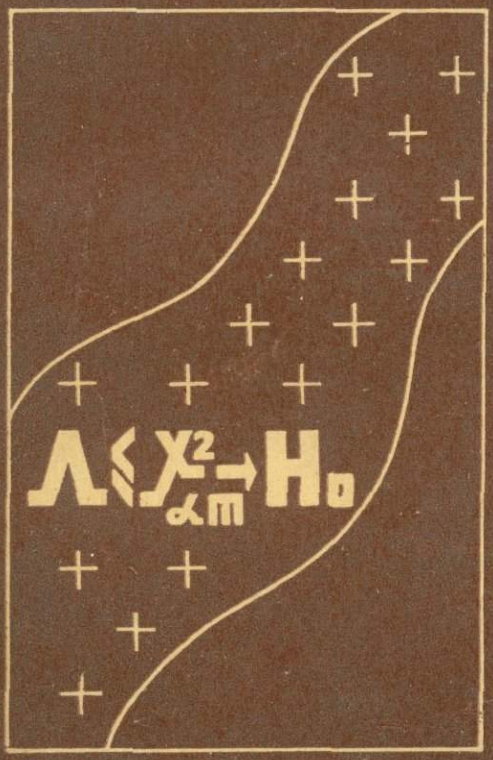
Р. И. КОГАН, Ю. П. БЕЛОВ, Д. А. РОДИОНОВ

СТАТИСТИЧЕСКИЕ РАНГОВЫЕ КРИТЕРИИ В ГЕОЛОГИИ



Р. И. КОГАН, Ю. П. БЕЛОВ, Д. А. РОДИОНОВ

СТАТИСТИЧЕСКИЕ РАНГОВЫЕ КРИТЕРИИ В ГЕОЛОГИИ

МОСКВА «НЕДРА» 1983

УДК 519.2 : 55

Коган Р. И., Белов Ю. П., Родионов Д. А. Статистические ранговые критерии в геологии. М., Недра, 1983. 136 с.

Проанализирован и обобщен материал по применению в геологии математических методов, основанных на статистических ранговых критериях. Описаны наиболее важные и актуальные методы классификации геологических объектов, корреляционного и регрессионного анализа, выделения информативнойкомбинации признаков. Предложены оптимальные сочетания ранговых и параметрических критериев для получения надежных выводов и их обоснованной геологической интерпретации. Даны примеры практического использования рекомендуемых методов для решения геологических задач.

Для геологов и других специалистов геологических организаций различного профиля, применяющих в своих исследованиях статистические методы.

Табл. 16, ил. 20, список лит. — 49 назв.

Рецензент — д-р геол.-минер. наук *В. Н. Бондаренко* (ИМГРЭ)

К $\frac{1904010000-235}{043(01)-83}$ 70—83

ПРЕДИСЛОВИЕ

В «Основных направлениях экономического и социального развития СССР на 1981—1985 годы и на период до 1990 года», утвержденных XXVI съездом КПСС, указаны важнейшие проблемы в области естественных и технических наук, на решении которых необходимо сосредоточить усилия в первую очередь. К числу таких проблем относится развитие математической теории, повышение эффективности ее использования в прикладных целях. В связи с этим актуальными задачами при геологических исследованиях являются широкое внедрение математических методов и ЭВМ в процессе получения и обработки геологической информации, разработка теоретических и практических вопросов применения математических методов и повышение их эффективности при решении конкретных геологических задач.

Актуальность работы по статистическим ранговым критериям в геологии обусловлена необходимостью использования в геологических исследованиях устойчивых статистических процедур обработки геологической информации с последующей интерпретацией статистических выводов.

При написании книги авторы ставили себе целью дать геологам представление об аппарате ранговой статистики в объеме, необходимом для самостоятельного его применения. Поэтому в книге дан обзор математических методов решения задач прогноза, поисков и разведки рудных месторождений, приведены общие сведения о классах математических методов и формальные модели геологических задач, описаны ранговые и параметрические методы классификации геологических объектов, ранговые и параметрические алгоритмы корреляционного и регрессионного анализа, а также выделения информативных комбинаций признаков.

Предисловие, введение, главы 1 и 2 написаны Р. И. Коганом и Д. А. Родионовым, глава 3 — Ю. П. Беловым, им же сделано большинство расчетов; глава 4 написана авторами совместно. Общее редактирование рукописи осуществлено проф. Д. А. Родионовым.

Авторы выражают признательность за помощь в исследованиях и при подготовке рукописи книги сотрудникам лаборатории математических методов в геологии ВИЭМС Т. Ф. Митрофановой, Н. В. Тимофеевой, В. Ю. Раевскому, В. А. Голубевой, а также Т. М. Злобиной, Е. А. Озерецковской, Т. А. Барской, Е. Г. Краевой, Т. Ю. Спивак, Л. В. Воробьевой, Г. А. Зыбиной. За ценные замечания, которые позволили улучшить содержание и изложение материала в книге, авторы особенно благодарны доктору геол.-минер. наук В. Н. Бондаренко.

Все замечания и пожелания по поводу содержания книги будут приняты авторами с признательностью.

ВВЕДЕНИЕ

Широкое применение статистических ранговых критериев и процедур обработки геологической информации становится все более актуальным в геологических исследованиях. Дело в том, что традиционные, классические параметрические статистические методы обработки геологических данных базируются на априорном, т. е. доопытном, знании вида функций распределения случайных величин, рассматриваемых в качестве математических моделей изучаемых геологических характеристик. В то же время более типичны ситуации, когда практически ничего не известно о виде функции распределения случайных величин, являющихся математическими моделями изучаемых геологических характеристик. Решению именно таких задач отвечают методы непараметрической статистики и ее важный раздел — статистические ранговые критерии.

Геологи-прикладники математическую статистику (критерии, алгоритмы, процедуры) часто рассматривают как систему, внутри которой совершенно неважно, что и как делается, но на выходе которой имеются различные статистические процедуры и инструкции по их использованию.

Геологическая практика все чаще выдвигает задачи, которые не укладываются в требования этих инструкций. Возникает сигнал «рассогласования», требующий развития самой математической статистики [43, с. 7].

Большая часть результатов классической параметрической математической статистики основана на предположении, что имеющейся у потребителя (геолога) информации достаточно для представления участвующих в задаче распределений случайных величин в виде некоторых функций с конечным числом параметров (очень часто функций нормального распределения). В геологической практике это априорное предположение нередко не выполняется.

Потребности в создании статистических процедур, не предполагающих знание вида функций распределения случайных величин, и отвечает специальная ветвь математической статистики — непараметрическая статистика, содержащая порядковые и ранговые статистические критерии [43, с. 8].

Ранговая статистика в самой исходной модели рассматривает только такие ситуации, в которых о функциональном виде распределений случайных величин ничего не известно. Единственной доступной потребителю-геологу априорной информацией является информация о характере случайных величин (непрерывны они или дискретны, некоторые самые общие предположения о функциях распределения типа: гладкость, сосредоточенность значений случайной величины внутри ограниченной области и т. п.), о типе различия между распределениями. Сами распределения для ста-

тистических ранговых критериев может быть какими угодно, геолога может интересовать лишь, сдвинуты ли они относительно друг друга (отличаются ли средними значениями), различаются ли они дисперсиями, т. е. характеристиками рассеяния вокруг средних, присутствует или отсутствует симметрия и т. п. Кроме того, ранговые критерии, как отмечал Л. Закс [18, с. 264], не требуют громоздких вычислений.

По мнению Ф. П. Тарасенко [43, с. 8], интерес практиков и теоретиков к непараметрической статистике объясняется не только тем, что она исходит из более широкой и реалистической модели, чем классические параметрические ветви математической статистики, не только тем, что получаемые ею процедуры очень просты в реализации и поэтому трудные математические проблемы часто решаются изящно и красиво, а тем, что непараметрические, включая ранговые, процедуры обладают высокой эффективностью, а именно:

потери эффективности при переходе от параметрических к непараметрическим процедурам (в случае истинности параметрической модели) незначительны;

эффективность непараметрической процедуры по сравнению с эффективностью фиксированной параметрической процедуры возрастает при отклонении истинных распределений от теоретических;

во многих случаях непараметрические процедуры оказываются асимптотически оптимальными.

Главным источником этих положительных качеств является то, что при непараметрическом подходе используются весьма общие, доступные прикладнику предположения. Предположение же о нормальности распределения тогда, когда оно на самом деле может быть не только не нормальным, а вообще полимодальным, в некотором смысле эквивалентно введению дополнительных случайных мешающих факторов.

По нашему мнению, основная роль статистических ранговых критериев — служить одним из средств проверки обоснованности заключений, которые в настоящее время обычно делаются лишь на основе параметрических статистических методов. Именно сочетание параметрических и ранговых критериев позволяет получать надежные статистические выводы и соответственно осуществлять их обоснованную геологическую интерпретацию. Одновременное принятие нулевой гипотезы или ее отклонение принципиально различными параметрическими и ранговыми тестами обусловливает надежное статистическое заключение. Расхождение в принятии или отклонении нулевой гипотезы с помощью параметрических и ранговых критериев — это сигнал о возможной необоснованности статистического заключения об изучаемых параметрах и соответственно последующей геологической интерпретации этих выводов. В данной работе авторы наряду с подробным разбором ранговых статистических критериев регулярно будут привлекать и их параметрические аналоги.

ГЛАВА I

РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ, ПРИМЕНЯЕМЫЕ ДЛЯ РЕШЕНИЯ ТИПОВЫХ ГЕОЛОГИЧЕСКИХ ЗАДАЧ

1.1. ОБЗОР МАТЕМАТИЧЕСКИХ МЕТОДОВ РЕШЕНИЯ ЗАДАЧ ПРОГНОЗА, ПОИСКОВ И РАЗВЕДКИ МЕСТОРОЖДЕНИЙ

В ряде обзорных работ [39, 37, 5, 24] обобщены сведения о математических методах, алгоритмах, применяемых при решении прогнозных, поисковых и геологоразведочных задач.

В автоматизированных системах, пакетах прикладных программ, а также во многих рядовых программах, разработанных в геологической отрасли, реализованы алгоритмы большинства разделов современной математики: теории вероятностей, математической статистики, математической логики, теории множеств, комбинаторного анализа, теории графов, линейной и нелинейной алгебры, топологии и др.

Нередко для решения типовых прогнозных, поисковых, геологоразведочных и геолого-экономических задач привлекаются одни и те же математические методы.

Среди множеств математических методов нами выделены 10 основных направлений (классов). Принципиальная структура этой классификационной схемы с детальным перечислением самих математических методов и оценкой их перспективности рассмотрена в ранее опубликованных работах [39, 21]. Схема эта не является универсальной или всеобъемлющей и не претендует на строгую логическую выдержанность принципов группирования критериев и алгоритмов. Но она хотя бы в частичном виде представляется полезной разработчикам автоматизированных систем обработки геологической информации.

Обсуждаемая схема классов алгоритмов выглядит следующим образом.

1-й класс. Алгоритмы описания геологических объектов (подбор моделей распределения изучаемых геологических характеристик, оценивание параметров и их погрешностей, проверка статистических гипотез о параметрах и др.).

2-й класс. Алгоритмы корреляционного, регрессионного, дисперсионного и факторного анализа. Возможно, здесь должны рассматриваться и методы планирования экспериментов.

3-й класс. Алгоритмы разграничения на однородные области совокупности многомерных количественных и качественных наблюдений над комплексом геологических свойств, кластер-анализ, таксономия и др. (априорно неизвестно, на сколько классов следует разделить изучаемую совокупность).

4-й класс. Алгоритмы стохастической и эвристической классификации (группирования) геологических объектов, охарактеризованных многомерными количественными и качественными наблюдениями над комплексом геологических свойств, методами многомерной статистики, распознавания образов, автоматической классификации, дискриминантный анализ и др. (априорно заданы все геологические объекты и известна принадлежность каждого многомерного наблюдения к конкретным объектам).

5-й класс. Статистические и эвристические алгоритмы поиска информативных комбинаций признаков (геологических свойств) при сопоставлении геологических объектов, а также алгоритмы снижения размерности пространства при исследовании каждого геологического объекта, охарактеризованного многомерными наблюдениями над комплексом геологических свойств.

6-й класс. Алгоритмы тренд-анализа, статистического анализа временных рядов, методы изучения случайных процессов.

7-й класс. Методы исследования операций, теоретико-игровые методы, адаптивные и неадаптивные методы стохастического программирования, методы экспертных оценок (Делфи, мозговой атаки и др.).

8-й класс. Классические (методы сечений, геологических, эксплуатационных блоков, объемный метод и др.) и нетрадиционные (аппроксимация функцией Лапласа, сплайн-функцией и др.) процедуры подсчета запасов твердых, жидких и горючих полезных ископаемых.

9-й класс. Методы реализации в фактографических и документальных автоматизированных информационно-поисковых системах процедур ввода, корректировки, контроля, поиска, выдачи геологической информации, т. е. реализации процедур ведения и эксплуатации массивов информации.

10-й класс. Вычислительные алгоритмы, предназначенные для повышения устойчивости результатов вычислений, обращения плохо обусловленных матриц и др.

Ф. Агтерберг [44] отмечает, что с применением математических методов геолог обычно сталкивается в пяти областях своей деятельности:

1) при сборе и обработке геологических данных;

2) при анализе геологических данных, связанных с выявлением тренда и зависимостей;

3) при статистическом обосновании решения практических задач сбора геологических данных;

4) при проверке предположений или моделей геологических процессов, включая моделирование на ЭВМ;

5) при количественном предсказании (прогнозировании) в геологических задачах.

Конечно, этот список областей приложения математических методов и ЭВМ неполный. В частности, не упомянута очень важная область автоматизации проведения подсчетов запасов, опти-

мизации кондиций, исследования степени разведанности месторождений и т. п.

В работе А. Б. Вистелиуса [10] охарактеризованы три следующих направления применения математических методов в геологии при решении разнообразных задач как принципиального, так и технического характера:

а) связанное идеей создания собственной математической дисциплины в цикле геологических наук (создание математической геологии);

б) независимое от математической геологии и связанное с информативной, т. е. с «научными» методами хранения, поиска и выдачи документов;

в) связанное с использованием математики для решения задач по подсчету запасов, с применением горной экономики, оптимальных методов разведки.

Из десяти вышеупомянутых классов алгоритмов, применяемых при обработке геологической информации, более подробно рассмотрены те, которые относятся к первым пяти классам, хотя частично затрагиваются математические процедуры 6, 7 и 10-го классов.

В настоящее время обобщение большого опыта обработки эмпирических данных в различных областях знаний (не только в геологии) привело исследователей к выработке новой схемы исследования численной информации, в которой на первый план выдвигается правильная первичная обработка исходных массивов данных. Кроме того, требование повышения надежности статистических выводов и их геологической интерпретируемости приводит к необходимости контроля промежуточных этапов вычислений.

В приводимой принципиальной блок-схеме статистического анализа геологических данных (рис. 1) показаны различные этапы проведения статистических исследований геологоразведочных данных, рекомендуемые при обработке количественной информации по рудным месторождениям полезных ископаемых. Схема охватывает в основном классы математических методов, охарактеризованных в главах 2—4.

1.2. ФОРМАЛЬНЫЕ МАТЕМАТИЧЕСКИЕ МОДЕЛИ ГЕОЛОГИЧЕСКИХ ЗАДАЧ

Как уже отмечалось [44], с применением математических методов геолог сталкивается при сборе и обработке геологической информации, в частности при анализе геологических данных, связанных с выявлением тренда и зависимостей, при количественном прогнозировании и т. д.

При решении прогнозных, поисковых и геологоразведочных задач успешное использование математических методов возможно лишь тогда, когда имеется формальная модель, адекватная изучаемым объектам, в условиях конкретной задачи.

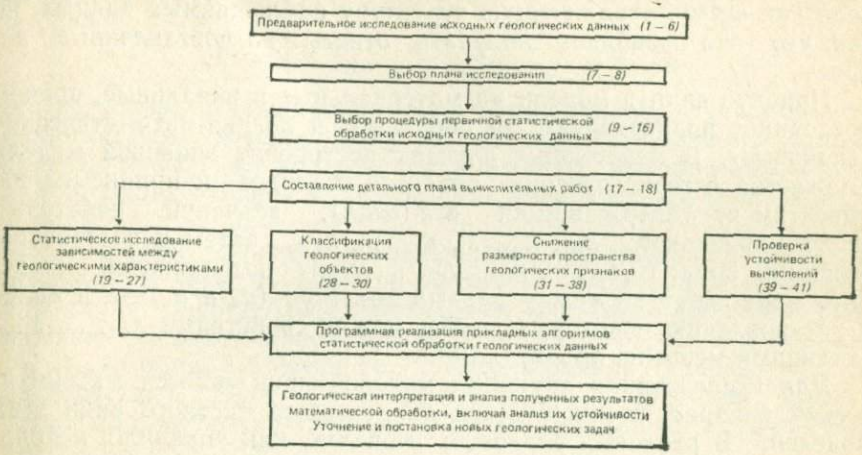


Рис. 1. Принципиальная блок-схема статистического анализа геологических данных:

1 — определение цели исследований на содержательном геологическом уровне; *2* — определение геологических объектов статистических исследований; *3* — определение геологических характеристик (признаков) геологических объектов; *4* — определение объемов и масштаба исследований; *5* — формализация геологической задачи; *6* — фиксирование формы записи геологических данных для их ввода в ЭВМ;
7 — составление детального плана сбора исходных геологических данных; *8* — составление схемы анализа исходных геологических данных;
9 — унификация записи геологических данных, унификация типа переменных; *10* — анализ аномальных геологических данных; *11* — восстановление пропущенных геологических данных; *12* — проверка статистической независимости геологических данных; *13* — оценивание параметров распределений случайных величин — моделей геологических признаков; *14* — анализ однородности геологических данных; *15* — анализ структуры геологических данных с помощью критериев согласия; *16* — проверка согласованности распределений исходных геологических данных с теоретическими моделями распределений;
17 — описание блок-схемы статистического анализа с указанием используемых прикладных алгоритмов обработки геологических данных; *18* — описание критериев качества математической обработки и выбор тестов с наибольшей устойчивостью (robust-процедур);
19 — параметрические и ранговые алгоритмы корреляционного анализа; *20* — параметрические и ранговые алгоритмы регрессионного анализа; *21* — алгоритмы конфлюэнтного анализа; *22* — алгоритмы планирования регрессионных экспериментов; *23* — алгоритмы дисперсионного анализа; *24* — алгоритмы ковариационного анализа; *25* — алгоритмы сплайн-аппроксимации; *26* — алгоритмы анализа временных рядов; *27* — алгоритмы зависимостей специального типа (марковские зависимости и др.);
28 — параметрические и ранговые алгоритмы разграничения геологических объектов; *29* — параметрические и ранговые алгоритмы дискриминантного анализа; *30* — параметрические и ранговые алгоритмы кластерного анализа, включая численную таксономию, автоматическую классификацию, методы группирования и др.;
31 — статистические методы поиска информативных комбинаций геологических признаков; *32* — алгоритмы поиска информационных весов в эвристических процедурах распознавания образов; *33* — алгоритмы определения весов в логико-информационных методах изучения геологических объектов; *34* — алгоритмы поиска наиболее информативных признаков в моделях регрессии; *35* — алгоритмы метода главных компонент; *36* — алгоритмы факторного анализа; *37* — алгоритмы метода многомерного шкалирования; *38* — алгоритмы метода экстремальной группировки параметров;
39 — алгоритмы решения линейных алгебраических систем с прямоугольными матрицами (обратные матрицы Мура и Пенроуза); *40* — алгоритмы приближенной ортогонализации строк матрицы (методы Е. И. Филипповича, А. В. Черного); *41* — алгоритмы построения псевдообратных матриц на основе процедуры окаймления матрицы и построения полуобратного оператора (метод А. А. Матвеева)

Под моделью понимается система, отображающая (воспроизводящая) объект исследования и способная заменить его в процессе исследования. У. Крамбейн [27, с. 8] рассматривает модель как схему, отражающую структуру наблюдаемых данных, но так, что она позволяет получать ответы на поставленные вопросы.

Принято делить модели на материальные и идеальные, причем последние подразделяют на образные и логико-математические (знаковые). В простейших случаях построение знаковой модели множества k геологических объектов сводится к приписыванию объектам u, содержащимся в k ($u \in k$), значений некоторых свойств-признаков: $x_{u1}, x_{u2}, \ldots, x_{uj}, \ldots, x_{um}$, затем к построению функции вида: $F(x_{u1}, \ldots, x_{um}, y_{u1}, \ldots, y_{um})$, позволяющей определять меру сходства между парами объектов (u_x и u_y) $\in k$ и далее к установлению некоторых связей между свойствами $x_{u1}, \ldots, x_{um}$, имеющими место на объектах $u \in k$.

Для исследования знаковых моделей привлекается математический аппарат, специфический для каждого частного вида этих моделей. В ранговых знаковых моделях, как правило, используются специальные функции (ранги, метки) от выбранных исходных признаков.

В качестве математической модели изучаемого набора геологических объектов без учета каких-либо конкретных их свойств (признаков) будем рассматривать множество T точек t ($t \in T$). Множество T рассматривается фиксированным и в соответствии с терминологией теории множеств будет называться пространством.

Обозначим через A_u произвольное множество точек t ($t \in A_u$), содержащееся в пространстве T ($A_u \subset T$), где $u \in L$, L обозначает множество индексов u.

Заметим, что здесь и ниже использованы следующие символы: $\in$ и $\subset$ — принадлежности, $\cup$ — объединение, сложение множеств, $\cap$ — пересечение, умножение множеств, $\varnothing$ — пустое множество, не содержащее ни одного элемента.

Фиксированные множества A_u и A_v точек t, пересечения которых есть пустые множества $A_u \cap A_v = \varnothing$, рассматриваются как математические модели u и v-го геологических объектов.

В случае, когда общее число объектов k составляет весь набор геологических объектов, соблюдаются следующие соотношения:

$$\bigcup_{u=1}^{k} A_u = T; \qquad A_u \cap A_v = \varnothing; \quad u \neq v; \qquad u, v \in L.$$

Таким образом, все множества разграничения в сумме составляют T, два любых множества A_u и A_v не имеют общих точек.

В качестве модели геологического наблюдения, охарактеризованного комплексом m признаков (например, петрохимический состав единичной силикатной пробы гранитов), рассматривается многомерная случайная величина ξ:

$$\xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\} \quad j = 1, 2, \ldots, m.$$

В результате каждой точке $t \in T$ можно поставить в соответствие многомерную случайную величину

$$\xi^t = (\xi_{t1}, \xi_{t2}, \ldots, \xi_{tj}, \ldots, \xi_{tm}),$$

что позволяет в качестве математической модели изучаемого u-го геологического объекта (ξ^{A_u}) рассматривать дискретное множество A_u точек t, на котором задана m-мерная случайная функция ξ^t, $t \in A_u$. Будем полагать, что множество A_u содержит n_u точек t.

Обозначим функцию распределения многомерной случайной величины ξ^t через $F_t(X)$, а ее плотность вероятности (если ξ непрерывная) через $f_t(X)$. Тем самым мы ставим в соответствие каждому элементу нашей модели геологического объекта функцию $F_t(X)$, которая обеспечивает этому элементу однозначную характеристику. Если F и f нормальны, то в ранговых моделях часто используются обратные функции нормального распределения.

В качестве модели средних значений изучаемых геологических характеристик j в геологическом объекте (ξ^A) будем рассматривать вектор средних (набор математических ожиданий случайных величин ξ_j)

$$a = \{a_j\} = M\{\xi_j\}, \quad j = 1, 2, \ldots, m.$$

В качестве модели степени изменчивости и зависимостей между изучаемыми геологическими признаками ξ_j в геологическом объекте (ξ^A) будем рассматривать ковариационную матрицу $\Sigma = \{\sigma_{ij}\}$; $i, j = 1, 2, \ldots, m$, т. е. наборы ковариаций $\mathrm{cov}_{ij} = \sigma_{ij}$, $i \neq j$ и дисперсий $D\xi_j = \sigma_j^2 = \sigma_{jj}$ случайных величин ξ_j и ξ_i.

Чаще всего в качестве эмпирической оценки вектора средних значений геологического объекта (ξ^A) используется вектор средних арифметических по каждому признаку

$$\hat{a} = \{\bar{x}_j\} = \left\{\frac{1}{n}\sum_{t=1}^{n} x_{tj}\right\}. \qquad j = 1, 2, \ldots, m.$$

В качестве эмпирической оценки ковариационной матрицы геологического объекта (ξ^A) используется несмещенная выборочная ковариационная матрица

$$\hat{\Sigma} = \{\hat{\sigma}_{ij}\} = \left\{\frac{1}{n-1}\sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)\right\}.$$

Естественно, что эмпирические данные, взятые по конкретному геологическому объекту, представляют собой прямоугольную матрицу X с n наблюдениями, анализами проб (строки матрицы X_t) над m геологическими признаками (столбцы матрицы X_j).

$$X = \{X_t\} = \{x_j\} = \begin{pmatrix} x_{11}, \ldots, x_{1m} \\ \ldots, x_{tj}, \ldots \\ x_{n1}, \ldots, x_{nm} \end{pmatrix}$$

$$t = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

где вектор-строка $X_t = (x_{t1}, \ldots, x_{tm})$, а вектор-столбец $x_j = (x_j')'$, $x_j' = (x_{1j}, \ldots, x_{nj})$.

$'$ — операция транспонирования (для удобства вектор-столбец x_j показан, как вектор-строка x'_j).

ГЛАВА 2

РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ МАТЕМАТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

2.1. ПОСТАНОВКА ЗАДАЧ КЛАССИФИКАЦИИ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

При металлогенических исследованиях, поисковых и геологоразведочных работах, а также при геолого-экономической оценке месторождений полезных ископаемых геолог сталкивается с необходимостью классифицировать геологические объекты. Такими объектами могут быть различные рудоносные и безрудные геологические образования, пегматиты, скарны, грейзены и другие породы, нефтегазоносные и пустые структуры и ловушки, участки рудных тел и зон с промышленной концентрацией и т. п.

При решении прогнозных задач геолог группирует (строит классификацию) изученные геологические объекты, а затем уточняет геологические свойства каждой полученной таким образом однородной классификационной группы. Если исследователь получает данные по новому объекту, то он должен отнести изучаемый объект к одной из априорно известных однородных классификационных единиц либо построить по имеющимся данным новую классификацию. Особое внимание в подобной ситуации обращают на геологические объекты, сходные с месторождениями или, напротив, значительно отличающиеся от них. Максимальная типичность и максимальная аномальность — это, как справедливо пишут А. Н. Бугаец и Л. Н. Дуденко [5, с. 26], единственно пригодные принципы прогнозирования при отсутствии информации по эталонным месторождениям и недостатке сведений о благоприятных признаках.

Исследователь обычно решает одну из двух задач классификации: задачу типизации, при которой изучаемую совокупность геологических наблюдений и объектов разбивают на сравнительно небольшое число областей группирования так, чтобы элементы одной области лежали друг от друга по возможности на небольшом расстоянии, или задачу выявления естественного расслоения исходных наблюдений и объектов на четко выраженные группы (кластеры), лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные друг от друга части. Задача типизации всегда имеет решение, а кластеризация не всегда, т. е. может существовать один-единственный кластер [1, с. 76].

Классификация — один из фундаментальных процессов в науке. Факты и явления должны быть упорядочены, прежде чем мы смо-

жем их понять и разработать общие принципы, объясняющие как их появление, так и наблюдаемый среди них порядок. По определению Р. Снита и Р. Сокала [21, с. 7], классификация — это упорядочение объектов по их схожести. Попытки разработать методы автоматической классификации сделали необходимым оценивать сходство количественно. Заметим, что И. Дж. Гуд [21, с. 66] предлагает теорию кластеров обозначать понятием «ботриология». Более известное в геологии понятие «таксономия» следует, по его мнению, относить лишь к биологическим объектам, тогда как термин «ботриология» предназначен для всех объектов при условии, что применительно к ним используются математические методы. Ботриологию можно считать частью более общей теории формулирования гипотез [21, с. 67].

Имеются три основных типа данных, используемых в кластерном анализе: многомерные данные, данные о близости и данные о кластерах. Термин «близость», следуя терминологии, предложенной Р. Н. Шепардом, относится к сходству, различию, корреляции, мере пересечения или же к любой другой переменной, используемой в качестве меры сходства или расстояния между двумя объектами одного вида [21, с. 21]. Чаще всего задача классификации заключается в том, чтобы разбить множество объектов или наблюдений на подмножества так, чтобы каждый объект принадлежал одному и только одному подмножеству разбивки и чтобы объекты, принадлежащие разным группам, различались между собой.

Классификацию объектов, с одной стороны, 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 специалистами (геологами соответствующих областей знаний: петрологами, геохимиками, геофизиками, геологоразведчиками и др.) на основании профессиональных знаний, опыта, интуиции.

Хотя бóльшая часть характеристик конкретного многомерного геологического наблюдения должна быть похожа на характеристики всех остальных наблюдений из одного и того же кластера, нет никакой необходимости в сходстве по всем признакам. Принадлежность к кластеру определяется «большинством голосов» (наибольшим числом общих значений признаков), и ни один из признаков не определяет принадлежность к данному кластеру. Вышеуказанное было сформулировано М. Бекнером как принцип классификации, а кластеры (классы наблюдений), определенные таким образом, получили название политетических. Большинство методов кластерного анализа и классификации направлено на получение политетических кластеров. В противоположных системах принадлежность определяется общими значениями всех (или, по крайней мере, некоторых) признаков. Это делает монотетические классификации полезными для построения таксономических

ключей, но порождаемая ими систематизация естественных объектов (конечно, и геологических объектов) часто оказывается неудовлетворительной [21, с. 9].

Следует отметить, что для сравнения различных процедур классификации оправдано привлечение искусственных многомерных моделей с известной таксономической структурой [12, с. 2].

Под термином «классификация» обычно понимается распределение предметов по заданным классам согласно наиболее существенным признакам, присущим предметам данного типа и отличающим их от предметов других типов. При этом каждый класс занимает в заданной системе определенное постоянное место и в свою очередь делится на подклассы [36]. Под термином «класс» в математике принято понимать некоторую совокупность объектов, обладающих хотя бы одним общим характеристическим признаком.

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.

2.2. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧ КЛАССИФИКАЦИИ

Осуществим необходимую конкретизацию формальной задачи классификации; введем необходимые понятия метрик, расстояния, мер или коэффициентов сходства, входящих в состав рабочих статистик рекомендованных статистических критериев решения задач классификаций геологических объектов.

Метрическим пространством называется пара (X, d), состоящая из некоторого множества (пространства) элементов (точек) X_u и расстояния d.

Однозначная неотрицательная вещественная функция, определенная для любых X_u и X_v из X, называется функцией расстояния (метрикой), если соблюдаются следующие аксиомы [17, с. 16].

1. $d(X_u, X_v) \geqslant 0$ для всех X_u и X_v из X.

2. Аксиома максимальной близости объекта с самим собой: $d(X_u, X_v) = 0$, тогда и только тогда, когда $X_u = X_v$.

3. Аксиома симметрии: $d(X_u, X_v) = d(X_v, X_u)$.

4. Аксиома треугольника $d(X_u, X_v) \leqslant d(X_u, X_z) + d(X_z, X_v)$.

Понятием, противоположным расстоянию $d(X_u, X_v)$, является понятие сходства между двумя объектами $r(X_u, X_v)$.

Однозначная неотрицательная вещественная функция, определенная для любых X_u и X_v из X, называется мерой сходства, или коэффициентом сходства, по В. Дюрану и П. Оделлу [17, с. 19], если соблюдаются следующие аксиомы.

1. $0 \leqslant r(X_u, X_v) \leqslant 1$ для всех $X_u \neq X_v$.

2. Аксиома симметрии $r(X_u, X_v) = r(X_v, X_u)$.

3. Аксиома максимального сходства объекта с самим собой: $r(X_u, X_v) = 1$, тогда и только тогда, когда $X_u = X_v$.

4. Аксиома монотонного убывания коэффициентов сходства $r(X_u, X_v)$ по $d(X_u, X_v)$, т. е. из $d(X_z, X_l) \geqslant d(X_u, X_v)$ должно с необходимостью следовать выполнение неравенства $r(X_z, X_l) \leqslant r(X_u, X_v)$.

Выбор расстояния (или коэффициента сходства) является узловым моментом исследования, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения.

Функция расстояния и мера сходства определяют понятие однородности объектов, которое является наименее формализованным в кластерном анализе. Если задана функция $d(X_u, X_v)$ или функция $r(X_u, X_v)$, то близкие в смысле этих метрик объекты считаются однородными, принадлежащими к одному классу. Естественно, при этом необходимо сопоставить $d(X_u, X_v)$ и $r(X_u, X_v)$ с некоторыми пороговыми значениями, определяемыми в каждом конкретном случае по-своему.

При формализации задач классификации типична ситуация: среди k изучаемых геологических объектов, по каждому из которых взято по n_u, $u=1, 2, ..., k$ наблюдений, охарактеризованных комплексом m геологических свойств x_j, $j=1, 2, ..., m$, необходимо выявить однородные группы объектов с аналогичным типом распределений исследуемых характеристик x_j в пределах каждой такой однородной группы. Любой объект должен входить в один и только один кластер.

Целесообразно различать три аспекта процедуры применения кластерного анализа.

А. Выбор функций расстояний d или мер сходства r между любыми парами многомерных наблюдений, т. е. между парами m-мерных векторов X_{ut} и X_{vt}.

Б. Выбор функций расстояний d или мер сходства r между любыми парами объектов, каждый из которых охарактеризован наборами n_u и n_v многомерных наблюдений X_{ut}, $t=1, 2, ..., n_u$ и X_{vt}, $t=1, 2, ..., n_v$. Иначе говоря, речь идет о выборе функций d или r между любыми парами матриц X_u и X_v размерности $n_u \cdot m$ и $n_v \cdot m$.

В. Выбор функций расстояния d или мер сходства r между любыми парами групп объектов, в том числе между объектом и группой объектов, состоящей минимум из двух других объектов. Другими словами, речь идет о функциях расстояний d или мер сходства r между матрицей X_u и парой матриц (X_{v1}, X_{v2}) или в общем виде между одной группой матриц $(X_{u1}, X_{u2}, ...)$ и другой $(X_{v1}, X_{v2}, ...)$.

С. А. Айвазян и ряд других исследователей [1, с. 98—110] задачи кластерного анализа подразделяют на два типа. К первому (I) типу они относят задачи классификации сравнительно небольших по объему совокупностей многомерных наблюдений, когда их несколько десятков, а ко второму (II) — задачи классификации достаточно больших массивов многомерных наблюдений, когда их сотни и тысячи.

С точки зрения априорной информации о числе классов, на которое следует разбить исследуемую совокупность многомерных наблюдений, задачи классификации подразделяются на три типа [1, с. 98—99]: а) количество классов априорно задано; б) количество классов неизвестно и его следует определить; в) количество классов неизвестно, но его определение не входит в условие задачи.

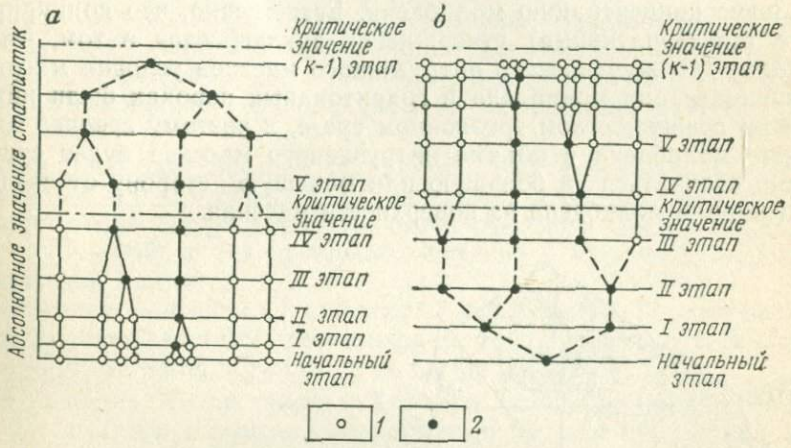


Рис. 2. Иерархическое дерево классификации объектов:
а — агломеративное; *б* — дивизимное
1 — объекты; *2* — однородные группы

Последние ситуации приводят к построению иерархических деревьев (дендрограмм). Существует два типа иерархических деревьев — агломеративное и дивизимное (рис. 2). В соответствии с вышеотмеченным разделением на типы выделяют три основные кластерные процедуры:

1) иерархические агломеративные и дивизимные — для решения задач типа в) и типа I;

2) параллельные — для решения задач сочетания типов Iа, Iб, реализуемых с помощью итерационных алгоритмов (на каждом шаге итерации одновременно привлекаются все наблюдения);

3) последовательные — для решения задач сочетания типов IIа, IIб, реализуемых с помощью итерационных алгоритмов, причем на каждом шаге итерации привлекается небольшая часть наблюдений (например, одно из исходных наблюдений или результат классификации на предыдущем шаге итерации).

2.3. РЕКОМЕНДУЕМЫЕ ОДНОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

Типичную картину постановки геологических задач выявления сходства или подтверждения различий в средних и дисперсиях проиллюстрируем на следующем примере.

Будем проверять, влияет ли фактор глубинности на среднюю концентрацию молибдена в эльджуртинских гранитах, с которыми, возможно, генетически связано крупное молибден-вольфрамовое месторождение Тырныауза. Воспользуемся данными, опубликованными В. В. Ляховичем [30, табл. 31, 108]. Можно было бы также проверить, сказывается ли фактор глубинности на качестве газа и нефти в некотором бассейне. Проверяемое геологическое предположение заключается в гипотезе, что глубинность не влияет на среднюю концентрацию молибдена. Естественно, что конкурирующие (альтернативные) утверждения заключаются в том, что на глубине эльджуртинского интрузивного массива условия миграции и концентрации молибдена в гранитоидных породах были иными, чем на поверхностном эрозионном срезе, и поэтому среднее содержание молибдена в объеме интрузивного массива будет существенно отличаться в большую или меньшую сторону от среднего содержания молибдена на поверхности массива.

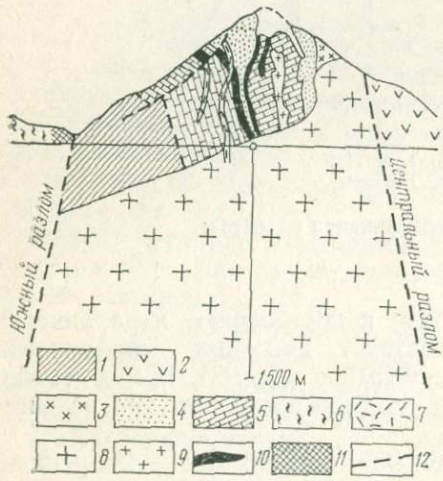


Рис. 3. Геологическое положение скв. 600 (Эльджуртинский гранитный массив). По В. В. Ляховичу

1 — мулуканская свита; *2* — вулканогенная свита; *3* — кварцевые плагиопорфиры (трондъемиты); *4* — биотитовые роговики; *5* — мраморы; *6* — мигматиты; *7* — липариты; *8* — эльджуртинские граниты; *9* — лейкократовые гранитоиды; *10* — скарны; *11* — кристаллические сланцы; *12* — тектонические контакты

Охарактеризуем формальный аналог этой геологической задачи. Будем рассматривать случайную величину ξ как модель (полезную абстракцию) содержаний молибдена в гранитах поверхности, а случайную величину η — как модель содержаний молибдена в гранитах, изученных по буровой скв. 600 (рис. 3). Математические ожидания (средние) и дисперсии этих случайных величин обозначим: $M\xi = a_1$, $M\eta = a_2$; $D(\xi) = \sigma_1^2$; $D(\eta) = \sigma_2^2$. С учетом введенных понятий случайных величин ξ и η проверяемую H_0 и альтернативные H_1 гипотезы можно записать как $H_0 : a_1 = a_2$ и $H_1 : a_1 \neq a_2$, где a_1 — среднее случайной величины — модели содержаний молибдена на поверхности эльджуртинского гранита; a_2 — то же, для эльджуртинского гранита, изученного по буровой скв. 600.

Для другого проверяемого геологического предположения, что глубинность не влияет на величину среднего рассеяния молибдена (дисперсию), следует поставить в соответствие следующие формальные гипотезы: $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$, где σ_1^2 — дисперсия случайной величины, рассматриваемой моделью содержаний молибдена на поверхности эльджуртинского гранита; σ_2^2 — то же, для эльджуртинского гранита, изученного по буровой скв. 600.

Для проверки гипотезы $H_0: a_1 = a_2$ рекомендуются два перспективных метода: ранговый W критерий Вилкоксона (нередко обозначаемый как критерий Манна и Уитни, Уилкоксона) [18, с. 270; 4, с. 147—149, 34, с. 341—343] и параметрический t-критерий, являющийся разновидностью статистики Вэлча [18, с. 248—249].

Для проверки гипотезы $H_0: \sigma_1^2 = \sigma_2^2$ рекомендуются три перспективных метода: ранговый R критерий Сиджела—Тьюки (иногда обозначаемый как ранговый критерий рассеяния Зигеля и Тьюки) [18, с. 264—266], а также параметрические F критерий Фишера и M критерий Бартлета.

Ранговые критерии Вилкоксона и Сиджела—Тьюки описаны в постановке известных статистиков Л. Н. Большева и Н. В. Смирнова [4, с. 147—149], а также Д. Б. Оуэна [34, с. 354].

В качестве геологического примера классификации набора k ($k \geqslant 2$) объектов с помощью иерархической агломеративной кластер-процедуры (попарных сравнений) воспользуемся данными В. В. Ляховича [30, табл. 9, 10, 11 и 13] о содержаниях молибдена в лейкократовых гранитоидах Тырныауза.

В разделе 2.3.2 настоящей работы осуществлена проверка геологических предположений, что средняя концентрация молибдена и ее разброс в лейкократовых гранит-порфирах «Паука» [30, табл. 9] и балки «Самолет» [там же, табл. 10], а также в лейкократовых аплит-порфирах [там же, табл. 11] и в вертикальном разрезе лейкократовых гранитов [там же, табл. 13] Северного участка одни и те же при альтернативе, что на некоторых участках они существенно различаются. Особенности геологического строения и вещественного состава эльджуртинских гранитов и лейкократовых гранитоидов Тырныауза, используемых в качестве геологических примеров, описаны достаточно подробно В. В. Ляховичем [30].

Проведем формализацию поставленных геологических задач. Случайные величины ξ_1, ξ_2, ξ_3 и ξ_4 будем рассматривать как модели содержаний молибдена в указанных четырех типах лейкократовых гранитоидов. Тогда нулевую и альтернативные гипотезы о среднем можно записать так: $H_0: a_1 = a_2 = a_3 = a_4 = a_0$, а $H_1: a_u \neq a_0$ хотя бы для одного объекта u $(u = \overline{1,4})$.

Соответственно нулевую и альтернативные гипотезы о дисперсиях можно записать как: $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_0^2$, $H_1: \sigma_u^2 \neq \sigma_0^2$ хотя бы для одного $u = 1, 2, 3, 4$.

Отметим два важных обстоятельства.

1. Ранговые критерии проверки гипотез о равенстве истинных средних в сопоставляемых объектах основаны на изучении так называемого параметра сдвига распределений. Отсутствие сдвига (равенство параметра сдвига нулю) свидетельствует о равенстве средних в рассматриваемых объектах. На все ранговые критерии проверки различий в параметре масштаба, т. е. в дисперсиях, накладывается весьма существенное ограничение. Как указывалось Л. Е. Мозесом, Я. Гаеком, З. Шидаком [11, с. 117], предположение о равенстве параметров сдвига (равенстве средних) обеих сравниваемых плотностей вероятностей существенно, без него никакой ранговый критерий для проверки различия в масштабе не имеет удовлетворительных свойств.

С увеличением разности между средними значениями сравниваемых генеральных совокупностей $|a_1 - a_2|$ уменьшается вероятность того, что нулевая гипотеза $H_0: \sigma_1^2 = \sigma_2^2$ при наличии действительной разницы в дисперсиях $\sigma_1^2 \neq \sigma_2^2$ будет отклонена. Чем бóльшая разница средних значений $|a_1 - a_2|$, тем больше вероятность появления ошибки второго рода *, т. е. тем меньше эффективность критериев.

