У·КРАМБЕЙН
М·КАУФМЕН
Р·МАК-КЕММОН

МОДЕЛИ ГЕОЛОГИЧЕСКИХ ПРОЦЕССОВ


ИЗДАТЕЛЬСТВО
МИР

MODELS OF GEOLOGIC PROCESSES

An Introduction to Mathematical Geology

Convener
P. FENNER

Lecturers
W. C. KRUMBEIN,
M. E. KAUFFMAN,
R. B. McCAMMON

1969
American Geological Institute,
Washington, D. C.

У. КРАМБЕЙН,
М. КАУФМЕН,
Р. МАК-КЕММОН

МОДЕЛИ ГЕОЛОГИЧЕСКИХ ПРОЦЕССОВ

ВВЕДЕНИЕ
В МАТЕМАТИЧЕСКУЮ
ГЕОЛОГИЮ

Перевод с английского
Р. И. КОГАНА

Под редакцией и с предисловием
д-ра геол.-мин. наук
Д. А. РОДИОНОВА

ИЗДАТЕЛЬСТВО «МИР»
МОСКВА 1973

УДК 550. 550.8. 519.2

Книга представляет собой цикл лекций по применению математических методов в геологии, прочитанных в Американском геологическом институте в Вашингтоне для геологов. В отличие от ранее изданных в русском переводе работ Р. Миллера и Дж. Кана «Статистический анализ в геологических науках» («Мир», 1965) и У. Крамбейна и Ф. Грейбилла «Статистические модели в геологии» («Мир», 1969) эта книга содержит общие принципы подхода к построению моделей геологических процессов, т. е. рассмотрена динамика. В ней освещены наиболее часто употребляемые в геологии методы многомерной математической статистики, например метод главных компонент, и вопросы применения цепей Маркова в геологических экспериментах.

Книга рассчитана на широкий круг геологов различных специальностей.

Редакция литературы по вопросам геологических наук

К $\frac{0291-139}{041(01)-73}$

ПРЕДИСЛОВИЕ

Список трудов, посвященных вопросам применения математических методов в геологических исследованиях, к настоящему времени уже достаточно обширен. Однако большинство монографических работ в этой области как советских, так и зарубежных авторов написано весьма сложно для понимания геологами, не имеющими специальной математической подготовки. Кроме того, изучению геологических процессов с помощью математических методов в этих работах уделяется недостаточно внимания.

Предлагаемый читателям цикл лекций У. Крамбейна, М. Кауфмена и Р. Мак-Кеммона написан значительно проще, чем более ранние публикации, и доступен для широкого круга специалистов-геологов. Хотя цикл лекций и называется «Модели геологических процессов», его тематика не ограничивается только этими вопросами. В лекциях рассматривается ряд весьма полезных для геологической практики методов обработки информации, например таких, как дисперсионный анализ, метод главных компонент и др.

Необходимо отметить, что методу главных компонент в нашей литературе, посвященной применению математики в геологии, уделялось очень мало внимания. Между тем это весьма эффективное средство предварительной обработки результатов многомерных наблюдений с целью сокращения числа анализируемых характеристик. В данном цикле лекций метод главных компонент рассмотрен достаточно подробно для непосредственного практического применения в геологии.

Весьма детально освещены также вопросы классификации. Особую ценность при этом представляют способы построения дендрограмм и дендрографов, которые лишь затронуты в книге У. Крамбейна и Ф. Грейбилла «Статистические модели в геологии» (см. русский перевод, М., «Мир», 1969).

Все описанные в книге методы хорошо иллюстрированы примерами из геологической практики, что в значительной степени облегчает усвоение материала.

Этот цикл лекций будет полезен геологам самых разных специальностей как при непосредственном применении математических методов в геологии, так и для ознакомления с потенциальными возможностями математики при решении геологических задач.

Д. Родионов

ПРЕДИСЛОВИЕ К АНГЛИЙСКОМУ ИЗДАНИЮ

Этот краткий курс лекций является составной частью программы изучения количественной геологии, разработанной Советом по обучению геологов при Американском геологическом институте. В эту программу входит также ознакомление с аннотированной «Библиографией статистических приложений в геологии». Дж. Говарда, «Инструкцией по геоматематике» Р. Осборна, циклом лекций У. Фокса «Лабораторные работы с применением вычислительной техники для геологов и океанографов старших курсов», а также с кратким обзором Д. Мак-Интайра «Матричный анализ на кухне». Программа обучения завершилась симпозиумом по количественной геологии, который состоялся 10 ноября 1969 г. в Атлантик-Сити в Американском геологическом обществе. Труды симпозиума предполагается опубликовать отдельным изданием. Весь цикл обучения построен так, чтобы предшествующие разделы служили основой для последующих.

Несмотря на то что ни одна из лекций цикла обучения не претендует на исчерпывающую полноту, авторы попытались построить их так, чтобы отразить современное состояние по основным вопросам количественной геологии.

Эти лекции будут полезны студентам и преподавателям геологического профиля, которые в своей работе сталкиваются с вопросами количественных исследований.

П. Феннер

ЛЕКЦИЯ 1

Последовательное моделирование и функции распределения в геологии

У. Крамбейн

СХЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ГЕОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ. МОДЕЛЬ В ГЕОЛОГИИ

Основной вопрос начальной стадии геологического исследования — выявление геологической задачи. Сведения, необходимые для ее постановки, можно получить разными способами, а сама задача может оказаться как очень узкой и четко сформулированной, так и весьма широко и расплывчато поставленной. Обычно, выражаясь словами Чемберлена, геолог формулирует одну или несколько «рабочих гипотез» и выделяет типы наблюдений, позволяющие решить поставленную задачу, или по крайней мере выбирает среди множества альтернативных гипотез наиболее приемлемые.

Если воспользоваться более формальными определениями, которых мы будем придерживаться, то рабочие гипотезы можно рассматривать как «концептуальные (понятийные) модели», отвечающие мыслимым картинам изучаемого явления. Концептуальные модели могут содержать описания возможных причинно-следственных связей, помогающие установить зависимые и независимые переменные.

В простейшей форме модель может быть просто утверждением или диаграммой, позволяющей более наглядно представить наблюдаемые геологические факты или выявить зависимости между изучаемыми характеристиками. На этой стадии начинают сбор данных с целью общей оценки ситуации, формы представления, предпочтений, возможностей и пытаются построить формальную модель путем получения ответов на ряд специальных вопросов относительно собранных данных. Иногда ответы на эти вопросы получают путем прямых наблюдений, как, например, об относительном возрасте двух даек по характеру их пересечения. Иногда же для ответа на поставленный вопрос необходима проверка статистической гипотезы, что в свою очередь приводит к построению формальной статистической модели, а для этого необходимо определить генеральную совокупность, выборочный план и др.

При решении некоторых геологических задач можно воспользоваться дифференциальными уравнениями; граничные условия, константы и параметры выбираются в соответствии с принятой теорией или особенностями результатов наблюдений. Такой подход, естественно, приводит к детерминированным моделям,

предсказание на основании которых оказывается точным, если модель верна. В некоторых случаях модель, даже если она и удовлетворительна, не обеспечивает точного предсказания. В подобной ситуации используются модели стохастических процессов, в основе которых лежит аппарат теории вероятностей, учитывающий различные виды случайных флуктуаций изучаемого процесса.

Формально модель можно рассматривать как схему, отражающую структуру наблюдаемых данных, но так, что она позволяет получать ответы на поставленные вопросы. Очень важно учитывать различия между понятиями «то, что требуется доказать» и «поставленный вопрос». В современном исследовании задача обычно заключается в проверке пригодности модели для описания или объяснения какого-либо реального геологического явления по результатам наблюдения. Мы используем модель для проверки обоснованности наших интуитивных представлений или теоретических предпосылок, на основании которых она построена.

Очень редко модель в своей исходной форме достаточно хорошо отражает реальную геологическую обстановку. Обычно необходимы некоторые усовершенствования для эффективного ее применения при проверке гипотез. Эти усовершенствования могут заключаться в учете более сложных особенностей изучаемого явления или в переходе к более высокому уровню представления структуры наблюдаемых данных. Таким образом, моделирование в геологии представляет собой последовательный процесс, на каждой стадии которого проверяется пригодность модели путем установления ее соответствия заданным требованиям.

Требования, которым должна удовлетворять модель, могут быть как весьма общими, так и очень специфичными. Так, например, если нужно проверить предположение, что средний размер галек убывает вниз по течению реки по экспоненциальному закону, для этого необходимо отобрать некоторое количество галек, измерить их и затем нанести полученные результаты на полулогарифмическую бумагу. Если расположение точек окажется близким к прямой линии, модель можно считать удовлетворительной даже при столь грубом способе представления данных. Можно пойти и дальше, приняв, что прямая линия соответствует функции $Y = Y_0 e^{-aX}$, где Y — средний размер галек в любой точке, X — характеризует положение вдоль по течению реки, Y_0 — исходный размер в начальной точке $X = 0$, a — константа, пропорциональная скорости убывания размеров, которая представляет собой обычную производную от выражения, описывающего данную модель, т. е. $dY/dX = -aX$. Таким образом, скорость убывания размера галек представляет собой функцию от самого размера, т. е. чем больше галька, тем быстрее она отложится в русле. Это так называемый закон Штернберга, сформулированный еще в 1875 г.

Впоследствии было установлено, что селективная транспортировка также является фактором, влияющим на наблюдаемое

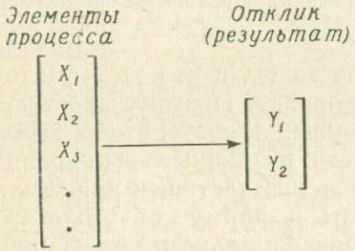
Таблица 1

Классификация геологических моделей

1. «*Концептуальная (понятийная) модель*». «Мысленная картина» геологического процесса или явления. Она может быть количественной, качественной или в виде диаграммы

2. *Диаграмма.* Схема, таблица или график, показывающие основные зависимости процесса или явления. Например, на приведенном ниже рисунке изображена гипотетическая модель соотношения источников сноса

3. *Модель типа процесс—отклик (результат).* Элементы процесса X (слева) и его результаты Y (справа) представляются в виде матриц



4. *Статистическая модель прогноза.* Для ее построения обычно используется метод наименьших квадратов. В приведенном ниже уравнении Y — величина, значения которой предсказываются по одной или более переменным X:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots$$

5. *Детерминированная модель.* Зависимость между функцией и аргументом обычно представляется в виде дифференциального уравнения, которое позволяет точно предсказывать значения Y по значениям X

6. *Стохастическая модель.* Вероятностная модель, в которой предсказание основано на условных вероятностях множества возможных исходов, например матрица вероятностей перехода в цепи Маркова

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Таблица 2

Схематическое представление этапов геологического исследования

Возвращение к соответствующей стадии

1. Выявление и постановка геологической задачи
2. Построение «концептуальной (понятийной) модели»
3. Выбор переменных, учитываемых в модели, а также определение функций и аргументов
4. Сбор и анализ данных в соответствии с требованиями выбранной модели
5. Выбор наиболее важных переменных (если это не было сделано теоретически в пункте 3)
6. Уточнение модели путем перевода ее в детерминированную, стохастическую или регрессионную (предсказывающую)
7. Применение уточненной модели для предсказания новых зависимостей, которые можно проверить при помощи новых видов наблюдений
8. Принятие, отклонение или изменение окончательной модели

уменьшение размеров галек, и что этот фактор может быть более существенным, чем абразия. Поэтому простую детерминированную модель следует усовершенствовать, чтобы учесть влияние этих двух факторов. Если учесть различные гидродинамические факторы, контролирующие движение частиц, и принять во внимание литологическую изменчивость и форму обломков, то станет очевидным, что построение дифференциального уравнения, описывающего распространенность частиц и их селективную транспортировку, — задача довольно сложная.

Этот пример наглядно показывает, что модель быстро усложняется по мере введения в нее дополнительных переменных. Более того, с увеличением числа переменных причинно-следственные зависимости становятся менее отчетливыми за счет зависимостей между исходными аргументами. Это не означает, что какие-то результаты процесса перестают определяться своими контролирующими факторами, а только указывает на то, что эти факторы значительно труднее «рассортировать», поставив их в точное соответствие наблюдаемым результатам.

В данной ситуации необходимо ввести в модель стохастические элементы. Если нельзя сказать, что данному значению X соответствует значение Y с вероятностью $p=1$, а вместо этого приходится говорить, что данному значению X соответствует значение Y с вероятностью $0<p<1$, то модель перестает быть детерминированной и становится стохастической.

Подробно на стохастических моделях мы остановимся позднее, а здесь необходимо лишь подчеркнуть, что при последовательном моделировании модель не только усложняется, но и меняется ее положение в множестве переходных моделей от полностью детерминированных до чисто случайных.

В табл. 1 и 2 приведены типы моделей и этапы их построения в представлении автора. Табл. 1 содержит перечень некоторых моделей, обычно использующихся в геологии, а табл. 2 — некоторые последовательные стадии геологического моделирования. Эта последовательность исходит из понятия простой концептуальной модели или рабочей гипотезы. Если одновременно возникает несколько таких гипотез, то их комбинация образует множество рабочих гипотез, из которых затем производится соответствующий выбор.

ИЗМЕРЕНИЯ В ГЕОЛОГИИ

Прежде чем применять модели в геологии, необходимо определить возможность их построения по наблюдаемым данным. На самом процессе измерения я остановлюсь весьма кратко. Возрастание роли вероятностных моделей в геологии приводит к необходимости выделения трех видов случайной изменчивости. Два из них при желании можно исключить из рассмотрения, тогда как третий имеет геологическое значение и может быть использован в большинстве геологических исследований независимо от того, рассматривается детерминированная модель процесса или стохастическая.

Существенная особенность числовых измерений заключается в том, что два измерения одного и того же свойства объекта одинаковы по точности только в том случае, когда результаты выражены одним и тем же достаточным числом десятичных знаков. Для единичного объекта в условиях данной процедуры измерения расхождения между такими результатами отражают ошибку измерения, представляемую как случайная компонента, которая присутствует во всех наблюдениях. Это первый источник изменчивости. Вторая важная особенность заключается в том, что при достаточной точности измерения две выборки объектов из данной совокупности будут характеризоваться различающимися выборочными характеристиками неизвестных параметров. Это так называемая ошибка опробования, и ее эффект может быть довольно большим. Она также рассматривается как случайная компонента.

Ошибку измерения можно уменьшить путем повышения точности процесса измерения, а ошибку опробования с помощью увеличения объема выборки и подбора соответствующей процедуры рандомизации. Третий источник изменчивости связан с различиями между геологическими объектами. Примером может служить изменчивость размеров частиц в осадочных образованиях.

Матрица вероятностей перехода выражается в том, что не все зерна, отложившиеся в данное время в одних и тех же условиях, имеют одинаковые размеры, форму и удельный вес. Эта «присущая» изменчивость результатов наблюдений реальна в том смысле, что она входит как элемент в комплекс процесс—отклик, действующий при образовании осадка. Мы не можем отнести каждую отдельную частицу к конкретному микрособытию, а вместо этого строим распределение размеров частиц и ставим его параметры в соответствие данному комплексу условий. Полученная при этом гистограмма характеризует распределение вероятностей, а определяемые значения диаметра каждой отдельной частицы рассматриваются как значения случайной величины в математическом смысле.

При первоначальном рассмотрении эту случайную компоненту можно не учитывать за счет усреднения, так как в виде характеристики пробы осадочной породы используется среднее значение. Однако в дальнейшем распределение размеров частиц следует изучать и связывать с процессом формирования осадка. Если предметом исследования являются зерна осадочных пород, свойства рельефа местности, мощности стратиграфических слоев и др., то возникает задача изучения распределения по результатам наблюдения в выборке, что приводит к понятию модели совокупности.

МОДЕЛИ СОВОКУПНОСТЕЙ В ГЕОЛОГИИ

Практически все виды исследований во всех отраслях геологии, связанные с изучением каких-либо явлений с помощью опробования, приводят в конечном итоге к распределению того или иного вида. Эти распределения эмпирические или выборочные, и одна из важных задач геологического анализа заключается в выборе для них соответствующих теоретических функций распределения, что обычно делается с помощью графических методов на вероятностной бумаге или путем проверки пригодности некоторой гипотетической функции распределения для описания выборочных данных с помощью критерия хи-квадрат. Обе процедуры эмпирические; пользуются ими при отсутствии каких-либо теоретических обоснований для априорного задания модели распределения. Но и в этом случае проводится соответствующая проверка пригодности выбранной теоретической модели.

Ниже кратко рассмотрены некоторые модели распределений, обычно использующиеся в геологических исследованиях, а также наиболее важные модели, применяемые в других областях науки. В табл. 3 приведен перечень функций плотности вероятности этих распределений. Некоторые виды распределений будут рассматриваться в последующих лекциях этого краткого курса, и их список будет дополнен.

Таблица 3

Краткий список некоторых моделей распределения, применяемых в геологии

Левая часть уравнения представляет собой обозначение функции плотности. Первая буква в скобках — аргумент функции; он отделен точкой с запятой от последующих букв, которыми обозначены параметры распределения. Подробное описание свойств этих распределений приведено в работах Муда и Грейбилла [68], Крамбейна и Грейбилла [50] и Гриффитса [29].

Дискретные распределения

Пуассона

$$P(x;\ \lambda) = \lambda^x e^{-\lambda}/x!, \quad x = 0,\ 1,\ 2, \ldots,\ \lambda > 0$$

Биномиальное

$$B(x;\ p) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{n-x}, \quad x = 0,\ 1,\ 2, \ldots, n$$

Геометрическое

$$Ge(x;\ p) = p(1-p)^{k-1}, \quad k = 1,\ 2,\ 3, \ldots, 0 < p < 1$$

Непрерывные распределения

Нормальное

$$N(x;\ \mu,\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Логнормальное

$$L(x;\ \mu_L,\ \sigma_L) = \frac{1}{x\sigma_L\sqrt{2\pi}}\, e^{-\frac{(\log x - \mu_L)^2}{2\sigma_L^2}} \qquad 0 < x < \infty$$

Гамма-распределение

$$G(x;\ r,\ \beta) = \frac{x^{r-1}\, e^{-\frac{x}{\beta}}}{\Gamma(r)\,\beta^r}, \quad x \geqslant 0$$

Экспоненциальное

$$Z = (x;\ \lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

Круговое нормальное

$$C(x;\ 0,\ \gamma) = k e^{\theta \cos(x-\gamma)}, \quad -\pi \leqslant x \leqslant \pi,$$

где $k = f(\theta)$ — такая функция, что ограниченная ею площадь равна 1.

В некоторых случаях вид плотности вероятности можно заранее предсказать, исходя из особенностей результатов измерения. Так, например, распределение числа зерен акцессорных минералов, когда компоненты относительно редки, будет близко к закону Пуассона, а распределение числа зерен более распространенных минералов будет приближаться к биномиальному. Обе эти модели дискретных распределений приведены в табл. 3. Третье дискретное распределение, геометрическое, связано с моделированием последовательностей независимых событий, а также с цепями Маркова с дискретным временем, которые будут рассмотрены ниже.

Непрерывные модели используются для описания распределений размеров частиц, формы, ориентации, мощности слоев и других свойств, результаты измерения которых непрерывны. Наиболее широко распространенная модель распределения в геологии — это нормальный закон, лежащий в основе многих статистических критериев. Заметим, что большинство статистических критериев являются равномерно наиболее мощными только в условиях распределений, близких к симметричным.

В некоторых случаях, когда распределения отличаются значительно выраженной асимметрией, можно подобрать такое преобразование наблюдаемых значений X, что полученное новое распределение будет близко к нормальному. Так, например, широко используется преобразование типа $\varphi = -\log_2 X$, обеспечивающее переход от логнормального распределения к нормальному. Более общий случай представляет собой простое преобразование $L = \log X$ при любом основании, что показано в табл. 3. Логнормальная модель применяется для описания распределений размеров зерен, мощностей слоев, длин русл первого порядка, содержаний редких элементов в породах и др. Даже в тех случаях, когда «истинные» распределения отличаются от логнормального, логарифмическое преобразование создает хорошие условия для проведения статистического анализа.

Гамма-распределение мало используется в геологии, хотя оно могло бы применяться шире, так как пригодность лежащей в его основе вероятностной модели для ряда геологических ситуаций очевидна. Так, в частности, его можно использовать для описания распределений длин русл гидросети вместо логнормальной модели. Вместо логнормального нередко можно привлекать некоторые гамма-распределения, причем выбор модели зависит от теоретических или интуитивных предпосылок об условиях возникновения функции данного вида.

Экспоненциальное распределение является непрерывным аналогом дискретного геометрического распределения и применяется в условиях экспериментов, моделями которых могут служить цепи Маркова с непрерывным временем. Если в гамма-распределении параметр r приравнять к 1, то соответствующий параметр экспо-

ненциальной плотности λ будет $\lambda = 1/\beta$. Эта зависимость также используется в вероятностных моделях.

Круговое нормальное распределение широко применяется при изучении «векторных характеристик» осадочных пород. На практике обычно используются векторное среднее и стандартное отклонения относительно этого среднего. Иногда представляется более целесообразным пользоваться непосредственно оценками для γ и θ в тех случаях, когда θ является лучшей мерой рассеяния, чем σ.

ЗНАЧЕНИЕ МОДЕЛЕЙ СОВОКУПНОСТЕЙ

Как уже отмечалось выше, в дальнейшем потребуется более детальное рассмотрение функций распределения, чем это сделано здесь, особенно с точки зрения теоретических построений моделей изучаемых геологических процессов. Читателю можно рекомендовать работу Гриффитса [29], в которой подробно рассмотрены условия возникновения логнормального распределения. Для геометрического распределения аналогичные условия нетрудно вывести, что будет рассмотрено в лекции 2, в которой детально описано применение геометрического распределения в некоторых видах моделей стохастических процессов.

ЛЕКЦИЯ 2

Детерминированные и вероятностные модели в геологии

У. Крамбейн

ВВЕДЕНИЕ

Нередко научное исследование представляет собой попытку «объяснения» явления с позиций возможности точного предсказания каждой стадии процесса (состояния системы) во времени или в пространстве. Классическим примером такого подхода является математическая физика, которая, базируясь на принятых постулатах, строит функциональные зависимости так, что каждый отклик (результат) процесса неодинаково связан с некоторыми контролирующими факторами, а ошибки измерения пренебрежимо малы.

Обычно при классическом подходе для выражения скорости изменения функции относительно изменений одного или нескольких аргументов используют одно или несколько дифференциальных уравнений при заданных граничных условиях и на основе наблюдаемых или теоретически рассчитанных данных получают различные характеристики. Конечный результат представляет собой зависимость типа $Y=f_Y(X_1, X_2, \ldots)$, которая позволяет точно предсказывать значения Y по заданным значениям $X_1, X_2, \ldots$.

В тех случаях, когда схему процесса легко представить, а число рассматриваемых переменных не очень велико, этот подход бывает весьма эффективным. Многие геологические задачи можно разделить на части, для каждой из которых такой подход будет приемлем, но, как только начинает возрастать число рассматриваемых переменных и уменьшается контроль человека над аргументами, действуют многочисленные усложняющие факторы. Среди этих факторов существенное влияние оказывают зависимости между аргументами, которые приводят к нарушению классического «причинно-следственного» подхода. Это значит, что с введением новых переменных некоторые исходные аргументы могут перестать быть таковыми и появятся новые, занимающие промежуточное положение.

На фиг. 1 изображен относительно простой случай, когда береговые течения и средний размер обломков занимают промежуточное положение как контролирующие факторы между процессами, оказывающими воздействие на берег, и соответствующим откликом (результатом) — углом наклона прибрежной части дна.

Этот отклик имеет обратную связь с основными характеристиками волн, в частности с углом подхода волны в прибрежной зоне и их ударным воздействием. Представленная на фиг. 1 система в значительной степени упрощена по сравнению с реальным процессом формирования угла наклона дна в прибрежной зоне, но тем не менее она наглядно показывает сложность задачи даже фрагментарного геологического исследования.

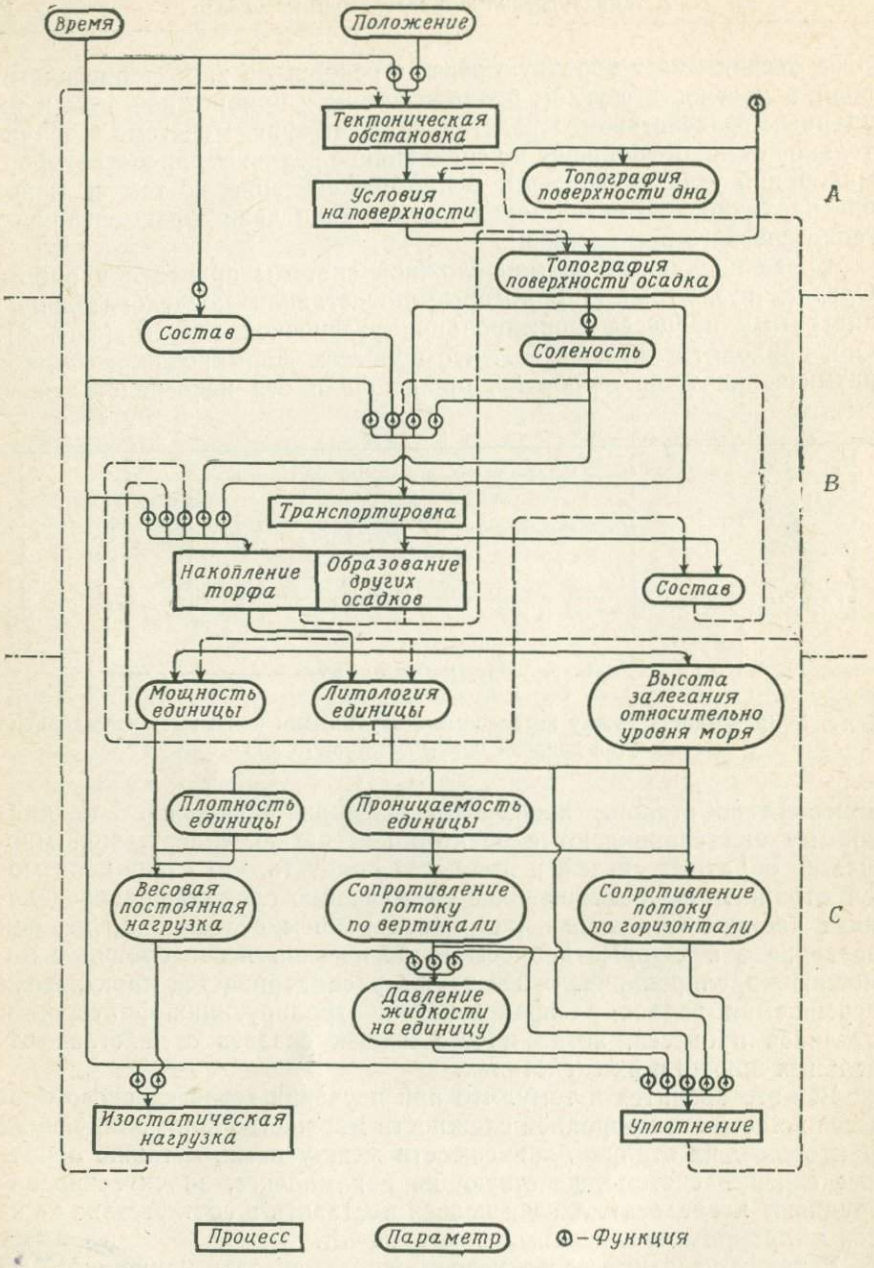
Более наглядный пример сложной системы приведен в работе Ортила и Уолтона [70], которые представили последовательную диаграмму процесса формирования угленосной дельты (фиг. 2). Они приходят к выводу, «... что процессы формирования осадков ритмических толщ, вероятно, происходили без непосредственного

Фиг. 1. Зависимость между переменными, изучаемыми в процессе исследования угла наклона прибрежной части дна.

контроля со стороны каких-либо внешних факторов». Эту диаграмму они сравнивают с электронной цепью, содержащей множество обратных связей, и пытаются показать, что такие цепи могут отражать свои изменения в «необычайно сложных распределениях частот». Отдельные части этой работы особенно ценны как введение в некоторые вопросы геологических исследований с помощью моделирования. В ней также рассматривается практически нерешаемая задача распределения контролирующих факторов и откликов процесса, когда ничего нельзя сказать о действии отдельных причин и их следствиях.

Все это сводится к тому, что при изучении геологического процесса во всей его природной сложности множество контролирующих факторов и их отклики, зависимости между ними, а также обратные связи в системе, действующие в комплексе, значительно затрудняют исследование, если нельзя поставить в соответствие каждой микропричине отдельный микроэффект.

В таких условиях часто невозможно высказать какие-либо утверждения об отдельных факторах, но вполне реальны конкретные заключения об их группах. Уотсон [95] рассматривает этот вопрос детально и определяет момент перехода от точной детерминированной постановки задачи к вероятностной.



Фиг. 2. Диаграмма процесса образования угленосной дельты.

Непрерывные линии отражают прямую зависимость, пунктирные — обратную связь. Термин «единица» соответствует призме горизонтально лежащего слоя осадка в условиях выбранной сети координат. *A* — географические условия; *B* — транспортировка, осадконакопление и эрозия; *C* — уплотнение и изостатическое воздействие.

После того как в рассмотрение вводятся вероятностные характеристики, меняются и особенности задачи исследования (обычно относительно аргументов) от анализа неслучайных величин к изучению случайных, а модель, естественно, изменяется от детерминированной до стохастической, которая, однако, может содержать детерминированные компоненты. В стохастической модели результату эксперимента соответствуют вероятности, связанные с множеством возможных событий, и изучаемый процесс рассматривается как случайный.

СЛУЧАЙНЫЕ ПРОЦЕССЫ В ГЕОЛОГИИ

Прежде чем перейти к рассмотрению случайных процессов, полезно остановиться на некоторых семантических трудностях, связанных с точным пониманием этого термина. Такой процесс многие геологи рассматривают как случайную, неорганизованную и непредсказуемую последовательность событий. Частично столь неправильное понимание возникает обычно из-за того, что не учитывается такой факт, что применение случайной величины в научных исследованиях также обоснованно, как и нестохастических переменных. Основное различие здесь заключается в том, что появление отдельного события нельзя предсказать с вероятностью, равной 1, и она принимает значения в интервале от 0 до 1. Тем не менее с помощью соответствующих методов можно оценить вероятность, а в некоторых случаях вывести ее значения теоретическим путем.

Вторая трудность, связанная с применением моделей случайных процессов в геологии, возникает благодаря мнению, что в случайных процессах все связи с более ранними, предшествующими «причинными» событиями каким-то образом утрачены, а это приводит к нарушению цепи процесс—отклик. Некоторые случайные процессы являются последовательностями независимых событий, как, например, бросание монеты или игральной кости, когда результат $(n-1)$-го бросания не оказывает влияния на исход n-го бросания. С другой стороны, некоторые случайные процессы характеризуются наличием зависимости последующих событий от предыдущих, и к процессам такого типа относятся цепи Маркова.

Выше было введено понятие частичной зависимости, когда одна и та же переменная может быть одновременно как причиной, так и следствием. В подобной ситуации появление отдельного события, связанного с этой переменной в момент времени t_n, может зависеть от появления или непоявления некоторого другого события в момент t_{n-1}. Такая схема приводит к понятию условной вероятности, которую также можно оценить (или иногда вывести теоретически) с помощью соответствующих методов. Один из методов изучения условных вероятностей — построение вероятностного дерева, пример которого будет приведен ниже.

Так как сами вероятности могут быть составной частью сложных методов, можно выделить полный набор моделей, начиная от чисто детерминированных, обеспечивающих точное предсказание (с вероятностью $p=1$), до моделей, полностью состоящих из независимых случайных событий, в которых отсутствуют какие-либо связи типа цепей.

В предыдущей лекции был приведен пример полностью детерминированной модели $Y=Y_0e^{-aX}$. Примером выражения модели, основанной на полностью независимых случайных событиях, может служить

$$P\{j,\ t_{n+1}\,|\,i,\ t_n\}=p_j.$$

Это выражение показывает, что вероятность события, заключающегося в том, что система в момент t_{n+1} будет находиться в состоянии j, при условии, что в момент t_n она находилась в состоянии i, равна p_j, где p_j — безусловная вероятность появления состояния j.

Если предположить, что процесс представляет собой цепь Маркова первого порядка, то вероятность появления состояния j в момент t_{n+1} будет зависеть от состояния, в котором находилась система в момент t_n. Тогда

$$P\{j,\ t_{n+1}\,|\,i,\ t_n\}=p_{ij},$$

где p_{ij} — условная вероятность, зависящая от предшествующего состояния системы. Число p_{ij} представляет собой вероятность перехода системы из состояния i в состояние j.

Последовательность независимых случайных событий обычно называют процессом без памяти, так как предшествующие события не влияют на появление последующих, а цепи Маркова — процессом с короткой памятью, так как в них только одно предшествующее состояние оказывает влияние на последующее. Аналогично детерминированный процесс можно назвать процессом с долгой памятью, так как в нем все предшествующие состояния, начиная с момента t_0, оказывают влияние на последующие, вплоть до момента t_{n+1}.

МОДЕЛИ СТОХАСТИЧЕСКИХ ПРОЦЕССОВ И ПАРА ПРОЦЕСС — ОТКЛИК

Какие бы соотношения ни наблюдались между предшествующими и последующими состояниями случайного процесса, они не дают сведений о наличии или отсутствии пары процесс—отклик. Дело в том, что данный микроэффект может быть следствием более чем одной микропричины, что исключает однозначное соответствие конкретного события и отдельного фактора. Рассмотрим, как начинается раздвоение верхней части растущего русла первого

порядка. Оно может быть вызвано вывороченным валуном, упавшим деревом, размывом звериной тропы; любое из этих событий может создать условия для раздвоения русла. В течение многих лет, даже столетия, возможно, не произойдет каких-либо существенных событий. Так же как и появление микропричины, событие, заключающееся в раздвоении русла, можно рассматривать как случайное. Если бы у нас было много данных о таких раздвоениях русл, соответствующих различным видам микрособытий, мы получили бы соответствующие вероятности, которые затем использовали бы для оценки правдоподобия предположения, что та или иная микропричина является основной. Таким образом, мы видим, что никакой из «законов природы» не нарушен применением модели случайного процесса — реки по-прежнему текут вниз, а овраги рассекают их берега. Все дело в том, что отдельную микропричину нельзя однозначно поставить в соответствие наблюдаемому событию. В данном случае каждая причина одинаково возможна и равноправна в их множестве.

Некоторые геологи (в том числе и я) считают, что в природе отдельные события происходят без влияния внешних микропричин, как, например, формирование ритмических осадочных толщ, которые можно рассматривать как системы с внутренними обратными связями. Более того, известны примеры полной пригодности моделей, основанных на независимых случайных событиях, для изучения геологических процессов, что позволяет провести весьма строгий анализ, даже более строгий, чем в условиях соответствующей детерминированной модели. Концепция Шрива [84, 85] бесконечной случайной гидросети дала рациональную теоретическую основу для некоторых законов Гортона, особенно для изучения числа русл. Работы Шрива — хороший пример применения методов комбинаторной математики при решении геологических задач.

В качестве примера стохастического процесса с короткой памятью рассмотрим применение модели марковского процесса в стратиграфическом анализе.

СХЕМЫ СТРАТИГРАФИЧЕСКОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Вполне очевидно, что для ритмических осадочных толщ можно построить полную схему последовательности образования пород, но нельзя точно предсказать, какая литологическая разновидность будет следовать за другой в разрезе. Как отмечал Вистелиус еще в 1949 г., этот тип моделирования приводит к использованию моделей марковских процессов. Детальное изучение стратиграфами ритмических последовательностей было начато в середине 60-х годов, и в настоящее время насчитывается более 25 работ, связанных с этим вопросом. Большинство из них, учитывая также другие работы по применению цепей Маркова в геологии, принадлежит Крамбейну и Дейси [53].

В стратиграфии использовалось несколько вариантов цепей Маркова, но здесь мы ограничимся только регулярной цепью, построенной на основе литологических наблюдений в коротких фиксированных вертикальных интервалах, расположенных сверху вниз по разрезу. В результате этих наблюдений была получена матрица переходных вероятностей [62], которую можно использовать при моделировании, а также и для того, чтобы проиллюстрировать более ранние толкования [95] частного события, противопоставленного группе событий.

Приведенная в табл. 1 матрица Маркова построена на основе результатов 981 наблюдения микрослоистости в двухфутовых интервалах разреза формации Офисина (миоцен) восточной Венесуэлы [86]. Осадочный цикл образуют четыре компонента (песчаник, сланец, глина и лигнит), обозначенные в таблице буквами от *A* до *D* и рассматриваемые как состояния системы. Мы сосредоточим внимание на лигните, который образует хорошо выраженные тонкие прослои. Наша задача заключается в рассмотрении некоторых аспектов вероятностного предсказания в сравнении с предсказанием, базирующимся на детерминированных моделях.

Таблица 1

Матрица переходных вероятностей разреза формации Офисина (миоцен), Венесуэла

		A	*B*	*C*	*D*
A — песчаник	*A*	0,79	0,07	0,07	0,07
B — сланец	*B*	0,05	0,79	0,06	0,10
C — глина	*C*	0,10	0,32	0,43	0,15
D — лигнит	*D*	0,18	0,39	0,13	0,30

Примечание. Для того чтобы воспользоваться матрицей переходных вероятностей, следует найти соответствующую строку, например *D*, и определить вероятность того, что произойдет переход системы из состояния *D* в какое-либо другое. Так, например, если система находится в состоянии *D*, то вероятность того, что система останется в этом состоянии, будет 1,00 — 0,30 = 0,70. Вероятность перехода в какое-либо конкретное состояние зависит от вероятностей, приведенных в соответствующих столбцах матрицы. Таким образом, вероятность перехода системы в состояние *A* будет равна 0,18, а в состояние *B* — 0,39.

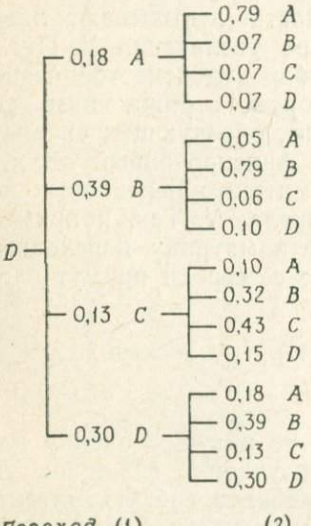
Поставим вопрос следующим образом: если в результате первого наблюдения зафиксировано состояние *D* (лигнит) и прослежен процесс, в котором переход (т. е. простейший элемент осадкообразования) происходит в каждой точке последовательности, то каковы вероятности того, что в конце данного множества точек лигнит снова будет следовать за глиной, за сланцем, за песчаником?

Если допустить, что матрица переходных вероятностей содержит не оценки, а «истинные» значения, то мы сможем предска-

зать рассматриваемое множество событий, причем это предсказание будет также научно обоснованным при вероятности P_i (где $0<P_i<1$, $\Sigma P_i=1$), как и предсказание с вероятностью $P=1$ в условиях соответствующей детерминированной модели.

ВЕРОЯТНОСТНОЕ ДЕРЕВО ЦЕПИ МАРКОВА

Для построения вероятностного дерева цепи Маркова необходимо, начиная с исходного состояния (в нашем случае лигнит — *D*), построить диаграмму, содержащую все возможные пары, которые образует *D* с другими состояниями с соответствующими им вероятностями первого перехода. Затем строятся пары, образованные каждым возможным состоянием после первого перехода и всеми возможными состояниями. Соответствующие этим парам вероятности будут показывать возможность второго перехода. Эта часть вероятностного дерева построена по матрице, приведенной в табл. 1, и изображена на фиг. 3.



Фиг. 3. Вероятностное дерево с двумя переходами, построенное по данным табл. 1.

