Высшее профессиональное образование

Ю.Г. Пузаченко

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЭКОЛОГИЧЕСКИХ И ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЯХ

Учебное пособие

Естественные науки

ACADEMIA

ВЫСШЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

Ю. Г. ПУЗАЧЕНКО

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЭКОЛОГИЧЕСКИХ И ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЯХ

Допущено
Учебно-методическим объединением по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений, обучающихся по географическим и экологическим специальностям

Москва
ACADEMA
2004

УДК 91
ББК 26.8
П882

Рецензенты:
вице-президент Международного географического союза,
чл.-кор. РАН *Н. Ф. Глазовский*;
д-р физ.-мат. наук, проф. *С. М. Семенов* (Институт глобального климата
и экологии Росгидромета и РАН)

Пузаченко Ю. Г.

П882 Математические методы в экологических и географических исследованиях: Учеб. пособие для студ. вузов / Юрий Георгиевич Пузаченко. — М.: Издательский центр «Академия», 2004. — 416 с.

ISBN 5-7695-1348-9

Пособие знакомит с базовыми приемами анализа экологических и географических данных, собираемых в полевых условиях.

Применение пакета статистических программ Statistica, SPSS, Systat, NCSS позволяет сократить формальные алгебраические выкладки и сконцентрировать внимание на семантических основаниях использования конкретного метода. Особое внимание уделяется анализу нестационарных, неравновесных систем и систем с выраженными нелинейными отношениями между переменными.

Для студентов высших учебных заведений, обучающихся по географическим и экологическим специальностям.

УДК 91
ББК 26.8

ISBN 5-7695-1348-9

ВВЕДЕНИЕ

Много лет назад, будучи студентом-зоологом, я обратил внимание на то, что в результате полевых исследований собирается огромный объем данных, но в конце концов изучаемое нами явление описывается лишь таблицами средних значений. Такая ситуация показалась мне крайне несправедливой как по отношению к моему труду, так и по отношению к диким животным, которых зоолог вынужден уничтожать для своих исследований.

С другой стороны, мой учитель, упорядочивая множество данных учета численности мышевидных грызунов или птиц, применял четкую логическую процедуру. На библиотечных карточках он в одной и той же последовательности записывал виды и их численность. Виды он раскрашивал в зависимости от численности в три цвета. Каждая карточка содержала информацию об одном учете или местообитании. Затем на большом столе раскладывал, как он говорил, «пасьянс», в котором упорядочивал карточки по сходству цветовой разметки. Проведя такую классификацию, он сравнивал полученные классы с аналогичным образом упорядоченными характеристиками среды и в наглядной форме устанавливал связи между численностью отдельных видов и средой.

Сама процедура пасьянса была очень увлекательна. Она требовала нескольких последовательных перекладываний (итераций). Конечный результат не всегда был строго однозначным. В нем часто сохранялась некоторая неопределенность: какую-то карточку в равной степени можно было отнести к двум разным хорошо отличимым подмножествам. Однако обычно «природа» демонстрировала стремление к некоторому порядку. При всем этом процедура, подчиняясь с полной очевидностью определенному правилу, занимала много времени. Вместе с тем было очевидно, что чисто логически задача взаимоупорядочивания данных разрешима только при относительно небольшом объеме наблюдений.

В то время «узкой специализации», будучи убежденным биологом, я считал себя полностью независимым от математики и не очень утруждался ее изучением. Однако поиск путей устранения очевидных для меня противоречий между «объемом данных и содержанием результатов», практически полной алгоритмизацией процедуры упорядочивания данных и необходимостью ручного перебора, сложность которого экспоненциально увеличивалась с уве-

личением объема выборки, заставили меня искать их разрешения в области математических методов.

Обратившись к учебнику «Биометрия» Н.А. Плохинского, я с первой страницы столкнулся с проблемами. Попытка понять смысл дисперсии и среднеквадратического отклонения вызвала большие проблемы, а распределение и действующие в нем законы полностью убедили в том, что либо я что-то не понимаю, либо в учебнике упущены какие-то содержательные вещи, без которых невозможно уяснить смысл всех последующих действий. Как показала практика, скорее всего, справедливым оказалось последнее. В то время учебники прикладной статистики в основном сводились к описанию правил действий, и места для рассмотрения их смысла не оставалось. Учебники были написаны по принципу «делай как я, и будет хорошо».

Все-таки после месяца усилий, приложенных к первым десяти страницам, я понял, о чем идет речь, и стал довольно бойко разбирать более сложные операции. Однако результаты классической статистики меня не удовлетворили. Почти вся статистика того времени сводилась к проверке альтернативных гипотез, сравнению средних и расчету коэффициента корреляции как меры отношений между двумя переменными. При этом детали этих отношений в расчет не принимались.

Исследование отношений между свойствами среды и свойствами какого-либо вида животных или растений — важнейшая цель полевого эколога. Классические методы статистики в большинстве своем не очень пригодны для решения такого рода задач. В то время они в большей степени были ориентированы на проверку отклика на воздействие в строго организованном сельскохозяйственном эксперименте, чем на полевые исследования в реальной природе.

Данный факт подвиг меня на расширенный поиск методов исследования отношений. Шестидесятые годы XX века открывали для этого широкие возможности. Это был период бурного развития кибернетики и внедрения строгих математических методов во все сферы человеческой деятельности. Математическая логика и статистическая теория информации дали неплохие основания для полуколичественного исследования отношений и позволили существенно усилить содержательные результаты полевых исследований.

В принципе в то время были разработаны и важнейшие нетрадиционные методы статистического анализа, однако их использование ограничивалось возможностями вычислительной техники. Только с появлением персональных компьютеров весь интеллектуальный потенциал этих методов стал полностью доступен для любого исследователя.

Однако на практике анализ данных в мировой экологической науке, оперирующей с полевыми данными, обычно ограничивается стандартными дисперсионным и факторным анализами. При

этом последний обычно используется там, где он теоретически неприменим. По моему мнению, причина — в сохранившейся слабой адаптации существующих учебных пособий к мышлению натуралиста. Как и прежде, в докомпьютерную эпоху, в руководствах по статистическим методам анализа больше внимания уделяется не смыслу действий, а процедуре, написанию формул как алгоритмов некоторых действий без должного объяснения их содержания.

Опыт убедил меня в том, что натуралист обладает специфическим предметным мышлением, при котором смысл имеет только то, действие чего абсолютно понятно. Для натуралиста ответ на вопрос «как предметы и явления соотносятся друг с другом» является полезным, но промежуточным. Его в первую очередь интересует ответ на вопрос «почему они находятся в таких отношениях и к чему это приводит». Вероятность и случайность, несмотря на то, что с ними необходимо считаться, являются его врагами, и все его действия направлены на выявление детерминированных отношений. Именно поиск смысла в наблюдаемом заставляет его углубляться в предмет, хотя часто приводит к недоказуемым результатам. Вообще талант исследователя-натуралиста определяется способностью его мышления распутывать клубок отношений и взаимодействий, выбирать из множества возможностей наиболее достоверные. Он от рождения, не подозревая об этом, владеет законами различных логик и применяет их, соизмеряясь с реальным объектом, ощущает случайность и риск ошибочного вывода, ищет максимально простое объяснение отношений, обладающее при этом наибольшей общностью, и «эстетическая сторона» объяснения является для исследователя существенным критерием истинности.

Вместе с тем, интуитивный натурализм сплошь и рядом граничит с волюнтаризмом и часто опирается на аргументы типа «нам кажется, мы считаем, мы думаем». При этом, как «считают» и каков ход мысли, обычно не разъясняется. Тем не менее практика любого добросовестного натуралиста показывает, что ему действительно часто «кажется», и желаемое часто искажает действительное. Таким образом, чем сложнее объект, чем больше рассматриваемых взаимосвязанных переменных, тем труднее распутать клубок множества отношений.

В этом случае на помощь приходят количественные методы анализа данных, владение которыми необходимо для современного исследователя, решающего существенно более сложные задачи, чем его учителя. Однако использование этих методов без понимания их смысла и возможностей не исключает ошибочности выводов. Дело в том, что каждый метод опирается на некоторую модель отношений, и на его основе можно «раскрыть реальность» в том и только в том случае, если его логические основания адекватны ее

свойствам. Иначе, каждый метод анализа можно рассматривать как отмычку, пригодную для определенного типа замков. Сконструировать универсальную отмычку теоретически невозможно.

В предлагаемом пособии сделана попытка обобщить накопленный опыт изучения и применения количественных методов анализа данных.

При таком подходе необходимо начинать изучение предмета с изложения методологических основ организации самих исследований и важнейших элементов системологии, определяющих всю последовательность дальнейших действий.

Следует обратить особое внимание на возможность использования современным исследователем пакетов статистических программ, освобождающих его от непосредственных трудоемких вычислений. Наиболее распространенными и универсальными являются следующие программные продукты: Statistica, Systat, NCSS, SPSS. В основном все пакеты совпадают, но существенно различаются в деталях. В разных пакетах одни и те же методы могут быть представлены в несколько отличных версиях и с разной полнотой. Пакеты могут отличаться по максимальному объему данных, которые можно включить в анализ, по способу построения изображений и т. п. Опыт показывает, что примерно 90 % типичных задач можно успешно решить, используя пакет статистических программ Statistica. Однако в некоторых случаях приходится обращаться и к другим программным продуктам. Кроме того, в Интернете можно найти множество частных статистических программ, адаптированных для решения относительно узких экологических задач. Обычно эти задачи можно решить, используя пакет общего назначения. Однако такое решение более трудоемко.

Практика показывает, что если исследователь хорошо разобрался со способами управления расчетами в одном пакете программ, то он легко справится и с другими. Кроме того, во всех пакетах существует файл «Помощь» (Help), что позволяет пользователю разобраться в содержании проблемы и путях ее решения. Содержание «Помощи» столь полно, что ее можно рассматривать как оперативный учебник по статистике, адаптированный к конкретному программному продукту.

Это позволяет при изучении конкретных методов анализа и правил действий опираться на один пакет программ, например на Statistica. Можно не сомневаться, что по мере усложнения задач исследования читатель самостоятельно разберется и в других пакетах.

Наконец, необходимо обратить внимание на тот факт, что в современном информационном мире передать через курс лекций или через учебник все знания о предмете невозможно. Важно, чтобы был освоен некоторый базис знаний и умение самообучаться, постоянно перерабатывая новую информацию по соот-

ветствующей предметной области. При этом не следует рассчитывать на то, что сразу же при первом прочтении все становится понятным. В некоторых случаях требуется многократный разбор содержания какого-либо текста или метода, иногда он становится понятным при обращении к другим источникам. Процесс расширения знаний непрерывен, и прекращение самообучения с полным основанием можно определить как моральную гибель любого исследователя. Каждый год вы должны узнавать и реализовывать что-то такое, чего не знали и не делали в прошлом. Постоянный поиск нового есть гарант вашего исследовательского здоровья.

Глава 1
ОБЩИЕ ПРЕДСТАВЛЕНИЯ О СИСТЕМАХ И СИСТЕМОЛОГИИ

1.1. Общая схема научного познания мира

У науки, как и у человека, связавшего с ней свою жизнь, существует единственная цель — получение новых знаний через взаимодействие с природой и применение их в технологиях, способных повысить устойчивость всего социума. Как шахтер добывает уголь, так и ученый добывает знания. Шахтер извлекает источник энергии, ученый — новую информацию. В конечном итоге труд ученого повышает эффективность работы шахтера. Но в обоих случаях источником и энергии, и знания является природа в широком смысле этого слова.

Наука как коллективное творчество ученых, более точно — исследователей, как результат их взаимодействия с природой определяет информационное развитие социума, открывает новые ресурсы и возможности. Собственно и сам социум, который можно отнести к природе, также является предметом исследования. Однако информационные отношения между человеком и окружающим миром не являются прерогативой только одной науки. Религия и искусство на иной основе выполняют фактически те же функции, отображая окружающий мир и самого человека в его сознание. При этом разрабатываются модели его организации, которые на определенном интервале времени и пространства определяют восприятие мира социумом и его социально-экономическое поведение.

Методологические различия между этими тремя информационными формами (наука, религия, искусство) взаимодействия с окружающим миром можно свести к следующему:

1) для науки и религии воспроизводимость явлений во времени и в пространстве принимается как обязательная. Для искусства воспроизводимость недопустима — она переводит искусство в ремесло;

2) наука и искусство ориентированы на постоянное открытие нового, ранее неизвестного и не используют категорию «веры». Религия воспринимает мировую систему как в той или иной форме замкнутую с точки зрения преобразования информацию;

3) религия не требует доказательства истинности основных своих положений и оперирует категорией веры. Наука и искусство на основе новых знаний постоянно стремятся верифицировать свои теории и постулаты и через новое знание или метатеории доказывать или отрицать их истинность. Постулаты, аксиомы и правила их преобразования, определяющие теорию для науки и искусства, по условию не являются незыблемыми. Если теория представляется истинной, то наука и искусство всегда ищут условия, в которых она оказывается ложной;

4) в отличие от религии и искусства наука ищет ответ на вопрос «почему это происходит» и стремится воспроизвести цепочку причинно-следственных отношений, понять механизм, порождающий конкретное явление.

Конечно, такое жесткое разделение науки, искусства и религии несколько условно. Однако существующие между ними различия делают их совершенно независимыми друг от друга и непротиворечиво совместимыми в одном человеке. В какой-то мере в результате этой непротиворечивости можно и в науке обнаружить яркие проявления религиозного мышления и искусства, а в религии — научные подходы.

Но в данном случае речь идет о научной форме общения с окружающим миром и в первую очередь о методологии этого общения.

Для того чтобы достаточно точно определить место нашего предмета «анализа данных» в научном поиске, полезно рассмотреть упрощенную схему познания человеком окружающего мира (рис. 1.1).

Будем постулировать мир открытым, чем-то очень большим, данным человеку в восприятии через его сенсоры — органы чувств.

Информация, или «что-то» наблюдаемое, существует лишь в том случае, если она отлична от фона, т.е. присутствует хотя бы в двух состояниях «наличие» — «отсутствие». Если «наличие» наблюдаемо с помощью сенсоров во времени и в пространстве единожды или всего лишь несколько раз, то в лучшем случае сознание,

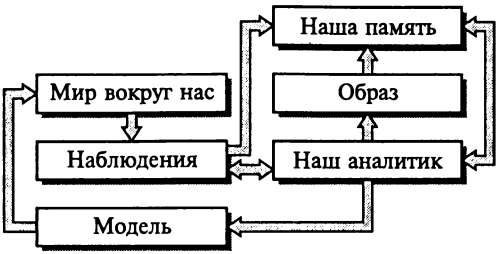


Рис. 1.1. Схема познания мира

взаимодействующее с сенсорами, фиксирует и запоминает его не более, чем как возможный факт, или даже вообще забывает его. Если же нечто наблюдаемо во времени и в пространстве с некоторой регулярностью или правильностью, то тогда оно воспринимается как «явление» или «свойство» природы. Таким образом, восприятие должно обязательно взаимодействовать с органом, обеспечивающим запоминание наблюдаемого, т.е. с памятью.

Совершенно очевидно, что для констатации некоторой пространственно-временно́й правильности необходим анализ результатов наблюдений, т.е. анализ данных или анализ правил возбуждения сенсора, фиксирующего факт «наличия». Этот простейший анализатор, обладающий некоторой оперативной памятью, становится способным на основе запоминания правила изменения состояния явления, пусть и в небольшой степени, предсказывать будущее.

Сенсоры могут различать одновременно несколько различных явлений. Анализатор изначально ничего о них не знает и потому априори считает их независимыми. Однако, осуществляя длительные измерения и все время общаясь со своей «стратегической» памятью, он устанавливает, что если наблюдается явление *A*, то через некоторый интервал времени и пространства наблюдается, пусть не всегда, но часто, некоторое явление *B*. Очевидно, что на этой основе прогностические возможности анализатора резко расширяются, и он может предсказывать состояние явления, опережая его появление во времени и в пространстве.

Итак, наш анализатор смог установить факт определенного правила отношений между явлениями. Очевидно, что эта способность резко увеличивает возможности приспособления организма к изменяющимся условиям среды. В частности можно сказать, что организм строит свои отношения с окружающим миром на основе «примет». При этом он ничего не знает о том, почему существуют такие отношения. Для прогноза достаточно знания факта их существования, т.е. ответа на вопрос «как состояния явления (или вещи) соотносятся во времени и в пространстве друг с другом». Ответ на этот вопрос есть результат анализа многомерных данных (наблюдений за разными явлениями). Но отношения между состояниями двух и большего числа явлений совершенно необязательно строго определенны. Факт этой неопределенности является мощным раздражителем, стимулирующим необходимость повышения качества (разрешающей способности) наблюдений, увеличения объема памяти, числа наблюдаемых явлений, а также числа свойств или переменных, описывающих одно и то же явление. При этом возможности памяти ограничены, и необходимо найти какой-либо способ более эффективного кодирования поступающей информации.

Первым таким формальным способом кодирования становится *классификация*, т.е. процедура выделения типов явлений и их со-

стояний. Очевидно, что реализация этой процедуры требует появления новых возможностей анализатора. Анализатор должен научиться оценивать «сходство-различие» между состояниями явлений обычно по нескольким описывающим их переменным и объединять их в одной «ячейке» памяти. Если некоторые явления и их состояния начинают представлять большое практическое значение для выживания, то множество их несколько различающихся состояний заменяется одним «образом». Образов много меньше, чем исходных состояний, и для их запоминания и установления отношений между ними требуется много меньше памяти. Каждому образу можно поставить в соответствие некоторое кодовое «слово». Появление слов открывает качественно новые возможности для организмов — организацию коллективного поведения на основе взаимообмена информацией.

Однако этого оказывается недостаточно, и анализатор должен научиться принимать решения о том, что будет в условиях неопределенности и что при этом необходимо делать. Факт отношения между любыми явлениями, при которых нельзя однозначно предсказывать последовательность их смены во времени и в пространстве, определяет становление образов «случайности, возможности, риска». Отношения этого типа можно определить как «мета», потому что они не связаны с каким-то конкретным явлением, а являются всеобщими. На этом этапе эволюции наш анализатор фактически строит первую модель мироздания — *модель случайного поведения.* Чтобы принимать решение нужно как-то оценивать неопределенность последствий и риск неудачи прогноза или ошибки. Наш анализатор еще ничего не знает о теории вероятностей, но он из опыта может ввести образы крайних состояний отношений — абсолютную зависимость (однозначность) отношений и абсолютную независимость (неоднозначность) отношений между явлениями. Сравнивая множество наблюдений с этими эталонами, он оценивает дистанции от них реальных отношений и может выбрать такие, в которых правило отношения существует почти наверняка и риск ошибочного прогноза невелик. С другой стороны, анализатор должен формулировать первые модели принятия решений в условиях неопределенности, и простейшее из них есть «*правило голосования*». Правило голосования требует запоминания большого количества правил отношений между состояниями многих явлений.

Пусть наблюдатель констатирует факт того, что произошли некоторые конкретные состояния нескольких явлений, находящихся в определенном соотношении с явлением Y, состояние которого требуется прогнозировать. Если большее число состояний этих явлений указывает на то, что произойдет состояние Y_i, а не какое-то другое, то принимается решение в его пользу. Фактически на этой основе формируется то, что в настоящее время принято называть *нейронными сетями.* Затем уже возникает логика как некото-

рая совокупность правил мышления, и, используя их, организм может достичь больших высот в прогнозе изменения среды и управления своим поведением. Все эти возможности появляются в результате ответа только на один вопрос «как явления и вещи соотносятся друг с другом». Вся совокупность методов анализа данных направлена на поиск ответа на два вопроса: 1) как явления и вещи соотносятся друг с другом; 2) что можно сказать о состоянии одних явлений, зная состояния других.

Повышение качества ответа на эти вопросы определяет неизбежность возникновения абстракций как образов или понятий, уже не относящихся к конкретным явлениям природы, а отображающих существующие свойства отношений между ними. Чтобы оперировать с этими абстракциями, необходимы некоторые идеальные состояния отношений, которые в чистом виде в природе уже практически не встречаются, но их можно получить как образы через упрощение реальности. Следовательно, на этом этапе возникают первые модели мировых отношений, первые зачатки того, что в настоящее время называют *теорией множеств*, *теорией информации*, *теорией принятия решений*, *статистикой* и, наконец, *логикой*.

Можно полагать, что в той или иной степени всеми этими атрибутами принятия решений обладают любые организмы. Однако, скорее всего в отличие от человека они воспринимают наблюдаемые отношения между явлениями как данные и не ставят перед собой вопроса «почему они существуют».

Поиск ответа на вопрос «почему» принципиально изменяет требования ко всей системе восприятия мира, анализа данных и принятия решений. Чтобы ответить на этот вопрос, необходимо целенаправленно изменить объект наблюдения, найти свойства, которые могли бы установить цепочку причинно-следственных отношений.

Рассмотрим на простом примере логику такого поиска.

Существует народная примета: если ласточки летают вечером высоко в небе, то завтра будет хорошая погода. Поиск причины строится по следующей логической схеме.

1. Вводится постулат: ласточки не просто летают, а ловят в воздухе мелких насекомых. Проверить данный постулат довольно просто. Для этого достаточно установить очевидный факт кормления птенцов при возвращении ласточек к гнезду.

2. Из этого следует вывод: если завтра будет хорошая погода, то мелкие насекомые — корм ласточек — поднимаются на очень большую высоту. Таким образом, есть объяснение причины очень большой высоты полета ласточек, но нет ответа на вопрос «почему насекомые, «предвидя» хорошую погоду, летают высоко над землей».

3. Априори существует две альтернативы: насекомые активно поднимаются на большую высоту или насекомых поднимает на высоту некоторая внешняя сила.

4. Первое допущение, имея в виду размеры насекомых, представляется маловероятным. Более того, дополнительные наблюдения показывают, что в то же время, когда ласточки летают высоко, дым из трубы или от костра поднимается почти вертикально вверх. Таким образом, можно почти наверняка утверждать, что насекомые поднимаются вверх вертикальными потоками воздуха.

5. Но требуется ответ на следующий вопрос «почему в этих условиях воздух поднимается вверх». Наблюдения показывают, что легкие предметы всплывают в воде, зола от костра, легкие перья в этих условиях могут подниматься довольно высоко над землей. Отсюда следует объяснение: в условиях, когда на следующий день будет хорошая погода, воздух наверху существенно более плотный, чем внизу, поэтому более «легкий» воздух устремляется вверх, поднимая за собой насекомых, а за насекомыми — и ласточек.

Но остается еще один вопрос «почему воздух наверху в этих условиях холоднее, чем внизу». Для ответа на этот вопрос у нашего гипотетического натуралиста уже нет никаких априорных оснований. Наблюдая только за ласточками, он не сможет дать ответ на этот вопрос. Необходимо существенно расширить область его исследований, установить и объяснить причинную связь существенно иных отношений. Однако, установив причинно-следственные отношения в рамках доступных для него наблюдений и экспериментов, он смог дать в принципе правильное объяснение.

Таким образом, поиск причинно-следственных отношений требует переопределения объекта наблюдения и с одной стороны более детального описания поведения ласточек, а с другой — привлечения внешне независимых наблюдений, обобщение которых приводит к формулировке некоторого более общего отношения.

Современный исследователь из школьных и университетских курсов физики и химии априори владеет базовыми знаниями метаотношений. Эти науки описывают метаотношения, инвариантные для широкого класса явлений, и формируют современный уровень понимания любым человеком окружающего мира. Для объяснения частных причинно-следственных отношений исследователь обычно явно или неявно привлекает эти наиболее общие модели миропонимания. Однако знания, накопленные за всю историю развития человечества, не могут исчерпывать все разнообразие мира и далеко не всегда достаточны для объяснения причинно-следственных связей, например, в сложных биологических и социальных явлениях. Но при всем этом, модели отношений, применимые для объяснения явлений самой разной природы, сохраняют свою непреходящую ценность. Если же их не хватает для объяснения наблюдаемых явлений, то исследователь ищет новые, более общие модели, снимающие частные противоречия. Эти новые модели, когда они обладают очень высокой общностью, получают статус «*открытия*».

Часто для поиска причинно-следственных отношений исследователь вынужден изобретать новые сенсорные системы, позволяющие увидеть ранее ненаблюдаемое, т.е. иначе говоря, позволяющие открыть новые явления. При этом обычно сенсор, сконструированный для решения какой-либо одной частной задачи, открывает широкий класс новых явлений, которые в свою очередь требуют объяснений.

Окружающий нас мир является открытой системой, из которой человек постоянно извлекает знания. Этот процесс строится по схеме: существующие знания → проверка адекватности их реальности → проблемы → новые средства и методы наблюдения → новые явления → новые отношения → новые модели и теории → расширение знаний → повторение цикла в более широкой области открытого мира.

На всех этапах этого цикла анализ данных — средство выявления сложных отношений и основа доказательства реальности наблюдаемого.

1.2. Основные системные понятия

Реальность, окружающая человека, огромна. Исследовать ее в целом и теоретически, и практически невозможно. Поэтому в каждом случае из окружающей человека реальности выделяется некоторая часть, которая получила название «*система*». Сразу же подчеркнем, что «система» есть некоторое, по определенным правилам организованное отражение реальности, но не сама реальность. В наиболее общем случае система — это отношение, определенное на множестве элементов.

Первое, важнейшее и имеющее огромное методологическое значение имеет понятие элемента, или материальной точки.

Под **элементом** понимается индивидуальный объект, который во всех последующих преобразованиях рассматривается как неизменный, несжимаемый, неделимый.

Это определение имеет прямое отношение к полевым исследованиям. Например, если геоботаник описывает растительные сообщества, которые он рассматривает как элементы какой-то региональной системы, то он должен использовать один и тот же способ выделения элемента, т.е. пробной площади, на которой он осуществляет все наблюдения. Эта пробная площадь во всех однотипных исследованиях должна быть одинаковой. Если площадь будет меняться произвольно и зависеть от настроения исследователя или от его нечетко сформулированных условий, то все собранные материалы будут несопоставимы и результаты их анализа и обобщения некорректными. Именно поэтому столь большое внимание обычно уделяется методам исследования. Единый способ представления элементов в однотипных исследованиях различными авто-

рами является важнейшим условием сравнимости и воспроизводимости результатов.

Обоснования элемента — не простая задача. Он должен быть соизмерим с уровнем иерархической организации объекта исследования, который принимается в качестве основного. При этом часто оказывается невозможным совместить различные аспекты одного и того же объекта в одном элементе. Так, например, очевидно, что если исследуется древесный ярус, то площадь элемента (конкретного описания) должна быть по крайней мере соизмерима со средней высотой деревьев, т.е. иметь линейные размеры не менее 25 м. Только в этом случае можно рассчитывать на отображение свойств некоторого участка леса, а не отдельных деревьев. Но такая площадь будет очень большой для описания кустарникового, травяного и, тем более, мохового ярусов. На этой площади будет представлено несколько различных сообществ трав, мхов, кустарников. Площадь описания травяного яруса, если есть необходимость исследовать правила его пространственного варьирования, должна быть также соизмерима с его высотой и соответственно может варьировать от 1 до 0,25 м^2. В зависимости от задач исследования может быть выбрано три схемы выделения элемента, включающего все ярусы сообщества: 1) описания всех ярусов подчиняются площади древесного яруса; 2) в пределах общей площади описания закладывается несколько площадок описания подчиненных ярусов; 3) рассматривается некоторый мыслимый конус пропускания света как основного фактора организации сообщества, и описания кустарникового, травяного и мохового яруса осуществляются на площадках, соответствующих их размерам, но закладываются в центре площадки описания верхнего древесного яруса. В результате получается нечто напоминающее сувенирную матрешку. В этой же центральной точке, в частности, можно сделать и описание почвы. В любом случае в качестве элемента принимается каждое описание, но в первом случае оно ориентировано в основном на отображение свойств древесного яруса и его местообитания, во втором — предусматривается возможность представления каждого яруса как отдельного элемента, а в третьем — описание ориентировано на некоторое соизмеримое, масштабированное отображение в одном элементе разных ярусов.

Если, например, исследуется распределение на некоторой площади деревьев по диаметрам, то элементом системы будут уже отдельные деревья. При описании почвенной ямы элементами описания становятся образцы определенного объема, взятые или с точно определенной глубины, или из генетических горизонтов.

При учете животных в зависимости от поставленной задачи элементом может быть интервал определенной длины, конкретная точка наблюдения (например, при учете птиц на круговых площадках) или конкретная особь животного конкретного вида. Во

всех случаях выбор элемента определяется целями исследования и априорными гипотезами о поведении объекта.

При этом каждый элемент отображается на множестве присущих ему **свойств** или признаков. Под свойством подразумевается нечто присущее объекту исследования, наблюдаемое и измеримое. Так, для сообщества растений свойствами являются: доля участия каждого вида, общее проективное покрытие, или сомкнутость, высота и т. п.; для почв: цвет, механический состав, текстура и структура, доля включений, новообразований и т.д.; для рельефа: крутизна, форма и экспозиция. Выбор свойств в конкретном исследовании определяется его целями и в меньшей степени техническими возможностями. Выбор наблюдаемых и измеряемых свойств обычно вызывает большие дискуссии. Значительное число свойств не поддается какому-либо анализу, а описание их занимает очень много времени. В то же время невключение в измеряемый перечень функционально важных свойств может привести к потере содержательности результатов. В общем, при отборе свойств всегда желательно стремиться свести их число к возможному минимуму, стараясь максимально точно оценить физический смысл их измерения. На критериях этого «смысла» остановимся несколько позже. После того как свойства отобраны, их удобно называть переменными.

Следующее важное понятие — **«состояние»**. Состояние в общем случае это то, что позволяет отличать элементы друг от друга. В физике состоянием называется положение элемента в координатах пространства X, Y, Z и вектора скорости по этим координатам. Это определение в полной мере пригодно и для экологии с той лишь разницей, что координатами становятся переменные, а вектора скорости, к сожалению, обычно трудно определимы. Правда эколог, изучающий поведение животных, часто вполне надежно может измерить и вектора скорости по каждой переменной.

Покажем, что определение «состояния» в физике полностью пригодно как для экологии, так и для географии. Действительно, если мы говорим «ельник зеленомошник», то подразумеваем существование каких-то других состояний леса, которые могут быть описаны в координатах «господствующая порода» и «напочвенный покров». Когда мы говорим «старый ельник зеленомошник», то вводим новую координату — возраст. А когда говорим «приспевающий ельник зеленомошник» или «перестойный ельник», то в неявном виде вводим и вектор скорости.

Следующим важным понятием является **«процесс»**. Процесс — это просто смена состояний в пространстве и во времени. Процесс сам по себе констатирует факт «движения», или динамики, но ни в коем случае не подразумевает объяснения его причины. Конечно, устанавливая правила протекания процесса, получаем некоторые основания для суждения о возможных его причинах. Однако последнее не обязательно.

Полезно различать равновесные и неравновесные процессы. *Равновесными* называются процессы, вектора скоростей которых близки к нулю, так что изменения внутренних свойств элементов успевают многократно измениться и подстроиться к очень медленно меняющимся условиям. Иначе говоря, время релаксации внутреннего состояния элемента много меньше скорости изменения положения его в координатах пространства. Следует отметить, что статистический анализ данных в основном отображает именно равновесные процессы.

Неравновесные — соответственно такие процессы, скорость которых близка к скорости релаксации внутренних свойств элемента. Равновесные процессы предсказуемы, неравновесные процессы по определению непредсказуемы.

Отношение — это то, что ставит в определенное соответствие друг другу элементы или состояния элементов одного множества, или элементы и состояния одного множества к другому, или нескольких множеств друг к другу.

Наиболее типичным отношением множества на само себя является отношение порядка (больше-меньше, выше-ниже и т.п.). Отношение порядка часто называют *структурой*. Действительно, если мы говорим «структура древостоя», то подразумеваем отношение порядка между деревьями или по возрасту, или по высоте, или по доли участия пород в насаждении, или сразу по трем этим переменным. Когда говорят «пространственная структура», то очевидно подразумевают определенный порядок смен состояния объекта, например крутизны и экспозиции склонов в пространстве.

В естественных науках широко используется априорное ведение систем как объектов исследования на основе отношений. Например, *популяция* есть множество особей одного вида, находящихся в генетическом родстве, заселяющих общую территорию и способных достаточно длительное время поддерживать на основе размножения свою численность. Здесь система задана отношением генетического родства, отношением самовоспроизводства во времени и единством территории. *Сообщество* есть множество популяций различных видов, заселяющих общую территорию, находящихся, по крайней мере, в одном из следующих возможных отношений: конкуренции, хищника-жертвы, паразита-хозяина, временной или постоянной коалиции, комменсализма, мутализма или нейтрализма и способных длительное время поддерживать свою структуру. В отличие от сообщества *биоценоз* по определению допускает все перечисленные отношения, кроме нейтрализма.

Наконец, *экосистема* определяется отношением сообщества организмов со средой. При этом имеются в виду в первую очередь отношения типа использования энергии, минерального питания, влаги и т.д. Обратное влияние сообщества на среду для экосистемы обычно не рассматривается как обязательное.

В отличие от экосистемы, в определение *биогеоценоза* вводится отношение однородности территории и влияния как абиотической среды на сообщества, так и сообщества на среду, т. е. рассматриваются прямые и обратные отношения.

Такого рода примеры можно продолжать разными типами отношений из одной и той же реальности, при этом как объекты исследования выделяются различные системы. Очевидно, что выделение с помощью отношений есть важный методологический прием, конкретизирующий объект исследования через определения отношений, которым в первую очередь уделяется внимание.

При рассмотрении элементов как описания состояния, например растительности и почвы в пространстве, также подразумеваются некоторые территориальные системы, или пространственно-временны́е системы, которые выделяются не на основе однотипности отношений, а на основе общности свойств или принадлежности элементов к одному типу. Такими системами, очевидно, являются растительность, почвы, птицы или почвенные беспозвоночные. В физической географии за некоторыми из них в частном случае закрепляется понятие компонент.

Существует два способа выделения системы: 1) на основе однотипности элементов и 2) однотипности отношений. Первый способ предусматривает описание отношений между элементами и их свойствами в ходе исследования, во втором они определены априори, и необходимо измерить их параметры или описать их характер.

Под **параметром** подразумевается нечто относительно неизменное, определяющее правило отношения. В уравнении $y = a + bx$, a и b — параметры. Если они известны, то, зная значения x, можно определить все значения y. Таким образом, само уравнение задает вид отношения (в данном случае — линейное), а параметры определяют правило отображения одного множества в другое. Полезно связывать понятие «параметра» с понятием «характер». Действительно, когда говорят о «характере» человека, подразумевают нечто относительно устойчивое, присущее данному индивидууму на достаточно большом интервале времени. Знание характера конкретного лица позволяет предсказать его поведение и снизить риск возможных конфликтов.

После уточнения важнейших системных понятий рассмотрим некоторые общие правила организации исследования систем (впервые предложены Дж. Клиром, 1990).

Эти правила образуют иерархию эпистемологических уровней системы (рис. 1.2). В общем случае эпистемология, или теория познания, — раздел философии, в котором изучается природа и сфера распространения знаний, их предпосылки и основы, а также критерии истинности знания. Соответственно, предлагаемая схема описывает иерархические уровни познания.

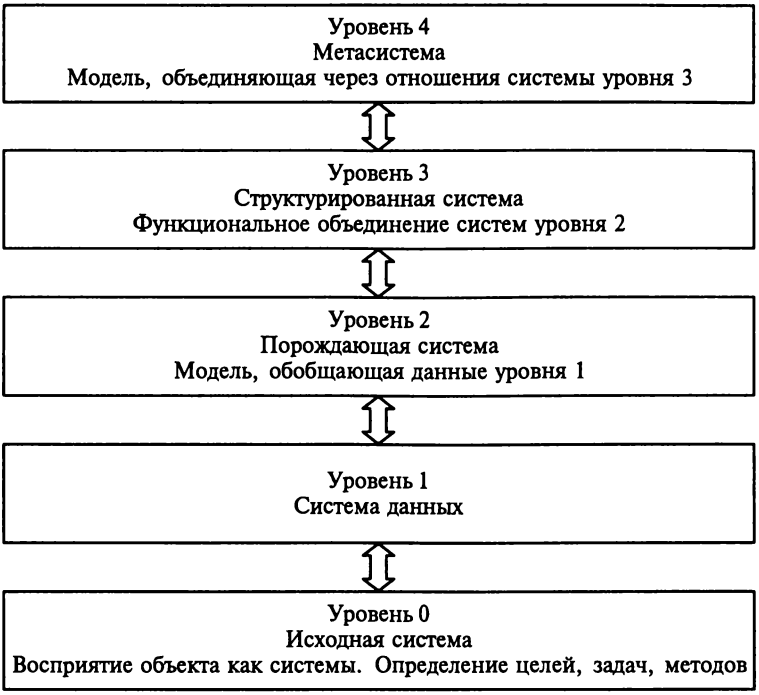


Рис. 1.2. Иерархия эпистемологических систем (Дж. Клир, 1990)

Эта схема позволяет при решении частных задач видеть стратегические масштабы всего исследования и давать прогноз на будущее.

Система нулевого уровня понимается как восприятие природы исследователем. Это восприятие определяется его базовыми знаниями, опытом и даже характером. Вполне понятно, что любое исследование, требующее часто значительных средств и времени, организуется для решения какой-либо проблемы. Проблема возникает тогда, когда исследователь не может исходя из базовых знаний описать, предсказать, понять и объяснить какое-либо явление или когда он подозревает, что существует какое-то явление, которое имеет большое значение в определении поведения наблюдаемого, но само не поддается прямым измерениям. Иными словами проблема — это всегда некоторая неопределенность восприятия и суждения.

В некоторых случаях исследователь нечетко понимает и формулирует проблему, которую он пытается решить в ходе исследования, эта проблема существует в его мышлении в неявном виде. Такое неявное представление затрудняет четкое определение базовых характеристик системы нулевого уровня: определение системы через элементы и переменные, формулировку целей, задач и обоснование методов исследования.

Сформулировать проблему помогает понимание источников неопределенности.

Перечислим некоторые наиболее типичные источники проблем:

• *неадекватность реальности*, принимаемой исследователем концепции, или доктрины, или базовых постулатов. Например, если в качестве абсолютного принимается постулат «в природе все связано», то исследователь будет всюду искать эти связи, и его огорчит, если он их не всегда сможет обнаружить. Достаточно изменить концепцию, приняв, что связь в природе не является всеобщей, как проблема снимается и возникает четкая задача установления факта наличия (отсутствия) связи и измерения ее величины;

• *недостаток или неадекватность понятийного аппарата и невозможность в результате отождествлять факты.* Достаточно типичным источником такой неопределенности является многозначность научного термина. Так, например, понятие «ландшафт» имеет много различных трактовок. Иерархическая схема организации ландшафта как система понятий может оказаться неадекватна реальности. Понятийного аппарата может не хватать и в метанауке. Так, например, для ландшафтоведения в начале его становления была актуальна проблема: объективны ли границы в природе или это лишь удобный способ выделения дискретного или разрывов на фоне реально непрерывной природы. Считалось, что один и тот же реальный объект может быть только или дискретным или непрерывным. Одновременное сосуществование этих двух моделей реальности не допускалось. В 90-х годах XX в. в естественные науки широко вошло понятие фрактального множества, объединяющего и непрерывность, и дискретность. С пониманием механизмов, порождающих такие множества, исчезла и проблема, и исследователи стали ориентироваться на измерение свойств природных тел, описываемых такими множествами;

• *неадекватность применяемой для объекта базовой модели, описывающей отношения между явлениями.* При этом модель может пониматься очень широко. Это может быть и математическая, и логическая, и неявно сформулированная понятийная модель. В свое время большие проблемы создавала асинхронность перемещения в пространстве границы леса и тундры в различных частях континента. Модель, лежащая в основе понимания процесса, подразумевала, что граница леса определяется температурой, а потепление или похолодание — явления глобальные. Проблема снимается, если признать, что граница леса определяется, по крайней мере, влиянием тепла и влаги и их соотношением. С учетом этого допущения при неизменном уровне осадков и потеплении граница леса в континентальных районах будет активно продвигаться на север, в то время как в условиях морского и океанического климатов положение границы может не измениться. Если же потепление

сопровождается увеличением количества осадков, то в этих регионах граница леса может сместиться на юг;

• *неадекватность методов полевых исследований и их пространственно-временно́й организации.* При этом следует иметь в виду то, что многие методы сами по себе вносят возмущения в измеряемые объекты.

Рассмотрим простейший пример — измерение дыхания почвы с помощью «ящиков». На почву устанавливают чашку-поглотитель CO_2 (обычно щелочь) и закрывают металлическим ящиком, который углубляют в почву. Через определенное время поглотитель титруют и рассчитывают количество поглощенного CO_2, ассоциируя его с интенсивностью дыхания почв. Однако поглощение CO_2 из среды заведомо изменяет парциальное давление и неизбежно увеличивает выход углекислого газа из почвы.

Вообще обоснование методов измерения очень тонкое и ответственное дело, и исследователь должен рассмотреть «все за и против» и степень адекватности метода для решения поставленной задачи;

• *сложное, нелинейное поведение систем или положение их в области неравновесных процессов.* Обычно исследователь ориентируется на детерминированное поведение систем и простые линейные отношения. Реальные системы в основном нелинейны, т. е. их параметры (характер) изменяются в зависимости от их собственного состояния. Такие системы в одних и тех же условиях среды могут реализовать существенно различные состояния. Если исследователь не представляет такую возможность, он всегда будет попадать в сложную ситуацию при объяснении наблюдаемых внешне противоречивых фактов. В условиях неравновесности система может генерировать состояния, труднообъяснимые и ее предысторией, и ее средой;

• *применение неадекватных методов анализа данных.* Каждый метод анализа опирается на определенную логико-математическую модель. Если она неадекватна свойствам изучаемой системы, то результаты анализа будут сильно, а иногда принципиально, искажать реальность (собственно именно разрешению этой проблемы и посвящено данное пособие);

• *действие неизвестных явлений или факторов.* Разрешение этой неопределенности возможно на основе поиска новых средств измерения или специальной организации исследования.

Итак, желательно, чтобы исследователь стремился осознать суть проблемы, которую он пытается решить, и, исходя из этого, определял систему нулевого уровня, или исходную систему. В соответствии с целями и задачами он должен найти элемент, перечень измеряемых переменных, точность и способ их измерения. Обычно полезно разделить переменные на входные и выходные. Входные переменные, строго говоря, не являются предметом исследо-

вания и рассматриваются как среда. Иногда полезно выделить внутренние переменные, т.е. такие, которые позволяют объяснить механизмы реакции переменных на выходе на изменения переменных на входе. Однако такое деление не должно быть слишком строгим, так как часто оказывается, что то, что воспринимается как среда, в действительности принадлежит самой системе.

Выбор типа переменных и способа их измерения является важнейшей задачей при исследовании систем, описываемых многими переменными. В общем случае, переменные могут быть непрерывными, дискретными, дискриптивными, четкими и нечеткими (лингвистическими). Они могут измеряться в каких-либо абсолютных величинах или с помощью специальных шкал и эталонов.

Принято считать, что наилучшим вариантом являются *непрерывные* измерения. Например, измерение температуры или прихода солнечной радиации. Однако в действительности измерения никогда не бывают строго непрерывными. Любые приборы имеют ограничения по точности и по пространственно-временно́му разрешению (инерционность). Чем выше точность, тем выше стоимость прибора, чем больше частота наблюдений во времени или разрешение в пространстве, тем больше массив данных, требующих последующего анализа. Точность и пространственное разрешение должны быть соизмерены с задачей. Если, например, ставится задача «исследовать пространственную структуру лесного покрова», то пространственное разрешение аэрофотосъемки должно в два раза превышать средние размеры деревьев, т.е. около 6—10 м на местности. Если разрешение больше, то будет получена информация уже не о сообществах, а о самих деревьях. Более крупный масштаб существенно повышает стоимость съемки и резко увеличивает массив данных. Но если предполагается для распознавания различных типов сообществ использовать строение крон деревьев и их взаимоупорядоченность в пологе, то разрешение съемки должно составлять уже примерно 0,25—0,5 м на местности. Однако исследователь должен четко понимать, с каким массивом данных ему придется работать во втором случае, и адекватно оценить свои технические возможности.

Дискретным можно называть множество, любая точка которого содержит информацию обо всем множестве. Строго дискретных переменных в природе обычно немного. Простейшей дискретной переменной является «наличие-отсутствие» какого-либо признака, например, вида. Обычно приходится иметь дело с дискретным представлением непрерывных переменных.

Дискриптивными (описательными) переменными называются множества, не имеющие внутреннего порядка. Например, переменная «доминирующая порода в лесном пологе» по определению является дискриптивной, так как априори виды деревьев нельзя упорядочить по количественной шкале.

Нечеткими называются переменные, полученные обычно в результате визуальной, качественной (квалиметрической) оценки. Например, молодой, средневозрастный, старый; высокосомкнутый, среднесомкнутый, низкосомкнутый; глина, суглинок, супесь, песок являются нечеткими переменными. Нечеткость их определяется тем, что реальные значения возраста, сомкнутости, гранулометрического состава не всегда соответствуют определенному состоянию. В некоторых случаях то, что оценено как глина, в действительности — суглинок. Нечеткие, или лингвистические, переменные широко используются в полевых исследованиях географов и экологов. Существуют самые различные приемы увеличения их «четкости», однако размытость при этом все-таки остается. Несмотря на нечеткость, применение этого метода измерения не только неизбежно, но и необходимо. Дело в том, что если исследуются отношения между большим числом переменных (например, между видами при описании растительных сообществ), то для выявления таких отношений требуется $(\log^* m + 1)^n$ независимых измерений (m — среднее число различимых измеренных значений, n — число видов). Допустим, что обилие каждого вида измеряется с высокой точностью, например, по шкале в 100 единиц, а таких видов около 20. Очевидно, что данный объем измерений просто не реализуем. Даже если обилие видов будет измерено в трех градациях, то формально для полного решения задачи потребуется около 2 млн независимых измерений. Конечно, существуют способы уменьшить этот объем, однако все они ведут к неизбежно неполному исследованию всех потенциально возможных отношений.

Таким образом, применение качественного оценивания переменных совершенно необходимо.

Этот качественный взгляд на мир нашел отражение в теории размытых (нечетких) множеств и квалиметрии, представляющих самостоятельные разделы математики, широко используемые при анализе данных. Конечно, при этом нужно стремиться, чтобы качественные оценки, выполняемые разными наблюдателями, были бы максимально сопоставимы. Для этого согласования существуют свои приемы. Там, где это возможно, чисто визуальные оценки заменяются применением соответствующих шкал с ограниченным числом градаций. Такой, например, является трехмерная шкала Манселла, применяемая для оценки цвета почвы. Существуют специальные палетки для оценки проективного покрытия трав, сомкнутости древесного яруса и т. п. Методы полевых исследований и измерений изложены в специальных руководствах, однако следует отметить, что изданные более 20—30 лет назад они, к сожалению, несколько устарели.

---

* Здесь и далее, если не указано, будем подразумевать основание логарифма равным 2.

Итак, учтя все возможные проблемы организации исследования, четко сопоставив их с целями и задачами, определяем исходную систему как объект исследования.

Следующий уровень системы — это система данных, организованная в соответствующую базу, — *уровень 1* (см. рис. 1.2).

В современных исследованиях эта система обычно включает и геоинформационную систему (ГИС), которая содержит на первом уровне всю информацию о размещении точек наблюдения в пространстве, космическую многоканальную съемку, топографические карты и т. п. Каждая точка в ГИС (элемент системы *уровня 0*) связана с базой данных, в которой хранятся результаты измерения всех переменных. Обычно такая база данных имеет покомпонентную организацию, и с ней связывается специальная система запросов.

Система «база данных» позволяет контролировать полноту и качество полевых исследований и является основой для системы *уровня 2 — порождающей системы*. Порождающим этот эпистемологический уровень называется потому, что именно здесь выявляются отношения между переменными исходной системы — на этом уровне сосредоточены основные задачи системного анализа. Конечной задачей анализа является представление отношений между переменными через некоторые параметры как инварианты отношений. Собственно параметрами может быть описана и каждая переменная. В частном случае в рамках этого уровня может осуществляться определение параметров для априори принятых моделей отношений. Например, для моделей жидкого стока или роста деревьев, или моделей сукцессионной динамики лесного сообщества.

В результате множество данных, полученных в рамках исходной системы и накопленных в базе данных, преобразуется в очень ограниченное число правил отображения одних переменных через другие с количественными значениями их параметров. На этом уровне в первую очередь получаем ответ на вопрос «как явления соотносятся друг с другом».

Однако, если исходная система организована таким образом, что она включает внутренние переменные, которые могут объяснить механизмы отношений, то исследование может содержать ответ и на вопрос «почему».

Приведем пример такой организации. Допустим, исследуется зависимость фотосинтеза и транспирации дерева (выходные переменные) от прихода солнечной радиации, температуры, влажности воздуха и почвы (входные переменные). В результате измерений и анализа данных можно получить вид и параметры эмпирических зависимостей выходов от входов. Чтобы приблизиться к объяснению этих зависимостей, необходимо организовать измерения сосущей силы корней, скорости и направления движения влаги в стволе, состояния устьиц и других измеримых внутренних переменных. Связывая внут-

ренние переменные с внешними (входы и выходы) можно найти механизм, определяющий общие отношения.

Третий системный уровень (*уровень 3*) подразумевает с одной стороны функциональное объединение нескольких подсистем и (или) сравнение полученных результатов с аналогичными исследованиями (см. рис. 1.2).

Объединение в систему структурных частей подразумевает, что исследование было организовано относительно независимо по отдельным функциональным блокам. Например, в рамках речного бассейна в одном из блоков исследовался твердый, жидкий и растворенный сток, в другом — перемещение влаги, катионов и анионов через почву, в третьем — варьирование в пространстве состава и продуктивности растительности и т.д. На этом системном уровне осуществляется структурно-функциональная интеграция моделей подсистем в единую систему.

Наконец, на *четвертом уровне* (*уровень 4*) на основе обобщения собственных результатов и результатов аналогичных исследований с привлечением существующих метамоделей базовых процессов строится общая модель исследованной реальности, приемлемая для отображения не только исходной системы, но и широкого класса систем того же типа.

Подходы к организации полевых исследований рассматриваются в специальных руководствах. Однако для большей предметности изложения методов анализа данных приведем их краткую систематизацию.

Полевые исследования в экологии и географии можно разделить на следующие основные группы:

• по типу выделяемых систем и целям:

1) исследование процессов — правило смены состояний;

2) исследование отношений — отношения между переменными:

a) во времени;

b) в пространстве;

c) в пространстве и во времени;

• по цели:

1) определение параметров процессов и отношений;

2) поиск механизмов, порождающих конкретный тип процесса или отношений;

3) параметризация моделей динамики;

• по объему системы:

1) монокомпонентные;

2) поликомпонентные;

• по числу включаемых уровней:

i — одноуровневые (вход—выход);

ii — многоуровневые (вход—внутренность—выход).

Продемонстрируем на конкретных примерах принципиальную схему организации исследования.

Тема 1. **«Динамика численности популяций мышевидных грызунов».**

Проблема. Механизмы, ответственные за динамику численности в пространстве и во времени.

Цель. Построить модель, описывающую динамику и создающую основу для прогноза.

Задачи.

1. Определить отношения вида к условиям среды на основе исследования отношений частоты поимки и изменения состояний переменных среды в пространстве.

2. Определить возможные механизмы этих отношений.

3. Найти параметры динамики во времени в различных условиях среды в отношении к изменениям во времени внешних переменных.

4. Определить механизмы, ответственные за динамику.

Система, которую предстоит исследовать, может быть представлена как поликомпонентная, многоуровневая.

Многоуровненность определяет выделение трех типов элементов.

1. Учетная линия ловушек стандартной длины со стандартным регулярным шагом с переменными, характеризующими местообитание, и датой учета.

2. Конкретная ловушка с описанием среды в радиусе, соответствующем половине расстояния между соседними, и датой учета.

3. Конкретный пойманный зверек с собственными характеристиками (вид, пол, возраст и т.п.).

Определение схемы размещения учетных линий в пространстве.

В общем случае по независимым градиентам среды:

1) градиент теплообеспеченности;

2) градиент увлажнения;

3) градиент сукцессионных смен или возрастных стадий развития растительного сообщества;

4) градиент «островной» изоляции (в частном случае).

На каждом полном градиенте должно быть выбрано минимум три-пять точек наблюдения (элементов первого уровня). Желательно, чтобы элементы отражали все возможные комбинации состояний градиента. Соответственно при трех градиентах минимальное общее число точек наблюдений — от 27 до 75. Для каждого сочетания состояний градиента необходимо иметь два-три наблюдения. Тогда оптимальное число измерений — от 75 до 200. Число элементов в исходной системе несколько меньше оптимального, так как градиенты обычно частично зависимы. Традиционно зоологи расставляют ловушки по линии с фиксированной дистанцией между ними (обычно пять шагов). Эта схема во многом определяется традицией учета численности и удобством поиска расставленных ловушек. Однако в общем случае, в том

числе и при учете численности, линейная расстановка ловушек не обязательна. Если используются линии как элемент системы, то каждая линия должна размещаться в гомогенной среде, соответствующей конкретному сочетанию состояний градиентов. Традиционная линия в 25 ловушек для решения поставленной задачи не очень удобна, так как с большой долей вероятности будет включать различные состояния градиентов. Более логично использовать линии в 10 ловушек.

Определение частоты измерений.

Формально частота измерения для воспроизведения процесса динамики должна быть равна половине длительности ожидаемого периода. Для мышевидных грызунов летом собственный период колебаний, определяемый размножением, составляет около 20—30 дней, соответственно в идеале наблюдения надо проводить в течение одного-двух дней раз в 10 дней. Обычно такую частоту реализовать затруднительно. В этом случае можно ставить задачу описания полного летнего цикла размножения и проводить наблюдения в пять сроков с периодичностью примерно раз в месяц. В крайнем случае, допустимо три срока с периодичностью одно наблюдение каждые полтора месяца. Желательно ставить ловушки в точно фиксированное место.

Оценить отношения видов к условиям среды в пространстве можно за один-два сезона наблюдений. Объяснение отношений к условиям среды можно получить на основе анализа частоты попадания животных в конкретную ловушку с известными состояниями среды и отношений переменных, характеризующих конкретные особи, к условиям местообитания в точке поимки.

Уточнение полученных представлений о механизмах потребует дополнительных, может быть экспериментальных, исследований, ориентированных на проверку конкурирующих гипотез.

Для оценки параметров динамики во времени необходимо не менее 10 лет наблюдений.

На основе полученных данных последовательно решаются следующие задачи.

1. Оцениваются параметры самой динамики с учетом градиентов среды.
2. В зависимости от полученных результатов проверяется гипотеза связи динамики с изменением среды во времени с учетом градиентов среды.
3. Проверяется гипотеза автохтонности динамики (составляющая динамики, не зависящая от изменения внешних условий).
4. С учетом полученных результатов рассматривается динамика на уровне конкретных ловушек и с учетом специфики связи с локальными условиями или изменениями этих связей и изменением собственных признаков особей во времени формулируются гипотезы о возможных механизмах динамики.

5. На основе отобранных значимых отношений во времени и в пространстве строится модель динамики.

6. Формулируются условия организации экспериментов и дополнительных наблюдений для верификации гипотез о механизмах.

В принципе это общая схема, которая легко может быть трансформирована для изучения динамики популяций птиц, растений и вообще любых видов организмов.

Тема 2. **«Отношения между свойствами какого-либо компонента в пространстве и во времени».**

Проблема. Отношения между переменными и их механизмы.

Ожидаемый результат. Модель динамики переменных в пространстве и во времени и объяснение ее природы.

В такой постановке эта тема тождественна при исследованиях отношений между видами и видов со средой в растительном сообществе (ординация), при исследовании пространственно-временно́й изменчивости различных свойств почв, жидкого и растворенного стока, при совместном исследовании растительности и почв как компонентов ландшафта.

Задачи.

1. Выделить переменные со статистически значимыми отношениями.

2. Найти виртуальные, независимые факторы, возможно имеющие реальную физическую природу, описывающие варьирование в пространстве и во времени, системно связанные переменные.

3. Выделить переменные с теснотой отношений, позволяющей рассматривать их как подсистему и определить их виртуальные факторы.

4. Определить природу виртуальных факторов.

5. Построить параметризированную модель, описывающую варьирование переменных как функции состояния виртуальных факторов.

Стандартная задача такого рода исследований — замена множества измеренных переменных существенно меньшим числом независимых факторов, которые часто называются виртуальными. Если задача решена корректно, то эти виртуальные факторы обычно имеют достаточно ясную природу.

Исследуемая система в такой постановке почти всегда одноуровневая. Построение исследования по многоуровневой схеме возможно, но технологически довольно сложно.

Элементом системы является описание или измерение в конкретной точке пространства в конкретный момент времени. Чаще при исследованиях этого типа время не рассматривается. Однако в принципе это вполне возможно.

Размещение элементов в пространстве при таком варианте оптимально организовать по регулярной схеме. В этом случае удается отразить пространственные связи между элементами.

Возможны два варианта организации:

1) трансект с регулярным шагом размещения элементов;

2) размещение по регулярной сетке квадратов (грид).

Трансект прокладывают на территории таким образом, чтобы он пересекал максимально возможное разнообразие форм рельефа и почвообразующих пород. Однако при этом он всегда должен быть строго линейным. Иногда могут закладываться параллельные или ортогональные трансекты. Расстояние между элементами или шаг опробования по трансекту определяется целью исследования. Чем меньше шаг, тем более мелкие пространственные структуры могут быть отражены в исследовании.

Размещение элементов по гриду возможно при использовании их спутниковой привязки к географической системе координат. Эта схема очень удобна, когда исследователь располагает дистанционной съемкой достаточно высокого разрешения, например Landsat 7 с разрешением 30 м на местности. Использование космического снимка добавляет новые переменные: значения яркостей в восьми полосах спектра, которые отражают физические свойства поверхности; их связь с измеренными значениями переменных в каждой точке создает условия для интерполяции результатов на пространство и позволяет строить исследования по гриду с размещением точек измерения на относительно больших расстояниях друг от друга (порядка 120 м).

В ходе анализа данных последовательно решаются следующие задачи:

1) изменение отношения между всеми переменными;

2) выделение относительно независимых подсистем;

3) оценка количества существующих независимых виртуальных факторов для каждой подсистемы;

4) расчет значений виртуальных факторов, оценка полноты описания каждой переменной набором виртуальных факторов и определение параметров этих описаний;

5) поиск физической интерпретации факторов;

6) исследование отношения между виртуальными факторами, описывающими различные подсистемы, и поиск общих для них виртуальных факторов, частично описывающих некоторые свойства каждой подсистемы;

7) составление общей модели отношений и оценка ее прогностических возможностей;

8) формирование гипотез о физически не ясных отношениях и определение систем нулевого уровня, которые позволили бы устранить существующую неопределенность.

Рассмотренные схемы организации исследований считаются наиболее типичными. На их основе можно построить самые разнообразные частные схемы организации исследований.

Подведем некоторые общие итоги.

1. Любое исследование направлено на решение какой-то проблемы. Явная формулировка проблемы и адекватная оценка ее источников является важной общей задачей, позволяющей более осмысленно организовать все этапы работы.

2. Выделение системы — методологический прием, направленный на однозначное определение области исследования и последовательности действий.

3. Базовые системные понятия формируют единый язык методологии исследований.

4. Планирование исследования включает обоснование правила размещения элементов системы в пространстве и во времени, выбор измеряемых переменных, способ их измерения, методы последующего анализа данных.

5. Анализ данных — часть системных исследований, выполняемых на эпистемологическом уровне порождающих систем. Возможности анализа данных во многом определяют сам объект исследований, организацию исходной системы сбора данных и организацию их в базы данных. С другой стороны, возможности и результаты анализа данных определяют качество реализации более высоких уровней познания.

Контрольные вопросы

1. Что такое система и каково ее отношение к реальности?
2. Что такое элемент?
3. Что означает понятие «отношение»?
4. Какова разница между переменной и параметром?
5. Что такое «проблема» и каковы ее возможные источники?

Контрольные задания

1. Определите систему для близкого вам объекта исследования.
2. Попытайтесь определить эпистемологические уровни систем для вашего объекта исследования.
3. Определите проблему и опишите ее содержание, а также возможные источники в близкой вам области исследования.
4. Разработайте общий план организации исследования, направленный на решение конкретной проблемы или достижение конкретной цели.

Глава 2
ОСНОВНЫЕ ПОЛОЖЕНИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

2.1. Теоретико-множественные и комбинаторные основания

Теория вероятностей — одна из самых общих моделей мироздания, обобщающая представления о феномене случайности. Интуитивно случайность воспринимается как нечто непредсказуемое, происходящее в какой-то степени вне связи с ожидаемым или желаемым. С другой стороны, человек активно использовал случайность в играх, конструируя специальные генераторы случайности: игральные карты, кости, домино, лотерея и т.п. При этом в играх он всегда стремился уменьшить роль случая и получить желаемый результат вопреки исходной случайности и непредсказуемости. Обсуждение природы случайности является традиционной областью философии. В общем случае случайность есть результат одновременного действия на явления множества частично независимых факторов. Но это только одно из возможных объяснений. Множество частных случайностей состояний явлений во времени и в пространстве есть хаос. Хаос воспринимается как полное отсутствие внутренней взаимозависимости между состояниями или событиями. Чем больше к системе подводится энергии, тем скорее ее поведение будет все более и более хаотическим. Такой хаос со случайным взаимоположением множества частиц называют «*тепловым шумом*». Случайность часто рассматривается как основа развития. Непредсказуемое, случайное соединение каких-то элементов с различными свойствами может дать совершенно новое, ранее не существовавшее качество.

С другой стороны, некоторую условность понятия случайности можно показать на следующем примере. Столкновение двух автомобилей в транспортном потоке, если рассматривать множество автомобилей во множестве потоков, т.е. ансамбль множества элементов, является очевидной случайностью. Однако любое конкретное столкновение рассматривается как строго детерминированное событие, в котором по условию подразумевается существование виновного.

Фактически в этом примере демонстрируются различные способы выделения систем. В одном случае системой является ансамбль

множества элементов, в другом — двух взаимодействующих элементов и среды.

Модель теории вероятностей ориентирована в первую очередь на описание поведения очень больших ансамблей, элементы которых или неразличимы, или, по крайней мере, их свойства остаются неизменными во времени и в пространстве. Это условие очень важно иметь в виду. В теории речь идет об идеальном, т.е. о своеобразной «норме» поведения.

Модель объясняет поведение «идеального». Сравнивая реальное поведение какого-либо объекта с его идеальным представлением, получаем некоторую оценку, которая называется *статистикой*. На основе статистики проверяются конкурирующие гипотезы — можно ли считать, что реальность соответствует «идеалу» или она отлична от него. В анализе данных в большинстве случаев решается именно эта задача. Статистические правила принятия решений сами по себе опираются на специальные идеальные модели теории вероятностей. Это обычно модели очень высокой общности.

Однако, как и во всех случаях, корректное использование моделей для решения практических задач требует хотя бы самого общего представления о лежащих в их основе правилах преобразования и, соответственно, условиях их применимости. Это определяет необходимость знания основ теории вероятностей и приемов построения ее моделей.

Чтобы достаточно корректно сформулировать основы теории вероятностей, необходимо ввести еще одну модель более высокой общности — *модель множества*. Выше слово «множество» рассматривалось как интуитивно понятное. Применение этого важного понятия к экологическим и географическим объектам исследования выявляет существование, по крайней мере, двух различных типов множеств. С одной стороны, можно говорить о множестве элементов (например, особей), а с другой — о множестве свойств или переменных, которыми характеризуются эти элементы. Кроме того, могут рассматриваться некоторые классы элементов, которые также объединяются в множества. О множестве переменных в первом случае можно говорить как о пространстве.

Первый тип множеств определяется системой аксиом Цермело — Френкеля. В ней утверждается, что всякое множество определяется своими элементами.

Второй тип множеств рассматривается в системе Геделя — Бернайса. В этой системе класс определяется как «коллекция» всех множеств, обладающих некоторыми общими свойствами или свойством. Эта система теории множеств использует два типа переменных — классы и множества, при этом элементом класса может быть только множество. Система Геделя — Бернайса как частный случай включает систему Цермело — Френкеля. Соответственно основные операции на множествах остаются общими.

Следующим важным общим понятием является *пространство*. Пространство в широком смысле — среда, в которой существуют множества, например в математике пространство — множество элементов.

Рассматривая множество как пространство, учитывают только те свойства, которые устанавливаются принятыми во внимание или введенными по определению отношениями. Эти отношения между элементами или классами элементов обусловливают геометрию пространства. Чаще всего в качестве отношения проводится анализ расстояний или дистанций между точками (классами), найденных на множестве рассматриваемых свойств. Введение отношения расстояния называется *метризацией*.

В теории вероятностей в основном используется модель множества Цермело—Френкеля. Однако при организации исследований и анализе данных в экологии и географии исследователю приходится иметь дело и с более мощной конструкцией Геделя—Бернайса.

Для понимания модели теории вероятностей необходимо вспомнить основные алгебраические преобразования множеств*.

Равенство множеств: множества A и B совпадают, если каждый элемент множества A принадлежит множеству B, и наоборот. Например, множество [ель, береза, сосна] совпадает с множеством [береза, сосна, ель], но только в том случае, если в качестве элемента не рассматривается их место в строчке записи.

Это означает, что множества характеризуются свойствами:

- рефлексивности: $A = A$;
- симметричности: если $A = B$, то $B = A$;
- транзитивности: если $A = B$ и $B = C$, то $A = C$.

Все это представляется очевидным. Но достаточно допустить, что если с течением времени число элементов во множестве изменяется, то свойство рефлексивности исчезает. В рамках данной логической конструкции это уже будут разные множества. Поэтому, применяя модели к реальному объекту с изменяющимся числом элементов, получаем искаженное отображение реальности.

Объединение множеств: объединением множеств A и B называется множество $A \cup B$, образованное всеми элементами, которые принадлежат хотя бы одному из множеств (рис. 2.1, *а*). Например, объединение двух множеств [ель, береза, сосна] и [ель, береза, осина] есть [ель, береза, сосна, осина].

Пересечение множеств: пересечением множеств A и B называется множество $A \cap B$, образованное элементами, которые принадлежат каждому из множеств. В рассматриваемом примере пересечением будет множество [ель, береза].

---

* Множества принято обозначать прописными буквами латинского алфавита, а их элементы — строчными.

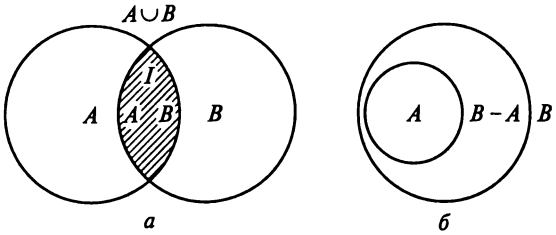


Рис. 2.1. Диаграммы (*а*, *б*) основных отношений двух множеств

Множество, не содержащее ни одного элемента, называется пустым и обозначается символом ∅.

Часть множества: произвольное множество A, каждый элемент которого принадлежит множеству B, называется частью множества B ($A \subseteq B$). Пустое множество $A = \varnothing$ по определению является подмножеством B. Множество $A = B$ является частью множества B. Пустое множество ∅ и само множество B называются несобственными частями множества B.

Если $A \subseteq B$ и $A \neq B$, то говорят, что множество A строго содержится в множестве B, и записывают $A \subset B$.

Дополнение: дополнением части A множества B называется часть $B - A$ множества B, образованная всеми элементами множества B, которые не принадлежат A. В нашем примере дополнением части A до множества B является множество [осина], рис. 2.1, *б*.

Обратим внимание на важное свойство: дополнение объединения двух множеств равно пересечению их дополнений, а дополнение пересечения двух множеств равно объединению их дополнений.

Объединение (логическая сумма) множеств: $A + B = (A \cup B) - (A \cap B)$ — объединение без пересечения.

Объединения и пересечения множеств обладают следующими свойствами:

- коммутативности $A \cup B = B \cup A$, $A \cap B = B \cap A$;
- ассоциативности $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$;
- дистрибутивности $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Сложение множеств коммутативно и ассоциативно:

- $A + B = B + A$;
- $(A + B) + C = A + (B + C)$.

Объединение трех множеств есть

$A \cup B \cup C = (A \cup B) \cup (A \cup C) \cup (B \cup C) - (A \cap B \cap C)$.

Противоположное множество: множество $-A$ противоположно множеству A и равно множеству A, т.е. $A + A = 0$, $-A = A$.

Произведение множеств: $A \times B = A \cap B$.

Теория множеств потребуется при обосновании некоторых типов дистанций. Но в данном случае важно, что именно на основе этих общих правил строится модель теории вероятности.

Интуитивные предпосылки теории вероятностей

Обычно при обосновании теории вводятся так называемые «интуитивные предпосылки», т.е. положения, которые во многом естественным образом вытекают из опыта.

Первоначальное понятие при изучении окружающего мира — понятие «*событие*». Событие определяется тем, происходит или не происходит некоторое явление. Таким образом, событие есть то же, что в системном подходе было определено как состояние. Абстрактное понятие события не подразумевает рассмотрения его физической природы. События будем обозначать буквами A, B, C, ... Каждому событию A можно поставить в соответствие противоположное событие «не A», которое обозначим $\bar{A}$. Одно событие может вызывать появление другого: A влечет за собой B, если при появлении события A обязательно происходит событие B, то будем записывать, как и в теории множеств, $A \subset B$ (очевидно, что A может трактоваться как часть множества B). Если A влечет за собой B, а B влечет A, то говорят, что события A и B *эквивалентны* ($A = B$).

Алгебра событий та же, что и алгебра теории множеств.

Операция пересечения: новое событие «A и B» происходит тогда и только тогда, когда происходят оба события A и B ($A \cap B$ или AB). Если события не могут произойти совместно ($AB = 0$), то говорят, что они взаимоисключающие или несовместные.

Операция объединения: событие «A или B» происходит тогда и только тогда, когда происходит по крайней мере одно из событий A или B или оба вместе.

Если A и B совместны, то записываем $A \cup B$, в противном случае справедлива запись $A + B$.

Все перечисленные операции и их записи те же, что и в теории множеств. Но разница состоит в том, что с одной стороны рассматриваются только события, а с другой — допускается появление одного события как следствия другого и сложного события как результата определенного отношения частных.

Действительно, модель теории вероятностей целиком опирается на теорию множеств, и все рассмотренные выше аксиомы теории множеств справедливы для теории вероятностей.

С точки зрения теории множеств Ω — пространство, в котором лежат множества A, B, C; $\varnothing$ — пустое множество; $\bar{A}$ — дополнение к множеству A; AB — пересечение; $A \cup B$ — объединение.

В научных исследованиях «события» разделяют на «условия» и «исходы». Условия эксперимента — это известные события или

события, которые тем или иным способом осуществляются экспериментатором. Исходы эксперимента — это события, которые могут произойти, когда осуществляются соответствующие условия. Все комбинации исходов по известным условиям, образуемые с помощью операций «не», «и», «или», также являются исходами.

В терминах теории множеств исходы эксперимента образуют поле (или алгебру множества). Условия эксперимента вместе с полем исходов составляют испытание.

Термины «условия», «исходы» и «испытания» хорошо ассоциируются с действиями экспериментатора. Эколог и географ чаще вынуждены использовать результаты эксперимента, осуществленного или осуществляемого самой природой. В таком варианте событиями («условиями») становится то, что в системе определено как входы, а «исходами» — как выходы, а вся система становится ни чем иным как «испытанием». Например, нужно установить связь состава древесного яруса с механическим составом почв: «условиями» здесь являются древесные породы, «исходами» — механический состав почв (глины, суглинки, супеси, пески). Все возможные «исходы» могут быть записаны символами теории множеств. В эксперименте допускается, что каждое испытание может осуществляться сколь угодно много раз. В географии и экологии это в большинстве случаев невозможно, но допущение множества независимых испытаний над одной и той же системой сохраняется. Именно это допущение является необходимым условием «научности» — воспроизводимости результата.

Введем пространство E как множество взаимоисключающих исходов. Такое пространство назовем *пространством элементарных событий*. Допустим, исследователь осуществляет подсчет деревьев на некоторой площадке. Элементарным событием будет конкретное дерево конкретного вида e_{si}. Тогда событие «сосна» произошло, если произошло хотя бы одно из элементарных событий, принадлежащих множеству A (множеству [сосна]).

Элементарное событие не обязательно состоит из одного описания. Можно представить, что деревья разных видов перебираются до тех пор, пока не появится сосна. Тогда элементарным событием будет последовательность стволов видов деревьев до тех пор, пока не появилась сосна.

Продемонстрировать различные варианты элементарных событий можно на примере бросания игральных костей или монеты. Каждая кость имеет шесть граней, пронумерованных от 1 до 6, монета имеет два состояния: 1 — «герб» и 0 — «решка». Элементарным событием может быть, например, двухкратное подбрасывание монеты. В этом варианте таких исходов может быть четыре (11, 01, 10, 00). Событием может быть подбрасывание монеты до тех пор, пока первый раз не выпадет «герб». Таких событий может

быть достаточно много. «Герб» может выпасть сразу при первом бросании (элементарное событие 1), при втором бросании (элементарное событие 01), при третьем бросании (001) и т.д. Общим является то, что все эти события несовместны.

Коль скоро они несовместны, можно ввести меру, такую, что

$$\sum_{e \in E} P(e) = 1,$$

где $P(e)$ — вероятность элементарного события e.

Вероятность события A есть сумма вероятностей всех элементарных событий, принадлежащих A, $P(A) = \sum_{e \in A} P(e)$. Например, вероятность события «сосна».

Если речь идет о модели, например об идеально правильной игральной кости, то вероятности любых событий, например при бросании двух костей, можно определить, не проводя никаких экспериментов. Точно также в лотерее можно рассчитать вероятность комбинации любых цифр на шарах, например 36 при выборке по шесть шаров. Такой расчет возможен в том случае, если шары идеально одинаковые и в лотерейном барабане происходит их полное перемешивание.

Так, в простейшем случае, если монета правильной формы, то при одном подбрасывании монеты возможны события: A — появление «герба»; B — появление «решки»:

$$P(A) = 1/2 \text{ и } P(B) = 1/2.$$

Если одновременно подбрасываются две монеты, то возможен следующий исход:

$$P(A,A) = 1/4,\ P(A,B) = 1/4,\ P(B,A) = 1/4,\ P(B,B) = 1/4.$$

Еще раз подчеркнем, что это условие постулируется не более как для геометрически правильных монет с несмещенным центром тяжести и не требует каких-либо доказательств.

Элементарные события совершенно необязательно должны принимать целочисленные значения. Допустим, что измеряются диаметры деревьев. Тогда каждый измеренный диаметр может рассматриваться как элементарное событие. Если, например, определяется механический состав почвы, то элементарным событием будет массовая доля каждой фракции. Таким образом, элементарное событие может быть как дискретным, так и непрерывным, т.е. принадлежать некоторому континууму.

Из курса теории множеств для вероятностного пространства естественным образом вводим следующие свойства:

1) $P(\varnothing) = 0$ — вероятность невозможного события равна 0;

2) $P(E) = 1$ — вероятность достоверного события равна единице;

3) $P(A+B) = \sum_{e \in A \cup B} P(e) + \sum_{e \in A} P(e) + \sum_{e \in B} P(e) - \sum_{e \in A \cap B} P(e) = P(A) +$

$+ P(B) - P(AB)$.

Вывод суммы вероятностей двух событий точно совпадает с диаграммой рис. 2.1 для пространства с нормой I.

В рассмотренном выше примере это будет вероятность выпадения или двух «гербов» (0,25), или «решки» с «гербом» (0,5):

$$P(A + B) = 0{,}5 + 0{,}5 - 0{,}25.$$

Заметим, что события (A, A) и (B, B) несовместны и их вероятность, как вероятность несовместных событий, будет равна их сумме, т.е. 0,5;

4) $P(\bar{A}) = 1 - P(A)$ — вероятность события «не A», дополняющего и соответственно несовместного с A;

5) если $A \subset B$, то $P(A) \le P(B)$;

6) в общем случае $P(A + B) \le P(A) + P(B)$, равенство существует в том случае, если события несовместны.

Для n событий

$$P\left(\bigcup_{k=1}^{n} A_k\right) \le \sum_{k=1}^{n} P(A_k),$$

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{k<l}^{n} P(A_k A_l) + \sum_{k<l<m}^{n} P(A_k A_l A_m) - \ldots - (-1)^{n-1} \times$$

$\times \sum_{k=1}^{n} P(A_1 \ldots A_n)$ — вероятность объединения n событий есть сумма вероятностей за вычетом всех пар, четверок, шестерок (четного числа вероятностей совместных событий) с суммированием вероятностей совместных событий в нечетных комбинациях.

Аксиома независимости. Два события называются независимыми, если их совместная вероятность равна произведению их вероятностей:

$P(A \cap B) = P(AB) = P(A)P(B)$ и для n независимых событий

$$P\bigcap_{i=1}^{n} A_i = \prod_{k=1}^{n} A_k.$$

Условная вероятность. $P(A/B)$ — вероятность того, что произойдет событие A, если произошло событие B, т.е. $P(A/B) = \dfrac{P(AB)}{P(B)}$.

Например, какова вероятность выпадения в следующем бросании монеты «герба», если перед этим выпала «решка». В данном

случае ответ тривиален. Если монета правильной формы, то 0,5. Если же монета имеет смещенный центр тяжести, то эта вероятность не будет равна 0,5.

Если события A и B независимы, то

$$P(A/B) = \frac{P(A)P(B)}{P(B)} = P(A).$$

Независимость является наиболее общей гипотезой отношения между событиями и отображаемыми ими явлениями, в сравнении с которой устанавливается существование зависимости.

Формула полной вероятности. Пусть A — некоторое событие; B_1, B_2, ..., B_n — попарно несовместные достоверные события, такие, что $A \subset \bigcup_{j=1}^{n} B_j$. Тогда справедлива формула полной вероятности

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

Формула полной вероятности используется, в частности, при расчете информационных мер связи.

Обобщением формулы полной вероятности является *формула Бейеса*, часто используемая при распознавании образов,

$$P(B_j/A) = \frac{P(B_j)P(A/B_j)}{\sum_{k=1}^{n} P(B_k)P(A/B_k)}.$$

Если события независимы, то числитель равен $P(B_j)P(A/B_j)$, а знаменатель равен $P(A)$.

Если существует зависимость между событиями, то формула Бейеса позволяет рассчитать условную вероятность любого из несовместных событий из B_j по их априорным вероятностям $P(B_k)$ и условным вероятностям события A по B_k.

Простейшая модель случайного процесса. Чтобы хотя бы в малой степени почувствовать, как строятся модели случайных процессов, рассмотрим простейшую из всех возможных — модель биномиального распределения.

Модель задается следующими правилами: имеем мешок с черными и белыми шарами. Шары идеально одинаковы и тщательно перемешаны в мешке.

Вероятность события «белый шар — 0» = p.

Вероятность события «черный шар — 1» = $1 - p = q$.

Будем вынимать по три шара. Каждое изъятие шаров будет означать элементарное событие. После оценки вынутой комбинации шаров их возвращают назад и вновь перемешивают.

Эта схема организации модели называется *схемой Бернулли.*

В конкретной модели существует восемь событий (табл. 2.1).

Таким образом, если нас не интересует порядок появления шаров в элементарном событии, то P(белый шар — 0 раз) = q^3, P(белый шар — 1 раз) = $3pq^2$, P(белый шар — 2 раза) = $3p^2q$, P(белый шар — 3 раза) = p^3.

Очевидно, что существует правило, по которому можно рассчитать число комбинаций, в которых на любом месте может быть встречено k шаров одного из двух свойств из N объектов, образующих элементарные события.

Это одно из правил комбинаторики: «число различимых последовательностей из N объектов, содержащих $k \leq N$ неразличимых объектов типа 1 и $N-k$ неразличимых объектов типа 2».

Необходимо напомнить базовые функции комбинаторики, так как они являются логической и математической основой многих дискретных распределений и непараметрических критериев, понимание правил синтеза которых весьма полезно (табл. 2.2).

Очевидно, что одни и те же выражения связываются с несколько различными логическими моделями сортировки объектов, однако все они отвечают случаю, когда или последовательности, или объекты, или сочетания, или размещения не зависят друг от друга. Соответственно, если с ними связывается некоторая мера вероятности, то совместные вероятности совместных

Таблица 2.1

Модель биномиального распределения

E	Шар 1	Шар 2	Шар 3	$P(e)$	Число белых шаров
e_1	0	0	0	qqq	0
e_2	1	0	0	pqq	1
e_3	0	1	0	qpq	1
e_4	0	0	1	qqp	1
e_5	1	1	0	ppq	2
e_6	0	1	1	qpp	2
e_7	1	0	1	pqp	2
e_8	1	1	1	ppp	3

событий будут произведениями частных вероятностей, а несовместных — их суммой.

Возвращаясь к нашей модели, можно записать:

$$P(k) = C_N^k p^k q^{(n-k)} = \frac{N!}{(N-k)!k!} p^k q^{(n-k)}.$$

Пусть $p = 0{,}5$, т.е. белых шаров ровно столько же, сколько черных.

Тогда $P(k = 0) = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} 0{,}5^0 (1 - 0{,}5)^3 = 1 \cdot 0{,}125 = 1/8;$

$$P(k = 1) = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} 0{,}5^1 \cdot 0{,}5^2 = 3 \cdot 0{,}125 = 3/8;$$

$$P(k = 2) = \frac{3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 1} 0{,}5^2 \cdot 0{,}5^1 = 3 \cdot 0{,}125 = 3/8;$$

$$P(k = 3) = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} 0{,}5^3 (1 - 0{,}5)^0 = 1 \cdot 0{,}125 = 1/8.$$

Очевидно, что расчет точно соответствует результатам табл. 2.1 при $p = 0{,}5$.

Так как биномиальные коэффициенты можно рассчитать для любого числа N изымаемых шаров и любого значения параметра p — доли (вероятности шаров белого цвета в мешке), то $p(k)$ — функция распределения числа шаров белого цвета при сколь угодно большом числе испытаний или в иной, игровой терминологии — вероятности успеха. Такое распределение называется *биномиальным*.

Распределение может быть записано в дифференциальной форме $f(k) = p(k)$ или в кумулятивной $F(k) = \sum_k p(k)$.

Для рассматриваемого примера можно поставить вопрос о том, сколько белых шаров k выпадет среди N шаров. Эту величину будем называть *математическим ожиданием* (например, успеха):

$$M(k) = Np.$$

Термин «математическое ожидание» имеет глубокий смысл, и его нужно очень точно понимать.

В реальной выборке, даже если шары идеальны и хорошо перемешаны, число «успехов» совершенно необязательно будет равно Np. Оно может весьма существенно отличаться от этой величины. Это ожидаемое число успехов, но оно может быть как больше, так и меньше. Значение Np рассчитывается на основе знания параметра p и числа выбираемых шаров N.

Таблица 2.2

Основные функции комбинаторики

Определение	Алгебраическая форма	Комментарий
Число различных перестановок из k различимых объектов	$k! = k(k-1)(k-2)\ldots$	Все элементарные события рассматриваются принадлежащими только себе, и их пересечение пусто
а) Число различимых последовательностей из N объектов, содержащих $k \le N$ неразличимых объектов типа 1 и $N-k$ неразличимых объектов типа 2. б) Число различимых разбиений последовательности из N объектов на два класса из $k \le N$ и $N-k$ объектов соответственно	$C_N^k = \binom{N}{k} = \frac{N!}{(N-k)!k!}$ — биномиальный коэффициент	Например, объекты (N) — число шаров в одном элементарном событии, k — число белых шаров, C_N^k — число сочетаний из N по k
а) Число различимых последовательностей из $N = N_1 + N_2 + \ldots + N_r$ объектов, содержащих N_1 неразличимых объектов типа 1, N_2 неразличимых объектов типа 2, N_r неразличимых объектов типа N_r. б) Число различимых разбиений последовательностей $N = N_1 + N_2 + \ldots + N_r$ различных объектов на r классов из N_1, N_2, ..., N_r объектов	$\frac{N!}{N_1!N_2!\ldots N_r!}$ — мультиномиальный коэффициент	Формула для определения этого коэффициента, отражающего всевозможное число комбинаций классов различных объектов, есть одна из распространенных формул оценки разнообразия. В частности, на ее основе выводится статистическая мера энтропии или информации
Число различимых неупорядоченных сочетаний из N объектов различного типа по n в каждом		
Объект каждого типа может встречаться не более одного раза в любом сочетании (сочетания без повторений)	$C_N^k = \binom{N}{n}$	

Объект каждого типа может встречаться 0, 1, 2, ..., n раз в любом сочетании (сочетания с повторением)	$\binom{N+n-1}{n} = \binom{N+n-1}{N-1}$	
Объект каждого типа должен встречаться по крайней мере один раз в каждом сочетании	$\binom{n-1}{N-1}$	
Число различимых выборок (размещений, упорядоченных рядов) объема n из совокупности N различного типа объектов		
Объект каждого типа может встречаться не более одного раза в любой выборке (выборки без возвращения, размещения без повторения)	$\binom{N}{n} n!$	Основа для моделей случайных серий
Объект каждого типа может встречаться 0, 1, 2, ..., n раз в каждой выборке (выборка с возвращением, размещения с повторением)	N^n	Основа модели энтропии в теории информации, где N — длина алфавита, n — длина слова
Число различимых размещений из n неразличимых объектов в N различимых ячейках (положениях)		
Нет ячейки, которая содержит более одного объекта	$\binom{N}{n}$	
Каждая ячейка может содержать 0, 1, 2, ..., n объектов	$\binom{N+n-1}{n} = \binom{N+n-1}{N-1}$	
Каждая ячейка должна содержать хотя бы один объект	$\binom{n-1}{N-1}$	
Число различимых размещений из n различных объектов в N различных ячейках		
Нет ячейки, которая содержит более одного объекта	$\binom{N}{n} n!$	
Каждая ячейка может содержать 0, 1, 2, ..., n объектов	N^n	

Практически в каждом статистическом пакете программ существует возможность построения распределений при заданных параметрах. Для биномиального распределения задаются параметры p и N.

На рис. 2.2, *а* показан вид распределения $f(k)$ для $N = 10$ и пяти значений параметра вероятности p, а также кумулятивное представление распределения $F(k)$ (рис. 2.2, *б*). Распределения демонстрируют вероятность 0, 1, 2, ..., 10 успехов при различной вероятности или доли p белых шаров в выборке.

Математическое ожидание успехов $M(k) = 10p$.

Однако если осуществить выборки конечного объема из мешка с идеальными шарами и с возвращением изъятых шаров, то реальные оценки среднего значения успехов будут отличаться от ожидаемого значения.

Интуитивно понятно, что если увеличивать объем выборки, то реальное среднее значение успехов будет все более близким к математическому ожиданию. Таким образом можно сформулировать *теорему Бернулли*: **если производить серию испытаний, то с вероятностью, близкой к единице, можно ожидать, что число k появлений события A будет очень близко к своему наивероятнейшему значению, отличаясь от него лишь на незначительную долю общего числа n произведенных испытаний**.

Формальная запись теоремы выглядит следующим образом:

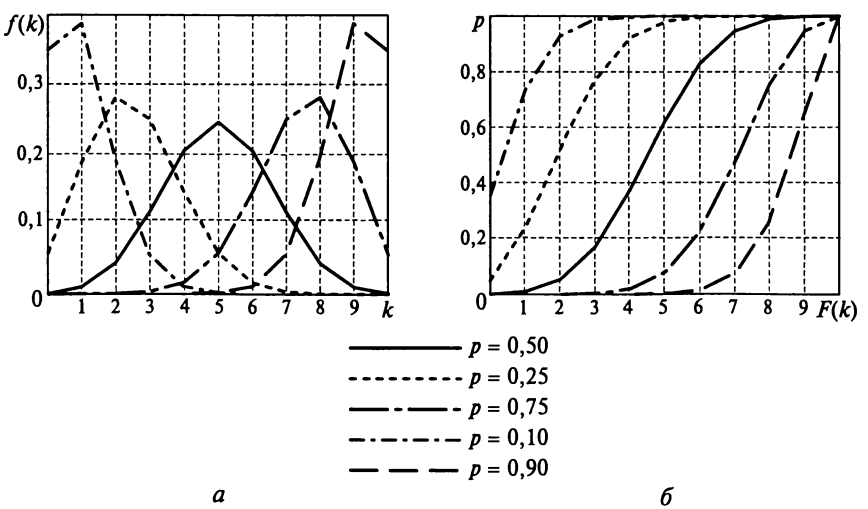


Рис. 2.2. Биномиальное распределение:

а — биномиальное распределение $f(k)$ при $N = 10$; *б* — кумулятивный вид функции биномиального распределения

$$\lim_{n\to\infty}(|k - np| > \varepsilon n) \to 0,$$

т.е. вероятность того, что разность между числом успехов k и математическим ожиданием успехов больше любого сколь угодно малого числа ε, умноженного на число испытаний n, становится при достаточно большом n равной 0.

Эта теорема строго доказывается на основе неравенства Чебышева. Доказательство очень красивое, и можно рекомендовать читателю ознакомиться с ним по любому учебнику теории вероятностей*.

Рассмотрим применение теоремы Бернулли на практике. Для реализации эксперимента воспользуемся генератором случайных чисел при заданных параметрах распределения и различном объеме выборки n и проанализируем, как проявляется сходимость оценки успехов по случайной выборке и математическому ожиданию. Генераторы случайных чисел реализованы в любом пакете прикладных программ.

Такой эксперимент полезен, так как в какой-то мере имитирует реальные исследования в природе. Например, пусть необходимо определить долю ели в каком-то типе насаждений. Важно, что в данном случае подразумевается, что тип насаждений однороден, т.е. в любой точке (элементе) отличие в соотношении пород не более чем случайно и не определяется какими-либо отличиями условий среды или историей развития насаждения. Далее имеется в виду, что полный пересчет деревьев с определением их видов технически невозможен, и исследователь для оценки доли их участия должен пользоваться ограниченной выборкой. Разница с машинным экспериментом состоит в том, что параметры распределения в машинном эксперименте заданы, а в природе они не известны. Подразумевается, что эти параметры необходимо оценить на основе выборки. Задачей статистики в частности и является оценка этих параметров.

Ставя эксперимент на ЭВМ и точно задавая параметры модели, можно допустить, что они нам неизвестны, и рассматривать результат оценок математического ожидания Np, как и в природе, на основе выборки. Компьютер в данном случае можно рассматривать как имитатор природы.

На рис. 2.3 демонстрируется реализуемость теоремы Бернулли в эксперименте.

Когда выборка очень мала ($N = 10$), отклонения оценок параметра от модели ($k/N - p$) могут достигать очень больших значений. По мере увеличения объема выборки значения отклонений становятся меньше, а при очень большом числе экспериментов

---

* *Гнеденко Б. В., Хинчин А. Я.* Элементарное введение в теорию вероятностей. — М.: Наука, 1970.

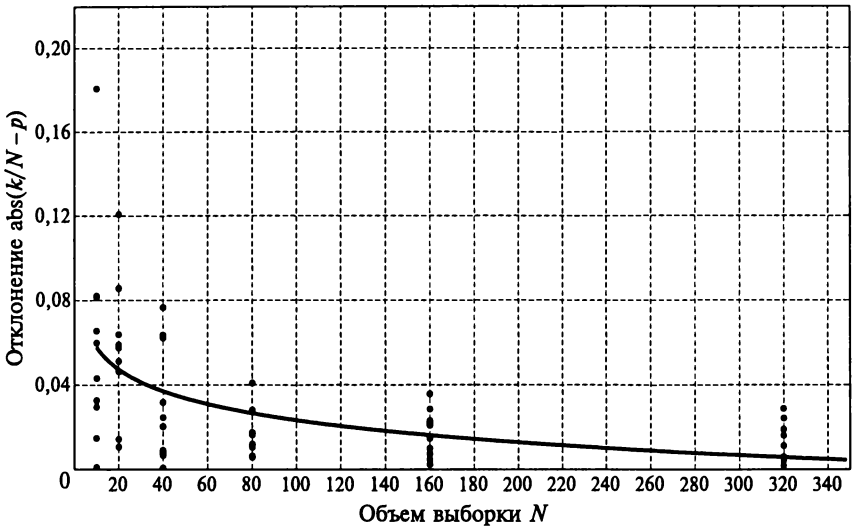


Рис. 2.3. Эксперимент на ЭВМ. Отклонения выборочной оценки параметра от заданного в модели как функции объема выборки

маловероятны. График показывает, что уже при 80 — 100 наблюдениях (например, оценки доли вида для выборки 80 — 100 деревьев) отклонения оценочного значения параметра от заданного в модели становятся сравнительно небольшими.

Таким образом, результаты эксперимента очевидно согласуются и с интуицией, и с теорий и, в какой-то степени показывают, что существует некоторое правило сходимости наблюдаемого с теоретическим, а в данном случае — с заданным и, следовательно, с реальным. Это правило становится важным объектом исследования и моделирования в теории вероятностей. Если оно известно, то очевидно можно предсказывать вероятность ошибок различного масштаба как функции объема выборки. В результате модели теории вероятностей становятся основой для оценки рисков ошибок, рисков принятия решений.

Модель биномиального распределения является самой простой и в какой-то степени базовой. Демонстрируя один из путей ее вывода, подчеркнем его полностью дедуктивный характер. Для экологов и географов она является моделью распределения дискретных классов элементов в условиях абсолютно однородной территории или среды. Если на территории каким-либо образом организована выборка, имитирующая схему Бернулли с возвращением, и частота успехов (частота выбора конкретной породы деревьев против всех остальных) соответствует биномиальному распределению, то можно с некоторым риском ошибки утверждать, что для деревьев рассматриваемая территория однородна.

Таким образом можно сделать следующие выводы. Опираясь на теорию множеств, последовательно, на самом общем уровне вводим модель теории вероятностей и строим наиболее простую и общую модель случайного события. Далее, с помощью генератора случайных чисел, имитирующего случайность в природе, убеждаемся в том, что параметр, оцененный по каждой конкретной выборке, в силу чистой случайности несколько отличается от реальности. Суждение о реальности на основе выборки неизбежно происходит с некоторой ошибкой, и величина ошибки тем меньше, чем больше объем выборки, по которой оценивается реальность.

В конечном итоге получаем, что случайное поведение подчиняется некоторым законам, знание которых даст в общем случае возможность судить о риске ошибок и о надежности полученных выводов. Именно на этой основе можно доказывать воспроизводимость результатов измерения в природе.

2.2. Распределения случайных событий

Прежде чем переходить собственно к статистическому оцениванию результатов измерений, необходимо рассмотреть модели распределения случайных событий или измерений. Эти модели воспроизводят распределения случайных событий и величин в природе и в общем случае важны для оценки возможности отнесения конкретных наблюдений к некоторым однородным условиям. Понятие однородности широко используется в экологии и географии. Так обычно допускается, что один тип сообщества растений или животных формируется в однородных условиях среды. Если на протяжении некоторого интервала времени условия не изменялись, то распределения среднегодовых температур и осадков должны варьировать как чисто случайные величины. С другой стороны, модель случайного процесса, порождающая конкретное распределение, строится при достаточно четко определенных условиях. Если какой-либо реальный случайный процесс описывается конкретным распределением, то можно полагать, что реальные механизмы, заложенные в его основу, соответствуют тем, которые заложены в модели. Таким образом, модели распределения создают первую основу для решения, или точнее — подхода к решению задач, стоящих перед любым исследователем.

Распределения обычно подразделяют на *дискретные* и *непрерывные*. В первом случае рассматриваются случайные события, которые однозначно различимы. Непрерывные распределения отражают распределения случайных величин, которые могут в принципе иметь любые значения, представимые через натуральные числа.

Распределения могут быть *одномерные* и *многомерные*. В первом случае случайная величина описывается лишь одним свойством,

например цветом, во втором — многими свойствами, например цветом, размером, плотностью и т.д.

Так, ежегодные состояния климата могут описываться отдельно распределением средних температур, суммой осадков и одновременно суммой температур и осадков. В реальных экологических и географических исследованиях чаще всего приходится иметь дело с многомерными распределениями.

Распределения описываются параметрами, т.е. некоторыми константами, зная которые, можно воспроизвести все их свойства. Выше уже был введен один параметр — математическое ожидание. Если расчет его аналога осуществлен для реальной выборки, то оно называется *выборочным средним*. Это различие существует в отношении всех других параметров. Параметр может быть задан в самой модели, а может быть оценен по выборке. Совершенно очевидно, что это качественно различные явления.

В табл. 2.3 приведены параметры распределений с необходимыми комментариями.

Каждый из параметров имеет свой смысл, который будем рассматривать применительно к конкретным распределениям.

Особым параметром дискретных распределений является *разнообразие*, или *энтропия*.

Разнообразие вводится на множестве несовместных подмножеств или классов как:

а) число различимых последовательностей из $N = n_1 + n_2 + ... + n_k$ объектов, содержащих n_1 неразличимых объектов типа 1, n_2 неразличимых объектов типа 2, n_k неразличимых объектов типа k;

б) число различимых разбиений последовательностей $N = n_1 + n_2 + ... + n_k$ различных объектов на k классов из $n_1, n_2, ..., n_r$.

Множество возможных последовательностей $C = n_1!n_2!n_3!...n_k!$ — все потенциально возможное разнообразие всех возможных комбинаций на множестве объема N. Подразумевается, что в природе может реализоваться любая из этих комбинаций, и некоторые из них окажутся устойчивыми и сохранятся во времени, например в форме сообщества растений, если классы есть виды, или в определенном механическом составе почв, если классы есть различимые фракции с определенными размерами элементов и т.д. Таким образом, подразумеваются потенциально возможные структуры, которые могут быть созданы на основе различных сочетаний элементов, принадлежащих разным классам.

Произведение факториалов — огромная величина, поэтому логично заменить ее суммой логарифмов от факториалов

$$I = \sum_{i=1}^{k} \log(n_k!).$$

При достаточно больших n_k в соответствии с формулой Стирлинга

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

имеем

$$\log(n!) = n \log n - n \log e + 0{,}5 \log n + 0{,}5 \log 2\pi \approx n \log n,$$

так как вклад всех членов кроме $n \log n$ ничтожен.

Соответственно можно записать

$$I = n_1 \log n_1 + n_2 \log n_2 + \ldots + n_k \log n_k.$$

Эта сумма уже вполне измеримая величина, однако она не удобна тем, что зависит от объема всего множества N. Чтобы найти величину, отражающую потенциальные возможности синтеза различных структур без учета объема множества, запишем

$$I = N \log N - N \log N + n_1 \log n_1 + n_2 \log n_2 + \ldots + n_k \log n_k.$$

Далее можно записать

$$-N \log N + n_1 \log n_1 + n_2 \log n_2 + \ldots + n_k \log n_k =$$

$$= -(n_1 \log n_1 / N + n_2 \log n_2 / N + \ldots + n_k \log n_k / N) =$$

$$= -(n_1 \log p_1 + n_2 \log p_2 + \ldots + n_k \log p_k).$$

Очевидно, что p_k — вероятность событий класса k.

Разделив и умножив сумму в скобках на N, получаем

$$-N \sum_{i=1}^{k} p_i \log p_i.$$

Очевидно, что параметр $H = -\sum_{i=1}^{k} p_i \log p_i$ содержит всю информацию о возможной сложности структур независимо от объема множества. Так же очевидно, что в отличие от множества она определена на вероятностном пространстве $\sum_{i=1}^{k} p_i = 1$.

Параметр H есть оценка разнообразия дискретного распределения или энтропия. В некоторых случаях эту оценку называют неопределенностью. Применяя это понятие, имеют в виду, что параметр показывает неопределенность исхода при случайном выборе, иначе говоря, неопределенность прогноза исхода испытания. Проведя испытание, уменьшаем неопределенность и соответственно получаем информацию. Чем больше разнообразие явления, тем больше объем получаемой информации в ходе его исследования. Если допустить, что система строго детерминирована и в результате исследования получена модель, обеспечивающая возможность однозначного прогноза ее поведения, то это означает, что получено количество информации, равное разнообразию ее состояний.

Таким образом, представление об информации прямо связано с разнообразием и неопределенностью.

Таблица 2.3

Основные параметры распределений

Тип параметра	Название	Формальная запись	Форма расчета по выборке	Определение	Комментарии
Характер положения	Центр распределения (математическое ожидание)	$M_x = \sum_x xp(x)$ — дискретное распределение $M_x = \int_{-\infty}^{\infty} xf(x)dx$ — непрерывное распределение	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Центр тяжести множества измерений. Характеризует положение величины x	Важнейший параметр, оценкой которого часто заканчиваются многие прикладные исследования
	Медиана	$X_{1/2}\ \ F(x) = 0{,}5$	$X(1/2) = \sum_{i=1}^{N/2} n(X_i) = N/2$	Числовое значение (X_i), делящее распределение пополам	
	Мода	$f(x) = \max$ — непрерывное распределение $f(x_{i-1}) < f(x) > f(x_{i+1})$ — дискретное распределение	Мода непрерывного распределения — точка максимума плотности распределения вероятности $f(x)$; дискретного — такое значение $P(k)$, что предшествующее и следующее за ним значения имеют вероятности меньше, чем $P(k)$	Распределение может иметь две и более мод, тогда распределение называется бимодальным или полимодальным	Мода — наиболее типичное состояние. Обычно бимодальные распределения свидетельствуют о неоднородности выборки
Характер рассеивания	Дисперсия	$D_x = \delta_x^2 = \sum_x (x - \bar{X})^2 p(x)$ — дискретное распределение $\int_{-\infty}^{\infty} (x - \bar{X})^2 f(x)dx$ — непрерывное распределение	$\delta_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$	Мощность явления — объем многомерного параллелепипеда со сторонами $(X_i - \bar{X})$. Характеризует рассеяние величины x	В общем случае чем больше дисперсия, тем больше мощность явления

	Среднее квадратическое отклонение	$\sigma_x = \sqrt{D_x}$	$\sigma_x = \sqrt{\sigma_x^2}$	Разброс распределения	
	Коэффициент вариации	v	$v = \frac{\sigma_x}{\bar{X}}$		
Моменты распределения	Центральные моменты порядка r	$\mu_r = M(x - \bar{X})^r =$ $= \sum_x [(x - \bar{X})^r p(x)]$ — дискретное распределение $\int_{-\infty}^{\infty} (x - \bar{X})^r f(x)dx$ — непрерывное распределение	$\mu_r = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^r$	Математическое ожидание отклонения значений от среднего в степени r	Если распределение симметрично относительно своего центра, то центральные моменты нечетного порядка равны нулю
	Математическое ожидание	$r = 1$	Преобразуйте по п.7		
	Дисперсия	$r = 2$			
	Коэффициент асимметрии	$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$		Мера отклонения моды от медианы или математического ожидания	
	Коэффициент эксцесса (куртозис)	$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$		Мера концентрации распределения около центра тяжести	
Квантили	Квантиль порядка P	Такое значение x_p случайной величины x, для которого $P(x < x_p) =$ $= F(x_p) = P$, $(0 < P < 1)$; $x_{1/2}$ есть медиана распределения	Квартили делят распределение на четыре части с границами $x_{1/4}$, $x_{1/2}$, $x_{3/4}$, децили — на 10, процентили — на 100 частей, попадания в которые имеют равные вероятности		

Разнообразие и неопределенность физически связаны с понятием энтропии в термостатике. В экологии, географии и других естественно-научных дисциплинах оценке разнообразия придают особое значение и рассматривают ее как одну из мер потенциальной ценности системы.