И это особенно справедливо при малых дисперсиях [24, с. 264]. Указанное обстоятельство приводит к конкретным алгоритмическим рекомендациям, связанным с центрированием и нормированием ранговых статистик для проверки гипотез о рассеянии. Они описаны ниже в разделе 2.3.1 и реализуются также в разделе 2.3.2. Речь идет о ранговом критерии Сиджела—Тьюки.

2. В фундаментальных методических пособиях по математической статистике [4, 18, 34] при описании критериев Вилкоксона, Сиджела—Тьюки и некоторых других рекомендуется вычислять ранговую сумму лишь меньшей по объему выборки. Это чревато, с нашей точки зрения, появлением ложных результатов при резкой разнице объемов в сравниваемых выборках, например 5 и 5000. Особенно это часто проявляется в случае применения иерархической кластерной процедуры набора выборок на базе вышеупомянутых ранговых критериев (см. раздел 2.3.2).

Для контроля возможности появления ошибочных заключений мы предлагаем воспользоваться одномерным вариантом многомерных ранговых статистик Пури—Сена—Тамуры (см. раздел 2.4.1), основанных на вычислении ранговых сумм по двум сравниваемым выборкам.

В случае расхождения результатов классификации набора объектов (выборок) на базе статистик Вилкоксона (Сиджела—Тьюки) и Пури—Сена—Тамуры, предпочтение следует отдавать результатам, полученным с помощью ранговых критериев Пури—Сена—Тамуры. Одномерная модификация статистик Пури—Сена—Тамуры приведена в разделе 2.4.1.

---

* Вероятность ошибки второго рода — вероятность ошибочного принятия нулевой гипотезы о равенстве проверяемого параметра, например дисперсии, когда на самом деле верна альтернатива о различиях этого параметра.

2.3.1. Ранговые и параметрические статистики Вилкоксона, Вэлча, Сиджела—Тьюки, Фишера, Бартлета проверки гипотез о среднем и дисперсии в двух объектах

В данном разделе детально охарактеризованы: ранговый критерий сдвига (среднего) Вилкоксона в постановке Л. Н. Большева и Н. В. Смирнова [4], комплексирующийся * с ним параметрический критерий Вэлча в постановке Л. Закса [18], ранговый критерий рассеяния (дисперсий) Сиджела—Тьюки, описанный у Л. Закса [18] и Д. Б. Оуэна [34], и комплексирующиеся * с ним параметрические критерии проверки дисперсий Фишера и Бартлета, описанные во всех руководствах по применению статистических методов.

2.3.1.1. Критерий Вилкоксона

Этот критерий предназначен для проверки гипотез: $H_0: a_1 = a_2$ против набора альтернатив $H_1: a_1 \neq a_2$, где a_1 и a_2 — истинные средние для первого и второго объектов. Являясь ранговой, статистика критерия Вилкоксона нечувствительна к нарушению условий нормальности распределения исходных геологических данных, а также к наличию аномальных значений и прочим стохастическим моментам. Предполагается, что элементы выборок взаимно независимы и подчиняются непрерывным распределениям.

Процедура применения критерия Вилкоксона такова. 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, ..., N. Равным значениям (совпадающим членам) соответствует 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)

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

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

Ситуация 1. Объемы наблюдений в выборках не превышают 25, т. е. n_1 и $n_2 \leqslant 25$.

В табл. 6.8 указанной работы [4, с. 419] обозначено: $n_1 = m$, $n_2 = n$, односторонний уровень значимости $\alpha/2 = Q$, удвоенное математическое ожидание статистики Вилкоксона — $2MW$.

---

* Описания сначала ранговых, а затем параметрических методов в монографиях, посвященных специально ранговым методам, приведены лишь для удобства изложения. В практических расчетах более естественно и оправданно сначала применять параметрические и потом лишь ранговые методы.

По этой таблице при заданных m, n, Q снимают показания нижнего критического значения 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 критических значений W_1 и W_2 из таблицы [4] остается правомерной. При этом, однако, уровень значимости используемого критерия будет несколько занижен.

Ситуация 2. Объемы наблюдений в выборках превышают 25, т. е. n_1 или (и) $n_2 > 25$.

Согласно [4] критические значения W_1 и W_2 определяют по следующим приближенным формулам:

$$W_1 \simeq \left\{0{,}5\,[m(m+n+1)-1] - t_{\alpha/2}\cdot\sqrt{\frac{1}{12}mn(m+n+1)}\right\},$$

$$W_2 \simeq m(m+n+1) - W_1,$$

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 принимает следующий вид:

$$W_1 \approx \left\{\frac{m(m+n+1)-1}{2} - t_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}\left(1 - \frac{\sum_{i=1}^{k}(t_i^3 - t_i)}{(m+n+1)(m+n)(m+n-1)}\right)}\right\},$$

где k — общее количество групп, состоящих из совпавших величин, принадлежащих разным выборкам, t_i — количество совпавших величин в группе с номером i, $i = 1, 2, \ldots, k$.

Совпадения, целиком состоящие из элементов какой-либо одной выборки, могут в расчет не приниматься, поскольку не влияют на величину статистики W.

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 $W > W_2$.

---

* Уровень значимости — это вероятность появления ошибки первого рода, т. е. вероятность ошибочного отклонения нулевой гипотезы и принятия альтернативы, когда на самом деле верна нулевая гипотеза.

Пример. Проверяется предположение, что фактор глубинности не влияет на среднюю концентрацию молибдена в гранитах Эльджуртинского массива.

Формальная постановка задачи рассматривалась выше $H_0: a_1 = a_2$ против набора альтернатив $H_1: a_1 \neq a_2$, где a_1 и a_2 — средние значения случайных величин, являющихся моделями содержаний молибдена в гранитах поверхности и скв. 600, пробуренной в эльджуртинских гранитах.

Таблица 1

Начало и окончание вариационных рядов исходных и центрированных данных о содержании молибдена в эльджуртинских гранитах и ранжирование по ранговым критериям Вилкоксона и Сиджела—Тьюки

Исходное содержание Мо, г/т		Критерий Вилкоксона		Критерий Сиджела—Тьюки					
						Центрированные медианами данные содержаний Мо			
Граниты поверхности $n_1=50$	Граниты скважины $n_2=61$	Ранг r	Скорректированный ранг $r_{ск}$	Ранг r	Скорректированный ранг $r_{ск}$	Граниты поверхности $n_1=50$	Граниты скважины $n_2=61$	Ранг r	Скорректированный ранг $r_{ск}$
0,4		1	1	1	1	—0,6		1	1
0,5		2	2	4	4	—0,5		4	4
0,6		3	3,5	5	6,5	—0,4		5	6,5
	0,6	4	3,5	8	6,5		—0,4	8	6,5
	0,7	5	19,5	9	38,5		—0,3	9	38,5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0,7		34	19,5	68	38,5	—0,3		68	38,5
0,8		35	36,5	69	72,5	—0,2		69	72,5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1,0*		56	55,5	111*	—	0*		111*	—
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2,2	110	110	3	3		1,2	3	3
	5,4	111	111	2	2		4,4	2	2
			Σ=2746,5		Σ=3044,9				Σ=3044,9

* Медиана общего нечетного вариационного ряда устраняется по критерию Сиджела—Тьюки, т. е. устраняется 111-й член.

Общий вариационный ряд, ранги, скорректированные ранги приведены в табл. 1. Наибольший ранг 50+61=111 соответствует максимальному содержанию молибдена в 5,4 г/т. Скорректированные ранги, например, для содержаний молибдена 0,7 г/т, занимающих в вариационном ряду места с 5 по 34, составят: (5+6+ +...+34) : 30=19,5.

Вычислим статистику $W=\sum_{i=1}^{50} r_i = 1+\ldots+105{,}5=2746{,}5$. Найдем критические значения W_1 и W_2. Учитывая, что $n_1=m=50>25$, а также $n_2=n=61>25$, имеем дело с ситуацией 2. Поэтому критические значения W_1 и W_2 при заданном уровне значимости $\alpha=0{,}05$ составят (с учетом совпадающих значений в группах)

$$\sum_{i=1}^{k}(t_i^3-t)=(2^3-2)+\ldots+(4^3-4)=32316,$$

$$W_1\approx\left\{0{,}5\,[50\,(50+61+1)-1]-\right.$$

$$\left.-1{,}96\sqrt{\frac{50\cdot 61\cdot 112}{12}\left[1-\frac{32\,316}{112\cdot 111\cdot 110}\right]}\right\}\approx 2472{,}7,$$

$$W_2=50\,(50+61+1)-2472{,}7\approx 3127{,}3.$$

Рассчитанное значение $W=2746{,}5$ не вышло за пределы, образованные допустимыми значениями $W_1=2472{,}7$ и $W_2=3127{,}3$. Поэтому проверяемая гипотеза $H_0: a_1=a_2$ должна быть принята как не противоречащая эмпирическим геологическим наблюдениям. Другими словами, нет оснований полагать, что глубинность влияет на среднюю концентрацию молибдена в эльджуртинских гранитах.

Для сравнения вычислим критические значения W_1 и W_2 без учета совпадений значений в выборках. Получим: $W_1\approx 2468$, $W_2\approx 3132$.

Как видим, для данного примера критические значения изменились незначительно, несмотря на то, что количество совпавших значений весьма велико (из 111 значений совпало 95). Это говорит о том, что при проверке гипотезы о средних, по-видимому, вводить поправки в критические значения W_1 и W_2 за счет совпадений значений необязательно.

2.3.1.2. Критерий Вэлча

Указанный критерий предназначен также для проверки гипотезы о среднем $H_0: a_1=a_2$ при наборе альтернатив $H_1: a_1\neq a_2$. Являясь параметрическим, критерий использует часто не выполнимые для геологической практики предположения о нормальности распределений случайных величин — моделей изучаемых геологических признаков в сравниваемых объектах, а также предположения об отсутствии аномальных наблюдений и некоторые другие. Поэтому наиболее рационально сочетать применение рангового критерия Вилкоксона и параметрического критерия Вэлча.

Строго говоря, когда истинные дисперсии неравны $\sigma_1^2\neq\sigma_2^2$ и исследователю неизвестны, мы сталкиваемся со сложной проблемой Беренса—Фишера, для которой точного решения не существует [18, с. 248].

Для практических целей проверки гипотезы о равенстве средних при не очень малых объемах наблюдений n_1 и n_2 в выборках можно воспользоваться статистикой Вэлча [24, с. 248]

$$t = |\bar{x} - \bar{y}| \Big/ \sqrt{s_1^2 \mid n_1 + s_2^2 \mid n_2}, \quad \text{где} \quad \bar{x} = \frac{1}{n_1}\sum_{t=1}^{n_1} x_t,$$

$$\bar{y} = \frac{1}{n_2}\sum_{t=1}^{n_2} y_t, \qquad s_1^2 = \frac{1}{n_1 - 1}\sum_{t=1}^{n_1}(x_t - \bar{x})^2, \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{t=1}^{n_2}(y_t - \bar{y})^2.$$

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

$$f = \{z\} = \left\{ \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 + 1} + \dfrac{(s_2^2/n_2)^2}{n_2 + 1}} - 2 \right\},$$

где символ $\{z\}$ означает взятие целой части от числа z. Для упрощения вычислений можно воспользоваться свойством близости распределения Стьюдента и нормального (гауссовского) для больших объемов наблюдений ($n > 50$). Поэтому полагаем, что в условиях нулевой гипотезы $H_0 : a_1 = a_2$ величина t приближенно распределена по стандартному нормальному закону.

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. $t \leqslant t_{\alpha, f}$, либо не превысит квантиль $t_{\alpha, f}$ нормального распределения: $t \leqslant t_{\alpha/2}$. Проверяемая гипотеза отклоняется и принимаются альтернативы о существенности различий в средних $H_1 : a_1 \neq a_2$, когда $t > t_{\alpha, f}$ или $t > t_{\alpha/2}$ (при $\alpha = 0{,}05$ квантиль $t_{0{,}025} = 1{,}96$).

Пример. Воспользуемся тем же геологическим примером проверки влияния фактора глубинности на среднюю концентрацию содержаний молибдена в эльджуртинских гранитах. Определим статистику t:

$$t = |\bar{x} - \bar{y}| \Big/ \sqrt{s_1^2 \mid n_1 + s_2^2 \mid n_2} =$$

$$= |1{,}04 - 1{,}17| \Big/ \sqrt{0{,}121 \mid 50 + 0{,}478 \mid 61} \simeq 1{,}29,$$

где $\bar{x} = 1/50\,(1{,}4 + \ldots + 1{,}4) \simeq 1{,}04$; $\bar{y} = 1/61\,(0{,}7 + \ldots + 0{,}7) \simeq 1{,}17$, $s_1^2 = 1/49[(1{,}4 - 1{,}04)^2 + \ldots + (1{,}4 - 1{,}04)^2] \simeq 0{,}121$, $s_2^2 = 1/60[(0{,}7 - 1{,}17)^2 + \ldots + (0{,}7 - 1{,}17)^2] \simeq 0{,}478$. Отсюда

$$f = \left\{ \frac{(0{,}121/50 + 0{,}478/61)^2}{\dfrac{(0{,}121/50)^2}{51} + \dfrac{(0{,}478/61)^2}{62}} - 2 \right\} \approx 103.$$

Квантиль $t_{\alpha, f}$ распределения Стьюдента при уровне значимости $\alpha=0{,}05$ и $f=103$ степенях свободы составляет 1,984.

Учитывая, что $t=1{,}29<t_{\alpha, f}=1{,}984$ (а также $t=1{,}29<t_{\alpha/2}=1{,}96$), следует нулевую гипотезу о равенстве проверяемых средних с помощью критерия Вэлча принять как подтвердившуюся. Другими словами, нет оснований полагать, что глубинность влияет на среднюю концентрацию молибдена в эльджуртинских гранитах.

Поскольку такое же заключение было сделано с помощью рангового критерия Вилкоксона, можно вполне обоснованно полагать, что глубинность не влияет на среднюю концентрацию молибдена в эльджуртинских гранитах. Одновременно следует считать, что различия в средних арифметических молибдена в гранитах поверхности (1,04 г/т) и керна скв. 600 (1,17 г/т) статистически незначимы и не вызваны более высоким истинным средним содержанием молибдена в объеме Эльджуртинского гранитного массива.

2.3.1.3. Критерий Сиджела—Тьюки

Этот критерий 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. Являясь ранговой, статистика Сиджела—Тьюки нечувствительна к нарушению условий нормальности распределения наблюдений, наличию аномальных значений и другим трудно устранимым для геологической практики стохастическим моментам.

Статистика Сиджела—Тьюки является полным аналогом статистики Вилкоксона [34, 18], но проверка осуществляется в этом случае относительно параметра масштаба (дисперсии), а не параметра сдвига (среднего).

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 [18] 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 корректного применения R-критерия Сиджела—Тьюки необходимо учитывать два обстоятельства:

убедиться в предположении о равенстве параметров сдвига (равенстве средних);

в случае отсутствия этого равенства произвести центрирование выборочных данных, например центрирование медианами.

В случае равенства средних можно пользоваться исходными геологическими данными и произвести корректировку рангов для совпадающих (связанных) членов общего вариационного ряда.

Корректирование рангов можно осуществить способом, аналогичным описанному при рассмотрении критерия Вилкоксона, т. е. рассчитать средние арифметические рангов для совпадающих их наблюдений (см. табл. 1). Имеются две схемы применения рангового критерия Сиджела—Тьюки.

Схема А.

1. С помощью критериев Вилкоксона и 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 или 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 ранжирование 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, 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 (аналогично учитываются две ситуации — см. раздел 2.3.1.1).

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 и тем самым принимается альтернатива, 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

$$t=\frac{2R-n_1(n_1+n_2+1)+\delta}{\sqrt{\frac{n_1n_2}{3}(n_1+n_2+1)}},\quad \text{где } \delta=\begin{cases}1, & \text{если } 2R>n_1(n_2+n_1+1),\\ -1, & \text{если } 2R\leqslant n_1(n_1+n_2+1).\end{cases}$$

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

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

Если больше чем пятая часть наблюдений связана равенствами, то вносятся коррективы в подкоренное выражение [18, с. 265]

$$t=\frac{2R-n_1(n_1+n_2+1)+\delta}{\sqrt{\frac{n_1n_2}{3}(n_1+n_2+1)-\frac{4n_1n_2(S_1-S_2)}{(n_1+n_2)(n_1+n_2-1)}}},$$

где $S_1=\Sigma r_{\text{св}}^2$ — сумма квадратов рангов зависимых наблюдений, $S_2=\Sigma r_{\text{ср}}^2$ — сумма квадратов средних рангов зависимых наблюдений.

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 $\alpha=0{,}05$ $t_{\alpha/2}=1{,}96$).

Пример. И в этом случае снова воспользуемся данными о содержаниях молибдена в гранитах Эльджуртинского массива [30], а проверять будем предположение, что фактор глубинности не влияет на меру рассеяния (дисперсию) содержаний молибдена в гранитах поверхности и скв. 600.

лась: $H_0: \sigma_1^2=\sigma_2^2$, $H_1: \sigma_1^2\neq\sigma_2^2$, где σ_1^2 и σ_2^2 — неизвестные дисперсии случайных величин, являющихся моделями содержаний молибдена в гранитах поверхности и скв. 600.

Для иллюстрации вычислений расчеты произведены по обеим схемам со всеми возможными коррективами.

По схеме А для рассматриваемого примера (см. табл. 1) наименьшему значению содержания молибдена (0,4 г/т) будет соответствовать ранг 1, наибольшему (5,4 г/т) — ранг 2, значению 2,2 г/т будет соответствовать ранг 3, значению 0,5 г/т — ранг 4 и т. д.

Медианное значение (56-й член) — 1,0 г/т устраняется, так как вариационный ряд нечетный ($n_1+n_2=50+61=111$ — нечетно). Равным значениям дается один и тот же средний ранг; например, для значений 0,7 г/т исправленный ранг будет равен: $(9+\ldots+68):30=38{,}5$.

Статистика R Сиджела—Тьюки представляет собой ранговую сумму исправленных рангов r, относящихся к меньшей по объему выборке, т. е. для гранитов поверхности. Учитывая равенство средних, произведем расчет ранговой суммы по исходным, нецентрированным данным.

Определенные выше (2.3.1.1) критические значения W_1 и W_2 статистики Вилкоксона, аналога критерия Сиджела—Тьюки составили: $W_1=2472{,}7$ (2468) и $W_2=3127{,}3$ (3132).

Так как статистика Сиджела—Тьюки $R=3044{,}9$ не вышла за пределы, образованные критическими значениями W_1 (W_1') и W_2 (W_2'), т. е. 2472,7 (2468) $<3044{,}9<3127{,}3$ (3132), то нет оснований отклонять нулевую гипотезу о равенстве дисперсий в сравниваемых объектах. Геологически интерпретируя этот статистический вывод, делаем заключение, что глубинность не влияет на степень содержания молибдена в эльджуртинских гранитах.

Проведем теперь подсчет ранговой суммы по центрированным данным. В нашем примере эта сумма 3044,9 (см. табл. 1) совпала с суммой, рассчитанной для центрированных данных (что не является характерным), так что и в этом случае с помощью рангового критерия Сиджела—Тьюки принимается гипотеза о равенстве дисперсий, т. е. одинаковой степени содержания молибдена на поверхностном эрозионном срезе и в глубине Эльджуртинского гранитного массива.

По схеме Б определяем t-статистику сначала без учета влияния совпавших значений содержаний молибдена в выборках

$$t=\frac{2R-n_1(n_1+n_2+1)-1}{\sqrt{\frac{n_1n_2}{3}(n_1+n_2+1)}}=\frac{2\cdot 3044{,}9-50(50+61+1)-1}{\sqrt{\frac{50\cdot 61}{3}(50+61+1)}}=1{,}448.$$

Так как $t=1{,}448<t_{\alpha/2}=1{,}96$, то с надежностью 0,95 следует принять нулевую гипотезу о равенстве дисперсий (равенстве содержаний молибдена в эльджуртинских гранитах), что согласуется с выводом об отсутствии влияния фактора глубинности, сделанном по схеме А.

Определим скорректированное на различающиеся объемы выборок значение статистики t^*

$$t^*=t+\left(\frac{1}{10n_1}+\frac{1}{10n_2}\right)(t^3-3t)=1{,}448+\left(\frac{1}{10\cdot 50}+\right.$$

$$\left.+\frac{1}{10\cdot 61}\right)(1{,}448^3-3\cdot 1{,}448)\simeq 1{,}448-0{,}005=1{,}443.$$

Полученное значение t^* практически совпадает с вышенайденным t, поэтому вновь следует принять нулевую гипотезу о равенстве дисперсий.

Найдем теперь скорректированное значение t, учитывающее большое количество равных наблюдений в выборках, т. е. связок,

$$S_1 = \sum r_{св}^2 = 5^2 + \ldots + 110^2 = 397\,891,$$

$$S_2 = \sum r_{ср}^2 = 6{,}5^2 + \ldots + 12{,}5^2 \cdot 4 = 387\,288,$$

$$S_1 - S_2 = 397\,891 - 387\,288 = 10\,602{,}98.$$

Вычисляем скорректированное на наличие связок значение статистики t

$$t = \frac{2 \cdot 3044{,}9 - 50\,(50 + 61 + 1) - 1}{\sqrt{\frac{50 \cdot 61}{3}\,(50 + 61 + 1) - \frac{4 \cdot 50 \cdot 61 \cdot 10\,602{,}98}{(50 + 61)\,(50 + 61 - 1)}}} \simeq 1{,}521.$$

Так как $t = 1{,}521 < t_{\alpha/2} = 1{,}96$, то вновь принимается нулевая гипотеза о равенстве дисперсий (содержание молибдена в эльджуртинских гранитах не зависит от глубинности).

2.3.1.4. *F*-критерий Фишера и критерий Бартлета

Теперь осуществим проверку гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ с помощью двух параметрических критериев — Фишера и Бартлета. 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 значительно более трудоемкий в расчетной части критерий Бартлета.

Применение F-критерия Фишера содержит предположение о нормальности распределения случайных величин ξ и η — в нашем примере моделей содержаний молибдена в гранитах поверхностных и керновых проб Эльджуртинского массива.

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

$$F = s_1^2/s_2^2, \quad \text{если} \quad s_1^2 \geqslant s_2^2 \quad \text{и} \quad F = s_2^2/s_1^2, \quad \text{если} \quad s_2^2 \geqslant s_1^2.$$

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 вычисленное F превысит критическое, т. е. при $F > F_{\alpha, f_1, f_2}$, нулевую гипотезу следует отклонить как противоречащую исходным дан-

ным и принять альтернативные гипотезы о существенности различий в истинных дисперсиях $H_1: \sigma_1^2 \neq \sigma_2^2$. Естественно, что в этом случае более высокую истинную дисперсию (степень рассеяния) следует ожидать у объекта с более высокой выборочной дисперсией s^2.

В нашем примере о содержаниях молибдена в эльджуртинских гранитах оценки дисперсий оказались следующими: $s_1^2 = 0{,}121$; $s_2^2 = 0{,}478$, $F = s_2^2/s_1^2 = 0{,}478/0{,}121 = 3{,}95$.

Так как найденное значение $F = 3{,}95$ превысило допустимое $F_{\alpha, f_2, f_1} = F_{0{,}05;\,60;\,49} = 1{,}59$, соответствующее уровню значимости $\alpha = 0{,}05$, $f_2 = n_2 - 1 = 61 - 1 = 60$ и $f_1 = n_1 - 1 = 50 - 1 = 49$ степеням свободы [4], то нулевую гипотезу следует отклонить как противоречащую исходным данным. Одновременно необходимо признать справедливыми альтернативные гипотезы о различиях в дисперсиях и тем самым сделать заключение, что степень рассеяния молибдена по гранитам скважины существенно выше, чем в эльджуртинских гранитах на поверхностном эрозионном срезе массива.

Для критерия Бартлета условия применения те же, что и для критерия Фишера. Ниже описана упрощенная процедура применения критерия Бартлета, более строгое изложение дано в фундаментальной работе Л. Н. Большева и Н. В. Смирнова [4].

Величина M для двух объектов определяется с помощью выражения

$$M \approx \frac{1+\beta}{1+2\beta}\left[(n_1-1)\ln\frac{s^2}{s_1^2} + (n_2-1)\ln\frac{s^2}{s_2^2}\right] \approx$$

$$\approx \frac{2{,}3026}{c}\left[(N-2)\lg s^2 - (n_1-1)\lg s_1^2 - (n_2-1)\lg s_2^2\right],$$

где $s^2 = \frac{1}{N-2}\left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right]$,

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

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. $M \leqslant \chi^2_{\alpha,\, f}$, соответствующего заданному двустороннему уровню значимости α и $f = 1$ 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.

Для вычисления статистики M Бартлета в примере с содержанием молибдена в эльджуртинских гранитах воспользуемся выражением с натуральными логарифмами

$$s_1^2 = 0{,}121; \quad s_2^2 = 0{,}478; \quad N = 50 + 61 = 111;$$

$$\beta = 1/3\left(\frac{1}{49} + \frac{1}{60} - \frac{1}{109}\right) \approx 0{,}01;$$

$$s^2 = \frac{1}{109}(49 \cdot 0{,}121 + 60 \cdot 0{,}478) = 0{,}318;$$

$$M = \frac{1 + 0{,}01}{1 + 0{,}02}\left(49 \ln \frac{0{,}318}{0{,}121} + 60 \ln \frac{0{,}318}{0{,}478}\right) \approx 22{,}42.$$

Так как численное значение $M = 22{,}42$ превысило критическое $\chi^2_{\alpha, f} = 3{,}84$, соответствующее уровню значимости $\alpha = 0{,}05$ и $f = 1$

Таблица 2

Результаты статистической проверки гипотез о равенстве средних и равенстве дисперсий по рудным, редким и петрогенным элементам в эльджуртинских гранитах Тырныауза

Номер п/п	Элемент	Проверка гипотез о среднем			Проверка гипотез о дисперсиях		
		Критерий Вилкоксона		Критерий Вэлча	Критерий Сиджела—Тьюки		Критерий Бартлета
		Статистика W	t-преобразование	Статистика t	Статистика R	t-преобразование	Статистика M
1	W	2429*	—2,20*	1,94	2582	—1,29	4,29*
2	Mo	2746,5	—0,32	1,21	3044,9	1,45	22,63*
3	Sn	2964	0,97	0,69	3100,5	1,78	134,55*
4	Pb	1958*	—4,99*	5,54*	3413,8*	3,64*	43,39*
5	Zn	2608	—1,14	0,71	3091,8	1,73	2,30
6	Cu	2798	—0,01	0,70	2809,1	0,05	23,4*
7	Nb	1953*	—5,02*	5,28*	1947*	—5,05*	1,54
8	Ta	2888,5	0,52	0,86	2901,7	0,60	149,08*
9	Zr	1867*	—5,52	5,98*	3050,8	1,49	1,07
10	Be	2408*	—2,32*	2,42*	3182*	2,26*	11,18*
11	Li	2863*	0,37	0,42	3303,9*	2,99*	29,69*
12	Cs	3865*	6,31*	5,69*	2620,2	—1,06	13,46*
13	B	3751,5	5,64*	5,89*	2173,8*	—3,71*	2,35*
14	F	3977*	6,69*	8,46*	2674,8	—0,34	0,43
15	Ti	3158*	2,12*	2,13*	2400,3*	—2,37*	6,72*
16	Fe	3829,5*	6,10*	6,87*	2534	—1,58	5,63*
17	Mg	3751,5*	5,64*	5,68*	2214,2*	—3,47*	19,22*
18	Ca	3254,5*	2,69*	2,57*	2825,7	0,15	0,09
19	Mn	4081,5*	7,59*	7,17*	1961*	—4,97*	40,69*

Звездочкой обозначены статистически значимые различия в средних и дисперсиях; жирным шрифтом выделены надежные заключения о различиях или сходстве в средних и дисперсиях, подтверждаемые одновременно ранговыми и параметрическими критериями.

Примечания. 1. *Объемы наблюдения:* пробы поверхностного эрозионного среза эльджуртинских гранитов $n_1 = 50$; керновые пробы вертикального сечения эльджуртинских гранитов $n_2 = 61$. 2. *Критические значения при уровне значимости* $\alpha = 0{,}05$: ранговых статистик Вилкоксона и Сиджела—Тьюки 2468,8—3131,2, их t-преобразований 1,96; параметрической статистики t Вэлча 1,96; параметрической статистики M Бартлета 3,84.

степени свободы, то нулевую гипотезу следует отклонить как неподтвердившуюся, т. е. признать справедливым альтернативное утверждение о существенности различий в дисперсиях и что степень рассеяния молибдена по гранитам скважины существенно выше, чем в гранитах поверхности Эльджуртинского массива.

Таким образом, ранговый критерий Сиджела—Тьюки не устанавливает различия в дисперсиях, в то время как параметрические критерии Фишера и Бартлета их подтвердили.

Расхождение в выводах приводит к необходимости большей осмотрительности при геологической интерпретации заключений о различиях в дисперсиях. Одновременно можно предположить, что ранговый критерий Сиджела—Тьюки оказался менее мощным, чем Фишера и Бартлета.

Результаты выявления сходства-различий в средних и дисперсиях по другим рудным, редким и петрогенным элементам в эльджуртинских гранитах Тырныауза приведены в табл. 2.

2.3.2. Иерархические агломеративные процедуры классификации набора объектов, основанные на критериях Вилкоксона, Вэлча, Сиджела—Тьюки и Бартлета

Агломеративная иерархическая кластерная процедура, использующая понятие постоянного порога, напоминает в основных чертах эффективную парагрупповую процедуру Р. Мак-Кеммона и Г. Венингера [5, с. 99—103; 1, с. 130—136], т. е. сохраняет все ее достоинства. Указанная процедура связана с вычислением ранговых и параметрических статистик (в нашем случае статистик Вилкоксона, Вэлча, Сиджела—Тьюки и Бартлета) между всеми парами объектов и объединением на каждом шагу той пары, для которой достигается минимум статистики, при условии, что эта величина меньше критического (порогового) значения.

Иерархическая кластерная процедура позволяет осуществлять законченную однозначную классификацию набора геологических объектов. В этом разделе проиллюстрирована процедура классификации геологических объектов относительно среднего и дисперсии одного изучаемого геологического признака (содержания элемента, минерала и др.).

Для проведения классификации объектов на основе ранговых тестов Вилкоксона и Сиджела—Тьюки потребовалось дополнительно осуществить с помощью центрирования и нормирования так называемое t-преобразование (гауссовское преобразование). Строго говоря, такое t-преобразование наиболее правомерно при больших объемах наблюдений ($n>30$). Приводимые примеры с лейкократовыми гранитоидами лишь частично удовлетворяют указанному условию, но это существенно не может сказаться на выводах, поскольку надежность, устойчивость этих заключений основываются на комплексировании ранговых и параметрических критериев.

Мы сочли возможным привести три (вместо одного) расчета по критерию Вэлча, так как пример с содержаниями молибдена оказался очень типичным для геологической практики — речь идет о наличии аномальных, «ураганных», содержаний в выборках геологических данных (в лейкократовых гранит-порфирах «Паука»).

2.3.2.1. Процедура классификации относительно среднего, основанная на ранговой статистике Вилкоксона

Задача заключалась в проверке предположения, что средняя концентрация молибдена в лейкократовых гранитоидах Тырныауза одна и та же в северном и южном блоках, на поверхностном эрозионном срезе и на глубине. Аргументированное возражение против указанного предположения может возникнуть после тщательного геологического и статистического анализа представительных данных о содержаниях молибдена во всех основных типах лейкократовых гранитоидов Тырныауза.

Формализация геологической задачи сводится к проверке гипотезы: $H_0: a_1=a_2=a_3=a_4=a_0$ при наборе альтернатив $H_1: a_u \neq a_0$ хотя бы для одного объекта u ($u=1, 2, 3, 4$), где a_u — математические ожидания — модели средних содержаний в u-х объектах.

Иерархическая процедура классификации набора k объектов, основанная на критерии Вилкоксона, заключается в следующем.

1. Для всех пар объектов (выборок) рассчитывают W статистики Вилкоксона (см. раздел 2.3.1) и соответствующие t-преобразования. Последние вычисляют по формуле [4]

$$t_{uv} = \frac{W_{uv} - 0{,}5n_u(n_u+n_v+1)}{\sqrt{\frac{n_u n_v}{12}(n_u+n_v+1)}}, \qquad n_u \leqslant n_v.$$

2. Из всех значений t выбирается минимальное, которое сравнивается с критическим значением — квантилем нормального распределения $t_{\alpha/2}$ (при уровне значимости $\alpha=0{,}05$ $t_{\alpha/2}=1{,}96$).

Если окажется $\min t_{uv} > t_{\alpha/2}$, то исследования прекращаются, принимается альтернатива и все k объектов (выборок) признаются статистически неоднородными относительно проверяемого параметра, в нашем случае относительно среднего.

Если же $\min t_{uv} \leqslant t_{\alpha/2}$, то та пара объектов ($u$, v), для которой $t_{uv}=\min\limits_{u,\,v} t$, объединяется в одну однородную группу. Однородная группа теперь будет иметь объем (n_u+n_v) наблюдений.

3. На втором этапе исследований используются ранее рассчитанные значения W и t между всеми парами объектов (выборок), кроме u, v, а также рассчитываются значения W и t между однородной группой и всеми остальными.

4. Из всей матрицы рассчитанных величин t выбирается минимальное $t_{u'v'}$ и вновь сравнивается с допустимым $t_{\alpha/2}$.

Если $\min t_{u'v'} > t_{\alpha/2}$, то исследования прекращаются и все объекты (выборки), кроме одной однородной группы, признаются

статистически неоднородными. Если же $\min t_{u'v'} \leqslant t_{\alpha/2}$, то в новую однородную группу объединяются объекты (u', v'). Заметим, что $\min t_{u'v'}$ может соответствовать какому-либо объекту u' и ранее выделенной однородной группе. Тогда на втором этапе нахождения однородных групп мы получим группу с объектами (u', v'), а именно объединенную группу с объектами (u, v, u').

5. Последовательная процедура проверки гипотез об однородности, использующая попарные объединения по минимальному значению преобразованной (из статистики Вилкоксона) величины t, продолжается до тех пор, пока не будет достигнут k—1 шаг объединения (т. е. принятие нулевой гипотезы) или на каком-либо k—h (h=2, 3, ..., k—2) шаге минимальное значение t не превзойдет допустимое $t_{\alpha/2}$. В первом случае исследователь с помощью критерия Вилкоксона получает подтверждение о существовании единственной однородной относительно среднего группы из k объектов, а во втором случае — h объектов выделены при классификации в однородные группы, а k—h объектов характеризуются существенно различающимися средними.

Пример. Как выше указывалось, проверяется предположение, что среднее содержание молибдена в лейкократовых гранитоидах Тырныауза одно и то же. Формальная постановка этой задачи также выше рассматривалась: $H_0: a_1 = a_2 = a_3 = a_4 = a_0$ при наборе альтернатив $H_1: a_u \neq a_0$, где a_1 — математическое ожидание (среднее) случайной величины — модели содержаний молибдена в лейкократовых гранит-порфирах «Паука» (n_1=27); a_2 — то же, для балки «Самолет» (n_2=49); a_3 — то же, для аплит-порфиров Северного участка (n_3=38); a_4 — то же, для вертикального разреза лейкократовых гранитов Северного участка (n_4=20).

Таблица 3

Значения W и t критерия Вилкоксона

Статистика	Сопоставляемые выборки (объекты)					
	1—2	1—3	1—4	2—3	2—4	3—4
W	1570,5	874	241	2308,5	498,5	252,5
t	5,76	—0,23	—5,14	5,45	—2,66	—5,52

Аналогично описанному в разделе 2.3.1 определяем между всеми парами объектов статистики W критерия Вилкоксона и их t-преобразования (табл. 3). Вместо процедуры t-преобразования можно было воспользоваться $t^2 = \chi^2$-статистикой [42]

$$\chi^2 = \frac{[W_{uv} - 0{,}5n_u(n_u + n_v + 1)]^2}{\frac{n_u n_v}{12}(n_u + n_v + 1)}$$

в условиях нулевой гипотезы, распределенной по закону Пирсона χ^2 с одной степенью свободы.

Как видно из табл. 3, минимальное значение $|t|=0{,}23$, соответствующее объектам (выборкам) 1 и 3, значительно меньше критического $t_{\alpha/2}=1{,}96$ при уровне значимости 0,05. Поэтому эти объекты подлежат объединению, так как они однородны относительно среднего. Вновь вычисляем значения W и t с учетом однородной группы, состоящей из 1-го и 3-го объектов (табл. 4).

Таблица 4

Значения W и t критерия Вилкоксона после объединения 1-го и 3-го объектов

Статистика	Сопоставляемые выборки (объекты)		
	(1+3)—2	(1+3)—4	2—4
W	1650	285	498,5
t	—6,68	—5,96	—2,66

Как видно из табл. 4, минимальное значение $|t|=2{,}66$ больше критического $t_{\alpha/2}=1{,}96$ и поэтому дальнейшее объединение объектов статистически невозможно.

Таким образом, с помощью описанной процедуры классификации на основе рангового критерия Вилкоксона устанавливается однородная группа, т. е. близкое среднее содержание молибдена в лейкократовых гранит-порфирах «Паука» и в лейкократовых аплит-порфирах Северного участка. В то же время средние содержания молибдена в вертикальном разрезе лейкократовых гранитов Северного участка и в лейкократовых гранит-порфирах балки «Самолет» существенно отличаются и друг от друга, и от содержаний молибдена в установленной однородной группе.

2.3.2.2. Процедура классификации относительно среднего, основанная на параметрической статистике Вэлча

Эта процедура аналогична вышерассмотренной. Также между всеми парами выборок рассчитываются t-статистики критерия Вэлча, хотя можно было вычислять χ^2-статистики [42].

Воспользуемся упрощенной t-статистикой, которая в условиях нулевой гипотезы асимптотически приближается при значительных объемах наблюдений к нормальному распределению (см. 2.3.1.2).

Учитывая, что параметрические критерии очень чувствительны к наличию аномальных наблюдений, а выборка 1 содержит ураганное значение молибдена (430 г/т), вычислительные процедуры осуществим по трем вариантам: с ураганной пробой, без нее, с заменой ураганного содержания (430 г/т) на максимальное в выборке (59 г/т).

При объединении объектов (выборок) потребуется использование оценок генерального среднего и генеральной дисперсии, определяемых по выборочным оценкам:

$$\bar{x} = \frac{1}{N}\sum_{u=1}^{2} \bar{x}_u n_u = \frac{1}{N}(\bar{x}_1 n_1 + \bar{x}_2 n_2),$$

$$\bar{s}^2 = \frac{1}{N-2}\sum_{u=1}^{2} s_u^2 (n_u - 1) = \frac{1}{N-2}[s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)],$$

$$N = \sum_{u=1}^{2} n_u = n_1 + n_2.$$

Воспользуемся вышеохарактеризованным геологическим примером с четырьмя типами лейкократовых гранитоидных пород. Оценки параметров в этих четырех выборках приведены в табл. 5.

Таблица 5

Таблица исходных данных по лейкократовым гранитоидам Тырныауза

Оценки параметров	Выборка 1			Выборка 2	Выборка 3	Выборка 4
	с ураганной пробой (вариант I)	без ураганной пробы (вариант II)	с заменной ураганной пробы на максимальную (вариант III)			
$\bar{x}$	36,66	21,53	22,92	2,98	36,91	1,35
s^2	6445,97	277,1	318,44	11,78	1765,49	1,56
n	27	26	27	49	38	20

Рассчитанные значения статистики t Вэлча приведены в табл. 6.

Таблица 6

Значения t критерия Вэлча

Вариант	Выборки					
	1—2	1—3	1—4	2—3	2—4	3—4
I	2,18	—0,01	2,28	—4,96	2,86	5,21
II	5,62	—2,03	5,15	—4,96	2,86	5,21
III	5,75	—1,83	6,25	—4,96	2,86	5,21

Как видно из табл. 6, согласно I и III вариантам объединению подлежат выборки 1 и 3, так как $\min t < t_{\alpha/2} = 1{,}96$, а по II варианту получаем крайне незначительное превышение критического значения: $|-2{,}03| > 1{,}96$, т. е. фактически попадаем в зону неопределенности в принятии или отклонении нулевой гипотезы

о равенстве средних. Авторы считают, что наиболее приемлем вариант III с заменой ураганного содержания молибдена на максимальное в оставшейся выборке. Это часто делается, например, в наиболее ответственных вычислительных операциях, связанных с подсчетом запасов в рудных месторождениях.

Выводы, полученные с помощью ранговой статистики Вилкоксона (см. табл. 3) и параметрического критерия Вэлча (см. табл. 6), убедительно свидетельствуют об однородности, близости средних содержаний молибдена в 1-м и 3-м объектах.

Объединим эти выборки и определим усредненные оценки параметров объединенной выборки, причем расчеты и выводы сделаем только по выбранному нами III варианту

$$n_{1+3} = n_1 + n_3 = 27 + 38 = 65,$$

$$\bar{x}_{1+3} = \frac{1}{n_{1+3}}(\bar{x}_1 n_1 + \bar{x}_3 n_3) = \frac{1}{65}(22{,}92 \cdot 27 + 36{,}91 \cdot 38) = 31{,}1,$$

$$s^2_{1+3} = \frac{1}{n_{1+3} - 2}[s_1^2 (n_1 - 1) + s_3^2 (n_3 - 1)] =$$

$$= \frac{1}{63}(318{,}44 \cdot 26 + 1765{,}49 \cdot 37) = 1168{,}29.$$

Находим значения статистики t критерия Вэлча (табл. 7).

Таблица 7

Значения t критерия Вэлча после объединения 1-го и 3-го объектов

Статистика	Выборки				
	(1+3)—2		(1+3)—4		2—4
	Вариант III	Вариант I	Вариант III	Вариант I	
t	6,59	(4,51)	7,0	(4,73)	2,86

Так как минимальное $t = 2{,}86$ превысило критическое значение $t_{\alpha/2} = 1{,}96$, то при уровне значимости 0,05 следует сделать заключение о статистической невозможности производить дальнейшее объединение объектов.

Таким образом, с помощью описанной процедуры классификации на основе параметрического критерия Вэлча устанавливается единственная однородная группа объектов 1 и 3 с выборочным средним содержанием молибдена $\bar{x}_{1+3} = 31{,}1$ г/т, которое значительно выше, чем во 2-м ($\bar{x}_2 = 2{,}98$ г/т) и 4-м ($\bar{x}_4 = 1{,}35$ г/т) объектах. Заключение совпадает с тем, которое получено на основе рангового критерия Вилкоксона, и поэтому представляется вполне убедительным, включая суждения о существенной разнице в истинных средних содержаниях молибдена во 2-м и 4-м объектах (в гранит-порфирах «Паука» и в вертикальном разрезе лейкократовых гранитов Северного участка).

2.3.2.3. Процедура классификации относительно степени рассеяния (дисперсии), основанная на ранговой статистике Сиджела—Тьюки

Геологическая постановка задачи сводится к проверке гипотезы, что степень рассеяния молибдена одна и та же в лейкократовых гранитоидах Тырныауза.

Формальный аналог этой задачи: $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_0^2$ при наборе альтернатив: $H_1: \sigma_u^2 \neq \sigma_0^2$ хотя бы для одного объекта u ($u=1, 2, 3, 4$), где σ_u^2 — дисперсии — модели степени рассеяния молибдена в u-х объектах.

Иерархическая процедура классификации набора k объектов, основанная на критерии Сиджела—Тьюки, сводится к следующим вычислительным операциям.

1. Для всех пар объектов (выборок) определяют R-статистики Сиджела—Тьюки (см. раздел 2.3.1) и соответствующие им по схеме Б t-статистики (или χ^2-распределенные случайные величины с одной степенью свободы).

2. Из всех значений t выбирается минимальное и сравнивается с критическим значением $t_{\alpha/2}$.

При $\min t_{uv} > t_{\alpha/2}$ исследования прекращаются и все k объектов признаются неоднородными с различающимися дисперсиями. Если $\min t_{uv} \leqslant t_{\alpha/2}$, то пара объектов u и v объединяется в одну однородную по дисперсии группу.

3. Вновь вычисляются R и t с учетом статистик R и t между ранее установленной однородной группой объектов и всеми оставшимися.

4. Выбирается минимальное $t_{u'v'}$ и сравнивается с критическим $t_{\alpha/2}$.

При $\min t_{u'v'} > t_{\alpha/2}$ исследования прекращаются и все k—2 объекта, кроме установленной однородной группы, признаются неоднородными, т. е. объектами с различающимися дисперсиями.