Шестнадцать возможных исходов, соответствующих первому и второму переходам

(Вероятности взяты из табл. 1)

Исход	Вероятность
DAA	0,1422
DAB	0,0126
DAC	0,0126
DAD	0,0126
DBA	0,0195
DBB	0,3081
DBC	0,0234
DBD	0,0390
DCA	0,0130
DCB	0,0416
DCC	0,0559
DCD	0,0195
DDA	0,0540
DDB	0,1170
DDC	0,0390
DDD	0,0900
	Сумма 1,0000

Полученные пары кодируются, как это показано внизу фиг. 3, и вычисляются вероятности каждого возможного исхода путем перемножения вероятностей, соответствующих данной паре. Нетрудно видеть, что сумма для всех пар равна 1. Теперь можно сформулировать ряд вопросов и получить на них ответы. Какова вероятность того, что лигнит останется как состояние системы

в результате двух шагов, т. е. перехода в другие состояния не произойдет? Такой исход обозначен символом *DDD* и ему соответствует вероятность 0,0900. Какова вероятность того, что в рассматриваемых тройках состояний, соответствующих двум переходам, будет наблюдаться последовательность лигнит, глина, лигнит. На фиг. 3 эта последовательность обозначена как *DCD* и ей соответствует вероятность 0,0195. Какова вероятность события, заключающегося в том, что в результате первого перехода система будет находиться в состоянии *A* (песчаник)? Этому событию соответствуют четыре последовательности *DAA*, *DAB*, *DAC* и *DAD*, сумма вероятностей которых равна 0,1800.

Вероятностное дерево только с двумя переходами устроено довольно просто, но оно быстро усложняется и число возможных исходов быстро возрастает с увеличением числа шагов в системе. Однако если для построения такого дерева использовать ЭВМ, то вопросы, аналогичные сформулированным выше, можно поставить для любого числа *N* переходов. Весьма интересно наблюдать, как при достаточно больших *N* дерево достигает «устойчивого состояния», или «равновесия», когда вероятности испытывают лишь очень незначительные изменения по мере увеличения *N*. После того как такое состояние достигнуто, набор вероятностей становится практически фиксированным вектором, отражающим «устойчивые пропорции» различных компонент, образующих систему. С помощью ЭВМ легко получить такой фиксированный вероятностный вектор простым вычислением матрицы переходных вероятностей для достаточно большого числа *N*. Так, например, если на основе данных табл. 1 рассчитать матрицу переходных вероятностей для сорокового шага, то все ее строки примут одно и то же фиксированное значение:

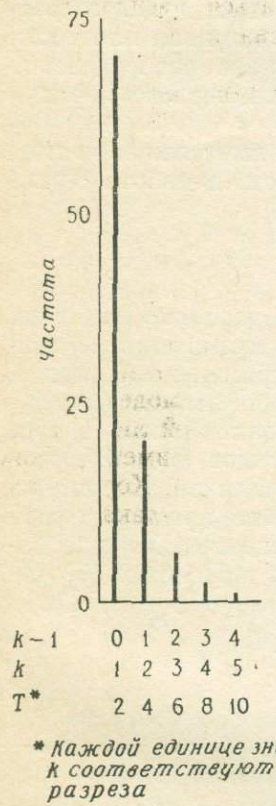
[0,27	0,49	0,12	0,12]
A	*B*	*C*	*D*

Если эти числа мы умножим на 100, то получим оценку для процентных соотношений компонентов в изучаемой совокупности.

Цепь Маркова представляет собой наиболее гибкую математическую модель, число областей применения которой в геологии будет все время увеличиваться.

При использовании таких моделей очень важно, чтобы изучаемое явление удовлетворяло тем же условиям, которые сформулированы для модели. Так, например, если в качестве модели используется цепь Маркова, то необходимо проверить, обладают ли наблюдаемые данные свойствами цепи Маркова [52]. При этом, чтобы удостовериться в применимости модели, следует убедиться, что вводимые в машину результаты наблюдения имеют то же самое распределение, что и результаты моделирования. Таким образом, если при изучении мощностей литологических единиц стра-

тиграфического разреза подходить к применению модели марковского процесса строго, необходимо проверить, согласуются ли их распределения с экспоненциальным законом (а в дискретном случае с геометрическим распределением). В настоящее время существует ряд факторов, свидетельствующих в пользу принятия в этих случаях логнормального распределения в качестве модели, наряду с экспоненциальным. В заключение покажем, что распределение мощности любого литологически однородного образования в условиях цепи Маркова будет подчиняться геометрическому закону в дискретном представлении.



Фиг. 4. Теоретическое распределение вероятностей для лигнита (состояние D), построенное по данным табл. 1.

Рассмотрим состояние D, вернувшись к табл. 1, где $P_{DD}=0{,}30$, а $(1-P_{DD})=0{,}70$. Напишем на карточках числа в интервале от 01 до 30, причем каждое из них будет обозначать, что система находится в состоянии D, тогда как числа в интервале от 31 до 100 — что система находится в каком-либо другом состоянии; произведем выборку с возвращением. Если наблюдение над одной карточкой рассматривать как переход, то появление числа, меньшего или равного 30, будет означать, что мощность лигнита увеличилась на некоторую постоянную величину, а числа, превышающего 30, — отсутствие приращения мощности. Будем рассматривать как «успех» событие, заключающееся в том, что система остается в состоянии D. Тогда k-му переходу должны предшествовать $k-1$ переходов, сохранивших состояние D.

Вычисление теоретических частот мощности лигнита при $n=100$

$k-1$	k	Вероятность $\times$ 100	
0	1	$(0{,}70)\ (0{,}30)^0$	70,0
1	2	$(0{,}70)\ (0{,}30)^1$	21,0
2	3	$(0{,}70)\ (0{,}30)^2$	6,3
3	4	$(0{,}70)\ (0{,}30)^3$	1,9
4	5	$(0{,}70)\ (0{,}30)^4$	0,6
	>5		0,2
			100,0

(Среднее значение $\bar{k}=1/\theta=1{,}43$; для $\bar{T}=2\bar{k}=2{,}86$ фута)

Теперь мы можем формально представить этот процесс, обозначив через P вероятность того, что мощность лигнита будет точно равна k единицам. В результате получим выражение

$$P\{T=k\}=(1-P_{DD})\,P_{DD}^{k-1}=(0{,}70)\,(0{,}30)^{k-1}.$$

Если $(1-P_{DD})$ обозначить через θ, то выражение примет следующий вид:

$$P\{T=k\}=\theta\,(1-\theta)^{k-1},$$

что является удобной формой представления геометрического распределения. На фиг. 4 приведена теоретическая кривая плотности вероятности, с которой должно согласовываться распределение мощности лигнита в условиях цепи Маркова первого порядка с дискретным временем.

На этой же фигуре приведена процедура вычисления частот. Формальное доказательство этого уравнения см. в работе Крамбейна и Дейси [53].

ЗАКЛЮЧЕНИЕ

В своих лекциях я не останавливался на рассмотрении ряда удобных статистических моделей, в настоящее время широко применяющихся в геологии, на таких, как регрессионный анализ, факторный анализ и др. В последующих лекциях эти вопросы будут подробно рассмотрены. Моя же задача заключалась в том, чтобы дать основу для изучения различных видов моделей, описание которых постоянно встречается в геологической литературе. Необходимо особо отметить, что моделирование применимо как к случайным, так и к детерминированным процессам. Хорошо иллюстрирует эту применимость работа Бриггса и Поллака [6], посвященная изучению эвапоритов.

ЛЕКЦИЯ 3

О количественных исследованиях в геологии

М. Кауфмен

ВВЕДЕНИЕ

Цель данного изложения заключается в том, чтобы показать *полезные* и *действенные* методы, которые в настоящее время достаточно широко освещены в программах обучения молодого поколения. Очевидно, мне придется нарушить традиции беглого сжатого доклада, которые были присущи ряду недавно проведенных симпозиумов по применению математических методов и вычислительной техники в геологии.

Я понимаю цель этого краткого курса не как стремление к строгому обоснованию количественных методов в геологии. Думаю, что в данном случае основная задача — увлечь вас количественными методами в геологии или по крайней мере познакомить с геологическими работами, связанными с применением математики, и с направлениями исследований в этой области, отраженными в литературе.

Если современный студент, просматривающий периодические издания, наталкивается на статью, переполненную производными, дифференциальными уравнениями или формулами математической статистики, он отбрасывает ее и берется за другую статью! Большинство геологов (включая и учителей этого студента), попав в подобную ситуацию, также трепещут перед математикой. Чаще всего при виде математических символов они думают, что статья «выше их понимания», так как в ней используется математика, воспринимаемая как незнакомый иностранный язык.

Этого страха или сомнений не возникнет, если уделить некоторое время изучению основ математического языка, познать «алфавит» которого можно довольно быстро (это не значит, что нужно навсегда стать «греком» среди несведущих в математике геологов) [45].

Я хочу познакомить геологов, «робких в области математики», с некоторыми примерами, характеризующими основные направления применения статистических и других математических методов в геологии. Ясно, что многие затронутые вопросы будут для большинства слушателей «стары, как мир», так как они привыкли изучать математику только с непосредственными вычислениями. Тем не менее факт остается фактом — многие геологи избегают

работ, связанных с применением математики, или не доверяют сделанным в них выводам, как и своим возможностям сделать такие выводы. Есть и такая категория геологов, которые не в состоянии использовать существующий мощный математический аппарат, который мог бы в значительной степени усилить различные области исследований.

Иногда применение математических методов в геологии было неудачным, но об этом в данный момент следует просто забыть. Рассмотрим некоторые обозначения.

Какой смысл вкладывается в символ «сигма»? В большинстве случаев прописной буквой Σ обозначают сумму нескольких слагаемых и в этом смысле ее используют в различных отраслях науки. Малой буквой σ обычно обозначают стандартное отклонение, квадрат которого называют дисперсией (об этом будет сказано подробнее в разделе, посвященном дисперсионному анализу). В качестве примера наиболее простого применения буквы Σ в смысле суммы рассмотрим процедуру вычисления среднего арифметического по данному набору n наблюдений $x_1, x_2, \ldots, x_i, \ldots, x_n$:

$$\text{Среднее арифметическое} = \frac{\sum_{i=1}^{n} x_i}{n},$$

что читается как сумма «иксов», деленная на n.

Если результаты наблюдения сгруппированы, получим

$$\bar{x} = \sum_{j=1}^{k} x_j f_j \Big/ \sum_{j=1}^{k} f_j,$$

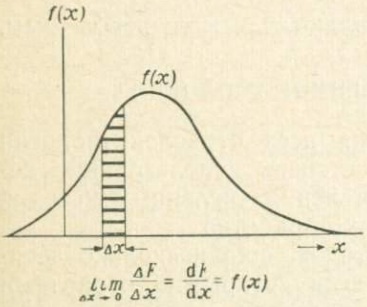
где k — число групп, f_j — число наблюдений в группе с номером j.

В геологии существуют различные варианты количественного подхода к решению задач, примеры которых мы сейчас рассмотрим. Начнем с интегрирования.

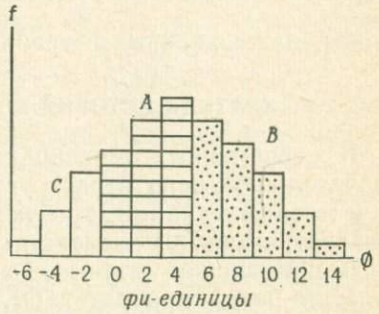
Если мы имеем дело с кривой, аналогичной изображенной на фиг. 1, то говорим, что кривая $f(x)$ есть функция x. В чем заключается смысл этой кривой? Пусть Δx — некоторый интервал на оси x и ΔF — вероятность того, что наблюдаемое в результате эксперимента значение случайной величины будет принадлежать интервалу Δx. Тогда отношение $\Delta F/\Delta x$ представляет собой среднюю плотность вероятности в интервале Δx. Если допустить, что Δx стремится к нулю, отношение $\Delta F/\Delta x$ будет приближаться к функции $f(x)$, которая называется плотностью вероятности в точке x.

Эта функция аналогична гистограмме, приведенной на фиг. 2, где интервал значений x разбит на классы, а также непрерывной кривой, представленной на фиг. 3. Если мы имеем дело с выбор-

кой, то ее всегда можно разделить на какое-то число классов. Так, в ряде случаев, связанных с исследованием осадочных пород, эти классы могут соответствовать размерам частиц. Например, класс C будет соответствовать гравию, класс B — песчаным

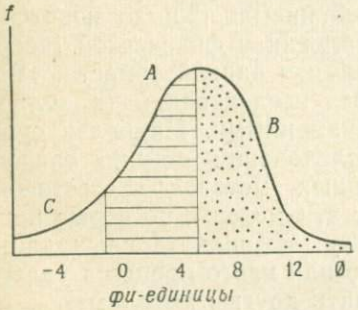


Фиг. 1. Кривая, иллюстрирующая $f(x)$ как функцию от x.

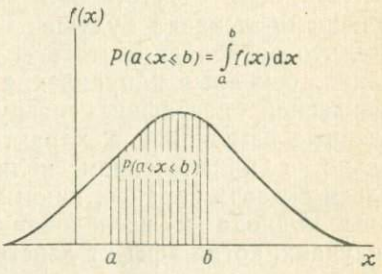


Фиг. 2. Гистограмма, построенная на основе классов размеров зерен.

частицам, а класс A — илистым осадкам или глинам. Каждому из них можно поставить в соответствие число, выражающее долю частиц в классе. Та или иная приведенная фигура характеризует распределение таких отношений в классах.



Фиг. 3. Кривая непрерывного распределения размеров частиц.



Фиг. 4. Функция плотности вероятности.

Для дальнейшего рассмотрения кривой плотности вероятности обратимся к фиг. 4. Вероятность события, заключающегося в том, что случайная величина в результате наблюдения примет значение x, принадлежащее интервалу от a до b, равна площади, ограниченной кривой $f(x)$ и ординатами в точках a и b. Эта площадь определяется как интеграл $\int_a^b f(x)\,dx$.

Общая площадь, ограниченная кривой $f(x)$, равна 1, и ее можно принять за 100%, что соответствует всей совокупности, которую представляет выборка. Это можно записать как интеграл $\int_{-\infty}^{\infty} f(x)dx=1$. Заметим, что в настоящее время в геологии интегралы, даже двойные и тройные, не являются чем-то необычным.

КРАТКАЯ ИСТОРИЯ КОЛИЧЕСТВЕННОЙ ГЕОЛОГИИ [1]

В современных тенденциях развития всех отраслей геологии значительную роль играет увеличение степени использования количественных данных, требующих применения численных методов или статистических доказательств при их обработке.

В применении математических методов еще несколько десятилетий назад приоритет был за самыми «старыми» отраслями геологии. Так, в седиментологии и в петрологии осадочных пород первые опыты применения математики относятся к 30-м годам. Эти работы (Крамбейна, Петтиджона и др.) посвящены в основном изучению распределения размеров частиц осадочных пород и их отсортированности. Некоторые методы изучения свойств осадочных пород были затем обобщены в работах Крамбейна и Петтиджона в 1938 г.

В то время когда книга Крамбейна и Петтиджона была уже опубликована, опыт применения математики при изучении форм рельефа пока отсутствовал. Однако Феннеман [23] до некоторой степени предсказал будущее геоморфологии и физической географии, что позволяет считать его пророком в данной области. «Возможно, основные направления научных исследований (в области физической географии) останутся неизменными. Наиболее существенные изменения в характере исследований обычно связаны только с привлечением количественных методов. Естественно, нельзя сказать заранее, какими будут эти методы, но новые принципы подхода к исследованиям могут возникнуть неожиданно, в случаях, когда эффект действия того или иного процесса удастся измерить и результаты противопоставить другим методам».

В 1945 г. Гортон опубликовал статью, посвященную вопросам применения гидродинамики в геоморфологии. Из всех имеющихся публикаций по данному вопросу ее можно рассматривать как классическую. Несмотря на то что она была опубликована в ведущем журнале (Bull. Am. Geol. Soc.), эта статья ускользнула от внимания большинства геологов, не связанных с данным разделом наук о Земле. Статья Гортона стимулировала ряд работ, содер-

[1] К сожалению, автор не упоминает работ советских исследователей (кроме статей А. Б. Вистелиуса, изданных на английском языке). Из наиболее важных следует упомянуть публикации М. А. Романовой, Ю. А. Воронина, А. Н. Бугайца, В. Н Бондаренко и др. — *Прим. ред.*

жащих применение количественных методов, группы исследователей Колумбийского университета. В течение двух последних десятилетий ведущим ученым в области количественной геоморфологии был Стралер, работы которого посвящены анализу форм рельефа. Из соавторов Стралера следует отметить Миллера, Смита, Кунса, Шумма, Милтона, Чорли, Морисава и Яссо. Необходимо также отметить ряд теоретических исследований в области количественной геоморфологии Шейдеггера, Шонберга и др.

Количественные исследования речных потоков в течение последнего десятилетия связаны с именами Леопольда, Уолляна, Хакка, Маддока, Миллера и др.

В области петрологии осадочных пород большая часть исследований, связанных с применением математических методов, принадлежит Гриффитсу и Фридману, а более ранняя — Ван дер Пласу, П. Аллену, Флеммингу и др. В этой области имеются также работы, связанные с палеогеографией, написанные Петтиджоном и его учениками Пеллитером, Йекилом, Мак-Брайдом, Уайтенером и др. Кроме того, необходимо отметить работы Д. Аллена, Боума, Хзу, Джулинского, Стенли и др.

В течение последнего десятилетия появился ряд работ Крамбейна и др., посвященных изучению процессов береговой эрозии, транспортировке частиц в процессе осадконакопления, а также сложным системам типа берег — океан — атмосфера. Серия таких статей недавно опубликована Крамбейном, Гаррисоном и др.

В области седиментологии следует отметить работы Агтеберга, Дугласа, Кроммелина, Браша, Лангбейна и Бокмана.

Исследования, связанные с применением математических методов в региональном стратиграфическом анализе, проводились Крамбейном, Шлоссом и их учениками в Северо-Западном университете. Результаты этих исследований применимы и в других отраслях геологии, как, например, методы анализа поверхностей тренда, использованные Крамбейном и Уиттеном, работы которых посвящены петрологическим и структурно-геологическим исследованиям.

В области геохимии, минералогии и петрографии следует отметить работы Аренса, Вистелиуса и Чейза.

В палеонтологии количественные методы стали применяться уже несколько десятилетий назад, но получили широкую известность только после классической работы Бринкманна, а также работ Барма, Миллера, Олсона и Симпсона. Несколько позднее Имбри и его соавторы начали исследования по числовому представлению палеонтологических данных и применению обобщенных статистических методов, в частности факторного анализа. Из других исследователей, применявших числовые методы в палеонтологии, следует отметить Мангиллера (Амстердам), Другера (Утрехт) и некоторых других голландских исследователей. В недавно опубликованной статье Джейна вероятностные методы используются для

построения видовых классификаций, что при наличии электронных вычислительных машин открывает перед геологами новые возможности.

Геофизика была «количественной» с самого начала, так что нельзя сказать, что эта область стала более «количественной». Математика применялась в ней столь широко, что нередко геофизику рассматривали как единственную «истинно научную отрасль» геологии. Конечно, с таким утверждением можно не соглашаться, но тем не менее верно, что в других областях геологии математические методы и вычислительная техника применялись в значительно меньших масштабах, чем в геофизике.

Таким образом, в развитии геологии *начало* математизации связано с применением физических и химических методов, более «математизированных» по сравнению с чисто «описательными» науками, к которым долгое время относилась геология (и биология).

ПРИРОДА КОЛИЧЕСТВЕННЫХ ИССЛЕДОВАНИЙ

Несмотря на то что в количественной геологии в значительной степени ощущается влияние статистических методов, можно привести ряд примеров и других вариантов количественного подхода к геологическим исследованиям.

Если количественные исследования в геологии расклассифицировать, то полученные в результате категории *не будут* взаимоисключающими. Так, например, *статистические* методы часто применяются при *геометрических* и *алгебраических* исследованиях. Можно назвать некоторые разновидности «геометрических» исследований, например такие, как изучение формы зерен, заполнения ими пространства, анализ форм рельефа и построение трендов поверхности. Аналогично можно рассматривать факторный или векторный анализ как «алгебраическое» исследование. «Статистические» методы нельзя рассматривать только как совокупность параметрических критериев проверки гипотез, противопоставляя их непараметрическим критериям, как это зачастую неправильно делается в геологической литературе. Некоторые параметрические критерии основаны на предположении, что наблюдения в выборках независимы, опробуемые совокупности характеризуются нормальным распределением, дисперсии равны (или известно их отношение), а результаты измерений представляют собой по крайней мере значения на «интервальной» шкале. Если какое-либо из этих условий не выполнено, то такие критерии нельзя использовать достаточно обоснованно. В данной ситуации их можно заменить непараметрическими критериями, которые, хотя и не являются столь эффективными, тем не менее представляются весьма полезными, так как их можно использовать там, где параметрические критерии неприменимы.

Гриффитс, а также Крамбейн и Грейбилл разделили геологические измерения на следующие четыре типа, каждому из которых соответствует своя шкала: 1) номинальная (или классификационная), 2) порядковая, 3) интервальная и 4) шкала отношений. Эти шкалы «кумулятивны» в том смысле, что для каждого более высокого уровня выполняются требования всех низших уровней с дополнением новых условий, характерных для шкалы данного уровня. Заметим, что для проверки гипотез по результатам наблюдений, выраженных на шкалах двух первых уровней, можно использовать только непараметрические критерии. По результатам наблюдений, соответствующих шкалам 3 и 4, можно проверять гипотезы как с помощью непараметрических, так и параметрических критериев.

Номинальная шкала характеризуется «эквивалентностью» объектов внутри выделенных категорий, например классов пород, минералов, ископаемых организмов и др.

Порядковая шкала *дополняет характеристику предыдущей со*отношением «больше чем». Наблюдения располагаются в порядке изменения значений (не обязательно равномерного изменения), как, например, шкала твердости Мооса, классы размеров частиц, неохарактеризованные количественно (гравий, песок, ил, глина), визуальные оценки сферичности, окатанности и др.

Интервальная шкала основана на знании соотношения двух интервалов. Каждый последующий класс характеризуется заданным значением, большим, чем предыдущее. Примерами результатов измерения, соответствующих этой шкале, могут служить показатели сферичности и окатанности Уейделла, температура, размеры зерен, изотопный состав и др.

Шкала отношений характеризуется наличием истинной нулевой точки, являющейся началом отсчета. Ей соответствуют измерения таких величин, как вес или масса, длина, площадь, объем, размер отдельных частиц, абсолютные отметки точек рельефа, скорость течения рек и др.

Естественно, природа геологических измерений зависит от вида статистических моделей, выбранных геологом для исследования.

ЛЕКЦИЯ 4

Дисперсионный анализ

М. Кауфмен

ВВЕДЕНИЕ

Одним из наиболее распространенных в геологической литературе за последние два десятилетия количественным методом является дисперсионный анализ. Число исследователей, оценивших потенциальные возможности этого метода и проводящих свои эксперименты в соответствии со схемами этого анализа, все время увеличивается. Тем не менее правильная постановка эксперимента — постоянный источник затруднений, возникающих у многих геологов, недостаточно четко воспринимающих количественные методы.

Дисперсия была определена выше как квадрат стандартного отклонения. Ее статистическая оценка по выборке представляет собой сумму квадратов отклонений результатов наблюдения от среднего арифметического, деленную на число наблюдений без единицы.

На результаты опробования может воздействовать несколько факторов одновременно. Цель дисперсионного анализа заключается в выявлении той доли общей изменчивости результатов наблюдения, которая обусловлена изменениями данного фактора, воздействующего на выборку или на особенности проводимых исследований. Таким образом, дисперсионный анализ — это метод совместной проверки гипотез о различиях между совокупностями и определения степени влияния каждого фактора на общую изменчивость.

Множество факторов, воздействующих на изменчивость в выборках, весьма разнообразно и содержит как самые обычные характеристики, например место и время, так и специфические, например климатические условия, свойства субстрата, локальные химические характеристики.

Обычно факторы разделяют на фиксированные и случайные, но иногда они «смешанные», т. е. производящие как фиксированное, так и случайное воздействие. Хотя определение характера факторов представляет интерес в математическом смысле, в данном курсе нет времени для подробного рассмотрения этого вопроса и мы приведем только соответствующую цитату из книги Крамбейна и Грейбилла [50].

«Важно следующее: если изучаемые факторы выбраны наудачу из некоторой совокупности факторов, то результаты исследования можно распространить в вероятностном смысле на всю их совокупность.

С другой стороны, если эти факторы выбраны неслучайно, то полученные результаты относятся только к данным факторам, и если они распространяются на другие факторы, то этот вывод следует расценивать как субъективное мнение исследователя.

Такой вывод необоснован в вероятностном отношении, но с точки зрения здравого смысла он может быть обоснованным и полезным. Все дело в том, что необоснованное применение схемы анализа с фиксированным эффектом вместо модели со случайными эффектами (или какого-либо другого способа) может внести путаницу в окончательную геологическую оценку статистических результатов».

«Дисперсионный анализ, хотя и существует значительное время, связан с рядом трудностей для изучающих его, когда они, чисто механически освоив алгебраические действия, не понимают смысла практического применения того, что делают» [32].

В основе дисперсионного анализа лежит утверждение, рассматриваемое как проверяемая гипотеза, что дисперсия суммы независимых случайных величин равна сумме дисперсий слагаемых. Это весьма важное утверждение означает, что два или более контролирующих факторов порождают изменчивость в наблюдаемой совокупности, измеряемую дисперсией, которая равна сумме соответствующих отдельных дисперсий.

ОДНОМЕРНЫЕ МОДЕЛИ И МЕТОДЫ

Самая простая ситуация, правда, редко применяемая в геологии, представляет собой случай, когда на совокупность или выборку влияет только одна переменная (фактор). При этом могут возникнуть разные принципы классификации наблюдений — по окраске, возрасту, принадлежности к различным видам и др.

Допустим, что заданы четыре классификационные категории, например биологические виды; из каждой совокупности, соответствующей этим категориям, взята выборка, на элементах которой измерено одно свойство. Пусть первым трем категориям соответствуют выборки, состоящие из трех наблюдений, а четвертой — из четырех.

Эти наблюдения можно представить в виде таблицы следующего типа:

Обобщенная таблица

	A	B	C	D	
	x_{11}	x_{21}	x_{31}	x_{41}	
	x_{12}	x_{22}	x_{32}	x_{42}	
	x_{13}	x_{23}	x_{33}	x_{44}	
Сумма	T_{1+}	T_{2+}	T_{3+}	T_{4+}	T_{++} Общая сумма
Среднее	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{4\cdot}$	$\bar{x}_{\cdot\cdot}$ Генеральное среднее

	Виды				
	A	*B*	*C*	*D*	
	8	6	7	2	
	7	5	4	3	
	5	7	6	6	
				4	
Сумма	20	18	17	15	70

Приведенный выше пример представляет собой типичную ситуацию, соответствующую *однофакторной* схеме дисперсионного анализа в условиях одномерной модели.

В данном случае проверяемая гипотеза заключается в том, что между выделенными категориями не существует различий. Естественно, что уровень значимости, при котором будет проверяться гипотеза, следует определить до начала эксперимента. Обычно задается 5-процентный уровень значимости (0,05). Однако в некоторых ситуациях *принимают уровень значимости*, равный 0,01 (1%) или даже 0,001 (0,1%), но во всех случаях его выбирают до начала проведения эксперимента.

Приведенные выше обозначения ($\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$) выборочных средних представляют собой оценки истинных, неизвестных нам средних значений признака для каждой категории (например, видов). Дисперсионный анализ дает ответ на вопрос — существуют ли реальные различия между истинными средними в выделенных группах? Эти различия признаются существенными, если изменчивость между категориями (видами) превышает изменчивость внутри них. Таким образом, если изменчивость между рассматриваемыми категориями (т. е. между выборками, соответствующими каждому виду) превышает изменчивость внутри этих групп (видов), мы можем сделать вывод, что они существенно отличаются *одна* от другой. Этот вывод нельзя было бы сделать, если бы изменчивость внутри групп превышала изменчивость между группами.

Изменчивость результатов наблюдений характеризуется *суммами их квадратов* или суммами квадратов отклонений от среднего. Общую сумму квадратов можно разделить на две части — сумму квадратов, обусловленную действием изучаемого фактора (главный эффект, в данном случае *между группами* или *их средними*), и сумму квадратов, связанную с изменчивостью результатов наблюдения (*внутри групп*).

Ниже приведена таблица одномерного однофакторного дисперсионного анализа, содержащая вычислительные процедуры для соответствующих сумм квадратов.

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	*F*-отношение
Главный эффект (средние)	$\sum_i \frac{T_i^2}{n_i} - \frac{\left(\sum_{ij} x_{ij}\right)^2}{N}$	$c-1$	$SS/c-1$	MS_1/MS_2
Изменчивость в выборках (внутри)	$\sum_{ij} x_{ij}^2 - \sum_i \frac{T_i^2}{n_i}$	$N-c$	$SS/N-c$	
Общая сумма	$\sum_{ij} x_{ij}^2 - \frac{\left(\sum_{ij} x_{ij}\right)^2}{N}$	$N-1$		
Среднее	16,99	3	5,66	$5{,}66 : 2{,}23 = 2{,}54$
Внутри групп	20,09	9	2,23	
Общая сумма	37,98	12		$F_{0{,}05;\ 3,\ 9} = 3{,}86$

Из приведенной выше таблицы видно, что вычислительные процедуры однофакторного дисперсионного анализа складываются из вычисления квадратов сумм для каждой группы, деленных на число наблюдений в группе ($\Sigma T_i^2/n_i$) подсчета, соответствующих сумм результатов наблюдений, квадрата общей суммы $(\Sigma x_{ij})^2$, сумм квадратов в группах и сравнения полученных результатов с помощью критерия Фишера (F-отношения).

Для приведенного примера сумма квадратов, отражающая главный эффект, вычисляется следующим образом:

$$\left(\frac{20^2}{3}+\frac{18^2}{3}+\frac{17^2}{3}+\frac{15^2}{4}\right)-\frac{70^2}{13}=393{,}91-376{,}92=16{,}99.$$

Так как число рассматриваемых групп равно 4, число степеней свободы будет $4-1=3$ и соответствующий средний квадрат окажется равным $16{,}99 : 3=5{,}66$.

Сумма квадратов, характеризующая изменчивость внутри групп, вычисляется следующим образом:

$$(8^2+7^2+5^2+6^2+5^2+7^2+7^2+4^2+6^2+2^2+3^2+6^2+4^2)-$$
$$-\left(\frac{20^2}{3}+\frac{18^2}{3}+\frac{17^2}{3}+\frac{15^2}{4}\right)=414-393{,}91=20{,}09.$$

Так как общее число наблюдений (N) равно 13, при четырех группах число степеней свободы будет $13-4=9$, а соответствующее значение среднего квадрата окажется равным 2,23.

F-ОТНОШЕНИЕ И ВЫВОДЫ

Проверяемая, или нулевая, гипотеза заключается в том, что между средними значениями в группах нет существенных различий. В условиях альтернативы (средние не равны) изменчивость *между* группами должна превышать изменчивость *внутри* групп и нулевую гипотезу в подобной ситуации следует отклонить. Задача заключается в том, чтобы определить, как должны быть велики различия для уверенного отклонения (или принятия) нулевой гипотезы.

Критерий Фишера (*F*-отношение), как и другие статистические критерии, позволяет выбрать уровень значимости при анализе изменчивости. Смысл его заключается в следующем. В результате применения статистического критерия мы можем допустить ошибки двух типов. Первая ошибка — нулевая гипотеза ошибочно отклоняется в том случае, когда она правильна; вторая — нулевая гипотеза ошибочно принимается в том случае, когда она ложна. Уровень значимости — это вероятность появления ошибки первого рода, которая в геологических задачах обычно принимается равной 0,05 (т. е. 5%), что дает удовлетворительные результаты.

F-отношение вычисляется путем деления среднего квадрата, соответствующего изменчивости между группами, на средний квадрат, характеризующий изменчивость внутри групп. В нашем примере это отношение 5,66/2,23=2,54. Теперь его следует сравнить с допустимым значением, которое берется из таблиц *F*-распределения при заданном уровне значимости и соответствующем числе степеней свободы. При уровне значимости 0,05, а также трех степенях свободы числителя и девяти степенях свободы знаменателя допустимое значение равно 3,86. Так как вычисленное значение 2,54 меньше, чем 3,86, нулевую гипотезу при данном уровне значимости следует принять и различия между средними признать несущественными.

ЛЕКЦИЯ 5

Многомерный анализ

М. Кауфмен

ВВЕДЕНИЕ

Многомерный анализ представляет собой совокупность методов исследования при более чем одной переменной. Андерсон [1] разделил многомерный анализ на пять главных проблем [см. 69]: а) корреляцию и регрессию; б) многомерные аналоги одномерных статистических методов проверки гипотез, включая t-критерий Стьюдента; в) задачи преобразования системы координат; г) задачи, в которых множество переменных разделено на группы, и д) зависимые наблюдения, включая корреляцию в их сериях.

Для иллюстрации рассмотрим ряд примеров многомерного анализа, опишем кратко корреляционный и регрессионный методы, которые будут детально проанализированы ниже, а также двухфакторный дисперсионный анализ с повторением и в заключение приведем пример многомерного дисперсионного анализа.

КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Две случайные величины могут изменяться таким образом, что между ними возникнет зависимость. Так, например, с увеличением значений одной величины значения другой могут уменьшаться. Аналогично возможна ситуация, когда с увеличением значений одной величины другая также проявляет тенденцию к увеличению значений. Мера силы зависимости между переменными, позволяющая судить о точности предсказания значений одной переменной по значениям другой, называется *коэффициентом корреляции*, выборочное значение которого мы будем обозначать через r.

Если выборочные значения двух случайных величин нанести на двумерный график (точечная диаграмма), точки могут группироваться около некоторой прямой линии. Такая зависимость называется сильной, если точки расположены близко от прямой, и слабой, если эта тенденция выражена не резко. Если коэффициент корреляции близок к нулю, точки беспорядочно рассеяны по всей диаграмме и зависимость отсутствует. Если его значение близко к $+1$, отчетливо выражена положительная корреляция (с увеличением значений одной переменной возрастают значения другой),

а если оно близко к —1, корреляция отрицательная (с увеличением значений одной величины значения другой уменьшаются).

«Линию регрессии», отражающую тенденцию изменения изучаемых величин, можно описать с помощью уравнения, которое в случае двух переменных имеет следующую общую форму: $y=a+bx$. В этом уравнении a — начальная точка на оси y при $x=0$, а b — величина, характеризующая угол наклона прямой по отношению к оси x. Величины a и b можно подсчитать по выборке, содержащей n пар значений x и y с помощью следующих выражений:

$$b=\frac{n(\sum x_i y_i)-(\sum x_i)(\sum y_i)}{n(\sum x_i^2)-(\sum x_i)^2},$$

$$a=\frac{\sum y_i - b(\sum x_i)}{n}.$$

Выборочный коэффициент корреляции r подсчитывается по следующей формуле:

$$r=\frac{n\sum x_i y_i-(\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}.$$

Пусть нам дана выборка, состоящая из девяти пар значений x и y, по которым, используя приведенные выше формулы, вычислим r и величины a и b уравнения регрессии.

x	y	x^2	y^2	xy
6,29	4,03	39,5641	16,2409	25,3487
6,22	4,18	38,6884	17,4724	25,9996
3,99	5,86	15,9201	34,3396	23,3814
4,21	4,89	17,7241	23,9121	20,5869
5,82	4,59	33,8724	21,0681	26,7138
3,86	5,87	14,8996	34,4569	22,6582
6,11	3,89	37,3321	15,1321	23,7699
6,22	4,09	38,6884	16,7281	25,4398
3,97	6,03	15,7609	36,3609	23,9391
Сумма 46,69	43,43	252,4501	215,7111	217,8354

$$b=\frac{n(\sum x_i y_i)-(\sum x_i)(\sum y_i)}{n(\sum x_i^2)-(\sum x_i)^2}=\frac{9\,(217{,}8354)-(46{,}69)\,(43{,}43)}{9\,(252{,}4501)-(46{,}69)^2}=$$

$$=\frac{1960{,}52-2027{,}75}{2272{,}05-2179{,}96}=\frac{-67{,}23}{92{,}09}=-0{,}81,$$

$$a=\frac{\sum y_i - b\left(\sum x_i\right)}{n}=\frac{43{,}43-(-0{,}81)\,(46{,}69)}{9}=$$

$$=\frac{43{,}43+37{,}84}{9}=\frac{81{,}27}{9}=9{,}03.$$

Уравнение регрессии будет иметь следующий вид: $y=9{,}03-0{,}81\,x$, где выборочный коэффициент корреляции

$$r=\frac{n\sum x_i y_i-\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2-\left(\sum x_i\right)^2}\sqrt{n\sum y_i^2-\left(\sum y_i\right)^2}}=$$

$$=\frac{9\,(217{,}8354)-(46{,}69)\,(43{,}43)}{\sqrt{9\,(252{,}4501)-(46{,}69)^2}\sqrt{9\,(215{,}7111)-(43{,}43)^2}}=$$

$$=\frac{1960{,}5186-2027{,}7467}{\sqrt{2272{,}0509-2179{,}9561}\sqrt{1941{,}3999-1886{,}1649}}=$$

$$=\frac{-67{,}2281}{\sqrt{92{,}0948}\sqrt{55{,}2350}}=\frac{-67{,}2281}{\sqrt{5086{,}8673}}=\frac{-67{,}2281}{71{,}32}=$$

$$=-0{,}943.$$

СХЕМА ДВУХФАКТОРНОГО АНАЛИЗА

В однофакторном дисперсионном анализе предполагалось, что общая изменчивость представляет собой сумму изменчивости внутри групп и между ними. Сравним наблюдаемые значения величины x с истинным средним μ. Обычно эти значения несколько отклоняются от μ. Такие отклонения можно рассматривать как состоящие из двух частей. Одна часть обусловлена принадлежностью к некоторой категории (положение в пространстве, условие эксперимента, время или какой-либо другой фактор). Это отклонение обозначим через A. Вторая часть называется «ошибкой», или случайной компонентой (e), отражающей изменчивость при условии, что действие изучаемого фактора исключено. Все это можно записать в виде следующего равенства:

$$(x-\mu)=A+e.$$

Если вместо одного рассматривать два фактора и их воздействие на изменчивость в выборке, то нельзя ограничиться анализом каждого фактора в отдельности, а необходимо учесть также порождаемый ими эффект смешанного взаимодействия. Это значит, что существует изменчивость в выборке, вызванная как действием каждого фактора в отдельности, так и их совместным воздействием. Это можно записать следующим образом (введя дополнительные обозначения: B — величина, аналогичная A, для второго

фактора и AB — изменчивость, обусловленная совместным действием двух факторов):

$$(x-\mu)=A+B+AB+e.$$

Если совместно изучаются три фактора, то аналогично можно выделить три главных эффекта A, B и C, три смешанных эффекта первого порядка AB, AC и BC, соответствующих всем возможным парам факторов, и один смешанный эффект второго порядка ABC, обусловленный совместным действием всех трех факторов. Это можно записать следующим образом:

$$(x-\mu)=A+B+C+AB+AC+BC+ABC+e.$$

СХЕМА ДВУХФАКТОРНОГО АНАЛИЗА С ПОВТОРЕНИЕМ

Рассмотрим схему двухфакторного дисперсионного анализа, но допустим, что выборки состоят из групп проб, представляющих собой дубликаты. Это «повторение» дает возможность получать дополнительную информацию об источниках изменчивости, воздействующих на анализируемую выборку. (Мы попытаемся охарактеризовать изменения в процедуре анализа по сравнению с обычной схемой без повторений.)

Общую характеристику изменчивости в условиях двухфакторного анализа можно разделить на четыре суммы квадратов:

1. Сумма квадратов, характеризующая главный эффект действия первого фактора.

Второй фактор (строки)	Первый фактор (столбцы)					
	1	2	3	. . .	c	
1	x_{111} x_{112} x_{113}	x_{121} x_{122} x_{123}	x_{131} x_{132} x_{133}	. . .	x_{1c1} x_{1c2} x_{1c3}	$\bar{x}_{1..}$
2	x_{211} x_{212} x_{213}	.	.	. . .	x_{2c1} x_{2c2} x_{2c3}	$\bar{x}_{2..}$
. . .	.	.	.	. . .	.	.
r	x_{r11} x_{r12} x_{r13}	.	.	. . .	x_{rc1} x_{rc2} x_{rc3}	$\bar{x}_{r..}$
	$\bar{x}_{.1.}$	$\bar{x}_{.2.}$	.	. . .	$\bar{x}_{.c.}$	$\bar{x}$ Генеральная средняя

2. Аналогичная сумма квадратов для второго фактора.

3. Сумма квадратов, обусловленная совместным действием обоих факторов.

4. Сумма квадратов, характеризующая случайную изменчивость после исключения влияния изучаемых факторов.

Обозначим через x_{ijk} измерение с номером k, соответствующее группе с номером j для первого фактора и группе с номером i для второго фактора.

