_x = 0$), а Q с M — его образ (т. е. совокупность элементов M, имеющих представление $y = Ax$, $x \in N$, $y \in M$). Рассмотрим матрицу X, удовлетворяющую условию $AXA = A$. Из этого условия следует, что $(AX)^2 = AX$; $(XA)^2 = XA$, следовательно, AX и XA — матрицы операторов проектирования: $AX : M \to Q$.

Обозначим через T подпространство пространства M, параллельно которому осуществляется проектирование. Матрица XA аннулирует векторы из подпространства P, потому что XA есть проектор пространства N на некоторое подпространство S параллельно P. Легко проверить, что если $y \in Q$, то $X_y \in S$ и $AX_y = y$, а если $z \in S$, то $Az \in Q$ и $XAz = z$, так что A и X осуществляют взаимно обратные линейные отображения S на Q и Q на S. Таким образом, условие $AXA = A$ вполне определяет действие матрицы X на векторы из Q. Для определения X на всем пространстве M нужно доопределить X на подпространстве T. На T оператор с матрицей X может действовать произвольным образом, только его значения должны попадать в P, так как при $y \in TAXy = 0$.

Если, кроме того, выполнено условие $XAX = X$, то при $y \in TXy = XAXy = 0$, т. е. X доопределен на T нулем. Тем самым X определен однозначно, если только S и T даны. Условия $AA^+ = (AA^+)^*$ м $A^+A = (A^+A)^*$ обозначают ортогональность проекторов соответственно AX и XA; первое из этих условий равносильно ортогональности T и R, второе — ортогональности S и P.

Если $X = A^+$, то подпространства S и T ортогональны к P и Q соответственно, A^+ осуществляет отображение Q на S, обратное отображению S на Q, определенному матрицей A, причем A^+ доопределяется на T нулем.

Если матрица A имеет полный ранг, т. е. $r = m$ или $r = n$, то геометрическая картина упрощается. При $k = m$ образ есть все пространство N, так что T состоит только из нуля и ортогональный проектор XA есть единичная матрица. Условия $XAX = X$ и $(XA)^* = XA$ в этом случае выполняются автоматически, так что матрица A^+ характеризуется лишь условиями $AA^+ = A$ и $AA^+ = (AA^+)^*$. При $k = n$ ядро A состоит только из Q, так что проектор AX есть единичная матрица.

Условия $XAX = X$ и $(AX)^* = AX$ выполняются автоматически, и A характеризуется условиями 1 и 4.

Понятие обобщенной обратной матрицы было введено Муром в 1920 г. Обобщенные обратные матрицы используются при решении несовместных систем линейных алгебраических уравнений, часто возникающих в практических приложениях.

Рассмотрим в общем случае несовместную систему линейных уравнений:

$$Ax = y,$$

где A — матрица известных значений размеров $[m \times n]$; y — вектор известных значений размерности m; x — вектор неизвестных значений размерности n.

Тогда решение x^0 системы $Ax = y$, полученное методом наименьших квадратов (МНК), т. е. из условия обращения в минимум выражения

$$\sum_{i=1}^{m}\left(y_i - \sum_{k=1}^{n} a_{ik}x_k\right)^2,$$

и обладающее минимальной длиной $\sum_{i=1}^{n} x_i^2$

среди всех МНК-решений, определяется выражением $x^0 = A^+y$.

Обобщение последнего результата на матричный случай. Рассмотрим матричное уравнение

$$AX = Y,$$

где A, Y — заданные матрицы размеров $m \times n$ и $m \times p$ соответственно, X — искомая матрица размеров $n \times p$. Норму матрицы A определим так:

$$\|A\|^2 = \sum_{i,k} a_{ik}^2,$$

Тогда решение матричного уравнения $AX = Y$, полученное методом наименьших квадратов из условия обращения в минимум выражения $\|Y - AX\|$ и обладающее минимальной нормой $\|X\|^2$ среди всех таких МНК-решений, определяется выражением $X^0 = A^+y$.