В дискретных распределениях энтропия максимальна, когда все классы или типы объектов равновероятны:

$$H_{\max} = \log k.$$

На этой основе вводится оценка выравненности

$$E = H/H_{\max}.$$

Выравненность тем меньше, чем более выражено доминирование какого-либо одного класса.

Для биномиального распределения выравненность максимальна при $p = 0{,}5$.

Рассмотрев основные параметры распределения, можно перейти к изучению наиболее важных из них для теории и практики.

Будем описывать распределения через порождающие их модели, графики, параметры и обычную область применения. Случайную величину или событие обозначим X. Модель, порождающую соответствующее распределение, будем связывать с генеральной совокупностью исходов, подразумевая под ней нечто абстрактное, отображающее некоторые свойства реальности. Генеральная совокупность — это то, что нельзя получить в прямых наблюдениях. Она мыслится как бесконечно большая и однородная.

Одномерные дискретные распределения

Биномиальное распределение (см. рис. 2.2). $P(x)$ есть:

1) вероятность того, что повторная, случайная выборка объема n содержит точно x элементов типа 1, если генеральная совокупность объема N содержит pN элементов типа 1;

2) вероятность появления события точно x раз в n независимых испытаниях по схеме Бернулли при условии, что вероятность события в каждом испытании равна p, т.е.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad (x = 0 < 1 < 2 < \ldots < n;\ 0 \le p \le 1).$$

Математическое ожидание $M_x = np$; дисперсия $D_x = np(1-p)$; третий центральный момент $\mu_3 = np(1-p)(1-2p)$; четвертый центральный момент $\mu_4 = np(1-p)[1 = 3(n-2)p(1-p)]$.

Типичная область применения в экологии и географии: оценка однородности сообщества по встречаемости видов отдельно для

каждого вида i и всех остальных не i по выборкам из n экземпляров.

Геометрическое распределение: $P(x)$ есть вероятность появления события типа 1 в первый раз после точно x испытаний по схеме Бернулли при его вероятности p, т.е.

$$P(x) = p(1-p)^x,\ (x = 0 < 1 < 2 < \ldots < n;\ 0 \le p \le 1).$$

Математическое ожидание $M_x = \dfrac{1-p}{p}$; дисперсия $D_x = \dfrac{1-p}{p^2}$; плотность распределения $f(x) = 1 - (1-p)^{x+1}$ есть вероятность того, что первый успех появится самое большее после x испытаний.

Таким образом, x — число испытаний до успеха, а M_x — математическое ожидание этого числа.

Распределение удобно, например, для оценки вероятности повреждения каких-либо плодов или других объектов. Действительно, собрав, например, много листьев какого-либо дерева, можно случайным образом брать листья до тех пор, пока не попадется лист с исследуемым типом повреждения, например с галлом. Повторив несколько раз такие выборки, можно оценить долю поврежденных листьев через выборочное среднее $\bar{x}$:

$$p = \frac{1}{1+\bar{x}},$$

где $\bar{x}$ — средний объем выборки до первого успеха.

В результате объем работы может быть существенно меньшим, чем при использовании выборок с фиксированным объемом n.

Гипергеометрическое распределение: $P(x)$ есть вероятность того, что случайная бесповторная выборка объемом n содержит точно x элементов типа 1, если эта выборка производится из генеральной совокупности N элементов, среди которых $N_1 = pN$ элементов принадлежат типу 1, т.е.

$$P(x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}},\ (x = 0, 1, 2, \ldots, n;\ N \ge n \ge 0;\ N \ge N_1 = pN > 0).$$

Математическое ожидание $M_x = \dfrac{nN_1}{N} = np$; дисперсия $D_x =$

$$= \frac{nN_1(N-N_1)}{N^2}\left(1 - \frac{n-1}{N-1}\right) = np(1-p)\left(1 - \frac{n-1}{N-1}\right).$$

Это распределение применяется для определения размера популяции N после выпуска в нее N_1 меченых особей. По истечении некоторого времени, достаточного для «перемешивания», осуще-

ствляется отлов n особей. Повторная выборка будет содержать x меченых особей, при этом $x < N_1$. Три величины в выражении для математического ожидания известны, и, соответственно, можно определить общую численность популяции. Такого рода оценки удобны для острова или для пространственно-замкнутой популяции. Для «открытой» популяции требуются дополнительные оценки возможной площади расселения.

Распределение Паскаля (*отрицательное биномиальное распределение*): $P(x)$ есть вероятность появления события типа 1 в m-й раз после точно $m + x - 1$ испытаний по схеме Бернулли при вероятности успеха p. $F(x)$ есть вероятность того, что m-й успех наступит самое большее после $m + x - 1$ испытаний. При $m = 1$ распределение Паскаля сводится к геометрическому распределению, т.е.

$$P(x) = (m_x + x - 1)\, p^m (1-p)^x, \quad (x = 0, 1, 2, \ldots, n;\ m = 1, 2, \ldots;\ 0 < p < 1).$$

Математическое ожидание $M_x = m\left(\frac{1-p}{p}\right)$; дисперсия $D_x = \left(\frac{1-p}{p^2}\right)$.

Этому распределению очень часто подчиняется распределение плотности многих видов животных, если x_i — число особей какого-либо вида в пробе i. Возможно, что модель, заложенная в распределении Паскаля, в какой-то степени имитирует их поведение в однородной среде. Передвигаясь в поисках пищи или благоприятных условий, особи задерживаются в точках пространства после некоторого числа успехов m, после $m + x - 1$ «шага». Ситуацию с некоторой величиной m собственных успехов в поиске пищи или условий среды они фиксируют как благоприятную для относительно длительного пребывания. Для такого распределения даже при довольно большой средней численности относительно большое число учетных площадок или проб оказываются пустыми.

Распределение Пуассона (*распределение редких событий*). Распределение Пуассона аппроксимирует гипергеометрическое и биномиальное распределения, когда $pN \to 1$, $N \to \infty$, $p \to 0$. Это приближение применяется обычно при $p < 0{,}1$, т.е.

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

Математическое ожидание и дисперсия в этом распределении равны: $M_x = D_x = \lambda$.

Третий момент распределения $\mu_3 = \lambda$, четвертый центральный момент распределения $\mu_4 = 3\lambda^2 + \lambda$.

Это распределение является основой для моделирования самых различных случайных процессов. Очень важным является его приложение в геометрической теории вероятности, где на его основе в частности выводится метод расчета суммы площадей сечений стволов деревьев релаксометром Битерлиха или призмой Анучина. На его же основе выводится формула подсчета численности животных на единицу площади по числу пересечения следов линейным маршрутом. Конечно, при этом допускается, что животные по территории передвигаются случайно (модель случайного блуждания).

Очень наглядное соотношение выводится на основе этого распределения для модели массового обслуживания, которая заслуживает специального рассмотрения.

Система массового обслуживания предполагает, что существует некоторый случайный пуассоновский поток требований с интенсивностью λ (например, среднее число людей, подходящих к кассе магазина за единицу времени). Система обслуживания работает с интенсивностью ν (число людей, которое может обслужить кассир в единицу времени).

Выясним, какова будет длина очереди, т.е. ожидание обслуживания, если интенсивность требований λ равна интенсивности обслуживания ν. Подавляющее большинство утверждает, что очередь, если и будет, то небольшой. В действительности же, если людям не надоест стоять в очереди, то при таких соотношениях параметров системы массового обслуживания очередь будет стремиться к бесконечности.

В соответствии с моделью массового обслуживания средняя длина очереди определяется по формуле

$$S = \frac{\lambda^2}{\nu(\nu - \lambda)}.$$

Очевидно, что если $\nu = \lambda$, то очередь будет бесконечна. Очередь будет тем меньше, чем в большей степени интенсивность обслуживания превосходит интенсивность требований.

Объясним природу такого соотношения. Предположим, что кассир на одного покупателя в среднем тратит 3 мин, тогда за один час он мог бы обслужить 20 клиентов. Однако это может иметь место только в том случае, если покупатели к кассе будут подходить строго регулярно, а длительность обслуживания строго постоянна. Если же поток и время обслуживания случайны, то в одни периоды кассир окажется свободным, а в другие перед кассой будет накапливаться очередь, т.е. кассир не работает все время, а работает только тогда, когда к нему подошел покупатель. С другой стороны, он на одного покупателя затратит много времени, а на другого меньше. Но покупатель, подошедший к кассе в первом случае, попадает в очередь, а за ним в очередь может попасть и

следующий покупатель. Иными словами, в конечном итоге время, которое затрачивает кассир на обслуживание, становится на 100 % используемым, только в том случае, если существует постоянная очередь. Таким образом, чтобы не было очередей, интенсивность обслуживания должна во много раз превосходить интенсивность требований.

В теории массового обслуживания определяются самые разнообразные параметры системы. Однако важно отметить, что эта модель имеет очень общее значение. Как систему массового обслуживания можно трактовать фильтрацию влаги в почве, где функцию «касс» выполняют поры, «очередей» — поверхностный сток, возникающий в тех случаях, когда подавляющее большинство пор заполнено водой, а вода продолжает поступать с осадками. Точно так же разложение опада в лесу можно рассматривать как систему массового обслуживания, где в качестве «кассиров» выступают почвенные беспозвоночные и микроорганизмы. Подстилку на поверхности почвы и, в какой-то степени, гумус можно трактовать как «очередь» в системе обслуживания. С этих позиций развитие болота или чернозема можно рассматривать как результат формирования бесконечной очереди.

Модели, в своей основе опирающиеся на модели случайных потоков, являются основой очень важных прикладных направлений теории вероятностей, теории надежности, восстановления, очередей и т. п. Их идеи широко используются при моделировании природных процессов.

Одномерные непрерывные распределения

Нормальное распределение можно определить как базовое для всей статистики. Моделью такого распределения может быть в частности функция $Y = a_1X_1 + a_2X_2 + ... + a_nX_n$, где X_i — независимые переменные, принимающие значения от 0 до n при примерно равных значениях коэффициентов a_i. Таким образом, модель описывает аддитивное действие на функцию множества равномощных независимых факторов.

Плотность распределения

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

где m_x — математическое ожидание; σ — среднее квадратическое отклонение.

Если положить, что $U = \frac{x - m_x}{\sigma}$, т. е. рассматривать вместо распределения реальных значений x значения, представленные в форме отклонения от среднего нормированного на среднее квадратиче-

ское, то получаем распределение стандартизованной нормальной величины:

$$f(U) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}U^2}.$$

Это распределение уже не зависит ни от среднего, ни от среднего квадратического или от дисперсии (мощности) и отражает свойства, присущие всем нормальным распределениям.

Следует особо отметить, что процедура стандартизации широко используется для того, чтобы сделать полностью соизмеримыми переменные с различными единицами измерения при самом различном масштабе собственного варьирования.

На рис. 2.4 приведен график плотности вероятности нормального распределения. По оси абсцисс графика отложены стандартизированные значения случайной величины, представленные в отклонениях от среднего на одну, две и три средних квадратических. Очевидно, что мода и медиана в нормальном распределении равны математическому ожиданию. Эксцесс и асимметрия, т.е. третий и четвертый центральные моменты распределения, равны нулю. Это фундаментальное свойство позволяет развить на основе законов нормального распределения весьма мощные средства анализа данных, которые часто называют параметрическими методами анализа. Эти методы будут важным предметом нашего изучения. Случайная величина, равная математическому ожиданию, имеет вероятность 0,4, т.е. такое значение можно ожидать в четырех случаях из десяти.

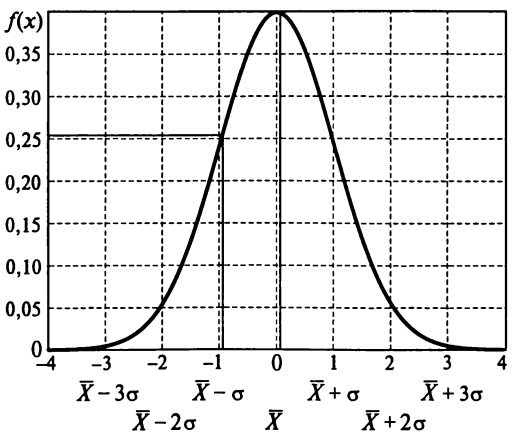


Рис. 2.4. График плотности вероятности нормального распределения

($f_{\mu}(U) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}U^2}$ — вероятность случайного непрерывного события U)

Случайная величина, на одну среднюю квадратическую меньше или больше среднего, произойдет с вероятностью 0,25. Однако любая случайная величина измеряется в некотором интервале, поэтому для оценки вероятности случайных величин удобнее пользоваться кумулятивной записью закона нормального распределения (рис. 2.5). В этом случае рассматривается накопленная вероятность случайных величин от 0 до 1, изменяющихся в интервале от $-\infty$ до U. Формально U может быть сколь угодно большим, но случайные величины со значениями, отличающимися от среднего на несколько средних квадратических, являются практически невозможными событиями.

Как следует из графика, 10 % случайных величин лежит в интервале от $-\infty$ до $-1{,}281552\sigma$ и 50 % — в интервале от $-\infty$ до математического ожидания. Соответственно в интервал от $-1{,}281552\sigma$ до 0 попадает 40 % случайных величин. Или случайные величины попадают в этот интервал с вероятностью 0,4. Случайные величины со значениями больше $-1{,}281552\sigma$ будут встречаться в интервале вероятностей 0,9; меньше $-1{,}281552\sigma$ — с вероятностью $1 - 0{,}9 = 0{,}1$. На рис. 2.5 левая шкала представлена в градациях децилей, а правая — квартилей. Соответственно, используя кумулятивный график, можно рассчитать вероятность попадания случайной нормально распределенной величины в любой интервал квартилей, децилей или процентилей и определить значение случайной величины для любого интервала вероятностей.

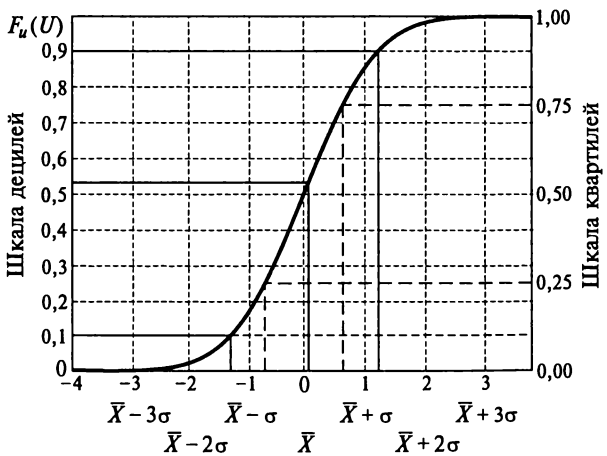


Рис. 2.5. Кумулятивная функция нормального распределения ($F_u(U) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{U} e^{-\frac{1}{2}U^2}$ — накопленная вероятность — сумма вероятностей случайных событий в заданном интервале случайных величин)

Следует обратить внимание на то, что в общем случае эти простые отношения нужно ощущать почти на уровне интуиции. Если необходимо выяснить, какова вероятность событий, превышающих по значению два средних квадратических отклонения, то надо сразу же понимать тот факт, что эта вероятность довольно мала и составляет приблизительно 0,023, а при отклонении на три средних квадратических она весьма мала и составляет около 0,013, а для четырех средних квадратических — около 0,00003. При этом очевидно, что знак отклонения не меняет этих оценок.

Нормальное распределение отражает равновесное состояние системы, когда варьирование ее состояний или положения ее элементов определяются только тепловым «шумом». В соответствии с этим его энтропия относительно всех других непрерывных распределений максимальна и прямо связана с дисперсией:

$$H(x) = 0,5 \log 2\pi e \sigma_x^2.$$

Так как энтропия не может быть отрицательной, имеем

$$\log 2\pi e + \log \sigma_x > 0.$$

Соответственно

$$\log \sigma_x > -2,83 \quad \text{и} \quad \sigma_x > 0,0587.$$

Это ограничение полезно при оценках информационных связей и разнообразия в системах с большим числом переменных.

Следует иметь в виду, что в природе собственно нормальные распределения встречаются крайне редко. Этот факт имеет вполне естественное объяснение. В природе мы постоянно имеем дело с активным преобразованием вещества, энергии, структуры, и, следовательно, действительно равновесные отношения встречаются крайне редко. Поэтому в лучшем случае распределения, с которыми мы имеем дело, приближаются к нормальным, а чаще существенно отличны от них. Все это будет накладывать определенные ограничения на применимость методов параметрического анализа данных, опирающихся на законы нормального распределения, и требовать от исследователя тщательного их анализа. С другой стороны, являясь моделью идеального равновесного состояния, случайного процесса, не подверженного действию какого-либо одного фактора, нормальное распределение является своеобразным идеалом. Сравнение реальности с идеалом равновесия позволяет оценить масштабы существующих отклонений. Если отклонения значительны, то очевидно в системе действует какой-либо фактор, смещающий ее из области равновесия. Часто задачей анализа и является поиск такого фактора.

Логнормальное распределение. Нормальное распределение подразумевает, что случайная величина определена на всей число-

вой оси от $-\infty$ до ∞. В природе таких событий, строго говоря, не существует. Даже температура по шкале Кельвина ограничена нулем. Так как нуль физически существует для всех явлений, это уже ограничивает применение нормального распределения. В какой-то степени выход из этого противоречия дает модель логнормального распределения, в котором закон нормального распределения применяется к логарифмам случайной величины. Точнее, случайной величиной становится логарифм измеренной величины.

Фактически эта модель генерируется уравнением $\log Y = a_1 \log X_1 + a_2 \log X_2 + \ldots + a_n \log X_n$.

В логнормальном распределении математическое ожидание определяется по формуле

$$M_x = m_x = \frac{1}{n} \sum_{i=1}^{n} \log X_i,$$

соответственно $\exp M_x = (X_1 \ldots X_n)^{\frac{1}{n}}$ есть среднегеометрическое; дисперсия

$$D_x = \sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\log X_i - m_x)^2.$$

В результате реальные отклонения от среднегеометрического в потенцированных значениях будут равны

$$\Delta = e^{t\sigma}.$$

Следовательно, если $t = -1, -2, -3$, то соответствующие отклонения при $\sigma_x = 1$ будут равны 0,369; 0,136; 0,05. В то же время при отклонении в положительную область $t = 1, 2, 3$ отклонения от среднегеометрического будут соответственно 2,71; 7,34; 19,98.

Таким образом, логнормальное распределение имеет резко асимметричную форму. Чем меньше среднеквадратическое отклонение, тем очевидно меньше асимметрия распределения, однако в нем всегда существует относительно большая вероятность больших значений X.

На рис. 2.6 приведено несколько вариантов логнормальных распределений, демонстрирующих изменения их вида при изменении параметров. Для логнормальных распределений обычно характерна левая асимметрия. При большом значении математического ожидания и небольших средних квадратических отклонениях логнормальное распределение приближается к нормальному.

Если исходные данные логарифмировать, то логнормальное распределение автоматически становится нормальным.

Логнормальное распределение часто хорошо соответствует реальным данным и широко используется для их преобразования

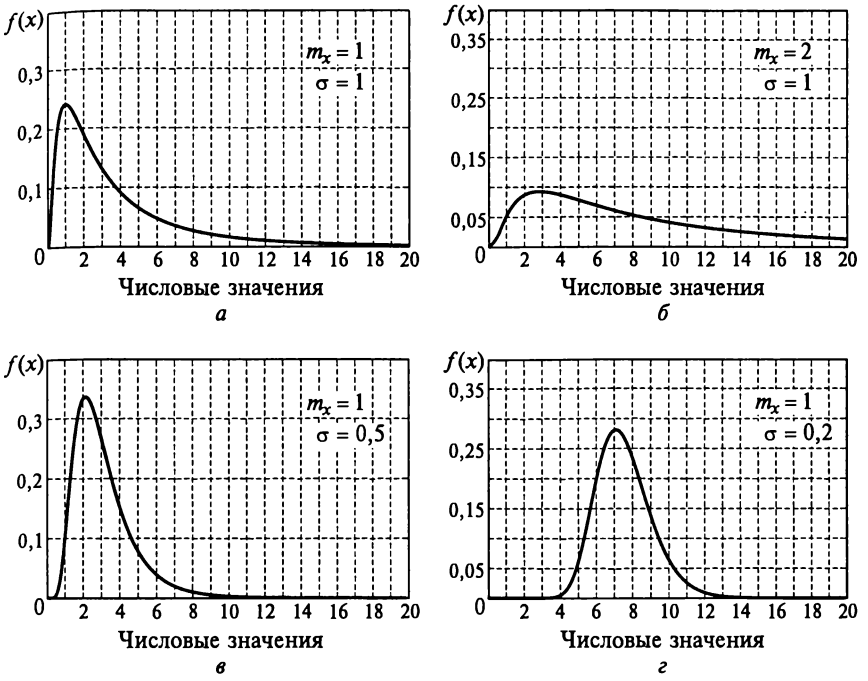


Рис. 2.6. Варианты (*а* — *г*) логнормальных распределений

при использовании параметрических методов анализа. Но самое важное, что оно часто удачно описывает распределение загрязнения, например в атмосфере или почве. Практическая важность использования логнормального закона определяется тем, что на его основе предсказывается отличная от нуля вероятность больших концентраций токсических веществ при вполне умеренных средних значениях. Если для тех же исходных данных использовать нормальное распределение, то очень высокие концентрации становятся почти невероятными событиями. Таким образом, если реальное распределение концентраций описывается логнормальным законом распределения, то практические следствия, вытекающие из этого факта, существенно изменяют представления о нормальном, допустимом уровне концентрации.

С другой стороны, необходимо обратить внимание на тот факт, что логнормальное распределение не может иметь нулевых значений переменной. Следовательно, оно не может описывать явление, для которого реально состояние «отсутствие». В случае же с загрязнением можно допустить существование сколь угодно малых, но отличных от нуля концентраций, лежащих за пределом точности измерения. В то же время, например, диаметр дерева не может быть сколь угодно малой величиной.

Гамма-распределение внешне часто очень похоже на логнормальное, но практика показывает, что содержательные различия часто оказываются весьма существенными.

Алгебраическая форма записи плотности распределения довольна сложна:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad (x > 0,\ \alpha > 0,\ \beta > 0),$$

где $\Gamma(\alpha) = \lim_{n\to\infty} \frac{n!\, n^{\alpha}}{\alpha(\alpha+1)\ldots(\alpha+n)}$ — гамма-функция.

Математическое ожидание $M_x = \frac{\alpha}{\beta}$, дисперсия $D_x = \frac{a}{\beta^2}$.

Существует и иная запись, используемая в частности в программе Statistica:

$$f(x) = \left(\frac{x}{b}\right)^{c-1} e^{-\frac{x}{b}} \left(\frac{1}{b}\tilde{\Gamma}(\tilde{n})\right),$$

где b — параметр масштаба (шкалирования); c — параметр формы.

Гамма-функция определяет некоторые фундаментальные свойства процессов, порождающих такие распределения. Эта функция называется трансцедентной, т.е. ненепрерывной. Если процесс описывается такой моделью, то он разрывный или фрактальный. Можно полагать, что такой процесс будет определять мозаичное пространственно-временно́е варьирование переменной. Например, в пространстве могут сочетаться пятна со значительными концентрациями с участками с нулевыми значениями. Аналогичное может иметь место и во времени. Гамма-функция, как следует из ее вида, фактически описывает комбинаторику возможных случайных событий. Сомножитель $e^{-\beta x}$ тем меньше, чем больше x. Например, чем больше концентрации, тем меньше их вероятность. С другой стороны, $x^{\alpha-1}$ растет с увеличением x. В результате получаем модель распространения выбросов от источника с мощностью, пропорциональной $x^{\alpha-1}$, с вероятностью переноса продуцируемого им вещества, пропорциональной $e^{-\beta x}$, по $\Gamma(\alpha)$ возможным траекториям.

Это, безусловно, упрощенная трактовка гамма-распределения. Однако распределение атмосферных осадков, аэрозолей, распределение химических веществ в почве, стоке, численности некоторых видов норных животных хорошо описываются именно этим распределением.

Гамма-распределение неплохо приводится к нормальному закону через извлечение квадратного корня из исходных данных.

Перечисленными распределениями, конечно, не исчерпывается все их разнообразие. Однако этот перечень включает распределения, с которыми наиболее часто приходится иметь дело.

Контрольные вопросы

1. Какова связь теории множества с теорией вероятности?
2. Опишите содержательные стороны модели случайного процесса.
3. Чем модель теории вероятностей отличается от реальности?
4. Что такое распределения и какова их природа?
5. Выпишите основные параметры распределений и их алгебраические формы записей и определите их физический смысл.
6. Используя компьютер, постройте модели различных распределений.
7. Исследуйте сходимость параметров моделей различных типов случайных процессов к их математическому ожиданию при увеличении объема выборки.

Глава 3
ОДНОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Настоящая глава включает формулировку основной задачи статистики, описание основной модели принятия решений, разбор специальных распределений, на основе которых принимаются решения, и решение типовых задач анализа одномерных выборок. Эти задачи являются базовыми и без их решения нельзя переходить к более интересным и содержательным задачам многомерного анализа.

3.1. Логические основания проверки статистических гипотез

Статистическая гипотеза есть непротиворечивое предположение о распределении случайной величины. Если она однозначно определяет распределение, то статистическая гипотеза H называется *простой*, если же она определяет некоторую область состояний, то она называется *сложной*.

Например, для врача множество возможных болезней может рассматриваться как вероятностное пространство с совместными и несовместными классами болезней и несовместным с ними событием «здоров». Все они в совокупности образуют распределения. На основе некоторых признаков и критериев врач должен соотнести состояние больного с его множествами признаков или с конкретной «точкой» — состоянием, или с некоторой областью, включающей различные болезни, или с областью конкретной болезни, или с областью «здоров». Следует отметить, что характер работы врача часто наглядно демонстрирует многие существенные стороны статистики. Однако точно такие же задачи стоят, например, перед экологом, оценивающим качество среды, или выявляющим субъект хозяйственной деятельности, допустивший нарушение. Более того, любое исследование в природе, по сути дела, можно свести к модели «врач — пациент», с той лишь разницей, что исследователь должен поставить диагноз, оп-

ределяющий природу исследуемого явления, а врач должен, кроме того, еще и лечить.

Итак, пусть дана некоторая фиксированная выборка объемом n; **критерий статистической гипотезы** H есть правило, позволяющее отвергнуть или не отвергнуть гипотезу H на основании выборки. Каждый критерий определяет **критическое множество (область)**. Гипотеза H отвергается, если выборка принадлежит критическому множеству, и не отвергается в противном случае.

Возвращаясь к модели «врач—пациент», определяем как выборку фиксированные результаты обследования пациента. Пусть в простейшем случае врач должен ответить на вопрос «болен» пациент или «здоров». Поскольку пациент обратился к нему сам, то исходная гипотеза «болен». Если выборка, полученная в ходе обследования, принадлежит некоторому множеству, которое ближе к состоянию «здоров», то врач отвергает гипотезу о болезни. Заметим, что если проводится профилактическое обследование, то проверяемая гипотеза — «здоров».

Однако такая схема принятия или отбрасывания гипотезы не дает ее логического доказательства или опровержения.

При этом возможны четыре случая.

1. Гипотеза H верна и принимается согласно критерию.
2. Гипотеза H неверна и отвергается согласно критерию.
3. Гипотеза H верна, но отвергается согласно критерию (*ошибка первого рода*).
4. Гипотеза H неверна, но принимается согласно критерию (*ошибка второго рода*).

Для любого множества измеренных значений вероятность отвергнуть проверяемую гипотезу по данной критической области определяется вероятностью принадлежности этих измеренных значений данной критической области, т.е. чем больше вероятность того, что измеренные значения принадлежат критическому множеству, тем больше вероятность отвергнуть гипотезу.

Если гипотеза H конкурирует лишь с одной альтернативной гипотезой $H1$ «болен—здоров», то вероятность P отвергнуть гипотезу H, когда верна гипотеза $H1$, называется *мощностью критерия*, определенного на критическом множестве по отношению к гипотезе $H1$. Вероятность не отвергнуть гипотезу H, т.е. $1 - P$, называют *оперативной характеристикой критерия*.

Таким образом, принятие решения во многом, если не в основном, зависит от того, как определена «критическая область». Если граница между «болен—здоров» нечеткая или область пересечения двух этих подмножеств очень большая, то вероятность отвергнуть гипотезу «болен» будет очень велика и наоборот.

Желательно применять такую критическую область, чтобы вероятность P была мала, если проверяемая гипотеза верна, и велика в противном случае.

Пусть проверяется гипотеза H_0 и пусть эта гипотеза верна. Тогда вероятность напрасно отвергнуть гипотезу H_0 (ошибки первого рода) обозначается как $P = \alpha$, где α называется *уровнем значимости данного критерия* (P-levels).

От этих общих положений о правилах принятия решений в условиях неопределенности можно перейти к решению основных задач статистики, к которым относят:

1) оценку параметров явления по выборке;

2) проверку гипотезы о принадлежности выборочного распределения к одному из стандартных распределений;

3) проверку гипотезы о принадлежности двух выборок к одной генеральной совокупности.

3.2. Описательные статистики

В любом пакете статистических методов существует раздел «Описательная статистика» (Descriptive Statistics), который обычно содержит параметры, приведенные в табл. 3.1.

Демонстрация одномерного анализа будет осуществляться на основе 100-летнего ряда наблюдения температур и выпадения атмосферных осадков по данным метеостанции «Рязань». Выбор климатических рядов наблюдений в качестве примера использования одномерного анализа определяется тем, что:

- почти каждому экологу приходится иметь дело с такого рода данными;
- все рассматриваемые приемы анализа могут быть легко распространены на практически любые наблюдения во времени и в пространстве;
- проблемы динамики климатических переменных весьма актуальны.

На рис. 3.1 показано изменение среднемесячных температур за 100 лет. Поверхность отражает сглаженное изменение температур.

В качестве элемента принимается среднемесячная температура, а в качестве генеральной совокупности — все множество среднемесячных температур за 100 лет. Это множество представлено 12 независимыми выборками. То, что эти выборки принадлежат одной генеральной совокупности с общим математическим ожиданием, дисперсией и другими параметрами, принимается в качестве нулевой гипотезы. Конечно, эта гипотеза априори несостоятельна. Критическая область между летними и зимними температурами пуста. Очевидно, что если среднемесячная температура отрицательна, то это могут быть только зимние месяцы. Однако температуры, например, декабря, января и февраля могут принадлежать одной генеральной совокупности. Критическая область между их распределениями очень велика. Это означает, что система в те-

Таблица 3.1

Основные параметры, используемые в описательной статистике

№	Наименование параметра		Формула вычисления	Комментарии
	русское	английское		
1	Значимые N	Valid N	$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$	Объем выборки
2	Среднее	Mean		
3	Медиана	Median		
4	Мода	Mode		
5	Дисперсия	Variance	$D_x = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	
6	Среднее квадратическое отклонение	Standard Deviations	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N-1}}$	
7	Средняя квадратическая ошибка среднего	Standard error of mean	$m = \frac{\sigma}{\sqrt{N}}$	
8	95%-й доверительный интервал	95 % Confidence limit of mean		Область, в которой с вероятностью 0,95 лежит математическое ожидание, оцененное по выборке
9	Минимум и максимум	Minimum & Maximum		

Окончание табл. 3.1

№	Наименование параметра		Формула вычисления	Комментарии
	русское	английское		
10	Амплитуда	Range		
11	Максимальная и минимальная квартили	Lower & upper quartiles		
12	Амплитуда по квартилям	Quartile range		
13	Асимметрия	Skewness	$\frac{m_3}{m_2^{3/2}}$	
14	Эксцесс	Kurtosis	$\frac{m_4}{m_2^2} - 3$	
15	Стандартная ошибка асимметрии	Standard error of Skewness		
16	Стандартная ошибка эксцесса	Standard error of Kurtosis		
17	Число степеней свободы	Freedom	$df = (N - 1)$	См. прим. к табл.

Примечание. Число степеней свободы в общем определяется объемом выборки N, оно описывает возможное множество ячеек, которые случайным образом могут быть заполнены N событиями. Допустим, что событий всего два. Два события с высокой вероятностью чисто случайно могут оказаться в одной ячейке и у нее остается еще одна свободная. Соответственно число степеней свободы для выборки объемом 2 равно 1. Таким образом, до тех пор, пока не оговорено иное, число степеней свободы есть объем выборки за минусом единицы.

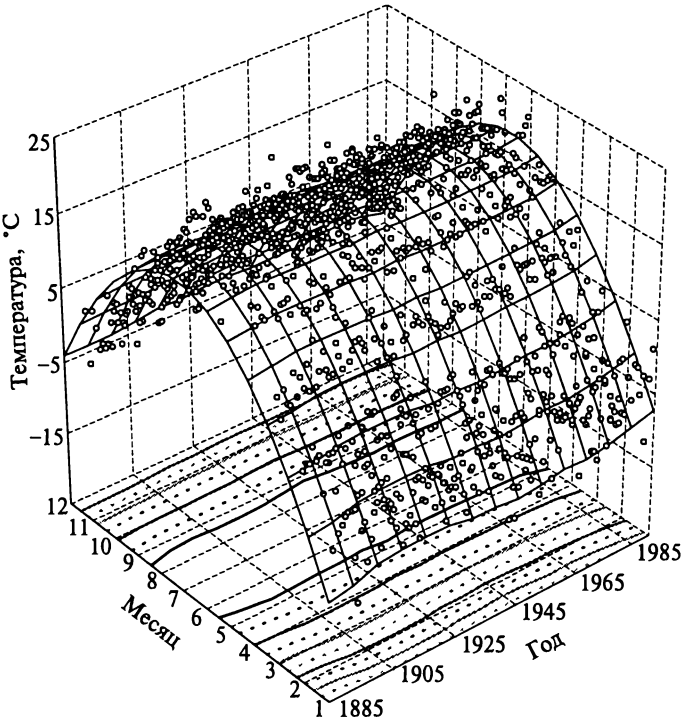


Рис. 3.1. Изменение среднемесячных температур за 100 лет (по данным метеостанции «Рязань»)

чение трех месяцев однородна и находится в стационарном режиме. Если же по оцененным параметрам риск ошибки второго рода при отнесении их к одной генеральной совокупности велик, или вероятность того, что они могут принадлежать к одной генеральной совокупности, очень мала, то это будет означать, что выборки взяты из разных генеральных совокупностей.

Отметим, что ту же систему можно задать иначе. Для этого достаточно в качестве элемента определить год, а температуры месяцев рассматривать как переменные, при этом система становится многомерной.

В табл. 3.2 приведены оценки параметров 12 одномерных выборок.

Естественно начать проверку гипотезы на основе наиболее важного параметра «выборочной средней». Выборочная средняя рассматривается совместно со связанными с нею параметрами. Например, средняя за 100 лет температура января в Рязани −10,385 °С, при этом в 95 % случаев она попадает в интервал от −11,188 до −9,582 °С. Медианная температура −9,8 °С несколько меньше средней, но лежит в 95%-м доверительном интервале. Верхняя кварти-

Таблица 3.2

Основные описательные статистики (Descriptive Statistics)

Наименование параметра	Переменная											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Значимые N Valid N	98,00	98,00	98,00	98,00	98,000	98,000	98,000	98,000	98,000	98,00	98,00	98,00
Среднее Mean	–10,385	–9,919	–4,526	4,896	13,058	17,253	19,014	17,477	11,715	4,694	–1,947	–7,439
Доверительный интервал нижний Confid. –95 %	–11,188	–10,698	–5,075	4,398	12,587	16,843	18,670	17,129	11,344	4,256	–2,446	–8,065
То же, верхний Confid. +95 %	–9,582	–9,141	–3,976	5,394	13,529	17,663	19,358	17,824	12,086	5,131	–1,448	–6,813
Медиана Median	–9,800	–10,400	–4,700	4,550	13,000	17,250	18,900	17,300	11,750	4,650	–1,800	–7,200
Сумма Sum	–1017,7	–972,1	–443,50	479,80	1279,700	1690,800	1863,400	1712,700	1148,100	460,000	–190,80	–729,0
Минимум Minimum	–20,900	–20,200	–10,700	–2,100	7,000	13,300	15,800	13,600	7,600	–1,500	–7,900	–15,800
Максимум Maximum	–3,600	–700	3,400	10,400	18,700	22,400	23,800	23,000	19,800	9,700	4,300	–0,800
Минимальная квартиля Lower quartil	–12,900	–12,300	–6,200	3,100	11,500	15,900	17,800	16,400	10,500	3,000	–3,600	–9,700
Максимальная квартиля Upper quartil	–7,500	–7,700	–2,600	6,300	14,400	18,500	19,900	18,500	12,600	5,800	–0,400	–5,100

Амплитуда Range	17,30	19,50	14,10	12,50	11,700	9,100	8,000	9,400	12,200	11,20	12,20	15,00
Ранг квартилей Quartile Range	5,400	4,600	3,600	3,200	2,900	2,600	2,100	2,100	2,100	2,800	3,200	4,600
Дисперсия Variance	16,049	15,091	7,525	6,179	5,519	4,182	2,942	3,012	3,422	4,759	6,184	9,755
Среднее квадратическое отклонение STD.DEV.	4,006	3,885	2,743	2,486	2,349	2,045	1,715	1,736	1,850	2,182	2,487	3,123
Средняя квадратическая ошибка Standart error of mean	0,405	0,392	0,277	0,251	0,237	0,207	0,173	0,175	0,187	0,220	0,251	0,316
Асимметрия (косость) Skewness	–0,514	–0,096	0,050	0,216	0,198	0,197	0,562	0,451	0,841	–0,009	–0,112	–0,258
Ошибка косости STD.ERR.SK	0,244	0,244	0,244	0,244	0,244	0,244	0,244	0,244	0,244	0,244	0,244	0,244
Эксцесс Kurtosis	–0,202	0,141	0,137	–0,027	0,067	–0,345	0,049	0,351	3,059	0,065	–0,149	–0,276
Ошибка эксцесса STD.ERR.	0,483	0,483	0,483	0,483	0,483	0,483	0,483	0,483	0,483	0,483	0,483	0,483

ля –12,9 °С, нижняя –7,5 °С с амплитудой 5,4 °С. Общая амплитуда температур составляет 17,3 °С. Средняя квадратическая ошибка ±0,405.

3.3. Параметрические критерии проверки гипотез

Из закона больших чисел и проведенного выше эксперимента следует, что средняя квадратическая ошибка будет уменьшаться как функция объема выборки. Она показывает возможное положение истинного среднего относительно выборочного. Чем меньше ошибка, тем лучше оценка математического ожидания генеральной совокупности по выборке. Следовательно, логично полагать, что с помощью средней квадратической ошибки можно образовать критическую область проверки гипотезы. Но для этого нужно связать среднее и среднюю квадратическую ошибку с вероятностным пространством, в котором и будет образована критическая область.

Это осуществляется с помощью распределения Стьюдента.

Доказывается, что отклонение между выборочным средним и его математическим ожиданием распределено следующим образом:

$$t = \frac{\bar{X} - M_x}{m},$$

где $\bar{X}$ — выборочное среднее; M_x — математическое ожидание; m — средняя квадратическая ошибка.

В общем случае чем больше значение t, тем меньше вероятность принадлежности выборочного среднего к генеральной совокупности. По умолчанию, когда о генеральной совокупности ничего не известно, ее математическое ожидание принимается равным нулю и, соответственно, t-критерий образует критическое множество относительно генеральной совокупности с математическим ожиданием, равным нулю.

При t = 0 критическая область определяет, что с вероятностью 1 выборочное среднее соответствует математическому ожиданию генеральной совокупности. Это, очевидно, может иметь место только в том случае, когда среднее равно нулю. При t = 1 вероятность принадлежности выборочного среднего p = 0,5 при очень малом объеме выборки или иначе «числе степеней свободы», а при выборке объемом больше 10 p = 0,16. При t = 2 p = 0,12 при малом объеме выборки и около 0,03 — при большом. При t = 3 p = 0,003, т. е. только в трех случаях из 1000 гипотеза о принадлежности среднего генеральной совокупности с математическим ожиданием, равным нулю, будет верна. Таким образом, если в этом случае принять гипотезу верной, то в подавляющем большинстве случаев это

будет ошибкой. Таким образом, t-критерий является критерием значимости, а уровень значимости определяется по связанному с ним распределению.

Очевидно возникает вопрос: «какой уровень значимости принять при проверке нулевой гипотезы». Выбор объема критического множества, а проще уровня критерия, определяется, с одной стороны, последствиями, к которым может привести отклонение нулевой гипотезы, если она на самом деле верна, а с другой стороны — затратами, которые необходимы для улучшения критерия, т.е. дополнительными измерениями или исследованиями. В чисто научных исследованиях потери, связанные с получением ложных выводов в общем случае не приводят к сколь-либо тяжелым последствиям, однако приводят к неоправданно большому «шуму» в потоке публикаций. Обычно в научных исследованиях принимается минимальный уровень значимости около 0,05 и лишь иногда допускают 0,1. При этом уровне значимости выявленные отношения принимаются достоверными. Однако в медицине, где ошибки связаны со здоровьем человека, нулевая гипотеза о том, что человек здоров, отвергается при уровне значимости 0,25.

Впрочем, в современном мире результаты научных исследований могут часто повлечь за собой далеко идущие экономические и социальные последствия. Допустим, например, что представления об антропогенном потеплении климата «ложны». Это означало бы, что все средства, затрачиваемые на уменьшение антропогенного загрязнения атмосферы, израсходованы по существу впустую. Но загрязнение атмосферы по статистически доказанной гипотезе прямо влияет на здоровье человека, поэтому средства, вложенные в снижение загрязнения, тратятся, скорее всего, оправданно.

Итак, выбор уровня значимости в конечном итоге имеет чисто практическую мотивацию, но понимать его относительность совершенно необходимо. Иногда и в научных исследованиях допустимо мотивированное снижение уровня критерия значимости.

Перейдем к проверке гипотезы о принадлежности двух средних одной генеральной совокупности. Предварительно отметим, что:

• сумма (разность) выборочного среднего двух независимых выборок равна сумме (разности) их выборочных средних:

$$\overline{X + Y} = \bar{X} + \bar{Y};$$

• выборочная дисперсия двух независимых выборок есть сумма их частных дисперсий:

$$\sigma^2_{(X,Y)} = \sigma^2_X + \sigma^2_Y;$$

• соответственно средняя квадратическая ошибка выборочного среднего двух объединенных выборок (вне зависимости от того, рассматривается их сумма или разность) равна

$$m_{xy} = \sqrt{m_x^2 + m_y^2}.$$

На основе этих предварительных замечаний можно поставить задачу проверки гипотезы о принадлежности двух выборочных средних одной генеральной совокупности. Действительно,

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{m_x^2 + m_y^2}},$$

с помощью t-критерия проверяем гипотезу о принадлежности математического ожидания их разности генеральной совокупности с математическим ожиданием, равным нулю. Соответственно можно полагать, что если t мало́, то выборки принадлежат одной генеральной совокупности. (Сразу же отметим, что две выборки могут не различаться по средним, но различаться по дисперсиям, и тогда они будут принадлежать к разным генеральным совокупностям).

В рассматриваемом примере проверяется гипотеза о принадлежности двух выборок одной генеральной совокупности только на основе сравнения их выборочных средних.

В статистических пакетах программ имеются стандартные процедуры расчета t-test для выборочных средних (табл. 3.3).

Из табл. 3.3 следует, что нулевая гипотеза о принадлежности средних одной генеральной совокупности может быть принята для средних температур января и февраля. В то же время зимний месяц декабрь достоверно теплее января и февраля. Средняя температура апреля не отличается от температуры октября, а температура мая — от температуры сентября. Средняя температура июня сопоставима с температурой августа, а июль достоверно теплее июня и августа.

В табл. 3.4 приведены соответствующие уровни вероятности отнесения средних к одной генеральной совокупности.

Из табл. 3.4 следует, что выделенные выше месяцы действительно с высокой вероятностью относятся к одной генеральной совокупности.

Современные программы дают хорошие возможности визуально тестировать значимость различий средних. Обычно эти графические средства используют до начала количественного анализа. Метод построения изображенного на рис. 3.2 графика называется бокс-плот (box-plot).

На рис. 3.2 все полученные результаты наблюдаемы визуально. Так как диапазон доверительных интервалов для 2 и 3-го уровней среднеквадратических ошибок полностью перекрывается для температур января и февраля, очевидно, что температуры этих месяцев достоверно неотличимы. Точно также очевидна и высокая симметричность сезонного хода.

Содержательным выводом из первого результата анализа является предположение о высокой стационарности системы в течение

Таблица 3.3

Различия средних по t-критерию (переменные строк вычтены из переменных столбцов)
(Mean Differences (row vars minus column vars))

Перемен-ная	Переменная											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
M1	0,0	**0,5**	–5,9	–15,0	–23,0	–28,0	–29,0	–28,0	–22,0	–15,0	–8,4	–2,9
M2	**0,5**	0,0	–5,4	–15,0	–23,0	–27,0	–29,0	–27,0	–22,0	–15,0	–8,0	–2,5
M3	5,9	5,4	0,0	–9,0	–18,0	–22,0	–24,0	–22,0	–16,0	–9,0	–2,6	2,9
M4	15,3	14,8	9,4	0,0	–8,0	–12,0	–14,0	–13,0	–7,0	**0,0**	6,8	12,3
M5	23,4	23,0	17,6	8,0	0,0	–4,0	–6,0	–4,0	**1,0**	8,0	15,0	20,5
M6	27,6	27,2	21,8	12,0	4,0	0,0	–2,0	–0,0	6,0	13,0	19,2	24,7
M7	29,4	28,9	23,5	14,0	6,0	2,0	0,0	2,0	7,0	14,0	21,0	26,5
M8	27,9	27,4	22,0	13,0	4,0	0,0	–2,0	0,0	6,0	13,0	19,4	24,9
M9	22,1	21,6	16,2	7,0	**–1,0**	–6,0	–7,0	–6,0	0,0	7,0	13,7	19,2
M10	15,1	14,6	9,2	**–0,0**	–8,0	–13	–14,0	–13,0	–7,0	0,0	6,6	12,1
M11	8,4	8,0	2,6	–7,0	–15,0	–19,0	–21,0	–19,0	–14,0	–7,0	0,0	5,5
M12	2,9	2,5	–2,9	–12,0	–20,0	–25,0	–26,0	–25,0	–19,0	–12,0	–5,5	0,0

Таблица 3.4

Т-тест для выбранных данных (уровни значимости) (T-test for Dependent Samples: p-levels)

Переменная	Переменная											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
M1	1,00	0,32	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
M2	0,32	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
M3	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
M4	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,59	0,00	0,00
M5	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
M6	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,34	0,00	0,00	0,00	0,00
M7	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00
M8	0,00	0,00	0,00	0,00	0,00	0,34	0,00	1,00	0,00	0,00	0,00	0,00
M9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00
M10	0,00	0,00	0,00	0,59	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00
M11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00
M12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00

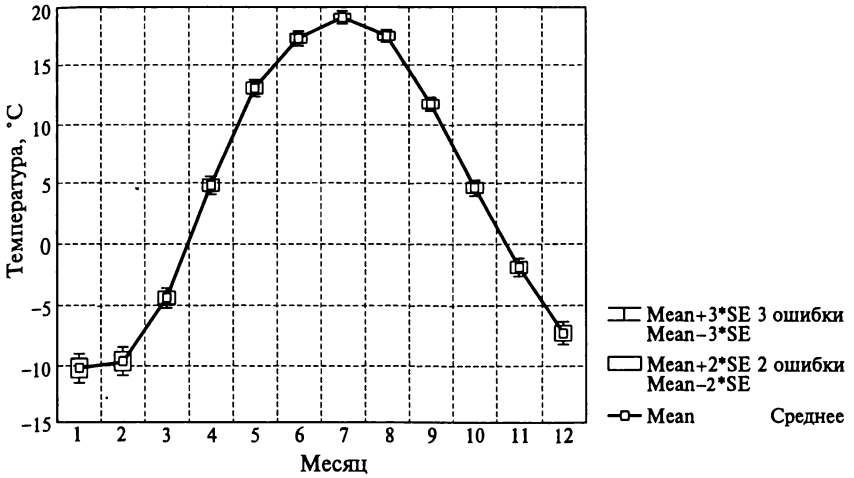


Рис. 3.2. Сезонный ход температур по осреднению за 100 лет (по данным метеостанции «Рязань»)

двух зимних месяцев и о достижении климатической системой в этот период некоторого состояния термодинамического равновесия. Можно полагать, что равновесное радиационное выхолаживание достигается в январе и сохраняется в феврале. Летом же такая равновесная область практически не выражена.

На следующем этапе анализа проверяется гипотеза стационарности или неизменности условий во времени на протяжении 100 лет. Нулевая гипотеза подразумевает, что если условия стационарны, то распределение температур за каждый месяц должно быть нормальным или должно приводиться к нормальному. Если распределения нормальны, то коэффициенты асимметрии и эксцесса не должны достоверно отличаться от нуля, а медиана должна быть равна среднему.

Отношение эксцесса и асимметрии к их ошибкам имеет также t-распределение и, используя его, можно в первом приближении выделить месяцы, в которые, скорее всего, выборки имеют ненормальное распределение (табл. 3.5).

Из табл. 3.5 следует, что асимметрия не соответствует гипотезе нормальности и скорее всего характерна для января, июля, августа и октября, а эксцесс — только для октября. По-видимому, наибольшие отличия от нормального распределения существуют в октябре. В остальные месяцы, в соответствии с рассматриваемым критерием распределения, они близки к нормальным.

Однако это лишь самый предварительный тест. Более сильную оценку можно получить на основе сравнения реального распределения с нормальным. При этом нормальное распределение строится по выборочным параметрам.

Таблица 3.5

Проверка гипотезы нормальности распределения среднемесячных температур за 100 лет (Descriptive Statistics)

Переменная	Разница между средней и медианой	t-критерий для асимметрии	Асимметрия Skewness	t-критерий для эксцесса	Эксцесс Kurtosis
M1	−0,585	−2,108	−0,514	−0,419	−0,202
M2	0,481	−0,395	−0,096	0,291	0,141
M3	0,174	0,204	0,050	0,284	0,137
M4	0,346	0,885	0,216	−0,057	−0,027
M5	0,058	0,811	0,198	0,139	0,067
M6	0,003	0,810	0,197	−0,714	−0,345
M7	0,114	2,304	0,562	0,102	0,049
M8	0,177	1,851	0,451	0,727	0,351
M9	−0,035	3,451	0,841	6,334	3,059
M10	0,044	−0,037	−0,009	0,135	0,065
M11	−0,147	−0,458	−0,112	−0,308	−0,149
M12	−0,239	−1,058	−0,258	−0,572	−0,276

Проверка гипотезы о принадлежности реального распределения к его модели осуществляется на основе специальных тестов. Эти тесты строятся также как и критерий Стьюдента на основе специальных распределений, описывающих правила сходимости распределения случайных величин к постулируемой норме.

Если критерий Стьюдента строился на основе двух параметров распределения «среднего и среднего квадратического», то эти критерии опираются уже на свойства всей выборки. Соответственно они называются *непараметрическими.*

При проверке гипотезы о принадлежности двух распределений одной генеральной совокупности обычно используется критерий хи-квадрат и критерий Колмогорова — Смирнова.

Типичная интерпретация распределения χ^2 («хи-квадрат») следующая:

если u_i (i = 1, 2, ..., m) взаимно независимых стандартизованных случайных величин имеют нормальное распределение, то сумма их квадратов $\chi^2 = \sum_{i=1}^{m} u_i^2$ имеет χ^2-распределение с m степенями свободы. Математическое ожидание такого распределения $M_y = 0$ и дисперсия $\sigma_y^2 = 1$.

Это распределение является важным во многих статистических задачах.

Если среднее значение распределения $\chi^2 = m$, то вероятность нулевой гипотезы близка 0,5 и распределения с высокой вероятностью принадлежат одной генеральной совокупности, при значении $\chi^2 = 2m$ при числе степеней свободы больше 10 вероятность нулевой гипотезы близка к 0,02.

При сравнении распределений необходимо определить число градаций или число степеней свободы, приемлемых при данном объеме выборки. Если градаций очень много, то при случайном размещении многие из них могут оказаться пустыми. Если градаций будет очень мало, то не будет использоваться вся информация, содержащаяся в выборке.

Обычно число градаций определяется как $df = \log N + 1$. Это правило прямо вытекает из возможных комбинаций из N. В программных средствах операция определения числа степеней свобод осуществляется автоматически и часто с дополнительным учетом параметров распределения.

Проверим гипотезу о нормальности распределения для октября с наибольшими значениями t-критерия для асимметрии и эксцесса.

На рис. 3.3 показан график распределения и расчетный вид нормального распределения, построенного по параметрам выборочного распределения. Из графика хорошо видны источники эксцесса и асимметрии. На рисунке выписаны результаты трех тестов. В соответствии с тестом χ^2-распределение с вероятностью 0,46 должно быть отнесено к нормальному. Нет отличия от нормального и

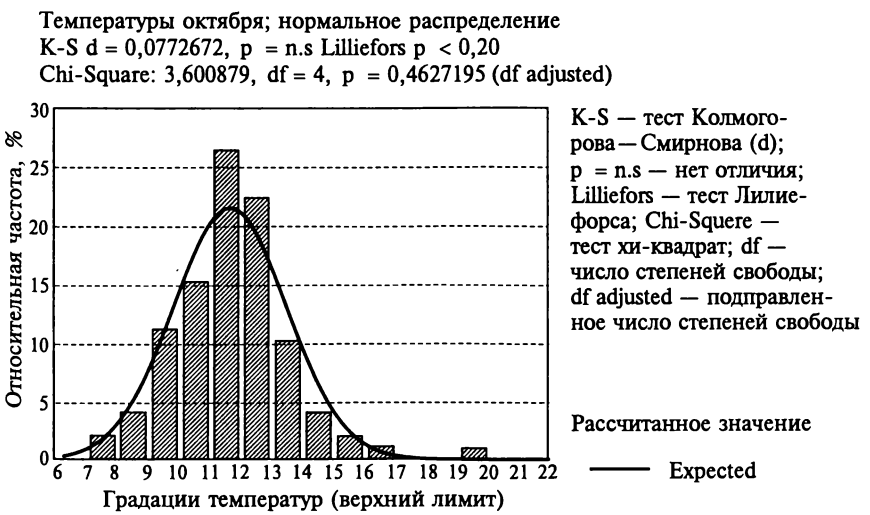


Рис. 3.3. Проверка гипотезы о нормальности распределения

по более сильному критерию Колмогорова — Смирнова и лишь по самому мощному тесту Лилиефорса вероятность нормальности распределения меньше 0,2.

Таким образом, можно полагать, что даже распределение с наиболее выраженным эксцессом и асимметрией является почти наверное нормальным и соответственно в рамках этих критериев можно полагать, что температурный режим климата на протяжении всех 100 лет измерений в районе метеостанции «Рязань» оставался неизменным. Следует отметить, что тесты распределений не учитывают информации об изменении температуры по годам. Вполне возможно, что использование более полной информации о явлении заставит нас отказаться от гипотезы неизменности многолетнего хода температур. Однако на данном этапе нам ничего не остается, как принять ее.

Не останавливаясь на технологии расчета хи-квадрат (эта процедура, как и другие, выполняется программными средствами), проведем сравнение дисперсий между наиболее близкими распределениями температур в январе и феврале. Если выборочные дисперсии мало отличимы, то можно с еще большим основанием утверждать о принадлежности выборок одной генеральной совокупности.

Вопрос о принадлежности выборочных дисперсий двух распределений одной генеральной совокупности решается на основе F-критерия, который в свою очередь строится на основе распределения Фишера.

Отношение большей дисперсии к меньшей двух распределений описывается F-распределением Фишера. Очевидно, что если отношение равно 1 (дисперсии равны), то выборки принадлежат одной генеральной совокупности.

Математическое ожидание F-распределения

$$M(\mathrm{F}) = \frac{m_1}{m_2 - 2}.$$

где m_1 — число степеней свободы для распределения с максимальным объемом выборки; m_2 — число степеней свободы для распределения с минимальным объемом выборки. Математическое ожидание дисперсии при $m > 30$

$$\sigma^2(\mathrm{F}) = \frac{m_1(m_1 + 1)}{(m_2 - 1)(m_2 - 2)}.$$

Таким образом, при достаточном объеме выборки математическое ожидание отношения большей дисперсии к меньшей близко к единице. При среднем объеме данных при F = 4 вероятность того, что дисперсии принадлежат одной выборке $P = 0{,}01$, а при F = 2 соответственно $P = 0{,}05$.

В табл. 3.6 приведен расчет критерия Фишера для двух месяцев со сходными средними значениями температур. Оценка показыва-

Таблица 3.6

Оценка вероятности принадлежности дисперсий одной генеральной совокупности

Дисперсия		F-критерий	p-значимость
Январь	Февраль		
4,006087	3,884722	1,063459	0,762513

ет, что с вероятностью 0,76 они принадлежат одной генеральной совокупности.

Дисперсия — важная физическая переменная, отражающая мощность системы или мощность воздействия на нее внешних сил.

На рис. 3.4 тем же методом бокс-плот показан масштаб различия средних квадратических и разброса температур в разные месяцы. Видно, что в зимние месяцы амплитуда варьирования температур в многолетнем ряду наблюдений существенно выше, чем в летние. Если это действительно так, то это означает, что колебания температур год от года зимой существенно выше. Следовательно, несмотря на общую стационарность зимы, действуют внешние факторы, порождающие большое разнообразие этой переменной. С другой стороны, для эколога эта информация подсказывает, где скорее всего следует искать факторы, ответственные за изменение, например, численности мышевидных грызунов. Это, конечно, не означает, что флуктуации зимних температур являются безусловным фактором. Но вполне логично полагать, что чем боль-

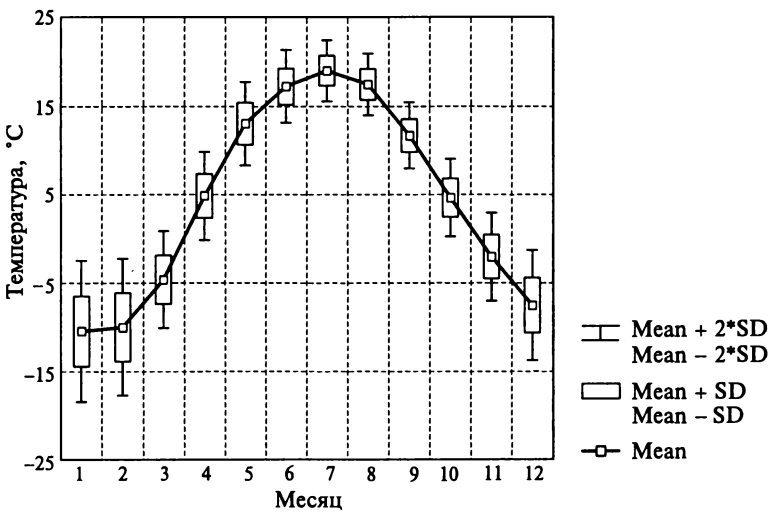


Рис. 3.4. Масштабы варьирования температур в различные месяцы

ше варьирует во времени некоторая переменная, тем скорее она может определять колебания потенциально зависящих от нее объектов и явлений.

На рис. 3.5 показаны значения F-критерия при попарном сравнении дисперсий всех 12 месяцев.

Большие значения F-критерия особенно характерны при сравнении дисперсий зимних и летних месяцев. Но и дисперсии летних месяцев различаются вполне достоверно ($F > 2$). Вместе с тем дисперсия в марте недостоверно отличается от дисперсии в апреле и мае, дисперсия апреля от дисперсии мая, дисперсия мая от дисперсии июня. Дисперсия в июне недостоверно отличается от дисперсии в августе, сентябре и октябре, дисперсия в июле — от дисперсии в августе и сентябре, дисперсия в августе — от дисперсии в сентябре, дисперсия сентября — от дисперсии октября, дисперсия в октябре — от дисперсии ноября. В то же время дисперсия температур в ноябре недостоверно отличается от дисперсии в марте, апреле и мае, а дисперсия в декабре достоверно отличается от дисперсии всех месяцев. Таким образом, на основе сравнения дисперсий можно констатировать, что в соседние месяцы чаще всего дисперсии подобны и в то же время циклическое изменение дисперсии от января—февраля к минимуму в июле вполне достоверно. Общий вывод сводится к следующему: летом варьирования тем-

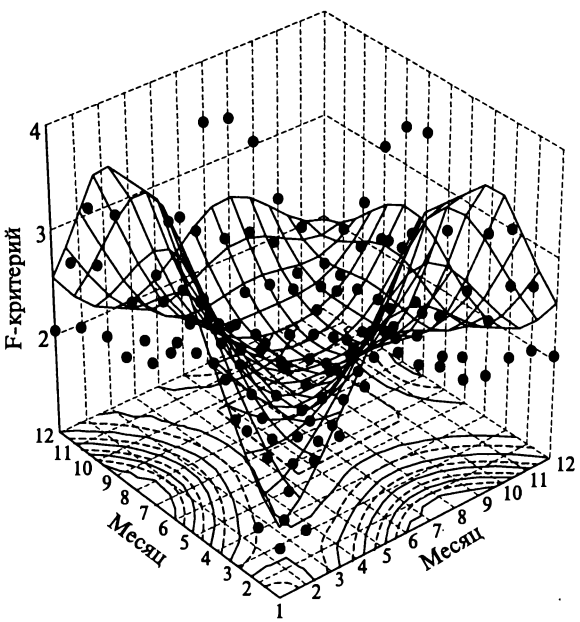


Рис. 3.5. Значения F-критерия при попарном сравнении дисперсий среднемесячных температур

пературы год от года различаются меньше, чем зимой. Этот вывод можно считать вполне достоверным. Вместе с тем получить какие-либо объяснения из этого индуктивно выведенного правила на основе варьирования самих температур невозможно.

Обратимся к анализу варьирования атмосферных осадков по измерениям на той же метеостанции. Прежде, чем осуществлять количественный анализ, проведем визуальную оценку распределений.

Для этого построим бокс-плоты, но не относительно выборочного среднего, а относительно медианы, с отображением границ квантилей, максимума и минимума, укладывающихся в рамки нормального распределения и точками, выходящими за его границы. Сезонный ход осадков очевиден, но очевидна также асимметрия распределения (рис. 3.6).

Для визуальной проверки гипотезы нормальности воспользуемся визуальными тестами «график нормальных вероятностей» [Normal Probability Plot (NPP)]. По оси абсцисс на графике откладывают наблюдаемые значения случайных событий от меньшего к большему, а по оси ординат — их расчетные стандартизованные значения, которые они имели бы при строго нормальном распределении. На рис. 3.7 приведены графики NPP для средних температур в январе, нормальность распределения которых доказана, и распределения осадков в том же месяце.

Идеально нормальным будет распределение, для которого точки, соответствующие наблюдаемым значениям, лежат точно на линии теоретической зависимости «стандартизированные расчетные значения при гипотезе нормальности распределения — наблюдаемые значения». Если распределение отличается от нормального, то обычно точки размещаются на графике по выпуклой тра-

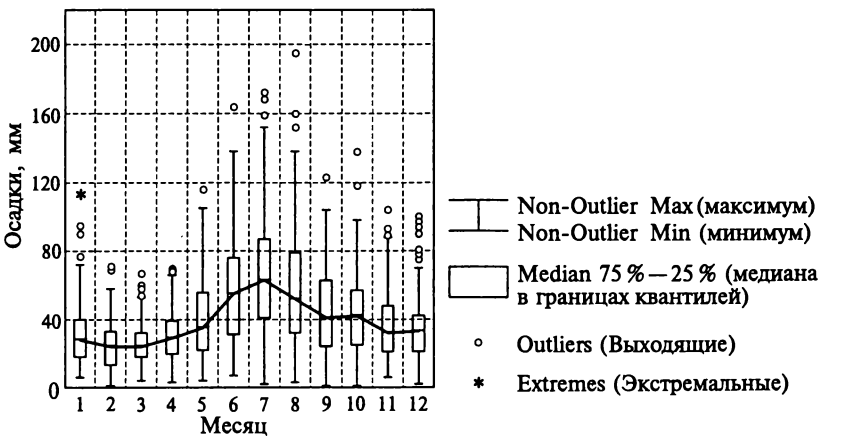
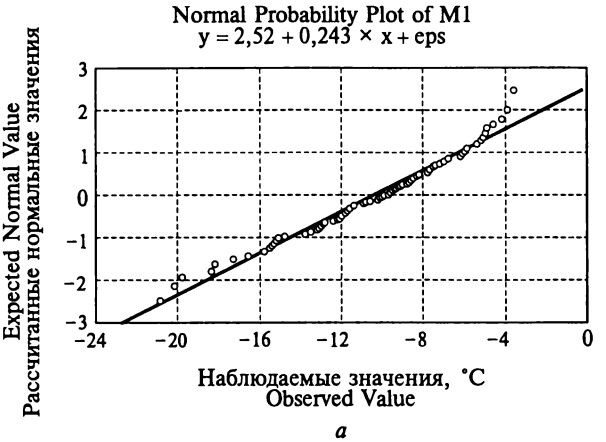
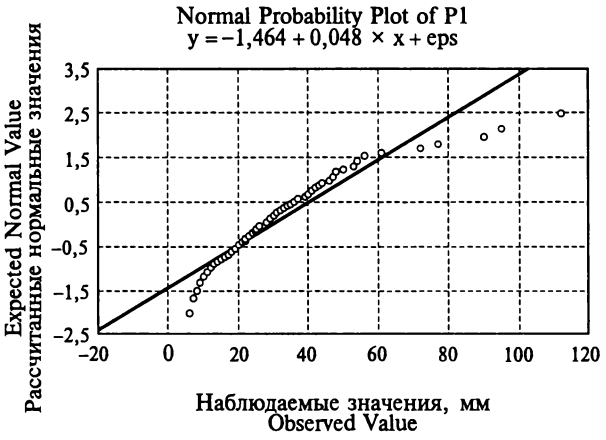


Рис. 3.6. Варьирование сезонного хода осадков с 1988 по 1989 г. (по данным метеостанции «Рязань»)



а



б

Рис. 3.7. Визуальная проверка гипотезы нормальности (вероятность нормального распределения, по данным метеостанции «Рязань»):

а — январская среднемесячная температура; *б* — среднее количество осадков за январь

ектории. Из рис. 3.7 следует, что, скорее всего, распределение осадков в январе не является нормальным. Этот визуальный тест в дальнейшем будет очень полезен при решении задач многомерной статистики. В данном случае констатируем наши предположения о существенной ненормальности распределения осадков.

Если распределения не носят нормальный характер, а подчиняются или логнормальному, или гамма-распределению, то это, также как и в случае с нормальным, позволяет принять гипотезу стационарности климатических условий на протяжении всего периода наблюдений, но при иной природе случайного процесса.

Какому теоретическому распределению ближе выборки априори неизвестно, поэтому необходимо одновременно тестировать нормальное, логнормальное и гамма (γ)-распределения (табл. 3.7).

Из табл. 3.7 с полной очевидностью следует, что распределение выпадения месячных осадков подчиняется закону гамма-распределения. При этом почти во все месяцы можно считать, что климатические условия по этой переменной были стационарны. Гипотезу о стационарности можно отвергнуть только для условий апреля и октября. Хотя и здесь вероятность ее истинности довольно велика. В последующем верифицируем гипотезу стационарности на основе более мощных тестов, а в данном случае будем констатировать необходимость преобразований распределений к нормальному виду. Для этого рассмотрим корни квадратные из сумм осадков. На примере распределения осадков в январе покажем эффект такого преобразования (табл. 3.8).

Из графиков соответствующих распределений (рис. 3.8) хорошо виден эффект преобразования данных при извлечении из них квадратного корня. Более того, преобразованные данные более близки к нормальному распределению, чем исходные к γ-распределению. Из табл. 3.8 следует, что γ-распределение имеет очень большую, безусловно статистически значимую косость и эксцесс.

Далее рассмотрим, какое практическое значение имеет доказанный факт γ-распределения в выпадении осадков на примере января. Снег, скапливающийся в зимнее время года на улицах, — большая проблема для городских служб. Допустим, известно, сколько нужно единиц снегоуборочной техники для того, чтобы выпавший снег не причинял значительных убытков экономике города или региона.

Если ориентироваться на среднее многолетнее количество осадков, то также как и в задаче, связанной с системой массового обслуживания, можно не сомневаться, что всегда будет проблема проезда транспортных средств и, соответственно, заторов. Значит, техники должно быть больше. Но сколько? Допустим, известно, сколько нужно техники, чтобы обеспечить очистку улиц от снега при среднемноголетнем выпадении осадков. (Предлагаем читателям самостоятельно найти способ рассчитать ожидаемое количество осадков за один день, если допустить, что в суточном ходе выпадения осадков сохраняются параметры того же гамма-распределения и нормального распределения преобразованных данных.)

Будем считать, что необходимое число единиц техники пропорционально сумме осадков за месяц (вообще говоря, ее нужно несколько больше, так как необходимо учитывать вероятность отказов, а чем больше единиц техники в целом, тем больше будет и число отказов. Вероятность остается неизменной, а число меняется). Будем считать средний срок службы техники, равным 10 годам, естественно предполагая, что в течение этого времени про-

Таблица 3.7

Проверка гипотез о соответствии выборочных распределений месячных сумм осадков за 100 лет для трех моделей распределения случайных событий

Месяц	Распределение								
	нормальное			логнормальное			гамма		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Январь	14,67034	5	0,0118809	9,521495	3	0,0231171	3,075650	4	0,5452549
Февраль	18,04772	7	0,0117729	18,20525	7	0,0110932	8,544972	7	0,2870329
Март	15,12648	7	0,0344413	7,619812	6	0,2673351	7,229478	6	0,3001778
Апрель	16,74362	8	0,0329241	28,31819	8	0,0004190	14,33404	8	0,0735077
Май	17,73023	7	0,0132658	10,92245	6	0,0908404	1,491526	5	0,9140415
Июнь	18,59941	8	0,0171777	22,14930	7	0,0023998	10,70021	7	0,1522859
Июль	6,394936	4	0,1715605	25,05462	5	0,0001365	3,746373	5	0,5864848
Август	19,62340	5	0,0014735	10,33622	4	0,0351492	3,882631	4	0,4221390
Сентябрь	9,379534	6	0,1533712	13,50063	6	0,0357668	8,611118	6	0,1966972
Октябрь	4,358861	6	0,6282376	22,02033	7	0,0025259	10,86752	6	0,0925957
Ноябрь	16,41345	5	0,0057659	1,008080	3	0,7992969	2,346050	4	0,6724004
Декабрь	13,76208	5	0,0172096	7,184741	5	0,2072925	4,664924	5	0,4581406

Таблица 3.8

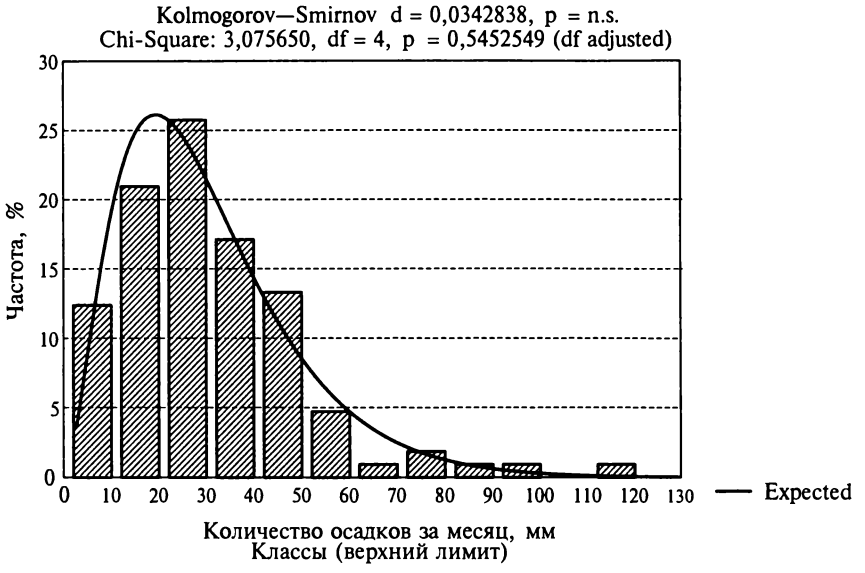
Оценки параметров распределения для исходных и нормализованных данных (суммы осадков в январе)

Параметр	Данные	
	исходные	нормализованные
Valid N	105,000	105,000
MEAN	30,341	5,261
Confid.–95 %	26,646	4,944
Confid.+95 %	34,036	5,578
Median	28,000	5,292
Minimum	6,000	2,449
Maximum	112,000	10,583
Lower quartil	18,000	4,243
Upper quartil	40,000	6,325
Range	106,000	8,134
Quartile Range	22,000	2,082
Variance	364,538	2,690
STD.DEV.	19,093	1,640
Standart error of mean	1,863	0,160
Skewness	1,574	0,535
STD.ERR.SK	0,236	0,236
Kurtosis	3,835	0,596
STD.ERR.K	0,467	0,467

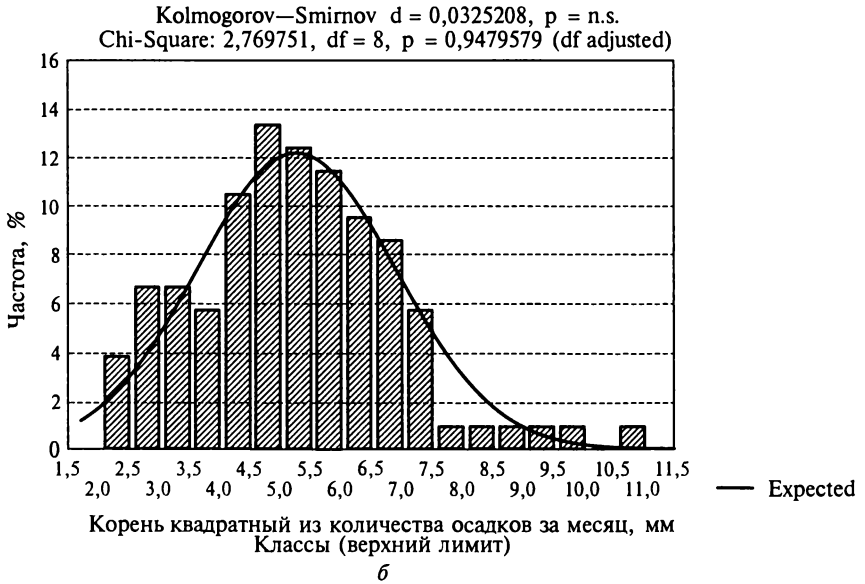
исходит ее некоторое обновление, поддерживающее ее в работоспособном состоянии. Таким образом, необходимо определить вероятность экстремально снежного января за расчетный период и по отношению к этому экстремуму определить необходимый объем техники. Эту задачу можно решить по таблицам нормального распределения для исходных и преобразованных данных. В современных статистических пакетах программ существуют разделы, где эта задача решается программными средствами.

Необходимо задать параметры распределения, приведенные в таблице, и вероятность $p = 0{,}1$. Для сопоставления приведем расчеты и для $p = 0{,}01$ (табл. 3.9).

Общий вывод вполне очевиден. Количество снегоуборочной техники должно быть ориентировано на сумму январских осадков около 50—60 мм. И оценки по нормальному распределению и по гамма-распределению (оценка по нормальному распределению с преобразованными данными с последующим возведением полученного значения в квадрат) дают практически одинаковые результаты. Оценка для гамма-распределения становится выше, чем по нор-



а



б

Рис. 3.8. Преобразование γ-распределения в нормальное:

а — исходные данные — γ-распределение; *б* — преобразованные данные — нормальное распределение

Таблица 3.9

Оценка возможных экстремальных сумм осадков в январе

Вид данных	Выборочное среднее	Среднее квадратическое отклонение	Возможный максимум осадков			
			за 10 лет		за 100 лет	
			нормальное распределение	гамма-распределение	нормальное распределение	гамма-распределение
Исходные	30,341	19,093	54,69	—	74,56	—
Преобразованные	5,261	1,640	7,36	54,18	9,075	82,35

мальному, для интервала в 100 лет. Очевидно, что чем больше косость гамма-распределения (параметр наклона), тем больше будут расходиться оценки вероятности редких событий, полученные по нормальному и гамма-распределению.

Конечно, полученная оценка не может исключить в течение десяти лет реализации редких событий, имеющих, допустим, вероятность 0,001 (105 мм осадков). Кстати в течение 100 лет наблюдений такой снежный год был один раз. Очевидно, чтобы противостоять таким редким событиям, нужно иметь техники в 2—3 раза больше нормы. Скорее всего затраты на ее содержание приведут к бóльшим экономическим потерям, чем возможная реализация очень редкого события.

Проблема оценки рисков от экстремальных явлений природы — особая задача, для решения которой используются специальные модели экстремальных событий, но существо дела приближенно соответствует описанной выше процедуре.

Важным результатом этого раздела анализа данных является демонстрация метода их нормализации. Если нормализация удается, то с высокой надежностью можно применять параметрические методы анализа, опирающиеся только на информацию, содержащуюся в двух первых параметрах распределений (среднего и дисперсии).

Если данные не поддаются нормализации, то для корректной проверки гипотез необходимо использовать непараметрические критерии.

3.4. Непараметрические критерии проверки гипотез

Рассмотрим смысл и применение непараметрических критериев на реальном примере. Учитывая, что настоящее пособие предназначено как для экологов, так и для географов, обратимся к существенно иному предмету.

В тропических сезонно-влажных лесах Южного Вьетнама ведущий сотрудник Института проблем экологии и эволюции им. А. Н. Северцова РАН Г. В. Кузнецов проводил учеты крыс.

Крыс отлавливали в течение 10 дней 100 ловушками, размещенными по одной прямой линии (схема трансекта) с шагом 25 м. За этот период было поймано 76 особей пяти видов крыс и один вид мышей: Maxomys surifer (Miller, 1900), Leopoldamys sabanus (Thomas, 1887), Berylmys berdmorei (Blyth, 1851), Berylmys bowersi (Anderson, 1879), Niviventer fulvescens (Gray, 1847), Mus pahari (Thomas, 1916). Три последних вида представлены всего одной особью.

Как и в первом случае попытаемся определить, соответствует ли их распределение какой-либо стандартной дискретной модели. В данном случае элементом исходной системы принимается каждая конкретная ловушка и конкретный вид крысы, с которыми связаны дискретные случайные события: не поймано ни одной крысы данного вида, пойманы 1, 2, 3 крысы. Тест показал, что во всех случаях в соответствии с критерием принимается распределение редких событий Пуассона (рис. 3.9). Заметим, что χ^2-тест применим только при четырех классах численности. При трех классах численности его расчет невозможен и гипотеза принимается только на основе критерия Колмогорова — Смирнова. Таким образом, на этом уровне анализа для четырех видов крыс рассматриваемая территория может быть признана однородной.

В табл. 3.10 приведены результаты проверки гипотезы о принадлежности распределений численности разных видов к одной или точнее к подобным генеральным совокупностям по параметрическому тесту, построенному на основе распределения Стьюдента. Применение этого теста к распределению редких дискретных событий заведомо некорректно. Однако все-таки полезно сравнить результаты этого тестирования с тестированием на основе непараметрических критериев.

Простейшим непараметрическим тестом для медианы является ***критерий знаков*** (Signal Test). Рассмотрим его логико-математические основания. Пусть существует последовательность из N независимых испытаний, в которых может быть два исхода (+) или (–). Общее количество положительных исходов, очевидно, есть случайная величина, подчиняющаяся биномиальному распределению.

Пусть имеются два сравниваемых ряда наблюдений. Упорядочим эти ряды от максимального значения к минимальному. Будем считать положительным исходом все пары, для которых значение первого ряда больше второго. Равные значения исключим из выборки. Если сравниваемые выборки принадлежат одной генеральной совокупности, то математическое ожидание вероятности сигнала — знака (+) будет очевидно $0{,}5N$ (при условии принадлежности к одной генеральной совокупности вероятность положитель-

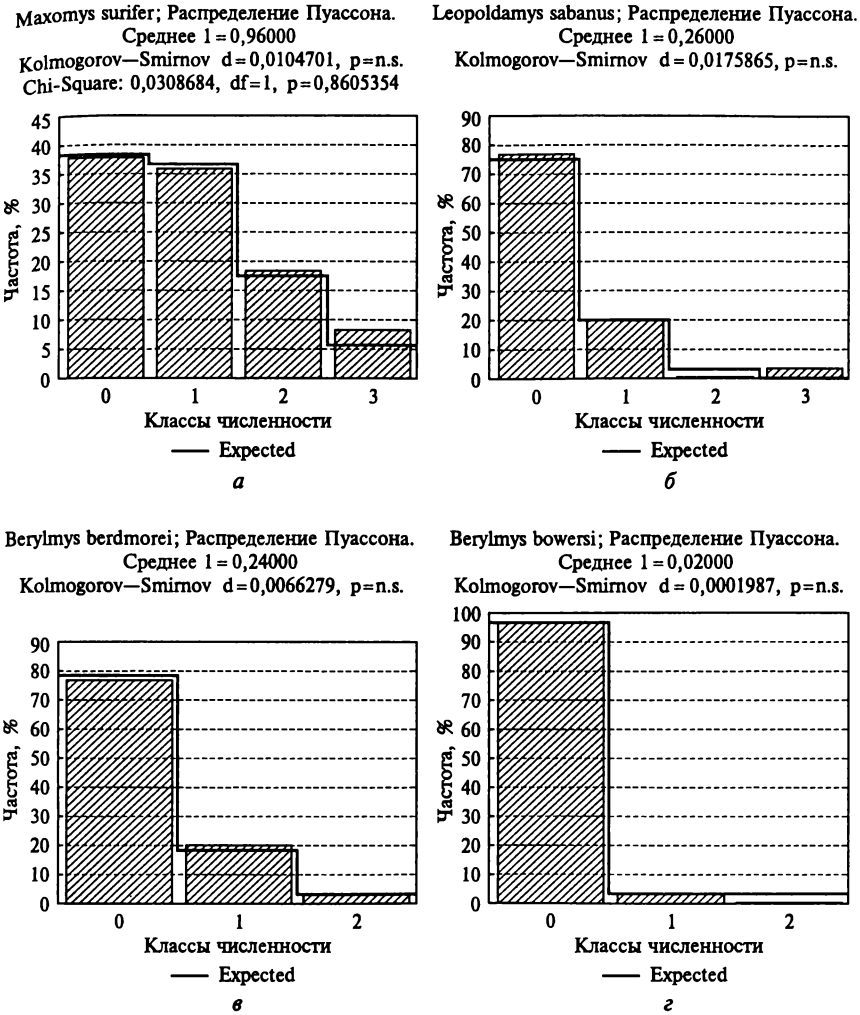


Рис. 3.9. Распределение классов численности крыс (*а* — *г*) по данным учета в сезонно-влажных тропических лесах Южного Вьетнама (по материалам Г. В. Кузнецова)

ных исходов равна 0,5). Теперь легко понять суть критерия. Если наблюдаемая вероятность $p(+) = n(+)/N < 0{,}01$, то вероятность принадлежности двух выборок одной генеральной совокупности мала и может быть отброшена в соответствии с этим критерием. Множество исходов $p(+)$ определяет критическую область.

Критерий знаков весьма удобен в расчетах и при визуальной оценке различия небольших выборок. Если, например, учитываются почвенные беспозвоночные методом раскопок, то объемы выборок неизбежно очень малы (4 — 8 проб). Допустим необходи-

Таблица 3.10

Проверка гипотезы о принадлежности видов одной генеральной совокупности распределений численности по t-тесту (MAX — Maxomys surifer; LEO — Leopoldamys sabanus; BER-Berylmys berdmorei; BERYL — Berylmys bowersi)

Вид крыс	Численность	MAX	LEO	BER	BERYL
MAX	0,96		4,31421	5,03597	7,110294
LEO	0,26	0,000077		0,20651	2,871061
BER	0,24	0,000007	0,837249		3,070191
BERYL	0,02	0,000000	0,006026	0,003484	

Примечание. Выше диагонали — t-тест, ниже диагонали — p-уровень значимости.

мо оперативно определить достоверно ли отличаются численности двух видов дождевых червей в двух сравниваемых выборках. Если во всех четырех пробах одной выборки численность одного вида выше, чем второго, то они могут быть отнесены к одной генеральной совокупности с вероятностью меньше 0,1; если численности различаются в трех пробах, то вероятность принадлежности подобным генеральным совокупностям двух разных видов — 0,25.

Критерий знаков демонстрирует наиболее простую и прозрачную модель построения непараметрического критерия. В конечном итоге большинство непараметрических критериев опираются на дискретные распределения и связанные с ними комбинаторные модели.

Чтобы закрепить понимание сущности непараметрических критериев и вообще непараметрических методов статистики, рассмотрим конструкцию *критерия Вилкоксона* (Wilcoxon Test).

Критерий предназначен для проверки гипотезы об однородности двух выборок. Предполагается, что элементы двух выборок взаимно независимы и подчиняются каким-то непрерывным распределениям. Основная гипотеза предполагает, что выборки извлечены из одной генеральной совокупности, и функции распределения их случайных величин одинаковы.

Пусть выборка x имеет объем n и выборка y объем m, а общий объем выборки равен $N = n + m$.

Расположим все наблюдения в один ряд от большего к меньшему и пронумеруем их:

x	x	y	x	x	x	y	y	y	...	x	y	y
1	2	3	4	5	6	7	8	9		$N-2$	$N-1$	N

Если события выборки x и выборки y принадлежат одной генеральной совокупности, то их положение в ряду, т.е. номер ранга, будет определяться одной из возможных перестановок из $N!$ и математическое ожидание сумм ранговых чисел будет и для ряда x, и для ряда y одинаково.

Соответственно можно построить комбинаторную модель, порождающую распределение сумм ранговых чисел при чисто случайном процессе. Здесь не так важно, какова эта модель. Важно то, что ее можно построить, а на ее основе ввести критическое множество и соответствующий критерий. Поэтому будем стремиться понять логические основания метода, а все вычислительные процедуры оставим на совести разработчиков соответствующих пакетов программ. Так, в частности, отметим, что критерий знаков использует информацию только о различии чисел, а критерий Вилкоксона использует информацию о величине числа. Чем больше информации о явлении используется в критерии, тем соответственно больше его мощность и выше надежность выводов, полученных на его основе.

Продемонстрируем проверку гипотезы принадлежности выборок по учетам численности к подобным генеральным совокупностям (табл. 3.11).

Сравнение результатов оценивания показывает, что параметрический тест во всех случаях дает существенно меньшую вероятность нулевой гипотезы, чем непараметрические тесты (см. табл. 3.10).

Таблица 3.11

Проверка гипотезы о принадлежности видов одной генеральной совокупности распределений численности по непараметрическим критериям

Вид крыс	MAX	LEO	BER	BERYL
MAX		0,000499	0,000126	0,000000
LEO	0,000353		1,000000	0,009375
BER	0,000063	0,849818		0,009375
BERYL	0,000001	0,009637	0,009637	

Примечание. Р-уровень значимости: выше диагонали — критерий знаков (Signal Test), ниже диагонали — критерий Вилкоксона (Wilcoxon Test).

В данном случае материал таков, что качественных ошибок при использовании параметрического теста проверки гипотезы принадлежности к одной выборке в сопоставлении с непараметрическим нет и численности всех видов, кроме одной пары, различаются по всем критериям вполне значимо.

Однако легко представить переходные ситуации, при которых по параметрическому тесту выборки значимо различаются, а по непараметрическим — нет. Что касается двух непараметрических тестов, то оценки вероятности принадлежности выборок одной генеральной совокупности различаются слабо. Формально тест Вилкоксона более мощный и поэтому более точен, чем критерий знаков.

В переходных ситуациях правильнее использовать несколько непараметрических тестов и в некоторых случаях принимать оценку, противоречащую желаемому. Дело в том, что в зависимости от целей исследования может быть привлекательна истинность нулевой гипотезы о подобии, а в другом случае, напротив, о различии. Иногда от этих оценок зависит весь результат исследования, на которое затрачено весьма много времени. Если ситуация окажется действительно принципиальной, то можно рекомендовать исследователю более углубленно изучить правила принятия решений и оценивать ошибки как второго, так и первого рода.

3.5. Одномерный дисперсионный анализ

Следующим важным методом одномерного анализа является одномерный дисперсионный анализ (Analysis of Variance, ANOVA).

Исходная система определяется элементарным случайным событием с дополнительным указанием его принадлежности к одному из r-классов или групп, априори определенных в исходной модели.

Например, измеряется влажность почвы в каждом генетическом горизонте. Элемент — проба с определенной влажностью в априори заданном классе — горизонте. Определяется численность какого-либо вида в каждом типе местообитания. Тип местообитания — априори заданный класс. Целью анализа является проверка гипотезы о принадлежности выборок в каждом классе по их математическим ожиданиям одной общей генеральной совокупности. Анализ строится на основе сопоставления дисперсий выборок, с учетом принадлежности их к классам, с общей дисперсий всей совокупности измерений. Использование при оценивании только одного параметра (дисперсии) требует, чтобы распределения не сильно отличались от нормальных. Следовательно, перед использованием дисперсионного анализа требуется обязательная нормализация выборки.

Исторически возникновение дисперсионного анализа связано с экспериментами, проводимыми в сельском хозяйстве, по ис-

пользованию различных доз удобрений. Требовалось доказать существование «отклика» урожая на точно определенную дозу удобрений в сравнении с контролем.

Одновариантный дисперсионный анализ широко используется в экологических исследованиях и часто является их завершающим шагом. С другой стороны, он входит составной частью в некоторые важные методы анализа данных. Все это делает целесообразным его достаточно подробное рассмотрение (табл. 3.12).

В табл. 3.12 в каждом классе (i) n_{i*} элементарных наблюдений со средними $\bar{X}_{i*}$ и дисперсиями σ^2_{i*} — средняя внутриклассовая выборочная дисперсия.

Средняя дисперсия внутри классов (внутриклассовое, внутригрупповое рассеивание) определяется по формуле

$$\sigma^2_G = \frac{1}{n-r}\sum_{i=1}^{r}(n_{i*}-1)\sigma^2_{i*},$$

где $\sum_{i=1}^{r}\sigma^2_{i*}$ — сумма дисперсий каждого класса; n_{i*} — число степеней свободы; $(n - r)$ — общее число степеней свободы (общий объем выборки минус число классов).

Дисперсию между выборками (межклассовое, межгрупповое рассеивание) находят по формуле

$$\sigma^2_{INT} = \frac{1}{r-1}\sum_{i=1}^{r}n_{i*}(\bar{X}_{i*}-\bar{X})^2,$$

где $\sum_{i=1}^{r}(\bar{X}_{i*}-\bar{X})^2$ — сумма квадратов разности среднего выборки в i классах и среднего для всей выборки; n_{i*} — объем выборки в каждом классе; r — число классов.

Таблица 3.12

Дисперсионный анализ. Схема размещения элементов по классам и основные параметры

Класс 1	Класс 2	...	Класс r	Параметры выборки в целом
x_{11}	x_{12}	...	x_{1r}	
x_{12}	x_{22}	...	x_{2r}	
⋮	⋮	...		
x_{1k}	x_{2k}	...	x_{kr}	
n_{1*}	n_{2*}	...	n_{r*}	n
$\bar{X}_{1*}$	$\bar{X}_{2*}$	...	$\bar{X}_{k*}$	$\bar{X}$
σ^2_{1*}	σ^2_{2*}	...	σ^2_{k*}	σ^2

Тогда формула для определения общей дисперсии для всей выборки будет следующей:

$$\sigma^2 = \sigma_G^2 + \sigma_{INT}^2 = \frac{1}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{k}(x_{ij} - \bar{X})^2.$$

Первая сумма в этом выражении, очевидно, есть сумма по всем строкам внутри столбцов и по всем столбцам квадрата отклонения значения любого элемента от среднего по всей выборке, т. е. стандартная формула измерения дисперсии. То, что она действительно равна сумме внутригрупповой и межгрупповой дисперсии вытекает из того, что дисперсия независимых выборок есть сумма их дисперсий. Выборки же по столбцам и по последней строке перпендикулярны друг другу, т. е. независимы.

Однако справедливость этого соотношения легко проверить в мыслимом эксперименте, представив два очевидных крайних варианта из множества возможных:

1) все измерения в одном классе абсолютно тождественны (какое-либо варьирование внутри класса вообще отсутствует, $\sigma_G^2 = 0$), но различаются по классам. Очевидно, что в этом случае общая дисперсия будет равна межклассовой;

2) средние во всех классах абсолютно тождественны $\sigma_{INT}^2 = 0$, но в каждом классе существует определенное рассеивание. В этом варианте общая дисперсия будет равна межгрупповой.

Вполне понятно, что реальность лежит где-то между этими двумя крайними ситуациями.

Проверим нулевую гипотезу о том, что вся выборка принадлежит одной генеральной совокупности.

Гипотеза принимается с вероятностью 1, когда $F = \frac{\sigma_{INT}^2}{\sigma_G^2} = 1$, т. е. все варьирование является внутригрупповым, а межгрупповое отсутствует. Гипотеза обычно считается неверной, и средние в разных классах или хотя бы между двумя классами из всех различны при $F > 2$ или $F > 4$. Если гипотеза принадлежности выборок в разных классах одной генеральной совокупности неверна, то логично полагать, что рассматриваемое явление зависит от принадлежности к классу или иначе между выборочными средними и классами существует зависимость.

Рассмотрим все операции на конкретном примере.

На стационаре Института проблем экологии и эволюции им. А. Н. Северцова, расположенном на расстоянии 40 км южнее Москвы и 5 км южнее г. Новотроицка на моренной слаборасчаленной возвышенности, перекрытой покровными суглинками с дерново-подзолистыми почвами, был заложен трансект длиной около 3 км с регулярным шагом измерения относительной влажности, кислотности, обменных оснований Ca, Mg, Na, K, P на

глубинах 5, 10, 20, 30, 40 см. Измерения проводились в августе после шести дней без осадков. Это позволяло надеяться на то, что разовые наблюдения в какой-то степени могут отражать более или менее стационарные отношения в почве.

В одновариантном дисперсионном анализе измерения по каждой переменной рассматриваются как самостоятельная система, элементом которой является образец со значением переменной и принадлежностью к определенной глубине (группе).

В итоге получено, что распределения, за исключением кислотности, близки к логнормальным, в связи с чем в анализе используются логарифмированные данные (табл. 3.13).

Таблица результатов рассчитана в пакете Statistica. В других пакетах схема выдачи результатов примерно такая же. Отметим, что внутригрупповая дисперсия здесь обозначается как ошибка. Тем самым подразумевается, что в идеале все разнообразие должно описываться межгрупповой дисперсией. В данном случае F-критерий очень высокий и влияние глубины отбора пробы на влажность почвы абсолютно достоверно. Конечно, сам по себе этот результат достаточно очевиден. Однако в рамках анализа можно получить важные уточнения.

Так, в программе обычно предлагаются тесты, несколько поразному использующие информацию, содержащуюся во внутригрупповой дисперсии при формировании критической области и дающие несколько различающиеся результаты в переходных ситуациях. В табл. 3.14 приведен LSD-тест, показывающий, что все горизонты по логарифмированным значениям влажности попарно различны, кроме горизонтов на глубине 30 и 40 см.

Таблица 3.13

Одновариантный дисперсионный анализ для относительной влажности почвы по слоям (Analysis of Variance)

Переменная	Сумма квадратов межгрупповая	Число степеней свободы	Межгрупповая дисперсия	Сумма квадратов внутригрупповая	Число степеней свободы	Внутригрупповая дисперсия	F-критерий	Уровень значимости
	SS Effect	df	MS Effect	SS Error	df	MS Error	F	p
Влажность почвы	68,56259	4	17,14065	41,18662	600	0,068644	249,7022	0,00

Таблица 3.14

LSD-тест попарный уровень значимости между всеми парами (M5 — средняя влажность на глубине 5 см)

Переменная	Переменная				
	M5 = 3,9417	M10 = 3,6415	M20 = 3,3069	M30 = 3,0950	M40 = 3,0629
G_1:1 {M5}		0,000000	0,000000	0,000000	0,000000
G_2:2 {M10}	0,000000		0,000000	0,000000	0,000000
G_3:3 {M20}	0,000000	0,000000		0,000000	0,000000
G_4:4 {M30}	0,000000	0,000000	0,000000		0,341386
G_5:5 {M40}	0,000000	0,000000	0,000000	0,341386	

LSD-тест (Post hoc Comparisons-LSD Test or Planned Comparison) эквивалентен t-критерию для независимых или зависимых выборок. Это приводит к повышенной чувствительности к объемам выборок в различных классах. В отличие от него тест «Post hoc Comparisons — Newman — Keuls test & Critical Ranges» строится на основе статистического диапазона распределения Стьюдента. При вычислении значения выстраиваются в порядке возрастания. Для каждой пары классов оценивается нулевая гипотеза различия средних рангов. Этот критерий хорошо учитывает число наблюдений в классах и менее чувствителен к их значительным различиям.

Вторая группа тестов проверяет гипотезу о тождественности (гомогенности) дисперсий в разных группах. Различия дисперсий подразумевают разную мощность процессов и, соответственно, могут иметь вполне определенный физический смысл. Не останавливаясь на алгебре теста, отметим лишь, что его логика сходна с логикой дисперсионного анализа, но не в отношении средних, а в отношении оценок масштабов варьирования самой дисперсии. Следует отметить также, что классы могут не различаться по среднему, но различаться по варьированию, что указывает на безусловное различие связанных с ними процессов (табл. 3.15).

Как следует из теста, варьирование в разных группах достоверно различается и гипотеза гомогенности не принимается.

На рис. 3.10 показаны параметры варьирования влажности почвы с глубиной, измеренные методом бокс-плот. Из графика, построенного относительно медианы, следует, что логарифмическое

Таблица 3.15

Тест на гомогенность дисперсий в группах (Brown-Forsythe Test of Homog. of Variances)

Переменная	Сумма квадратов межгрупповая	Число степеней свободы	Межгрупповая дисперсия	Сумма квадратов внутригрупповая	Число степеней свободы	Внутригрупповая дисперсия	F-критерий	Уровень значимости
	SS Effect	df	MS Effect	SS Error	df	MS Error	F	p
Влажность почвы	1,911742	4	0,477935	21,88568	600	0,036476	13,10269	0,000000

преобразование не смогло нормализовать выборку. Распределения асимметричны, и экстремальные значения, особенно на глубинах 5—10 см, далеко выходят за допустимые границы. Вместе с тем, в целом изменение влажности вполне естественное и постепенное уменьшение среднего квадратического отклонения с глубиной (выявленного тестом на гомогенность) хорошо демонстрирует источник воздействия. Очевидно, что варьирование влажности в пространстве в верхних горизонтах выше, чем в нижних. Но то, что величина варьирования действительно отражает мощность процесса и источник воздействия, полезно для содержательной трактовки дисперсии в любых других наблюдениях.

На рис. 3.11 показано варьирование влажности почвы в пространстве. Очевидно, что максимальные значения влажности связаны в основном с долинами ручьев, которые, конечно, никак не могут быть отнесены к генеральной совокупности склонов и водораздельных поверхностей. Генезис влажности здесь принципиально иной. При использовании параметрических методов анализа эти точки желательно не рассматривать. Иными словами, нужно исключать все точки, принципиально нарушающие нормальность распределений трансформированных данных.

Используя возможности дисперсионного анализа, проведем оценку отклика всех измеренных переменных на глубину отбора почвенного образца (табл. 3.16).

Из табл. 3.16 следует, что все переменные связаны с глубиной отбора образца, но максимальный отклик демонстрируют влаж-

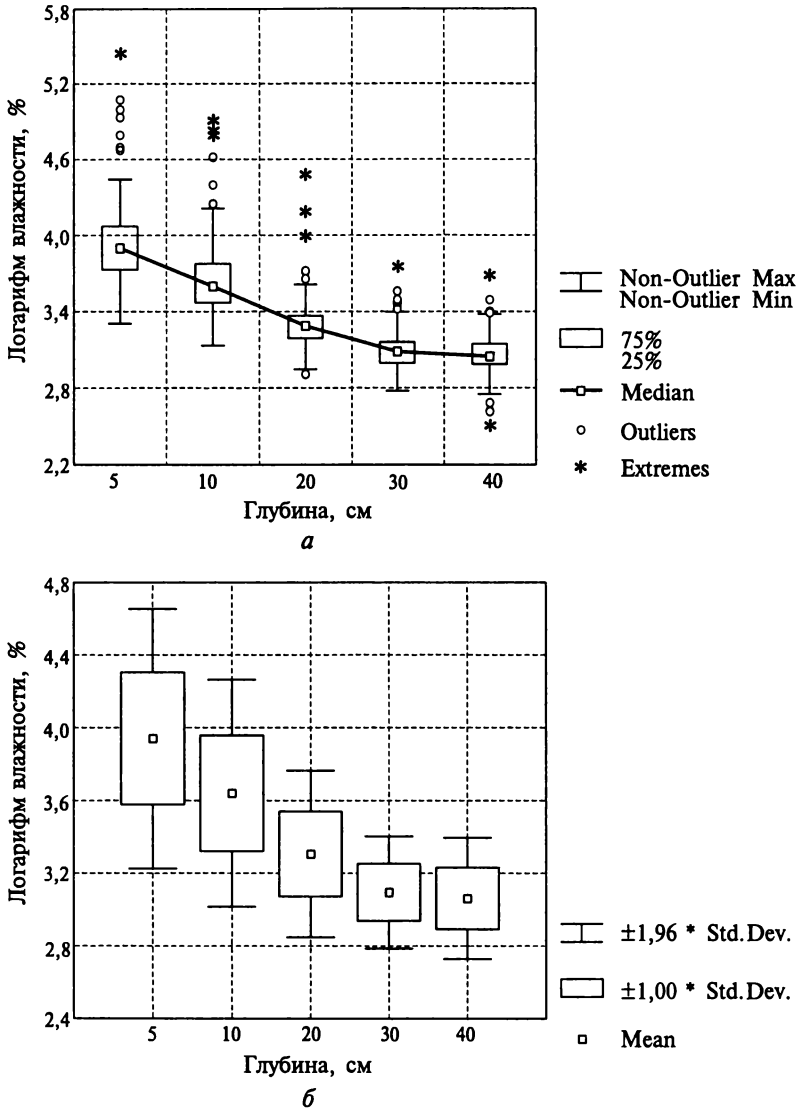


Рис. 3.10. Изменение параметров варьирования влажности почв с глубиной (*а*, *б*)

ность и концентрация магния, минимальный — показатель кислотности и концентрация натрия.

Рассмотрим тест различия средних для кислородного показателя pH (кислотности).

Из табл. 3.17 следует, что на очень высоком уровне значимости кислотность на глубине 5 см отличается от кислотности на глубине 30 и 40 см и существенно с меньшим уровнем значимости отлича-

Таблица 3.16

Одновариантный дисперсионный анализ для относительной влажности почвы, кислотности и обменных оснований по слоям (Analysis of Variance)

Переменная	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Влажность	68,5626	4	17,14065	41,1866	600	0,068644	249,7022	0,000000
pH	3,2576	4	0,81439	90,4512	600	0,150752	5,4022	0,000281
P	49,8894	4	12,47235	346,2579	600	0,577097	21,6122	0,000000
K	142,3380	4	35,58449	244,7099	600	0,407850	87,2490	0,000000
Na	1,7556	4	0,43890	16,2575	600	0,027096	16,1981	0,000000
Mg	151,9848	4	37,99620	182,5144	600	0,304191	124,9092	0,000000
Ca	103,5766	4	25,89416	166,3708	600	0,277285	93,3848	0,000000

ется от кислотности на глубине 10 и 20 см. Кислотность на глубинах 10—30 см остается фактически неизменной. Точно так же не отличаются значения кислотности на глубинах 30 и 40 см. В целом же кислотность с глубиной медленно нарастает.