Если $\min t_{u'v'} \leqslant t_{\alpha/2}$, то новая пара объектов (u', v') или три объекта (u', u, v) объединяются в однородную группу.

5. Последовательная процедура проверки гипотез об однородности относительно дисперсий продолжается либо до k—1 шага объединения и принятия нулевой гипотезы, либо до k—h шага, когда окажется, что $\min t > t_{\alpha/2}$. В первом случае мы подтвердим наличие одной однородной группы объектов, а во втором — h объектов при классификации объединились в однородные группы, в то время как k—h объектов характеризуются существенно различающимися дисперсиями.

В качестве геологического примера проверяется гипотеза, что дисперсии (степени рассеяния молибдена) в лейкократовых гранитоидах Тырныауза одни и те же. Другими словами, формальный аналог этой геологической задачи имеет вид: $H_0 = \sigma_1^2 = \sigma_4^2 = \sigma_3^2 = \sigma_4^2 = \sigma_0^2$ при наборе альтернатив $H_1: \sigma_u^2 \neq \sigma_0^2$, $u=1, 2, 3, 4$, где σ_1^2 — дисперсия случайной величины, рассматриваемой моделью содержаний молибдена в гранит-порфирах «Паука»;

σ_2^2 — то же, для балки «Самолет»; σ_3^2 — то же, для аплит-порфиров Северного участка и σ_4^2 — для вертикального разреза гранитов Северного участка.

Аналогично описанному в разделе 2.3.1 найдем по центрированным данным для всех пар объектов значения статистик R и t (табл. 8).

Таблица 8

Значения R и t критерия Сиджела—Тьюки

Статистики	Сопоставляемые выборки					
	1—2	1—3	1—4	2—3	2—4	3—4
R	463	1173	690,2	778	923	964
t (по схеме Б)	—11,21	3,76	4,53	—7,65	2,96	6,13

Медианы выборок: 1-й—20 г/т, 2-й—1,8 г/т, 3-й—25,5 г/т, 4-й—0,7 г/т Mo.

Как видно из табл. 8, минимальное значение $t=2{,}96$ между вторым и четвертым объектами превысило допустимый квантиль нормального распределения $t=1{,}96$, поэтому на основе ранговой статистики Сиджела—Тьюки принимается статистическое решение о невозможности объединения выборок в какие-либо однородные группы. Другими словами, лейкократовые гранитоиды Тырныауза характеризуются существенно различающейся между собой изменчивостью (степенью рассеивания) содержаний молибдена.

2.3.2.4. Процедура классификации относительно степени рассеяния (дисперсии), основанная на параметрической статистике Бартлета

В основных чертах она совпадает с теми процедурами, которые базировались на критериях Вилкоксона, Вэлча, Сиджела—Тьюки.

Вновь воспользуемся сравнением дисперсий молибдена в четырех типах лейкократовых гранитоидов Тырныауза. Нулевая гипотеза H_0 и набор альтернатив были сформулированы выше.

Учитывая, что в первой выборке (в гранит-порфирах «Паука») содержится аномальное, ураганное, содержание молибдена (430 г/т), существенно влияющее на оценку s^2 дисперсии и соответственно на мощность критерия Бартлета, вычисления производим по трем вариантам: с ураганной пробой, без ураганной пробы, с заменой ураганной пробы (430 г/т) на максимальное содержание молибдена без нее (59 г/т). Вычисляем между всеми парами объектов значения статистики Бартлета (см. раздел 2.3.1).

Анализ табл. 9 показал, что минимальные значения статистики Бартлета 12,69, 19,12 и 18,3 оказались больше критического значения $\chi^2_{\alpha,\, f\approx 1}=3{,}84$. Поэтому на базе критерия Бартлета нет

оснований к выделению однородных по дисперсии групп объектов, т. е. должно быть признано, что у всех геологических объектов существенно различная изменчивость в распределении молибдена. Наименьшей изменчивостью характеризуются лейкократовые граниты в вертикальном разрезе Северного участка (оценка дисперсии молибдена — 1,56), наибольшей (если заменить ураган-

Таблица 9

Значения M критерия Бартлета

Варианты	Сопоставляемые выборки					
	1—2	1—3	1—4	2—3	2—4	3—4
I	222,14	12,69	130,59	168,50	19,12	108,23
II	78,98	19,48	72,13	168,50	19,12	108,23
III	84,51	18,30	71,67	168,50	19,12	108,23

ную пробу в гранит-порфирах «Паука») — лейкократовые аплит-порфиры в поверхностном эрозионном срезе на том же Северном участке (оценка дисперсии — 1756,5).

Указанное заключение полностью совпало с тем, которое было получено с помощью ранговой статистики Сиджела—Тьюки, поэтому представляется вполне обоснованным и надежным.

2.3.3. Параметрические и ранговые статистические методы Ватсона—Вильямса и Вилера—Ватсона—Ходжеса обработки угловых наблюдений

Успех статистической обработки угловых наблюдений, играющих большую роль при геологическом картировании, изучении литологии и стратиграфии, структурной геологии (включая трещинную тектонику), исследовании явлений палеомагнетизма, обеспечивается применением специальных математических процедур.

Попытка применить обычные методы обработки данных к угловым наблюдениям может привести к парадоксальным результатам. Например, среднее арифметическое азимутов падения пластов 1°, 359° равно 180°, тогда как геологическая интуиция справедливо подсказывает, что среднее должно быть 0°. Кроме того, наблюдаются серьезные искажения при использовании обычной выборочной дисперсии для оценки изменчивости угловых наблюдений. В том же примере классическая оценка дисперсии и стандартного отклонения, произведенная по формулам $s^2 = \frac{1}{n-1} \times$ $\times \sum_t (x_t - \bar{x})^2$; $s = \sqrt{s^2}$, приведет к нелепым результатам. В то же время ясно, что отклонение от среднего направления равно примерно 1°.

В настоящем разделе охарактеризованы наиболее простые ранговые и параметрические критерии проверки гипотез по данным

угловых наблюдений (гипотез о средних направлениях). Читателей, желающих углубленно изучить проблему обработки угловых наблюдений, мы отсылаем к фундаментальной работе К. Мардиа [31]. Из этой работы для обработки угловых наблюдений в двух геологических объектах нами отобраны два эффективных метода для проверки гипотезы о равенстве средних направлений по угловым данным: критерий Ватсона—Вильямса и ранговый критерий равномерных меток Вилера—Ватсона—Ходжеса.

Полагаем, что иерархическая процедура (см. раздел 2.3.2) классификации набора объектов, охарактеризованных угловыми наблюдениями на основе указанных двух выборочных критериев, составит серьезную конкуренцию аналогичным процедурам, описанным К. Мардиа на основе многовыборочных критериев [31, с. 169—173, 204—205].

Для изложения статистических критериев обработки угловых наблюдений потребуются некоторые специфические понятия, кратко приводимые ниже.

Использовав окружность единичного радиуса, отдельное угловое наблюдение ν_t ($0° < \nu_t \leqslant 360°$) можно представить в виде угла, заключенного между вектором единичной длины и положительным направлением оси абсцисс и измеряемого в направлении против движения часовой стрелки. Декартовые координаты вектора будут $\cos \nu_t$ и $\sin \nu_t$, а полярные: 1 и ν_t.

Графические негруппированные угловые наблюдения можно изображать точками на окружности или векторами, соединяющими эти точки с началом координат (рис. 4, *а*). Группированные данные удобно изображать в виде круговых гистограмм (там же, *б*), линейных гистограмм (там же, *в*) и диаграмм розы наблюдений (там же, *г*). С уточняющими моментами в их построении можно ознакомиться у К. Мардиа [31, с. 12—14], у которого и были заимствованы схематические рисунки.

Введем некоторые обозначения. Предположим, что $\nu_1, \ldots, \nu_t, \ldots, \nu_n$ — выборка объема n на окружности, соответствующая функции распределения $F(\theta)$ некоторых параметров θ, рассмотренных ниже. Каждый из углов ν_t, образованный радиусами OP и OP_t (см. рис. 4, *а*), соответствует дуге окружности единичного радиуса. Дуга отсчитывается против часовой стрелки; за начало отсчета принимается точка пересечения окружности с положительной полуосью абсцисс. Пусть P_t — точка на окружности единичного радиуса, соответствующая углу ν_t ($t = 1, 2, \ldots, n$). Тогда выборочное среднее направление определится как направление суммы единичных векторов $\overrightarrow{OP}_1, \ldots, \overrightarrow{OP}_n$. Декартовы координаты точки P_t будут $(\cos \nu_t, \sin \nu_t)$, $t = 1, 2, \ldots, n$. Если всем этим точкам приписать одинаковую «массу» $1/n$, то координаты «центра масс» $(\bar{c}, \bar{s})$ можно выразить формулами [31, с. 24]

$$\bar{c} = \frac{1}{n} \sum_{t=1}^{n} c_t = \frac{1}{n} \sum_{t=1}^{n} \cos \nu_t \quad \text{и} \quad \bar{s} = \frac{1}{n} \sum_{t=1}^{n} s_t = \frac{1}{n} \cdot \sum_{t=1}^{n} \sin \nu_t,$$

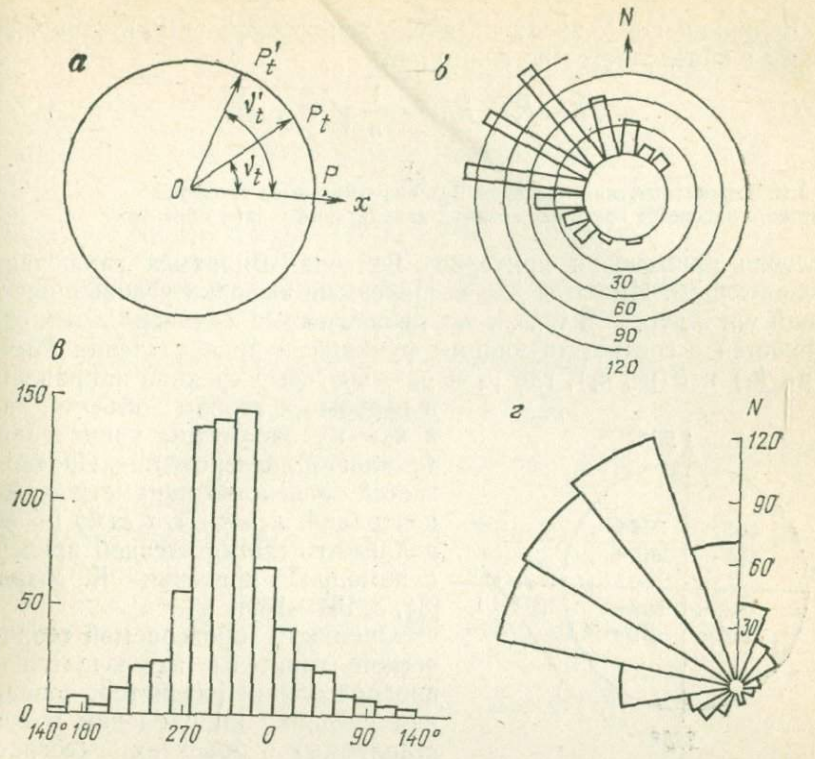


Рис. 4. Графическое представление угловых наблюдений:

а — исходные несгруппированные данные; *б* — круговая гистограмма сгруппированных наблюдений; *в* — линейная гистограмма сгруппированных наблюдений; *г* — диаграмма розы сгруппированных наблюдений

где $c_t = \cos v_t$ и $s_t = \sin v_t$;

$$c = \sum_{t=1}^{n} \cos v_t \quad \text{и} \quad s = \sum_{t=1}^{n} \sin v_t.$$

Если предположить $r = \bar{R} = \sqrt{\bar{c}^2 + \bar{s}^2}$, то $R = rn = \bar{R}n$ будет представлять собой длину вектора $\overrightarrow{OP}_1, \ldots, \overrightarrow{OP}_n$, направление которого — выборочное круговое среднее направление m, являющееся решением любого из двух уравнений: $\bar{c} = r \cos m$ или $\bar{s} = r \sin m$, т. е., другими словами, $m = \arccos(\bar{c}/r)$, или $m = \arcsin(\bar{s}/r)$, или $m = \operatorname{arctg}(\bar{s}/\bar{c})$.

Согласно К. Мардиа [39, с. 26], выборочную круговую дисперсию направлений $v_1, \ldots, v_n$ (обозначим ее $0 \leqslant v \leqslant 1$) определяют по формуле

$$v(m) = 1 - r = 1 - \bar{R}, \quad \text{причем} \quad 0 \leqslant v \leqslant 1.$$

Величина $r=\overline{R}$ носит название выборочной результатирующей длины и определяется по формуле

$$r=\overline{R}=R/n=\frac{1}{n}\sqrt{\bar{c}^2+\bar{s}^2}.$$

2.3.3.1. Параметрический критерий Ватсона—Вильямса проверки гипотез о равенстве круговых средних направлений в двух объектах

Условия применения критерия Ватсона—Вильямса заключается в следующем. Имеются две независимые выборки угловых наблюдений $v_{11}, \ldots, v_{1n_1}$ и $v_{21}, \ldots, v_{2n_2}$, извлеченные из генеральных совокупностей с соответствующими функциями распределения Мизеса $F(\mu_1, k_1)$ и $F(\mu_2, k_2)$, где μ_1 и μ_2 — круговые средние направления в первом и втором объекте, а k_1 и k_2 — их параметры концентраций (аналог дисперсий). Предполагается равенство параметров концентраций: $k_1=k_2$, что само по себе подлежит статистической проверке с помощью критерия К. Мардиа [31, с. 164—169].

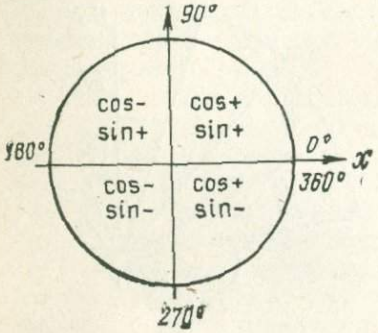


Рис. 5. Знаки тригонометрических функций

Сущность проверяемой геологической гипотезы заключается в предположении равенства, совпадения средних направлений в сопоставляемых объектах (согласно конкурирующему утверждению эти средние направления существенно различаются). Если с каким-то направлением связано оруденение, то статистические сопоставления с указанным направлением (допустим, с направлением потенциально рудоносных трещин) вооружат геолога дополнительными прямыми или косвенными поисковыми признаками на изучаемый тип минерализации и оруденения.

В формальном виде нулевая гипотеза H_0 и набор альтернатив H_1 имеют следующий вид: $H_0: \mu_1=\mu_2$, $H_1: \mu_1\neq\mu_2$.

Естественно, что истинные значения параметров μ_u и k_u $(u=1,2)$ геологу неизвестны, и он имеет дело только с оценками этих и других величин.

Процедура вычислений следующая.

1. С учетом знаков (рис. 5) по угловым данным каждой выборки находят косинусы и синусы, рассчитывают их суммы (т. е. величины c и s), а также координаты «центра масс» $\bar{c}$ и $\bar{s}$:

$$c_1=\sum_{t=1}^{n_1}\cos v_{1t};\quad s_1=\sum_{t=1}^{n_1}\sin v_{1t};\quad \bar{c}_1=\frac{1}{n_1}c_1;\quad \bar{s}_1=\frac{1}{n_1}s_1;$$

$$c_2=\sum_{t=1}^{n_2}\cos v_{2t};\quad s_2=\sum_{t=1}^{n_2}\sin v_{2t};\quad \bar{c}_2=\frac{1}{n_2}c_2;\quad \bar{s}_2=\frac{1}{n_2}s_2.$$

2. Определяем по каждой выборке суммарные величины R_1 и R_2, а также выборочные результатирующие длины $\bar{R}_1=r_1$ и $\bar{R}_2=r_2$:

$$R_1=\sqrt{c_1^2+s_1^2}\ ; \quad \bar{R}_1=\frac{1}{n_1}R_1;$$

$$R_2=\sqrt{c_2^2+s_2^2}\ ; \quad \bar{R}_2=\frac{1}{n_2}R_2.$$

3. Определяем по каждой выборке круговые средние направления m_1 и m_2 и дополнительно круговые дисперсии v_1 и v_2:

$$m_1=\arccos(\bar{c}_1/\bar{R}_1); \quad m_2=\arccos(\bar{c}_2/\bar{R}_2);$$

$$v_1=1-\bar{R}_1; \quad v_2=1-\bar{R}_2.$$

4. Определяем векторную величину $R: R=\sqrt{R_1^2+R_2^2+2R_1R_2\cos(m_2-m_1)}$.

5. Вычисляем статистику Ватсона—Вильямса $\bar{R}'$, а также величины $\bar{R}$, $\hat{k}$, η:

$$\bar{R}'=\frac{R_1+R_2}{n}, \quad \text{где} \quad n=n_1+n_2;$$

$$\bar{R}=\frac{R}{n}; \quad \eta=\frac{n_1}{n}; \quad \hat{k}=A^{-1}(\bar{R}),$$

где параметр концентрации распределения Мизеса берется из [3, прил. 3].

6. В зависимости от значения статистики R расчет производится по четырем вариантам (схемам).

6а. Случай, когда $0<\bar{R}<0{,}4$ (дополнительно предполагается, что $\frac{1}{3}<\eta<\frac{1}{2}$).

Критические значения статистики Ватсона и Вильямса $\bar{R}'_{кр}$ рассчитываются по специальным номограммам [31, прил. 9а и 9б] с помощью интерполяции по значениям $\bar{R}'$ при $\eta=\frac{1}{3}$ и $\eta=\frac{1}{2}$.

6б. Случай, когда $0{,}4<\bar{R}<0{,}7$ (и при умеренных значениях η). Используются критические значения критерия $\bar{R}'$ Ватсона и Вильямса по упомянутому приложению 9а [31].

6в. Случай, когда $0{,}7<\bar{R}<0{,}98$.

Пользуются скорректированной F-статистикой

$$F_{скор}=\left(1+\frac{3}{8\hat{k}}\right)\frac{(n-2)(R_1+R_2-R)}{n-R_1-R_2},$$

которая при больших $\hat{k}$ приближенно подчиняется F-распределению Фишера с $f_1=1$ и $f_2=n-2$ степенями свободы.

6г. Случай, когда $\bar{R}>0{,}98$.

Можно воспользоваться нескорректированной F-статистикой

$$F = \frac{(n-2)(R_1 + R_2 - R)}{n - R_1 - R_2},$$

имеющей в условиях нулевой гипотезы вышеуказанное асимптотическое F-распределение.

7. Нулевая гипотеза о равенстве круговых средних направлений в двух объектах принимается как подтвердившаяся, если $\bar{R}' \leqslant \bar{R}'_{кр}$ или $F \leqslant F_{q, f_1=1, f_2=n-2}$, и отклоняется т. е. принимается альтернатива о существенности различий в круговых средних направлениях, когда $\bar{R}' > \bar{R}'_{кр}$ или $F > F_{q, f_1, f_2}$. Опыт показывает, что при массовой обработке угловых геологических данных оправданно пользоваться в основном нескорректированной F-статистикой (см. 6г).

Пример. Пусть в распоряжении геолога имеются замеры азимутов падения рудных и безрудных прожилков, сосредоточенные соответственно в рудоносной и безрудной частях редкометального штокверка: в рудоносной части v_{1t} — 75, 75, 80, 80, 80, 95, 130, 170, 216°; $n=9$; в безрудной части v_{2t} — 10, 50, 55, 55, 65, 90, 285, 285, 325, 355°; $n=10$.

Утверждение геолога заключается в том, что средние направления рудоносных и пустых кварцевых прожилков одни и те же в штокверке, другими словами, — простирание и падение прожилков не могут служить диагностирующим оруденение (прямым или косвенным) признаком. Согласно противоположному суждению в рудоносной части штокверка среднее направление (средний азимут падения) кварцевых прожилков существенно отличается от их среднего направления в безрудной части штокверка и признак «направление» может играть роль индикатора оруденения.

Формальный аналог этой задачи имеет вид: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, где μ_1 — истинное среднее случайной величины — модели азимутов падения кварцевых прожилков с оруденением, μ_2 — то же, для пустых прожилков.

Алгоритм вычислений следующий.

1. Определяем сумму косинусов и синусов азимутов падения в градусах: $c_1 = -1{,}4854$, $s_1 = 6{,}2342$, $c_2 = 5{,}5304$ и $s_2 = 1{,}8917$.

2. Вычисляем средние величины: $\bar{c}_1 = -1{,}4854/9 = -0{,}1650$; $\bar{s}_1 = 6{,}2342/9 = 0{,}6927$; $\bar{c}_2 = 5{,}5304/10 = 0{,}5530$ и $\bar{s}_2 = 1{,}8917/10 = 0{,}1892$.

3. Находим величины R_1 и R_2, а также результирующие длины $\bar{R}_1$ и $\bar{R}_2$:

$$R_1 = \sqrt{c_1^2 + s_1^2} = \sqrt{(-1{,}4854)^2 + (6{,}2342)^2} = 6{,}4087;$$

$$R_2 = \sqrt{c_2^2 + s_2^2} = \sqrt{(5{,}5304)^2 + (1{,}8917)^2} = 5{,}8450;$$

$$R_2 = \frac{R_1}{n_1} = \frac{6{,}4087}{9} = 0{,}7121; \qquad \bar{R}_2 = R_2/n_2 = 5{,}845/10 = 0{,}5845.$$

4. Определяем выборочные круговые средние направления m_1 и m_2 и дисперсии v_1 и v_2:

$$m_1 = \arccos(\bar{c}_1/\bar{R}_1) = \arccos\left(\frac{-0{,}1650}{0{,}7121}\right) \simeq 103{,}4°;$$

$$m_2 = \arccos(\bar{c}_2/\bar{R}_2) = \arccos\left(\frac{0{,}5530}{0{,}5845}\right) \simeq 18{,}9°;$$

$v_1 = 1 - \bar{R}_1 = 1 - 0{,}7121 = 0{,}2879$ и $v_2 = 1 - \bar{R}_2 = 1 - 0{,}5845 = 0{,}4155.$

5. Вычисляем векторную величину R:

$$R = \sqrt{\bar{R}_1^2 + \bar{R}_2^2 + 2R_1R_2\cos(m_2 - m_1)} =$$

$$= 6{,}4087^2 + 5{,}845^2 + 2 \cdot 6{,}4087 \cdot 5{,}845\cos(18{,}9° - 103{,}4°) = 9{,}0783.$$

6. Вычисляем статистику Ватсона—Вильямса $\bar{R}'$, а также величины $\bar{R}$, η и $\widehat{k}$:

$$\bar{R}' = (R_1 + R_2)/n = (6{,}4087 + 5{,}845)/19 = 0{,}64;$$

$$\bar{R} = R/n = 9{,}0783/19 \simeq 0{,}48;$$

$$\eta = n_1/n = 9/19 = 0{,}47;$$

$$\widehat{k} = A^{-1}(\bar{R}) = A^{-1}(0{,}48)\,\bar{k} = 1{,}09788 \quad [31, \text{прилож. } 3].$$

7. Вычисляем F-статистику и скорректированную $F_{\text{скор}}$-статистику:

$$F = \frac{(n-2)(R_1 + R_2 - R)}{n - R_1 - R_2} = \frac{(19-2)(6{,}4087 + 5{,}8450 - 9{,}0783)}{19 - 6{,}4087 - 5{,}845} \simeq 8{,}00,$$

$$F_{\text{скор}} = \left(1 + \frac{3}{8\widehat{k}}\right) \qquad F = \left(1 + \frac{3}{8 \cdot 1{,}09788}\right)8{,}0 \simeq 10{,}72.$$

8. По схемам 6б, 6в и 6г осуществим проверку нулевой гипотезы $H_0: \mu_1 = \mu_2$, хотя наиболее обоснованной является схема 6б (так как $0{,}4 < \bar{R}' = 0{,}48 < 0{,}7$).

По схеме 6б, поскольку $\bar{R}' = 0{,}48 > 0{,}4$, используем критические значения двухвыборочного критерия Ватсона и Вильямса [31, прилож. 9а], сведенные в номограмму с серией кривых для объемов наблюдений $n = 12$, 14, 16, 20, 24, 30, 40, 60, 120 и 240 при уровне значимости 0,05. По оси абсцисс отложены величины $\bar{R}$, по оси ординат — критические значения статистики Ватсона—Вильямса $\bar{R}'$ (предполагается также $n_1 \approx n_2$).

В нашем примере при $n = 19$, $\bar{R} = 0{,}48$ критическое значение $\bar{R}'_{\text{кр}}$ составляет $\approx 0{,}57$. Так как вычисленное значение статистики Ватсона и Вильямса превысило допустимое ($\bar{R}' = 0{,}64 > \bar{R}'_{\text{кр}} = 0{,}57$), нулевая гипотеза должна быть отклонена и принята альтернатива о существенности различий в круговых средних. Другими словами, среднее направление (азимут падения) кварцевых прожилков в рудной части штокверка существенно отличается в бóльшую сторону ($m_1 = 103{,}4°$, а $m_2 = 18{,}9°$) от среднего направления кварцевых прожилков в безрудной части, а следовательно, этот вывод

можно использовать для диагноза оруденелости прожилков (ориентироваться на азимуты падения в 100—105°).

По схеме 6в, поскольку вычисленная скорректированная статистика $F_{скор}=10{,}72$ превысила критическое F_{q, f_1, f_2}, равное при уровне значимости $q=0{,}05$, $f_1=1$ и $f_2=n-2=17$ степенях свободы 4,45, т. е. $10{,}72>4{,}45$, то вновь следует отклонить нулевую гипотезу и признать альтернативу о существенности различий в азимутах падения рудных и безрудных кварцевых прожилков в редкометальном штокверке.

По схеме 6г нескорректированная статистика $F=8{,}0$ также превысила критическое значение $F_{0{,}05;\,1;\,17}=4{,}45$, поэтому вновь отклоняется нулевая гипотеза и принимается альтернатива о статистически значимых различиях в азимутах падения рудных и безрудных прожилков.

Таким образом, с помощью наиболее простой схемы 6г сделано правильное заключение, хотя наиболее, казалось бы, оправданным было использовать для геологического примера схему 6б.

2.3.3.2. Ранговый критерий равномерных меток Вилера—Ватсона—Ходжеса проверки гипотез о равенстве круговых средних направлений в двух объектах

Представим угловые наблюдения обеих выборок точками на окружности единичного радиуса (на рис. 6 для 1-й выборки объема n_1 — черные точки, для 2-й выборки объема n_2 — белые, причем $n_1 \leqslant n_2$).

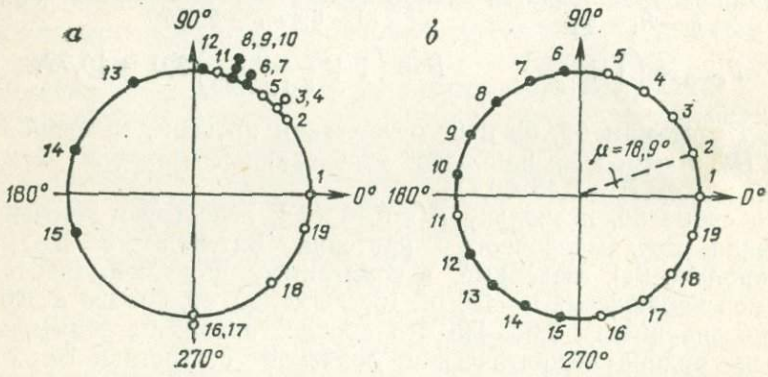


Рис. 6. Графическое представление угловых наблюдений по критерию равномерных меток Вилера—Ватсона—Ходжеса:
а — исходные данные двух выборок; *б* — равномерно расположенные данные двух выборок

Составим вариационный ряд в порядке возрастания угловых наблюдений и расположим их в виде точек на окружности. Обозначим через R_{1t}, $t=1, 2, \ldots, n_1$ — ранги (номера) угловых наблюдений 1-й выборки в вариационном ряду объема $n=n_1+n_2$, а через R_{2t}, $t=1, 2, \ldots, n_2$ — ранги 2-й выборки (см. рис. 6, *а*).

Изменим расстояния между последовательными точками так, чтобы все расстояния были одинаковыми и равными $2\pi/n$ (см. рис. 6, *б*), т. е. заменим угловые наблюдения в объединенной выборке точками $2\pi t/n$, $t=1, 2, \ldots, n$ на окружности так, что первой выборке будут соответствовать точки с углами β_t^0

$$\beta_t^0 = \frac{360° R_t}{n}, \qquad t = 1, 2, \ldots, n_1.$$

Ранги R_t угловых данных 1-й выборки объема $n_1=9$, изображенные на рис. 6, равны 6, 7, 8, 9, 10, 12, 13, 14 и 15.

С. Вилером и Г. С. Ватсоном, по предложению Дж. Л. Ходжеса [31; с. 197], рассмотрен критерий, основанный на функциях от рангов R_t, $t=1, 2, \ldots, n_1$, который характеризуется инвариантностью (независимостью) относительно поворотов и изменения направления упорядочения [31, с. 196].

Постановка геологической задачи проверки равенства средних направлений в сопоставляемых объектах и ее формализация полностью идентичны описанным в разделе о применении критерия Ватсона—Вильямса, поэтому они здесь не приводятся.

Процедура вычисления ранговой статистики B Вилера—Ватсона—Ходжеса такова.

1. Вышеуказанным способом (см. рис. 6) составляется на окружности единичного радиуса равномерная последовательность угловых наблюдений, которые ранжируются (нумеруются) с рангами R_{1t}, $t=1, 2, \ldots, n_1$ для 1-й выборки.

2. Рассчитываются углы β_t^0, $t=1, 2, \ldots, n_1$ для 1-й выборки

$$\beta_t^0 = \frac{360° R_t}{n}, \qquad n = n_1 + n_2.$$

3. По угловым данным 1-й выборки β_{1t} находят их косинусы и синусы и рассчитываются их суммы c_1, s_1

$$c_1 = \sum_{t=1}^{n_1} \cos\beta_{1t}, \qquad s_1 = \sum_{t=1}^{n_1} \sin\beta_{1t}.$$

4. Определяют статистику B: $B = R_1^2 = c_1^2 + s_1^2$.

5. Находят статистику R^*: $R^* = \frac{2(n-1)B}{n_1 n_2}$.

6. В зависимости от объема наблюдения $n=n_1+n_2$ учитываются две ситуации.

6а. Случай, когда $n \leqslant 20$.

Критические значения статистики B Вилера—Ватсона—Ходжеса берутся из специальной таблицы [31, прилож. 14].

6б. Случай, когда $n > 20$.

В условиях нулевой гипотезы о равенстве круговых средних направлений в двух объектах статистика R^* распределена приближенно, как χ^2 с двумя степенями свободы [31, с. 198].

7. Нулевая гипотеза о равенстве круговых средних направлений в двух объектах принимается как подтвердившаяся, если

$B \leqslant B_{кр}$ или $R^* \leqslant \chi^2_{q,\ f=2}$, и отклоняется, т. е. принимаются альтернативы о существенности различий в круговых средних направлениях, когда $B > B_{кр}$ или $R^* > \chi^2_{q,\ f=2}$.

В качестве геологического примера воспользуемся вышеприведенными данными замеров азимутов падения оруденелых и безрудных кварцевых прожилков в редкометальном штокверке. На рис. 6 приведены именно эти данные ($n_1=9$, $n_2=10$, $n=19$).

Алгоритм вычислений следующий.

1. Определяем углы β^0_t, косинусы и синусы этих углов

$$\beta^0_t = \frac{360^\circ R_t}{n} = \frac{360^\circ R_t}{19} = 18{,}95^\circ R_t,$$

где ранги — R_t, $t = 1, 2, \ldots, n$.

2. Определяем суммы косинусов c_1 и синусов s_1 для 1-й выборки ($n_1=9<n_2=10$)

$$c_1 = \sum_{t=1}^{n_1} \cos\beta_{1t} = (-0{,}4019 + \ldots + 0{,}2462) \simeq -4{,}8454,$$

$$s_2 = \sum_{t=1}^{n_1} \sin\beta_{1t} = (0{,}9157 + \ldots + (-0{,}9692) \simeq -1{,}4919.$$

3. Определяем статистику B

$$B = R_1^2 = c_1^2 + s_1^2 = (-4{,}8454)^2 + (-1{,}4919)^2 \simeq 25{,}7.$$

4. Находим статистику R^*

$$R^* = \frac{2(n-1)B}{n_1 n_2} = \frac{2(19-1)25{,}7}{9 \cdot 10} \simeq 10{,}28.$$

5. По обеим схемам 6а и 6б осуществим проверку нулевой гипотезы $H_0: \mu_1 = \mu_2$.

Так как статистика B критерия Вилера—Ватсона—Ходжеса превысила критическое $B_{кр}$ при уровне значимости $\alpha=0{,}05$, равное 14,58 [31, прилож. 14]: $B=25{,}7>B_{кр}=14{,}58$, то нулевую гипотезу следует отклонить и принять альтернативные гипотезы о существенности различий в круговых средних направлениях (в средних азимутах падения рудных и безрудных кварцевых прожилков). Аналогичное заключение следует сделать по схеме 6б, так как статистика $R^*=10{,}28$ превысила допустимое $\chi^2_{\alpha,\ f=2}=5{,}99$, т. е. $10{,}28>5{,}99$.

Учитывая, что заключение о различии средних азимутов падения рудных и безрудных кварцевых прожилков в редкометальном штокверке ранее было сделано с помощью параметрической статистики Ватсона—Вильямса, это суждение следует полагать надежным, т. е. его можно использовать для прогноза и диагноза рудоносности рассматриваемых прожилков. Другими словами, следует контролировать азимуты падения прожилков в 100—105° как возможный индикатор рудоносности.

2.4. РЕКОМЕНДУЕМЫЕ МНОГОМЕРНЫЕ РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ МАТЕМАТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

В разделе описаны некоторые ранговые и параметрические математические методы выявления сходства или различий многомерных средних и ковариационных матриц — характеристик изменчивости и коррелированности комплекса геологических признаков, изученных на двух и более геологических объектах*.

Проиллюстрируем типичную картину постановки геологической задачи выявления сходства или подтверждения различий в многомерных средних и ковариационных матрицах. Для простоты ограничимся двумя геологическими признаками в двух объектах, т. е. двумерными средними. Заметим, что это совершенно не отразится на общности ниже высказанных рассуждений.

Будем проверять, влияет ли фактор глубинности на средние концентрации вольфрама и молибдена в эльджуртинских гранитах. Вновь воспользуемся исходными данными В. В. Ляховича [30].

Проверяется предположение, что глубинность не влияет одновременно на средние концентрации вольфрама и молибдена (в отношении последнего это уже доказано — см. раздел 2.3.1). Альтернативная гипотеза заключается, естественно, в обратном, т. е. в том, что средние содержания вольфрама и молибдена в объеме гранитов Эльджуртинского интрузивного массива существенно отличаются от средних содержаний этих рудных компонентов в поверхностном эрозионном срезе.

Формальный аналог указанной геологической задачи имеет следующий вид.

Будем рассматривать многомерную (m-мерную) случайную величину $\xi = \{\xi_1, \ldots, \xi_m\}$ как математическую модель содержаний вольфрама и молибдена (в общем случае также и свинца, цинка, меди и т. д.) в гранитах поверхности, а случайную величину $\eta = \{\eta_1, \ldots, \eta_m\}$ — как модель рассматриваемых признаков в гранитах, изучаемых по буровой скважине. Тогда проверяемую H_0 и альтернативную H_1 гипотезы можно записать как $H_0 : M\xi = M\eta$ и $H_1 : M\xi \neq M\eta$, где $M\xi$ и $M\eta$ — математические ожидания m-мерных случайных величин ξ и η, причем $M\xi_1, \ldots, M\xi_m$ — математические ожидания (средние) одномерных случайных величин — моделей содержаний вольфрама и молибдена на поверхности Эльджуртинского массива, а $M\eta_1, \ldots, M\eta_m$ — то же, для гранита, изученного по скважине.

Исходные данные о содержаниях молибдена и вольфрама в гранитах Эльджуртинского массива приведены в работе В. В. Ляховича [30].

---

* Порядок описания сначала ранговых, а затем параметрических методов классификаций диктуется лишь спецификой настоящей работы, посвященной ранговым критериям.

Для другого проверяемого предположения, что глубинность не влияет на величину среднего рассеяния и коррелированности (зависимости) вольфрама и молибдена, следует поставить в соответствие следующие гипотезы:

$$H_0: \Sigma_\xi = \Sigma_\eta, \quad H_1: \Sigma_\xi \neq \Sigma_\eta,$$

где $\Sigma_\xi = \{\sigma_{ij}^{(\xi)}\}$; $i, j = 1, 2, \ldots, m$ — неизвестные ковариации и дисперсии случайных величин — моделей содержаний вольфрама и молибдена на поверхностном эрозионном срезе Эльджуртинского массива; $\Sigma_\eta = \{\sigma_{ij}^{(\eta)}\}$ $i, j = 1, 2, \ldots, m$ — то же, для эльджуртинского гранита, изученного по буровой скважине.

Для проверки гипотезы $H_0: M\xi = M\eta$ рекомендуется использовать два критерия: ранговый Λ критерий Пури—Сена—Тамуры [45, 3] и параметрический информационный $2I$ критерий Джеймса—Сю [29, 22].

Для проверки гипотезы $H_0: \Sigma_\xi = \Sigma_\eta$ также рекомендуются два критерия: ранговый Λ критерий Пури—Сена—Тамуры и параметрический информационный $2I_0$ критерий Кульбака [29].

При проведении законченной классификации k ($k \geqslant 2$) геологических объектов, охарактеризованных набором m геологических свойств, привлекаются функции расстояния и мер сходства между двумя объектами. Наиболее полная теория выбора стратегий классификации содержится в работах Г. Н. Ланса и В. Т. Вильямса [5, 1, 36], которые выделяют две большие группы алгоритмов классификации: иерархические и кластерные. Первые оптимизируют межгрупповые, а вторые — внутригрупповые характеристики сходства [1, с. 122—135].

При решении классификационных геологических задач рекомендуется эффективная пороговая иерархическая алгомеративная процедура кластерного анализа попарных сравнений, напоминающая парагрупповой алгоритм Р. Мак-Кеммона и Г. Венингера, т. е. сохраняющая все его достоинства [5, 36]. Эта процедура классификации набора геологических объектов, охарактеризованных комплексом геологических признаков, базируется на следующих типах функций расстояний и мер сходства между двумя объектами:

на ранговых статистиках Пури—Сена—Тамуры проверки гипотез о равенстве многомерных средних и о равенстве ковариационных матриц;

на параметрической статистике Кульбака, предназначенной для проверки гипотез о равенстве ковариационных матриц;

на коэффициенте сходства, основанного на взвешенном евклидовом расстоянии;

на коэффициенте подобия при разномасштабных признаках, как непрерывных, так и альтернативных;

на коэффициентах сходства для бинарных (качественных) данных Р. Сокала и С. Миченера, Д. Роджерса и Т. Танимото, П. Снита и др. [17].

В качестве геологического примера классификации набора k ($k \geqslant 2$) объектов, охарактеризованных комплексом m ($m \geqslant 2$) геологических свойств, на основе указанной иерархической кластерной процедуры воспользуемся данными В. В. Ляховича [30] о содержаниях вольфрама и молибдена в лейкократовых гранитоидах Тырныауза.

В разделе 2.4.1 осуществлена проверка геологических гипотез о том, что средние концентрации вольфрама и молибдена, а также характеристики их рассеяния и зависимости в лейкократовых гранит-порфирах «Паука», балки «Самолет», в аплит-порфирах Северного участка и в лейкократовых гранитах вертикального разреза Северного участка одни и те же, при альтернативе, что на некоторых участках они существенно различаются. Приведем формальные аналоги вышепоставленных задач. Многомерные (в нашем случае двумерные) случайные величины

$$\xi = \{\xi_1, \ . \ . \ ., \xi_m\}, \quad \eta = \{\eta_1, \ . \ . \ ., \eta_m\},$$

$$\tau = \{\tau_1, \ . \ . \ ., \tau_m\} \quad \text{и} \quad \gamma = \{\gamma_1, \ . \ . \ ., \gamma_m\}$$

будем рассматривать как модели содержаний вольфрама и молибдена в четырех указанных типах лейкократовых гранитов. Тогда нулевую и альтернативные гипотезы относительно многомерных средних можно записать как $H_0 : M\xi = M\eta = M\tau = M\gamma = a_0$ и $H_1 : M\xi$ (или $M\eta$, или $M\tau$, или $M\gamma$) $\neq a_0$.

Другое проверяемое предположение заключается в том, что характеристики рассеяния и зависимости содержания вольфрама и молибдена в четырех типах лейкократовых гранитов одни и те же при наборе альтернатив, что хотя бы в одном типе пород они существенно иные, т. е. $H_0 : \Sigma_\xi = \Sigma_\eta = \Sigma_\tau = \Sigma_\gamma = \Sigma_0$, $H_1 : \Sigma_\xi$ (или Σ_η, или Σ_τ, или Σ_γ) $\neq \Sigma_0$, где Σ — соответствующие ковариационные матрицы.

2.4.1. Ранговые и параметрические статистики Пури—Сена—Тамуры, Джеймса—Сю и Кульбака проверки гипотез о равенстве многомерных средних и ковариационных матриц в двух объектах

В настоящем разделе описаны процедуры применения статистических методов проверки гипотезы о многомерном среднем в двух объектах с помощью рангового критерия Пури—Сена—Тамуры в сочетании с параметрическим информационным критерием Джеймса—Сю, а также проверки гипотезы о ковариационных матрицах в двух объектах с помощью рангового метода Пури—Сена—Тамуры в сочетании с параметрической информационной статистикой Кульбака.

2.4.1.1. Ранговый критерий Пури—Сена—Тамуры проверки гипотез о равенстве многомерных средних в двух объектах

Этот ранговый критерий устойчив относительно нарушения условия предполагаемой нормальности (и даже унимодальности) рас-

пределения изучаемых случайных величин, а также относительно наличия в сопоставляемых выборках аномальных наблюдений.

Согласно Пури и Сену [45, с. 94—96], в качестве меток рабочих статистик критерия для проверки многомерных средних можно привлечь:

1) ранги $E_{ij}=R_{ij}$; тогда критерий Пури—Сена—Тамуры следует рассматривать как многомерный аналог одномерного критерия Вилкоксона (см. 2.3.1);

2) нормальные метки, аналогичные участвующим в одномерном критерии Фишера—Иэйтса—Терри—Гёфдинга [11, с. 106];

3) метки в виде обратных функций нормального распределения $E_{ij}=\Phi^{-1}(R_{ij}/(N+1))$, аналогичные тем, которые участвуют в одномерном критерии Ван дер Вардена [11, с. 107].

Так как наиболее простая рабочая статистика критерия Пури—Сена—Тамуры имеет вид ранговых сумм, аналогичных рекомендованному нами к широкому применению в геологических исследованиях критерию Вилкоксона, ниже описан именно этот подход (первый вид меток).

Ниже показано применение критерия Пури—Сена—Тамуры на примере проверки гипотезы о равенстве двухмерных средних, в качестве которых рассматриваются содержания вольфрама и молибдена в эльджуртинском граните. Другими словами, геологическая задача заключалась в проверке предположения об отсутствии влияния фактора глубинности на средние содержания вольфрама и молибдена в указанных гранитах [30]. Подтверждение или отклонение этой гипотезы содействовало бы более обоснованному суждению о рудогенерирующей способности кислого расплава в пределах Тырныаузского рудного поля.

Отметим, что проверка гипотезы о равенстве многомерных средних не заменяет проверки гипотезы о равенстве одномерных средних, и наоборот. Поэтому как один из видов комплексной обработки геологических данных рекомендуется сочетание одномерных и многомерных статистических методов, в данном случае — ранговых критериев Пури—Сена—Тамуры и Вилкоксона.

Ранее (см. раздел 2.3.1) в отношении молибдена было установлено, что глубинность не влияет на его среднюю концентрацию в эльджуртинском граните.

Проверяемая гипотеза о равенстве двухмерных (в общем случае m-мерных) средних в двух объектах и альтернативные гипотезы имеют следующий вид: $H_0: M\xi=M\eta$ и $H_1: M\xi\neq M\eta$, где $M\xi$ — математическое ожидание (среднее) двухмерной случайной величины ξ — модели содержаний вольфрама и молибдена в граните поверхности Эльджуртинского интрузивного массива, а $M\eta$ — то же, для эльджуртинского гранита по керну скв. 600.