Если $Y = E$ (единичная матрица), то $X^0 = A^+$.

Для нахождения A^+ может быть использован последовательный метод Гревиля.

ПЛОХО ОБУСЛОВЛЕННАЯ СИСТЕМА ЛИНЕЙНЫХ УРАВНЕНИЙ

Плохо обусловленная система линейных алгебраических уравнений — это система, малое изменение исходных данных которой приводит к большому изменению решения.

Необходимость решать плохо обусловленные системы линейных уравнений часто возникает в приложениях. Существует ряд способов, позволяющих отыскивать решение такой системы: применить

преобразование системы, повышающее ее число обусловленности (масштабирование, приближенную ортогонализацию и т. п.), отбросить часть уравнений, линейно зависящих от остальных, использовать дополнительную информацию о возможном решении (например, его положительность, ограниченность и другие сведения). Часто при этом вместо решения получаются лишь интервальные оценки для них. Как правило, решение плохо обусловленной системы не единственно, но среди них имеется нормальное решение с наименьшей длиной. Наиболее удобен при решении плохо обусловленных систем метод окаймления [33]. Очень эффективный метод интервального оценивания решений плохо обусловленной системы дан в работе [33].

Число обусловленности — это характеристика системы линейных алгебраических уравнений, описывающая чувствительность системы по отношению к малому изменению исходных данных.

Остановимся на тех факторах, которые определяют чувствительность по отношению к изменению элементов A и b решения системы уравнений $Ax = b$, где A — $n \times n$-матрица; x—n-мерный вектор-столбец неизвестных; b — заданная правая часть. Предположим, что правая часть системы b определена с ошибкой δb. Считая, что при этом вектор-столбец решения системы определяется с ошибкой δx, получаем систему уравнений $A(x + \delta x) = b + \delta b$, которая вместе с заданной системой $Ax = b$ дает после вычитания их друг из друга $\delta x = A^{-1}\delta b$. Отсюда вытекает неравенство

$$\frac{\|\delta x\|}{\|\delta b\|} \leqslant \|A^{-1}\|,$$

где $\|A\|$ — норма матрицы A, равная

$$\sup_{X \neq 0} \frac{\|AX\|}{\|X\|},$$

$$\|x\| = \sqrt{X_1^2 + \ldots + X_n^2}.$$

Рассмотрим относительную ошибку в определении вектора x:

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\|A^{-1}\|}{\|A\|^{-1}} \frac{\|\delta b\|}{\|b\|} = \frac{\|A\|\|A^{-1}\|\|\delta x\|}{\|b\|}.$$

Число $\|A\|A^{-1}\|$ можно рассматривать как число обусловленности задачи. Если это число очень велико, то задача плохо обусловлена для большинства b и δb, т. е. для большинства b и δb будем иметь $\|X\| \gg \|A\|^{-1}\|b\|$.

Оценим теперь относительную ошибку решения системы уравнений при возмущении матрицы A. Получим $(A + \delta A)(X + \delta x) = b$, откуда выводим

$$(A + \delta A)\delta x = -\delta A \cdot x.$$

Даже если A — невырожденная матрица, то $A + \delta A$ может быть вырожденной (определитель ее может быть близок к нулю).

Во избежание этого явления наложим ограничение $\| A^{-1} \cdot \delta A \| < 1$. Тогда $A + \delta A = A (E + A^{-1}\delta A)$.
Теперь мы имеем

$$\delta x = -(E + A^{-1}\delta A)^{-1} A^{-1} \delta A \cdot x$$

и далее

$$\| \delta x \| \leqslant \frac{\| A^{-1} \cdot \delta A \| \| x \|}{1 - \| A^{-1} \cdot \delta A \|} \leqslant \frac{\| A^{-1} \| \| \delta A \| \| x \|}{1 - \| A^{-1} \| \| \delta A \|},$$