Исходные данные для анализа представляются в таком виде, как они показаны в таблице на стр. 42.

Схема двухфакторного дисперсионного анализа с повторением приведена ниже. Обозначим через n число повторений наблюдения. Если повторений нет, то F-отношение представляет собой частное от деления соответствующего главного эффекта на средний квадрат смешанного эффекта.

Для иллюстрации вычислительной процедуры двухфакторного анализа с повторением (два наблюдения в ячейке) ниже приведен пример.

Дисперсионный анализ

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат
Главный эффект первого фактора (столбцы)	$\frac{1}{nr}\sum_i T_i^2 - \frac{T^2}{nrc}$	$c-1$	$SS/c-1$
Главный эффект второго фактора (строки)	$\frac{1}{nc}\sum_j T_j^2 - \frac{T^2}{nrc}$	$r-1$	$SS/r-1$
Смешанный эффект обоих факторов	$\frac{1}{n}\sum_{ij} T_{ij}^2 - \frac{1}{nc}\sum_i T_i^2 - -\frac{1}{nr}\sum_j T_j^2 - \frac{T^2}{nrc}$	$(r-1)(c-1)$	$SS/(r-1)(c-1)$
Общая промежуточная сумма	$\frac{1}{n}\sum_{ij} T_{ij}^2 - \frac{T^2}{nrc}$		
Остаточный эффект	$\sum_{ijk} x_{ijk} - \frac{1}{n}\sum_{ij} T_{ij}^2$	$rc(n-1)$	$SS/rc(n-1)$
Общая сумма	$\sum_{ijn} x_{ijn} - \frac{T^2}{nrc}$	$rcn-1$	

Фактор B (строки)	Фактор A (столбцы)				
	a	b	c	d	
1	2,5	2,8	3,3	2,6	
	2,4	2,5	3,2	2,4	
2	3,0	3,1	3,8	2,8	
	2,5	3,0	3,5	2,5	
3	2,9	3,6	3,9	2,4	
	2,9	3,6	3,8	3,1	
	Таблица сумм в ячейках			Суммы по строкам	
	4,9	5,3	6,5	5,0	21,7
	5,5	6,1	7,3	5,3	24,2
	5,8	7,2	7,7	5,5	26,2
Суммы по столбцам	16,2	18,6	21,5	15,8	72,1 Общая сумма

Вычисления начинаются с суммирования («уплотнения») данных внутри клеток таблицы, затем вычисляются суммы по строкам и столбцам и, наконец, подсчитывается общая сумма. При этом используются следующие обозначения:

n — число повторений (в нашем примере $n=2$);

r — число строк ($r=3$);

c — число столбцов ($c=4$);

A — сумма квадратов всех наблюдений:

$$A=\sum_{ijk} x_{ijk}^2=225{,}53;$$

B — сумма квадратов сумм наблюдений в ячейках, деленная на число повторений n:

$$B=\frac{1}{n}\sum_{ij} T_{ij}^2=\frac{449{,}84}{2}=224{,}92;$$

C — сумма квадратов сумм в столбцах, деленная на nr:

$$C=\frac{1}{nr}\sum_{i} T_{i}^2=\frac{1338{,}06}{6}=223{,}01;$$

D — сумма квадратов сумм в строках, деленная на nc:

$$D=\frac{1}{nc}\sum_{j} T_{j}^2=\frac{1767{,}66}{8}=220{,}96;$$

E — общая сумма квадратов, деленная на число всех наблюдений nrc:

$$E=\frac{T^2}{nrc}=\frac{5267{,}86}{24}=219{,}49.$$

Дисперсионный анализ

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	F-отношение
Столбцы	$(C-E)=3,52$	$(c-1)=3$	1,17	19,5
Строки	$(D-E)=1,47$	$(r-1)=2$	0,73	12,2
Смешанный эффект обоих факторов	$(B-E)-(C-E)-(D-E)=0,44$	$(c-1)(r-1)=6$	0,07	11,47
Общая промежуточная сумма	$(B-E)=5,43$	11		$F_{0,05;\ 6,\ 12}=3,00$
Остаточный эффект	$(A-B)=0,61$	12	0,05	$F_{0,05;\ 2,\ 18}=3,55$
Общая сумма	$(A-E)=6,04$	$rnc-1=23$		$F_{0,05;\ 3,\ 18}=3,16$

Сначала F-отношение вычисляется для смешанного и остаточного эффекта. Если это отношение превысит критическое значение, т. е. окажется существенным, то эффекты по строкам и столбцам сравниваются со смешанным эффектом. Если же оно окажется меньше критического (как в данном примере), то суммы квадратов, соответствующие смешанному и остаточному эффектам, складываются и делятся на сумму чисел степеней свободы $[0{,}44+0{,}61=1{,}05;\ 1{,}05:(6+12)=0{,}06]$. Эта величина затем используется как знаменатель при вычислении F-отношения для каждого из главных эффектов ($1{,}17:0{,}06=19{,}5;\ 0{,}73:0{,}66=12{,}2$). Такой способ обеспечивает более надежный результат проверки гипотезы за счет большего числа степеней свободы знаменателя. Вычисленные значения F можно сравнить с критическими, которые приведены в таблице.

В нашем примере проверяемая гипотеза отсутствия влияния обоих факторов на изучаемую величину уверенно отклоняется при уровне значимости 0,05. (Необходимо отметить, что в данном примере она отклоняется и при уровне значимости 0,01; это только усиливает сделанный вывод.)

Интерпретация таких результатов требует не только математического «обоснования», но и, что более важно, «здравого геологического смысла» при формулировании окончательных выводов. Заметим также, что подобные эксперименты могут привести к определению новых задач последующего анализа.

МНОГОМЕРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ В УСЛОВИЯХ МОДЕЛЕЙ С МНОГИМИ УРОВНЯМИ ИЗМЕНЧИВОСТИ

Результаты применения гнездовой схемы анализа при изучении состава пород впервые были опубликованы Крамбейном и Тьюки [51]. Эта тема рассматривалась затем Кроммелином [15], материалы которого мы используем в качестве примера для иллюстрации модели с рядом уровней изменчивости.

Кроммелин воспользовался методом Крамбейна и Тьюки для изучения минерального состава пород двух регионов, из которых было отобрано по две серии проб. Задача заключалась в получении ответа на вопрос: существуют ли какие-нибудь значимые различия в содержаниях четырех изучаемых минералов на этих площадях?

При изучении процентных величин возникают трудности, свойственные «закрытым системам». В этих системах с увеличением значений одной величины значения какой-либо другой уменьшаются в связи с тем, что их сумма представляет собой константу (100%). Вопросы, связанные с так называемой «ложной», отрицательной, корреляцией, рассматривались в работе Чейза и Маккензи [13], а также многими другими авторами.

Прежде чем приступить к обработке данных этого типа, их следует преобразовать. Для преобразования берется арксинус квадратного корня исходных данных [51].

Преобразованные данные используются затем для получения соответствующих сумм квадратов, необходимых для дисперсионного анализа.

Район ($i=1, 2$)		x		y		x	y	$x+y$
Выборка ($j=1, 2$)		a	b	a	b	$a+b$	$a+b$	$a+b$
Минералы ($m=1, 2, 3, 4$)	m_1	45,0	42,7	26,6	28,0	87,7	54,6	142,3
	m_2	30,0	31,3	15,3	12,9	61,3	28,2	89,5
	m_3	23,6	24,4	30,0	32,0	48,0	62,0	110,0
	m_4	17,5	18,4	43,9	42,1	35,9	86,0	121,9
$\sum_m x =$		116,1	116,8	115,8	115,0			
$\sum_{jm} x =$		232,9		230,8				
$\sum_{ijm} x =$		463,7						

$\sum_{ijm} x = 463,7$

Вычисление сумм квадратов

III $$\frac{\sum_{ij}\left(\sum_m x\right)^2}{4_j} = \frac{116,2^2 + \ldots + 115,0^2}{4} = 13439,0$$

II $$\frac{\sum_{i}\left(\sum_{jm} x\right)^2}{2 \cdot 4} = \frac{232,9^2 + 230,8^2}{2 \cdot 4} = 13438,9$$

I $$\frac{\left(\sum_{jim} x\right)^2}{2 \cdot 2 \cdot 4} = \frac{463,7^2}{2 \cdot 2 \cdot 4} = 13438,6$$

C $$\sum_{ijm} (x)^2 + 45,0^2 + \ldots + 42,1^2 = 15040,8$$

B $$\frac{\sum_{im}\left(\sum_{jm} x\right)^2}{2} = \frac{87,7^2 + \ldots + 86,0^2}{2} = 15029,1$$

A $$\frac{\sum_{m}\left(\sum_{ij} x\right)^2}{2 \cdot 2} = \frac{142,3^2 + \ldots + 121,9^2}{2 \cdot 2} = 13804,8$$

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	F-отношение
Между районами	$II - I = 0,3$	$2 - 1 = 1$	0,3	
Между выборками	$III - II = 0,1$	$2(2 - 1) = 2$	0,05	
Между минералами	$A - I = 366,2$	$4 - 1 = 3$	122,0	
Смешанный эффект (район×минерал)	$(I + B) - (II + A) = 1224,0$	$(2 - 1)(4 - 1) = 3$	408,0	214,7
Смешанный эффект (выбора×минерала)	$(II + C) - (III + B) = 11,6$	$2(2 - 1)(4 - 1) = 6$	1,9	
Общая сумма	$C - I = 1602,2$	$2 \cdot 2 \cdot 4 - 1 = 15$		

С помощью F-отношения было проведено сравнение изменчивости, вызванной действием двух смешанных эффектов: (район×минерал) и (выборка×минерал).

ИЗУЧЕНИЕ ТЕРРАСЫ РЕКИ САСКУЭХАННА

Истоки реки Саскуэханна находятся в южной части штата Нью-Йорк, откуда она течет в Пенсильванию, затем снова возвращается в штат Нью-Йорк и опять уходит в Пенсильванию близ Тованды и далее течет по территории этого штата на протяжении 500 км по направлению к границе штата Мэриленд, а оттуда к Чесапикскому заливу, где впадает в Атлантический океан. Она пересекает Аппалачское плато, провинцию Хребтов и Бассейнов, Пидмонт и небольшой участок Береговой равнины.

В результате деятельности ледников в плейстоценовое время в северной Пенсильвании и южной части штата Нью-Йорк образовались моренные отложения. Берущие начало в этой области водные потоки несут большое количество обломочного материала и образуют гравелитовые отложения.

Современное положение этих террас было закартировано при более ранних исследованиях. Пельтер [71] сопоставил отложения этих террас с различными гляциальными образованиями, используя для их группировки превышения над урезом воды, глубины эродированности и литологический состав. Средние значения превышений над уровнем воды для различных террас в значительной степени совпадают. Вопрос о применимости такого показателя, как глубина эродированности для определения возраста террас, обсуждался многими геоморфологами [67]. Останцы исходных террас при сопоставлении оставляют широкую область для различных предположений.

Цель исследования заключалась в получении ответа на вопрос: существуют ли различия между речными террасами, связанными с различными стадиями ледниковой деятельности? Если такая зависимость существует, то можно предполагать, что в характеристиках разных террас имеются различия. Такие различия позволят проводить в дальнейшем сопоставление террас на основе количественного изучения их физических свойств.

План выборки был разработан так, чтобы можно было выявить источники наибольшей изменчивости физических свойств отложений, слагающих террасы. Последующее применение дисперсионного анализа позволило выявить источники наибольшей изменчивости характеристик этих отложений. Так, например, была рассмотрена изменчивость между пробами, отобранными в одном участке, между опробованными участками и между различными террасами при фиксированном положении относительно течения реки и при разных положениях вдоль русла. Решение поставленной задачи сводилось к проверке гипотезы, заключающейся в том, что параметры распределения характеристик различных террас отличаются несущественно, при альтернативе, что различия между их статистическими оценками значимы.

Статистическое изучение результатов наблюдения состава пород требует применения многомерных методов анализа, позволяющих рассматривать совместно несколько характеристик, по которым проводится классификация.

В табл. 1 приведены результаты многомерного дисперсионного анализа состава пород, учитывающие как все четыре террасы, так и четыре участка опробования. Результаты анализа проб, приведенные в этой таблице, соответствуют точкам опробования. Эти результаты представлены величинами, в сумме составляющими 100% для каждого элемента классификации таблицы.

Таким образом, результаты анализа состава приведены в табл. 1 под соответствующим наименованием пункта опробования. Названия террас расположены от наиболее древней в верхней части таблицы до наиболее молодой в нижней. Состав пород представлен следующими основными типами: Q — кварцитовые песчаники, F — полевошпатовые песчаники, аркозы, граувакки, S — илы и сланцы, C — роговики и I — изверженные и метаморфические породы, включая метаморфические кварциты. К данным исходной таблицы применили преобразование (арксинус квадратного корня исходных данных) и полученные результаты использовали в дисперсионном анализе [5].

В основу анализа была положена модель, учитывающая эффект действия как факторов по столбцам и строкам, так и типов пород; эта модель заимствована из неопубликованных работ Крамбейна, на что получено его разрешение [44].

Общая схема дисперсионного анализа приведена в нижней части табл. 1.

Для проверки гипотезы об однородности результатов наблюдения был применен F-критерий Фишера, значения которого вычислялись по выборочным дисперсиям соответствующих смешанных произведений. В результате было установлено, что существенных расхождений в совместных эффектах действия, положении во времени и между типами пород нет в связи с тем, что вычисленное значение критерия меньше допустимого при уровне значимости 0,05 и степенях свободы 12 и 36.

Наоборот, существенное значение F-отношения было получено для суммарного эффекта действия двух факторов — времени образования террас и типов их состава.

Данные, представленные в виде процентных величин в табл. 1, основаны на результатах подсчета числа галек в выборке, которую можно считать хорошо перемешанной. Поэтому третья проверяемая гипотеза была сформулирована как утверждение об однородности значений результатов преобразования (арксинус корня квадратного исходных данных). Дисперсия биномиального распределения (остаточная или стабильная), соответствующая преобразованным величинам, постоянна и равна $821/n$. При анализе состава число изучаемых галек в среднем составляло в выборке $n = 95$.

Таблица 1

Многомерный дисперсионный анализ результатов изучения состава четырех террас на четырех участках опробования (данные пересчитаны на 100%)

Терраса		Пункты опробования			
		Санбери	Ливерпул	Гаррисберг	Мариетта
Иллинойсская	*Q*	18	4	17	11
	F	14	3	1	3
	S	42	34	54	48
	C	26	55	25	37
	I	0	4	3	1
Олеанская	*Q*	15	15	18	11
	F	11	2	1	4
	S	54	29	19	62
	C	13	52	50	22
	I	7	2	12	1
«Бингемтон»	*Q*	7	29	42	50
	F	13	14	6	3
	S	68	50	33	23
	C	5	5	14	15
	I	7	2	5	9
«Валли-Хедз»	*Q*	9	10	22	13
	F	6	4	16	14
	S	71	64	48	37
	C	7	17	5	21
	I	7	5	9	15

Дисперсионный анализ

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	*F*-отношение
Между пунктами опробования	6,18	3		
Между террасами	21,30	3		
Остаточный эффект (пункты опробования×террасы)	24,60	9		
Общая промежуточная сумма	52,08	15		
Между типами	9 285,47	4	2321,34	
Внутри пунктов опробования и типов	1 074,99	12	89,58	1,28
Внутри террас и типов	2 117,91	12	176,49	2,52
Внутри типов и остаточного эффекта	2 518,01	36	69,94	8,10
Общая сумма	14 996,38	79		

$F_{0,95;\ 12,\ 36} = 2,03,$ $F_{0,05;\ 36,\ \infty} = 1,42.$

Следовательно, дисперсия будет равна 821 : 95 = 8,64. Полученное значение дисперсии, деленное на остаточную дисперсию для типов пород, дало значение F-отношения, которое оказалось весьма существенным. Это позволяет сделать вывод, что, хотя состав террас и однороден в региональном смысле, между террасами существует локальная неоднородность.

Второй пример применения дисперсионного анализа для изучения состава террас основан на рассмотрении данных о составе террас реки Висконсин, опробованных в шести различных участках (табл. 2). Полученное значение F-отношения значительно превышает критическое, что свидетельствует о существенных различиях в составе для разных участков опробования террас.

Сравнение остаточных дисперсий в условиях биномиальной модели также дало очень большое значение F-отношения. Такие различия были доказаны для так называемых «очищенных» данных по Иллинойсской террасе и для трех террас реки Висконсин, рассматриваемых отдельно. В связи с этим можно сделать вывод, что значительные различия в составе относительно точки опробования свойственны для всех четырех террас, включая и удаленную от реки Висконсин Иллинойсскую террасу.

Таблица 2

Многомерный дисперсионный анализ результатов изучения состава трех террас реки Висконсин на шести участках опробования (данные пересчитаны на 100%)

Терраса		Пункты опробования					
		Вайсокс	Танханнок	Питстон	Санбери	Ливерпул	Хайспайр
Олеанская	*Q*	12	38	7	16	15	18
	F	1	12	16	11	2	1
	S	73	33	68	52	29	19
	C	8	3	1	14	52	50
	I	6	14	8	7	2	12
«Бингемтон»	*Q*	1	2	22	7	29	50
	F	10	18	16	13	14	3
	S	89	66	44	68	50	23
	C	0	7	6	5	5	15
	I	0	7	12	7	2	9
«Валли-Хедз»	*Q*	6	7	15	18	11	13
	F	1	13	11	10	4	14
	S	88	69	50	61	63	37
	C	1	5	5	3	17	21
	I	4	6	19	8	5	15

Дисперсионный анализ

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	*F*-отношение
Между пунктами опробования	153,71	5		
Между террасами	8,06	2		
Остаточный эффект (пункты опробования × × террасы)	69,27	10		
Общая промежуточная сумма	231,04	17		
Между типами	13 357,02	4	3339,26	
Внутри пунктов опробования и типов	4 585,45	20	229,27	3,48
Внутри террас и типов	953,82	8	119,23	1,81
Внутри типов и остаточного эффекта	2 638,80	40	65,97	7,65
Общая сумма	21 766,13	89		

Биномиальная дисперсия = 821 : 95 = 8,64, $F_{0,05;\ 20,\ 40} = 1,84$, $F_{0,05;\ 40,\ \infty} = 1,39$, $F_{0,05;\ 8,\ 40} = 2,18$.

Для того чтобы изучить зависимость между составом террас «Бингемтон» и «Валли-Хедз», был также проведен дисперсионный анализ, результаты которого приведены в табл. 3. Получено весьма

Таблица 3

Многомерный дисперсионный анализ состава террас «Бингемтон» и «Валли-Хедз» на шести участках опробования (данные пересчитаны на 100%)

Терраса		Пункты опробования					
		Вайокс	Танкханнок	Питстон	Санбери	Ливерпул	Хайспайр
«Бингемтон»	*Q*	1	2	22	7	29	50
	F	10	18	16	13	14	3
	S	89	66	44	68	50	23
	C	0	7	6	5	5	15
	I	0	7	12	7	2	9
«Валли-Хедз»	*Q*	6	7	15	18	11	13
	F	1	13	11	10	4	14
	S	88	69	50	61	63	37
	C	1	5	5	3	17	21
	I	4	6	19	8	5	15

Дисперсионный анализ

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	F-отношение
Между пунктами опробования	187,19	5		
Между террасами	6,53	1		
Остаточный эффект (пункты опробования × × террасы)	13,26	5		
Общая промежуточная сумма	206,98	11		
Между типами	11 137,33	4		
Внутри пунктов опробования и типов	3 268,78	20	163,44	3,41
Внутри террас и типов	160,60	4	40,15	<1
Внутри типов и остаточного эффекта	1 042,44	20	52,12	6,04
Общая сумма	15 609,15	59		

Биномиальная дисперсия $=821 : 95 = 8{,}64$; $F_{0{,}05;\ 20,\ 20} = 2{,}12$, $F_{0{,}05;\ 20,\ \infty} = 1{,}57$.

существенное значение F-отношения для положений участков опробования и остаточного эффекта относительно биномиальной дисперсии, тогда как отношение дисперсий для террас и остаточного эффекта оказалось незначительным, что свидетельствует об однородности составов террас в любой точке вдоль долины.

Подводя итог проведенному дисперсионному анализу состава террас, необходимо отметить, что в том случае, когда в анализе участвовали все четыре террасы, были выявлены существенные различия для Иллинойсской террасы и террас реки Висконсин, что приводит к значимым различиям между террасами. Однако, после того как из рассмотрения была исключена Иллинойсская терраса, различия стали не очень большими. Сходство в составе террас «Бингемтон» и «Валли-Хедз» приводит к несущественному значению F-отношения между террасами. В результате анализа трех террас реки Висконсин были получены аналогичные выводы. Такое сходство литологического состава террас можно объяснить тем, что образовавшие их ледники двигались через геологически похожие территории, транспортируя обломки сходных типов пород.

ЛЕКЦИЯ 6

Многомерные методы в геологии

Р. Мак-Кеммон

МАТРИЦЫ

Допустим, что у нас есть некоторое множество элементов, поведение которых определяется заданным набором аксиом. В геологии под этим понимают совокупность результатов наблюдений, по которым необходимо сделать те или иные выводы. Форма представления этих данных может быть самой разнообразной. Так, например, в палеонтологии это результаты определения наличия или отсутствия представителей отдельных видов изучаемого комплекса ископаемой фауны; в петрографии — значения содержаний различных окислов в породах; в геоморфологии — результаты измерения углов наклона долины или террасы вдоль течения реки; в геохимии — содержания редких элементов в породах, выраженные в млн$^{-1}$. В каждом из этих примеров набор имеющихся данных можно рассматривать в качестве совокупности элементов, которой соответствует группа аксиом, определяющих их поведение как математических характеристик. Определим матрицу

$$\mathbf{A}=\begin{bmatrix} a_{11} & \dots & a_{1c} \\ \cdot & \cdot & \cdot \\ a_{r1} & \dots & a_{rc} \end{bmatrix}$$

как прямоугольную таблицу с определенным расположением образующих ее элементов. Любой элемент матрицы обозначим a_{ij}, где первый нижний индекс i означает номер строки, а второй j — номер столбца. Число строк и столбцов определяет порядок матрицы, который для приведенного выше примера будет $r \times c$.

Частный случай матрицы — скаляр, обычно воспринимаемый как отдельное число. Действительно, матрица порядка 1×1 есть скаляр. Так, например, единичное измерение коэффициента окатанности кварцевой гальки представляет собой скалярную величину, но если мы располагаем набором измерений этого коэффициента для галек различного состава, то такой набор значений можно записать в виде матрицы.

Матрица, состоящая из одной строки или одного столбца, называется вектором. Обычно n-мерный вектор-столбец записывается следующим образом:

$$\mathbf{x}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix},$$

а n-мерная вектор-строка — как

$$\mathbf{x}'=[x_1, \ \ldots \ x_n].$$

Штрих в данном случае означает, что $\mathbf{x}'$ — транспонированный вектор-столбец $\mathbf{x}$. Вообще если матрица $\mathbf{A}$ порядка $r\times c$, то ее транспозицией будет матрица $\mathbf{A}'$ порядка $c\times r$, образуемая путем замены строк матрицы $\mathbf{A}$ на столбцы. Так, транспозицию приведенной выше матрицы $\mathbf{A}$ можно записать в следующей форме:

$$\mathbf{A}'=\begin{bmatrix} a_{11} \ \ldots \ a_{r1} \\ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \\ a_1 \ \ldots \ a_{rc} \end{bmatrix}.$$

Если матрица квадратная и равна своей транспозиции, то ее называют симметричной. В такой матрице $a_{ij}=a_{ji}$ для всех i и j. Элементы a_{ii} квадратной матрицы образуют главную диагональ и называются диагональными элементами.

Наиболее часто встречающиеся типы матриц следующие. Единичная матрица

$$\mathbf{I}=\begin{bmatrix} 1 \ \ldots \ 0 \\ \cdot \ \cdot \ \cdot \\ 0 \ \ldots \ 1 \end{bmatrix}$$

представляет собой квадратную матрицу, диагональные элементы которой равны 1, а все остальные — нулю. Диагональная матрица порядка $p\times p$

$$\mathbf{D}\,(d_i)=\begin{bmatrix} d_1 \ \ldots \ 0 \\ \cdot \ \cdot \ \cdot \ \cdot \\ 0 \ \ldots \ d_p \end{bmatrix}$$

также является квадратной матрицей, диагональные элементы которой равны $d_1, d_2, \ldots, d_p$, а все остальные равны нулю. За-

метим, что некоторые d_i могут быть также равны нулю. Верхнетреугольная матрица порядка $p \times p$

$$\mathbf{T}(i<j)=\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1p} \\ 0 & t_{22} & \dots & t_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & t_{pp} \end{bmatrix}$$

содержит ненулевые элементы выше главной диагонали (включая последнюю), а все остальные ее элементы равны нулю. Не исключено, что некоторые из элементов, расположенные выше главной диагонали, будут равны нулю. Транспозиция $\mathbf{T}'$ $(i>j)$ верхнетреугольной матрицы $\mathbf{T}$ $(i<j)$ называется нижне-треугольной матрицей, также имеющей порядок $p \times p$.

В дальнейшем нам потребуются понятия единичного вектора

$$\mathbf{e}'=[1,\ 1,\ \dots,\ 1]$$

и матрицы, состоящей только из единиц:

$$\mathbf{E}=\begin{bmatrix} 1 & \dots & 1 \\ \cdot & \cdot & \cdot \\ 1 & \dots & 1 \end{bmatrix}.$$

Как и для скалярных величин, для матриц можно определить операции сложения, вычитания, умножения и деления. Две матрицы $\mathbf{A}$ и $\mathbf{B}$ порядка $r \times c$ равны, если для всех i и j выполнено равенство

$$a_{ij}=b_{ij}.$$

Элементы матрицы, представляющей сумму двух матриц одного порядка, определяются как суммы соответствующих элементов складываемых матриц. Если

$$\mathbf{A}=\begin{bmatrix} a_{11} & \dots & a_{1c} \\ \cdot & \cdot & \cdot \\ a_{r1} & \dots & a_{rc} \end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix} b_{11} & \dots & b_{1c} \\ \cdot & \cdot & \cdot \\ b_{r1} & \dots & b_{rc} \end{bmatrix},$$

то их сумма будет

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix} a_{11}+b_{11} & \dots & a_{1c}+b_{1c} \\ \cdot & \cdot & \cdot \\ a_{r1}+b_{r1} & \dots & a_{rc}+b_{rc} \end{bmatrix}.$$

Операция вычитания матриц определяется аналогично.

Сложение матриц коммутативно и ассоциативно. Нетрудно доказать, что

$$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A},$$

$$\mathbf{A}+(\mathbf{B}+\mathbf{C})=(\mathbf{A}+\mathbf{B})+\mathbf{C}.$$

Операция умножения матрицы $\mathbf{A}$ на скаляр d записывается следующим образом:

$$d\mathbf{A}=\begin{bmatrix} da_{11} & \dots & da_{1c} \\ \cdot & \cdot & \cdot \\ da_{r1} & \dots & da_{rc} \end{bmatrix},$$

т. е. каждый элемент a_{ij} матрицы $\mathbf{A}$ умножается на d.

Умножение матрицы $\mathbf{A}$ порядка $m\times p$ на матрицу $\mathbf{B}$ порядка $p\times n$ производится по следующей формуле:

$$\mathbf{AB}=\begin{bmatrix} \sum_{j=1}^{p} a_{1j}b_{j1} & \dots & \sum_{j=1}^{p} a_{1j}b_{jn} \\ \cdot & \cdot & \cdot \\ \sum_{j=1}^{p} a_{mj}b_{j1} & \dots & \sum_{j=1}^{p} a_{mj}b_{jn} \end{bmatrix},$$

где матрица $\mathbf{AB}$ порядка $m\times n$, каждый элемент которой с номером ij образуется путем суммирования произведений элементов i-ой строки матрицы $\mathbf{A}$ на элемент j-ого столбца матрицы $\mathbf{B}$. Нетрудно видеть, что перемножать можно только матрицы, одна из которых имеет число столбцов, равное числу строк другой матрицы. Можно также показать, что для операции умножения матриц справедливы следующие соотношения:

$$\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC},$$

$$\mathbf{A}(\mathbf{BC})=\mathbf{AB}(\mathbf{C}).$$

Необходимо отметить, что для квадратных матриц равенство $\mathbf{AB}=\mathbf{BA}$ неверно.

Используя правила матричного умножения, легко показать, что произведение вектора-строки и вектора-столбца, имеющих одинаковое число элементов, есть скаляр, т. е.

$$\mathbf{x}'\mathbf{y}=[x_1, \dots, x_p]\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_p \end{bmatrix}=\sum_{i=1}^{p} x_i y_i.$$

Этот скаляр представляет собой сумму произведений соответствующих пар элементов, образующих векторы. Из этого следует, что

$$\mathbf{x}'\mathbf{x}=\sum_{i=1}^{p} x_i^2,$$

т. е. произведение вектора-строки и его транспозиции является суммой квадратов элементов вектора. Таким образом, если по результатам p измерений $\mathbf{x}=\{x_1, \ldots, x_p\}$ требуется подсчитать выборочную дисперсию, то можно воспользоваться следующим векторным выражением:

$$s^2=\frac{1}{(p-1)}\left(\mathbf{x}-\frac{1}{p}\mathbf{E}\mathbf{x}\right)'\left(\mathbf{x}-\frac{1}{p}\mathbf{E}\mathbf{x}\right)=\frac{\sum_{i=1}^{p}(x_i-\bar{x})^2}{p-1},$$

где матрица $\mathbf{E}$ порядка $p\times p$, все элементы которой равны 1. Если результаты измерения образуют связанные пары $[\{x_1, y_1\}, \ldots, \{x_p, y_p\}]$ и для них требуется оценить коэффициент корреляции, то его оценку можно вычислить по формуле

$$r_{xy}=\frac{\left(\mathbf{x}-\frac{1}{p}\mathbf{E}\mathbf{x}\right)'\left(\mathbf{y}-\frac{1}{p}\mathbf{E}\mathbf{y}\right)}{\left[\left(\mathbf{x}-\frac{1}{p}\mathbf{E}\mathbf{x}\right)'\left(\mathbf{x}-\frac{1}{p}\mathbf{E}\mathbf{x}\right)\right]^{\frac{1}{2}}\left[\left(\mathbf{y}-\frac{1}{p}\mathbf{E}\mathbf{y}\right)'\left(\mathbf{y}-\frac{1}{p}\mathbf{E}\mathbf{y}\right)\right]^{\frac{1}{2}}}.$$

Умножение диагональной матрицы $\mathbf{P}$ (d_i) с элементами главной диагонали $d_1, \ldots, d_r$ на матрицу $\mathbf{A}$ порядка $r\times c$ приводит к новой матрице, каждый элемент которой, принадлежащий i-ой строке, представляет произведение $d_i a_{ij}$, т. е.

$$\mathbf{D}(d_i)\,\mathbf{A}=\begin{bmatrix} d_1a_{11} & \ldots & d_1a_{1c} \\ \cdot & \cdot & \cdot \\ d_ra_{r1} & \ldots & d_ra_{rc} \end{bmatrix}.$$

В результате умножения матрицы $\mathbf{A}$ порядка $r\times c$ на диагональную матрицу порядка $c\times c$ будет получена новая матрица, каждый элемент которой, принадлежащий j-ому столбцу, представляет произведение $d_j a_{ij}$. Произведение нижне-треугольной матрицы $\mathbf{T}$ $(i>j)$ порядка $c\times c$ и матрицы $\mathbf{A}$ порядка $r\times c$ дает в результате матрицу

порядка $r \times c$, элементами которой будут различные линейные комбинации элементов j-ого столбца исходных матриц, т. е.

$$\mathbf{T}(i>j)\,\mathbf{A}=\begin{bmatrix} t_{11}a_{11} & \dots & t_{11}a_{1c} \\ t_{21}a_{21}+t_{22}a_{21} & \dots & t_{21}a_{1c}+t_{22}a_{2c} \\ \cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot & & \\ \sum_{j=1}^{r} t_{rj}a_{j1} & \dots & \sum_{j=1}^{r} t_{rj}a_{jc} \end{bmatrix}.$$

Умножение матрицы **A** на верхне-треугольную матрицу порядка $c \times c$ приводит к линейным комбинациям, образованным элементами i-ой строки.

Матричные операции, аналогичные делению, мы рассмотрим ниже.

МОДЕЛИРОВАНИЕ ПРОЦЕССА ЭВОЛЮЦИИ ПЕСЧАНЫХ ОТЛОЖЕНИЙ

Рассмотрим простой пример, который может служить хорошей иллюстрацией применения действий над матрицами. Этот пример заимствован из статьи Жизба [40], посвященной моделированию процесса эволюции песчаных отложений.

В качестве первого приближения состав песчаного слоя можно представить в виде матрицы, в которой столбцы будут соответствовать различным минералам, а строки — гранулометрическим фракциям. Элемент этой матрицы с индексом ij будет представлять собой содержание i-ого минерала в j-ой фракции, выраженное в процентах от веса пробы. Таким образом, каждое значение состава можно рассматривать как величину x_{ij}, имеющую двойной индекс ij.

Если рассматривается n минералов и m фракций, то для всей системы можно записать

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}=1.$$

Для удобства вместо процентов мы будем считать, что значения x_{ij} находятся в интервале от нуля до единицы, т. е. $0<x_{ij}<1$. Определение вероятностей, соответствующих гранулометрическим фракциям, Жизба [40] проводит по следующей формуле:

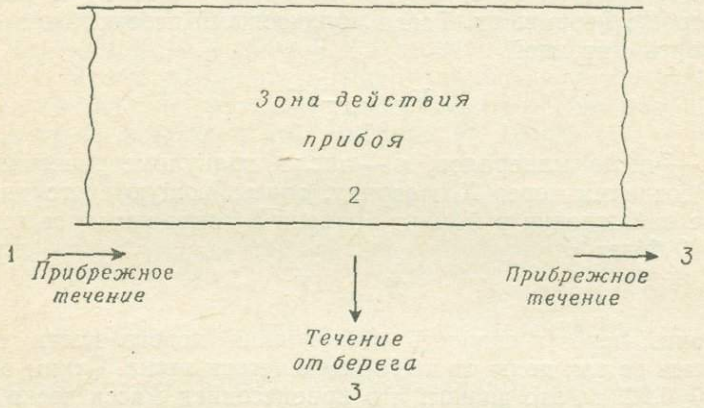
$$y_i=\sum_{i=1}^{n} x_{ij},$$

а для определения вероятностей, соответствующих минеральным видам, он использует выражение

$$\gamma_i = \sum_{j=1}^{m} x_{ij}.$$

Для удобства предположим, что размер зерен выражен в фи-единицах и каждая гранулометрическая фракция соответствует изменению размера на одну фи-единицу.

Можно было бы более полно охарактеризовать состав песка, введя дополнительные факторы, например такие, как форма зерен и их структурное расположение. Это привело бы к увеличению порядка матрицы и несколько усложнило бы процедуру обработки



Фиг. 1. Упрощенная модель зоны действия прибоя [40].

данных. Чтобы не вводить усложнений, ограничимся только двумя факторами — размером зерен и их минеральным составом.

Итак, мы дали формальное описание песчаной залежи в статическом положении. Однако, как известно, условия залегания песка претерпевают непрерывные изменения и предметом нашего исследования является взаимодействие между условиями залегания и песчаным телом. Такое взаимодействие лучше всего представить в терминах матричной алгебры.

В качестве примера рассмотрим упрощенную модель зоны действия прибоя, заимствованную из работы Жизбы [40], которая приведена на фиг. 1. В этой модели введен ряд ограничений, которые упрощают ее, но позволяют наглядно показать действия над матрицами, характеризующими динамику прибрежной части песчаных отложений и их перемещения. Основные допущения сводятся к следующим:

1. Привнос нового песка приливным течением происходит с постоянной скоростью, и состав привносимого песка постоянный.

2. Результат действия прибоя на сортировку песка по размерам частиц не зависит от его минерального состава.

3. Совместное действие прилива и отлива вызывает перемещение осадка с той же скоростью, что и привнос нового песка, и оно сказывается главным образом на размерах и минеральном составе. Из сделанных допущений следует, что масса песка в точке наблюдения остается постоянной, т. е. что дополнительный привнос или вынос песка отсутствует. Наша задача заключается в изучении изменений состава песка во времени при данных исходных условиях. Для простоты мы будем считать время дискретным, а интервалы времени между наблюдениями постоянными.

Начнем с предположения, что скорость вновь поступающего песка постоянна и он однороден по своему составу и размерам, т. е. хорошо перемешан. Тогда привносимый песок можно охарактеризовать матрицей

$$\mathbf{Y}=\left(\frac{1}{mn}\right)\mathbf{E},$$

где m — число минералов, n — число гранулометрических фракций. Обозначим через $\mathbf{X}_t$ матрицу, описывающую состояние песка в точке наблюдения в момент времени t, тогда исходное состояние при $t=0$ будет

$$\mathbf{X}_{t=0}=\mathbf{Y}.$$

Обозначим через r скорость поступления нового песка и будем выражать ее как долю от всего имеющегося песка. Таким образом, если $r=0{,}20$, то это значит, что привнесенная масса песка составляет 20% от массы песка, имевшегося в наличии. Обозначив через $\mathbf{U}_1$ состояние песка в данный момент, мы можем записать

$$\mathbf{U}_1=r\mathbf{Y}+\mathbf{X}_t.$$

В результате действия прибоя песок сортируется по размерам частиц. Так как при этом частицы песка истираются и дробятся, можно считать, что изменение фракций имеет тенденцию к возрастанию значений фи-единиц, т. е. к уменьшению размера зерен. Если не учитывать минерального состава, то можно записать

$$\mathbf{U}_2=\mathbf{U}_1\mathbf{S},$$

где $\mathbf{S}$ — верхне-треугольная матрица порядка $n\times n$, элементы которой s_{ij} принимают значения в интервале от 0 до 1, т. е. $0<s_{ij}<1$. Кроме того,

$$\sum_{j>1} s_{ij}=1$$

для всех i. Это означает, что масса песка остается неизменной. Результат умножения матрицы $\mathbf{U}_1$ на матрицу $\mathbf{S}$ будет характеристикой эффекта сортировки песка по размеру зерен независимо от минерального состава.

Теперь мы рассмотрим перемещение песка, вызванное действием отлива и непрерывно действующего прибрежного течения. Заметив, что перемещение песка сказывается прежде всего на минеральном составе и размерах зерен, мы умножим диагональную матрицу $\mathbf{R}$ порядка $m \times m$, отражающую различия в перемещении для минеральных видов на матрицу $\mathbf{U}_2$, а полученный результат умножим на диагональную матрицу $\mathbf{T}$ порядка $n \times n$, которая характеризует различия в перемещении для гранулометрических фракций. Таким образом,

$$\mathbf{V}_1 = \mathbf{R}\mathbf{U}_2\mathbf{T},$$

где $\mathbf{V}_1$ — матрица, характеризующая состояние перемещенного песка. Для элементов матриц $\mathbf{R}$ и $\mathbf{T}$ имеют место неравенства $r_{ii} > 0$, $t_{ii} < 1$. Чтобы сохранить процентные соотношения, в матрицу $\mathbf{V}_1$ следует ввести весовые коэффициенты. Обозначим через $s(\mathbf{V}_1) = s$ сумму всех элементов матрицы $\mathbf{V}_1$. Тогда

$$\mathbf{V}_2 = \frac{r}{s}\mathbf{V}_1.$$

Каждое последующее наблюдение состояния песка можно представить в виде матрицы

$$\mathbf{X}_{t+1} = \mathbf{U}_2 - \mathbf{V}_2 = (r\mathbf{Y} + \mathbf{X}_t)\,\mathbf{S} - \frac{r}{s}\mathbf{R}\,(r\mathbf{Y} + \mathbf{X}_t)\,\mathbf{S}\mathbf{T}.$$

Это выражение показывает, что каждое последующее состояние песка является функцией предыдущего состояния и воздействия процесса, приводящего к изменениям.

При изучении геологических процессов наибольший интерес представляют не кратковременные изменения (в данном случае в размерах зерен и в составе), а эффект, полученный в результате длительного течения процесса. Так, например, могут возникнуть следующие вопросы. Будет ли в условиях данной модели достигнуто равновесие? Что произойдет в результате изменения скорости поступления нового песка? Каков будет результат, если исходную матрицу, описывающую состояние песка, представить не набором констант, а случайными величинами? Как изменяется средний размер зерен и среднее значение показателя отсортированности во времени при данных исходных условиях? Для получения ответов на эти вопросы лучше всего воспользоваться моделированием, основанным на последовательных итерациях в условиях данной модели.