Процедура применения критерия Пури—Сена—Тамуры для проверки гипотезы о равенстве многомерных средних такова.

1. По каждому геологическому признаку в отдельности составляется общий вариационный ряд в порядке возрастания его членов, аналогично тому, как это производится при процедуре при-

менения критерия Вилкоксона (см. раздел 2.3.1). Все члены нумеруются 1, 2, ..., (n_1+n_2), т. е. определяются метки-ранги $E_{tj}=R_{tj}$; $t=1, 2, ..., N$; $(N=n_1+n_2)$; $j=1, 2, ..., m$. Для рассматриваемого геологического примера имеем двухмерную матрицу рангов $R=\{R_{tj}\}$, $t=1, 2, ..., N$ (табл. 10).

Таблица 10

Начальная и конечная части вариационных рядов содержаний вольфрама и молибдена в эльджуртинском граните с определением меток по ранговому критерию Пури—Сена—Тамуры

Исходные данные о содержаниях вольфрама, г/т		Ранги (метки) R_{t1}	Скорректированные ранги R^*_{t1}	Исходные данные о содержаниях молибдена, г/т		Ранги (метки) R_{t2}	Скорректированные ранги R^*_{t2}
на поверхности $n_1=50$	в скважине $n_2=60$			на поверхности $n_1=50$	в скважине $n_2=61$		
0,7		1	13,5	0,4		1	1
0,7		2	13,5	0,5		2	2
0,7		3	13,5	0,6		3	3,5
:	:	:	:		0,6	4	3,5
0,7		17	13,5		0,7	5	19,5
0,7		18	13,5		0,7	6	19,5
	0,7	19	13,5		0,7	7	19,5
	0,7	20	13,5	:	:	:	:
	0,7	21	13,5		1,8	101	102
:	:	:	:	1,8		102	102
3,4		103	104	1,8		103	102
	3,4	104	104		1,9	104	105,5
	3,4	105	104		1,9	105	105,5
	3,6	106	106		1,9	106	105,5
	5	107	107	1,9		107	105,5
5,1		108	108		2,0	108	108
5,1		109	109		2,0	109	109
	5,6	110	110		2,2	110	110
	8,5	111	111		5,4	111	111

2. Так же аналогично критерию Вилкоксона равным значениям, в нашем примере одним и тем же содержаниям вольфрама и молибдена, ставится в соответствие скорректированный ранг (метка) — среднее арифметическое из рангов R_{tj} (см. табл. 10).

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 (в геологическом примере двухмерных) вектора средних меток-рангов T_1 и T_2:

$$T_1 = (T_{11}, T_{12}, \ldots, T_{1j}, \ldots, T_{1m}); \quad j = 1, 2, \ldots, m,$$
$$T_2 = (T_{21}, \ldots, T_{2j}, \ldots, T_{2m});$$

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

4. Определяется m-мерный (в нашем примере двухмерный) вектор средних меток-рангов по всей объединенной выборке объема $N = n_1 + n_2$

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

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

5. Составляется ковариационная матрица меток V размерностью $m \times m$ (в геологическом примере 2×2):

$$V = \{v_{ij}\}, \quad i, j = 1, 2, \ldots, m,$$

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_{ti} первой (второй, ..., N-й) пробы по признаку i не зависят от рангов R_{tj} первой (второй, ..., N-й) пробы по признаку j.

6. С помощью стандартной программы на ЭВМ находим обратную матрицу $V^{-1} = \{v_{ij}\}$.

7. Находим статистику Пури—Сена—Тамуры для проверки гипотезы о равенстве многомерных средних в двух объектах, представляющую собой квадратичную форму

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

где $'$ — знак транспонирования разностей векторов $(T_u - \bar{E})$.

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.

Осуществим вычислительные операции по критерию Пури—Сена—Тамуры на примере сравнения средних содержаний вольфрама и молибдена в эльджуртинском граните. Составление вариационных рядов содержаний вольфрама и молибдена в эльджуртинском граните, ранжирование членов этих рядов R_{tj} $j = 1, 2$; $t = 1, 2, \ldots, n_j$; нахождение скорректированных рангов R^*_{tj} для

Таблица 11

Вычислительные операции по нахождению элементов v_{ij} ковариационной матрицы меток V

Исходные данные о содержаниях вольфрама, г/т		Ранги (метки) R_{i1}	Скорректированные ранги R^*_{i1}	$(R^*_{i1}-\overline{E_1})$, $\overline{E_1}=56$	$(R^*_{i1}-\overline{E_1})^2$	Исходные данные о содержаниях молибдена, г/т		Ранги (метки) R_{i2}	Скорректированные ранги R^*_{i2}	$(R^*_{i2}-\overline{E_2})$, $\overline{E_2}=56$	$(R^*_{i2}-\overline{E_2})^2$	$(R^*_{i1}-\overline{E_1})\times(R^*_{i2}-\overline{E_2})$
на поверхности $n_1=50$	в скважине $n_2=61$					на поверхности $n_1=50$	в скважине $n_2=61$					
0,7		1	13,5	—42,5	1806,25	1,4		85	88,5	32,5	1056,25	—1381,25
0,7		2	13,5	—42,5	1806,25	0,8		35	36,5	—19,5	380,25	828,75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0,7		18	13,5	—42,5	1806,25	0,7		8	19,5	—36,5	1332,25	1551,25
	0,7	19	13,5	—42,5	1806,25		0,7	9	19,5	—36,5	1332,25	1551,25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0,7	26	13,5	—42,5	1806,25		1	52	55,5	—0,5	0,25	21,25
1		27	32	—24	576	0,7		10	19,5	—36,5	1332,25	876
1		28	32	—24	576	0,7		11	19,5	—36,5	1332,25	876
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	3,6	106	106	50	2500		1,4	92	88,5	32,5	1056,25	1625
	5	107	107	51	2601		1	60	55,5	—0,5	0,25	—25,5
5,1		108	108	52	2704	1,8		103	102	46	2116	2392
5,1		109	109	53	2809	1,9		107	105,5	49,5	2450,25	2623,5
	5,6	110	110	54	2916		1,3	84	80	24	576	1296
	8,5	111	111	55	3025		2,2	110	110	54	2916	2970
				$\Sigma=111549,5$						$\Sigma=111\,251,5$		$\Sigma=27\,811,25$

равных значений содержаний вольфрама и молибдена видно из табл. 11 (см. также о критерии Вилкоксона в разделе 2.3.1).

$$T_{11} = \frac{1}{50}(13{,}5 + 18 + \ldots + 109) = 48{,}58,$$

$$T_{12} = \frac{1}{50}(1 + \ldots + 105{,}5) = 54{,}93,$$

$$T_{21} = \frac{1}{61}(13{,}5 + 8 + \ldots + 111) = 62{,}08,$$

$$T_{22} = \frac{1}{61}(3{,}3 + \ldots + 111) = 56{,}88,$$

$$\bar{E}_1 = \bar{E}_2 = \frac{1}{111}(1 + \ldots + 111) = 56.$$

Определим теперь элементы ковариационной матрицы меток размерности 2×2. Табл. 11 позволяет уточнить принцип нахождения сомножителей. Следует помнить, что упорядочить (составить вариационный ряд) мы можем в этом случае только первый признак — содержания вольфрама, а связанные с первым признаком в пробах второй, третий и другие признаки (у нас только второй — молибден) будут характеризоваться неупорядоченными, бессистемными рангами. Например, первому минимальному содержанию вольфрама 0,7 г/т со скорректированным рангом 13,5 соответствует отнюдь не минимальное (0,4 г/т) содержание молибдена, а 1,4 г/т со скорректированным рангом 88,5.

Элементы ковариационной матрицы меток составят:

$$v_{11} = \frac{1}{N}\sum_{t=1}^{N}(R^*_{t1} - \bar{E}_1)^2 =$$

$$= \frac{1}{111}[(13{,}5 - 56)^2 + \ldots + (111 - 56)^2] = 1004{,}95;$$

$$v_{22} = \frac{1}{N}\sum_{t=1}^{N}(R^*_{t2} - \bar{E}_2)^2 =$$

$$= \frac{1}{111}[(88{,}5 - 56)^2 + \ldots + (110 - 56)^2] = 1002{,}27;$$

$$v_{12} = v_{21} = \frac{1}{N}\sum_{t=1}^{N}(R^*_{t1} - \bar{E}_1)(R^*_{t2} - \bar{E}_2) =$$

$$= \frac{1}{111}[(13{,}5 - 56)(88{,}5 - 56) + \ldots + (111 - 56)(110 - 56)] =$$

$$= 250{,}55.$$

Определяем элементы обратной матрицы меток:

$$v_{11}^{-1} = \frac{v_{22}}{v_{11}v_{22} - (v_{12})^2} = \frac{1002{,}27}{1004{,}95 \cdot 1002{,}27 - (250{,}55)^2} = 0{,}00106;$$

$$v_{22}^{-1} = \frac{v_{11}}{v_{11}v_{22} - (v_{12})^2} = \frac{1004{,}95}{1004{,}95 \cdot 1002{,}27 - (250{,}55)^2} = 0{,}00106;$$

$$v_{12}^{-1} = v_{21}^{-1} = \frac{-v_{12}}{v_{11}v_{22} - (v_{12})^2} =$$

$$= \frac{-250{,}55}{1004{,}95 \cdot 1002{,}27 - (250{,}55)^2} = -0{,}00026.$$

Вычисляем статистику Λ Пури—Сена—Тамуры:

$$\Lambda = \sum_{u=1}^{2} n_u (T_u - \bar{E}) V^{-1} (T_u - \bar{E})' = 50\,(48{,}58 -$$

$$-56 \quad 54{,}93 - 56) \begin{pmatrix} 0{,}00106 & -0{,}00026 \\ -0{,}00026 & 0{,}0016 \end{pmatrix} \begin{pmatrix} 48{,}58 - 56 \\ 54{,}93 - 56 \end{pmatrix} +$$

$$+61(62{,}08 - 56 \quad 58{,}88 - 56) \begin{pmatrix} 0{,}00106 & -0{,}00026 \\ -0{,}00026 & 0{,}0016 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 62{,}08 - 56 \\ 58{,}88 - 56 \end{pmatrix} \approx 2{,}86 + 2{,}27 \approx 5{,}13.$$

Критическое значение $\chi^2_{\alpha,\, f=2}$ при уровне значимости $\alpha = 0{,}05$ и двух степенях свободы составляет 5,99.

Так как вычисленное значение $\Lambda = 5{,}13$ не превысило критическое 5,99, то следует принять как подтвердившуюся нулевую гипотезу о равенстве двухмерных средних (средних содержаний вольфрама и молибдена) в эльджуртинском граните. Другими словами, с помощью рангового критерия Пури—Сена—Тамуры подтверждается предположение, что глубинность не влияет на двухмерные средние содержания указанных рудных компонентов в Эльджуртинском гранитном массиве.

Интересно, что аналогичный результат получается и при использовании упрощенной модификации статистики Пури—Сена—Тамуры (2.4.1.3), а именно:

$$\Lambda = n_1 \left(1 + \frac{n_1}{n_2}\right) (T_1 - \bar{E}) V^{-1} (T_1 - \bar{E})^1 \approx 5{,}2 < \chi^2_{0{,}05;\,2} = 5{,}99,$$

2.4.1.2. Параметрический критерий Джеймса—Сю проверки гипотез о равенстве многомерных средних в двух объектах

Это параметрический метод без предположений о равенстве ковариационных матриц. Критерий Джеймса—Сю базируется на предположении о многомерном нормальном распределении случайных величин (моделей геологических признаков) и отсутствии аномальных наблюдений [29, 21]. В комплексе с ранговым критерием Пури—Сена—Тамуры указанные методы позволяют получать надежные геологические интерпретируемые результаты.

В качестве примера вновь воспользуемся проверкой гипотезы о равенстве средних содержаний вольфрама и молибдена, т. е. предположением об отсутствии влияния фактора глубинности на

распределение содержаний вольфрама и молибдена в указанных породах. Формальный аналог этой задачи выше приводился при описании критерия Пури—Сена—Тамуры. Используются исходные аналитические данные по гранитам поверхности Эльджуртинского массива и по керновым пробам.

Процедура применения критерия Джеймса—Сю такова.

1. По двум исходным m-мерным (в нашем примере двухмерным) выборочным данным объема n_1 и n_2 соответственно рассчитываются векторы средних арифметических $\overline{X}^{(1)}$ и $\overline{X}^{(2)}$ и оценки ковариационных матриц $S^{(1)}$ и $S^{(2)}$ по каждой выборке:

$$\overline{X}^{(u)} = \{\overline{x}_j^{(u)}\}, \quad \text{где} \quad \overline{x}_j^{(u)} = \frac{1}{n_u} \sum_{t=1}^{n_u} x_{tj}^{(u)}, \quad j = 1, 2, \ldots, m; \quad u = 1, 2;$$

$$S^{(u)} = \{\widehat{\sigma}_{ij}^{(u)}\}, \quad u = 1, 2, \quad \text{где} \quad \widehat{\sigma}_{ij}^{(u)} = \frac{1}{n_1 - 1} \sum_{t=1}^{n_1} (x_{tj}^{(u)} - \overline{x}_i^{(u)}) \times$$

$$\times (x_{tj}^{(u)} - \overline{x}_j^{(u)}), \quad i, j = 1, 2, \ldots, m; \quad u = 1, 2.$$

2. Рассчитываются разности векторов средних арифметических

$$\overline{X}^{(1)} - \overline{X}^{(2)} = \{\overline{x}_j^{(1)} - \overline{x}_j^{(2)}\}, \quad j = 1, 2, \ldots, m.$$

3. Рассчитывается оценка обобщенной ковариационной матрицы

$$S = \{\widehat{\sigma}_{ij}\} = \frac{S_1}{n_1} + \frac{S_2}{n_2} = \left\{\left(\frac{\widehat{\sigma}_{ij}^{(1)}}{n_1} + \frac{\widehat{\sigma}_{ij}^{(2)}}{n_2}\right)\right\}, \qquad i, j = 1, 2, \ldots, m.$$

4. Рассчитывается статистика Джеймса—Сю, представляющая собой квадратическую форму,

$$2I = (\overline{X}^{(1)} - \overline{X}^{(2)})^1 \, S^{-1} \, (\overline{X}^{(1)} - \overline{X}^{(2)}).$$

В условиях нулевой гипотезы о равенстве многомерных средних в двух объектах статистика $2I$ асимптотически распределена по закону Пирсона χ^2 с m-степенями свободы *.

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 $2I > \chi_{\alpha, m}$, то нулевая гипотеза должна быть отклонена как противоречащая эмпирическим данным и приняты альтернативные гипотезы о существенности различий в многомерных средних сравниваемых двух объектов.

Для сопоставления средних содержаний вольфрама и молибдена в эльджуртинском граните произведем вычисления по алгоритму Джеймса—Сю.

---

* В работах [29, 21] даны более точные распределения статистики $2I$ критерия Джеймса—Сю в условиях нулевой гипотезы.

Элементы двухмерных векторов средних арифметических и выборочных ковариационных матриц оказались равными:

$$\bar{x}_{W}^{(1)} = 1{,}462, \qquad \hat{\sigma}_{WW}^{(1)} = 0{,}969,$$
$$\bar{x}_{Mo}^{(1)} = 1{,}048, \qquad \hat{\sigma}_{WMo}^{(1)} = 0{,}193,$$
$$\bar{x}_{W}^{(2)} = 1{,}887, \qquad \hat{\sigma}_{MoMo}^{(1)} = 0{,}121,$$
$$\bar{x}_{Mo}^{(2)} = 1{,}170, \qquad \hat{\sigma}_{WW}^{(2)} = 1{,}718,$$
$$\hat{\sigma}_{WMo}^{(2)} = 0{,}248,$$
$$\hat{\sigma}_{MoMo}^{(2)} = 0{,}478.$$

Находим разности векторов средних арифметических

$$\bar{X}^{(1)} - \bar{X}^{(2)} = (\bar{x}_j^{(1)} - \bar{x}_j^{(2)}) = (-0{,}425 \quad -0{,}122).$$

Определяем оценку обобщенной ковариационной матрицы

$$S = \begin{pmatrix} \frac{0{,}969}{50} + \frac{1{,}718}{61} & \frac{0{,}193}{50} + \frac{0{,}248}{61} \\ \frac{0{,}193}{50} + \frac{0{,}248}{61} & \frac{0{,}121}{50} + \frac{0{,}478}{61} \end{pmatrix} = \begin{pmatrix} 0{,}048 & 0{,}008 \\ 0{,}008 & 0{,}010 \end{pmatrix}.$$

Находим определитель двухмерной симметричной ковариационной матрицы:

$$\Delta = \hat{\sigma}_{11}\hat{\sigma}_{22} - (\hat{\sigma}_{12})^2 = 0{,}048 \cdot 0{,}01 - (0{,}008)^2 = 0{,}000416.$$

Находим элементы обратной матрицы S^{-1}

$$S^{-1} = \begin{pmatrix} \frac{\hat{\sigma}_{22}}{\Delta} & -\frac{\hat{\sigma}_{12}}{\Delta} \\ -\frac{\hat{\sigma}_{12}}{\Delta} & \frac{\hat{\sigma}_{11}}{\Delta} \end{pmatrix} = \begin{pmatrix} \frac{0{,}010}{0{,}000416} & \frac{-0{,}008}{0{,}000416} \\ \frac{-0{,}008}{0{,}000416} & \frac{0{,}048}{0{,}000416} \end{pmatrix} \simeq$$

$$\simeq \begin{pmatrix} 24{,}13 & 4-18{,}634 \\ -18{,}634 & 111{,}862 \end{pmatrix}.$$

Теперь определяем статистику Джеймса—Сю для нашего двухмерного случая

$$2I = (-0{,}425 \quad -0{,}122) \begin{pmatrix} 24{,}134 & -18{,}634 \\ -18{,}634 & -111{,}862 \end{pmatrix} \begin{pmatrix} -0{,}425 \\ -0{,}122 \end{pmatrix} \simeq 4{,}1.$$

Так как вычисленная статистика Джеймса—Сю $2I = 4{,}1$ не превысила допустимое $\chi^2_{\alpha,\, f=2} = 5{,}99$, то следует принять нулевую гипотезу о равенстве двухмерных средних (содержаний вольфрама и молибдена в эльджуртинском граните) подтвердившейся, т. е. не противоречащей исходным данным. Поэтому можно полагать, что глубинность не влияет на среднюю концентрацию указанных рудных компонентов.

Заключение, полученное с помощью параметрической статистики Джеймса—Сю, совпало с ранее сделанным на основе рангового критерия Пури—Сена—Тамуры и поэтому должно рассматриваться как надежное.

2.4.1.3. Ранговый критерий Пури—Сена проверки гипотезы о равенстве ковариационных матриц в двух объектах

Согласно Пури и Сену [45, с. 94—96], в предположении, что многомерные случайные величины (модели комплекса m геологических признаков в сопоставляемых объектах) имеют одинаковые медианы, можно в качестве меток статистик рассматриваемого критерия для проверки гипотезы о равенстве ковариационных матриц использовать:

1) модуль нормированных и центрированных рангов $E_{ij}=\left|\frac{R_{ij}}{N+1}-0{,}5\right|$, и тогда критерий Пури—Сена—Тамуры выступает как многомерный аналог одномерного критерия масштаба Ансари—Бредли [11, с. 121];

2) квадрат нормированных и центрированных рангов $E_{ij}=\left(\frac{R_{ij}}{N+1}-0{,}5\right)^2$, и тогда критерий Пури—Сена—Тамуры следует рассматривать как многомерный аналог одномерного критерия масштаба Муда [11, с. 123];

3) квадрат обратной функции нормального распределения $E_{ij}=\left\{\Phi^{-1}\left(\frac{R_{ij}}{N+1}\right)\right\}^2$, и тогда критерий Пури—Сена—Тамуры выступает как многомерный аналог одномерного критерия Клотца [11, с. 122];

4) квадрат j-й порядковой статистики выборки объема N_i из стандартного нормального распределения [11, с. 106].

Второй вид меток (аналог критерия Муда) является наиболее простым, и поэтому подход рекомендуется для широкого применения в геологических исследованиях и описан ниже.

Обращаем внимание, что условие равенства медиан означает необходимость центрирования исходных геологических данных медианами по каждому признаку аналогично тому, как рекомендовано при применении одномерного критерия масштаба Сиджела—Тьюки (см. раздел 2.3.1).

В качестве примера вновь воспользуемся данными о содержаниях вольфрама и молибдена в эльджуртинском граните и проверим предположение об отсутствии влияния фактора глубинности на ковариационную матрицу вольфрама и молибдена в указанных гранитах.

Задача имеет следующий формальный вид: $H_0:\Sigma_1=\Sigma_2$, $H_1:\Sigma_1\neq\Sigma_2$, где Σ_1 — неизвестная ковариационная матрица двухмерной случайной величины — модели содержаний вольфрама и молибдена в граните поверхности Эльджуртинского массива, а Σ_2 — то же, для эльджуртинского гранита по керну скв. 600.

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: $Me^{(1)}_{(w)} = 1{,}125$ г/т, $Me^{(1)}_{(Mo)} = 1$ г/т, $Me^{(2)}_{(w)} = 1{,}5$ г/т, $Me^{(2)}_{(Mo)} = 1$ г/т.

2. Центрируем исходные данные медианами (в табл. 12 и 13 они показаны в скобках), т. е., например, для содержаний вольфрама во 2-й выборке: 0,7—1,5=—0,8; 0,7—1,5=—0,8 и т. д.

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 (см. раздел 2.3.1). Все члены нумеруются 1, 2, ..., t, ..., N ($N=n_1+n_2$), т. е. определяются ранги R_{tj}, $t=1, 2, ..., N$, $j=1, 2, ..., m$. Для рассматриваемого геологического примера получим двухмерную матрицу рангов $R=\{R_{tj}\}t=1, 2, ..., N$, $j=1,2$ (см. табл. 12).

4. Аналогично статистике Муда для каждого ранга R_{tj} находим соответствующую ему метку E_{tj}: $E_{tj} = (R_{tj}/(N+1) - 0{,}5)^2$.

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. Например, центрированным исходным данным содержаний молибдена 0,8 г/т (нецентрированным — 1,8 г/т) с метками 0,161; 0,169 и 0,176 будет соответствовать скорректированная средняя метка $(0{,}161+0{,}169+0{,}176) : 3 \simeq 0{,}169$ (см. табл. 12).

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, многомерный вектор средних меток $\bar{E}$ по всей объединенной выборке объема N, ковариационную матрицу меток V и обратную матрицу V^{-1}. Характер вычислительных операций по нахождению этих величин иллюстрируется табл. 13.

10. Находим статистику Пури—Сена—Тамуры для проверки гипотезы о равенстве ковариационных матриц в двух объектах

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

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 откло-

Таблица 12

Начальная и конечная части вариационных рядов центрированных медианами содержаний вольфрама и молибдена в эльджуртинском граните с ранжированием и определением меток по ранговому критерию Пури—Сена—Тамуры проверки ковариационных матриц

Исходные данные о содержаниях вольфрама, г/т		Ранги центрированных данных R_{f1}	$\Delta = \left(\frac{R_{f1}}{N+1} - 0{,}5\right)$	Метки $E_{f1} = \Delta^2$	Скорректированные метки E^*_{f1}	Исходные данные о содержаниях молибдена, г/т		Ранги центрированных данных R_{f2}	$\Delta = \left(\frac{R_{f2}}{N+1} - 0{,}5\right)$	Метки $E_{f2} = \Delta^2$	Скорректированные метки E^*_{f2}
на поверхности $n_1=50$	в скважине $n_2=61$					на поверхности $n_1=50$	в скважине $n_2=61$				
	0,7(—0,8)	1	—0,491	0,241	0,212	0,4(—0,6)		1	—0,491	0,241	0,241
	0,7(—0,8)	2	—0,482	0,232	0,212	0,5(—0,5)		2	—0,482	0,232	0,232
	:	:	:	:	:	0,6(—0,4)		3	—0,473	0,224	0,219
	0,7(—0,8)	8	—0,429	0,184	0,212		0,6(—0,4)	4	—0,464	0,215	0,219
0,7(—0,55)		9	—0,420	0,176	0,120		0,7(—0,3)	5	—0,455	0,207	0,112
0,7(—0,55)		10	—0,411	0,169	0,120		0,7(—0,3)	6	—0,446	0,199	0,122
:	:	:	:	:	:	:	:	:	:	:	:
	5,6(4,1)	110	0,482	0,232	0,232		2,2(1,2)	110	0,482	0,232	0,232
	8,5(7)	111	0,491	0,241	0,241		5,4(4,4)	111	0,491	0,241	0,241

Примечание. В скобках—центрированные медианами исходные данные.

Таблица 13

Вычислительные операции по нахождению элементов v_{ij} ковариационной матрицы для меток по центрированным медианами исходным данным

Исходные данные о содержаниях вольфрама, г/т: на поверхности $n_1=50$	Исходные данные о содержаниях вольфрама, г/т: в скважине $n_2=61$	Метки E_{t1}	Скорректированные метки E^*_{t1}	$\left(E^*_{t1}-\overline{E_1}\right)$, $\overline{E_1}=0,095$	$\left(E^*_{t1}-\overline{E_1}\right)^2$	Исходные данные о содержаниях молибдена, г/т: на поверхности $n_1=50$	Исходные данные о содержаниях молибдена, г/т: в скважине $n_2=61$	Метки E_{t2}	Скорректированные метки E^*_{t2}	$\left(E^*_{t2}-\overline{E_2}\right)$, $\overline{E_2}=0,092$	$\left(E^*_{t2}-\overline{E_2}\right)^2$	$\left(E^*_{t1}-\overline{E_1}\right)\times\left(E^*_{t2}-\overline{E_2}\right)$
	0,7(—0,8)	0,241	0,212	0,117	0,0137		0,7(—0,3)	0,120	0,120	0,028	0,0008	—0,0032
	0,7(—0,8)	0,232	0,212	0,117	0,0137		1,3(0,3)	0,046	0,046	—0,046	0,00021	0,0054
	:	:	:	:	:		:	:	:	:	:	:
	0,7(—0,8)	0,184	0,212	0,117	0,0137		1,0(0,0)	0,0004	0,001	—0,091	0,0085	—0,0107
0,7(—0,55)		0,176	0,120	0,025	0,0006	1,4(0,4)		0,067	0,084	—0,008	0,0072	—0,0002
:	:	:	:	:	:		:	:	:	:	:	:
	5(3,5)	0,207	0,207	0,112	0,0125		1,0(0,0)	0,001	0,001	—0,091	0,0085	—0,0102
5,1(3,85)		0,216	0,220	0,125	0,0156	1,8(0,8)		0,161	0,169	0,074	0,0059	0,0092
5,1(3,85)		0,224	0,220	0,125	0,0156	1,9(0,9)		0,184	0,195	0,103	0,0106	0,0129
	5,6(4,1)	0,232	0,232	0,137	0,0188		1,3(0,3)	0,046	—0,046	0,046	0,0021	—0,0063
	8,5(7)	0,241	0,241	0,146	0,0213		2,2(1,2)	0,232	0,232	—0,140	0,0196	0,020
Сумма					0,1877						0,1292	0,0215

Примечание. В скобках—центрированные медианами исходные данные.

нить и принять альтернативные — о существенности различий в ковариационных матрицах сравниваемых объектов. Другими словами, меры рассеяния и зависимости геологических характеристик в сравниваемых геологических объектах статистически значимо различаются.

Произведем расчеты по сопоставлению содержаний вольфрама и молибдена в эльджуртинском граните (на поверхности и по керну скв. 600).

Находим двухмерные векторы средних меток по каждой из выборок T_1 и T_2, а также по объединенной выборке $\bar{E}$: $T_1=(T_{11}, T_{12})$; $T_2=(T_{21}, T_{22})$; $\bar{E}=(\bar{E}_1, \bar{E}_2)$.

Они оказались равными:

$$T_{11}=\frac{1}{50}(0{,}120+\ \ldots\ +0{,}220)=0{,}0942;$$

$$T_{12}=\frac{1}{50}(0{,}241+\ \ldots\ +0{,}195)=0{,}0929;$$

$$T_{21}=\frac{1}{61}(0{,}212+\ \ldots\ +0{,}241)=0{,}0949;$$

$$T_{22}=\frac{1}{61}(0{,}219+\ \ldots\ +0{,}241)=0{,}0906;$$

$$\bar{E}_1=\frac{1}{111}(0{,}212+\ \ldots\ +0{,}241)=0{,}0950;$$

$$\bar{E}_2=\frac{1}{111}(0{,}241+\ \ldots\ +0{,}241)=0{,}0920.$$

Определим ковариационную матрицу меток $V=\{v_{ij}\}$; $i, j=1,2$ и обратную к ней $V^{-1}=\{v^{ij}\}$; $i, j=1,2$ (см. табл. 13):

$$v_{11}=\frac{1}{N}\sum_{t=1}^{N}(E_{t1}^{*}-\bar{E}_1)^2=\frac{1}{111}\cdot[(0{,}241-0{,}095)^2+$$

$$+(0{,}241-0{,}095)^2]=0{,}0016911;$$

$$v_{12}=v_{21}=\frac{1}{N}\sum_{t=1}^{N}(E_{t1}^{*}-\bar{E}_1)(E_{t2}^{*}-\bar{E}_2)=\frac{1}{111}\cdot[(0{,}241-0{,}095)^2\times$$

$$\times(0{,}120-0{,}092)+\ \ldots\ +(0{,}241-0{,}095)(0{,}232-0{,}092)]=$$

$$=0{,}000194;$$

$$v_{22}=\frac{1}{N}\sum_{t=1}^{N}(E_{t2}^{*}-\bar{E}_2)^2=\frac{1}{111}[(0{,}120-0{,}0092)^2+\ \ldots\ +(0{,}232-$$

$$-0{,}092)^2=0{,}001166];$$

$$\Delta=v_{11}v_{22}-(v_{12})^2=0{,}0000019;$$

$$v^{11} = \frac{v_{22}}{\Delta} = \frac{0{,}001166}{0{,}0000019} = 613{,}684;$$

$$v^{12} = v^{21} = \frac{-v_{12}}{\Delta} = \frac{-0{,}000194}{0{,}0000019} = -102{,}105;$$

$$v^{22} = \frac{v_{11}}{\Delta} = \frac{0{,}001691}{0{,}0000019} = 890{,}0.$$

Вычисляем теперь статистику Λ_Σ Пури—Сена—Тамуры:

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

$$= 50\,[(0{,}0942 - 0{,}095),\ (0{,}0929 - 0{,}092)] \times$$

$$\times \begin{bmatrix} 613{,}684 & -102{,}105 \\ -102{,}105 & 890{,}0 \end{bmatrix} \cdot \begin{bmatrix} 0{,}0942 - 0{,}095 \\ 0{,}0929 - 0{,}092 \end{bmatrix} +$$

$$+ 61\,[(0{,}0949 - 0.095),\ (0{,}0906 - 0{,}092)] \times$$

$$\times \begin{bmatrix} 613{,}684 & -102{,}105 \\ -102{,}105 & 890{,}0 \end{bmatrix} \cdot \begin{bmatrix} 0{,}0949 - 0{,}095 \\ 0{,}0906 - 0{,}092 \end{bmatrix} =$$

$$= 0{,}063 + 0{,}105 = 0{,}168.$$

Критическое значение $\chi^2_{\alpha,\, f=m}$, как мы выше установили, составляет при $\alpha = 0{,}05$ и $m = 2$ $\chi^2_{\alpha,\, f} = 5{,}991$. Учитывая, что вычисленное $\Lambda_\Sigma = 0{,}168 \ll \chi^2_{0{,}05;\ 2} = 5{,}991$, следует принять нулевую гипотезу о равенстве ковариационных матриц в сравниваемых объектах как подтвердившуюся. Другими словами, с помощью рангового критерия Пури—Сена—Тамуры нет оснований полагать, что характеристики рассеяния и зависимости содержаний вольфрама и молибдена различаются по данным опробования керна (т. е. на глубине) и по поверхностным пробам в эльджуртинском граните.

Рабочие статистики Λ критерия Пури—Сена—Тамуры для проверки гипотез о равенстве многомерных средних (2.4.1.1) и проверки гипотез о равенстве ковариационных матриц (2.4.1.3) и имеющие вид

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

для рассматриваемого нами случая сравнения двух объектов можно упростить. После несложных преобразований нетрудно получить следующее выражение:

$$\Lambda(\Lambda_\Sigma) = n_1 \left(1 + \frac{n_1}{n_2}\right) (T_1 - \bar{E}) V^{-1} (T_1 - \bar{E})'.$$

Расчеты на ЭВМ по этим формулам свидетельствуют о их полной идентичности.

Обращаем внимание, что в качестве T_1, T_2, $\bar{E}$ можно использовать и ранги, и метки, описанные в разделах 2.4.1.1 и 2.4.1.3, т. е. упрощенное выражение статистики Λ критерия Пури—Сена—Тамуры применять для проверки гипотез как о равенстве многомерных средних, так и о равенстве ковариационных матриц в двух объектах.

Одномерная модификация упрощенной статистики Пури—Сена—Тамуры будет иметь вид

$$\Lambda = n_1\left(1+\frac{n_1}{n_2}\right)(T_1-\bar{E})\,V^{-1}\,(T_1-\bar{E})' =$$

$$= \frac{N}{n_1 n_2}\,\frac{\left(N\sum\limits_{t=1}^{n_1}R_t - n_1\sum\limits_{t=1}^{N}R_t\right)^2}{\left[N\sum\limits_{t=1}^{N}R_t^2-\left(\sum\limits_{t=1}^{N}R_t\right)^2\right]},\ \text{где } N=n_1+n_2.$$

Комплексная программа по кластерному анализу, составленная Ю. П. Беловым на ФОРТРАН-IV для ЕС ЭВМ относительно ранговых статистик Пури—Сена—Тамуры, реализует вышеописанные модифицированные алгоритмы.

2.4.1.4. Параметрический критерий Кульбака проверки гипотез о равенстве ковариационных матриц в двух объектах

Этот критерий является своеобразным многомерным аналогом одномерного критерия Бартлета (см. раздел 2.3.1) и учитывает не только дисперсии, но и ковариации признаков (т. е. характеристики их связи). Ограничения в применении критерия Кульбака полностью аналогичны вышеупомянутым при описании критерия Джеймса—Сю.

Воспользуемся тем же примером. Будем проверять предположение, что глубинность не влияет на ковариационные матрицы содержаний вольфрама и молибдена в керновых пробах и на поверхности Эльджуртинского массива.

Формальный аналог этой геологической задачи будет следующим: $H_0: \Sigma_1 = \Sigma_2$ и $H_1: \Sigma_1 \neq \Sigma_2$, где Σ_1 — ковариационная матрица случайных величин ξ_1 и ξ_2 — моделей содержаний вольфрама и молибдена в гранитах поверхности Эльджуртинского массива; Σ_2 — ковариационная матрица случайных величин η_1 и η_2 — моделей тех же рудных элементов по керновым пробам.

Процедура применения критерия Кульбака такова.

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:

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

3. С помощью стандартных программ на ЭВМ рассчитываются определители трех выборочных ковариационных матриц S_1, S_2 и S.

4. 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 *.

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.

Для вышеуказанного примера сопоставления характеристик рассеяния и зависимости содержаний вольфрама и молибдена в эльджуртинском граните произведем вычисления по алгоритму Кульбака. Элементы выборочных ковариационных матриц, определенные по алгоритму Джеймса—Сю, оказались равными:

$$\widehat{\sigma}^{(1)}_{WW} = 0{,}969, \qquad \widehat{\sigma}^{(2)}_{WW} = 1{,}718,$$

$$\widehat{\sigma}^{(1)}_{WMo} = \widehat{\sigma}^{(1)}_{MoW} = 0{,}193, \qquad \widehat{\sigma}^{(2)}_{WMo} = \widehat{\sigma}^{(2)}_{MoW} = 0{,}248,$$

$$\widehat{\sigma}^{(1)}_{MoMo} = 0{,}121, \qquad \widehat{\sigma}^{(2)}_{MoMo} = 0{,}478.$$

Определяем оценку общей ковариационной матрицы:

$$S = \begin{pmatrix} \dfrac{49\cdot 0{,}969 + 60\cdot 1{,}718}{50 + 61 - 2} & \dfrac{49\cdot 0{,}193 + 60\cdot 0{,}248}{109} \\ \dfrac{49\cdot 0{,}193 + 60\cdot 0{,}248}{109} & \dfrac{49\cdot 0{,}121 + 60\cdot 0{,}478}{109} \end{pmatrix} \approx \begin{pmatrix} 1{,}382 & 0{,}223 \\ 0{,}223 & 0{,}318 \end{pmatrix}.$$

Находим детерминанты трех матриц $|S_1|$, $|S_2|$, $|S|$:

$$|S_1| = \begin{vmatrix} 0{,}969 & 0{,}193 \\ 0{,}193 & 0{,}121 \end{vmatrix} = 0{,}08,$$

$$|S_2| = \begin{vmatrix} 1{,}718 & 0{,}248 \\ 0{,}248 & 0{,}478 \end{vmatrix} = 0{,}7597,$$

$$|S| = \begin{vmatrix} 1{,}382 & 0{,}223 \\ 0{,}223 & 0{,}318 \end{vmatrix} = 0{,}3897.$$

* С более точным распределением статистики $2I_0$ в условиях нулевой гипотезы следует ознакомиться в [29, 35].

Вычисляем статистику $2I_0$ критерия Кульбака:

$$2I_0 = 49 \cdot \ln \frac{0{,}3897}{0{,}08} + 60 \cdot \ln \frac{0{,}3897}{0{,}7597} = 37{,}5.$$

Так как вычисленное значение превысило допустимое, т. е. $2I_0 = 37{,}5 > \chi^2_{0{,}05;\ 3} = 7{,}81$, следует нулевую гипотезу отклонить как неподтвердившуюся и принять альтернативные гипотезы о неравенстве ковариационных матриц в сравниваемых объектах. Этот вывод не совпадает с ранее сделанным с помощью рангового критерия Пури—Сена—Тамуры. Поэтому следует с осторожностью полагать, что степень рассеяния и коррелированность значимо различаются в гранитах поверхности и на глубине Эльджуртинского массива, что подтверждается только параметрическим критерием.

2.4.2. Функции расстояний и мер сходства между объектами

Для решения задач классификации геологических объектов весьма перспективным является применение статистических алгоритмов кластерного анализа, относящихся к иерархическим агломеративным процедурам типа б) (см. раздел 2.2), с включением пороговой (критической) постоянной величины, с которой сравниваются функции расстояния (или меры сходства) между геологическими объектами и многомерными наблюдениями. Ниже кратко рассматриваются или упоминаются иерархические процедуры кластерного анализа, наиболее широко применяемые (или могущие найти применение) в геологических исследованиях.

Более подробно с этими процедурами, а также с вопросами применения параллельных и последовательных процедур кластерного анализа и вопросами сочетания методов кластерного анализа и многомерного шкалирования заинтересованному читателю следует ознакомиться в ранее опубликованных работах [1, 21, 17].

А. Как отмечал Р. Сокал [21, с. 11], обычной математической основой для классификации объектов служит вычисление некоторых функций на каждой паре многомерных наблюдений, в результате чего получаются матрицы коэффициентов сходства или различия между всеми возможными парами. Эти коэффициенты бывают в основном трех типов.

Первый тип — коэффициенты расстояния имеют общий вид

$$d_r(X_{ut},\ X_{vt}) = \left(\frac{1}{m} \sum_{j=1}^{m} | x_{utj} - x_{vtj} |^r \right)^{1/r},$$

где u, v — индекс объектов, каждый из которых охарактеризован набором $t = 1, 2, \ldots, n_u$ (n_v) многомерных наблюдений X_{ut} и X_{vt}; x_{utj} и x_{vtj} — значения признака j ($j = 1, 2, \ldots, m$) для u (v)-го объектов; m — число признаков; r — положительное целое число. Два случая особенно полезны: при $r = 1$ имеем дело с Манхэттенским расстоянием и при $r = 2$ получаем таксономическое расстояние.

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 [21, с. 11]:

$$r = (X_{ut},\ X_{vt}) = \sum_{j=1}^{m} w_{uvj} s_{uvj} \Big/ \sum_{j=1}^{m} w_{uvj}, \quad j = 1, 2, \ldots, m,$$

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 между этими наблюдениями по значениям признаков.

Если для пары векторов — многомерных наблюдений можно оценить и их ковариационные матрицы, то в кластерном анализе привлекаются обобщенные расстояния.

Среди наиболее употребительных функций для расстояний d и коэффициентов сходства r между парой многомерных наблюдений X_{ut} и X_{vt} следует отметить:

1) обычное евклидово расстояние [1, с. 78];

2) взвешенное евклидово расстояние [1, с. 78];

3) согласно Б. Дюрану и П. Оделлу [17, с. 17], некоторые коэффициенты типа расстояния: L_1 — норма, супремум — норма, L_m — норма;

4) обычное расстояние Махаланобиса [17, с. 17];

5) взвешенное расстояние Махаланобиса [1, с. 77];

6) расстояние типа коэффициента расового сходства Пирсона и его модификаций — расстояний Родионова, Гавришина [21, с. 12; 40; 12, с. 2—3, 5—6];

7) расстояние Минковского [21, с. 150—151];

8) генетические расстояния Сангхви, Нея, расстояние Хеллингера для качественных данных [21, с. 150—151];

9) расстояние Минковского, классификационный индекс, квадратическая дифференциальная метрика Рао для количественных данных [21, с. 153—154];

10) Хеммингово расстояние как мера различия наблюдений, задаваемых дихотомическими, альтернативными, бинарными (0 и 1) признаками [1, с. 78];

11) 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:

мера отдаленности Джеффриса—Матуситы [17, с. 17—18];

коэффициент дивергенции [17, с. 17];

информационный радиус Джардайна и Сибсона [21, с. 155];

12) расстояния, задаваемые с помощью потенциальных функций, три из которых приводятся С. А. Айвазяном и др. [1, с. 79];

13) коэффициент сходства, основанный на взвешенном евклидовом расстоянии [5, с. 121];

14) коэффициент подобия при разномасштабных признаках, как непрерывных, так и альтернативных (0 и 1) [5, с. 122];

15) серия наиболее известных коэффициентов сходства для бинарных данных, т. е. булевых (0 и 1) векторов многомерных наблюдений [17, с. 20]:

коэффициент сходства Сокала и Миченера,

коэффициент сходства Рао и Рассела,

коэффициент сходства Хаммана,

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.;

16) аналогично функциям расстояний, задаваемых с помощью потенциальных функций, существуют разнообразные коэффициенты сходства на основе этих потенциальных функций [1].

Заметим, что с функциями расстояний типа Махаланобиса для типичной в геологических исследованиях ситуации наличия зависимых признаков в многомерных наблюдениях связаны наиболее перспективные алгоритмы классификаций объектов, а именно те, которые основаны на статистиках Джеймса—Сю, Кульбака, Готелинга, Пури—Сена—Тамуры и др.

Б. Теперь рассмотрим наиболее употребительные функции для расстояний d и коэффициентов сходства r между двумя объектами, т. е. матрицами наблюдений X_u и X_v. Наиболее полная теория выбора стратегий кластерного анализа содержится в работах Г. Ланса и В. Вильямса, которые выделяют две большие группы алгоритмов классификации: иерархические и кластерные [5, с. 122—135]. Первые оптимизируют межгрупповые характеристики сходства, а вторые — внутригрупповые характеристики.