Таблица 3.17

LSD-тест уровней значимости между всеми парами (M5 — средняя кислотность почвы на глубине 5 см)

Переменная	Переменная				
	M5 = 3,8887	M10 = 3,7633	M20 = 3,7752	M30 = 3,7081	M40 = 3,6737
G_1:1 {M5}		0,012281	0,023369	**0,000322**	**0,000019**
G_2:2 {M10}	0,012281		0,811645	0,269189	0,073206
G_3:3 {M20}	0,023369	0,811645		0,179339	0,042482
G_4:4 {M30}	**0,000322**	0,269189	0,179339		0,491254
G_5:5 {M40}	**0,000019**	0,073206	0,042482	0,491254	

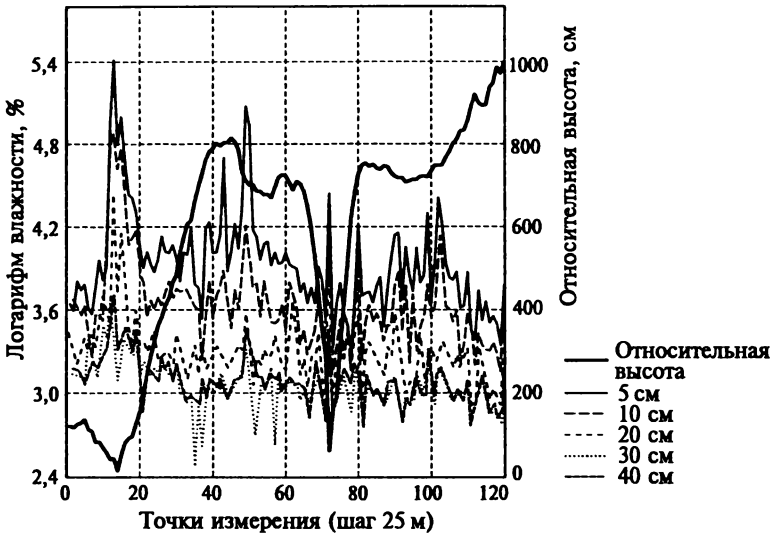


Рис. 3.11. Пространственное варьирование влажности дерново-подзолистой почвы на покровных суглинках (10—12 августа)

Тест на гомогенность (табл. 3.18) показывает, что наиболее резко различаются дисперсии по горизонтам для влажности и концентраций калия, кальция и магния. Натрий, концентрации кото-

Таблица 3.18

Тест на гомогенность дисперсий в группах для всех измеренных переменных (Brown-Forsythe Test of Homog. of Variance)

Переменная	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Влажность	1,911742	4	0,477935	21,8857	600	0,036476	13,10269	0,000000
pH	1,638099	4	0,409525	65,4230	600	0,109038	3,75579	0,004994
P	4,192200	4	1,048050	147,5546	600	0,245924	4,26168	0,002079
K	8,022136	4	2,005534	126,4497	600	0,210750	9,51620	0,000000
Na	0,088264	4	0,022066	11,7051	600	0,019509	1,13110	0,340781
Mg	5,717177	4	1,429294	75,9254	600	0,126542	11,29499	0,000000
Ca	2,940284	4	0,735071	68,3039	600	0,113840	6,45707	0,000043

рого в почве невелики, во всех горизонтах имеет сходную дисперсию.

Дисперсионный анализ при всей своей наглядности, как и любой параметрический метод статистики, весьма чувствителен к нормальности распределения. При прочих равных условиях дисперсия распределения, отличающегося от нормального, обычно больше дисперсии нормального распределения. В результате критическая область, на которой проверяется гипотеза, оказывается непропорционально большой и повышает вероятность истинности нулевой гипотезы о принадлежности данных в разных группах одной генеральной совокупности. Так как вклад во внутригрупповую дисперсию каждого класса существенно зависит от объема выборки, при больших различиях в их объемах оценки вообще могут быть сильно искажены.

Этих недостатков лишены непараметрический одновариантный анализ Краскала и медианный тест.

Метод Краскала (Краскал-тест) основывается на распределении рангов.

Ранги рассчитываются следующим образом:

• вся выборка упорядочивается от меньшего значения к большему, и всем одинаковым значениям переменной присваивается номер ранга $k + m$ (одиночные значения пропускаются):

X	[56	56	56]	[57	57	57	57]	[58	58]	60	[61	61	61	61	61]	70
Ранг	1	2	3	4	5	6	7	8	9		10	11	12	13	14	
Средний ранг		2			5,5			8,5					12			

• затем для каждого класса определяется сумма рангов как сумма принадлежащих им средних значений, определенных выше. При этом одно и то же значение «средней» может относиться к разным классам. Сумма средних значений рангов, очевидно, содержит информацию и о среднем, и о масштабах рассеивания в каждом классе (табл. 3.19).

Таблица 3.19

Суммы рангов для влажности

Горизонт (глубина)	Число наблюдений	Сумма рангов
1 (5 см)	121	61868,50
2 (10 см)	121	51460,50
3 (20 см)	121	34281,00
4 (30 см)	121	18745,00
5 (40 см)	121	16960,00

Максимальная сумма рангов наблюдается в первом горизонте, для которого характерна наибольшая влажность почвы, минимальная — в пятом (наименьшая влажность). Очевидно, что если перемешать значения переменной чисто случайно, то суммы средних значений рангов в разных классах различались бы очень мало.

Далее все на той же комбинаторной основе строится распределение H, отражающее варьирование по классам квадрата суммы рангов каждого класса, деленного на число наблюдений в классе n_i:

$$H = \sum_{i=1}^{r} \frac{\left(\sum_{i=1}^{n} R\right)_i^2}{n_i},$$

которое имеет χ^2-распределение с числом степеней свободы, равным числу классов минус единица ($r - 1$).

На основе этого критерия определяется критическая область.

Для рассматриваемого примера Краскал-тест (Kruskal-Wallis ANOVA by Ranks) есть: H (df = 4, N = 605) = 424,4900, p = 0,000 — как и в дисперсионном параметрическом анализе выборки в разных классах не принадлежат одной генеральной совокупности.

Используем этот тест для проверки нулевой гипотезы в отношении наиболее неопределенной переменной — кислотности (табл. 3.20).

Тест показывает (см. табл. 3.20), что максимальные суммы рангов — в первом горизонте, минимальные — в пятом. Суммы рангов отражают изменение кислотности по профилю, отличное от того, что было получено по средним значениям в дисперсионном анализе. Здесь наблюдается локальный минимум кислотности на глубине 10 см, с последующим повышением на глубине 20 см. В остальном же тенденции, выявленные в дисперсионном анализе, не меняются.

Таблица 3.20

Краскал-тест вариации для кислотности (Kruskal-Wallis ANOVA by Ranks) Kruskal-Wallis test: H (df = 4, N = 605) = 53,71896, p = 0,0000

Горизонт	Valid N	Sum of Ranks
1	121	44985,50
2	121	38298,00
3	121	40803,50
4	121	31858,00
5	121	27370,00

Используя ранговый анализ вариаций, целесообразно проверить нулевую гипотезу для выборок, относимых на основе параметрического дисперсионного анализа к одной генеральной совокупности: между кислотностью в горизонтах 10 и 15 см, 20 и 30 см, 30 и 40 см.

Оценка проводится по выборке для двух сравниваемых классов.

1. Для пары 10 и 15 см получаем суммы рангов соответственно 14398,50 и 15004,50, H (df = 1, N = 242) = 0,3098352, p = 0,5778 — выборки принадлежат одной генеральной совокупности.

2. Для пары 20 и 30 см сумма рангов 16872,50 и 12530,50; H (df = = 1, N = 242) = 15,91024, p = 0,0001 — выборки не принадлежат одной генеральной совокупности.

3. Для пары 30 и 40 см суммы рангов 15915,00 и 13488,00; H (df = = 1, N = 242) = 4,970437, p = 0,0258 — выборки скорее всего принадлежат разным генеральным совокупностям.

Итак, в двух случаях из трех то, что было неразличимо в дисперсионном параметрическом анализе достоверно различимо в непараметрическом.

Второй широко применяемый тест — **медианный** (Median Test). Его смысл легко понять из табл. 3.21, где он рассчитан для кислотности почвы.

Медиана делит выборку пополам и рассчитывается как общая для всех групп выборки. В первой информационной строчке приводится число наблюдений, имеющих значения меньше медианы; во второй — расчетное значение для каждого класса, исходя из гипотезы нормальности; в третьей — разность между наблюдаемым числом данных и расчетным. Соответственно, на глубине 5 см отклонение от расчетного отрицательное и максимальное по абсолютному значению. На глубинах 10 и 20 см оно отрицательное и одинаковое, а на глубинах 30 и 40 см — положительное. В следующих трех строчках аналогичные оценки приведены для значений, превышающих медиану. Отклонение наблюдаемого значения от расчетного подчиняется χ^2-распределению с (G – 1) степенями свободы (G — число групп).

Итоговая оценка сводится к следующему: Median Test (медиана по всей выборке) = 3,68, Chi-Square = 43,26252, df = 4, p = 0,0000 — отклик кислотности на глубину безусловно достоверен.

Итак, непараметрические критерии существенно уточняют оценки в области неопределенности, возникающей по параметрическому критерию. При этом Краскал-тест вариаций особенно эффективен для выборок с резко различающимися объемами наблюдений в разных классах.

Таким образом, в целом дисперсионный анализ дает весьма наглядные результаты, которые часто позволяют оценить качество материала, некоторые наиболее общие отношения и сформулировать рабочие гипотезы для более глубоких методов анализа. Так, из

Таблица 3.21

Переменная	Group 1 (pH — 5 см)	Group 2 (pH — 10 см)	Group 3 (pH — 20 см)	Group 4 (pH — 30 см)	Group 5 (pH — 40 см)	Число наблюдений
Медиана (<): наблюдаемая (obs)	43,000	54,000	54,000	76,000	87,000	314,000
расчетная (exp)	62,800	62,800	62,800	62,800	62,800	
Разность (obs – exp)	–19,800	–8,800	–8,800	13,200	24,200	
Медиана (>): наблюдаемая (obs)	78,000	67,000	67,000	45,000	34,000	291,000
расчетная (exp)	58,200	58,200	58,200	58,200	58,200	
Разность (obs – exp)	19,800	8,800	8,800	–13,200	–24,200	
Число наблюдений	121,000	121,000	121,000	121,000	121,000	605,000

рассмотренного примера следует, что результаты измерения влажности и катионов в почве позволяют определить почву как систему с входами и выходами (дисперсии на поверхности существенно выше, чем на глубине) и дают основания искать функциональные отношения, описывающие стационарные параметры вертикальной миграции.

Приведенные выше непараметрические критерии проверки нулевой гипотезы ни в коем случае не исчерпывают их разнообразия. Рассмотрены лишь наиболее употребимые и относительно легко объясняемые. Исследователь, хорошо владеющий непараметрическими методами оценивания, может выбрать наиболее мощные критерии, соответствующие особенностям его данных. На их основе он может доказать различия там, где это невозможно сделать по более слабым критериям. Этот факт, в частности, иногда позволяет утверждать, что на основе статистики можно доказать «все что угодно». Но это не более, чем слова. Хорошее владение методами и обоснованное их применение, наоборот, позволяет получить надежные результаты и не позволяет доказать недоказуемое.

Подведем итог методам одномерного анализа. В их рамках последовательно решаются задачи, организуемые обычно по следующей схеме.

1. Визуальный просмотр данных с использованием графиков бокс-плот, гистограмм, вероятностного графика для тестирования типа распределения.

2. Оценка параметров распределения с тестированием их на нормальность.

3. Обоснование наиболее приемлемой модели распределений.

4. Трансформация данных для приближения их распределения к нормальным.

5. Проверка гипотезы принадлежности выборки одной генеральной совокупности по средним (достоверность различия средних) на основе параметрических и непараметрических статистик.

6. Проведение дисперсионного параметрического и непараметрического анализа.

В итоге могут быть получены следующие содержательные результаты.

1. Ответ в первом приближении на вопрос: является ли система стационарной (равновесной) в пространстве-времени.

2. Образована она одной (подобными) или несколькими генеральными совокупностями с достоверно различными параметрами.

3. Существует ли в системе отклик на априорно определенные типы воздействий.

4. В каких условиях у системы наблюдается наибольшая мощность процессов (наибольшая амплитуда при смене состояний во времени и в пространстве).

5. Можно ли считать систему ориентированной (на уровне гипотезы).

На основе одномерного анализа решаются следующие практические задачи.

1. Определение количества вещества, или объектов, или измеренного свойства в целом для системы или для отдельных априорно выделенных классов с оценкой доверительных интервалов.

2. Оценка вероятности редких и экстремальных состояний системы.

Первая задача в конечном итоге позволяет рассчитать запасы ресурсов с определением доверительных интервалов, в которых с известной вероятностью будет находиться их реально существующая в природе величина.

Вторая задача связана с проблемами оценок рисков, возникающих при реализации редких событий. Как природа, так и социум в пределах конкретной территории подстраиваются чаще всего к наиболее обычным событиям, к нормальной области случайных флуктуаций внешних переменных. Поэтому любые экстремальные отклонения, как очень редкие события (даже не выходящие за рамки

случайного процесса) оказывают (по крайней мере на отдельные объекты) негативное воздействие, а в некоторых случаях воспринимаются как катастрофа.

Контрольные вопросы

1. Рассмотрите логическую модель принятия гипотезы.
2. Разберите, какова логика введения критической области критерия.
3. Почему специальные распределения определяют критические области?
4. Тщательно разберите логическую модель параметрического дисперсионного анализа.
5. В чем состоит различие параметрических и непараметрических критериев?

Глава 4
МНОГОМЕРНЫЙ АНАЛИЗ

Наиболее содержательные разделы статистического исследования связаны с многомерным анализом данных. В отличие от одномерного в многомерном анализе элемент системы описывается множеством переменных. Число этих переменных в наиболее сложных случаях может исчисляться сотнями, но обычно не превышает несколько первых десятков. Именно в рамках многомерного анализа осуществляется поиск ответов на вопросы: «как явления соотносятся друг с другом во времени и в пространстве», «какие реалистичные гипотезы можно сформулировать о механизмах этих отношений», «как проверить эти гипотезы». В этой главе дано общее представление о многомерном пространстве, приведена модель многомерного нормального непрерывного и многомерного дискретного распределения. На этой основе описываются методы измерения степени сопряженности между переменными и методы построения статистических моделей для ориентированных систем. Далее рассматриваются методы снижения размерности пространства и отображения переменных в системе независимых (ортогональных) факторов, обычно допускающих достаточно ясную физическую интерпретацию. В конечном итоге в результате многомерного анализа получаем статистическую модель, описывающую равновесные отношения между переменными и обладающую определенными прогностическими возможностями.

4.1. Представления о многомерном пространстве и размерности

В соответствии с системным подходом, говоря о пространстве, имеем в виду не реальность, а отражающую ее некоторые свойства модель. При этом различные модели могут отражать и различные свойства реальности. Пренебрежение этим фактом приводит к неоправданному распространению частных отношений на всю реальность, что может приводить к фатальным ошибкам при решении

чисто практических проблем отношения человека со средой. Иными словами, получив некоторый результат, нужно всегда стремиться определить условия его истинности и практической применимости.

Математическая энциклопедия, вышедшая в 1984 г. в издательстве «Советская энциклопедия», которую с полным основанием можно считать концентрированным изложением высшего достижения мышления человека в конце XX века, определяет пространство как «логически мыслимую форму (или структуру), служащую средой, в которой осуществляются другие формы и те или иные конструкции. Например, в элементарной геометрии плоскость или пространство служит средой, где строятся разнообразные фигуры. В большинстве случаев в пространстве фиксируются отношения, сходные по формальным свойствам с обычными пространственными отношениями (расстояния между точками, равенство и подобие фигур и др.) так, что о таких пространствах можно сказать, что они отражают логически мыслимые пространственно-подобные формы».

Это общее представление хорошо согласуется с некоторыми логическими конструкциями экологии. Например, экологическое пространство может мыслиться как отображение взаимоположения экологических ниш видов, где экологические ниши в свою очередь мыслятся как множества, образующие некоторые фигуры (например, многомерные окружности или эллипсы), к каждой из которых принадлежат особи соответствующих видов.

В таком представлении экологического пространства не рассматриваются ресурсы и условия среды. Ресурсы и условия среды можно определить в рамках самостоятельного пространства, в котором каждый ресурс будет находиться в некотором отношении к любому другому ресурсу. Например, приход солнечной радиации и поле температур будут ограничивать в логически мыслимом пространстве некоторую область возможных сочетаний их значений.

Экологическое пространство и пространство среды можно погрузить в некоторое общее пространство, в котором будут определяться отношения между нишами и сочетаниями состояний ресурсов и условий среды. В этом варианте обычно рассматриваются сочетания различных состояний видовых популяций с состояниями условий и ресурсов среды.

В физической географии, в частности в ландшафтоведении, пространство в неявном виде задается в шкале отношений между выделяемыми типами объектов по степени их подобия. В частности иерархически организованная легенда любой ландшафтной карты в неявном виде подразумевает ландшафтное пространство.

Следует сразу же отметить, что теория конструирования пространств в экологии и географии практически не разработана. В большинстве случаев пространство понимается как множество измерен-

ных переменных факторов (условий среды), потенциально определяющих свойства исследуемых объектов (обилие видов, свойства почв и т. п.). Строго говоря, это не пространство, а ориентированная система, пространство которой не определено. В частном случае при применении статистических методов она де-факто определяется как вероятностное пространство. Но в общем случае использование вероятностного пространства совершенно не обязательно.

Чтобы показать потенциальное значение корректного определения пространства, введем понятие метрического пространства.

Метрическое пространство есть множество X с некоторой метрикой ρ на нем.

Метрика есть расстояние на множестве X, — функция ρ с неотрицательными действительными значениями, удовлетворяющая при любых $x, y \in X$ условиям:

1) $\rho(x, y) = 0$ тогда и только тогда, когда $x = y$ (аксиома тождества);

2) $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$ (аксиома треугольника);

3) $\rho(x, y) = \rho(y, x)$ (аксиома симметрии).

Метрика с разным способом измерения расстояния ρ фактически определяет разные типы отношений и модель пространства. В экологии и географии используется очень широкий набор метрик, однако обоснования целесообразности представления отношений в избранной метрике обычно не приводятся. Любой человек, взаимодействуя с окружающим миром, неизбежно метризует его. То, что он представляет отношения между объектами расстояниями на поверхности земли — очевидно. Но даже здесь он использует различные метрики. Например, для того чтобы измерить расстояние между объектами в городе, где нельзя передвигаться произвольно, человек неявно использует метрику, получившую название «городское расстояние» (City block Manhattan distance). Если расстояние измеряется без учета существующих препятствий, ограничивающих передвижение, то используется классическое геометрическое расстояние в метрике Евклида; если препятствий много, например, в горах, то расстояние в основном вводится через время, необходимое для перемещения из одной точки в другую. Соответственно, пространства, определяемые этими тремя разными метриками, будут иметь разные свойства. Более того, в горах, если измерять расстояние через время, обычно может не выполняться аксиома симметрии, так как для подъема требуется больше времени, чем для спуска.

Неоднозначность определения расстояния отражается в частности в известной пословице: «прямо пять, в объезд четыре».

Физический смысл расстояния меняется в зависимости от того, велико оно или мало́. Так, очевидно, что расстояния 1 и 2 км существенно различны, а 100 и 101 км, в рамках ощущения расстояния, различаются очень мало. В этом варианте метризация вво-

дится как логарифм расстояния Эвклида. Отсюда, в частности, получаем, что для пешехода «близко» это примерно 2—3 км, «средне» — 4—6 км, «далеко» — 8—12 км. Для велосипедиста шкала расстояний будет уже иная.

Еще большее разнообразие метрик человек применяет при оценке отношений к людям, работе, предметам и явлениям. Поведение системы может быть совершенно не понятно, если наблюдатель задал неадекватную метрику. Допустим, что некоторый внешний наблюдатель контролирует движение людей в городе, но не видит улиц и домов. Измеряя движение пешеходов с помощью метрики Евклида, он увидит, что элементы, удивительным образом нарушая все законы нормальной динамики, передвигаются не по кратчайшим расстояниям, а по довольно сложным ступенчатым траекториям. Если наблюдатель не введет коррективы в метрику, то он не сможет построить адекватную модель их передвижения.

Точно также и исследователь, введя неадекватную метрику при исследовании конкретного объекта, может получить ложные отображения реальности.

Далее, насколько это возможно, будем максимально аккуратно определять пространство, в котором строится отображение, и оценивать порождаемые им свойства.

Понятие **размерности** интуитивно воспринимается в двух отношениях: размерность как связь между различными величинами, из которой вытекают различные системы измерения (например, СГС — сантиметр, грамм-масса, секунда), и размерность как число степеней свободы или число мыслимых перпендикулярных друг другу координат (свойств), определяющих объем, в котором можно найти отношение точек или объектов друг к другу. В данном случае будет рассматриваться второй вариант размерности. Первый вариант не менее важен для экологов и географов, но он в рамках этих наук вообще не разработан. Точнее, он разработан ровно настолько, насколько физика описывает в географической и экологической областях знаний реальность.

Введем понятие *размерности по А. Лебегу*: n-мерный в смысле элементарной геометрии куб Q^n при любом положительном числе ε может быть покрыт конечным числом замкнутых множеств (даже кубов) диаметром, меньшим ε, таким образом, что кратность этого покрытия равна $n + 1$. Это означает, что линию можно плотно покрыть замкнутыми отрезками сколь угодно малого размера так плотно, что щелей между отрезками не будет, но при этом соприкасаться друг с другом будут не более чем два отрезка. Соответственно размерность линии равна единице, а размерность точки равна нулю. Плоскость можно покрыть без щелей плитками, из которых соприкасаться друг с другом могут не более трех (размерность 2), размерность куба соответственно 3 и т.д.

Таким образом, когда говорят «n-мерное пространство», имеют в виду, что в него могут быть плотно упакованы n-мерные кубы, между которыми есть не более $n + 1$ контактов. Заметим, что пространства с нецелочисленной размерностью, разнообразие которых несоизмеримо больше пространств с целочисленной размерностью, имеют сколь угодно много незаполненных, в том числе и бесконечно тонких щелей. Такое пространство и связанное с ним множество называют *фрактальным*.

Эти формальные конструкции человек воспринимает как полную реальность, также как представляется ему полной реальностью трехмерное *пространство Евклида*. Однако, как показала практика и на существующем уровне обобщила теория, это только отображение не более чем некоторых свойств реальности. Экологические и географические свойства реальности по существу только начинают исследоваться и нет никаких априорных оснований утверждать, что они не содержат свойств, не нашедших отражения в существующих моделях пространства.

Но пока эти свойства не открыты, мы вынуждены рассматривать объекты в рамках существующих модельных представлений.

Так как обычно нас интересует положение объектов в n-мерном пространстве, естественно рассматривать объекты в рамках многомерной геометрии.

В рамках многомерной геометрии Евклидово пространство произвольного числа измерений (размерности) определяется как такое пространство, в котором выделены подмножества — прямые и плоскости, и имеются отношения: принадлежности ($\subset$, $\in$), порядка ($<$, $>$) конгруэнтности (определены расстояния ρ, или движения) и определяются обычные аксиомы теории множеств, кроме следующей: две плоскости с общей точкой имеют, по крайней мере, еще одну. Если это условие выполнено, то пространство трехмерно. В противном случае, т. е. в случае двух плоскостей с единственной общей точкой, пространство как минимум четырехмерно.

Понятие плоскости обобщается в многомерном пространстве следующим образом: *плоскостью* называется такое множество точек, которое вместе с любыми двумя своими точками содержит и проходящую через них прямую. В этом смысле все пространство является плоскостью. Пересечение всех плоскостей, содержащих данное множество точек M, будет плоскостью, «натянутой на M». Если плоскость натягивается на $(m + 1)$ точку, но не натягивается на меньшее их число, то она называется m-мерной. Пространство называется m-мерным, если оно является m-мерной плоскостью.

В n-мерном *Евклидовом пространстве* E_n через любую точку можно провести n и не более взаимно перпендикулярных прямых и соответственно прямоугольных координат $x_1, \ldots, x_n$.

При этом длина любого отрезка (расстояние между точками) есть

$$[XY] = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}.$$

В этом пространстве определены все операции на векторах и вектор имеет n составляющих.

Если обратиться к экологическим исследованиям, то трудно утверждать, что в них используются модели пространства E_n. Обычно координаты пространства явно не заданы. Если же в качестве координат рассматриваются некоторые факторы или условия среды, то, с одной стороны, они заведомо не ортогональны и формально не могут быть координатами E_n, а с другой — часто трудно измеримы, например, минеральное питание, или варьирование во времени режима увлажнения. Более доступны в измерениях расстояния, определенные в пространстве измеренных переменных, а собственно ортогональная система координат, к которой они принадлежат в модели E_n, пространства априори не известны.

Возможно для экологических и географических исследований более соответствует модель *псевдоевклидова пространства* E_n^{n-m} — множество, в которое введены координаты x_1, ..., x_n и интервалы между точками X, Y: $[(x_1 - y_1)^2 + \ldots + (x_m - y_m)^2] - [(x_{m-1} - -y_{m-1})^2 + \ldots + (x_n - y_n)^2]$. Геометрическими считаются определения и утверждения, формулируемые через отношения интервалов.

В частной теории относительности пространство-время определяется как E_4^1, при этом для нахождения расстояний используют формулу

$$D = c^2(t_1 - t_2)^2 - \left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2\right],$$

где $(t_1 - t_2)$ — расстояние по координате времени; $(x_1 - x_2)$, $(y_1 - y_2)$, $(z_1 - z_2)$ — расстояния в координатах пространства (x, y, z).

В этом пространстве, если интервал во времени равен интервалу в пространстве, то интервал между объектами равен нулю, т.е. объект принадлежит самому себе. Если задать такую или подобную дистанцию в палеогеографии и определить некоторую средневзвешенную константу c, то интервал нуль будет определять некоторую норму преобразования, а все отклонения от него — различные темпы палеогеографической эволюции.

Частным случаем псевдоевклидова пространства является *пространство Лобачевского* E_{n+1}^n. Важным фактом, заставляющим задуматься о применимости этой модели в экологии является то, что зрительное восприятие близких областей пространства человеком порождает эффект обратной перспективы. В прямой перспективе более удаленные объекты должны расходится, а в обратной перспективе близкие объекты более удалены друг от друга, чем

удаленные. Этот эффект объясняется тем, что геометрия этих областей восприятия пространства близка к геометрии пространства Лобачевского с радиусом кривизны около 15 м. Можно полагать, что восприятие пространства во многом определяет и взаимодействия объектов. Если считать доказанным восприятие пространства человеком по модели Лобачевского, то существует достаточно оснований допускать такое же восприятие окружающего мира и другими организмами.

Обобщением многомерной геометрии является «геодезическая геометрия» — геометрия многомерных пространств (G-*пространств*), которая определяется единственностью продолжения геодезических линий, определяемых как локально кратчайшие. Сама по себе геодезическая линия — геометрическое понятие, обобщающее понятие прямой (или отрезка прямой) евклидовой геометрии на случай пространств более общего вида. Определения геометрической линии в различных пространствах зависят от того, какая из структур (в частности метрика) лежит в основе геометрии рассматриваемого пространства.

G-пространство характеризуется следующей системой аксиом:

1) G есть метрическое пространство; $\rho(x, y)$ — расстояние в нем;

2) в G-пространстве ограниченные бесконечные множества имеют предельные точки, т.е. точки, для которых в любой сколь угодно малой окрестности существуют точки, принадлежащие тому же множеству (аксиома конечной компактности).

Здесь полезны некоторые, может быть, не очень строгие экологические и географические комментарии. Представление об экологической нише автоматически подразумевает существование предельной точки, в которой концентрируется информация о ее положении в пространстве. В какой-то степени предельной точке может быть поставлен в соответствие центр тяжести ниши как отражение понятия экологического оптимума вида. В географии существование предельной точки может быть связано в частности с широким применением классификаций, т.е. допущением того, что есть такая точка, свойства которой могут быть распространены на подмножество точек, относимых к тому же классу;

3) G-пространство выпукло, т.е. для точек $x \neq y$ есть отличная от них точка z такая, что $\rho(x, z) + \rho(z, y) = \rho(x, y)$. Это достаточно очевидное требование, подразумевающее, что как бы ни было мало расстояние между двумя точками, между ними всегда найдется место для третьей;

4) для каждой точки в шаре любого и сколь угодно малого радиуса найдутся две неравные точки и отличная от них третья;

5) если верна аксиома 4, то из $\rho(y, z_1) = \rho(y, z_2)$ следует, что $z_1 = z_2$.

Таким образом, G-пространство определяется весьма общими и достаточно естественными свойствами. Однако при дополнении

к этим общим аксиомам более частных строятся конструкции самых разнообразных пространств. Так например, G-пространства, в которых продолжение геодезической линии возможно в целом и любой ее участок остается кратчайшим, называются *прямыми пространствами*. К прямым пространствам, в частности, относятся пространства Евклида, Лобачевского, Минковского, любые римановы пространства неположительной кривизны. В прямом пространстве геодезическая линия определяется двумя точками. В общих G-пространствах, в отличие от прямых, сфера не всегда выпукла, а перпендикуляр не обязательно симметричен относительно линии или плоскости.

Вопрос о том, отвечает ли исследуемая экологами реальность представлениям о прямых пространствах или относится к пространствам более общего вида, можно считать фундаментальной проблемой науки.

Безусловно, теория пространства и теория измерения — будущее географии и экологии. Это — то, что можно с полным основанием определить как проблему: измеряя в природе множество переменных, получаем некоторые внешне осмысливаемые результаты, понимая, что наши действия не подчиняются достаточно обоснованным правилам и во многом скользят по самой поверхности изучаемых явлений. Происходит постоянное мультиплицирование множества частных отношений и ускользает то, что называется инвариантом: общие отношения, из которых через преобразования, подчиняющиеся определенным правилам, выводится множество частных. Это — будущее, путь в которое идет через настоящее.

Однако прежде чем обсуждать реалии в области известного, необходимо напомнить основные представления векторного пространства и векторной алгебры.

Векторное пространство есть **линейное** пространство. Оно определяется множеством элементов (точек) ($x, y \in E$) и полем ($\lambda, \mu, 1 \in K$). Поле есть класс объектов a, b, c, ..., для которых определены бинарные операции умножения и сложения, содержащие по крайней мере один элемент, отличный от нуля. При этом для каждого элемента $a \neq 0$ существует мультипликативный элемент a^{-1}. Элементы, принадлежащие полю K, называются *скалярами*. В векторном пространстве отображаются самые разнообразные линейные пространства и в том числе Евклидово пространство. Однако особо следует отметить, что векторное пространство способно описывать только линейные отношения на множествах. Если алгебра векторного пространства применяется к реальности с сильно нелинейными отношениями, когда параметр системы есть функция ее состояния, то ее отображение будет заведомо искажено, а строго говоря, бессмысленно. Несмотря на это при аккуратном использовании векторное пространство и векторная алгебра позво-

ляют в достаточно широком диапазоне реалий получать содержательные результаты.

Вектор — геометрически направленный отрезок прямой Евклидова пространства, у которого один конец называется началом, другой — концом. Длина отрезка без учета направления есть модуль вектора.

Вектор, длина которого равна единице, называется *единичным вектором* или *ортом*. Два вектора называются коллинеарными, если они лежат на одной прямой или параллельных прямых. Два вектора называются равными, если они имеют одинаковые модули и одинаково направлены. Такие вектора называются также свободными, так как по условию их начальная точка может быть выбрана свободно.

В экологических исследованиях исходная система с множеством измерений в векторном пространстве может быть задана двумя способами.

1. Каждая точка пространства есть элемент системы i, которому ставится в соответствие значение измеренных n переменных, трактуемых как n-мерный вектор $\mathbf{x}^i$, элементы которого имеют начало в нулевой точке системы координат:

$(x_1^i, x_2^i, \ldots, x_n^i)$, $i = 1, 2, \ldots, n$ — точки наблюдения (элементы) в пространстве-времени.

2. Каждая измеренная переменная или свойство (признак) рассматривается как элемент системы, которому ставится в соответствие m-мерный вектор $\mathbf{x}_j$, элементами которого являются значения переменной j, измеренные в m элементах системы:

$(x_j^1, x_j^2, \ldots, x_j^m)$, $j = 1, 2, \ldots, m$ — переменные (элементы).

Первый способ представления системы часто называют пространством наблюдений (R-пространство), а второй — пространством признаков (Q-пространство). Пространству наблюдений ставят в соответствие R-анализ, а пространству переменных — Q-анализ. Очевидно, что два пространства должны быть взаимно отображаемы.

Например, наблюдения в конкретный момент времени на конкретной метеостанции рассматриваются как элементы, а значения измеренных переменных (давление, температура, влажность и т. п.) как вектор (R-пространство). Напротив, измеренные переменные рассматриваются как элементы, каждому из которых соответствует вектор их значений на каждой метеостанции в конкретный срок наблюдений (Q-пространство).

Далее для упрощения записи будем обозначать векторы малыми латинскими буквами жирного начертания: **a**, **b**, **c**, ... При этом способ определения пространства не рассматривается.

Операция ***сложения*** векторов обладает следующими свойствами:

- *коммутативности*

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a};$$

• *ассоциативности*:

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c});$$

• допускает наличие *нулевого*

$$(\mathbf{a} + 0) = \mathbf{a}$$

и *противоположного*

$$\mathbf{a} + (-\mathbf{a}) = 0$$

элементов.

Разностью векторов **a** и **b** называется вектор **x** такой, что **x** + + **b** = **a**.

Произведением вектора **a** на число λ ($\lambda \neq 0$ и $\mathbf{a} \neq 0$) называется вектор **b**, коллинеарный **a**, модуль которого равен $|\lambda||\mathbf{a}|$ и который направлен в ту же сторону, что и вектор **a**, если $\lambda > 0$ и в противоположную, если $\lambda < 0$ (при $\lambda = 0$ или(и) $\mathbf{a} = 0$, $\mathbf{b} = 0$).

Умножение на скалярную величину подчиняется законам ассоциативности, коммутативности и дистрибутивности, в результате умножения на единицу имеем: $1 \cdot \mathbf{a} = \mathbf{a}$.

В анализе данных особо важное значение имеет понятие **линейной зависимости** векторов, что ассоциируется с линейной зависимостью физически измеренных переменных.

Вектора **a**, **b**, ..., **c** называются *линейно зависимыми*, если существуют числа α, β, ..., γ, из которых хотя бы одно отлично от нуля, такие, что справедливо равенство

$$\alpha\mathbf{a} + \beta\mathbf{b} + \ldots + \gamma\mathbf{c} = 0.$$

Отсюда следует, что вектора являются независимыми, если они лежат на одной или параллельных плоскостях.

Пусть каждый вектор имеет размерность k, тогда:

$$\begin{cases} \alpha\mathbf{a}_1 + \beta\mathbf{b}_1 + \ldots + \gamma\mathbf{c}_1 = 0; \\ \alpha\mathbf{a}_2 + \beta\mathbf{b}_2 + \ldots + \gamma\mathbf{c}_2 = 0; \\ \qquad\qquad \ldots \\ \alpha\mathbf{a}_k + \beta\mathbf{b}_k + \ldots + \gamma\mathbf{c}_k = 0. \end{cases}$$

Решение k-мерной системы уравнений позволяет определить коэффициенты, при которых вектора линейно зависимы, т.е. их значения определяют друг друга. Если нельзя найти коэффициенты, отличные от нуля, то вектора — независимы. Число линейно независимых векторов меньше или равно числу векторов пространства. Линейно независимые вектора принадлежат перпендикулярным друг другу плоскостям.

Из условия независимости вытекает важное положение, лежащее в основе анализа всех линейных систем. Понимание ниже определяемого факта является необходимым условием осмысленного использования всех методов многомерного анализа.

Совокупность линейно независимых векторов $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ многомерного пространства образует базис пространства.

В этом базисе любой вектор определяется так:

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \ldots + a_n\mathbf{e}_n,$$

где a_i (i = 1, 2, ..., n) — координаты (компоненты) вектора $\mathbf{a}$ в данном базисе. Соответственно записывают $\mathbf{a} = \{a_i\}$.

Два вектора $\mathbf{a} = \{a_i\}$ и $\mathbf{b} = \{b_i\}$ равны, когда равны их соответствующие координаты.

Необходимым и достаточным условием компланарности двух векторов является пропорциональность их соответствующих координат $\{a_i\} = \{\lambda b_i\}$ или определителя матрицы, образуемой координатами векторов

$$\Delta = \begin{pmatrix} a_1 & a_2 & a_n \\ b_1 & b_2 & b_n \\ . & . & . \\ c_1 & c_2 & c_n \end{pmatrix} = 0.$$

Если определитель матрицы координат векторов равен нулю, то вектора линейно зависимы.

В многомерных измерениях, представленных в пространстве переменных, все полученные значения можно рассматривать как значения базиса пространства или значения координат. Если определитель матрицы, образуемой координатами, равен нулю, то все переменные взаимозависимы и представимы как линии одной плоскости на одном общем базисе.

Представление об определителе возникло из решения систем линейных уравнений в рамках линейной алгебры:

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_{1n} = b_1; \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_{1n} + \ldots + a_{nn}\,x_{nn} = b_n. \end{cases}$$

Например, измерены значения a_{ij} и b_j и требуется определить значения x_{ij}. Из курса алгебры понятно, что задача разрешима, если число уравнений равно числу переменных. Множество коэффициентов a_{ij} называется матрицей, а множество b_j — ее собственными значениями. Когда число n линейных уравнений равно числу аргументов, матрицу называют **квадратной матрицей порядка *n***.

В матричной форме уравнение записывается следующим образом

$$Ax = b \text{ и } x = b/\Delta A,$$

где ΔA — определитель (детерминант) матрицы A.

Если определитель равен нулю, то система уравнений не имеет решений, точнее этими решениями являются любые значения из X. Действительно, например, вектора, принадлежащие одной плоскости или параллельным плоскостям, можно как угодно перемещать относительно друг друга так, что изменение координат одного вектора не приведет к изменениям координат других векторов.

Вернемся к рассмотрению операций с векторами.

Скалярным произведением ненулевых векторов $(\mathbf{a},\mathbf{b})$ называется произведение их модулей (т.е. их значений без учета направлений) на косинус угла $|\varphi|$ между их направлениями:

$$(\mathbf{a},\mathbf{b}) = |\mathbf{a}||\mathbf{b}|\cos\varphi.$$

Для вычисления скалярных произведений векторов используют декартову прямоугольную систему координат, т.е. координаты векторов в базисе (ортах).

Если $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ и $\mathbf{b} = (b_1, b_2, \ldots, b_n)$, то

$$(\mathbf{a},\mathbf{b}) = a_1b_1 + a_2b_2 + \ldots + a_nb_n.$$

Модули векторов $\mathbf{a}$ и $\mathbf{b}$

$|\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}$ и $|\mathbf{b}| = \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}$ — очевидно есть просто длины векторов в декартовой системе координат и

$$\cos\varphi = \frac{(\mathbf{a},\mathbf{b})}{|\mathbf{a}||\mathbf{b}|}.$$

Значение угла между векторами исключительно информативно и содержательно. Если косинус многомерного угла равен нулю, то, очевидно, что вектора тождественно равны; если они имеют общую нулевую точку в системе координат или если они не имеют такой общей точки, — параллельны. Если же косинус угла между векторами равен 1, то вектора перпендикулярны друг другу. Первый вариант соответствует понятию абсолютной зависимости между переменными, представленными в векторном пространстве, а второй, напротив, полной независимости. Таким образом, получаем метрику, позволяющую сформулировать две альтернативные гипотезы (H_1 — абсолютная зависимость, H_2 — абсолютная независимость), по отношению к которым можно рассматривать реальные отношения.

Заметим, что Евклидово пространство есть конечномерное, действительное, векторное пространство. Все алгебраические преобразования тождественны, а различия состоят лишь в том, что в пространстве Евклида изучаются объекты трех типов: точки, линии и плоскости, а в векторном пространстве — только линии.

Итак, рассмотрены все необходимые операции в линейном векторном пространстве, являющиеся в конечном итоге основой всех многомерных параметрических методов анализа.

Приведем результаты, на которые необходимо обратить особое внимание:

- на основе исходных данных в пространстве переменных или пространстве точек измерения можно определить их ортогональный базис или систему ортогональных координат пространства Евклида, которые однозначно будут описывать значения каждой переменной;
- между двумя переменными, если рассматривать их элементы с измерениями как координаты многомерного векторного пространства, всегда можно определить косинус многомерного угла, который указывает их взаимоположение в многомерном линейном пространстве.

4.2. Многомерные распределения случайных событий

Если случайное событие описывается упорядоченным набором действительных чисел x_1, x_2, ..., x_n, то этот набор представляет значения n-мерной случайной величины $X = (X_1, X_2, ..., X_n)$. Можно также говорить о системе случайных величин или о n-мерном случайном векторе с элементами из x_i. При этом имеют в виду, что X_i есть случайные значения координаты x_i.

В общем случае могут рассматриваться непрерывные, дискретные, дискриптивные (номинативные) переменные.

Рассмотрим поведение *трехмерной случайной величины*. Пусть случайное событие описывается значениями температур на метеостанции «Рязань» в июне, июле и августе (рис. 4.1). В трех координатах векторного пространства значений температур показаны положения случайных событий (элементов) в системе относительно друг друга. Первый вектор — среднемесячные температуры в июне за 100 лет наблюдений, второй — в июле и третий — в августе. В пространстве трех координат точки образуют разреженное облако с некоторым сгущением в области температур около 18—19 °С (рис. 4.1, *а*). На рис. 4.1, *б—г* показаны те же данные, но уже не в векторном, а в вероятностном пространстве частот событий в форме двухмерных распределений. Естественно, что на графике можно построить лишь двухмерные проекции из трехмерного пространства случайных величин. На графиках двухмерных распределений хорошо выражена мода, рассеивание вокруг которой демонстрирует конкретную структуру двухмерных распределений.

Рассмотрим модель двухмерного нормального распределения.

Фактически оно описывается теми же параметрами, что и одномерное, но с одним существенным дополнением.

Математическое ожидание двухмерного распределения случайных величин X_1 и X_2, называется *центром рассеивания* и определя-

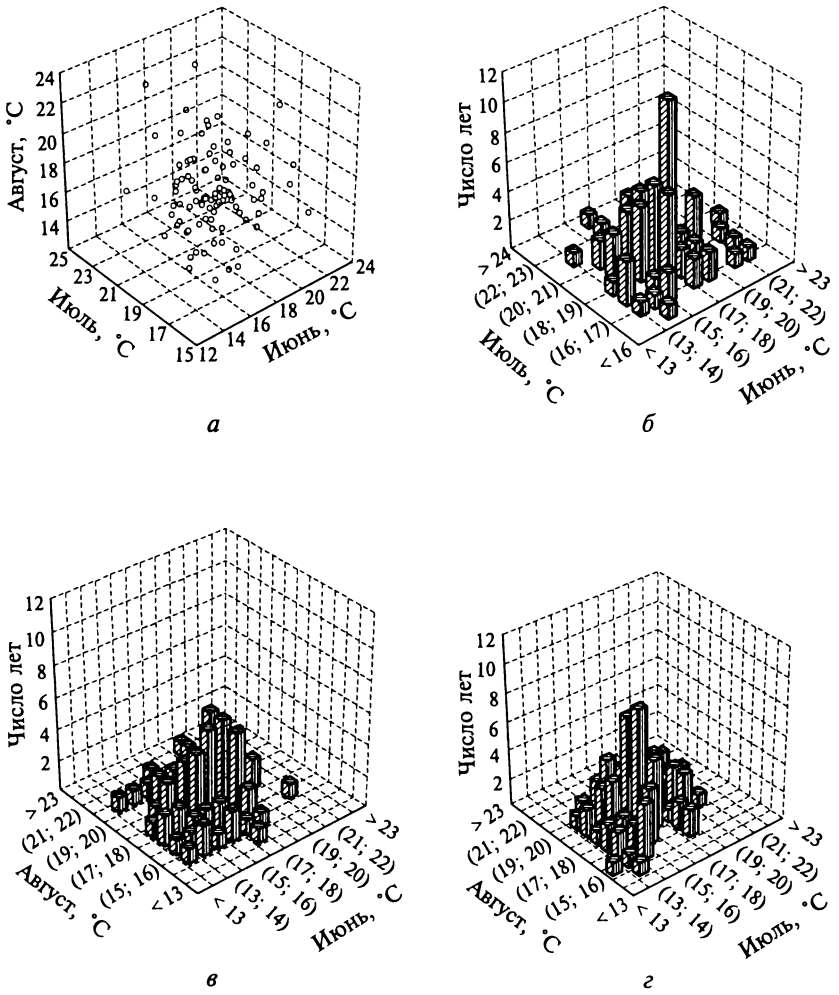


Рис. 4.1. Двухмерные распределения температур в летние месяцы (по данным метеостанции «Рязань»). Положение точек в трехмерном пространстве температур июня, июля, августа (*а*); июня, июля (*б*); июня, августа (*в*); июля, августа (*г*)

ется математическими ожиданиями первого и второго распределений M_{x_1} и M_{x_2}.

Второй центральный момент μ двухмерного распределения определяется тремя параметрами:

1) дисперсией первой переменной $\sigma^2_{x_1}$;

2) дисперсией второй переменной $\sigma^2_{x_2}$;

3) математическим ожиданием скалярного произведения векторов случайных величин $(x_1 - M_{x_1})$ и $(x_2 - M_{x_2})$, называемого *ковариацией*

$$K_{x_1x_2} = \frac{1}{n}\sum_{i=1}^{n}(x_1 - M_{x_1})(x_2 - M_{x_2}).$$

Математическое ожидание произведения стандартизованных отклонений одномерных распределений называется *коэффициентом корреляции*

$$R_{X_1X_2} = \frac{K_{x_1x_2}}{D_{x_1}D_{x_2}} = \frac{\operatorname{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

Если вспомнить, что среднее квадратическое отклонение определяется по формуле

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_{x_i})^2},$$

то легко увидеть, что коэффициент корреляции есть просто косинус угла между двумя n-мерными векторами ($\cos\varphi$).

Если коэффициент корреляции равен единице, то две случайные величины параллельны, если же он равен нулю, то они перпендикулярны друг другу (ортогональны) и независимы.

Для упрощения записи будем обозначать случайные нормально распределенные переменные через X и Y. Плотность их распределения есть

$$f(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - r_{XY}^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1 - r_{XY}^2)}\left[\left(\frac{x - \bar{X}}{\sigma_X}\right)^2 - 2r_{XY}\frac{x - \bar{X}}{\sigma_X}\frac{y - \bar{Y}}{\sigma_Y} + \left(\frac{y - \bar{Y}}{\sigma_Y}\right)^2\right]\right\}.$$

Другая форма записи, менее принятая, но более наглядная

$$f(X,Y) = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2 - R_{XY}^2}} \times$$

$$\times \exp\left(-\frac{\sigma_Y^2(x - \bar{X})^2 - 2R_{XY}(x - \bar{X})(y - \bar{Y}) + \sigma_X^2(y - \bar{Y})^2}{2(\sigma_X^2\sigma_Y^2 - R_{XY}^2)}\right),$$

где R_{XY} — корреляционный момент распределения $R_{XY} = r_{XY}\sigma_X\sigma_Y$.

Из простого анализа этого отношения видно, что если коэффициент корреляции равен единице, то двухмерное распределение вырождается в одномерное, но с совместной дисперсией, равной произведению частных дисперсий, т.е. с большим рассеиванием. Если коэффициент корреляции равен нулю, то плотность двухмерного распределения определяется независимым варьированием каждой переменной, т.е. суммой их дисперсий.

На рис. 4.2 показано двухмерное нормальное распределение с коэффициентом корреляции, равным 0,5.

Проекции изолиний равных значений $f(X, Y)$ образуют эллипс, большой диаметр которого проходит под углом $\varphi = 45°$ ($\cos\varphi = 0{,}5$). При отрицательном коэффициенте корреляции большой диаметр поворачивается на 90°. Средства любого пакета статистических программ позволяют построить модели нормальных распределений, что и предлагается проделать читателю самостоятельно.

Обобщим случай двухмерного распределения на многомерный.

Нормальное n-мерное распределение, так же как и двухмерное, полностью описывается математическими ожиданиями каждой переменной (вектора), их дисперсиями и попарными корреляциями. Однако дисперсии и коэффициенты корреляции представляются как квадратная матрица порядка n, в которой на главной диагонали записаны дисперсии, а в ячейках $i \neq j$ — соответствующие значения корреляций r_{ij}.

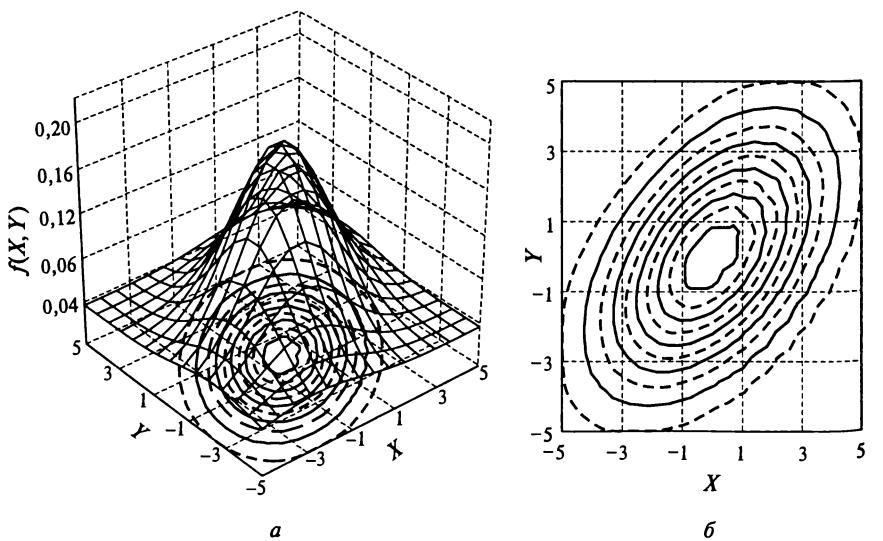


Рис. 4.2. Двухмерное нормальное распределение (*а*, *б*) с положительным коэффициентом корреляции ($r_{XY} = 0{,}5$)

$$\Delta(R_{jk}) = \begin{pmatrix} \sigma_{11}^2 & \cdots & R_{1n} \\ \vdots & \ddots & \vdots \\ R_{n1} & \cdots & \sigma_{nn}^2 \end{pmatrix}.$$

Определитель этой матрицы $\Delta(R_{jk})$ содержит всю необходимую информацию о нормальном распределении. Если определитель равен нулю, то все случайные переменные параллельны, если единице, то, напротив, ортогональны. Представления о матрице и определителе являются исключительно важными для понимания смысла решения всех задач многомерного параметрического анализа, а в конечном итоге — и непараметрических методов.

Собственно все задачи можно решить на основе операций векторной алгебры.

Прежде всего обратим внимание на то, что парные коэффициенты корреляции накладывают некоторые ограничения друг на друга. Если, например, совместное распределение двух случайных переменных A и B описывается их парной корреляцией, отличной от нуля, и если A коррелирует с C также с некоторым коэффициентом корреляции, отличным от нуля, то возможные комбинации значений B и C ограничены. При этом, чем больше значения корреляций, тем больше эти ограничения. Если коэффициент корреляции между A и B и между A и C равен единице, то он будет равен единице и для B и C.

С другой стороны, матрица вторых моментов может рассматриваться как система уравнений, каждое из которых можно приравнять нулю. Как следует из векторной алгебры, решения этого уравнения дадут значения координат ортогонального базиса системы. В частном случае некоторые из них будут равны нулю. Если определитель такой системы будет равен нулю, то система уравнений не имеет решений и все переменные в многомерном пространстве будут параллельны. Если же определитель такой матрицы равен единице, то все координаты в ортогональном базисе будут отличны от нуля.

Итак, плотность n-мерного нормального распределения по аналогии с двухмерным определяется по формуле

$$f(X_1, \ \ldots, X_n) = \frac{1}{\sqrt{2\pi\Delta(R_{ij})}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_i - M_{x_i})(x_j - M_{x_j})}{\Delta(R_{ij})} \right).$$

Энтропия многомерного нормального распределения

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} (0,5 \log 2\pi\sigma_{X_i}) - [-\log(\Delta(R_{X_1, \ldots, X_n})].$$

Под знаком суммы стоит энтропия n независимых случайных переменных. Логарифм от определителя является мерой взаимо-

сопряженности всех переменных и отражает взаимоограничение их варьирования. Если иметь в виду, что энтропия по смыслу близка к разнообразию, то можно сказать, что логарифм определителя есть мера ограничения разнообразия состояний системы. С другой стороны, ограничения разнообразия есть информация и соответственно эта же мера может трактоваться как количество информации всех переменных друг о друге, содержащееся в системе. Чем ближе определитель к нулю, тем больше сопряженность между переменными системы.

Замечание. Вместо термина «сопряженность» часто используются термины «связь» или «зависимость». Однако в русском языке «зависимость» не явно подразумевает некоторое физическое воздействие, а «связь» подразумевает процессы во времени и физическую среду, в которой происходит воздействие одного объекта на другой. При оценке ограничения разнообразия ни физическая среда, ни механизмы воздействия не рассматриваются. Следовательно в этом случае более оправданно применять более нейтральное понятие «сопряженность».

Вполне понятно, что многомерные распределения непрерывных переменных не обязательно нормальны. В отличие от нормального, они так же, как в одномерном случае должны иметь ненулевые значения третьих и четвертых моментов, что требует построения очень сложных моделей. На практике такие распределения рассматриваются крайне редко и в весьма специальных задачах.

Корреляция Пирсона и ее статистическое оценивание

Второй момент многомерного нормального распределения обычно называют *коэффициентом корреляции Пирсона* или просто корреляцией Пирсона.

Как и в одномерном случае, оценка корреляции осуществляется на основе выборки из генеральной совокупности.

Выборочное значение r_{XY} является несмещенной оценкой математического ожидания.

Выборочная дисперсия коэффициента корреляции $D_r = \dfrac{(1-r^2)^2}{n}$.

Среднеквадратическая ошибка коэффициента корреляции

$$m_{r_{XY}} = \frac{1-r^2}{\sqrt{n}}.$$

Отношение выборочного коэффициента к среднеквадратической ошибке есть t-распределение Стьюдента

$$\mathrm{t} = \frac{r}{m_{r_{XY}}}.$$

Соответственно интервал, к которому с заданной вероятностью принадлежит математическое ожидание коэффициента корреляции r при оценке по выборке, есть $r \pm tm_r$.

На основе коэффициента корреляции проверяется гипотеза значимости линейной связи.

Коэффициент корреляции считается значимо отличным от нуля, если выполняется неравенство

$$r^2 > \left(1 + \frac{n-2}{t^2}\right).$$

Как обычно гипотеза проверяется при t = 2 (первый уровень значимости) и t = 3 (второй уровень значимости).

Если $r \neq 0$, то используется критерий на основе z-распределения

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Уже при небольших объемах выборки z-распределение подобно нормальному с математическим ожиданием

$$M_z = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-1)}$$

и дисперсией

$$D_r = \sigma_r^2 = 1/(n-3).$$

Гипотеза $r = 0$ принимается по z-распределению при заданном математическом ожидании и стандартном отклонении. В современных программах уровень значимости принятия гипотезы о независимости (p-levels) обычно рассчитывается автоматически вместе с коэффициентом корреляции.

Коэффициент корреляции Пирсона адекватно отражает реальность в том и только в том случае, если распределения близки к нормальным, а сами отношения линейны.

Продемонстрируем значение этого требования на конкретном примере.

Вернемся к данным об относительной влажности почв. Известно, что распределения реально измеренных переменных существенно отличны от нормальных, логарифмированные значения переменных приближаются к нормальным, но все-таки содержат максимальные значения, далеко выходящие за доверительные интервалы. Рассмотрим, каково значение коэффициента корреляции между первым и вторым горизонтами в трех вариантах представления данных: исходных, логарифмированных, логарифмированных без выбросов (рис. 4.3, *а—в*).

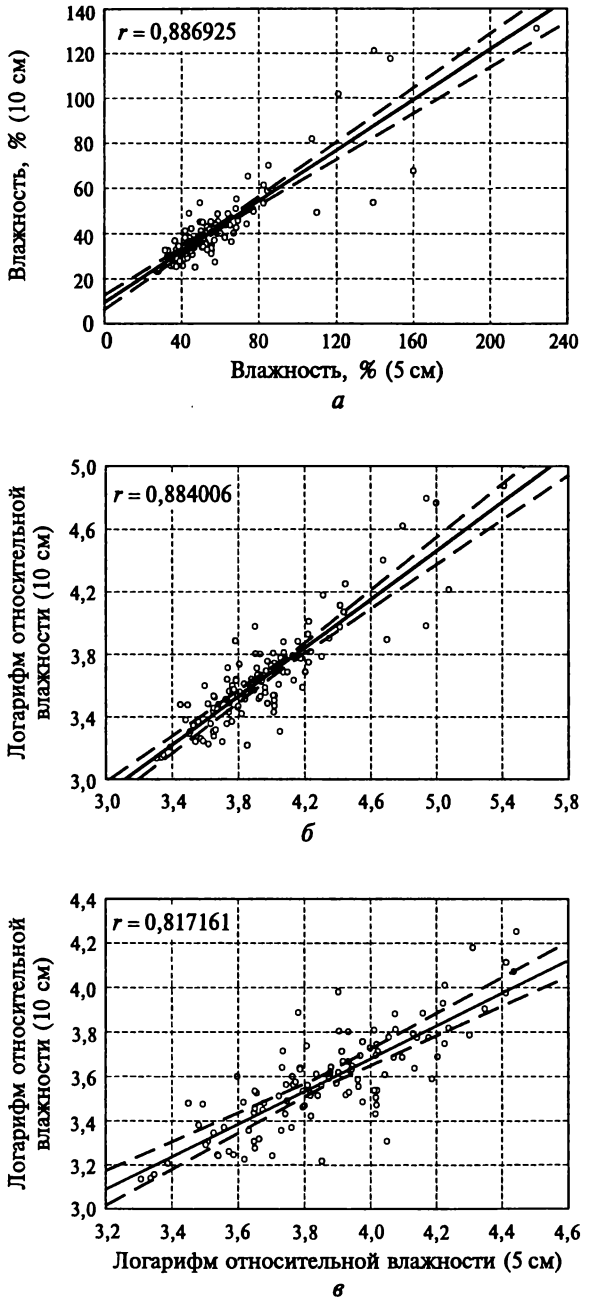


Рис. 4.3. Графики зависимости коэффициента корреляции от способа представления данных:

а — исходных; *б* — логарифмированных; *в* — логарифмированных без выбросов

Распределения логарифмированных значений без экстремальных значений влажности в долинах ручьев имеют нормальное распределение и минимальное значение корреляции. Если не исключать максимальные значения, то корреляция увеличивается и становится максимальной при расчете для исходных данных. Очевидно, что экстремальные данные существенно завышают корреляцию, причем в данном случае выборочные оценки корреляции различаются статистически значимо.

Продемонстрировать влияние экстремальных отклонений на величину коэффициента корреляции Пирсона можно на следующем примере (рис. 4.4):

1) в генераторе случайных чисел создадим две независимые выборки;

2) добавим к ним только одну, но достаточно удаленную точку *1*.

Корреляция между двумя случайными выборками практически равна нулю, а корреляция с добавлением всего одной экстремальной точки становится весьма значительной и статистически значимой в соответствии с критерием.

Совершенно очевидно, что реально такой корреляции не существует и для таких данных нельзя применять корреляцию Пирсона. Здесь оценка корреляции просто не имеет смысла. Реакцию коэффициента корреляции Пирсона на экстремальные отклонения легко понять, если обратиться к самой схеме расчета ковариации и дисперсий. Очевидно, что экстремальные значения входят

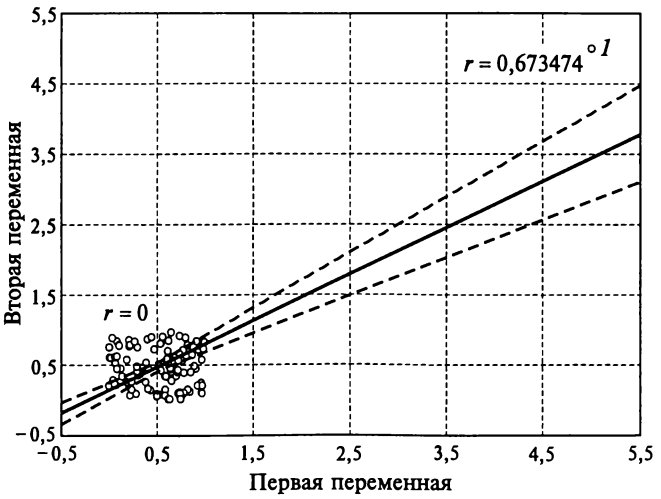


Рис. 4.4. Реакция коэффициента корреляции Пирсона на добавление к двум независимым переменным одной общей удаленной от остальных точки *1*

в оценки с очень большим весом, пропорциональным величине их отклонения от среднего, и в результате корреляция, рассчитанная на их основе, становится для реальных данных существенно завышенной.

Рассмотрим, как влияет на коэффициент корреляции нелинейность отношений. Обратимся к реальному примеру. На лесной станции Ковита в Южных Аппалачах (Coweeta, Северная Каролина, США) много лет действует одна из лучших в мире систем экологического мониторинга, включающая измерение прихода осадков над конкретными речными бассейнами, содержания в них катионов и анионов основных биогенных элементов, расхода воды в соответствующих бассейнах и содержания в стоке катионов и анионов. Вполне понятно, что одной из задач исследования является определение параметров речного бассейна в преобразовании воздействия на входе (поступление из атмосферы) в состояние выхода (речной сток).

В данном случае рассмотрим корреляцию между модулем стока и концентрацией в нем катиона кальция по среднемесячным данным. Распределения переменных близки к гамма-распределению и приближаются к нормальному при логарифмировании исходных данных. Как видно из графика (рис. 4.5), логарифмирование, хотя и не полностью, но переводит нелинейную зависимость модуля стока и концентрации катиона кальция в почти линейную. Коэффициент корреляции Пирсона для линеаризованных данных почти на 0,1 больше, чем для исходных, и формально это увеличение достоверно. Эффект уменьшения коэффициента корреляции при нелинейной зависимости в сравнении с линейной определяется тем, что при прочих равных условиях дисперсия распределения, отличающего от нормального, существенно больше, чем у нормального, и соответственно меньше отношение ковариации и средних квадратических, т.е. коэффициента корреляции.

Можно показать, что даже для полностью детерминированных нелинейных отношений можно получить нулевой коэффициент корреляции.

Для этого достаточно допустить, что некоторая переменная есть параболическая функция какого-либо фактора

$$y = ax - bx^2.$$

Такие отношения часто описывают обилие какого-либо вида от некоторого фактора среды (например, влажности).

На рис. 4.6 показан вид этой абсолютной неслучайной зависимости, у которой коэффициент корреляции Пирсона естественно равен нулю. Это происходит по той простой причине, что любому значению y соответствуют два противоположных по величине значения x. Поэтому в оценку ковариации они внесут равный вклад,

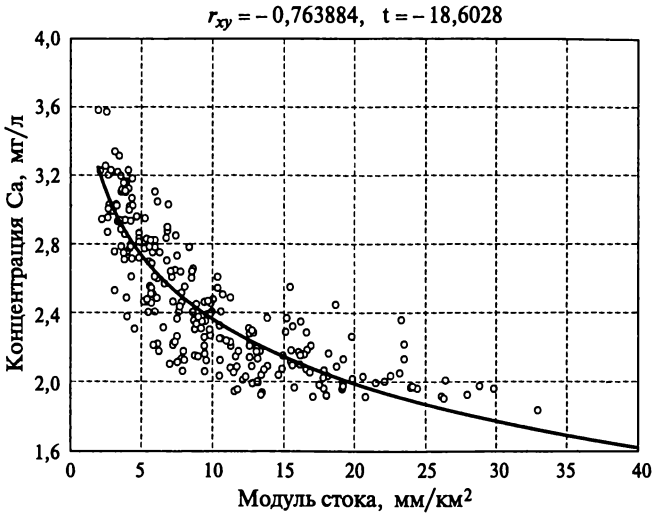


Рис. 4.5. Оценки корреляции Пирсона для исходного (*а*) и линеаризованного (*б*) отношений

но с разными знаками, что при суммировании по всем n произведениям отклонений обратит ковариацию в нуль

$$\sum (x_i - \bar{X})(y_i - \bar{Y}) \approx 0.$$

Столь большое внимание свойствам коэффициента корреляции Пирсона уделяется потому, что, несмотря на все предуп-

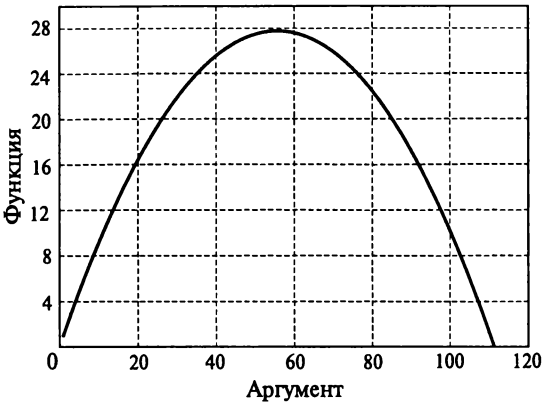


Рис. 4.6. График зависимости вида $y = x - 0{,}009\,x^2$ ($r_{xy} = 0$)

реждения о строгих ограничениях его применения, его часто используют без всякого контроля типа распределения измеренных переменных и типа их отношений. Из приведенных примеров следует, что это совершенно недопустимо. В лучшем случае будут получены оценки корреляции, находящиеся в неизвестно каком отношении к реальности, а в худшем будет установлена связь там, где ее не существует, и напротив, не будет установлена там, где она абсолютная, но нелинейная. Поэтому коэффициент корреляции Пирсона без предварительной оценки свойств распределений и их максимально возможной нормализации и оценки типа отношений между переменными просто не содержит никакой информации.

4.3. Регрессионная модель и параметрический регрессионный анализ

Если дано распределение двух случайных величин X и Y, то регрессией Y по X называется любая функция $g(X)$, приближенно представляющая статистическую зависимость Y от X

$$Y = g(X) + \varepsilon(X, Y),$$

где $\varepsilon(X, Y)$ — ошибка, нормально распределенная случайная величина с математическим ожиданием, равным нулю, т.е.

$$g(X) \equiv M(y/x) = \begin{cases} \int y f_{y/x}(y/x)dy \text{ — для непрерывного распределения;} \\ \sum_y y p_{y/x}(y/x) \text{ — для дискретного распределения.} \end{cases}$$

Таким образом, регрессия есть математическое ожидание Y при известном значении из X.

Для дискретного распределения — это сумма произведений условных вероятностей $p(y/x)$ при одном и том же $x_i \in X$, умноженная на соответствующие значения из Y.

Для непрерывных переменных $f_{y/x}$ — функция плотности условных распределений.

Наиболее просто регрессионные модели строятся для систем с линейными отношениями и нормальными распределениями.

В алгебраической форме безразлично, какую переменную рассматривать как аргумент, а какую — как функцию. В реальных исследованиях наиболее оправданно применение регрессионной модели для ориентированных систем по схеме «вход (X) — выход (Y)».

Линейную регрессию функции от одного аргумента определяют по формуле

$$y = a + bx + \varepsilon,$$

где $b = \rho \dfrac{\sigma_y}{\sigma_X}$ — регрессионный коэффициент; $a = M_y - \rho \dfrac{\sigma_y}{\sigma_X} M_x$; ε — средняя квадратическая ошибка модели; $M_x(M_y)$ — математическое ожидание X (Y).

Модель регрессии прямо вытекает из векторной алгебры. Коэффициент корреляции r определяет угол между векторами Y и X, а отношение средних квадратических масштабирует рассеивание Y по X.

Обратим внимание на физический смысл отношения средних квадратических отклонений. Если оно меньше единицы, то среда, через которую проходит сигнал от входа к выходу, уменьшает его амплитуду, а если больше единицы, то, наоборот, увеличивает. Это позволяет трактовать отношение средних квадратических как меру среды (регулятор).

Конечно, это регулирование совершенно необязательно осуществляет среда. Механизмы регулирования могут принадлежать и самой системе. Однако если переменные измерены в одной системе, то факт различия средних квадратических заслуживает поиска физической трактовки. Кроме того, коэффициент уравнения регрессии b можно трактовать как коэффициент чувствительности Y к X.

Чувствительность b в наиболее общей трактовке есть отношение производной $\mathrm{d}Y$ к производной $\mathrm{d}X$, т.е. она показывает, во сколько раз изменится Y по отношению к изменению X:

$$b = \mathrm{d}y/\mathrm{d}x \text{ или } \mathrm{d}y = b\mathrm{d}x.$$

Представление о регуляторе и чувствительности как о параметрах регрессионной модели иногда способствует семантической трактовке полученных отношений.

Ранее речь шла о модели регрессии. Собственно регрессионный анализ оперирует с оценками параметров модели по выборке. В рамках регрессионного анализа все оценки строятся на основе уже описанной регрессионной системы. В связи с чем вполне оправдано рассматривать всю систему оценивания модели на конкретном примере.

В качестве такого примера дадим оценку параметров одномерного уравнения регрессии для измерений влажности почвы в горизонте 5 см (X) и в горизонте 10 см (Y). В данном случае естественно принять в качестве входа влажность почвы в верхнем слое по отношению к нижнему. Предыдущий анализ показал, что корректное построение линейной регрессионной модели возможно только для ряда дерново-подзолистых почв на покровном суглинке, без включения в анализ измерений в приручьевых, переувлажненных фациях.

В табл. 4.1 приведены параметры полученной модели. Коэффициент детерминации показывает, какую долю варьирования влажности на глубине 10 см описывает влажность на глубине 5 см.

Подправленный коэффициент детерминации вводится с учетом числа степеней свободы, что в большинстве случаев не имеет принципиального значения. Он становится информативным только при малых выборках. Критерий Фишера более подробно развернут в специальной таблице дисперсионного анализа (табл. 4.2), но в данном случае на его основе констатируем, что вклад регрессионной модели в описание варьирования можно считать абсолютно значимым.

Таблица 4.1

Отчет о регрессионной модели по отношению к зависимой переменной

Regression Summary for Dependent Variable: влажность 10 см; коэффициент корреляции r = 0,82767561; коэффициент детерминации R^2 = = 0,68504691; подправленный коэффициент детерминации (Adjusted) R^2 = 0,68225971; критерий Фишера для модели $F(1,113)$ = 245,78; уровень значимости модели p < 0,00000; оценка средней квадратической ошибки модели Std. Error of estimate: 0,13381

Переменная	BETA	Std. Err. of BETA	b	Std. Err. of b	t(113)	p-level
Константа Intercept	—	—	0,760255	0,181212	4,19540	0,000054
Влажность на глубине 10 см	0,827676	0,052794	0,729503	0,046532	15,67749	0,000000

Таблица 4.2

Дисперсионный анализ для модели регрессии (Analysis of Variance)

Характеристика	Сумма квадратов Sums of Squares	Число степеней свободы df	Среднее квадратическое отклонение Mean Squares	Критерий Фишера F	Уровень значимости p-level
Модель регрессии (Regress)	4,400750	1	4,400750	245,7836	0,000000
Остатки (Residual)	2,023263	113	0,017905		
В целом (Total)	6,424013				

Стандартная ошибка показывает, что все значения переменной (логарифм влажности на глубине 10 см по влажности на глубине 5 см) предсказываются с общей ошибкой ±0,13381.

В табл. 4.1 приведены следующие параметры модели: BETA — стандартизированное значение коэффициента модели регрессии b (коэффициент чувствительности, величина которого не зависит от варьирования переменной); Std. Err.of BETA — средняя квадратическая ошибка BETA; b — значения параметров линейной модели $y = a + bx + \varepsilon$ (a — константа (Intercept); b — наклон (slope) линии регрессии; t(113) — t-критерий Стьюдента при числе степеней свободы df = 113; p-level — уровень значимости, вероятность принадлежности параметра генеральной совокупности с нулевым значением его математического ожидания.

Таким образом, очевидно, что параметры линейной регрессии статистически значимы на высоком уровне качества описания.

Обратим внимание на тот факт, что отношение сумм квадратов модели регрессии к общей сумме квадратов (4,400750/6,424013 = = 0,685) равно коэффициенту детерминации R^2, т.е.

$$R^2 = \frac{\sigma^2_{\text{модель}}}{\sigma^2_{\text{общая}}}.$$

Таким образом, модель может быть записана в следующей форме:

$$\ln(\text{M10}) = 0{,}760255 + 0{,}729503 \ln(\text{M5}) \pm 0{,}13381.$$

Для того чтобы признать модель полной, т.е. исключающей зависимость влажности почвы на глубине 10 см от каких-либо других факторов, кроме влажности почвы на глубине 5 см, необходимо исследовать остатки (Residual). Можно принять гипотезу отсутствия

иных факторов, если остатки имеют нормальное распределение и их величина не зависит от значения функции или наблюдаемого варьирования влажности на глубине 10 см.

Из рис. 4.7 следует, что распределение остатков с высокой вероятностью нормальное. Однако на рис. 4.8 видно, что величина остатков есть функция наблюдаемых значений зависимой переменной. При этом коэффициент корреляции $r = 0{,}56120682$ при коэффициенте детерминации $R^2 = 0{,}31495309$. Это означает, что модель в среднем занижает максимальные значения. Возможно, это указывает на существование некоторого квадратичного вклада в модель зависимой переменной:

$$\ln(\mathrm{M10}) = a + b\ln(\mathrm{M5}) + c(\ln(\mathrm{M5}))^2.$$

Однако это несоответствие реальным соотношениям не позволяет отвергнуть полученной модели. С практической точки зрения различия относительно небольшие.

Коэффициент регулирования в модели, равный 0,91285 при ошибке 0,18, не указывает на существование при переходе от одной глубины к другой какого-либо регулирующего эффекта.

Напомним, что регрессионная модель построена в логарифмической шкале измерения. Потенцируя, получаем степенную модель вида

$$\mathrm{M10} = \exp(a + b\ln \mathrm{M5} \pm \varepsilon),$$

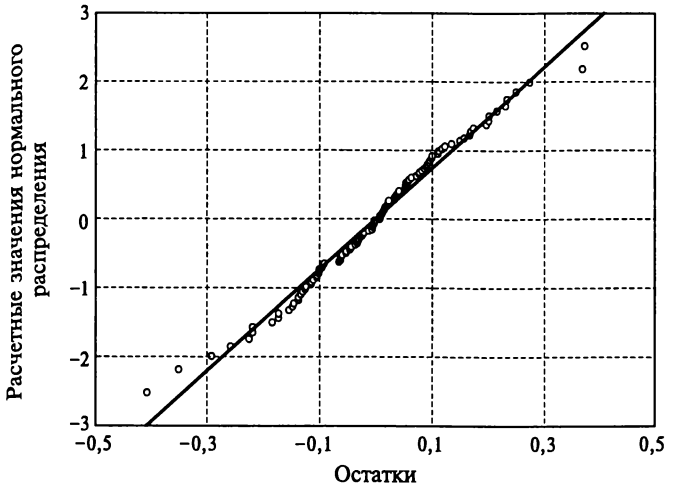
$$\mathrm{M10} = 2{,}1388\mathrm{M5}^{0{,}7295}(0{,}875 \leftrightarrows 1{,}143).$$

Чтобы отразить эффект регулирования в степенной модели, разделим левую и правую части последнего равенства на M10 и получим

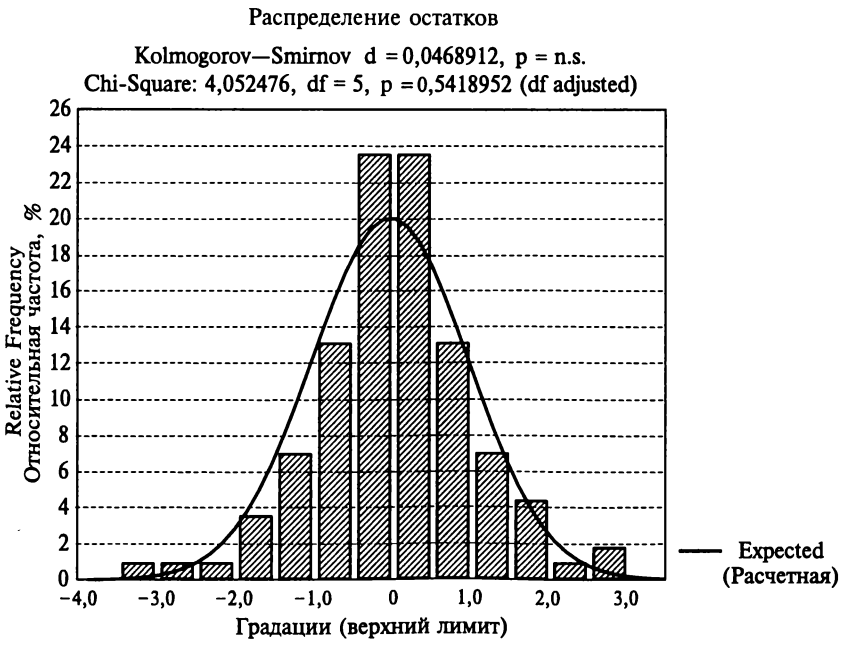
$$\mathrm{Рег} = \mathrm{M10}/\mathrm{M5} = (2{,}1388/\mathrm{M5}^{0{,}27})\ (0{,}875 \leftrightarrows 1{,}143).$$

На рис. 4.9 это отношение показано на графике. Полученное соотношение вполне логично: чем больше влаги в верхнем горизонте, тем относительно меньшая ее доля достигает нижнего. Обычно относительная влажность в гумусовом горизонте дерново-подзолистых почв не бывает ниже 20 %, поэтому отношения при меньших значениях влажности чисто гипотетические. В целом вид функции соответствует моделям теории очередей. Полученное соотношение описывает вероятность p-перехода влаги из верхнего горизонта в нижний.

Надо полагать, что доля влаги, дополняющая вероятностное пространство до единицы $(1 - p)$, определяет среднестатистический вклад транспирации в интервале глубин 5—10 см. Анализируемый результат имел бы строгую физическую интерпретацию, если бы измерялось абсолютное содержание влаги на единицу объема. Однако качественное описание отношений вполне содержательно и имеет физический смысл.



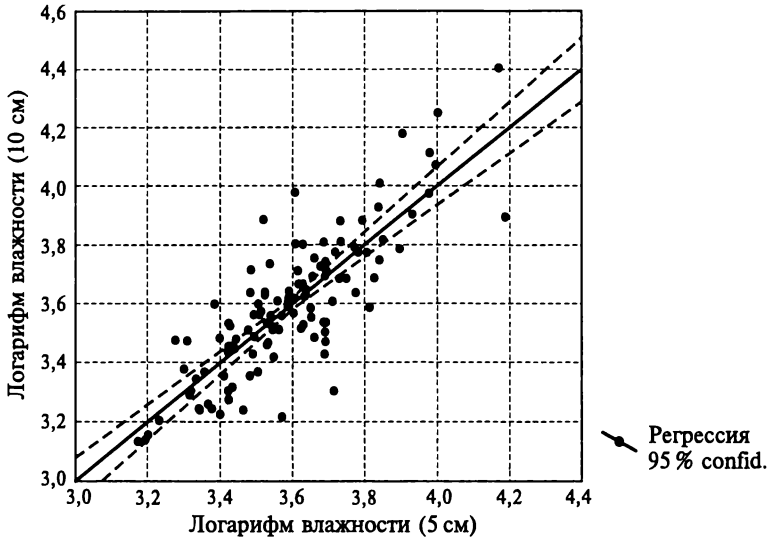
а



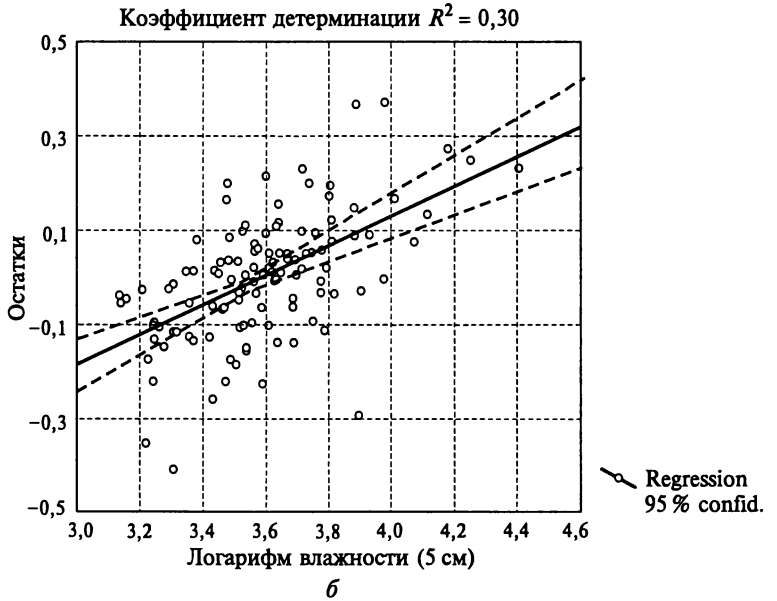
б

Рис. 4.7. Тесты на нормальность остатков регрессионной модели влажности почвы на глубине 10 см по влажности почвы на глубине 5 см:

а — график вероятностей; *б* — распределение остатков



а



б

Рис. 4.8. Регрессионная модель: логарифм влажности почвы на глубине 10 см как функция логарифма влажности почвы на глубине 5 см (*а*); регрессия между стандартизованными значениями остатков и определяемой переменной (*б*)

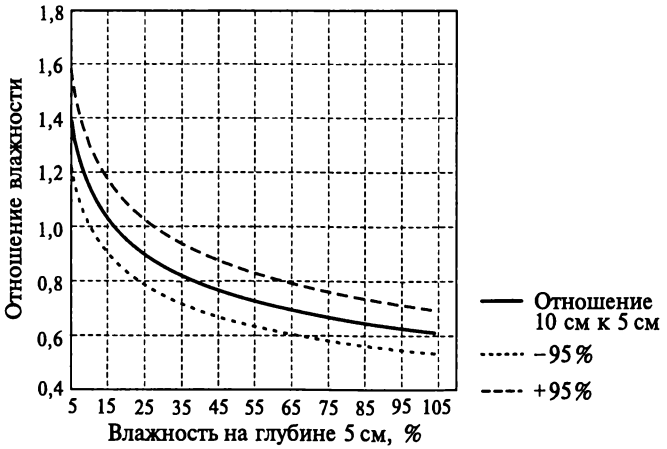


Рис. 4.9. Отношение относительной влажности почвы на глубине 10 см к влажности на глубине 5 см (эффект регулирования)

Многомерная регрессионная модель

Перейдем к определению параметров ориентированной системы, имеющей один выход и несколько входов, т.е.

$$Y = f(X_1, X_2, ..., X_n).$$

Общую модель регрессии можно представить формулой

$$Y = a + \sum_{k \neq 1} \beta_{ik}(x_k - \mu_k),$$

где $\beta_{ik} = -\dfrac{A(K(X_1, ..., X_n))}{A(\sigma^2_{X_1,...,X_n})}$ — коэффициенты уравнения регрессии при аргументах, определяемые так же как и в одномерном случае из соотношений ковариационной матрицы $A(K(X_1, ..., X_n))$ и матрицы дисперсий — $A(\sigma^2_{X_i})$, т.е. через соотношения вторых моментов многомерного распределения. Коэффициенты определяются из решения системы k уравнений, в которых известны значения из Y и X и требуется определить параметры β_i при аргументах X_i.

В общем случае многие из $\beta_i = 0$, т.е. соответствующие им переменные не связаны с аргументом. Если какие-то переменные не определяют Y и включены в модель, то они не повлияют на отношение дисперсии, описанной моделью, к общей дисперсии, но понизят в результате увеличения числа степеней свободы значение F-критерия, т.е. снизят качество модели.

В связи с вышеизложенным можно предложить следующую процедуру построения модели множественной регрессии.

1. Строим частные парные модели регрессий ($y = a + b_iX_i$), качество которых будет определяться значением F-критерия;

2. Принимаем некоторое пороговое значение F или уровень p, который будет ограничивать включение переменной в модель.

3. Выбираем частную регрессионную модель из $y = a + b_i X_i$ с максимальным F-критерием и добавляем в модель следующую переменную со вторым по величине F-критерием, рассчитав новые ее параметры. При этом возможны следующие варианты:

- качество модели повысилось и F-критерий стал выше на некоторую заранее принятую пороговую величину;
- качество модели не повысилось на заданную пороговую величину.

В первом случае сохраняем новую переменную в модели, во втором — исключаем.

Следует отметить, что частная модель исключенной переменной могла быть очень неплохого качества, однако это качество определялось не ее собственной связью с Y, а высокой связью с первой переменной. Эта собственная связь переменной-вход с переменной-выход с исключением связи, определяемой скоррелированностью ее с другими переменными-входами, измеряется частным коэффициентом корреляции

$$r_{12|3} = \frac{r_{12} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}},$$

где r_{ij} — парные коэффициенты корреляции, записанные в символике математического ожидания. Для выборочных оценок используется символ r или R (relation — отношение).

Таким образом, используя F-критерии и частные корреляции, можно последовательными итерациями добиться того, что в модели останутся только переменные-входы, значимо и относительно независимо определяющие состояние переменной-выхода.

Можно построить процедуры от обратного. Сначала создать общую регрессионную модель от всех переменных, а затем исключать те переменные, вклад которых по частной корреляции невысок и исключение которых не снижает качества модели при заданном уровне F-критерия.

Первый метод называют «шаг вперед — Forward», а второй — «шаг назад — Backward».

В целом эти методы построения регрессионных моделей называются *пошаговой регрессией.*

Теперь, располагая достаточными методическими ресурсами, попытаемся решить задачу реального уровня сложности. Вернемся к материалам по станции «Ковита» (США). Применяя методы множественной регрессии, всегда полезно иметь некоторую гипотезу о возможном влиянии входов на выходы. Нас будет интересовать концентрация катиона кальция в стоке. Факт нелинейной зависимости концентрации кальция от модуля стока установлен ранее.

Такая связь может иметь физическое объяснение, но не исключена и ее опосредованная косвенная природа. В общем случае можно полагать, что концентрация кальция в стоке есть функция прихода его с атмосферными осадками, причем не обязательно текущего времени. Можно допустить существенное запаздывание, связанное с затратами времени на фильтрацию, а также влияние на концентрацию кальция температуры среды, так как растворимость любого вещества — логарифмическая функция температуры. В принципе и во влиянии температур может быть некоторое запаздывание. Концентрация кальция может также зависеть от функционирования растительности, однако состояние этого фактора нам неизвестно.

Так как во всех случаях распределения близки к гамма или логнормальным, примем простейшее и наиболее удобное логарифмическое преобразование всех переменных, кроме суммы осадков за месяц, для которого оптимально преобразование через корень квадратный из данных. Исходные данные: концентрация катиона кальция в мг/л (CA), поступление кальция с осадками в мг/км^2 (CAPR), количество осадков в мм (PR), модуль стока в мм/км^2 (FLOW), температура по шкале Кельвина (T); PR1 — осадки в текущем месяце, PR2 — в предшествующем и аналогично для CAPR и T.

В табл. 4.3 приведена общая стандартная модель множественной регрессии. Статистическую значимость влияния, или формально значимость отличия выборочного значения параметра от нуля, можно рассматривать как меру его влияния на функцию выхода. Из модели следует очевидно высокая значимость влияния модуля стока. В целом, чем больше осадков, тем меньше концентрация кальция в стоке, при этом вклад в модель статистически не значим в месяц наблюдений и максимален от осадков, выпавших в третий предшествующий месяц. Сумма выпавшего с осадками на территорию кальция практически не влияет на концентрацию его в стоке. Наибольшее положительное влияние на концентрацию кальция в стоке оказывает температура текущего месяца и несколько меньшую — двух предшествующих. В табл. 4.4 приведены результаты модели пошаговой регрессии методом «шаг вперед» при входном значение F = 4 и исключении переменной из модели при F = = 3,999. Следует отметить, что в данном случае методы «шаг вперед» и «шаг назад» дают тождественные результаты.

Как следует из модели, исключение из нее очень большого числа малосвязанных и косвенно связанных с концентрацией кальция в стоке переменных практически не снижает ее качества: коэффициент детерминации почти не изменился, средняя квадратическая ошибка практически та же, а критерий Фишера в связи со значительным уменьшением числа степеней свободы существенно выше.

Таблица 4.3

Параметры стандартной модели множественной регрессии концентрации кальция в стоке

Коэффициент корреляции r = 0,96112941; коэффициент детерминации R^2 = 0,92376975; подправленный коэффициент детерминации (Adjusted) R^2 = 0,91727479; критерий Фишера F(19,223) = 142,23; уровень значимости нулевой гипотезы p < 0,0000; стандартная ошибка модели (Std.Error of estimate): 0,05037

Переменная	Стандартизованный коэффициент BETA	Ошибка Std.Err.of BETA	Параметры модели b_i	Ошибки параметра Std.Err.of b	Критерий Стьюдента t(223)	Уровень значимости p-level
Константа Intercept			–23,9703	1,894426	–12,6531	0,000000
FLOW	–0,479055	0,042475	–0,0877	0,007775	–11,2785	0,000000
PR1	–0,022706	0,029380	–0,0011	0,001399	–0,7729	0,440428
PR2	–0,047362	0,036768	–0,0023	0,001764	–1,2881	0,199043
PR3	–0,108194	0,031309	–0,0052	0,001502	–3,4557	0,000657
PR4	–0,089781	0,029832	–0,0043	0,001428	–3,0095	0,002917
PR5	–0,031877	0,029625	–0,0015	0,001420	–1,0760	0,283091
PR6	–0,033712	0,029556	–0,0016	0,001421	–1,1406	0,255251
PR7	0,036351	0,028430	0,0018	0,001370	1,2786	0,202364
CAPR1	0,032856	0,033231	0,0079	0,008027	0,9887	0,323873
CAPR2	0,023065	0,036428	0,0056	0,008803	0,6332	0,527282
CAPR3	0,030084	0,035417	0,0073	0,008617	0,8494	0,396557
CAPR4	0,021858	0,035552	0,0053	0,008651	0,6148	0,539306
CAPR5	0,004347	0,035516	0,0011	0,008642	0,1224	0,902700
CAPR6	0,051670	0,035714	0,0126	0,008695	1,4468	0,149366
CAPR7	–0,055039	0,032369	–0,0134	0,007874	–1,7003	0,090462
T1	0,298998	0,055191	2,1271	0,392644	5,4175	0,000000
T2	0,121637	0,067047	0,8625	0,475415	1,8142	0,070990
T3	0,157290	0,067448	1,1148	0,478047	2,3320	0,020592
T4	0,025446	0,055892	0,1806	0,396689	0,4553	0,649355

Прежде чем обсуждать физический смысл модели, рассмотрим дополнительные оценки ее качества. Стандартная таблица анализа вариаций дает общую оценку качества моделей (табл. 4.5).

Это уже известный нам одновариантный дисперсионный анализ.

Таблица 4.4

Модель множественной пошаговой регрессии концентрации стока от совокупности заданных входных переменных

Коэффициент корреляции r = 0,95856843; коэффициент детерминации R^2 = 0,91885343; подправленный коэффициент детерминации (Adjusted) R^2 = 0,91679038; критерий Фишера F(6,236) = 445,39; уровень значимости нулевой гипотезы p < 0,0000; средняя квадратическая стандартная ошибка (Std.Error of estimate): 0,05052

Переменная	Стандартизованный коэффициент BETA	Ошибка Std.Err. of BETA	Параметры модели b_i	Ошибки параметра Std.Err.of b	Критерий Стьюдента t(236)	Уровень значимости p-level
Intercept			–23,3961	1,096296	–21,3410	0,000000
FLOW	–0,527028	0,026691	–0,0965	0,004886	–19,7452	0,000000
T1	0,360744	0,023470	2,5664	0,166970	15,3705	0,000000
T3	0,227875	0,025393	1,6151	0,179979	8,9738	0,000000
PR3	–0,121386	0,025506	–0,0058	0,001224	–4,7591	0,000003
PR4	–0,066978	0,020021	–0,0032	0,000958	–3,3453	0,000956
CAPR3	0,070946	0,025343	0,0173	0,006166	2,7995	0,005542

Частный коэффициент корреляции (табл. 4.6) показывает корреляцию каждого аргумента с переменной вне зависимости от ее собственной корреляции с другими переменными, которые косвенно могут увеличивать или уменьшать значение коэффициента корреляции, рассчитываемого просто по варьированию функции и аргумента. В соответствии с этими оценками частная корреляция

Таблица 4.5

Дисперсионный анализ (Analysis of Variance)

Характеристика	Сумма квадратов Sums of Squares	Число степеней свободы df	Среднее квадратическое Mean Squares	Критерий Фишера F	Уровень значимости p-level
Модель регрессии (Regress)	6,847372	6	1,141229	450,7090	0,00
Остатки (Residual)	0,605166	239	0,002532		
В целом (Total)	7,452538				

Таблица 4.6

Оценка частного влияния каждой переменной

Переменная	Стандартизованный коэффициент BETA	Частная корреляция Partial Cor.	Получастная корреляция Semipart Cor.	Чувствительность Tolerance	Частный коэффициент детерминации R_i^2	Коэффициент Стьюдента t(236)	Уровень значимости p-level
FLOW	–0,527418	–0,789536	–0,366605	0,483155	0,516845	–19,7452	0,000000
T1	0,360953	0,707899	0,285600	0,626060	0,373940	15,3705	0,000000
T3	0,228692	0,505813	0,167087	0,533806	0,466194	8,9738	0,000000
PR3	–0,119832	–0,293772	–0,087578	0,534125	0,465875	–4,7591	0,000003
PR4	–0,065825	–0,209411	–0,061027	0,859524	0,140476	–3,3453	0,001074
CAPR3	0,069852	0,176809	0,051190	0,537047	0,462953	2,7995	0,005918

максимальна для модуля стока и температуры воздуха. У остальных аргументов она существенно ниже.

Получастный коэффициент корреляции содержит ту же информацию, что и частный, но он нормирован на варьирование определяемой (зависимой) функции. Если получастная корреляция очень мала, но частная корреляция относительно велика, то соответствующая переменная может предсказывать уникальный «кусок» изменчивости зависимой переменной, который не находит отражения в других переменных. Однако, с практической точки зрения, этот «кусок» может быть очень мал и представлять только очень малую долю полной изменчивости. Частный коэффициент детерминации показывает долю варьирования зависимой переменной, описываемой конкретной независимой i-переменной без косвенного влияния всех других.

В результате получаем оценку варьирования зависимой переменной (функции) каждым конкретным аргументом. Соответственно в данном случае можно констатировать, что наибольшее собственное значение в описании варьирования функции имеют модуль стока, температуры, наблюдавшиеся два месяца назад, и приход в это же время кальция с атмосферными осадками. Чувствительность (толеранс) T рассчитывается по формуле $T = 1 - R_i^2$ и показывает степень некоррелируемости независимых переменных. Большое значение толеранса указывает на высокую взаимозависимость аргументов и большую возможную ошибку соответствующего коэффициента регрессии. Такую переменную обычно не рекомендуется включать в модель. В данном случае в целом все независимые переменные (аргументы) вносят вполне достоверный вклад в регрессионную модель. Информация, представлен-

ная в табл. 4.7, позволяет в случае неопределенности искать более надежный вариант модели.

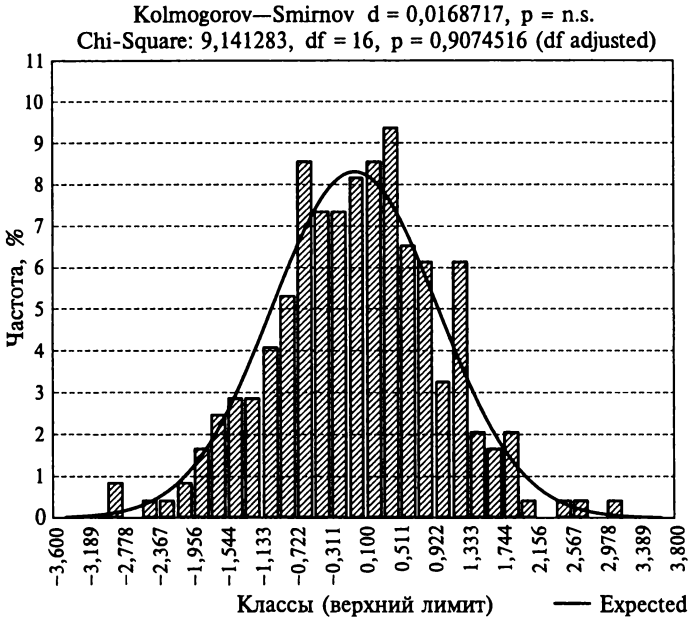
Из табл. 4.7 следует, что все переменные, включаемые в модель, имеют высокие значения частных коэффициентов корреляции, при этом значения их коэффициентов детерминации могут быть и не очень высокими. Так, например, температуры второго и четвертого месяца и поступление кальция с осадками описывают бóльшую долю варьирования переменной, но частные коэффициенты их корреляции очень малы, т.е. вся информация о концентрации кальция, содержащаяся в этих переменных, является косвенной.

Оценим качество модели по внешним критериям. Если регрессионная модель адекватно описывает реальные данные, то распределение остатков (разница между расчетным и реальным значениями) должно подчиняться нормальному закону. На рис. 4.10

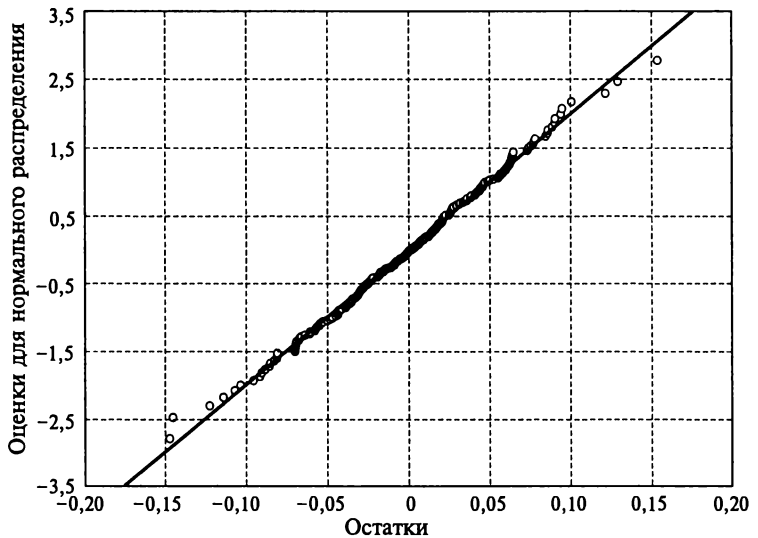
Таблица 4.7

Корреляционные оценки переменных, включенных и не включенных в модель

Переменная	Параметр				
	Tolerance T	R_i^2	Partial Cor.	Semipart Cor.	Статус переменной
FLOW	0,482628	0,517372	–0,789257	–0,366134	Включенные в модель
T1	0,624217	0,375783	0,707296	0,285015	
PR3	0,528538	0,471462	–0,295918	–0,088248	
PR4	0,857765	0,142235	–0,212775	–0,062032	
T3	0,533229	0,466771	0,504392	0,166400	
CAPR3	0,535369	0,464631	0,179278	0,051911	
PR1	0,816162	0,183838	0,030158	0,008591	Не включенные в модель
PR2	0,559203	0,440797	–0,086876	–0,024748	
PR5	0,887189	0,112811	–0,071627	–0,020404	
PR6	0,877732	0,122268	0,004104	0,001169	
PR7	0,919582	0,080418	0,003738	0,001065	
CAPR1	0,771590	0,228410	0,102409	0,029173	
CAPR2	0,563575	0,436425	–0,000429	–0,000122	
CAPR4	0,361094	0,638906	0,079913	0,022764	
CAPR5	0,778775	0,221225	0,029371	0,008367	
CAPR6	0,737372	0,262628	0,086554	0,024656	
CAPR7	0,775570	0,224430	–0,010866	–0,003095	
T2	0,079211	0,920789	0,113897	0,032445	
T4	0,122979	0,877021	0,030517	0,008693	



а



б

Рис. 4.10. Оценка нормальности распределения остатков регрессионной модели:

а — нормальность распределения; *б* — нормальность по регрессии

показаны результаты соответствующего тестирования, на основе которого можно утверждать, что действительно распределение остатков может быть признано нормальным и по этому критерию модель может быть признана удовлетворительной. На рис. 4.11 представлена связь между расчетными (рис. 4.11, *а*) и измеренными (рис. 4.11, *б*) значениями с 95%-м доверительным интервалом и связь остатков с наблюдаемым значением переменной. Второй график (см. рис. 4.11, *б*) показывает, что при больших значениях переменных значения остатков несколько больше, чем при малых. Это означает, что модель не может воспроизвести всей амплитуды варьирования переменной и недостаточно описывает максимальные значения. Достоверность корреляции между остатками и измеренными значениями концентрации кальция показывает, что модель не учитывает какой-то неизвестный фактор.

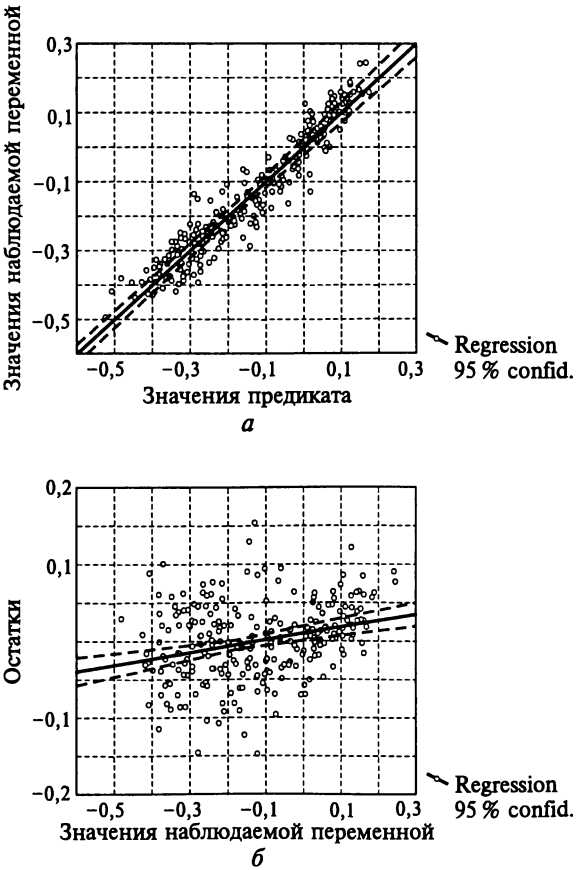


Рис. 4.11. Связь реальных данных с расчетными (*а*) и остатков с измеренными (*б*) значениями функции

На рис. 4.12, *а* приведена динамика измеренных и рассчитанных по модели концентраций кальция в стоке во времени, на рис. 4.12, *б* — изменение во времени стандартизованных остатков от модели регрессии. Из графиков следует, что лишь в трех случаях отклонения достигают трех средних квадратических, т.е. уровня очень редких событий. Однако визуальный просмотр графиков приводит к необходимости принять гипотезу о том, что во времени величина ошибки изменяется не случайно. Эта неслучайность подчеркивается линией полиномиального тренда пятой степени. Анализу временны́х рядов будет посвящен отдельный раздел пособия (см. гл. 9). Здесь же будем констатировать возможность построения регрессионной модели от полинома степени n для временно́го ряда:

$$Y = a + b_1 t + b_2 t^2 + \ldots + b_n t^n,$$

где t — время (в данном случае единица измерения времени — один месяц).