В табл. 1 приведены результаты моделирования процесса эволюции песчаных отложений, полученные после 70 итераций,

Таблица 1

Результаты моделирования процесса эволюции песчаных отложений

Матрица R

0,900	0,0	0,0	0,0
0,0	0,800	0,0	0,0
0,0	0,0	0,400	0,0
0,0	0,0	0,0	0,200

Матрица S

0,900	0,050	0,030	0,020	0,0	0,0	0,0	0,0
0,0	0,900	0,050	0,050	0,0	0,0	0,0	0,0
0,0	0,0	0,950	0,030	0,020	0,0	0,0	0,0
0,0	0,0	0,0	0,950	0,030	0,020	0,0	0,0
0,0	0,0	0,0	0,0	0,970	0,030	0,0	0,0
0,0	0,0	0,0	0,0	0,0	0,970	0,020	0,010
0,0	0,0	0,0	0,0	0,0	0,0	0,980	0,020
0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,000

Матрица T

0,800	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,0	0,600	0,0	0,0	0,0	0,0	0,0	0,0
0,0	0,0	0,400	0,0	0,0	0,0	0,0	0,0
0,0	0,0	0,0	0,200	0,0	0,0	0,0	0,0
0,0	0,0	0,0	0,0	0,200	0,0	0,0	0,0
0,0	0,0	0,0	0,0	0,0	0,400	0,0	0,0
0,0	0,0	0,0	0,0	0,0	0,0	0,600	0,0
0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,900

Скорость поступления песка=0,2

После 0 итераций
Средний размер=2,50, показатель отсортированности=2,29
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,250
0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,250
0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,250
0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,250

После 10 итераций
Средний размер=2,69, показатель отсортированности=1,74
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,003	0,006	0,014	0,034	0,037	0,017	0,007	0,002	0,121
0,004	0,007	0,017	0,038	0,042	0,021	0,009	0,004	0,142
0,012	0,019	0,039	0,063	0,066	0,046	0,027	0,017	0,289
0,023	0,034	0,061	0,084	0,085	0,071	0,049	0,042	0,448

После 20 итераций
Средний размер=2,73, показатель отсортированности=1,64
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,005	0,011	0,028	0,031	0,013	0,005	0,002	0,097
0,003	0,006	0,013	0,032	0,036	0,016	0,007	0,003	0,116
0,010	0,016	0,033	0,064	0,074	0,045	0,022	0,013	0,276
0,020	0,030	0,063	0,102	0,117	0,090	0,051	0,039	0,511

После 30 итераций
Средний размер=2,75, показатель отсортированности=1,61
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,004	0,010	0,026	0,029	0,012	0,005	0,001	0,090
0,003	0,006	0,013	0,030	0,033	0,015	0,007	0,002	0,108
0,009	0,015	0,030	0,061	0,074	0,044	0,020	0,012	0,267
0,019	0,029	0,061	0,107	0,133	0,099	0,051	0,037	0,535

После 40 итераций
Средний размер=2,76, показатель отсортированности=1,60
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,004	0,010	0,025	0,028	0,012	0,005	0,001	0,088
0,003	0,005	0,012	0,029	0,033	0,015	0,006	0,002	0,105
0,009	0,015	0,030	0,060	0,073	0,043	0,020	0,012	0,261
0,019	0,028	0,060	0,108	0,140	0,103	0,051	0,037	0,545

После 50 итераций
Средний размер=2,76, показатель отсортированности=1,60
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,004	0,010	0,025	0,028	0,012	0,005	0,001	0,087
0,003	0,005	0,012	0,029	0,032	0,015	0,006	0,002	0,105
0,009	0,015	0,030	0,059	0,072	0,042	0,020	0,012	0,259
0,019	0,028	0,059	0,108	0,143	0,104	0,051	0,036	0,549

После 60 итераций
Средний размер=2,76, показатель отсортированности=1,60
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,004	0,010	0,025	0,028	0,012	0,005	0,001	0,087
0,003	0,005	0,012	0,029	0,032	0,015	0,006	0,002	0,104
0,009	0,015	0,030	0,059	0,072	0,042	0,020	0,012	0,258
0,019	0,028	0,059	0,108	0,145	0,105	0,051	0,036	0,551

После 70 итераций
Средний размер=2,76, показатель отсортированности=1,60
(фи-единицы)

—1	0	1	2	3	4	5	6	
0,002	0,004	0,010	0,025	0,028	0,012	0,005	0,001	0,087
0,003	0,005	0,012	0,029	0,032	0,015	0,006	0,002	0,104
0,009	0,015	0,029	0,059	0,072	0,042	0,020	0,012	0,257
0,019	0,028	0,059	0,108	0,145	0,105	0,051	0,036	0,552

а также промежуточные результаты после каждых 10 итераций. Скорость привноса песка выбрана 0,2. Для иллюстрации процесса изменения матрицы, характеризующей состояние песка в результате последовательных итераций с помощью метода моментов, вычислялись средние значения размера частиц и показателя их отсортированности. Оба средних значения выражены в фи-единицах. Процедура моделирования процеса эволюции песчаных отложений начата в условиях однородного состава. Ясно, что средний размер частиц с течением времени уменьшается, чему соответствуют увеличения среднего значения фи-единиц. Кроме того, песок постепенно становится лучше отсортированным, что видно в крайнем правом столбце таблицы, соответствующем четвертому минералу, который постепенно становится доминирующим.

Несмотря на то что эта модель значительно упрощена по сравнению с реальным процессом переноса, действующим в условиях прибрежной фации, она наглядно показывает возможности матричной алгебры для описания сложного процесса осадконакопления, играющего очень важную роль в геологии.

ГЕОМЕТРИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ

После того как мы рассмотрели основы матричной алгебры, можно перейти к геометрическому представлению действий над матрицами. При этом будет показано, что не только большинство многомерных статических процедур можно наглядно выразить с помощью геометрических построений, но и большинство задач сокращенного представления геологических данных по существу можно свести к задачам аналитической геометрии. Начнем с рассмотрения векторов в многомерном пространстве.

Вектор $\mathbf{v}$ можно представить как набор значений координат $(x_1, \ldots, x_n)$ точки P. В данной постановке началу координат будет соответствовать точка $(0, \ldots, 0)$. Значение каждой координаты — скалярная величина, а совокупность координатных осей образует систему координат. В связи с этим при изменении системы координат вектору $\mathbf{v}$ будет соответствовать новый набор значений $(y_1, \ldots, y_n)$.

Правила сложения, вычитания и умножения векторов остаются теми же, что и для матриц, так как вектор является частным случаем матрицы. Используя геометрические представления, можно определить длину вектора, угол его поворота, угол между двумя векторами и проекцию одного вектора на другой. Все операции определения этих характеристик можно представить в алгебраической форме для любого числа координат пространства.

Длина вектора $\mathbf{v}$, которую мы обозначим $\| \mathbf{v} \|$, определяется как

$$\| \mathbf{v} \| = (x_1^2 + \ldots + x_n^2)^{\frac{1}{2}}$$

или, используя матричные обозначения, $\| \mathbf{v} \| = (\mathbf{v}'\mathbf{v})^{\frac{1}{2}}$, где $\mathbf{v}$ — вектор-столбец порядка $n \times 1$. Длина единичного вектора равна 1, т. е. $\| \mathbf{v} \| = 1$. Простейшим примером единичного вектора будет набор значений (1, 0, ..., 0). Угол θ между двумя векторами $\mathbf{v}$ и $\mathbf{w}$ с координатами $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_n)$ определяется выражением

$$\cos\theta = \frac{(x_1y_1 + \ldots + x_ny_n)}{\| \mathbf{v} \| \cdot \| \mathbf{w} \|},$$

или в матричной записи

$$\cos\theta = \frac{\mathbf{v}'\mathbf{w}}{(\mathbf{v}'\mathbf{v})^{\frac{1}{2}} (\mathbf{w}'\mathbf{w})^{\frac{1}{2}}},$$

где $\mathbf{v}$ и $\mathbf{w}$ — векторы-столбцы порядка $n \times 1$. Если два вектора $\mathbf{v}$ и $\mathbf{w}$ образуют угол 90°, то

$$\cos\theta = 0,$$

поскольку $(x_1y_1 + \ldots + x_ny_n) = 0$, что равносильно $\mathbf{v}'\mathbf{w} = 0$. Это равенство выполняется даже и в том случае, если векторы $\mathbf{v}$ и $\mathbf{w}$ не являются нулевыми векторами (0, ..., 0). Два ненулевых вектора называются ортогональными, если их произведение равно 0. Этот результат можно обобщить на любое число векторов $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$, которые мы будем называть взаимно ортогональными, если каждый из них ортогонален по отношению ко всем остальным $r - 1$ векторам. Нетрудно видеть, что множество единичных векторов взаимно ортогонально. В дальнейшем мы воспользуемся одним из вариантов такого набора единичных (базисных) векторов:

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \ldots, 0), \\ \mathbf{e}_2 &= (0, 1, \ldots, 0), \\ &\cdots\cdots\cdots \\ \mathbf{e}_n &= (0, 0, \ldots, 1), \end{aligned}$$

который образует матрицу $\mathbf{I}$.

Таким образом, вектор $\mathbf{v}$ с координатами $(x_1, x_2, \ldots, x_n)$ определяется в этой системе координат выражением

$$\mathbf{v} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_n\mathbf{e}_n$$

или

$$\mathbf{I} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} + \ldots + x^n \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}.$$

В итоге становится очевидной скалярная природа координат.

В результате мы пришли к понятию системы координат. Ясно, что рассмотренный выше набор единичных векторов представляет собой систему координат, с помощью которой любой вектор $\mathbf{v}$ с координатами $(x_1, \ldots, x_n)$ можно описать однозначно. Векторы, образующие координатную систему, называются базисными. Размерность пространства, описываемого данной системой координат, определяется числом базисных векторов, которое равно максимальному числу линейно независимых векторов этого пространства. Набор базисных векторов дает возможность однозначно представить любой вектор в этом пространстве.

ПРЕОБРАЗОВАНИЕ СИСТЕМЫ КООРДИНАТ

Допустим, что нам нужно изменить первоначально выбранную систему координат. Это может быть вызвано необходимостью перевода данных в другую форму, более удобную для интерпретации, или устранения избыточной информации, не нужной для дальнейших вычислений. Кроме того, для удобства обобщения данных можно соответствующим образом выбирать систему координат, а также преобразовывать данные путем изменения системы координат.

Рассмотрим две n-мерные системы координат, заданные в одном и том же пространстве. Обозначим через $\mathbf{e}_1, \ldots, \mathbf{e}_n$ и $\mathbf{f}_1, \ldots, \mathbf{f}_n$ соответствующие им наборы базисных векторов. Координаты вектора $\mathbf{v}$ в первой системе обозначим $(x_1, \ldots, x_n)$, а во второй $(y_1, \ldots, y_n)$. Тогда можно записать

$$\mathbf{v}=x_1\mathbf{e}_1+\ldots+x_n\mathbf{e}_n=y_1\mathbf{f}_1+\ldots+y_n\mathbf{f}_n.$$

Допустим, что соотношение между базисными векторами определено следующими уравнениями:

$$\mathbf{e}_1=P_{11}\mathbf{f}_1+P_{12}\mathbf{f}_2+\ldots+P_{1n}\mathbf{f}_n,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\mathbf{e}_n=P_{n1}\mathbf{f}_1+P_{n2}\mathbf{f}_2+\ldots+P_{nn}\mathbf{f}_n.$$

Следовательно, мы можем записать

$$y_1=P_{11}x_1+P_{21}x_2+\ldots+P_{n1}x_n,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$y_n=P_{1n}x_1+P_{2n}x_2+\ldots+P_{nn}x_n,$$

что определяет координаты вектора $\mathbf{v}$ относительно второй системы координат в терминах первой системы. Это можно записать и в матричной форме

$$\mathbf{y}=\mathbf{P}'\mathbf{x},$$

где $\mathbf{y}$ и $\mathbf{x}$ — векторы-столбцы порядка $n\times 1$, а $\mathbf{P}'$ — матрица порядка $n\times n$.

Теперь предположим, что нам нужно выразить координаты вектора **v** относительно первой системы координат в терминах второй системы. Рассмотрим случай, когда матрица $\mathbf{P}'$ есть единичная матрица **I**. Тогда можно записать

$$y_1 = x_1,$$
$$y_2 = \ldots x_2,$$
$$\ldots\ldots$$
$$y_n = \ldots x_n,$$

или

$$x_1 = y_1,$$
$$x_2 = \ldots y_2,$$
$$\ldots\ldots$$
$$x_n = \ldots y_n.$$

Если в уравнении $\mathbf{y}=\mathbf{P}'\mathbf{x}$ положим $\mathbf{P}'=\mathbf{I}$ и умножим обе части равенства на **I**, то получим

$$\mathbf{Iy}=\mathbf{I}(\mathbf{P}'\mathbf{x})=\mathbf{I}(\mathbf{Ix})=(\mathbf{II})\mathbf{x}$$

или

$$\mathbf{y}=\mathbf{x}.$$

В нашей исходной задаче матрица $\mathbf{P}'$ — любая матрица порядка $n\times n$, и нам нужно найти такую матрицу **P** порядка $n\times n$, умножение которой на $\mathbf{P}'$ дало бы единичную матрицу. Эта матрица является обратной матрице $\mathbf{P}'$ и обозначается $(\mathbf{P}')^{-1}$, т. е.

$$(\mathbf{P}')^{-1}\mathbf{P}'=\mathbf{I}.$$

Операция умножения на обратную матрицу — аналог деления. Так, например, для матрицы порядка 1×1 получим

$$\frac{1}{p_{11}}p_{11}=1$$

при $p_{11}\neq 0$. Если операция обращения для матрицы более высокого порядка существует, такая матрица называется неособой. Эквивалентом равенства $p_{11}=0$ для матриц более высокого порядка является отсутствие обратной матрицы. Такие матрицы называются особыми. Любое преобразование пространства, сохраняющее его размерность, называется неособым преобразованием, тогда как преобразование, приводящее к проектированию исходного пространства в подпространство с меньшим числом измерений, является особым преобразованием.

В нашем примере можно записать

$$(\mathbf{P}')^{-1}\mathbf{y}=(\mathbf{P}')^{-1}\mathbf{P}'\mathbf{x}=\mathbf{Ix}=\mathbf{x}$$

или

$$x_1 = p^{11}y_1 + p^{21}y_2 + \ldots + p^{n1}y_n,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$x_n = p^{1n}y_1 + p^{2n}y_2 + \ldots + p^{nn}y_n,$$

где верхние индексы используются для обозначения элементов обратной матрицы. Теперь уместно сказать, что для нахождения обратной матрицы существуют численные методы.

Наибольший интерес представляет матрица **P**, для которой выполнено соотношение

$$\mathbf{P}'\mathbf{P} = \mathbf{I},$$

где **I** — единичная матрица. Матрица, обладающая этим свойством, называется ортогональной. В матрице **I** как векторы, так и векторы-столбцы образуют наборы единичных векторов.

Если матрица $\mathbf{P}'$ приведенного выше примера получена в результате пространственного преобразования ортогональной матрицы, то такое преобразование называется ортогональным. Это преобразование сохраняет длину всех векторов, определенных в данном пространстве, и углы между ними. Ортогональное преобразование является эквивалентом поворота координатных осей относительно начала координат или поворота с последующим отражением.

Наиболее распространенной формой представления данных, собираемых геологом, является таблица. Петрографы, например, представляют в виде таблиц содержания различных окислов, полученные в результате анализа образцов; седиментологи — содержания различных гранулометрических фракций, а стратиграфы — процентные соотношения различных типов пород в разрезах. В каждом случае сумма значений в строке таблицы, соответствующей одному наблюдению, постоянна. Если рассмотрение ограничивается только тремя компонентами, большинство исследователей для их наглядного представления воспользуются треугольной диаграммой. Хотя это и создает иллюзию трехмерного представления, мы знаем, что имеем дело с проекцией на плоскость. В случае n компонент такая проекция будет расположена на $(n-1)$-мерной гиперплоскости. Хотя в действительности невозможно нанести наши данные на $(n-1)$-мерную гиперплоскость в n-мерном пространстве, мы можем провести соответствующее преобразование исходных данных с целью выбора подходящей системы координат.

Для этого рассмотрим вектор $(x_1, \ldots, x_n)$, представляющий n-мерное наблюдение, которое мы изображаем в n-мерном пространстве, и введем для него следующее ограничение:

$$\sum_{i=1}^{n} x_i = 1,$$

где $0<x_i<1$. Изменим исходную систему координат путем преобразования

$$y_i=x_i-\frac{1}{n}$$

для всех $i=1, 2, \ldots, n$. Положив, что $\mathbf{y}$ есть n-мерный вектор-столбец с координатами $(y_1, \ldots, y_n)$ относительно исходной системы координат, проведем следующее преобразование:

$$\mathbf{z}=\mathbf{P}'\mathbf{y},$$

где

$$\mathbf{P}'=\begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{bmatrix}.$$

Эту матрицу называют матрицей Хельмерта. Нетрудно доказать, что она ортогональна. Результатом соответствующего ортогонального преобразования будет n-мерный вектор-столбец $\mathbf{z}$, координаты которого определяются выражениями

$$z_1=0,$$

$$z_2=\frac{1}{\sqrt{2}}(y_2-y_1),$$

$$z_3=\frac{1}{\sqrt{6}}(2y_3-y_1-y_2),$$

$$\ldots\ldots\ldots\ldots$$

$$z_n=\frac{1}{\sqrt{n(n-1)}}[(n-1)y_n-(y_1+\ldots+y_{n-1})].$$

Необходимо отметить, что первая строка матрицы Хельмерта представляет собой координатную ось, образующую прямой угол с плоскостью, на которую спроектированы данные. В табл. 2 приведен пример преобразования Хельмерта для содержаний окислов химических элементов в типичных кислых и основных породах.

Таблица 2

Результаты преобразования Хельмерта

Преобразование Хельмерта

Матрица Хельмерта

0,289	0,289	0,289	0,289	0,289	0,289	0,289	0,289	0,289	0,289	0,289	0,289
0,707	—0,707	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,408	0,408	—0,816	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,289	0,289	0,289	—0,866	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,224	0,224	0,224	0,224	—0,894	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,183	0,183	0,183	0,183	0,183	—0,913	0,0	0,0	0,0	0,0	0,0	0,0
0,154	0,154	0,154	0,154	0,154	0,154	—0,926	0,0	0,0	0,0	0,0	0,0
0,134	0,134	0,134	0,134	0,134	0,134	0,134	—0,935	0,0	0,0	0,0	0,0
0,118	0,118	0,118	0,118	0,118	0,118	0,118	0,118	—0,943	0,0	0,0	0,0
0,105	0,105	0,105	0,105	0,105	0,105	0,105	0,105	0,105	—0,949	0,0	0,0
0,095	0,095	0,095	0,095	0,095	0,095	0,095	0,095	0,095	0,095	—0,953	0,0
0,087	0,087	0,087	0,087	0,087	0,087	0,087	0,087	0,087	0,087	0,087	—0,957

Тип пород

Кислые изверженные породы [1]

SiO_2	Al_2O_3	Fe_2O_3	FeO	CaO	MgO	Na_2O	K_2O	TiO_2	MnO_2	P_2O_5	H_2O
x (1)											
0,7208	0,1386	0,0086	0,0167	0,0133	0,0052	0,0308	0,0546	0,0037	0,0006	0,0018	0,0053
Вектор координат											
0,000	0,412	0,344	0,236	0,186	0,159	0,111	0,074	0,113	0,104	0,093	0,082

Если вы хотите продолжать вычисления, то введите «yes», если хотите прекратить вычисления, то «по».

Yes

Тип пород

Основные породы [1]

SiO_2	Al_2O_3	Fe_2O_3	FeO	CaO	MgO	Na_2O	K_2O	TiO_2	MnO_2	P_2O_5	H_2O
x_1											
0,4820	0,1560	0,0300	0,0780	0,1050	0,0820	0,0260	0,0090	0,0190	0,0017	0,0030	0,0080
Вектор координат											
0,000	0,231	0,236	0,125	0,073	0,081	0,120	0,120	0,096	0,102	0,091	0,079

Если вы хотите продолжать вычисления, то введите «yes», если хотитет прекратить вычисления, то «по».

No

** 235. XEQ „STOP".

[1] Содержания окислов даны в долях единицы; если приведенные числа умножить на 100, получим содержание в процентах.

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Несколько раньше мы уже рассмотрели формулу для вычисления выборочного коэффициента корреляции двух непрерывных случайных величин x и y. Этот коэффициент является мерой силы линейной зависимости между изучаемыми случайными величинами. Для n двумерных наблюдений $[(x_1, y_1), \ldots, (x_n, y_n)]$ выборочный коэффициент корреляции вычисляется по следующей формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

где x и y — средние арифметические, вычисленные по выборочным значениям x и y. Выше мы уже отмечали, что коэффициент корреляции принимает значения в интервале от —1 до 1, т. е. $-1 \leqslant r_{xy} \leqslant 1$.

Теперь мы дадим коэффициенту корреляции геометрическую интерпретацию, которая нам потребуется позднее при рассмотрении многомерных методов классификации. Пусть **x** и **y** — векторы в n-мерном евклидовом пространстве с началом координат в точке, соответствующей выборочным средним $\bar{x}$ и $\bar{y}$. Пусть также вектор **u** имеет координаты

$$\{(x_1 - \bar{x}), \ldots, (x_n - \bar{x})\},$$

а вектор **v** —

$$\{(y_1 - \bar{y}), \ldots, (y_n - \bar{y})\}.$$

Углу θ между этими векторами в n-мерном пространстве соответствует

$$\cos\theta = \frac{\mathbf{u}'\mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|},$$

где $\mathbf{u}'\mathbf{v}$ — скалярное произведение векторов, а $\|\mathbf{u}\|$ и $\|\mathbf{v}\|$ — их длины. Угол θ выражен в радианах. Первая часть этого равенства представляет собой коэффициент корреляции, и мы можем записать

$$r = \cos\theta$$

или

$$\theta = \arccos r.$$

Проведя преобразование $\arccos r$, мы получим функцию от коэффициента корреляции, являющуюся мерой рассеяния. Если эту

функцию обозначить как $\theta(x, y)$, можно показать, что $\theta(x, y)$ обладает всеми свойствами метрики, а именно:

1) $\theta(x, y)=0$ при $x=y$,

2) $x \neq y$ при $\theta(x, y)>0$,

3) $\theta(x, y)=\theta(y, x)$,

4) $\theta(x, y)<\theta(x, z)+\theta(z, y)$.

Доказательство последнего неравенства можно найти в работе Блюменталя [4]. Основная особенность меры $\theta(x, y)$ заключается в том, что последовательные арифметические операции, зависящие от метрических свойств используемых стандартных величин, можно существенно изменить. Так, например, среднее значение коэффициента корреляции $r_{ср}$ для группы случайных величин $(x_1, \ldots, x_p)$ можно определить выражением

$$r_{ср}=\cos\frac{\sum\limits_{i}\sum\limits_{j<}\arccos\left(r_{x_i x_j}\right)}{\binom{p}{2}},$$

где $\binom{p}{2}$ — число сочетаний из p по 2. Ниже мы воспользуемся этим выражением.

ЛЕКЦИЯ 7

Главные компоненты

Р. Мак-Кеммон

ВВЕДЕНИЕ

При обработке результатов наблюдений нередки случаи, когда число рассматриваемых характеристик (переменных) бывает большим. В ряде задач не всегда удается непосредственно выбрать комплекс переменных для последующей обработки. Это важно, когда отсутствуют хорошо обоснованные математические модели или четко сформулированные гипотезы, а цель эксперимента заключается в выявлении смысла зависимостей между переменными, чтобы использовать последние в дальнейшем. Если число изучаемых переменных велико, то нередко возникают трудности, связанные с выбором среди них функций и аргументов. Так, например, факторами, определяющими сложную химическую реакцию в кристаллизующемся расплаве, могут быть как состав участвующих в ней реагентов, так и физические условия. Возрастание числа переменных обычно приводит к смешению причинных и следственных связей.

Для последующей интерпретации результатов многомерных наблюдений полезно провести их преобразование, используя метод главных компонент. Это преобразование устраняет возможную нелинейность зависимостей между переменными и служит для дальнейших исследований. Иногда такое сокращение весьма существенно. Кроме того, метод главных компонент нередко позволяет получить дополнительную информацию об источниках изменчивости изучаемых данных. Выявленные главные компоненты можно использовать как для описания, так и для классификации результатов наблюдений.

Метод главных компонент уже давно известен в литературе и неоднократно рассматривался с различных точек зрения рядом авторов, например Андерсоном [1] и Кендаллом [46]. Вопросы практических приложений метода главных компонент и интерпретации результатов его применения рассмотрены в работе Рао [77]. Наша задача заключается в краткой характеристике этого метода и в иллюстрации возможностей его применения в конкретных геологических ситуациях. В дальнейшем мы рассмотрим основные уравнения этого метода и проведем его сравнение с обычным регрессионным анализом, затронув и геометрическую интерпретацию.

Кроме того, будут рассмотрены два практических примера.

Основная задача метода главных компонент формулируется очень просто. Требуется найти такое линейное преобразование m случайных величин $(x_1, \ldots, x_m)$ в новый набор p случайных величин $(z_1, \ldots, z_p)$, который бы обладал заданными статистическими свойствами. Эти свойства заключаются в независимости случайных величин, образующих набор $(z_1, \ldots, z_p)$, и в расположении величин z_i в порядке уменьшения дисперсий. Каждая новая случайная величина представляет собой линейную комбинацию m исходных случайных величин и называется главной компонентой. Вообще число главных компонент равно числу исходных случайных величин. Однако несколько первых главных компонент обычно учитывает большую часть общей изменчивости. В практике число p обычно выбирается значительно меньшим, чем m. Таким образом, метод главных компонент позволяет значительно сократить число случайных величин без существенной потери информации об изменчивости. Геометрическая интерпретация главных компонент весьма проста. Главные компоненты являются осями координат m-мерного эллипсоида, образованного точками, соответствующими результатам наблюдений в m-мерном пространстве. Выбор первых p главных компонент из набора m переменных соответствует ортогональной проекции результатов наблюдения на p-мерное подпространство, осями координат которого являются p главных компонент.

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Метод главных компонент проще понять, если сравнить его с методом линейной регрессии, который рассмотрен выше. Простейшая регрессионная модель определяется уравнением

$$y = \alpha + \beta x,$$

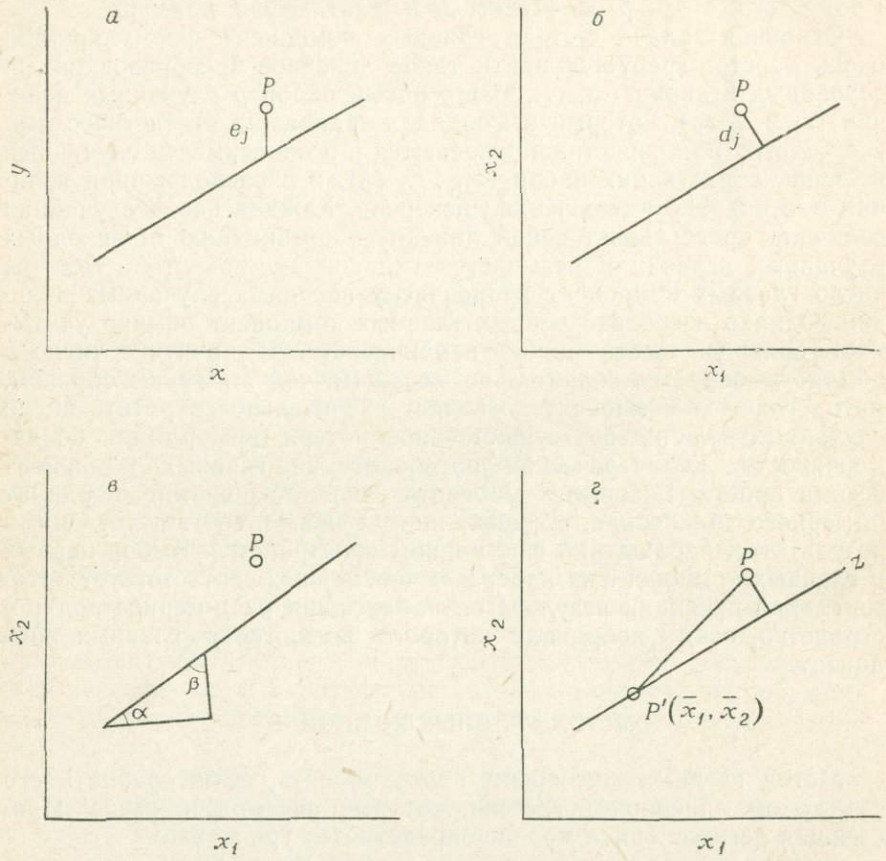
где y — функция, x — аргумент, а α и β — неизвестные параметры. В случае n наблюдений это уравнение для наблюдения с номером j примет следующий вид:

$$y_j = a + bx_j + e_j,$$

где e_j — случайная величина с некоторым заданным распределением. Величины a и b являются оценками неизвестных параметров α и β, которые находят методом наименьших квадратов путем минимизации суммы:

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

Величина e_j, соответствующая одному наблюдению, показана графически на фиг. 1. Главные компоненты являются линейными



Ф и г. 1. Геометрическая интерпретация метода главных компонент.

а — линейная регрессионная модель; *б* — модель главной компоненты; *в* — косинусы направления; *г* — ортогональная проекция.

функциями исходных случайных величин, и для случая двух переменных описывающая их модель примет следующий вид:

$$z = \alpha x_1 + \beta x_2, \tag{1}$$

где z — главная компонента, x_1 и x_2 — исходные случайные величины, а α и β — неизвестные параметры. В случае n наблюдений квадрат результата наблюдения с номером j можно представить следующей формулой:

$$h_j^2 = (ax_{1j} + bx_{2j})^2 + d_j^2,$$

где

$$h_j^2 = x_{1j}^2 + x_{2j}^2,$$

а d_j — случайная компонента, соответствующая наблюдению с номером j. Коэффициенты a и b, являющиеся оценками для α и β, находят путем минимизации суммы квадратов:

$$\sum_{j=1}^{n} d_j^2 = \sum_{j=1}^{n} [h_j^2 - (ax_{1j} + bx_{2j})^2]. \qquad (2)$$

Графическая интерпретация величины d_j приведена на фиг. 1, *б*. Для того чтобы избежать неоднозначных решений, необходимо ввести условие

$$a^2 + b^2 = 1.$$

Это приводит к тому, что величины a и b можно рассматривать как косинусы углов, показанных на фиг. 1, *в*. Прямая линия, изображенная на этом чертеже, соответствует уравнению первой главной компоненты и проходит через центр тяжести фигуры, описывающей область расположения точек, соответствующих результатам наблюдений. Таким образом, уравнение (1) можно записать в более удобной форме

$$z = \alpha(x_1 - \mu_1) + \beta(x_2 - \mu_2), \qquad (3)$$

где μ_1 и μ_2 — неизвестные истинные средние значения случайных величин x_1 и x_2 соответственно. Из уравнения (3) следует, что уравнение (2) можно записать в следующем виде:

$$\sum_{j=1}^{n} d_j^2 = \sum_{j=1}^{n} \{h_j^2 - [a(x_{1j} - \bar{x}_1) + b(x_{2j} - \bar{x}_2)]^2\}, \qquad (4)$$

где

$$h_j^2 = (x_{1j} - \bar{x}_1)^2 + (x_{2j} - \bar{x}_2)^2.$$

Это отклонение изображено графически на фиг. 1, *г*. Так как $\sum_{j=1}^{n} h_j^2$ является константой для данных n наблюдений, минимизация $\sum_{j=1}^{n} d_j^2$ в уравнении (4) равносильна максимизации выражения

$$\sum_{j=1}^{n} [a(x_{1j} - \bar{x}_1) + b(x_{2j} - \bar{x}_2)]^2.$$

Нетрудно видеть, что максимизируемая функция есть не что иное, как выборочная дисперсия случайной величины z, определенной выражением (3), так как $E(z) = 0$. Найденные значения коэффициентов a и b соответствуют максимальной изменчивости данного набора наблюдений.

Для случая m переменных уравнение (4) можно записать в обобщенном виде

$$\sum_{j=1}^{n} d_j^2 = \sum_{j=1}^{n} \sum_{i=1}^{m} (x_{ij} - \bar{x}_i)^2 - \left[\sum_{i=1}^{m} a_i (z_{ij} - \bar{x}_i) \right]^2, \tag{5}$$

где

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

и

$$\sum_{i=1}^{m} a_i^2 = 1.$$

Так как в выражении (5) величина

$$\sum_{j=1}^{n} \sum_{i=1}^{m} (x_{ij} - \bar{x}_i)^2$$

является константой для данной выборки, мы можем свести задачу к максимизации величины

$$\left[\sum_{i=1}^{m} a_i (z_{ij} - \bar{x}_i) \right]^2,$$

являющейся выборочной дисперсией линейной комбинации

$$z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m, \tag{6}$$

где x_i — случайные величины с математическим ожиданием, равным нулю, а для коэффициентов α_i выполнено равенство

$$\sum_{i=1}^{m} \alpha_i^2 = 1.$$

Таким образом, мы получили обобщенное выражение первой главной компоненты.

Эти зависимости можно значительно проще выразить в матричной форме. Если $\mathbf{x}$ и $\boldsymbol{\alpha}$ рассматривать как p-мерные векторы-столбцы, то первую главную компоненту можно представить так:

$$z = \boldsymbol{\alpha}' \mathbf{x},$$

где $\boldsymbol{\alpha}'$ — транспозиция $\boldsymbol{\alpha}$. Дисперсия, которую мы максимизируем, определяется следующим выражением:

$$\operatorname{var}(z) = (\boldsymbol{\alpha}' \mathbf{x})^2 = \boldsymbol{\alpha}' \mathbf{x} \mathbf{x}' \boldsymbol{\alpha},$$

при условии, что

$$\alpha'\alpha=1.$$

Необходимо отметить, что произведение $\mathbf{xx}'$ представляет собой выборочную ковариационную матрицу, элементы которой вычислены по результатам n наблюдений. Напомним также, что каждое наблюдение является m-мерным вектором и наша выборка представляет собой матрицу порядка $m\times n$.

Процедуру вычисления второй главной компоненты мы рассмотрим на примере двумерной случайной величины. Эту компоненту мы запишем в виде следующего выражения:

$$z_2=\beta_1 x_1+\beta_2 x_2,$$

где x_1 и x_2 — те же случайные величины, что и при вычислении случайной величины z_1, которая определена равенством

$$z_1=\alpha_1 x_1+\alpha_2 x_2$$

и является первой главной компонентой. На коэффициенты α и β накладываются следующие ограничения:

1) $\alpha_1^2+\alpha_2^2=1$,

2) $\beta_1^2+\beta_2^2=1$,

3) $\alpha_1\beta_1+\alpha_2\beta_2=0$.

Эти условия означают, что векторы (α_1, α_2) и (β_1, β_2) ортогональны. Вторая главная компонента определяется как линейная функция x_1 и x_2, имеющая максимальную дисперсию при условии независимости от первой главной компоненты. Как следствие этого сумма квадратов расстояний выборочных точек на плоскости, определенной базисными векторами α и β, будет минимальна. Это можно записать в виде следующего выражения:

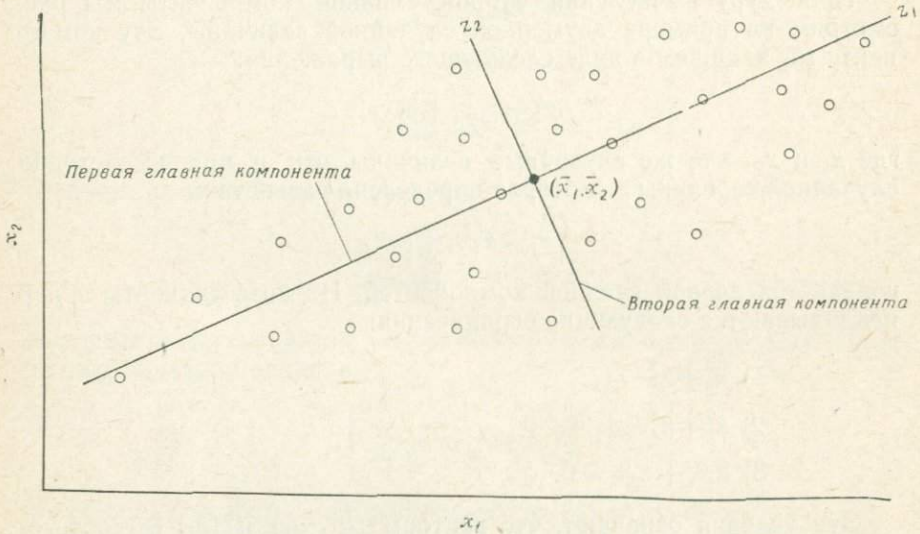
$$\min\sum_{j=1}^{n} d_j^2=\min\sum_{j=1}^{n}\{h_j^2-[a_1(x_{1j}-\bar{x}_1)+a_2(x_{2j}-\bar{x}_2)]^2-$$
$$-[b_1(x_{1j}-\bar{x}_1)+b_2(x_{2j}-\bar{x}_2)]^2\}, \qquad (7)$$

где ограничения для векторов $\mathbf{a}$ и $\mathbf{b}$ те же, что и для векторов $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$. В данном случае, когда мы имеем дело только с двумя случайными величинами, коэффициенты b_1 и b_2 можно получить сразу после вычисления параметров первой главной компоненты. Соотношение первой и второй главных компонент показано на фиг. 2, из которой видно, что процедура их нахождения сводится к повороту системы координат.

Для m случайных величин уравнение (7) можно записать в следующем обобщенном виде:

$$\min \sum_{j=1}^{n} d_j^2 = \min \sum_{j=1}^{n} \left\{ \sum_{i=1}^{n} (x_{ij} - \bar{x}_i)^2 - \left[\sum_{i=1}^{m} a_i (x_{ij} - \bar{x}_i) \right]^2 - \right.$$

$$\left. - \left[\sum_{i=1}^{m} b_i (x_{ij} - \bar{x}_i) \right]^2 \right\}, \qquad (8)$$



Фиг. 2. Оси, представленные главными компонентами в случае двумерных данных.

где x_{ij} — результаты j-ого измерения i-ой случайной величины. Налагаемые ограничения записываются в следующем виде:

1) $\sum a_i = 1$,
2) $\sum b_i = 1$,
3) $\sum a_i b_i = 0$.

Как и в случае первой главной компоненты, определенной равенством (6), можно свести задачу к максимизации двух последних слагаемых в первой части уравнения (8) и записать модель второй главной компоненты в следующем виде:

$$z = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

где набор коэффициентов β_i максимизирует дисперсию

$$\mathrm{var}(z) = (\beta' \mathbf{x})^2 = \beta' \mathbf{x}\mathbf{x}' \beta$$

при условии

$$\beta'\beta=1$$

и

$$\beta'\alpha=0.$$

Теперь мы рассмотрим наиболее общий случай нахождения p главных компонент для m случайных величин по выборке объема n. Обозначим через a_{ik} коэффициент для случайной величины с номером i в главной компоненте с номером k. Оценку для α_{ik} обозначим a_{ik}. Тогда общее выражение, которое требуется минимизировать, будет

$$\sum_{j=1}^{n} d_j^2=\sum_{j=1}^{n}\sum_{i=1}^{p}(x_{ij}-\bar{x}_i)^2-\sum_{k=1}^{p}\left[\sum_{i=1}^{k}a_{ik}(x_{ij}-\bar{x}_i)\right]^2 \qquad (9)$$

при условии

$$\sum_{i=1}^{m}a_{ij}a_{ik}=\delta_{jk},$$

где

$$\delta_{ik}=\begin{cases}1 \text{ при } j=k,\\ 0 \text{ при } j\neq k.\end{cases}$$

Геометрическая интерпретация этих выражений сводится к тому, что сумма квадратов расстояний в линейном подпространстве, определенном первыми k главными компонентами, будет удовлетворять уравнению (9), но не обеспечит выполнения условия независимости случайных величин.