если выполнено условие $\| A^{-1} \| \| \delta A \| < 1$. Оценим относительную ошибку:

$$\frac{\| \delta x \|}{\| x \|} \leqslant \| A \| \| A^{-1} \| \frac{\| \delta A \|}{\| A \|} \frac{1}{[1 - \| A \| \| A^{-1} \| (\| \delta A \| / \| A \|)]}.$$

Мы видим, что решающую роль в оценке обусловленности матрицы играет число $\| A \| \| A^{-1} \|$. Можно использовать другую норму $\| A \|_2$, равную максимальному собственному значению матрицы A.

Число $k (A) = \| A \|_2 \| A^{-1} \|_2$ назовем спектральным числом обусловленности задачи. Заметим, что число обусловленности не меняется при ортогональных преобразованиях матрицы A (ортогональным называется линейное преобразование $y = Ax$, матрица A которого удовлетворяет условию $AA^T = E$),

$$\| A \|_2 = \sqrt{\max_{i,} \sum_{i=1}^{n} a_{ij}^2}.$$

Для симметричной матрицы A с собственными значениями λ_i, причем $\lambda_{i-1} \geqslant \lambda_i$ $(i = 2, 3, \ldots, n)$, спектральное число $k (A)$

$$k (A) = | \lambda_1 | / | \lambda_n |.$$

Если нормировать систему $Ax = b$ так, чтобы λ_1 равнялось единице, то будем иметь

$$k (A) = 1 / | \lambda_n |.$$

Следовательно, симметричная нормированная матрица плохо обусловлена лишь тогда, когда собственное число λ_n мало.

Приведем еще ряд выражений для чисел обусловленности. Все они получаются при оценке ошибок округления при обращении матриц в качестве множителя. Таково, например, число Тюринга

$$\frac{1}{n} N (A) N (A)^{-1},$$

где $N (A) = \sqrt{Tr (A^T A)}$, $Tr (A^T A)$ — след матрицы $A^T A$, т. е. сумма ее диагональных элементов или ее собственных значений, или число обусловленности вида

$$nM (A) M (A^{-1}),$$

где

$$M (A) = \max_{i, j} | a_{i, j} |.$$

Очень удобно использовать в качестве числа обусловленности величину $\delta (A) = \det (A^N)$, где A^N — нормированная матрица с элементами

$$a_{ij}^N = a_{ij} / \| A \|_2, \qquad i, j = 1, \ldots, n,$$

$\det A^N$ — определитель матрицы A^N.

В связи с тем что нахождение собственных значений матрицы является очень трудоемкой задачей, целесообразно иметь оценку числа обусловленности матрицы через ее элементы. Одна из таких оценок, например, имеет вид

$$k \geqslant \max \left(\sqrt{\frac{\max \sum_{j=1}^{n} a_{ij}^2}{\min \sum_{j=1}^{n} a_{ij}^2}}, \sqrt{\frac{\max \sum_{i=1}^{n} a_{ij}^2}{\min \sum_{i=1}^{n} a_{ij}^2}} \right).$$

Это неравенство показывает, что большое различие в суммах квадратов элементов матрицы по строкам или по столбцам характеризует ее плохую обусловленность. Иногда целесообразно перед решением системы уменьшить указанное различие путем умножения уравнений системы на некоторые множители или путем введения некоторых масштабных множителей в неизвестные. Однако для плохой обусловленности матрицы указанное различие необязательно — это лишь достаточное условие. Такое явление, как значительное превышение по абсолютной величине элементов строки или столбца матрицы над элементами других строк, довольно часто встречается в приложениях. Прежде чем решать систему линейных уравнений с такими данными, необходимо предварительно преобразовать ее.

Прием, позволяющий в некотором смысле уравновесить значения элементов различных матриц, называется масштабированием. Под масштабированием понимается переход от системы $Ax = b$ к эквивалентной системе $D_1 A D_2 y = D_1 b$, где $x = D_2 y$, а D_1 и D_2 — две диагональные матрицы с положительными элементами.