Следует отметить следующие функции [1, 29]:

1) минимальное локальное расстояние, измеряемое по принципу «ближайшего соседа» [1, с. 82]; геометрическое представление этого расстояния для двумерных наблюдений приведено на рис. 7, *а*;

2) максимальное локальное расстояние, измеряемое по принципу «дальнего соседа» [1, с. 82]; его геометрическое представление см. на рис. 7, *б*;

3) центроидное расстояние, измеряемое по «центрам тяжести» таксонов [1, с. 82]; его геометрическое представление см. на рис. 7, *в*;

4) медианное расстояние — разновидность центроидного расстояния;

5) среднее расстояние, измеряемое по принципу «средней связи»; его геометрическое представление приведено на рис. 7, *г*;

6) расстояние Хаусдорфа — разновидность максиминной метрики, учитывающей минимальное и одновременно максимальное локальные расстояния; геометрическое представление его дано на рис. 7, *д* [36];

7) обобщенное, по Колмогорову, расстояние, основанное на понятии степенного среднего, включает многие из вышерассмотренных видов расстояний [1, с. 83—84];

8) расстояния центроидного типа:

Махаланобиса при равных ковариационных матрицах [33, с. 262—263],

Джеймса—Сю для любых ковариационных матриц [29, 21],

Готелинга для любых ковариационных матриц,

Пури — Сена — Тамуры (ранговая метрика) [3],

Свейна—Фу для любых ковариационных матриц [17, с. 29];

9) меры близости (сходства), основанные на потенциальной функции [1, с. 79—82];

10) неиерархический метод *k* средних [21, с. 131];

11) неиерархический адаптивный метод «Isodata», являющийся своеобразным аналогом метода средних [21, с. 131];

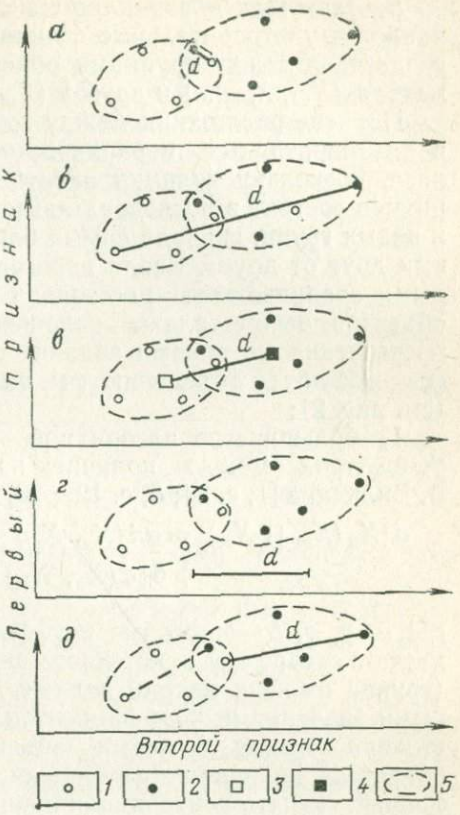


Рис. 7. Расстояния между парами объектов:

а — по принципу «ближайшего соседа»; *б* — по принципу «дальнего соседа»; *в* — между центрами тяжести; *г* — по принципу «средней связи»; *д* — хаусдорфово расстояние

1 — двухмерные точки наблюдения первого объекта; *2* — двухмерные точки наблюдения второго объекта; *3* — «центр тяжести» двухмерных точек наблюдения первого объекта; *4* — «центр тяжести» двухмерных наблюдений второго объекта; *5* — выборочные совокупности двухмерных наблюдений

12) метод, основанный на дисперсионном критерии оптимизации с двумя разновидностями [21, с. 131].

Б. Дюран и П. Оделл [17, с. 28—29] также рассматривают другие многомерные меры расстояний и их метрические свойства. В частности, кроме вышеотмеченных, проанализированы метрики Крамера—Мизеса, Колмогорова—Смирнова, вариационное расстояние Колмогорова, Бхаттачария, Джеффриса—Матусита, информационная мера Кульбака—Либлера, Самуэль—Бахи, Кифера—Вольфовиц, дивергенция и др.

В. Завершая обзор кластерных процедур, рассмотрим теперь наиболее употребительные функции расстояния d и коэффициенты сходства r между группами объектов, т. е. между одной группой матриц $\{X_{u1}, X_{u2}, \ldots\}$ и другой $\{X_{v1}, X_{v2}, \ldots\}$.

Понятие расстояния между группами объектов особенно важно в агломеративных иерархических процедурах кластерного анализа, поскольку принцип работы таких классифицирующих алгоритмов состоит в последовательном объединении объектов, а затем и целых групп, сначала самых близких, а затем все более отдаленных друг от друга. Вводя пороговые величины, т. е. такие, с которыми сравниваются расстояния между объектами и группами объектов, осуществляют законченную классификацию (в случае превышения метриками заданных значений). В противном случае все объекты объединяются на $k-1$ стадии в одну группу (см. рис. 2);

1) большинство алгоритмов классификации в иерархических процедурах можно получить из общей формулы Г. Ланса и В. Вильямса [1, с. 84; 8, с. 122; 23, с. 38]

$$d[X_u(X_{v1} \cup X_{v2})] = \alpha d(X_u, X_{v1}) + \beta d(X_u, X_{v2}) + \gamma d(X_{v1}, X_{v2}) + \\ + \delta\,|\,d(X_u, X_{v1}) - d(X_u, X_{v2})\,|,$$

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 и δ, легко получаем расстояния 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, расстояние для объединения групп объектов с использованием расстояния, основанного на понятии степенного среднего [1, с. 84];

3) расстояние по Уорду (модификация Уишартом общей формулы Г. Ланса и В. Вильямса) [1, с. 85; 17, с. 39];

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 [5, с. 125—126]);

5) процедура В. Н. Елкиной и Н. Г. Загоруйко [5, с. 126—127];

6) весьма эффективная парагрупповая процедура Р. Мак-Кеммона и Г. Венингера представляет частный случай агломеративной иерархической процедуры кластерного анализа [5, с. 127, 130—131]. В одну группу объединяются такие объекты X_u и X_v, на которых достигается минимум функционала $T_{uv} = (s_{uu} + s_{vv} - s_{uv})/c^2_{nu+nv}$, где s_{uu} и s_{vv} — величины, характеризующие расстояние внутри u- и v-й групп; s_{uv} — величина, характеризующая межгрупповое расстояние, nu и nv — объемы групп; c^2_{nu+nv} — число сочетаний из $nu+nv$ 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. 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 и подробно рассматривается ниже в разделе 2.4.3.

Завершая обзор процедур кластерного анализа, отметим следующее:

трудно установить точные правила кластерного анализа, применяемые во всех ситуациях, трудно построить объективный критерий для сравнения кластеров, полученных с помощью различных процедур [21, с. 157];

в качестве приближения к решению проблемы оценки процедур кластерного анализа можно, по-видимому, воспользоваться оценочными индексами, введенными Дж. Меззихом [21, с. 132]. Речь идет о величине внешнего критерия значимости, величине внутреннего критерия значимости и мере воспроизводимости.

Дж. Меззих измеряет внешний критерий значимости как процент совпадения предсказаний экспертов и результатов процедуры кластерного анализа. В качестве внутреннего критерия значимости предлагается кофенетический коэффициент корреляции, введенный Р. Сокалом и Ф. Рольфом [21, с. 132]. Мера воспроизводимости также, в сущности, является специальным коэффициентом корреляции;

в книгах описаны самые разнообразные методы кластерного анализа. Н. Джардайн и Р. Сибсон [21] создали общую теорию кластерного анализа, основанную на наборе аксиом.

2.4.3. Иерархические агломеративные процедуры классификации набора объектов, основанные на многомерных критериях

В настоящем разделе на том же примере о содержаниях молибдена и вольфрама в лейкократовых гранитоидах Тырныауза охарактеризована иерархическая агломеративная классификация объектов относительно многомерного (в данном примере двухмерного) среднего и ковариационной матрицы порядка $m \times m$ (в нашем примере 2×2).

В качестве статистических критериев предлагаются ранговые и параметрические многомерные критерии: Пури—Сена—Тамуры и Джеймса—Сю для многомерного среднего и Пури—Сена—Тамуры и Кульбака для ковариационных матриц.

2.4.3.1. Иерархическая процедура классификации набора объектов относительно многомерного среднего, основанная на ранговой статистике Пури—Сена—Тамуры

По четырем выборкам содержаний вольфрама и молибдена в лейкократовых гранитоидах Тырныауза проверим гипотезу об однородности относительно средних содержаний этих рудных элементов. Исходные данные приведены в ранее опубликованной работе [30]. Проверяемую гипотезу запишем как $H_0: M\xi_1 = M\xi_2 = M\xi_3 = M\xi_4 = M\xi_0$ при множестве альтернатив $H_1: M\xi_u \neq M\xi_0$ хотя бы одного u ($u = 1, 2, 3, 4$), где M — символ математического ожидания; ξ_1 — двухмерная случайная величина — модель содержаний вольфрама и молибдена в лейкократовых гранит-порфирах «Паука» [30, табл. 9]; ξ_2 — то же, в гранит-порфирах балки «Самолет»; ξ_3 — то же, в аплит-порфирах Северного участка Тырныауза [30, табл. 11]; ξ_4 — то же, для вертикального разреза лейкократовых гранитов Северного участка Тырныауза [30, табл. 13].

Для каждой пары (u, v) объектов рассчитывается статистика $\Lambda_{u,v}$ Пури—Сена—Тамуры согласно процедуре, описанной в разделе 2.4.1. Результаты приводятся ниже.

Сопоставляемые выборки

	1—2	1—3	1—4	2—3	2—4	3—4
$\Lambda_{u,v}$	40,15	33,09	27,99	31,12	8,03	30,81

Как видно из приведенных данных, минимальное значение $\min \Lambda_{u,v} = 8{,}03$, соответствующее объектам 2 и 4, превысило критическое $\chi^2_{\alpha=0{,}05;\ f=2} = 5{,}99$, поэтому ни одной однородной группы объектов статистически обоснованно по ранговому критерию Пури—Сена—Тамуры выделять нельзя. Все лейкократовые гранитоиды Тырныауза характеризуются различающимися значениями средних содержаний вольфрама и молибдена.

2.4.3.2. Иерархическая процедура классификации набора объектов относительно многомерного среднего, основанная на параметрическом критерии Джеймса—Сю

По тем же четырем типам лейкократовых гранитоидов Тырныауза осуществим их классификацию относительно средних содержаний вольфрама и молибдена с помощью критерия Джеймса—Сю. Проверяемое предположение об однородности указанных пород и формализация геологической задачи аналогичны тем, которые рассматривались выше при классификации геологических объектов с помощью рангового критерия Пури—Сена—Тамуры.

Для каждой пары (u, v) объектов рассчитывается значение $2I_{u,v}$ критерия Джеймса—Сю согласно процедуре, описанной в разделе 2.4.1. Результаты изложены ниже.

Сопоставляемые выборки

	1—2	1—3	1—4	2—3	2—4	3—4
$2I_{u,v}$	5,19	4,54	5,77	25,15	8,69	27,59

Как видно из приведенных данных, минимальное значение $\min 2I_{u,v}=4{,}54$, соответствующее объектам 1 и 3, не превысило критическое $\chi^2_{\alpha=0{,}05;\ f=2}=5{,}99$, поэтому эти объекты подлежат объединению.

Снова рассчитаем статистики Джеймса—Сю после объединения 1-го и 3-го объектов.

Сопоставляемые выборки

	(1+3)—2	(1+3)—4	2—4
$2I_{u,v}$	20,8	23,28	8,69

Минимальное значение $\min 2I_{u,v}=8{,}69$, соответствующее 2-му и 4-му объектам, превысило допустимое $\chi^2_{\alpha=0{,}05;\ f=2}=5{,}99$, и поэтому дальнейший поиск однородных групп прекращается. Таким образом, с помощью параметрического критерия Джеймса—Сю устанавливается единственная однородная группа (1+3) относительно средних содержаний вольфрама и молибдена, соответствующая гранит-порфирам «Паука» и аплит-порфирам Северного участка Тырныауза. Эта группа статистически отличается (по W и Mo) от гранит-порфиров балки «Самолет» и лейкократовых гранитов вертикального разреза Северного участка Тырныауза.

Таблица 14

Статистические характеристики основных рудных компонентов лейкократовых гранитоидов Тырныауза

Тип гранитоидов	Элемент	Статистические характеристики					Согласованность с законами распределений
		x и α	$\pm\lambda_{\bar{x}}$ и $\pm\lambda_\alpha$	s^2 и b^2	s и b	v и v_α	
Гранит-порфиры «Паука» ($n_1=27$)	W	7,14	5,01	176,56	13,29	186,2	≠н., л.
	Mo	36,66	30,28	6 445,9	80,29	219,0	л.
		35,94	3,24	5 775,7	76,0	211,5	
Гранит-порфиры «Самолет» ($n_2=49$)	W	8,10	5,56	394,86	19,87	245,3	≠н., л.
	Mo	2,98	0,96	11,78	3,43	115,3	л.
		2,87	0,08	9,14	3,02	105,2	
Аплит-порфиры Северного участка ($n_3=38$)	W	59,03	48,8	23 565,3	153,5	260,1	л.
		46,55	3,82	10 623,0	103,1	221,5	
	Mo	36,91	13,36	1 765,5	42,0	113,8	≠н., л.
Лейкократовые граниты разреза Северного участка ($n_4=20$)	W	3,63	1,46	11,17	3,34	91,9	л.
		3,58	0,14	11,08	3,33	93,0	
	Mo	1,35	0,55	1,56	1,25	92,2	≠н., л.

Условные обозначения. x—среднее арифметическое; α и b^2—максимально-правдоподобные оценки среднего и дисперсии, по Ачисону и Брауну [42]; $\lambda_{\bar{x}}$ и λ_α—точности среднего арифметического $\bar{x}$ и оценки α при надежности 0,95; s^2—выборочная дисперсия; v и v_α—выборочные коэффициенты вариации; л.—согласованность с логнормальным законом; ≠н. л.—отсутствие согласованности с нормальной и логнормальной моделями.

Результат, полученный по параметрическому критерию Джеймса—Сю, отличается от ранее полученного по ранговому критерию Пури—Сена—Тамуры, согласно которому ни одной однородной группы выделять нельзя. Учитывая, что выборки содержали ураганные значения полезного компонента, в частности, первая выборка содержала 430 г/т Mo [30, табл. 9], а третья выборка — 900 г/т W [30, табл. 11], что влияло на мощность параметрических критериев, следует заключение об отсутствии однородных групп объектов, полученное по ранговому критерию Пури—Сена—Тамуры, считать более обоснованным.

Основные статистические характеристики четырех типов лейкократовых гранитоидов приведены в табл. 14.

2.4.3.3. Иерархическая процедура классификации набора объектов относительно ковариационных матриц, основанная на ранговом критерии Пури—Сена—Тамуры

Вновь на примере лейкократовых гранит-порфиров и аплит-порфиров «Паука», балки «Самолет» и Северного участка Тырныауза охарактеризуем процедуру классификации гранитоидов относительно характеристик рассеивания и коррелированности содержаний вольфрама и молибдена.

Проверяется гипотеза об однородности относительно ковариационных матриц. Нулевая гипотеза имеет вид: $H_0: \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \Sigma_0$ при наборе альтернативных гипотез $H_1: \Sigma_u \neq \Sigma_0$ хотя бы для одного u ($u = 1, 2, 3, 4$), где Σ — символ ковариационной матрицы многомерной случайной величины, причем Σ_1 — ковариационная матрица двухмерной случайной величины — модели содержаний вольфрама и молибдена в лейкократовых гранит-порфирах «Паука»; Σ_2 — то же, для гранит-порфиров балки «Самолет»; Σ_3 — то же, для аплит-порфиров Северного участка Тырныауза; Σ_4 — то же, для вертикального разреза лейкократовых гранитов Северного участка Тырныауза.

Для каждой пары (u, v) объектов вычисляется значение критерия Λ_Σ (u, v) Пури—Сена—Тамуры согласно процедуре, описанной в разделе 2.4.1. Результаты их следующие.

	Сопоставляемые выборки					
	1—2	1—3	1—4	2—3	2—4	3—4
Λ_Σ (u, v)	46,68	22,86	27,77	60,84	13,44	29,76

Как видно из приведенных данных, минимальное значение $\min \Lambda\Sigma = 13{,}44$, соответствующее объектам 2 и 4, значительно превысило критическое $\chi^2_{\alpha=0,05;\ f=2} = 5{,}99$ и поэтому ни одной однородной группы объектов статистически обоснованно выделить нельзя по ранговому критерию Пури—Сена—Тамуры. Все лейкократовые гранитоиды Тырныауза характеризуются различными дисперсиями и зависимостями содержаний вольфрама и молибдена.

2.4.3.4. Иерархическая процедура классификации набора объектов относительно ковариационных матриц, основанная на параметрическом критерии Кульбака

На примере указанных четырех типов лейкократовых гранитоидов Тырныауза осуществим их классификацию относительно ковариационных матриц на основе критерия Кульбака. Постановка геологической задачи и ее формализация аналогичны вышерассмотренным при классификации объектов с помощью рангового критерия Пури—Сена—Тамуры.

Для каждой пары (u, v) объектов рассчитывается статистика $2I_0(u, v)$ критерия Кульбака согласно процедуре, описанной в разделе 2.4.1. Результаты ее изложены ниже.

Сопоставляемые выборки

	1—2	1—3	1—4	2—3	2—4	3—4
$2I_0\ (u, v)$	233,71	115,36	195,96	285,1	98,07	260,27

Как видно из приведенных данных, минимальное значение $\min 2I_0(u, v) = 98{,}07$, соответствующее объектам 3 и 4, значительно превысило критическое $\chi^2_{\alpha, f} = 7{,}81$ при $\alpha = 0{,}05$ и $f = 0{,}5\,m\,(m+1) = 3$. Поэтому ни одной однородной группы относительно ковариационных матриц на основе критерия Кульбака выделить нельзя, т. е. все лейкократовые гранитоиды Тырныауза характеризуются различающейся коррелированностью и степенью рассеивания вольфрама и молибдена. Этот вывод совпал с ранее сделанным на основе рангового критерия Пури—Сена—Тамуры, поэтому может рассматриваться как вполне обоснованный.

ГЛАВА 3

РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА ПРИ РЕШЕНИИ ПРОГНОЗНЫХ ГЕОЛОГИЧЕСКИХ ЗАДАЧ

3.1. ПОСТАНОВКА ПРОГНОЗНЫХ ГЕОЛОГИЧЕСКИХ ЗАДАЧ

Прогнозные задачи в практике геологического исследования занимают весьма важное место. Проблема прогноза многоаспектна, множество задач, рассматриваемых как прогнозные, весьма обширно. Процесс прогнозирования геологических характеристик с применением математических методов можно разбить на пять этапов: 1) установление факта наличия статистической зависимости изучаемых геологических характеристик; 2) определение формы зависимости геологических характеристик; 3) определение параметров уравнения связи изучаемых характеристик; 4) определение условий возможности переноса найденной закономерности на интересующий исследователя регион; 5) получение количественных значений прогнозируемых характеристик.

В настоящей главе рассмотрены вопросы первого и третьего этапов прогнозирования. Отдельные аспекты задач второго этапа рассмотрены в главе 4. Подробнее вопросы определения формы зависимости изучаемых величин описаны в ряде опубликованных ранее работ [2, 15, 19, 41].

Типы статистической зависимости. Как уже отмечалось, наблюдаемые геологические характеристики при решении конкретной геологической задачи могут иметь вероятностный характер. В связи с этим можно выделить различные типы зависимостей между переменными величинами [19, с. 373; 2, с. 53—58; 15, с. 115].

М. Кендалл и А. Стьюарт [19, с. 373] рассматривают две группы зависимостей: статистическую и функциональную. В рамках статистической зависимости выделяются два случая: общий — когда все переменные величины случайны, и важный частный — когда лишь одна переменная случайна, а остальные детерминированы и заданы без ошибок. В общем случае формулируются две задачи: о взаимозависимости случайных величин и зависимости одной случайной величины от других случайных величин.

Выделяют три вида информации, содержащейся в наблюденных данных: 1) количественную, т. е. сами измеренные значения геологических характеристик; 2) ранговую, отражающую порядковые соотношения между наблюдаемыми величинами; 3) качественную, содержащую информацию о классификации исходных наблюдений на категории.

Взаимозависимость случайных величин анализируется методами теории корреляции. Если используется количественная или ранговая информации (или обе одновременно), то вычисляются соответствующие коэффициенты (параметрические и ранговые) парной и частной корреляции. Для ранговой информации в качестве единой меры связи совокупности геологических характеристик может быть найден коэффициент конкордации. Для категоризованных данных вычисляются коэффициенты сопряженности. Для перечисленных характеристик связи могут быть найдены как точечные, так и интервальные оценки и проверены гипотезы об их статистической значимости.

При изучении зависимости одной случайной величины от других случайных величин вычисляются множественный коэффициент корреляции, выражающий меру линейности такой зависимости, и корреляционное отношение — мера нелинейности связи. Кроме точечных оценок, для этих коэффициентов можно построить доверительные интервалы и проверить гипотезы о статистической значимости точечных оценок. Установив статистически значимую связь, изучают зависимость одних случайных величин от других. Для этой цели привлекаются регрессионные методы, с помощью которых оцениваются параметры зависимости, определяется их статистическая значимость, находятся соответствующие доверительные интервалы и совместная доверительная область для всех параметров регрессии, проверяется статистическая значимость построенного уравнения регрессии, а также находится для него доверительная область.

В частном случае, когда только одна переменная (зависимая) случайная, а остальные (независимые) детерминированы, задачи взаимозависимости и определения частной и множественной корреляции теряют статистический смысл и рассматриваются вопросы оценки регрессионной поверхности.

В рамках функциональной зависимости выделяются также два случая: собственно функциональная зависимость, когда все переменные неслучайны, но содержат ошибки измерений, и структурная зависимость, когда переменные случайны и, кроме того, подвержены ошибкам измерений. Заметим, что оба случая функциональной зависимости имеют статистическую природу.

Собственно функциональная зависимость изучается регрессионными методами, а структурная — методами корреляции и регрессии.

С. Айвазян [2, с. 53—58] выделяет четыре группы (схемы) зависимостей между переменными величинами.

Схема A — зависимость между неслучайными переменными; это обычный случай функционально связанных величин.

Схема B — зависимость случайной переменной η от неслучайных $x_1, \ldots, x_m$; в этом случае значения η при каждом фиксированном наборе значений $x_1, \ldots, x_m$ подвержены некоторому случайному разбросу и изучается усредненная по $x_1, \ldots, x_m$ зависимость, т. е. закономерность в изменении условного математического ожи-

дания $M(\eta|x_1, ..., x_m)$ в зависимости от изменений $x_1, ..., x_m$*.

Схема C_1 — зависимость между случайными по своей физической сущности величинами $\eta, \xi_1, ..., \xi_m$, зависящими от совокупностей неконтролируемых по своей физической сущности факторов. В этом случае изучается усредненный закон поведения η в зависимости от $\xi_1, ..., \xi_m$, т. е. изменение условного математического ожидания $M(\eta|\xi_1=x_1, \xi_2=x_2, ..., \xi_m=x_m)$ случайной величины η от значений $x_1, ..., x_m$, принимаемых случайными величинами $\xi_1, ..., \xi_m$. Кроме того, в этой схеме исследуется теснота связи между случайными величинами $\eta, \xi_1, ..., \xi_m$.

Схема C_2 — зависимость между неслучайными функционально связанными величинами, измеренными с некоторыми случайными ошибками, причем наличие случайных ошибок не позволяет непосредственно определить функциональную зависимость, так как в качестве исходных данных выступают выборочные значения случайных величин и вместо функциональной зависимости структурных величин приходится анализировать стохастическую связь между соответствующими случайными величинами, т. е. условное среднее случайной величины η при условии, что случайные величины $\xi_1, ..., \xi_m$ соответственно принимают значения $x_1, ..., x_m$: $M(\eta|\xi_1=x_1, ..., \xi_m=x_m)$ **.

В условиях этой схемы оценки параметров связи $M(\eta|\xi_1, ..., \xi_m)$, найденные согласно схемам B или C_1, теряют многие оптимальные свойства, такие, как несмещенность, эффективность и даже состоятельность.

Поставленные вопросы в схеме B решаются методами регрессионного анализа, в схеме C_1 — методами корреляционного анализа и в схеме C_2 — методами конфлюэнтного анализа. Задачи схемы A не являются статистическими и решаются методами вычислительной математики с привлечением других дисциплин высшей математики, таких, как высшая алгебра, математический анализ, функциональный анализ и т. п.

3.2. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧ ПРОГНОЗИРОВАНИЯ ГЕОЛОГИЧЕСКИХ ХАРАКТЕРИСТИК

При формальной постановке задач прогнозирования геологических характеристик будем руководствоваться схемами прогнозирования, описанными в 3.1.

---

* В силу неслучайности $x_1, ..., x_m$ правильнее говорить о безусловном математическом ожидании $M(\eta)=M(\eta|x_1, ..., x_m)$.

** В схеме C_2 [2, с. 53—58] структурными названы переменные модели с неслучайными величинами, содержащими ошибки измерений. Однако более популярно [19, с. 503; 15, с. 133; 41, с. 205] определение структурной модели как модели со случайными величинами, которые подвержены ошибкам измерений. Заметим, что в классификации типов зависимостей А. Айвазяна [2, с. 53—58] случай структурной зависимости не нашел отражения.

3.2.1. Схема корреляционного анализа

Имеем m-мерную случайную величину $\Xi=\{\xi_1, ..., \xi_m\}$ с известной m-мерной функцией распределения $F(x_1, ..., x_m)=P(\xi_1\leqslant x_1, ..., \xi_m\leqslant x_m)$. В этой схеме практически хорошо разработан только случай многомерного нормального распределения. Как известно, m-мерное нормальное распределение величины Ξ полностью определяется вектором математических ожиданий $\mu=\{\mu_i\}$, $i=1, 2, ..., m$ одномерных случайных величин ξ_i (правильнее говорить о маргинальных распределениях Ξ с функцией распределения $F(\infty, ..., x_i, ..., \infty)=P(\xi_1\leqslant\infty, ..., \xi_i\leqslant x_i, ..., \xi_m\leqslant\infty)$) и ковариационной матрицей $\Sigma=\{\sigma_{ij}\}$, $ij=1, 2, ..., m$, элементы которой являются ковариациями случайных величин ξ_i и ξ_j. В качестве исходных геологических данных могут выступать как сами параметры μ, Σ (такая ситуация, когда μ и Σ заданы, встречается крайне редко), так и матрица $X=\{x_{ti}\}$, $t=1, 2, ..., n$, $i=1, 2, ..., m$, где x_{ti} — выборочное (наблюдаемое) значение геологической характеристики с номером i в наблюдении с номером t. Здесь предполагается, что выборка произведена из совокупности с m-мерным нормальным распределением и что моделью геологической характеристики с номером i, т. е. x_i, является нормальная случайная величина ξ_i. В последнем случае исходя из X определяют оценки μ и Σ. Зная $\hat{\mu}$ и $\hat{\Sigma}$, вычисляют необходимые коэффициенты корреляции (парные, частные, множественные), корреляционное отношение, параметры уравнений регрессии. Затем оценивают значимость полученных оценок, строят различные доверительные области и проверяют необходимые гипотезы. Соответствующие определения и расчетные формулы приводятся в 3.3 и 3.4.1.

3.2.2. Схема регрессионного анализа

Линейная регрессия. Имеет место линейная модель $y=X\beta+\varepsilon$, где $y=\{y_t\}$, $t=1, 2, ..., n$ — вектор случайных наблюдений прогнозируемой геологической характеристики, называемой также зависимой переменной, или откликом; $X=\{x_{ti}\}$, $t=1, 2, ..., n$, $i=1, 2, ..., m$ — матрица известных (неслучайных) коэффициентов, столбцы этой матрицы x_i есть векторы неслучайных наблюдений прогнозирующих геологических характеристик, называемых независимыми переменными, или регрессорами; $\beta=\{\beta_i\}$, $i=1, 2, ..., m$ — вектор коэффициентов регрессии, подлежащей оценке; $\varepsilon=\{\varepsilon_t\}$, $t=1, 2, ..., n$ — вектор случайных ошибок, относительно которых предполагается, что они некоррелированны, имеют нулевые средние и постоянные дисперсии. Для оценки точности регрессии и доверительного оценивания дополнительно предполагается, что ошибки имеют совместное нормальное распределение. Задача прогноза в схеме регрессионного анализа заключается в оценке параметров линейной модели и прогнозных значений y, оценке точности найденных величин и их доверительного оценивания. Все необходимые для этого определения и формулы приводятся в 3.4.

3.3. РЕКОМЕНДУЕМЫЕ РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ ОЦЕНКИ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Оценки коэффициентов корреляции производятся для установления наличия или отсутствия интересующих исследователя линейных связей между отдельными или всеми геологическими характеристиками. Выявление статистически связанных величин позволяет надеяться, что они связаны также функциональной зависимостью, нахождение которой является центральной задачей прог-

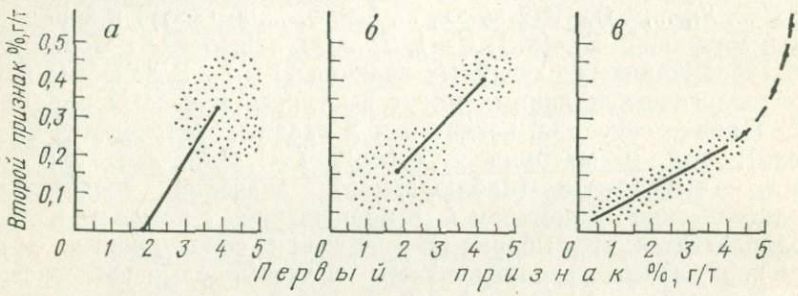


Рис. 8. Примеры ложных корреляционных зависимостей:
а) $r_n > r_\alpha$, $r_{n-1} < r_\alpha$; *б)* $r_n > r_\alpha$, $r_{n_1} < r_\alpha$, $r_{n_2} < r_\alpha$; *в)* $r_n < r_\alpha$, $r_{n-3} > r_\alpha$

нозирования геологических характеристик. При этом необходимо иметь в виду, что статистически значимая корреляционная связь может оказаться ложной. На рис. 8 представлено несколько случаев ложной корреляционной связи двух геологических характеристик, где через r_n обозначена оценка коэффициента корреляции, полученная по n точкам, а через r_{n-1}, r_{n_1}, r_{n_2}, r_{n-3} — соответствующие оценки по $n-1$, n_1, n_2, $n-3$ точкам. Через r_α обозначены соответствующие критические значения коэффициентов корреляции.

Несмотря на то, что $r_n > r_\alpha$ (см. рис. 8, *а*, *б*) или $r_{n-3} > r_\alpha$ (см. рис. 8, *в*), значимая корреляционная связь является ложной, поскольку $r_{n-1} < r_\alpha$ (см. рис. 8, *а*), $r_{n_1} < r_\alpha$, $r_{n_2} < r_\alpha$ (см. рис. 8, *б*), $r_n < r_\alpha$ (см. рис. 8, *в*). В настоящем разделе рассмотрены наряду с параметрическими коэффициентами парной, частной и множественной корреляции также их ранговые аналоги. Параметрические оценки коэффициентов корреляции как меры связи случайных величин статистически обоснованы лишь в условиях корреляционной схемы, когда предполагается наличие многомерного нормального распределения. В условиях отклонения от нормального распределения эти коэффициенты частично теряют свои оптимальные свойства, однако в ряде случаев они все же применимы.

Значительно более устойчивыми к нарушению предположения нормальности являются ранговые коэффициенты корреляции.

Кроме того, примененные в условиях нормального распределения, они, как правило, дают близкие результаты с оптимальными в этих условиях параметрическими коэффициентами. Все это позволяет рекомендовать ранговые коэффициенты корреляции для установления статистической связи анализируемых геологических признаков. Параметрические коэффициенты корреляции являются мерами линейной связи и могут указывать лишь на такую связь. Для выявления нелинейных зависимостей применяется корреляционное отношение. Ранговые коэффициенты корреляции, будучи независящими от системы мер, могут указать на наличие нелинейной зависимости, например логарифмической, экспоненциальной, квадратической, кубической и т. п., и таким образом реализуют функции корреляционного отношения.

3.3.1. Коэффициент парной корреляции

В качестве меры линейной парной зависимости случайных величин ξ и η применяется коэффициент парной корреляции $\rho(\xi, \eta)$ [19, с. 384; 6, с. 75—80].

Как мера линейной зависимости коэффициент парной корреляции обладает рядом замечательных свойств, в частности, если ξ и η независимы, то $\rho(\xi, \eta) = 0$.

Наряду с положительными свойствами коэффициент парной корреляции обладает рядом отрицательных [18, с. 367—368]: в условиях распределений, отличающихся от нормального, он часто теряет свои свойства как мера линейной зависимости, а именно сильно чувствителен к наличию аномальных значений (выбросов); зависит от системы мер, если заменить ξ на $F(\xi)$, где $F(\xi)$ — монотонная функция ξ (например, строго возрастающая), то $\rho[F(\xi), \eta] \neq \rho(\xi, \eta)$.

Эти свойства парного коэффициента корреляции отсутствуют у его ранговых аналогов.

Оценкой коэффициента корреляции по выборочным данным является выборочный коэффициент корреляции r [18, с. 376]

$$r = \frac{\sum_{t=1}^{n} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{\sqrt{\sum_{t=1}^{n} (x_{ti} - \bar{x}_i)^2 \sum_{t=1}^{n} (x_{tj} - \bar{x}_j)^2}}.$$

При малом объеме выборки значение r получается несколько заниженным. В этом случае ($n < 10$) лучше использовать для ρ другую оценку [18, с. 376]

$$r^* = r\left[1 + \frac{1 - r^2}{2(n-3)}\right].$$

Проверка значимости полученного значения коэффициента корреляции, т. е. проверка гипотезы о том, может ли выборочный

коэффициент корреляции иметь случайные отклонения от нуля для генеральной совокупности с параметром $\rho=0$, проводится одним из следующих способов.

1. С помощью специальных таблиц [18, с. 392] для любых n, особенно при $n<10$, когда другие методы проверки значимости становятся некорректными.

2. Вычислением статистики $t=r\sqrt{n-2}/\sqrt{1-r^2}$ [18, с. 391], распределенной в условиях проверяемой гипотезы по закону Стьюдента с $n-2$ степенями свободы. Если $t\geqslant t_{\alpha, n-2}$, где $t_{\alpha, n-2}$ — критическое значение распределения Стьюдента при уровне значимости α и $n-2$ степенях свободы, то гипотеза отклоняется.

3. Вычислением статистики $F=r^2(n-2)/(1-r^2)$, [18, с. 391], имеющей в условиях проверяемой гипотезы F-распределение с 1 и $n-2$ степенями свободы.

4. Вычислением статистики $F=(1+r)/(1-r)$, [18, с. 391], имеющей в условиях проверяемой гипотезы F-распределение с $n-2$ и $n-2$ степенями свободы.

5. С помощью Z-преобразования Фишера (при $n>10$) [18, с. 393]

$$Z=\sqrt{n-3}\left(\frac{1}{2}\ln\frac{1+r}{1-r}\right).$$

Статистика Z имеет стандартное нормальное распределение.

6. С помощью преобразования Хотелинга (при $n<50$) [18, с. 393]

$$Z_{\text{н}}=\sqrt{n-1}\,[Z-(3Z+r)/4n].$$

Статистика $Z_{\text{н}}$ имеет стандартное нормальное распределение.

7. С помощью графика 95%-ных доверительных границ для коэффициента корреляции [18, с. 390]. Рассчитав r, снимаем с графика значение нижней и верхней границ 95%-ного доверительного интервала. Если доверительный интервал не содержит нулевого значения, то гипотеза $\rho=0$ отвергается. Доверительный интервал для коэффициента корреляции может быть найден с помощью любой статистики, применяемой для проверки его статистической значимости.

Например, используя Z-преобразование Фишера, получим 95%-ный доверительный интервал для ρ

$$\frac{e^{\frac{2}{\sqrt{n-3}}(Z-1,96)}-1}{e^{\frac{2}{\sqrt{n-3}}(Z-1,96)}+1}<\rho<\frac{e^{\frac{2}{\sqrt{n-3}}(Z+1,96)}-1}{e^{\frac{2}{\sqrt{n-3}}(Z+1,96)}+1}.$$

На рис. 9, *а* и 10, *а* представлены результаты расчета на ЭВМ значений коэффициента парной корреляции на 19 редких, рудных и петрогенных элементов по данным опробования поверхности и скважины в эльджуртинском граните Тырныауза, а на рис. 11, *а*,

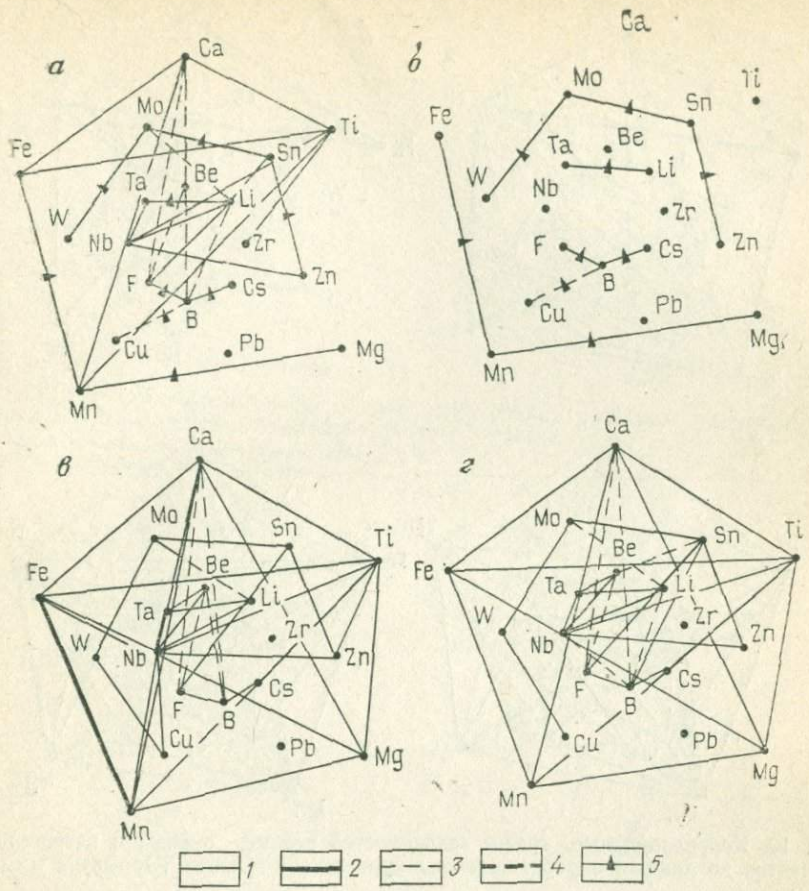


Рис. 9. Корреляционные графы зависимостей редких, рудных и петрогенных элементов по данным опробования поверхности эльджуртинского гранита Тырныауза:

а — парная зависимость; *б* — частная зависимость; *в* — ранговая зависимость по Спирмену; *г* — ранговая зависимость по Кендаллу

Значимые связи: *1* — положительные; *2* — сильные положительные; *3* — отрицательные; *4* — сильные отрицательные; *5* — устойчивые (связь значимая и парная и частная)

12, *а*, 13, *а*, 14, *а* — результаты расчета на 13 редких и рудных элементов по данным опробования гранит-порфиров «Паука», балки «Самолет», аплит-порфиров и вертикального разреза гранитов Северного участка Тырныауза.

3.3.2. Коэффициент ранговой корреляции Спирмена

Коэффициентом ранговой корреляции Спирмена $\rho_s(\xi, \eta)$ двух дискретных случайных величин ξ и η, являющихся моделями геологических характеристик, называется коэффициент парной кор-

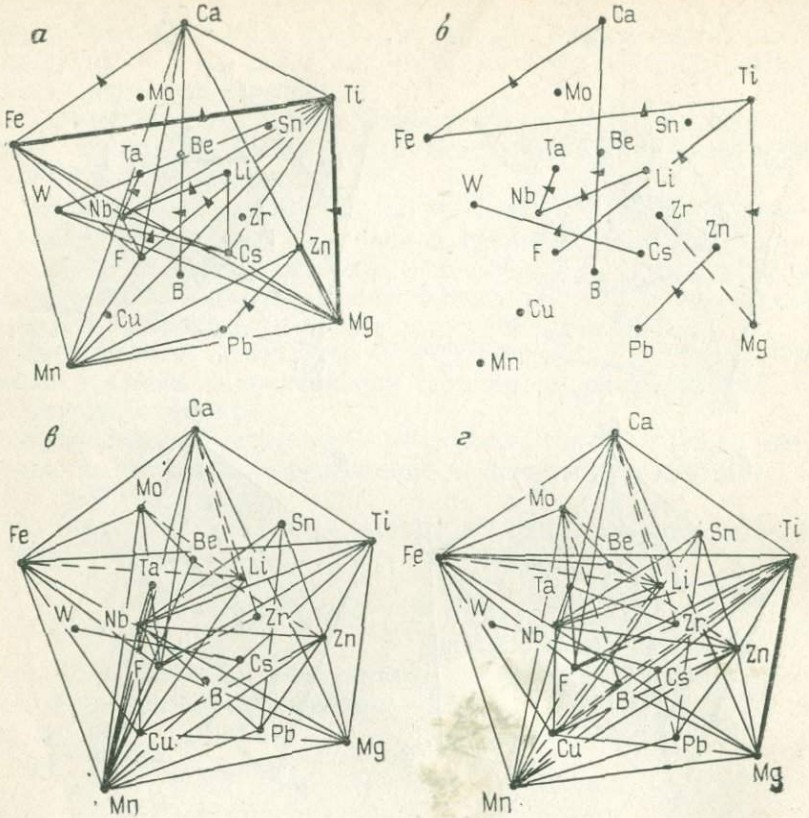


Рис. 10. Корреляционные графы зависимостей редких, рудных и петрогенных элементов по данным опробования эльджуртинского гранита Тырныауза в скважине:

а — парная зависимость; *б* — частная зависимость; *в* — ранговая зависимость по Спирмену; *г* — ранговая зависимость по Кендаллу.
Условные обозначения те же, что и к рис. 9

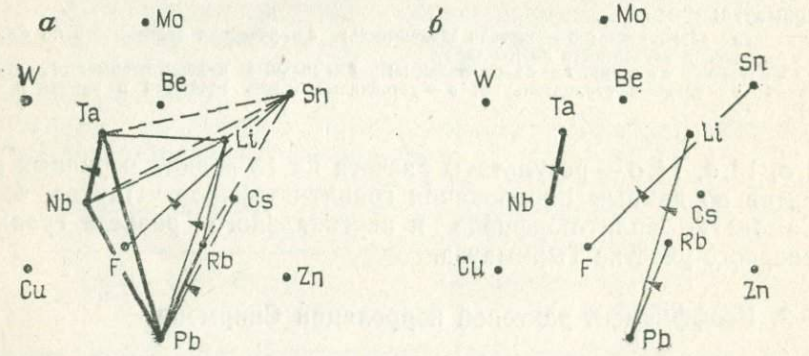


Рис. 11. Корреляционные графы зависимостей редких и рудных элементов по данным опробования гранит-порфиров «Паука» в скважине:

а — парная зависимость; *б* — частная зависимость.
Условные обозначения те же, что и к рис. 9

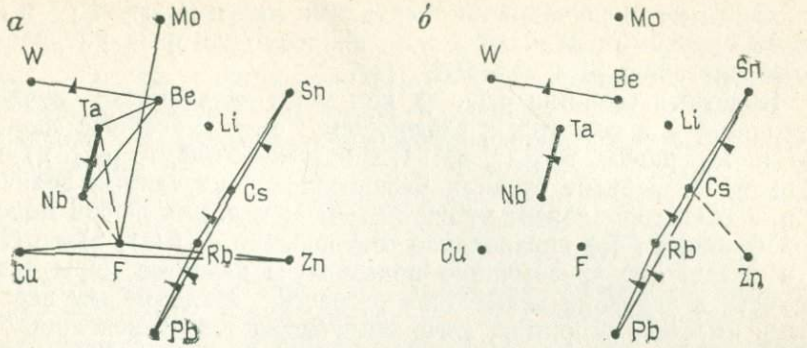


Рис. 12. Корреляционные графы зависимостей редких и рудных элементов по данным опробования гранит-порфиров балки «Самолет»:

а — парная зависимость; *б* — частная зависимость.
Условные обозначения те же, что и к рис. 9

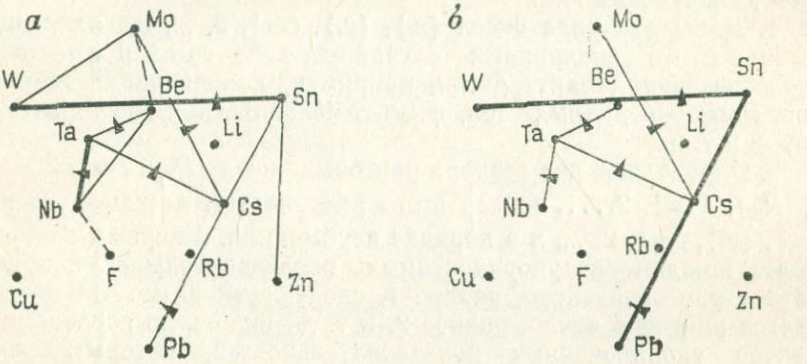


Рис. 13. Корреляционные графы зависимостей редких и рудных элементов по данным опробования аплит-порфиров Северного участка Тырныауза:

а — парная зависимость; *б* — частная зависимость.
Условные обозначения те же, что и к рис. 9.