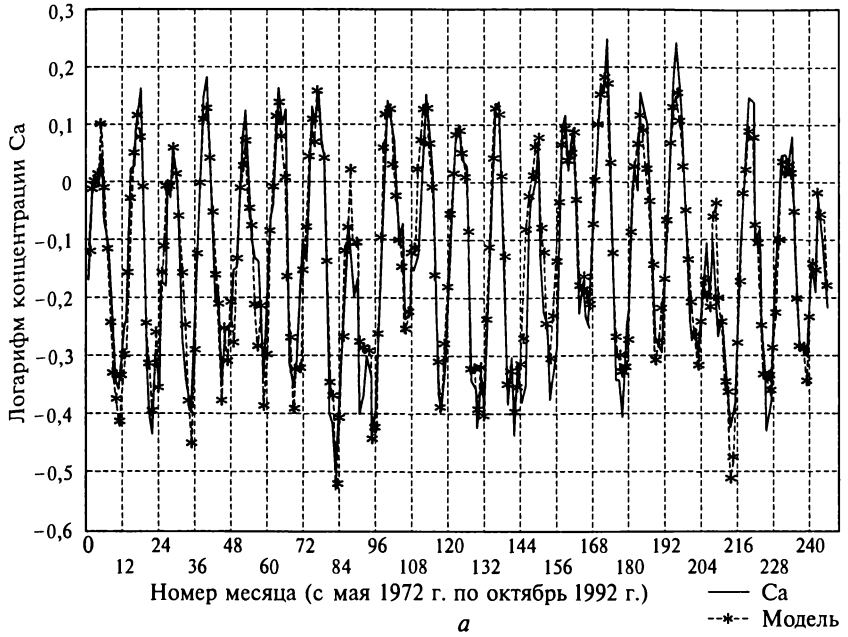
Если отклонение от модели есть функция времени, то это означает существование еще какого-то неизвестного фактора, влияющего на концентрацию, который сам по себе является функцией времени.

Для проверки гипотезы можно создать $n - 1$ новых переменных (квадрат времени, куб времени и т.д.) и воспользоваться той же программой множественного регрессионного анализа. Но обычно в статистических пакетах программ есть специальные модули, решающие задачи построения регрессионной модели от полинома степени n.

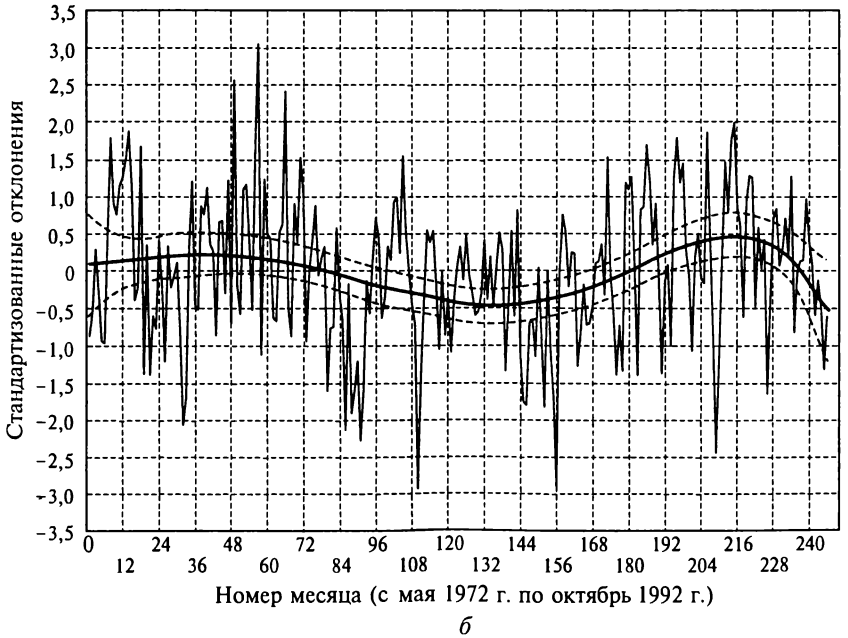
Была задана модель пятого порядка и, как следует из табл. 4.8, она вполне статистически значима и описывается полиномом пятой степени, представленным на рис. 4.12, *б*. Следовательно, есть все основания считать, что существует какой-то неизвестный фактор, изменяющийся во времени и описывающий всего лишь 6 % варьирования остатков.

На рис. 4.13 приведен еще один тест — дистанция Махаланобиса и дистанция Кока, показывающие масштабы отклонений измеренных значений от расчетных. Дистанция Махаланобиса в общем случае рассчитывается на основе ковариационных матриц, но в данном случае она тождественна обычной дистанции Евклида. Дистанция Кока обладает очень высокой чувствительностью к отклонениям. Совместно эти дистанции показывают максимумы отклонений модели от реальных измерений, которые скорее всего нарушают стационарность и равновесность отношений, описываемых регрессионной моделью.

Итак, с одной стороны, в целом регрессионная модель вполне приемлема и описывает около 91 % варьирования переменных, а с другой — все-таки оправдана гипотеза о существовании двух типов малозначащих, но достоверных неизвестных факторов. Что



а



б

Рис. 4.12. Динамика концентрации кальция в стоке — реальные данные и прогноз по модели (*а*); стандартизованная ошибка регрессионной модели и ее полиномиальный тренд (*б*)

Таблица 4.8

Дисперсионный анализ для модели полиномиального тренда остатков

Коэффициент корреляции r = 0,276783; коэффициент детерминации R^2 = 0,076609; подправленный коэффициент детерминации R^2 = 0,065162

Переменная	Сумма квадратов Sums of Squares	Число степеней свободы df	Среднее квадратическое отклонение Mean Squares	Критерий Фишера F	Уровень значимости p-level
Модель регрессии (Regress)	18,30958	3	6,103192	6,692508	0,000234
Остатки (Residual)	220,6904	242	0,911944		

касается фактора, который определяет экстремальные значения концентраций, не воспроизводимых моделью, то, возможно, они связаны с экстремальными температурами в режиме декад или

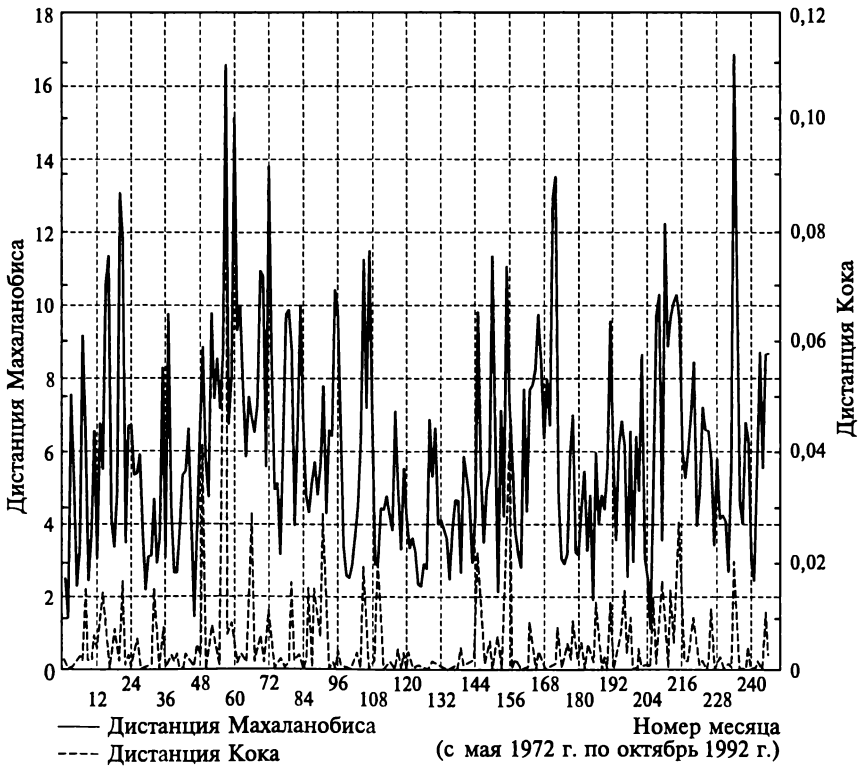


Рис. 4.13. Дистанции Махаланобиса и Кока между исходными данными и регрессионной моделью

более коротких периодов. Для проверки этой гипотезы нужно рассматривать не среднемесячные, а среднедекадные данные. Что касается фактора, определяющего слабые флуктуации концентраций во времени, то для формулировки гипотезы о его природе требуется существенная дополнительная информация. В общем, в его воздействии намечается некоторая периодичность, что позволяет предполагать возможность влияния циклических преобразований самой растительности.

Таким образом, рассмотренная реальная задача показывает, что, с одной стороны, удается построить статистическую модель, описывающую более 90 % варьирования выхода по значениям входов, а с другой — почти наверное существуют хотя и малозначащие, но скорее всего реально действующие, неизвестные факторы.

Запишем уравнение для модели, полученной на основе множественной пошаговой регрессии:

$$\ln Ca = -23{,}3961 - 0{,}0965 \ln FLOW + 2{,}5664 \ln T1 + 1{,}6151 \ln T3 - $$
$$- 0{,}0058 PR3^{0{,}5} - 0{,}0032 PR4^{0{,}5} + 0{,}0173 \ln CAPR3 \pm 0{,}05052.$$

Преобразуем уравнение в натуральную форму

$$Ca(мг/л) = 0{,}0000000000691 Flow -$$
$$-^{0{,}0965} T1^{2{,}5664} T3^{1{,}6151} CAPR3^{0{,}0173} 0{,}9942^{PR3\hat{}0{,}5} 0{,}9968^{PR4\hat{}0{,}5} (0{,}951-1{,}052).$$

Физическая трактовка полученной модели достаточно естественна:

1. Чем меньше модуль стока, тем выше концентрация, что можно объяснить большим временем, необходимым для установления равновесия между раствором и твердой почвенной фазой;

2. Это подтверждается аналогичным по сути действием атмосферных осадков: большое количество осадков приводит к снижению концентрации или большему «разбавлению раствора», при этом запаздывание в их воздействии составляет три-четыре месяца. Это запаздывание, по-видимому, определяется временем, необходимым для перехода атмосферных осадков в сток;

3. Поступление кальция в сток зависит от его прихода из атмосферы, но также с запаздыванием в три месяца;

4. Температура воздуха и соответственно почвы, по-видимому, прямо влияет на растворимость примерно пропорционально квадрату температуры. При этом и здесь на концентрацию влияет температура не только текущего месяца, но и предшествующего.

Таким образом, концентрация определяется растворимостью и «разбавляемостью». Естественно, что такие жесткие соотношения между внешними переменными и концентрацией возможны только при определенном дефиците кальция в обменном комплексе.

Общие замечания по применению метода множественной регрессии

Еще раз повторим, что метод множественной регрессии в полной мере применим в том случае, если распределения близки к нормальным, а зависимости приведены к линейной форме. Исследование остатков позволяет сформулировать гипотезы как о степени линейности модели, так и о возможных неизвестных факторах. Нелинейность модели ведет к существенному отклонению вероятностного графика остатков от расчетного нормального распределения. Ситуацию можно исправить, введя переменные, для которых подразумевается нелинейная форма влияния в степенной форме. Иногда оказывается существенна роль произведения или отношения двух факторов (неаддитивный, несуммируемый эффект). Найти простым перебором неаддитивный эффект довольно трудно. Желательно, чтобы у исследователя была более или менее естественная гипотеза о неаддитивности. Например, исследуется продукция растительности или интенсивность фотосинтеза как функция температуры и увлажнения. Из общих соображений известно, что график зависимости фотосинтеза от температуры имеет форму, приближающуюся к параболической: низкие и высокие температуры неблагоприятны и существует некоторый температурный оптимум. Точно такую же зависимость можно предположить и по отношению к увлажнению. Однако известно, что скорее всего, кроме собственно температуры и собственно влаги на продукционный процесс действует их соотношение. Поэтому в модели множественной регрессии каждая переменная должна быть представлена в первой и второй степени и, кроме того, должно быть введено их произведение или отношение. В общем всегда желательно, чтобы у исследователя существовала гипотеза о возможных отношениях.

При этом желательно, чтобы модели множественной пошаговой регрессии «шаг вперед» и «шаг назад» давали бы тождественные результаты. Однако модели, выполненные двумя методами, иногда различаются. Часто эти отличия определяются недостаточным объемом выборки при большом числе независимых (входных) переменных. Модель пошаговой регрессии «назад», используя в качестве входной матрицу всех парных коэффициентов корреляции, может обнаружить случайные, но при ограниченном объеме данных с формальных позиций статистически значимые комбинации факторов. Если модели, полученные двумя методами, не совпадают, то исследователю нужно очень аккуратно анализировать частные коэффициенты корреляции, частные коэффициенты детерминации и толерансы для того, чтобы убедиться, что модель «назад» включает действительно значимые переменные.

При небольшом объеме данных (не превышающих 50—100 измерений) включать большое число переменных в модель крайне

нежелательно. Если же условия эксперимента не позволяют получить большой массив данных, то регрессионную модель нужно строить с большой осторожностью, комбинируя различные переменные и отбирая те из них, вклад которых в описание варьирования сохраняется для любых комбинаций. Затем к отобранным наиболее «надежным» переменным можно постепенно добавлять менее «надежные». При этом добавление новых переменных должно улучшать качество описания (увеличивать коэффициент детерминации и снижать среднюю квадратическую ошибку или исправлять нелинейность), но при этом незначительно снижать значение критерия Фишера.

Все эти и другие приемы исследователь осваивает по мере накопления опыта анализа. Но во всех случаях он должен быть разумным скептиком и рассматривать полученный результат как с использованием статистических критериев, так и с позиции физического смысла полученных отношений.

4.4. Другие методы построения статистических моделей «вход—выход»

Существуют статистические методы, позволяющие строить модели отношений между переменными по заданным нелинейным функциям вида

$$y = a + b(x + d)^c; \quad y = ab^x; \quad y = a/(d + cx) \text{ и т. п.}$$

При этом зависимость может быть функцией от нескольких аргументов, однако функция задается в строго определенном виде и фактически отражает гипотезу исследователя о форме отношения.

В программе функция трансформируется в линейную форму и подбор коэффициентов осуществляется итерационно так, чтобы сумма квадратов ошибки была минимальной. В некоторых программах большое значение имеет установка стартовых значений определяемых параметров, но в большинстве пакетов это происходит автоматически. В частном случае реальные данные могут быть несовместимы с заданной моделью отношений. Программа показывает эту несовместимость, указывая на то, что ряд не сходится или параметры не могут быть рассчитаны. Ниже использование этой процедуры (nonlinear estimation) будет продемонстрировано на конкретном примере.

В модели множественной регрессии исследователь рассматривает систему как «черный ящик»: известны значения переменных на входах в систему и на выходе, но отсутствуют сведения о том, как связаны эти переменные входы внутри системы и как они влияют друг на друга. Если последовательно применять схему множественной регрессии к различным переменным, учитывая при этом их

частные корреляции, то можно выявить некоторые элементы структуры самой системы. Когда есть гипотезы о связях между входными переменными в их воздействии на выход, то иногда говорят о «сером ящике». При наличии разумных предположений о связях между переменными их использование существенно снижает потенциальную комбинаторику отношений и позволяет построить надежную регрессионную модель при меньшем исходном объеме данных или получить представление о неизмеренных факторах, существование которых физически необходимо, но измерение их по каким-либо причинам невозможно.

Эти задачи решаются методом структурного моделирования (SEPATH). Схема анализа сводится к следующему:

1) отношения между измеренными переменными, включенными в модель, определяются ковариационной матрицей. Структура этой матрицы определяется реальными отношениями между переменными;

2) гипотеза об отношении между переменными задается в форме диаграммы путей, описывающих представление исследователя о взаимодействиях («белый ящик»);

3) рассчитывается ковариационная матрица, учитывающая введенные ограничения, и сравнивается с общей матрицей;

4) программа диагностирует полученные соотношения и выводит различные критерии соответствия и линейную модель регрессии;

5) на основе этих критериев принимаются решения об изменении структуры (в случае необходимости).

Если в структуре системы заданы латентные (скрытые) переменные, то для них, исходя из общей схемы линейной алгебры, могут быть рассчитаны возможные значения.

Этот метод весьма эффективен, но его использование требует высокого уровня владения методами статистики в целом, специально поставленных и весьма осмысленных полевых исследований, поэтому ознакомление и работа с ним предполагают хорошее владение как теорией, так и навыком применения более простых методов статистического анализа. Читатель, зная о существовании столь эффективного метода, на определенном этапе сам выйдет на уровень знания и опыта, необходимых для его использования.

Глава 5
МНОГОМЕРНЫЙ ПАРАМЕТРИЧЕСКИЙ АНАЛИЗ

5.1. Метод главных компонент

В моделях регрессионного анализа рассматривались направленные или ориентированные системы, для которых существовали естественные предположения о входах и выходах. Более типично, когда система определяется переменными, для которых сформулировать такие гипотезы практически невозможно. Например, система определена через элементы, каждый из которых характеризуется обилием различных видов растений (типичная задача фитоценологии) или значениями концентраций каких-либо веществ (типичная задача геохимии), или мощностями горизонтов почв и их цветом по шкале Манселла (типичная задача почвоведения) и т.п. Конечно, мы предполагаем, что существуют какие-то внутренние или внешние факторы, определяющие состояния этих переменных в каждом элементе. Однако сами эти факторы не измерены. Если же и допускается измерение мыслимых факторов, например, определяющих обилие отдельных видов растений, то потребовалось бы построить очень большое число частных регрессионных моделей, и описание системы было бы чрезвычайно громоздким и мало информативным.

Вместе с тем из основных положений векторной алгебры следует, что если существует квадратная матрица, элементами которой являются ковариации или коэффициенты корреляции, а главная диагональ соответственно заполнена или дисперсиями, или (в случае корреляций) единицами, то всегда можно найти ее отображение в ортогональной системе координат.

Для того чтобы найти эти координаты, достаточно решить n совместных уравнений относительно нулевых значений функций, определив значения ортогональных векторов e_{ij}, т. е.

$$\begin{bmatrix} e_1\sigma_{11}^2 & \dots & e_nK_{1n} = 0; \\ \vdots & \ddots & \vdots \\ e_1K_{n1} & \cdots & e_n\sigma_{nn}^2 = 0. \end{bmatrix}$$

Задача не имеет решения, если все вектора матрицы параллельны друг другу и определитель равен нулю. Это будет просто означать, что существует только один вектор, описывающий все переменные. Решениями уравнений будут значения *n* координат, определяющих положение каждой переменной в векторном пространстве. Это решение может иметь физический смысл только в том случае, если распределения близки к нормальным, а отношения между переменными практически независимы. Анализ на основе алгебраического преобразования ковариационных или корреляционных матриц называется *методом главных компонент*: отображение векторов в системе независимых координат (компонент матрицы).

Рассмотрим все этапы применения метода главных компонент на конкретном примере.

Определение системы. На территории Европы на 117 станциях (рис. 5.1) в течение 17 лет измеряли концентрации ионов SO_4, NO_3, NH_4, Cl, Ca, K, Mg, Na в атмосферных осадках, а также их кислотность, кондуктивную проводимость и собственно сумму осадков. Элементами системы являются данные на каждой станции за каждый месяц с 1982 г. по 1998 г. Требуется выяснить факторы, определяющие пространственно-временно́е варьирование измеренных переменных. Распределения значений переменных по те-

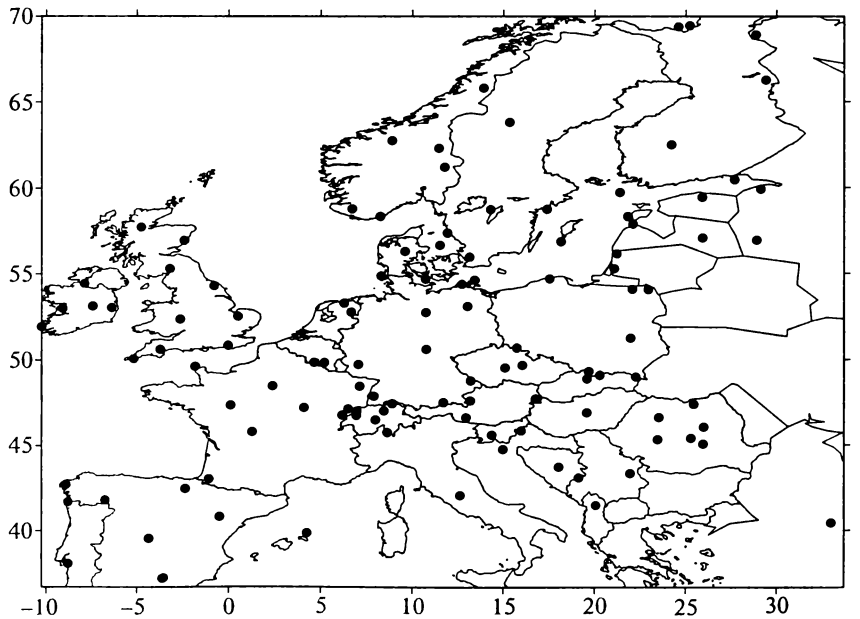


Рис. 5.1. Места расположения станций наблюдения за содержанием ионов в атмосферных осадках

Рис. 5.2. Матрица корреляций между парами переменных

сту Колмогорова — Смирнова за редким исключением близки к логнормальным. На рис. 5.2 и в табл. 5.1 показаны распределения точек в двухмерных пространствах переменных и белыми линиями — типы парных зависимостей. В основном зависимости можно считать близкими к линейным. Только кислотность имеет, скорее всего, нелинейную зависимость с некоторыми катионами, и очень слабо связаны со всеми переменными месячные осадки.

Корреляционная матрица является основой всех дальнейших расчетов.

Из табл. 5.2 следует, что вся совокупность переменных в наибольшей степени описывает концентрации Cl, Na, Mg, SO_4 и электрическую проводимость (COND). Практически независимы от всей системы осадки (мм) и слабо зависят от нее калий и кислотность (pH). Остальные ионы занимают промежуточное положение.

В табл. 5.3 приведены значения дисперсии каждого из 11 факторов и значение определителя всей матрицы в логарифмической форме по основанию 2. Определитель в логарифмической форме точно равен количеству информации как меры сопряженности, существующей в системе.

Разнообразие системы отношений может быть определено следующим образом

$$H = 11\log\pi e - \log\Delta = 21{,}57 \text{ бит.}$$

Таблица 5.1

Парные коэффициенты корреляции Пирсона между переменными (общий объем наблюдений — 8601)

Переменная	Переменная										
	pH	мм	SO_4	NO_3	NH_4	Cl	COND	Ca	K	Mg	Na
pH	1,00	0,00	–0,13	–0,29	–0,11	0,08	–0,22	0,37	0,15	0,24	0,13
мм	0,00	1,00	–0,28	–0,34	–0,25	–0,04	–0,25	–0,22	–0,19	–0,07	–0,05
SO_4	–0,13	–0,28	1,00	0,68	0,66	0,32	0,79	0,61	0,51	0,44	0,29
NO_3	–0,29	–0,34	0,68	1,00	0,74	0,05	0,57	0,40	0,25	0,14	0,02
NH_4	–0,11	–0,25	0,66	0,74	1,00	0,02	0,49	0,42	0,25	0,14	–0,01
Cl	0,08	–0,04	0,32	0,05	0,02	1,00	0,61	0,30	0,46	0,81	0,92
COND	–0,22	–0,25	0,79	0,57	0,49	0,61	1,00	0,46	0,51	0,60	0,55
Ca	0,37	–0,22	0,61	0,40	0,42	0,30	0,46	1,00	0,54	0,57	0,28
K	0,15	–0,19	0,51	0,25	0,25	0,46	0,51	0,54	1,00	0,53	0,48
Mg	0,24	–0,07	0,44	0,14	0,14	0,81	0,60	0,57	0,53	1,00	0,79
Na	0,13	–0,05	0,29	0,02	–0,01	0,92	0,55	0,28	0,48	0,79	1,00

Таблица 5.2

Совместное определение всеми переменными каждой отдельно взятой. Коэффициент детерминации R^2

Переменная	R-Square	Переменная	R-Square
pH	0,478644	Cl	0,874160
мм	0,146305	COND	0,805135
SO_4	0,785895	Ca	0,691367
NO_3	0,691077	K	0,452417
NH_4	0,619291	Mg	0,796927

Эти дополнительные оценки не имеют прямого отношения к методу главных компонент, но они полезны для измерения разнообразия отношений во всей системе. Суммарная дисперсия равна числу факторов, что указывает на представление варьирования ком-

Таблица 5.3

Расчет дисперсий по факторам. Собственные значения

Extraction: Principal components $\log \Delta = -12{,}409$ бит

Номер фактора (компоненты)	Дисперсия Eigenval	Дисперсия % Variance, % total	Накопленная дисперсия Cumul Eigenval	Накопленная дисперсия, % Cumul, %
1	4,764851	43,31683	4,76485	43,3168
2	2,347301	21,33910	7,11215	64,6559
3	1,307904	11,89004	8,42006	76,5460
4	0,848509	7,71372	9,26857	84,2597
5	0,552538	5,02307	9,82110	89,2828
6	0,365275	3,32068	10,18638	92,6034
7	0,265058	2,40962	10,45144	95,0131
8	0,205344	1,86676	10,65678	96,8798
9	0,140578	1,27798	10,79736	98,1578
10	0,127313	1,15739	10,92467	99,3152
11	0,075329	0,68481	11,00000	100,0000

понент в стандартизированном виде. Как следует из табл. 5.3, первые пять компонент описывают варьирование всей системы на 90 %. На остальные шесть компонент приходится лишь 10 %.

Возникает естественный вопрос: сколько компонент может быть признано достаточно значимым для включения их в модель? Или иначе, сколько ортогональных факторов описывает существенную часть варьирования всех переменных?

Как следует из чисто алгебраических построений, если все переменные независимы, то число главных компонент должно быть точно равно числу переменных. Если все переменные однозначно связаны друг с другом, то существует только одна компонента. В реальной системе истина лежит где-то посередине. Один из методов определения достаточного числа компонент или размерности пространства переменных опирается на правило увеличения дисперсии с уменьшением номера фактора как функции вида

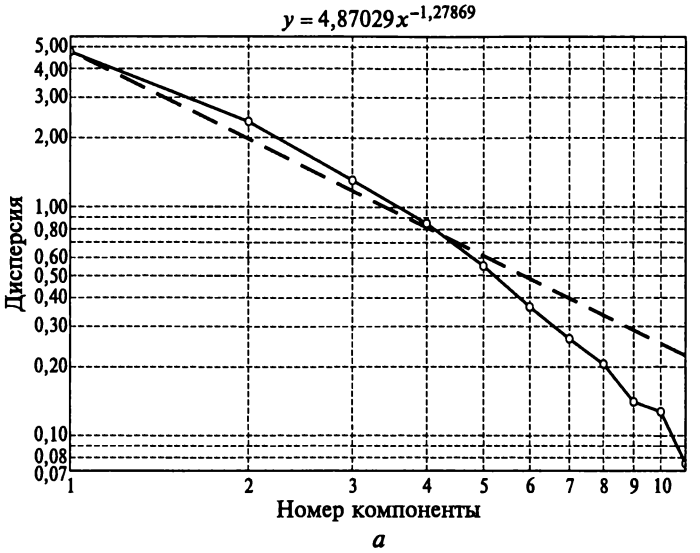
$$\sigma^2 = a(\mathrm{NPCA})^b,$$

где NPCA — номер главной компоненты при чисто случайном отношении между переменными; a, b — параметры уравнения.

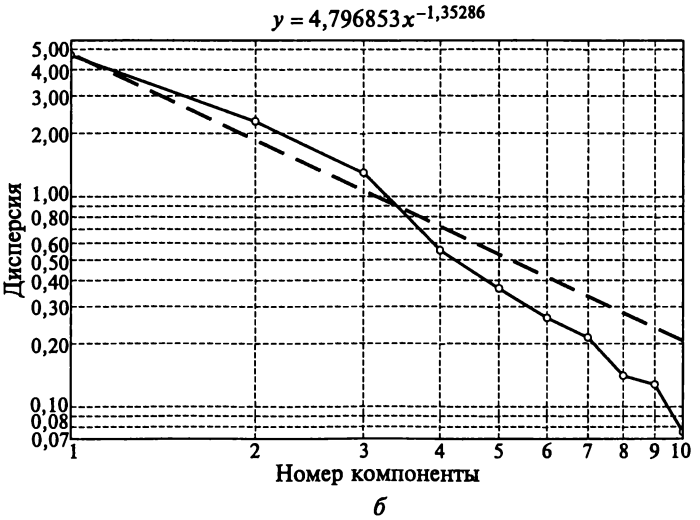
Вид этой теоретической зависимости можно получить, аппроксимируя реальные значения нелинейной моделью. Эту задачу решают методом нелинейной регрессии. На рис. 5.3 показано соотношение между моделью, описывающей случайный процесс, и реальными значениями дисперсий, полученных методом главных компонент. Точка пересечения двух графиков показывает оптимальную размерность пространства всех переменных, которая равна четырем ортогональным координатам. (Если исключить из анализа почти независимые от всех переменных атмосферные осадки, то размерность становится равной 3.) Можно полагать, что эти четыре координаты определяют основную часть варьирования. В соответствии с табл. 5.3 четыре первые компоненты описывают 84 % варьирования всех переменных.

Коэффициенты в табл. 5.4 можно трактовать по-разному. Чаще всего их называют нагрузками на компоненту. Они вычисляются из исходной корреляционной матрицы и фактически являются коэффициентами корреляции между независимыми компонентами и переменными. Так как они рассчитаны при стандартизованных значениях средних квадратических, они автоматически являются коэффициентами, отражающими косинус угла между переменной и координатой, т.е. являются коэффициентами чувствительности переменной к координате или к некоторому, возможно связанному с ней, физическому фактору. В целом же совокупность коэффициентов описывает линейную регрессионную модель каждой переменной от 11 координат.

Если переменные имеют близкие значения коэффициентов, то они положительно коррелируют друг с другом. Таким образом, из исходной корреляционной матрицы рассчитана система уравнений относительно 11 ортогональных компонент.



а



б

Рис. 5.3. Оценка достаточного числа факторов в модели метода главных компонент:

а — с осадками; *б* — без осадков

Таблица 5.4

Нагрузки на компоненты или координаты переменных в пространстве главных компонент (Factor Loadings)

Extraction: Principal components

Переменная	Номер фактора (компоненты)										
	1	2	3	4	5	6	7	8	9	10	11
pH	0,049306	–0,447637	**0,819387**	0,033626	0,202946	0,166305	0,154638	–0,174941	–0,033080	0,010631	–0,009771
мм	–0,314847	–0,327400	–0,189679	**0,866878**	–0,054826	0,020989	–0,010815	–0,050704	0,004750	0,007270	0,003822
SO_4	**0,821285**	0,393545	0,001128	0,121660	–0,064414	–0,155377	0,268919	–0,038063	0,084002	–0,214147	–0,028945
NO_3	0,578689	**0,696269**	–0,070592	0,052687	0,119963	0,059615	–0,263434	–0,289798	–0,024589	–0,024543	–0,006075
NH_4	0,548939	**0,650408**	0,104447	0,198638	0,215825	0,345038	0,023284	0,242008	0,000719	0,022691	0,003464
Cl	**0,702816**	–0,575615	–0,300858	–0,065877	0,137328	0,078951	–0,034835	–0,005423	0,110575	0,043603	–0,198493
COND	**0,872505**	0,126058	–0,299758	0,015152	0,003222	–0,089475	0,232033	–0,079762	–0,106435	0,226692	0,042714
Ca	**0,713648**	0,042728	0,551149	0,152610	–0,013740	–0,309159	–0,161628	0,099661	0,136127	0,105561	0,018123
K	**0,709050**	–0,146579	0,183038	–0,022292	–0,624714	0,215934	–0,055658	–0,004103	–0,040018	0,000401	–0,012279
Mg	**0,789640**	–0,481593	0,010262	0,043986	0,148871	–0,125812	–0,114752	0,105095	–0,258840	–0,115262	0,002973
Na	**0,679540**	**–0,612019**	–0,258639	–0,092604	0,108332	0,134742	–0,036451	–0,027988	0,145307	–0,049519	0,180610
Дисперсия	4,764851	2,347301	1,307904	0,848509	0,552538	0,365275	0,265058	0,205344	0,140578	0,127313	0,075329
Доля	0,433168	0,213391	0,118900	0,077137	0,050231	0,033207	0,024096	0,018668	0,012780	0,011574	0,006848

Примечание. Полужирным шрифтом выделены ведущие компоненты для каждой переменной.

В целом модель сводится к следующему: первый фактор в существенной степени описывает концентрации SO_4, Cl, Ca, Mg, K, Na, электрическую проводимость, второй — концентрации анионов азота и натрия. Кислотность фактически описывается совершенно независимой координатой, которая с достаточным уровнем описывает также только некоторую часть варьирования Ca. Месячные суммы осадков вообще связаны с собственной координатой, практически не влияющей на остальные переменные.

Первый напрашивающийся вывод: пересмотреть определение системы и исключить из нее осадки, как мало связанные с остальными переменными.

На рис. 5.3, *б* приведен график изменения дисперсии как функции номера компоненты для второго варианта определения системы, т. е. без осадков. Здесь перегиб в значениях дисперсий выражен резче, чем в первом случае (см. рис. 5.3, *а*), и размерность пространства несколько ниже. Однако все-таки ее логично принять равной четырем.

Определитель этой новой системы $\log\Delta = -12{,}181$ почти не отличается от определителя системы с включенными месячными осадками. Малое отличие определителя вполне естественно, так как из анализа исключена практически независимая переменная.

Разнообразие системы при этом варианте: $H = 10\log\pi e - 12{,}181 =$ $= 18{,}71$ бит, что существенно меньше разнообразия первого варианта представления системы.

Из табл. 5.5 следует, что первый фактор определяет варьирование большинства переменных, и в целом исключение осадков существенно не изменило отношений переменных к факторам. Несколько выросло влияние четвертого фактора на калий, что в частности делает целесообразным рассматривать пространство из четырех факторов. Кислотность, так же как в первом варианте, согласуется по третьему фактору с кальцием. Весьма характерно, что натрий и хлор зависят практически тождественно от всех первых четырех факторов, что прямо указывает на генетическое единство изменчивости их концентраций. Анионы азота также имеют подобную зависимость при втором ведущем факторе и существенной роли первого.

Положение переменных в ортогональном векторном пространстве демонстрирует рис. 5.4. В сущности здесь в графической форме отображено их положение относительно друг друга, показанное в табл. 5.5.

Теперь желательно рассчитать значения факторов для каждого элемента системы. Эти значения нам неизвестны, но зато известны коэффициенты линейных уравнений регрессий между известными значениями переменных и неизвестными значениями факторов для каждого элемента системы (станция, год, месяц наблюдений) (табл. 5.6).

Таблица 5.5

Нагрузки на компоненты или координаты переменных в пространстве главных компонент (Factor Loadings)

Extraction: Principal components

Переменная	Номер фактора (компоненты)									
	1	2	3	4	5	6	7	8	9	10
pH	0,055191	–0,471995	**0,805358**	–0,199814	–0,170143	–0,149070	–0,181872	–0,034839	–0,008653	–0,010533
SO_4	**0,812337**	0,427309	0,032791	0,055981	0,159341	–0,269076	–0,037968	0,080458	0,216271	–0,028765
NO_3	0,557533	**0,715958**	–0,042042	–0,117602	–0,065801	0,272228	–0,284209	–0,028829	0,029309	–0,007770
NH_4	0,533633	**0,678249**	0,148662	–0,227730	–0,338081	–0,034109	0,255194	0,003189	–0,025303	0,004088
Cl	**0,721555**	–0,547836	–0,316071	–0,134171	–0,080684	0,035334	–0,006482	0,110730	–0,041326	–0,198903
COND	**0,871162**	0,162760	–0,285779	–0,004376	0,089254	–0,228812	–0,091016	–0,104019	–0,227424	0,041840
Ca	**0,710938**	0,057445	**0,572252**	0,004259	0,312282	0,159047	0,102204	0,138481	–0,104973	0,017854
K	**0,711075**	–0,131808	0,183093	**0,627130**	–0,213257	0,053533	0,005312	–0,040638	–0,000003	–0,012638
Mg	**0,806050**	–0,452411	0,009294	–0,154224	0,128287	0,110675	0,116366	–0,259318	0,110570	0,004186
Na	**0,697783**	–0,589181	–0,278704	–0,102303	–0,137957	0,038295	–0,033122	0,144920	0,050736	0,180697
Дисперсия	**4,685787**	2,277087	1,293873	0,554332	0,365932	0,265376	0,213856	0,140686	0,127596	0,075476
Доля	0,468579	0,227709	0,129387	0,055433	0,036593	0,026538	0,021386	0,014069	0,012760	0,007548

Примечание. Ведущие компоненты для каждой переменной выделены полужирным шрифтом.

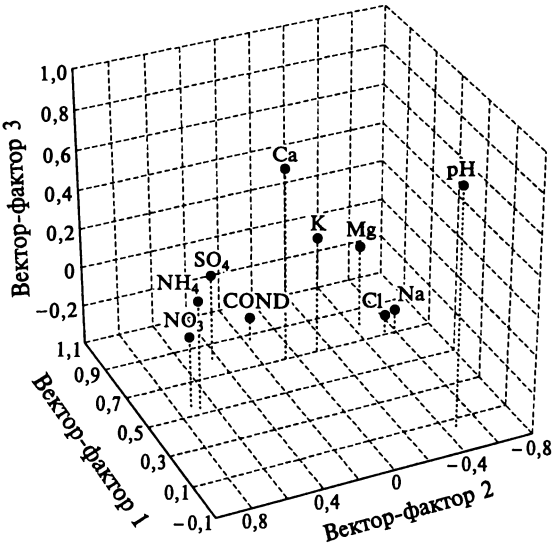


Рис. 5.4. Положение переменных в ортогональном векторном пространстве

В табл. 5.6 Y_j^i — измеренные величины концентраций иона (переменной j) на конкретной станции в конкретный год и месяц. В таблицу можно ввести из исходных данных конкретные их значения. Коэффициенты α_j^{Fi} — коэффициенты чувствительности или коэффициенты уравнения регрессии, полученные в рамках метода главных компонент из корреляционной или ковариационной матриц; X_i — неизвестное значение координаты i конкретного элемента в пространстве Евклида размерности n (в данном случае $n = 4$). Очевидно, что из десяти уравнений с 50 известными значениями можно определить четыре неизвестных. Таким образом, можно рассчитать значения координат для всех элементов (точек наблюдения). Так как подразумевается, что существуют некоторые факторы, определяющие неслучайное отношение измеренных переменных друг с другом, то можно надеяться, что эти координаты отражают какие-то физические факторы, определяющие пространственно-временну́ю флуктуацию измеренных переменных. При этом, по условию, четыре фактора должны с достаточной полнотой описывать варьирование переменных в пространстве-времени. Если взять 10 факторов, то это описание при нормальном распределении и линейных отношениях между переменными по условию будет абсолютным.

В рамках программных средств метода главных компонент значения координат для каждого элемента рассчитываются автоматически. При этом по условию каждая координата или фактор будет

Таблица 5.6

Значения факторов для каждого элемента системы

Переменная	Значение Y_j^i для элемента i	Фактор 1		Фактор 2		Фактор 3		Фактор 4	
		α_j^1	X_1	α_j^2	X_2	α_j^3	X_3	α_j^4	X_4
pH	Y_1^i	0,055191	X_1	−0,471995	X_2	**0,805358**	X_3	−0,199814	X_4
SO_4	Y_2^i	0,**812337**	X_1	0,427309	X_2	0,032791	X_3	0,055981	X_4
NO_3	Y_3^i	0,557533	X_1	**0,715958**	X_2	−0,042042	X_3	−0,117602	X_4
NH_4	Y_4^i	0,533633	X_1	**0,678249**	X_2	0,148662	X_3	−0,227730	X_4
Cl	Y_5^i	**0,721555**	X_1	−0,547836	X_2	−0,316071	X_3	−0,134171	X_4
COND	Y_6^i	**0,871162**	X_1	0,162760	X_2	−0,285779	X_3	−0,004376	X_4
Ca	Y_7^i	**0,710938**	X_1	0,057445	X_2	**0,572252**	X_3	0,004259	X_4
K	Y_8^i	**0,711075**	X_1	−0,131808	X_2	0,183093	X_3	**0,627130**	X_4
Mg	Y_9^i	**0,806050**	X_1	−0,452411	X_2	0,009294	X_3	−0,154224	X_4
Na	Y_{10}^i	**0,697783**	X_1	−0,589181	X_2	−0,278704	X_3	−0,102303	X_4

Примечание. Ведущие компоненты для каждой переменной выделены полужирным шрифтом.

определять в первую очередь те же переменные, что и в векторном пространстве.

Необходимо сделать важное замечание: если расчет параметров векторного пространства осуществляется на основе матрицы корреляции, то полученные значения координат стандартизированы по среднему квадратическому отклонению, т.е. имеют средние, равные нулю, и дисперсию, равную единице.

Если расчет параметров векторного пространства проведен на основе матрицы ковариации, то координаты не стандартизованы и дисперсия каждой соответствует ее весу в описании варьирования переменных при нулевом значении средних. Для того чтобы перейти от стандартизованных координат к нестандартизованным, достаточно каждую из них умножить на соответствующее ей значение среднего квадратического, вычисляемого из соответствующих значений факторных нагрузок (дисперсии).

Физический смысл координат или факторов раскрывается при сопоставлении их с независимыми переменными. Естественно полагать, что концентрации ионов есть функции года и месяца наблюдений, поэтому можно допустить, что они связаны с высотой местности и, конечно, должны зависеть от географического положения точки наблюдения. Провести такого рода оцен-

ку можно в простейшем случае на основе дисперсионного анализа.

Исследуем, как наши координаты или факторы связаны с годом наблюдения. Известно, что в Европе были предприняты значительные усилия по снижению выбросов, и если они имели какие-либо последствия, то по крайней мере значения хотя бы одного фактора должны быть связаны с годом наблюдения.

Из табл. 5.7 следует, что в соответствии с интегральным тестом Вилкоксон-ламбда существует гипотеза независимости между переменными для каждого элемента и факторами (компонентами), изменчивость последних минимально связана с годом наблюдения и месячными суммами осадков, средней — с месяцем наблюдений и в наибольшей степени — с высотой станции над уровнем моря. Тест Вилкоксон-ламбда строится на отношении матриц вторых моментов в группах и в целом в выборке и изменяется от 0 (зависимость абсолютная) до 1 (зависимость отсутствует). Приближенно он равен единице минус коэффициент детерминации, т.е. $1 - R^2$.

Отметим, что для возможности исследования связи высоты станции наблюдения с факторами данные высот были преобразованы по следующей формуле:

$$H_{\text{tr}} = \text{trunc}[\ln(H_{\text{н.у.м}}^*/100)].$$

Таким образом, получаем дискретное отображение высот. Операция trunc означает: принять с точностью до целого. Тем же методом были выделены и классы осадков.

С другой стороны, как следует из критерия Фишера, первый фактор (общее загрязнение) в наибольшей степени определяется высотой, месяцем года и собственно годом наблюдения; второй фактор имеет наиболее выраженный сезонный ход и тесно связан с высотой и суммой осадков; третий — с высотой над уровнем моря и сезонным ходом, а четвертый — с высотой и годом наблюдения.

Рассмотрим последовательно возможные механизмы пространственно-временно́й изменчивости трех основных факторов, первый из которых определяет основную часть концентрации анионов и, в первую очередь, SO_4, а также общую минерализацию, второй — NH_4 и NO_3 и в существенной степени Na и Cl, третий — pH и, во многом, содержание катиона Ca.

Для отображения отношений будем использовать графики бокс-плот и аппроксимировать отношения методом наименьших квадратов. Обычно эта процедура аппроксимации подключена к модулю построения графиков.

Из рис. 5.5, *а* с полным основанием следует, что за рассматриваемый период общая минерализация осадков уменьшилась при-

---

* $H_{\text{н. у. м}}$ — высота над уровнем моря.

Таблица 5.7

Одновариантный дисперсионный анализ отношения факторов к году и месяцу наблюдений, абсолютной высоте над уровнем моря и суммам осадков

Независимая переменная Wilks' Lambda= 1 – R^2	Номер фактора	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F-критерий	p-уровень
Год	1	359,3539	16	22,45962	8240,646	8584	0,960001	**23,39542**	0,000000
	2	99,5286	16	6,22054	8500,471	8584	0,990269	6,28166	0,000000
	3	87,6541	16	5,47838	8512,346	8584	0,991653	5,52450	0,000000
0,911176	4	233,0778	16	14,56736	8366,922	8584	0,974711	14,94531	0,000000
Месяц	1	320,4650	11	29,13318	8279,535	8589	0,963970	30,22209	0,000000
	2	698,6921	11	63,51746	7901,308	8589	0,919933	**69,04572**	0,000000
	3	394,8061	11	35,89146	8205,194	8589	0,955314	37,57032	0,000000
0,828862	4	117,4142	11	10,67402	8482,586	8589	0,987610	10,80793	0,000000
Высота над уровнем моря	1	4664,480	107	43,59328	3935,520	8493	0,463384	94,0759	0,00
	2	4785,815	107	44,72724	3814,185	8493	0,449098	99,5936	0,00
	3	5575,388	107	52,10643	3024,612	8493	0,356130	**146,3130**	0,00
0,613236	4	3970,060	107	37,10336	4629,940	8493	0,545148	68,0611	0,00
Сумма осадков	1	506,9193	5	101,3839	8093,081	8595	0,941603	**107,6715**	0,000000
	2	456,5814	5	91,3163	8143,419	8595	0,947460	96,3801	0,000000
	3	53,1266	5	10,6253	8546,873	8595	0,994401	10,6851	0,000000
0,881638	4	17,3019	5	3,4604	8582,698	8595	0,998569	3,4653	0,003944

Примечание. Полужирным шрифтом выделены ведущие компоненты.

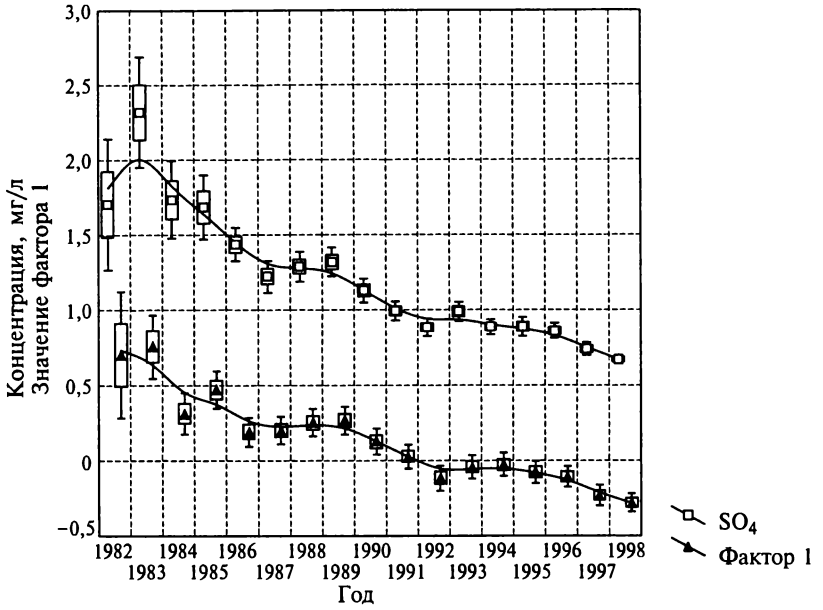
мерно в два раза по SO_4, а по общему загрязнению, индуцируемому первым фактором, даже несколько больше. Очевидно, что выявленная статистическая связь вполне значима. Это позволяет полагать, что фактор 1 сам по себе есть функция хозяйственной активности. То что это действительно так, в полной мере подтверждает максимальное значение фактора в марте, т.е. в конце отопительного сезона (рис. 5.5, *б*).

Концентрация ионов зимой в среднем в 1,5 раза больше, чем летом. При этом фактор, обобщая сезонную динамику многих элементов, демонстрирует более четкую зависимость, чем собственно концентрация SO_4.

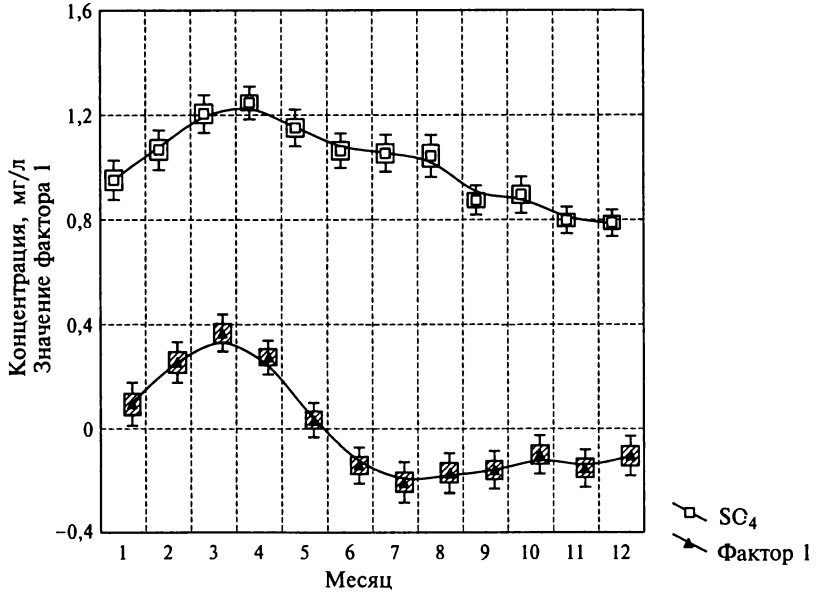
Высота над уровнем моря в наибольшей степени определяет варьирование первого фактора в пространстве (рис. 5.5, *в*). Общая тенденция сводится к заметному снижению значения фактора 1 и соответственно концентраций с высотой. В наиболее высоких точках наблюдения концентрация ионов в 4,5 раза меньше, чем на низменностях и равнинах (максимальная высота станции 3500 м н.у.м). Такое соотношение вполне объяснимо: осадки «вымывают» химические соединения из атмосферы и на низких высотах, где размещается основная промышленность, их концентрация в осадках существенно больше. Значительных высот в атмосфере достигает лишь относительно малая часть продуктов техногенеза, и осадки на больших высотах существенно чище. Вместе с тем график демонстрирует существование статистически значимой волновой структуры, что, возможно, отражает высотные уровни рельефа и среднестатистические локальные преобразования воздушных масс.

Связь месячных сумм осадков (рис. 5.5, *г*) с первым фактором весьма проста — чем больше сумма осадков, тем меньше концентрация. Фактически это тот же эффект «разбавления», который был получен для концентраций ионов в речном стоке. Следует отметить, что при очень большой сумме осадков варьирование концентраций несколько увеличивается, но это не меняет принципиальную схему зависимости. Влияние осадков и высоты над уровнем моря на значение первого фактора вполне самостоятельно можно показать на основе их частных корреляций, рассчитав в программе множественной регрессии. Частные корреляции высоты над уровнем моря и суммы месячных осадков (мм) в отношении значений первого фактора соответственно: –0,179089 при t-критерии 16,8239 и –0,201342 при t-критерии 18,9977, что указывает на независимую их сопряженность с первым фактором.

По множеству значений в каждой точке наблюдения с помощью крейгинг-метода можно построить карту изменения значения фактора 1 в пространстве (рис. 5.6). Методы построения таких карт рассматриваются в рамках особого направления анализа данных «геостатистика». Варьирование первого фактора в пространстве хорошо выделяет светлым тоном основные промышленные

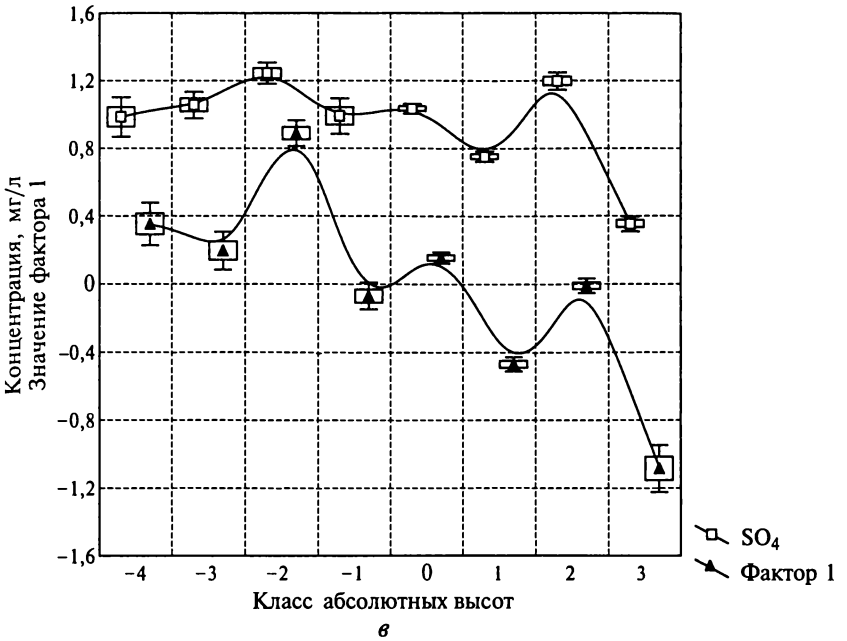


а

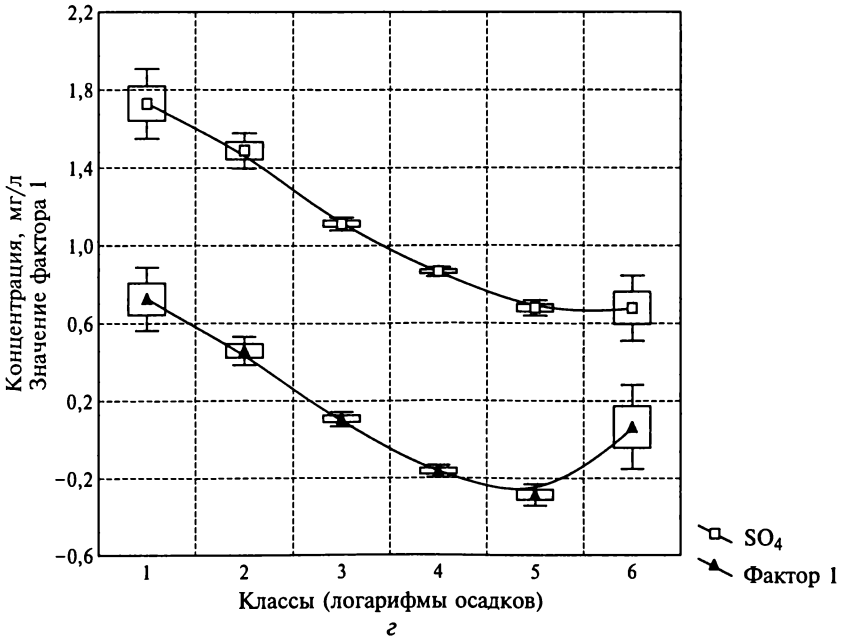


б

Рис. 5.5. Связь фактора 1



в



г

с независимыми переменными (*а—г*)

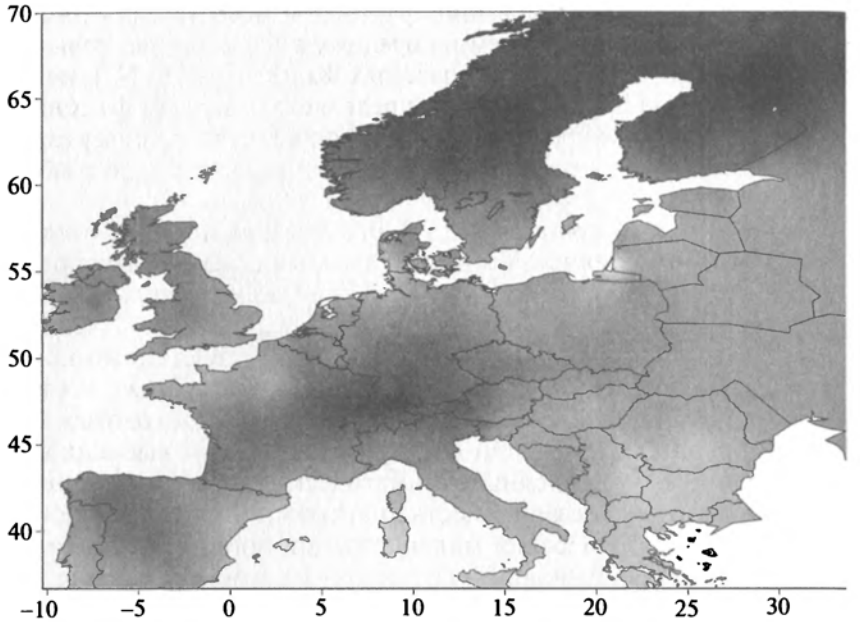


Рис. 5.6. Варьирование фактора 1, отражающего общее содержание ионов в атмосферных осадках на территории Европы (светлые тона — высокое значение фактора)

регионы, а темным — горные массивы и возвышенности. По-видимому, определенное значение имеет и западный перенос, так как осадки на территории, прилегающей к Атлантике, не имеют экстремально высоких уровней загрязнения.

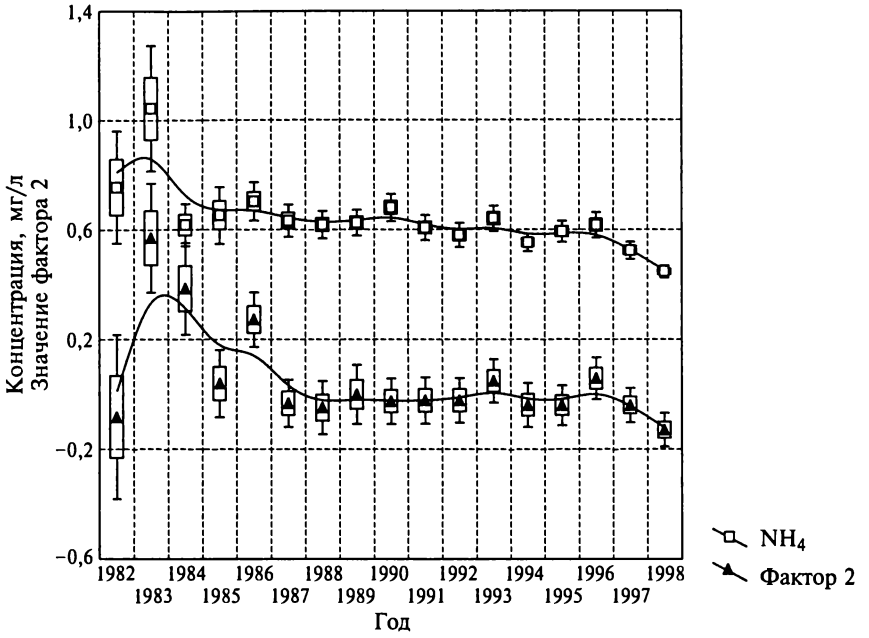
Таким образом, получаем, что **фактор 1** исследованной системы обобщает интенсивность поступления ионов и интенсивность их выноса с осадками из атмосферы. Поступление в атмосферу во многом определяется антропогенным воздействием, а вынос — суммой осадков и длительностью их воздействия на воздушную массу. Воздушные массы, достигающие значительных высот над уровнем моря, уже в существенной степени очищены осадками и концентрация продуктов техногенеза в них невелика.

Фактор 2, как следует из факторных нагрузок, ассоциируется в первую очередь с концентрациями NO_3 и NH_4. В наибольшей степени этот фактор определяется сезонностью. Действительно, на протяжении всего периода наблюдений изменения значений этого фактора не существует, но в сезонном ходе его максимум приходится на май—июль (рис. 5.7). Вместе с тем концентрация NH_4 фактически имеет два сезонных максимума — зимний и летний, что заставляет предполагать генетическую неоднородность эмиссий соединений азота. С высотой над уровнем моря связь фактора

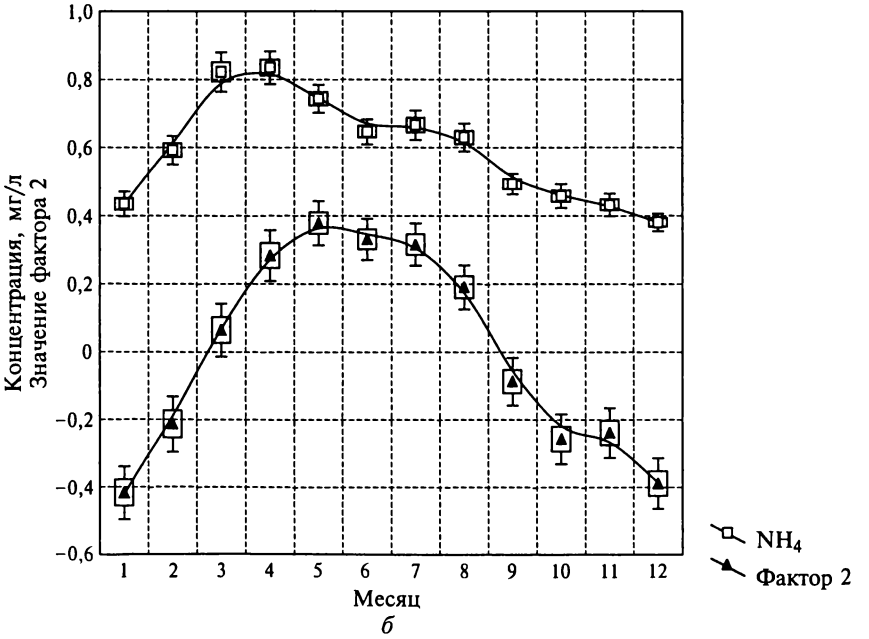
весьма неопределенна: значение фактора велико на малых высотах, потом понижается, затем на средних высотах вновь повышается и минимально на больших высотах. Концентрации NH_4 испытывают колебания, скорее всего определяемые первым фактором. Сумма месячных осадков вполне однозначно связана с фактором 2 и демонстрирует в целом рассмотренное выше правило разбавления.

Итак, генетическая природа второго фактора не столь очевидна, как первого. Среднемноголетнее пространственное варьирование значения фактора показывает, что максимумы приходятся на средние широты с умеренными температурами летом и относительно большим количеством осадков (рис. 5.8). В Средиземноморье и на севере Европы его значения существенно ниже. Сезонность фактора, практическая неизменность его среднегодовых значений за период наблюдений и приуроченность его высоких значений к вполне определенным климатическим условиям заставляет предполагать его независимость от антропогенной активности и связывать его с процессами минерализации органического вещества. Последнее предположение исходит из того, что просто нет других мыслимых источников эмиссии соединений азота.

Логика этого вывода в сущности отражает и сами возможности генетической интерпретации отношений, полученных в результате статистического анализа, который показал существование высокой связи этого фактора с сезоном. Связь с климатическими условиями лета сформулирована на уровне гипотезы. В рамках статистических методов ее, конечно, можно проверить, исследовав статистические отношения между фактором с одной стороны и гидротермическими условиями (в конкретных условиях измерения или в среднемноголетнем) — с другой. Но никакой прямой информации собственно о механизмах поступления соединений азота в атмосферу в этих отношениях не будет. Подтвердить или отвергнуть эту гипотезу можно только на основе специальных измерений концентраций соединений азота в атмосфере на серии станций, удаленных от прямого воздействия хозяйственной деятельности, для различных типов ландшафта и климатических условий. Однако, даже если после проведения этих специальных измерений будет доказано, что эмиссия соединений азота в атмосферу не зависит от человека и связана с гидротермическим режимом и свойствами ландшафта и экосистемы, даже если на основе прямых измерений будут выявлены типы среды, дающие максимальные эмиссии, механизмы действия этого фактора останутся все равно не доказаны. Доказательство возможно только в том случае, если будут выявлены микроорганизмы, ответственные за выделение соединений азота, и доказано их прямое влияние на концентрацию ионов азота в атмосфере. Следовательно, статистический анализ чаще дает основания для

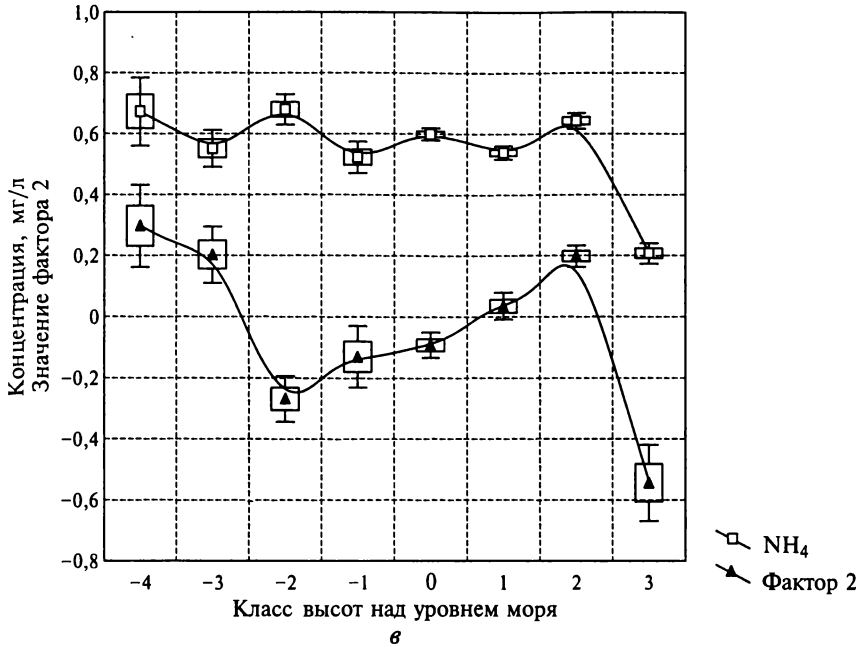


а

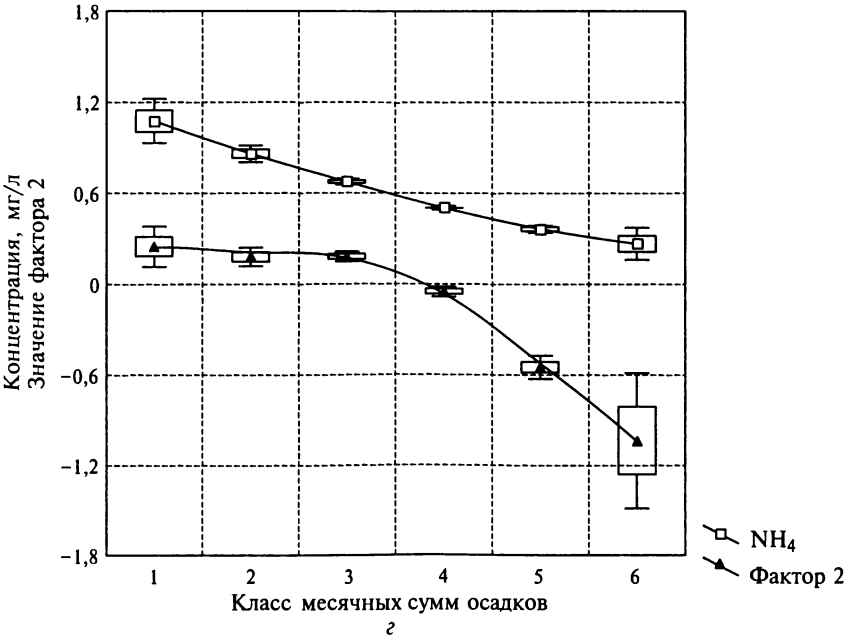


б

Рис. 5.7. Связь фактора 2



в



г

с независимыми переменными (*а*—*г*)

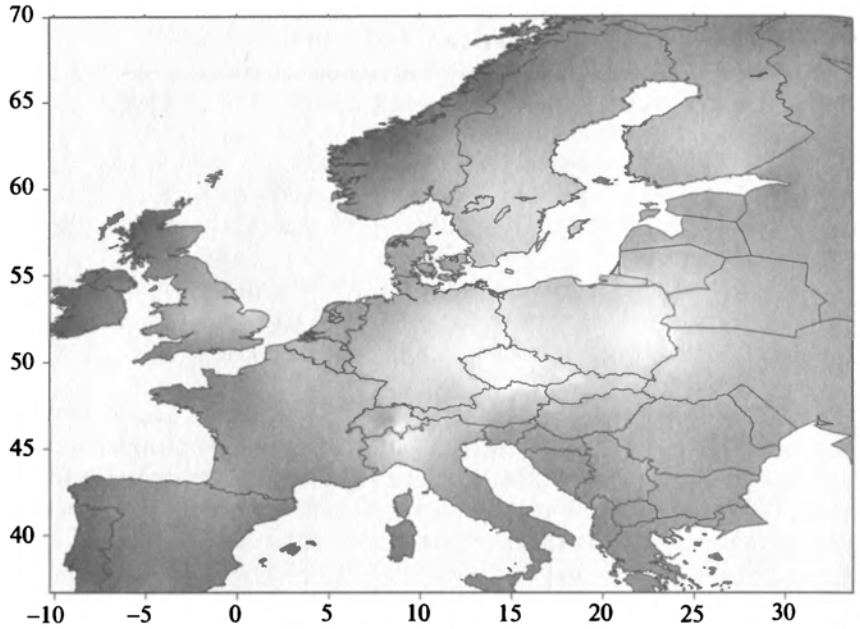


Рис. 5.8. Варьирование фактора 2, отражающего общее содержание ионов в атмосферных осадках на территории Европы (светлые тона — высокое значение фактора)

постановки новых вопросов и реже способен дать прямые ответы на вопросы о механизмах, лежащих в основе измеренных отношений.

Фактор 3 связывается в первую очередь с пространственно-временны́м варьированием pH и катиона кальция. На основе всего массива данных (8601 измерение) можно утверждать, что среднее значение pH равно 4,982395±0,007046 и математическое ожидание с вероятностью 0,95 лежит в интервале от 4,968583 до 4,996208, минимальное наблюдавшееся значение pH — 3,3, а максимальное — 7,24 при стандартном отклонении 0,653489.

В соответствии с общими представлениями о кислотных дождях и их связи с антропогенной активностью можно было бы ожидать, что pH определяется первым фактором, индуцирующим общее загрязнение осадков. Однако его пространственно-временно́е варьирование практически не зависит от года наблюдения. Фактор 3 связывается в большей степени с высотой над уровнем моря и в меньшей — с сезонностью. Действительно, как следует из рис. 5.9, *а—г*, существует слабый временно́й тренд, увеличивающий его значения и указывающий на наличие очень слабого многолетнего уменьшения кислотности осадков. В 1982 г. средняя кислотность по 31 наблюдению составляла 4,799677±0,114895

и в 1998 г. — 4,997053±0,017041. Хотя различия средних статистических достоверны, но очевидно, что они невелики.

Максимум значений фактора 3 приходится на май—август. Однако кислотность имеет два локальных максимума: в апреле и июле, при минимуме зимой.

Начиная со средних высот значение фактора 3 резко возрастает и pH на больших высотах достоверно выше, чем над уровнем моря: на уровне моря — 4,737541±0,02161, на высоте 3000 м н.у.м. — 5,819667±0,051279.

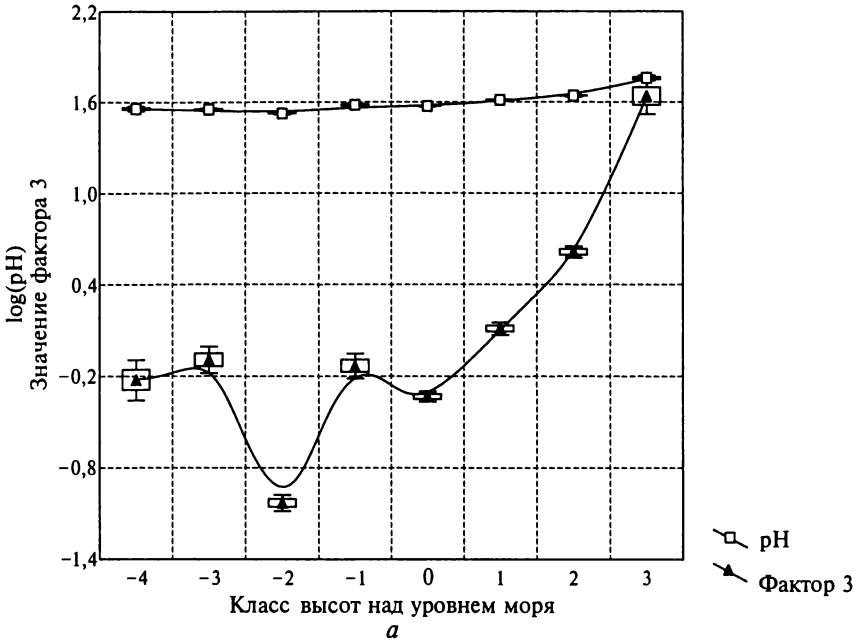
Наконец, чем больше сумма осадков, тем меньше значение фактора, меньше концентрация катиона кальция, наиболее связанного с кислотностью, и несколько выше сама кислотность (меньше значение pH).

На карте, представленной на рис. 5.10, видно, что значения фактора 3 в среднем существенно выше в Средиземноморье и особенно в средиземноморском секторе Испании, в бассейне Адриатики, на значительной части Альп и Карпат. Относительно высокие значения фактора характерны и для стран Прибалтики. Соответственно в выделенных регионах кислотность осадков наименьшая, а в большинстве стран Западной Европы, включая Германию (без Альп), Польшу, Чехию и Словакию (без Татр), отмечена наибольшая.

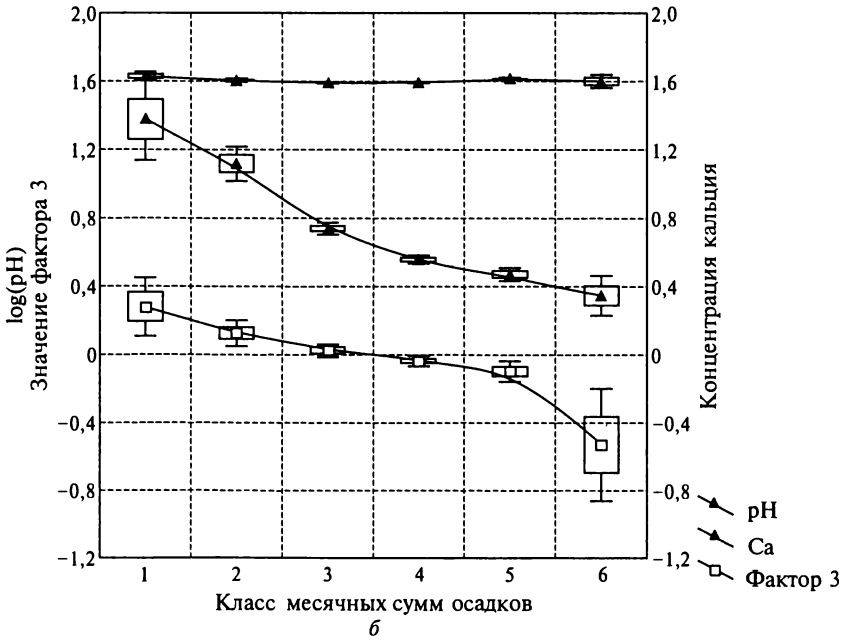
Какова возможная интерпретация третьего фактора? Подберем логические основания для формулировки гипотезы:

- варьирование pH частично связано с варьированием катиона кальция, что указывает на роль последнего в подщелачивании осадков;
- достоверно существует летний максимум, что можно связать с особенностью летней циркуляции атмосферы;
- достоверно больше значения на больших высотах, что может отражать связь с эффектами дальнего переноса;
- достоверно меньше значения при увеличении месячных осадков, что можно связать с разбавлением концентрации кальция и снижением эффекта подщелачивания;
- очень слабое уменьшение кислотности во времени указывает на некоторую, но слабую, зависимость от хозяйственной деятельности;
- достаточно строгая приуроченность больших значений фактора к бассейну Средиземного моря.

Из вышеизложенного вытекает единственно реалистичное предположение: фактор 3 в основном индуцируется выносом воздушных масс, насыщенных пылью и обогащенных кальцием с севера Африки. Однако этой гипотезе «дальнего трансграничного переноса» противоречит локальное увеличение значений третьего фактора в Прибалтике. Возможно, что этот фактор имеет полигенетическую природу. С другой стороны, если бы существовали анало-

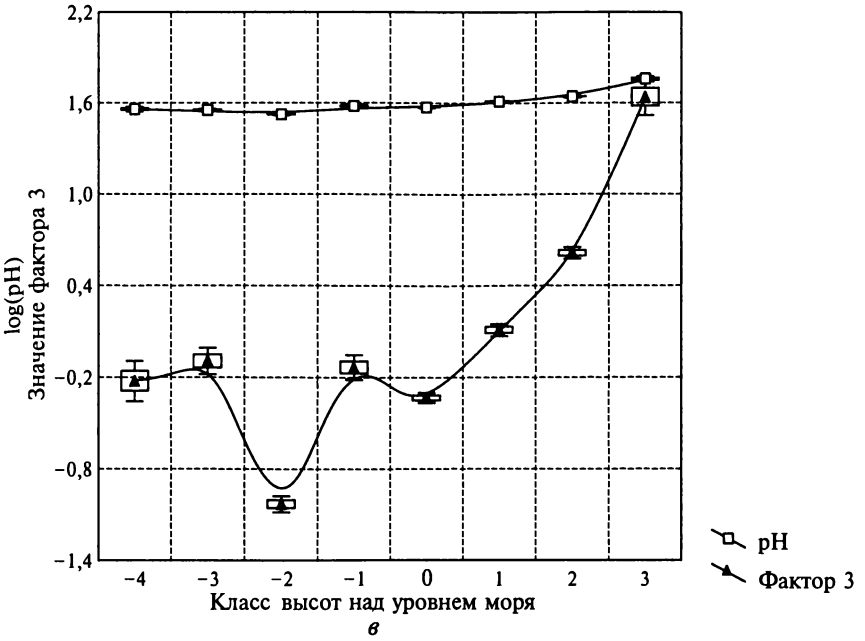


а

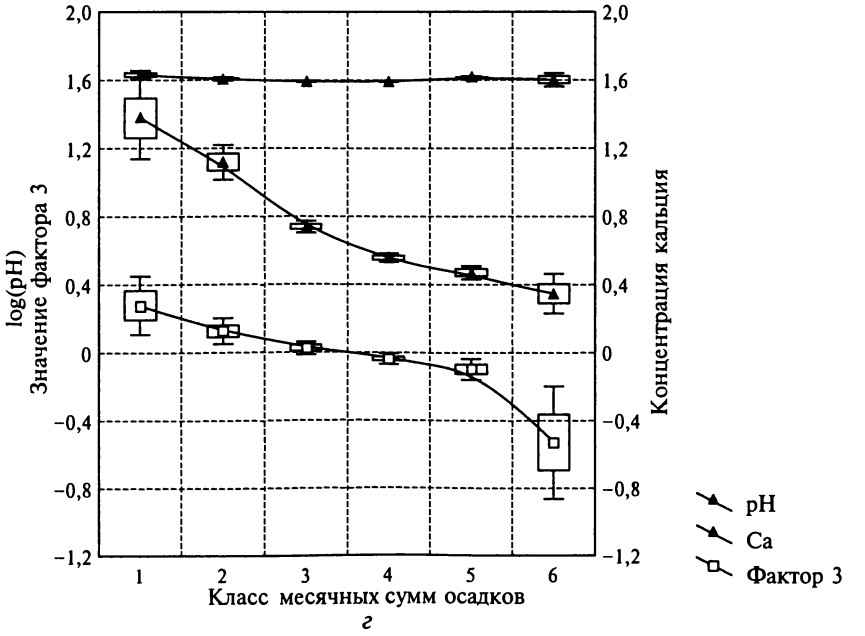


б

Рис. 5.9. Связь фактора 3



в



г

с независимыми переменными (*а—г*)

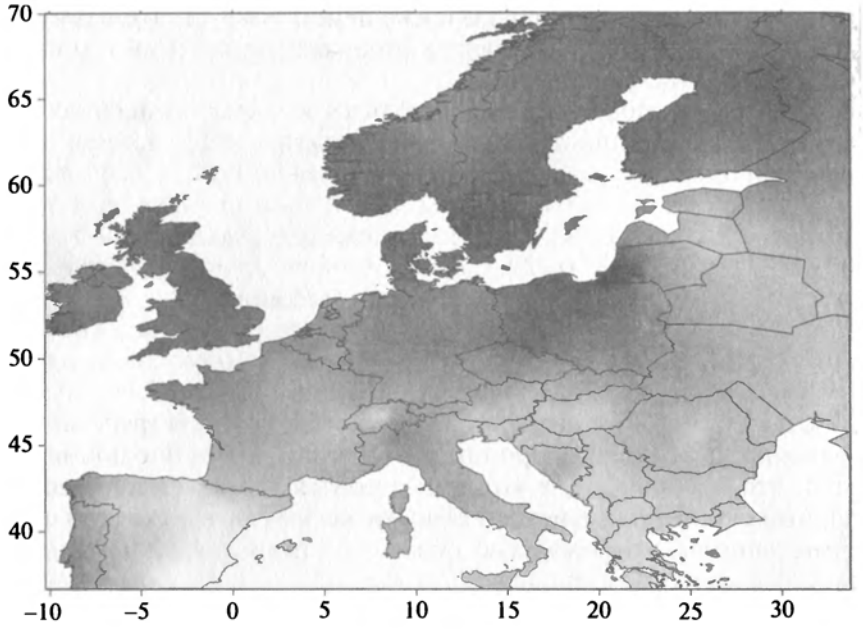


Рис. 5.10. Варьирование фактора 3, отражающего общее содержание ионов в атмосферных осадках на территории Европы (светлые тона — высокое значение фактора)

гичные измерения на территории России, то формулировка гипотезы была бы более определенна.

Рассмотренная логика разбора результатов факторного анализа практически тождественна при исследовании любых объектов. Сами по себе виртуальные факторы выделяют независимые составляющие пространственно-временно́го поведения системы. Для их объяснения необходимо привлекать независимо измеренные переменные или условия наблюдения и общие априорные представления, не включенные в исходную систему.

Однако в данном методе возможно и иное отображение переменных в пространстве факторов, иногда дающее важную дополнительную информацию. Прямое вычисление факторных нагрузок по корреляционной или ковариационной матрице отображает систему в ортогональных координатах-факторах таким образом, что на фактор 1 приходится наибольшая нагрузка и он содержит как бы наиболее интегральную информацию о поведении системы, на фактор 2 нагрузка несколько меньше и т.д. Величины нагрузок определяются парными значениями корреляций в матрице.

Пространство можно преобразовать таким образом, чтобы каждой переменной, насколько это возможно, соответствовал бы свой собственный фактор. Такое преобразование часто существенно об-

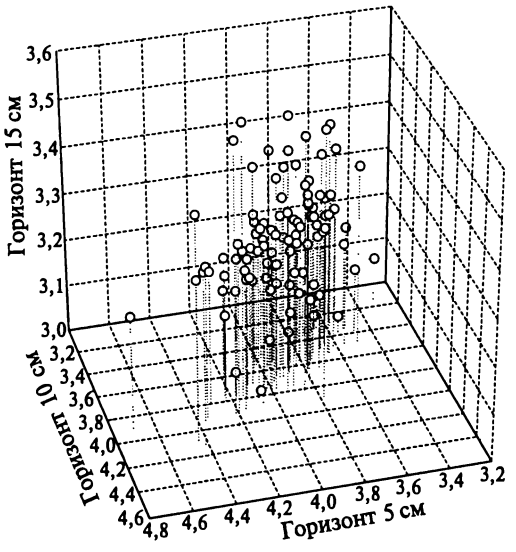
легчает трактовку факторов, а также демонстрирует возможности построения различных вариантов отображений явлений в многомерном пространстве.

Общая идея преобразования сводится к такому вращению ортогональных координат вокруг центра тяжести всей системы, при котором проекции каждой переменной максимальны для какой-либо одной координаты. Как для любой многомерной геометрической фигуры, так и для всего множества наблюдений всегда можно найти центр тяжести, координаты которого будут соответствовать значениям их математических ожиданий.

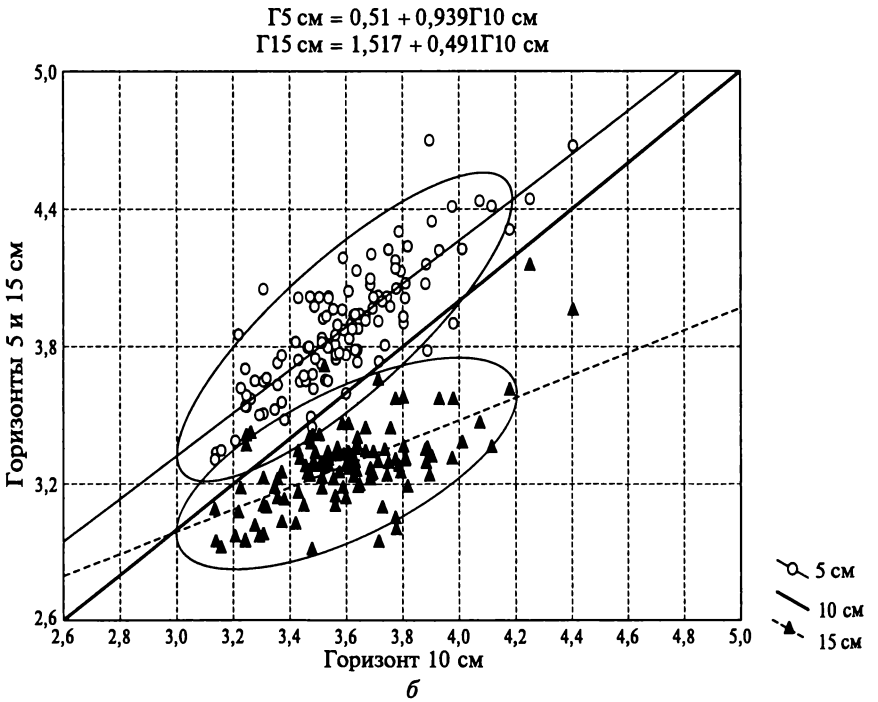
Приведем геометрическое представление последовательности преобразований, проведя анализ изменения влажности почв в трех горизонтах из рассмотренного выше примера. На рис. 5.11, *а* показано положение точек наблюдений в трехмерном пространстве измеренных переменных. При таком изображении условно принимается, что координаты, в которых строится график, перпендикулярны друг другу. В действительности же элементы исходной системы занимают очень малую область от гипотетически возможной, что собственно и свидетельствует об их взаимозависимости. На рис. 5.11, *б* те же данные представлены в двухкоординатной системе пар переменных. По оси абсцисс отложены значения влажности в горизонте 10 см, а по оси ординат — влажности двух других горизонтов. Уравнение регрессии показывает, что линия регрессии — влажность между горизонтами 5 и 10 см, почти точно параллельна линии регрессии горизонта 10 см с самим собой (жирная линия), так что горизонты принадлежат почти параллельным плоскостям в многомерном векторном пространстве. Линия регрессии 15—10 см наклонена к первым двум под углом примерно 30°.

Очевидно, можно найти некоторую средневзвешенную координату, с которой все три переменные будут образовывать примерно одинаковый угол и которая будет описывать бóльшую часть их варьирования. Далее можно найти ортогональную вторую координату, которая будет наилучшим образом описывать оставшееся варьирование какой-то пары переменных, и третью, которая будет описывать варьирования остатков от всех трех переменных.

Именно эту операцию на основе чисто алгебраических преобразований и выполняет метод главных компонент. В табл. 5.8 приведены результаты отображения переменных в векторном пространстве методом главных компонент. Из рис. 5.12, *а* видно, что фактор 1 в среднем одновременно хорошо описывает все три переменные. Пространства наблюдений 5 и 10 см почти параллельны. Горизонт 15 см имеет другой угол наклона, но также достаточно хорошо описывается первым фактором. Как следует из табл. 5.8, фактор 2 в наибольшей степени описывает третью и первую переменные и в минимальной степени — вторую. Графики, показанные на рис. 5.12, *б*,



а



б

Рис. 5.11. Взаиморасположение горизонтов 5 и 15 см по влажности относительно горизонта 10 см

выходят примерно из одной точки координатного пространства и расходятся лучами, причем вторая переменная имеет минимальный угол наклона. Наконец, фактор 3 в основном описывает первые две переменные (рис. 5.12, *в*) и практически не содержит информации о третьей, так как она почти полностью определена двумя первыми факторами. Как и по отношению к первым двум факторам, линии регрессии стремятся выходить из одной точки.

Проанализируем результаты различных вращений (рис. 5.13). Будем вращать трехмерное пространство, образованное координатами двух первых главных компонент факторов и значением влажности почв на глубине 10 см. Наш взгляд на изображение будем рассматривать как двухмерную плоскость, на которую проецируются точки вращающегося пространства. Рамка, в которую погружен рисунок, рассматривается как двухкоординатное пространство, на которое проецируется при различных вращениях трехмерное пространство. Стрелки показывают масштабы варьирования пространства наблюдений по двум координатам плоскости. Очевидно, что разные повороты позволяют увидеть различные масштабы варьирования измеренных значений в двухмерном пространстве главных компонент.

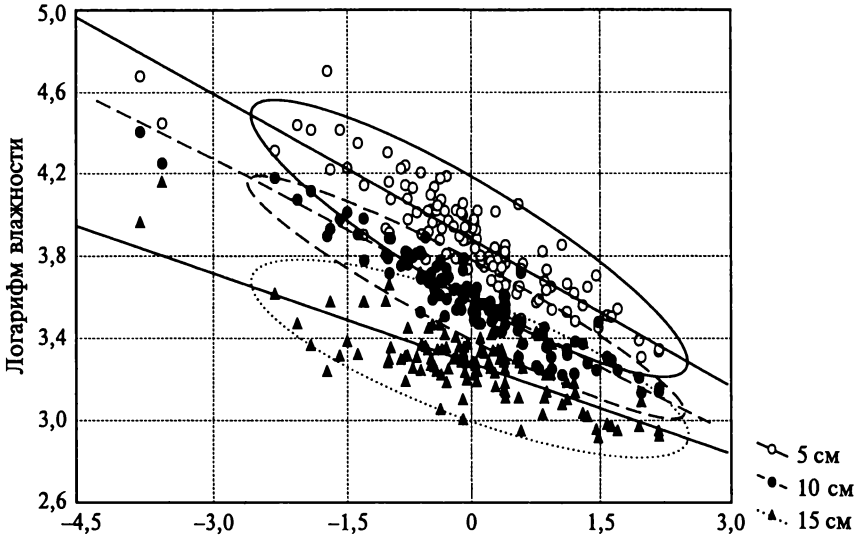
В первом случае (рис. 5.13, *а*) это почти компактная группа точек, с примерно равными проекциями на две координаты плоскости; во втором, при повороте на 90° по часовой стрелке (рис. 5.13, *б*) — почти нормально распределенная совокупность точек, в которой нельзя усмотреть никакой связи с двумя координатами пространства главных компонент; в третьем, при повороте на 90° против часовой стрелки (рис. 5.13, *в*) — вновь компактное множество. Наконец, при повороте с изменением наклона (рис. 5.13, *г*) проекция представляет собой компактное множество с максимальной нагрузкой на одну ось и с минимальной — на другую.

Примерно по этой схеме можно организовать многомерное вращение с поиском интересующей нас проекции. Одним из критериев вращения может быть максимизация дисперсии на каждый фактор без искажения взаимоположения точек (элементов системы) в пространстве. Существует несколько критериев: одни из них максимизируют дисперсию и ее аналоги, другие — скалярные произведения векторов. Чаще всего они дают сходные результаты.

Из табл. 5.9 следует, что после вращения нагрузки на три фактора стали весьма близки и каждому фактору соответствует своя основная переменная, в то время как остальные содержат лишь относительно небольшую дополнительную информацию.

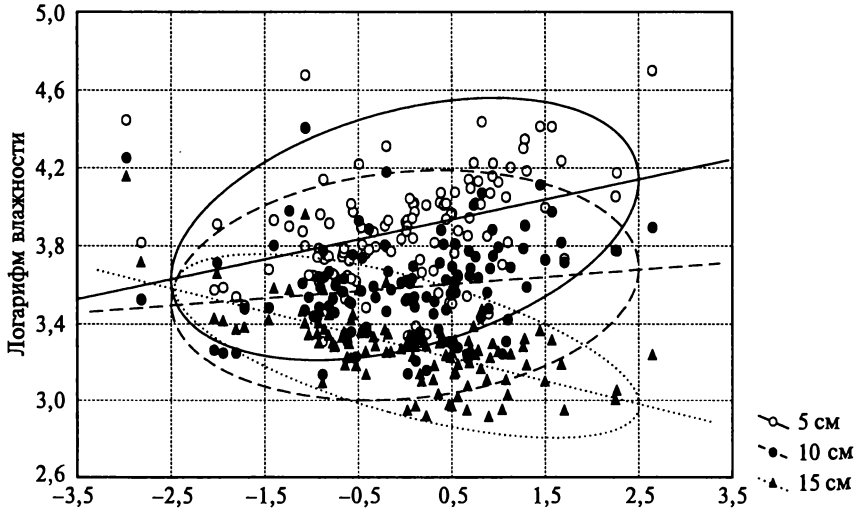
Взаимоположение элементов нашей системы в результате вращений не изменилось. Вращение в полном смысле слова позволило увидеть ту же систему, но под другим углом зрения, что показано на рис. 5.14. Здесь каждому фактору соответствует своя переменная, маркируемая узким эллипсом рассеивания с доверительным интервалом 95 %.

Г5 см = 3,885 – 0,24 Фактор 1
Г10 см = 3,594 – 0,224 Фактор 1
Г15 см = 3,28 – 0,148 Фактор 1



а

Г5 см = 3,885 + 0,103 Фактор 2
Г10 см = 3,594 + 0,037 Фактор 2
Г15 см = 3,28 – 0,119 Фактор 2



б

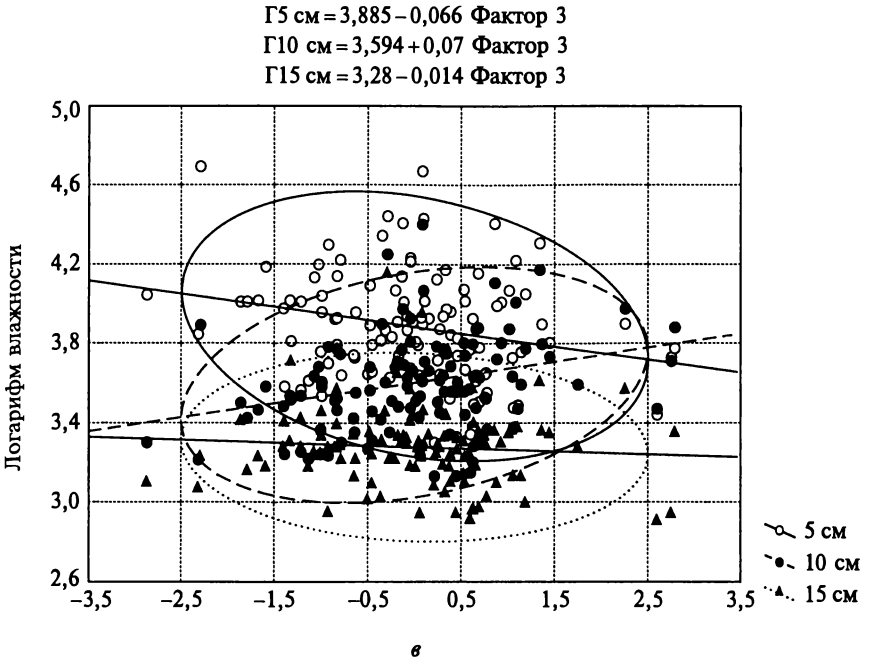


Рис. 5.12. Отображение переменных факторами, полученными методом главных компонент без вращения:

а — фактор 1; *б* — фактор 2; *в* — фактор 3

Таблица 5.8

Факторные нагрузки, полученные методом главных компонент

Переменная	Номер фактора		
Горизонт (глубина)	1	2	3
1 (5 см)	**–0,891594**	**0,380705**	**–0,245201**
2 (10 см)	**–0,942969**	–0,155302	**0,294432**
3 (15 см)	**–0,775992**	**–0,626141**	–0,076058
Дисперсия	2,286294	0,561107	0,152598
Доля дисперсии	0,762098	0,187036	0,050866

Примечание. Полужирным шрифтом выделены ведущие компоненты.

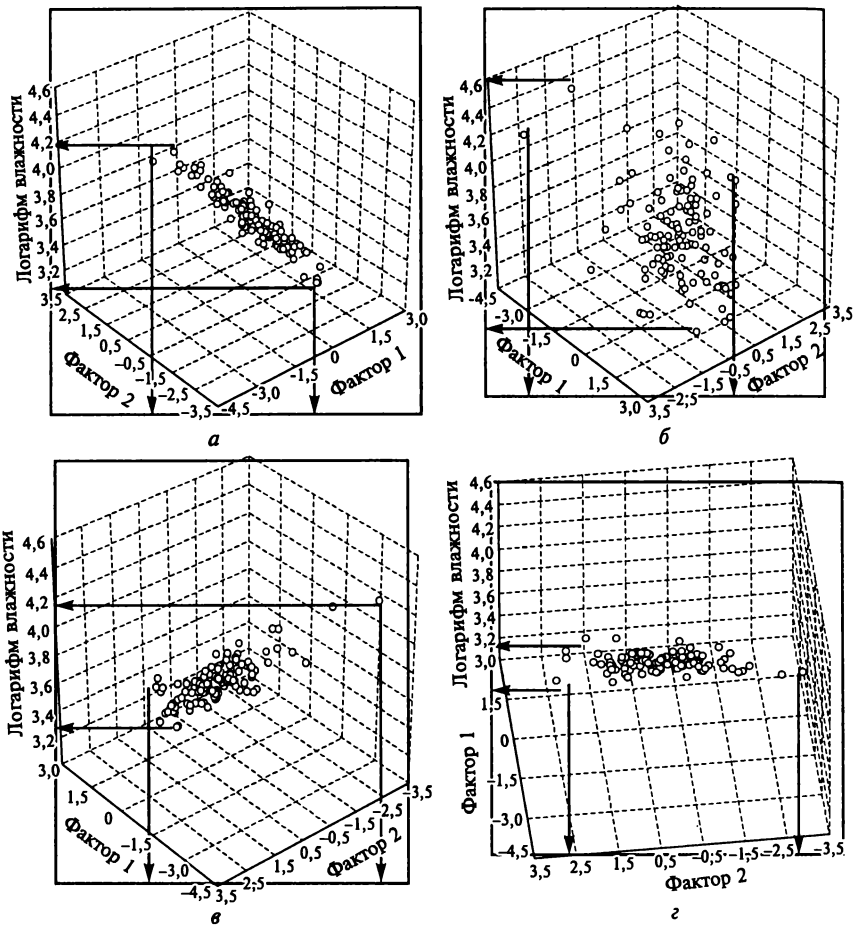


Рис. 5.13. Визуализация процедуры вращения:

а — первое отображение; *б* — поворот на 90° по часовой стрелке; *в* — поворот на 90° против часовой стрелки; *г* — поворот с изменением наклона

Теперь рассмотрим под иным углом зрения отображения отношений между факторами и концентрациями ионов в атмосферных осадках в Европе, получаемые после операции вращения факторного пространства (табл. 5.10).

Очевидно, что отображение существенно изменилось. Фактор 1 после вращения стал определять ионы натрия, магния и хлора (в исходном варианте они были выражены нечетко). Фактор 2 определяет практически с тем же весом, что и первый, ионы SO_4, NO_3, NH_4 и в меньшей степени — электрическую проводимость. Фактор 3, так же как и в первом случае, определяет pH и содержание катиона кальция, но с существенно большей однозначнос-

Таблица 5.9

Факторные нагрузки после вращения методом веримакс (Varimax normalized)

Переменная	Номер фактора		
Горизонт (глубина):	1	2	3
1 (5 см)	**0,906136**	0,214333	0,364663
2 (10 см)	0,516825	0,339654	**0,785829**
3 (15 см)	0,201371	**0,950774**	0,235539
Дисперсия	1,128741	1,065275	0,805984
Доля дисперсии	0,376247	0,355092	0,268661

Примечание. Полужирным шрифтом выделены ведущие компоненты.

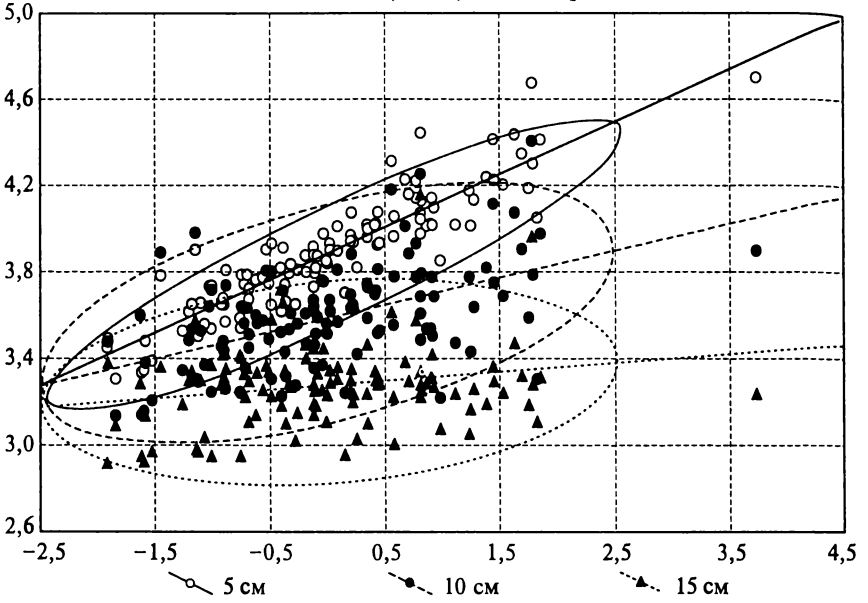
Таблица 5.10

Факторные нагрузки для системы «Концентрация ионов в атмосферных осадках Европы» после вращения

Переменная	Номер фактора			
	1	2	3	4
pH	0,066359	–0,200809	**0,931999**	0,031662
SO_4	0,269699	**0,803771**	–0,010132	0,357498
NO_3	–0,000727	**0,901018**	–0,148909	0,070919
NH_4	–0,037808	**0,901900**	0,062286	0,003761
Cl	**0,959187**	0,035642	0,010738	0,131311
COND	0,598164	**0,629116**	–0,193264	0,275979
Ca	0,232012	0,524443	**0,589483**	0,399829
K	0,332098	0,195305	0,120923	**0,886995**
Mg	**0,844376**	0,191533	0,292020	0,208166
Na	**0,945690**	–0,014449	0,044917	0,160067
Дисперсия	3,127829	3,058763	1,381651	1,242836
Доля дисперсии	0,312783	0,305876	0,138165	0,124284

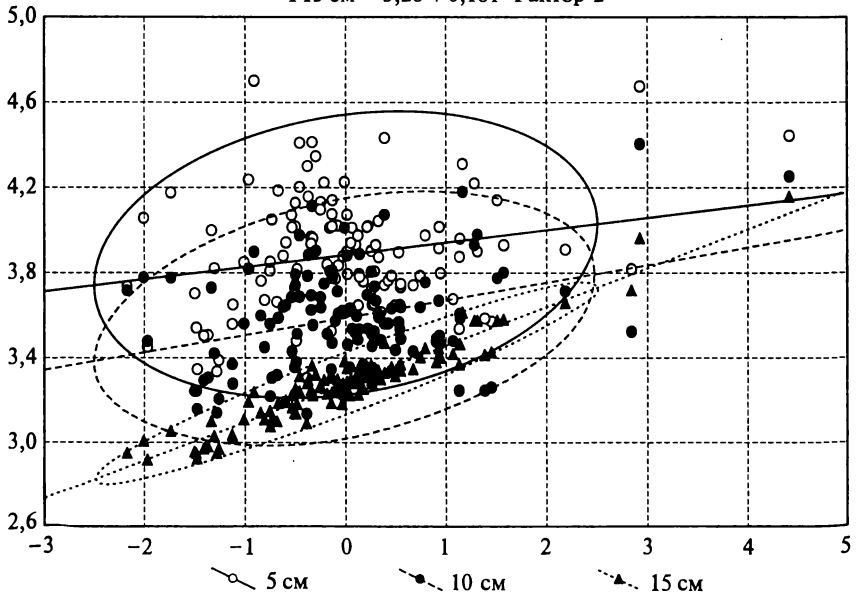
Примечание. Полужирным шрифтом выделены ведущие компоненты.