МЕТОДЫ РЕШЕНИЯ

Для того чтобы показать, как определяются главные компоненты, мы сначала рассмотрим процедуру нахождения первой главной компоненты в наборе m случайных величин по выборке объема n. Из выражения (9) следует, что функцией, которую нужно максимизировать, будет

$$\sum_{j=1}^{n}\sum_{i=1}^{m}a_i(x_{ij}-\bar{x}_i)^2 \qquad (10)$$

при условии, что

$$\sum_{i=1}^{m}a_i^2=1. \qquad (11)$$

Для того чтобы определить коэффициенты a_i, которые максимизируют выражение (10) и в то же время удовлетворяют

условию (11), введем множитель Лагранжа λ и из выражения (10) получим следующее уравнение:

$$F=\sum_{j=1}^{n}\left[\sum_{i=1}^{m} a_i(x_{ij}-\bar{x}_i)\right]^2+\lambda\left(1-\sum_{i=1}^{m} a_i^2\right). \qquad (12)$$

Взяв частные производные из F по a_i и приравняв их к нулю, получим следующую систему уравнений:

$$\frac{\partial F}{\partial a_1}=a_1\left[\sum(x_{1j}-\bar{x}_1)^2-\lambda\right]+\ldots+a_m\sum(x_{1j}-\bar{x}_1)(x_{mj}-\bar{x}_m)=0,$$

$$\frac{\partial F}{\partial a_2}=a_1\sum(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)+\ldots+a_m\sum(x_{2j}-\bar{x}_2)\times$$

$$\times(x_{mj}-\bar{x}_m)=0, \qquad (13)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\frac{\partial F}{\partial a_m}=a_1\sum(x_{1j}-\bar{x}_1)(x_{mj}-\bar{x}_m)+\ldots+a_m\left[\sum(x_{mj}-\bar{x}_m)^2-\lambda\right]=0.$$

Полученная система однородных линейных уравнений имеет нетривиальное решение при условии, что матрица, образованная коэффициентами при неизвестных, имеет детерминант, равный нулю. Обозначив

$$\mathbf{C}=\{c_{ij}\}=\{\sum(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)\}, \qquad (14)$$

получим характеристическое уравнение матрицы **C**:

$$|\mathbf{C}-\lambda\mathbf{I}|=0,$$

где **I** — единичная матрица. В данном случае матрица **C** является положительно определенной, из чего следует, что существует m действительных положительных корней λ. Каждому из характеристических корней соответствует характеристический вектор матрицы **C**. Далее мы выбираем наибольшее значение $\lambda_{\max}$. Это является следствием того, что если мы умножим каждое уравнение с номером i в системе (13) на a_i, а затем сложим все уравнения, то получим выражение, которое требуется максимизировать в равенстве (10), т. е.

$$\lambda=\sum_{j=1}^{n}\left[\sum_{i=1}^{m} a_i(x_{ij}-\bar{x}_i)\right]^2. \qquad (15)$$

Иными словами, наибольший корень является дисперсией первой главной компоненты.

Для того чтобы найти вторую главную компоненту, требуется определить другой набор коэффициентов, которые минимизируют

выражение (12) при условии, что вторая главная компонента не зависит от первой. Это условие можно определить выражением

$$\sum_{j=1}^{n}\left[\sum_{i=1}^{m} a_i(x_{ij}-\bar{x}_i)\right]\left[\sum_{i=1}^{m} b_i(x_{ij}-\bar{x}_i)\right]=0.$$

Тогда функция, которую требуется минимизировать, примет следующий вид:

$$F=\sum_{j=1}^{n}\left[\sum_{i=1}^{m} b_i(x_{ij}-\bar{x}_i)\right]^2+\lambda\left(1-\sum_{i=1}^{m} b_i\right)^2+$$
$$+\mu\sum_{j=1}^{n}\left[\sum_{i=1}^{m} a_i(x_{ij}-\bar{x}_i)\right]\left[\sum_{i=1}^{m} b_i(x_{ij}-\bar{x}_i)\right], \qquad (16)$$

где λ и μ — множители Лагранжа, а b_i — набор коэффициентов второй главной компоненты. Взяв частные производные от F по каждому b_i, приравняв их нулю и решив полученные уравнения относительно b_i, получим набор коэффициентов для второй главной компоненты. Нетрудно показать [1], что $\mu=0$ и что набор коэффициентов, удовлетворяющий уравнению (16), есть собственный вектор, соответствующий второму наибольшему значению матрицы **C**, которым является λ, определенная выражением (16). Таким образом, последовательно подбирая собственные значения и связанные с ними собственные векторы матрицы **C**, можно вычислить оценки параметров всех главных компонент. Иногда встречается ситуация, когда удается определить не все главные компоненты, например в случае кратных собственных значений, что для практики нетипично.

Значительно проще сформулировать задачу нахождения главных компонент, используя матричные обозначения. Определим матрицу **C**, согласно равенству (14), следующей формулой:

$$\mathbf{C}=\{c_{ij}\}=\frac{1}{n-1}\sum_{k}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$$

и получим ковариационную матрицу, собственные значения и собственные векторы которой определяются преобразованием

$$\mathbf{P}'\mathbf{C}\mathbf{P}=\mathbf{D},$$

где **D** — диагональная матрица, диагональными элементами которой являются собственные числа матрицы **C**, а соответствующие им собственные векторы представлены векторами-столбцами матрицы **P**. Для вычисления матриц **P** и **D**, соответствующих данной матрице **C**, существует ряд широко известных алгоритмов.

Рассмотрим ситуацию, когда случайные величины центрированы и нормированы, т. е. имеют средние значения, равные нулю, и дисперсии, равные единице. Кроме того, для таких величин ковариационная матрица равна корреляционной.

Рассмотрим произвольную линейную комбинацию, образованную набором случайных величин $\mathbf{y}$; соответствующий этой комбинации вектор-столбец коэффициентов обозначим через $\mathbf{b}$. Коэффициент корреляции между любой случайной величиной x_i и набором случайных величин $\mathbf{y}$ определяется выражением

$$\rho(x_i\mathbf{y})=\frac{\mathbf{b}'\mathbf{c}_i}{\mathbf{b}'\mathbf{C}\mathbf{b}}, \tag{17}$$

где $\mathbf{c}_i$ — вектор-столбец ковариационной матрицы $\mathbf{C}$, который соответствует x_i. Если квадраты коэффициентов корреляции, соответствующие случайным величинам x_i и набору случайных величин $\mathbf{y}$, сложить, то получим квадрат обобщенного коэффициента корреляции

$$T=\sum_{i=1}^{m}\alpha^2(x_i,\ \mathbf{y})=\sum_{i=1}^{m}\frac{(\mathbf{b}'\mathbf{c}_i)^2}{\mathbf{b}'\mathbf{C}\mathbf{b}}=\frac{\mathbf{b}'\mathbf{C}\mathbf{C}\mathbf{b}}{\mathbf{b}'\mathbf{C}\mathbf{b}}. \tag{18}$$

Разделив выражение (18) на m, получим средний квадрат коэффициента корреляции между $\mathbf{y}$ и исходными случайными величинами. Рао [77] показал, что линейная комбинация, которая максимизирует это выражение, является первой главной компонентой корреляционной матрицы. В терминах корреляционной матрицы первую главную компоненту можно определить выражением

$$\lambda_{\max}=\mathbf{a}'\mathbf{C}\mathbf{a}.$$

В этом случае

$$\mathbf{b}'\mathbf{c}_i=a_i\lambda$$

и

$$\mathbf{b}'\mathbf{C}\mathbf{b}=\lambda,$$

где a_i — i-ый элемент собственного вектора, соответствующего наибольшему собственному значению, которое равно λ.
Тогда

$$T_{\text{гл. комп.}}=\sum_{i=1}^{m}\frac{a_i^2\lambda^2}{\lambda}=\lambda$$

как максимальное значение T. Так как $1<\lambda<m$, λ/m представляет собой количественную меру силы взаимной корреляции внутри исходных случайных величин.

ИЗУЧЕНИЕ ИЗМЕНЧИВОСТИ ЧЕРЕПАХ

В статье Джоликойера и Мосимена [43] предложены результаты изучения изменчивости черепах вида *Chrysemus picta marginata*. Основная цель статьи — исследование возможностей применения метода главных компонент для изучения изменчивости размера и формы в группах ныне живущих организмов. Часть собранных данных представляет собой результаты измерения длины, ширины и высоты панцирей 24 самцов (см. табл. 1). Эти данные можно с успехом использовать для иллюстрации применения метода главных компонент.

Таблица 1

Результаты измерений панциря 24 самцов черепах, мм

Длина	Ширина	Высота	Длина	Ширина	Высота
93	74	37	116	90	43
94	78	35	117	90	41
96	80	35	117	91	41
101	84	39	119	93	41
102	85	38	120	89	40
103	81	37	120	93	44
104	83	39	121	95	42
106	83	39	125	93	45
107	82	38	127	96	45
112	89	40	128	95	45
113	88	40	131	95	46
114	86	40	135	106	47

Итак, мы располагаем выборкой, состоящей из 24 измерений трех характеристик панциря. Обозначим через x_1 длину, через x_2 — ширину и через x_3 — высоту панциря. Вычисленные по этим данным средние значения и ковариационные матрицы приведены в табл. 2. Элементы выборочной ковариационной матрицы вычислялись по следующей формуле:

$$c_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j),$$

где $\bar{x}_i$ и $\bar{x}_j$ — средние арифметические, соответствующие признакам с номерами i и j.

Для определения главных компонент мы сначала находим собственные векторы ковариационной матрицы, которые вместе с соответствующими собственными числами приведены в табл. 3. Дисперсия каждой главной компоненты равна соответствующему собственному значению, которое характеризует суммарный эффект изменчивости каждой главной компоненты. Для нас наибольший

Таблица 2

Средние и ковариационные матрицы результатов измерений панциря 24 самцов черепах, мм

Длина	Ширина	Высота
113,37	88,29	40,71
138,77	79,15	37,38
79,15	50,04	21,65
37,38	21,65	11,26

интерес представляют главные компоненты с большими дисперсиями. Поэтому в табл. 3 главные компоненты расположены в порядке уменьшения дисперсий. Если каждое собственное значение разделить на общую дисперсию, равную сумме дисперсий исходных случайных величин, мы получим доли изменчивости, приходящиеся на каждую главную компоненту. Эти значения в виде процентов приведены в нижней части табл. 3.

Таблица 3

Главные компоненты результатов измерения панцирей черепах, мм

	P_1	P_2	P_3
Длина	0,840	—0,488	—0,236
Ширина	0,492	0,869	—0,047
Высота	0,228	—0,077	0,970
λ	195,28	3,69	1,10
Изменчивость, %	97,6	1,8	0,6

Для интерпретации полученных главных компонент следует проанализировать элементы каждого из собственных векторов. В нашем примере первая главная компонента охватывает 98% общей изменчивости и обладает только положительными коэффициентами. Это дает основание считать, что первая главная компонента является обобщенной мерой размеров панциря и содержит наибольшую долю изменчивости. Собственному вектору, характеризующему вторую главную компоненту, соответствует менее 2% общей изменчивости. Кроме того, два ее наибольших компонента имеют противоположные знаки, а третий коэффициент весьма близок к нулю. Это позволяет считать, что вторая главная компонента является характеристикой формы панциря и учитывает главным образом его длину и ширину. На долю третьей главной ком-

поненты приходится менее 1% общей изменчивости, и ее также можно рассматривать как характеристику формы панциря, учитывающую в основном различия между его длиной и высотой.

В процессе применения метода главных компонент полезно подсчитать значения координат результатов наблюдений в новой системе, которую образуют главные компоненты. Так, например, для наблюдения с номером k i-ая координата z_{ik} будет определена выражением

$$z_{ik}=\sum_{j=1}^{m} P_{ij}(x_{jk}-\bar{x}_j),$$

где P_{ij} — коэффициент j-ой случайной величины в i-ой главной компоненте. Эти значения координат вычисляются при условии, что их начало расположено в точке, соответствующей набору выборочных средних. В табл. 4 приведены величины z_{ik}, соответствующие исходным значениям x_{ik}. Построенные по ним гистограммы или точечные диаграммы на плоскости могут дать дополнительную информацию, которую можно использовать при классификационных построениях.

Таблица 4

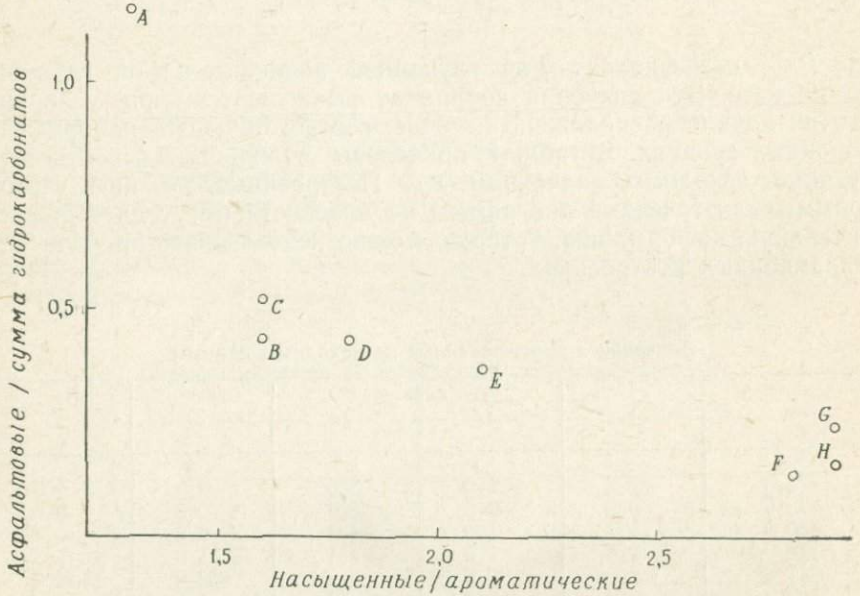
Значения z_{ik}, вычисленные по исходным данным

z_1	z_2	z_3	z_1	z_2	z_3
—25,0	—2,2	1,9	3,6	0,0	1,5
—22,6	0,9	—0,5	4,0	—0,3	—0,6
—20,0	1,7	—1,0	4,4	0,6	—0,7
—12,9	2,4	1,5	7,1	1,3	—1,3
—11,8	2,9	0,2	5,8	—2,6	—2,3
—13,2	—1,0	—0,8	8,6	0,6	1,4
—10,9	0,1	0,8	10,0	2,0	—0,9
—9,2	—0,9	0,3	13,1	—1,9	1,2
—9,1	—2,2	—0,8	16,2	—0,3	0,6
—1,0	1,3	—0,4	16,6	—1,6	0,4
—0,6	0,0	—0,6	19,3	—3,2	0,6
—0,8	—2,2	—0,7	28,3	4,4	0,2

АНАЛИЗ СЫРОЙ НЕФТИ

В процессе изучения сырой нефти возникла задача, связанная с классификацией восьми проб. Из восьми взятых проб две были заведомо отобраны из двух различных продуктивных зон — верхней и нижней, и задача заключалась в том, чтобы каждую из оставшихся проб отнести к верхней или нижней зоне. Состав двух проб нефти из этих зон отчетливо различался, что позволило провести дальнейшую классификацию. Результаты анализа восьми

проб, обозначенных буквами от A до H, приведены в табл. 5. В каждой пробе измерялось два физических свойства и определялись содержания двадцати различных составляющих нефти, которые выражены в весовых процентах. Эти двадцать два свойства можно рассматривать как случайные величины. Результаты более раннего исследования данных табл. 5 приведены на фиг. 3. При построении этой фигуры были использованы признаки с 5 по 21. По оси абсцисс откладывались значения отношений содержания насыщенных смол к содержанию ароматических смол, а по оси



Фиг. 3. Зависимость между отношением насыщенных фракций к ароматическим и отношением асфальтовых фракций к сумме гидрокарбонатов.

ординат — отношения содержаний асфальтов и гидрокарбонатов. Выбор этих отношений для изучения проб сырой нефти основан на априорных сведениях и более ранних исследованиях. На фиг. 3 отчетливо видно, что восемь точек, соответствующих пробам, разделились на три группы. Пробы F, G и H образовали группу, соответствующую нижней продуктивной зоне, пробы B, C, D и E — группу из верхней зоны, а проба A, отделившаяся от них, позволяет предположить наличие неописанной третьей продуктивной зоны.

В данном случае задача заключается в выявлении возможности построения такой характеристики эмпирических данных, которую можно было бы использовать для классификации. Просмотр всей имеющейся информации по этим пробам показывает, что

Таблица 5

Физические и химические свойства в восьми пробах сырой нефти (по Кампу)

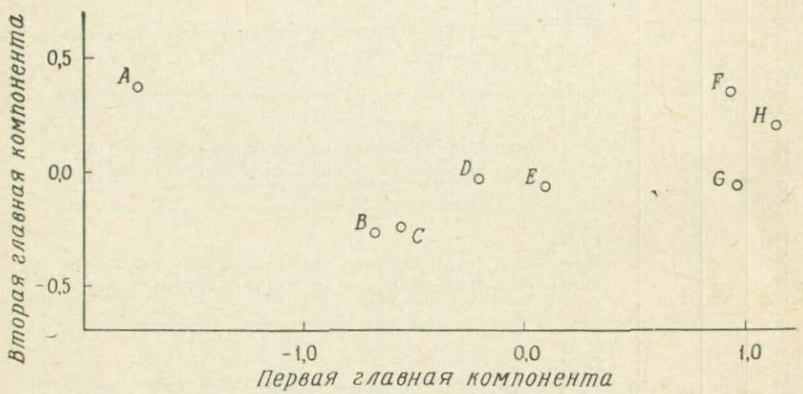
Свойства	Пробы							
	A	*B*[1]	*C*	*D*	*E*	*F*[2]	*G*	*H*
1. Удельный вес	0,9516	0,8936	0,8774	0,8805	0,8680	0,8231	0,8370	0,8161
2. R. I.	1,5400	1,5010	1,4920	1,4935	1,4860	1,4615	1,4675	1,4570
3. Общее содержание серы	2,39	1,46	1,37	0,52	0,34	0,49	0,11	0,34
4. Насыщенные смолы	89,6	83,8	77,0	81,8	82,7	67,1	73,3	60,4
5. Легкие смолы	27,8	15,0	19,0	17,8	17,2	6,4	12,1	8,3
6. Тяжелые смолы	11,5	9,6	9,1	8,9	8,6	4,6	6,0	4,6
7. Асфальтены	14,5	5,3	6,2	3,3	1,3	0,6	1,2	0,2
8. Жидкие углеводороды	44,7	66,8	61,4	65,4	65,7	83,4	73,4	79,7
9. Твердые углеводороды	1,5	3,3	4,3	4,6	7,2	5,0	7,3	7,2
10. Щелочи	6,1	15,9	17,2	17,8	20,2	25,3	29,2	27,9
11. Неконденсированные циклощелочные соединения	30,2	28,2	29,1	30,6	30,6	31,9	30,3	31,4
12. Конденсированные циклощелочные соединения	19,9	17,4	15,8	16,6	16,6	16,3	15,0	15,1
13. Бензолы	11,7	10,9	10,7	10,2	9,0	8,1	8,4	8,6
14. Инданы	5,8	4,6	4,9	4,6	4,1	3,2	3,1	2,7
15. Динартенбензолы	3,2	2,5	2,4	2,1	2,1	1,6	1,4	1,5
16. Нафталины	5,8	5,8	5,6	5,1	5,0	3,9	4,0	4,0
17. Фенантрены	2,6	2,2	2,2	2,0	1,8	1,5	1,3	1,3
18. Пирены	3,5	2,9	2,4	2,5	2,3	1,6	1,5	1,5
19. Хризены	1,1	0,8	0,9	0,7	0,6	0,5	0,5	0,5
20. Аценафтины	5,5	4,7	4,6	4,3	4,1	3,5	2,9	3,2
21. Аценафтены	4,6	3,9	3,8	3,5	3,4	2,5	2,3	2,3
22. Сера связанная	6,8	6,0	6,7	5,7	5,0	3,7	3,7	3,4

[1] Верхняя продуктивная зона.
[2] Нижняя продуктивная зона.

сделанные выводы можно несколько расширить или по крайней мере сократить возможность пропуска какой-либо зависимости.

В данном примере изучаемые свойства измеряются различными единицами. Однако для применения метода главных компонентов необходимо все измерения сделать сопоставимыми. Это можно сделать путем деления каждого наблюдаемого значения случайной величины на соответствующее стандартное отклонение. Ковариационная и корреляционная матрицы для таких нормированных величин совпадают.

В табл. 6 приведены собственные векторы для первой и второй главных компонент. Эти компоненты совместно учитывают 93% общей изменчивости. Результаты анализа проб, пересчитанные для



Фиг. 4. Рассчитанное положение точек относительно двух первых главных компонент.

За единицу измерения принята величина стандартного отклонения первой главной компоненты.

системы координат, из этих двух главных компонент графически представлены на фиг. 4. Сравнивая фиг. 3 и 4, легко видеть, что относительное расположение точек на них приблизительно одинаково. Первая главная компонента, учитывающая 88% общей изменчивости, позволяет классифицировать пробы с потерей только 12% информации об изменчивости, но при этом число переменных сокращается с 22 до 1.

Можно ли дать какое-либо физическое истолкование первой главной компоненте? Нетрудно видеть, что ее значения хорошо коррелируются с отношением насыщенных составляющих к ароматическим. Если это отношение рассматривать как меру относительного количества парафинов, содержащихся в нефти, тогда и первую главную компоненту можно считать характеристикой парафиноносности. Теперь рассмотрим коэффициенты, соответствующие первой главной компоненте, приведенные в табл. 6. Свойства ко-

Таблица 6

Первая и вторая главные компоненты для результатов анализа проб сырой нефти

Свойства	Главные компоненты		Свойства	Главные компоненты	
	P_1	P_2		P_1	P_2
1	—0,22	0,60	12	—0,20	0,40
2	—0,22	0,11	13	—0,22	—0,12
3	—0,21	0,17	14	—0,22	—0,03
4	—0,20	—0,14	15	—0,22	0,07
5	—0,21	0,03	16	—0,21	—0,29
6	—0,22	—0,16	17	—0,22	—0,03
7	—0,21	0,25	18	—0,23	—0,04
8	0,21	—0,04	19	—0,22	0,04
9	0,20	—0,22	20	—0,22	0,04
10	0,22	—0,13	21	—0,22	—0,05
11	0,14	0,68	22	—0,22	—0,21

торым соответствуют положительные коэффициенты первой главной компоненты, положительно коррелируются с общим содержанием парафинов. Из этого следует, что пробы сырой нефти так расположены в табл. 5, что соответствующие им значения первой главной компоненты возрастают слева направо. Свойства с номерами от 8 до 11 обладают тенденцией к увеличению значений слева направо в ряду проб, тогда как для всех остальных свойств характерна обратная тенденция. Таким образом, компоненты состава с 8 по 11 ведут себя как парафины, а остальные — как непарафины. Рассмотрим, например, свойство с номером 2 — показатель преломления нефти. Хорошо известно, что увеличение содержания парафинов снижает значения этой характеристики. Аналогичные объяснения можно дать и для поведения других свойств.

Подводя итог, можно отметить, что полученная нами характеристика для классификации проб сырой нефти основана на содержании парафинов. Она учитывает главные различия в физических свойствах и химическом составе изучаемой сырой нефти, и можно показать, что аналогичные зависимости наблюдаются у весьма широкого класса сырых нефтей.

КАНОНИЧЕСКИЕ ПЕРЕМЕННЫЕ

В ряде задач нередко необходимо изучить обобщенные зависимости между двумя наборами случайных величин. В частности, может потребоваться найти в обоих наборах такие линейные комбинации, которые обладали бы максимальным коэффициентом корреляции. Такая информация весьма ценна для получения выводов

по комплексу случайных величин, когда структура этого комплекса определена нечетко. Метод решения этой задачи разработан Хотелингом [38] и называется методом канонической корреляции.

Рассмотрим только линейную комбинацию одного набора случайных величин, обладающую максимальным коэффициентом корреляции, с линейной комбинацией другого набора случайных величин, который в частном случае может быть представлен одной случайной величиной. На этих линейных комбинациях, называемых первыми каноническими переменными, мы и сосредоточим наше внимание. Сначала рассмотрим общий метод получения первых канонических переменных, а затем пример, когда один из наборов случайных величин представлен только одной величиной. Подробное рассмотрение этого вопроса можно найти в книге Андерсона [1].

Рассмотрим p-мерный случайный вектор $\mathbf{x}$, ковариационную матрицу которого мы обозначим Σ. Допустим, что все случайные величины, образующие $\mathbf{x}$, имеют математическое ожидание, равное нулю, т. е. $E(\mathbf{x})=0$. Разделим вектор $\mathbf{x}$ на два подвектора $\mathbf{x}^{(1)}$ и $\mathbf{x}^{(2)}$ с числом случайных величин p_1 и p_2 соответственно, т. е.

$$\mathbf{x}=\begin{bmatrix}\mathbf{x}^{(1)}\\ \mathbf{x}^{(2)}\end{bmatrix}.$$

В результате такого разбиения вектора $\mathbf{x}$ мы получим два набора случайных величин $\mathbf{x}^{(1)}$ и $\mathbf{x}^{(2)}$. В соответствии с этим ковариационная матрица Σ будет разделена на четыре подматрицы, т. е.

$$\Sigma=\begin{bmatrix}\Sigma_{11}\,\Sigma_{12}\\ \Sigma_{21}\,\Sigma_{22}\end{bmatrix}.$$

Рассмотрим две произвольные линейные комбинации: $U=\boldsymbol{\alpha}'\mathbf{x}^{(1)}$, образованную компонентами вектора $\mathbf{x}^1$, и $V=\boldsymbol{\beta}'\mathbf{x}^{(2)}$, соответствующую $\mathbf{x}^{(2)}$. Из множества возможных линейных комбинаций U и V нужно выбрать такие, для которых коэффициент корреляции максимален. Так как коэффициент корреляции не зависит от любых линейных изменений размерности $\mathbf{x}^{(1)}$ и $\mathbf{x}^{(2)}$, выберем $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$ так, чтобы U и V обладали дисперсиями, равными 1, т. е.

$$\begin{aligned}\operatorname{var}(U)&=(\boldsymbol{\alpha}'\mathbf{x}^{(1)})^2=\boldsymbol{\alpha}'\,\Sigma_{11}\,\boldsymbol{\alpha}=1,\\ \operatorname{var}(V)&=(\boldsymbol{\beta}'\mathbf{x}^{(2)})^2=\boldsymbol{\beta}'\,\Sigma_{22}\,\boldsymbol{\beta}=1.\end{aligned}\qquad(19)$$

Так как $E(\mathbf{x}^{(1)})=E(\mathbf{x}^{(2)})=0$, $E(U)=E(V)=0$, что позволяет определить коэффициент корреляции для U и V выражением

$$\rho(U,\ V)=(\boldsymbol{\alpha}'\mathbf{x}^{(1)})(\boldsymbol{\beta}'\mathbf{x}^{(2)\prime})=\boldsymbol{\alpha}'\,\Sigma_{12}\,\boldsymbol{\beta},$$

которое требуется максимизировать при условии (19). Все вышеизложенное можно записать в виде простого выражения

$$\psi=\alpha'\,\Sigma_{12}\,\beta-\frac{1}{2}\lambda\,(\alpha'\,\Sigma_{11}\,\alpha-1)-\frac{1}{2}\mu\,(\beta'\,\Sigma_{22}\,\beta-1), \qquad (20)$$

где λ и μ — множители Лагранжа.

Андерсоном [1] показано, что, продифференцировав выражение (20) по α и β и приравняв полученные производные к нулю, получим матричные уравнения

$$\begin{aligned} -\lambda\,\Sigma_{11}\,\alpha+\Sigma_{12}\,\beta&=0,\\ \Sigma_{21}\,\alpha-\lambda\,\Sigma_{22}\,\beta&=0, \end{aligned} \qquad (21)$$

которые можно решить относительно λ, α и β с помощью итерационных методов. Нам необходимо рассмотреть только частный случай, когда $p_1=1$, что позволяет получить прямое решение.

Действительно, при $p_1=1$ как α, так и Σ_{11} являются скалярными величинами, и, кроме того,

$$\alpha=\Sigma_{11}^{\frac{1}{2}}.$$

Подставив это выражение в уравнения (21), можно определить β. Соответственно

$$\beta=\frac{\Sigma_{22}^{-1}\,\Sigma_{21}}{\Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}}.$$

Выражение для λ можно записать в общем виде:

$$\lambda=\left(\frac{\Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{11}}{\Sigma_{11}}\right)^{\frac{1}{2}},$$

где λ является мерой связи между двумя наборами случайных величин $\mathbf{x}^{(1)}$ и $\mathbf{x}^{(2)}$. В нашем случае при $p_1=1$ λ представляет собой множественный коэффициент корреляции. Так как $0<\lambda<1$, его можно использовать как меру близости набора аргументов и их функции.

ЛЕКЦИЯ 8

Классификация

Р. Мак-Кеммон

ЗНАЧЕНИЕ КЛАССИФИКАЦИИ

Задачу классификации можно рассматривать как поиск решения, позволяющего наилучшим образом разделить множество объектов, охарактеризованных многомерными наблюдениями, на более мелкие однородные подгруппы. Необходимо отметить, что задача в такой постановке коренным образом отличается от задачи дискриминантного анализа, заключающейся в классификационном отнесении изучаемого объекта к одной из заданных групп, число которых заранее определено. В нашей задаче число групп не известно и его требуется определить.

При решении задач классификации необходимо располагать некоторой мерой сходства одного объекта с другим. Чаще всего она определяется как мера сходства изучаемых свойств. Множество таких мер, пригодных как для количественных, так и для качественных данных, уже было описано в литературе [82, 2]. Допустим, что мы располагаем соответствующей мерой сходства.

Наиболее простой подход к решению задачи классификации набора объектов, охарактеризованных комплексом измеренных характеристик, заключается в выборе функции, отражающей внутреннюю однородность групп, полученных в результате решения изучаемого множества объектов для каждого варианта такого распределения. Для фиксированного числа групп можно выбрать такой вариант разделения совокупности объектов, который обеспечивает максимальную однородность в группах. Вариант распределения совокупности объектов, который обеспечивает максимальную однородность в группах, называется оптимальным. Однако, чтобы найти оптимальный вариант, необходимо перебрать все возможные варианты разделения. Число способов, которыми совокупность n объектов можно разделить на m групп, равно числу Стирлинга второго рода $S(n, m)$, определяемому [78] выражением

$$S(n, m)=\frac{1}{m!}\sum_{j=0}^{m}(-1)^j\binom{m}{j}(m-j)^n.$$

Даже для умеренных значений n величина $S(n, m)$ велика. В табл. 1 приведены примеры значений $S(n, m)$ для различных

чисел n объектов, разделенных на четыре группы. Из этой таблицы видно, что уже для совокупности, состоящей из 20 объектов, перебирать все варианты разделения на четыре части затруднительно даже электронной вычислительной машине. Ясно, что если число объектов принимает значения от нескольких сотен до тысячи, то такой подход явно неприемлем.

Таблица 1

Число вариантов разделения n объектов на четыре группы

n объектов	Число вариантов
5	10
10	34 105
15	42 335 950
20	45 232 115 901

Вместо набора всех вариантов разделения совокупности объектов на заданное число групп лучше воспользоваться последовательной процедурой, для которой найден оптимальный шаг. Надо полагать, что такая процедура потребует рассмотрения меньшего числа вариантов разбиения. Известно несколько таких классификационных процедур [58, 21, 91, 83, 60], рассмотрение которых и является нашей задачей [57, 55]. С решением задач классификации больших совокупностей объектов связаны и другие работы [57, 55], в которых наиболее интересны иерархические методы, базирующиеся на последовательном объединении объектов в группы. Исходное предположение при данном подходе к решению задач классификации заключается в том, что каждый объект рассматривается как отдельная группа. Таким образом, для n объектов в начале процедуры будет n групп. Первый шаг классификационной процедуры заключается в сокращении числа исходных групп на единицу за счет объединения двух наиболее сходных объектов в одну группу. Для того чтобы найти наиболее сходную пару из n объектов, необходимо рассмотреть $\binom{n}{2} = \frac{n(n-1)}{2}$ возможных пар, для которых следует провести соответствующие сравнения. После этого у нас останутся $n-1$ групп. После второго шага классификационной процедуры останутся $n-2$ групп. Необходимо отметить, что теперь нам придется сравнивать группы, состоящие из более чем одного наблюдения, и, следовательно, нужна мера сходства между группами. Допустим, что такая мера существует. Тогда процесс последовательного объединения групп будет продолжаться до тех пор, пока все n объектов объединятся в одну группу или пока ни будет достигнуто некоторое заданное критическое значение числа групп. Согласно теоремы Феллера [22], если проведены все $n-1$ шагов последовательного объединения, то это потребует $\binom{n+1}{3} = (n+1)n(n-)/6$ попарных сравнений. Такое число сравнений значительно меньше, чем при прямом решении задачи классификации.

В табл. 2 приведены значения числа парных сравнений, необходимых для объединения групп объектов объемом от 10 до 1000. Однако такие числа парных сравнений еще очень велики, и для проведения этой процедуры в больших совокупностях объектов потребуется очень много машинного времени. Кроме того, при работе с электронными вычислительными машинами возникает задача запоминания $n(n-1)/2$ значений меры сходства, вычисляемой по имеющимся данным на каждом этапе. В процедуре классификации эти значения потребуются для части или всех вычислений в зависимости от используемого алгоритма.

Таблица 2

Иерархические методы группировки

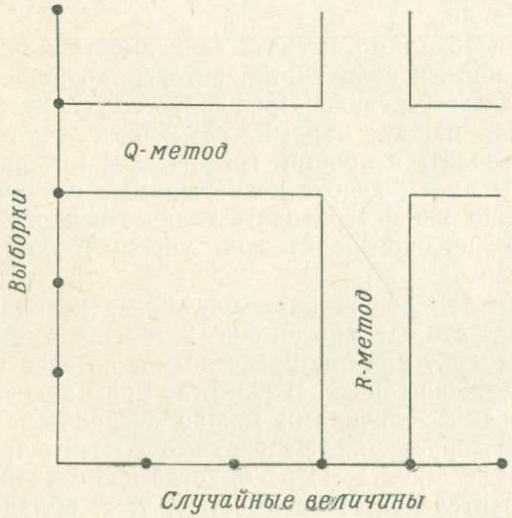
Число объектов n	Число рассмотренных пар $\binom{n+1}{3}$	Объем памяти $\binom{n}{2}$
10	165	45
50	20 825	1 225
100	166 650	4 950
200	1 333 300	19 900
500	20 444 250	124 750
1000	166 666 500	499 500

Таким образом, описанные методы классификации используют перебор вариантов. Несколько иной подход к решению задачи классификации связан с сравнением центров выборочных групп. При этом используется факторный анализ матрицы, элементами которой являются значения меры сходства, вычисленной для всех возможных пар объектов. В случае n объектов это будет матрица порядка $n \times n$. Такой подход называется Q-методом факторного анализа и детально рассмотрен в работе Кателла [9]. Несмотря на то что существует еще ряд задач, связанных с масштабным представлением данных [27], этот метод может служить основой для классификации объектов. Однако он требует введения ряда ограничений, связанных с порядком матрицы.

Теперь мы рассмотрим такой метод классификации, который не зависит от объема выборки и ликвидирует ограничения, накладываемые на число наблюдений в матрице исходных данных. Для того чтобы упростить рассмотрение метода, мы проведем его на примере фациальной классификации осадочных образований. На этом примере мы покажем основные принципы, лежащие в основе указанного метода. Тем не менее необходимо отметить, что его применимость не ограничивается рамками данного примера, а охватывает широкий класс задач классификации.

АССОЦИАТИВНЫЙ АНАЛИЗ

Как было показано выше, большинство обычных методов классификации ограничено в своем применении числом наблюдений в выборке. Эти методы можно рассматривать как относящиеся к группе *Q*-методов, что показано графически на фиг. 1. При этом сравнение результатов наблюдения основано на сходстве изучаемых свойств, которые рассматриваются как случайные величины. Наоборот, можно воспользоваться приемами, относящимися к группе *R*-методов, в которых сравниваются системы случайных величин на основе сходства между результатами наблюдения.



Фиг. 1. Диаграмма, иллюстрирующая основные методы корреляции.

На последнем методе мы и сосредоточим свое внимание, построив на его основе метод классификации, который не зависит от объема выборки.

При изучении экологии растений было установлено, что можно проводить классификацию выборочных данных на основании обобщенных типичных представителей различных видов. В связи с этим заимствованное из работы Гуделла [26] понятие специфической ассоциации получило дальнейшее развитие в методе группирования и классификации растений. Такой подход к решению задачи классификации был назван ассоциативным анализом и подробно описан Вильямсом и Ламбертом [93]. При этом растительные сообщества рассматриваются как комплексный организм, составными частями которого являются виды растений, взаимодействующие с различными характеристиками, определяющими общие экологические

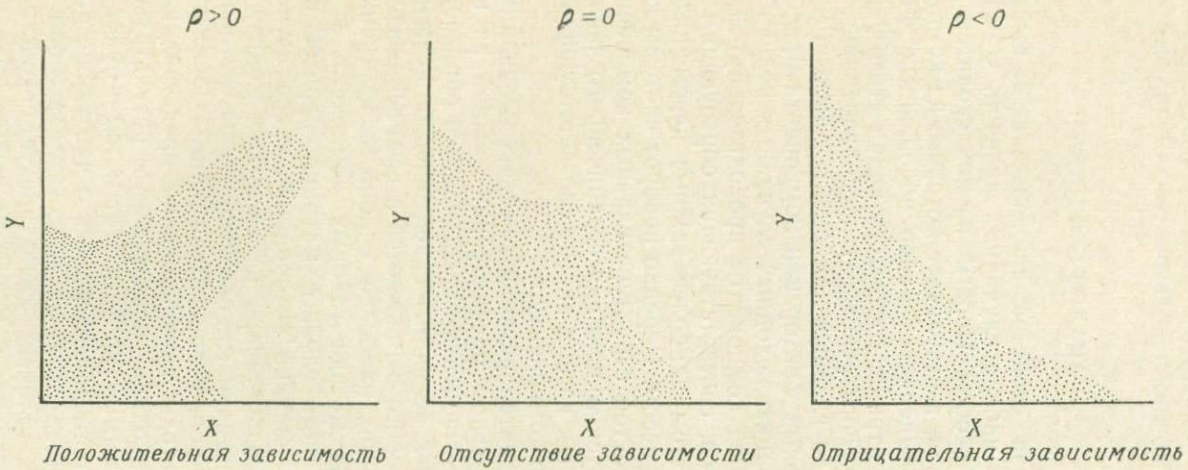
условия. Таким образом, эти характеристики определяются соответствующими взаимодействиями. Задача ассоциативного анализа заключается в выявлении этих взаимодействий, если они существуют.

Понятие комплексного организма можно перенести на структурные свойства и состав осадочных пород. Хотя последние и не являются организмами, их свойства можно рассматривать как характеристики, определяющие взаимодействие пород с условиями их образования, что позволяет провести аналогию с растительными видами. Конечно, было бы наиболее правильным использовать этот подход применительно к биогенным составляющим породы, которые характеризуют биологическую деятельность в момент осадкообразования.

Итак, нам необходимо выявить зависимости между рассматриваемыми случайными величинами, которые возникли за счет действия фациальных условий. Прежде всего следует рассмотреть взаимодействия между парами случайных величин, для чего можно воспользоваться простой графической интерпретацией. Допустим, что мы имеем дело с непрерывными случайными величинами и их можно охарактеризовать количественно. Это требование не обязательно, но оно значительно упрощает дальнейшее изложение.

Зависимость для любой пары случайных величин можно изобразить так, как это сделано на фиг. 2. Масштаб на этом чертеже не указан, поскольку для нас представляет интерес только относительное расположение точек. В качестве меры зависимости между двумя случайными величинами можно воспользоваться обычным коэффициентом корреляции. Необходимо отметить три группы значений этого коэффициента: 1) $\rho>0$, когда две случайные величины связаны положительной зависимостью, т. е. обладают одинаковыми тенденциями в сходных фациальных условиях; 2) $\rho=0$, т. е. случайные величины независимы; 3) $\rho<0$, когда случайные величины связаны отрицательной зависимостью, что отражает противоположные тенденции их поведения. Приведенные на фиг. 2 графики в значительной степени идеализированы и в таком виде на практике не встречаются. Однако при выборках большого объема получаемые точечные диаграммы оказываются похожими на графики, приведенные на фиг. 2.

Сформулируем нашу задачу в более общем виде, распространив ее на n случайных величин, поскольку такая модель лучше отражает ситуацию при фациальных исследованиях. Однако для представления n случайных величин потребуется использовать понятие n-мерного пространства, а это в свою очередь создает трудности применения обычных методов графического представления данных. Однако расположение точек в таком многомерном пространстве можно представить наглядно. В этом пространстве точки представляются в виде групп, характеризующихся определенным



Фиг. 2. Основные виды точечных диаграмм.