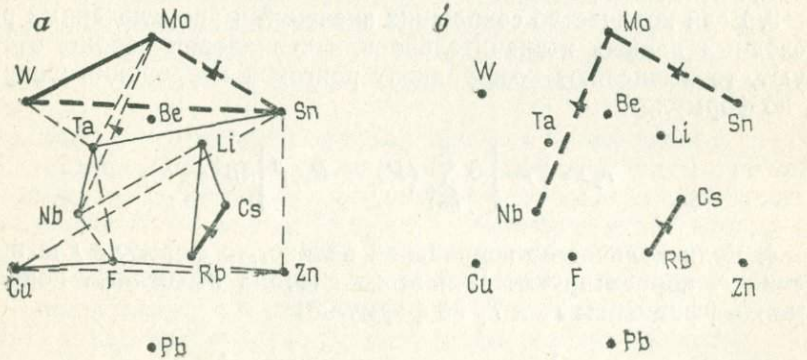


Рис. 14. Корреляционные графы зависимостей редких и рудных элементов по данным опробования гранитов из вертикального разреза Северного участка Тырныауза:

а — парная зависимость; *б* — частная зависимость.
Условные обозначения те же, что и к рис. 9.

реляции между случайными величинами R_ξ и R_η, где R_ξ и R_η — ранги случайных величин ξ и η, значения которых упорядочены по возрастанию, $\rho_s(\xi, \eta) = \rho(R_\xi, R_\eta)$.

Полезные свойства $\rho_s(\xi, \eta)$ как меры зависимости случайных величин ξ и η сводятся к следующему: если случайные величины R_ξ и R_η равны, то $\rho_s(\xi, \eta) = 1$, а если обратны, то $\rho_s(\xi, \eta) = -1$; $\rho_s(\xi, \eta)$ не зависит от вида распределения случайных величин ξ и η, а также не зависит от системы мер, т. е. для любой монотонной функции $F(\xi)$ справедливо соотношение $\rho_s(F(\xi), \eta) = \rho_s(\xi, \eta)$, что делает его чрезвычайно полезным в качестве меры зависимости в условиях нелинейной регрессии, когда между переменными имеется, например, логарифмическая или экспоненциальная зависимость или, вообще, когда с увеличением одной переменной другая в среднем или непрерывно растет, или непрерывно убывает.

Алгоритм вычисления коэффициента ранговой корреляции Спирмена и *проверки гипотезы* о его статистической значимости имеет следующий вид.

1. Даны два ряда чисел $\{x_t\}$, $\{y_t\}$, $t = 1, 2, \ldots, n$, являющихся выборочными значениями случайных величин ξ и η — моделей геологических характеристик, например, концентраций компонентов, размеров рудных тел, физико-механических свойств горных пород и т. п.

2. Определяют два вектора ранговых чисел $\{R_{x_t}\}$, $t = 1, 2, \ldots, n$, $\{R_{y_t}\}$, $t = 1, 2, \ldots, n$. Для этого располагают каждый из рядов $\{x_t\}$, $\{y_t\}$, $t = 1, 2, \ldots, n$ в порядке неубывания, и первым элементам рядов каждой из упорядоченных последовательностей присваивается ранговое число, равное 1, следующим элементам присваивается ранговое число, равное 2, и т. д., вплоть до последних элементов упорядоченных последовательностей, которым присваивается ранг n. Равным значениям исходных последовательностей присваиваются средние ранговые числа. В качестве R_{x_t} и R_{y_t} берут номера x_t и y_t соответственно в упорядоченных последовательностях исходных данных.

3. Если количество совпавших значений в упорядоченных рядах исходных данных незначительно или совпадения вообще отсутствуют, то вычисляют коэффициент ранговой корреляции Спирмена r_s по формуле

$$r_s = 1 - \left[6 \sum_{t=1}^{n} (R_{x_t} - R_{y_t})^2\right] / (n^3 - n).$$

4. Если количество совпадений велико, то формулу для вычисления r_s корректируют. Для этого сначала вычисляют поправочные коэффициенты T_x и T_y по формулам

$$T_x = \frac{1}{12} \sum_{l=1}^{l_x} (u_l^3 - u_l) \quad \text{и} \quad T_y = \frac{1}{12} \sum_{l=1}^{l_y} (v_l^3 - v_l),$$

где l_x и l_y соответственно числа групп равных ранговых чисел в рядах $\{R_{x_t}\}$, $t=1, 2, \ldots, n$ и $\{R_{y_t}\}$, $t=1, 2, \ldots, n$, а u_l и v_l — число членов в группах равных ранговых чисел с номером соответственно в рядах $\{R_{x_t}\}$, $t=1, 2, \ldots, n$ и $\{R_{y_t}\}$, $t=1, 2, \ldots, n$.

Значение коэффициента ранговой корреляции Спирмена находят по формуле

$$r_s = 1 - \frac{6[S(d^2) + T_x + T_y]}{n(n-1)}.$$

5. Для проверки значимости r_s выбирают уровень значимости α и если $n \leqslant 10$, то с помощью таблицы [18, с. 369] для выбранного уровня значимости α и фиксированного n находят критическое значение $r_s^{\text{кр}}$. Вычисленное значение r_s сравнивают с $r_s^{\text{кр}}$. Если $r_s \geqslant r_s^{\text{кр}}$, то гипотеза $H_0: \rho_s = 0$ отклоняется, в противном случае она принимается как подтвердившаяся. Следует подчеркнуть, что все же в $100\alpha\,\%$ случаев гипотеза H_0 будет отвергнута, когда на самом деле она верна (последнее замечание в равной степени относится и к пп. 6 и 7).

6. Если $10 \leqslant n \leqslant 30$, то снова можно воспользоваться той же таблицей и повторить п. 5. Однако часто удобнее и быстрее оценить значимость r_s с помощью величины

$$t = |r_s| \sqrt{(n-2)/(1-r_s^2)},$$

которая в условиях проверяемой гипотезы распределена по закону Стьюдента с $n-2$ степенями свободы. Поэтому если $t \geqslant t_{\alpha, n-2}$, то гипотеза $H_0: \rho_s = 0$ отклоняется при уровне значимости α, в противном случае H_0 принимается как подтвердившаяся.

7. Если $n \geqslant 30$, то для оценки значимости r_s можно снова привлечь t-статистику из п. 6, однако более простой будет процедура с использованием величины $Z = |r_s| \sqrt{n-1}$, которая в условиях проверяемой гипотезы распределена приблизительно нормально со средним значением, равным 0, и дисперсией, равной 1. Поэтому если $Z \geqslant Z_\alpha$, то гипотеза $H_0: \rho_s = 0$ отклоняется при уровне значимости α, в противном случае H_0 принимается как подтвердившаяся.

Проиллюстрируем процедуру вычисления коэффициента ранговой корреляции Спирмена на примере содержаний рудных элементов вольфрама и молибдена по данным опробования поверхности в эльджуртинском граните Тырныауза. Вписываем эти данные во вторую и третью колонки табл. 15. Располагая каждый из рядов значений содержаний вольфрама и молибдена в порядке возрастания, приписываем каждой пробе два ранговых числа. Первое число — номер пробы в упорядоченном ряду значений вольфрама, второе число — аналогичный номер в ряду значений молибдена. Записываем эти ранговые числа в четвертую и пятую колонки таблицы. Затем находим связанные ранги, присваиваем пробам,

Таблица 15

Процедура вычисления коэффициента ранговой корреляции Спирмена для вольфрама и молибдена по данным опробования поверхности в эльджуртинском граните Тырныауза

Номер п/п	W	Mo	R^W	R^{Mo}	$R^W_{св}$	$R^{Mo}_{св}$	ΔR	$(\Delta R)^2$
1	0,7	1,4	1	41	9,5	41,5	—32,0	1024
2	1,5	1,3	28	37	29,5	38,5	—9	81
3	1,5	0,9	29	16	29,5	19,5	10	100
4	5,1	1,8	49	48	49,5	48,5	1	1
5	1,7	0,9	35	17	36,5	19,5	17	289
6	1,5	1,0	30	24	29,5	26,5	3	9
7	1,7	0,9	36	18	36,5	19,5	17	289
8	1,9	1,8	41	49	41,5	48,5	—7	49
9	1,6	1,3	32	38	33	38,5	—5,5	30,25
10	1,8	1,5	39	43	39,5	44,5	—5	25
...	...	...	...	...	...	...	...	...
41	2,1	1,3	45	39	44,5	38,5	6	36
42	1,7	1,3	38	40	36,5	38,5	—2	4
43	1,6	1,0	34	29	33	26,5	6,5	42,25
44	0,7	1,1	17	36	9,5	33	—23,5	552,25
45	2,8	1,7	46	47	46,5	47	—0,5	0,25
46	5,1	1,9	50	50	49,5	50	—0,5	0,25
47	3,4	0,8	48	15	48	13,5	34,5	1190,25
48	0,7	0,7	18	11	9,5	7,5	2	4
49	1,0	0,9	22	23	20,5	19,5	1	1
50	2,8	1,4	47	42	46,5	41,5	5	25

имеющим одинаковые содержания соответствующего элемента, равные ранговые числа. Для этого ранги, соответствующие одинаковым содержаниям рудных элементов, суммируются и определяется среднее ранговое число как среднее арифметическое несвязанных рангов. Например, рассматривая содержания вольфрама по данным опробования поверхности в эльджуртинском граните, легко убедиться, что минимального значения, равного 0,7, содержания вольфрама достигнут в 1, 14, 16, 19, 20, 21, ..., 48-й пробах. Всем этим пробам присваивается одно и то же общее среднее ранговое число, которое вычисляется следующим образом:

$$R = \frac{1 + 14 + 16 + 19 + 20 + 21 + \ldots + 48}{18} = 9{,}5.$$

Ряды связанных рангов заносятся в шестую и седьмую колонки табл. 15. Затем вычисляется ряд разностей связанных рангов (8-я колонка) и ряд квадратов разностей связанных рангов (9-я колонка). Девятая колонка суммируется и определяется значение $S(d^2) = 1024 + \ldots + 25 = 11\,813$.

Вычисляются поправки, учитывающие наличие связей рангов:

$$T_x = \frac{1}{12}\sum(u^3 - u) = \frac{1}{12}(18^3 - 18 + \ldots + 2^3 - 2) = 504{,}5.$$

$$T_y = \frac{1}{12}\sum(v^3 - v) = \frac{1}{12}(8^3 - 8 + \ldots + 2^3 - 2) = 145{,}5.$$

Напомним формулу для вычисления коэффициента ранговой корреляции Спирмена

$$r_s = 1 - \frac{6[S(d^2) + T_x + T_y]}{n(n^2 - 1)},$$

$$r_s = 1 - \frac{6[11\,813 + 504{,}5 + 145{,}5]}{50(50 \cdot 50 - 1)} = 0{,}41.$$

При проверке значимости r_s учитываем, что число проб равно 50 ($50 \geqslant 30$). Следовательно, можно использовать статистику $\hat{Z} = |r_s|/\sqrt{n-1}$, распределенную в условиях нулевой гипотезы по стандартному нормальному закону. Для 5%-ного уровня значимости находим соответствующее значение стандартного нормального распределения $r_{0,05} = 1{,}96$. Вычислим значение статистики $Z = 0{,}41 \cdot \sqrt{50-1} = 2{,}87$.

Поскольку $2{,}87 = \hat{Z} > r_{0,05} = 1{,}96$, то гипотеза о равенстве нулю коэффициента ранговой корреляции Спирмена при 5 %-ном уровне значимости отклоняется.

Для подтверждения полученного результата приведем значение параметрического коэффициента парной корреляции для тех же исходных данных: $r = 0{,}56$.

Определим 95%-ный доверительный интервал с помощью графика [18, с. 390]. Для $n = 50$ и $r = 0{,}56$ получаем доверительный интервал для ρ: $0{,}35 \leqslant \rho \leqslant 0{,}72$.

Поскольку доверительный интервал не включает значения $\rho = 0$, то можно говорить о наличии значимой корреляции ($\rho \neq 0$) между данными опробования по вольфраму и молибдену.

Результаты расчета на ЭВМ коэффициентов ранговой корреляции Спирмена для каждой пары геологических характеристик на 19 редких, рудных и петрогенных элементов по данным опробования поверхности и скважины в эльджуртинском граните Тырныауза представлены в виде корреляционных графиков на рис. 9, *в*, 10, *в*. Нанесены лишь статистически значимые связи.

3.3.3. Тетрахорический коэффициент корреляции Бломквиста

Этот критерий, иногда называемый методом дробового выстрела, предназначен для проверки гипотезы о независимости двух сравниваемых случайных величин, которые в геологических исследованиях могут быть математическими аналогами двух изучаемых геологических характеристик, например концентраций элементов или минералов в породах, размеров геологических объектов, физических свойств горных пород, абсолютного возраста пород и т. п.

Коэффициент Бломквиста нашел широкое применение в качестве критерия для проверки выводов об отсутствии или наличии связи между изучаемыми геологическими характеристиками. Следует еще раз подчеркнуть, что ранговые критерии (в том числе коэффициент Бломквиста) устойчивы по отношению к нарушению условий нормальности распределений. В практических ситуациях наиболее оправданно применение коэффициента Бломквиста при $n>50$, где n — объем двухмерной выборки исходных геологических данных.

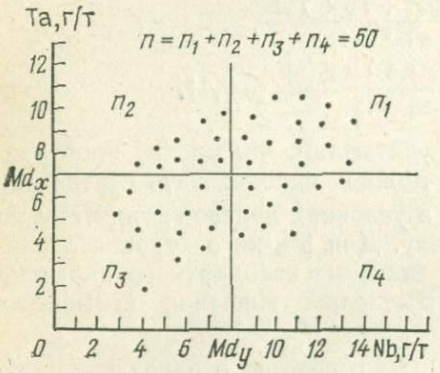


Рис. 15. Корреляционное поле по тетрахорическому коэффициенту корреляции Бломквиста

Коэффициент Бломквиста описан в многих работах [18, 19]. При вычислении этого тетрахорического коэффициента определяется число двухмерных точек наблюдений в соответствующих квадрантах двухмерного корреляционного графика, осями которого являются соответствующие медианы.

Весьма просто реализуется процедура вычисления коэффициента корреляции Бломквиста.

Пусть x_t и y_t, $t=1, 2, ..., n$ — исходные данные двух сопоставляемых геологических признаков x и y, образующих двухмерную выборку объема n.

1. По исходным данным x_t и y_t составляется двухмерная точечная диаграмма.

2. Проводятся две медианы Md_x и Md_y. Естественно, что $n_1+n_2\simeq n_3+n_4$, $n_1+n_4\simeq n_2+n_3$ и $n_1+n_2+n_3+n_4=n$.

Объемы n_1, n_2, n_3 и n_4 означают число двухмерных точек наблюдений, попавших соответственно в 1, 2, 3 и 4-й квадранты, образованные медианами Md_x и Md_y.

При попадании точек на медианные оси Md_x и Md_y точки пропорционально разносятся в разные квадранты (рис. 15).

3. Рассчитывают величину

$$R=\frac{(n_1+n_3)-(n_2+n_4)}{n},$$

которая в условиях гипотезы о независимости случайных величин ξ и η асимптотически нормально распределена со средним, равным 0 и дисперсией, равной 1.

Поэтому гипотеза о независимости (отсутствии связи) случайных величин ξ и η принимается, если $R\sqrt{n-1}\leqslant t_\alpha$, и отклоняется ($\xi$ и η коррелированы), если $R\sqrt{n-1}>t_\alpha$, где t_α — критическое значение при заданном уровне значимости α.

Для примера на рис. 15 $R=0{,}2$; $R\sqrt{n-1}=1{,}4$; $t_{0{,}05}=1{,}96$; и принимается гипотеза об отсутствии связи.

3.3.4. Частная и множественная корреляция. Конкордация

Если изучаемые случайные величины ξ_i и ξ_j рассматривать совместно с остальными случайными величинами $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_m$, следует учитывать то обстоятельство, что изменение ξ_i и ξ_j вызывается в какой-то степени изменением других величин. Влияние других случайных величин на ξ_i и ξ_j может как усилить, так и ослабить парную корреляционную связь между ξ_i и ξ_j.

Для устранения этого влияния обычно используют так называемый частный коэффициент корреляции, который оценивает связь между ξ_i и ξ_j после устранения изменений, вызванных влиянием остальных случайных величин. Для этого коэффициента существует и ранговый аналог.

Коэффициент частной корреляции. Рассмотрим m случайных величин $\xi_1, \xi_2, \ldots, \xi_m$, являющихся моделями геологических характеристик $x_1, x_2, \ldots, x_m$. Пусть q — набор индексов $1, 2, \ldots, m$ без i и j. Тогда коэффициентом частной корреляции между ξ_i и ξ_j при фиксированных $m-2$ оставшихся величинах называется величина

$$\rho_{ij\cdot q}=\frac{-C_{ij}}{\sqrt{C_{ii}C_{jj}}},$$

где C_{ij} — алгебраическое дополнение, соответствующее элементу ρ_{ij} в определителе корреляционной матрицы

$$C=\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \ldots & \rho_{1m} \\ \rho_{21} & 1 & \rho_{23} & \ldots & \rho_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{m1} & \rho_{m2} & \rho_{m3} & \ldots & 1 \end{pmatrix}.$$

Справедливо соотношение $C_{ij}=|C|C^{ij}$, где C^{ij} — элемент матрицы C^{-1}, обратной C.

Обозначим через p набор индексов $1, 2, \ldots, m$ без i, j, k. Тогда $\rho_{ij\cdot q}$ можно выразить через коэффициенты частной корреляции на единицу меньших порядков

$$\rho_{ij\cdot q}=\frac{\rho_{ij\cdot p}-\rho_{ik\cdot p}\rho_{jk\cdot p}}{\sqrt{(1-\rho^2_{ik\cdot p})(1-\rho^2_{jk\cdot p})}}.$$

Выборочным коэффициентом частной корреляции $r_{ij\cdot q}$ случайных величин ξ_i и ξ_j, являющихся моделями геологических характеристик x_i и x_j, при фиксированных $m-2$ оставшихся величинах называется отношение

$$r_{ij\cdot q}=-\frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}},$$

где B_{ij} — алгебраическое дополнение выборочной корреляционной матрицы $\{r_{ij}\}$ случайных величин $\xi_1, \ldots, \xi_m$, соответствующее элементу r_{ij}.

Распределение $r_{ij \cdot q}$, построенное по n наблюдениям X_t, совпадает с распределением выборочного коэффициента парной корреляции r_{ij} с заменой n на $n-m+2$, так что для оценки значимости коэффициента частной корреляции применим аналогичный критерий для коэффициента парной корреляции с уменьшением числа степеней свободы на $m-2$.

На рис. 9, *б*, 10, *б*, 11, *б*, 12, *б*, 13, *б*, 14, *б* представлены в виде корреляционных графов результаты расчета на ЕС ЭВМ коэффициентов частной корреляции для каждой пары геологических характеристик, очищенных от влияния остальных признаков. На графах нанесены только статистически значимые частные связи. На рис. 9, *б* и 10, *б* представлены результаты расчета на 19 редких, рудных и петрогенных элементов по данным опробования эльджуртинского гранита Тырныауза с поверхности и в скважине, а на рис. 11, *б*, 12, *б*, 13, *б*, 14, *б* — результаты расчета на 13 редких и рудных элементов по данным опробования гранит-порфиров «Паука», балки «Самолет», аплит-порфиров и гранитов вертикального разреза Северного участка Тырныауза.

Ранговый коэффициент частной корреляции. Рассмотрим наряду со случайными величинами $\xi_1, \ldots, \xi_m$ случайные величины $R^1, \ldots, R^m$, являющиеся рангами $\xi_1, \ldots, \xi_m$.

Определим частный ранговый коэффициент корреляции $\rho_{ij \cdot k}$, $i, j, k = 1, 2, \ldots, m$ следующим образом.

Расположим один под другим три ряда значений рангов:

$$R^k = R_1^k, \ R_2^k, \ \ldots, \ R_n^k,$$
$$R^i = R_1^i, \ R_2^i, \ \ldots, \ R_n^i,$$
$$R^j = R_1^j, \ R_2^j, \ \ldots, \ R_n^j.$$

Образуем всевозможные пары из n точек: (1,2), (1.3), ..., $(n-1, n)$ (число таких пар равно $C_n^2 = \frac{n(n-1)}{2}$) и составим таблицу размерности $3C_n^2$:

	(1, 2)	(1,3),	. . .,	$(n-1, n)$
R^k	+	+	. . . . .	+
R^i	+	—	. . . . .	+
R^j	—	—	. . . . .	+

Первую строку этой таблицы, соответствующую R^k, считаем стандартной и заполняем ее целиком знаками «+». При заполнении 2 и 3-й строк этой таблицы руководствуемся следующим правилом: если пара рангов $i_1 i_2$ (или $j_1 j_2$) расположена в том же порядке, что и пара рангов $k_1 k_2$, то соответствующая клетка запол-

няется знаком «+», если же порядок рангов обратный, то знаком «—».

Затем составляем классификационную таблицу 2×2 (табл. 16).

Частным ранговым коэффициентом корреляции $\dot{\rho}_{ij\cdot k}$ случайных величин R^i и R^j с R^k называется величина

$$\dot{\rho}_{ij\cdot k} = \frac{(ad - bc)}{(a+b)(c+d)(a+c)(b+d)}.$$

Определенный таким образом коэффициент принимает значение от —1 до +1, с его помощью можно измерить степень согласованности последовательностей R^i и R^j в совпадении порядка их рангов с порядком рангов в R^k.

Таблица 16

Ранги i	Ранги j: Число пар	Пары с «+»	Пары с «—»	Сумма
	Пары с «+»	a	b	$a+b$
	Пары с «—»	c	d	$c+d$
	Сумма	$a+c$	$b+d$	N

Если частный ранговый коэффициент корреляции принимает значение +1, то это означает, что указанные ранги полностью согласуются между собой, если он равен —1, то предпочтения противоположные (рассогласованы), если же он обращается в 0, так что $a/b = c/d$, то предпочтения оказываются совершенно независимыми.

где $N = C_n^2 = a + b + c + d$.

Введем еще одну ранговую меру парной корреляции — ранговый коэффициент корреляции Кендалла

$$\dot{\rho}_{ij} = \frac{(a+d) - (b+c)}{N},$$

обладающий такими же полезными свойствами, как и коэффициент ранговой корреляции Спирмена.

Тогда коэффициент частной ранговой корреляции $\dot{\rho}_{ij\cdot k}$ может быть выражен через коэффициенты парной ранговой корреляции Кендалла $\dot{\rho}_{ij}$, $\dot{\rho}_{ik}$, $\dot{\rho}_{jk}$ следующим образом:

$$\dot{\rho}_{ij\cdot k} = \frac{\dot{\rho}_{ij} - \dot{\rho}_{kj}\dot{\rho}_{ki}}{\sqrt{(1-\dot{\rho}_{kj}^2)}\sqrt{(1-\dot{\rho}_{ki}^2)}}.$$

Для удобства пользования выпишем формулы для $\dot{\rho}_{kj}$ и $\dot{\rho}_{ki}$

$$\dot{\rho}_{kj} = \frac{(a+c) - (b+d)}{N}, \quad \dot{\rho}_{ki} = \frac{(a+b) - (c+d)}{N}.$$

Методы проверки статистической значимости ранговых коэффициентов частной корреляции до настоящего времени не разработаны. Непригодной оказывается здесь проверка по критерию χ^2,

поскольку между некоторыми рангами, входящими в величины a, b, c, d, существует взаимная зависимость, например, если ранг A меньше ранга B, а ранг B меньше ранга C, то ранг A должен быть меньше ранга C. Однако это условие часто бывает нарушено. «Здесь мы сталкиваемся с тем разделом статистической теории, который требует дальнейших исследований» [20, с. 135].

Коэффициент множественной корреляции. Для характеристики зависимости одной случайной величины ξ_i — модели одной геологической характеристики от совокупности других случайных величин $\xi_1, \xi_2, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$ служит коэффициент множественной корреляции. Пусть k — набор индексов $1, 2, \ldots, i-1, i+1, \ldots, m$. Коэффициентом множественной корреляции $\rho_{i \cdot k}$ случайной величины ξ_i от набора случайных величин $\xi_1, \xi_2, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$ называется величина

$$\rho_{i \cdot k} = \sqrt{1 - \frac{1}{C^{ii}}},$$

где C^{ii} — диагональный элемент матрицы C^{-1}, обратной корреляционной матрице C.

Справедливо соотношение

$$\rho_{i \cdot k} = \sqrt{1 - \prod_{j=1,\, j \neq i}^{m} (1 - \rho^2_{ij \cdot L_j})},$$

связывающее коэффициент множественной корреляции с набором из $m-1$ коэффициентов частной корреляции $\rho_{ij \cdot L_j}$ между ξ_i и ξ_j при фиксированных величинах $\xi_1, \xi_2, \ldots, \xi_{j-1}$, но без ξ_i (таким образом, L_j есть множество индексов $1, 2, \ldots, j-1$, но без i).

Коэффициент множественной корреляции $\rho_{i \cdot k}$ не меньше модуля произвольного коэффициента частной корреляции $\rho_{ij \cdot l}$, $j \in k$, $l \in k$, $j \notin l$; $\rho_{i \cdot k} \geqslant |\rho_{ij \cdot l}|$.

Если $\rho_{i \cdot k} = 0$, то обращаются в нуль все коэффициенты частной корреляции $\rho_{ij \cdot l}$ и, следовательно, ξ_i полностью не коррелированы как со всем набором оставшихся случайных величин $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$, так и с любой частью этих величин.

Если $\rho_{i \cdot k} = 1$, то ξ_i является строго линейной функцией от $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$.

Выборочным коэффициентом множественной корреляции $R_{i \cdot k}$ между величиной ξ_i и набором $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m$ называется величина

$$R_{i \cdot k} = \sqrt{1 - \frac{1}{\widehat{C}^{ii}}},$$

где $\widehat{C}^{ii}$ — диагональный элемент матрицы $\widehat{C}^{-1}$, обратной матрице выборочных коэффициентов корреляции.

Для проверки гипотезы H_0 о равенстве нулю коэффициента множественной корреляции $H_0: \rho_{i \cdot k} = 0$ при множестве альтернатив H_1: $\overset{2}{}_{i \cdot k} \neq 0$ вычисляется величина

$$F = \frac{n-m}{m-1} \cdot \frac{R^2_{i\cdot k}}{1-R^2_{i\cdot k}},$$

имеющая в условиях нулевой гипотезы F-распределение с $m-1$ и $n-m$ степенями свободы.

При уровне значимости α по таблицам F-распределения находят $F_{\alpha, m-1, n-m}$ — критическое значение с $m-1$ и $n-m$ степенями свободы. Тогда, если $F \geqslant F_{\alpha, m-1, n-m}$, то гипотеза H_0 отклоняется, в противном случае она принимается как подтвердившаяся.

Ранговый коэффициент множественной корреляции. Ранговый аналог множественного коэффициента корреляции определяется в обозначениях частной ранговой корреляции [20, с. 135] по формуле

$$\dot{\rho}_{k\cdot ij} = \sqrt{1-(1-\dot{\rho}^2_{kj})(1-\dot{\rho}^2_{ki\cdot j})}.$$

Следует сказать, что, как и для коэффициентов частной ранговой корреляции, методов проверки статистической значимости коэффициентов множественной корреляции пока не существует.

Конкордация. Это многомерная характеристика согласованности ранжированных рядов, параметрический аналог которой отсутствует. Ранжированные ряды получаются следующим образом. Берется столбец с номером i, $i=1, 2, ..., m$, матрицы $X=\{x_{ti}\}$, $t=1, 2, ..., n$, $i=1, 2, ..., m$ (см. раздел 3.3.1) выборочных значений геологических характеристик x_i, $i=1, 2, ..., m$. Элементы столбца упорядочиваются в порядке неубывания. Каждому элементу упорядоченного ряда ставится в соответствие натуральное число от 1 до n, причем первому элементу ряда ставится в соответствие 1, второму — 2 и т. д., и последнему — число n. Если упорядоченный ряд содержит последовательности равных значений, то каждому члену такой последовательности ставится в соответствие одно и то же ранговое число, равное среднему арифметическому из соответствующих ранговых чисел. В результате процесса ранжирования каждому номеру t исходного ряда x_{ti}, $t=1, 2, ..., n$, где i — фиксировано, $i=1, 2, ..., m$, будет соответствовать определенное ранговое число.

Определим для каждого t, $t=1, 2, ..., n$ сумму ранговых чисел

$$R_t = \sum_{i=1}^{m} R_i^t$$

и среднее значение суммы рангов для одного наблюдения

$$\bar{R} = m(n+1)/2.$$

Найдем S — сумму квадратов отклонений R_t и $\bar{R}$

$$S = \sum_{t=1}^{n} (R_t - \bar{R})^2.$$

Тогда коэффициентом конкордации, или коэффициентом согласованности, будет величина

$$W = \frac{12S}{m^2(n^3 - n)}.$$

Величина W может принимать значение от 0 до 1. Заметим, что коэффициент конкордации W и совокупность коэффициентов Спирмена ρ_s для каждой пары геологических характеристик связаны соотношением

$$\bar{\rho}_s = \frac{mW - 1}{m - 1},$$

где $\bar{\rho}_s$ — среднее значение коэффициента Спирмена для всех возможных пар геологических характеристик.

Если некоторые последовательности рангов содержат связи, то значение W будет заниженным и для нахождения коэффициента конкордации следует воспользоваться формулой

$$W = \frac{S}{m^2(n^3 - n)/12 - m\sum_i T_i},$$

где

$$T_i = \frac{1}{12}\sum_{k=1}^{L}(u_k^3 - u_k),$$

и L — число групп равных рангов в ряду R^i, а u_k — число равных ранговых чисел в группе с номером k.

Для проверки статистической значимости W при малых n ($n \leqslant 7$) следует воспользоваться специальными таблицами [20, с. 201—203].

При $n > 7$ вычисляют величину

$$V = \frac{S}{mn(n+1)/12}$$

(если связи рангов отсутствуют или количество связанных рангов невелико) или величину

$$V = \frac{S}{mn(n+1)/12 - \sum_i T_i(n-1)}$$

(если количество связанных рангов значительно велико), которые в условиях гипотезы $H_0: W = 0$ при множестве альтернатив $H_1: W \neq 0$ распределены как χ^2 с $n-1$ степенями свободы.

Поэтому при уровне значимости α из таблиц χ^2-распределения находят критическое значение $\chi^2_{\alpha, n-1}$, соответствующее $n-1$ степеням свободы. Тогда выполнение условия $V > \chi^2_{\alpha, n-1}$ ведет к отклонению проверяемой гипотезы, а нарушение условия — к принятию ее.

Если гипотеза о равенстве нулю коэффициента конкордации отклонена, то можно полагать, что в рассмотренных ранжированных рядах наблюдается определенная согласованность.

3.3.5. Выявление нелинейной связи двух геологических характеристик. Корреляционное отношение

Равенство нулю коэффициента корреляции $\rho(\xi, \eta)$ или коэффициентов регрессии $\beta_{1\xi}$ и $\beta_{1\eta}$ еще не доказывает, что между случайными величинами ξ и η отсутствует зависимость, которая может быть не только линейной, но и нелинейной. Характеристикой, указывающей на наличие как линейной, так и нелинейной корреляционной зависимости случайных величин ξ и η, являющихся моделями геологических характеристик, является корреляционное отношение.

Для произвольного фиксированного значения величины $\eta = Y$ распределение ξ называется ξ-сечением и является условным распределением ξ при заданном η. Среднее этого условного распределения называется регрессией ξ по η и обозначается как $\overline{\xi} = \overline{M}(\xi/\eta)$.

Различают ν_ξ — корреляционное отношение ξ по η

$$\nu_\xi^2 = \frac{D[M(\xi/\eta)]}{\sigma_\xi^2}$$

(здесь в числителе стоит дисперсия относительно $\overline{Y}$ регрессии ξ по η, а в знаменателе — дисперсия ξ) и ν_η — корреляционное отношение η по ξ

$$\nu_\eta^2 = \frac{D[M(\eta/\xi)]}{\sigma_\eta^2}.$$

Справедливы важные соотношения: $0 \leqslant \rho^2(\xi, \eta) \leqslant \nu_\xi^2 \leqslant 1$ и $0 \leqslant \rho^2(\xi, \eta) \leqslant \nu_\eta^2 \leqslant 1$.

Равенства $\rho^2(\xi, \eta) = \nu_\xi^2 \equiv 1$ $(\rho^2(\xi, \eta) = \nu_\eta^2 = 1)$ с необходимостью и достаточностью свидетельствуют о строгой линейной функциональной зависимости ξ от η (η от ξ).

Соотношения $\rho^2(\xi, \eta) < \nu_\xi^2 = 1$ $(\rho^2(\xi, \eta) < \nu_\eta^2 = 1)$ с необходимостью и достаточностью свидетельствуют о строгой нелинейной функциональной зависимости ξ от η (η от ξ).

Соотношения $\rho^2(\xi, \eta) = \nu_\xi^2 < 1$ $(\rho^2(\xi, \eta) = \nu_\eta^2 < 1)$ с необходимостью и достаточностью свидетельствуют о строгой линейности регрессии ξ по η (η по ξ) и об отсутствии функциональной зависимости.

Неравенства $\rho^2(\xi, \eta) < \nu_\xi^2 < 1$ $(\rho^2(\xi, \eta) < \nu_\eta^2 < 1)$ указывают на отсутствие функциональной зависимости и на существование некоторой нелинейной регрессии ξ по η (η по ξ), которая лучше, нежели любая линейная регрессия, осуществляет подгонку исходных данных.

Соотношения $\nu_\xi^2 = \rho^2(\xi, \eta) < \nu_\eta^2$ $(\nu_\eta^2 = \rho^2(\xi, \eta) < \nu_\xi^2)$ свидетельствуют о строгой линейности регрессии ξ по η (η по ξ) и нелинейности регрессии η по ξ (ξ по η).

Оценкой корреляционного отношения ν^2, полученной по выборочным данным $t=1, 2, \ldots, n$, является выборочное корреляционное отношение:

$$e_x = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{k} \sum_{t=1}^{n} (x_{ti} - \bar{x})^2} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i^2 - n\bar{x}^2}{\sum_{i=1}^{k} \sum_{t=1}^{n} x_{ti}^2 - n\bar{x}^2}.$$

где x_{ti}, $i=1, 2, \ldots, k$, $t=1, 2, \ldots, n_i$ — выборочные значения случайной величины ξ, классифицированные каким-либо образом в k групп; n_i — число наблюдений в i-й группе, n — общее число наблюдений; x_i — среднее по ξ для i-й группы; $\bar{x}$ — общее среднее.

Аналогичная формула может быть записана для e_y — выборочного корреляционного отношения y по x.

Справедливы соотношения $0 \leqslant r^2 \leqslant e_x^2 \leqslant 1$ и $0 \leqslant r^2 \leqslant e_y^2 \leqslant 1$.

Оценим значимости $e_x (e_y)$. При условии $\xi = 0$ статистика

$$F = \frac{e_x^2}{1 - e_x^2} \cdot \frac{n-k}{k-1}$$

распределена по закону Фишера с $k-1$ и $n-k$ степенями свободы.

Статистика F может быть использована для проверки гипотезы $H_0 : \nu = 0$ или, что то же самое, проверки условия, что кривая регрессии ξ по η представляет собой прямую, параллельную оси y при альтернативе H_1, что кривая регрессии может быть любой формы, отличной от указанной в формулировке гипотезы.

Для проверки H_0 вычисляют статистику F и, задавшись определенным уровнем значимости α, находят по таблицам F-распределения значение $F_{\alpha, k-1, n-k}$. Тогда условие $F \geqslant F_{\alpha, k-1, n-k}$ ведет к отклонению гипотезы H_0, в противном случае гипотеза H_0 принимается как подтвердившаяся.

Фактически данная гипотеза проверяет предположение, что кривая регрессии есть полином степени $k-1$ и все коэффициенты полинома, кроме константы, равны нулю.

Для проверки гипотезы H_0 о том, что кривая регрессии ξ по η является полиномом $(k-1)$-й степени и все коэффициенты полинома, кроме константы и коэффициента при линейном члене, равны нулю, используется статистика

$$F = \frac{(e_x^2 - r^2)(n-k)}{(1 - e_x^2)(k-2)},$$

распределенная по закону Фишера с k—2 и n—k степенями свободы. Условие $F \geqslant F_{\alpha, k-2, n-k}$, где $F_{\alpha, k-2, n-k}$ — критическое значение F-распределения, ведет к отклонению нулевой гипотезы. В противном случае гипотеза H_0 принимается как подтвердившаяся.

Ранговые аналоги корреляционного отношения в литературе не описаны, однако, как было показано при описании коэффициентов ранговой корреляции (см. разделы 3.3.2, 3.3.3), коэффициенты ранговой корреляции служат мерой не только линейной, но и многих нелинейных зависимостей.

3.4. РЕКОМЕНДУЕМЫЕ РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ РЕГРЕССИОННЫЕ МОДЕЛИ

В настоящем разделе рассматриваются два вида регрессии: регрессия как условное математическое ожидание и регрессия как безусловное математическое ожидание.

В рамках модели регрессии как условного математического ожидания описывается линейная (по регрессорам) регрессия в предположении многомерного нормального распределения случайных величин — моделей геологических характеристик (см. раздел 3.4.1). Модель применима, когда все переменные модели — случайные величины.

Описана также линейная (по параметрам) модель регрессии как безусловного математического ожидания (см. раздел 3.4.2). Использование этой модели предполагает детерминированность регрессоров и нормальность распределения ошибок. Указано, к каким ошибкам приводит недобор и перебор факторов в регрессии. При недоборе факторов (регрессоров) оценки коэффициентов регрессии приобретают смещение и становятся несостоятельными, а при переборе факторов те же оценки сохраняют свойства несмещенности и состоятельности, хотя точность оценок понижается. Поэтому недобор считается более серьезной ошибкой, нежели перебор, и не следует стремиться к слишком малым значениям ошибки первого рода.

Корректное использование моделей регрессии предполагает выполнение ряда теоретических предпосылок. В практике геологических исследований эти предположения часто нарушены, и в этом случае можно попытаться устранить эти нарушения путем изменения модели определенным образом, например, преобразуя регрессоры и зависимую переменную [16, 41].

Другим способом проведения исследований в условиях нарушения основных предположений относительно постулируемой модели является замена критерия, согласно которому определяются оценки коэффициентов регрессии. Построенные таким образом регрессии называют устойчивыми, робастными (robust) [15, с. 175—182].

Еще одним способом построения уравнения регрессии в условиях отклонения от нормального распределения случайных величин — моделей геологических характеристик или случайных оши-

бок является процедура построения ранговых моделей регрессии. Ранговые модели регрессии часто весьма успешно применяются там, где нарушаются основные предположения параметрических моделей. В то же время в условиях применимости параметрических моделей ранговые модели часто дают результаты, близкие параметрическим [43]. Это свойство ранговых моделей позволяет рекомендовать их наряду с параметрическими моделями для решения задач геологического прогнозирования.

Следует отметить, однако, что нахождение оценок коэффициентов множественного уравнения регрессии, построенного на рангах или метках, представляет собой в вычислительном плане сложную задачу минимизации функции многих переменных, для решения которой не предложено специальных методов минимизации функций, более экономичных, нежели общие методы минимизации, например метод Ньютона, градиентный и т. п. [15, с. 245—256].

Для парного уравнения регрессии (случай одного регрессора) в разделе 3.4.3.3. описан алгоритм Адичи нахождения оценок коэффициентов регрессии, имеющий высокую сходимость и не слишком сложный в вычислении.

3.4.1. Линейная регрессия в условиях нормальной модели

Как и в разделе 3.2, рассмотрим m-мерную случайную величину

$$\Xi = \{\xi_1, \ldots, \xi_m\}$$

с известным законом распределения. Обозначим через Ξ^i компоненты m—1-мерной случайной величины, получаемой из Ξ удалением компоненты ξ_i, т. е. $\Xi^i = \{\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m\}$. Регрессией случайной величины ξ_i на Ξ называется условное математическое ожидание $\mu(\xi_i|\Xi^i)$ как функция Ξ^i, т. е. $\mu(\xi_i|\Xi^i) = f(\Xi^i) = f(\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_m)$. Линейной называется такая регрессия, для которой функциональное соотношение $f(\Xi^i)$ линейно по $\xi_1, \ldots, \xi_m$. Условная дисперсия $\sigma^2(\xi_i|\Xi^i) = M[(\xi_i - \mu(\xi_i|\Xi^i))^2|\Xi^i]$ как функция Ξ^i называется остаточной дисперсией ξ_i на Ξ^i. Здесь M — оператор математического ожидания. Остаточная дисперсия является важнейшей характеристикой точности приближения ξ_i регрессией $\mu(\xi_i|\Xi^i)$.

Если регрессия существует и линейна, то коэффициенты регрессии можно определять из условия минимизации остаточной дисперсии. Найденная из этого условия по методу наименьших квадратов линейная регрессия называется средней квадратической, а значение остаточной дисперсии для такой регрессии называется средней квадратической остаточной дисперсией. Для коэффициентов уравнения средней квадратической регрессии $\xi_i = \beta_0 + \beta_1\xi_1 + \ldots + \beta_{i-1}\xi_{i-1} + \beta_{i+1}\xi_{i+1} + \ldots + \beta_m\xi_m$ справедливы соотношения $\beta_0 = \mu_i - \beta_1\mu_1 - \ldots - \beta_{i-1}\mu_{i-1} - \beta_{i+1}\mu_{i+1} - \ldots - \beta_m\mu_m$, $\beta_j = -\sigma^{ij}/\sigma^{ii}$, $j \neq i$ ($\sigma^{ii} \neq 0$), где σ^{ij} — элементы матрицы $\Sigma^{-1}\{\sigma^{ij}\}$, $i, j = 1, 2, \ldots, m$, обратной ковариационной матрице $\Sigma = \{\sigma_{ij}\}$, $i, j = 1, 2, \ldots, m$ компо-

нент случайной величины Ξ. Средняя квадратическая остаточная дисперсия $\sigma^2_{i\cdot 12\ \ldots\ i-1,\ i+1\ \ldots\ m}$ в этом случае равна $1/\sigma^{ii}$.

В случае m-мерного нормального распределения случайной величины Ξ *условное математическое ожидание и дисперсия* ξ_i при условии Ξ^i совпадают соответственно со средней квадратической регрессией и средней квадратической остаточной дисперсией ξ_i на Ξ^i. Таким образом, при весьма сильном предположении о нормальности распределения Ξ задача нахождения регрессии полностью решается, а именно регрессия существует, она линейна относительно компонент Ξ, ее коэффициенты и средняя ошибка прогноза определяются по указанным формулам.

Если исходные геологические данные представляют собой выборку из совокупности с m-мерным нормальным распределением, то, подставив в формулу для коэффициентов регрессии соответствующие выборочные аналоги генеральных коэффициентов, можно проверить гипотезы о статистической значимости рассчитанных коэффициентов и всего уравнения регрессии в целом.

В частности, проверка гипотезы относительно равенства нулю угла наклона прямой регрессии с осью координат (т. е. обращения в нуль коэффициента регрессии при линейном члене для двухмерной случайной величины) эквивалентна проверке гипотезы об отсутствии парной корреляции между этими величинами. Проверка гипотезы относительно равенства нулю коэффициента множественной регрессии равносильна проверке гипотезы об обращении в нуль соответствующего частного коэффициента корреляции, проверка гипотезы о значимости всего уравнения регрессии эквивалентна проверке значимости множественного коэффициента корреляции. Проверка таких гипотез относительно соответствующих коэффициентов корреляции приводится в разделе 3.3.