Г5 см = 3,885 + 0,244 Фактор 1
Г10 см = 3,594 + 0,123 Фактор 1
Г15 см = 3,28 + 0,038 Фактор 1



а

Г5 см = 3,885 + 0,58 Фактор 2
Г10 см = 3,594 + 0,81 Фактор 2
Г15 см = 3,28 + 0,181 Фактор 2



б

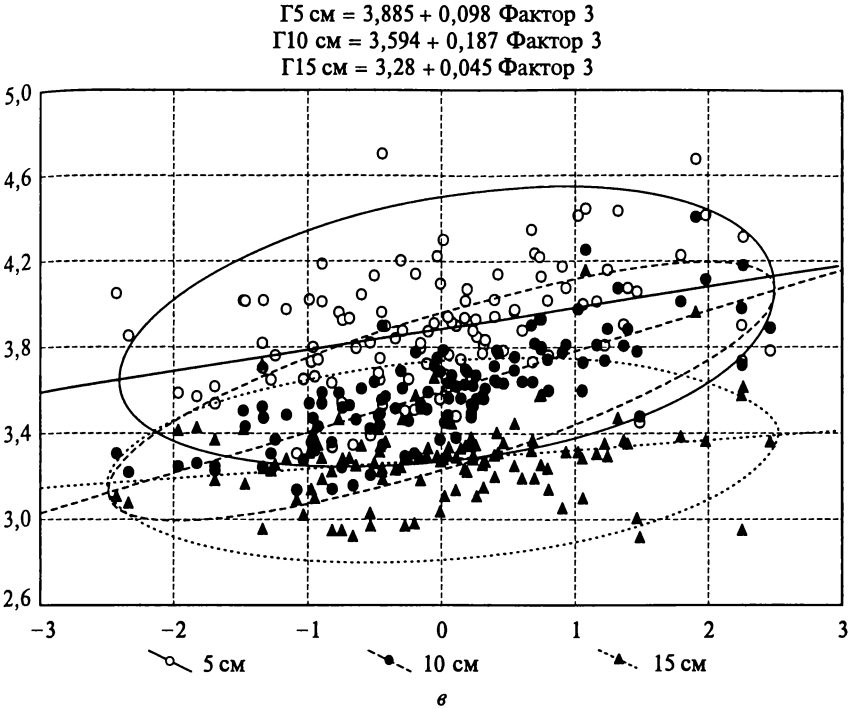


Рис. 5.14. Отображение переменных факторами, полученными методом главных компонент после вращения:

а — фактор 1; *б* — фактор 2; *в* — фактор 3

тью, а фактор 4, так же как и без вращения, но более однозначно, определяет калий.

В конечном итоге каждая переменная для своего полного описания требует рассмотрения всех факторов, но все-таки в этом отображении влияние одного из них (специфического) максимально. Итак, содержание двух первых факторов принципиально изменилось, а у двух остальных (относительно слабых по влиянию) — осталось практически неизменным.

Повторим процедуру дисперсионного анализа факторов в отношении независимых переменных (табл. 5.11).

Вполне естественно, что связь переменных с факторами в обеих системах отображения тождественна, но сами факторы связаны с ними иначе. В факторном пространстве, полученном на основе вращения с годом наблюдения, существенно связан фактор 1, отображающий содержание натрия, магния и хлора. Сезонный ход наиболее четко выражен у факторов 1 и 2, связанных с анионами серы и азота. С высотой особо тесно связан фактор 1 и, как в первой системе, фактор 3 (pH); с суммой осадков — факторы 1 и 2.

Таблица 5.11

Одновариантный дисперсионный анализ отношения факторов после вращения к году и месяцу наблюдений, абсолютной высоте над уровнем моря и суммам осадков

Внешняя переменная и Wilks' Lambda	Номер фактора	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Год	1	107,6911	16	6,73069	8492,309	8584	0,989318	6,80336	0,000000
	2	186,8456	16	11,67785	8413,154	8584	0,980097	11,91499	0,000000
	3	120,5708	16	7,53567	8479,429	8584	0,987818	7,62861	0,000000
0,911176	4	364,5069	16	22,78168	8235,493	8584	0,959400	**23,74575**	0,000000
Месяц	1	812,6447	11	73,87679	7787,355	8589	0,906666	**81,48181**	0,000000
	2	570,6652	11	51,87866	8029,335	8589	0,934839	**55,49473**	0,000000
	3	83,6771	11	7,60701	8516,323	8589	0,991538	7,67192	0,000000
0,828862	4	64,3904	11	5,85367	8535,610	8589	0,993784	5,89029	0,000000
$H_{\text{н.у.м}}$	1	1727,039	7	246,7198	6797,795	8537	0,796274	**309,8427**	0,00
	2	386,022	7	55,1460	8172,573	8537	0,957312	57,6050	0,00
	3	1103,619	7	157,6599	7437,524	8537	0,871210	**180,9665**	0,00
0,613236	4	393,835	7	56,2621	8180,627	8537	0,958255	58,7130	0,00
Месячная сумма осадков	1	104,9818	5	20,9964	8495,018	8595	0,988367	21,2435	0,000000
	2	749,7097	5	149,9419	7850,290	8595	0,913355	**164,1660**	0,000000
	3	38,9741	5	7,7948	8561,026	8595	0,996047	7,8258	0,000000
0,881638	4	140,2636	5	28,0527	8459,736	8595	0,984263	28,5013	0,000000

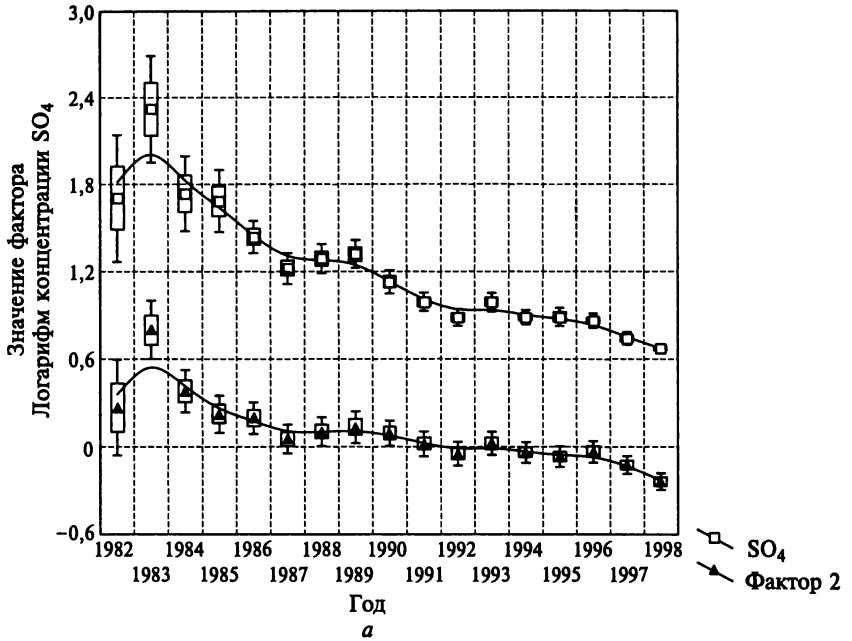
Сохраняя последовательность изложения принятого при рассмотрении результатов анализа данных методом главных компонент, проанализируем сначала характер связи с независимыми переменными фактора 2 после вращения, индуцирующего концентрации анионов серы и азота и электрическую проводимость. Как следует из рис. 5.15, *а—г* и карты (рис. 5.16), результаты здесь полностью тождественны отображению в прямом методе главных компонент с той лишь разницей, что анионы азота ведут себя так же, как и анионы серы, и, по-видимому, целиком связываются с антропогенным загрязнением среды. Небольшие различия в двух отображениях не могут существенно повлиять на выводы и вытекающие из них гипотезы.

Фактор 1, полученный в результате вращения, строго выделяет концентрации катионов натрия, магния и хлора. Многолетняя динамика (рис. 5.17, *а—г*) выражена очень слабо, но сезонная, напротив, весьма строго. Максимум концентрации рассматриваемых соединений наблюдается зимой, минимум — летом. Существует некоторая предпочтительная высота над уровнем моря (около 100—200 м), на которой осадки содержат максимальные концентрации этих анионов и соответственно имеют максимальное значение фактора.

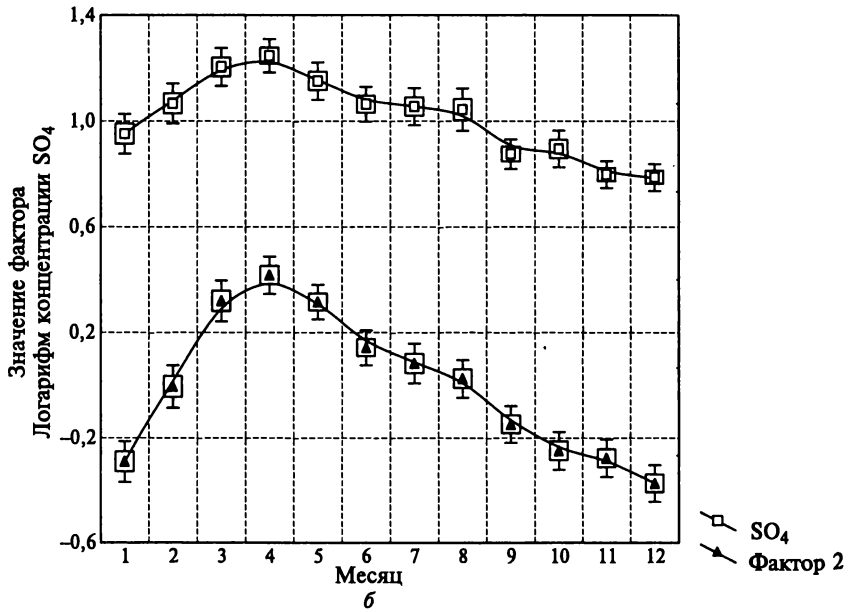
На больших высотах их концентрации уменьшаются. Наконец, достоверна, но не очень понятна обратно параболическая связь этого фактора с суммой осадков. Максимум концентрации при большой дисперсии наблюдается как при наименьших, так и при наибольших месячных суммах осадков. В целом же можно почти наверное утверждать, что этот фактор отражает перенос на континент морских воздушных масс. Все территории, открытые для западного переноса и расположенные на небольшом расстоянии от акваторий морей, имеют повышенное значение этого фактора (рис. 5.18).

Фактор 3 в обеих моделях описывает pH и концентрацию кальция. Многолетний ход изменения этого фактора в двух отображениях практических тождественен, но сезонный ход различается весьма существенно. Если в первой модели сезонный максимум этого фактора отмечается летом, то во второй — в марте и мае. В то же время отношения с высотой над уровнем моря и суммой осадков практически тождественны в обеих моделях. Различия между изображениями на картах невелики.

Итак, несмотря на то что оба способа отображений во многом подобны, каждый выявляет некоторые особенности: первая модель — существование летнего максимума концентрации некоторых соединений и pH, вторая модель — возможную связь концентраций натрия, магния и хлора с морскими воздушными массами. Существуют также и некоторые различия в описании динамики pH.

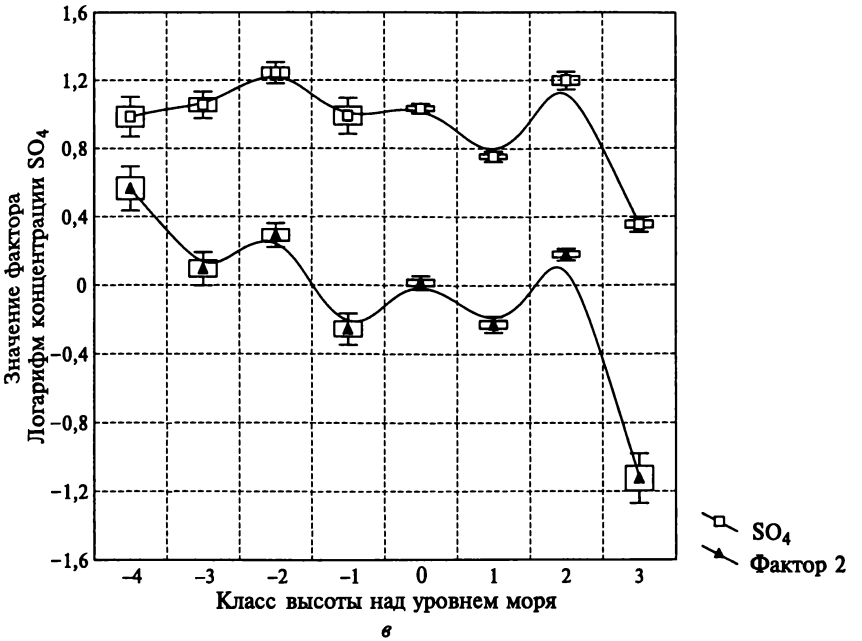


а

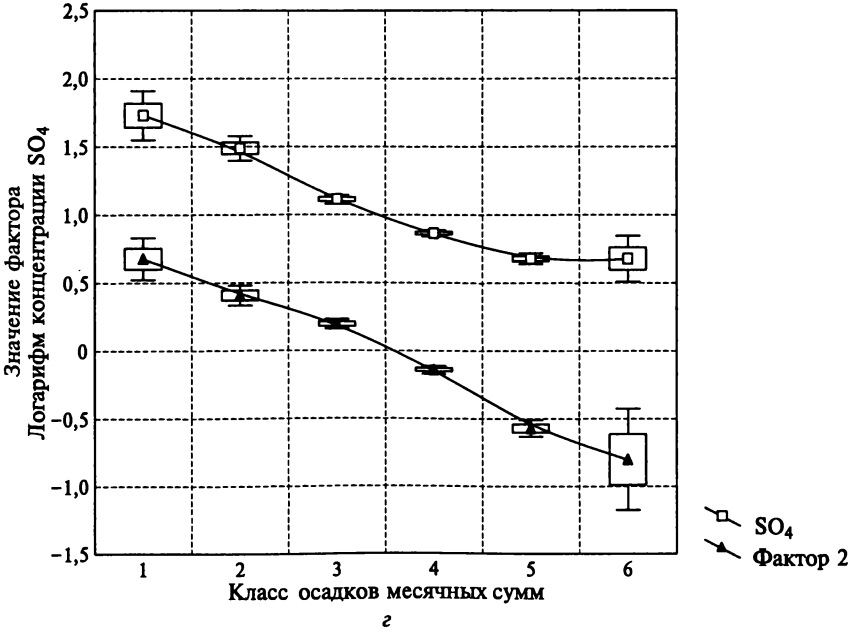


б

Рис. 5.15. Связь фактора 2, полученного на основе вращения,



в



г

с независимыми переменными (*а—г*)

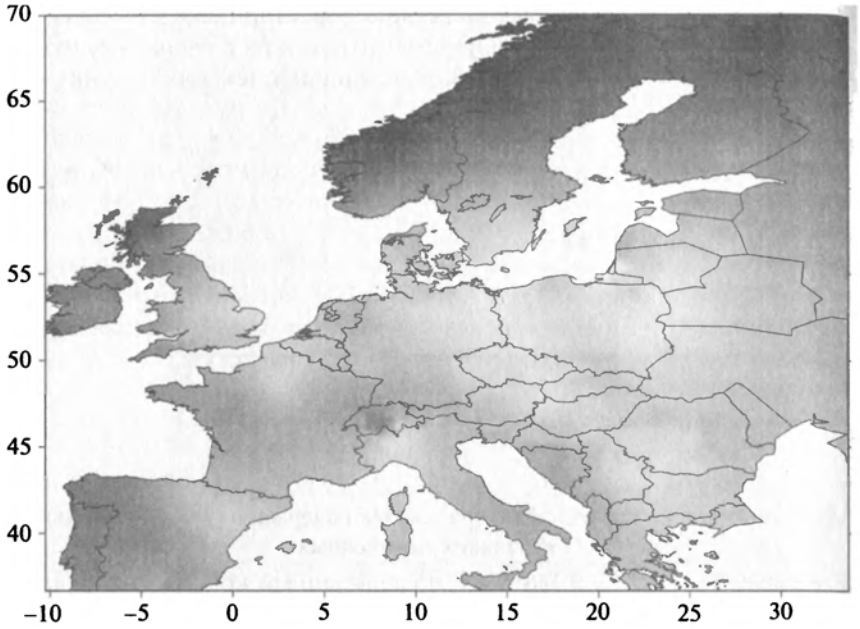


Рис. 5.16. Варьирование фактора 2 с поворотом в системе координат, отражающего содержание ионов SO_4, NO_3, NH_4 в атмосферных осадках и их электрическую проводимость на территории Европы. Светлые тона — высокое значение фактора

Возникает естественный вопрос: в какой мере чисто алгебраические преобразования корреляционной матрицы дают содержательную информацию об изучаемом многомерном явлении? Если в этом есть смысл, то он должен определяться какими-то фундаментальными свойствами самой природы, которые лишь моделируют линейные алгебраические преобразования. Ответ на этот вопрос определяет наше отношение к целесообразности применения метода главных компонент.

При этом необходимо иметь в виду следующее:

- метод главных компонент как любой статистический метод отображает равновесные, стационарные отношения между переменными;
- метод отображает только линейные отношения нормально распределенных переменных.

Допустим, что реальность отвечает этим условиям. Тогда можно полагать, что в системе существует некоторый порядок в отношениях между ее свойствами и связанными с ней частями системы. Любой порядок есть ограничение потенциального разнообразия, т.е. ограничение независимости между свойствами. Эти ограничения суть функциональные связи между переменными. Функцио-

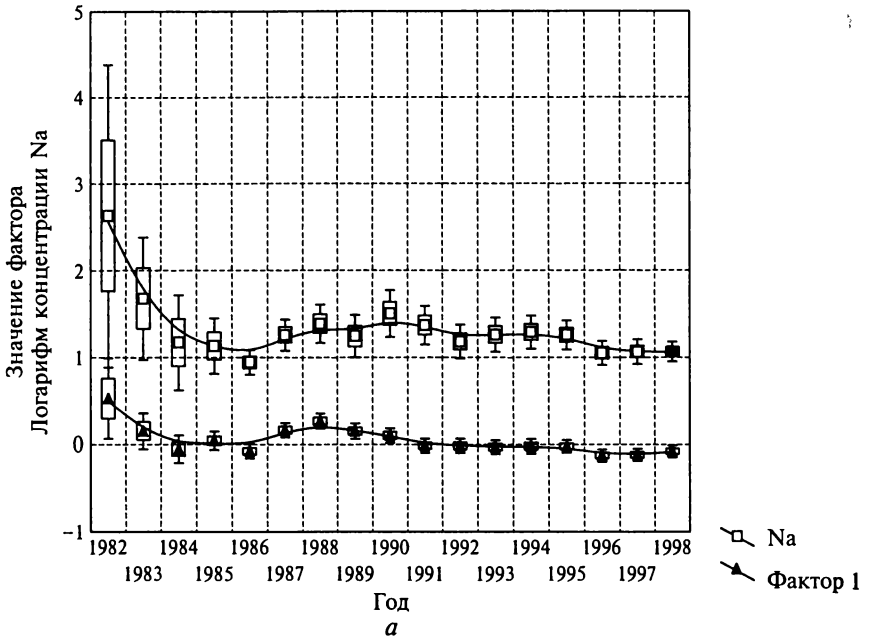
нальные связи по схеме «все со всеми» в равной степени приводят фактически к независимости переменных и к их случайному относительно друг другу изменению. В реальных системах в функциональных зависимостях любое свойство — функция ограниченного числа других переменных. При этом почти всегда можно выделить один-два ведущих фактора на фоне ограниченного числа подчиненных. Свойства-факторы могут как принадлежать самой системе, так и находиться вне ее. Если свойства сопряжены друг с другом, но прямо функционально не связаны, то логично полагать, что они управляются общими свойствами-факторами. Таким образом, любые ограничения независимости можно трактовать как действие некоторого ограниченного числа факторов.

Таблица 5.12

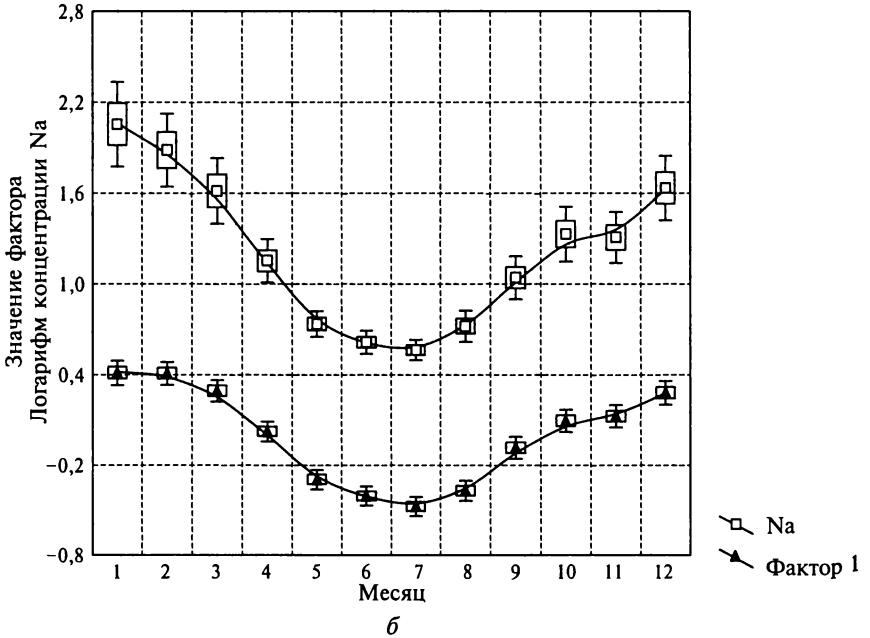
Модель регрессии первого фактора (общее содержание ионов в осадках) от внешних переменных

r = 0,39516243; R^2 = 0,15615334; подправленный коэффициент детерминации R^2 = 0,15536249; F(8,8536) = 197,45; p < 0,0000; Std.Error of estimate: 0,91872

Переменная	BETA	Std.Err. of BETA	*b*	Std.Err. of *b*	t(8536)	p-level
Константа			19,98959	5,137181	3,8912	0,000101
Высота над уровнем моря	–0,20751	0,010270	–0,15450	0,007646	–20,2066	0,000000
Квадрат высоты над уровнем моря	0,03895	0,010045	0,01532	0,003951	3,8772	0,000106
Год	–1,40564	0,456239	–0,34570	0,112206	–3,0809	0,002070
Квадрат года	1,17997	0,456200	0,00158	0,000612	2,5865	0,009711
Месяц (порядковый номер)	–0,24947	0,043970	–0,07253	0,012784	–5,6737	0,000000
Квадрат порядкового номера месяца	0,13212	0,043875	0,00287	0,000952	3,0114	0,002608
Осадки	–0,50371	0,052978	–0,54897	0,057739	–9,5079	0,000000
Квадрат осадков	0,32559	0,052996	0,05241	0,008531	6,1437	0,000000

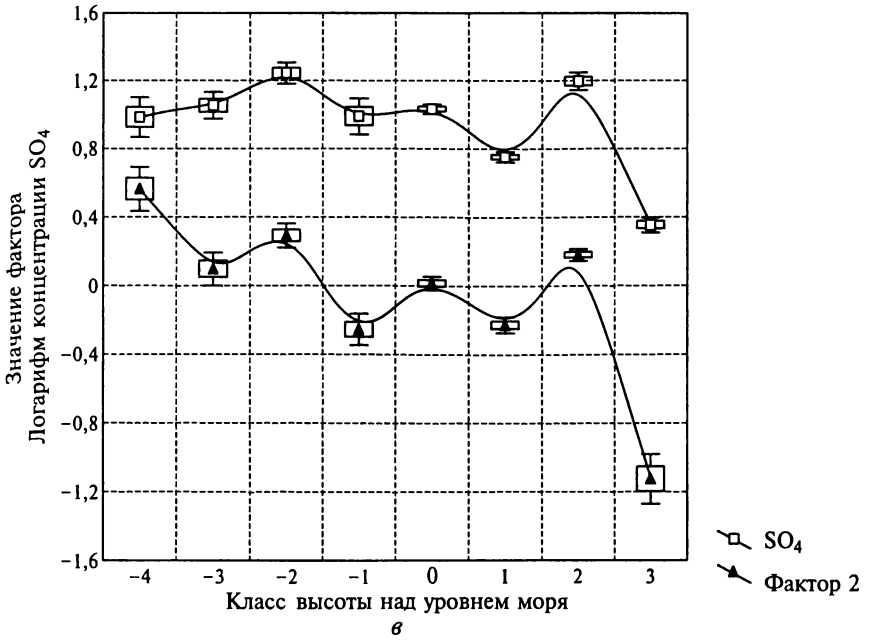


а

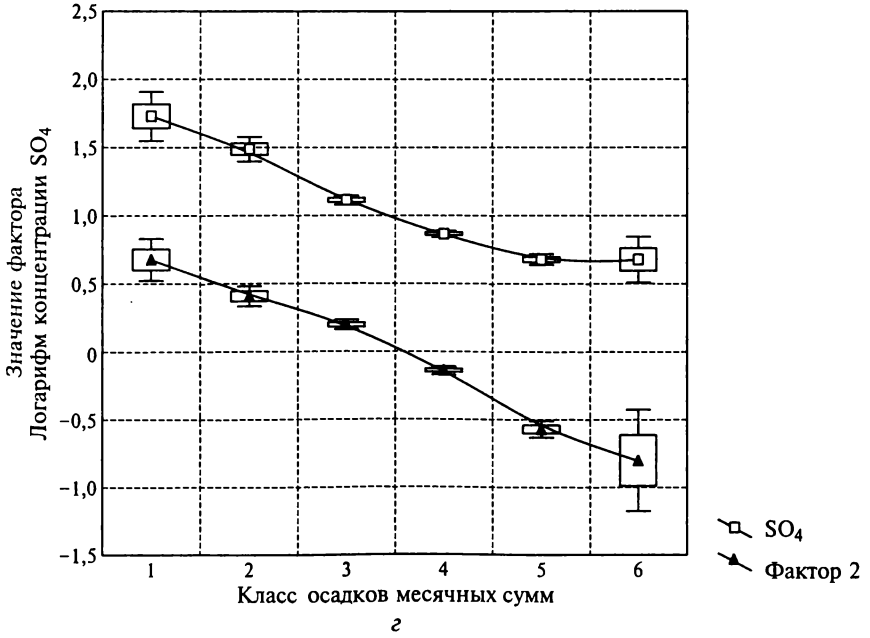


б

Рис. 5.17. Связь фактора 1, полученного на основе



в



г

вращения, с независимыми переменными (*а*—*г*)

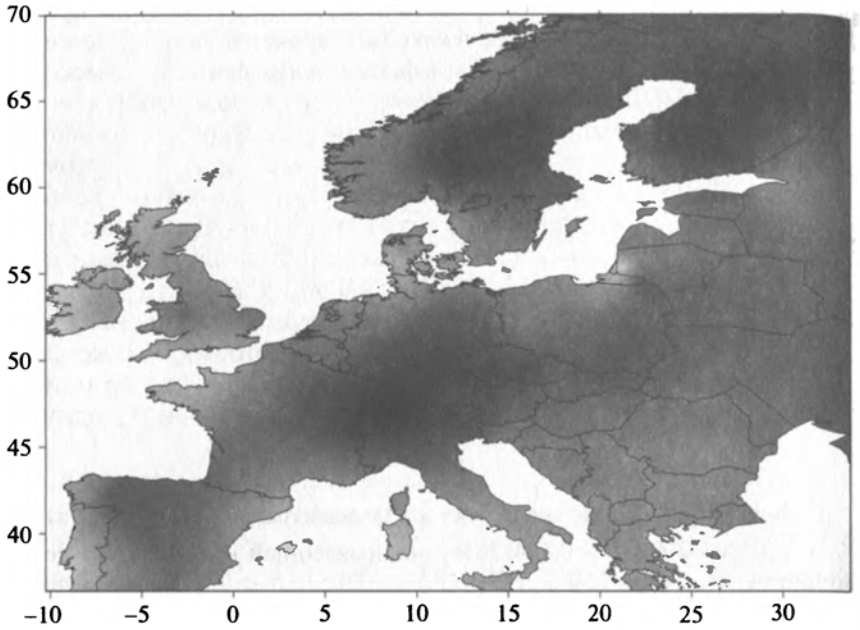


Рис. 5.18. Варьирование фактора 1 с поворотом в системе координат, отражающего содержание ионов Na, Mg, Cl в атмосферных осадках на территории Европы (светлые тона — высокое значение фактора)

Метод главных компонент при указанных выше условиях позволяет выявить эти латентные факторы, ограничивающие разнообразие отношений между свойствами системы. Если реальность соответствует модели, то выявленные латентные факторы также принадлежат реальности. Вращение системы относительно центра тяжести позволяет максимизировать наблюдаемость наиболее характерных отношений, скрываемых более общими, получаемыми при прямом преобразовании.

Однако определение физического смысла этих факторов в общем случае непростая задача. Фактор может не соответствовать какому-либо «простому», далее не расчленяемому свойству. Он сам по себе может быть продуктом сложных системных отношений. Приведем простой пример. Человек воспринимает «тепло» как соотношение собственно температуры, влажности воздуха и интенсивности турбулентного теплообмена. Если мысленно представить различные свойства, определяющие поведение человека или его организма, то, скорее всего, выявится некоторый обобщенный фактор, который будет интегрировать в себе, по крайней мере, три физические переменные. Если попытаться связать этот виртуальный фактор только с температурой среды, то последняя будет, очевидно, описывать только часть его варьирования.

Сходная ситуация имеет место и в рассмотренном выше примере. Один из факторов, скорее всего, определяет поступление ионов с продуктами техногенеза в атмосферу и вымывание их атмосферными осадками. Очевидно, что он отражает для каждого измерения соотношение между поступлением ионов в атмосферу в данной точке и вероятностью выноса их с осадками, т. е. некоторый баланс между «приходом и расходом». Сезонный характер изменения значения фактора есть результатирующая сезонной изменчивости поступления некоторых ионов в атмосферу. В конечном итоге значения каждого фактора — функция нескольких внешних переменных.

В табл. 5.12—5.14 приведены модели множественной регрессии трех факторов от внешних переменных, из которых, так же как и из дисперсионного анализа, следует, что каждый фактор описывается в некоторой степени и статистически значимо от одних и

Таблица 5.13

Модель регрессии второго фактора (концентрация анионов азота)

r = 0,38044209; R^2 = 0,14473619; подправленный коэффициент детерминации R^2 = 0,14393463; F(8,8536) = 180,57; p < 0,0000 Std. Error of estimate: 0,92550

Перемення	BETA	Std. Err. of BETA	*b*	Std. Err. of *b*	t(8536)	p–level
Константа			11,42418	5,175088	2,2075	0,027303
Высота над уровнем моря	0,12668	0,010339	0,09438	0,007702	12,2532	0,000000
Квадрат высоты над уровнем моря	0,08100	0,010113	0,03188	0,003980	8,0093	0,000000
Год	–1,03378	0,459315	–0,25441	0,113034	–2,2507	0,024430
Квадрат года	0,98553	0,459276	0,00132	0,000617	2,1458	0,031915
Месяц (порядковый номер)	1,07576	0,044267	0,31297	0,012879	24,3017	0,000000
Квадрат порядкового номера месяца	–1,13445	0,044171	–0,02464	0,000959	–25,6833	0,000000
Осадки	0,28071	0,053335	0,30613	0,058165	5,2631	0,000000
Квадрат осадков	–0,52764	0,053353	–0,08499	0,008594	–9,8894	0,000000

Таблица 5.14

Регрессионная модель третьего фактора (кислотность осадков)

r = 0,53813424; R^2 = 0,28958846; подправленный коэффициент детерминации R^2 = 0,28900595; F(7,8537) = 497,14; p < 0,0000; Std.Error of estimate: 0,84389

Переменная	BETA	Std. Err. of BETA	*b*	Std. Err. of *b*	t(8537)	p-level
Константа			–30,5947	4,718360	–6,4842	0,000000
Высота над уровнем моря	0,46496	0,009400	0,3466	0,007006	49,4667	0,000000
Квадрат высоты над уровнем моря	0,23161	0,009216	0,0912	0,003629	25,1311	0,000000
Год	2,61073	0,418589	0,6428	0,103066	6,2370	0,000000
Квадрат года	–2,53710	0,418553	–0,0034	0,000562	–6,0616	0,000000
Месяц (порядковый номер)	0,94881	0,040182	0,2762	0,011696	23,6127	0,000000
Квадрат порядкового номера месяца	–0,93630	0,040100	–0,0203	0,000871	–23,3490	0,000000
Осадки	–0,18107	0,009410	–0,1976	0,010267	–19,2427	0,000000

тех же внешних переменных, но различным образом. Предлагаем читателю самостоятельно сравнить отображение отношения факторов в моделях регрессии и дисперсионном анализе. Переведите «язык» моделей регрессии в формат обычного текста, описывающего влияние и форму зависимости.

Очевидно, что описание варьирования факторов через переменные (год наблюдений, месяц наблюдений, высота над уровнем моря и сумма осадков) не может быть полным, так как существенная часть варьирования определяется географическим положением станции наблюдения относительно переноса. Но в то же время каждый из факторов интерпретируется одними и теми же внешними переменными, что доказывает не более чем их объективность и географическую обусловленность. Совершенно очевидно, что эти внешние переменные лишь отчасти позволяют понять физическую природу каждого фактора, но не дают его физической интерпретации. Физический смысл виртуальных факторов может быть определен как гипотеза о природе выявленных отношений.

Для доказательства этих гипотез необходимо, как было показано выше, ставить специальную систему наблюдений и строить динамические модели «эмиссия — перенос — вынос с осадками». Некоторые «подсказки» к конструкции такой модели и ограничения на ее параметры можно извлечь из результатов многомерного анализа.

Таким образом, виртуальные факторы есть отображение интегрального действия нескольких физических, генетически разных или тождественных механизмов, различающихся по их проявлению в отношении различных свойств системы.

Вместе с тем статистическая модель, получаемая методом главных компонент, позволяет с достаточной надежностью определить эффективность усилий, затраченных на снижение загрязнения атмосферы, выделить свойства, не зависящие от хозяйственной деятельности, найти высотные уровни с наименьшим загрязнением атмосферы.

Обычно такие общие оценочные результаты можно получить при корректном применении метода главных компонент к широкому классу явлений, отражающих отношения между физико-химическими переменными, которые относительно легко линеанизируются и нормализуются.

5.2. Многомерный факторный анализ

Формально целью метода главных компонент является отображение исходной системы в ортогональных координатах, которые при определенных условиях могут рассматриваться как виртуальные факторы. Прямой целью факторного анализа является отображение исходной системы с размерностью (числом переменных) m в пространстве ограниченного числа k ортогональных факторов, так что $k < m$. Иными словами, целью факторного анализа является отображение варьирования переменных в пространстве с размерностью, существенно меньшей, чем число переменных, т.е. уменьшение размерности.

Основные идеи метода опираются на описанные выше алгебраические операции с корреляционными и ковариационными матрицами переменных исходной системы.

В отличие от метода главных компонент модель факторного анализа предполагает, что каждая переменная y_i может быть представлена следующим образом:

$$y_i = a_i + b_{ij} f_j + \varepsilon_i,$$

где a_i — константа; b_{ij} — константы при f_j факторе ($j = 1, 2, ..., k$); ε_i — собственная изменчивость переменной i и/или нормально распределенные ошибки, определяемые шумом.

Приведем одну из наиболее распространенных схем факторного анализа:

1) определяем полный ортогональный базис и рассчитываем значения факторов (по аналогии с методом главных компонент);

2) строим регрессионные модели каждой переменной от первого фактора с наибольшей описываемой дисперсией;

3) вычитаем из реальных значений измеренных переменных рассчитанные значения по уравнениям регрессии и получаем «новые переменные»;

4) на основе «новых переменных» рассчитываем новую корреляционную (ковариационную) матрицу;

5) вновь применяем к ней процедуру метода главных компонент и получаем новые факторы;

6) строим уравнения регрессии для исходных значений переменных от первого фактора, полученного на первом этапе, и на основе новой корреляционной матрицы;

7) вновь вычитая из измеренных значений рассчитанные, получаем переменные, представленные через остатки;

8) далее процедуру повторяем.

Очевидно, что, проведя такие процедуры $m - 1$ раз, можно получить факторы, аналогичные определенным в методе главных компонент. Поэтому принципиальным является своевременная остановка процедуры расчета факторов. Используемый критерий окончания процедуры в основном определяет конкретный метод факторного анализа.

Общая идея остановки опирается на идею проверки нулевой гипотезы о вероятности принадлежности получаемой новой корреляционной матрицы к исходной. Очевидно, что если в остатках от регрессионных моделей остается только «случайный шум», то исходная и рассчитанная при ограниченном числе факторов матрицы корреляции будут принадлежать одной генеральной совокупности. Соответственно, если k-факторов воспроизводят исходные переменные с такой полнотой, что рассчитанная на основе предсказываемых k-факторами значениях переменных и исходная корреляционная матрицы различаются статистически незначимо. Критериями сравнения матриц может быть χ^2-критерий наибольшего правдоподобия, прямое сравнение корреляционных матриц по ошибкам выборочных корреляций и т. п. Соответственно можно рассчитывать число факторов при различных уровнях значимости или различных минимальных значениях дисперсии, описываемой последним фактором.

Необходимо отметить, что, если система линейна, то различия результатов, найденных разными методами, обычно ничтожны. В «*методе наибольшего правдоподобия*» может быть определен тест оценки статистической значимости полученного числа факторов. Часто в качестве наилучшего метода рекомендуется именно метод

наибольшего правдоподобия, так как в нем осуществляются процедуры дополнительной подгонки новой корреляционной матрицы к исходной. При этом полученная система отображения дает наиболее равномерное распределение значений дисперсий по факторам и не требует применения вращений.

Однако далеко не всегда отображение, получаемое на основе вращений, является наиболее информативным. Вместе с тем весьма полезно то, что этот метод дает дополнительный критерий для нахождения оптимальной размерности системы (число независимых факторов, определяющее варьирование всех переменных).

Несколько особняком стоит *метод центроидов*, который опирается на геометрическую интерпретацию многомерных отношений. Согласно этому методу, сначала ищется общий центр тяжести системы, через который проводится ось или координата, в наибольшей степени коррелирующая со всеми остальными. Далее определяется выходящая из центра тяжести, перпендикулярная первой вторая координата, которая в наибольшей степени коррелирует с переменными и т.д. Часто этот метод трактуется как наиболее современный и менее чувствительный к небольшим нелинейностям отношений.

Возвращаясь к примеру, рассмотренному при описании метода главных компонент, имеем, что различные методы факторного анализа дают тождественные результаты. Если распределения нормальны и отношения близки к линейным, то тождественность результатов — обычна и естественна.

Поэтому можно рекомендовать при анализе реальных данных выбирать тот метод, который при минимуме факторов дает максимально полное описание варьирования переменных. С другой стороны, если результаты применения различных методов сильно разнятся, то это является важным доводом о существенной нелинейности отношений, которую необходимо выявить и исследовать. Конечно, решение этой задачи требует хорошего уровня владения статистическими методами.

В заключение рассмотрим соотношения методов многомерного факторного анализа с общей схемой иерархических эпистомологических уровней систем.

Для того чтобы начать многомерный анализ, необходимо исследовать на основе одномерного анализа свойства переменных (в первую очередь их распределения) и привести их, насколько это возможно, к «нормальному виду». В целом эти преобразования можно соотнести с первым уровнем или системой данных.

Второй уровень описывает отношения между переменными и реализуется на основе корреляционных и ковариационных матриц и моделей регрессии.

Факторный анализ может быть соотнесен с третьим уровнем системологической схемы: с уровнем структурированных систем,

в которых определены достаточно мощные инварианты — базовые факторы, характеризующие варьирование всего множества переменных и аккумулирующие множество частных статистических отношений.

Их физическая интерпретация создает основу для построения модели метасистемы как модели процессов, справедливой для широкого класса условий, что приближает исследователя к пониманию базовых «сущностей» изучаемого им явления.

С другой стороны, обобщенное отображение отношений дает само по себе широкие возможности для решения важнейших прикладных задач:

1. Описание нормы отношений и прогноза допустимых, равновесных соотношений состояния всех переменных.

2. Выделение переменных, наиболее чувствительных к независимым факторам, которые можно рассматривать как индикаторы состояния системы (задачи мониторинга).

3. Прогноз равновесной динамики системы.

Контрольные вопросы

1. Проследите связь между векторной алгеброй и базовыми преобразованиями в многомерном параметрическом анализе.

2. Рассмотрите связь между моделями множественной регрессии и методом главных компонент.

3. Назовите геометрическую фигуру, проекции которой на координаты пространства не изменяются при вращении.

4. Постройте трехмерную фигуру, проекции которой на координаты существенно менялись бы при их вращении.

5. Почему нелинейные отношения невозможно без специальных преобразований и без искажений отобразить средствами метода главных компонент?

Глава 6
МНОГОМЕРНЫЙ НЕПАРАМЕТРИЧЕСКИЙ АНАЛИЗ

Цель многомерного непараметрического анализа (также как и методов факторного анализа и главных компонент) — найти виртуальные независимые факторы, описывающие варьирование всех переменных системы, при глубоком нашем желании того, чтобы эти факторы имели бы физический смысл.

В параметрических методах фактически используется только одна единственная метрика: коэффициент корреляции как мера подобия. Вместе с тем принятый метод измерения «сходства—различия» или в общем случае дистанции должен отражать содержательную сторону взаимодействия между различными свойствами и объектами. Если метрика не адекватна реальным отношениям, то их отображение будет неизбежно искажено. Именно поэтому методы главных компонент и факторного анализа с их корреляционной метрикой не могут адекватно отображать системы с большой ролью нелинейных отношений.

6.1. Метризация пространства и меры расстояния

Типизация метрик или различных способов измерения расстояний в пространстве переменных может быть проведена по следующим основаниям:

1. Дистанция отражает в первую очередь различия:

а) размеров многомерных геометрических тел;

б) формы многомерных геометрических тел.

2. Переменные не имеют структуры (номинальные переменные).

3. Переменные имеют собственную, естественную структуру порядка (порядковые переменные):

а) единую размерность (принадлежат к единой системе измерения);

б) все или несколько переменных измеряются в собственной системе измерения;

в) переменные измерены баллами (квалиметрически) или являются лингвистическими (размытыми).

Два типа дистанций: дистанция объема и дистанция подобия являются естественным отражением того факта, что изменение размеров может происходить независимо от изменения формы и наоборот. Так как эти изменения и связанные с ними отношения могут быть независимы, они могут определяться действием различных независимых факторов или, иначе говоря, иметь различную природу.

Первый тип метрики, отражающей различия объемов в общем случае, обобщается метрикой Минковского

$$D_{ij} = \left(\sum_{i=1}^{n} | (x_i - x_j)^p | \right)^{1/q}, \quad p(q) = 1, 2, 3, 4$$

и различными модификациями на ее основе.

Второй тип метрик строится на основе различных способов расчета коэффициента корреляции r_{ij}

$$D_{ij} = 1 - r_{ij}.$$

Метрики этих двух типов в полной мере применимы для порядковых переменных. Хотя метрики на основе корреляций при соответствующих способах измерения подобия могут использоваться и для номинальных переменных. Для номинальных переменных метрики строятся на теоретико-множественной основе и отражают одновременно как подобие, так и объем.

Дистанция в общем случае определяется числом элементов, принадлежащих разным множествам, т. е. числом несовпадающих элементов

$$D_{ij} = \sum_{i=1}^{n} (X_i \cup X_j) - \sum_{i=1}^{n} (X_i \cap X_j),$$

где $\sum_{i=1}^{n} (X_i \cup X_j)$ — сумма элементов, принадлежащих хотя бы одному из множеств (объединение); $\sum_{i=1}^{n} (X_i \cap X_j)$ — сумма элементов, которые принадлежат одновременно обоим множествам (пересечение).

При этом можно полагать, что признаки представлены как 1 и 0 (наличие, отсутствие), так и через число элементов в данном классе. Например, первому варианту соответствуют переменные, представляющие собой списки видов, в которых вид может присутствовать или отсутствовать. Во втором случае может быть список с указанием числа элементов для каждого вида (численности особей для каждого вида), при этом дистанцию определяют по формуле

$$D_{ij} = \sum_{i=1}^{n} \max(X_i, X_j) - \sum_{i=1}^{n} \min(X_i, X_j),$$

где $\sum_{i=1}^{n} \max(X_i, X_j)$ — сумма максимальных значений для каждого элемента из двух сравниваемых перечней; $\sum_{i=1}^{n} \min(X_i, X_j)$ — сумма минимальных значений из каждой пары.

Для двух переменных, представленных номинальными классами, может быть введена *информационная дистанция Кульбака* или мера расхождения на основе распределений вероятности их совместной встречаемости

$$D(x_i y_i) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{i=1}^{n} p(y_i) \log p(y_i) + \sum_{i=1}^{n} p(x_i y_i) \log p(x_i y_i),$$

где $p(x_i)$, $(p(y_i))$ — вероятность состояния i-й выборки $X(Y)$; $p(x_i y_i)$ — совместная вероятность i-го состояния в X и Y. Таким образом, эта мера прямо связана с измерением неопределенности для полиномиальной выборки и соответственно проверяет гипотезу независимости. Если две сравниваемые переменные X и Y подобны, то их совместная энтропия равна минимальной энтропии одной из двух сравниваемых переменных, и дистанция равна нулю. Если переменные взаимонезависимы, то дистанция равна максимальной энтропии двух сравниваемых переменных. Следует обратить внимание на то, что эта дистанция по содержанию ближе к мере подобия, и является псевдометрикой, так как для нее не выполняется аксиома неравенства треугольников.

Приведенные метрики можно рассматривать как базовые, но на их основе строятся разные модификации, отображающие различные варианты возможных отношений.

Если используется метрика на основе различия формы, то для нее безразлично, в каких единицах измеряется переменная. Но для дистанции, измеряющей различия количества или объема, способ измерения весьма существенен.

Допустим, что в каждой точке надо измерить температуру, количество осадков [мм], влажность воздуха [%]. При этом ставится задача измерить дистанции по этим переменным между парами точек или между самими этими переменными. Очевидно, что результаты таких измерений просто не имеют смысла. Соответственно, их необходимо привести к одной шкале. Это можно сделать несколькими способами:

1) разделить отклонения от среднего каждой переменной, на среднее квадратическое $\frac{x - M_x}{\sigma}$. Такая операция называется стандартизацией;
2) разделить все значения на среднее;
3) разделить все значения на медиану;
4) разделить все значения на максимум;
5) разделить все значения на минимум;
6) разделить все отклонения от среднего на среднее $\frac{x - M_x}{M_x}$;
7) разделить все отклонения от среднего на медиану;
8) разделить все отклонения от среднего на максимум;

9) разделить все отклонения от среднего на минимум.

Если последние два преобразования умножить на 100, то получим нормировку данных, представленную в процентах;

10) присвоить каждому значению ранг от 1 — для минимального значения, до N — для максимального с присвоением одинаковых значений одного и того же ранга или случайным образом разных рангов для равных значений, соответствующих порядку их выпадения.

Каждое из преобразований несколько изменяет свойства пространства. Наиболее нейтральное преобразование — деление отклонений от среднего на среднее квадратическое. Это преобразование не изменяет пропорции размаха. Деление отклонения от среднего на максимум автоматически приводит все переменные к одному масштабу варьирования сверху (от 1 до $-\infty$). Деление отклонения на минимум приводит к противоположной форме трансформации. Точно так же нелинейно трансформируется переменная при ее делении на свое среднее, минимум или максимум. Фактически это приводит к вариантам нелинейных преобразований. Впрочем, линеаризовать это преобразование можно с помощью логарифмирования. Какое из преобразований использовать зависит от того, какая часть ряда представляется для исследователя более важной. Если он ожидает, что принципиально важные отношения возникают при больших значения переменных, то оптимально нормирование на минимум, если же наоборот, наиболее существенные отношения предполагаются при малых численных значениях, то оптимально нормирование на максимум. Эффект нормирования по среднему и медиане существенно зависит от свойств распределения.

Эффект стандартизации определяется также во многом свойствами распределения. Если исследователь не располагает какими-либо конструктивными гипотезами, конкретизирующими способ выбора нормирования, то более оправданно использование преобразования отклонения по среднему значению или медиане.

Рассмотрим детально наиболее употребимые метрики и оценим возможные области их применения. При этом обратим внимание на то, что способ представления переменной и выбранная метрика фактически во многом предопределяют свойства пространства и все последующие результаты анализа.

Дистанция Минковского получается в результате комбинации целочисленных значений p и q. При $p = q = 1$ имеем «Сити блок манхеттен дистанс», при $p = q = 2$ — классическую дистанцию Евклида.

Будем рассматривать изменения соотношения между разными вариантами дистанции Минковского на примере уже использованных выше данных по влажности почв.

На рис. 6.1, *а—г* показано, как четыре варианта метрики Минковского, рассчитанные по исходным нелогарифмированным данным, искривляют пространство относительно метрики Евклида,

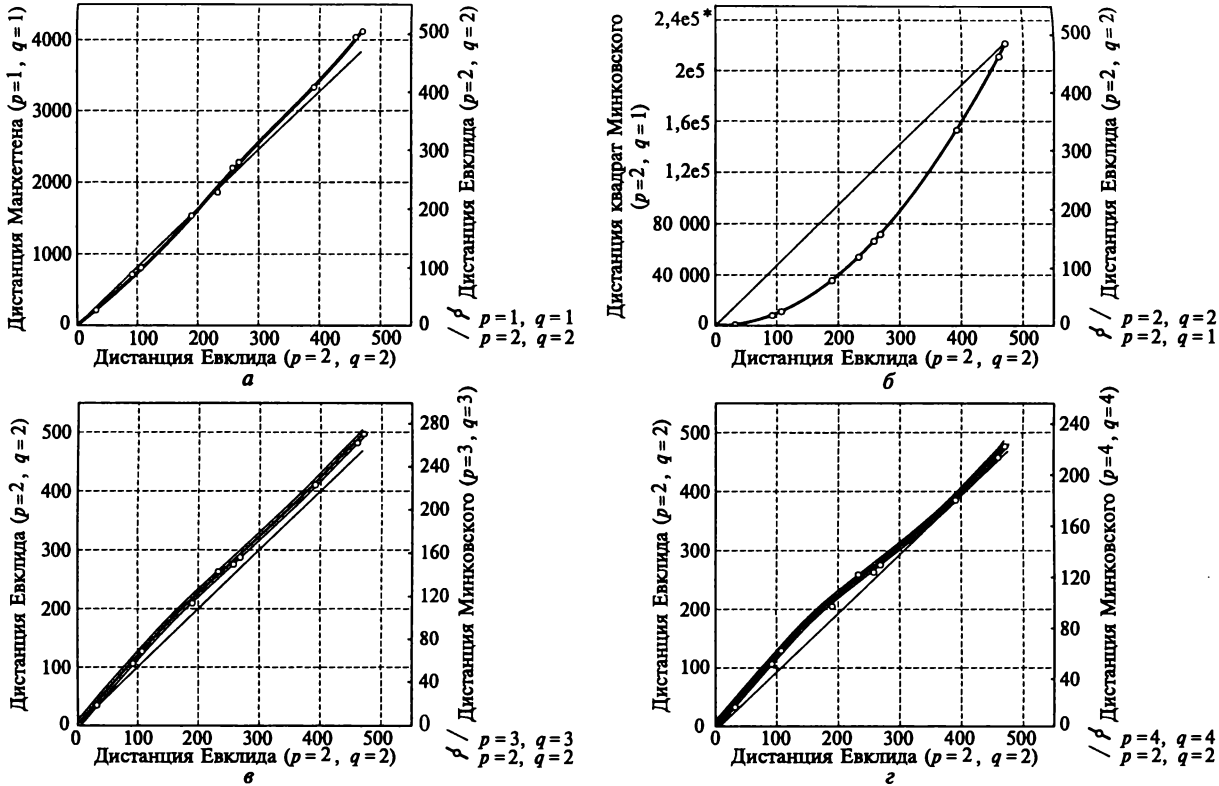


Рис. 6.1. Сравнение четырех вариантов (*а*—*г*) метрик Минковского с метрикой Евклида

* 2,4e5 = $2{,}4 \cdot 10^5$ — форма записи больших чисел

представленной на графиках центральной прямой линией. Очевидно, что отклонение дистанции Манхеттена относительно дистанции Евклида больше при больших значениях расстояний и меньше при малых. Еще в большей степени эффект относительного увеличения дистанций при больших расстояниях по Евклиду и относительное уменьшение этих значений при малых дистанциях проявляется у квадрата дистанции Евклида. Напротив, дистанции с $p = q = 3$ и $p = q = 4$ растягивают пространства относительно метрики Евклида при малых значениях и стягивают при больших (см. рис. 6.1, *в*, *г*).

Таким образом, общая схема трансформации пространства различными метриками вполне понятна. Метрики типа $p > q$, включая $p = q = 1$, превращают пространство в форму постепенно расширяющегося рога и существенно увеличивают значимость больших расстояний.

Пространства с метриками $p \leq q$, относительно пространства $p = q = 2$, напротив, растягивают пространство при малых дистанциях и стягивают при больших.

Рассмотрим, как влияет логарифмирование данных на форму пространства.

Очевидно, что дистанция Минковского от логарифмированных данных

$$D_{ij} = \left(\sum_{1}^{n} | (\log x_i - \log x_j)^p | \right)^{1/q} = \left(\sum_{1}^{n} \left| \left(\log \left(\frac{x_i}{x_j} \right) \right)^p \right| \right)^{1/q},$$

есть по сути логарифм отношения и это уже совершенно иное пространство, имеющее по смыслу мало общего с пространством, конструируемым по нелогарифмированным данным. Фактически — это пространство некоторых индексов или отношения многомерных гипотенуз геометрических фигур. С физической точки зрения, логарифмирование подразумевает, что отношения в системе между переменными описываются функцией типа $y = ax^b$, а не как $y = a + bx$.

На рис. 6.2 сравниваются измерения дистанций с использованием метрики Евклида для логарифмированных и исходных данных. Очевидно, что положительную связь между этими метриками можно рассматривать не более как общую тенденцию. Отношения существенно изменяют строение пространства, так как в одних областях относительно метрики Евклида происходит увеличение дистанции, а в других, напротив, уменьшение.

Исследуем дистанции, измеряющие различия форм многомерных фигур, т. е. строящихся на основе различных коэффициентов корреляции: Пирсона, Спирмена, Кендалла (τ) и гамма (γ).

Дистанцию Пирсона можно определить по формуле

$$D = 1 - \rho_{ij},$$

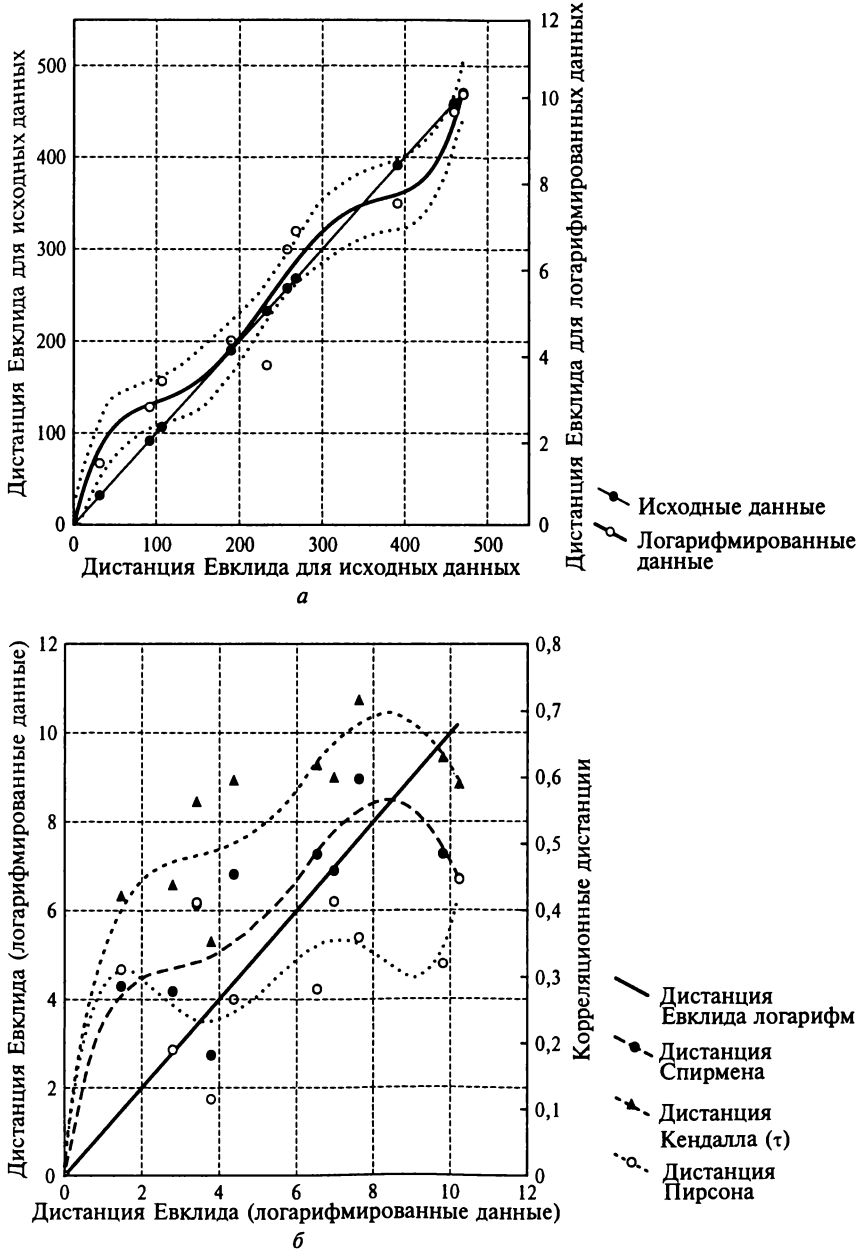


Рис. 6.2. Соотношение пространства с метриками на основе логарифмированных данных (*а*) и на основе метрик подобия (*б*) с метрикой Евклида

где ρ_{ij} — обычный коэффициент корреляции, используемый для нормального распределения зависимости.

Три остальные коэффициента корреляции — непараметрические или ранговые и мало чувствительны к формам распределения и зависимости.

Ранговые коэффициенты корреляции учитывают только порядок величины, т. е. для их расчета исходные данные преобразуются в ранги по описанной выше схеме. У каждой переменной наименьшее значение получает ранг 1, наибольшее — ранг N, где N — общее число наблюдений. Естественно полагать, что если ранги всех переменных совпадают, то корреляция между переменными равна единице.

Ранговый коэффициент корреляции Спирмена близок по содержанию стандартному коэффициенту корреляции

$$R_{ij}^{Sp} = 1 - \frac{6\sum_{i=1}^{n}(r_i - r_j)^2}{n^2 - n},$$

где r — ранг переменной $i(j)$; n — объем выборки.

В отличие от рангового коэффициента корреляции Спирмена в *коэффициенте корреляции Кендалла* (τ) не используется собственно значение разности рангов, а учитывается только ее знак

$$\tau_{ij} = \frac{2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\operatorname{sign}(r_j - r_i)}{n(n-1)},$$

где $\operatorname{sign}(r_j - r_i) = 1$, если $r_i > r_j$; $\operatorname{sign}(r_j - r_i) = -1$, если $r_i < r_j$; $\operatorname{sign}(r_j - r_i) = 0$, если $r_i = r_j$.

Коэффициент корреляции гамма (γ) отличается от тау (τ) тем, что совпадающие численные значения учитываются только один раз и все повторные аналогичные сочетания из расчета исключаются. В результате, например, многократно повторяющиеся значения нулей, отражающих в исходных данных отсутствие какого-либо вида растений при их повторной совместной встречаемости (отсутствие обоих видов), при расчете корреляции не учитывается. Естественно, что при этом объем выборки n уменьшается на величину исключенных сочетаний. Сразу же отметим, что коэффициент корреляции γ является наиболее адекватной метрикой для оценки дистанции между сообществами или видами растений и животных, оценивающей различия их структуры (соотношения видов, место вида в сообществе и др.).

Ранговый коэффициент конкордации Кендалла показывает степень согласованности значений всех переменных. Этот непараметрический критерий может рассматриваться как обобщенная мера

организованности системы. При этом под организованностью понимается существование какого-то порядка в отношениях между переменными, а в случае сообществ — во взаиморазмещении видов в пространстве.

Для расчетов по всем исходным значениям приписывается ранг, общий для всей выборки, и для каждой пары рассчитывается ранговый коэффициент корреляции — τ или γ.

Среднее значение коэффициента корреляции

$$M_\tau = \frac{mW - 1}{m - 1}.$$

Коэффициент конкордации определяют по формуле

$$W = \frac{1 + M_\tau(m - 1)}{m},$$

где m — число переменных.

Оценки коэффициента конкордации в рамках многих программ проводятся совместно с непараметрическим методом одновариантного дисперсионного анализа.

Из табл. 6.1 следует, что согласованность влажности в горизонтах очень высокая. Средние значения рангов и их суммы значительно более контрастно подчеркивают различия влажности в горизонтах, чем собственно средние значения.

Таблица 6.1

Одновариантный дисперсионный анализ Фридмана и коэффициент конкордации Кендалла (Friedman ANOVA and Kendall Coeff. of Concordance). Дисперсионный анализ (ANOVA) для логарифмированных значений влажности почв по горизонтам

Chi Sqr. (N = 121, df = 4) = 441,2836, p < 0,00000; коэффициент конкордации Coeff. of Concordance = 0,91174; средний ранговый коэффициент корреляции (Aver. rank) r = 0,91101

Переменная	Среднее ранга	Сумма ранга	Среднее	Среднеквадратическое отклонение
Влажность на глубине, см	Average Rank	Sum of Ranks	Mean	Std.Dev.
5	4,966942	601,0000	3,941682	0,364139
10	3,991735	483,0000	3,641521	0,319050
20	2,950413	357,0000	3,306880	0,234331
30	1,623967	196,5000	3,094980	0,157724
40	1,466942	177,5000	3,062907	0,170421

Итак, четыре меры корреляции преобразуются в дистанции по указанному выше правилу.

Рассмотрим, как эти вновь введенные дистанции связаны с пространством Евклида (см. рис. 6.1, *б*).

Очевидно, что все три корреляционные метрики дают существенно иную конфигурацию пространства, чем метрика Евклида. Это прямо указывает на существенную нелинейность отношений в этой системе. (Здесь использованы все данные, включая и экстремальные отклонения значений влажности.) В целом разные корреляционные метрики дают сходную картину, но метрика на основе коэффициента корреляции Пирсона существенно занижает дистанции и по конфигурации существенно отличается от дистанций на основе ранговых метрик. Дистанция на основе метрики Кендалла при прочих равных условиях дает наибольшие значения, но по конфигурации почти тождественна дистанции Спирмена.

Таким образом, в зависимости от способа преобразования и использования различных метрик, для одних и тех же данных получаем пространства с существенно различными свойствами. Поэтому в процессе исследования желательно подобрать пространство, наиболее адекватно отражающее отношения между переменными, характеризующими изучаемые явления. Однако сразу же необходимо отметить, что проблема выбора в некотором смысле наилучшей или адекватной метрики практически не разработана, так что в дальнейшем придется ограничиться простейшими рекомендациями.

Рассмотрение различных типов метрик завершим оценкой дистанции по информационной мере на основе таблиц кросс-табуляции или сопряженности. В качестве примера проанализируем сопряженность двух пород деревьев — березы и ели, для которых в каждой точке трансекта с шагом 10 м измерены суммы площадей сечений. Измерения проведены в Центрально-лесном биосферном заповеднике. Общее число измерений 662 (табл. 6.2).

Необходимо ответить на вопрос: можно ли отвергнуть гипотезу независимости в размещении в пространстве этих двух видов и какова мера связи между ними или, напротив, мера различия в их размещении в пространстве.

Исходные данные суммы площадей сечений (BSA) преобразованы в дискретные баллы как trunk[log(BSA + 1)] — округленное до целого значение логарифма по основанию два от измеренной величины. Единица добавляется к измеренной величине для того, чтобы избежать логарифма от нулевого значения.

В табл. 6.2 приведены условные распределения рангов обилия березы при фиксированном ранге обилия ели. В соответствии с базовым положением теории вероятности можно записать, что, если событие B_i не зависит от события E_j, то

$$p(B_i/E_j) = p(B_i).$$

Таблица 6.2

Кросс-табуляция отношений обилия березы и ели

Переменная		Береза (B). Балл обилия — i. Условная вероятность $p(B_i/E_j)$					H(B/ /E_j)	J(B/E_j)	$p(E)$	$p(E)$ J(B/E)
		1	2	3	4	5				
Ель (E). Балл обилия — j	1	**0,3684**	0,1579	0,1579	0,0526	**0,2632**	2,1021	0,0419	0,0287	0,0012
	2	0,1207	**0,3103**	0,3276	0,2069	0,0345	2,0573	0,0867	0,0877	0,0076
	3	0,0840	0,2101	**0,4118**	0,2101	**0,0840**	2,0734	0,0706	0,1800	0,0127
	4	0,1010	**0,2648**	0,3275	0,2334	0,0732	2,1352	0,0088	0,4342	0,0038
	5	0,0805	0,1724	**0,3563**	**0,3046**	**0,0862**	2,0875	0,0565	0,2632	0,0149
	6	**0,4000**	0,2000	0,2000	0,2000	0,0000	1,9219	0,2221	0,0076	0,0017
$p(B)$		0,1042	0,2311	0,3444	0,2402	0,0801		0,0000	1,0000	0,0419

Примечание. Полужирным шрифтом выделены ведущие компоненты.

Следовательно, распределения условных вероятностей по E_1, E_2, ..., E_6, если обилие березы не зависит от обилия ели (два вида в своем размещении не зависят друг от друга), должны быть тождественно подобны друг другу и подобны априорному распределению событий по баллам обилия березы.

Мера отличия или мера влияния каждого балла обилия ели на березу оценивается следующим образом:

$$\mathrm{J}(B/E_j) = \mathrm{H}(B) - \mathrm{H}(B/E_j) \quad —$$

информативность состояния j-й ели для березы;

$$\mathrm{H}(B) = -\sum_i p(B_i)\log p(B_i) \quad —$$

неопределенность (разнообразие состояний) березы (B);

$$\mathrm{H}(B/E_j) = -\sum_i p(B_i/E_j)\log p(B_i/E_j) \quad —$$

условная неопределенность березы при известном состоянии ели.

Чем больше информативность, тем больше условное распределение отличается от априорного и тем больше условная сопряженность состояний березы с избранным состоянием ели. Выделим ячейки, в которых $p(B_i/E_j) > p(B_i)$. Эти ячейки можно рассматривать как сочетания состояний, в которых формируется положительная сопряженность между состояниями двух пород. Будем называть такие сочетания состояний «характерными». Анализ таблицы показывает, что при отсутствии ели характерно как отсутствие, так и высокое обилие березы. При малом обилии ели характерно малое обилие березы. При среднем обилии ели (строка 3) характерно среднее обилие березы, а также в некоторой степени очень высокое. При обилии ели (строка 4) характерно низкое обилие березы. Затем при высоком обилии ели (строка 5) характерно оби-

лие березы от среднего до очень высокого, при наиболее характерном «высоком», а при очень высоком обилии ели (строка 6) береза вообще отсутствует. Таким образом, намечается в целом нелинейная сопряженность ели и березы, проявляющаяся в том, что очень высокое обилие березы существует при не очень высоком обилии ели, и точно так же, при очень высоком обилии березы характерно отсутствие ели.

Следует отметить, что $2n_j\mathrm{J}$ — χ^2-распределение с $(k-1)$ числом степеней свободы, где n_j — объем выборки по строке; k — число состояний (в данном случае березы).

Общая сопряженность в системе вычисляется по формулам:

$$\mathrm{I}(B,E) = \sum_j p(E_j)\,\mathrm{J}(B/E_j) -$$

суммирование по строке j или

$$\mathrm{I}(B,E) = \mathrm{H}(B) - \mathrm{H}(B/E);$$

$$\mathrm{H}(B/E) = \sum_j p(E_j)\,\mathrm{H}(B/E_j) -$$

условная неопределенность березы по ели.

Обобщая эти отношения, запишем:

$$\mathrm{I}(B,E) = \mathrm{H}(B) - \mathrm{H}(B/E) = \mathrm{H}(E) - \mathrm{H}(E/B) =$$

$$= \mathrm{H}(E) + \mathrm{H}(B) - \mathrm{H}(E,B).$$

Здесь $2n\mathrm{I}(B,E)$ — χ^2-распределение с $(k-1)(m-1)$ числами степеней свободы.

Соответственно, можно ввести дистанцию как

$$\mathrm{DI}(E/B) = 1 - \mathrm{I}(B,E)/\mathrm{H}(E);$$

$$\mathrm{DI}(B/E) = 1 - \mathrm{I}(B,E)/\mathrm{H}(B);$$

$$\mathrm{DI}(B,E) = 1 - \mathrm{I}(B,E)/(\mathrm{H}(B) + \mathrm{H}(E)).$$

Таким образом получили, что дистанция несимметрична, поэтому в частности она и является псевдометрикой.

В данном случае $\mathrm{H}(B) = 2{,}1440$ бит; $\mathrm{H}(E) = 1{,}983335$; $\mathrm{I}(B,E) = 0{,}0419$ бит.

Очевидно, что сопряженность между обилием березы и ели очень мала и соответственно дистанция между ними велика. При этом, если виды полностью независимы, то дистанция между ними максимальна.

В табл. 6.3 приведены различные оценки сопряженности обилия ели и березы. Хотя сопряженность, безусловно, низкая, но все-таки в соответствии с критерием χ^2 гипотеза независимости отвергается.

По всем трем ранговым коэффициентам корреляции фиксируется слабая, но статистически значимая, положительная сопряженность, что в целом соответствует генеральной структуре отно-

Таблица 6.3

Оценки сопряженности в пространстве ели и березы по различным критериям

Критерий	Оценка		
Хи-квадрат Pearson Chi-square	Chi-square = = 45,652	df = 20	p = 0,00090
Кендалла (τ) двусторонний Kendall's tau b&c	b = 0,07542	c = 0,06855	
Кендалла (τ)	0,075425	z = 2,903269	p = 0,003693
Гамма (Gamma)	0,103312	z = 2,903269	p = 0,003693
Спирмена Spearman Rank R	0,089	t = 2,3015	p = 0,02168
Коэффициент неопределенности	E = 0,02108	B = 0,01949	(E,B) = 0,0099
Информационная дистанция	DI(E/B) = 0,9789	DI(B/E) = = 0,9851	DI(B,E) = 0,9901

шений между состояниями в таблице сопряженности. Очевидно, что и по корреляционным метрикам, и по информационной псевдометрике дистанции между видами в определяемом ими многомерном пространстве очень большие.

Итак, существует широкий диапазон возможностей определения расстояний между переменными или элементами, что позволяет исследовать соотношения между переменными, измеренными практически любыми способами, имеющими или не имеющими собственную структуру порядка. Все они в конечном итоге создают основу для решения практически любых задач многомерного непараметрического анализа.

6.2. Многомерное непараметрическое шкалирование

Задача многомерного шкалирования та же, что и у факторного анализа: найти минимальное число факторов, имеющих физическую интерпретацию и описывающих варьирование всех включенных в систему переменных. Однако логические основания метода совершенно иные. Суть метода можно продемонстрировать на следующем примере.

Допустим, мы располагаем данными об измерении расстояний между n городами. Очевидно, что это $n(n-1)/2$ пар измерений. Требуется определить положения этих городов в ортогональ-

ных координатах X, Y — аналогах широте и долготе. Рассмотрим геометрический способ решения задачи. Для этого возьмем два города s1 и s2, между которыми расстояние наибольшее, и отложим его в виде прямой на бумаге. Затем циркулем с радиусом d(s1,s3) проведем первую окружность с центром в точке s1 и с радиусом d(s2,s3) — вторую с центром s2. Точка пересечения окружностей определит положение города s3 на плоскости, т.е. в ортогональной системе координат. Теперь, используя радиусы d(s1,s4), d(s2,s4) и d(s3,s4), нанесем четвертый город s4. Три окружности лишь в частном случае пересекутся в одной точке, так как с одной стороны расстояния измерены с определенной точностью, а с другой — они есть расстояния, полученные на поверхности сфероида, т.е. трехмерного пространства. Обычно пересечение окружностей образует то, что можно назвать треугольником погрешности. Можно полагать, что наиболее вероятное положение города s4 соответствует центру тяжести этой фигуры. Примем этот центр тяжести за положение города s4 и, опираясь на него как на первую точку, построим радиусами, соответствующими измеренным, новые окружности с радиусами d(s1,s4), d(s2,s4) и d(s3,s4). Эти новые окружности будут образовывать треугольники погрешности для каждого из трех городов. Определим центр тяжести для треугольника с наибольшей погрешностью и вновь повторим построение окружностей. Затем ту же операцию применим для двух оставшихся городов. Измерим полученные новые расстояния и оценим сумму квадратов ошибок как сумму квадратов разности исходных и новых дистанций. Повторим всю процедуру с новыми дистанциями. Треугольники погрешности, скорее всего, сохранятся, но их площадь будет меньше, чем в первоначальной конфигурации. Будем продолжать процедуру до тех пор, пока сумма квадратов ошибок не станет равной нулю или минимальной. Примем эту конфигурацию как наилучшую. По индукции очевидно, что ту же процедуру можно выполнить для трехмерного пространства. Логично полагать, что в этом случае новые дистанции, определяющие положения городов после подгонки, будут меньше отличаться от исходных, чем в двухмерном пространстве.

Так как любые геометрические преобразования могут быть описаны на языке алгебры, можно найти алгебраические способы поиска наилучшего соответствия расчетных дистанций исходным для пространства любой размерности.

В нашем примере с городами желательно, чтобы рассчитанные координаты в общем соответствовали бы географическим. Очевидно, что это всегда можно достигнуть поворотом пространства на некоторый угол.

В общем случае нас интересует интерпретируемость координат. Эту интерпретируемость как и в факторном анализе можно полу-

чить на основе вращения координат таким образом, чтобы переменные, наиболее удаленные друг от друга, маркировали бы первую ось; переменные, в основном или в существенной степени определяемые второй координатой и по этой координате максимально удаленные друг от друга, маркировали бы вторую ось и т.д. Фактически это означает, что на первую координату приходится максимальная дисперсия значения переменных, на вторую, ортогональную первой, несколько меньшая и т.д. При этом всегда возможно добиться того, чтобы математическое ожидание значений переменных по каждой координате было бы очень близко к нулю.

Методы многомерного шкалирования делятся на метрические и неметрические. *Метрические методы* по сути опираются на модель факторного анализа, но при произвольной исходной дистанции между переменными или объектами. *Неметрические методы* используют только порядок размещения пар объектов по дистанции, начиная от наиболее близких к более удаленным. Алгебраические преобразования поиска наилучшей конфигурации объектов в пространстве заданной размерности можно осуществлять различными способами, набор которых, используемый в стандартных пакетах статистических программ, совершенно определенно не исчерпан. Геометрический аналог метода показывает принцип реализуемости такой задачи, а рассмотрение конкретных методов выходит за пределы возможностей данного пособия. Читатель, ищущий более глубокого понимания технологических аспектов процедуры многомерного шкалирования, может познакомиться с ними по литературе, приведенной в заключительной части пособия. Здесь же ограничимся необходимым минимумом сведений, позволяющим грамотно использовать этот наиболее общий метод многомерного анализа.

Оценка соответствия конфигурации, получаемой в результате многомерного шкалирования, осуществляется на основе функции напряжения — «стресса» (stress) и/или коэффициента отчуждения (alienation). Использование именно этих слов связывается не более чем с традицией, определяемой историей развития метода многомерного шкалирования в социологических исследований. Так, например, в рамках этой традиции переменные обычно называются стимулами.

В основном используются две **функции стресса**:

1. *Стресс-функции Краскала*

$$S_1 = \left(\frac{\sum_{ij} (d_{ij} - \bar{d}_{ij})^2}{\sum_{ij} \bar{d}_{ji}^2} \right)^{1/2}, \quad S_2 = \left(\frac{\sum_{ij} (d_{ij} - \bar{d}_{ij})^2}{\sum_{ij} (\bar{d}_{ij} - \bar{\bar{d}})^2} \right)^{1/2},$$

где d_{ij} — дистанции, измеренные на основе исходных данных или на одном из шагов итерационной процедуры поиска наилучшей конфигурации; $\bar{d}_{ij}$ — дистанция, полученная по конечной конфигурации; $\bar{d}$ — средняя дистанция по всей выборке для конечной конфигурации.

2. *Коэффициент отчуждения k*

$$k = (1 - M_x^2)^{1/2}$$

$$M_x = \frac{\sum d_{ij}\,\bar{d}_{ij}}{\left[\left(\sum_{ij} d_{ij}^2\right)\left(\sum_{ij} \bar{d}_{ij}^2\right)\right]^{1/2}}.$$

Очевидно, что стресс по существу есть нормированная среднеквадратическая ошибка аппроксимации расстояний в исходной конфигурации новой конфигурацией, рассчитанной в многомерном шкалировании.

Если стресс равен нулю, то новая конфигурация абсолютно точно воспроизводит исходные дистанции. Коэффициент отчуждения — просто дистанция, получаемая на основе расчета скалярного произведения векторов между исходными и рассчитанными дистанциями. Если коэффициент отчуждения равен нулю, то конфигурации дистанций полностью подобны.

Исходная система для многомерного шкалирования определяется так же, как и для факторного анализа.

Элементом системы является точка наблюдений в пространстве-времени с измеряемыми свойствами или переменными. Положение свойств в пространстве координат векторного пространства отражает их чувствительность к этим координатам (неизвестным латентным факторам). Положение точек наблюдения в координатах Евклидова пространства отражает состояния этих латентных факторов в каждой точке пространства-времени.

Последовательность действий при использовании методов многомерного непараметрического шкалирования сводится к следующим основным шагам:

Шаг первый. Выбирается и обосновывается метрика, наиболее соответствующая свойствам объекта;

Шаг второй. Рассчитываются матрицы дистанций между переменными;

Шаг третий. Осуществляется оценка размерности или числа независимых координат, достаточных для отображения существенной части пространственно-временно́го варьирования;

Шаг четвертый. Строится отображение переменных в k-мерном векторном пространстве и осуществляется экспертная оценка интерпретируемости полученных отношений;

Шаг пятый. Осуществляется расчет значений координат точек (элементов) в k-мерном факторном пространстве;

Шаг шестой. Проводится оценка качества отображения переменных (свойств) в многомерном факторном пространстве;

Шаг седьмой. Осуществляется поиск оснований для интерпретации физического смысла координат-факторов.

Рассмотрим основные методы решения важнейших задач многомерного непараметрического шкалирования. В качестве примера возьмем приведенную выше систему, отображающую варьирование в пространстве для пяти различных глубин (1—5, 2—10, 3—15, 4—20, 5—30, 6—40 см) относительной влажности почвы (M1, M2, M3, M4, M5), кислотности (pH1, pH2, pH3, pH4, pH5), обменных оснований фосфора (P1, P2, P3, P4, P5), калия (K1, K2, K3, K4, K5), кальция (Ca1, Ca2, Ca3, Ca4, Ca5) и магния (Mg1, Mg2, Mg3, Mg4, Mg5) — всего 30 переменных. Так как распределения близки к гамма или логнормальным и характеризуются значительными выбросами, то исходные данные логарифмированы. В отличие от параметрических методов экстремально большие значения из выборки не исключаются.

Шаг первый. Очевидно, что переменные измеряются различными методами и их числовые значения не имеют общей размерности. Влажность измерялась в процентах, pH — в условных единицах, концентрации элементов — в миллиграммах, эквивалентах. Кроме того, различные элементы представлены с весьма разными средними концентрациями.

Если бы все переменные были измерены в одной системе измерений, то логично использовать метрику Евклида. В этом случае величина дистанции будет во многом функцией средней концентрации и переменные с малыми средними концентрациями будут наиболее удаленными от остальных. Если цель исследования — получить правила соотношений концентраций веществ, то метрика Евклида — оптимальна. Если необходимо исследовать отношения между переменными и правила их варьирования в пространстве, то метрика Евклида не подходит. В этом случае можно применить несколько способов:

1. Использовать нормированную метрику Евклида

$$D_{i,i+k,j} = \frac{\sqrt{\sum_j (x_{ij} - x_{i+kj})^2}}{\sqrt{\sum_j (x_{ij}^2 + x_{i+kj}^2)}}$$

или ее нормированные аналоги метрик Минковского.

2. Нормировать исходные значения таким образом, чтобы они были соизмеримы по амплитуде варьирования. Наиболее приемлема процедура стандартизация всех данных: деление отклонений от

среднего на среднее квадратическое. Если нас интересует отображение роли экстремальных значений, то такая схема нормирования переменных — оптимальна. Далее к стандартизированным данным можно применять метрику Евклида. В этом варианте представления данных она во многом будет отображать подобие пространственного варьирования переменных. Процедура стандартизации данных существует во всех пакетах статистических программ и легко выполнима.

3. Подобие отношений можно эффективно оценить, используя метрики на основе корреляций. В частности, в данном случае наиболее приемлема метрика на основе ранговой корреляции Спирмена. Если отношения в системе линейны, то метрика Евклида по стандартизованным данным и метрика Спирмена должны давать тождественные результаты.

Шаг второй. Матрицы дистанций обычно рассчитываются в рамках меню, предлагаемом в используемом пакете программ. Пакеты существенно различаются по перечню включенных в них метрик. Далеко не во всех программах вводятся дистанции на основе ранговых корреляций, поэтому приходится сначала рассчитать корреляционную матрицу в разделе «корреляция», затем вычесть все ее значения из 1, превращая ее в матрицу дистанций. В большинстве пакетов есть соответствующие средства, позволяющие преобразовать всю матрицу корреляционных значений в матрицу дистанций. В частности, эта операции легко выполнима в Excel.

Следует обратить внимание на то, что в разных программных средствах матрицы представлены по-разному. В одних программах используются диагональные матрицы с заполнением верхней или нижней части, в других — полная матрица с соответствующей кодовой строкой, индуцирующей, является ли эта матрица матрицей подобия или различия. Поэтому сначала в справке необходимо посмотреть способ записи матрицы.

Целью отображения систем в различных метриках является либо выбор в некотором смысле оптимальной дистанции, либо исследования вклада и характера нелинейных отношений. Однако оба этих важных аспекта многомерного анализа практически не разработаны и их изложение выходит за рамки настоящего пособия.

Поэтому ограничимся демонстрацией многомерного шкалирования для дистанции Евклида при стандартизированных данных.

Шаг третий. После того как матрица дистанций рассчитана, переходим собственно к операции многомерного шкалирования переменных. Организация многомерного шкалирования в различных пакетах существенно различается. Наиболее полно она представлена в пакете Statistica. Программа демонстрирует все этапы вычисления, рассчитывает все виды стресс-функции и содержит полезную дополнительную информацию и хорошие графические сред-

ства. Недостатком ее является то, что максимальная размерность, при которой можно рассчитывать координаты переменных в векторном пространстве, равна девяти, а число переменных ограничивается 70. Другие пакеты могут быть свободны от этих недостатков и содержать набор различных методов многомерного шкалирования, но обычно дают менее наглядное представление результатов.

В данном случае будем решать задачу на основе пакета Statistica.

Общая схема анализа сводится к следующему. Сначала задаем максимальную размерность и рассчитываем соответствующую ей конфигурацию. Программа дает оценки первого стресса Краскала в двух вариантах: относительно исходных данных Hat-stress и относительно дистанций в исходной конфигурации, которая обычно рассчитывается по схеме факторного анализа Star-stress. Далее приводится коэффициент отчуждения и второй стресс Краскала. Последовательно уменьшая размерность, получаем новые конфигурации с соответствующими оценками функций стресса и отчуждения. Естественно, что чем меньше размерность, тем больше значения стресса. На рис. 6.3 приведен график, демонстрирующий зависимость величины стресса от размерности.

В данном случае все функции, отражающие качество конфигурации, с уменьшением размерности изменяются подобно, при этом более монотонны функции отчуждения и общего стресса.

Самым контрастным является стресс 1 (Hat). Следует отметить, что такое соотношение типично. Хотя встречаются ситуации, когда коэффициент отчуждения меняется по несколько иному правилу, чем функции стресса. Однако практически можно ограничиться функцией первого стресса. В логарифмической шкале видно, что стресс меняется в общем как нелинейная функция от размерности и лишь при размерности 4 и 3 заметен перегиб. Чтобы уточнить оценку размерности, можно построить гладкую модель функции «стресс-размерность», которая бы имитировала чисто случайный процесс.

В зависимости от исходных распределений и типов метрик гладкая функция имеет вид

$$1 - \log(\text{Stress}) = a + b\log \text{D}$$

или

$$2 - \log(\text{Stress}) = a + b\log(\text{D} + c).$$

В данном случае модель соответствует функции 1 (см. рис. 6.3, табл. 6.4). Как видно из графика, функция стресса практически точно аппроксимируется моделью случайного процесса, пересекаясь с моделью при размерности 4 и размерности 7. Следует отметить, что система может быть иерархически организована и в ней могут существовать две оптимальные размерности. Но так или ина-

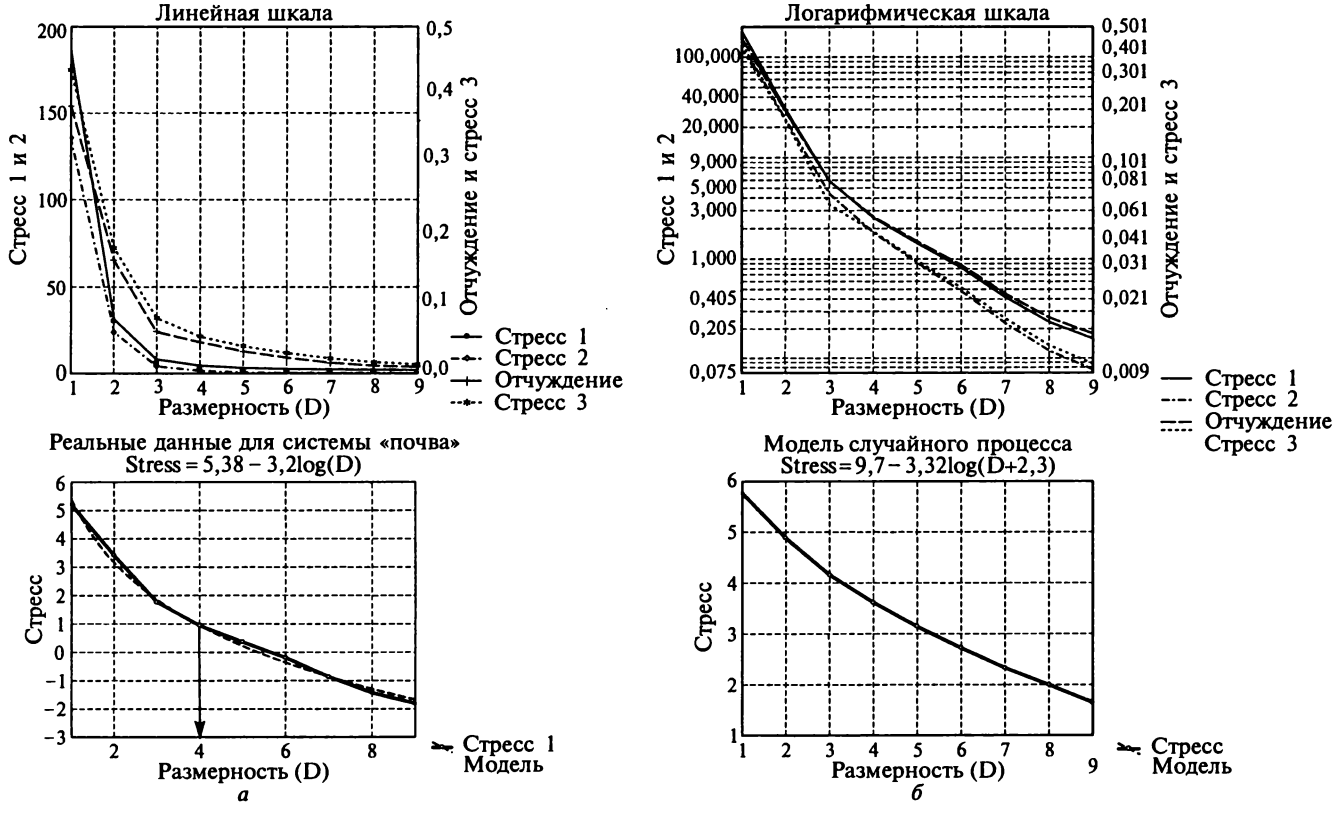


Рис. 6.3. Функции стресса и отчуждения для дистанции Евклида в двух формах отображения:
а — линейной; *б* — логарифмической

че ситуация в данном случае весьма неопределенна и возможно мы имеем дело с системой с весьма независимыми переменными и потому очень близкой к случайной модели. Необходимо также отметить, что обычно реальный стресс четко отличается от модели и в выборе размерности по точке его пересечения с линией модели не вызывает затруднения.

Уточненное решение задачи можно получить, создав с помощью генератора случайного процесса 30 нормально распределенных стандартизированных переменных. Рассчитав для них матрицу Евклида и проведя оценки стресса при уменьшении размерностей, получим модель чисто случайного процесса, имитирующего реальность.

К строго случайному процессу применима только модель 2-го типа (табл. 6.4). Как следует из рис. 6.3 и табл. 6.4, модель случайного процесса однозначно описывается монотонной функцией, от линии которой вообще нет отклонений. Теперь на основе уравнения регрессии стресса для реальных отношений и модели можно уточнить минимально допустимую размерность пространства. Из рис. 6.4 видно, что значения измеренного стресса достоверно отличаются от модели при размерности 4 и 3, что позволяет признать минимально допустимую размерность пространства, равную 4.

Таким образом, задачу третьего шага можно считать решенной.

Шаг четвертый. Отображение переменных в четырехмерном пространстве получаем на основе расчета конфигурации векторного четырехмерного пространства методом многомерного шкалирования. Абсолютные значения коэффициентов отражают чувствитель-

Таблица 6.4

Оценка параметров показательных регрессионных моделей для функции «стресс — размерность»

Объект	R^2, ошибка	Параметры	a	b	c
Стресс исследуемой системы	0,99537 0,19831	Оценка Estimate	5,38509	–3,2092	
		Ошибка Std.Err.	0,13037	0,0827	
		t-критерий	41,30487	–38,7887	
		p-level	0,00000	0,0000	
Стресс модели случайного процесса	0,99980 0,00296	Estimate	9,7388	–3,324	2,30877
		Std.Err.	0,0732	0,027	0,05187
		t-критерий	133,0156	–124,862	44,51501
		p-level	0,0000	0,000	0,00000

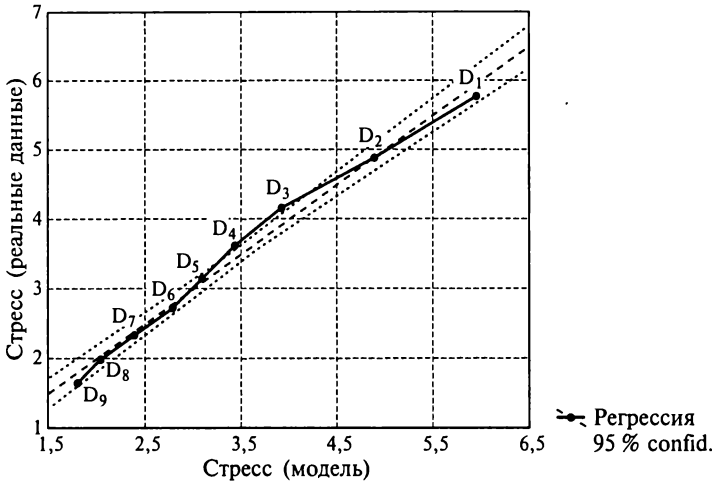


Рис. 6.4. Регрессия между стрессами по реальным данным и по модели случайного процесса с заведомо независимыми переменными

ность переменных к каждому виртуальном фактору. Знак при коэффициенте показывает, увеличиваются или уменьшаются значения переменных в Евклидовом пространстве при изменении значений пока не известных координат, в которых определяется положение точек наблюдения.

Опишем общие правила трактовки отношений по положению переменных в векторном пространстве.

1. Если две переменные зависят в наибольшей степени от разных факторов (по отношению к одному из них абсолютное значение коэффициента чувствительности у одной из переменных максимально, а по отношению к другому близко к нулю, а для второй переменной — наоборот), то эти переменные заведомо независимы и их варьирование в пространстве зависит от различных физических факторов.

2. Если две переменные зависят от одной координаты (большие абсолютные значения коэффициентов чувствительности), но различаются по знаку, то они заведомо зависимы, но реакция их на изменения значения координаты в Евклидовом пространстве — противоположна.

3. Две переменные с близкими значениями коэффициентов чувствительности неизбежно коррелируют в пространстве и их корреляция тем выше, чем более сходны значения их коэффициентов.

4. Физический смысл координаты нужно искать в области физических процессов, определяющих варьирование наиболее чувствительной к ней переменной.

5. Если переменная зависит в основном от одного фактора, будем называть зависимость «очень простой», от одного основного

фактора и второго с коэффициентом примерно в два раза ниже первого — «простой», от трех факторов при наличии хотя бы одного значительного коэффициента чувствительности — «средней», при наличии трех факторов со средним уровнем значения коэффициентов — «сложной» и при отсутствии четких ведущих факторов — «очень сложной». Если все коэффициенты по абсолютному значению невелики, будем называть такую зависимость — «неполной». Характеризуя таким образом отношения переменной к координатам, подразумеваем сложность механизмов, лежащих в основе ее пространственного варьирования.

На рис. 6.5 представлены коэффициенты чувствительности к четырем осям в форме диаграммы. Опираясь на приведенные выше критерии, попытаемся извлечь из отражаемых ими отношений содержательную информацию. Анализ рис. 6.5, а также дополнительных данных, приведенных в таблице 6.5, показывает что: 1) пространственное варьирование переменных от влажности до концентрации калия практически не зависит от состояния других переменных и скорее всего определяется сходными механизмами, отображаемыми первой координатой; 2) концентрация фосфора слабо связана с состояниями других элементов и описывается механизмами, отображаемыми второй координатой; 3) остальные

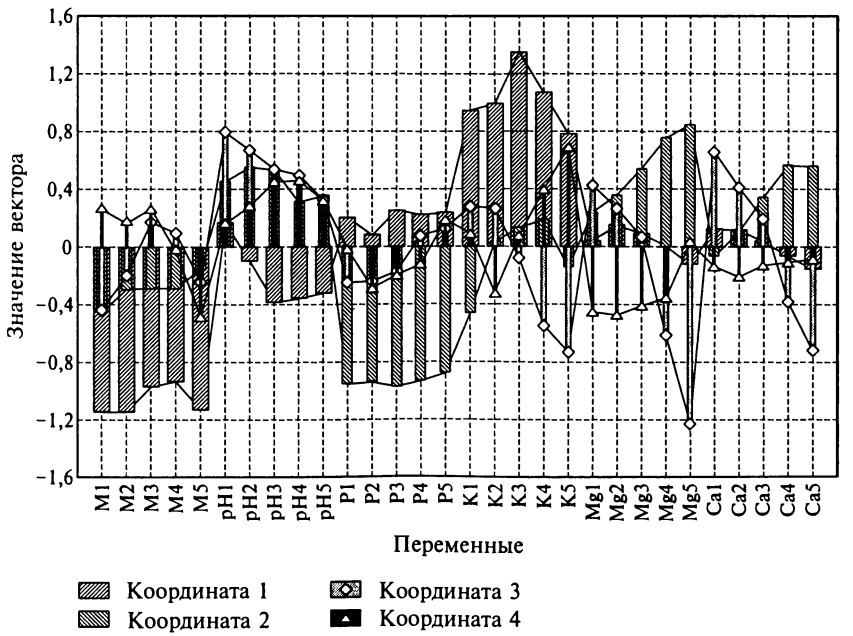


Рис. 6.5. Диаграмма коэффициентов чувствительности переменных к координатам векторного пространства

Таблица 6.5

Значения коэффициентов чувствительности по отношению к четырем координатам векторного пространства

Переменные	Координата 1	Координата 2	Координата 3	Координата 4	Сложность зависимости
M1	**–1,14588**	–0,453275	–0,43635	0,271494	Простая
M2	**–1,14624**	–0,294594	–0,19972	0,163420	Простая
M3	**–0,97113**	–0,287536	0,17079	0,263999	Простая
M4	**–0,93454**	–0,287394	0,09585	–0,031978	Простая
M5	**–1,13010**	–0,155643	–0,24523	*–0,486471*	Простая
pH1	0,16237	*0,451661*	**0,79187**	0,161445	Средняя
pH2	–0,09956	*0,551510*	**0,66696**	0,282752	Средняя
pH3	–0,38535	*0,531391*	*0,53638*	*0,449997*	Очень сложная
pH4	–0,36124	0,308971	*0,49688*	*0,457448*	Очень сложная
pH5	–0,32188	0,350502	0,32494	0,322132	Очень сложная
P1	0,20308	**–0,953547**	–0,24805	–0,018221	Очень простая
P2	0,07782	**–0,941085**	–0,24109	–0,286962	Очень простая
P3	0,25231	**–0,973295**	–0,17196	–0,198037	Очень простая
P4	0,21993	**–0,934246**	0,07493	–0,123231	Очень простая
P5	0,23396	**–0,879516**	0,12571	0,188368	Очень простая
K1	***0,94548***	**–0,464844**	0,28019	0,089471	Простая
K2	***0,99213***	0,031818	0,26348	–0,322508	Простая
K3	***1,35138***	0,129774	–0,08203	0,075675	Простая

K4	***1,06605***	0,185781	*–0,54678*	0,397276	Простая
K5	***0,78227***	–0,145302	*–0,72969*	**0,691007**	Средняя
Mg1	0,04186	0,237826	0,42442	–0,451716	Сложная
Mg2	0,14563	0,357377	0,26451	–0,473808	Сложная
Mg3	0,09092	0,539025	0,06782	–0,410337	Сложная
Mg4	0,00418	***0,752111***	–0,61414	–0,353847	Средняя
Mg5	–0,12451	***0,847077***	**–1,22563**	0,008156	Средняя
Ca1	0,12088	–0,065882	*0,65632*	–0,136460	Простая (неполная)
Ca2	0,11517	0,102841	0,41054	–0,208356	Простая (неполная)
Ca3	0,03555	0,341651	0,19168	–0,127957	Средняя (неполная)
Ca4	–0,06494	*0,561696*	–0,38195	–0,107322	Средняя (неполная)
Ca5	–0,15558	*0,555147*	*–0,72066*	–0,085432	Средняя (неполная)
Характеристика	Определяет варьирование влажности и концентрации калия. При этом чем больше влажность, тем меньше концентрация калия.	Определяет варьирование концентрации фосфора и магния в нижних горизонтах, в меньшей степени pH и концентрацию кальция. Чем выше концентрация фосфора, тем меньше концентрация магния и кальция и ниже pH.	Определяет pH в верхних и средних горизонтах, в меньшей степени — концентрацию калия и магния в нижних горизонтах и кальция в верхнем и (с противоположным знаком) нижнем горизонтах.	Определяет концентрацию калия в нижнем горизонте и pH в средних горизонтах и в горизонтах с отрицательным знаком концентрации магния.	

Примечание. Полужирный шрифт — отрицательное высокое влияние координаты; полужирный курсив — положительное высокое влияние координаты; подчеркнуто — отрицательное среднее влияние; курсив — положительное среднее влияние.

переменные зависят одновременно от нескольких координат и механизмы, описывающие их пространственное варьирование, почти наверное имеют многофакторную природу. Можно также отметить, что на глубине 40 см практически у всех переменных форма зависимости сложнее, чем в вышележащих горизонтах. Если иметь в виду, что эта глубина в среднем соответствует генетическому горизонту A2B, то такое отношение, скорее всего, имеет определенный физический смысл.

Шаг пятый. Определить координаты положения точек для пространства Евклида можно двумя методами:

1) поиском конфигурации точек наблюдений по матрице дистанций, определенной для всего их множества, используя непосредственно технологию многомерного непараметрического шкалирования для пространства с определенной выше размерностью;

2) на основе решения системы уравнений

$$y_i^j = \alpha_{1i}x_1^j + \alpha_{2i}x_2^j + \alpha_{3i}x_3^j + \alpha_{4i}x_4^j,$$

где y_i^j — известное значение переменной i в точке (элементе) j; $\alpha_{1i}...\alpha_{4i}$ — значение коэффициента чувствительности переменной i при координате 1 или 2 или 3 или 4, определенное в результате решения задачи многомерного шкалирования на предыдущем шаге анализа; $x_1^j...x_4^j$ — искомое значение координаты 1 или 2 или 3 или 4 для точки j.

Первый метод, реализованный с той же метрикой, что и на предшествующем шаге анализа по тождественно трансформированным данным, дает наилучшую подгонку конфигурации к реальности. Однако в большинстве программ введено ограничение на размер матрицы или на возможности расчета, которые весьма объемны и существенно ограничены мощностью компьютера. Второй метод не содержит никаких ограничений на объем наблюдений, но дает оценки координат лишь как средневзвешенные из системы i уравнений. Соответственно, он весьма чувствителен и к неизбежным ошибкам измерений и обычно не может описать экстремальные значения координат. Впрочем, последний недостаток легко корректируется аппроксимацией переменных в регрессионных моделях от квадратичных форм координат. Эта операция не изменяет физического содержания координат, но позволяет отображать большую амплитуду варьирования переменных. Вторым недостатком этого метода является то, что он требует написания специальной программы, которую, впрочем, можно скомпилировать в Excel.