расположением относительно начала координат. В каждой из этих групп отмечается положительная зависимость между случайными величинами, что можно использовать для диагностики отдельных фаций.

АНАЛИЗ СХОДСТВА

Задача анализа сходства заключается в выявлении связей внутри заданного набора случайных величин. Обычно ее решают одним из двух способов. Первый из них сводится к применению факторного анализа, в результате которого из комплекса коррелируемых величин выделяется набор обобщенных факторов. Если ковариационная матрица, соответствующая этим факторам, диагональна, то мы получим результат преобразования, эквивалентный применению метода главных компонент. Каждый, выявленный таким образом главный фактор или компонента представляет собой линейную комбинацию исходных случайных величин. Интерпретация каждой главной компоненты проводится путем анализа весовых коэффициентов, соответствующих образующим ее случайным величинам. При этом наиболее важные переменные будут обладать наибольшим весом. Нередко поворот системы координат приводит к очень простой структуре главных компонент [61]. Интерпретация проводится только для ограниченного числа переменных, хотя вообще каждая из них вносит соответствующую долю изменчивости в главную компоненту.

Второй способ интерпретации комплекса случайных величин заключается в отдельном рассмотрении каждой из них. При этом каждая главная компонента представляется образованной подмножеством случайных величин. Такой способ называют анализом групп, задача которого заключается в выявлении групп случайных величин, обладающих наиболее сильными внутренними связями.

Как и в случае метода главных компонент, в анализе групп исходным материалом служит матрица парных коэффициентов корреляции, на основании которой из набора изучаемых случайных величин последовательно исключаются подгруппы, обладающие сильными внутренними связями. Вычислительные процедуры этого метода подробно описаны в книге Сокала и Сниса [83]. Иерархические методы группирования представляют собой интерес, так как позволяют представлять результаты группировки в форме дендрограммы [72]. Хотя этот метод и будет подробно рассмотрен в следующем разделе, здесь о нем также полезно сказать несколько слов. Дендрограмма представляет собой график, напоминающий дерево, который характеризует сходство между рассматриваемыми объектами. В нашем случае объектами являются изучаемые случайные величины. Дендрограмма имеет иерархическую структуру, в которой каждый иерархический уровень отражает степень внутренней однородности групп. Заметим, что

взаимно связанные группы в дендрограмме представлены как соседние. Используя этот метод, исследователь может довольно быстро получить любой из возможных вариантов группировки. Ограничение метода заключается в том, чтобы любая из случайных величин могла принадлежать только одной группе, а следовательно, группы являются взаимно исключающими. Однако, как будет показано ниже, можно так перестроить ковариационную матрицу, что окажется возможным выявить перекрывающиеся группы.

Описанные два типа анализа сходства можно так скомбинировать, что получится новый метод, который называется множественным анализом компонент. Сначала мы воспользуемся анализом групп для определения числа и состава положительно связанных групп случайных величин, а затем применим метод главных компонент для описания каждой из полученных групп. С помощью этого метода можно классифицировать неограниченное число наблюдений. При этом наше основное предположение заключается в том, что число выявленных фаций приблизительно равно числу положительно связанных групп случайных величин. Очевидно, что нейтральные случайные величины будут прежде всего характеризовать отдельные фации. Однако в процессе классификации станет ясно, что такие случайные величины увеличивают неопределенность в выборе групп, а следовательно, их необходимо устранить из рассмотрения.

МНОЖЕСТВЕННЫЙ АНАЛИЗ КОМПОНЕНТ

Методы группировки в факторном анализе широко применялись для определения структуры изучаемого комплекса случайных величин. Как правило, эти методы, рассмотренные Харманом [35], зависят от априорных предположений, сделанных относительно изучаемых случайных величин. Преимущество этих методов заключается в том, что вычисление значений изучаемых факторов производится за один прием, в связи с чем устанавливается необходимость поворота координатной системы для достижения простейшей структуры. Это объясняется тем, что состав групповых факторов задан заранее. Все это справедливо и для множественного анализа компонент, так как число и состав различных подгрупп взаимно связанных переменных заранее определены с помощью анализа групп. После этого для каждой полученной подгруппы случайных величин вычисляются оценки параметров первой главной компоненты. Так как в данном случае главные компоненты, соответствующие различным группам, строятся по разному числу случайных величин, их дисперсии различны и необходимо провести соответствующее нормирование, чтобы дисперсии стали равными единице. Полученные после такого преобразования главные компоненты можно представить геометрически

как оси, по направлению которых расположены выявленные группы случайных величин в исходном пространстве. После этого вычисляются значения координат каждой пробы относительно нормированных главных компонент. Каждая проба относится к той группе, для которой получено наибольшее значение координат. Таким способом проводится последовательная классификация всех результатов наблюдений. При этом нет никаких ограничений для общего числа наблюдений, подвергаемых классификации.

Все изложенное выше можно обобщить следующим образом. Допустим, что в результате применения анализа групп было получено из n исходных случайных величин k наборов. Для каждой такой группы, обозначенной G_i и состоящей из m_i случайных величин, можно определить

$$T_i = \sum_{j=1}^{m_i} (x_{ji};\ j \in G_i),\ i = 1,\ 2,\ \ldots,\ k$$

как совокупность случайных величин x_j, причем каждая величина x_j имеет среднее, равное нулю, и дисперсию, равную единице. Для каждой совокупности T_i подсчитывается первая главная компонента P_i, представленная как вектор-столбец порядка $n \times 1$, определенный из ковариационной матрицы $\mathbf{C}_i$, соответствующей совокупности случайных величин T_i.

Таким образом,

$$\mathbf{C}_i \mathbf{P}_i = \lambda_i P_i,$$

где λ_i — наибольшее собственное значение матрицы $\mathbf{C}_i$. Для каждого результата наблюдения $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)$ вычисляется k-мерный вектор $\mathbf{z} = (z_1, z_2, \ldots, z_k)$,
где

$$z_i = \lambda_i^{-\frac{1}{2}} \sum x_i P_i,\ i \in G_i.$$

Данное наблюдение $\mathbf{x}$ относится к группе с номером j, для которой $z_j = \max(z_1, z_2, \ldots, z_k)$. Эта операция проводится для всех наблюдений, в результате чего все они будут отнесены к одной из k групп.

КЛАССИФИКАЦИЯ ОСАДОЧНЫХ ПОРОД БАГАМСКИХ ОСТРОВОВ

Рассмотрим реальный пример, позволяющий провести сравнение метода множественного анализа компонент с Q-методом факторного анализа. В нашем примере рассматривается классическое изучение карбонатных отложений [75, 76], задача которого заключается в том, чтобы на количественной основе разработать си-

стему классификации осадочных пород, взятых из различных фаций. В работе Парди [74] приведен полный перечень всех 216 отобранных образцов. В табл. 3 перечислены 12 составляющих, наблюдаемых в породах, которые были использованы в исследовании в качестве фациальных характеристик. Эти характеристики детально описаны Парди [75], и их можно разделить на три группы — скелетные формы, нескелетные формы и характеристики размеров частиц. Относительное значение каждой составляющей определялось путем точечных подсчетов в шлифах. Содержание тонкообломочного материала ($<1/8$ мм), выраженное в весовых процентах, в пробах определялось с помощью ситового анализа. Эти данные послужили исходной информацией, на основе которой была построена система классификации.

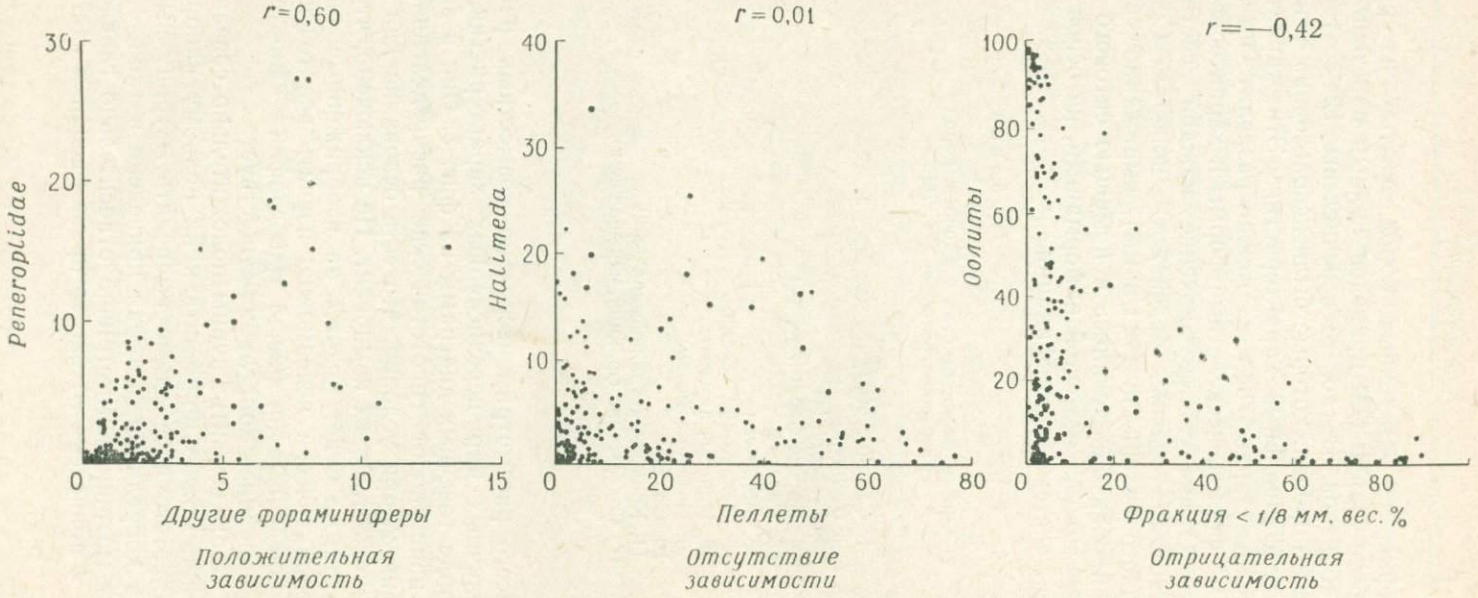
Таблица 3

Список характеристик осадочных пород Багамских островов [74, 75]

1. *Coralline algae*
2. *Halimeda*
3. *Peneroplidae*
4. Другие фораминиферы
5. Кораллы
6. Моллюски
7. Пеллетовые осадки
8. Илистые агрегаты
9. Натечные формы
10. Оолиты
11. Криптокристаллические зерна
12. Содержание фракций, меньших чем $^1/_8$ мм

Прежде всего рассмотрим несколько совместных распределений для различных пар перечисленных характеристик, которые показаны в виде точечных диаграмм на фиг. 3. Они в значительной степени напоминают гипотетические распределения, изображенные на фиг. 2, при условии, что есть основание для применения метода множественных компонент. Не исключено, что каждая точечная диаграмма будет похожа на изображенные на фиг. 2. Кроме того, практически невозможно изучить все точечные диаграммы, а приведенные на фиг. 3 три графика свидетельствуют о том, что обоснованная корреляция существует.

Для того чтобы выявить группы положительно связанных случайных величин, можно применить к их полному набору анализ сходства. Для этого необходимо вычислить матрицу выборочных коэффициентов корреляции, которая приведена в табл. 4. В связи с тем, что эта матрица симметрична относительно главной диагонали, в табл. 4 приведена только ее правая верхняя часть. В этой



Ф и г. 3. Сравнение точечных диаграмм осадочных пород Багамских островов.

Таблица 4

Корреляционная матрица характеристик осадочных пород Багамских островов

Характеристики	1	2	3	4	5	6	7	8	9	10	11	12
1	1,00	0,33	0,09	0,24	0,70	0,14	—0,19	0,03	0,09	—0,21	0,10	—0,12
2		1,00	0,14	0,30	0,37	0,42	0,01	0,22	—0,21	—0,36	—0,05	0,16
3			1,00	0,60	—0,01	0,45	0,30	0,32	—0,25	—0,40	—0,17	0,67
4				1,00	0,11	0,46	0,12	0,30	—0,12	—0,53	—0,14	0,53
5					1,00	0,29	—0,16	0,06	—0,12	—0,21	0,02	—0,14
6						1,00	0,13	0,23	—0,14	—0,55	0,01	0,45
7							1,00	0,52	—0,38	—0,37	—0,43	0,66
8								1,00	—0,32	—0,40	—0,25	0,55
9									1,00	—0,31	0,64	—0,39
10										1,00	—0,44	—0,42
11											1,00	—0,36
12												1,00

таблице нумерация строк и столбцов совпадает с номерами характеристик табл. 3.

В связи с тем что все случайные величины до начала исследования равноценны в смысле пригодности для классификации, их можно сгруппировать методом Сокала и Сниса [83], который называется методом невзвешенных парных групп. Чтобы избежать рассмотрения отрицательных значений, воспользуемся преобразованием arccos, в результате чего получим меру расстояния между векторами. Это расстояние и будет служить критерием классификации.

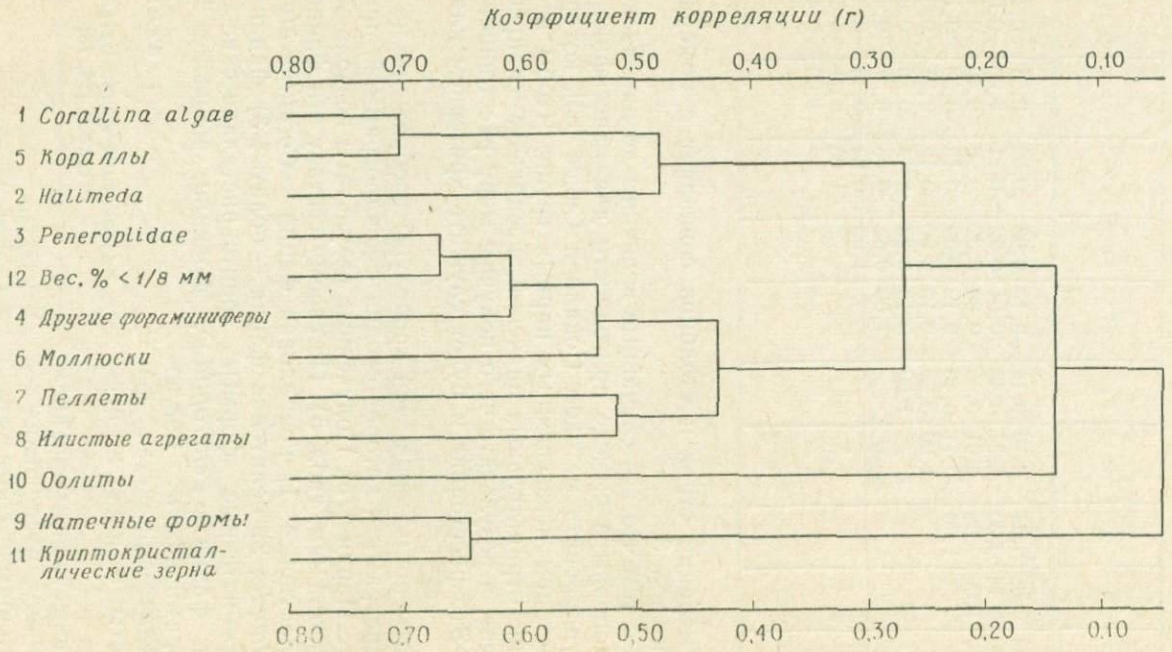
Результаты применения этой процедуры изображены на фиг. 4 в виде дендрограммы, которая сходна с дендрограммой, приведенной в работе Парди [75]. Полученные результаты позволяют выделить пять фаций (см. табл. 5). По своему составу эти фации представляют следующее: 1) коралловая фация — *Coralline algae,* кораллы, *Halimeda*; 2) оолитовая фация — оолиты; 3) фация натечных форм — натечные формы, криптокристаллические зерна; 4) илистая фация — *Peneroplidae*, фракции, меньшие $^1/_8$ мм, другие фораминиферы, моллюски; 5) пеллетово-илистая фация — пеллеты и илистые агрегаты.

Теперь перестроим корреляционную матрицу, расположив ее строки и столбцы в соответствии с дендрограммой. Такая перестроенная матрица приведена в табл. 6, в которой сохранены

Таблица 5

Карбонатные фации Багамских островов [74, 75]

1. Кораллы
2. Оолиты
3. Натечные фации
4. Илы
5. Пеллетовые илы



Фиг. 4. Иерархические группы, полученные по составляющим осадочных пород Багамских островов.

Таблица 6

Корреляционная матрица осадочных пород Багамских островов после перестановки строк и столбцов

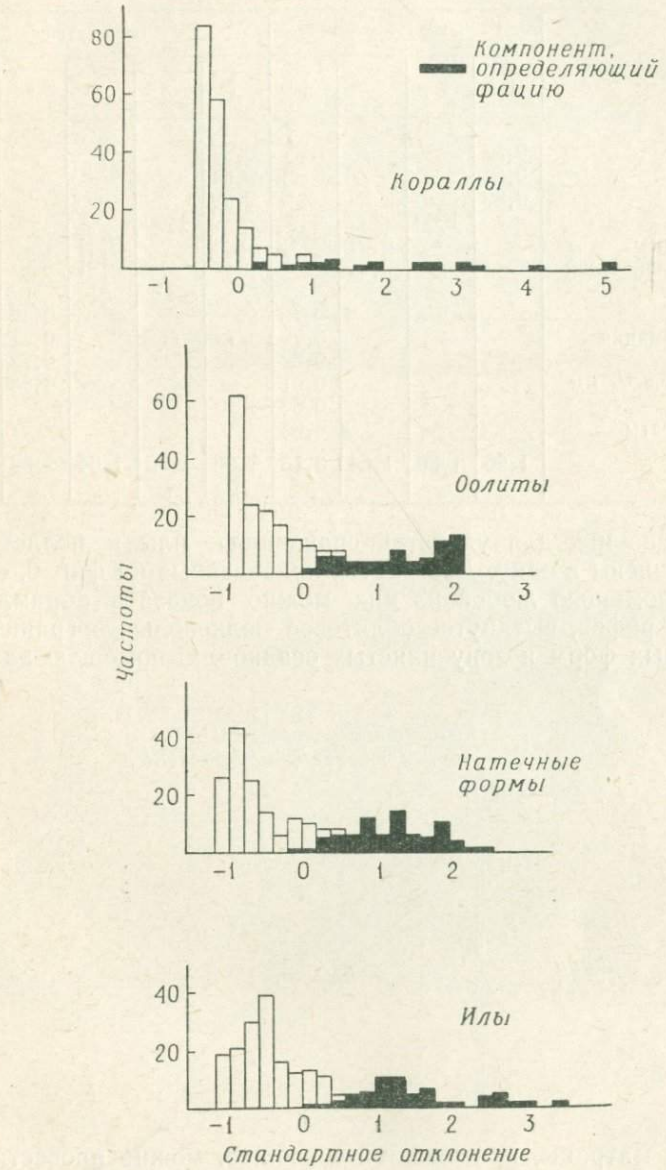
Характеристики	1	2	5	10	9	11	3	4	6	12	7	8
1	1,00	0,33	0,77	—0,21	—0,09	0,10	0,09	0,24	0,14	—0,12	—0,19	0,03
2	0,33	1,00	0,37	—0,36	—0,21	—0,05	0,14	0,30	0,42	0,16	0,01	0,22
5	0,70	0,37	1,00	—0,21	—0,12	0,02	—0,01	0,11	0,29	—0,14	—0,16	0,06
10	—0,21	—0,36	—0,21	1,00	—0,31	—0,44	—0,40	—0,53	—0,55	—0,42	—0,37	—0,40
9	—0,09	—0,21	—0,12	—0,31	1,00	0,64	—0,25	—0,12	—0,14	—0,39	—0,38	—0,32
11	0,10	0,05	0,02	—0,44	0,64	1,00	—0,17	0,14	0,01	—0,36	—0,43	—0,25
3	0,09	0,14	—0,01	—0,40	—0,25	—0,17	1,00	0,60	0,45	0,67	0,30	0,32
4	0,24	0,30	0,11	—0,53	—0,12	0,14	0,60	1,00	0,46	0,53	0,12	0,30
6	0,14	0,42	0,29	—0,55	—0,14	0,01	0,45	0,46	1,00	0,45	0,13	0,23
12	—0,12	0,16	—0,14	—0,42	—0,39	—0,36	0,67	0,53	0,45	1,00	0,66	0,55
7	—0,19	0,01	—0,16	—0,37	—0,38	—0,43	0,30	0,12	0,13	0,66	1,00	0,52
8	0,03	0,22	0,06	—0,40	—0,32	—0,25	0,32	0,30	0,23	0,55	0,52	1,00

номера строк и столбцов исходной матрицы. Группы случайных величин расположены в том же порядке, что и в табл. 5. Нетрудно видеть, что наибольшие положительные коэффициенты корреляции оказались расположенными близ главной диагонали, что отражает внутреннюю однородность выявленных групп. Кроме того, из данной матрицы видно, что структурный параметр охарактеризован как илистой, так и пеллетово-илистой фациями. На дендрограмме это не видно. В рассмотренном нами примере встречаются перекрывающиеся группы. Естественно, что такая ситуация может встретиться в любом геологическом исследовании. При этом весьма важно, что построение классификации основано на перестановке строк и столбцов ковариационной матрицы в соответствии с дендрограммой.

При анализе дендрограммы, изображенной на фиг. 4, можно сделать вывод, что в пробах представлены только четыре фации. Вполне возможно, что илистые фации неотличимы от пеллетово-илистых фаций и этот случай отражен в матрице, приведенной в табл. 6, где внутренние корреляционные связи между случайными величинами, входящими в каждую из двух групп, еще достаточно высоки. Для удобства сравнения допустим, что имеется всего четыре различных фации, после чего рассмотрим результаты классификации проб с помощью Q-метода и метода множественных компонент.

Допустив существование четырех фаций, разделим матрицу, приведенную в табл. 6, на четыре подматрицы. Для каждой подматрицы вычислим первую главную компоненту и нормируем ее, чтобы получить дисперсию, равную единице. Коэффициенты первых главных компонент, приведенных в табл. 7, выбраны так, что сумма квадратов коэффициентов каждой компоненты равна соответствующему собственному значению, приведенному в нижней части табл. 7. Для каждого из 216 результатов наблюдения было вычислено значение главных компонент по каждой фации. Результат наблюдения относился к той фации, для которой значение главной компоненты достигало максимума. Таким способом были расклассифицированы все имеющиеся наблюдения. На фиг. 5 приведены гистограммы распределения значений главных компонент для каждой из четырех фаций. Области наибольших значений, по которым производилась классификация, зачернены. В некоторых гистограммах (в зачерненной части) отмечается бимодальность. Таким образом, описанная классификационная процедура, основанная на множественных компонентах, соответствует сделанным относительно нее предположениям.

Фациальную карту, полученную на основе четырех множественных компонент, соответствующих различным фациям, можно сравнить с фациальной картой, построенной Парди [75] с помощью обычных методов. Обе эти карты приведены на фиг. 6. В связи с тем что в начальной стадии исследования было выделено пять



Фиг. 5. Гистограммы распределения компонент, определяющих фации.

Таблица 7

Множественные компоненты

Характеристики	Четыре группы				Пять групп				
	1	2	3	4	1	2	3	4	5
Coralline algae	0,87				0,87				
Halimeda	0,64				0,64				
Кораллы	0,89				0,89				
Оолиты		1,00				1,00			
Натечные формы			0,91				0,91		
Криптокристаллический гранит			0,91				0,91		
Peneroplidae				0,79				0,86	
Другая информация				0,70				0,81	
Моллюски				0,62				0,72	
Фракция мельче $^1/_8$ мм				0,91				0,83	0,88
Пеллеты				0,62					0,86
Илистые агрегаты				0,66					0,80
λ	1,96	1,00	1,64	3,16	1,96	1,00	1,64	2,59	2,16

фаций, две из них для удобства сравнения — илы и пеллеты — были объединены в одну. Две карты, приведенные на фиг. 6, очень сходны. С помощью любой из них можно показать обрамление коралловых рифов, вытянутое оолитовое мелководье, пограничную зону натечных форм и зону илистых осадков. С помощью класси-

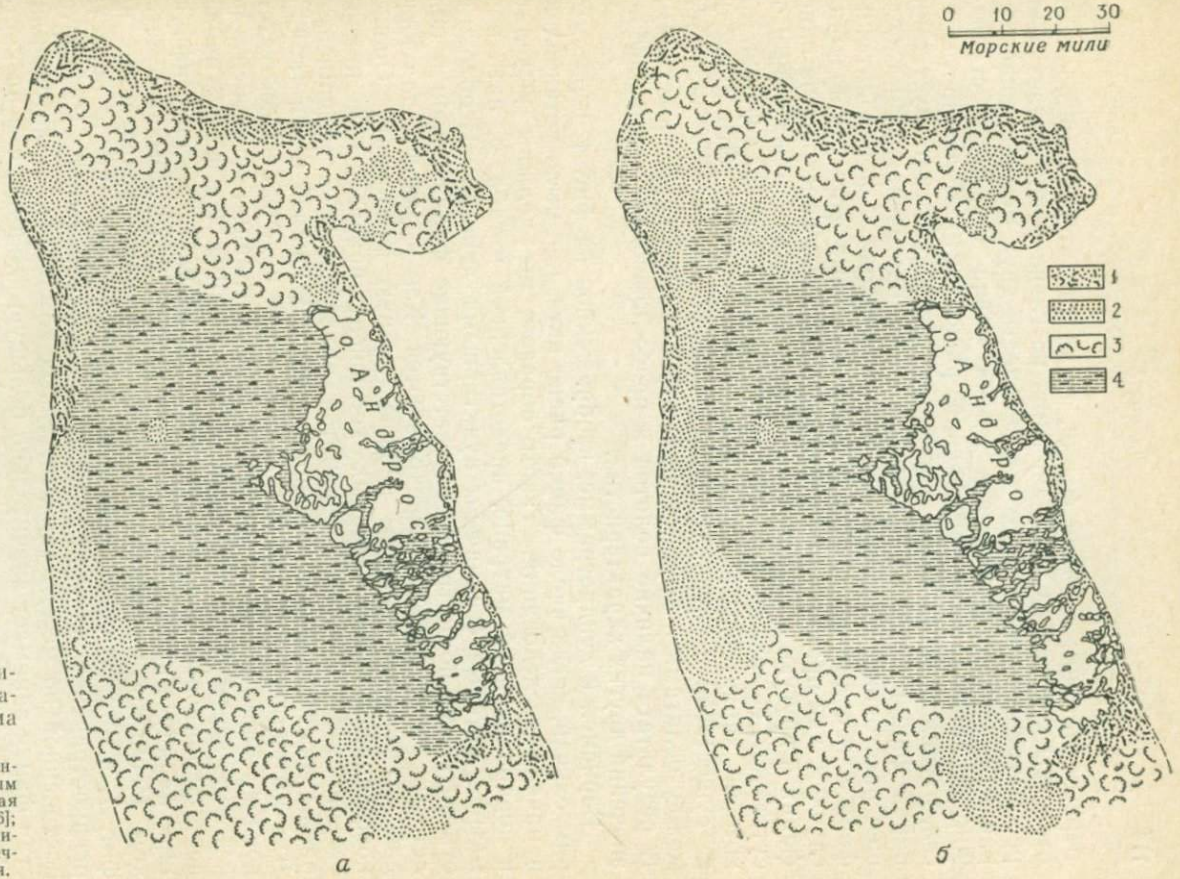
Таблица 8

Классификационная матрица. Множественные компоненты

Фации	1	2	3	4
1	20	0	7	4
2	1	64	5	1
3	0	1	63	0
4	0	2	0	48

Парди [75, 76]

фикацинной матрицы, приведенной в табл. 8, можно провести количественное сравнение этих методов. В этой матрице по главной диагонали помещены значения числа наблюдений, одинаково классифицированных обоими методами. Величины, находящиеся не на главной диагонали, показывают число наблюдений, отнесенных



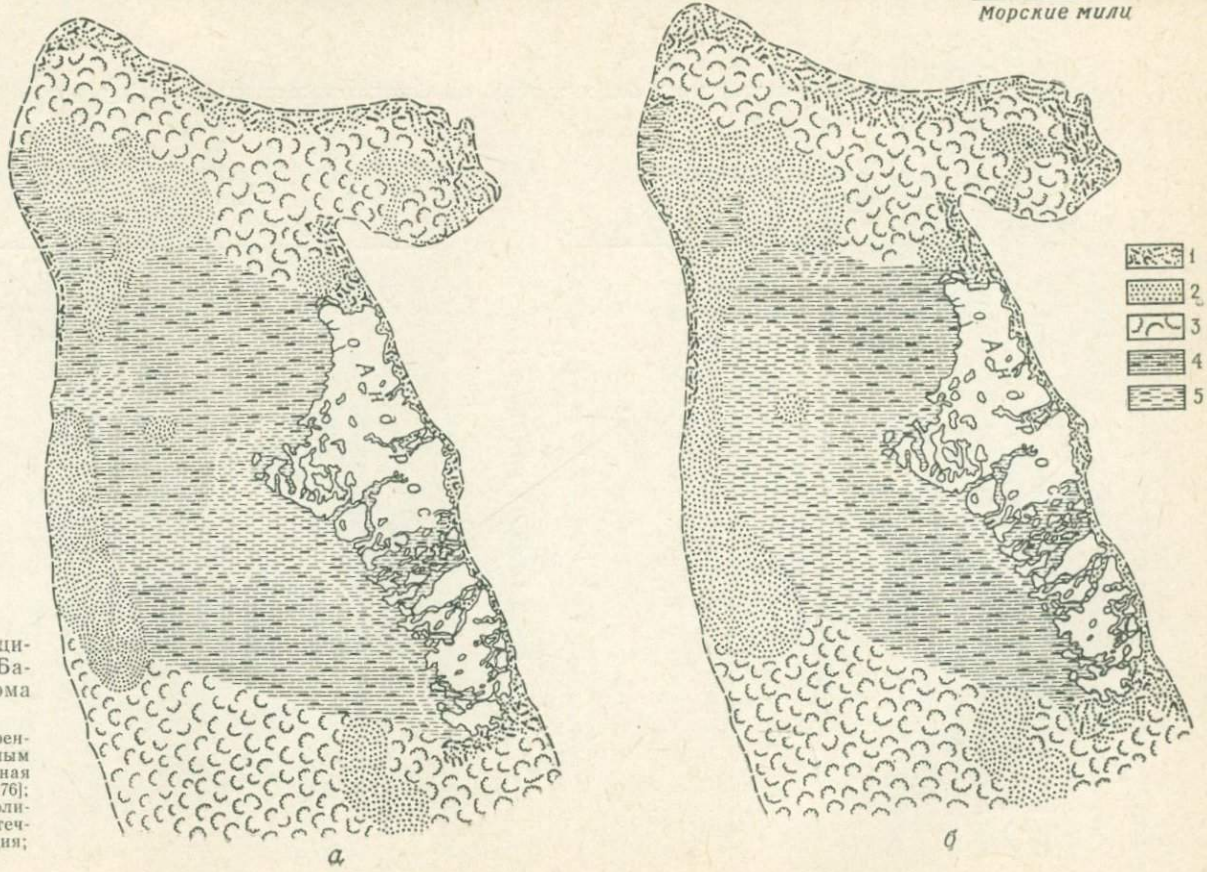
Фиг. 6. Сравнение фациальных карт Большой Багамской банки (платформа Андрос).

а — фациальная карта, построенная по четырем множественным компонентам; *б* — фациальная карта по данным Парди [75, 76]; *1* — коралловая фация; *2* — оолитовая фация; *3* — фация натечных форм; *4* — илистая фация.

первым методом к одной фации, тогда как вторым методом они отнесены к другой. Сумму значений по главной диагонали можно рассматривать как меру согласованности двух методов классификации. В нашем примере эта сумма равна 195, что составляет 90% от общего числа наблюдений, равного 216. Большинство чисел, расположенных не на главной диагонали, соответствует таким парам фаций, границы между которыми протягиваются на большие расстояния. Например, коралловая фация и фация натечных форм или оолитовая фация и фация натечных форм. В подобных случаях, как правило, отмечаются постепенные переходы из одной фации в другую, а следовательно, четкие границы между ними отсутствуют. Исключение представляет только оолитовая фация, для которой характерны высокие содержания оолитов. Это приводит к существенной отрицательной корреляционной зависимости между оолитами и другими рассматриваемыми характеристиками. В результате наличие оолитов в пробе приводит к уменьшению значений других характеристик.

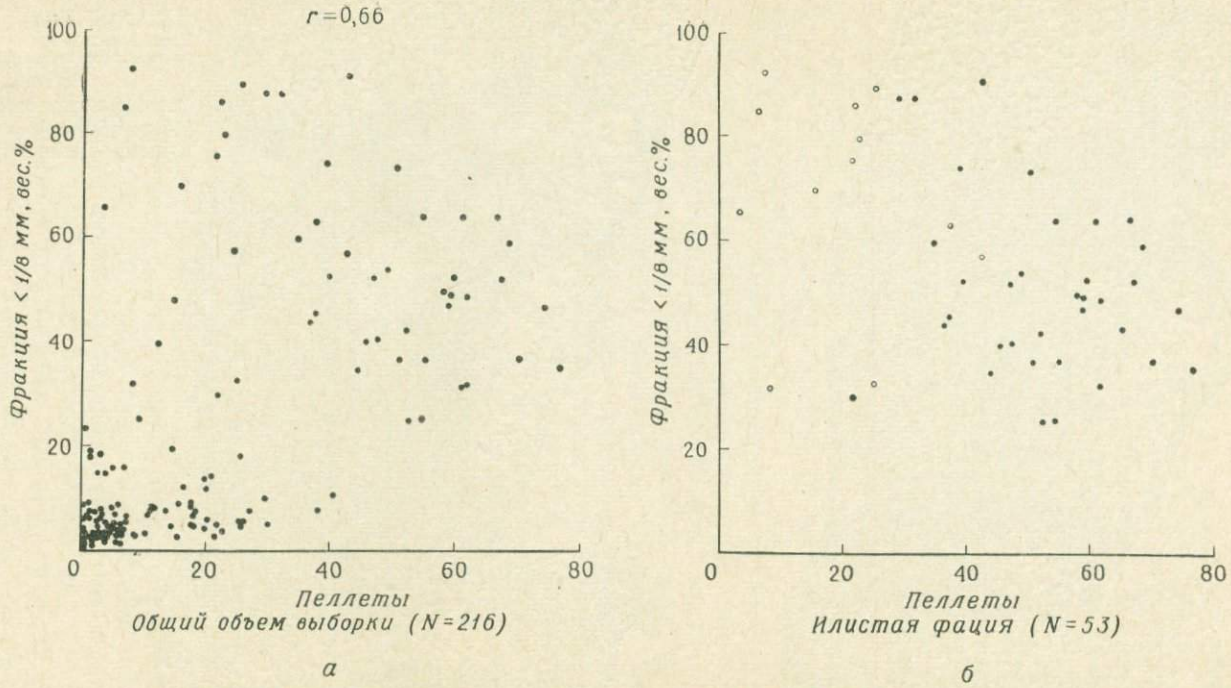
Вернемся к дендрограмме, изображенной на фиг. 4, и рассмотрим случай, когда число фаций равно пяти, как было установлено в начале исследования. Таким образом, мы должны выделить в ковариационной матрице, представленной в табл. 6, пять подматриц, которые ограничены пунктирными линиями. Необходимо отметить, что подматрицы, соответствующие фациям илов и пеллетовых илов, частично перекрываются. Как и раньше, для каждой подматрицы построим первую главную компоненту. Оценки коэффициентов этих главных компонент приведены в правом столбце табл. 7. Суммы квадратов приведенных коэффициентов равны собственным числам, приведенным в нижней части табл. 7. Для имеющихся выборочных данных были подсчитаны соответствующие нормированные значения первых главных компонент, по которым затем была проведена классификация. На фиг. 7 приведена фациальная карта, построенная по результатам классификации, наряду с картой, построенной Парди [75, 76] обычным методом. Несмотря на общее сходство двух карт, приведенных на фиг. 7, в них отмечаются и существенные различия в области распространения илистой и пеллетово-илистой фаций. На карте, построенной по методу множественных компонент, пеллетово-илистая фация занимает значительно большую площадь, чем на карте Парди. Таким образом, согласованность этих карт в значительной степени меньше, чем карт, построенных по четырем фациям.

В табл. 9 приведена классификационная матрица для пяти фаций, в которой сумма чисел, расположенных на главной диагонали, равна 179, что составляет 83% выборки объемом 216 наблюдений. Такое снижение согласованности двух карт, построенных разными методами, вызвано в основном несовпадением классификации результатов наблюдения пеллетово-илистой и илистой фаций. Объяснение этого расхождения наглядно представлено на



Фиг. 7. Сравнение фациальных карт Большой Багамской банки (платформа Андрос).

а — фациальная карта, построенная по пяти множественным компонентам; *б* — фациальная карта по данным Парди [75, 76]: *1* — коралловая фация; *2* — оолитовая фация; *3* — фация натечных форм; *4* — илистая фация; *5* — пеллетовая фация.



Фиг. 8. Сравнение точечных диаграмм осадочных пород Багамских островов.

фиг. 8. На фиг. 8, *а* приведена точечная диаграмма зависимости между содержанием тонкозернистого материала (< 1/8 мм) и пеллет, построенная по всей выборке объемом 216 наблюдений. Эти характеристики легче всего диагностируются в обеих фациях и, следовательно, являются наилучшими для установления принадлежности наблюдения к одной из них. Те результаты наблюдения, которые оказались сгруппированными около начала координат, принадлежат другим фациям. После установления точек, соответ-

Таблица 9

Классификационная матрица

		Множественные компоненты				
Фации		1	2	3	4	5
Парди [75, 76]	1	19	0	6	4	2
	2	0	63	3	0	5
	3	0	1	63	0	0
	4	0	0	0	12	15
	5	0	0	0	1	22

ствующих этим результатам наблюдения, была получена другая точечная диаграмма (фиг. 8, *б*). Из фиг. 8, *б* видно, что области расположения двумерных точек рассматриваемых двух фаций перекрываются; это позволяет объяснить причину появления сомнительных результатов при классификационном отнесении наблюдений к той или иной фации. Такой вывод останется правомерным до тех пор, пока исследователь не будет располагать большим числом данных, или до тех пор, пока не будут установлены новые характеристики, более пригодные для диагностики.

СРАВНЕНИЕ КЛАССИФИКАЦИЙ

В наших исследованиях может возникнуть вопрос: какой из методов классификации предпочтительнее? Прежде всего необходимо подчеркнуть, что во всех ситуациях мы заинтересованы в выборе лучшего метода. Из этого следует, что нам необходимо ограничиться одним типом задач классификации.

Если допустить, что установлены три типа связи, представленные на фиг. 2, то приемлемой классификацией будет та, при которой результаты наблюдений группируются более плотно в группах при заданном числе групп. Мерой, характеризующей это свойство, является дисперсия расчленения, или общая сумма квадратов отклонений от средних внутри групп. Такая мера уже использовалась [21] в качестве критерия для группировки. Однако нам

не нужно проверять с помощью этого критерия существование групп, так как у нас уже имеются результаты группировки, представленные в виде фациальной карты.

Для n результатов наблюдений, каждый из которых представляет собой набор значений ρ случайных величин, причем n наблюдений разделено с помощью некоторой процедуры классификации на r групп, общая сумма квадратов отклонений внутри групп (TSS) будет равна

$$TSS=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{p}(x_{ijk}-x_{i\cdot k})^2,$$

где $x_{i\cdot k}$ — выборочное среднее случайной величины с номером k для группы с номером i, вычисленное по n_i результатам наблюдения.

По имеющимся в нашем распоряжении данным были подсчитаны дисперсии расчленения как для четырех, так и для пяти групп. Заметим, что, помимо описанных здесь двух методов классификации, можно назвать еще два, которые использовались для сравнения и описаны в работе Мак-Кеммона [62].

В табл. 10 приведены соответствующие значения дисперсий расчленения. Как в случае четырех, так и пяти осадочных фаций наименьшая дисперсия расчленения соответствует методу множественных компонент. Этот факт представляет интерес в связи с тем, что дисперсия расчленения является функцией, минимизируемой на каждом шаге иерархического метода группировки. Однако, несмотря на то что эта функция минимизируется на каждом шаге процесса группировки, это не означает, что она обеспечивает общий минимум. Так, для случая четырех групп она привела к максимальной дисперсии для метода Уорда [91], а для случая пяти групп — к максимуму для метода Парди [74, 75]. Из табл. 10 следует, что из всех четырех рассмотренных методов метод множественных компонент можно считать лучшим.