Для случая идеально обусловленной системы уравнений $\delta (A)$ равно единице. Сравнение $\delta (A)$ с единицей характеризует не только обусловленность системы, но и возможную потерю точности при решении системы уравнений. Метод приближенной ортогонализации приводит к замене исходной плохо обусловленной системы другой, число обусловленности которой близко к единице. Эта система уже решается с помощью стандартных приемов линейной алгебры. Возвращаясь обратными преобразованиями к исходной системе, получаем ее решение. Метод приближенной ортогонализации матриц разработан Е. И. Филипповичем [43].

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. М., Статистика, 1974.
2. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Исследование зависимостей. М., Финансы и статистика, 1985.
3. *Андерсон Т.* Введение в многомерный статистический анализ. М., Физматиздат, 1963.
4. *Андерсон Т.* Статистический анализ временных рядов. М., Мир, 1976.
5. *Аронов В. И.* Методы математической обработки геологических данных на ЭВМ. М., Недра, 1977.
6. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. М., Мир, 1982.
7. *Белонин М. Д., Голубева В. А., Скублов Г. Т.* Факторный анализ в геологии. М., Недра, 1982.
8. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., Наука, 1983.
9. *Бугаец А. Н., Дуденко Л. Н.* Математические методы при прогнозировании месторождений полезных ископаемых. Л., Недра, 1976.
10. *Бухштабер В. М., Маслов В. К.* Факторный анализ и экстремальные задачи на многообразиях Грассмана.— В кн.: Математические методы решения экономических задач. М., 1977, с. 85—102.
11. *Вальд А.* Последовательный анализ. М., Физматиздат, 1960.
12. *Вистелиус А. Б.* Основы математической геологии. Л., Наука, 1980.
13. *Волков А. М.* Решение практических задач геологии на ЭВМ. М., Недра, 1980.
14. *Гиндикин С. Г.* Алгебра логики в задачах. М., Наука, 1972.
15. *Гриффитс Дж.* Научные методы исследования осадочных пород. М., Мир, 1971.
16. *Давид М.* Геостатические методы при оценке запасов руд. Л., Недра, 1980.
17. *Дмитриев А. Н.* Новые тестовые разработки в задачах прогнозирования рудоносности (на примере трапповых интрузий).— В кн.: Математические методы при прогнозе рудоносности. М., 1977, с. 104—163.
18. *Дюран Б., Оделл П.* Кластерный анализ. М., Статистика, 1977.
19. *Закс Л.* Статистическое оценивание. М., Статистика, 1976.
20. *Информационные* системы общего назначения. Пер. с англ. под ред. В. Л. Ющенко. М., Статистика, 1970.
21. *Йереског К.Г., Клован Д. И., Реймент Р. А.* Геологический факторный анализ Л., Недра, 1980.
22. *Кендалл М.Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды. М., Наука, 1976.
23. *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи. М., Наука, 1973.
24. *Классификация и кластер*/Под ред. Дж. Вен Райзина. М., Мир, 1980.
25. *Коган Р. И.* Интервальные оценки в геологических исследованиях. М., Недра, 1986.
26. *Коган Р. И., Белов Ю. П., Родионов Д. А.* Статистические ранговые критерии в геологии. М., Недра, 1983.
27. *Кондаков Н. И.* Введение в логику. М., Наука, 1967.
28. *Кульбак С.* Теория информации и статистика. М., Изд-во иностр. лит., 1960.
29. *Леман Э.* Проверка статистических гипотез. М., Наука, 1979.