Результаты расчетов, осуществляемых этими двумя методами, в общем случае не обязаны совпадать. Из простых алгебраических оснований очевидно, что второй метод точно преобразует отобра-

жение в векторном пространстве в Евклидово, но совершенно необязательно, что это отображение будет соответствовать результатам прямого поиска конфигурации на основе матрицы дистанций между элементами (точками) системы методом многомерного шкалирования. Поэтому целесообразно одновременно рассчитывать координаты Евклидова пространства различными методами, включая и методы факторного анализа. Окончательный выбор отображения будет определяться физической интерпретируемостью координат (табл. 6.6).

Очевидно, что прямой факторный анализ дает почти такое же отображение, как многомерное шкалирование по матрице дистанций между элементами исходной системы. Меняется лишь направление первой и четвертой координаты, что не имеет принципиального значения. Координаты, рассчитанные на основе коэффициентов векторного пространства, совпадают по абсолютным значениям с тремя координатами факторного анализа, но первая координата в прямом факторном анализе и, соответственно, полученная методом многомерного шкалирования, не имеют прямого соответствия в системе на основе преобразования векторного пространства. В свою очередь, четвертая координата этого пространства не имеет прямого отображения в первых двух способах расчета. Отображение, получаемое на основе вращения факторного анализа, вообще не имеет аналогов. Если пытаться описывать первую координату факторного анализа координатами, полученными из векторного пространства, то это возможно только на 57 %, а четвертую координату, полученную через векторное пространство, — только на 7 % координатами факторного.

Таким образом, имеем два несколько отличных отображения наблюдений в пространстве Евклида. Третье отображение, полученное на основе вращения, по информации тождественно исходному факторному и отличается от него только ориентацией осей. Поскольку два варианта отображения различаются только парой координат, то их сравнение не представляет особой сложности. Из рис. 6.6 следует, что первая координата, определенная на основе многомерного шкалирования и факторного анализа, фактически описывает синхронное изменение всех переменных и в первую очередь значения pH, концентрации магния и кальция. Четвертая координата, полученная на основе векторного пространства, вносит детали в описание концентрации магния, кальция и pH. Очевидно, отличия двух систем содержательны и заслуживают параллельного рассмотрения.

Шаг шестой. Оценку качества отображения переменных (свойств) в многомерном факторном пространстве осуществляем на основе многомерной регрессии. Опуская детали уже хорошо знакомого метода, укажем, что в каждой из систем отображения все переменные

Таблица 6.6

Оценка с помощью коэффициента парных корреляций подобия отображений координат точек в пространстве Евклида при различных методах анализа

Переменная	Способ расчета значений координат пространства Евклида															
	Факторный анализ с вращением				Факторный анализ				Прямое многомерное шкалирование по матрице дистанций между элементами				Расчет на основе коэффициентов векторного пространства			
	FR1	FR2	FR3	FR4	F1	F2	F3	F4	M1	M2	M3	M4	E1	E2	E3	E4
F1	0,75	0,49	0,25	0,38	1,00	0,00	–0,00	0,00	**–0,99**	–0,15	–0,10	0,39	–0,03	0,12	0,38	–0,23
F2	–0,15	0,63	–0,76	0,00	0,00	1,00	0,00	–0,00	–0,01	**0,94**	0,18	–0,09	**0,94**	–0,34	–0,08	–0,15
F3	–0,40	0,60	0,58	–0,38	–0,00	0,00	1,00	0,00	–0,00	–0,16	**0,97**	–0,06	–0,33	**–0,92**	–0,12	0,01
F4	–0,51	0,05	0,15	0,84	0,00	–0,00	0,00	1,00	–0,05	–0,15	–0,07	**–0,81**	–0,04	0,12	**–0,90**	–0,24
M1	–0,71	–0,49	–0,26	–0,41	**–0,99**	–0,01	–0,00	–0,05	1,00	0,15	0,10	–0,34	0,02	–0,12	–0,33	0,19
M2	–0,12	0,42	–0,86	–0,12	–0,15	**0,94**	–0,16	–0,15	0,15	1,00	0,05	–0,07	**0,95**	–0,21	0,00	0,08
M3	–0,45	0,64	0,39	–0,46	–0,10	0,18	**0,97**	–0,07	0,10	0,05	1,00	–0,07	–0,14	**–0,98**	–0,11	0,02
M4	0,75	0,06	0,01	–0,51	0,39	–0,09	–0,06	**–0,81**	–0,34	–0,07	–0,07	1,00	–0,06	0,04	**0,94**	–0,22
E1	–0,02	0,38	**–0,92**	0,08	–0,03	**0,94**	–0,33	–0,04	0,02	**0,95**	–0,14	–0,06	1,00	–0,03	–0,02	–0,08
E2	0,45	–0,71	–0,23	0,49	0,12	–0,34	**–0,92**	0,12	–0,12	–0,21	**–0,98**	0,04	–0,03	1,00	0,08	–0,00
E3	**0,81**	0,01	–0,05	–0,57	0,38	–0,08	–0,12	**–0,90**	–0,33	0	–0,11	**0,94**	–0,02	0,08	1,00	0,03
E4	–0,03	–0,21	0,03	–0,29	–0,23	–0,15	0,01	–0,24	0,19	0,08	0,02	–0,22	–0,08	–0,00	0,03	1,00

Примечание. Полужирным шрифтом выделены почти линейные отношения между координатами в различных системах расчета.

Рис. 6.6. Специфические отношения переменных к факторам в трех системах построения многомерного пространства

описываются координатами в среднем с коэффициентом детерминации 0,7 при минимуме 0,5 и максимуме 0,8, что можно считать вполне приемлемым. При этом следует иметь в виду, что 30 исходных переменных описываются всего четырьмя координатами, а полевые измерения никогда не бывают свободны от ошибок.

Шаг седьмой. При поиске физической интерпретации координат естественно исходить из того, что рассматриваемые значения переменных так или иначе отражают, в первую очередь, характер и масштабы миграции влаги, катионов и анионов. Естественно полагать, что миграция может определяться формой рельефа в микро-, мезо- и макромасштабе, вертикальным и горизонтальным варьированием механического состава и текстуры почвы, сформировавшихся в процессе почвообразования, особенностями функционирования растительного покрова.

В данном случае для интерпретации полученных координат мы располагаем только профилем высот, что позволяет рассчитывать на получение лишь самых общих оценок.

Рельеф может быть в простейшем случае представлен высотой, параметром крутизны склона, первой производной от высоты (с сохранением знака учитывается экспозиция, при абсолютном значении только крутизна) и профилем склона (вторая производная). Так как шаг опробования составляет 25 м, эти характеристики со-

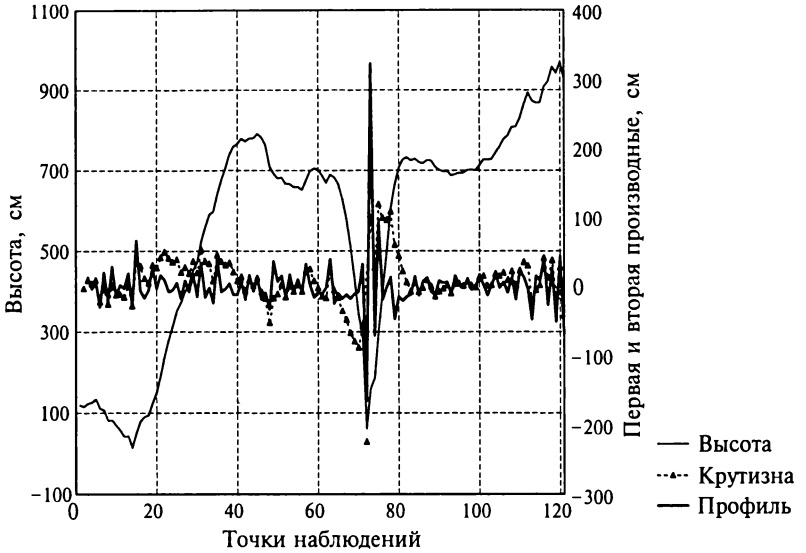


Рис. 6.7. Характеристики рельефа по трансеку

измеримы с микрорельефом (рис. 6.7). Южная экспозиция маркируется положительным значением крутизны, северная — отрицательным. Особо обратим внимание на то, что выпуклые формы маркируются отрицательным значением второй производной, а вогнутые — положительным. Такое соотношение ассоциируют обычно со знаком вектора движения влаги: минус — растекание, плюс — слияние.

Наиболее общую оценку интерпретируемости координат почвенной системы через рельеф можно получить на основе регрессионных моделей. Эти общие оценки показывают, что почвенные координаты описывают варьирование высоты рельефа почти на 40 % (табл. 6.7), причем по факторному анализу наиболее связан с рельефом первый фактор, а в модели векторного пространства — четвертый. Из рис. 6.8 следует, что почвенная система координат в обоих случаях хорошо воспроизводит общие закономерности изменения макрорельефа.

Система, полученная через векторное пространство (рис. 6.8, *б*), несколько лучше отображает рельеф, чем факторная (рис. 6.8, *а*). Однако в обоих отображениях пики рассчитанных значений высот чаще совпадают с микроповышениями в рельефе, а минимумы — с микропонижениями.

Интерпретацию смысла координат также осуществляем, используя метод множественной регрессии (табл. 6.8).

В общем все координаты статистически значимо связаны с переменными рельефа. Наибольшую связь имеет первая координата,

Таблица 6.7

Параметры регрессионных моделей «высота на профиле — система координат почвы»

Система координат	Переменная	BETA	Std. Err. of BETA	*b*	Std. Err. of *b*	t(116)	p–level
Факторный анализ $R^2 = 0{,}3921$	Константа			560,198	19,49782	28,73133	0,000000
	F1	–0,538906	0,072391	–145,752	19,57890	–7,44433	0,000000
	F3	–0,274755	0,072391	–74,310	19,57890	–3,79541	0,000236
	F2	0,135459	0,072391	36,636	19,57890	1,87121	0,063836
	F4	0,088547	0,072391	23,948	19,57890	1,22316	0,223747
Векторное пространство $R^2 = 0{,}4248$	Константа			576,814	27,82859	20,72738	0,000000
	E4	0,434272	0,074872	159,610	27,51794	5,80023	0,000000
	$E1^2$	–0,396678	0,093498	–172,346	40,62227	–4,24265	0,000045
	E1	0,311597	0,084879	145,705	39,69022	3,67105	0,000370
	E2	0,155918	0,077736	64,249	32,03267	2,00574	0,047274
	E3	–0,254010	0,090646	–103,653	36,98949	–2,80223	0,005973
	$E2^2$	0,124270	0,084357	53,521	36,33124	1,47314	0,143492
	$E3^2$	0,136805	0,112589	40,624	33,43338	1,21508	0,226867

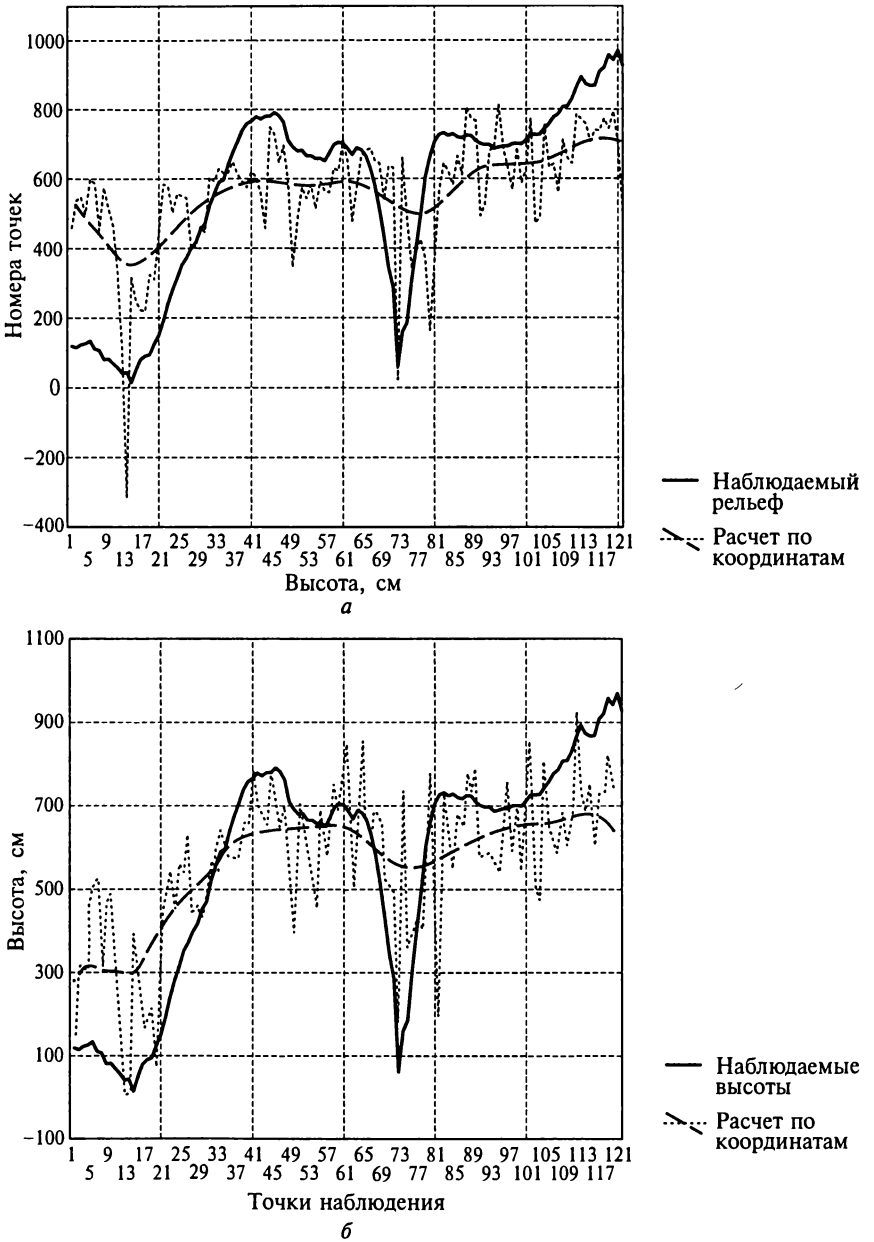


Рис. 6.8. Воспроизведение рельефа координатами почвенной системы на основе:

а — факторного анализа и многомерного шкалирования по матрице дистанций между точками наблюдений; *б* — расчета координат через векторное пространство

Таблица 6.8

Интерпретация координат через переменные рельефа

Координата почвенной системы	Содержание координаты	Коэффициент детерминации	Переменная рельефа	BETA	t-критерий	Комментарии
Координата 1 факторного анализа	Общее содержание обменных оснований и влаги	0,44309387	Высота	–0,518	–7,08137	Концентрации тем выше, чем ниже поверхность на крутых склонах вообще и на южных в особенности, при выпуклой форме рельефа
			Крутизна с экспозицией	0,218	2,63516	
			Крутизна	0,290	3,97876	
			Профиль микрорельефа	–0,210	–2,53765	
Координата 1 при расчете по векторному пространству	Определяет варьирование влажности и концентрации калия. Чем больше влажность, тем меньше концентрация калия. Максимум координаты—максимум концентрации калия	0,30644767	Высота	0,299	3,71052	На относительно крутых склонах и особенно южных экспозициях и больших высотах концентрация калия выше, а влажность — меньше
			Крутизна с экспозицией	0,411	5,27984	
			Крутизна	0,338	4,17695	
То же, координата 2	Минимум координаты—максимум концентрации фосфора. Чем больше концентрация фосфора, тем меньше концентрация магния и кальция и ниже pH	0,09939875	Высота	0,192	2,07203	Концентрация фосфора ниже на северных экспозициях, больших высотах, большой крутизне и на вогнутых элементах микрорельефа
			Крутизна с экспозицией	–0,303	–2,88088	
			Крутизна	0,113	1,22646	
			Профиль микрорельефа	0,189	1,80253	
То же, координата 3	Определяет pH в верхних и средних горизонтах—максимум значения координаты—максимум pH	0,18839326	Высота	–0,169	–1,94302	Чем круче склон и меньше высота, тем выше pH
			Крутизна	0,356	4,08947	
То же, координата 4	Определяет концентрацию калия в нижнем и pH в средних горизонтах и с отрицательным знаком — концентрацию магния	0,15464282	Высота	0,372011	4,16971	pH тем выше, чем больше высота и крутизна склона, особенно при южной экспозиции
			Крутизна с экспозицией	–0,142406	–1,65525	
			Крутизна	0,185922	2,07933	

полученная в факторном анализе, отражающая генеральные закономерности варьирования исследуемых свойств почвы. Поскольку высота влияет отрицательно, более высокие поверхности имеют более кислые и сухие почвы с меньшим содержанием обменных оснований.

Крутизна как без, так и с учетом экспозиции положительно связана с координатой, в соответствии с чем можно утверждать, что на южных крутых склонах более типичны менее кислые почвы, богатые обменными основаниями.

Отношения со второй производной — обратные, т.е. на выпуклых поверхностях с отрицательным значением производной концентрации обменных оснований выше, но, конечно, имея в виду слабое влияние этого фактора, это справедливо в полной мере для небольших высот и достаточно крутых склонов. По этой схеме трактуется содержание и остальных координат. Более детальное обсуждение соотношений, представленных в таблицах, выходит за рамки настоящего пособия. В данном случае важна демонстрация общей схемы, которая естественно тождественна использованной в факторном анализе.

Таким образом, многомерное непараметрическое шкалирование позволяет подойти к исследованию систем, описываемых большим числом переменных, при любых распределениях и любой форме представления данных. Вместе с тем при его применении существует большая неопределенность в выборе метрики, формы представления данных, метода самого многомерного шкалирования и перехода от векторного пространства, коэффициентов чувствительности к пространству координат.

Все это определяет необходимость исследования различных вариантов отображения исходной системы и максимально аргументированного выбора наилучшего или взаимодополняющих методов.

Для сложных нелинейных систем различные отображения могут нести свой собственный физический смысл. Используя эту технологию анализа, исследователь получает возможность рассматривать различные отображения многомерного явления, что, конечно, требует от него достаточно хороших базовых знаний об исследуемом явлении.

Специалист, идеально владеющий самим аппаратом статистического анализа, но не владеющий предметом, аккуратно используя различные подходы, может получить наиболее устойчивый результат, однако, не имея внешней системы оценки полученных отношений, он не сможет обсуждать и рассматривать его содержание. Поэтому только исследователь, имеющий базовые представления об изучаемом явлении, может наиболее эффективно осуществить анализ конкретных отношений, содержащихся в исходной системе данных.

Контрольные вопросы

1. Что такое метрика, каковы способы ее конструирования и выдвигаемые ею требования?

2. Попытайтесь самостоятельно сконструировать метрику, которая одновременно отражала бы два свойства объекта и докажите, что это действительно метрика.

3. Разберите логические основания процедуры многомерного непараметрического шкалирования.

4. С помощью программных средств постройте ряды n-независимых переменных для различных распределений, определите дистанцию для переменных каждого типа и на основе операций многомерного шкалирования постройте функцию «стресс — размерность».

Глава 7
ПРИМЕНЕНИЕ МНОГОМЕРНОГО ШКАЛИРОВАНИЯ ПРИ РЕШЕНИИ ЗАДАЧИ ЭКОЛОГИЧЕСКОЙ ОРДИНАЦИИ

7.1. Общие представления об экологических нишах, экологическом пространстве и взаиморазмещении видов

Традиционной задачей экологии является исследование правил размещения видов по отношениям к некоторым градиентам среды. Методы, позволяющие решать эту задачу, называются *ординацией*. Эти градиенты иногда рассматриваются как координаты экологического пространства. Следовательно, в общем случае задача ординации может быть переформулирована в терминах экологической ниши.

Экологическая ниша — абстрактное понятие, возникшее в экологии для отображения различий в отношениях видов к ресурсам, условиям среды и друг к другу. Согласно Г.Хатчинсону (G.Hutchinson, 1991), «ниша — область в многомерном пространстве всех потенциальных переменных, так или иначе определяющих существование каждого вида и их численность». Эта трактовка объединяет представления о нише как о местообитании (условия среды), о трофических отношениях (переменные, связанные с ресурсами), о конкуренции, об отношении к площади (островной аспект, подвижность) и ко времени (эфемеры, виды с саморегулируемой суточной и сезонной активностью).

В соответствии с этими представлениями под *экологическим пространством* можно понимать потенциально открытое многомерное множество отношений видов друг с другом и окружающей средой. Открытость подразумевает возможность эволюции экологического пространства. Эволюция может вести к увеличению размерности пространства, расширению области используемых условий и ресурсов, видообразованию и т.п.

Такая общая трактовка экологического пространства носит скорее философский, концептуальный, чем конструктивный характер. Поэтому в реальном исследовании, ориентированном на конкретную естественную, функционально подобную группу организмов, конкретную территорию и интервал времени, можно говорить о локальном или частном экологическом пространстве.

Конечно, для экологического пространства в целом можно конструировать подпространства с различными характерными террито-

риально-временны́ми иерархически соподчиненными размерами. Эта естественная, но в существенной степени умозрительная процедура, популярная в современной экологии, может быть распространена и на экологическое пространство, в частности иерархическая структура пространства-времени может быть предметом специального исследования. Однако при исследовании реальных отношений речь может идти только о частном экологическом пространстве.

Таким образом, при решении конкретных задач экологии будем рассматривать только частное экологическое пространство, специально определяемое для каждого случая порождающей системой.

Если общее экологическое пространство может иметь потенциально бесконечную размерность, то любое частное экологическое пространство по условию замкнуто и конечномерно. Однако число мыслимых переменных, потенциально определяющих численность каждого вида рассматриваемой группы столь велико, что задача определения отношения к ним разнообразных условий среды очень громоздка. С другой стороны, многие мыслимые и реальные переменные трудно измеримы или вообще неизмеримы, например измерение термического режима с учетом многообразия его изменчивости — весьма трудная задача.

Единственно, что надежно может измерить эколог, это факт наличия особей каждого вида на конкретной пробной площади или в общем случае в элементе учета и менее надежно — показатели их обилия (число особей, их массу, сумму площадей сечений для деревьев, проективное покрытие, обилие и т. п.).

Исходя из представлений об экологической нише как о многомерном пространстве, можно считать, что обилие каждого вида в каждой пробе есть функция неизвестных переменных — координат пространства, межвидовых отношений и некоторого случайного процесса.

Это утверждение формально справедливо только для случая равновесных отношений. Если вид находится в неравновесном, нестационарном состоянии, то его обилие может не содержать информации о состоянии переменных в конкретной точке. Эта вполне реальная, но в силу относительной быстротечности, весьма редкая ситуация, которая часто может быть выявлена как не соответствующая гипотезе равновесия. Вместе с тем нельзя исключить случаи, когда нестационарные отношения охватывают всю исследуемую группу видов на обширной территории. И это всегда необходимо иметь в виду при анализе результатов измерения.

В равновесном случае измерение обилия видов на пробах будет отображать изменение состояний неизвестных внешних и внутренних переменных.

Если тем или иным способом измерить дистанции между видами по множеству точек наблюдения, то можно определить некоторое минимальное число независимых, базовых переменных, меру

влияния каждой из них на каждый вид, положение каждой пробы (точки наблюдения) в ортогональной системе соответствующих координат.

Таким образом, если известны расстояния между видами в некоторой системе неизвестной размерности, то для пространства заданной размерности можно определить положение каждого вида в соответствующей системе координат. Полученное в результате отображение положения видов определяет их чувствительность к каждой координате векторного многомерного пространства. На основе векторов коэффициентов чувствительности видов и значений их обилия в каждой пробе можно, как это было сделано выше при исследовании почвенной системы, обозначить положение каждой пробы в Евклидовом пространстве, сопряженном с векторным. Множество значений обилия каждого вида будет описываться как функция от этих координат и, соответственно, отображать их положение в многомерном пространстве. Таким образом, получаем отображение видовых экологических ниш в многомерном пространстве абстрактных координат.

Эта общая модель может реализоваться различными методами. Логико-математический аппарат метода должен быть адекватен предполагаемым отношениям в исследуемой системе, принятому способу измерения дистанций и размерности пространства и не искажать исследуемую реальность.

Итак, для решения задачи необходимо обосновать выбор метода и метрики и найти способы определения реалистичного числа осей (координат) экологического пространства.

При исследовании положения видов в экологическом пространстве полезно использовать общие представления об отношении видов к градиентам среды и модели их равновесного размещения в сообществе. Эти представления и модели можно трактовать как проверяемые гипотезы.

Практически любая функция, отображающая реакцию измеряемой биологической переменной на действие внешнего фактора, в большинстве случаев имеет мультипликативную, логистическую или подобную им нелинейную форму. Такую же форму имеет широко известный феноменологический закон ощущения Вебера в психологии.

В общем случае можно записать

$$y = a\,(x - c)^b,$$

где y — измеряемая биологическая или экологическая переменная ($y \geq 0$); x — измеряемая внешняя переменная; c — порог, при превышении которого начинается реакция биологической системы на действие внешнего фактора ($c < x$); b — коэффициент чувствительности функции к действию внешней переменной ($b < 1$).

После логарифмирования получаем:

$$\log y = \log a + b\log(x - c),$$

a — константа, определяющая соизмеримость y и x.

Если форма связи y с внешней переменной отрицательна, то имеем

$$y = a(d - x)^b, \text{ где } d > x \ (d = \text{const}).$$

В случае действия переменной на биологическую функцию и положительно, и отрицательно

$$y = a(x - c)^{b_1}(d - x)^{b_2}.$$

В результате получаем типичную экологическую параболическую зависимость, например численности видов от некоторого фактора среды (рис. 7.1). Такую форму действия можно определить как двухканальную, мультипликативную. Она подразумевает, что один и тот же фактор одновременно действует на две различные функциональные части биологической системы, определяющие состояние целого. Один из типичных примеров — влияние содержания влаги в почве на фотосинтез: по мере увеличения запасов влаги обеспечивается устойчивая транспирация и регулирование температуры поверхности листа, но ухудшаются условия дыхания корневых систем. Этот тип отношений может быть выявлен в действии почти любого мыслимого внешнего фактора. Вообще можно полагать, что нет ни одного значимого фактора, воздействующего на биологическую систему, который не имел бы такой двойственной природы. Иначе говоря, «мало — плохо и много — плохо», опти-

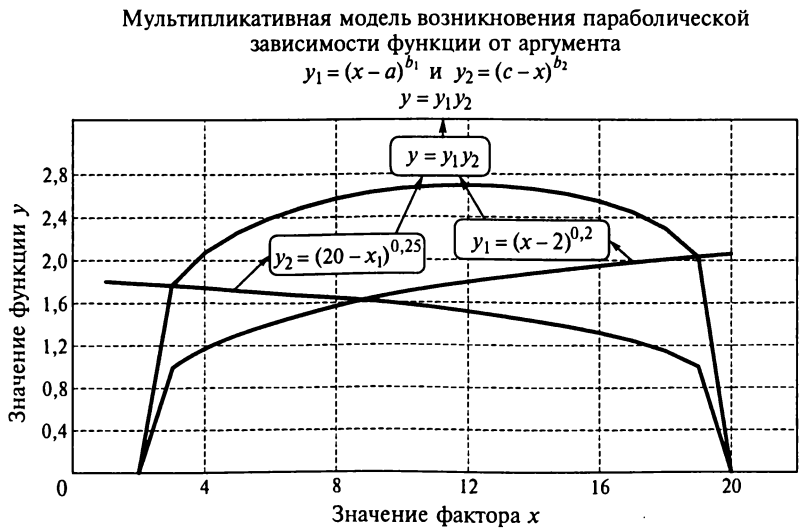


Рис. 7.1. Мультипликативная модель одномерной экологической ниши

мальный вариант находится в области средних значений любого фактора, но специфических для каждого вида.

Если допустить, что существует некоторый общий коэффициент чувствительности k целостной системы к функционированию двух ее частей, то

$$y = a[(x - c)^{b_1}(d - x)^{b_2}]^k.$$

При $k > 1$ вид функции имеет форму нормального или логнормального распределения; при $k < 1$ парабола становится более плоской (рис. 7.2).

Совместное действие двух и большего числа факторов может быть чисто мультипликативное (в логарифмической форме — аддитивно) (рис. 7.3):

$$y = a\,[(x_1 - c_1)^{b_1}(d_1 - x_1)^{b_2}]^{k_1}[(x_2 - c_2)^{b_1}(d_2 - x_2)^{b_2}]^{k_2}.$$

Если действует соотношение двух факторов, например тепла и влаги, то

$$y = a\,[(x_1 - c_1)^{b_1}(d_1 - x_1)^{b_2}]^{k_1}[(x_2 - c_2)^{b_1}(d_2 - x_2)^{b_2}]^{k_2} \times$$
$$\times[(x_1x_2 - c_3)^{b_3}(d_3 - x_1x_2)^{b_4}]^{k_3}.$$

Таким образом очевидно, что мультипликативный феноменологический закон функции и факторов позволяет отобразить все возможные формы реальных отношений, при этом коэффициенты c и d, определяя область существования функции, отражают

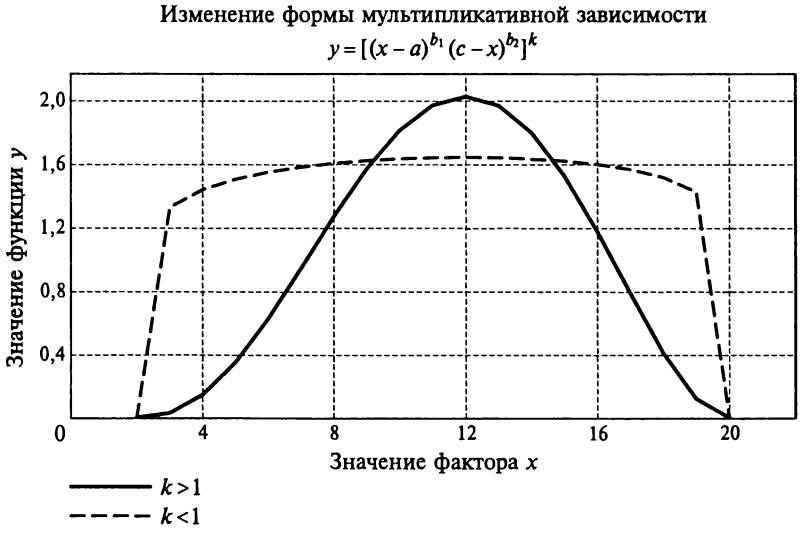
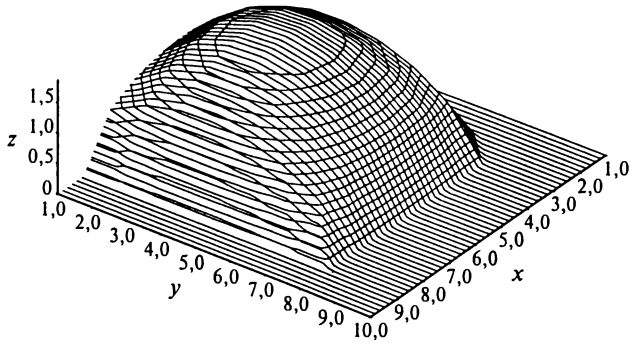
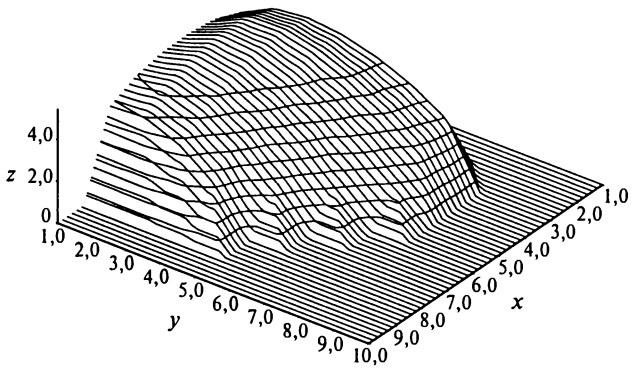


Рис. 7.2. Трансформация формы мультипликативной ниши общим параметром чувствительности



а



б

Рис. 7.3. Двухмерная экологическая ниша:
а — $z = 0{,}1[(x-2)^{0{,}2}(9-x)^{0{,}25}y^{0{,}3}((8-y)^{0{,}35})^2]$;
б — $z = 0{,}1[(x-2)^{0{,}2}(9-x)^{0{,}25}y^{0{,}3}(8-y)^{0{,}35}(xy-1)^{0{,}1}(40-xy)^{0{,}7}]$

действие по логике закона минимума Либиха, а сама мультипликативная форма — закон компенсаторных отношений.

Общетеоретическим основанием реализуемости этих отношений может быть закон пропускной способности канала связи в теории информации, доказанный К. Шенноном:

$$C = \omega \log[1 + P/(\omega N_0)],$$

где C — скорость передачи информации; ω — полоса частот, в которой проходит сигнал; P — мощность сигнала; N_0 — случайный шум на единицу полосы частот.

Поскольку пропускная способность канала на единицу полосы частот $\omega_0 = P/N_0$ — полоса частот, в которой мощность сигнала равна мощности шума, имеем

$$C/\omega_0 = (\omega/\omega_0)/\log(1 + \omega_0/\omega).$$

Следовательно, при увеличении полосы частот пропускная способность быстро растет до тех пор, пока мощность шума не сравняется с мощностью сигнала, после чего медленно стремится к пределу, равному $\log_2 e = 1{,}443$ бит.

Частоту можно определить по формуле

$$\omega = 1/T,$$

где T — время передачи информации.

Для биологических систем T может быть ассоциировано с характерным временем жизни или средним временем вступления в размножение, связанными с размерами особей по аллометрическому закону. Следовательно, для организмов одного размера максимальная полоса частот $\omega \leq 1$.

Полоса частот для биологических систем может быть прямо ассоциирована с понятием специализации: чем меньше специализация, тем больше полоса частот.

В результате получаем естественные общебиологические следствия:

1) возможная деспециализация (универсальность) имеет предел;

2) с ростом деспециализации увеличивается потенциальная продуктивность (r-стратегия);

3) связь функции с условиями среды есть мультипликативная зависимость вида

$$y = 2^C = [1 + P/(\omega N_0)]^{\omega} \approx ax^{\omega};$$

4) допуская среднее равенство экологических ниш $C_i \approx \text{const}$, соответственно $\omega_1 \approx \omega_2 \approx \omega_3 \approx \ldots \approx \omega_k$, и общую полосу пропускания всей системы $\Omega = \sum \omega_i = k\omega$ (k — общее число видов), имеем

$$\text{const} = \omega_i \log[1 + P/(N_0\omega_i)] \approx \omega_i \log[P/(N_0\omega_i)] =$$

$$= \omega_i \log(P/N_0) - \omega_i \log \omega_i.$$

Поскольку $\text{const} = \Omega/k[\log(P/N_0)] - \Omega/k[\log \Omega] - \Omega/k[\log k]$ и число видов k определяется как функция мощности сигнала и общей полосы частот

$$k - \Omega \log k \approx (\Omega / \text{const}) \log(P/N_0),$$

окончательно в упрощенном виде получаем одну из типичных зависимостей числа видов от условий среды, площади или времени наблюдений

$$k = a + b \log Y.$$

Степенная зависимость числа видов от благоприятности среды

$$k = aY^b$$

получается, если $\text{const} \approx (P/N_0)^{\Omega/k}$.

Таким образом, из теории информации прямо выводится стандартная зависимость числа видов от площади или времени наблюдений и в общем случае — от благоприятности среды.

Естественность следствий, вытекающих из информационного закона пропускной способности канала связи, позволяет допускать его применимость для биологических систем. В целом вывод этого закона основывается на обобщении поведения статистических ансамблей. Так как для большинства случайных процессов масштаб случайных флуктуаций сигнала тем выше, чем больше его среднее значение, с увеличением мощности сигнала растет ошибка его декодирования и приращение пропускной способности к росту мощности сигнала стремится к нулю, а сама пропускная способность — к некоторому пределу. Для снижения ошибки требуется бóльшая специализация, что само по себе приводит к уменьшению пропускной способности.

Таким образом, феномен мультипликативной связи биологических и экологических систем с условиями среды, скорее всего, есть следствие общих законов поведения статистических ансамблей. Опираясь на эти основания, можно сформулировать гипотезу о некоторых фундаментальных свойствах экологического пространства.

Так, приведенное обоснование отношений биологических функций к условиям среды позволяет ассоциировать экологическое пространство с конструкцией пространства Хаусдорфа, образуемого непересекающимися открытыми множествами. Открытость множеств подразумевает существование подмножеств границы, а непересекаемость — отсутствие перекрытий ниш в условиях равновесия.

Естественной метрикой в этом случае является h-мера Хаусдорфа. Она вводится как минимальное число всех возможных пересечений окрестностей сравниваемых плотных множеств (множество называется плотным, если оно не содержит изолированных точек). Причем сами множества могут как пересекаться, так и не пересекаться. В соответствии с этим вводится непрерывная, возрастающая функция $h(t)$, заданная на положительной полуоси t такая, что $h(0) = 0$:

$$h(t) = \gamma(d)t^d,$$

где d — фиксированное положительное, не обязательно целое, число; $\gamma(d)$ — положительная константа, зависящая только от d.

В частности,

$$\gamma(d) = \frac{[\Gamma(1/2)]^d}{[\Gamma(1+d/2)]2^d},$$

если множества мыслятся как шары единичного радиуса; $\Gamma(x)$ — гамма-функция.

Таким образом, в этом достаточно общем случае через соотношение гамма-функций вводится возможное число комбинаций, по которым рассматривается перекрытие окрестностей сравниваемых множеств.

Параметр d является Хаусдорфовой или фрактальной размерностью пространства. В отличие от топологической размерности фрактальная размерность может быть нецелочисленной. Нецелочисленность размерности отражает одновременную непрерывность и разрывность множества. Непрерывность подразумевает, что для любой точки множества в окрестности любого радиуса найдется точка, принадлежащая этому множеству, а, с другой стороны, найдется и пустое множество, не содержащее его точек. В качестве примера типичной фрактальной структуры часто приводят кусочек пепла от бумаги, который при увеличении представляется как комплект вложенных друг в друга сит с отверстиями различных размеров. Если бы кусочек пепла был строго фрактален, то он имел бы нулевую массу, так как состоял бы из сколь угодно больших и бесконечно малых отверстий. И хотя масса пепла действительно непропорционально мала относительно объема, но все-таки она отлична от нуля. Соответственно, фрактальная модель применима к природным явлениям лишь до некоторых пределов.

Размещение особей любого вида в пространстве, даже с учетом их пространственных взаимодействий, хорошо ассоциируется с представлениями о фрактальном множестве. Такое распределение даже в однородном пространстве должно состоять из сгущений разной плотности и почти свободных дырок и обычно описываться гамма-распределением. Эта разрывная структура сама по себе допускает взаимососуществование видов с относительно близкими требованиями к условиям среды, которое наблюдатель может воспринимать как сообщества. Система совокупности видов также будет фрактальна, а коль скоро это так, то она, с одной стороны, является непрерывной, а с другой — в ней всегда можно обнаружить «границы» как области разрыва непрерывности разного масштаба.

Однако фрактальность в полной мере не может объяснить правила размещения видов в сообществе. Известный экологам закон Гауза, полученный в 30-е годы XX в. при изучении взаимного размещения видов дрожжей в культурах, развивающихся в пробирке, утверждает, что в одной экологической нише не может сосуществовать несколько видов. Этот закон, названный конкурентным исключением, был подтвержден рядом экспериментов с близкими видами мучных червей и другими. Существует математическая модель, показывающая, что сосуществование двух видов в одной нише имеет очень узкую область устойчивости. С другой стороны, любое сообщество сложено многими видами, например растений, которые в конечном итоге используют весьма сходные ресурсы.

В 70-е годы XX в. в связи с программами жизнеобеспечения космических полетов в Институте биофизики Сибирского отделения РАН проводили обширные экспериментальные исследования, целью которых было создание длительно устойчивых сообществ в микрокосме при постоянном подводе некоторых ресурсов или вообще в замкнутой системе. Экспериментальные исследования на культурах различных видов грибов и микроорганизмов велись совместно с работами молодых и талантливых математиков. В аннотированным списке рекомендуемой литературы будут даны соответствующие ссылки, здесь же приведем один из фундаментальных полученных ими законов равновесного сосуществования видов в гомогенном пространстве, получившем полное подтверждение в экспериментах.

Допустим, что в систему равномерно поступает n взаимозаменяемых пищевых субстратов x_1, x_2, x_3, ..., x_n, которые используются k видами. «Взаимозаменяемость» подразумевает, что каждый вид в принципе способен развиваться на любом ресурсе. Субстраты в пространстве микрокосма полностью перемешаны, поступают равномерно и потому независимы. Однако эффективность их использования видоспецифична и численность или биомасса вида i определяется как

$$n_{ij} = \beta_{i1}x_1 + \beta_{i2}x_1 + \beta_{i2}x_2 + \ldots + \beta_{in}x_n,$$

где β_{ij} — коэффициент чувствительности вида i к субстрату j ($\beta_{ij} \geq 0$), чем более он отличен от нуля, тем более эффективно используется субстрат (на единицу субстрата производится бóльшая биомасса).

Фактически это уравнение определяет экологическую нишу вида i при линейной форме зависимости биомассы от концентрации ресурса.

Тогда сообщество можно выразить через систему ниш-уравнений:

$$\begin{cases} n_{1j} = \beta_{11}x_1 + \beta_{12}x_1 + \beta_{12}x_2 + \ldots + \beta_{1n}x_n \\ n_{2j} = \beta_{21}x_1 + \beta_{22}x_1 + \beta_{22}x_2 + \ldots + \beta_{2n}x_n \\ n_{3j} = \beta_{31}x_1 + \beta_{32}x_1 + \beta_{32}x_2 + \ldots + \beta_{3n}x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ n_{kj} = \beta_{k1}x_1 + \beta_{k2}x_1 + \beta_{k2}x_2 + \ldots + \beta_{kn}x_n. \end{cases}$$

Таким образом, получаем уже хорошо знакомую запись системы уравнений в векторном пространстве.

Очевидно, что, если скалярные произведения векторов равны 0, т.е. многомерный угол между плоскостями-нишами равен 90°, биомассы видов не зависят друг от друга, а пересечение ниш отсутствует. При многомерной геометрической трактовке ниши полностью тождественны, если косинус угла между ними (скалярное произведение векторов) равен 1.

Было показано, что в такой системе может длительно и устойчиво сосуществовать число видов, равное числу ресурсов. При этом необходимо выполнение одного из трех условий:

$$1)\ \beta_{kk} > (N-1)\sum_{i=1}^{N}\beta_{ik},\ i \neq k, \qquad k = 1, 2, ..., N,$$

где β_{kk} — коэффициент чувствительности к видоспецифичному, предпочитаемому субстрату он в $(N-1)$ раз больше суммы коэффициентов чувствительности β_{ik} ко всем другим субстратам (жесткое условие);

$$2)\ \beta_{jj} \geq \beta_j + \beta_{1j} \geq ... \geq \beta_{nj} \geq \beta_{1j} \geq \beta_{2j} \geq ... \geq \beta_{j-1j},\ j = 1, 2, ..., N.$$

Это условие допускает равномерное потребление всех ресурсов. Однако требуется монотонная упорядоченность предпочтительности ресурсов, приводящая к различию у конкурентов как наиболее, так и наименее предпочтительных ресурсов;

$$3)\ \beta_{kk} > \max_{i \neq k} b_{ik},\ k = 1, 2, ..., N.$$

Это более слабое условие означает, что различаются только самые предпочтительные ресурсы.

Таков самый общий и вместе с тем фундаментальный результат, полученный сибирскими биофизиками.

Из общей модели выведен и экспериментально подтвержден интересный эффект «хищника». Если в систему добавить хищника, то число устойчиво сосуществующих жертв увеличивается на один вид, т. е. «хищник» становится как бы новым субстратом, к которому можно приспособиться с несколько различной эффективностью.

С другой стороны, если в систему поставлять различные ресурсы неравномерно во времени, то число совместно сосуществующих видов также увеличивается.

Примерно в то же время автор настоящего пособия исследовал связь развития компонентов лесной растительности для лесной зоны с климатом и рельефом. Из очень большого числа описаний растительности, выполненных в разных регионах нашей страны, были отобраны данные, характеризующие развитие лесных ярусов, доминирующих видов в древесном и кустарниковом ярусах и жизненных форм в травяно-кустарниковом и мохово-лишайниковом ярусах, а также сопутствующие характеристики рельефа и климата (сумма температур больше 10 °С, продолжительность периода с температурами выше 10 °С, средние температуры января, сумма летних и зимних осадков, относительная влажность воздуха в час дня). Весь этот массив данных был обработан описанным выше методом кросс-табуляции (таблицы сопряженности) с расчетом информационных индексов сопряженности. Так как переменные, характеризующие растительность, можно рассматривать как функцию климатических переменных и

рельефа, следуя простейшей логике, естественно было ожидать, что между переменными растительности в среднем сопряженность будет выше, чем между переменными климата и рельефа.

В действительности все оказалось наоборот: сопряженность между характеристиками растительности была много меньше, чем между климатическими переменными. Систематизация этих мер сопряженности показала, что развитие двух ярусов тем меньше сопряжено, чем больше различается структура их связи с климатическими факторами. Таким образом, получилось, что компоненты сообщества, имея вектора чувствительности к внешним переменным, косинус угла между направлениями которых близок к нулю, делают среду, образованную существенно зависимыми факторами, ортогональной.

В конечном итоге это означает, что компоненты экологической системы за счет различного приспособления к ресурсам и условиям среды максимизируют независимость друг от друга. Соответственно, они могут встречаться во времени и в пространстве в широком диапазоне соотношений друг с другом, точно также как ель сочетается с березой в рассмотренном выше примере, связанном с лесами Центрально-лесного заповедника. Максимизация независимости практически тождественна максимизации устойчивости системы во времени и в пространстве. Если допустить, что некоторая реальная система независима от множества условий, то она никак не реагирует на их флуктуации и в этом смысле максимально устойчива (инвариантна).

Таким образом, в общем случае модель сообщества может строиться не только на основе разнообразия пищевых ресурсов (субстратов), но и на разнообразии условий среды. Здесь полезно прокомментировать содержательное различие между ресурсами и условиями. Оно является полезным при исследовании правил организации сообществ, но не является общепринятым. Под *ресурсом* полезно понимать все вещественно-энергетические отношения между организмами и средой, а под *условиями среды* — все свойства, которые прямо не участвуют в вещественно-энергетических отношениях, но изменяют величину их эффективности. С этих позиций температура среды, влажность воздуха являются условиями, а не ресурсом. Условиями среды являются также ее мозаичность, сочетание функционально по-разному используемых местообитаний для животных и т.п. Конечно, такое разделение не абсолютно. Например, влага в почве для растений является и ресурсом, и условием. Как вода она участвует в вещественно-энергетических обменах, но как влажность среды определяет интенсивность дыхания корней растений. Такая двойственность отношений определяет и двойственное влияние почвенной влаги, например, на процессы фотосинтеза и биологической продуктивности.

Кроме ресурсов и условий среды в качестве особых факторов необходимо рассматривать время и пространство, к которым виды могут обладать также различной чувствительностью. Феномен влияния *пространства* можно увидеть в существовании видов растений с различными линейными размерами. Если линейные размеры пространства, необходимые для существования особей двух разных видов, различаются более чем в два раза, то эти виды по условию не зависят друг от друга и их ниши ни при каких других условиях не пересекаются. Вид, требующий большего пространства, является некоторым фоном для второго, но сильные взаимодействия между ними невозможны.

Точно такие же отношения определяются и *временем*: если собственный жизненный цикл особей одного вида в два раза больше жизненного цикла второго, то они заведомо независимы. Эти соотношения являются прямым следствием теории колебаний и проявляются в природе в частности в существовании жизненных форм растений с разными линейными размерами и различной продолжительностью жизни особи. При этом абсолютной связи между линейными размерами и продолжительностью жизни индивидуума не существует. Вместе с тем имеются жизненные формы, мало зависящие от пространства (типичный пример — лианы) и от времени (деревья, продолжительности жизни которых есть во многом функция скорости роста, а не собственно циркадного времени). Эти дополнительные условия необходимо учитывать при исследовании правил организации сообществ. Два вида, принадлежащих к различным «размерным» жизненным формам, могут иметь практически тождественные отношения к условиям и ресурсам среды и устойчиво сосуществовать друг с другом.

Все эти соображения позволяют обобщить модель сообщества для системы факторов гетерогенной среды, включающих и ресурсы, и условия, и градиенты их изменения в пространстве-времени. Включение гетерогенности подразумевает, что коэффициенты чувствительности видов определяются не только в положительной, но и в отрицательной области, т.е. могут различаться не только по величине, но и по знаку. В такой системе виды с разными знаками по коэффициенту чувствительности взаимодополняют друг друга по градиенту среды и в области контакта могут конкурировать. Очевидно также, что в модель необходимо добавить кроме прямого и мультипликативное влияние факторов (например, соотношения тепла и влаги). Реакции на произведение значений двух факторов создают дополнительные возможности для максимизации независимости, так как произведение заведомо лишь частично связано с каждым исходным фактором и их суммой. Вполне понятно, что нелинейная зависимость вида от любого условия среды входит в систему уравнений квадратичным членом и создает дополнительные условия для максимизации независимости.

Таким образом, можно записать модель равновесного сообщества в виде

$$\cup n_{ij} = \sum_{j=1}^{n} \beta_{ij} x_j + \sum_{j=1}^{n} \beta_{ijk} x_i x_k \quad \text{— объединение ниш } i \text{ видов.}$$

Продемонстрируем независимое размещение трех видов в двухмерном пространстве:

1) первый вид зависит только от фактора 1;

2) второй вид зависит только от фактора 2;

3) третий вид по мультипликативной форме зависит от факторов 1 и 2.

Отображение ниш приведено на рис. 7.4 (*а* — *в*).

Корреляции между «численностями видов» образуют матрицу:

	Второй вид	Третий вид
Первый вид	0,00	0,3
Второй вид		0,3

Первые два вида по условию независимы, а третий коррелирует с каждым лишь с коэффициентом корреляции 0,3, т.е. практически независим от них (совместно два первых вида описывают лишь 17 % его варьирования). Если провести факторный анализ и осуществить вращение, то получаем, что заданное двухкоординатное пространство трансформируется такими нелинейными отношениями видов к факторам в трехмерное (рис. 7.4, *г*, табл. 7.1).

Итак очевидно, что в двухмерном гомогенном пространстве, используя мультипликативные отношения, можно разместить три вида. Чтобы получить гетерогенное пространство, нужно просто добавить к нему симметричное пространство, в котором максимум численности первого вида будет соответствовать минимальному, а не максимальному значению первой координаты, максимум второй вида — минимуму второй координаты, а большая ось эллипса ниши третьего вида будет перпендикулярной к описанному мультипликативному виду. Таким образом, в гетерогенном двухмерном пространстве может сосуществовать 6 видов.

Имея в виду эту комбинаторику, можно рассчитать оптимальное число видов в пространстве любой размерности. Для гомогенного трехмерного пространства число устойчиво сосуществующих видов будет соответственно: три основных, каждый из которых связан с одной координатой, и еще три, каждый из которых мультипликативно связан с комбинацией каждой пары координат. С формальных позиций может существовать еще два вида, которые мультипликативно связаны с тремя координатами, однако реальное восприятие одновременного действия трех факторов по ряду соображений довольно проблематично.

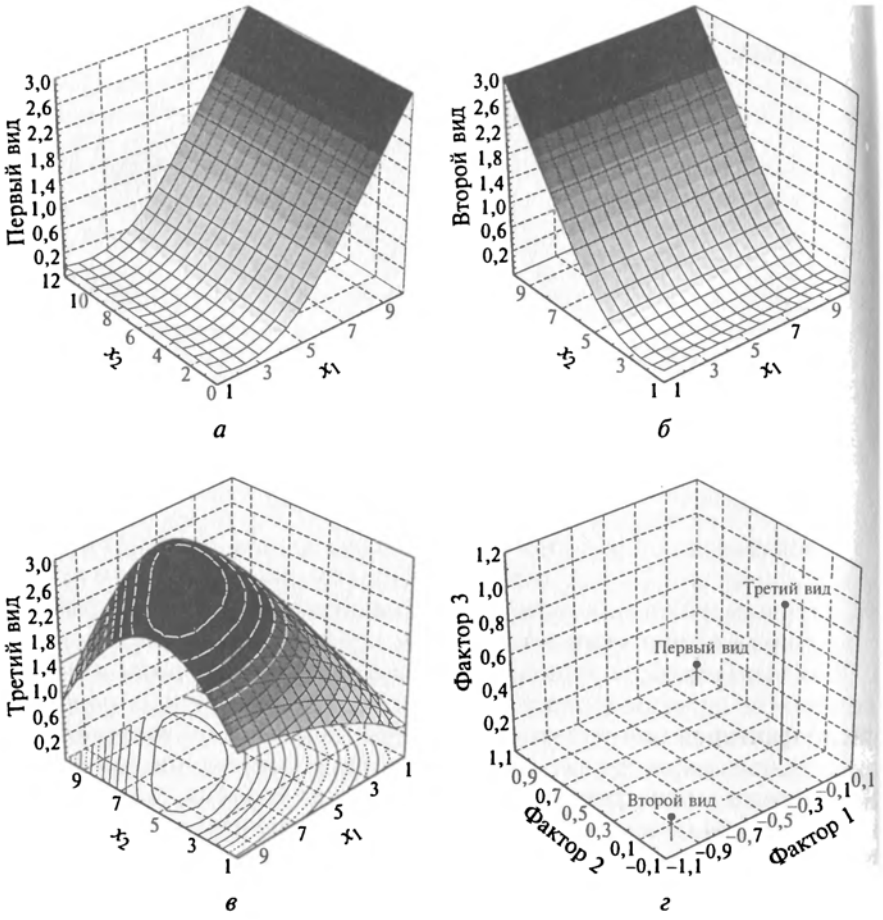


Рис. 7.4. Взаиморасположение независимых ниш в двухмерном пространстве с градиентом значения факторов:

а — вид нелинейно зависит от фактора 1; *б* — вид зависит от фактора 2; *в* — вид зависит от мультипликативного действия; *г* — положение ниши по коэффициентам чувствительности в трехфакторном пространстве

Если пространство четырехмерное, то возможное число видов в гомогенной области уже равно 10.

Можно оценить насколько эти соотношения реальны. Так, для гомогенного растительного сообщества, размерность пространства может быть определена близкой к пяти: свет, минеральное питание, влага, пространство и время. Соответственно в пределе в таком пространстве для гомогенных условий может устойчиво сосуществовать $5 + (4 \cdot 5)/2 = 15$ видов. Фактически в таком гомогенном пространстве каждому виду соответствует своя координата формального факторного пространства.

Таблица 7.1

Факторное пространство ниш трех видов, заданных в двухмерном пространстве

Переменная	Номер фактора		
	1	2	3
Первый вид	0,005750	**0,989492**	0,144472
Второй вид	**–0,988587**	–0,005810	0,150536
Третий вид	–0,156158	0,150282	**0,976233**
Нагрузка	1,001724	1,001713	0,996563

Примечание. Полужирным шрифтом выделены ведущие факторы.

В локальном гетерогенном пространстве возможное число сосуществующих видов существенно больше (говоря о локальном пространстве имеем в виду относительно небольшой градиент среды, в пределах которого взаимодополняют друг друга два вида).

Число взаимодополняющих видов с противоположным положением хотя бы по одной координате точно равно 2^n, где n — число факторов при мультипликативном действии и число ниш в гомогенном пространстве. Соответственно в двухмерном локально гетерогенном пространстве может быть $2^3 = 8$ видов, в трехмерном — $2^6 = 64$, в четырехмерном — $2^{10} = 1024$. Очевидно, что рост числа видов с увеличением размерности пространства происходит очень быстро и в первую очередь за счет мультипликативного увеличения неаддитивных отношений. Конечно, рассмотренная модель демонстрирует предельные возможности зависимости числа экологических ниш от размерности экологического пространства, для реализации которых требуется мощный поток ресурсов и значительное эволюционное время. Однако модель равновесного сообщества показывает, с одной стороны, широкие возможности в упаковке видов в экологическом пространстве, а с другой — полную невозможность использования для ординации сообществ классического факторного анализа и метрик, в основе которых лежит корреляция Пирсона.

Конечно, к любой модели нужно относится осторожно, однако она может служить гипотезой, конкретизирующей задачу исследования. Применительно к сообществу эта задача в соответствии с идеологией рассмотренной модели может быть определена как «исследование правил размещения видов в экологическом пространстве и поиск инвариантов или общих правил» такого размещения.

7.2. Анализ экологического пространства методом многомерного шкалирования

Методам ординации видов в сообществе посвящено большое количество литературы. Так, например, М. Пальмер (2000) сформулировал свойства решения (под решением понимается нахождение оптимального типа взаиморазмещения), которые заключаются в следующем:

1) метод восстанавливает градиент по каждому внешнему фактору без искажений;

2) если в пространстве есть группировки видов, то они отражаются в результате ординации;

3) в результате анализа не появляются «ложные», несуществующие группировки;

4) для одной и той же выборки данных при повторном анализе результат полностью воспроизводим;

5) существует единственно возможное решение;

6) экологическое подобие отражает сходство в размещении видов в пространстве;

7) масштабирование осей (факторов) отражает β-разнообразие (чем больше β-разнообразие, тем больше амплитуда значений осей);

8) метод не чувствителен к «шуму»;

9) «сигнал» и «шум» хорошо разделяются;

10) количество осей не задается априори, а определяется, исходя из реальных данных;

11) решение не зависит от физической размерности переменных, образующих массив данных (не зависит от способа измерения обилия или оценки состояния видов);

12) независимо от способа организации наблюдений (пробная площадь, описание в точке) метод дает одинаковые результаты;

13) поиск решения не занимает много компьютерного времени;

14) метод дает хорошие результаты для коротких и длинных рядов, при низком и высоком уровне «шума», при высокой и низкой плотности, для большого и маленького объема выборки (числа описаний, пробных площадей), для богатых и бедных по числу видов сообществ;

15) математически изящный метод;

16) доступный, недорогой и легко понимаемый метод.

К этим требованиям необходимо добавить по крайне мере два весьма существенных:

а) результат не зависит от формы распределения (распределение не обязательно нормально);

б) результат не зависит от того, линейны или нелинейны отношения видов друг с другом и с реальными, априори неизвестными, факторами среды.

Читатель может более детально ознакомиться с проблемами ординации на сайтах Интернета. Для этого достаточно записать в окне «поиск» любой поисковой системы «ordination + ecology». Здесь же обсудим возможности решения задачи ординации на основе многомерного непараметрического шкалирования.

Объектом анализа будут сообщества древесного яруса смешанных лесов Центрально-лесного биосферного заповедника. Обилие каждого вида деревьев оценивалось в виде суммы площадей сечений на гектар (BSA). Сумма площадей сечений измерялась с помощью релаксометра (вилки) Битерлиха. Измерения проводились на трансектах с шагом опробования 10 и 25 м. Анализ данных выполнен студентом кафедры физической географии и ландшафтоведения географического факультета МГУ Т. В. Орловым.

К геоботаническим материалам такого рода можно применять два типа дистанций: дистанцию Минковского с различными значениями (p, q) и дистанцию на основе коэффициента корреляции гамма или на основе скалярного произведения векторов. Только эти метрики адекватно отражают отсутствие какого-либо вида в точках наблюдения. Однако метрика Минковского резко повышает значение господствующих видов, в результате чего размещение редких видов практически выпадает из анализа. Метрика на основе скалярного произведения также весьма чувствительна к обилию и не позволяет сопоставимо отразить место редкостей в экологическом пространстве. Только гамма-корреляция, исключающая из расчета совпадающие ранги обилия, позволяет сопоставить размещение редких и многочисленных видов. При этом расчет может быть выполнен для видов, встреченных даже всего лишь в двух точках. Хотя суждения по двум находкам весьма рискованны, априорное исключение таких редких объектов из анализа нецелесообразно. Достаточно помнить, что в этом конкретном случае мы имеем дело с редкими событиями. С формальных позиций гамма-корреляция вида с двумя находками по отношению, по крайней мере, к некоторым другим видам может быть статистически значима.

В табл. 7.2 приведен видовой состав древостоя. Всего в анализ включено 1052 точек наблюдений, однако следует отметить, что устойчивые результаты можно получить и по 60—100 точкам.

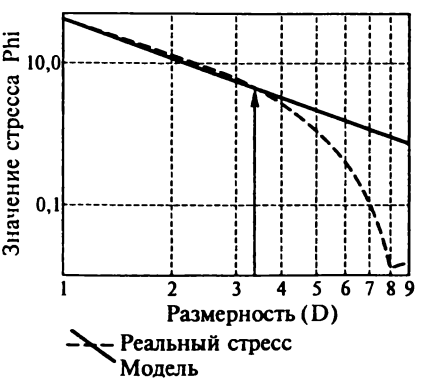


Рис. 7.5. Оценка размерности экологического пространства по функции стресса при матрице дистанций на основе гамма-корелляции

На рис. 7.5 продемонстрированы результаты оценки размерности пространства по стрессу. В общем случае имеет смысл рассматривать отношения в системе трех (максимум — четырех) факторов. Последующий анализ показал, что четвертый фактор почти не дает «новой» информации, т.е. в существенной степени не определяет ни один из видов. В связи с чем размерность пространства была принята равной трем.

В трехмерном пространстве может быть восемь подобластей (табл. 7.3).

Виды, занимающие различные подобласти по одному или большему числу факторов, взаимно дополняют друг друга. Если размещение видов в пространстве факторов соответствует теории, то виды, принадлежащие одной подобласти, должны различаться по ведущим факторам. В данном регионе дуб и ель занимают одну общую подобласть, но различаются по ведущим факторам: у ели — первый, у дуба — второй. Береза занимает собственную подобласть,

Таблица 7.2

Встречаемость видов в древесном ярусе на трансектах

Вид	Встречаемость	
	Количество описаний	Доля, %
Ель (Picea abies (L.) Karst.)	1021	97,0
Береза (Betula pendula Roth)	857	81,4
Осина (Populus tremula L.)	599	56,9
Рябина (Sorbus aucuparia L.)	335	31,8
Ива (Salix caprea L.)	227	21,5
Ольха серая (Alnus incana (L.) Moench)	187	17,7
Клен (Acer platanoides L.)	137	13,0
Липа (Tilia cordata Mill.)	137	13,0
Вяз (Ulmus laevis Pall.)	130	12,3
Ольха черная (Alnus glutinosa (L.) Gaertn.)	70	6,6
Сосна (Pinus sylvestris L.)	43	4,0
Ясень (Fraxinus excelsior L.)	9	0,8
Черемуха (Padus avium Mill.)	9	0,8
Дуб (Querqus robur L.)	2	0,1
Всего описаний	1052	

Таблица 7.3

Коэффициенты чувствительности видов к виртуальным факторам экологического пространства

(полужирные линии разделяют экологическое пространство на восемь подобластей)

Вид	Номер фактора			Индикаторы гомогенных подобластей		
	1	2	3	Фактор 1	Фактор 2	Фактор 3
Ель (Picea)	**0,772**	0,353	0,374	+	+	+
Дуб (Querqus)	0,407	**1,168**	0,288	+	+	+
Береза (Betula)	**0,537**	0,394	–0,454	+	+	–
Ольха черная (Alnus glutinosa)	**0,727**	**–0,873**	0,044	+	–	+
Осина (Populus)	0,288	–0,008	**0,978**	+	–	+
Сосна (Pinus)	**1,390**	–0,238	–0,243	+	–	–
Рябина (Sorbus)	0,063	**–0,898**	–0,344	+	–	–
Липа (Tilia)	**–0,888**	0,028	0,365	–	+	+
Черемуха (Padus)	–0,337	0,435	**–0,876**	–	+	–
Серая ольха (Alnus incana)	–0,486	**0,659**	–0,040	–	+	–
Клен (Acer platanoides)	–0,455	**–0,585**	**0,417**	–	–	+
Вяз (Fraxinus)	**–0,808**	–0,242	0,234	–	–	+
Ива (Salix)	–0,275	–0,125	**–0,644**	–	–	–
Ильм (Ulmus)	**–0,935**	–0,066	–0,100	–	–	–

Примечание. Полужирным шрифтом выделены ведущие факторы.

но она почти в равной степени зависит от всех трех факторов, что прямо указывает на ее широкое распространение и возможность комбинаций с разными видами. Ольха черная и осина занимают одну подобласть, но их ведущие факторы принципиально различны. При этом ольха черная почти в равной степени зависит от двух факторов, что прямо указывает на относительно малый объем в многомерном пространстве ее экологической ниши. Продолжая анализировать табл. 7.3, убеждаемся, что правило: «в одной «гомогенной» нише виды должны различаться по ведущим факторам» полностью реализуется.

Зная отношения видов к условиям среды, можно в первом приближении определить физический смысл факторов экологического пространства. Почти не вызывает сомнения тот факт, что фактор 1 отражает отношение видов к теплу. Положительная область занята бореальными видами — елью, березой, осиной, а отрицательная — типичными широколиственными видами: ильм, вяз, липа. Такое закономерное соотношение несколько нарушает дуб, попадающий в группу бореальных видов, но, помня о том, что в анализе всего два случая его обнаружения, этим фактом можно пренебречь. С другой стороны, фактор 1 не является для дуба ведущим.

Существенно труднее понять смысл фактора 2: с одной стороны, с ним в наибольшей степени положительно связаны дуб и ольха серая, а отрицательно — ольха черная, рябина и в меньшей степени — клен. Два вида одного рода ольхи резко дифференцированы по этому фактору. Это позволяет предположить, что фактор 2 отражает отношение к режиму увлажнения, с одной стороны, в пределе к застойному, а с другой — с постоянно хорошим дренажом.

Положительную область третьего фактора определяет в первую очередь осина и клен, а отрицательную, безусловно, виды, связанные с распадом основного древесного полога: черемуха и ива козья. Береза также существенно зависит от отрицательных значений этого фактора. Клен — безусловно вид старых лесов; береза — пионерный вид. Эти соотношения нарушает осина. Обычно осина рассматривается как пионерный вид. Однако на рассматриваемой территории осины в первом ярусе характерны для старых еловых лесов, а молодые леса с доминированием осины практически отсутствуют.

Если принять это условие, то фактор 3 отображает возраст сукцессионной стадии леса или освещенность. Такая идентификация физического смысла факторов, конечно, не очень корректна. Она опирается на априорные знания требований видов к условиям среды, что само по себе противоречит смыслу анализа, в результате которого необходимо выявить эти требования. Поэтому такого рода оценки «смысла факторов» следует рассматривать не более как гипотезы.

Расчет координат пространства Евклида через векторное пространство для такого массива данных может быть осуществлен только методом наименьших квадратов для исходных, логарифмированных и ранжированных данных. Естественно, что в каждом варианте значения осей несколько отличаются, хотя в целом они подобны друг другу.

Обратим внимание на то, что в действительности экологические ниши могут иметь весьма сложные конфигурации. В векторном пространстве их координаты определены относительно их не-

которого центра тяжести. Соответственно при переходе в пространство Евклида получаем практически линейное отображение факторов. Следовательно, оправданно аппроксимировать обилие видов от осей полиномиальной формой второй степени с мультипликативными частями. Задача решается методом пошаговой регрессии, но для области существования вида.

В табл. 7.4 приведены соответствующие модели значения коэффициента детерминации R^2.

Из таблицы следует, что большинство видов описывается тремя факторами вполне удовлетворительно. При этом велико мультипликативное влияние факторов 3 и 1 и факторов 1 и 2. Только черемуха и ольха серая существенно определяются квадратичной частью уравнения.

Собственно форму экологических ниш можно представить несколькими способами:

1. Используя программу Surfer, построить методами триангуляции или «естественного ближайшего соседа» двухмерные отображения.

2. Описать ниши с помощью метода нелинейной аппроксимации в статистических пакетах или Surfer.

3. Построить трехмерные ниши в треугольной системе координат по уравнению, приведенному в табл. 7.4 (программа Statistica).

Первый способ дает локально сглаженные поверхности и наиболее полное отображение проекции; во втором — двухмерная поверхность строится методом итерационной оценки параметров в статистических пакетах программ. Возможность представления на рисунке и сглаженной поверхности, и реальных точек наглядно демонстрирует, как реальные наблюдения описываются регрессионной моделью. Аппроксимирующий двухмерный полином в этом случае выбирается в соответствии с табл. 7.4, описывающей видовые ниши.

Отображение в треугольнике строится по схеме, показанной на рис. 7.6, где трем факторам даны условные названия, соответствующие принятой гипотезе. Значения факторов ранжируются от нуля до единицы. Вертикально к этому треугольнику воспроизводится ось, отражающая обилие вида. Модель, описывающая нишу, вводится в соответствии с результатами, приведенными в табл. 7.4. Этот способ отображе-

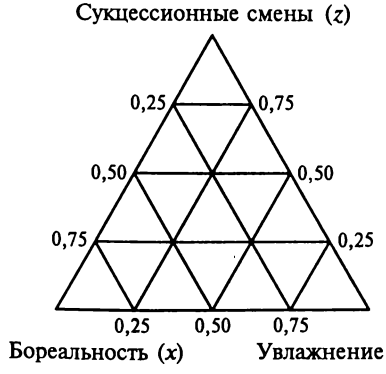


Рис. 7.6. Схема отображения экологической ниши $Y = b_0 + b_1x + b_2y + b_3z + b_4x^2 + b_5y^2 + b_6z^2 + b_7xy + b_8xz + b_9yz$

Таблица 7.4

Описание факторами экологического пространства сумм площадей сечения видов деревьев

$Y_i = b_0 + b_1X_{11} + b_2X_2 + b_3X_3 + b_4X_1^2 + b_5X_2^2 + b_6X_3^2 + b_7X_1X_2 + b_8X_1X_3 + b_9X_2X_3$ ($b_{0, 1, 2, \ldots, 9}$ — параметры модели)

Сумма площадей сечений (Y_i)	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	R^2
Picea	1,34	1,56	1,15	0,74	—	—	–0,45	—	—	—	0,62
Betula	1,30	0,54	—	–1,00	—	1,23	0,29	2,59	—	—	0,62
Alnus glutinosa	0,56	1,98	–1,73	—	—	2,13	—	—	—	–2,39	0,85
Populus	0,89	—	—	0,67	—	—	0,41	—	1,29	—	0,76
Pinus	0,72	0,90	—	–0,66	1,20	–1,57	—	–1,37	–1,75	5,90	0,99
Sorbus	1,27	—	–1,63	–0,24	—	—	—	1,24	—	—	0,51
Tilia	1,55	–1,62	—	—	—	—	—	—	–1,65	2,13	0,65
Padus	0,58	1,96	–0,49	—	–10,30	9,87	—	—	—	2,44	0,97
Alnus incana	0,82	—	1,62	—	–1,18	1,92	—	–3,14	—	—	0,69
Acer	0,96	—	–0,54	0,31	—	—	—	—	—	–1,16	0,29
Fraxinus	1,14	—	—	—	—	—	—	–7,17	—	—	0,42
Salix	1,00	—	—	–0,71	—	—	0,72	—	—	—	0,25
Ulmus	1,34	–1,16	—	—	—	—	—	—	—	—	0,39
Общая сумма площадей сечений	10,1	—	6,10	—	28,87	43,09	10,42	—	—	—	0,64

ния трехмерных пространств включен только в пакет Statistica. Демонстрация разных способов отображения экологических ниш показывает возможности визуализации данных при представлении результатов анализа.

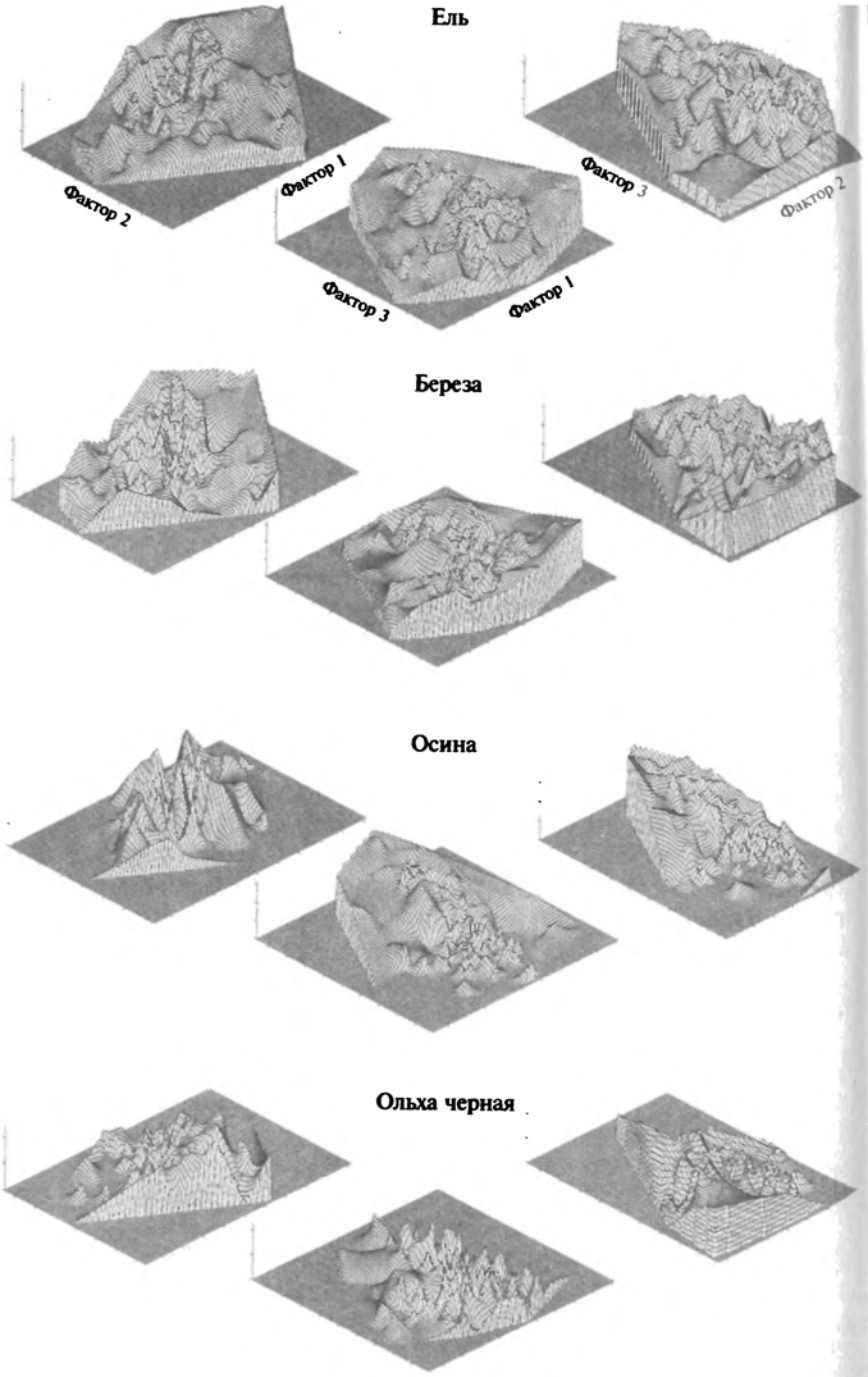
На рис. 7.7 показаны проекции экологических ниш бореальных видов из трехмерного пространства в двухмерные. Обратим внимание на то, что структура проекций сама по себе показывает, от каких факторов зависят виды. Если в проекции нет ориентированной в пространстве поверхности, то вид не зависит, по крайней мере, от одного из образующих ее факторов. Так, например, совершенно очевидно, что ниша осины не зависит от факторов 1 и 2 и в основном определяется фактором 3. Ель зависит в первую очередь от факторов 1 и 2 и в меньшей степени от фактора 3. Проекции весьма наглядно демонстрируют правила расхождения ниш видов: различие в ведущей роли хотя бы по одному фактору. Ниши пересекаются, что и отражает сочетание видов в разных пропорциях в пространстве, но области оптимума у них никогда не совпадают.

Среди широколиственных пород (рис. 7.8) наиболее близкие ниши у липы и вяза, но при этом область оптимума у липы больше, чем у вяза. Ниша клена, который в основном распространен во втором ярусе, по своей пространственной структуре весьма неопределенна.

На рис. 7.9—7.13 показаны отображения ниш двухмерными полиномами средствами Statistica. Это удается сделать для большинства видов с достаточно высокой надежностью, так как их ниши (за исключением березы) почти двухмерные, т.е. описываются в основном двумя факторами. В принципе отображения тождественны первым. Однако здесь хорошо видно, что ниша рябины описывается на уровне тенденции. Далеко не идеально описывается и ниша вяза.

Отображения в треугольнике (рис. 7.14, 7.15) как в трехмерном пространстве несколько хуже аппроксимируют данные, но вместе с тем наглядно отражают схему дифференциации экологических ниш. Здесь дополнительно к отображениям, полученным первыми двумя методами, показана ниша ивы, которая также весьма неопределенна. В целом те виды, которые имеют низкие значения коэффициента детерминации, естественно хуже отражаются и на графике.

Вместе с тем различные способы описания и отображения экологических ниш показывают, что в векторном пространстве хорошо отображаются их центры тяжести, обобщенные в коэффициентах чувствительности, а в Евклидовом пространстве — их форма. Оба отображения демонстрируют четкий общий принцип взаиморазмещения видов: каждый вид имеет свой экологический оптимум в трехмерном пространстве, размещение их такого, что они


Ель
Фактор 2
Фактор 1
Фактор 3
Фактор 3
Фактор 2
Фактор 1
Береза
Осина
Ольха черная

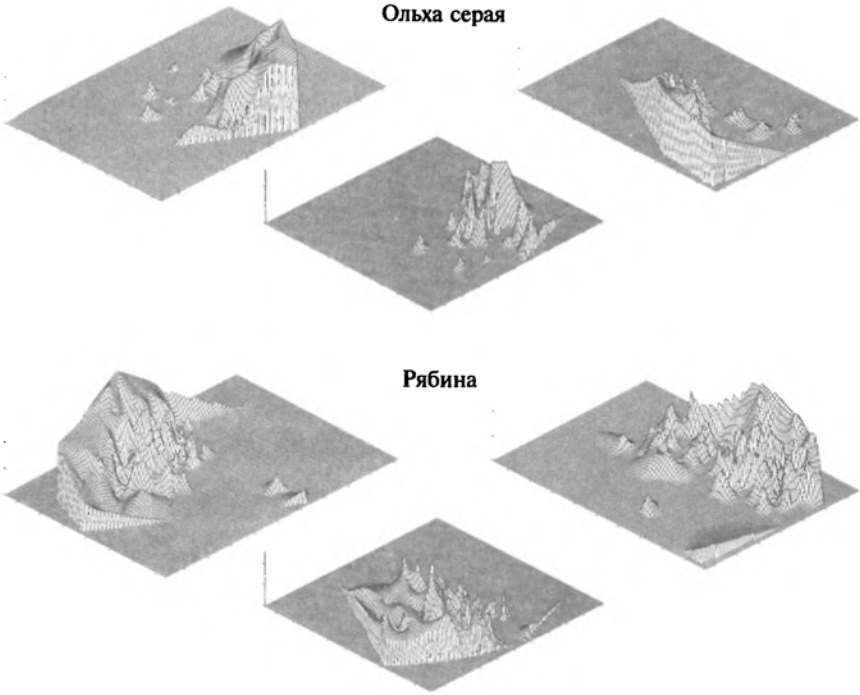


Рис. 7.7. Двухмерные проекции экологических ниш бореальных видов

слабо связаны друг с другом (за исключением черной и серой ольхи), образуя самые различные комбинации. В результате покров, формируемый различными видами деревьев, становится непрерывным.

Идентификация содержания факторов осуществляется на основе сопоставления их с независимо измеренными переменными. На рис. 7.16 показан рельеф в сопоставлении со значениями первого фактора. Связь на уровне тенденции очевидна: максимум значения фактора, т.е. бореальные леса, соответствует понижениям, минимум — возвышенностям, что отражает увеличение роли широколиственных пород на склонах и водораздельных поверхностях моренной гряды. В каждой точке трансекта по бурению до глубины 1,0 м определялся механический состав почв, который ранжировался в баллах от 9 (песок) до 1 (тяжелый суглинок) и 0 — органогенный горизонт. Кроме того, в каждой точке измерялась глубина вскипания почвы. Рельеф, используя метод главных компонент, можно разложить на микро-, мезо- и макроформы. Для этого ряд смещается относительно самого себя на 7 точек (анализ структуры рельефа будет рассмотрен ниже в соответствующем разделе. Здесь же примем полученный результат без дополнительных объясне-

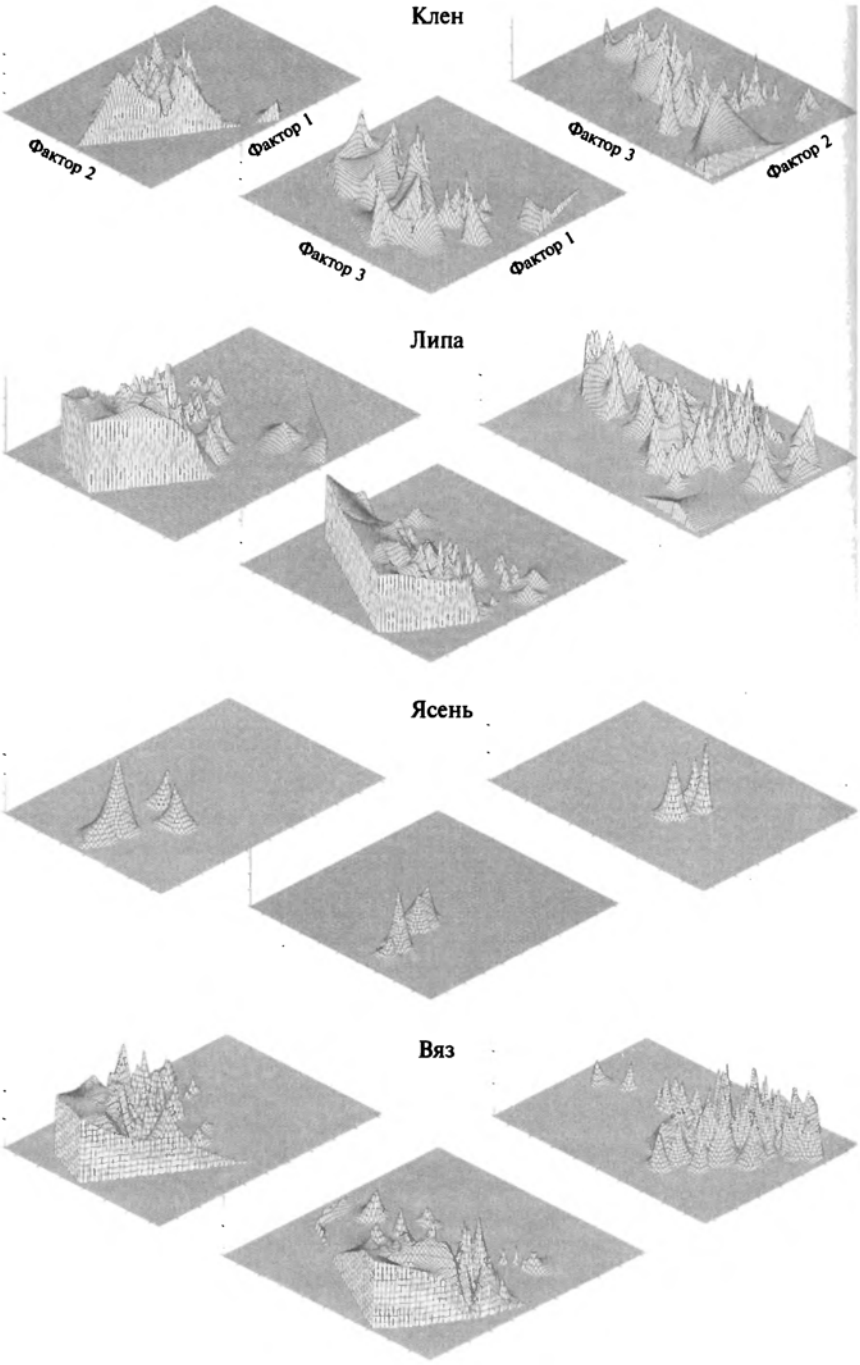


Рис. 7.8. Двухмерные проекции экологических ниш широколиственных видов

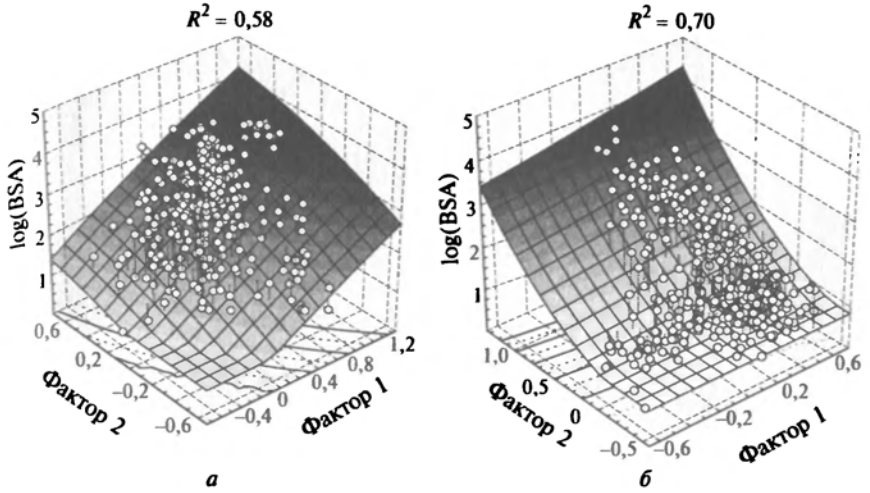


Рис. 7.9. Отображение экологической ниши ели (*а*) и осины (*б*) относительно факторов 1 и 2

ний). Каждая форма представляется через собственную высоту, крутизну и профиль поверхности. Таким образом, получаем возможность использовать три иерархических уровня организации рельефа в качестве косвенных характеристик физических свойств местообитания. Очевидно, что чем круче склон и более выпуклая форма рельефа, тем лучше дренаж и меньше увлажнение почвы.

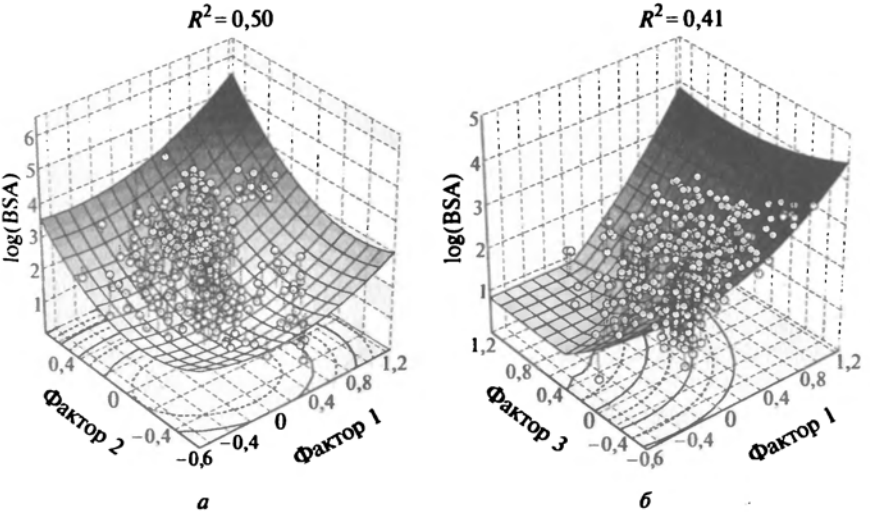


Рис. 7.10. Отображение экологической ниши березы по отношению к факторам 1 и 2 (*а*), факторам 1 и 3 (*б*)

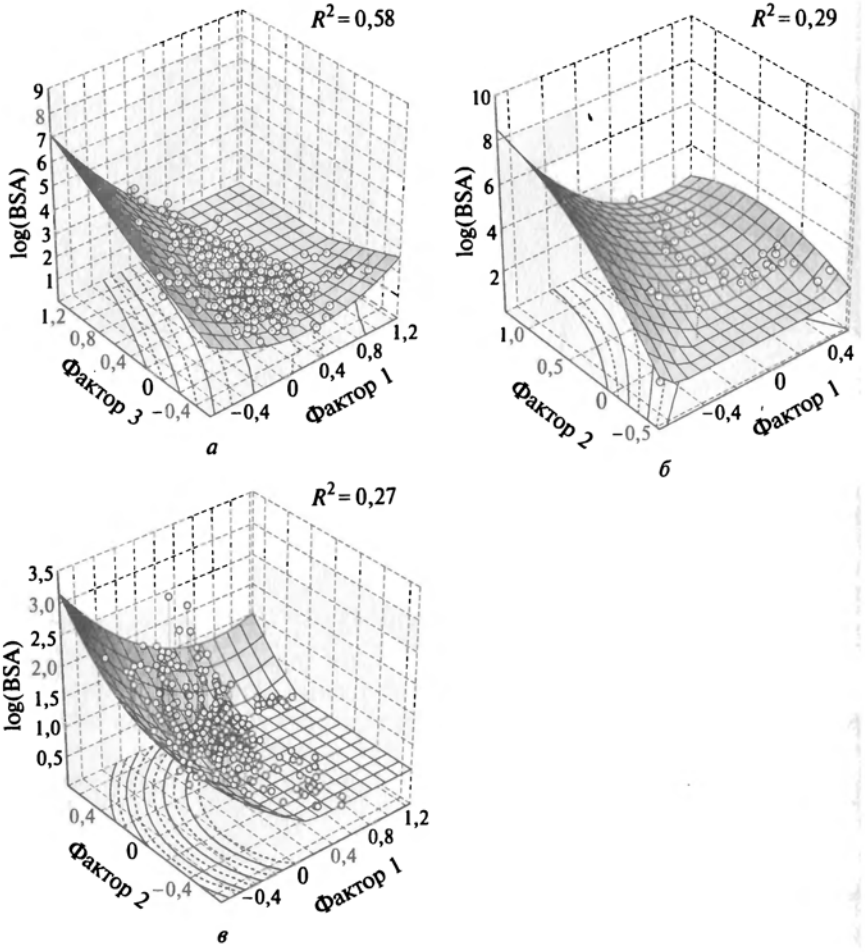


Рис. 7.11. Отображение экологических ниш широколиственных пород:

а — липа по отношению к факторам 1 и 3; *б* — клен по отношению к факторам 1 и 2; *в* — вяз по отношению к факторам 1 и 2

Однако следует иметь в виду, что наша информация о среде чрезвычайно неполна. Крутизна и профиль склонов определены только по направлению трансекта, а не по максимальному градиенту реальной поверхности, поэтому информация, содержащаяся в таких характеристиках, не более чем на 50 % отражает реальность. Кроме того, содержание влаги в почве зависит не только от характеристик рельефа и почвообразующих пород в конкретной точке бурения, но и в существенной степени — от их сочетания на некоторой площади, соизмеримой с площадью оценки состояния растительности.

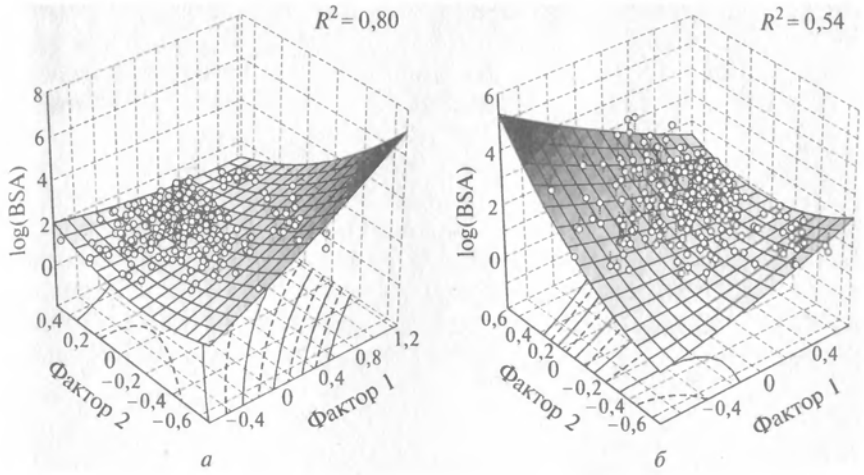


Рис. 7.12. Отображение экологических ниш ольхи черной (*а*) и ольхи серой (*б*) по отношению к факторам 1 и 2

Наблюдения проведены на моренных отложениях сложного генетического состава, перекрытых покровным суглинком, чем объясняется большая мозаичность местообитаний. При всех этих условиях можно рассчитывать лишь на качественную оценку физического смысла координат экологического пространства.

Можно было бы привести более простые и легко интерпретируемые отношения, например, для горных условий или зандровых равнин. Однако выбор этой, наиболее сложной, ситуации оправдан целесообразностью демонстрации того минимума, на который можно рассчитывать, применяя методы ординации на основе многомерного шкалирования для наиболее сложных территорий.

На рис. 7.17 показаны результаты факторного разложения рельефа, позволившие выделить три иерархических уровня его организации, которые условно можно назвать макро-, мезо- и микрорельеф. Теперь, используя какой-либо статистический пакет программ или Excel, возьмем первую и вторую производные от этих трех форм рельефа

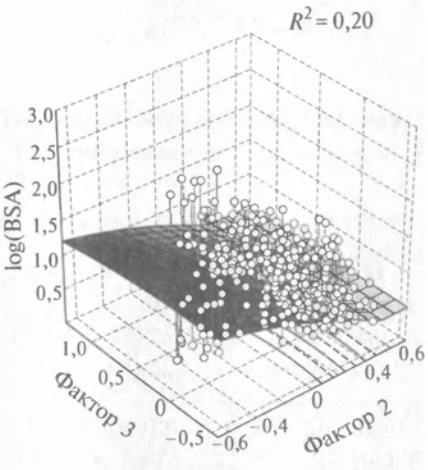


Рис. 7.13. Отображение экологической ниши рябины по отношению к факторам 2 и 3

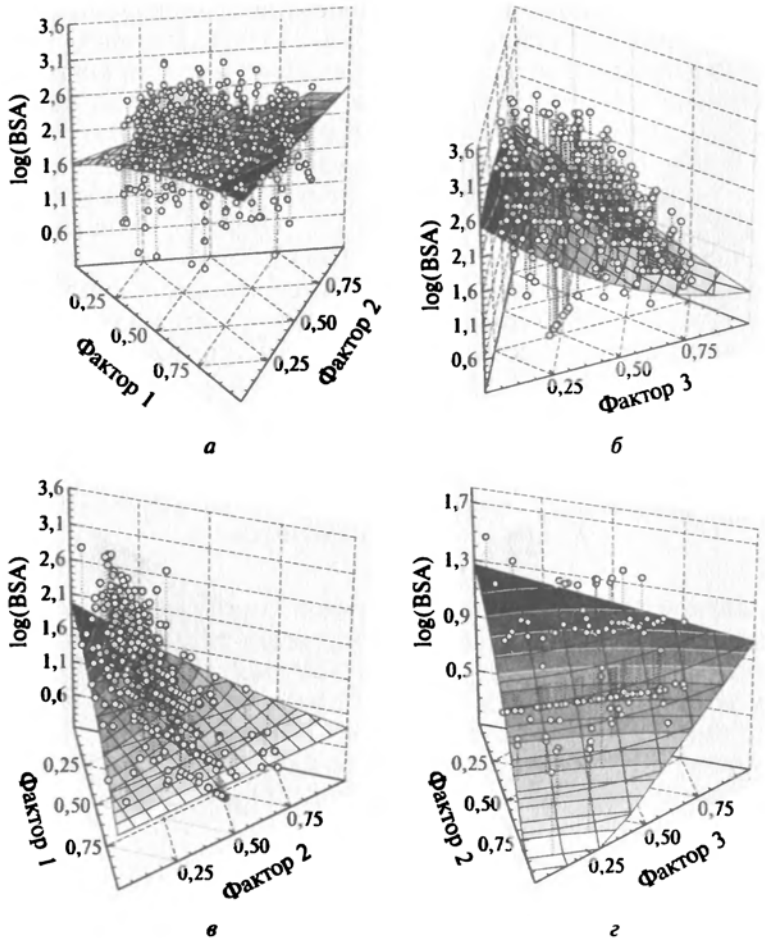


Рис. 7.14. Экологические ниши ели (*а*), березы (*б*), осины (*в*) и рябины (*г*) в трехмерном пространстве координат

и получим показатели крутизны и экспозиции склонов разных уровней и их профиль.

В качестве факторов будем рассматривать все характеристики рельефа и почв, используя метод пошаговой регрессии. Будем оценивать, как эти переменные «среды» отражаются в координатах экологического пространства. В табл. 7.5 показаны результаты такого отображения.

Из сопоставления первой координаты экологического пространства с переменными среды следует, что заметное участие широколиственных лесов характерно в основном для возвышенных форм макро- и мезорельефа с хорошо выраженным склоном макрорельефа и в меньшей степени определяется тяжелым суг-

линком в подзолистом горизонте почвы, но подстилаемом на глубине легкими грунтами. Следует обратить внимание на то, что глубина вскипания почв не влияет на значения этой координаты экологического пространства. Действительно, прямые наблюдения показывают, что широколиственные породы четко приурочены к склонам моренных гряд, для которых типично вскипание почв на глубине 60—150 см. Для плоских, иногда слабовогнутых ступеней гряд, с редким, близким к поверхности вскипанием почвы, характерны типичные бореальные леса. Это заставляло полагать, что распространение широколиственно-хвойных лесов определяется карбонатными моренными отложениями. Однако проведенный анализ заставляет усомниться в справедливости этого

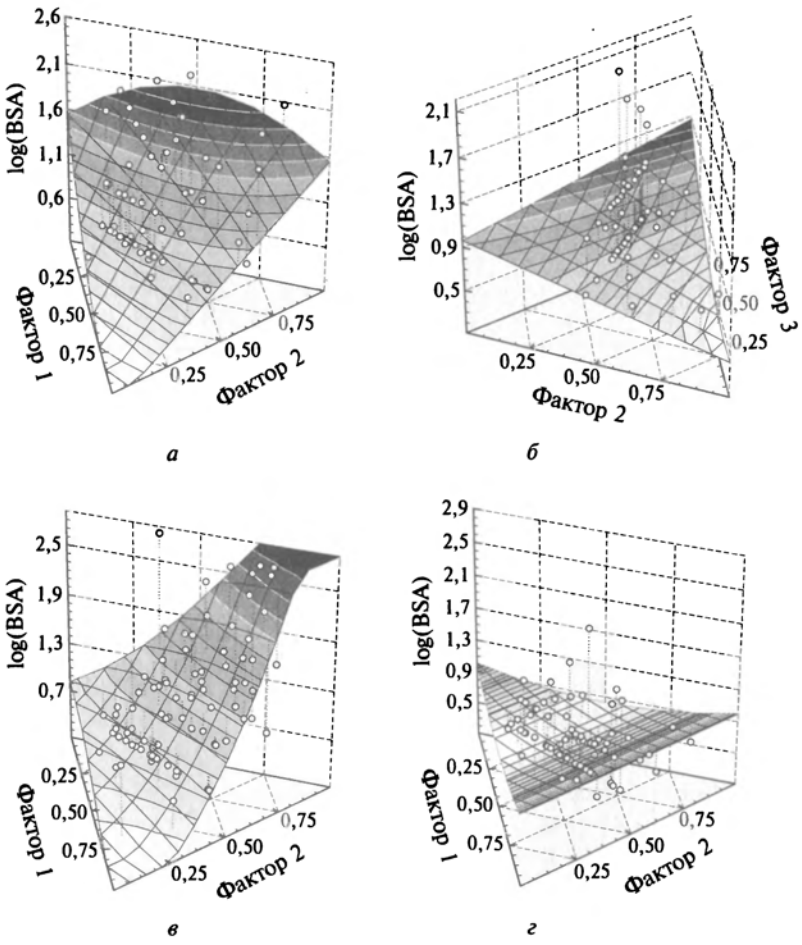


Рис. 7.15. Экологические ниши липы (*а*), клена (*б*), ольхи серой (*в*) и ивы (*г*) в трехмерном пространстве координат

Таблица 7.5

Параметры регрессионных отношений между координатами пространства и характеристиками среды

Координата	Коэффициент детерминации и F-критерий	Переменная	Параметры регрессии			Комментарий
			BETA	t(417)	p-level	Бореальные виды:
Первая (положительная область — бореальные, отрицательная — широколиственные породы)	0,259	Макрорельеф	–0,231594	–5,66797	0,000000	В понижениях
	24,292	Макросклон	–0,214980	–5,04204	0,000001	На плоских поверхностях
		Мезорельеф	–0,236967	–5,83236	0,000000	В понижениях
		Микрорельеф	0,106782	2,63906	0,008625	В повышениях
		М* 25 см	–0,208874	–4,87099	0,000002	Легкие почвы
		М 80 см	0,092329	2,21546	0,027268	Тяжелые почвы
Вторая (положительная область — серая ольха, дуб, отрицательная — черная ольха)	0,0971	Макрорельеф	–0,173358	–3,57221	0,000396	На повышениях серая ольха
	5,5788	М 5 см	0,110145	2,19284	0,028874	Положительная область:
		М 20 см	0,201958	2,11179	0,035302	легкие почвы
		М 25 см	–0,312756	–2,30507	0,021655	тяжелые почвы
		М 30 см	0,258974	2,35458	0,019009	легкие почвы
		М 70 см	0,197487	2,30281	0,021783	
		М 75 см	–0,337907	–3,09169	0,002124	тяжелые почвы
		М 80 см	0,126192	1,43073	0,153261	легкие почвы

* М — механический состав почв на глубине.

предположения. Становится вполне реалистичной гипотеза в основном о температурном содержании этой координаты. Можно полагать, что для пониженных территорий типичны зимние инверсии температуры, которые и ограничивают распространение широколиственных пород. Гипотезу легко проверить, установив по градиенту фактора несколько микроклиматических станций. Оперативные измерения температуры почвы термощупом в середине дня в июне подтверждает эту гипотезу.