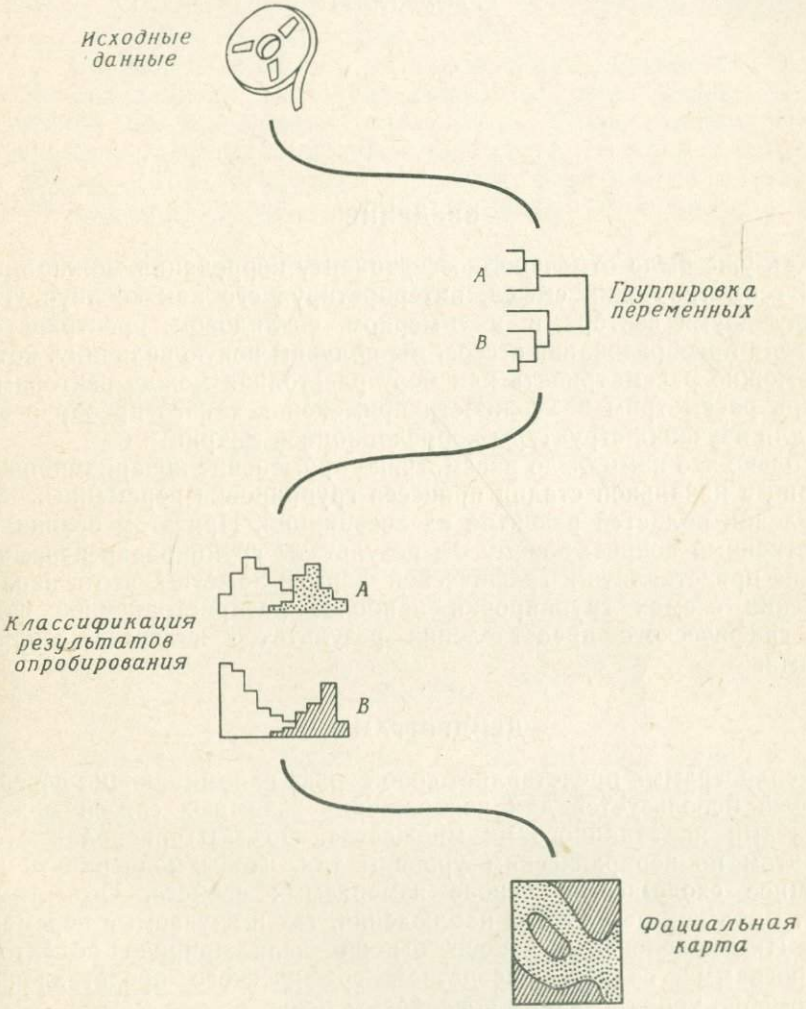
Таблица 10

Дисперсия расчленения
(общая сумма квадратов отношений от средних внутренних групп)

	Число групп	
	4	5
Множественные компоненты	1245,4	1085,6
Парди [74, 75]	1305,2	1206,3
Главные компоненты	1264,4	1168,6
Уорд [91]	1356,3	1136,1

Мы рассмотрели ряд примеров фациального исследования, когда можно использовать метод множественных компонент. Было показано, что в результате такой классификации получаются однородные подгруппы общей выборки. На фиг. 9 приведена схема последовательных стадий обработки данных при классификации

с помощью этого метода. После того как собраны исходные данные, определяются подмножества взаимно связанных случайных величин. Для каждого такого подмножества строится первая глав-



Фиг. 9. Схема процесса получения фациальной карты с помощью метода множественных компонент.

ная компонента. Каждой первой главной компоненте ставится в соответствие определенная фация. Для выбора пробы вычисляются значения всех первых главных компонент, и она относится к той фации, для которой достигнуто наибольшее значение. На основе этой информации строится фациальная карта, которая используется для интерпретации процесса осадкообразования.

ЛЕКЦИЯ 9

Графическое представление корреляции

Р. Мак-Кеммон

ВВЕДЕНИЕ

Как уже было отмечено, коэффициенту корреляции можно придать геометрический смысл, интерпретируя его как косинус угла между двумя векторами в n-мерном евклидовом пространстве. Проведя преобразование arccos, мы получим новую величину, которую можно рассматривать как меру расстояния между векторами. Теперь рассмотрим возможности применения такого преобразования для изучения структуры корреляционной матрицы.

Только что нами было рассмотрено применение ковариационных матриц в начальной стадии процесса группировки переменных при выявлении областей развития их ассоциаций. При этом возникает естественный вопрос: почему бы результаты группировки переменных не представлять в графической форме? Заметим, что в иерархических схемах группировки наиболее распространенной формой графического представления результатов является дендрограмма.

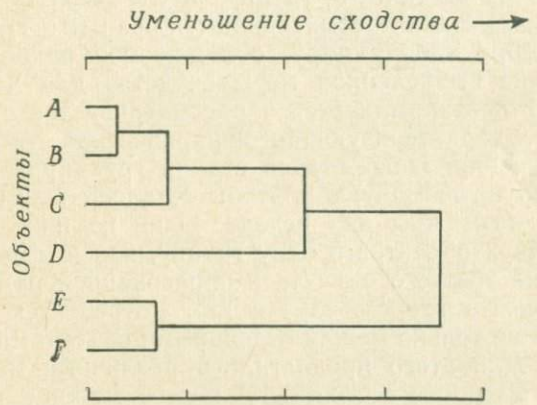
ДЕНДРОГРАММЫ

Дендрограмма представляет собой граф, напоминающий дерево, который используется для изображения взаимных связей между объектами из заданного их множества. Объекты располагаются при этом по иерархическим уровням так, чтобы подчеркнуть их взаимное сходство на основе измеряемых свойств. Объектами могут быть как результаты наблюдения, так и изучаемые переменные. Иными словами, какова бы ни была природа объектов, дендрограмма является средством графического представления многомерно упорядоченной последовательности.

В общем виде дендрограмма представлена на фиг. 1, где шесть объектов, обозначенных буквами от A до F, образуют гнездовую иерархическую структуру. Объекты A, B и C рассматриваются как сходные, образующие одну группу (аналогично объекты E и F), в то время как объект D отличается от всех групп объектов, которые выражены в структуре на более низком иерархическом уровне. Таким образом, сходные объекты образуют группы на более высоких иерархических уровнях, чем объекты менее сходные, которые

образуют группы на более низких иерархических уровнях. Объединение в группы объектов имеет смысл только в условиях высокой степени сходства. Компактная группировка свидетельствует о силе взаимных связей между объектами, тогда как некомпактное группирование указывает на слабую зависимость.

Группировка начинается с определения для каждой пары объектов набора величин, характеризующих сходство, которые можно вычислить по результатам наблюдений. Такие характеристики показывают, насколько каждый объект сходен со всеми остальными объектами. Если объектами являются непрерывные случайные величины, то в качестве меры близости между ними обычно исполь-



Фиг. 1.

зуется коэффициент корреляции. В тех случаях, когда объекты представляют собой результаты наблюдения, охарактеризованные множеством измеренных свойств, в качестве меры сходства пар используются некоторые специальные характеристики. По вопросу выбора характеристик сходства как для качественных, так и количественных исходных данных существует обширная литература [83, 2].

Для нас существенно, что для любых двух объектов, которые мы обозначим x_i и x_j, имеется характеристика сходства (s_{ij}), отражающая степень близости между x_i и x_j на основе исходных выборочных данных. Если число объектов n, то число их возможных пар составит $\binom{n}{2}$, или $\frac{n(n-1)}{2}$. Соответствующие этим парам характеристики сходства можно расположить в виде матрицы S, порядка $n \times n$, для элементов которой (s_{ij}) будет выполнено равенство $s_{ij}=s_{ji}$, из чего следует, что матрица S симметрична. Эта матрица образует основу для последующего применения любого метода группировки объектов во взаимно связанные группы.

Группировка объектов — результат анализа групп (кластерный анализ). Несмотря на то что метод группировки может изменяться, он должен соответствовать определенному типу классификации [55], а именно иерархическому агломеративному. (Подробно с различными методами анализа групп можно познакомиться во многих работах [57, 55, 41, 80]).

Исходное предположение иерархического агломеративного метода заключается в том, что каждый объект в начале процедуры представляет собой самостоятельную группу. Так, в случае n объектов будет n групп. Первая стадия группировки заключается в отыскании двух наиболее сходных объектов с последующим объединением их в единую группу, после чего останется $n-1$ групп. В первой стадии рассматриваются все $n(n-1)/2$ пар объектов, чтобы выбрать одну пару, обладающую наибольшим сходством. Значение вычисленной меры сходства для двух наиболее сходных объектов запоминается и рассматривается в дальнейшем как величина, характеризующая иерархический уровень первой стадии группировки. Цель второй стадии группирования — сокращение числа групп на одну, так, чтобы осталось $n-2$ групп. В данной ситуации возможны два исхода. Один из них — это простое объединение двух объектов в одну группу, как и в первой стадии, или дополнение третьего объекта к образованной ранее группе из двух объектов. Во втором случае нам потребуется определение меры сходства не только между отдельными объектами, но и между их группами. Для этого предлагались различные меры, которые были сведены в единую таблицу [55]. В этой таблице представлены метод ближайшего соседа, центрирование, нахождение медианы, среднего значения в группе и др. Каждая из этих характеристик отчетливо определяет стратегию процесса группировки, которая приводит к различным результатам в иерархическом расположении групп. Для успешного проведения процедуры иерархического объединения необходимо выполнить единственное требование, заключающееся в том, чтобы функция, характеризующая иерархический уровень объединения, изменялась монотонно.

В качестве мер группового сходства удобно рассматривать только такие характеристики, которые можно выразить в метрической форме. Одна из таких мер — arccos коэффициента корреляции. Для ряда других мер сходства также можно найти соответствующие преобразования, приводящие к новым мерам — метрическим или по крайней мере почти метрическим. Определенный таким способом набор элементов матрицы **S** будет обладать следующими свойствами: $s_{ii}=0$ и $s_{ij}>0$ для всех i и j. Таким образом, матрица **S** — не отрицательная. Если элементы матрицы **S** имеют свойства метрики, то для любых i, j и k будет справедливо неравенство $s_{ij}<s_{ik}+s_{kj}$.

Таким образом, на второй стадии группировки по значению меры расстояния между группами находятся две наиболее близкие

группы, которые объединяются в одну. В результате останется $n-2$ групп. Заметим, что любой из двух, отмеченных выше исходов, может привести к подобному объединению. Чтобы определить, для какой пары групп необходимо провести объединение, требуется рассмотреть $\binom{n-1}{2}$, или $\frac{(n-1)(n-2)}{2}$ пар. Значение меры близости между объединяемыми группами на второй стадии является характеристикой второго иерархического уровня группировки.

Теперь процедура группировки четко определена. На каждой последующей стадии группировки число групп сокращается на одну за счет объединения двух «наиболее близких» групп. На всех стадиях значение меры близости, соответствующее двум объединяемым группам, характеризует иерархический уровень группировки. Этот процесс продолжается до тех пор, пока все объекты будут объединены в одну группу. Для n объектов эта процедура потребует $n-1$ стадий объединения. Все вычисления легко провести на электронной вычислительной машине, так что даже для весьма больших значений n потребуется умеренное количество машинного времени. По результатам вычислений можно построить дендрограмму.

Теперь полезно рассмотреть числовой пример. При этом необходимо определить меру близости групп, которая будет использована при последовательном анализе дендрограммы. Мы воспользуемся теми данными, по которым построена дендрограмма, приведенная на фиг. 1. Будем считать шесть объектов, обозначенных буквами от A до F, непрерывными случайными величинами, результаты измерения которых образуют выборку. По данным этой выборки были вычислены оценки коэффициентов корреляции, матрица которых приведена в табл. 1. В связи с тем что корреляционная матрица симметрична, оценки коэффициентов корреляции приведены только в правой верхней части таблицы. В ее левой нижней части приведены соответствующие значения arccos коэффициентов корреляции, рассматриваемые как мера близости между изучаемыми величинами. Значения arccos являются исходными для последующего проведения анализа групп.

Таблица 1

Матрица коэффициентов корреляции и мер близости для шести переменных

Переменные	*A*	*B*	*C*	*D*	*E*	*F*
A		0,990	0,854	—0,338	—0,934	—0,933
B	0,142		0,882	—0,213	—0,943	—0,939
C	0,547	0,491		—0,095	—0,819	—0,884
D	1,916	1,785	1,666		—0,100	—0,201
E	2,776	2,802	2,530	1,671		0,946
F	2,773	2,791	2,655	1,773	0,330	

Сначала мы предполагаем, что у нас шесть групп. На первой стадии группировки в матрице значений arccos отыскивается наименьшее как соответствующее максимальному расстоянию между объектами. Нетрудно видеть, что из всех $\binom{6}{2}=15$ пар наименьшее значение, равное 0,142, соответствует паре AB. Эти случайные величины мы объединяем в одну группу, иерархический уровень которой охарактеризован числом 0,142. В результате у нас осталось пять групп, и теперь необходимо определить меру сходства между группами. Представляется естественным в качестве такой меры принять усредненное внутригрупповое расстояние между парами, что соответствует групповой усредненной мере, описанной Лэнсом и Вильямсом [55], и отражает однородность внутри группы. Эта мера имеет существенное преимущество, заключающееся в том, что на нее не очень сильно влияют объекты, введенные в группу на поздних стадиях группирования. Следовательно, вопрос о наличии смещения оценки за счет недоучета весовых коэффициентов снимается. Более того, если от этой усредненной меры мы возьмем косинус, то получим среднее значение коэффициента корреляции в группе. На каждой последующей стадии группирования объединяются те две группы, которые в результате объединения дают наиболее однородную группу, или, что равноценно, наименьшее среднее внутригрупповое расстояние между парами. Эта мера монотонно возрастает при уменьшении числа групп. Таким образом, основное требование монотонности меры выполнено. Теперь вернемся к нашему примеру.

На второй стадии группирования мы располагаем $\binom{5}{2}=10$ вариантами попарного сравнения пяти имеющихся групп. Для каждой пары подсчитывается среднее внутригрупповое расстояние. Например, если случайная величина C объединяется с группой, образованной величинами A и B, то среднее расстояние между парами в группе ABC можно подсчитать следующим образом: $(0{,}142+0{,}547+0{,}491):3=0{,}393$. Рассмотрев остальные девять возможных комбинаций, мы определим, что минимальное значение среднего расстояния между парами, равное 0,330, соответствует объединению переменных E и F. После объединения E и F у нас останется четыре группы, для которых повторяется аналогичная процедура. Результаты всего процесса группирования приведены в табл. 2, по которой нетрудно проследить процедуру построения дендрограммы, приведенной на фиг. 1.

ИЕРАРХИЧЕСКИЕ СТРУКТУРЫ ДЕНДРОГРАММ

Основной недостаток дендрограммы заключается в неоднозначности ее древовидной структуры. Рассмотрим еще один пример. На последней стадии группирования, показанной в нижней части табл. 2, все шесть переменных расположены в таком порядке, ко-

торый получился в результате построения дендрограммы. Однако данный вариант расположения переменных не единственный. Нетрудно показать, что при иерархическом группировании n объектов существует 2^{n-1} эквивалентных дендрограмм. В нашем случае, когда рассматриваются шесть переменных, можно построить 32 дендрограммы, которые будут эквивалентны в смысле иерархической структуры.

Таблица 2

Анализ сходства по шести переменным

Шаг	Группа	Переменная	Среднее расстояние, $\bar{z}$	$\cos \bar{z}$
1			0,142	0,990
	1	*A, B*		
	2	*C*		
	3	*D*		
	4	*E*		
	5	*F*		
2			0,330	0,946
	1	*A, B*		
	2	*C*		
	3	*D*		
	4	*E, F*		
3			0,393	0,924
	1	*A, B, C*		
	2	*D*		
	3	*E, F*		
4			1,091	0,462
	1	*A, B, C, D*		
	2	*E, F*		
5			1,776	—0,204
	1	*A, B, C, D, E, F*		

Ясно, что такое обилие эквивалентных дендрограмм нежелательно. Заметим, что дендрограммы чаще всего используются для отыскания осмысленных групп переменных среди их комплекса. В связи с тем что это обычно делается с помощью визуального изучения переменных, такая неопределенность вполне приемлема.

Таким образом, нам нужен не только алгоритм для группирования объектов, но и правило для расположения их в таком порядке, чтобы полученная затем дендрограмма обеспечивала возможность наглядной интерпретации. Это можно сделать, приняв простое правило. Сначала выясним, какова цель интерпретации дендрограммы визуальными методами. Ясно, что основные свойства дендрограммы связаны с ее пирамидальной структурой. Большинство возможных стратегий построения групп связано с добавлением

одного объекта к уже образованным группам до тех пор, пока группы достигнут критического размера и объединятся на более высоком уровне. Это и объясняет подход к задаче упорядочивания объектов. В начале процедуры мы устанавливаем некоторое исходное расположение группируемых объектов. Предлагаемое нами правило упорядочивания объектов достаточно обще, но не полностью независимо от их исходного расположения. Следовательно, если существует какой-либо исходный порядок предпочтения, то его целесообразно проверить на практике. В большинстве случаев он будет оказывать только весьма слабое влияние на окончательное расположение объектов.

Допустим, что мы хотим провести группировку n объектов с помощью иерархического агломеративного метода. Обозначим порядковый номер каждого объекта и отдельно определим группу, к которой он принадлежит. Таким образом, i-й объект x_i принадлежит группе с номером i, которую мы обозначим G_i и определим как упорядоченный набор объектов. В этом примере G_i содержит только один объект. Обозначим через $R(G_i)$ порядковый номер G_i, а через $N(G_i)$ число объектов, содержащихся в G_i. Тогда для исходного состояния можно записать

$$R(G_i)=i \qquad \text{для всех } i=1, 2, \ldots, n,$$
$$N(G_i)=1 \qquad \text{для всех } i=1, 2, \ldots, n.$$

Теперь нам потребуется еще одно дополнительное определение, которое будет использоваться в процессе группирования. Обозначим через G_i^* группу, соответствующую G_i, но имеющую обратное расположение объектов. Для исходного состояния

$$G_i^*=G_i.$$

Для объединения двух групп G_i и G_i запишем

$$G_\alpha=G_i \bigcup G_j, \quad \alpha=\min(i, j),$$

где G_α представляет собой новую группу, образованную упорядоченными объектами из G_j, расположенными справа от упорядоченных объектов G_i. Ясно, что

$$G_\beta=G_j \bigcup G_i, \quad \beta=\min(i, j),$$

не равна G_α, хотя $R(G_\alpha)=R(G_\beta)$ и $N(G_\alpha)=N(G_\beta)=N(G_i)+N(G_j)$. Теперь мы можем сформулировать правило упорядочивания объектов в процессе группировки при условии, что полученная в результате дендрограмма будет пирамидальной.

Таким образом, в первой стадии группировки необходимо определить, какие из двух объектов являются наиболее сходными, т. е. какие из двух групп нужно объединить в одну. Пусть каждая из двух групп представляет отдельный объект. Обозначим эти группы

G_U и G_V. Объединяя их, получим $G_i = G_U \supset G_V$, где $U < V$, и множество

$$R(G_{j+1}) = j \text{ для } j = V,\ V+1,\ \ldots,\ n-1,$$

которое представлено $n-1$ группами.

Для k-ой стадии группирования мы примем следующее правило объединения двух наиболее сходных групп G_U и G_V. Сначала определим для группы, образованной G_U и G_V, следующие равенства:

$$G_i = \begin{cases} G_U \bigcup G_V^*, & N(G_U) > N(G_V), \\ G_U \bigcup G_V^*, & N(G_U) = N(G_V),\ U < V,\ i = \min(U, V), \\ G_V \bigcup G_U^*, & N(G_U) < N(G_V). \end{cases}$$

Множество

$$R(G_{j+1}) = j \qquad \text{для } j = V,\ V+1,\ \ldots,\ n-k$$

будет представлено $n-1$ группами. Другими словами, это правило говорит, что при объединении двух групп группа, насчитывающая меньшее число объектов, располагается справа от группы с большим числом объектов, а элементы первой группы располагаются в обратном порядке. Если две группы содержат одинаковое число объектов, то слева располагается группа, в которой меньший порядок предпочтения. В данном случае исходный порядок предпочтения будет влиять на окончательный результат расположения объектов, а во всех остальных ситуациях это влияние отсутствует.

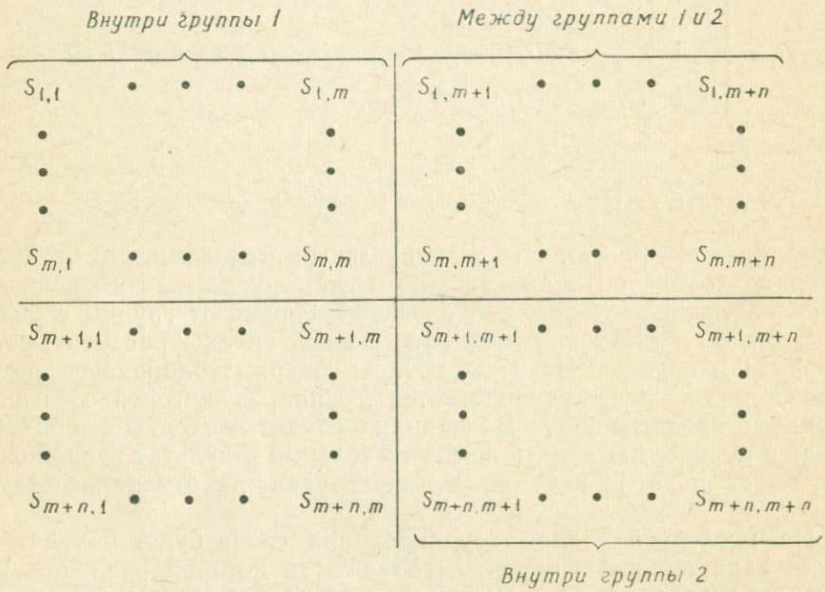
На последней стадии группирования слева будет находиться такая единственная группа G_1, объекты в которой расположены так, что дают дендрограмму пирамидальной формы. Эта процедура обеспечивает выбор из всех возможных вариантов расположения объектов наилучший для графической интерпретации.

ДЕНДРОГРАФЫ

Рассмотрим дендрограмму как одномерный граф, имеющий единственную ось, по которой откладываются значения характеристик иерархических уровней. Другая ось, по которой откладываются упорядоченные объекты, не имеет существенного значения. При этом объекты располагаются на равном расстоянии друг от друга, выбранном произвольно. Замысел дендрограммы заключается в том, что ветви дерева характеризуют иерархический порядок объектов. Однако не следует думать, что при этом отражаются и иерархические зависимости между объектами. Их можно учесть, если на второй оси расстояние между объектами сделать неравнозначным. Таким образом, мы введем новую координатную ось

и полученную двумерную дендрограмму будем называть дендрографом. Дендрограф описывает зависимости как внутри групп объектов, так и между группами. В результате мы получаем более наглядное графическое представление связей между объектами. Таким образом, дендрограф обеспечивает меньшие искажения, чем дендрограмма.

Для того чтобы построить дендрограф, необходимо определить неотрицательную меру внутригруппового сходства. Дело в том, что расстояние между объектами не может быть отрицательным. Такой

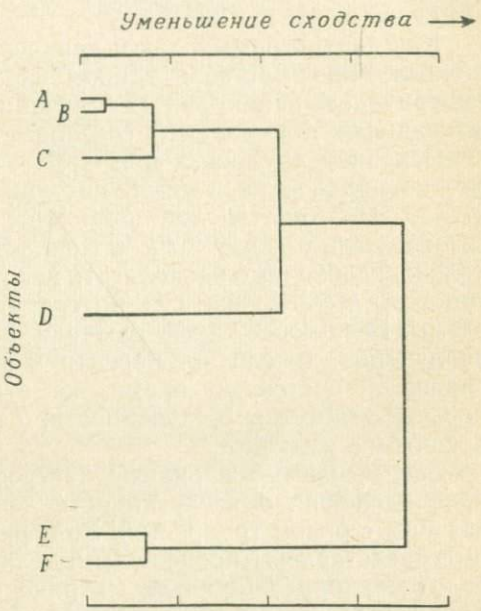


Фиг. 2.

мерой, согласующейся с определенным ранее средним внутригрупповым расстоянием, является среднее расстояние между группами. Последнее удовлетворяет основному требованию (неотрицательности) и характеризует степень сходства между двумя заданными группами. Если к этой мере применить преобразование cos, то получим среднюю характеристику корреляционной зависимости между группами, которая является аналогом средней характеристики корреляции внутри групп. Таким образом, дендрограф описывает как внутригрупповую, так и межгрупповую зависимости.

Вычислительная процедура среднего значения расстояния между группами показана на фиг. 2, где приведена часть матрицы коэффициентов, характеризующих попарное сходство между объектами. Эта часть матрицы соответствует двум группам объектов,

которые могут быть объединены в результате применения описанного метода. Приведенные в ней коэффициенты определены как мера расстояния. Среднее расстояние между m объектами первой группы и n объектами второй группы представляет собой среднее значение набора коэффициентов, образующего правую верхнюю часть матрицы. Таким образом, усредняется $m \times n$ значений расстояния. Иерархический уровень первой группы определяется как среднее расстояние по m объектам, характеристики которых расположены в верхней левой части матрицы. Для второй группы он определяется аналогично по характеристикам правой нижней подматрицы. Среднее межгрупповое расстояние не зависит от средних расстояний внутри групп. Иерархический уровень такого объединения будет охарактеризован средним внутригрупповым расстоянием, вычисленным по $\binom{m+n}{2}$, или $(m+n)\ (m+n-1)/2$, коэффициентам, образующим справа вверху треугольную подматрицу.



Фиг. 3.

Дендрограф отличается от дендрограммы только тем, что при его построении учитываются характеристики расстояния между группами, к которым принадлежат объекты. Процесс же группирования, а также правило упорядочения объектов остаются теми же, что и при построении дендрограммы. Однако в дендрографе среднее расстояние между группами используется как мера разделения двух объединенных групп. Кроме того, полученное последовательное расположение объектов сохраняется и в пространстве.

Вернемся к числовому примеру, приведенному в табл. 2. В правом крайнем столбце этой таблицы для каждой стадии группировки приведены значения среднего расстояния между группами. В данном примере оно монотонно возрастает с уменьшением числа групп, но такое возрастание не обязательно. Дендрограф, построенный по данным табл. 2, приведен на фиг. 3. Он соответствует дендрограмме, изображенной на фиг. 1, и отличается от нее только тем, что расстояние между объектами стало переменной величиной,

а не константой. Необходимо отметить, что отрезки, соответствующие расстояниям внутри групп, перекрываются, тогда как для отрезков, характеризующих расстояния между группами, перекрытия отсутствуют. Для того чтобы избежать чрезмерного растягивания дендрографа, следует выбрать соответствующий коэффициент пропорциональности между масштабом межгруппового и внутригруппового расстояний. В данном примере этот коэффициент равен 0,5.

СЛУЧАЙНЫЕ СХЕМЫ

Как дендрограмму, так и дендрограф строят с целью выявления смысла зависимостей в данном наборе объектов. Что бы ни рассматривалось как объекты — случайные величины или пробы, — они объединяются во взаимно связанные группы.

Основная трудность при этом заключается в том, что невозможно заранее распознать ситуацию, когда схема связей отсутствует. Поэтому вполне возможно, что при условии отсутствия связей между объектами возникает случайная схема. Следовательно, необходимо располагать некоторым критерием для проверки гипотезы о случайности до интерпретации полученной графической схемы. Если после такой проверки окажется, что полученная схема неслучайна, то следует перейти к решению задачи выбора числа групп. В настоящее время не существует простых критериев проверки гипотезы о случайности, и их разработку следует считать задачей на будущее.

Все мыслимые варианты построения критериев для проверки предположения о неслучайности дендрограммы или дендрографа связаны с рядом трудностей. Во-первых, многие геологические данные представлены процентными величинами, что в результате приводит к корреляционной матрице закрытой числовой системы. Трудности, связанные с применением методов корреляционного анализа в условиях этих систем, описаны в работе Чейза и Краскла [14]. Во-вторых, если группируемые объекты представляют собой пробы, мера сходства для различных пар может вычисляться по разному количеству измеряемых свойств. В результате число степеней свободы при проверке гипотезы о случайности становится неопределенным. При проверке предполагается, что большая часть групп не являются случайными группировками, а выбраны целенаправленно, по максимуму внутригрупповой однородности. Задачи построения методов проверки гипотезы о случайности группировок в дендрограммах и дендрографах требуют особого внимания.

Тем не менее интересно проанализировать процесс возникновения случайных схем в форме дендрографа. С этой целью рассматривались 20 независимых случайных величин, из которых каждая равномерно распределена в интервале (0,1). Над этими случайными величинами были проведены наблюдения, которые

составили выборки объемом 100, 200, 500 и 1000. Для каждого варианта выборочного объема подсчитывалась корреляционная матрица, элементы которой подвергались преобразованию arccos, что давало в результате меру расстояния. Для группирования 20 случайных величин использовался иерархический агломеративный метод, результаты применения которого для всех четырех вариантов объема выборки приведены на фиг. 4 в форме дендрографов. Для удобства интерпретации внутригрупповая мера расстояния была преобразована в корреляционную характеристику, так что различные иерархические уровни охарактеризованы значениями коэффициента корреляции.

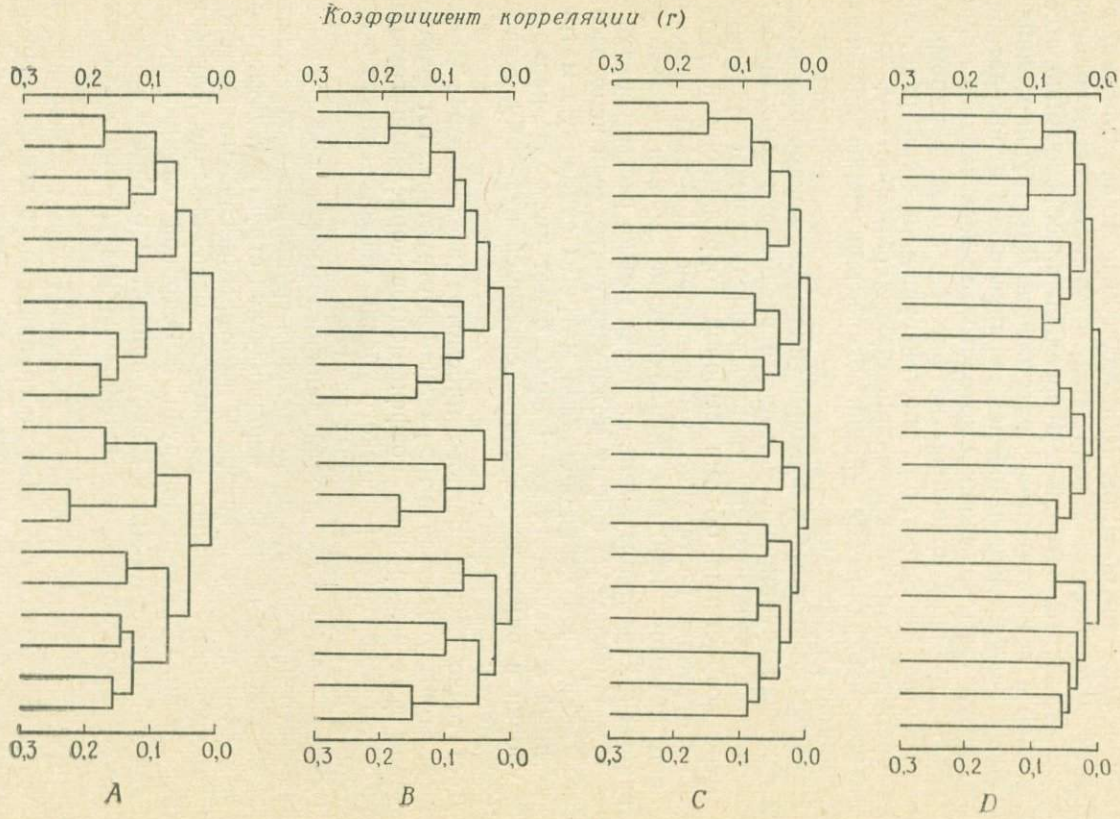
Наиболее отчетливой характеристикой всех дендрографов является исходная группировка, которая в рассматриваемых примерах соответствует очень низкому уровню корреляционной зависимости. По мере возрастания объема выборок корреляционный уровень исходной группы постепенно уменьшается. Действительно, при выборке объемом 1000 все иерархические уровни практически одинаковы и характеризующие их числа можно рассматривать как константу. Предельную форму можно получить, если использовать вместо выборочной корреляционной матрицы единичную матрицу. Полученный в результате дендрограф будет иметь все иерархические уровни в одной точке, соответствующей нулевому значению коэффициента корреляции.

Достаточно весьма беглого просмотра всех четырех дендрографов, чтобы убедиться в их различной структуре. Тот факт, что значения всех выборочных коэффициентов корреляции группируются около одного иерархического уровня, свидетельствует о независимости изучаемых случайных величин.

Главное достоинство любой дендрограммы и любого дендрографа — это наглядность представления группировки объектов, которая выражена в гнездовой иерархической структуре. Даже в случайных схемах, если пренебречь некоторыми связями (на фиг. 4 они пропущены), можно наблюдать расположение случайных величин в виде групп и подгрупп. В подобной ситуации внешний вид дендрографа обманчив и может привести к ошибкам в выводах, поскольку какой-либо смысл в наблюдаемых группировках отсутствует.

Отсутствие соответствующих количественных критериев для проверки гипотезы о случайности группировок в дендрографе приводит к тому, что в виде качественной характеристики выборки при данном объеме используются корреляционные уровни исходных групп, а также пространственное расположение объектов. Правомерность выделения в дендрографе данного числа групп можно оценить при помощи сопоставления длины промежутков между иерархическими уровнями и между группами.

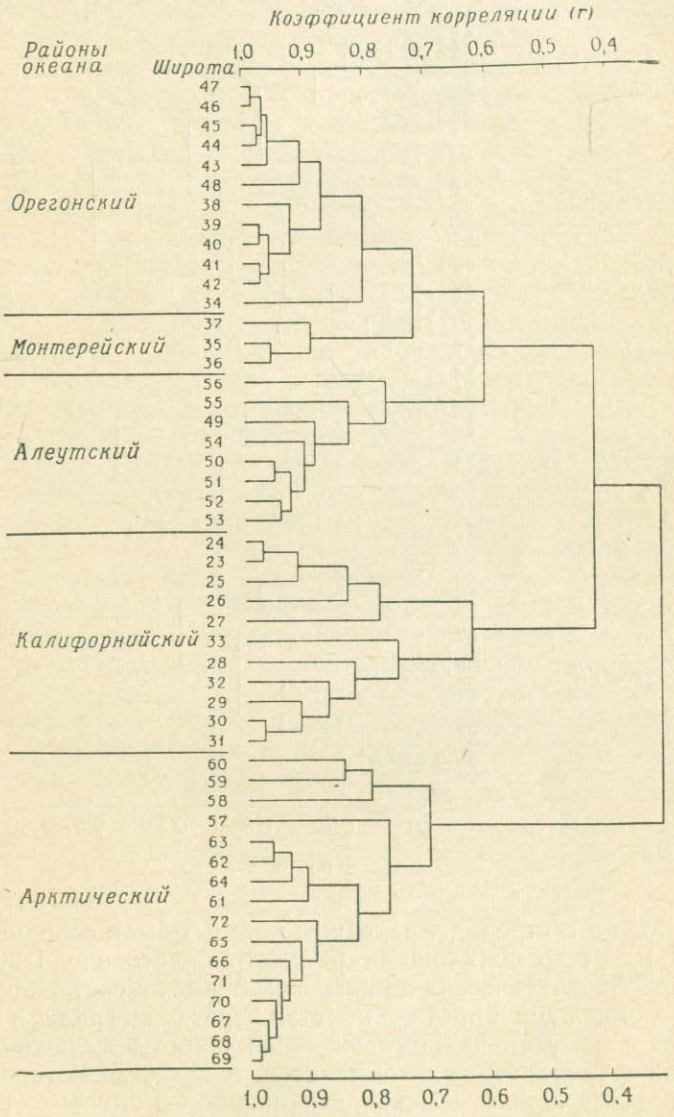
Для сравнения дендрографа и дендрограммы вернемся к нашему примеру, заимствованному из литературы.



Фиг. 4.

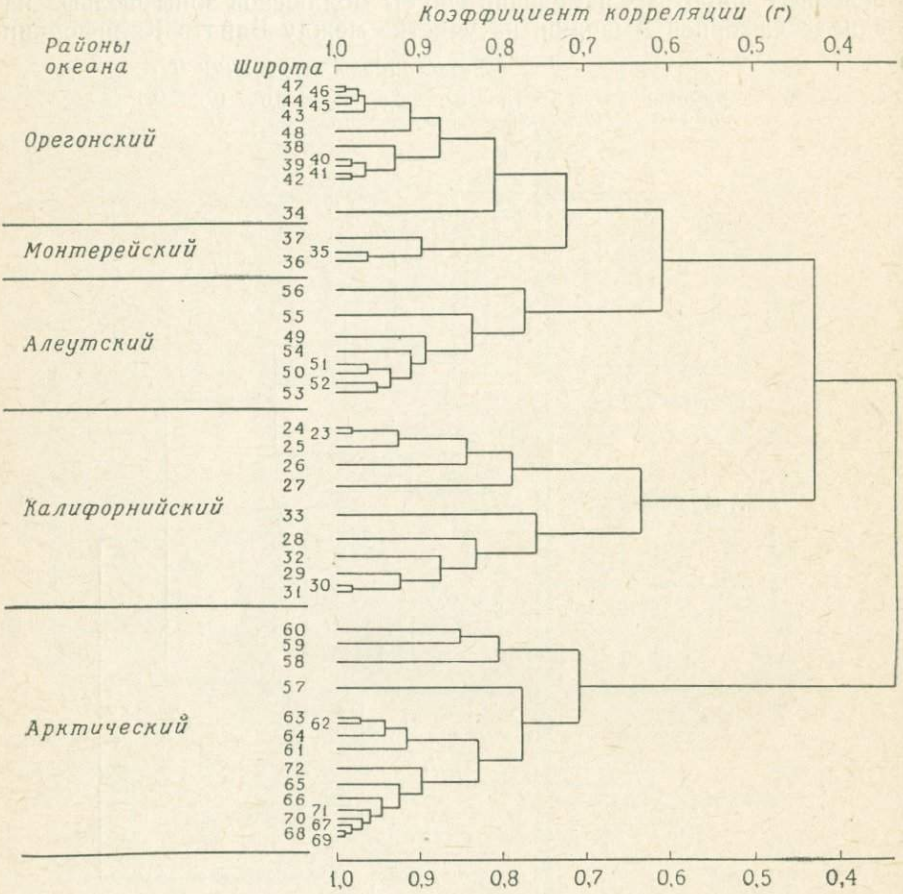
ЭКОЛОГИЧЕСКИЕ ПРИЛОЖЕНИЯ

Этот пример не геологический, но он так наглядно показывает эффективность применения дендрографа, что мы рассмотрим его. В работе Валентина [88] предпринята попытка описать природу основных широтных изменений фауны моллюсков зоны шельфа на западе Северной Америки, на участке между Байа в Калифорнии



Фиг. 5.

и Пойнт-Барроу на Аляске, с помощью числовых методов анализа. В результате этих исследований было предложено внести ряд изменений в систему морских провинций. Список изучаемой фауны включал бентосные формы моллюсков, принадлежащих к классам



Фиг. 6.

пелециподa и гастропода, раковины которых были собраны с глубин до 183 м. Всего по площади рассматривалось около 2077 форм. Метод сбора фактических данных заключался в составлении списка фауны для каждой широтной отметки через один градус в интервале 23 — 72° с. ш. Для каждого значения широты вычислялась мера сходства фаунистических комплексов, в качестве которой использовался коэффициент Джеккарда (подробное описание его можно найти в работе Сокала и Сниса [83]). Этот коэффициент, соответ-

ствующий паре выборок, представляет собой частное от деления числа видов, наблюдаемых в обоих выборках, на общее число видов в этих выборках. Данный коэффициент, как и коэффициент корреляции, принимает значения в интервале от нуля до единицы. Таким образом, каждому значению широты можно поставить в соответствие некоторую случайную величину (число видов), а матрица, образованная значениями коэффициента фаунистического сходства для всех возможных пар этих случайных величин, создаст основу для дальнейшего группового анализа.