30. *Марголин А. М.* Методы геометризации разведуемых запасов полезных ископаемых. Усовершенствованная процедура крайгинга. М., 1983 (ВИЭМС).
31. *Мардиа К.* Статистический анализ угловых наблюдений. М., Наука, 1978.
32. *Мартин Дж.* Организация баз данных в вычислительных системах. Пер. с англ. под ред. А. Л. Щерса. М., Мир, 1978.
33. *Матвеев Л. А.* Об одном алгоритме псевдообращения матриц.— Ж. вычисл. математики и матем. физики, 1974, т. 14, № 2.
34. *Математические* методы решения прогнозных задач нефтяной геологии. Новосибирск, 1978, с. 36—77.
35. *Матерон Ж.* Основы прикладной геостатистики. М., Мир, 1968.
36. *Новиков П. С.* Элементы математической логики. М., Наука, 1973.
37. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей. Основное понятие. Предельные теоремы. Случайные процессы. М., Наука, 1967.
38. *Распознавание* образов в задачах качественного прогноза рудных месторождений. Новосибирск, Наука, 1980.
39. *Родионов Д. А.* Статистические решения в геологии. М., Недра, 1981.
40. *Себер Дж.* Линейный регрессионный анализ. М., Мир, 1980.
41. *Сиротинская С. В.* Логические методы анализа геологической информации. М., Недра, 1985.
42. *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М., Мир, 1978.
43. *Фадеев Д. К., Фадеева В. Н.* Вычислительные методы линейной алгебры.— Зап. науч. семинаров ЛОМИ, 1975, т. 54.
44. *Философский* энциклопедический словарь. М., Советская энциклопедия, 1983.
45. *Харбух Дж., Бонэм-Картер Г.* Моделирование ЭВМ в геологии. М., Мир, 1974.
46. *Шеффе Г.* Дисперсионный анализ. Изд. 2-е. М., Наука, 1980.
47. *Энциклопедия* кибернетики/Под ред. В. М. Глушкова. Киев, 1974, (Главная редакция украинской советской энциклопедии).
48. *Яблонский С. В.* Введение в дискретную математику. М., Наука, 1979.
49. *Burrus W., Rust B. W., Cope J. E.* Constrained interval estimatiens for linear models with ill — conditioned equations.— Information Lincage Appl. Math. anb Ind. «2 Proc. Annu. Workshop», Monterey, Calif., 1980, p. 1—38.
50. *Journel A., Huijbrechts Ch.* Mining Geostatistics. Academic Press, 1978.
51. *Puri M. L., Sen P. K.* Nonparametric methods in multivariate analysis. N.—Y.— London—Sydney—Toronto, J. Wiley and sons, 1971.
52. *Rao C. K., Mitra S. K.* Generalized inverse of matrices and its applications. N.—Y.— London—Sydney—Toronto, J. Wiley and sons, 1971.

ОГЛАВЛЕНИЕ

СПРАВОЧНИК СПЕЦИАЛИСТА

Дмитрий Алексеевич Родионов, Роберт Иосифович Коган
Валентина Алексеевна Голубева, Борис Иванович Смирнов
Сусанна Валерьяновна Сиротинская

СПРАВОЧНИК ПО МАТЕМАТИЧЕСКИМ МЕТОДАМ В ГЕОЛОГИИ

Редактор издательства *А. И. Федотова*
Технические редакторы *Е. В. Воробьева, О. А. Колотвина*
Корректоры *М. Е. Лукина, В. Т. Юдович*

ИБ № 5698

Сдано в набор 26.11.86. Подписано в печать 13.04.87. Т-11006. Формат 60 × 90 1/16. Бумага типографская № 1. Гарнитура Литературная. Печать высокая. Усл.-печ. л. 21,0. Усл. кр.-отт. 21,0. Уч.-изд. л. 21,28. Тираж 8600 экз. Заказ 3002/200—1. Цена 1 р. 60 к.

Ордена «Знак Почета» издательство «Недра»,
125047, Москва, пл. Белорусского вокзала, 3.

Ленинградская типография № 4 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли.
191126, Ленинград, Социалистическая ул., 14.