Вторая координата в очень слабой степени, но достоверно описывается переменными среды. В общем, положительная область координаты, которой соответствует серая ольха и дуб, характерна для возвышенных поверхностей с многочленными, легкими сверху почвообразующими породами. Противоположные условия типичны для заведомо устойчивой к застойному увлажнению черной ольхе. Скорее всего, в данном случае приемлема гипотеза, трактующая эту координату экологического пространства как «режим увлажнения». Проверка этой гипотезы, очевидно, также возможна — достаточно в течение вегетационного сезона провести прямые измерения влажности почвы примерно в 20—30 точках градиента координаты примерно один раз в десять дней. Более частые измерения необходимо

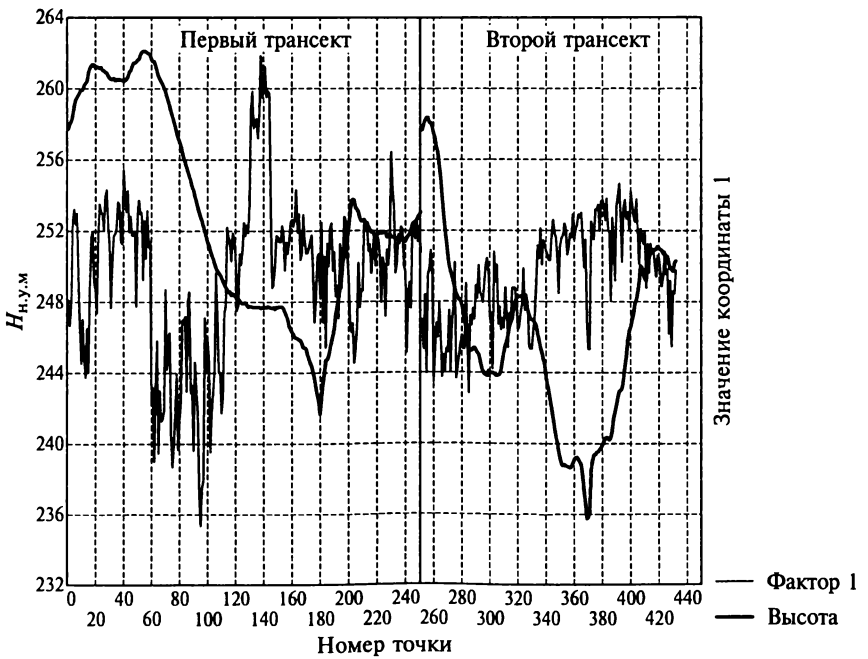


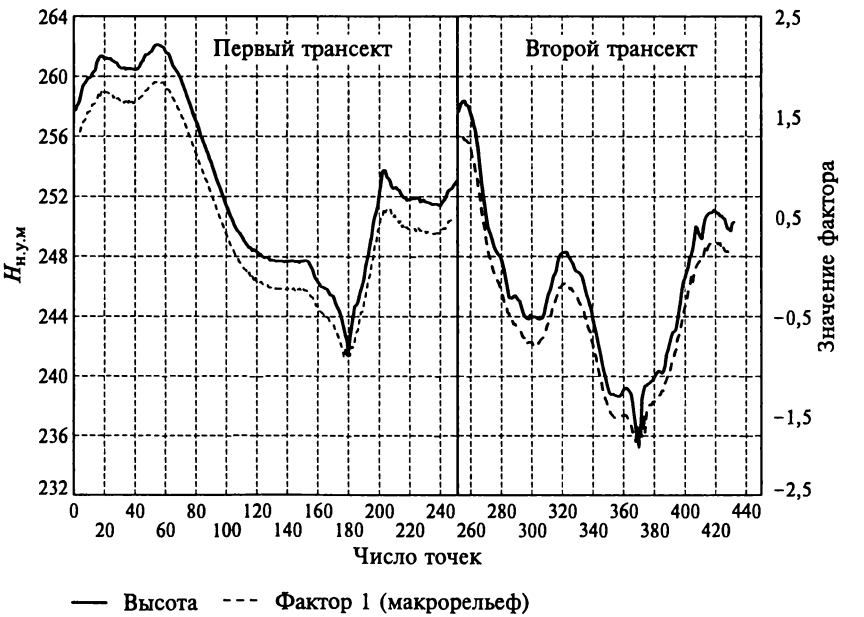
Рис. 7.16. Соотношение координаты 1 экологического пространства и рельефа

Таблица 7.6

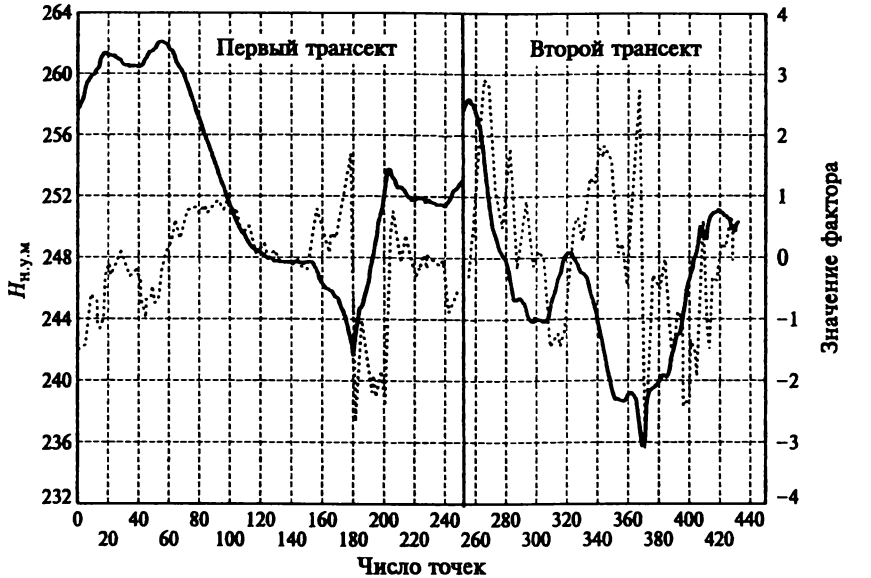
Регрессионная модель третьей координаты экологического пространства от независимых характеристик древесного яруса и характеристик среды

(коэффициент детерминации $R^2 = 0{,}39867$; F-критерий (4,114) = 18,895; стандартная ошибка — 0,32764)

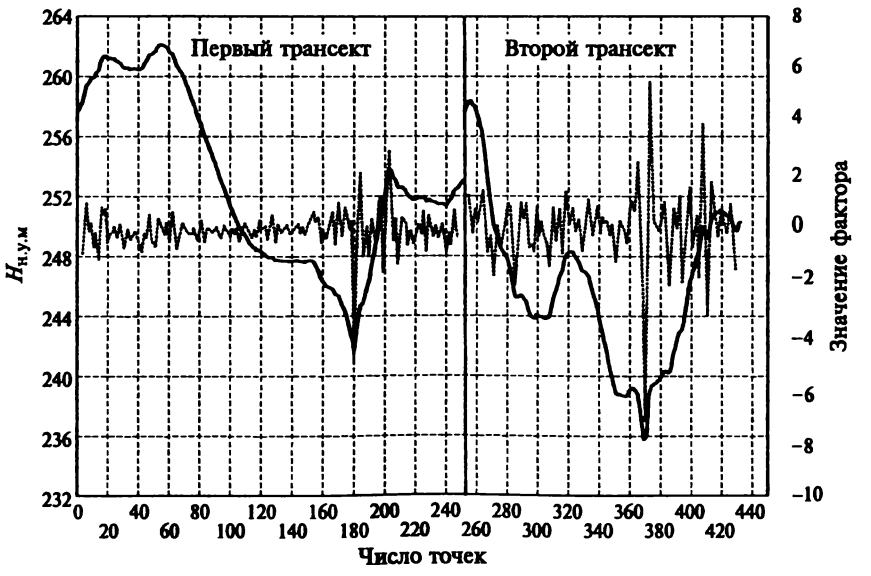
Переменная	BETA	Std. Err. of BETA	*b*	Std. Err. of *b*	t-критерий	Уровень значимости p-level
Константа			–0,377233	0,110309	–3,41979	0,000870
Общая сомкнутость	0,249535	0,076824	0,398835	0,122788	3,24815	0,001526
Диаметр первого яруса	0,415752	0,081966	0,018903	0,003727	5,07225	0,000002
Высота мезорельефа	0,418324	0,126121	0,137095	0,041333	3,31684	0,001222
Форма мезорельефа	0,355198	0,124227	0,450058	0,157403	2,85928	0,005051



а



б



в

Рис. 7.17. Разложение рельефа на иерархические уровни организации методом факторного анализа:

а — фактор 1; *б* — фактор 2; *в* — фактор 3

проводить после схода снегового покрова и затяжных летних дождей, чтобы учесть помимо содержания влаги скорость снижения переувлажнения.

Высокая связь третей координаты с диаметром деревьев в первом ярусе и общей сомкнутостью древостоя показывает, что действительно третья координата экологического пространства описывает сукцессионные смены или иначе сукцессионный возраст сообщества. Некоторые коррективы в этот процесс вносит мезорельеф, отражающий факт большой устойчивости старых лесов на возвышенных, но слегка вогнутых поверхностях мезорельефа (табл. 7.6).

Оценим полученные результаты с общих позиций:

1. Положения большинства видов удовлетворительно описываются координатами экологического пространства.

2. Соотношение координат экологического пространства с независимо измеренными характеристиками среды позволяют высказать проверяемые гипотезы об их физической природе.

Конечно, желательно получить более однозначное определение параметров древесных пород к изменению тепла и влаги. Если известны параметры видов по отношению к теплу, влаге и свету, то относительно легко, используя технологию так называемых гэп-моделей (моделирование сукцессионной динамики через мозаичный распад полога древостоя), описать динамику сообществ при постоянном и флуктуирующем климате. Такие модели, очевидно, имеют практическое значение. Правда для того, чтобы связать динамику с варьированием гидротермического режима в пространстве, необходимо построить модели тепло-влагообеспеченности поверхности от рельефа, что является в принципе разрешимой, но непростой задачей. Однако так как существующий состав древесного полога с учетом его сукцессионной стадии сам по себе хорошо отражает гидротермический режим местообитаний, то поверхность местообитания через породный состав можно описать двумя собственными координатами экологического пространства, на основе которых можно строить собственно модели динамики. Индикация гидротермического режима через видовой состав более легко решаемая задача, чем специальное построение поверхностей гидрологического и теплового режимов. Так или иначе, экологическая ординация прокладывает прямой путь к моделированию динамики.

Рассмотрим возможные причины плохого отображения некоторых видов в экологическом пространстве и вообще существенных отклонений наблюдаемого обилия вида от воспроизводимого по вычисленным координатам. Эти причины можно разделить на три типа:

1. Неполное соответствие выбранной метрики реальности и неадекватная геометрическая модель пространства.

2. Неравновесное текущее состояние объекта исследования.

3. Прямые ошибки полевых измерений.

Первая причина в простейшем варианте разрешается поиском наилучшей метрики и способа представления данных. Однако это самый простой случай. Источник искажений может быть связан с тем, что в существенно различных местообитаниях, например на дренируемых поверхностях и в условиях слабого дренажа, само отношение видов к одним и тем же физическим факторам среды может быть различно. Объединяя местообитания с такими различными отношениями в одну систему, неизбежно будем получать осредненные отношения с большими ошибками для каждой частной ситуации. Это означает, что принята неадекватная геометрическая модель пространства.

Леса территории, рассмотренной в данном примере, тридцать лет назад существенно пострадали от вырубок. Кроме того, для них характерны периодически повторяющиеся сплошные и локальные ветровалы. Различные сообщества, охваченные наблюдениями, находятся совершенно на разных стадиях сукцессоных преобразований. Такие преобразования по условию проходят очень быстро и в локальных точках существенно случайны. Так как любые статистические методы отражают только равновесные, среднестатистические отношения, они заведомо не могут описать неравновесные процессы. Скорее всего, именно с этим связано относительно плохое предсказание экологическими координатами клена, древовидной козьей ивы и в какой-то степени вяза. Клен типичен в окнах старых еловых лесов, а эти окна весьма непостоянны во времени. Ива характерна для участков, формирующихся на ветровалах в условиях достаточного увлажнения.

Неравновесные состояния невозможно отличить от элементарных ошибок измерения, неизбежных в любых полевых исследованиях. Неизбежны и ошибки при вводе больших массивов информации в базу данных. Читатель должен понимать, что не существует человека, который мог бы ввести данные из полевых записей в базу данных на персональном компьютере без ошибок. Более того, некоторых исследователей, причем очень хороших, вообще нельзя допускать до этой очень важной и ответственной работы, поскольку они психологически не приспособлены к ее выполнению. Далеко не всегда эти ошибки можно выявить прямым просмотром данных. Типичные ошибки на порядок величины (введена лишняя цифра) можно выявить прямым просмотром графика. Однако неизбежны ошибки, не выводящие величину за пределы разумного. Такие ошибки можно выявить только на основе проведения полного, продемонстрированного выше анализа непосредственно вернувшись в точки, где были получены не укладывающиеся в норму отношений результаты. Получив результат анализа, выделив наиболее статистически значимые отклонения реальных значений от расчетных и проверив их по полевой документации, можно устра-

нить ошибки ввода данных. Проведя повторный анализ, выделяют новые статистически значимые отклонения, реальность которых необходимо проверить повторными измерениями в соответствующей точке.

Работа на трансектах облегчает проведения таких проверок. Однако, используя GPS, можно с необходимой точностью вернуться в любую точку наблюдений. Следует особо обратить внимание на то, что исключение отклонений без их проверки в поле недопустимо. Если эти отклонения реальны, то они являются прямым свидетельством неравновесных процессов в конкретном элементе исходной системы. Степень же неравновесности всей территории является ее очень важной, но до настоящего времени практически неиспользуемой характеристикой объекта исследования.

Контрольные вопросы

1. Почему отношения видов к условиям среды всегда описываются нелинейными зависимостями?
2. Каковы условия сосуществования видов со сходными требованиями к ресурсам и условиям в пределах гомогенной территории?
3. Почему для ординации видов в пространстве факторов не применимы многомерные параметрические методы анализа?
4. Как показать правило размещения видов в векторном пространстве?
5. Назовите основные источники «ошибок» при ординации.
6. Какова область фундаментальных и прикладных исследований в экологии, для которых эффективен метод ординации на основе многомерного непараметрического шкалирования?

Глава 8
КОЛИЧЕСТВЕННЫЕ МЕТОДЫ КЛАССИФИКАЦИИ (КЛАСТЕР-АНАЛИЗ)

8.1. Общие представления о классификации

Рассмотренные выше методы многомерного анализа позволяют отобразить исходную систему, определенную на множестве признаков-переменных и множестве элементов в системе координат конечной размерности в непрерывной форме. Непрерывная форма отображения создает некоторую основу для построения динамических моделей изучаемого явления. Однако она в полной мере применима только к действительно непрерывным множествам и заведомо будет искажать реальность, если множества дискретны. С другой стороны, когда переменные измеряются квалиметрически или с очень невысокой точностью непрерывное отображение их в многомерном пространстве часто содержит избыточную информацию. Все это определяет естественный переход к другой форме отображения реальности через выделение дискретных образов на основе классификации.

Выделение классов или образов, различимых состояний явления в рамках эволюции познания мира, выше трактовалось как способ сжатия информации или уменьшения ее начального разнообразия. В идеале желательно, чтобы состояния и сопоставляемые с ними классы или образы были дискретны, т.е. каждый элемент, принадлежащий конкретному образу, содержал в себе информацию обо всех прочих элементах этого множества. Иначе говоря, различия свойств элементов, принадлежащих одному образу, чисто случайны и все его элементы относятся к одной генеральной совокупности. При этом варьирование свойств этих элементов описывается каким-либо одним, имеющим физический смысл, каноническим распределением случайного поведения.

Классификация явлений окружающего мира — фундаментальное свойство мышления и реализуется любым человеком. В результате общений членов социума в нем формируется единая схема классификации, общая система образов, которая адекватно воспринимается каждым индивидуумом. При этом образы формируются не только относительно реально дискретных, но и непрерывных явлений. Наглядным примером обучения распознавания

образов может быть известное стихотворение В. В. Маяковского «Что такое хорошо и что такое плохо». Через систему признаков вводится разделение множества возможных форм поведения на два непересекающихся образа. Причем все критерии «хорошего» и «плохого» вводятся таким образом, что некоторое возможное «среднее» не рассматривается. Этот принцип формирования адекватного восприятия мира всеми членами социума по существу является всеобщим. Члены социума, реализующие отличную от других схему выделения образов, обладают, с точки зрения других, неадекватным восприятием мира. Эта общая преамбула необходима для понимания формальных классификационных процедур, которые фактически воспроизводят различные формы реально существующего адаптивного поведения.

Научное познание мира на всех этапах своего развития неизбежно также опиралось на классификации, выделяя из всего разнообразия классы явлений и классы их состояний. Наиболее полной моделью научной классификации как наиболее сложного явления является таксономическая классификация форм жизни. Вполне понятно, что и до К. Линнея животные и растения у каждого народа имели свои названия. Однако эти названия часто объединяли организмы, существенно различные с точки зрения ученого. Для человека, не связанного с биологией, образ мыши объединяет широкую группу организмов и позволяет отличить их, например, от крыс. Для практической жизни большинства людей большего и не нужно. Как только человек начинает изучать некоторое явление более детально, он различает больше дискретных образов, для которых стремится выделить отличительные признаки и дать соответствующее название.

Часто оказывается, что рамках известного набора признаков множество организмов можно отнести к одной генеральной совокупности, но при более детальном «описании» могут появиться новые ранее неизвестные признаки, по которым множество разбивается на два или большее число однородных подмножеств. В идеале в таксономической классификации признаки должны быть бинарные (признак есть или его нет) или иметь четкий хиатус (разрыв). Например, длина зубного ряда у одного вида мышевидных грызунов меньше некоторой величины, а у другого — обязательно больше и между этими величинами существует разрыв. Далее, что было априори очевидно, признаки имеют различную степень общности, т. е. принадлежат большей или меньшей группе организмов.

По принципу этой общности естественно различать уровни классификации. Грубо говоря, все грызуны обладают одним общим признаком — сильно развитыми передними резцами, что позволяет объединить соответствующие организмы в один таксон достаточно высокого ранга. Все организмы, выкармливающие потомство моло-

ком, естественно объединить в таксон очень высокого ранга — млекопитающие. Очевидно, что такая классификация строится по наблюдаемым, т. е. физиономическим признакам.

Однако естественный опыт человека подсказывал, что сходство «физиономий» с высокой вероятностью свидетельствует о родстве. Родство, очевидно, это то, что связано с генезисом таксонов и потому более содержательно, чем простое сходство. С другой стороны, тот же опыт показывал, что если какие-то явления образуют достаточно дискретный набор состояний, то можно полагать, что они порождены и разными причинами. На этой основе возникает естественная связь между уровнем физиономического сходства и родством или общностью генезиса. Вполне понятно, что генетическая классификация, отражающая связь между явлениями во времени или по общности породивших их причин, существенно более содержательна, чем исходная физиономическая классификация. В результате построение генетических классификаций как средства познания стало целью всех естественных наук. Именно эта генетическая идея нашла отражение в названиях наиболее наблюдаемых иерархических уровней таксономической классификации «род» и «семейство».

К сожалению, в реальности все оказалось не так просто. Более детальные исследования, выявлявшие четкие бинарные признаки, заставляли пересматривать ранее построенные таксономические схемы. Биохимические, хромосомные и другие признаки молекулярного уровня, ставшие доступными для измерения, также трансформировали ранее созданные классификационные системы. С расширением системы признаков, сменой логических оснований таксономические классификации периодически изменялись и неизбежно будут изменяться и в дальнейшем.

Начиная с 60-х годов XX в. постоянно осуществлялись разработки формализованных таксономических моделей классификации. Эти работы, включая и программные средства, способные заменить стандартные определители, активно осуществляются и в настоящее время. С формальных позиций, создание определительной системы в рамках конечного множества признаков вполне возможно, однако число этих признаков огромно, а логика отношений между ними оказывается не всегда очевидной. Вполне естественно идет поиск формальных алгоритмов построения таксономической системы в замкнутом признаковом пространстве. Здесь фактически возникают те же проблемы, что и в определителе. Отличие алгоритмизированной схемы классификации от работы реального таксономиста состоит в первую очередь в том, что любые программные средства (алгоритмы) могут преобразовывать по заданному правилу только конечное множество признаков. Таксономист же ищет новые, ранее неизвестные признаки, осуществляя при этом огромный перебор наблюдаемых свойств организма. Поскольку это множество, априори, бесконечно, ника-

кой алгоритм и никакая вычислительная машина не сможет заменить исследователя, выполняя лишь функции его помощника.

Не менее важно и то, что логика классификации, осуществляемой человеком, не всегда не противоречива. Точнее, она несколько изменяется при работе с разными подмножествами или таксонами. Эти изменения часто отражают суть дела, различные механизмы дифференциации в подмножествах, однако для того чтобы обобщить эти изменения логических оснований при переходе от одного таксона к другому, требуется некоторая метамодель, разработка которой — дело будущего. Если нет правила (алгоритма), то и не может существовать формализованного метода анализа вообще и классификации в частности.

Есть все основания полагать, что эти и другие проблемы, с которыми сталкивается таксономист при классификации организмов, типичны для всех естественных наук. Однако необходимость воспроизводимости, которая особенно очевидна в задачах классификации, заставляет искать адекватные формализованные методы и формулировать жесткие логические основания. Без классификации невозможно исследовать окружающую нас реальность, без классификации, устанавливающей стандарты состояния, невозможно решать и чисто практические проблемы отношения человека со средой.

Вместе с тем, используя формальные методы, необходимо четко понимать, что каждый из них может упорядочить лишь некоторые реально существующие свойства объекта, часто не те, которые в рамках своих гипотез о природе явления ожидает исследователь. В некоторых случаях эти свойства интересны, в других — бессмысленны. И хотя можно ввести формальные оценки качества классификации, однако и они не могут являться абсолютным критерием.

Необходимо ясно понимать, что создание единой классификации какого-либо крупного явления теоретически невозможно. Любая формальная и неформальная классификация отражает лишь некоторую «норму» его восприятия, отвечающую текущему уровню знаний и уровню развития восприятия социумом окружающего мира. Преимущества формальных классификаций состоят только в том, что они реализуются по однозначно описанным правилам и, соответственно, воспроизводимы. Для выбора конкретного правила для конкретного случая можно дать лишь очень общие рекомендации.

8.2. Формальные основания классификации

Два множества A и B принадлежат одному классу, если их объединение $A \cup B = A$ или $A \cup B = B$.

Два множества A и B не принадлежат одному классу, если их объединение $A \cup B = m(A, B)$ есть новое множество, а пересечение $A \cap B = \varnothing$.

Если $A \cap B < \varepsilon$, т.е. меньше некоторого малого числа элементов ε, то имеются в виду нечеткие множества. Класс, тип, группа, кластер, таксон рассматриваются как тождественные понятия и, если не оговорено особо, между ними не усматривается никаких семантических различий.

Система как и во всех случаях задается элементами (конкретными наблюдениями) и множеством признаков (характеристик, переменных), каждый из которых определен на множестве собственных состояний, измеряемых или оцениваемых в ходе наблюдений. Эти множества могут быть представлены так же как в многомерном шкалировании.

Процедура классификации сводится к образованию групп (непересекающихся подмножеств) из множества элементов, которые по логическим отношениям, используемым при решении задачи классификации, можно разделить на следующие взаимодополняющие типы.

Исключающие/неисключающие. В исключающей классификации один элемент может принадлежать только одному классу (подмножеству). В неисключающей классификации один элемент может принадлежать разным классам или подмножествам. Типичный пример неисключающей классификации — ключевые слова для поиска статей. В практике экологических и географических исследований обычно применяются исключающие классификации. Неисключающие классификации типичны для поисковых систем. Может быть определена формальная задача поиска наилучшего набора «ключевых слов», обеспечивающих кратчайший путь поиска.

В экологических исследованиях неисключающие классификации практически не проводятся. Однако задача поиска наилучших индикаторов состояния среды фактически сводится к поиску признаков и таких их состояний, которые содержали бы информацию о множестве других признаков. Поиск территорий с интересующим состоянием среды по каждому отдельно взятому индикатору или их набору по смыслу тождественен поиску статей по ключевым словам.

Внутренние/внешние. Внутренние классификации подразумевают, что все признаки рассматриваются как принадлежащие одному множеству. Внешние классификации один или несколько признаков не включают в классификацию и определяют их как «внешние». Задача состоит в том, чтобы на основе только внутренних признаков получить классификацию, наилучшим образом отражающую внешние признаки. Фактически этот тип классификации соответствует представлению о направленных системах с входами и выходами.

В экологических и географических исследованиях такие классификации используются, когда некоторый признак имеет особое функциональное значение, но измерен далеко не во всех точках, когда внешний признак (признаки) рассматриваются как возможные факторы для организации множества внутренних признаков или, напротив, как их функция. Часто это признак, для которого необходимо получить прогнозные оценки по схеме «если, то». Использование таких схем достаточно типично и фактически эта логика применялась при идентификации физического смысла виртуальных координат (факторов) в факторном анализе и многомерном непараметрическом шкалировании.

Иерархические/неиерархические. Иерархия — операция включения на множестве. Система иерархически организована, если каждое множество является в принятом смысле (в рассматриваемом признаковом пространстве) подмножеством другого множества, которое в свою очередь является подмножеством множества следующего иерархического уровня и так до тех пор, пока объединение подмножеств не приводит к множеству всех элементов. В иерархической классификации классы рассматриваются попарно как возможные кандидаты на объединение. При этом, если особо не оговорено, принимается, что существует наилучшая в некотором смысле попарная схема их объединения в классы (множества) более высокого иерархического уровня.

Обычно критерием объединения служит такая конфигурация, при которой однородность классов более высокого уровня — максимальная из всех возможных. Или, наоборот, разнообразие каждого класса в принятой конфигурации для следующего иерархического уровня должно быть минимально относительно всех других возможных конфигураций. Неиерархическая классификация выделяет классы таким образом, чтобы каждый класс был максимально однороден и в идеале — дискретен.

Использование обоих вариантов классификаций достаточно типично для экологии и географии. Чаще всего иерархия строится по двоичному основанию, но в общем случае это не обязательно. Строго говоря, допущение существования иерархической классификации неявно подразумевает фрактальность природных явлений. Только одновременная реализация свойства непрерывности-разрывности допускает корректность процедуры иерархической сборки множества из его элементов. Для строго дискретных множеств иерархическая классификация неприменима. В этом случае используют неиерархическую процедуру, выделяющую предпочтительно дискретные однородные подмножества, каждое из которых может рассматриваться как самостоятельная генеральная совокупность.

Агломеративные/дивизионные. Разница состоит в альтернативном подходе к множеству. В агломеративных классификациях про-

цедура начинается от каждого элемента, который объединяется по своим свойствам с каким-либо другим (или другими) элементами, далее все пары (группы) объединяются с какой-нибудь другой парой (группой) и так далее до тех пор, пока не будет исчерпано все множество. Такую классификацию логично назвать классификацией снизу. В дивизионной классификации (классификация сверху) исходное множество элементов разделяется сверху без парного перебора на два или большее число подмножеств, каждое из которых вновь дробится на подмножества. Очевидно, что могут быть иерахически-агломеративные, иерархически-дивизионные, неирархические-агломеративные и неиерархически-дивизионные классификации.

Дивизионная классификация осуществляется любым человеком, когда очевиден общий признак классификации. Например, контролер, используя простейшие критерии, делит некоторую продукцию (детали) на две группы: кондиционная и некондиционная. Затем кондиционную группу деталей другой контролер разделяет по некоторому тесту по сортам, а в некондиционной группе по простейшим критериям отделяют детали, которые еще могут быть использованы после доработки. Дивизионную классификацию осуществляет любой географ при взгляде на космический снимок или аэрофотоснимок. Он не видит конкретных элементов изображения, но различия светлых и темных тонов на первом же уровне почти надежно позволяют ему выделить, например, лесные и безлесные территории, по структуре рисунка — горные и равнинные и т. п.

Логика агломеративной классификации может быть продемонстрирована на примере типизации изображения и выделения полигонов (однородных подмножеств территориально соседствующих точек) того же аэрофотоснимка. В этом варианте исследователь неявно определяет для себя некоторый элемент A (фрагмент изображения обычно очень небольшого размера площади) и, выбрав некоторый исходный элемент, начинает сравнивать с ним соседние элементы — B и C. Обычно одному элементу A ставятся в соответствие два расположенных на одной с ним линии B и C. По оптической плотности, текстуре и структуре сравниваемых элементов исследователь решает, к какому из элементов (A или C) ближе элемент B: если к A, то он объединяет его в подмножество AB и сравнивает его с элементами C и D. Если же он относит элемент B к C, то между A и B проводится граница, а множеству BC, ставятся в соответствие элементы D и G. Таким образом, исследователь постепенно заполнит все поле изображения и получит множество полигонов, после этого он определяет эти полигоны как новые элементы и осуществляет с ними ту же операцию классификации. Так происходит до тех пор, пока не будет исчерпано все изображение.

В идеале классификации, осуществленные по логике «снизу» и логике «сверху», должны совпадать, однако обычно этого не происходит даже при жесткой формализации всех процедур. Несовпадение определяется самой сутью большинства явлений, которые допускают существование независимых и по сути адекватных траекторий как объединения, так и разделения множеств.

Монотетические/политетические. В монотетических классификациях деление производится на основе одного признака, который часто называют ведущим. В политетической классификации признаки и/или их комбинации учитываются в равной степени. Типичной монотетической классификацией является таксономия организмов. Правда в ней часто существуют «виртуальные» классы, не имеющие таксономического статуса, что, однако, не меняет сути дела. Обычно классификации растительности, почв, ландшафтов строятся по монотетической схеме. При этом признаку, по которому выделяются более высокие иерархические уровни, придается, как более общему, большее функциональное значение. С другой стороны, все агломеративные классификации по условию политетические.

Классификации можно проводить как на основе сходства, так и на основе различий между элементами или группами. Однако в агломеративных процедурах чаще используется мера различия, т. е. расстояние (дистанция), определенное на множестве признаков или множестве элементов.

Классификация признаков (R-анализ) по содержанию адекватна расчету коэффициентов чувствительности признаков к координатам векторного пространства в факторном анализе и многомерном шкалировании. Классификация элементов (Q-анализ) по смыслу тождественна определению элементов в системе координат пространства Евклида. Разница состоит в дискретной и непрерывной форме отображения явления.

Свойства классификации и получаемый результат во многом определяются выбором метода и типом дистанции. Как и при многомерном шкалировании все отношения между парами объектов классификации представлены в форме обычной симметричной матрицы дистанций.

Пять возможных подходов без учета множества нюансов, вариантов представления исходных данных и дистанций, приводят к большому разнообразию классификационных схем. Если же принять во внимание множество критериев объединения элементов в одно подмножество или разделения их на подмножества, то возможное число классификаций становится необозримым. При некотором навыке не составляет особого труда придумать алгоритм классификации, в чем-то отличающийся от существующих, приспособленный для решения конкретной задачи для определенного объекта.

Внутри каждого из подходов существуют свои стратегии объединения, порождающие часто содержательно различные результаты. Полное изложение теории и методов кластер-анализа не входит в задачу данного пособия. Продемонстрируем результаты и объясним суть некоторых наиболее часто используемых методов. Затем обсудим некоторые аспекты качества классификации и оставим читателю возможности для совершенно необходимого самообразования.

8.3. Методы кластер-анализа

Демонстрацию и обсуждение методов кластер-анализа будем осуществлять на примере анализа состава того же древесного яруса.

Начнем с классификации переменных, сравнивая результаты по двум метрикам: Евклида и на основе гамма-корреляции.

Метод ближайшего (ближнего) соседа (метод одиночного сцепления-связи) — первый практический метод иерархической агломеративной классификации. Алгоритм метода сводится к следующему: на основе матрицы дистанций определяются два наиболее близких объекта, которые объединяются в один кластер. На следующем шаге выбирается объект, который также будет включен в этот кластер. Это объект, имеющий наименьшую дистанцию хотя бы с одним из объектов этого класса. Если существуют два объекта, которые имеют меньшую дистанцию, то они образуют отдельный независимый кластер. Процедуру такой классификации лучше рассмотреть на реальной дендрограмме. По оси ординат отложены дистанции, показывающие расстояние между объединяемыми кластерами (рис. 8.1). Обычно в программах классификации приводится специальная таблица последовательности включения объектов в кластеры (табл. 8.1).

Первый кластер образуют вяз (FRAX) и дуб (QUERCUS) при минимальной дистанции соединения 2,676, с дистанцией 2,808 к ним подсоединяется черемуха (PADUS), затем клен (ACER_PL), сосна (PINUS), рябина (SORBUS), ива (SALIX), ольха черная (ALNUS_GL). В этом месте монотонность нарушается и весь этот кластер присоединяется к ильму (ULMUS), так как последний ближе к нему, а ольха черная находится от всех остальных на существенно большем расстоянии. Далее к кластеру последовательно подсоединяется серая ольха (ALNUS_IN), липа (TILIA), осина (POPULUS). Береза (BETULA) и ель (PICEA) оказываются ближе друг к другу, чем к любому объекту большого кластера, в результате чего они образуют изолированный кластер. Очевидно, что логика классификации отражает не более как общее обилие видов на трансекте. Соединения начинаются с наиболее редких ясеня и дуба и заканчиваются наиболее многочисленными березой и елью. Вполне понятно, что ничего другого при использовании дистанции

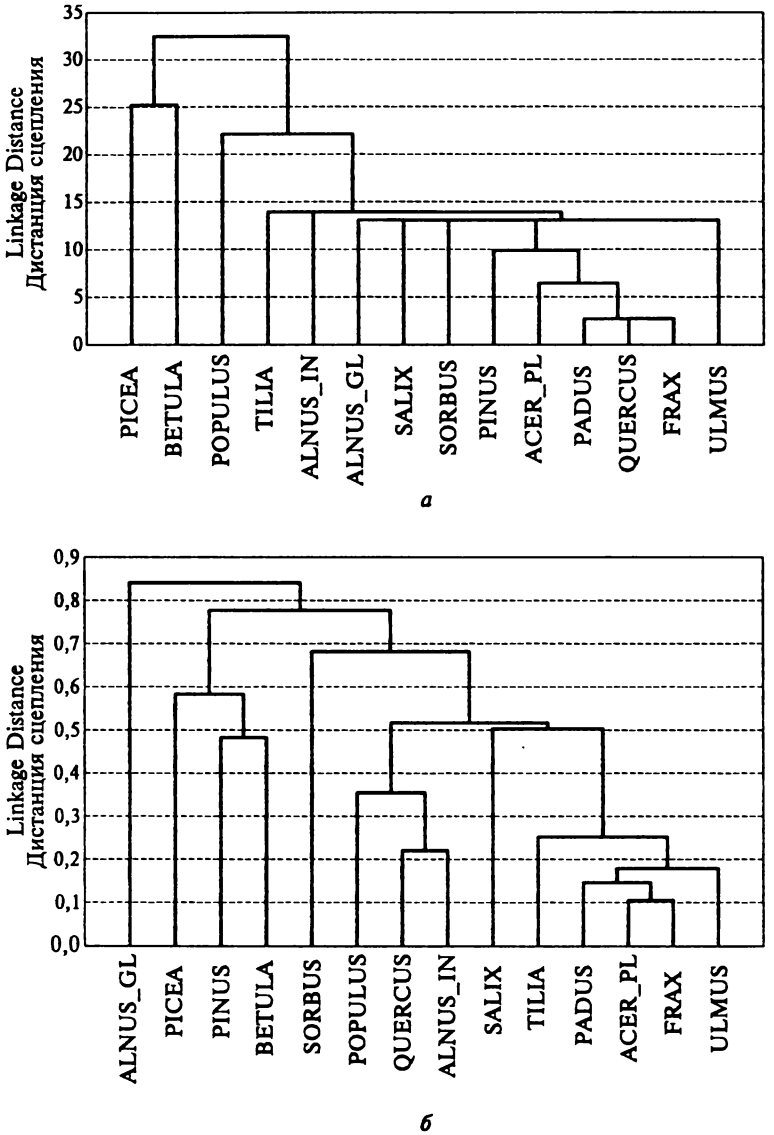


Рис. 8.1. Дендрограмма классификации видов методом ближайшего соседа. Одиночное сцепление (Single linkage):

а — дистанция Евклида; *б* — гамма-дистанция

Евклида, отражающей в первую очередь различия по обилию, ожидать не приходится.

Как видно из рис. 8.1, классификация в метрике на основе гамма-корреляции дает существенно иной результат (табл. 8.2).

Совершенно иначе происходит объединение видов в классе в пространстве с метрикой на основе меры подобия.

Наиболее подобны в парном территориальном размещении ясень и клен, затем с ними объединяется черемуха и все три вида присоединяются к вязу, образуя вместе относительно замкнутую группу. Через шаг к ним присоединяется липа. Но подобие в размещении серой ольхи и дуба оказалось существенно больше, чем подобие в размещении липы и других широколиственных пород, что и нарушило монотонность. В результате образовалась компактная группа широколиственных пород, включившая в себя дополнительно черемуху, а на следующем шаге — группа из ольхи серой, дуба и осины. Монотонность вновь нарушается относительно большим подобием в размещении березы и сосны. Далее группа широколиственных пород объединяется с ивой и на следующем шаге — с серой ольхой, дубом и осиной. Монотонность вновь нарушается классом «бореальных видов»: береза, сосна, ель. Затем к классу широколиственных пород, ольхи серой и осины, подсоединяется рябина. После этого «широколиственный» и «бореальный» классы объединяются и на самом последнем шаге к ним подсоединяется наиболее своеобразно распределенная в пространстве ольха черная. Совершенно очевидно, что эта классификация дает общее отображение тех отношений, которые в несколько иной форме были получены в ординации и которые, безусловно, более содержательны, чем отношения, найденные на основе метрики Евклида. В связи с этим следующие методы кластер-анализа будем рассматривать только для метрики гамма-корреляции.

Метод наиболее удаленного (дальнего) соседа (метод полного сцепления-связи — Complete Linkage). В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. «наиболее удаленными соседями»).

Суть этого метода хорошо видна из рис. 8.2, *а* и табл. 8.3 последовательности сцеплений.

На всем множестве дистанций расстояние между двумя группами определяется как расстояние между двумя самыми удаленными представителями двух групп. Алгоритм выбирает пару самых близких элементов и фиксирует их как первый кластер. В данном случае, как и в методе ближайшего соседа, им оказались ясень и клен. Затем он просматривает все парные дистанции и образует новый кластер в том случае, если дистанции между элементами меньше, чем со всеми другими и с элементами выделенного первого класса, т.е. в отличие от метода ближайшего соседа, когда каждый элемент сравнивается только с первым уже выделенным классом против всех остальных, здесь оцениваются дистанции пар. В данном случае таким вторым классом оказались серая ольха и дуб — комбинация, известная нам и из ординации, и из метода

Таблица 8.1

Последовательность сцепления (включения) объектов в кластеры (Amalgamation Schedule) на основе метрики Евклида методом ближайшего соседа (одиночного сцепления)

Дистанция соединения	Номер объекта						
	1	2	3	4	5	6	7
2,676074	FRAX	QUERCUS					
2,808364	FRAX	QUERCUS	PADUS				
6,454337	FRAX	QUERCUS	PADUS	ACER_PL			
9,885604	FRAX	QUERCUS	PADUS	ACER_PL	PINUS		
13,04334	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS	
13,04526	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS	SALIX
13,15728	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS	SALIX
13,21825	**ULMUS**	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS
13,89546	**ULMUS**	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS
13,92286	**ULMUS**	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS
22,11737	**ULMUS**	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS
25,24111	**BETULA**	**PICEA**					
32,47277	ULMUS	FRAX	QUERCUS	PADUS	ACER_PL	PINUS	SORBUS

Окончание табл. 8.1

Дистанция соединения	Номер объекта						
	8	9	10	11	12	13	14
2,676074							
2,808364							
6,454337							
9,885604							
13,04334							
13,04526							
13,15728	ALNUS_GL						
13,21825	SALIX	ALNUS_GL					
13,89546	SALIX	ALNUS_GL	ALNUS_IN				
13,92286	SALIX	ALNUS_GL	ALNUS_IN	TILIA			
22,11737	SALIX	ALNUS_GL	ALNUS_IN	TILIA	POPULUS		
25,24111							
32,47277	SALIX	ALNUS_GL	ALNUS_IN	TILIA	POPULUS	BETULA	PICEA

Таблица 8.2

Последовательность сцепления (включения) объектов методом ближнего соседа в кластеры (Amalgamation Schedule) на основе метрики гамма-корреляции методом ближайшего соседа (одиночного сцепления)

Дистанция соединения	Номер объекта						
	1	2	3	4	5	6	7
0,1051005	FRAX	ACER_PL					
0,1461596	FRAX	ACER_PL	PADUS				
0,1787642	**ULMUS**	FRAX	ACER_PL	PADUS			
0,2205882	**ALNUS_IN**	**QUERCUS**					
0,2520385	ULMUS	FRAX	ACER_PL	PADUS	TILIA		
0,3546798	ALNUS_IN	QUERCUS	POPULUS				
0,4822592	**BETULA**	**PINUS**					
0,5027581	ULMUS	FRAX	ACER_PL	PADUS	TILIA	SALIX	
0,5162602	ULMUS	FRAX	ACER_PL	PADUS	TILIA	SALIX	ALNUS_IN
0,5827176	**BETULA**	**PINUS**	**PICEA**				
0,6816767	ULMUS	FRAX	ACER_PL	PADUS	TILIA	SALIX	ALNUS_IN
0,7771215	ULMUS	FRAX	ACER_PL	PADUS	TILIA	SALIX	ALNUS_IN
0,8409574	ULMUS	FRAX	ACER_PL	PADUS	TILIA	SALIX	ALNUS_IN

Окончание табл. 8.2

Дистанция соединения	Номер объекта						
	8	9	10	11	12	13	14
0,1051005							
0,1461596							
0,1787642							
0,2205882							
0,2520385							
0,3546798							
0,4822592							
0,5027581							
0,5162602	QUERCUS	POPULUS					
0,5827176							
0,6816767	QUERCUS	POPULUS	SORBUS				
0,7771215	QUERCUS	POPULUS	SORBUS	BETULA	PINUS	PICEA	
0,8409574	QUERCUS	POPULUS	SORBUS	BETULA	PINUS	PICEA	ALNUS_GL

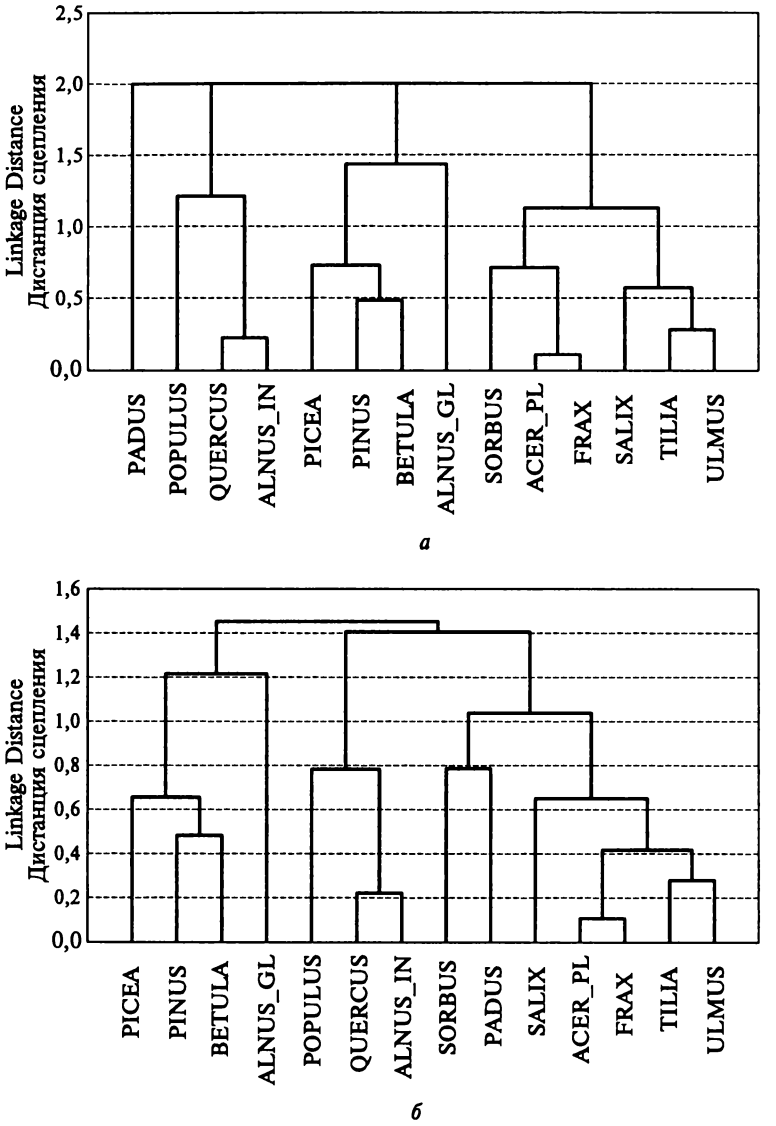


Рис. 8.2. Методы классификации по дистанции на основе гамма-корреляции:

а — наиболее удаленного соседа; *б* — невзвешенного попарного среднего

ближайшего соседа. В результате на первом уровне выделяются четыре группы, элементы которых наиболее удалены друг от друга, образующие в данном случае вполне естественные объединения. После того как расстояния между любой оставшейся парой оказываются больше, чем расстояние хотя бы одного элемента от эле-

ментов выделенных кластеров, начинается выделение кластеров по три. Когда возможности такого объединения исчерпаны, объединения берутся по четыре и т.д. В результате получаем компактно выделяемые собственно широколиственные породы с примкнувшими к ним ивой и рябиной, класс серой ольхи, осины, дуба и группу черной ольхи, березы, сосны и ели. Принципиальным меняющим смысл классификации в сравнении с первым методом является объединение черной ольхи с бореальными видами и объединение их на следующем шаге с широколиственными видами, а уже затем с серой ольхой, осиной и дубом. В данной схеме черемуха оказалась самым удаленным элементом и объединилась с остальными лишь на самом последнем шаге.

Сравнивая два метода, трудно определить, какой из них лучше — и один, и другой подчеркивают несколько различные, но вполне реалистичные аспекты пространственных отношений. Часто отмечают, что метод ближнего соседа стягивает пространство, а метод дальнего соседа — растягивает. Считается, что метод дальнего соседа работает очень хорошо, когда объекты происходят на самом деле из различных реально существующих достаточно дискретных компактных групп. Если же объекты в многомерном пространстве образуют эллипсовидные подмножества или цепочки, то он сильно искажает отношения. Напротив, метод ближайшего соседа рекомендуют применять в тех случаях, когда имеется предположение о существовании «волокнистых» или цепочечных структур по схеме «самый близкий родственник, близкий родственник» и т.д.

Метод невзвешенного попарного среднего (метод средней связи — Unweighted pair-group average). Если в рассмотренных выше методах учитывались только попарные дистанции и с ними не осуществлялось никаких преобразований, то в этом методе, после образования класса из двух и более элементов, сравнение его с остальными элементами осуществляется по среднему значению дистанций между образующими его элементами.

При этом принцип объединения в кластеры подобен методу дальнего соседа. Из рис. 8.2 и табл. 8.4 видно, что, как и во всех иерархических агломеративных методах, формирование кластеров начинается от наиболее подобных элементов (ясеня и клена). Затем выделяются те же три класса, что и в методе дальнего соседа и образуется класс строго широколиственных пород, береза объединяется с сосной и т.д. Все объединения почти полностью повторяют метод дальнего соседа с той лишь разницей, что виды с относительно неопределенным положением (черемуха и рябина) образуют самостоятельный элементарный кластер. Обычно этот метод признается эффективным практически для любых типов конфигураций элементов множества в многомерном пространстве.

Метод взвешенного попарного среднего (Weighted pair-group average) идентичен методу невзвешенного попарного среднего, за

Таблица 8.3

Последовательность сцепления (включения) объектов методом дальнего соседа в кластеры (Amalgamation Schedule) на основе метрики гамма-корреляции

Дистанция соединения	Номер объекта						
	1	2	3	4	5	6	7
0,1051005	FRAX	ACER_PL					
0,2205882	ALNUS_IN	QUERCUS					
0,2784110	ULMUS	TILIA					
0,4822592	BETULA	PINUS					
0,5715525	ULMUS	TILIA	SALIX				
0,7115010	FRAX	ACER_PL	SORBUS				
0,7274970	BETULA	PINUS	PICEA				
1,127642	ULMUS	TILIA	SALIX	FRAX	ACER_PL	SORBUS	
1,210117	ALNUS_IN	QUERCUS	POPULUS				
1,434345	ALNUS_GL	BETULA	PINUS	PICEA			
2,000000	ULMUS	TILIA	SALIX	FRAX	ACER_PL	SORBUS	ALNUS_GL
2,000000	ULMUS	TILIA	SALIX	FRAX	ACER_PL	SORBUS	ALNUS_GL
2,000000	ULMUS	TILIA	SALIX	FRAX	ACER_PL	SORBUS	ALNUS_GL

Окончание табл. 8.3

Дистанция соединения	Номер объекта						
	8	9	10	11	12	13	14
0,1051005							
0,2205882							
0,2784110							
0,4822592							
0,5715525							
0,7115010							
0,7274970							
1,127642							
1,210117							
1,434345							
2,000000	BETULA	PINUS	PICEA				
2,000000	BETULA	PINUS	PICEA	ALNUS_IN	QUERCUS	POPULUS	
2,000000	BETULA	PINUS	PICEA	ALNUS_IN	QUERCUS	POPULUS	PADUS

Таблица 8.4

Последовательность сцепления (включения) объектов методом невзвешенного попарного среднего в кластеры (Amalgamation Schedule) на основе метрики гамма-корреляции

Дистанция соединения	Номер объекта						
	1	2	3	4	5	6	7
0,1051005	FRAX	ACER_PL					
0,2205882	ALNUS_IN	QUERCUS					
0,2784110	ULMUS	TILIA					
0,4166946	ULMUS	TILIA	FRAX	ACER_PL			
0,4822592	BETULA	PINUS					
0,6503198	ULMUS	TILIA	FRAX	ACER_PL	SALIX		
0,6551073	BETULA	PINUS	PICEA				
0,7823983	ALNUS_IN	QUERCUS	POPULUS				
0,7854730	PADUS	SORBUS					
1,038050	ULMUS	TILIA	FRAX	ACER_PL	SALIX	PADUS	SORBUS
1,215232	ALNUS_GL	BETULA	PINUS	PICEA			
1,405929	ULMUS	TILIA	FRAX	ACER_PL	SALIX	PADUS	SORBUS
1,453105	ULMUS	TILIA	FRAX	ACER_PL	SALIX	PADUS	SORBUS

Окончание табл. 8.4

Дистанция соединения	Номер объекта						
	8	9	10	11	12	13	14
0,1051005							
0,2205882							
0,2784110							
0,4166946							
0,4822592							
0,6503198							
0,6551073							
0,7823983							
0,7854730							
1,038050							
1,215232							
1,405929	ALNUS_IN	QUERCUS	POPULUS				
1,453105	ALNUS_IN	QUERCUS	POPULUS	ALNUS_GL	BETULA	PINUS	PICEA

исключением того, что при вычислениях размер соответствующих кластеров (число объектов, содержащихся в них) используется в качестве весового коэффициента. Этот метод применяют при классификации большого числа объектов. В нашем случае оба метода дают тождественный результат.

Невзвешенный центроидный метод (Unweighted pair-group centroid). В этом методе расстояние между двумя кластерами определяется как расстояние между их геометрическими центрами тяжести. Не останавливаясь на деталях, отметим, что в этом методе черная ольха, как и в методе «дальнего соседа», занимает особое положение (рис. 8.3, *а*). Основные ядра кластеров остаются неизменными, а виды с неопределенным положением (черемуха, рябина, ива) присоединяются к кластеру широколиственных пород.

Метод Варда (Ward's method). Этот метод отличается от всех других, так как он использует при оценке расстояний между кластерами элементы дисперсионного анализа, т. е. второй момент распределения.

На первом шаге предполагается, что каждый кластер состоит из одного объекта. Как и в рассмотренных выше методах, первоначально в класс объединяют два ближайших элемента. Далее группируются объекты, которые минимизируют сумму квадратов дистанций для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге.

На каждом шаге элементы объединяются таким образом, чтобы приращение внутри кластерной дисперсии дистанций V_k было бы минимальным, т. е. на каждом шаге отбираются наиболее «плотные» группы элементов:

$$V_k = \sum_{i=1}^{n_k} (d_i - \bar{d}_k)^2 ,$$

где $\sum (d_i - \bar{d}_k)^2$ — сумма квадратов отклонений дистанции i-го элемента от средней дистанции в кластере k; n_k — число элементов в кластере k.

То, что алгоритм использует вторые моменты и автоматически информацию о среднем делает этот метод более эффективным чем методы, опирающиеся только на средние и медианы. Однако по логическим основаниям он независим от методов ближайшего и дальнего соседа. Каждый из рассмотренных методов лучше отражает несколько различные свойства системы.

На рис. 8.3, *б* и в табл. 8.5 приведены результаты классификации методом Варда. Дерево классификации в данном случае полностью подобно полученному методом невзвешенного попарного среднего, но изображение дерева выглядит более компактным. Такое совпадение совершенно не обязательно. Довольно часто метод Варда

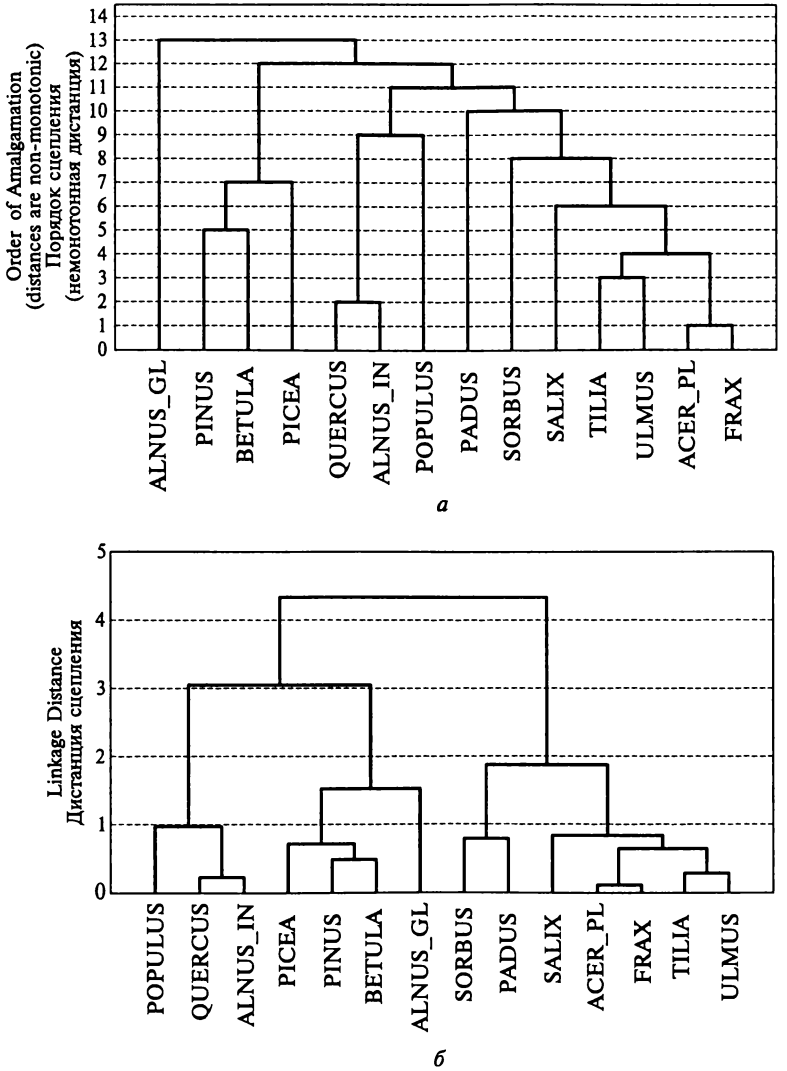


Рис. 8.3. Методы классификации по дистанции на основе гамма-корреляции:
а — невзвешенный центроидный; *б* — Варда

дает результаты, подобные полученным по методу дальнего соседа. Все в конечном итоге определяется конфигурацией данных.

Внимательный читатель, сравнивая результаты, полученные разными методами алгоритмизации элементов, и помня результаты многомерного непараметрического шкалирования, показавшего, что распределение видов в пространстве фактически образуют трехмерный континуум, заметил, что устойчивые комбинации об-

Таблица 8.5

Последовательность сцепления (включения) объектов методом Варда (Ward's method) в кластеры (Amalgamation Schedule) на основе метрики гамма-корреляции

Дистанция соединения	Номер объекта						
	1	2	3	4	5	6	7
0,1051005	FRAX	ACER_PL					
0,2205882	ALNUS_IN	QUERCUS					
0,2784110	ULMUS	TILIA					
0,4822592	BETULA	PINUS					
0,6416336	ULMUS	TILIA	FRAX	ACER_PL			
0,7127233	BETULA	PINUS	PICEA				
0,7854730	PADUS	SORBUS					
0,8354828	ULMUS	TILIA	FRAX	ACER_PL	SALIX		
0,9696683	ALNUS_IN	QUERCUS	POPULUS				
1,524102	ALNUS_GL	BETULA	PINUS	PICEA			
1,873197	ULMUS	TILIA	FRAX	ACER_PL	SALIX	PADUS	SORBUS
3,048606	ALNUS_GL	BETULA	PINUS	PICEA	ALNUS_IN	QUERCUS	POPULUS
4,335587	ULMUS	TILIA	FRAX	ACER_PL	SALIX	PADUS	SORBUS

Окончание табл. 8.5

Дистанция соединения	Номер объекта						
	8	9	10	11	12	13	14
0,1051005							
0,2205882							
0,2784110							
0,4822592							
0,6416336							
0,7127233							
0,7854730							
0,8354828							
0,9696683							
1,524102							
1,873197							
3,048606							
4,335587	ALNUS_GL	BETULA	PINUS	PICEA	ALNUS_IN	QUERCUS	POPULUS

разуют элементы, занимающие наиболее крайние положения в матрице коэффициентов чувствительности по какой-либо одной координате: ясень, клен и вяз с липой — отрицательная область первой координаты векторного пространства, серая ольха и дуб — положительная область второй координаты. Если виды зависят от многих факторов или сильно изолированы (ольха черная), то их положение становится менее устойчиво и они с высоким значением дистанции сцепления входят в разных методах в кластеры с различными «устойчиво ассоциирующими» группами видов.

В рамках методов иерархической алгоритмизации осуществляется попытка упорядочить элементы из многомерного пространства в линию, что с формальных позиций без искажений отношений не реализуемо. Единственно, к чему можно стремиться — это уменьшить масштаб таких искажений. С этих позиций наиболее эффективным можно считать метод Варда, однако и здесь объединениям на высоких уровнях не следует придавать особого значения.

С другой стороны, если провести классификацию разными методами и выделить устойчивые комбинации элементов, не зависящих от метода отображения, то можно с высокой степенью достоверности утверждать, что они экологически действительно наиболее близки и с наибольшей вероятностью будут встречаться в пространстве совместно.

Однако такое толкование прямо связано с идеологией экологической ординации.

Более общую трактовку можно пояснить на примере классификации переменных почвенной системы (рис. 8.4).

Следует отметить, что если распределения близки к нормальным, то классификация по дистанции Евклида для стандартизованных данных и дистанции по метрике коэффициента корреляции Пирсона тождественны.

На рис. 8.4 приведена классификация, хорошо соответствующая результатам многомерного анализа и физическому смыслу. Практически каждый тип переменной имеет свою ветвь кластера. Исключением являются кальций и магний. Эти элементы образуют два отдельных блока, слои 4 и 5 (30 и 40 см) и верхние слои. При этом верхние слои ближе к кластеру «кислотности», чем к нижним слоям с одноименными переменными. Та же неоднозначная схема поведения в пространстве катионов кальция и магния отражалась и в многомерном анализе. Очень малые значения дистанций сцепления на высоком уровне заставляют полагать, что почти наверное блоки «калий, фосфор, кальций и магний в нижних слоях», «кальций и магний в верхних слоях вместе с кислотностью» и блок «влажности» практически независимы. Структура кластеров на нижнем уровне показывает, что у разных почвенных переменных третий слой (20 см), соответствующий в среднем под-

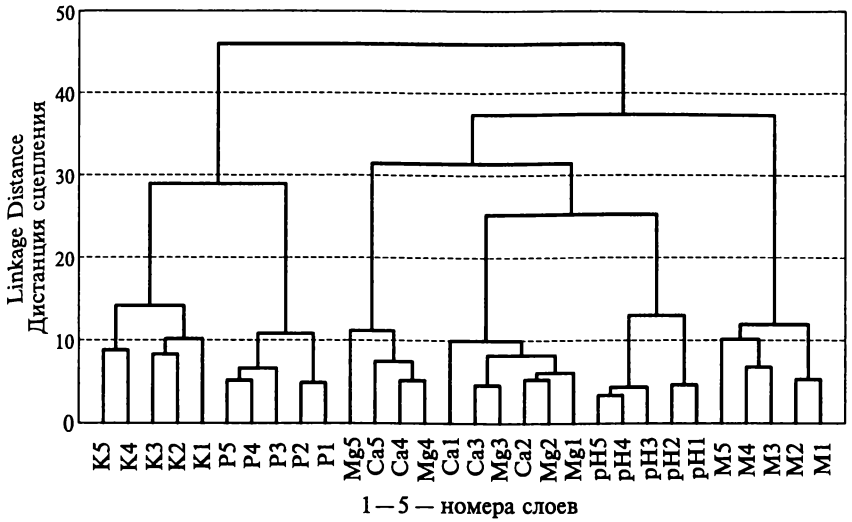


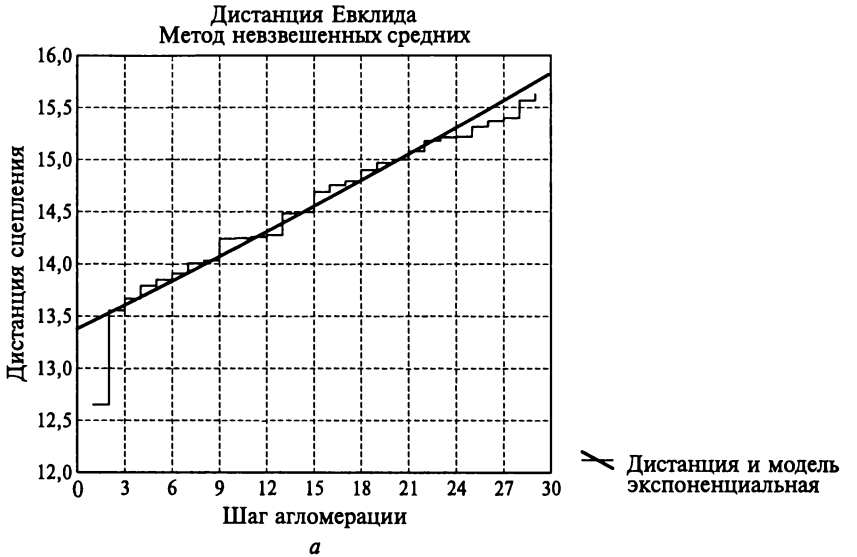
Рис. 8.4. Дендрограмма классификации почвенной системы (влажность, pH, обменные основания) по методу Варда. Дистанция Евклида, стандартизованные данные

золистому горизонту присоединяется или к двум верхним или к двум нижним слоям, что, скорее всего, отображает особенности миграции различных элементов, или как для pH — комплексное влияние многих переменных и, возможно, в первую очередь особенностей миграции гуминовых кислот.

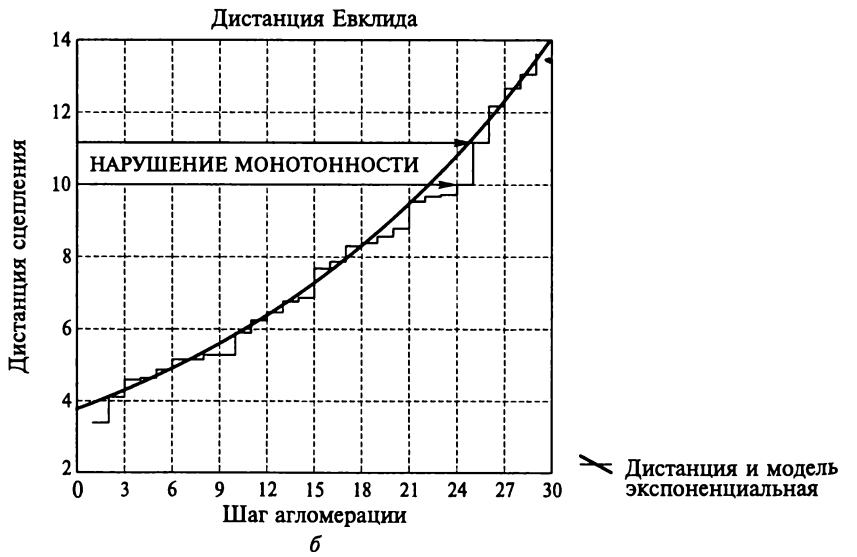
Этот пример показывает, что в общем случае кластер-анализ допускает трактовку отдельных кластеров как подсистем различного иерархического уровня большой системы. Чем выше уровни сцепления, тем более независимы подсистемы. Если бы система была одномерна, то кластеры соответствовали бы подсистемам с наиболее тождественной реакцией на внешнее воздействие, а наиболее удаленные реагировали бы на внешнее воздействие с разным знаком. Но такие простые отношения в реальности практически невозможны.

То, что упорядочивание элементов в кластеры методом иерархической агломерации осуществляется из многомерного пространства, позволяет использовать этот метод для независимой оценки целочисленной размерности системы. Если классификация осуществляется для независимых случайно распределенных переменных, то функция: «дистанция сцепления — шаг агломерации» должна быть монотонна, а для метода невзвешенного среднего при метрике Евклида — линейна. Линейность очевидным образом определяется использованием в процедуре агломерации средних значений. На рис. 8.5 сравнивается вид функций для модели (рис. 8.5, *а*) и

реальных данных (рис. 8.5, *б*). Наличие структурной организации в реальных данных отражает нелинейность зависимости «дистанция сцепления — шаг агломерации». На уровне дистанции сцепления около 10 монотонность функции резко нарушается. Эту дистанцию



а



б

Рис. 8.5. Оценка размерности пространства на основе кластер-анализа:
а — независимые переменные, модель случайного процесса; *б* — почвенные переменные

можно считать границей, определяющей число классов, объекты которых независимы друг от друга. На рис. 8.6 показано, где лежит эта граница. Линия, соответствующая этой дистанции, пересекает шесть ветвей дендрита. Соответственно, с учетом числа степеней свободы (из одномерного пространства всегда можно получить два класса) размерность пространства может быть оценена равной пяти.

По методу многомерного шкалирования оценка размерности была принята равной четырем. Если учесть относительно близкое положение к границе дистанции сцепления для переменных «калий», то результаты можно считать тождественными. Таким образом, используя метод невзвешенного среднего кластер-анализа, можно получить дополнительную оценку самого важного параметра многомерной системы — размерности.

Метод k-средних относится к дивизионным методам классификации (классификация сверху). В этом методе число желаемых k-классов определяет сам исследователь. На первом шаге алгоритм выбирает из всего множества k-точек, каждая из которых описывается n-переменными. Выбор точек может осуществляться по различным правилам. Чаще всего используется правило случайного выбора, или k-первых, или k-точек с максимальными дистанциями друг от друга. Последний метод при очень большом числе элементов требует больших затрат машинного времени, так как алгоритм должен несколько раз сравнить все пары дистанций. В зависимости от версии используются различные метрики дистанций, но чаще всего дистанция Евклида.

После того как выбраны k-элементов, все остальные последовательно сравниваются по дистанции с каждым из k и присоеди-

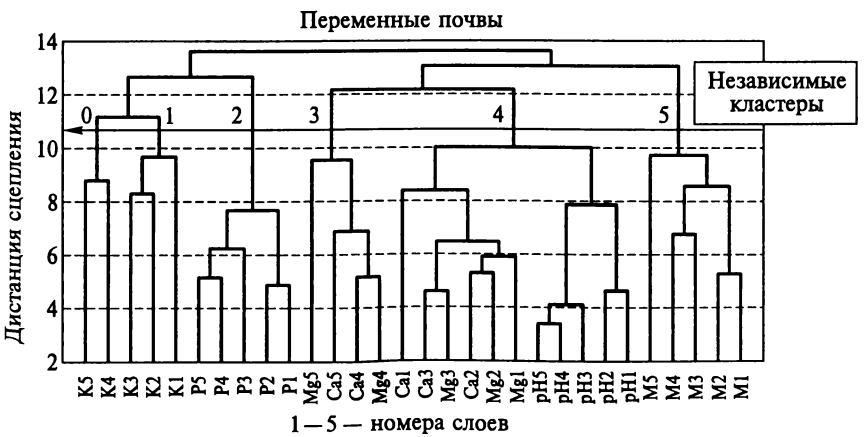


Рис. 8.6. Оценка размерности на основе кластер-анализа при методе невзвешенного среднего для свойств почвы в сравнении с результатами рис. 8.5

няются к тому, к которому оказываются ближе, образуя k-кластеров с различным числом элементов. Для каждого класса рассчитывают средние значения переменных и фиктивную точку, соответствующую центру их тяжести. Эти точки принимаются в качестве «новых». Вся процедура повторяется по отношении к этим новым точкам. Образуется новые сочетания точек, для которых вновь рассчитывают среднее и т.д. Такая процедура в конечном итоге сводится к некоторым наиболее устойчивым k-центрам тяжести, вокруг которых максимально плотно группируются все точки. Процедура их поиска может останавливаться по различным критериям. Например, если после некоторой по счету итерации элементы, входящие в класс, не меняются или расстояния между центрами тяжести в i и $(i+1)$-й итерации отличаются на величину, во много раз меньшую точности измерения переменных, или вообще не меняются, или не происходит изменение ковариационных матриц в каждом классе. В некоторых версиях алгоритма средние рассчитываются не после окончания полной сортировки точек в ходе конкретной итерации, а сразу же после ее присоединения к какому-либо классу. Результаты этих двух версий обычно несколько отличаются. Не всегда совпадают результаты классификации при различных стартовых конфигурациях. Эти несовпадения возникают из-за того, что в многомерном пространстве сгущение может быть относительно плотным как по одной координате, так и по ортогональной к ней. В зависимости от того, куда попали стартовые точки, кластеры могут собрать подмножества из разных ортогональных подпространств. Хотя такие ситуации встречаются очень редко, но их все-таки надо иметь в виду, и повторять классификацию одного и того же множества несколько раз. Если во всех случаях получены тождественные отображения, то результат можно признать успешным. В противном случае полезно разобраться в природе несоответствия. Дело в том, что каждая из этих двух несовпадающих классификаций имеет физический смысл.

После того как получены классы в первом приближении, качество классификации и ее физический смысл можно оценивать методом одномерного дисперсионного анализа.

Одной из общих проблем применения метода k-средних является выбор числа классов. Если тем или иным методом осуществлена оценка размерности, то число классов должно быть равно $d+1$, где d — размерность. При недостаточном числе классов можно оценить размерность каждого из выделенных подмножеств и для каждого отдельно повторить классификацию.

И иерархические, и дивизионные классификации можно применять как к переменным (R-анализ), так и к элементам (Q-анализ).

Рассмотрим применение метода k-средних в рамках Q-анализа для классификации на примере данных по древесному ярусу, т.е.

по отношению к элементам. Так как размерность была принята равной трем, классов должно быть четыре ($k = 4$).

На рис. 8.7 показаны характеристики выделенных классов и их распределение в пространстве. В табл. 8.6 даны самые общие оценки вклада переменных в разделение классов, полученные на основе дисперсионного анализа, при различных методах кластерного анализа. Первый метод минимизирует дисперсию в классах на основе метрики Евклида. Второй — при той же метрике ищет оптимальную конфигурацию на основе множества случайных независимых стартов при различном числе задаваемых классов с приведением их в конечном итоге к четырем. Третий метод минимизирует многомерную площадь (объем) кластера и потому отчасти учитывает коррелируемость переменных.

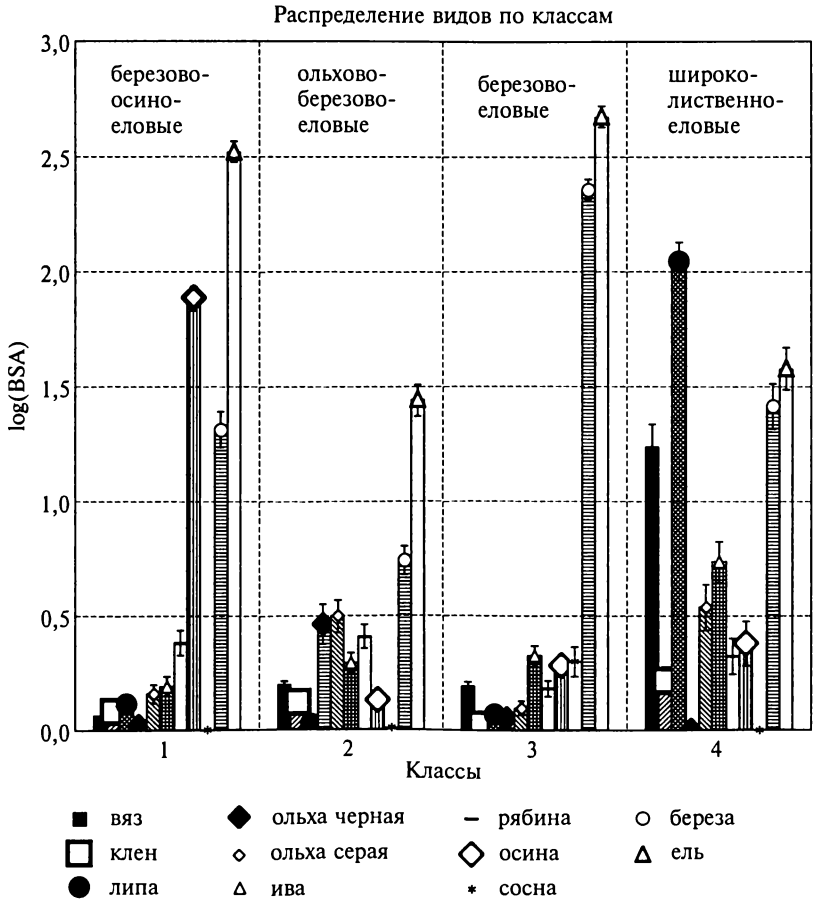


Рис. 8.7. Классификация древесного полога по составу методом k-средних (дистанция Евклида)

Из табл. 8.6 следует, что три метода классификации существенно различаются по вкладу видов в разделение классов. В целом они примерно одинаково полно отражаются в экологическом пространстве, но различаются вкладом координат в это разделение (отметим, что классификация точек наблюдения построена только на основе сумм площадей сечений, представленных на каждой из них видов). Качество третьего метода классификации в целом в соответствии с F-критерием наиболее высокое, причем для его почти полного отображения достаточно всего двух первых координат экологического пространства. Судя по значениям F-критерия, первый метод классификации наиболее «равномерно» отображает свойства экологического пространства.

В табл. 8.7 приведены средние значения логарифмов сумм площадей сечений в каждом классе. Названия классов даны в соответствии с традицией: первый вид наименее обильный из всех наиболее важных для выделения класса, последний — наиболее обильный.

Первые две классификации более близки друг другу и принципиально отличаются от третьей. В третьей классификации выделяются в отдельный класс леса с господством черной ольхи и леса со значительным участием сосны. Здесь же достаточно надежно выделяются леса с большим участием широколиственных пород, которые четко отличаются от осиново-березово-еловых лесов. Первая классификация в большей степени строится на различии участия в классах наиболее широко распространенных видов: ели, березы, осины и липы. Вторая — кроме липы включает в классификацию вяз, клен и ясень, в результате чего возникает три класса с разным участием широколиственных пород. Первая классификация, учитывающая в основном господствующие виды, разбивает множество на подмножества с не очень сильно различающимся числом элементов и, соответственно, показывает относительно высокое разнообразие лесов. Вторая классификация демонстрирует существенно меньшее разнообразие, а третья — выделяет огромный по объему класс собственно мелколиственно-еловых лесов, противопоставляя им очень небольшие по объему классов основном с широколиственными видами, господством черной ольхи и со значительным участием сосны. Несмотря на то что эта классификация в наилучшей степени минимизирует отношение внутригрупповой и межгрупповой дисперсии, она дает очень невысокое разнообразие, что не может считаться ее достоинством.

Число классов определялось исходя из минимально допустимой размерности пространства. Большая размерность пространства, возможно, приведет, с одной стороны, к выделению черноольховых и сосновых лесов по первой и второй классификациям, с другой — к детализации мелколиственно-еловых по третьей классификации. Однако проблемы несоответствия классификаций, пост-

Таблица 8.6

Одномерный дисперсионный анализ соотношения классов с древесными породами

Вид и координаты пространства Евклида	Первый метод (дистанция Евклида)		Второй метод (случайный поиск наилучшей конфигурации)		Третий метод (минимизация многомерной площади кластера)	
	F-критерий	p-уровень	F-критерий	p-уровень	F-критерий	p-уровень
Ильм	77,3008	0,000000	**175,5089**	0,000000	89,082	0,000000
Ясень	5,2364	0,001473	17,7737	0,000000	3,867	0,009477
Клен	5,5308	0,000985	**261,5895**	0,000000	4,208	0,005974
Липа	**537,3240**	0,000000	**165,0375**	0,000000	**468,551**	0,000000
Дуб	0,9294	0,426422	0,8968	0,442768	0,109	0,954787
Ольха черная	16,9812	0,000000	24,5037	0,000000	**1951,151**	0,000000
Ольха серая	15,7198	0,000000	28,2392	0,000000	8,574	0,000015
Ива	13,2313	0,000000	24,4167	0,000000	19,226	0,000000
Черемуха	1,0574	0,367011	36,5651	0,000000	0,191	0,902580
Рябина	4,8257	0,002579	1,5998	0,188765	2,395	0,067797
Осина	**326,3446**	0,000000	27,6353	0,000000	9,340	0,000005
Сосна	13,4665	0,000000	5,5438	0,000968	**891,592**	0,000000
Береза	127,8612	0,000000	38,5134	0,000000	9,335	0,000005
Ель	116,6322	0,000000	**128,7347**	0,000000	21,970	0,000000
Сумма площадей сечений	108,3670	0,000000	83,0061	0,000000	16,873	0,000000
Координата 1	**224,7230**	0,000000	**187,3435**	0,000000	**187,142**	0,000000
Координата 2	58,7392	0,000000	42,4477	0,000000	**103,176**	0,000000
Координата 3	**189,1694**	0,000000	14,2831	0,000000	6,764	0,000182
F-критерий	82,313	0,000000	76,0666	0,000000	214,92	0,00000

Примечание. Полужирным шрифтом выделены переменные, осуществляющие наибольший вклад в разделение множества на классы.

Таблица 8.7

Характеристика классов трех различных версий классификации методом *k*-средних

Переменная	Классификация на основе							
	дистанции Евклида				минимизации случайного поиска лучшей конфигурации			
	Класс 1 (N = 114)		Класс 2 (N = 129)		Класс 3 (N = 139)		Класс 4 (N = 50)	
	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка
Вяз	0,041	0,018	0,174	0,040	0,167	0,041	**1,215**	0,121
Ясень	0,000	0,000	0,009	0,009	0,000	0,000	0,075	0,046
Клен	0,081	0,027	0,121	0,032	0,028	0,013	0,218	0,061
Липа	0,114	0,035	0,025	0,012	0,072	0,024	**2,046**	0,083
Дуб	0,006	0,006	0,000	0,000	0,000	0,000	0,000	0,000
Ольха черная	0,019	0,019	**0,464**	0,087	0,049	0,026	0,000	0,000
Ольха серая	0,160	0,040	**0,500**	0,070	0,098	0,030	0,538	0,100
Ива	0,194	0,042	0,295	0,044	0,325	0,045	**0,736**	0,089
Черемуха	0,000	0,000	0,026	0,016	0,018	0,013	0,000	0,000
Рябина	0,383	0,055	0,410	0,052	0,183	0,035	0,325	0,079
Осина	**1,884**	0,052	0,131	0,029	0,283	0,039	0,380	0,097
Сосна	0,020	0,015	0,005	0,005	0,301	0,065	0,000	0,000
Береза	**1,314**	0,077	**0,742**	0,062	**2,359**	0,044	**1,414**	0,099
Ель	**2,522**	0,045	**1,440**	0,069	**2,676**	0,046	**1,579**	0,092
Название класса	Березово-осиново-еловые		Ольхово-березово-еловые		Березово-еловые		Вязово-березово-елово-липовые	

Переменная	Классификация на основе															
	минимизации случайного поиска лучшей конфигурации								минимизации многомерного объема кластеров							
	Класс 1 (N = 22)		Класс 2 (N = 60)		Класс 3 (N = 233)		Класс 4 (N = 117)		Класс 1 (N = 325)		Класс 2 (N = 62)		Класс 3 (N = 26)		Класс 4 (N = 19)	
	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка	среднее	ошибка
Вяз	0,394	0,137	**1,337**	0,102	0,050	0,015	0,092	0,028	0,119	0,021	**1,151**	0,108	0,042	0,042	0,000	0,000
Ясень	0,188	0,108	0,012	0,012	0,000	0,000	0,000	0,000	0,003	0,003	0,060	0,037	0,000	0,000	0,000	0,000
Клен	1,141	0,103	0,104	0,032	0,015	0,007	0,041	0,015	0,075	0,015	0,212	0,054	0,089	0,066	0,000	0,000
Липа	0,528	0,196	**1,571**	0,128	0,080	0,020	0,033	0,015	0,043	0,011	**1,821**	0,100	0,053	0,037	0,000	0,000
Дуб	0,000	0,000	0,000	0,000	0,000	0,000	0,006	0,006	0,002	0,002	0,000	0,000	0,000	0,000	0,000	0,000
Ольха черная	0,050	0,050	0,000	0,000	0,018	0,011	**0,544**	0,098	0,013	0,005	0,000	0,000	**2,493**	0,112	0,000	0,000
Ольха серая	0,105	0,078	0,615	0,098	0,078	0,018	**0,563**	0,076	0,262	0,033	0,593	0,095	0,053	0,037	0,000	0,000
Ива	0,370	0,103	0,833	0,084	0,238	0,031	0,244	0,042	0,285	0,028	0,748	0,082	0,122	0,059	0,000	0,000
Черемуха	0,268	0,112	0,000	0,000	0,000	0,000	0,000	0,000	0,015	0,008	0,018	0,018	0,000	0,000	0,000	0,000
Рябина	0,475	0,128	0,275	0,066	0,285	0,034	0,383	0,053	0,332	0,030	0,334	0,071	0,371	0,117	0,000	0,000
Осина	**0,787**	0,193	0,406	0,093	**0,974**	0,062	0,183	0,039	**0,791**	0,051	0,415	0,093	0,131	0,077	0,210	0,098
Сосна	0,000	0,000	0,000	0,000	0,186	0,040	0,012	0,008	0,010	0,006	0,000	0,000	0,053	0,037	**2,107**	0,149
Береза	**1,154**	0,222	**1,508**	0,088	**1,846**	0,058	**0,840**	0,072	**1,478**	0,055	**1,435**	0,093	**1,064**	0,142	**2,479**	0,058
Ель	**1,891**	0,159	**1,714**	0,086	**2,656**	0,032	**1,375**	0,068	**2,251**	0,045	**1,620**	0,091	**1,494**	0,124	**2,811**	0,149
Название класса	Осиново-клено-во-березово-еловые		Вязово-березово-липово-еловые		Осиново-березово-еловые		Ольхово-березово-еловые		Осиново-березово-еловые		Вязово-березово-елово-липовые		Березово-елово-черноольховые		Сосново-березово-еловые	

Примечание. Полужирным шрифтом выделены переменные, осуществляющие наибольший вклад в разделение множества на классы.

роенных разными методами и на основе различных метрик, все равно сохранятся. Более того, очевидно, что для разных целей преимущества имеет каждая из трех классификаций. В этом примере можно видеть аналогию с классификациями растительности, выполняемыми на основе участия видов в сообществе (российская школа геоботаники) и на основе обычных видов с компактным экологическим ареалом (верные или характерные виды в классификации Браун-Бланкé). Первая классификация, осуществленная методом *k*-средних, по логике ближе к классификации российской школы, третья, строящаяся фактически с учетом корреляций, ближе к схеме Браун-Бланкé, а вторая занимает промежуточное положение. Такая неоднозначность классификаций естественна при исследовании биологических систем, пространство которых сильно нелинейно, многомерно, а виды размещаются в пространстве весьма независимо. В результате этого вид полученных линеаниризованных отображений сильно зависит от метрики. Для систем, в основе которых лежат физико-химические механизмы взаимодействий, неоднозначность классификаций существенно меньше.

Каждую из классификаций можно рассматривать как приемлемую, однако всегда остается открытым вопрос: не существует ли лучшей классификации?

Для того чтобы ответить на него, повторим источники различий классификаций:

1) способ представления переменных;
2) форма метризации;
3) метод классификации;
4) неизвестная геометрическая структура многомерного пространства.

Приступая к классификации, как и в случае многомерного непараметрического анализа, нужно в первую очередь с содержательной точки зрения обосновать способ представления данных: исходные данные (нестандартизированные и стандартизированные), логарифмированные данные (нестандартезированные и стандартизированные), ранжированные данные и т.п. Затем необходимо разобраться в методах и метриках, которые представляют соответствующие программные средства, понять содержание положенного в их основу алгоритма и обосновать выбор, исходя из целей исследований. Конечно, это не всегда удается сделать однозначно. Однако сузить область неопределенности все-таки возможно.

Геометрия многомерного пространства сама по себе должна быть предметом исследования и априорные представления в настоящее время сформировать фактически невозможно. Можно лишь еще раз отметить, что при исследовании биологических систем почти всегда приходится иметь дело с многомерными про-

странствами с сильной кривизной. Развитие методов их исследования, описания и трактовки — важная задача современной экологии.

Собственные критерии качества классификации в основном сводятся к получению классов с минимумом внутриклассовой дисперсии как по дистанциям, так и по признакам, или к минимизации дистанции Махаланобиса элементов от центров тяжести классов. Используя эти обычные критерии, можно сравнить качество различных методов кластер-анализа с близкими логическими основаниями.

8.4. Дискриминантный анализ

Более полную проверку качества классификации можно осуществить на основе дискриминантного анализа.

При этом дискриминантный анализ является одним из наиболее доступных методов классификации с обучением. Под обучением подразумевается, что исследователь располагает эталонами классов с соответствующими им элементами и переменными, и имеются измеренные переменные для других элементов, принадлежность которых к классам неизвестна. Необходимо по значениям переменных отнести их к какому-либо из эталонов.

Целью дискриминантного анализа является построение гиперплоскости размерности $k-1$ (k — число классов), разделяющей многомерное пространство на k подобластей таким образом, чтобы элементы каждого класса принадлежали бы только одной подобласти. На рис. 8.8 показано положение классов, полученных для древесного яруса (первый метод, дистанция Евклида) по отношению к трем координатам экологического пространства. Одномерный дисперсионный анализ показал (см. табл. 8.7), что координаты экологического пространства с высоким уровнем значимости отображаются в классах, полученных на основе кластерного анализа. Это соответствие хорошо видно на рис. 8.9, отображающем положение классов в трехмерном экологическом пространстве в целом и в его двухмерных проекциях. В двух вариантах этих проекций (координаты 1 и 3) и (координаты 2 и 3) одномерную «гиперплоскость»-линию можно провести визуально.

Используя операции с моногомерными пространствами на основе алгебраических преобразований, можно построить линейную дискриминантную функцию

$$D = a + b_1x_1 + \ldots + b_nx_n,$$

которая максимизирует F-критерий Фишера

$$\mathrm{F} = T(d)/W(d),$$

Рис. 8.8. Положение классов на трансекте

где $T(d)$ — общая дисперсия; $W(d)$ — внутригрупповая дисперсия в пространстве n-переменных.

Отсюда следует, что классическая линейная версия дискриминантного анализа опирается на идеи дисперсионного и является строго параметрической. С другой стороны, очевидно, что могут существовать нелинейные и непараметрические модели построения гиперплоскости, которые в настоящее время активно разрабатываются, но не получили пока широкого распространения и стандартного программного обеспечения.

Параметрическая основа дискриминантного анализа ограничивает область его применения многомерными нормальными распределениями и линейными пространствами или отношениями.

Если в пределе значения всех переменных определяются только принадлежностью к какому-либо классу, то внутригрупповая корреляция между ними в идеальном случае должна быть равна нулю, а средние с некоторой случайной ошибкой должны однозначно определяться принадлежностью к конкретному классу или, иначе говоря, к k различным в идеале непересекающимся многомерным нормальным генеральным совокупностям.

Соответственно задача сводится к построению обычной регрессионной модели (функция — номер класса, аргументы — переменные) и отображение ее по схеме метода главных компонент в

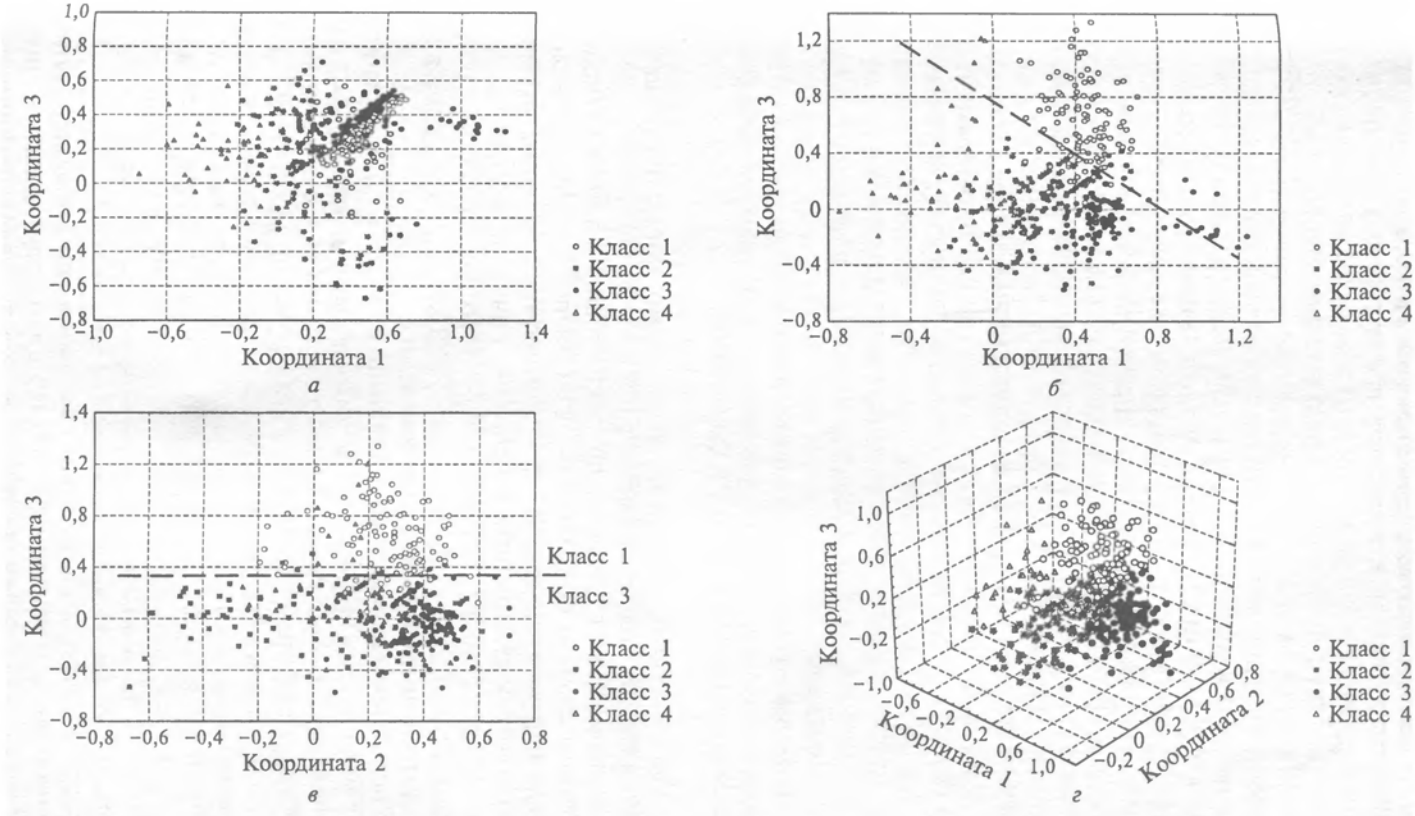


Рис. 8.9. Положение четырех групп (*а*—*г*), полученных методом *k*-средних по дистанции Евклида

ортогональную систему $k-1$ координат. Если читатель разобрался в алгебраическом содержании метода множественной пошаговой регрессии и факторного анализа, то для него не представит особого труда понять и процедуру дискриминантного анализа.

Так же как и при многомерной пошаговой регрессии, в программных средствах дискриминантного анализа используются три версии: стандартная — с учетом всех переменных и пошаговая в двух версиях (вперед и назад). Аналогично методу пошаговой регрессии задаются уровни по F-критерию или вероятности включения в дискриминантный анализ переменной и ее исключения.

Рассмотрим последовательно все результаты дискриминантного анализа на примере классификации методом k-средних по дистанции Евклида (табл. 8.8).

Напомним смысл критериев, приведенных в табл. 8.8:

1) Вилкоксон-ламбда (Wilks' lambda) — статистика отношения определителя матриц внутригрупповой дисперсии/ковариации к определителю матрицы общей дисперсии/ковариации, т.е. критерий, минимизируемый в ходе пошагового дискриминантного анализа. При Wilks' lambda = 0 дискриминация абсолютно полная, при 1 — отсутствует;

2) частный Вилкоксон-ламбда (Partial lambda) — вклад в дискриминацию конкретной переменной, очищенный от эффекта совместного влияния с другими факторами;

Таблица 8.8

Общая оценка качества дискриминантного анализа для четырех классов по 15 переменным, при F-критерии на входе 4 и выходе 3,9. Модель включила 7 переменных

Критерий Вилкоксон-ламбда (Wilks' lambda): 0,01811; F-критерий 175,65; уровень значимости p < 0,0000

Переменная	Критерии					
	Wilks' Lambda	Partial Lambda	F-remove (3,422)	p-level	Tolerance	1-Toler. (R-Sqr.)
Вяз	0,020572	0,880402	19,1089	0,000000	0,939950	0,060050
Липа	0,065037	0,278483	364,4502	0,000000	0,971358	0,028642
Ольха черная	0,019206	0,943044	8,4957	0,000017	0,947663	0,052337
Ольха серая	0,019354	0,935808	9,6491	0,000004	0,910433	0,089567
Осина	0,053196	0,340468	272,4897	0,000000	0,963249	0,036751
Береза	0,030695	0,590057	97,7283	0,000000	0,959899	0,040101
Ель	0,024368	0,743243	48,5941	0,000000	0,978953	0,021047

3) критерий Фишера для уровня удаления переменных (F-remove) — стандартный F-критерий, связанный при расчетах с частным критерием Вилкоксон-ламбда и определяющий «чистый вес» (роль) переменной в разделении классов;

4) толеранс (Tolerance) — ценность переменной, вычисленная как $1 - R^2$ есть мера оценки избыточности соответствующей переменной. Например, если $T = 0{,}10$, то переменная на 90 % избыточна, так как в существенной степени зависит от других;

5) коэффициент детерминации ($R^2 = 1 - T$) — мера зависимости переменной от всех остальных, включенных в модель.

В данном случае два метода пошаговой регрессии (назад и вперед) дают тождественные результаты. Ведущее значение в разделении классов по частному критерию Вилкоксон-ламбда и непосредственно связанному с ним F-критерию имеет липа. Второе место занимает осина. Эти же виды минимально связаны с остальными. В целом же зависимость видов друг от друга по линейному коэффициенту детерминации очень низкая. Обратим внимание на то, что в данном случае результаты дисперсионного анализа совпадают с результатами дискриминантного. Однако так бывает далеко не всегда. Если переменные сильно коррелируют друг с другом, то оценки в дискриминантном анализе по частному ламбда и F-критерию можно рассматривать как более корректные, так как они характеризуют влияние конкретной переменной, «очищенное» от косвенного действия других.

В табл. 8.9 приводится оценка расстояний (различий), оцениваемых по F-критерию и дистанции Махаланобиса, между классами.

Таблица 8.9

Оценки расстояния между классами

Классы	Класс 1	Класс 2	Класс 3	Класс 4
1		161,3660	116,5051	272,2266
2	19,08842		118,7950	245,8135
3	13,31345	12,70047		237,5614
4	56,54581	49,24956	46,64142	

Примечание. Над диагональю значения F-критерия при различении классов (F-values), под диагональю — квадрат дистанции Махаланобиса.

Из табл. 8.9 следует, что наиболее удален от всех четвертый класс, т.е. класс, для которого характерны широколиственные породы. В некотором смысле центральным для других трех классов можно рассматривать второй класс (ольхово-березово-еловых лесов). К нему ближе всего третий класс (березово-еловых лесов). В данном случае F-критерий однозначно упорядочивает классы по расстоянию друг от друга: класс 1, класс 2, класс 3, класс 4. Наиболее удаленными друг от друга являются осиново-еловые и широколиственно-еловые леса.

Следующая таблица (табл. 8.10) фактически отражает цель дискриминантного анализа — получение регрессионных моделей, на основе которых можно рассчитать, к какому классу принадлежит конкретное сочетание значений обилия видов.

Для каждого класса существует уравнение со своими константой и коэффициентами. Подставляя в каждое уравнение значения переменных, будем получать значения классов в дробных величинах относительно номера класса. Поскольку модель параметрическая и предполагается нормальное распределение, можно оценить, с какой вероятностью конкретное значение может быть отнесено к классу, который описывает конкретное уравнение регрессии.

Эти оценки выводятся в специальных таблицах и могут быть запомнены как новые переменные.

В табл. 8.11 приведен фрагмент таблицы оценок принадлежности наблюдаемого класса к классу, полученному на основе расчетов по дискриминантному уравнению.

В первом столбце табл. 8.11 приводится номер точки в ряду наблюдений (номер строки в исходных данных), во втором столб-

Таблица 8.10

Классифицирующая дискриминантная функция при учете объема класса (*p* — доля элементов класса в выборке)

Переменная	Класс 1	Класс 2	Класс 3	Класс 4
	$p = 0{,}26389$	$p = 0{,}29861$	$p = 0{,}32176$	$p = 0{,}11574$
Вяз	1,5618	1,12982	2,5509	21,0659
Липа	8,8798	1,17017	2,0541	1,6652
Ольха черная	3,4564	1,55882	5,6241	3,8772
Ольха серая	7,2833	4,11101	7,5582	5,6529
Осина	1,6602	0,90538	1,3958	6,6003
Береза	0,6349	1,75572	–0,2272	0,9438
Ель	0,5216	1,77949	–0,0367	0,2594
Константа	–21,3311	–5,76889	–18,3673	–35,4891

це — класс, полученный на основе классификации, в третьем столбце — наиболее вероятный класс, получаемый на основе решения дискриминантного уравнения, в четвертом — менее вероятный и т.д. Следующие четыре столбца (с 7-го по 10-й) показывают, с какой вероятностью конкретная точка на местности по состоянию наблюдаемых переменных может быть отнесена к каждому классу. Так, первая точка может быть с вероятностью 0,97 отнесена к действительно наблюдаемому третьему классу, с $p = 0{,}002$ — к первому классу, с $p = 0{,}00005$ — ко второму классу и с $p = 0{,}0000$ — к четвертому. Следующая точка в некотором смысле граничная: с вероятностью 0,76 она относится к наблюдаемому четвертому классу, а с $p = 0{,}23$ — к третьему. Наиболее неопределенно положение точки 8: с $p = 0{,}57$ она относится к наблюдаемому классу 3 и с $p = 0{,}43$ — к классу 4. Точка 11 по дискриминантному уравнению с $p = 0{,}88$ относится к четвертому классу, хотя принадлежит по исходной классификации к первому классу. Такую ситуацию можно трактовать как «ошибку» классификации или дискриминации. Так как в данном случае классификация выполнена формальными методами, ошибочность классификации по дискриминантной модели может быть в первую очередь связана с жесткими параметрическими критериями самой модели дискриминации. Однако при всех условиях к какому бы классу в конечном итоге не относить эту точку, ясно, что она лежит в области границ первого и четвертого классов.

Таким образом, распределения вероятностей принадлежности каждой точки к какому-либо классу могут быть содержательной оценкой для выделения точек или сообществ, положение которых в существующей классификации не соответствует критериям линейной параметрической модели дискриминации. Эти точки могут быть интересны тем, что именно с ними связаны основные эффекты нелинейности отношений или нестационарности и неравновесности конкретных состояний системы.

В целом качество дискриминации оценивается по общей таблице ошибок (табл. 8.12). Дискриминантная функция дает только около 4,5 % ошибок. Второй класс распознается на 100 %. Три точки первого класса относятся по дискриминантной функции к четвертому, а четыре — к третьему классам. Очевидно, что во всех случаях — это своеобразные граничные точки. В целом же классификация может быть признана вполне удовлетворительно отображаемой через линейные дискриминантные функции.

Следует обратить внимание на то, что дискриминантные функции можно строить при двух условиях: с учетом объема каждого класса и считая классы примерно одинакового объема. В данном случае, когда число элементов в каждом классе отличается не очень значительно, различия между этими двумя расчетными схемами невелики. Однако в том случае, когда существуют непропорцио-

Таблица 8.11

Фрагмент таблицы результатов решения дискриминантных уравнений для каждого класса

Элемент	Наблюдаемый класс G_i:j	Порядок предпочтительности решения				Вероятность (*p*) принадлежности к классу			
		1	2	3	4	1	2	3	4
1	2	3	4	5	6	7	8	9	10
1	G_3:3	G_3:3	G_1:1	G_2:2	G_4:4	0,021905	0,000053	**0,978042**	0,000000
2	G_4:4	G_4:4	G_3:3	G_2:2	G_1:1	0,000118	0,000270	0,233657	**0,765954**
3	G_3:3	G_3:3	G_1:1	G_2:2	G_4:4	0,139556	0,005604	**0,854841**	0,000000
4	G_3:3	G_3:3	G_1:1	G_2:2	G_4:4	0,000031	0,000006	**0,999963**	0,000000
5	G_3:3	G_3:3	G_2:2	G_1:1	G_4:4	0,000029	0,000030	**0,999941**	0,000000
6	G_3:3	G_3:3	G_2:2	G_1:1	G_4:4	0,000069	0,000432	**0,999498**	0,000000
7	G_3:3	G_3:3	G_1:1	G_2:2	G_4:4	0,006599	0,000037	**0,993364**	0,000000
8	G_3:3	G_3:3	G_4:4	G_2:2	G_1:1	0,000030	0,000123	**0,572183**	0,427663
9	G_4:4	G_4:4	G_3:3	G_1:1	G_2:2	0,000000	0,000000	0,000000	**1,000000**
10	G_4:4	G_4:4	G_1:1	G_3:3	G_2:2	0,000000	0,000000	0,000000	**1,000000**
Ошибка 11	G_1:1	G_4:4	G_1:1	G_2:2	G_3:3	0,115854	0,000022	0,000017	**0,884107**
12	G_4:4	G_4:4	G_1:1	G_3:3	G_2:2	0,000000	0,000000	0,000000	**1,000000**
13	G_4:4	G_4:4	G_2:2	G_3:3	G_1:1	0,000000	0,000000	0,000000	**1,000000**
14	G_4:4	G_4:4	G_3:3	G_2:2	G_1:1	0,000000	0,000000	0,000000	**1,000000**

Примечания: 1. G_i:j — номер или название класса. 2. Полужирным шрифтом выделено *максимальное значение вероятности.*

Таблица 8.12

Оценка качества классификации (метрика Евклида) по частоте ошибочной дискриминации

Класс	Процент корректной дискриминации	Класс 1	Класс 2	Класс 3	Класс 4
		$p = 0{,}26389$	$p = 0{,}29861$	$p = 0{,}32176$	$p = 0{,}11574$
1	93,8596	107	0	4	3
2	100,0000	0	129	0	0
3	92,8058	5	3	129	2
4	96,0000	0	0	2	48
Всего	95,6019	112	132	135	53

Примечание. Строка: наблюдаемый класс; столбец: рассчитанный класс.