3.4.2. Модель линейной регрессии

Данное выше определение регрессии как условного математического ожидания требует точного знания исходного распределения. В практических исследованиях это распределение обычно неизвестно. Поэтому, прилагая аппарат регрессионного анализа для прогнозирования геологических характеристик, существенно ослабляют требования к определению регрессии, так что под линейной регрессией понимают упрощенную (но все же достаточно реалистическую) линейную модель $y=\beta_0=\beta_1 x_1+\ldots+\beta_m x_m+\varepsilon$, которую удобно записать в матричном виде $y=X\beta+\varepsilon$, где $y=(y_1, \ldots, y_n)'$ — вектор случайных наблюдений зависимой переменной (отклика) y в n точках; $X=\{x_{ti}\}$, $t=1, 2, \ldots, n$, $i=0, 1, 2, \ldots, m$ — матрица известных коэффициентов (неслучайных величин), причем x_{ti} — значение независимой геологической характеристики (регрессора) с номером i в точке с номером t, $x_{t0}=1$, $t=1, 2, \ldots, n$; $\beta=(\beta_0, \beta_1, \ldots, \beta_m)'$ — вектор коэффициентов регрессии; $\varepsilon=(\varepsilon_1, \ldots, \varepsilon_n)'$ — вектор случайных ошибок.

В линейной модели относительно ошибок ε делаются предположения, что ошибки некоррелированны, имеют нулевые средние и постоянные дисперсии, т. е. $M(\varepsilon)=0$, $V(\varepsilon)=\sigma^2 I$, где $M(\varepsilon)$ — вектор математических ожиданий ошибок, $V(\varepsilon)$ — ковариационная матрица ошибок, I — единичная матрица.

Следует пояснить, что при определении регрессии как условного математического ожидания имеется в виду регрессия, линейная по регрессорам, $\eta=\beta_0+\beta_1\xi_1+\ldots+\beta_m\xi_m$.

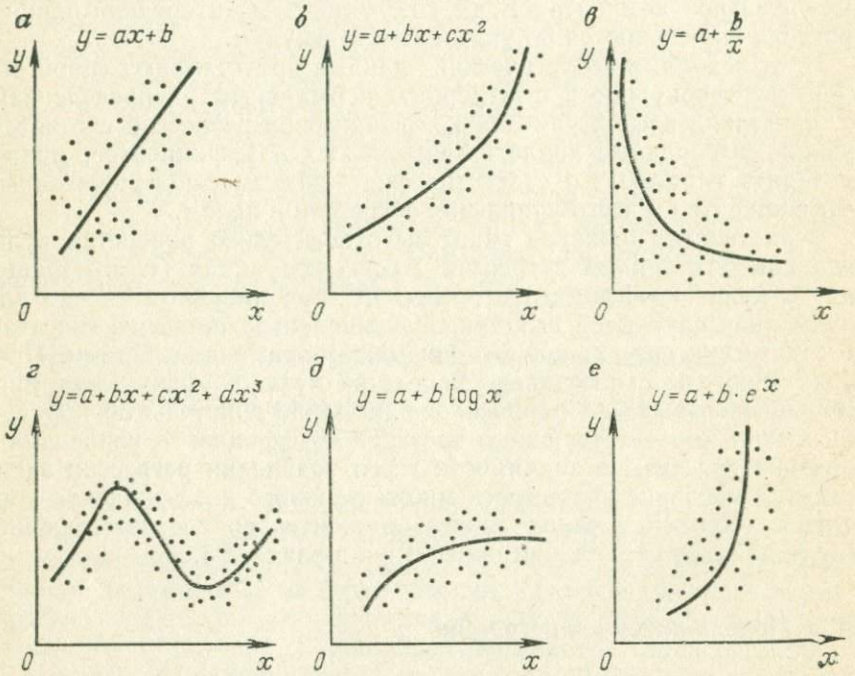


Рис. 16. Простейшие виды регрессионных зависимостей:
а — линейной; *б* — квадратической параболической, *в* — гиперболической, *г* — кубической параболической, *д* — линейной логарифмической, *е* — экспоненциальной

С введением упрощенного определения регрессии как безусловного математического ожидания линейность регрессии удобно рассматривать по параметрам β_i. Тогда линейными будут следующие уравнения регрессии: полиномиальные — $y=\beta_0+\beta_1 x+\ldots+\beta_m x^m+\varepsilon$ и $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1^2+\beta_4 x_2^2+\beta_5 x_1 x_2+\varepsilon$, гиперболическая — $y=\beta_0+\beta_1/x+\varepsilon$, логарифмическая — $y=\beta_0+\beta_1 \ln x+\varepsilon$, экспоненциальная — $y=\beta_0+\beta_1 e^x+\varepsilon$, тригонометрическая — $y=\beta_0+\beta_1 \sin x+\beta_2 \cos x+\varepsilon$ и др. (рис. 16).

Суммой квадратов ошибок называется скалярная величина

$$\Sigma=\varepsilon'\varepsilon=(y-X\beta)'(y-X\beta).$$

Существуют различные методы оценки параметров β, σ^2 линейной модели. Наиболее известны методы наименьших квадратов и максимального правдоподобия.

Метод наименьших квадратов заключается в минимизации скалярной суммы квадратов ошибок Σ. Значение вектора β, обращающее Σ в минимум, находится из решения системы нормальных уравнений $\partial\Sigma/\partial\beta = 2X'(y - X\beta) = 0$ и равно $b = \hat{\beta} = (X'X)^{-1}X'y$ в предположении, что матрица $X'X$ *не вырождена.*

Предсказанные значения зависимой переменной y с помощью линейной модели с коэффициентами $\hat{\beta}$ обозначаются $\hat{y}$ и равны $\hat{y} = X\hat{\beta}$. Остатками e называют вектор $e = \hat{\varepsilon} = y - \hat{y} = \{e_t\}$, $t = 1, 2, \ldots, n$, где $\hat{y} = X\hat{\beta}$ есть предсказанные значения зависимой переменной с помощью линейной модели с коэффициентами $\hat{\beta}$.

Суммой квадратов остатков называют величину

$$\mathrm{SSE} = \sum_{t=1}^{n} e_t^2 = y'y - \hat{\beta}'X'y.$$

Несмещенной оценкой s^2 дисперсии σ^2 является

$$s^2 = \hat{\sigma}^2 = \mathrm{SSE}/(n - m - 1).$$

Общей суммой квадратов называют величину

$$\mathrm{SST} = \sum_{t=1}^{n} y_t^2 - n\bar{y}^2.$$

Точность регрессионного анализа оценивается при дополнительном предположении, что ошибки ε подчиняются нормальному распределению с нулевыми средними и равными дисперсиями σ^2, что символически записывается как $\varepsilon \sim N(0, \sigma^2 I)$, где I — единичная матрица. Следствиями такого допущения являются следующие утверждения: $y \sim N(X\beta, \sigma^2 I)$, $\hat{\beta} \sim N(\beta, \sigma^2 (X', X)^{-1})$.

Проверка общей линейной гипотезы $H: A\beta = 0$, где A — известная матрица размерности $k \times (m+1)$ ранга $k \leqslant m+1$, эквивалентна проверке предположения, что коэффициенты регрессии удовлетворяют системе линейных уравнений с матрицей A. Выразим k параметров β_i через остальные $m+1-k$. Получим новую линейную модель, для которой число оцениваемых коэффициентов регрессии уменьшилось и равно $m+1-k$. Найдем суммы квадратов остатков для исходной и новой моделей — SSE и $\mathrm{SSE}_н$. Тогда если гипотеза H верна, то статистика

$$F = \frac{(\mathrm{SSE}_н - \mathrm{SSE})/k}{\mathrm{SSE}/(n - m - 1)}$$

имеет F-распределение с k и $n-m-1$ степенями свободы. Так что если $F > F^{\alpha}_{k,\, n-m-1}$, где $F^{\alpha}_{k,\, n-m-1}$ — верхняя 100α-процентная точка распределения $F_{k,\, n-m-1}$, то гипотеза H отклоняется с уровнем

значимости α, в противном случае H принимается как подтвердившаяся.

Если гипотеза H отклонена, то следует выяснить, почему это произошло. Для этого можно проверить выполнение каждого ограничения, наложенного на коэффициенты регрессии. Каждое такое (линейное) ограничение задается одной строкой матрицы A и принимает вид $a_i\beta = c_i$, где a_i — вектор-строка с номером i матрицы A, c_i — константа. Частные гипотезы H_i могут быть заданы в виде $H_i : a_i\beta = c_i$.

Для проверки H_i вычисляется статистика

$$t_i = (a_i\hat{\beta} - c_i)/(s\sqrt{a_i X'X a_i}),$$

распределенная в условиях гипотезы по закону Стьюдента с $n-m-1$ степенями свободы t_{n-m-1}. Так что если $|t_i| \geqslant t^{\alpha/2}_{n-m-1}$, где $t^{\alpha/2}_{n-m-1}$ — верхняя $100(\alpha/2)$-процентная точка распределения t_{n-m-1}, то гипотеза H_i отклоняется с уровнем значимости α. В противном случае H_i принимается как подтвердившаяся.

Чрезвычайно важными частными случаями общей линейной гипотезы являются следующие:

1) проверка значимости отдельных коэффициентов регрессии β_i сводится к проверке гипотезы $H_i : \beta_i = 0$;

2) проверка значимости всего уравнения регрессии сводится к проверке гипотезы $H : \beta_1 = \beta_2 = \ldots = \beta_m = 0$.

Для проверки гипотезы H_i вычисляется статистика

$$t_i = b_i \sqrt{s^2 (X'X)^{-1}_{ii}},$$

где $(X'X)_{ii}$ — диагональный элемент с номером i матрицы, обратной X^1X; подчиняется статистика t_i в условиях гипотезы распределению Стьюдента t_{n-m-1}, так что если $|t_i| \geqslant t^{\alpha/2}_{n-m-1}$, где $t^{\alpha/2}_{n-m-1}$ — верхняя $100(\alpha/2)$-процентная точка распределения Стьюдента t_{n-m-1}, то гипотеза H_i отклоняется с уровнем значимости α. В этом случае говорят, что коэффициент β_i значим и переменной x_i пренебрегать, вообще говоря, нельзя.

Для проверки гипотезы H вычисляется статистика

$$F = \frac{(\mathrm{SST} - \mathrm{SSE})/m}{\mathrm{SSE}/(n - m - 1)},$$

которая в условиях гипотезы имеет F-распределение с m и $n-m-1$ степенями свободы, так что если $F > F^{\alpha}_{n-m-1,\,m}$, где $F^{\alpha}_{n-m-1,\,m}$ — верхняя 100α-процентная точка распределения $F_{n-m-1,\,m}$, то гипотеза H отклоняется с уровнем значимости α. В противном случае гипотеза считается подтвердившейся.

Отклонение гипотезы H свидетельствует о значимости всего уравнения регрессии и утверждает, что, вообще говоря, переменными x_i, $i = 1, 2, \ldots, m$ пренебрегать нельзя. Однако значимость уравнения регрессии еще не свидетельствует об его адекватности и пригодности для целей прогнозирования. Действительно, если

размах величин, полученных с помощью уравнения регрессии, близок или лишь немного превосходит величину стандартной ошибки, то такое уравнение будет предсказывать только ошибки и по этой причине непригодно для прогноза. Высказано мнение [16, с. 74], что для пригодности уравнения регрессии для прогнозирования необходимо четырехкратное превышение F по отношению к $F^{\alpha}_{n-m-1,\ m}$.

Проверка гипотезы $H_i: \beta_i=0$ тесно связана с проблемой оптимального выбора регрессоров. При этом можно совершить два типа ошибок:

1) истинное значение параметра $\beta_i=0$, а мы делаем вывод, что $\beta_i\neq 0$ и включаем x_i в модель регрессии. Такую ошибку называют ошибкой первого рода. Вероятность такой ошибки задается заранее и равняется принятому уровню значимости α. Ошибка такого рода ведет к перебору признаков в регрессии;

2) истинное значение параметра $\beta_i\neq 0$, а мы делаем вывод, что $\beta_i=0$ и исключаем x_i из уравнения регрессии. Такую ошибку называют ошибкой второго рода. Ошибка такого рода ведет к недобору признаков в регрессии.

При переборе факторов в регрессии МНК-оценки параметров регрессии несмещены и состоятельны, однако имеют меньшую точность [15, с. 92—98]. Потери точности оценок можно избежать, если использовать процесс ортогонализации. При недоборе факторов картина качественно меняется, МНК-оценки приобретают смещения и становятся несостоятельны. По этой причине недобор считается более серьезной ошибкой, нежели перебор, и не следует стремиться к слишком малым значениям α.

Большой интерес для практического применения линейной модели для целей прогнозирования геологических характеристик представляет проверка гипотез относительно истинных значений математического ожидания отклика в произвольной точке t и относительно произвольного теоретического значения отклика. Также значительный интерес при прогнозировании геологических характеристик могут представлять так называемые доверительные интервалы, которые с заданной доверительной вероятностью накрывают истинное значение оцениваемого параметра. Оценивание параметра с помощью доверительного интервала называют интервальным оцениванием в отличие от рассмотренного ранее точечного оценивания. Естественным обобщением доверительного интервала являются доверительные области, с заданной доверительной вероятностью накрывающие одновременно несколько теоретических параметров или некоторую функцию от них. Соответствующие формулы приводятся в [2, 15, 16, 19, 28, 41].

3.4.3. Ранговые модели регрессии

В разделе 3.4.3.1 описана модель многомерной линейной регрессии, построенной на рангах или метках, как безусловного математического ожидания. В зависимости от предполагаемого (допускаемого) вида распределения зависимой случайной величины —

модели прогнозируемой геологической характеристики (или распределения случайных ошибок) выбирается соответствующий вид меток. Коэффициенты уравнения регрессии определяются из условия обращения в минимум определенной функции многих переменных, зависящей от значений регрессоров и рангов или меток остатков. Численное решение задачи оценивания коэффициентов уравнения проводится общими методами минимизации функции многих переменных, например градиентным методом Ньютона и т. п.

В работах И. Адичи и других исследователей [47, 48] не предложено специальных алгоритмов минимизации, более экономных, нежели общие методы.

Описывается процедура [3.4.3.2.] проверки гипотезы об одновременном обращении в нуль всех коэффициентов регрессии. Процедура не требует предварительного нахождения численных значений коэффициентов регрессии, что позволяет сначала проверить эту гипотезу и в случае ее принятия отказаться от вычисления коэффициентов регрессии. Полезную информацию можно получить, изменяя уровень значимости и найдя то предельное значение уровня значимости, начиная с которого гипотеза отклоняется.

В случае отклонения гипотезы об одновременном обращении в нуль всех коэффициентов регрессии определяются коэффициенты *линейного парного рангового* уравнения регрессии для одного регрессора с помощью итеративного алгоритма Адичи или оцениваются коэффициенты множественного рангового уравнения регрессии методом, описанным в разделе [3.4.3.1.].

3.4.3.1. Модель многомерной регрессии, построенной на рангах

Рассматривается линейная по параметрам модель [48, с. 1328—1338] $\mu(y)=\beta_0+\beta_1 x_1+\ldots+\beta_m x_m$, где $x_i=\{x_{ti}\}$, $t=1, 2, \ldots, n$, $i=1, 2, \ldots, m$ — вектор фиксированных (неслучайных) значений *регрессора*; x_i — *модель геологической характеристики*; m — количество регрессоров; n — число измерений каждой характеристики; $y=\{y_1, \ldots, y_n\}$ — вектор независимых случайных величин (y — модель зависимой прогнозируемой геологической характеристики) с непрерывными функциями распределения

$$F_t(z)=F(z-\beta_0-\beta_1 x_{t1}-\ldots-\beta_m x_{tm}),$$

$$t=1, \ldots n;\ \beta_0, \beta_1, \ldots, \beta_m -$$

коэффициенты регрессии, которые необходимо оценить.

Считая коэффициенты $\beta_1, \ldots, \beta_m$ *известными* (например, приравняв их произвольным числам в качестве начального приближения), можно вычислить ряд остатков

$$\{y_t-\beta_1 x_{t1}-\ldots-\beta_m x_{tm}\},\quad t=1, 2, \ldots, n$$

и ряд рангов $\{R_t\}$, $t=1, 2, \ldots n$ этих остатков. Ряд рангов определяется путем расположения ряда остатков в порядке неубывания

и определения номера в упорядоченном ряде остатков каждого элемента неупорядоченного ряда.

Затем находится вектор $S_n = \{S_{ni}\}$, $i = 1, 2, ..., m$,

$$S_{ni} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (x_{ti} - \bar{x}_t) R_t,$$

где $\bar{x}_i$ — среднее арифметическое x_i.

Из компонент вектора S_n составляется выражение

$$S = \sum_{i=1}^{m} | S_{ni} | ,$$

которое является нелинейной функцией параметров $\beta_1, ..., \beta_m$.

Те значения параметров $\beta_1, ..., \beta_m$, на которых достигается минимум S, дают искомые оценки коэффициентов линейной ранговой множественной регрессии. Минимизация S может быть осуществлена одним из общих методов минимизации функции многих переменных, например методом Ньютона, градиентным и т. п. В работах И. Адичи и других [47, 48] специальной процедуры минимизации S не приводится.

Найденные оценки $\beta_1, ..., \beta_m$ теперь могут быть использованы в качестве начальных приближений для коэффициентов регрессии, и аналогично будет найдено следующее приближение оценок $\beta_1, ..., \beta_m$.

Процесс получения итераций может быть продолжен до тех пор, пока приращения оценок $\beta_1, ..., \beta_m$ не станут меньше наперед заданных положительных чисел.

Оценка $\widehat{\beta_0}$ свободного члена β_0 уравнения регрессии находится по формуле $\widehat{\beta_0} = (\beta_0^{\bullet} + \beta_0^{\bullet\bullet})/2$ [48, с. 1337], где

$$\beta_0^{\bullet} = \sup \{b,\ S_n^+ > 0\}, \quad \beta_0^{\bullet\bullet} = \inf \{b,\ S_n^+ < 0\},$$

$$S_n^+ = \sum_{t=1}^{n} \operatorname{sgn} (y_t - b - \widehat{\beta}_1 x_{t1} - \ . \ . \ . - \widehat{\beta}_m x_{tm}) R_t,$$

$\{R_t\}$, $t = 1, 2, ..., n$ — вектор рангов величины $\{y_t - b - \widehat{\beta}_1 x_{t1} - - ... - \widehat{\beta}_m x_{tm}\}$, $t = 1, 2, ..., n$; $\widehat{\beta}_1, ..., \widehat{\beta}_m$ — найденные выше оценки коэффициентов регрессии.

Функция $\operatorname{sgn}(z)$ определяется по формуле

$$\operatorname{sgn}(z) = \begin{cases} -1, & \text{если } z < 0 \\ 0, & \text{если } z = 0 \\ +1, & \text{если } z > 0. \end{cases}$$

Операция взятия верхней грани $\sup\{b, S_n^+ > 0\}$ заключается в нахождении такого числа b, для которого величина $S_n^+ > 0$ принимает наименьшее значение. Аналогично операция взятия ниж-

ней грани $\inf\{b, S_n^+ < 0\}$ состоит в определении такого числа b, для которого величина $S_n^+ < 0$ принимает наибольшее значение.

В работе [48] приводится обобщение рассмотренной модели *линейной* ранговой регрессии путем замены рангов остатков на выбранные каким-либо образом метки. Выбор системы меток производится в соответствии с предполагаемым видом распределения случайных ошибок. Системой меток $a_n(t)$, $t=1, 2, ..., n$, порожденной функцией $\varphi(u)$, $0<u<1$, называется либо вектор $\{M\varphi(V^{(t)})\}$, $t=1, 2, ..., n$, либо вектор $\{\varphi(t/n+1)\}$, $t=1, 2, ..., n$ где V^t — порядковая с номером t статистика в выборке объема n из равномерного распределения на (0,1). Обычным выбором $\varphi(u)$ является

$$\varphi(u) = \varphi(u, f_0) = -f_0'(F^{-1}(u))/f_0(F^{-1}(u)),$$

где f_0 — предполагаемая (допускаемая) плотность, обладающая ограниченной информацией Фишера ($\varphi(u, f_0)$, тогда и только тогда не убывает, когда f_0 строго унимодальна). Знак $M\varphi$ — математическое ожидание φ, f_0' — производная от функции f_0, F^{-1} — функция, обратная F.

Таким образом, если есть основание предполагать, что случайные ошибки имеют плотность распределения f_0, где f_0 отлична от плотности нормального распределения (но обладает ограниченной информацией Фишера), то, определив $\varphi(u)$, можно найти ряд меток $a_n(t)$, $t=1, 2, ..., n$ и, заменив в выражениях для S и β_0 ранги R_t на метки $a_n(t)$, определить оценки коэффициентов линейной множественной ранговой регрессии.

3.4.3.2. Проверка гипотез о ранговых моделях регрессии

В работе М. Пури и П. Сена [49, с. 462] приводится процедура проверки гипотезы о равенстве нулю коэффициентов регрессии, осуществляемая без вычисления самих коэффициентов регрессии, нахождение которых является сложной вычислительной задачей. Проверяемая гипотеза H_0 формулируется так: $H_0: \beta_1=\beta_2=...=\beta_m=0$ при альтернативе H_1, что хотя бы для одного i, $i=1, 2, ..., m$; $\beta_i \neq 0$.

Для проверки H_0 вычисляется статистика

$$L_n = \frac{n-1}{n} A_n^2 S_n' \tilde{C}_n S_n,$$

где вектор-столбец S_n определяется так же, как в разделе 3.4.3.1, $\tilde{C}_n$ — матрица, обобщенная обратная к матрице сопряженности $C_n = \{C_n(i, j)\}$, $i, j=1, 2, ..., m$ $C_n(ij) = \sum_{t=1}^{n} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)$; $\bar{x}_i \bar{x}_j$ — средние арифметические x_i, x_j, A_n^2 — скалярная величина, равная $A_n^2 = \frac{1}{n} \sum_{t=1}^{n} (R_t - \bar{R})^2$, причем R_t означает то же, что в разделе 3.4.3.1, $\bar{R}$ — среднее арифметическое ранговых чисел. Стати-

стика L_n в условиях H_0 распределена по закону χ^2 с m-степенями свободы.

Проверка H_0 на любом уровне значимости α проводится путем сравнения L_n и критического значения $\varkappa^2_{\alpha,m}$, взятого из таблиц.

В той же работе [49, с. 462] предлагается модификация статистики L_n путем замены рангов остатков $\{R_t\}$, $t=1, 2, \ldots n$ на систему меток $a_n(t)$, $t=1, 2, \ldots, n$. Метки полезно использовать, когда имеется дополнительная информация о предполагаемом (допускаемом) виде распределения ошибок линейной модели (подробнее см. в разделе 3.4.3.1).

В указанной работе на с. 463 проверяется также гипотеза о равенстве нулю всех коэффициентов регрессии (включая свободный член).

3.4.3.3. Итеративный алгоритм Адичи определения коэффициентов уравнения парной линейной регрессии, построенной на рангах

В случае одного регрессора $x=\{x_1, \ldots, x_n\}$ и одной зависимой переменной $y=\{y_1, \ldots, y_n\}$ уравнение линейной парной регрессии записывается в виде $\mu(y)=\beta_0+\beta_1 x$. Сначала определим оценку $\widehat{\beta}_1$ [47, с. 894—897].

Начальный шаг. Положим в качестве начального приближения $\widehat{\beta}_1^{(0)}=0$. Примем $\Delta=1$, $\Delta_0=0{,}1$.

Шаг итерации с номером $k: k=1,2, \ldots$. Определяем ряд остатков $e^{(k)}=\{y_t-\widehat{\beta}_1^{(k-1)}x_t\}$, $t=1, 2, \ldots, n$ и ряд рангов $R_t^{(k)}$ элементов последнего ряда. Вычисляем статистику $T^{(k)}=\dfrac{1}{n(n+1)}\sum\limits_{t=1}^{n}(x_t-\bar{x})R_t^{(k)}$ и сравниваем ее с нулем. Если $T^{(k)}=0$, то процесс приближения $\widehat{\beta}_1$ заканчиваем, полагаем $\widehat{\beta}_1=\widehat{\beta}_1^{(k)}$ и переходим к оцениванию свободного члена уравнения регрессии. В противном случае находим

$$\delta^{(k)}=\begin{cases}-1, & \text{если } T^{(k)}<0\\ +1, & \text{если } T^{(k)}>0\end{cases}$$

и, начиная с $k=2$, сравниваем $\delta^{(k)}$ и $\delta^{(k-1)}$. Если они не равны, то уменьшаем Δ вдвое и сравниваем полученное значение с Δ_0. Если $\Delta\leqslant\Delta_0$, то итерационный процесс завершаем. Для этого вычисляем $\widehat{\beta}_1^{(k)}=\widehat{\beta}_1^{(k-1)}+\delta^{(k)}\Delta$ принимаем $\widehat{\beta}_1=(\widehat{\beta}_1^{(k)}+\widehat{\beta}_1^{(k-1)})/2$ и переходим к вычислению $\widehat{\beta}_0$. Если $\Delta>\Delta_0$ или если $\delta^{(k)}=\delta^{(k-1)}$, то полагаем $\widehat{\beta}_1^{(k)}=\widehat{\beta}_1^{(k-1)}+\delta^{(k)}\Delta$. На этом шаг с номером k заканчиваем и переходим к шагу $k+1$.

Определение свободного члена регрессии $\widehat{\beta}_0$ проводится после нахождения оценки $\widehat{\beta}_1$ по формуле [47, с. 896]

$$\widehat{\beta}_0=\operatorname*{Me}_{t\leqslant t'}\frac{1}{2}[y_t+y_{t'}-\widehat{\beta}_1(x_t-x_{t'})],$$

где $\mathrm{Me}\{z_t\}$ — медиана числового ряда $\{z_t\}$, $t=1, 2, \ldots, n$.

Заметим, что в работе Д. Адичи [47] не приводится конкретный алгоритм для нахождения $\widehat{\beta}_1$. Там лишь рекомендовано определять $\widehat{\beta}_1$ из условия обращения статистики T в нуль. Нами введены в алгоритм параметры δ, Δ и Δ_0, позволяющие построить эффективную процедуру оценивания $\widehat{\beta}_1$. Параметр δ определяет знак приращения оценки $\widehat{\beta}_1$, параметр Δ — величину приращения, а параметр Δ_0 — абсолютную точность оценивания $\widehat{\beta}_1$. Предусмотрена возможность автоматического изменения величины Δ в зависимости от скорости сходимости итерационного процесса.

Алгоритм можно модифицировать очевидным образом, заменив критерий окончания процесса по величине абсолютной погрешности на окончание по величине относительной погрешности коэффициента регрессии.

Проиллюстрируем процедуру определения $\widehat{\beta}_1$ на геологическом примере. Имеем две характеристики: содержание x и запасы y металла в условных единицах в некотором рудном теле для четырех наблюдений.

$$x = 1 \quad 2 \quad 5 \quad 8,$$
$$y = 8 \quad 10 \quad 15 \quad 20.$$

Предполагая, что запасы металла и его содержание находятся в линейной зависимости $\mu(y)=\widehat{\beta}_0+\widehat{\beta}_1 x$, определим угол пересечения прямой регрессии с осью x, т. е. оценим коэффициент регрессии $\widehat{\beta}_1$.

В качестве начального приближения положим $\widehat{\beta}_1^{(0)}=0$ и примем $\Delta=1$, $\Delta_0=0{,}1$. Вычислим: $\bar{x}=4$, $\{x-\bar{x}\}=\{-3, -2, 1, 4\}$.

Первый шаг. Определим ряд $e^{(1)}=\{y_t-\widehat{\beta}_1^{(0)} x_t\}=\{8, 10, 15, 20\}$ и ряд рангов $R_t^{(1)}=\{1, 2, 3, 4\}$. Вычислим статистику $T^{(1)}=(-3\cdot 1-2\cdot 2+1\cdot 3+4\cdot 4)/4\cdot 5=12/20$. Поскольку $T^{(1)}>0$, то полагаем $\delta^{(1)}=+1$ и находим $\widehat{\beta}_1^{(1)}=0+1\cdot 1=1$.

Второй шаг. Аналогичным образом находим: $e^{(2)}=\{7, 8, 10, 12\}$, $R_t^{(2)}=\{1, 2, 3, 4\}$, $T^{(2)}=12/20>0$, $\delta^{(2)}=+1$, $\widehat{\beta}_1^{(2)}=1+1\cdot 1=2$.

Третий шаг. $e^{(3)}=\{6, 6, 5, 4\}$, $R_t^{(3)}=\{3, 4, 2, 1\}$.

Заметим, что осреднять ранги, соответствующие равным значениям компонент вектора e, недопустимо, поскольку от этого нарушается монотонность функции T по оцениваемому параметру. Наличие такой монотонности — непременное условие хорошей сходимости предлагаемого алгоритма.

$$T^{(3)}=-11/20<0, \quad \delta^{(3)}=-1, \quad \delta^{(3)}\neq\delta^{(2)}, \quad \Delta=0{,}5;$$

$$\widehat{\beta}_1^{(3)}=2+(-1)\cdot 0{,}5=1{,}5.$$

Четвертый шаг. $e^{(4)}=\{6{,}5; 7; 7{,}5; 8\}$, $R_t^{(4)}=\{1, 2, 3, 4\}$, $T^{(4)}=12/20>0$, $\delta^{(4)}=+1$, $\Delta=0{,}25$, $\widehat{\beta}_1^{(4)}=1{,}5+1\cdot 0{,}25=1{,}75$.

Пятый шаг. $e^{(5)}=\{6{,}25;\ 6{,}5;\ 6{,}25;\ 6\}$, $R_t^{(5)}=\{2,\ 4,\ 3,\ 1\}$, $T^5=-7/20<0$, $\delta^{(5)}=-1$, $\Delta=0{,}125$, $\widehat{\beta}_1^{(5)}=1{,}75-1\cdot 0{,}125=1{,}625$.

Шестой шаг. $e^{(6)}=\{6{,}375;\ 6{,}750;\ 6{,}875;\ 7\}$, $R_t^{(6)}=\{1,\ 2,\ 3,\ 4\}$, $T^{(6)}=12/20>0$, $\delta^{(6)}=+1$, $\Delta=0{,}0625$. Поскольку $\Delta\leqslant\Delta_0=0{,}1$, то итерационный процесс завершается; вычисляем $\widehat{\beta}_1^{(6)}=1{,}625+0{,}0625=1{,}6875$ и принимаем $\widehat{\beta}_1=(\widehat{\beta}_1^{(6)}+\widehat{\beta}_1^{(5)})/2=(1{,}6875+1{,}625)/2=1{,}65625$. Для сравнения вычислим выборочный параметрический коэффициент

$$\widehat{\beta}_1^{\text{парам}}=\left[\sum_{t=1}^{n}(y_t-\bar{y})(x_t-\bar{x})\right]\sqrt{\frac{\sum_{t=1}^{n}(y_t-\bar{y})^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}}.$$

Подставив в эту формулу значения x_t, y_t, $t=1,\ 2,\ 3,\ 4$ и $\bar{x}$, $\bar{y}$, получим $\widehat{\beta}_1^{\text{парам}}\approx 1{,}7$. Таким образом, $|\widehat{\beta}_1^{\text{парам}}-\widehat{\beta}_1|=0{,}04375<0{,}1=\Delta_0$.

ГЛАВА 4

АНАЛИЗ ФАКТОРОВ В ЗАДАЧЕ ВЫДЕЛЕНИЯ ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ПРИЗНАКОВ

4.1. ПОСТАНОВКА ЗАДАЧ ВЫЯВЛЕНИЯ ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ГЕОЛОГИЧЕСКИХ ПРИЗНАКОВ

Проблема выбора информативных комбинаций признаков многоаспектна (см. рис. 1). В настоящее время имеется несколько подходов к решению проблемы снижения числа изучаемых геологических признаков. В данной главе рассматриваются: статистические методы поиска информативных комбинаций геологических признаков, обусловливающих различия двух сопоставляемых геологических объектов (см. раздел 4.3.1); анализ проблемы поиска наиболее информативных признаков в моделях регрессии для решения задач прогнозирования (см. раздел 4.3.2). Первый подход (статистические методы) наиболее широко освещен в работах Д. А. Родионова [40, 42, 35] и в ряде публикаций других авторов [6, 7].

При попарном сопоставлении геологических объектов, охарактеризованных комплексом геологических признаков (например, содержаниями породообразующих и акцессорных минералов и элементов, физико-механическими свойствами горных пород и т. п.), исследователь, естественно, стремится выявить из их состава такую комбинацию признаков, которая бы наиболее контрастно различала сравниваемые объекты.

В качестве объектов сопоставления целесообразно использовать рудоносные и безрудные интрузивные массивы, метаморфические и метасоматические зоны, нефтегазоносные, водоносные и пустые пласты, ловушки и другие геологические образования. При таком выборе объектов исследований будет оправдано нахождение прямых и косвенных поисковых признаков среди установленной информативной комбинации.

В то же время одновременно выявленная неинформативная или малоинформативная комбинация геологических признаков будет свидетельствовать о чертах близости, сходства сравниваемых объектов, что может оказаться важным при выяснении генетических вопросов образования геологических объектов. Мы уже отмечали, что максимальное сходство и максимальное различие — важные показатели при прогнозировании рудоносных объектов. Таким образом, в геологических исследованиях представляет интерес выбор как информативных, так и неинформативных комбинаций признаков.

Поиск информативной и неинформативной комбинаций признаков (первый подход) осуществляется при сопоставлении двух геологических объектов, охарактеризованных комплексом m свойств.

Напротив, в задачах прогнозирования одного геологического свойства (второй подход) по комплексу других мы сталкиваемся с задачей снижения числа признаков в одном конкретно изучаемом объекте. Поиск информативной комбинации геологических признаков в этом случае позволяет без большой потери информации устранить признаки, мало влияющие на предсказуемый признак, т. е. оставить группу признаков, наиболее корреляционно связанную с прогнозируемой геологической характеристикой.

Если для первого подхода наиболее приемлемы классификационные процедуры (см. гл. 2), то для второго — корреляционные и регрессионные (см. гл. 3).

Следует сказать несколько слов о типах стратегий поиска комбинаций информативных и неинформативных признаков. В методической литературе известны два подхода — поиск «снизу» и поиск «сверху», т. е. процедура последовательного поиска первого, второго, ..., l-го информативного признака либо процедура определения первого, второго, ..., $(m-l)$-го неинформативного признака с автоматическим их отделением от информативной комбинации.

Первая стратегия использовалась во многих алгоритмах Д. А. Родионова, А. В. Гаранина, Л. А. Верховской, Р. И. Когана и др. [40, 35, 6], а вторую применяли Н. Н. Поплавский, С. В. Гольдин, А. Н. Бугаец и другие [13, 5], а также и Д. А. Родионов. От выбора стратегии и конкретного алгоритма нередко зависит эффективность выбора комбинаций информативных признаков. Заметим, что существует и третья стратегия — полный перебор комбинаций признаков, по-видимому, самый эффективный, но и самый трудоемкий.

Поиск тех или иных комбинаций признаков, когда общее число изученных геологических характеристик превышает 5—6 ($m>5$), без ЭВМ становится непосильной задачей, так как приходится решать системы линейных уравнений, находить детерминанты и обратные матрицы большой размерности, что вручную становится чрезвычайно затруднительно (хотя теоретически возможно).

4.2. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧ ВЫБОРА ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ПРИЗНАКОВ

Охарактеризуем формальную постановку задачи выбора информативных комбинаций признаков при сравнении двух геологических объектов.

Когда число изучаемых признаков m велико, возникает задача о сокращении без ущерба для конечного результата числа изучаемых геологических характеристик. Другими словами, желательно выявить такую комбинацию признаков, которая при меньшем числе входящих в нее характеристик обеспечивала бы ту же или бóльшую надежность различения сравниваемой пары геологических объектов. В практике геологических исследований представляют интерес не только те признаки (геологические свойства),

которые несут информацию о различиях между сравниваемыми объектами, но также и наборы признаков, которые такими различиями не обладают. Такие неинформативные комбинации признаков характеризуют близость, сходство сравниваемых объектов.

Приведем формальные определения этих понятий. В качестве удобных моделей наборов изучаемых геологических характеристик в двух сопоставляемых объектах рассматриваются две m-мерные случайные величины ξ и η:

$$\xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\},$$
$$\eta = \{\eta_1, \eta_2, \ldots, \eta_j, \ldots, \eta_m\},$$

и пусть θ_1 и θ_2 их параметры соответственно. В качестве этих параметров обычно рассматриваются: 1) многомерные векторы средних; 2) ковариационные матрицы; 3) векторы средних и ковариационные матрицы одновременно.

1. $\theta_1 = a_1 = \{a_j^{(1)}\} = \{a_1^{(1)}, a_2^{(1)}, \ldots, a_j^{(1)}, \ldots, a_m^{(1)}\}$,
   $\theta_2 = a_2 = \{a_j^{(2)}\} = \{a_1^{(2)}, a_2^{(2)}, \ldots, a_j^{(2)}, \ldots, a_m^{(2)}\}$.
2. $\theta_1 = \Sigma_1 = \{\sigma_{ij}^{(1)}\}$, $i, j = 1, 2, \ldots, m$,
   $\theta_2 = \Sigma_2 = \{\sigma_{ij}^2\}$, $i, j = 1, 2, \ldots, m$.
3. $\theta_1 = \{a_j^{(1)}, \Sigma_1\}$, $\theta_2 = \{a_j^{(2)}, \Sigma_2\}$.

В разделе 4.3.1. описана процедура поиска информативных комбинаций признаков относительно многомерных средних (т. е. ситуация 1) на основе ранговой статистики Пури—Сена—Тамуры, а также с помощью параметрического критерия Джеймса—Сю. Приведена процедура поиска информативных комбинаций признаков относительно ковариационных матриц (т. е. ситуация 2) на основе другой ранговой статистики Пури—Сена—Тамуры, а также с помощью параметрической статистики Кульбака.

Обозначим множество всех значений индекса j через M и будем полагать, что I_k — произвольное множество в M, содержащее k элементов. Это произвольное множество I_k будем называть комбинацией k признаков.

Дополнение множества I_k до M обозначим I_{m-k}, которое представляет собой комбинацию m—k признаков. Обозначим также через $\theta_1(I_k)$ и $\theta_2(I_k)$ наборы элементов многомерных параметров θ_1 и θ_2, соответствующие комбинации признаков I_k, и дадим определение неинформативной комбинации.

Неинформативной комбинацией I_k относительно многомерного параметра θ будем называть любую комбинацию k признаков I_k, для которой $\theta_1(I_k) = \theta_2(I_k)$. Если под θ понимается многомерное среднее, т. е. $\theta_1 = a_1$, $\theta_2 = a_2$, то неинформативная комбинация признаков будет определена относительно k-мерных векторов средних. Если же $\theta_1 = \Sigma_1$, $\theta_2 = \Sigma_2$, то неинформативная комбинация будет определена относительно подматриц Σ_1^k и Σ_2^k порядка $k \times k$ матриц Σ_1 и Σ_2 соответственно.

Любую комбинацию k признаков I_k, для которой $\theta_1(I_k) \neq \theta_2(I_k)$, будем называть информативной комбинацией относительно многомерного параметра θ и обозначать I^*.

4.3. РЕКОМЕНДУЕМЫЕ РАНГОВЫЕ И ПАРАМЕТРИЧЕСКИЕ ПРОЦЕДУРЫ ВЫБОРА ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ГЕОЛОГИЧЕСКИХ ПРИЗНАКОВ

4.3.1. Выбор информативных комбинаций признаков относительно многомерных средних и ковариационных матриц на основе ранговых и параметрических критериев Пури—Сена—Тамуры, Джеймса—Сю и Кульбака при сопоставлении двух геологических объектов

Выбор информативных комбинаций признаков при сравнении двух геологических объектов осуществляется обычно относительно двух параметров — многомерного среднего и ковариационной матрицы. Как и в задачах классификаций геологических объектов, повышение надежности статистических заключений и последующей их геологической интерпретации обеспечивается рациональным комплексированием ранговых и параметрических методов.

Для поиска информативных комбинаций геологических признаков относительно многомерного среднего без каких-либо ограничений на характер ковариационных матриц (т. е. требования их равенства, диагонального вида) в качестве рабочих статистик целесообразно привлекать ранговый критерий Пури—Сена—Тамуры Λ и параметрический критерий Джеймса—Сю $2I$ (см. раздел 3.4.1).

Для поиска информативных комбинаций относительно ковариационных матриц рекомендуется использование соответствующего рангового метода Пури—Сена—Тамуры Λ_ε и параметрического критерия Кульбака $2I_0$ (см. раздел 3.4.1).

В отношении применения параметрических критериев для совместного поиска полной информативной и полной неинформативной комбинаций признаков, а также для выбора наилучшей информативной комбинации можно воспользоваться рекомендациями Д. А. Родионова [35, гл. 7]. Для практических целей, по нашему мнению, представляет наибольший интерес первая часть вычислительной процедуры ранжирования признаков по их информативности, предложенная А. В. Гараниным и использованная Д. А. Родионовым. Она послужила основой для упрощенных процедур поиска полной и наилучшей информативной комбинаций признаков, базирующихся на ранговых и параметрических методах Пури—Сена—Тамуры, Джеймса—Сю, Кульбака, для проверки гипотез о многомерном среднем и ковариационных матрицах. Предпочтение отдается стратегии «поиска снизу» информативных комбинаций признаков, а не «поиска сверху» неинформативных комбинаций (см. раздел 4.1).

Как отмечалось выше (см. раздел 4.2), исходному комплексу признаков или свойств геологических объектов можно поставить в соответствие модель, представляющую m-мерную случайную величину.

Выбор информативной комбинации признаков для различения двух многомерных случайных величин ξ и η может быть сведен к выделению из имеющегося набора m признаков такой комбинации из g признаков, которая дает максимальную различающуюся информацию между двумя рассматриваемыми случайными величинами относительно многомерного среднего или ковариационных матриц [40, 39, 35, 36, 37].

Таким образом, в качестве меры различающей информации рассматриваются: ранговая статистика Λ критерия Пури—Сена—Тамуры в комплексе с параметрической статистикой $2I$ критерия Джеймса—Сю, по которым строятся соответствующие критерии для проверки гипотезы о равенстве математических ожиданий двух многомерных случайных величин ξ и η, а также другая ранговая статистика Λ_Σ критерия Пури—Сена—Тамуры в комплексе с параметрической статистикой $2I_0$ критерия Кульбака — для проверки гипотезы о равенстве ковариационных матриц двух многомерных случайных величин ξ и η.

Факт принятия нулевой гипотезы можно интерпретировать как отсутствие различающей информации относительно многомерного среднего или ковариационных матриц в подавляющем большинстве признаков, участвующих в проверке. Если же нулевая гипотеза о равенстве математических ожиданий или ковариационных матриц отвергается, то это значит, что в проверяемой комбинации признаков присутствуют такие признаки, которые несут определенную различающую информацию относительно указанных параметров.

Чтобы упростить поиск информативной комбинации признаков, проводят прежде всего упорядочение (ранжирование) признаков по мере уменьшения заложенной в них различающей информации. Наиболее информативным признаком считается тот (j_1), для которого одномерные критерии проверки гипотез о равенстве математических ожиданий и дисперсий (т. е. критерии Вилкоксона W, Вэлча t, Сиджела—Тьюки R и Бартлета M — а также одномерные варианты критериев Пури—Сена—Тамуры Λ и Λ_Σ) дали наибольшее значение статистики различающей информации *. Вторым по информативности признаком выбирается тот (j_2), в совокупности с которым наиболее информативный признак j_1 дал при проверке гипотез о равенстве двухмерных математических ожиданий или двухмерных ковариационных матриц опять же наибольшие значения статистики различающей информации, и т. д. Например, признаком k-м по величине информативности будет тот (j_k),

---

* Следует также учитывать, что с точностью до множителя одномерный вариант критерия Джеймса—Сю совпадает с критерием Вэлча, а одномерный вариант критерия Кульбака — с критерием Бартлета.

в совокупности с которым найденные на предыдущем шаге наиболее информативные признаки j_1, j_2, ..., j_{k-1} дали при проверке гипотез о равенстве k-мерных средних или k-мерных ковариационных матриц наибольшие значения статистики различающей информации.

Статистики, по которым проводится упорядочение признаков, несут в себе информацию о различиях между математическими ожиданиями или ковариационными матрицами двух случайных величин ξ и η, представленных той или иной комбинацией признаков.