В данном примере значения меры Джеккарда были преобразованы в arccos. Полученные преобразованные характеристики рассматривались как значения меры расстояния и на их основе с применением процедуры группирования были построены дендрограммы и дендрограф, которые приведены на фиг. 5 и 6. Эти результаты исключительно наглядны. При выделении главных фаунистических провинций возникают небольшие трудности. Широтные группы образуются на высоких уровнях корреляции и переходят к низшим уровням. Дендрограф более нагляден, чем дендрограмма, в связи с тем, что широкие интервалы между группами разделяют фаунистические провинции, а более узкие интервалы характерны для соотношений внутри провинций. Заметим, что значения широт расположены последовательно внутри провинций, хотя такой порядок не всегда точно соблюдается. Это связано главным образом с изменением направления береговой линии относительно параллелей, что иногда приводит к флуктуациям в выборках. Так, например, значение широты 34° представляет исключение, поскольку оно отнесено к группе более северных значений. Валентин [88] отмечает, что фаунистические изменения непосредственно к югу от Пойнт-Консепшн (около 34-й параллели) связаны с наличием северных видов, свойственных большинству областей распространения холодных течений. Следовательно, комплекс фауны на широте 34° представляет собой смесь видов, распространенных на севере, с видами, присущими данным широтам. Таким образом, по интервалам между группами в дендрографе можно получить хотя бы этот вывод.

ЛЕКЦИЯ 10

Дискриминантный анализ

Р. Мак-Кеммон

ЛИНЕЙНАЯ ДИСКРИМИНАНТНАЯ ФУНКЦИЯ

Выше при рассмотрении вопросов постановки задач классификации была сформулирована задача, заключающаяся в том, каким образом отнести данный объект к одной из заданного множества групп. Такую задачу классификации называют задачей дискриминантного анализа, и в ней заранее предполагается, что функции распределения для групп заданы. Таким образом, она сводится к построению решающего правила, позволяющего относить некоторый наблюдаемый объект к одной из заданных совокупностей, когда заранее неизвестно, к какой из них он принадлежит. Для случая двух совокупностей линейная дискриминантная функция, являющаяся искомым решающим правилом, была построена Фишером [24]. Рассмотрим ее подробнее. Допустим, что на изучаемом объекте сделано p измерений $x_1, x_2, \ldots, x_p$, на основании которых требуется отнести этот объект к одной из двух заданных совокупностей. Идея Фишера заключалась в сокращении числа p измеряемых признаков до одной величины, представляющей собой линейную функцию

$$z = l_1 x_1 + l_2 x_2 + \ldots l_p x_p, \tag{1}$$

и в отнесении объекта к одной из двух совокупностей в зависимости от того, какое значение примет z (больше или меньше критического). Предполагается, что каждая совокупность охарактеризована одним и тем же набором измерений, проведенных на n_1 и n_2 объектах, которые взяты из этих совокупностей, а коэффициенты $(l_1, \ldots, l_p)$ соответствуют максимуму выражения

$$F = \frac{(\bar{z}_1 - \bar{z}_2)}{s^2(z)}, \tag{2}$$

где $\bar{z}_1$ и $\bar{z}_2$ — выборочные средние для обеих совокупностей, а $s(z)$ — оценка обобщенного стандартного отклонения, вычисляемая по формуле

$$s(z) = \left[\frac{n_1}{n_1 + n_2} s_1^2(z) + \frac{n_2}{n_1 + n_2} s_2^2(z)\right]^{\frac{1}{2}}, \tag{3}$$

где $s_1(z)$ и $s_2(z)$ — оценки стандартных отклонений для каждой совокупности. Выражение (2) представляет собой отношение квадрата разности между оценками средних значений в совокупностях к обобщенной оценке дисперсии внутри совокупностей. При определении критического значения z_c Фишер исходит из предположения, что вероятности принадлежности изучаемого объекта к той или другой совокупности равны и что стоимости потерь за счет ошибочного классификационного отнесения объекта так же равны. В этих условиях

$$z_c = \frac{s_2(z)\,\bar{z}_1 + s_1(z)\,\bar{z}_2}{s_1(z) + s_c(z)}. \tag{4}$$

Если стандартные отклонения равны, то равенство (4) можно записать в следующем виде:

$$z_c = \frac{\bar{z}_1 + \bar{z}_2}{2}. \tag{5}$$

Фишер показал, что набор коэффициентов $\{l_1, \ldots, l_p\}$, соответствующий максимуму выражения (2), пропорционален набору коэффициентов $\{l_1, \ldots, l_p\}$, совместно удовлетворяющих p линейным уравнениям

$$\begin{aligned} &s_{11}l_1 + s_{12}l_2 + \ldots + s_{1p}l_p = cd_1, \\ &s_{21}l_1 + s_{22}l_2 + \ldots + s_{2p}l_p = cd_2, \\ &\cdot \\ &\cdot \\ &\cdot \\ &s_{p1}l_1 + s_{p2}l_2 + \ldots + s_{pp}l_p = cd_p, \end{aligned} \tag{6}$$

где

$$d_i = \bar{x}_{i1} - x_{i2}, \qquad i = 1, 2, \ldots, p,$$

$$s_{ij} = \sum_{q=1}^{2} \sum_{k=1}^{n_l} (x_{ikq} - \bar{x}_{iq})(x_{jkq} - \bar{x}_{jq}), \quad i, j = 1, \ldots, p,$$

а величина c в выражении (6) является константой. Не теряя общности, можно записать

$$\begin{aligned} &l_1 = s^{11}d_1 + s^{12}d_2 + \ldots + s^{1p}d_p, \\ &\cdot \\ &\cdot \\ &\cdot \\ &l_p = s^{p1}d_1 + s^{p2}d_2 + \ldots + s^{pp}d_p, \end{aligned} \tag{7}$$

где s^{ij} — элементы матрицы, обратной матрице, элементами которой являются s_{ij}.

Подставив вычисленные значения $l_1, \ldots, l_p$ в выражение (1), получим решающее правило для отнесения изучаемого объекта к одной из двух заданных совокупностей. Критическое значение можно определить с помощью выражения (4) или (5), если дисперсии в обеих совокупностях равны.

Если допустить, что изучаемые совокупности описываются многомерными нормальными распределениями с одной и той же ковариационной матрицей, то для построения решающего правила можно воспользоваться отношением правдоподобия. Пусть эти распределения будут $N(\mu^{(1)}, \Sigma)$ и $N(\mu^{(2)}, \Sigma)$, где $\mu^{(i)\prime} = (\mu_1^{(i)}, \ldots, \mu_p^{(i)})$ является вектором средних значений для совокупности с номером $i=1, 2$, а Σ — ковариационная матрица, одинаковая для обеих совокупностей. Согласно Андерсону [1], который ссылается на Вальда [90], первым использовавшего этот подход, можно записать i-ю плотность вероятности в следующем виде:

$$P_i(\mathbf{x}) = \frac{1}{2\pi^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu^{(i)})' \Sigma^{-1} (\mathbf{x}-\mu^{(i)})\right], \tag{8}$$

где $|\Sigma|$ — детерминант матрицы Σ. Отношение правдоподобия будет представлено выражением

$$\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\mu^{(1)})' \Sigma^{-1} (\mathbf{x}-\mu^{(1)})\right]}{\exp\left[-\frac{1}{2}(\mathbf{x}-\mu^{(2)})' \Sigma^{-1} (\mathbf{x}-\mu^{(2)})\right]}. \tag{9}$$

Прологарифмировав это выражение и приведя подобные члены, получим

$$\mathbf{x}'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}). \tag{10}$$

Первый член этого выражения идентичен дискриминантной функции, определенной Фишером. В основе решающего правила лежит предположение, что изучаемый объект может равновероятно относиться к первой или второй совокупности, а цены ошибочных решений равны. Решающее правило для объекта, охарактеризованного вектором $\mathbf{x}$, будет иметь следующий вид:
объект относится к первой совокупности, если

$$\mathbf{x}' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geqslant \frac{1}{2}(\mu^{(1)}+\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)});$$

объект относится ко второй совокупности, если

$$\mathbf{x}' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) < \frac{1}{2}(\mu^{(1)}+\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}).$$

Если параметры распределения совокупностей оцениваются по n_1 и n_2 пробам соответственно, то оценкой для $\boldsymbol{\mu}^{(i)}$ $(i=1, 2,)$ будет среднее арифметическое

$$\bar{x}^i=\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_j^i,$$

а оценка $\mathbf{S}$ для ковариационной матрицы Σ будет определена следующим выражением:

$$(n_1+n_2-2)\,\mathbf{S}=\sum_{j=1}^{n_1}(\mathbf{x}_j^{(1)}-\bar{x}^{(1)})(\mathbf{x}_j^{(1)}-\bar{x}^{(1)})'+$$

$$+\sum_{j=1}^{n_2}(\mathbf{x}_j^{(2)}-\bar{x}^{(2)})(\mathbf{x}_j^{(2)}-\bar{x}^{(2)})'.$$

Подставив значения $\bar{x}^{(1)}$, $\bar{x}^{(2)}$ и $\mathbf{S}$ в выражение (10), получим

$$\mathbf{x}'\mathbf{S}^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})-\frac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})'\mathbf{S}^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}). \qquad (11)$$

КВАДРАТИЧНАЯ ДИСКРИМИНАНТНАЯ ФУНКЦИЯ

В том случае, когда ковариационные матрицы распределений, соответствующих двум рассматриваемым совокупностям, не равны, возникают некоторые трудности в построении дискриминантной функции.

В этих условиях логарифм отношения правдоподобия, определенный выражениями (9) и (10), примет следующий вид:

$$\ln\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})}=(\mathbf{x}-\boldsymbol{\mu}^{(2)})'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}^{(2)})-(\mathbf{x}-\boldsymbol{\mu}^{(1)})'\,\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}^{(1)})+$$

$$+\ln|\Sigma_2^{-1}|-\ln|\Sigma_1^{-1}|, \qquad (12)$$

где Σ_1 и Σ_2 — две ковариационные матрицы. Выражение (12) представляет собой сумму разностей между двумя квадратичными формами и константами. Оценки параметров распределения по двум выборкам объемом n_1 и n_2 вычисляются, как и в предыдущем случае, за исключением двух оценок $\mathbf{S}_i$ для Σ_i, которые вычисляются по следующей формуле:

$$(n_i-1)\,\mathbf{S}_i=\sum_{j=1}^{n_i}(\mathbf{x}_j^{(i)}-\bar{x}^{(i)})(\mathbf{x}_j^{(i)}-\bar{x}^{(i)})', \quad (i=1, 2).$$

Решающее правило заключается в том, что объект, охарактеризованный вектором $\mathbf{x}$, относится к первой совокупности, если

выражение (12) после замены параметров их оценками и подстановки в него $\mathbf{x}$ примет значение больше нуля или равное нулю. Во всех остальных случаях объект будет относиться ко второй совокупности. Купер [17] показал, что квадратичные дискриминантные функции весьма надежны в условиях большинства одновершинных симметричных распределений. В работе Барнеби [7] описан способ такого преобразования двух случайных величин, имеющих неравные ковариационные матрицы, который приводит к приблизительно равным детерминантам, что сводит дискриминантную функцию (12) к сумме двух квадратичных форм и избавляет от необходимости вычисления детерминантов двух обратных матриц.

НЕПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

До сих пор мы имели дело с совокупностями, для которых функции распределения известны. Однако нередко наблюдаются случаи (например, дискретные данные или малые выборки из совокупностей с непрерывными распределениями), когда плотности вероятности заранее определить нельзя и, следовательно, невозможно получить оценки параметров распределения для построения решающего правила. В подобной ситуации мы вынуждены рассматривать распределение как весьма произвольное, зависящее непосредственно от наблюдаемых данных.

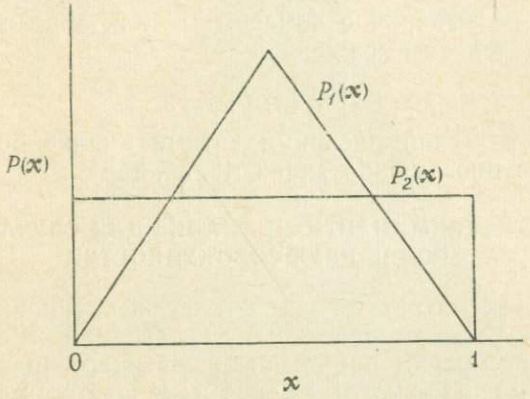
Один из вариантов решения этой задачи заключается в непосредственном получении оценки отношения правдоподобия. Для этого требуется определить оценку функции плотности в точке, соответствующей наблюдаемому вектору $\mathbf{x}_0$, при условии, что до этого были проведены n_1 и n_2 наблюдений в двух известных совокупностях. Если обозначить через $P_1(\mathbf{x}_0)$ плотность вероятности для первой совокупности в точке $\mathbf{x}_0$, а через $P_2(\mathbf{x}_0)$ — для второй, то отношение правдоподобия L можно представить в виде следующей формулы:

$$L = \frac{P_1(\mathbf{x}_0)}{P_2(\mathbf{x}_0)}. \tag{13}$$

Вектор $\mathbf{x}_0$ будет принадлежать первой совокупности, если $L \geqslant 1$; во всех остальных случаях мы будем относить $\mathbf{x}_0$ ко второй совокупности. На фиг. 1 приведен весьма простой случай треугольного и равномерного распределений одномерной случайной величины, принимающей значения в интервале от нуля до единицы. Пусть треугольное распределение соответствует первой совокупности, а равномерное — второй. Тогда для любого наблюдаемого значения $\mathbf{x}_0$ объект, охарактеризованный $\mathbf{x}_0$, будет принадлежать первой совокупности, если $^1/_4 \leqslant \mathbf{x}_0 \leqslant \frac{3}{4}$, поскольку в этом случае $L > 1$. Наоборот, если $0 \leqslant \mathbf{x}_0 < \frac{1}{4}$ или $\frac{3}{4} < \mathbf{x}_0 < 1$, то объект следует отнести ко второй совокупности, так как для этих интервалов $L < 1$.

В результате мы получили решающее правило. Теперь необходимо построить метод для получения оценки функции плотности вероятности.

Допустим, что мы располагаем n_1 наблюдениями над объектами из совокупности с треугольным распределением, показанным на фиг. 1, и аналогично n_2 наблюдениями над объектами из совокупности с равномерным распределением. Следует ожидать, что наблюдаемые значения будут скорее обладать тенденцией к группировке около средней области значений $\mathbf{x}$, чем около крайних участков. Если новое наблюдение $\mathbf{x}_0$ попало в эту среднюю зону, то скорее всего окажется, что среднее расстояние между $\mathbf{x}_0$



Фиг. 1. Плотности вероятности треугольного и равномерного распределений.

и набором n_1 наблюдений, взятых из совокупности с треугольным распределением, будет меньше среднего расстояния между $\mathbf{x}_0$ и набором n_2 проб из совокупности с равномерным распределением. Для результата наблюдения, попавшего в крайние области, наиболее вероятно обратное соотношение. Таким образом, эти средние расстояния можно использовать как оценки величины, обратной значениям плотности вероятности, и, следовательно, для отнесения изучаемого объекта с помощью выражения (13) к одной из двух совокупностей. Этот способ коренным образом отличается от метода, предложенного Махалонобисом [59], который заключается в рассмотрении оценки обобщенного расстояния D^2, применяемой в условиях многомерных распределений. В данном случае мы не делаем никаких предположений о форме распределения. Таким образом, мы можем непосредственно получить оценки плотности вероятности, а для определения меры расстояния применить схему обратного взвешивания.

Если нужно построить решающее правило для классификационного отнесения изучаемого объекта к одной из двух совокупностей по значениям более чем одной случайной величины, а информация

о виде распределения отсутствует, то понятие расстояния следует обобщить на многомерное пространство. Тогда, если $\mathbf{x}_0' = (x_{10}, x_{20}, \ldots, x_{p0})$ — транспонированный наблюдаемый p-мерный вектор-столбец $\mathbf{x}_0$, можно определить оценку плотности вероятности $P_i(\mathbf{x}_0)$ p-мерной случайной величины в точке $\mathbf{x}_0$ с помощью следующего выражения:

$$P_i(\mathbf{x}_0) \simeq \frac{1}{kn_i} \sum_{j=1}^{n_i} \exp\left[-(\mathbf{x}_0 - \mathbf{x}_j^{(i)})'(\mathbf{x}_0 - \mathbf{x}_j^{(i)})\right], \tag{14}$$

где $\mathbf{x}_j^{(i)}$ — p-мерный вектор, являющийся выборочным значением с номером j из выборки с номером $i=1, 2$, а величина k — константа, определенная выражением

$$\int P_i(\mathbf{x})\, d\mathbf{x} = 1. \tag{15}$$

Непараметрический дискриминантный анализ более подробно рассмотрен во многих работах [82, 48, 36, 25].

ДИСКРИМИНАНТНЫЙ АНАЛИЗ ДЛЯ СЛУЧАЯ БОЛЕЕ ДВУХ СОВОКУПНОСТЕЙ

Выше мы рассматривали задачу классификационного отнесения изучаемого объекта к одной из двух заданных совокупностей. Теперь мы рассмотрим построение решающего правила, по которому объект можно отнести к одной из k заданных совокупностей. При увеличении числа совокупностей до $k>2$ в построении решающего правила не возникает существенных трудностей. Число решений необходимо дополнить в соответствии с числом совокупностей, которым может принадлежать изучаемый объект.

Если мы воспользуемся линейной дискриминантной функцией Фишера, определенной выражением (10), сделав те же предположения относительно стоимости ошибок классификации, нормальности распределений и равенства ковариационных матриц, то построение решающего правила будет определено следующей функцией:

$$u_{ij}(\mathbf{x}) = \ln \frac{P_i(\mathbf{x})}{P_j(\mathbf{x})} = \left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(j)})\,\Sigma^{-1}(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)})\right], \tag{16}$$

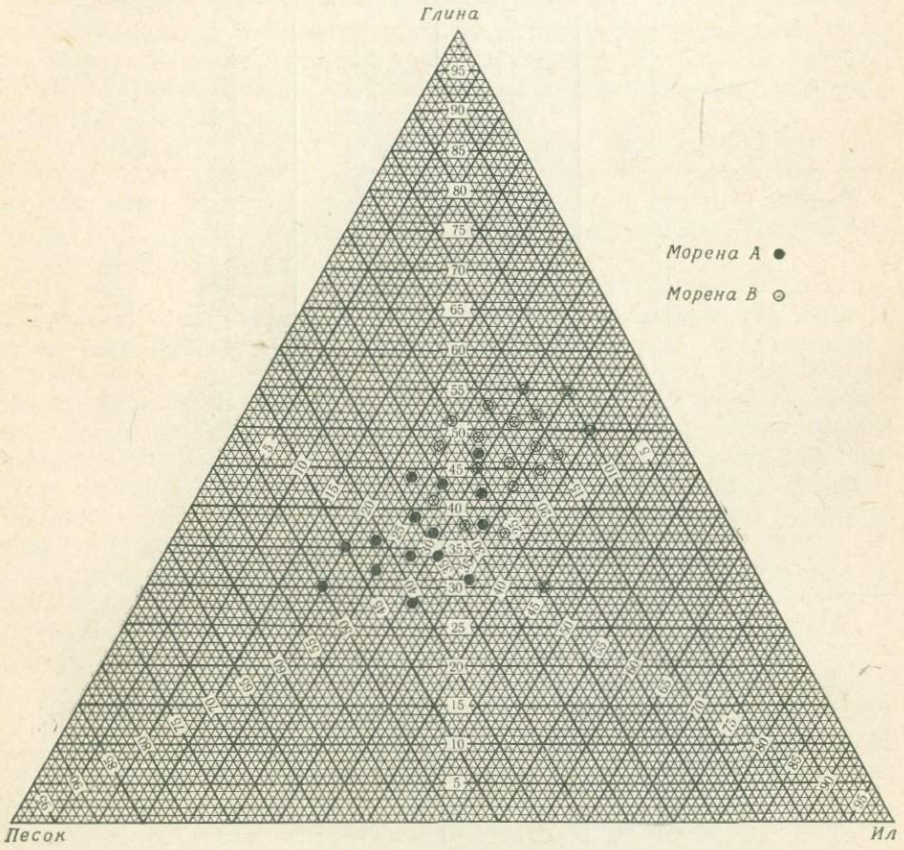
удовлетворяющей в случае принадлежности значения $\mathbf{x}$ совокупности с номером j следующему условию:

$$u_{ij}(\mathbf{x}) > u_{i\nu}(\mathbf{x}), \quad \nu = 1, \ldots, k, \ \nu \neq j.$$

Более детально эта задача рассмотрена в работе Андерсона [1]. Таким образом, в случае k совокупностей в процессе принятия решения требуется провести $\binom{k}{2}$, или $k(k-1)/2$, сравнений значений u_{ij}.

Если ковариационные матрицы, соответствующие совокупностям, не равны, то зависимость будет аналогична выражению (12). Необходимо отметить, что в данной ситуации классификационное отнесение результата наблюдения $\mathbf{x}_0$ к совокупности с номером j осуществляется по минимуму функции

$$(\mathbf{x}_0 - \mu^{(j)})' \Sigma_j^{-1} (\mathbf{x}_0 - \mu^{(j)}) + \ln \left| \Sigma_j^{-1} \right| . \qquad (17)$$



Фиг. 2. Диаграмма состава двух морен.

В качестве примера рассмотрим задачу классификационного распознавания двух морен по структурным характеристикам их отложений. Обозначим эти морены через A и B соответственно и допустим, что различия между ними установлены на основании полевых наблюдений. Предположим, что из морены A взято 15, а из морены B — 20 проб и в каждой из них проведен гранулометрический анализ. Результаты этого анализа представляют собой

Таблица 1

Исходные данные и результаты вычислений линейной дискриминантной функции для моренных отложений

	Песок	Ил	Глина
	Морена *A*		
1	50	20	30
2	45	20	35
3	43	35	22
4	41	31	28
5	41	23	36
6	38	28	34
7	35	26	39
8	34	29	37
9	35	31	34
10	33	36	31
11	33	23	44
12	30	27	43
13	28	36	36
14	26	32	42
15	24	29	47
	Морена *B*		
1	32	27	41
2	30	32	38
3	26	37	37
4	28	24	48
5	25	30	45
6	25	24	51
7	23	31	46
8	22	35	43
9	20	40	40
10	21	33	46
11	20	27	53
12	18	37	45
13	17	37	46
14	18	31	51
15	15	38	47
16	15	33	52
17	15	30	55
18	10	35	55
19	25	35	40
20	10	40	50
Координаты дискриминантной линии			
29	29	38	
26	40	33	

Классификационная матрица

	A	*B*
A	13	2
B	2	8

значения содержаний в процентах трехразмерных фракций — песчаной, илистой и глинистой. Для наглядности полученные данные можно нанести на треугольную диаграмму в виде 35 точек, что показано на фиг. 2. В этом примере мы ограничимся рассмотрением только линейной дискриминантной функции. Задача заключается в том, чтобы найти такую прямую линию, которая наилучшим образом разделяла бы два типа моренных отложений, представленные на диаграмме. Даже беглого взгляда на фиг. 2 достаточно, чтобы убедиться, что такую прямую линию, разделяющую два типа моренных отложений на неперекрывающиеся области, построить нельзя. Однако можно построить линию, обеспечивающую минимальную область перекрытия.

В таком случае следует воспользоваться методом Фишера для построения линейной дискриминантной функции, который был описан выше. Как только такая линия будет построена и нанесена на фиг. 2, любую новую пробу, нанесенную на эту диаграмму, легко классифицировать.

В табл. 1 приведены исходные данные и полученные по ним результаты вычислений. В этой же таблице приведено два набора координат, соответствующих двум точкам на фиг. 2. Линия, проходящая через эти точки в пределах границ треугольника, будет наилучшим приближением к линейной дискриминантной функции в условиях равных ковариационных матриц. Точки, соответствующие результатам наблюдений, располагающиеся на диаграмме по одну сторону от этой линии, относятся к морене *A*, а точки, находящиеся по другую сторону, — к морене *B*. Для того чтобы охарактеризовать эффективность классификации, в табл. 1 приведена классификационная матрица, показывающая результаты классификации известных 35 наблюдений с помощью полученной дискриминантной функции. Нетрудно видеть, что для 31 наблюдения из 35 классификация проведена правильно.

Таким образом, можно определить эффективность как $(31 : 35) \times \times 100 = 88\%$. Это значит, что при равных шансах появления проб из морен *A* и *B* при последующей классификации приблизительно 88% от их общего числа будет классифицировано правильно. Несмотря на то что данные, использованные в этом примере, представляют собой искусственную модель, он наглядно показывает, как описанный алгоритм использовать в реальных задачах.

СПИСОК ЛИТЕРАТУРЫ

1. *Anderson T. W.,* Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York, 374, 1958. (Имеется русский перевод: *Андерсон Т.,* Введение в многомерный статистический анализ. М., Физматгиз, 1963.)
2. *Ball G. H.,* Data analysis in the social sciences what about the details, Proc. Fall Joint Computer Conference, **27**, Pt. I, 533—559, 1965.
3. *Behrens E. W.,* Environment reconstruction for a part of the Glen Rose limestone, central Texas, Sedimentology, **4**; 65—111, 1965.
4. *Blumenthal L. M.,* Theory and applications of distance geometry, Oxford Univ. Press, London, 347, 1953.
5. *Bonham-Carter,* A numerical method of classification using qualitative and semi-quantitative data, as applied to the facies analysis of limestone, Can. Pet. Geology Bull., **13**, № 4, 482—502, 1965.
6. *Briggs L. I., Pollack H. N.,* Science, **155**, 453, 1967.
7. *Burnaby T. P.,* Growth-invariant discriminant functions and generalized distances, Biometrics, **22**, 96—110, 1966.
8. *Burnaby T. P.,* Distribution-free quadratic discriminant functions in paleontology, Kansas Comput. Contrib., **7**, Kansas Geol. Survey, 70—77, 1966.
9. *Cattell R. B.,* Factor analysis: An introduction to essentials, II. The role of factor analysis in research, Biometrics, **21**, N° 2, 405—435, 1965.
10. *Chayes F.,* Classification in a ternary diagram by means of discriminant functions, Amer. Mineral., **50**, 1618—1633, 1965.
11. *Chayes F.,* On locating field boundaries in simple phase diagrams by means of discriminant functions, Amer. Mineral., **53**, 359—371, 1968.
12. *Chayes F., Velde D.,* On distinguishing basaltic lavas of circumoceanic and oceanic — island type by means of discriminant functions, Amer. Jour. Sci., **263**, 206—222, 1965.
13. *Chayes F., Mackenzie W. S.,* On correlations between variables of constant sum, Jour. Geophys. Res., **65**, 4185—4193, 1960.
14. *Chayes F., Kruskal W.,* An approximate test for correlations between proportions, Jour. Geology, **74**, 692—702, 1966.
15. *Crommelin R. D.,* De Krumbein — Tukey methode toegepast op de orderlinge vergelijking van Pleistocene Zanden, Geologie en Mijnbouw, **44**, 242—250, 1965.
16. *Cooley W. W., Lohnes P. R.,* Multivariate procedures for the behaviorial sciences, John Wiley & Sons, New York, 211, 1962.
17. *Cooper P. W.,* Statistical classification with quadratic forms, Biometrika, **50**, 439—448, 1963.
18. *Cooper P. W.,* Quadratic discriminant functions in pattern recognition, IEEE Trans. on Information Theory, **11**, 313—315, 1965.
19. *Cornfield J.,* Discriminant functions, Rev. Internat. Stat. Inst., **35**, 142—153, 1967.
20. *Dempster A. P.,* Continuous multivariate analysis, Addison — Wesley Reading Mass., 388, 1969.
21. *Edwards A. W. F., Cavalli-Sforza L. L.,* A method for cluster analysis, Biometrics, **21**, N° 2, 362—375, 1965.
22. *Feller W.,* An introduction to probability theory and its application. I. 2nd ed., John Wiley & Sons, New York, 461, 1957.

23. *Fenneman N. M.,* The rise of physiography, Geol. Soc. of Amer. Bull., **50**, 349—359, 1939.
24. *Fisher R. A.,* The use of multiple measurements in taxonomic problems, Ann. of Eugenics, **7**, 179—188, 1936.
25. *Freeman J. J.,* Experiments in discrimination and classification, Pattern Recognition, **1**, 207—218, 1969.
26. *Goodall D. W.,* Objective methods for the classification of vegetation, I. The use of positive interspecific correlation, Aust. Jour. Bot. **1**, 39—63, 1953.
27. *Gower J. C.,* Some distance of latent root and vector methods used in multivariate analysis, Biometrika, **53**, 325—338, 1966.
28. *Griffiths J. C.,* Application of discriminant function as a classification tool in the geosciences, Kan. Comput. Contrib., 7, Kansas Geol. Survey, 48—52, 1966.
29. *Griffiths J. C.,* Scientific method in analysis of sediments, McGraw-Hill, New York, 1967. (Имеется русский перевод: Гриффитс Дж., Научные методы исследования осадочных пород, М., «Мир», 1971.)
30. *Hagmeier E. M., Stults C. D.,* A numerical analysis of the distributional patterns of North American Mammals, Systematic Zoology, **13**, 125—155, 1964.
31. *Horton R. E.,* Erosional development of streams and their drainage basins; Hydrophysical approach to quantitative morphology, Geol. Soc. of Amer. Bull., **56**, 275—370, 1945.
32. *Huitson A.,* The Analysis of Variance, Hafner Pub., N. Y., 1966.
33. *Harbaugh J. W., Demirmen F.,* Application of factor analysis to petrologic variations of Americus limestone (Lower Permian), Kansas and Oklahoma, Kan. Geol. Survey Dist. Pub., **15**, 50, 1964.
34. *Harbuagh J. W., Merriam D. F.,* Computer applications in stratigraphic analysis, John Wiley & Sons, New York, 282, 1968.
35. *Harman H. H.,* Modern factor analysis, Univ. of Chicago Press, Chicago, Illinois, 474, 1967.
36. *Hills M.,* Discrimination and allocation with discrete data, Appl. Stat., **16**, 237—250, 1967.
37. *Hodson F. R., Sneath P. H. A., Doran J. E.,* Some experiments in the numerical analysis of archaeological data, Biometrika, **53**, 311—324, 1966.
38. *Hotelling H.,* Analysis of a complex of statistical variable into principal components, J. Educ. Psych., **24**, 417—441, 498—520, 1933.
39. *Imbrie J., Purdy E. G.,* Classification of modern Bahamian carbonate sediments, Am. Assoc. Petrol. Geologists, Mem. 1, 253—272, 1962.
40. *Jizba Z. V.,* Sand evolution simulation, Jour. Geology, **74**, 734—743, 1966.
41. *Johnson S. C.,* Hierarchical clustering schemes, Psychometrika, **32**, 241—254, 1967.
42. *Jones T. A.,* J. Sed. Petrology, **38**, **61**, 1968.
43. *Jolicoeur P., Mosimann J. E.,* Size and shape variation in the Painted Turtle, A principal component analysis, Growth, **24**, 339—354, 1960.
44. *Kauffman M. E.,* Statistical analysis of certain characteristics of the Susquehanna River terrace deposits, Pennsylvania, unpublished M. S. thesis, Northwestern University, 89, 1957.
45. *Kauffman M. E.,* Using quantitative methods in geology, Geologie en Mijnbouw, **45**, 231—237, 1966.
46. *Kendall M. G.,* A Course in Multivariate Analysis, Charles Griffin & Company, Ltd., London, England, 185, 1957.
47. *Kendall M. G.,* A course in the geometry of n dimensions, Griffin's Statistical Monograph Num. 8, Chas. Griffin & Company, London, England, 63, 1961.
48. *Kendall M. G., Stuart A.,* The advanced theory of statistics, **3**, Hafner Publ. Co., New York, 552, 1966.
49. *Klovan J. E.,* The use of factor analysis in determining depositional environments from grain-size distributions, Jour. Sed. Petrology, **36**, N° 1, 115—125, 1966.
50. *Krumbein W. C., Graybill F. A.,* An Introduction to Statistical Models in Geology, McGraw-Hill Co., N. Y., 475, 1965. (Имеется русский перевод: *Крам-*

бейн У., Грейбилл Ф., Статистические модели в геологии, М., «Мир», 1969.)
51. *Krumbein W. C., Tukey J. W.*, Multivariate analysis of mineralogic, lithologic, and chemical composition of rock bodies, Jour. Sed. Petrology, **26**, 322, 337, 1946.
52. *Krumbein W. C.*, Kans. Geol. Survey, Computer Contrib., **13**, 1967.
53. *Krumbein W. C., Dacey M. F.*, Jour. Int. Assoc. Math. Geol., **1**, № 1, 1969.
54. *Klovan J. E., Billings G. K.*, Classification of geological samples by discriminant-function analysis, Bull. Can. Pet. Geol., **3**, 313—330, 1967.
55. *Lance G. N., Williams W. T.*, Computer programs for hierarchical polythetic classification ("similarity analyses"), The Computer Journal, **9**, 60—64, 1966.
56. *Lance G. N.*, A general theory of classificatory sorting strategies, 1. Hierarchical systems, The Computer Journal, **9**, 373—380, 1967.
57. *Macnaughton-Smith P.*, Some statistical and other numerical techniques for classifying individuals, Studies in the causes delinquency and the treatment of offenders, 6. H.M.S.O., London, England, 33, 1965.
58. *MacQueen J. B.*, Some methods for classification and analysis of multivariate observations, West. Manag. Sci. Inst. Working Paper, 96, UCLA, 27, 1966.
59. *Mahalanobis P. C.*, On the generalized distance in statistics, Proc. Nat. Inst. Sci. India, **12**, 49—55, 1936.
60. *Mattson R. L., Dammann J. E.*, A technique for determining and coding subclasses in pattern recognition problems: IBM Jour., **9**, 294—302, 1965.
61. *McCammon R. B.*, Principal component analysis and its application in large scale correlation studies, Jour. Geology, **74**, № 5, pt. 2, 721—733, 1966.
62. *McCammon R. B.*, Multiple component analysis and its application in classification of environments, Am. Assoc. Petroleum Geologists, **52**, 2178—2196, 1968a.
63. *McCammon R. B.*, The dendrograph. A new tool for correlation, Geol. Soc. Amer. Bull., **79**, 1663—1670, 1968b.
64. *Merriam D. F., Sneath P. H. A.*, Quantitative comparison of contour maps. Jour. Geophysical Research, **71**, 1105—1115, 1966.
65. *Morrison D. F.*, Multivariate statistical methods, McGraw-Hill, New York, 338, 1967.
66. *Murdoch D. C.*, Linear algebra for undergraduates, John Wiley & Sons, New York, 239, 1957.
67. *Moss J. H., Ritter D. F.*, New evidence regarding the Binghamton substage in the region between the Finger Lakes and Catskills, New York. Amer. Jour. of Science, **260**, 81—106, 1962.
68. *Mood A. M., Graybill F. A.*, Introduction to the theory of statistics, McGraw-Hill, New York, 1963.
69. *Miller R. L., Kahn J. S.*, Statistical analysis in the geological sciences, John Wiley and Sons, New York, 1962. (Имеется русский перевод: *Миллер Р., Кан Дж.*, Статистический анализ в геологических науках, М., «Мир», 1965.)
70. *Oertel G., Walton E. K.*, Sedimentology, **9**, 157, 1967.
71. *Peltier L. C.*, Pleistocene Terraces of the Susquehanna River, Pennsylvania, Penna. Geol. Surv. Bull. G23, 1949.
72. *Parks J. M.*, Cluster analysis applied to multivariate data, Jour. Geology, **74**, № 5, pt. 2, 703—715, 1966.
73. *Pearson K.*, On lires and planes of closest fit to systems of points in space, Phil. Mag. 2 (6th Series), 559—572, 1901.
74. *Purdy E. G.*, Recent calcium carbonate facies of the Great Bahama Bank; Ph. D. dissertation, Columbia Univ., 174, 1960.
75. *Purdy E. G.*, Recent calcium carbonate facies of the Great Bahama Bank, 1. Petrography and reaction groups, Jour. Geology, **71**, № 3, 334—355, 1963a.
76. *Purdy E. G.*, Recent calcium carbonate facies of the Great Bahama Bank, 2. Sedimentary Facies, Jour. Geology, **71**, № 4, 472—479, 1963b.
77. *Rao C. R.*, The use and interpretation of principal component analysis in applied research, Sankhya, **26**, 329—358, 1964.

78. *Riordan J.*, An introduction to combinatorial analysis, John Wiley & Sons, New York, 244, 1958.
79. *Rowell A. J.*, A numerical taxonomic study of the Chonetacean brachiopods, in Essays in paleontology and stratigraphy, R. C. Moore Commemorative Volume, Dept. Geology, Univ. Kansas Spec. Publication, 2, The University of Kansas Press, Lawrence, 626, 1967.
80. *Rubin J.*, Optimal classification into groups, An approach for solving the taxonomy problem, Jour. Theoretical Biology, **15**, 103—144, 1967.
81. *Seal H. L.*, Multivariate statistical analysis for biologists, John Wiley & Sons, Inc., New York, 21, 1964.
82. *Sebestyen G. S.*, Decision-making processes in pattern recognition., The Macmillan Co., New York, 162, 1962.
83. *Sokal R. R., Sneath P. H. A.*, Principles of numerical taxonomy, W. H. Freeman, San Francisco, 359, 1963.
84. *Shreve R. L.*, Jour. Geology, **75**, 178, 1967.
85. *Shreve R. L.*, Jour. Geology, **77**, 397, 1968.
86. *Scherer W.*, Unpubl. ms. on Markov chains (to be published in early 1969 as ONR report), 1968.
87. *Toomey D. F.*, Application of factor analysis to a facies study of Leavenworth limestone (Pennsylvanian — Virginian) of Kansas and environs, Kan. Geol. Survey Spec. Dist. Pub., **27**, 27, 1966.
88. *Valentine J. W.*, Numerical analysis of marine molluscan ranges on the extratropical northeastern Pacific shelf, Limnology and Oceanography, **11**, 198—211, 1966.
89. *Valentine J. W., Peddicord R. G.*, Evaluation of fossil assemblages by cluster analysis, Jour. Paleontology, **41**, 502—507, 1967.
90. *Wald A.*, On a statistical problem arising in the classification of an individual into one two groups, Ann. Math. Stat., **15**, 145—162, 1944.
91. *Ward J. H.*, Hierarchical grouping to optimize an objective function, Am. Stat. Assoc., 58—236—244, 1963.
92. *Wilks S. S.*, Mathematical Statistics, John Wiley & Sons, Inc., New York, 564—568, 1962.
93. *Williams W. T., Lambert J. M.*, I. Association-analysis in plant communities, Jour. Ecology, **47**, 83—101, 1969.
94. *Williams W. T., Lambert J. M., Lance G. N.*, Multivariate methods in plant ecology, V. similarity analyses and information-analysis, Jour. Ecology, **54**, 428—445, 1966.
95. *Watson R. A.*, Jour. Geology, **77**, 488, 1969.

СОДЕРЖАНИЕ

УВАЖАЕМЫЙ ЧИТАТЕЛЬ!

Ваши замечания о содержании книги, ее оформлении, качестве перевода и другие просим присылать по адресу: 129820, Москва, И-110, ГСП, I Рижский пер., д. 2, издательство «Мир».

У. Крамбейн, М. Кауфмен, Р. Мак-Кеммон

МОДЕЛИ ГЕОЛОГИЧЕСКИХ ПРОЦЕССОВ

Редактор Г. П. Романович
Художник Г. Д. Конякина
Художественный редактор Ю. С. Урманчеев
Технический редактор Г. Б. Алюлина
Корректор Е. Г. Литвак

Сдано в набор 10/X 1972 г. Подписано к печати 21/II 1973 г. Бумага № 2 60×90¹/₁₆, бум. л. 4,75, печ. л. 9,50. Уч.-изд. л. 9,21. Изд. № 5/6579
Цена 93 коп. Зак. 483.

ИЗДАТЕЛЬСТВО «МИР»
Москва, 1-й Рижский пер., 2

Ленинградская типография № 8
«Союзполиграфпрома» при Государственном
комитете Совета Министров СССР
по делам издательств, полиграфии
и книжной торговли.
190000, Ленинград, Прачечный пер., 6

93 коп.

ИЗДАТЕЛЬСТВО МИР