Выбрав уровень значимости α, можно для каждой комбинации признаков (включающей любое число первых признаков из упорядоченной последовательности: j_1, j_2, ..., j_m) ответить на вопрос, является ли эта комбинация неинформативной для заданной вероятности ошибки первого рода α. Если все комбинации признаков оказываются неинформативными, то вопрос о выборе информативных признаков для различения данных двух случайных величин отпадает. В противном случае можно отметить комбинацию из g признаков (первых $g \leqslant m$ признаков из упорядоченного ряда), прибавление к которой любого из оставшихся признаков делает комбинацию неинформативной (когда принимается нулевая гипотеза). Эту комбинацию из g признаков назовем полной информативной комбинацией.

Если все комбинации признаков оказались информативными, то, значит, полной информативной комбинацией будет вся совокупность m исходных признаков. Однако полная информативная комбинация признаков может содержать в себе и неинформативные признаки, добавление которых к первым в упорядоченном ряду информативных признаков все же оставляет всю комбинацию в классе информативных комбинаций. Чтобы выделить из полной информативной комбинации признаков наилучшую комбинацию полностью информативных признаков, необходимо оценить ту долю различающей информации, которую вносит каждый добавляемый признак. Для этого следует построить ряд из приращений значений статистик, по которым проводилось упорядочение признаков. Так как ранжирование ряда признаков проводилось по уменьшению различающей информации, вносимой каждым последующим признаком, то ряд, составленный из приращений значений статистик, будет убывающим.

Комбинацию из g_0 признаков ($g_0 \leqslant g$) назовем наилучшей информативной комбинацией («ядром» информативности), если для признаков, номера которых в ранжированном ряду больше g_0, приращения статистик различающей информации незначимы*.

Таким образом, комбинация из g первых признаков (j_1, j_2, ..., j_g) упорядоченного ряда признаков будет полной информативной ком-

---

* В главе 4 приводятся лишь упрощенные процедуры поиска полной и наилучшей комбинаций признаков. С более строгими алгоритмами следует ознакомиться в работе Д. А. Родионова [35, гл. 7].

бинацией, если для комбинации из i признаков, где $i>g$, принимается нулевая гипотеза H_0 о равенстве математических ожиданий или о равенстве ковариационных матриц в двух объектах, т. е. если

$$V_i \geqslant c_{g,f=i} \text{ для } i=1, 2, \ldots, g, \text{ поэтому } H_1: M\xi \neq M\eta \text{ или } H_1: \Sigma_\xi \neq \Sigma_\eta$$

и

$$V_i \leqslant c_{g,f=i} \text{ для } i>g, \text{ поэтому } H_0: M\xi = M\eta \text{ или } H_0: \Sigma_\xi = \Sigma_\eta,$$

где V_i — соответствующие многомерные ранговые или параметрические статистики критериев Пури—Сена—Тамуры, Джеймса—Сю или Кульбака, а также одномерные статистики Вилкоксона, Сиджела—Тьюки, Вэлча и Бартлета, $c_{g,f=i}$ — соответствующие критические значения, $M\xi$ и $M\eta$ — математические ожидания случайных величин ξ и η — моделей комплекса m геологических характеристик в сопоставляемых геологических объектах, Σ_ξ и Σ_η — ковариационные матрицы (характеристики изменчивости и коррелированности признаков) в сопоставляемых геологических объектах.

Комбинацию из g_0 ($g_0 \leqslant g$) первых признаков (j_1, j_2, ..., j_{g_0} назовем наилучшей информативной комбинацией («ядром» информативности), если для признаков, номера которых в ранжированном ряду больше g_0 (т. е. $i>g_0$), приращения статистик различающей информации ΔV незначимы, т. е. принимается нулевая гипотеза H_0 для разности параметров нецентральности

$$\Delta V_i > \chi^2_{\alpha,f=i} \text{ для } i=1, 2, \ldots, g_0, \text{ поэтому } H_1: \Delta\beta_i \neq 0,$$

$$\Delta V_i \leqslant \chi^2_{\alpha,f=i} \text{ для } i>g_0, \text{ поэтому } H_0: \Delta\beta_i = 0,$$

где ΔV_i — приращения вышеуказанных статистик, $\Delta\beta_i = \beta_i - \beta_{i-1}$ — разность параметров нецентральности для распределения x^2 с одной степенью свободы, $\chi^2_{\alpha,f=1}$ — критическое значение χ^2 распределения центрального с одной степенью свободы (при $\alpha=0{,}05$, $\chi^2_{\alpha,f=1}=3{,}84$).

Для обоснования аппроксимации закона распределения статистики ΔV_i можно воспользоваться представлением квадратических форм статистик Λ и Λ_Σ критериев Пури—Сена—Тамуры, статистики $2I$ критерия Джеймса—Сю, а также отношением определителей матриц — статистикой $2I_0$ критерия Кульбака (которые все асимптотически имеют χ^2-распределение) через сумму квадратов ортогональных комбинаций исходных случайных величин [28].

Ниже на примере ранговой статистики Λ критерия Пури—Сена—Тамуры подробно описан упрощенный алгоритм поиска полной и наилучшей информативных комбинаций признаков.

Случаи с другими рабочими статистиками описаны схематично, так как процедуры полностью аналогичны указанной со статистикой Λ критерия Пури—Сена—Тамуры.

4.3.1.1. Выбор полной и наилучшей комбинаций признаков, информативных относительно многомерных средних, на основе рангового критерия Пури—Сена—Тамуры

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 (см. раздел 2.4.1), т. е. $\Lambda_j (j=1, 2, ..., m)$, которые имеют χ^2-распределение с одной степенью свободы. Выбираем то значение $j=j_1$, для которого Λ_j максимальна, т. е. $\Lambda_{j1}=\max \Lambda_j$.

2. Вычисляется $m-1$ величин ранговой статистики Λ критерия Пури—Сена—Тамуры для двухмерных величин, образованных j_1 признаком и одним из оставшихся $j\neq j_1$. Выбирается то значение j_2, для которого максимальна статистика $\Lambda_{j_1 j_2}$, т. е. $\Lambda_{j_1 j_2}=\max \Lambda_{j_1, j}$.

3. Процедура повторяется для всех $j=2, 3, ..., m$. Так, для $j=j_g \leqslant m$ вычисляется $(m-g)$ значений ранговой статистики Λ критерия Пури—Сена—Тамуры для g-мерных величин, образованных $(j_1 j_2, ..., j_{(g-1)})$ признаками на предыдущем шаге и одним из оставшихся $j\neq(j_1, j_2, ..., j_{(g-1)})$. Вновь выбирается то значение j_g, для которого максимальна статистика $\Lambda_{j_1, j_2, ..., j_g}$, т. е. $\Lambda_{j_1, j_2, ..., j_g}=\max \Lambda_{j_1 j_2, ..., j_{(g-1)}, j}$. При $g=m$ полученная последовательность номеров признаков $j_1, j_2, ..., j_m$ будет соответствовать их расположению от наилучшего к наихудшему.

4. Ранжируя с помощью приведенного критерия имеющийся в распоряжении геолога ряд из m признаков, получим набор статистик:

$$\max_j \Lambda_j^{(1)},\ \max_j \Lambda_{j_1, j}^{(2)},\ \max_j \Lambda_{j_1, j_2, j}^{(3)},\ \ldots,$$

$$\max_j \Lambda_{j_1, j_2, \ldots j_{(g-1)}, j}^{(g)},\ \ldots,\ \Lambda^{(m)},$$

поставленный в соответствие с упорядоченным рядом самих признаков $j_1, j_2, ..., j_g, ..., j_m$. В обозначении $\Lambda_{j_1, j_2, ..., j_{(g-1)}, j}^{(g)}$ 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. Как отмечалось (см. разд. 3.4.1), 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 (рис. 17).

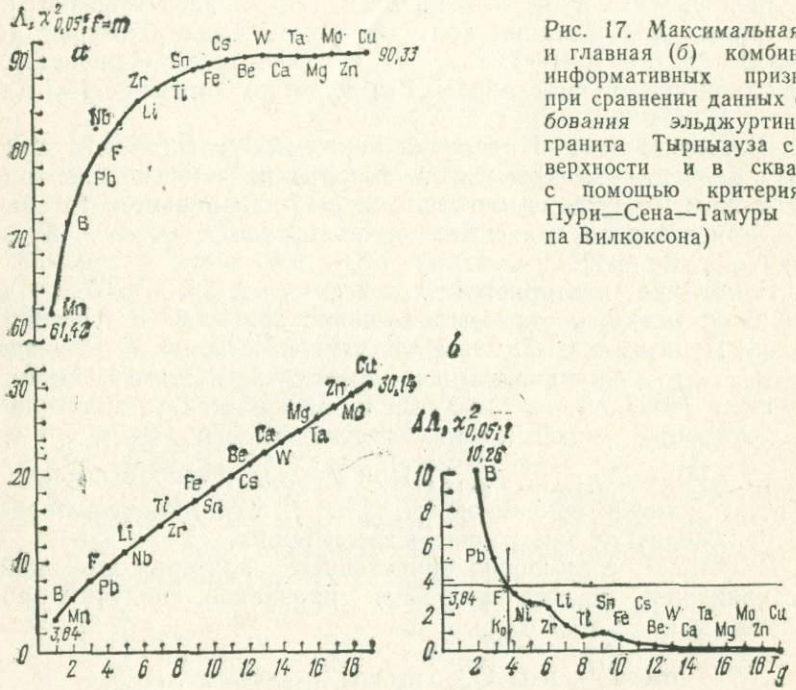


Рис. 17. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия Λ Пури—Сена—Тамуры (типа Вилкоксона)

6. Комбинация из g первых признаков (j_1, j_2, ..., 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

$$\Lambda_i > \chi^2_{\alpha, f=i} \quad \text{для } i = 1, 2, \ldots, g\,(H_1 : M\xi \neq M\eta),$$

а

$$\Lambda_i \leqslant \chi^2_{\alpha, f=i} \quad \text{для } i > g\,(H_0 : M\xi = Mh).$$

7. Если полная информативная комбинация признаков состоит более чем из одного признака ($g \geqslant 2$), то с помощью ряда из приращений значений статистик

$$\Delta\Lambda_i = \Lambda_i - \Lambda_{i-1} \quad (i = 2, 3, \ldots, g)$$

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 $\Delta\Lambda_i$ статистически уменьшаются с увеличением номера i. Первые g_0 признаков будут наилучшей информативной комбинацией, если

$$\Delta\Lambda_i > \chi^2_{\alpha,f=1} \quad \text{для } i = 2, 3, \ldots, g_0 (H_1 : \Delta\beta_i \neq 0)$$

и

$$\Delta\Lambda_i \leqslant \chi^2_{\alpha,f=1} \quad \text{для } i > g_0 (H_0 : \Delta\beta_i = 0).$$

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_i в ранжированном ряду.

10. Комбинацию признаков $i=m-g$, для которой принимается нулевая гипотеза о равенстве многомерных средних

$$\Lambda_i \leqslant \chi^2_{\alpha,f=i} \quad \text{для } i > g,$$

следует полагать неинформативной или малоинформативной, т. е. обеспечивающей, скорее всего, черты сходства сопоставляемых объектов.

Пример. Проиллюстрируем процедуру упорядочения (или ранжирования) геологических признаков по степени убывания их информативности, поиск полной и наилучшей информативной, а также неинформативной комбинаций признаков на примере сопоставления содержаний 19 рудных, редких и петрогенных элементов в эльджуртинском граните на поверхностном эрозионном срезе и по керну скважины, воспользуясь данными, приведенными в работе В. В. Ляховича [30, табл. 31, 107, 108].

1. По каждому элементу в отдельности (т. е. по 19 элементам) с помощью ранговой одномерной модификации критерия Пури—Сена—Тамуры определяем статистики Λ_j ($j=1, 2, \ldots, 19$). Выбираем из них максимальное значение: максимальное значение $\Lambda_{19}=61{,}42 \gg 3{,}84$ оказалось для содержаний марганца — 19-го признака.

Таким образом, на первом шаге вычислений устанавливаем, по крайней мере, один информативный признак — содержание марганца.

2. Определяем с помощью рангового критерия Пури—Сена—Тамуры 18 двухмерных статистик Λ_2 по данным о 19-м признаке (содержаниях марганца) и одном из оставшихся. Выбираем из них максимальное. Максимальное значение оказалось для пары: марганец—бор, т. е. $\Lambda_{19;\,13}=71{,}68 \gg \chi^2_{0,05;\ 2}=5{,}99$.

3. Вновь с помощью рангового критерия Пури—Сена—Тамуры определяем 17 трехмерных статистик Λ_3 по данным о содержаниях 19-го признака (марганца), 13-го признака (бора) и одного из оставшихся. Выбираем из них максимальную статистику: $\Lambda_{19,\,13,\,4}=76{,}97 \gg \chi^2_{0,05;\ 3}=7{,}81$. Максимальным значением оказа-

лась статистика $\Lambda_{19,13;4}$, соответствующая 19, 13 и 4-му признакам, т. е. сочетанию содержаний марганца, бора и свинца.

4. Последовательно продолжаются этапы ранжирования признаков.

В результате получен следующий упорядоченный по степени убывания информативности ряд признаков:

Mn	B	Pb	F	Nb	Li	Zr	Ti	Sn	Fe
19	13	4	14	7	11	9	15	3	16
61,42	71,68	76,97	80,09	82,82	85,35	86,74	87,63	88,70	89,49

Cs	Be	W	Ca	Ta	Mg	Mo	Zn	Cu
12	10	1	18	8	17	2	5	6
89,85	90,1	90,22	90,28	90,31	90,32	90,33	90,34	90,35

5. Определяем полную и наилучшую информативную и неинформативную (малоинформативную) комбинации признаков. Поскольку сочетание всех ранжированных 19 элементов, включая неинформативные элементы Mg, Mo, Zn, Cu, дает все же значение различающей информации Λ, превышающей соответствующее критическое значение, т. е. $\Lambda_{19}=90{,}35 \gg \chi^2_{0,05;19} = 30{,}14$, то полную информативную комбинацию составят все 19 элементов.

Ранжированный по убыванию информативности ряд признаков (элементов) находится выше линии критических значений (см. рис. 17, *а*).

Полученная с помощью анализа приращения статистик ($\Delta\Lambda$) наилучшая информативная комбинация признаков (ядро информативности по критерию Пури—Сена—Тамуры) состоит из трех наиболее информативных признаков: содержаний марганца, бора и свинца (см. рис. 17, *б*), так как

$$\Delta\Lambda_{2-1} = 10{,}26 > \chi^2_{0,05;1} = 3.84,$$

$$\Delta\Lambda_{3-2} = 5{,}29 > \chi^2_{0,05;1} = 3{,}84,$$

а

$$\Delta\Lambda_{4-3} = 3{,}12 < \chi^2_{0,05;1} = 3{,}84,$$

то же для последующих

$$\Delta\Lambda_{5-4},\ \Delta\Lambda_{6-5},\ \ldots,\ \Delta\Lambda_{19-18} < 3{,}84.$$

4.3.1.2. Выбор полной и наилучшей комбинаций признаков, информативных относительно многомерных средних, на основе параметрического критерия Джеймса—Сю

Процедура ранжирования геологических признаков (свойств) от наиболее информативного к наименее информативному, приемы поиска полной I_g и наилучшей I_g информативной, а также неинформативной I_{m-g} комбинаций признаков на основе параметрического критерия Джеймса—Сю полностью идентичны тем, которые описаны в предыдущем разделе 4.3.1.1. Другими словами, 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 (см. раздел 2.4.1) для g-мерных

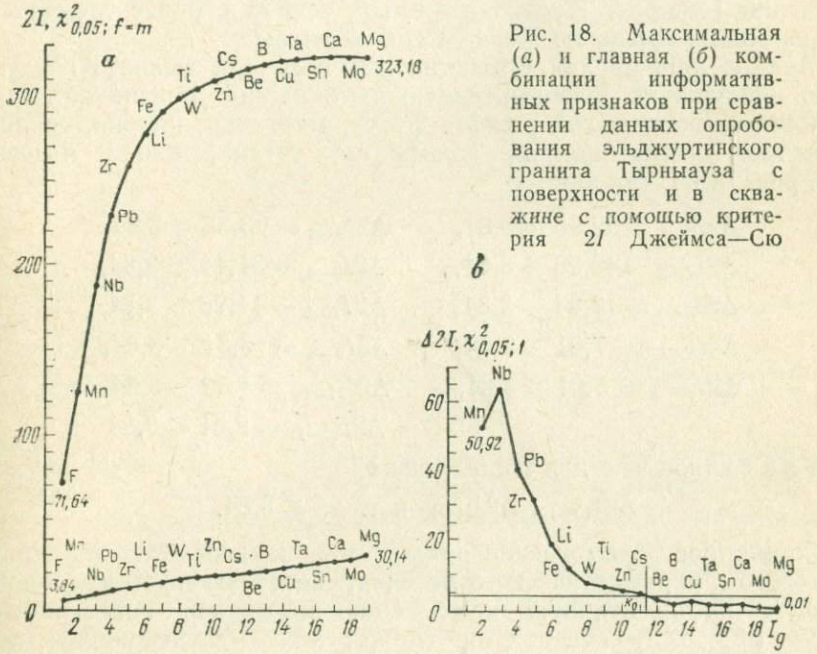


Рис. 18. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия $2I$ Джеймса—Сю

величин, образованных $(j_1, j_2, \ldots, j_{(g-1)})$ признаками 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 $2I = \max 2I_{j_1, j_2, \ldots, j_{g-1}}$.

В качестве примера вновь воспользуемся данными о содержаниях 19 рудных, редких и петрогенных элементов в эльджуртинском граните (см. раздел 4.3.1.1).

В результате вычислений на ЭВМ получен следующий упорядоченный по степени убывания информативности ряд признаков (рис. 18):

F	Mn	Nb	Pb	Zr	Li	Fe	W	Ti	Zn
14	19	7	4	9	11	16	1	15	5
71,61	124,56	188,11	228,4	259,84	278,48	290,24	297,25	303,56	308,57

Cs	Be	B	Cu	Ta	Sn	Ca	Mo	Mg
12	10	13	6	8	3	18	2	17
312,89	315,4	317,4	319,14	320,49	321,46	322,62	323,17	323,18

Так как сочетание всех 19 элементов, упорядоченных по убыванию информативности, дает значение различающей информации $2I$, значительно превышающее соответствующее критическое значение, т. е. $2I = 323{,}18 \gg \chi^2_{0{,}05;\ 19} = 30{,}14$, то необходимо сделать вывод о том, что полную информативную комбинацию составляют все 19 элементов. Ранжированный по убыванию информативности ряд признаков (*элементов*) *находится выше линии* критических значений (см. рис. 18, *а*). Этот вывод совпал с ранее сделанным по ранговому критерию Пури—Сена—Тамуры.

Анализ приращений статистик $\Delta 2I$ позволил выявить следующую наилучшую информативную комбинацию признаков (ядро информативности): содержания фтора, марганца, ниобия, свинца, циркония, лития, железа, вольфрама, титана, цинка и цезия (см. рис. 18, *б*), так как

$$\Delta 2I_{2-1} = 52{,}95 > 3{,}84, \qquad \Delta 2I_{3-2} = 63{,}55 > 3{,}84,$$
$$\Delta 2I_{4-3} = 40{,}29 > 3{,}84, \qquad \Delta 2I_{5-4} = 31{,}44 > 3{,}84,$$
$$\Delta 2I_{6-5} = 18{,}64 > 3{,}84, \qquad \Delta 2I_{7-6} = 11{,}76 > 3{,}84,$$
$$\Delta 2I_{8-7} = 7{,}01 > 3{,}84, \qquad \Delta 2I_{9-8} = 6{,}31 > 3{,}84,$$
$$\Delta 2I_{10-9} = 5{,}01 > 3{,}84, \qquad \Delta 2I_{11-10} = 4.32 > 3{,}84,$$
$$\Delta 2I_{12-11} = 2{,}51 < 3{,}84$$

и то же соотношение для последующих

$$\Delta 2I_{13-12},\ \Delta 2I_{14-13} \text{ и т. д. } < 3{,}84.$$

Совместное рассмотрение результатов, полученных с помощью рангового и параметрического методов, свидетельствует, что в наилучшую информативную комбинацию признаков, по-видимому, входят марганец, свинец, фтор (может также бор и ниобий), а в наименее информативную комбинацию — молибден, магний, кальций (может также медь, цинк). Остальные элементы образуют неустойчивые малоинформативные комбинации признаков.

Из двух важнейших рудных элементов — W и Mo — более информативен вольфрам. *Информативность бора* по критерию Пури—Сена—Тамуры не подтверждается с помощью параметрического критерия Джеймса—Сю.

В настоящей работе не приведена обстоятельная геолого-геохимическая интерпретация статистических выводов, а также не дан анализ результатов, полученных по другой стратегии, — по возрастанию информативности.

4.3.1.3. Выбор полной и наилучшей комбинаций признаков, информативных относительно ковариационных матриц, на основе рангового критерия Пури—Сена—Тамуры

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 кри-

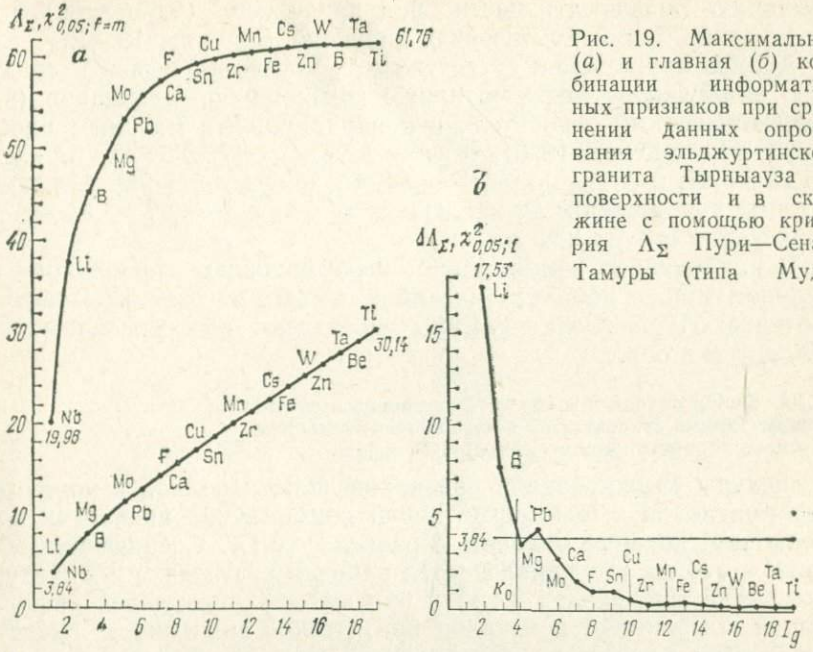


Рис. 19. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия Λ_Σ Пури—Сена—Тамуры (типа Муда)

терия Пури—Сена—Тамуры полностью идентичны тем, которые описаны в разделе 4.3.1.1. Естественно, что и в этом случае следует воспользоваться рабочей статистикой Λ_Σ критерия Пури—Сена—Тамуры.

В качестве примера вновь используем данные о содержаниях рудных, редких и петрогенных элементов в эльджуртинском граните, т. е. оценим влияние фактора глубинности на информативность (контрастность) указанных компонент относительно их коррелированности и степени рассеивания (относительно ковариационных матриц).

В результате вычислений на ЭВМ получен следующий упорядоченный по степени убывания информативности ряд признаков (рис. 19).

Nb	Li	B	Mg	Pb	Mo	Ca	F	Sn
7	11	13	17	4	2	18	14	3
19,98	37,51	45,3	48,99	53,16	55,92	57,42	58,43	59,99

Cu	Zr	Mn	Fe	Cs	Zn	W	Be	Ta	Ti
6	9	19	16	12	5	1	10	8	15
60,3	60,3	60,64	61,06	61,29	61,53	61,66	61,72	61,76	61,76

Сочетание всех 19 элементов, упорядоченных по убыванию информативности от ниобия до титана, дает значения различающей информации Λ_Σ, значительно превышающие соответствующее критическое значение χ^2, т. е. $\Lambda_\Sigma = 61{,}76 > \chi^2_{0{,}05;\ 19} = 30{,}14$. Поэтому необходимо сделать вывод, что полную информативную комбинацию составляют все 19 элементов. Ранжированный по

убыванию информативности ряд признаков (элементов) находится выше линии критических значений (см. рис. 19, *а*).

Анализ приращений статистик $\Delta\Lambda_\Sigma$ позволил выявить следующую наилучшую информативную комбинацию признаков (ядро информативности относительно ковариационных матриц): ниобий, литий, бор (см. рис. 19, *б*), так как $\Delta\Lambda_{\Sigma(2-1)}=17{,}53>3{,}84$, $\Delta\Lambda_{\Sigma(3-2)}=7{,}79>3{,}84$, а $\Delta\Lambda_{\Sigma(4-3)}=3{,}59<3{,}84$, то же соотношение для последующих приращений, за исключением $\Delta\Lambda_{\Sigma(5-4)}=4{,}27>3{,}84$ за счет флуктуации статистики Λ_Σ.

В наилучшую комбинацию информативных признаков при желании можно добавить магний и свинец, но основную часть по критерию Пури—Сена—Тамуры составляют вышеуказанные ниобий, литий и бор.

4.3.1.4. Выбор полной и наилучшей комбинаций признаков, информативных относительно ковариационных матриц, на основе параметрического критерия Кульбака

Процедуры ранжирования признаков, поиска полной и наилучшей информативной и неинформативной комбинаций признаков идентичны тем, которые описаны в разделе 4.3.1.1. Специфические отличия — это то, что привлекается рабочая статистика $2I_0$ критерия Кульбака (см. раздел 2.4.1) и при выборе наилучшей информативной комбинации признаков критические значения χ^2 представляют 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 в разделах 4.3.1.1, 4.3.1.2 и 4.3.1.3.

В качестве примера рассматривается та же геологическая задача о влиянии фактора глубинности на содержания рудных, редких и петрогенных элементов в эльджуртинском граните.

В результате вычислений на ЭВМ получен следующий упорядоченный по степени убывания информативности ряд признаков (рис. 20):

Ta	Sn	Mn	Pb	Li	Mo	Cu
8	3	19	4	11	2	6
149,08	286,66	330,85	378,72	414,9	454,31	481,13
Mg	Ti	F	Zn	Cs	Fe	B
17	15	14	5	12	16	13
507,78	557,78	603,41	634,84	673,48	708,3	743,31
Be	W	Ca	Zr	Nb		
10	1	18	9	7		
777,35	813,06	845,42	875,95	891,42		

Так как сочетание всех 19 элементов, упорядоченных по убыванию информативности относительно ковариационных матриц, характеризуется статистикой различающей информации $2I_0$, значительно превышающей соответствующее критическое значение, т. е. $2I_0=891{,}42\gg\chi^2_{0,05;\,190}=223{,}2$ при степенях свободы $f=0{,}5\,m\,(m+1)=0{,}5\cdot19\cdot20=190$, то следует сделать заключение, что полную информативную комбинацию составляют все 19 элементов.

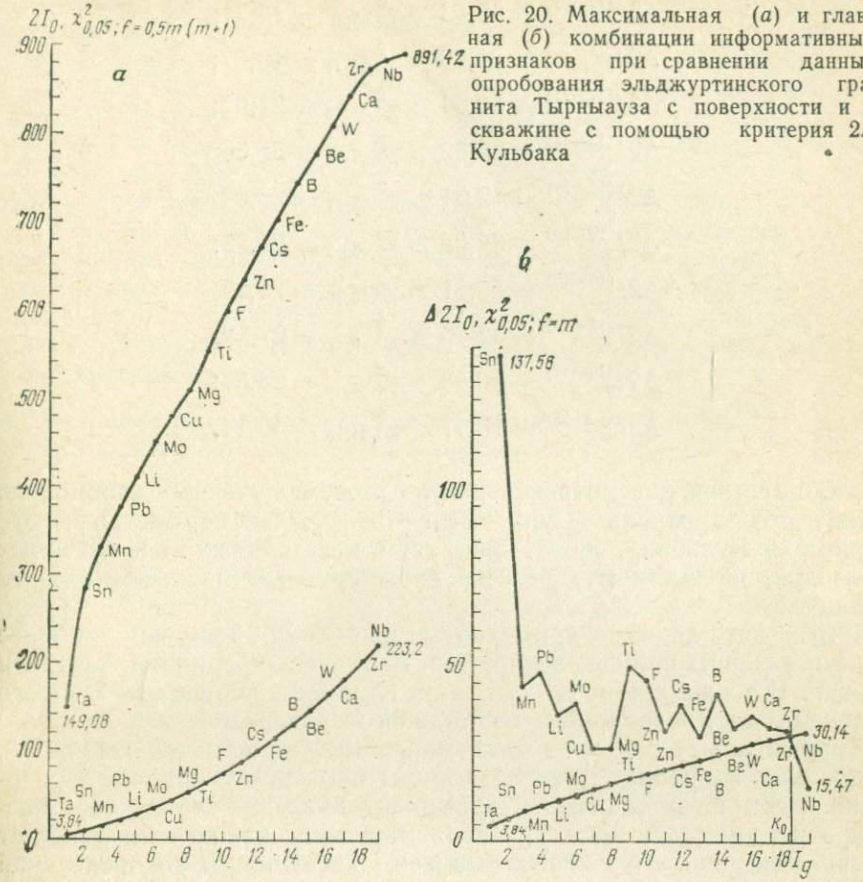


Рис. 20. Максимальная (*а*) и главная (*б*) комбинации информативных признаков при сравнении данных опробования эльджуртинского гранита Тырныауза с поверхности и в скважине с помощью критерия $2I_0$ Кульбака

Ранжированный по убыванию информативности ряд признаков (элементов) находится выше линии критических значений χ^2 (см. рис. 20, *а*).

Этот вывод совпал с ранее сделанным по ранговому критерию Пури—Сена—Тамуры. Анализ приращений статистик $\Delta 2I_0$ позволил выявить наилучшую информативную комбинацию признаков, состоящую из всех, кроме ниобия, элементов. Другими словами, тантал, олово, марганец, свинец, литий, молибден, медь, магний, титан, фтор, цинк, цезий, железо, бор, бериллий, вольфрам, кальций и цирконий входят по критерию Кульбака в состав главной информативной комбинации признаков (см. рис. 20, *б*), так как

$$\Delta 2I_0^{(2-1)} = 137{,}58 > \chi^2_{0,05;2} = 5{,}99; \quad \Delta 2I_0^{(3-2)} = 44{,}19 > \chi^2_{0,05;3} = 7{,}81;$$

$$\Delta 2I_0^{(4-3)} = 47{,}87 > \chi^2_{0,05;4} = 9{,}49; \quad \Delta 2I_0^{(5-4)} = 36{,}18 > \chi^2_{0,05;5} = 11{,}07;$$

$$\Delta 2I_0^{(6-5)} = 39{,}41 > \chi^2_{0,05;6} = 12{,}59; \quad \Delta 2I_0^{(7-6)} = 26{,}82 > \chi^2_{0,05;7} = 14{,}07;$$

$$\Delta 2I_0^{(8-7)} = 26{,}65 > \chi^2_{0,05;8} = 15{,}51; \quad \Delta 2I_0^{(9-8)} = 50{,}0 > \chi^2_{0,05;9} = 16{,}92;$$

$$\Delta 2I_0^{(10-9)} = 45{,}63 > \chi^2_{0{,}05;10} = 18{,}31;$$
$$\Delta 2I_0^{(11-10)} = 31{,}43 > \chi^2_{0{,}05;11} = 19{,}67;$$
$$\Delta 2I_0^{(12-11)} = 38{,}64 > \chi^2_{0{,}05;12} = 21{,}02;$$
$$\Delta 2I_0^{(13-12)} = 29{,}82 > \chi^2_{0{,}05;13} = 22{,}36;$$
$$\Delta 2I_0^{(14-13)} = 42{,}01 > \chi^2_{0{,}05;14} = 23{,}69;$$
$$\Delta 2I_0^{(15-14)} = 32{,}04 > \chi^2_{0{,}05;15} = 24{,}99;$$
$$\Delta 2I_0^{(16-15)} = 35{,}71 > \chi^2_{0{,}05;16} = 26{,}29;$$
$$\Delta 2I_0^{(17-16)} = 32{,}36 > \chi^2_{0{,}05;17} = 27{,}59;$$
$$\Delta 2I_0^{(18-17)} = 30{,}53 > \chi^2_{0{,}05;18} = 28{,}87;$$
а $$\Delta 2I_0^{(19-18)} = 15{,}47 < \chi^2_{0{,}05;19} = 30{,}14.$$

Совместное рассмотрение результатов, полученных с помощью рангового критерия Пури—Сена—Тамуры и параметрического критерия Кульбака, свидетельствует о неустойчивости полученных выводов, по-видимому, за счет сложнораспределенных исходных данных.

В самом деле, по критерию Пури—Сена—Тамуры наиболее информативны ниобий, литий, бор, наименее — бериллий, тантал и титан. В то же время по критерию Кульбака (мощность которого зависит от многомерного нормального распределения) наиболее информативны тантал, олово, марганец, а наименее — кальций, цирконий и ниобий. Анализ таблицы проверки гипотез о дисперсиях с помощью одномерного рангового критерия Синджела—Тьюки и параметрического критерия Бартлета также подтверждает появление противоречивых выводов, касающихся распределений тантала, ниобия, бора, кальция и ряда других элементов.

Таким образом, совместное рассмотрение результатов поиска информативных комбинаций признаков относительно ковариационных матриц с помощью параметрических и ранговых методов не позволяет сделать скоропалительных, но ненадежных заключений, получаемых на основе только параметрических или только ранговых методов, поскольку исследователю априорно неизвестен вид распределений многомерных случайных величин — моделей изучаемых геологических свойств. Конечно, более просто применять лишь один метод — параметрический или ранговый — и геологически интерпретировать полученные статистические результаты, но это не значит, что мы получим надежное заключение. Если геологическая закономерность существует, то она в среднем в большом числе расчетов будет достаточно надежно выявляться соответствующим сочетанием параметрических и ранговых статистических методов.

В приведенных примерах не встретилась ситуация с отделением полной информативной от неинформативной комбинации

признаков, так как в состав полной информативной комбинации входили все m признаки. Указанная ситуация рассмотрена в более ранних работах Р. И. Когана, Д. А. Родионова и других [35, 6].

В настоящей книге содержится много геологических примеров, иллюстрирующих сложную для интерпретации ситуацию, когда результаты, полученные с помощью параметрических методов, существенно отличаются от тех, которые получены на основе ранговых критериев. Это приведено не для того, чтобы противопоставлять ранговые и параметрические критерии, а напротив, чтобы подчеркнуть необходимость их комплексирования и очень бережного отношения к ответственной операции по геологической интерпретации полученных статистических выводов. Мы полагаем, что вместе с параметрическими почти всегда должны привлекаться и ранговые критерии.

4.3.2. Выбор информативных комбинаций геологических характеристик для одного геологического объекта

Наряду с задачей выявлений информативных, контрастных, геологических характеристик при сопоставлении двух геологических объектов, описанной в разделе 4.3.1, не меньшее практическое значение имеет задача определения наиболее важных информативных геологических характеристик для одного геологического объекта. Эта задача относится ко второму этапу прогнозирования геологических характеристик — определению формы регрессионной зависимости.

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.

Эта задача рассматривается во многих работах по регрессионному анализу [2, 15, 16, 19, 23, 41], в которых описаны различные процедуры последовательного поиска наилучшего в смысле принимаемого критерия подмножества регрессоров. Таких критериев описано несколько, особенно много их в монографии Дж. Себера [41]. В настоящей работе задача выбора какого-либо критерия из известного их числа не рассматривается. Однако, задавшись таким критерием, встает задача выбора процедуры поиска наилучшего подмножества регрессоров. Можно предложить три такие процедуры.

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 условия 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.

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 подмножества регрессоров. Анализ практических результатов, полученных 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 [16, 41] для определения так называемого ядра информативности, а затем экономную жесткую процедуру. Использование шагового метода вместо полного перебора позволит сделать процедуру более экономной.

Для выбора информативной комбинации геологических характеристик необязательно нахождение оценок коэффициентов соответствующих уравнений регрессии. Процедуры проверки значимости коэффициентов регрессии, приведенные в главе 3, как для параметрической, так и для ранговой регрессии не требуют определения самих оцениваемых параметров. Поэтому при выборе информативной комбинации геологических характеристик для одного геологического объекта могут быть рекомендованы ранговая процедура, использующая описанный выше прием поиска наилучшего подмножества регрессоров с нахождением ядра информативности, и описанный в разделе 3.4.3.2 метод проверки значимости коэффициентов множественной ранговой линейной регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. М., Статистика, 1974.

2. *Айвазян С. А.* Статистическое исследование зависимостей. М., Металлургия, 1968.

3. *Белов Ю. П., Голубева В. А., Коган Р. И.* Применение методов корреляционного, дисперсионного и регрессионного анализа в геологических исследованиях. М., ВИЭМС, 1976 (Экспресс-информация. Мат. методы исслед. в геологии, № 10).

4. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., ВЦ АН СССР, 1965.

5. *Бугаец А. Н,. Дуденко Л. Н.* Математические методы при прогнозировании месторождений полезных ископаемых. Л., Недра, 1976.

6. *Верховская Л. А., Голубева В. А., Коган Р. И.* Выбор информативной комбинации признаков для различения двух геологических объектов. М., ВИЭМС, 1972 (Обзор. Мат. методы исслед. в геологии).

7. *Верховская Л. А., Коган Р. И.* Выбор информативной комбинации признаков относительно ковариаций для различения двух геологических объектов. М., ВИЭМС, 1972 (Экспресс-информация. Мат. методы исслед. в геологии, № 5).

8. *Виноградов С. А., Созон-Ярошевич Е. А.* Применение математических методов и ЭВМ для решения задач расчленения и корреляции геологических разрезов. М., ВИЭМС, 1979 (Обзор. Мат. методы исслед. в геологии).

9. *Виноградов С. А., Созон-Ярошевич Е. А.* Автоматизированное решение задач расчленения и корреляции геологических разрезов с применением методов непараметрической статистики. — В кн.: Тр. ВНИГРИ. Л., 1978, с. 23—43.

10. *Вистелиус А. Б.* Математическая геология — ее основные направления и задачи. — Сов. геология, 1977, № 1, с. 11—34.

11. *Гаек Я., Шидак З.* Теория ранговых критериев. М., Наука, 1971.

12. *Гавришин А. И.* Процедуры классификации геологических объектов с помощью многомерного критерия. М., ВИЭМС, 1978 (Экспресс-информация, № 1).

13. *Гольдин С. В.* О проверке однородности совокупностей геологических объектов. — В кн.: Мат. методы при геологических исслед. в Западной Сибири. Тюмень, 1968 (Тр. ЗапСибНИГНИ).

14. *Давид М.* Геостатистические методы при оценке запасов руд. Пер. с англ. Л., Недра, 1980.

15. *Демиденко Е. З.* Линейная и нелинейная регрессии. М., Финансы и статистика, 1981.

16. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М., Статистика, 1973.

17. *Дюран Б., Оделл П.* Кластерный анализ. М., Статистика, 1977.

18. *Закс Л.* Статистическое оценивание. М., Статистика, 1976.

19. *Кендалл М., Стюарт А.* Статистические выводы и связи. М., Наука, 1973.

20. *Кендалл М.* Ранговые корреляции. М., Статистика, 1975.

21. *Классификация* и кластер. Под ред. Дж. Вэн Райзина. М., Мир, 1980.

22. *Коган Р. И., Анохин В. В., Жогина Г. Г.* Особенности применения многомерных статистик при решении классификационных геологических задач. М., ВИЭМС, 1972 (Обзор. Мат. методы исслед. в геологии).

23. *Комаров И. С., Хайме Н. М., Бабенышев А. П.* Многомерный статистический анализ в инженерной геологии. М., Недра, 1976.

24. *Константинов Р. М.* Математические методы количественного прогноза рудоносности. М., Недра, 1979.

25. *Коуден Д.* Статистические методы критерия качества. М., Физматгиз, 1961.

26. *Крамбейн У., Грейбилл Ф.* Статистические модели в геологии. М., Мир, 1969.

27. *Крамбейн У., Кауфман М., Мак-Кеммон Р.* Модели геологических процессов. М., Мир, 1973.

28. *Крамер Г.* Математические методы статистики. М., Мир, 1975.
29. *Кульбак С.* Теория информации и статистика. М., Наука, 1967.
30. *Ляхович В. В.* Связь оруденения с магматизмом (Тырныауз). М., Наука, 1976.
31. *Мардиа К.* Статистический анализ угловых наблюдений. М., Наука, 1978.
32. *Овчинников Л. Н., Коган Р. И.* О применении методов математической статистики в абсолютной геохронологии. — В кн.: Статистические методы геол. исследований. М., ИМГРЭ, 1971.
33. *Миллер Р., Кан Дж.* Статистический анализ в геологических науках. М., Мир, 1965.
34. *Оуэн Д. Б.* Сборник статистических таблиц. М., ВЦ АН СССР, 1966.
35. *Родионов Д. А.* Статистические решения в геологии. М., Недра, 1981.
36. *Родионов Д. А., Коган Р. И., Белов Ю. П.* Статистические методы классификации геологических объектов. М., ВИЭМС, 1979 (Обзор. Мат. методы исслед. в геологии).
37. *Родионов Д. А., Коган Р. И., Белов Ю. П.* Сравнительный анализ алгоритмов и программ для прогнозирования геологических свойств и рудоносных объектов. — В кн.: Использование мат. методов при прогнозе рудоносности. М., Наука, 1977, с. 67—73.
38. *Родионов Д. А., Коган Р. И., Голубева В. А.* Комплексирование статистических методов при обработке геологоразведочных данных по рудным месторождениям. — В кн.: Горнорудный Пршибрам в науке и технике. Пршибрам (ЧССР), 1981, с. 216—224.
39. *Родионов Д. А., Коган Р. И., Белов Ю. П.* Анализ современных математических методов решения типовых задач геологии. — В кн.: Горнорудный Пршибрам в науке и технике. Пршибрам (ЧССР), 1973, с. 103—162.
40. *Родионов Д. А.* Статистические методы разграничения геологических объектов по комплексу признаков. М., Недра, 1968.
41. *Себер Дж.* Линейный регрессионный анализ. М., Мир, 1980.
42. *Статистические* методы при геохимических поисках месторождений. М., ИМГРЭ, 1973.
43. *Тарасенко Ф. П.* Непараметрическая статистика. Томск, Изд-во Томск. ун-та, 1976.
44. *Agterberg F. P.* Geomathematics, Elsevier, 1974, The Netherlands.
45. *Puri M. L., Sen P. K.*
N.-Y, J. Wiley, 1971.
45. *Puri M. L., Sen P. K.* Nonparametric methods in Multivariate Analysis. N.-Y, J. Wiley, 1971.
46. *Tamura R.* Multivariate nonparametric several-sample tests. — Ann. Math. Statist., vol. 37, 1966, p. 611—618.
47. *Adichie J. N.* Estimates of regression parameters based on rank Test. — Ann. Math. Statist, vol. 38, 1967, p. 894—904.
48. *Jurečkova J.* Nonparametric estimation of regression coefficients. — Ann. Math. Statist, vol. 42, 1971, p. 1328—1338.
49. *Puri M. L., Sen P. K.* Distribution — Free Approaches to General Linear Models. — In: A survey of Statistical Design and Linear Models. Amsterdam—N.-Y., 1975, p. 459—474.

ОГЛАВЛЕНИЕ

Роберт Иосифович Коган
Юрий Павлович Белов
Дмитрий Алексеевич Родионов

СТАТИСТИЧЕСКИЕ РАНГОВЫЕ КРИТЕРИИ В ГЕОЛОГИИ

Редактор издательства *Л. М. Старикова*
Обложка художника *А. Е. Григорьева*
Художественный редактор *Е. Л. Юрковская*
Технический редактор *М. Е. Карева*
Корректор *А. А. Передерникова*
ИБ № 4894

Сдано в набор 10.01.83 Подписано в печать 19.05.83 Т-08482 Формат 60×90¹/₁₆ Бумага типографская № 2. Гарнитура «Литературная». Печать высокая. Усл.-п. л. 8,5. Усл. кр.-от. 8,88. Уч.-изд. л. 9,54. Тираж 3400 экз. Заказ 39/8883-14. Цена 50 коп.

Ордена «Знак Почета» издательство «Недра», 103633, Москва, К-12, Третьяковский проезд, 1/19

Московская типография № 6 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли.
109088, Москва, Ж-88, Южнопортовая ул., 24.

103

1

НЕДРА