нально большие по объему классы, различия между двумя схемами могут быть весьма существенны.

Рассмотрим отображение двух других версий классификации в дискриминантном анализе (табл. 8.13).

Из табл. 8.13 следует, что это — совершенно иная классификация, в которой точки разделились на классы в первую очередь с учетом обилия клена, ильма, а затем липы и ели, с включением таких весьма неопределенных видов, как черемуха, ива и ясень.

Но качество классификации не менее высокое, чем в первом случае (табл. 8.14).

Третья классификация существенно отличается от двух предыдущих (табл. 8.15).

Здесь разделение на классы вообще не зависит от многочисленных видов и строится на основе относительно редкой черной ольхи, сосны, вяза и липы. Качество же классификации наиболее высокое (табл. 8.16).

Таким образом, еще раз убеждаемся в том, что одно и то же множество можно разделить на принципиально различные, но в каждом случае практически не пересекающие классы. Последнее определяет их формальную реалистичность. Иными словами, существует не менее трех одинаковых по качеству, но различных по содержанию, слабо связанных друг с другом почти дискретных подмножеств.

Следующий этап проведения дискриминантного анализа опирается на особый вид многомерного анализа, называемый *каноническим*. В своей основе он может рассматриваться как вариант множественной регрессии и факторного анализа, объединяемых для

Таблица 8.13

Дискриминантный анализ классификации на основе итерационной процедуры

Wilks' Lambda: 0,01988, F = 103,83, $p < 0,0000$

Переменная	Критерий					
	Wilks' Lambda	Partial Lambda	F-remove (3,418)	p-level	Toler.	1-Toler. (R-Sqr.)
Клен	0,052474	0,378835	228,4614	0,000000	0,905571	0,094429
Вяз	0,033851	0,587247	97,9320	0,000000	0,888202	0,111798
Ель	0,028453	0,698647	60,0996	0,000000	0,929077	0,070923
Липа	0,032969	0,602963	91,7476	0,000000	0,913925	0,086075
Черемуха	0,026062	0,762753	43,3384	0,000000	0,880100	0,119900
Ясень	0,022978	0,865146	21,7185	0,000000	0,895324	0,104676
Осина	0,023636	0,841033	26,3360	0,000000	0,861156	0,138844
Береза	0,023866	0,832934	27,9468	0,000000	0,877675	0,122325
Ива	0,021678	0,917003	12,6109	0,000000	0,931358	0,068642
Ольха серая	0,022577	0,880495	18,9110	0,000000	0,904428	0,095572
Ольха черная	0,022134	0,898131	15,8037	0,000000	0,929122	0,070878

Таблица 8.14

Оценка качества классификации (итерационная процедура) по частоте ошибочной дискриминации

Класс	Процент корректной дискриминации	Класс 1	Класс 2	Класс 3	Класс 4
		$p = 0,25000$	$p = 0,25000$	$p = 0,25000$	$p = 0,25000$
1	100,0000	22	0	0	0
2	96,6667	0	58	1	1
3	97,4249	0	2	227	4
4	97,4359	0	0	3	114
Всего	97,4537	22	60	231	119

Примечание. Строка: наблюдаемый класс; столбец: рассчитанный класс. Третья классификация существенно отличается от двух предыдущих (см. табл. 8.15).

Таблица 8.15

Дискриминантный анализ классификации, минимизирующей многомерный объем кластеров

Wilks' Lambda 0,00171; F = 706,22; $p < 0,0000$

Переменная	Критерий					
	Wilks' Lambda	Partial Lambda	F-remove	p-level	Toler.	1-Toler. (R-Sqr.)
Ольха черная	0,024689	0,069385	1895,619	0,000000	0,997882	0,002118
Сосна	0,012253	0,139807	869,582	0,000000	0,993681	0,006319
Липа	0,005556	0,308307	317,085	0,000000	0,977618	0,022382
Вяз	0,002117	0,809077	33,351	0,000000	0,976186	0,023814
Ива	0,001779	0,962817	5,458	0,001089	0,987442	0,012558

исследования отношений между двумя списками переменных. Одни из этих переменных могут рассматриваться как вход в систему, вторые — как выход. Все теоретические основания те же, что и во всех многомерных параметрических методах, поэтому на них не будем останавливаться и перейдем непосредственно к обсуждению результатов.

Таблица 8.16

Оценка качества классификации, минимизирующей многомерный объем кластеров

Класс	Процент точно определенных классов	Класс 1	Класс 2	Класс 3	Класс 4
		$p = 0,25000$	$p = 0,25000$	$p = 0,25000$	$p = 0,25000$
1	99,07692	322	2	0	1
2	90,32258	6	56	0	0
3	92,30769	2	0	24	0
4	94,73684	1	0	0	18
Всего	97,22222	331	58	24	19

Примечание. Строка: наблюдаемый класс; столбец: рассчитанный класс.

В рамках дискриминантного анализа получено четыре дискриминантные функции, которые можно рассматривать как переменные значения «выхода» и «входа», которые измерены собственно в поле. Канонический анализ отображает их отношения в ортогональной системе координат, которые в его терминологии по традиции называются осями (*корнями*) (табл. 8.17).

По условию для отображения четырех переменных достаточно три оси (или в общей терминологии — факторов). Все тесты табл. 8.17 хорошо известны: нагрузка устанавливает, какая доля варьирования переменных описывается каждой осью и соответственно скрытыми за ними четырьмя дискриминантными функциями; каноническая корреляция показывает множественную корреляцию каждого «корня» с дискриминантными функциями; тест Вилкоксон-ламбда отражает отношение определителей соответствующих матриц ковариации; тест хи-квадрат — значимость каждого корня.

Следующая таблица (табл. 8.18) тождественна по смыслу аналогичным таблицам метода главных компонент и в большей степени метода многомерного шкалирования. Она отражает отношение переменных, в данном случае видов, к осям (корням). Порядок разбора этой таблицы тот же, что и в методе многомерного шкалирования. Вяз и липа в основном определяются первым фактором, осина в большей степени — вторым фактором, в меньшей степени — третьим и в минимальной степени — первым и т.д. Переменные, различающиеся по знаку, занимают противоположные области пространства. Таким образом, канонический анализ совместно с дискриминантным дает еще одну возможность определять положение видов в векторном пространстве на основе параметрического анализа.

Таблица 8.17

Тесты значимости трех осей (корней) канонического анализа

Ось (корни)	Нагрузка (дисперсия) Eigenvalue	Каноническая корреляция Canonic R	Wilks' Lambda	Хи-квадрат Chi-Sqr.	Число степеней свободы	Уровень значимости p-level
0	4,750364	0,908899	0,018112	1706,766	21	0,00
1	2,613488	0,850446	0,104149	962,454	12	0,00
2	1,657175	0,789722	0,376340	415,826	5	0,00

Таблица 8.18

Стандартизованные коэффициенты чувствительности видов к ортогональным корням канонического анализа

Переменная	Root 1	Root 2	Root 3
Вяз	**0,356768**	−0,169576	0,045528
Липа	**0,917407**	−0,256058	0,012521
Ольха черная	−0,030035	0,195585	0,225416
Ольха серая	0,026391	0,166126	0,283067
Осина	**−0,285307**	**−0,782274**	**0,529491**
Береза	0,023115	**−0,363291**	**−0,728706**
Ель	−0,099600	**−0,469644**	−0,389378
Нагрузка (Eigenvalue)	4,750364	2,613488	1,657175
Накопленная доля влияния (Cum.Prop)	0,526588	0,816299	1,000000

Примечание. Полужирным шрифтом выделен ведущий фактор.

Положение центров тяжести классов по отношению к корням (факторам, координатам) определяют по табл. 8.19.

Из таблицы видно, что в трехмерном пространстве по первой оси наиболее удален от остальных четвертый класс. По второй оси противоположное положение занимают второй и первый классы, а по третьей — первый и третий. Это размещение естественно можно

Таблица 8.19

Средние значения корней (центры тяжести) по отношению к каждому классу

Класс	Root 1	Root 2	Root 3
1	−1,45223	−2,01437	1,12748
2	−0,42590	2,20224	0,84728
3	−0,52980	−0,12097	−1,83129
4	5,88272	−0,75272	0,33437

показать в трехмерном пространстве. Оно дает представление о связи элементов каждого класса с каждым корнем.

Соответственно, как и во всех случаях, можно показать положения точек наблюдений (элементов системы) в трехмерном пространстве (рис. 8.10). Эти графики фактически отражают качество разделения классов в многомерном пространстве. Оси или корни имеют тот же смысл, что и факторы в факторном анализе и координаты в многомерном шкалировании. Однако важно особо отметить, что они получены из дискретного отображения данных через классы в непрерывное. Из рис. 8.10 видно, что хотя элементы, принадлежащие к подмножеству, действительно образуют достаточно компактные множества точек, однако они все-таки, имея в виду доверительные интервалы, частично перекрываются. Иными словами, они выделены из континуума. Корни канонического анализа с высокой надежностью описывают обилие тех видов, которые в первую очередь определяют выделение классов. Такие виды, как липа, осина, береза и ель, отображаются этими осями с коэффициентом детерминации 0,8—0,9. Размещение остальных видов в пространстве описывается очень слабо. Для каждого метода классификации существует свой набор видов, распространение которых хорошо отображается осями (корнями) дискриминантного анализа. Оси, полученные в результате дискриминантного анализа, в каждом из трех вариантов описываются координатами экологического пространства, полученными на основе многомерного шкалирования, в уравнении регрессии с коэффициентом детерминации от 0,7 до 0,9.

Таким образом, дискриминантный анализ позволяет перейти от дискретного отображения свойств изучаемого объекта, полученного с помощью кластерного анализа, к непрерывному. При этом они частично соответствуют координатам, полученным в рамках непрерывной многомерной схемы анализа. Чем бóльшую роль в системе играют линейные отношения, тем эти два типа отображения более тождественны. Однако, исследуя биологические системы, предпочтительнее все-таки использовать метод многомерного шкалирования, а классификацию, в случае необходимости, лучше строить на основе координат экологического пространства.

В табл. 8.20 приведены результаты дискриминантного анализа для классов, выделенных методом k-средних по метрике Евклида, на основе координат экологического пространства, полученных методом многомерного шкалирования. Высокое качество отображения координатами обилия каждого вида определяется тем, что классы с достаточно высокой надежностью разделяются собственно по обилию видов, хотя классификация построена на осях экологического пространства (табл. 8.21). В табл. 8.22 дается описание классов, полученных в результате этой последней версии классификации.

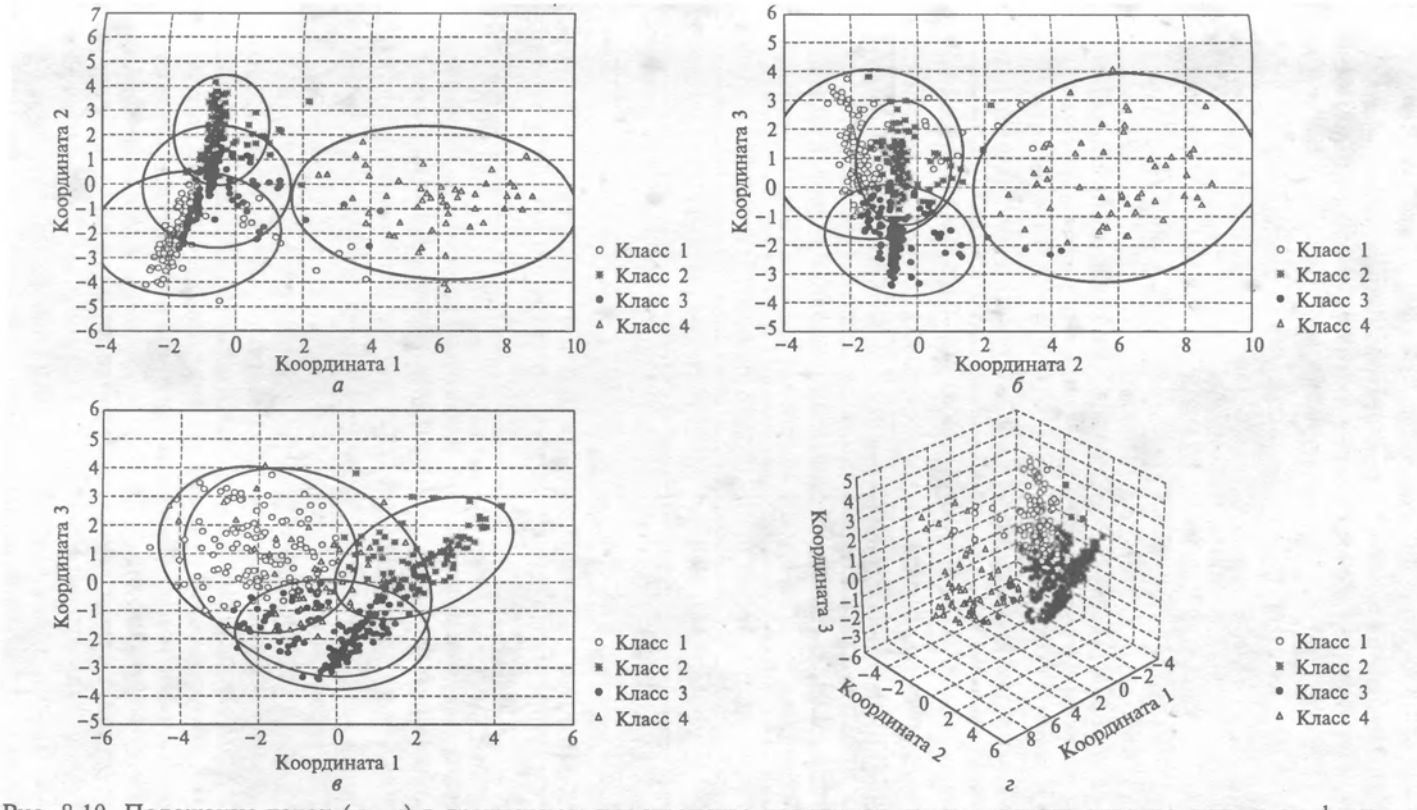


Рис. 8.10. Положение точек (*а*—*г*) в трехмерном пространстве координат канонического анализа для классификации методом *k*-средних по метрике Евклида (эллипсы — 95%-й доверительный интервал)

Таблица 8.20

Дискриминантный анализ классификации, полученный при метрике Евклида по экологическим координатам — по переменным, описывающим обилие видов

Wilks' Lambda 0,04738; F-критерий = 74,910; $p < 0,0000$

Переменная	Wilks' Lambda	Partial Lambda	F-remove (3,419)	p-level	Toler.	1-Toler. (R-Sqr.)
Осина	0,119892	0,395183	213,7559	0,000000	0,918338	0,081662
Ель	0,070493	0,672112	68,1359	0,000000	0,885219	0,114781
Береза	0,076346	0,620584	85,3901	0,000000	0,889713	0,110287
Вяз	0,058008	0,816767	31,3326	0,000000	0,799867	0,200133
Черная ольха	0,063191	0,749778	46,6107	0,000000	0,883021	0,116979
Рябина	0,055416	0,854970	23,6919	0,000000	0,934675	0,065325
Серая ольха	0,052738	0,898381	15,7982	0,000000	0,967872	0,032128
Ива	0,051860	0,913601	13,2083	0,000000	0,950396	0,049604
Клен	0,048303	0,980878	2,7227	0,044005	0,947278	0,052722
Липа	0,048230	0,982352	2,5091	0,058354	0,802966	0,197034

Таблица 8.21

Качество дискриминации классов, полученных на основе классификации по осям экологического пространства, обилием видов деревьев

Класс	Процент точно определенных классов	Класс 1	Класс 2	Класс 3	Класс 4
		$p = 0,18981$	$p = 0,22222$	$p = 0,34722$	$p = 0,24074$
1	79,26830	65	9	6	2
2	95,83334	2	92	2	0
3	94,00000	2	4	141	3
4	94,23077	1	0	5	98
Всего	91,66666	70	105	154	103

Таблица 8.22

Средние значения логарифма сумм площадей сечений древесных пород для классов классификации по экологическим координатам (дистанция Евклида)

Переменная	Черноольхово-еловые леса	Березово-осиново-еловые леса	Березово-еловые леса	Широколиственно-еловые леса
	Класс 1, N = 70	Класс 2, N = 105	Класс 3, N = 154	Класс 4, N = 103
Вяз	0,085179	0,063213	0,071588	**0,839980**
Ясень	0,026795	0,000000	0,000000	0,025376
Клен	0,190430	0,107761	0,025808	0,094441
Липа	0,167177	0,294186	0,066264	**0,735710**
Дуб	0,000000	0,007220	0,000000	0,000000
Ольха черная	**0,790406**	0,000000	0,013863	0,019995
Ольха серая	0,042265	0,156912	0,170796	**0,760688**
Ива	0,142127	0,146859	0,290919	0,699226
Черемуха	0,028080	0,000000	0,016566	0,010564
Рябина	**0,714049**	0,277887	0,198279	0,224471
Осина	0,332778	**1,954213**	0,418781	0,118334
Сосна	0,016906	0,000000	0,284961	0,006665
Береза	0,642765	**1,120437**	**2,267443**	**1,382566**
Ель	**1,737077**	**2,535403**	**2,671468**	**1,323886**

Примечание. Полужирным шрифтом выделены наибольшие значения для каждого класса.

Эта классификация ближе к версии, полученной на основе исходных переменных по метрике Евклида, но, безусловно, несколько иная, хотя и столь же логичная, как три предыдущие. На рис. 8.11 показано отображение классов в координатах корней дискриминантного анализа. Все классы по координатам как лучи расходятся из общего центра координатной системы, где они практически все пересекаются друг с другом. Первая ось (корень) противопоставляет леса с участием осины лесам с участием широколиственных пород, но не клена, что, скорее всего, отражает сукцессионные смены, так как широколиственные породы все-таки более характерны для молодых и средневозрастных лесов. По второй оси господствующие березово-еловые леса противопоставляются всем

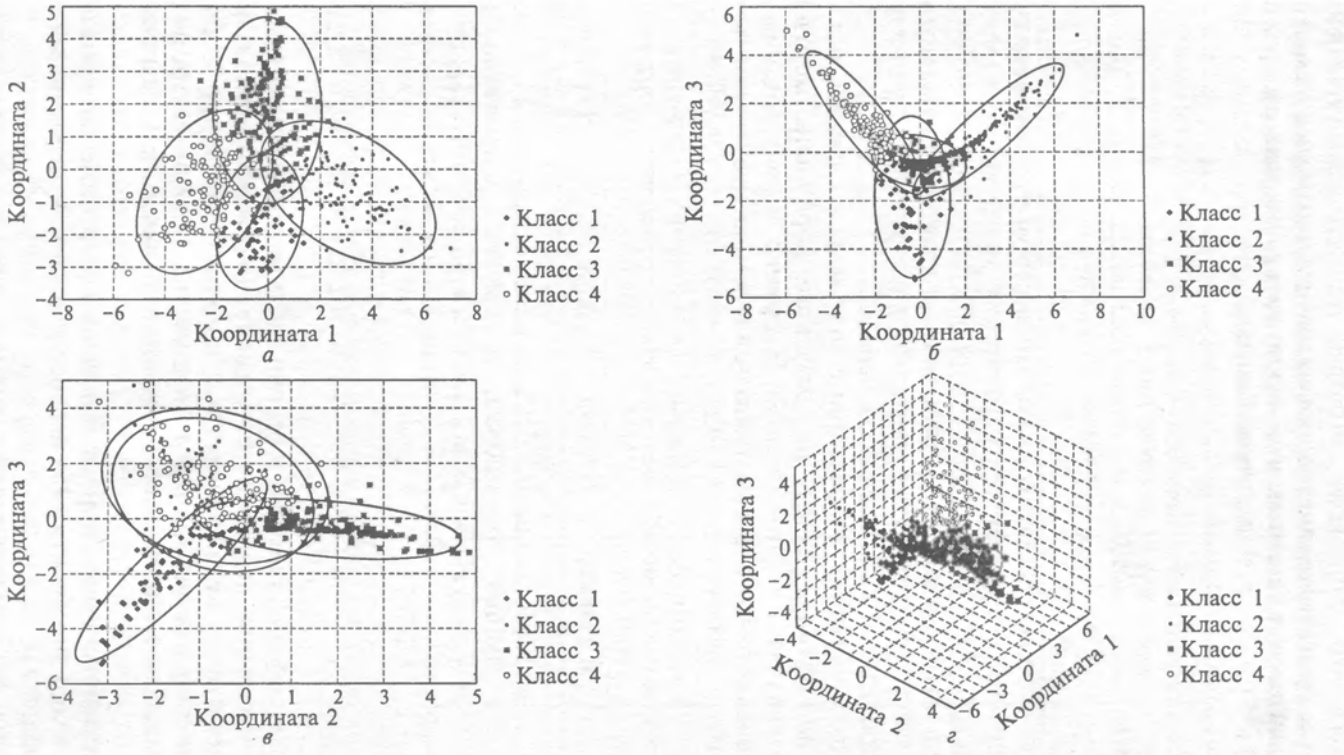


Рис. 8.11. Положение классов (*а*—*г*), полученных на основе классификации по осям экологического пространства в координатах (корнях) дискриминантного анализа (класс 1 — черноольхово-словые леса; класс 2 — березово-осиново-еловые леса; класс 3 — березово-еловые леса; класс 4 — широколиственно-еловые леса)

остальным, что, скоре всего, отражает температурный градиент. Третья ось противопоставляет леса со значительным участием черной ольхи всем остальным, что можно трактовать как отображение градиента увлажнения. Размещение классов в пространстве очень четко и компактно ориентировано по градиентам. Структура изображения в сравнении, например, с аналогичной классификацией по обилию древесных пород (см. рис. 8.10) значительно компактнее и логичнее, что позволяет считать ее более удачной. Однако это не более, чем внешний, эвристический, но полезный критерий качества.

Таким образом, несмотря на то, что дискриминантный анализ является весьма жестким параметрическим методом, его использование при решении задач многомерной ординации можно считать полезным, потому что в сочетании с кластерным анализом часто помогает более четко выявить основные факторы, организующие пространство исследуемого явления.

Дискриминантный анализ применяют и в том случае, когда классы или группы получены не в результате формальных методов кластер-анализа, а заданы априори. Например, в почвоведении в качестве классов могут рассматриваться генетические горизонты и в рамках дискриминантного анализа анализируется полнота их разделения какими-либо характерными для них переменными. В качестве классов могут также выступать типы местообитания, погодные условия и т. п.

Вернемся к примеру с почвами. Глубины, на которых взяты образцы почвы для анализа, можно рассматривать как классы. Требуется с помощью дискриминантного анализа определить, в какой степени глубина отбора образца описывает варьирование измеренных свойств почв и каковы закономерности этого варьирования в пространстве. В табл. 8.23 приведены результаты соответствующего анализа.

Из табл. 8.23 очевидно, что связь глубины отбора образца со значениями переменных вполне достоверна (она максимальна для влажности и минимальна для фосфора).

Данные табл. 8.24 показывают, что в целом глубина описывает 72 % варьирования значений всех элементов. При этом минимальная надежность классификации для горизонта глубиной 30 см. С другой стороны, все ошибочно идентифицированные классы с наибольшей вероятностью соседствуют с истинным классом глубины.

Теперь проведем оценку значимости корней (осей) дискриминантного анализа (табл. 8.25) и чувствительности переменных к осям (табл. 8.26).

Из табл. 8.25 можно сделать вывод о том, что статистически значимы только две оси и в рамках дискриминантного метода система представляется двухмерной.

Таблица 8.23

Результаты дискриминантного анализа глубины отбора образцов с значениями измеренных переменных

Wilks' Lambda 0,11872; approx. F (24,2076) = 72,865; $p < 0,0000$

Переменная	Wilks' Lambda	Partial Lambda	F-remove (4,595)	p-level	Toler.	1-Toler. (R-Sqr.)
Влажность	0,246379	0,481848	159,9575	0,000000	0,820727	0,179273
Кислотность	0,138317	0,858299	24,5579	0,000000	0,642149	0,357851
Фосфор	0,122785	0,966871	5,0969	0,000482	0,763694	0,236306
Калий	0,154782	0,766996	45,1884	0,000000	0,735811	0,264189
Магний	0,132879	0,893424	17,7443	0,000000	0,240696	0,759304
Кальций	0,130876	0,907094	15,2352	0,000000	0,180104	0,819896

Из табл. 8.26 следует, что первая ось с отрицательным знаком определяет влажность и концентрацию фосфора и с положительным — кислотность. Вторая ось положительно связана с концентрациями магния и фосфора и отрицательно — с концентрацией калия. Кислотность и концентрация магния в наибольшей степени зависят от третьей оси, а четвертая ось в существенной степени определяет концентрацию кальция. Однако вклад этих двух осей в описание варьирования всех элементов минимален и статистически незначим. Весьма характерно, что варьирование концентраций

Таблица 8.24

Качество отображения глубин по варьированию значений измеренных переменных

Глубина, см	Процент правильно определенных	Глубина, см				
		5	10	20	30	40
		$p = 0,20000$	$p = 0,20000$	$p = 0,20000$	$p = 0,20000$	$p = 0,20000$
5	72,72727	88	32	0	1	0
10	70,24793	17	85	11	6	2
20	76,03306	2	9	92	17	1
30	56,19835	0	1	20	68	32
40	85,95042	0	2	1	14	104
Всего	72,23141	107	129	124	106	139

Таблица 8.25

Оценка значимости осей дискриминантного анализа

Chi-Square Tests with Successive Roots Removed (mlk_col.sta)

Ось (корни)	Eigen — value	CanonicR	Wilks' Lambda	Chi-Sqr.	df	p-level
0	3,015087	0,866568	0,118717	1275,411	24	0,000000
1	1,059870	0,717309	0,476659	443,461	15	0,000000
2	0,011694	0,107513	0,981856	10,959	8	0,204087
3	0,006707	0,081621	0,993338	4,001	3	0,261426

магния и кальция существенно независимо от других переменных, что было в частности получено и в кластер-анализе. В целом же практически каждая переменная определяется своим ведущим фактором и имеет собственные, отличные от других, правила пространственного варьирования.

Закономерности организации вертикальной структуры почвенного профиля по измеренным переменным в наиболее простой

Таблица 8.26

Стандартизированные значения чувствительности переменных к осям (корням)

Переменная	Root 1	Root 2	Root 3	Root 4
Влажность	**–0,903376**	0,178716	–0,38954	0,216589
Кислотность	**0,738698**	–0,061762	**–1,63892**	0,023134
Фосфор	**–0,501958**	**0,496879**	–0,09332	0,219611
Калий	–0,144847	**–0,621136**	**–0,58248**	**–0,609854**
Магний	0,054811	**0,965567**	**1,70907**	–0,133953
Кальций	0,079113	–0,258815	0,20735	**–0,747004**
Нагрузка (Eigenvalue)	3,015087	1,059870	0,01169	0,006707
Накопленная доля (Cum. Prop)	0,736580	0,995505	0,99836	1,000000

Примечание. Полужирным шрифтом выделены ведущие факторы для каждой оси.

Таблица 8.27

Положение центров тяжести классов глубин в осях дискриминантного анализа

Глубина, см	Root 1	Root 2	Root 3	Root 4
5	−2,41616	0,97089	−0,077226	−0,065074
10	−1,34858	−0,18254	0,097543	0,129969
20	−0,01090	−1,65960	0,050385	−0,087864
30	1,59927	−0,32128	−0,172233	0,057014
40	2,17637	1,19253	0,101520	−0,034044

форме демонстрирует табл. 8.27. Значение первой оси минимально в слое глубиной 5 см и максимально на глубине 40 см. Очевидно, что значения описывают почти прямую линию, и в соответствии с табл. 8.26, отражают уменьшение влажности и увеличение концентрации магния с глубиной. Вторая ось, очевидно, отражает генетические горизонты. При этом подзолистый горизонт маркируется отрицательным значением оси, индуцирующим относительный минимум pH на этой глубине и минимум кальция и калия. Соответственно, pH максимально на глубине 5 и 40 см, где максимальна концентрация обменных оснований, в первую очередь кальция. Однако таково было бы «чистое» влияние этой оси, если бы не было ведущего влияния основной, первой, описывающей бóльшую часть варьирования. В целом почти все свойства почвы изменяются пропорционально изменению с глубиной значений первой оси, в то время как вторая вносит некоторые коррективы в соответствии со знаком и мерой влияния, увеличивая или уменьшая значения, определяемые первой осью. Эти общие закономерности варьирования свойств почв с глубиной представлены на рис. 8.12.

Множественная регрессия показывает, что две первые оси описывают бóльшую часть варьирования всех переменных, кроме кислотности и концентрации фосфора. Для отображения варьирования этих переменных требуется рассмотрение в целом малозначащих третьей и четвертой осей.

Очевидно, что дискриминантный анализ зависимости значений переменных от глубины горизонта, с одной стороны, весьма прост, а с другой — достаточно нагляден и эффективен. При многомерном анализе относительно простых почти линейных систем дискриминантный анализ можно рассматривать как весьма эффективный метод.

Подведем общий итог методов многомерного анализа.

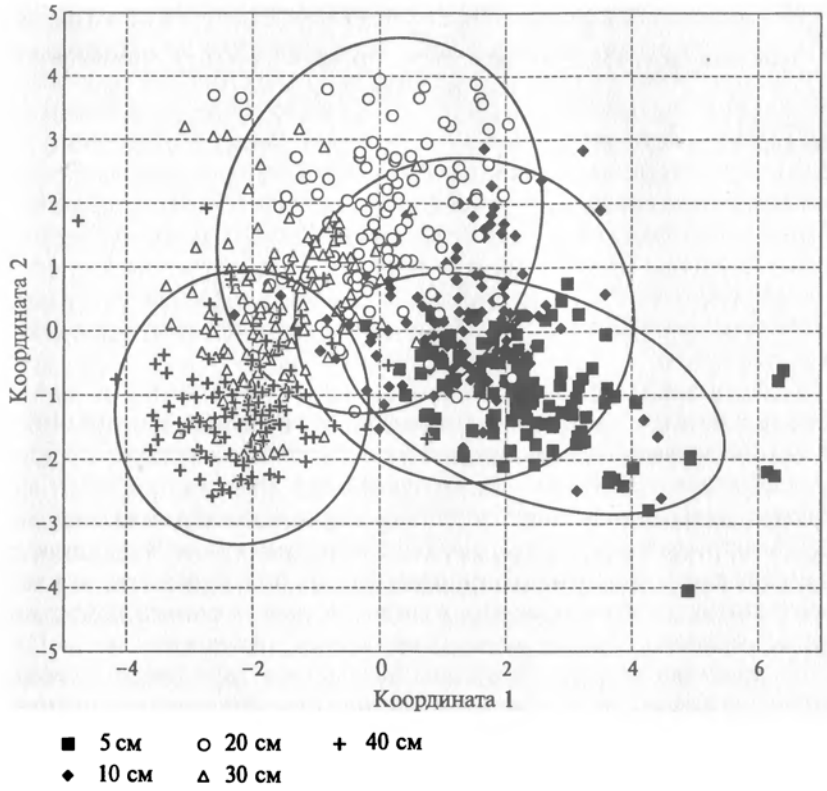


Рис. 8.12. Варьирование свойств почв в зависимости от глубины отбора образцов почвы в двухмерном пространстве

Методы многомерного анализа можно разделить на четыре большие группы:

I. Параметрические. Понимание их логико-математических основ и возможностей прямо связывается с базовыми представлениями теории матриц и методов решения систем *n*-уравнений. При анализе используется информация о среднем и дисперсии (ковариации). Все методы этой группы идеально пригодны для нормальных распределений и линейных отношений между переменными.

II. Непараметрические многомерные. К этим методам относится метод многомерного шкалирования, корректное применение которого требует понимания свойств различных метрик. Метод реализует поиск наилучшего взаимоположения элементов системы в многомерном пространстве, минимально искажающих измеренные между ними дистанции. Существенной проблемой является определение размерности пространства. При интерпретации результатов необходимо учитывать возможность нелинейной зависимости переменных от координат пространства.

III. Кластерный анализ. Эти методы позволяют выделить относительно дискретные компактные подмножества из континуума. Результат прямо зависит от используемой метрики и метода классификации. Эффективное применение возможно при условии понимания свойств метрик и логических оснований конкретного метода классификации. Проблемой является существование нескольких мало отличающихся по качеству, но существенно различающихся по содержанию классификаций. Выбор конечной версии, кроме формального критерия минимума внутригрупповой дисперсии, определяется целями и взаимоотношениями кластеров в многомерном пространстве, порождающих строго взаимоупорядоченные структуры.

Если системы линейны и нормальны, то результаты их многомерного анализа любыми методами практически тождественны.

IV. Дискриминантный анализ. Эти системы позволяют перейти из дискретного отображения явления в непрерывное. Классы или группы, рассматриваемые как непересекающиеся множества, могут быть введены априори или получены на основе классификации. Для систем с линейными или линеаризованными отношениями и нормальными распределениями метод в весьма наглядной форме отражает структуру многомерного пространства.

В целом многомерный анализ дает основу для формулировки гипотез о механизмах отношений и построения дедуктивных моделей пространственно-временной динамики изучаемого явления.

Контрольные вопросы

1. Чем отличаются методы параметрического и непараметрического многомерного анализа? Назовите области их применения.
2. Что является общей основной параметрических методов анализа?
3. Чем содержательно отличаются метрики, построенные на основе схемы Минковского, корреляции и их информации?
4. Какими способами можно оценить размерность пространства?
5. Какими способами можно определить физический смысл виртуальных факторов?

Глава 9
АНАЛИЗ ВРЕМЕННЫ́Х И ПРОСТРАНСТВЕННЫХ РЯДОВ НАБЛЮДЕНИЙ

9.1. Общие замечания

До сих пор при анализе изменения состояний явлений во времени и в пространстве элементы исходной системы рассматривались как полностью независимые. Вместе с тем эти элементы располагались во времени или в пространстве на равном интервале друг от друга по координате времени или координате пространства, т.е. регулярно. Хотя такая регулярно организованная выборка и не обязательна, но она имеет свои преимущества. Эти преимущества состоят в том, что есть основания полагать, что состояние элемента i содержит некоторую информацию о состоянии $i + k$ и эти изменения состояний в пространстве и во времени подчиняются одному или нескольким правилам, иначе говоря, имеют некоторый порядок, или структуру.

Таким образом, *временны́м* (*пространственным*) *рядом* называется упорядоченная совокупность значений переменных, измеряемых через строго постоянный шаг (лаг).

Задача анализа временно́го ряда — установление правил отношений между состояниями явления во времени (в пространстве).

Цель анализа — сформулировать гипотезы и построить дедуктивные модели, воспроизводящие этот порядок. Более частной целью является прогноз будущего состояния на основе знания предшествующих.

Нулевой гипотезой во всех случаях является представление о чисто случайном варьировании состояний в пространстве или во времени, подчиняющемуся к тому же нормальному распределению. Такое случайное нормальное варьирование называется «*белым шумом*». Белым он называется по ассоциации с листом бумаги, отражающим свет почти равномерно на всем диапазоне волн в видимой части спектра. Строго говоря, это не только аналогия.

В действительности каждый ряд может содержать в себе кроме «белого шума» фрактальную составляющую колебаний, регулярные гармонические колебания, тренд и выбросы. На рис. 9.1 показана одна из возможных композиций этих составляющих.

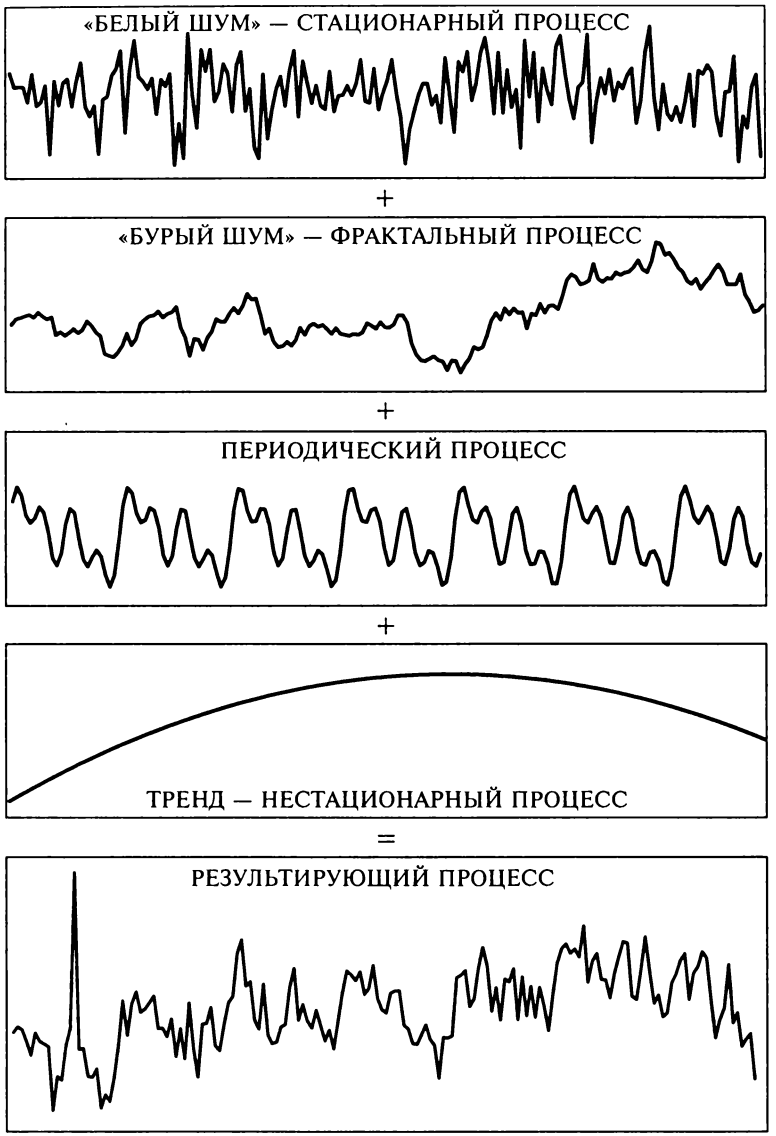


Рис. 9.1. Составляющие случайного процесса

Каждый пространственно-временно́й процесс характеризуется своим механизмом, т. е. имеет свою природу. «Белый шум» обычно связывается со случайными флуктуациями типа «теплового шума». «Тепловой шум» и его аналоги присутствуют в любой системе, его дисперсия (мощность) пропорциональна подводимой к системе энергии. «Тепловой шум» строго стационарный процесс, т. е. его дисперсия и среднее, два первых момента распределения не изме-

няются во времени (пространстве) для всего ряда наблюдений. Фрактальный процесс обычно связывается со случайным блужданием или диффузией и их аналогами. Он в той или иной степени проявляется всюду, где существуют движения и взаимодействия. Движение влаги в почве, в гидрогеологической системе; формирование эрозионной сети, разрывных нарушений в результате тектонических движений, трещин во льдах и тому подобное обычно описывается фрактальным процессом. Фрактальный процесс, строго говоря, не стационарен, так как его амплитуда тем больше, чем больше интервал наблюдений во времени или в пространстве. Вместе с тем фрактальный процесс самоподобен и вид изменения состояний в мелком масштабе подобен виду того же процесса в более крупном. Однако такое самоподобие возможно, если и в крупном и малом масштабе процесс имеет одну и ту же природу.

Периодические (регулярные, гармонические) процессы являются стационарными, если их среднее, дисперсия и автоковариация неизменны во времени и в пространстве.

Автоковариация — это та же ковариация, что и в общем случае, но рассчитываемая для двух рядов, полученных сдвигом исходного ряда на определенный шаг (лаг) относительно самого себя. В результате ковариация оценивается для значений при сдвиге на один шаг:

Первый ряд	x_1,	x_2,	x_3,	...,	x_i,	...,	x_n
Второй ряд	x_2,	x_3,	x_4,	...,	x_{i+1},	....	

Аналогичным образом оценивается и автокорреляция.

Если мысленно смещать ряд с гармоническими колебаниями относительно самого себя, то легко увидеть, что при сдвиге на четверть периода корреляция будет равна нулю, при сдвиге на половину периода — минус единице, на три четверти периода — вновь нулю и на полный период — плюс единица. Но на всей длине стационарного ряда автоковариация варьирует в некотором строго определенном диапазоне. Именно в этом смысле ряд может рассматриваться как стационарный. Вообще говоря, стационарным в широком смысле называется ряд, для которого механизм, определяющий флуктуации переменных, не меняется во времени и в пространстве.

Природа гармонических колебаний обычно описывается моделью линейного или нелинейного осциллятора. Под *осциллятором* понимается система, которая при подводе к ней энергии генерирует автоколебания. Типичным осциллятором является маятник часов любой конструкции. В общем случае автоколебания возникают в результате того, что система состоит из взаимодействующих частей, в передаче сигналов между которыми неизбежно существует запаздывание. Величина запаздывания определяется свойствами самой системы и, в частности, ее массой. Период автоколебаний связан с моментом инерции системы.

В линейных осцилляторах амплитуда колебаний не зависит от периода, в нелинейных — чем больше амплитуда (больше подвод энергии), тем меньше период. В результате нелинейная система начинает генерировать колебания в некотором диапазоне периодов. При этом при продолжении увеличения амплитуды, что происходит при увеличении подвода энергии, происходит так называемый триггерный эффект и система начинает колебаться с двумя основными периодами и множеством маломощных дополнительных. При дальнейшем увеличении подвода энергии возникают колебания с третьим значением периода и система в целом генерирует колебания с несколькими длинами волн. При очень большом подводе энергии колебания начинают происходить практически на всех длинах волн, и динамика становится хаотичной.

Таким образом, любая система имеет собственные колебания, которые выявляются при вводе энергии или при действии некоторой периодической (апериодической) внешней силы. При воздействии внешней гармонической силы возможны следующие три случая:

1) внешние возбуждающие колебания имеют периодичность, существенно меньшую, чем период собственных колебаний системы. На такие воздействия система обычно вообще не реагирует;

2) возбуждающие колебания по периоду близки собственным колебаниям системы. Это соотношение определяет возможность резонанса, и амплитуда колебаний системы во времени обычно растет. Такие колебания имеют очень сложную структуру и не являются стационарными. В пределе рост амплитуды колебаний может разрушить систему, а если она нелинейна — привести к колебаниям на нескольких гармониках. В этом варианте система может не разрушаться, так как часть энергии возбуждения «сбрасывается» на более короткие периоды колебаний;

3) период возбуждающих колебаний более чем в два раза превышает период собственных колебаний и система испытывает медленные колебания с периодом, равным возмущающей силе. Такие колебания обычно порождают тренды, они являются нестационарными.

При единовременном возмущении система по условию нестационарна и ее динамика определяется типом ее устойчивости. Если система устойчива по Ляпунову, то, осуществляя гармонические колебания с периодом, близким к собственному, она возвращается к своему стационарному состоянию в точку, в которой она флуктуирует на уровне «белого шума». Если система устойчива по Лапласу, то она постепенно уменьшает амплитуду колебаний и переходит в область собственных циклических автоколебаний. Если система устойчива по Пуассону, то она постепенно, после снятия возмущения переходит в область стохастических колебаний с конечной дисперсией.

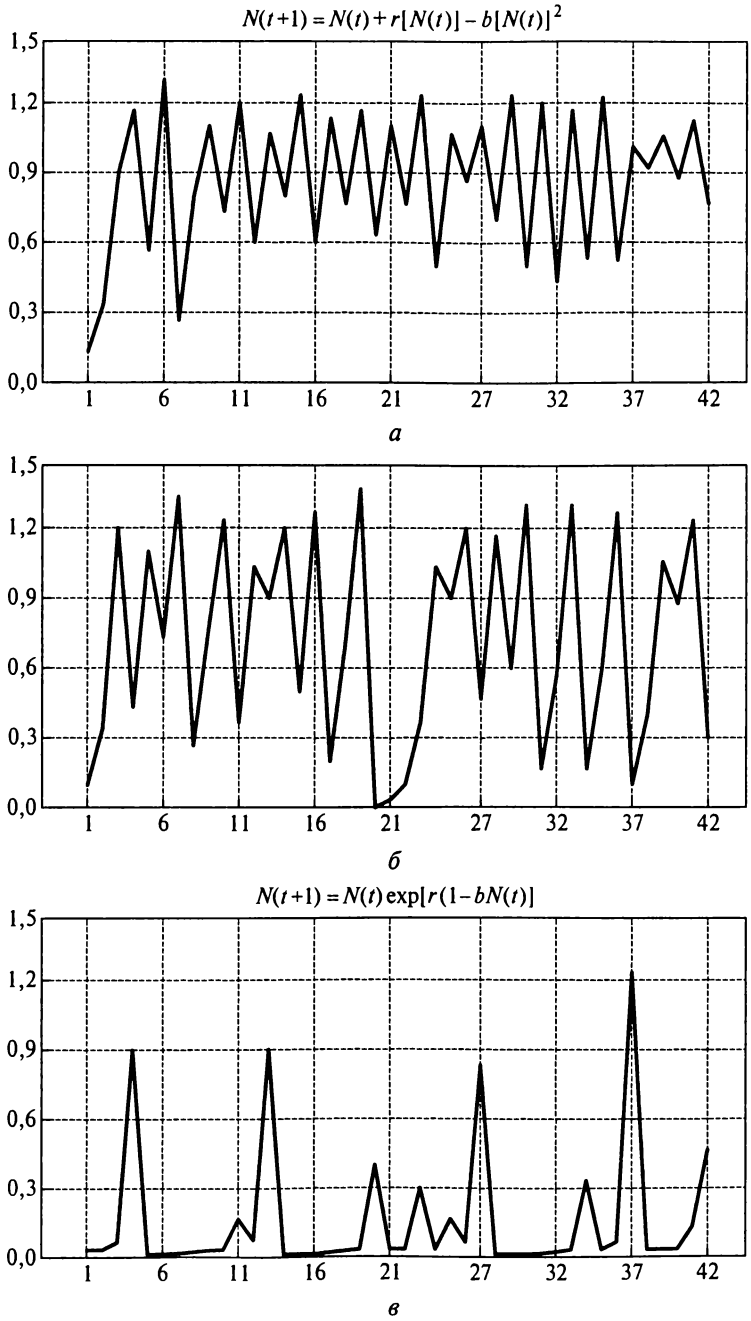


Рис. 9.2. Модели динамики с запаздыванием. Имитация динамики численности по моделям с самоингибированием случайной составляющей:
а — среднее $r = 2{,}5$; *б* — $r = 3{,}0$; *в* — $r = 5{,}0$

Подобная модель колебаний была предложена Мак-Артуром для популяций. Эта классическая логистическая модель, но с запаздыванием, равным среднему времени вступления в размножение. Модель достаточно детально разобрана Ю. М. Свирежевым и Д. И. Лагофетом*.

Собственный период колебаний в таких моделях равен примерно $4T$, (T — среднее время вступления особей в размножение). Характер колебаний определяется соотношением коэффициента размножения r и самоингибирования (рис. 9.2). Если коэффициент размножения мал, а самоингибирование велико, то популяция ведет себя как устойчивая по Ляпунову (рис. 9.2, *а*). При увеличении коэффициента размножения ее поведение соответствует модели Лапласа (рис. 9.2, *б*), а при очень большом коэффициенте размножения или очень малом коэффициенте самоингибирования она ведет себя как устойчивая по Пуассону (рис. 9.2, *в*).

Представления о возможных моделях динамики изучаемых явлений чрезвычайно полезны, так как переводят исследования из плоскости чистых статистических описаний поведения системы в область поиска механизмов наблюдаемого поведения. Однако прежде чем ставить такую задачу, необходимо корректно расчленить ряд на четыре составляющих: 1) «белый шум», 2) фрактальный процесс, 3) гармонические колебания, 4) тренд и выбросы и описать статистические модели этих процессов.

9.2. Методы исследования структурной организации временно́го (пространственного) ряда

Прежде, чем перейти к решению этой задачи, определим условия ее разрешения и основные связанные с ней понятия.

Пусть имеется ряд наблюдений длины L. Периодом колебания будем называть элемент ряда длиной P, за который функция пробегает один раз все свои возможные значения и приходит к исходному состоянию. Каждому периоду колебаний будем ставить соответствующую гармонику. Номер гармоники, отсчитываемый от максимально возможного периода длиной L, назовем *волновым числом* k.

$$k = L/P.$$

Частота колебаний есть величина, обратная периоду

$$\omega = 1/P.$$

---

* *Свирежев Ю. М., Лагофет Д. И.* Устойчивость биологических сообществ. — М.: Наука, 1978. — 352 с.

Круговая частота $\lambda = 2\pi\omega$.

Условие воспроизведения состояния ряда определяет теорема отсчета (теорема Вудворда — Котельникова):

Если функция не содержит частот выше ω, *она полностью определяется своими мгновенными значениями в моменты, отстающие друг от друга на интервал* $1/2\omega$.

Иными словами, если необходимо воспроизвести процесс периодом P, то необходимо, чтобы наблюдения отстояли друг от друга на шаг $\Delta t = P/2$. Следовательно, если имеется ряд длиной L, то на нем воспроизводимы функции с периодами в интервале $P_{\min} = 2$ до $P_{\max} = L/2$ или в полосе частот от 0,5 до $1/L$. Частота 0,5 называется *частотой Найквиста.*

Например, если имеются среднемесячные значения температур за 100 лет, то воспроизводим весь диапазон колебаний от 2 месяцев до 50 лет. При этом какие колебания происходят с периодами меньше двух месяцев по такому ряду установить невозможно. Все колебания с меньшими периодами будут воспроизводимы в ряду как шум.

Таким образом, теорема отсчета определяет частоты процессов, воспроизводимые на основе ряда заданной длины. Это общее представление дает основание для обобщенного описания свойств ряда, которое называется *спектральным анализом.* Так как в соответствии с теоремой отсчета временной ряд полностью определяется мгновенными значениями наблюдений, отстающими во времени или в пространстве на шаг, определяемый частотой, его можно описать системой уравнений регрессии, образующих ортогональный базис для каждой гармоники с соответствующей круговой частотой, как функцию времени t или, в общем случае, функцию номера члена ряда:

$$x_t^{\lambda} = a_0 + a\cos(\lambda t) + b\sin(\lambda t), \qquad (9.1)$$

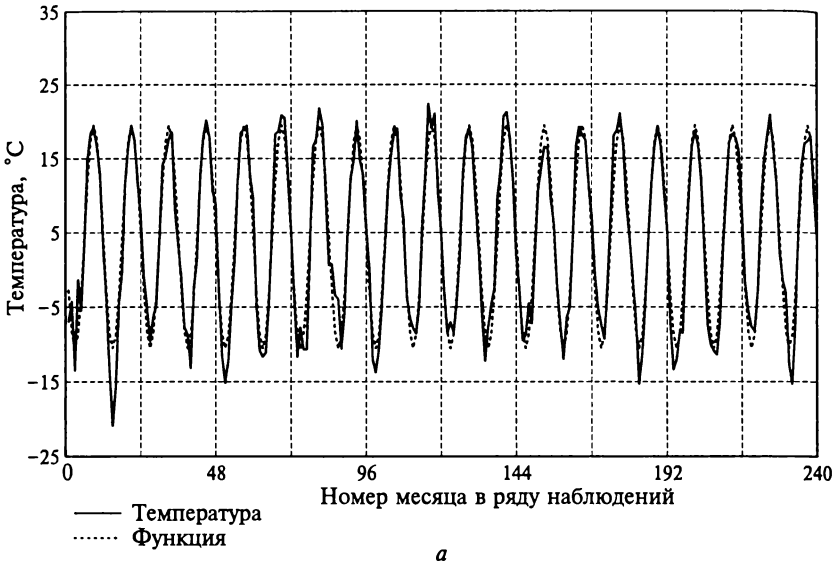
где a_0, a, b — неизвестные константы; λ — круговая частота, априори вводимый параметр; t — номер члена ряда (аргумент), $t = 0, 1, 2, 3, \ldots, L$.

Возьмем в качестве примера ряд измерения среднемесячных температур на метеостанции «Рязань» (рис. 9.3).

Длина ряда $L = 1176$ дат (месяцев). Данные охватывают интервал от декабря 1892 г. по ноябрь 1989 г.

Рассчитаем уравнение регрессии для периода в 12 месяцев и соответственно для частоты $\omega = 1/12$ с круговой частотой $\lambda = 2\pi/12$. Для этого, пользуясь стандартными средствами, построим функции косинуса и синуса от этой частоты и рассчитаем параметры уравнения регрессии для всего ряда (табл. 9.1).

Соответственно функция для периода 12 месяцев, почти на 94 % описывающая варьирование температур всего ряда, есть



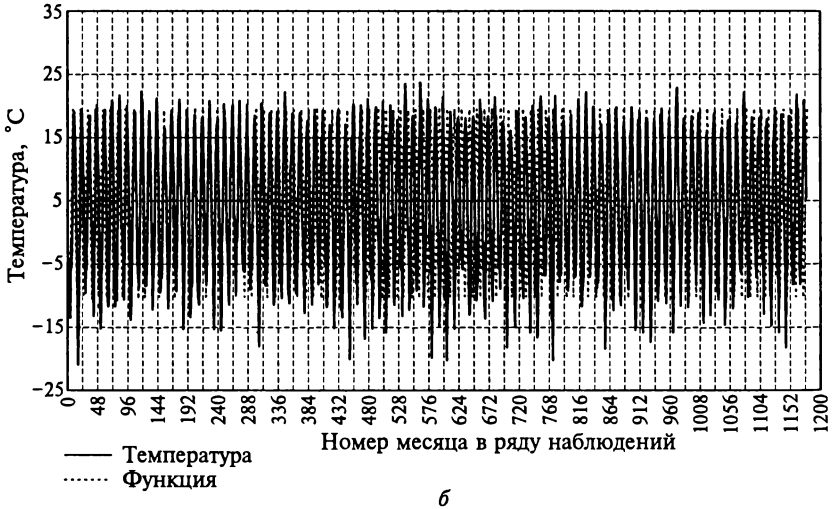


Рис. 9.3. Аппроксимация ряда месячных температур гармонической функции с периодом 12 месяцев (по данным метеостанции «Рязань»):

а — фрагмент ряда среднемесячных температур; *б* — ряд среднемесячных температур

$$x_t^{\lambda} = 4,49 + 0,37\cos(t \cdot 2\pi / 12) - 15,02\sin(i \cdot 2\pi / 12).$$

На рис. 9.3 показан результат такой аппроксимации для фрагмента (рис. 9.3, *а*) и всего ряда (рис. 9.3, *б*). Естественно, что такую

Таблица 9.1

Модель регрессии «ряд температур — косинус и синус» с периодом 12 месяцев

Коэффициент детерминации $R^2 = 0{,}9369$; критерий Фишера F = 8710,9; уровень значимости p < 0,0000; стандартная ошибка 2,7606

Переменная	*b*	Средняя квадратическая ошибка	t(1173)	p-уровень (p-level)
Константа	4,4910	0,080500	55,789	0,000000
Косинус	0,3700	0,113844	3,250	0,001186
Синус	−15,0219	0,113844	−131,951	0,000000

модель множественной регрессии можно построить для всех периодов от 2 до $L/2$.

Все множество косинусов и синусов однозначно опишет любой временной ряд. Вклад гармоники в описание ряда определяется величиной коэффициентов *a* и *b*. Заметим, что коэффициенты a_λ при косинусах и коэффициенты *b* при синусах — это *коэффициенты регрессии*, показывающие степень, с которой соответствующие функции коррелируют с данными. Отметим, что сами синусы и косинусы на различных частотах некоррелированны или, иначе говоря, ортогональны. Таким образом, имеем дело с частным случаем разложения ряда по ортогональному базису. Так как каждой гармонике ставится в соответствие два аргумента с косинусом и синусом, а число гармоник равно половине длины ряда, фактически рассматривается модель множественной регрессии, в которой число аргументов равно его длине. Такое число аргументов, очевидно, полностью воспроизводит все значения ряда и содержит всю информацию о нем.

Квадрат амплитуды, или величину дисперсии, связанные с конкретной частотой или периодом, можно определить по формуле

$$P_\omega = a_\omega^2 + b_\omega^2 .$$

Функция дисперсии от частоты или периода называется *периодограммой*. Ее величина, связанная с конкретной частотой или периодом, прямо показывает вклад связанных с ними гармоник в описание варьирования значений исследуемого временного ряда. На рис. 9.4 периодограмма представлена в логарифмической шка-

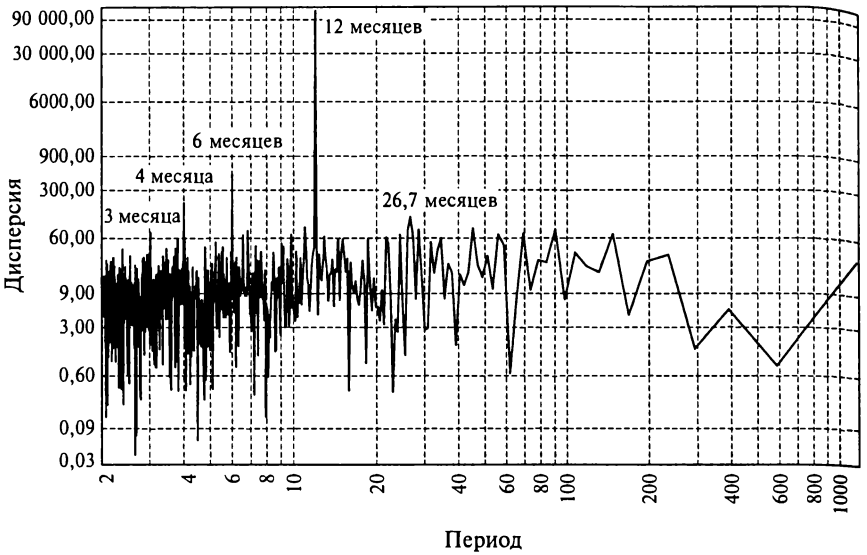


Рис. 9.4. Периодограмма для ряда среднемесячных температур (по данным метеостанции «Рязань»)

ле, что необходимо, так как на гармонику с периодом 12 месяцев приходится бóльшая часть всего варьирования и при отображении в натуральной шкале не было бы видно дисперсий, соответствующих остальным гармоникам. В логарифмическом представлении хорошо видны и другие экстремумы, выделяющие гармоники, которые вносят существенный вклад в варьирование временного ряда.

Периодограмма весьма чувствительна к локальным флуктуациям, поэтому часто используются ее осредненные значения, полученные с помощью различных вариантов сглаживания. При этом обычно задается размер окна, для которого определяется осреднение.

Статистическую значимость каждого максимума дисперсии можно определить обычными методами в регрессионной модели. Допустим, нас интересует, имеют ли статистическую значимость периоды 27,349; 26,73; 26,13 месяцев с дисперсиями соответственно 64,68; 87,63; 65,31.

Построим соответствующую регрессионную модель. Полученные на ее основе оценки ($F = 0,35$; $p < 0,9$) показывают, что вклад этих гармоник в варьирование ряда статистически незначим и не превышает уровня шума. Вклад гармоники с периодом 4 также недостоверен и на пределе значимости можно принять достоверным вклад в варьирование гармоники с периодом 6 месяцев ($F = 2,08$; $p < 0,12$).

Временно́й ряд можно разделить на составляющие типы процессов.

Шаг первый. *Определить, не является ли весь ряд «белым шумом».* Если ряд отображает процесс «белого шума», то его автокорреляционная функция (АКФ), как общая, так и частная (ЧАКФ), нигде не выходит за доверительные интервалы.

Рассмотрим в качестве примера январский ряд среднемесячных температур на метеостанции «Рязань». В январе в соответствии с одномерным анализом распределение было нормальным и по этому критерию ряд — стационарный. Из рис. 9.5 (АКФ) следует, что коэффициент корреляции нигде не выходит за границы доверительных интервалов, Q-статистика (Box-Ljung statistic) по смыслу идентичная χ^2, показывает, что всюду с вероятностью 0,9 автокорреляционная функция отсутствует. Тот же результат дает и частная автокорреляционная функция (ЧАКФ). Следовательно, можно однозначно утверждать, что на протяжении века колебания среднемесячных температур в январе в Рязани не выходили за рамки «белого шума» и условия в январе на протяжении всего времени наблюдения были стационарны. Стационарность сохраняется вплоть до мая, но в мае картина резко меняется.

Автокорреляция, соответствующая шагу в 4 года и в 8 лет (рис. 9.6), резко выходит за границу случайного процесса, маркируя существование гармоники с периодом около 4 года. При этом в интервале от 8 до 24 лет в соответствии с Q-статистикой процесс не является чисто случайным. Следует отметить, что в остальные месяцы за исключением ноября колебания температур не выходят за рамки «белого шума».

Периодограмма ряда выделяет период в 4,66 года. Регрессионная модель для этой гармоники статистически достоверна с $R^2 = 0{,}13$ и $F = 7{,}3898$ при $p < 0{,}00104$ (рис. 9.7). Соответственно в этот месяц в динамике температур, с большой вероятностью, существует гармоническая составляющая.

Совершенно иной вид имеет АКФ для сумм осадков в январе (рис. 9.8). Она демонстрирует, что ряд статистически значимо отличен от «белого шума». При этом величина корреляции почти линейно уменьшается по мере увеличения шага сдвига ряда относительно самого себя до шага в 34 года. Такой вид автокорреляционной функции однозначно указывает на существование в ряду трендовой составляющей, причем возможно соответствующей полиному степени больше двух.

Рассмотрев три принципиально различных вида функции автокорреляции: «белый шум», периодическое достоверное увеличение коэффициента корреляции при определенном шаге, постепенное уменьшение значения автокорреляции, индуцирующее существование тренда, перейдем к решению следующей задачи.

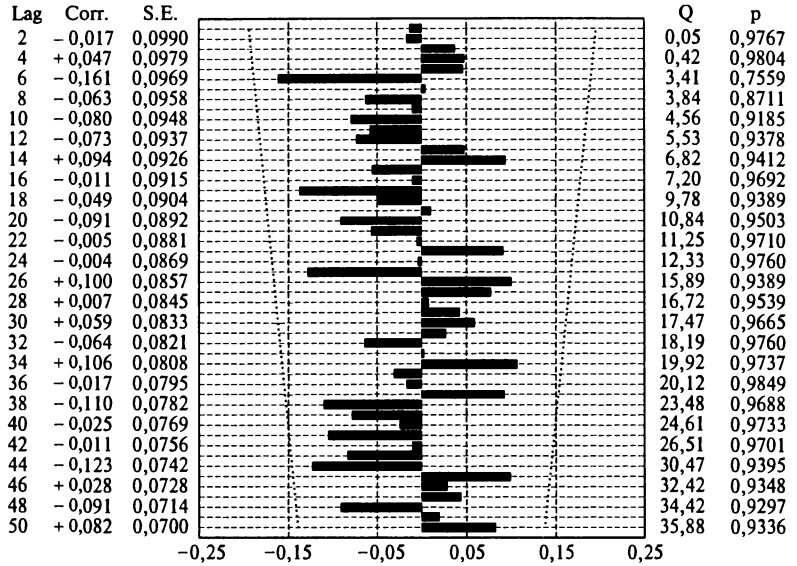


Рис. 9.5. Автокорреляционная функция для ряда средних температур января (по данным метеостанции «Рязань»). Стандартная ошибка — оценка границы «белого шума»

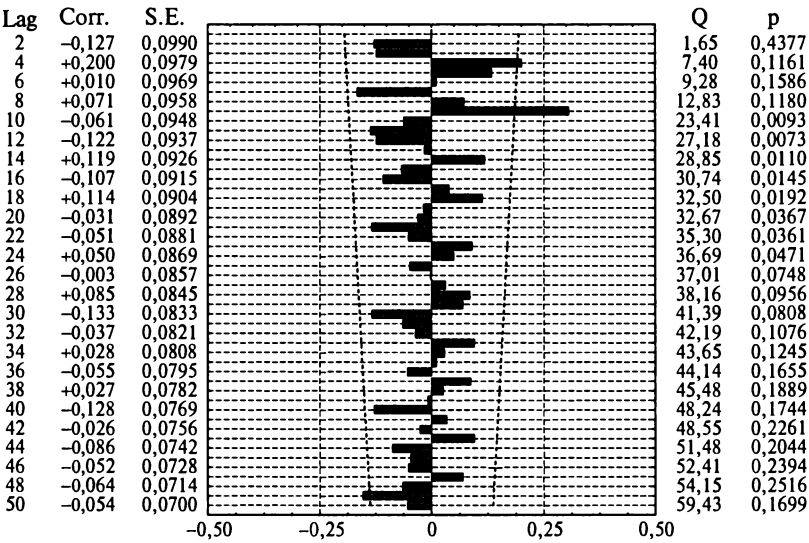


Рис. 9.6. Автокорреляционная функция для ряда средних температур мая (по данным метеостанции «Рязань»). Стандартная ошибка — оценка границы «белого шума»

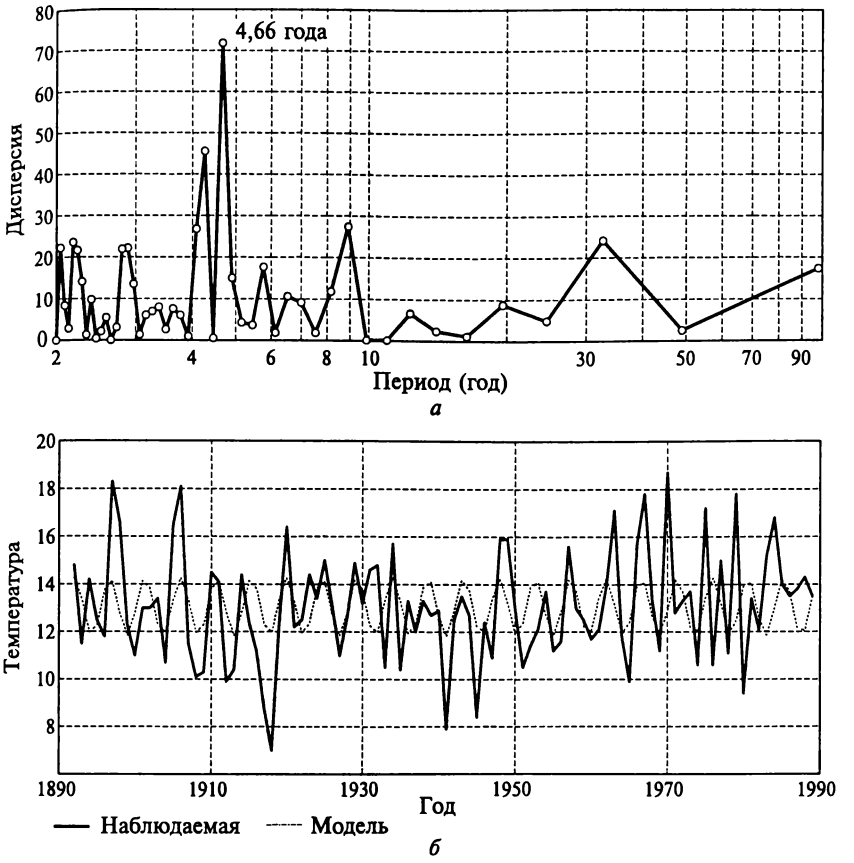


Рис. 9.7. Периодограмма (*а*) ряда средних температур мая и аппроксимация ряда гармоникой с периодом 4,66 года (по данным метеостанции «Рязань»)

Шаг второй. *Исключение тренда.* С помощью оперативных средств анализа можно установить — действительно ли существование полиномиального тренда (рис. 9.9). Для более точной оценки можно использовать специальные программные блоки, позволяющие рассчитать полином, определить его параметры и отделить от полиномиального тренда варьирование ряда, определяемого иными механизмами (табл. 9.2).

Из табл. 9.2 следует, что многолетняя динамика осадков описывается на 19 % полиномом четвертой степени с параметрами регрессионной модели, представленными в табл. 9.3.

Очевидно, что построение полинома осуществляется по модели многомерной регрессии и так же, как и в общем случае, здесь может использоваться модель пошаговой регрессии.

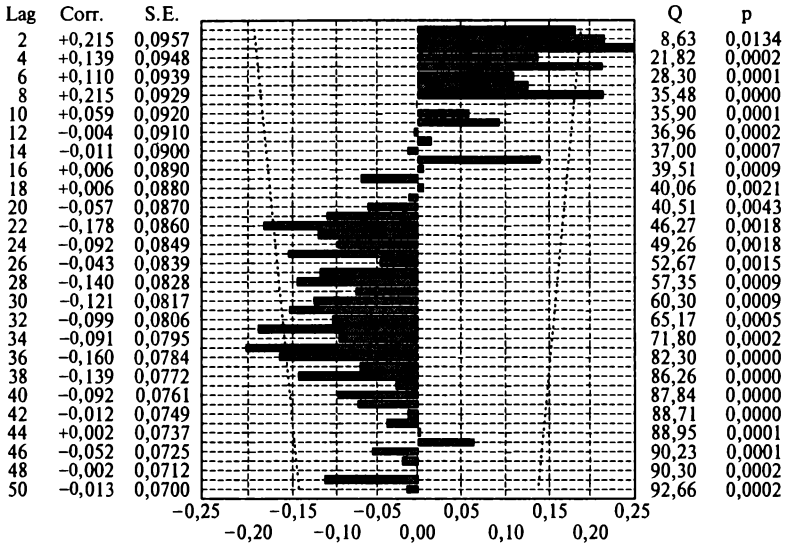


Рис. 9.8. Автокорреляционная функция ряда месячных сумм осадков января (по данным метеостанции «Рязань»). Стандартная ошибка — оценка границы «белого шума»

Таким образом, динамика осадков в январе за сто лет является нестационарным процессом с полиномиальным трендом четвертой степени. Возможно, что этот тренд отражает гармоническое

Таблица 9.2

Тест полиномиальной модели четвертой степени для корня квадратного из суммы январских осадков

R^2 = 0,190535; F = 5,88; p = 0,000271; ошибка 2,26
Effective hypothesis Decomposition

Переменная	Сумма квадратов	Число степеней свободы	Среднее квадратическое отклонение	F-критерий	Уровень значимости p-уровень
Константа	26,1337	1	26,13366	11,54119	0,000978
Год	30,5595	1	30,55946	13,49571	0,000386
(Год)^2	31,7568	1	31,75679	14,02448	0,000302
(Год)^3	27,6956	1	27,69560	12,23097	0,000703
(Год)^4	22,6478	1	22,64778	10,00174	0,002071
Ошибка	226,4383	100	2,26438		

Таблица 9.3

Параметры полинома четвертой степени, описывающего варьирование месячных сумм осадков

Переменная	Параметр	Средняя квадратическая ошибка	t-критерий	Уровень значимости	Доверительный интервал	
					–95,00 %	+95,00 %
Константа	2,959704	0,871211	3,39723	0,000978	1,231247	4,688161
Год	0,397949	0,108325	3,67365	0,000386	0,183035	0,612863
(Год)^2	–0,014980	0,004000	–3,74493	0,000302	–0,022916	–0,007044
(Год)^3	0,000193	0,000055	3,49728	0,000703	0,000084	0,000303
(Год)^4	–0,000001	0,000000	–3,16255	0,002071	–0,000001	–0,000000

колебание с периодом около 60 лет, однако ряд слишком короток, чтобы на его основе строго доказать существование такой периодичности.

Шаг третий. *Анализ остатков.* Если из значений ряда вычесть значения тренда, то получим остатки, для которых также необходимо проверить гипотезу соответствия образуемого ими ряда «белому

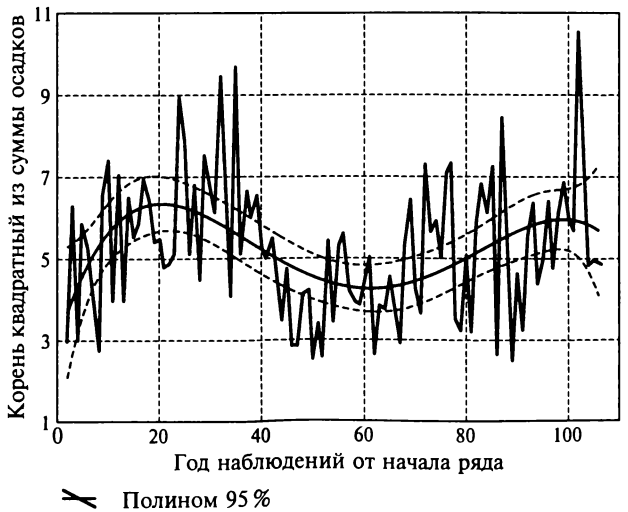


Рис. 9.9. Вековая динамика сумм осадков в январе и его полиномиальный тренд (начало наблюдений — 1885 г.)

шуму». Как следует из рис. 9.10, коэффициент автокорреляции практически нигде не выходит за границы случайной ошибки и его варьирование можно трактовать как «белый шум». Стандартизованные отклонения только очень редко достигают значения 3 (маловероятные события), что вполне допустимо для ряда длиной 100 лет. Соответственно, можно полагать, что в ряду нет выбросов. Однако если рассматривать производную от ряда, т.е. приращение осадков от одного года к другому, то ряд имеет строгую и достоверную регулярность. Суть ее сводится к следующему: если осадки в прошлом году увеличились по отношению к предыдущему, то на будущий год они, с большой вероятностью, уменьшатся (отрицательный коэффициент корреляции при шаге в один год (рис. 9.11)).

Шаг четвертый. *Выделение регулярной составляющей.* Таким образом, в производных наблюдаемого ряда есть, безусловно, регулярная составляющая.

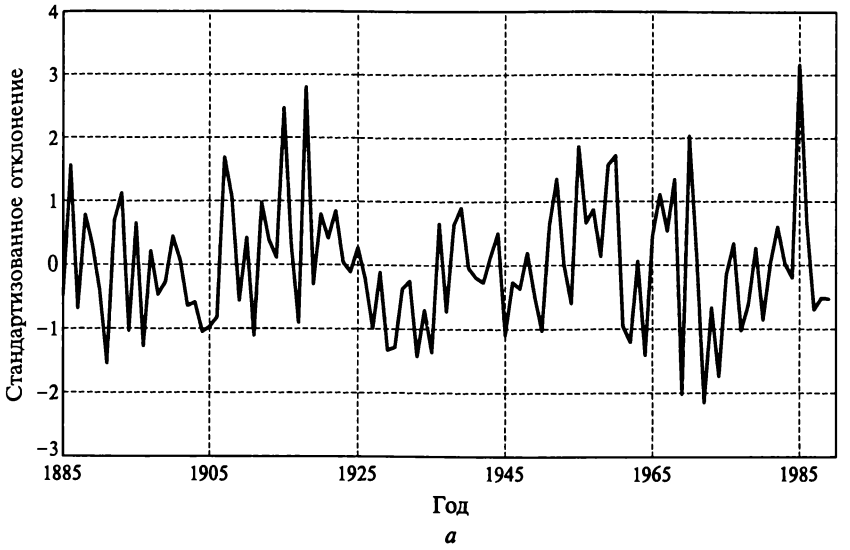
Периодограмма (рис. 9.12, *а*) показывает, что основная мощность колебаний приращения приходится на период 2,6 года. На рис. 9.12, *б* приведена сглаженная периодограмма, которую обычно называют спектром, при окне сглаживания 10 лет методом Хемминга. В подписи к графику приведены значения весов для 1, 2, ..., 10 лет, с которыми осуществляется осреднение. Осреднение убирает лишние пики и дает более наглядное отражение возможной структуры гармонических колебаний. Регрессионная модель (рис. 9.12, *в*) показывает что гармоники 2,6; 2,66 и 3,7 лет статистически значимы и совместно описывают 30 % варьирования при $F = 5{,}12$ со статистическим уровнем значимости $p < 0{,}00003$.

Выделение сезонной составляющей. Частным случаем регулярных колебаний является обычная для временных рядов сезонная составляющая. От строго гармонической составляющей она отличается тем, что может описываться несколькими гармониками.

Выделение сезонной составляющей методически близко к дисперсионному анализу. В качестве классов рассматриваются месяцы при показателе сезонности 12, а в общем случае — значения ряда с классами, выделяемыми номерами, кратными заданному интервалу сезонности. В рамках анализа временных рядов при сезонной декомпозиции, кроме сезонной составляющей, обычно выделяется остаток, сезонно сглаженный тренд и нерегулярная компонента, обычно отражающая «белый шум» или высокочастотную часть.

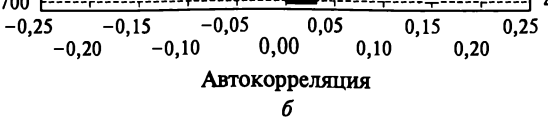
Рассмотрим эту операцию на примере месячных сумм осадков для метеостанции «Рязань» (табл. 9.4).

На рис. 9.13 показаны компоненты, обычно выделяемые при сезонном анализе ряда. Рассмотрим периодограммы выделенных рядов. Спектр собственно сезонной динамики осадков показывает, что этот ход может быть описан гармониками с периодами 12, 6, 4, 3 и 2,4 месяца. Соответственно частоты этих гармоник 0,08333(3), 0,1666(6), 0,25(4), 0,333(3), 0,4166(6) (рис. 9.14). Раз-



а

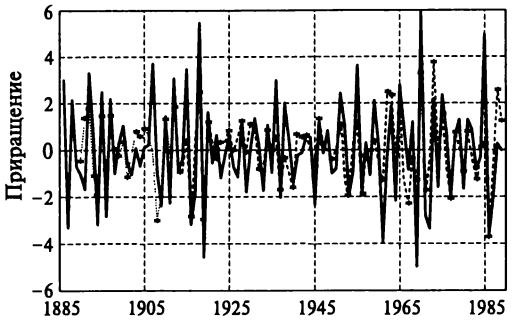
Lag	Corr.	S.E.	Q	p
2	+0,032	0,0957	0,12	0,9439
4	−0,044	0,0948	3,62	0,4603
6	−0,056	0,0939	4,38	0,6257
8	+0,107	0,0929	5,70	0,6810
10	−0,070	0,0920	9,43	0,4914
12	−0,122	0,0910	11,32	0,5020
14	−0,106	0,0900	13,73	0,4698
16	−0,066	0,0890	15,48	0,4900
18	−0,039	0,0880	18,32	0,4349
20	−0,060	0,0870	18,96	0,5246
22	−0,146	0,0860	23,00	0,4020
24	−0,028	0,0849	23,68	0,4802
26	+0,067	0,0839	25,08	0,5143
28	−0,017	0,0828	25,15	0,6199
30	+0,040	0,0817	26,31	0,6592
32	+0,062	0,0806	26,91	0,7221
34	+0,069	0,0795	27,74	0,7669
36	−0,020	0,0784	28,31	0,8159
38	−0,017	0,0772	29,58	0,8337
40	+0,016	0,0761	31,78	0,8197
42	+0,082	0,0749	33,10	0,8353
44	+0,076	0,0737	34,55	0,8456
46	+0,006	0,0725	38,68	0,7694
48	+0,046	0,0712	39,39	0,8073
50	+0,032	0,0700	40,95	0,8155



б

Рис. 9.10. Вид остатков ряда сумм осадков в январе после исключения полиномиального тренда (по данным метеостанции «Рязань»):

а — стандартизованные отклонения от полинома четвертой степени; *б* — автокорреляционная функция для отклонений. Стандартная ошибка — граница «белого шума»

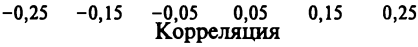


а

Lag	Corr.	S.E.	Q	p
2	−0,061	0,0962	28,44	0,0000
4	−0,156	0,0952	34,89	0,0000
6	−0,085	0,0943	36,82	0,0000
8	+0,192	0,0933	41,15	0,0000
10	+0,031	0,0923	45,47	0,0000
12	−0,063	0,0914	46,46	0,0000
14	−0,111	0,0904	48,04	0,0000
16	−0,041	0,0893	52,38	0,0000
18	+0,057	0,0883	53,99	0,0000
20	−0,005	0,0873	54,04	0,0001
22	−0,064	0,0862	54,64	0,0001
24	+0,036	0,0852	54,89	0,0003
26	+0,098	0,0841	57,55	0,0004
28	−0,063	0,0830	58,15	0,0007
30	+0,008	0,0819	58,97	0,0012
32	+0,076	0,0808	60,50	0,0017
34	+0,105	0,0797	63,24	0,0017
36	−0,035	0,0785	64,57	0,0024
38	−0,121	0,0774	68,97	0,0016
40	−0,051	0,0762	71,42	0,0017
42	+0,044	0,0750	71,86	0,0028
44	−0,026	0,0738	72,16	0,0047
46	−0,087	0,0725	75,93	0,0036
48	+0,066	0,0713	76,82	0,0052
50	+0,107	0,0700	82,02	0,0029

б

Lag	Corr.	S.E.	Q	p
2	−0,037	0,0980	0,15	0,9272
4	−0,095	0,0970	1,54	0,8193
6	−0,071	0,0960	3,22	0,7804
8	+0,146	0,0950	5,58	0,6938
10	−0,042	0,0939	8,12	0,6173
12	−0,079	0,0929	8,89	0,7122
14	−0,027	0,0918	9,48	0,7989
16	−0,002	0,0907	13,63	0,6262
18	+0,018	0,0897	14,02	0,7280
20	+0,020	0,0886	14,25	0,8173
22	−0,119	0,0874	17,30	0,7464
24	−0,029	0,0863	17,49	0,8271
26	+0,071	0,0852	18,21	0,8679
28	+0,021	0,0840	18,30	0,9182
30	+0,060	0,0828	19,04	0,9392
32	+0,000	0,0816	19,15	0,9645
34	+0,001	0,0804	19,16	0,9809
36	−0,082	0,0792	21,00	0,9781
38	−0,028	0,0780	21,70	0,9844
40	−0,060	0,0767	23,69	0,9811
42	−0,011	0,0754	23,75	0,9895
44	+0,053	0,0741	24,48	0,9924
46	−0,003	0,0728	27,55	0,9858
48	−0,005	0,0714	27,63	0,9919
50	−0,019	0,0700	28,68	0,9933



в

Рис. 9.11. Вид первой производной (приращение) сумм осадков января (по данным метеостанции «Рязань»):

а — динамика приращения осадков; *б*, *в* — АКФ для производной ряда остатков от полинома (*б*) и от модели (*в*). Стандартная ошибка — граница «белого шума»

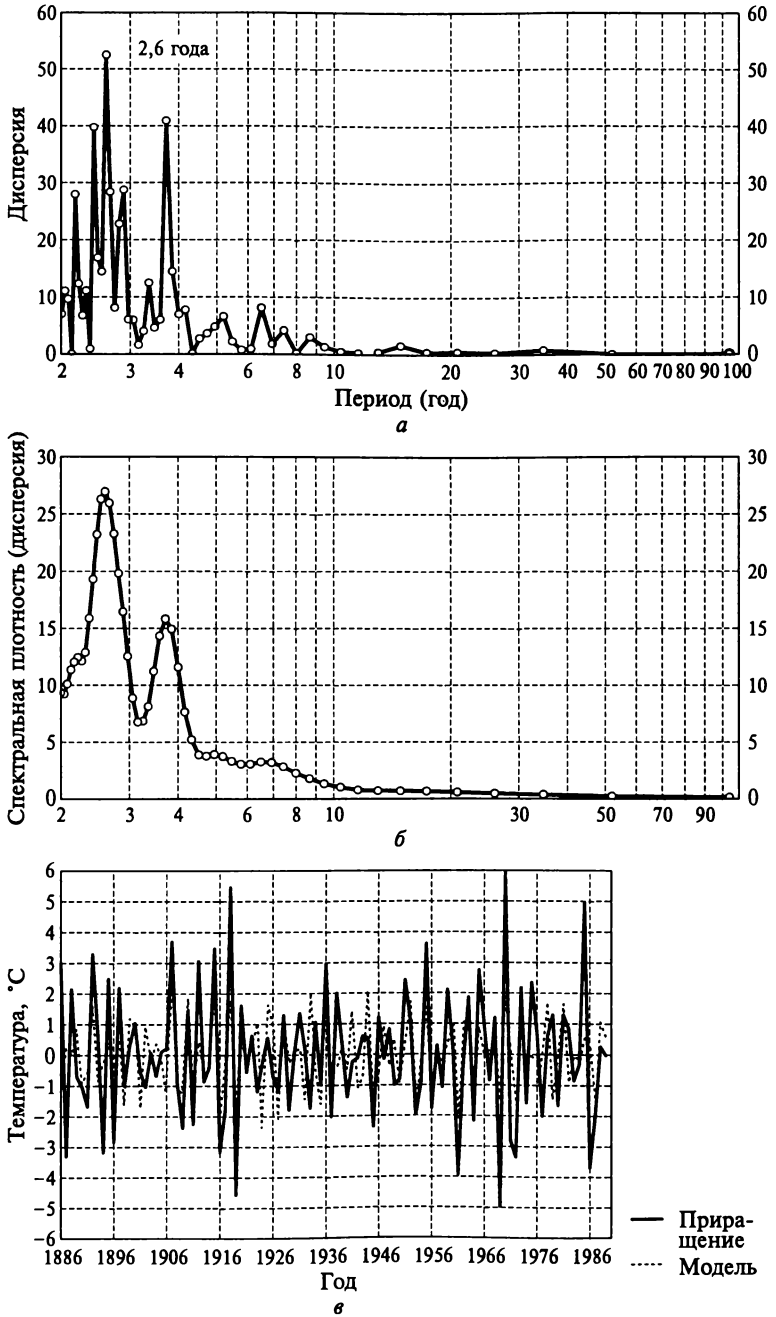
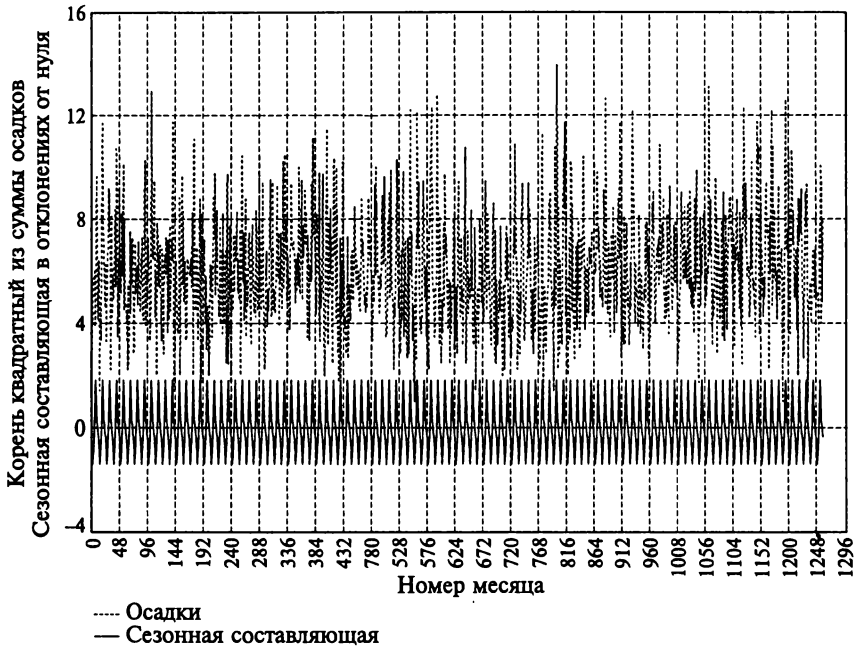
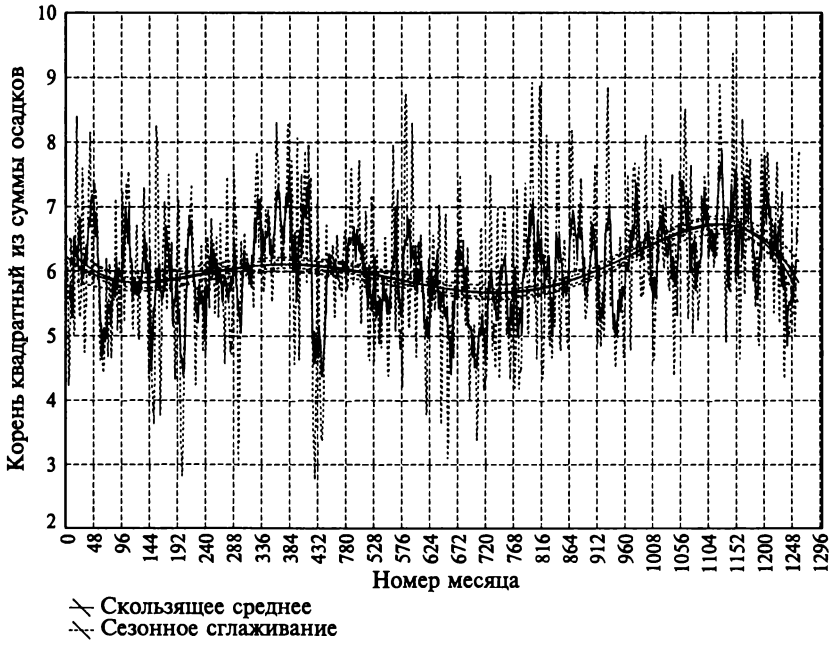


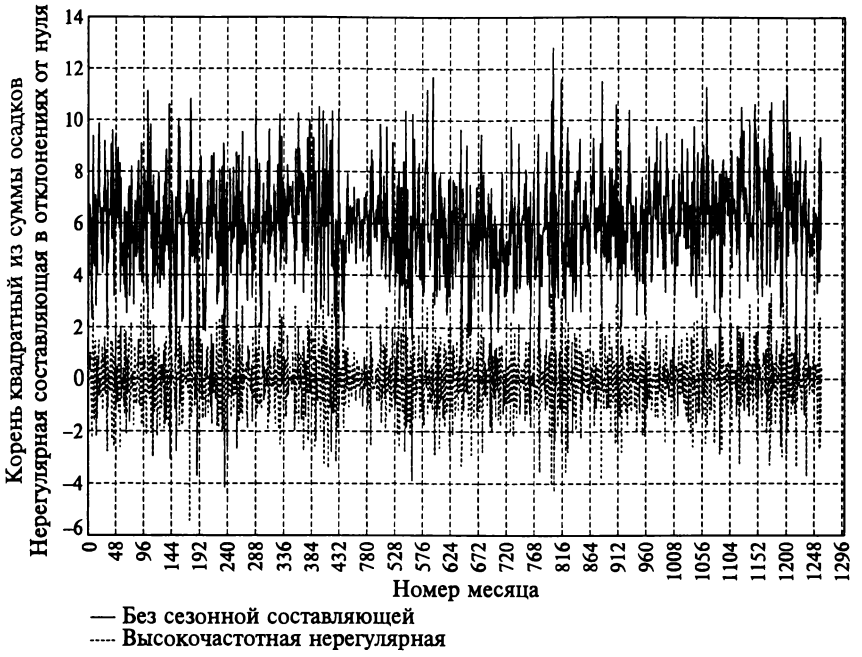
Рис. 9.12. Периодограмма (*а*), слаженный спектр (*б*) и гармоническая составляющая (*в*) ряда производных осадков за январь (по данным метеостанции «Рязань»)


Корень квадратный из суммы осадков
Сезонная составляющая в отклонениях от нуля
16
12
8
4
0
-4
0 48 96 144 192 240 288 336 384 432 780 528 576 624 672 720 768 816 864 912 960 1008 1056 1104 1152 1200 1248 1296
Номер месяца
Осадки
Сезонная составляющая

а


Корень квадратный из суммы осадков
10
9
8
7
6
5
4
3
2
0 48 96 144 192 240 288 336 384 432 780 528 576 624 672 720 768 816 864 912 960 1008 1056 1104 1152 1200 1248 1296
Номер месяца
Скользящее среднее
Сезонное сглаживание

б



в

Рис. 9.13. Разложение ряда среднемесячных сумм осадков на составляющие методом сезонной декомпозиции (по данным метеостанции «Рязань»):

а — месячные осадки; *б* — сглаживание; *в* — компоненты ряда

Таблица 9.4

Вклад в варьирование ряда среднемесячных сумм осадков различных его составляющих и вариантов сглаживания

Переменная	R^2	F-критерий	Уровень значимости p-уровень	Средняя квадратическая ошибка
Сглаживание скользящим средним с окном 12 месяцев	0,28970250	114,25	< 0,00000	2,0313
Сезонная составляющая	0,17596621	266,29	< 0,0000	1,9265
Остаток сезонной составляющей	0,73305833	3424,4	0,0000	1,0965
Сезонное сглаживание	0,39935735	829,11	0,0000	1,6448
Нерегулярная составляющая	0,52872816	1399,0	0,0000	1,4569

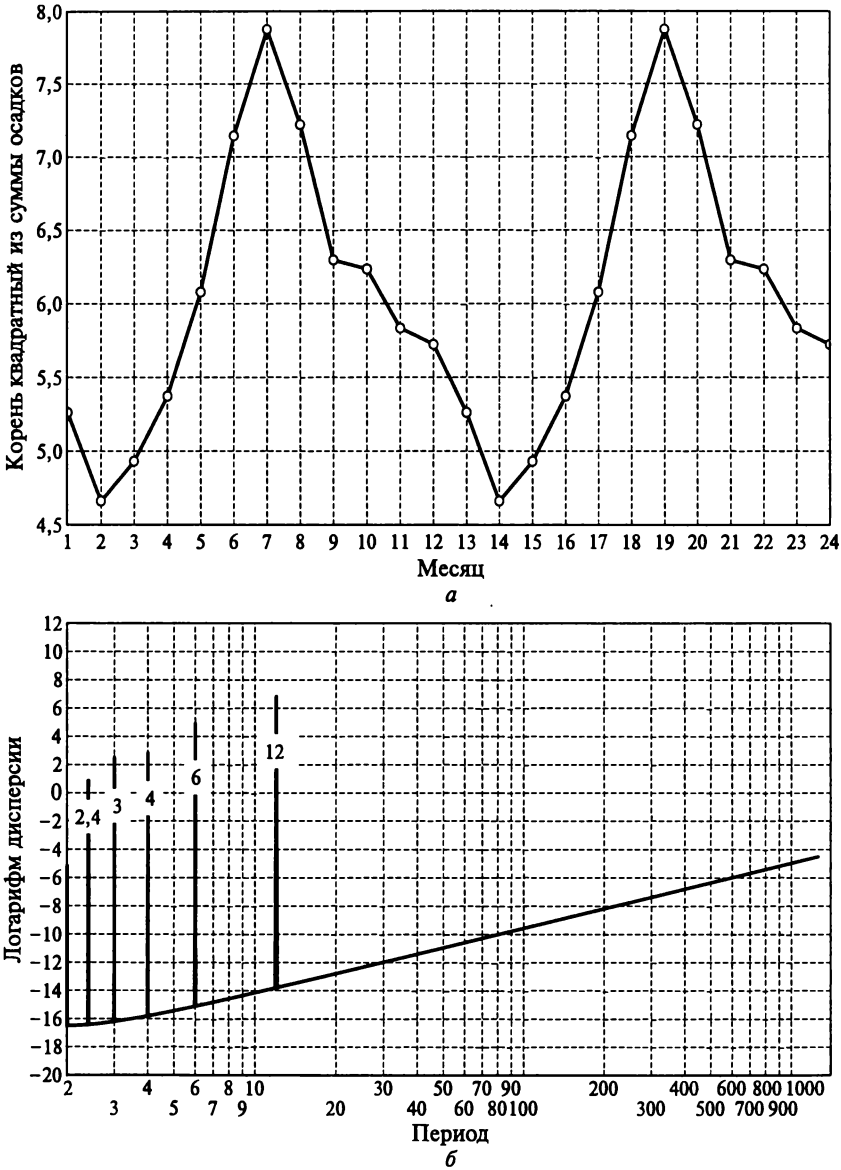


Рис. 9.14. Сезонная составляющая ряда осадков для периода 24 месяца (*а*) и его периодограмма (*б*) — по данным метеостанции «Рязань»

ность между соседними частотами есть константа, равная 0,08333(3). По существу этот ряд описывает рассмотренное выше важнейшее правило отношения гармоник при нелинейных колебаниях. Каждая следующая волна имеет частоту, отличающуюся от предыдущей на постоянную величину.

Таким образом, если основная частота равна 0,08833, то все остальные отличаются от нее на $0{,}0833n$ (n — номер волны). Собственно разложение ряда на ортогональные составляющие показывает, что для его полного воспроизведения нужно суммировать с соответствующими значениями коэффициентов пять уравнения вида (9.1) при $\lambda = 2\pi \cdot 0{,}08333$, $\lambda = 2\pi \cdot 0{,}16666$, $\lambda = 2\pi \cdot 0{,}25$, $\lambda = 2\pi \cdot 0{,}33333$ и $\lambda = 2\pi \cdot 0{,}4166$.

Следует иметь в виду, что эти гармонические составляющие могут иметь вполне определенный физический смысл.

Сглаживание методом скользящего среднего — стандартная операция, позволяющая снизить вклад в варьирование ряда высокочастотной и шумовой составляющих. На рис. 9.15, *а*, *б* показан сам ряд скользящего среднего и его периодограмма в логарифмическом масштабе. Из периодограммы следует, что сглаженный ряд описывает важное правило колебаний осадков: чем больше период, тем в среднем больше амплитуда колебаний, или дисперсия. Такое свойство типично для нелинейных колебаний и фрактального процесса.

Судя по периодичности изменения дисперсии колебаний в данном случае правильнее говорить о нелинейных колебаниях с ярко проявляющимся эффектом самоподобия.

На основе зависимости «логарифм спектральной плотности — логарифм частоты колебания» можно определить и фрактальную размерность:

$$\log \mathrm{Sp} = a + b \log(1/P),$$

где Sp — значение дисперсии по периодограмме; a и b — константы; P — период колебаний.

Фрактальная размерность определяется по формуле $\mathrm{D} = (5 - b)/2$ (табл. 9.5).

Размерность D = 1,56 практически точно соответствует тому, что принято называть «бурым шумом». При «черном шуме» размерность D близка к единице, а график ряда имеет форму сочетания разномасштабных «холмов». При размерности, близкой к двум, определяемой как «розовый шум», график напоминает композицию вложенных друг в друга разномасштабных зубцов. «Бурый шум» типичен для процессов диффузии и случайного блуждания, что возможно близко к механизмам формирования атмосферных осадков.

Как следует из рис. 9.15, *б*, изменение дисперсии, безусловно, есть гармоническая функция от частоты колебаний или волнового числа. На рис. 9.16, *а*, *б* представлены остатки от рассмотренного выше уравнения регрессии «логарифм спектральной плотности — логарифм частоты колебания» и периодограмма этих остатков. Периодограмма остатков выявляет константы, на которые сдвигаются локальные максимумы спектральной плотности. Наиболее выраженная величина сдвига равна 104 волновым числам. Максимумы гар-

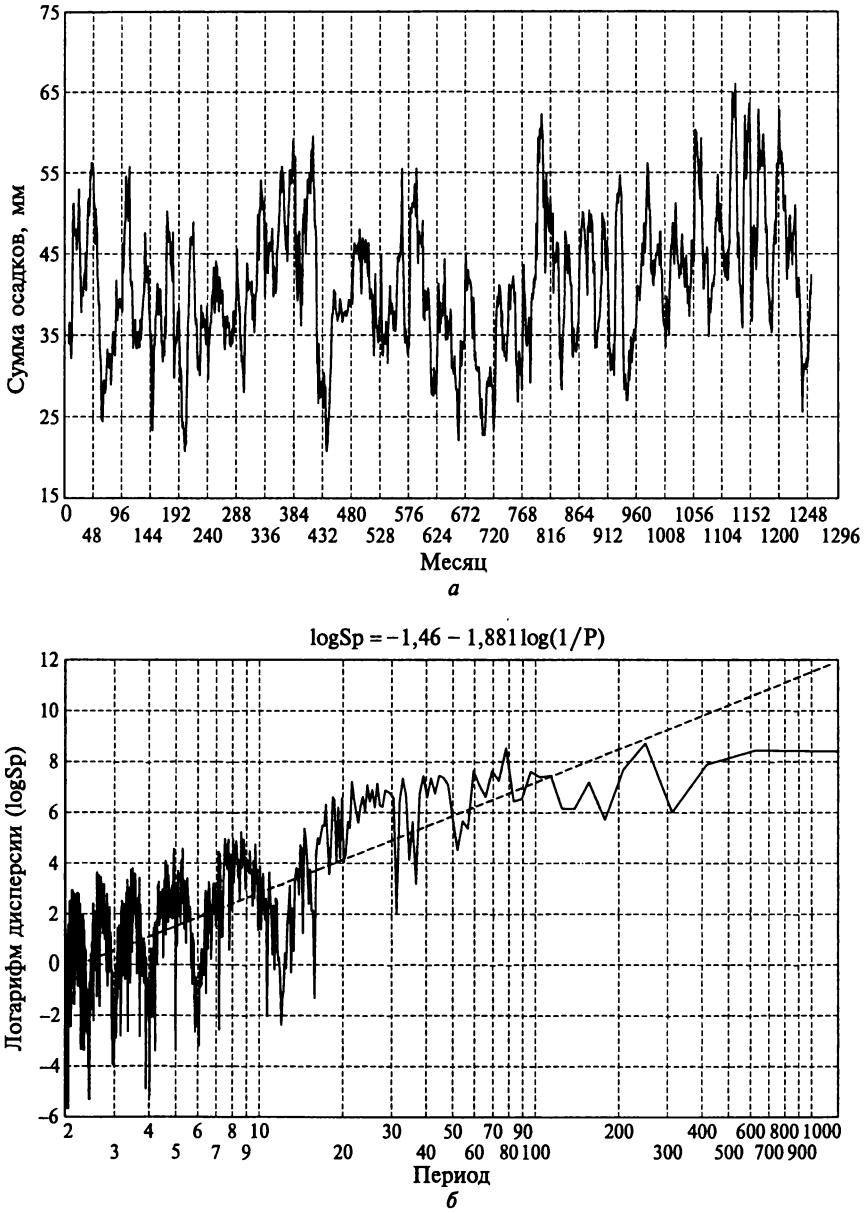


Рис. 9.15. Ряд осадков, сглаженный 12-летним скользящим средним (*а*), и его периодограмма (*б*)

моник исходного сглаженного ряда соответственно приходятся на волновые числа 52, 52 + 104 = 156, 260, 364, 468, 572. Соответственно, периоды колебаний ($P = L/k$) находят по формуле

Таблица 9.5

Оценка параметров уравнения регрессии «логарифм спектральной плотности — логарифм частоты колебания» и фрактальной размерности

R^2 = 0,4894; фрактальная размерность D = (5 – 1,8798)/2 = 1,56

Переменная	*a*	*b*
Оценка параметров	–1,46302	–1,8798
Средняя квадратическая ошибка	0,15012	0,0770
t-критерий	–9,74544	–24,4167
Уровень значимости (p-level)	0,00000	0,0000

$$P_i = L/(k_1 + k_i),$$

где k_1 — волновое число первой гармоники с первым максимумом дисперсии; i = 2, 3, 4, ... — номера волн с локальным максимумом дисперсии.

Те же соотношения можно выразить и через частоту колебаний ω.

Таким образом, в сглаженном скользящем среднем ряду существуют волны со следующими периодами:

первая волна 1248/52 = 24 месяца; вторая волна — 8 месяцев; третья волна — 4 месяца; пятая волна — 3,42 месяца; шестая волна — 2,66(6) месяцев, седьмая волна — 2,18(18) месяцев.

Если допустить, что процесс самоподобен на очень большом интервале времени, например на интервале длиной 2496 месяцев (208 лет), то первый период колебаний с максимальной амплитудой составлял бы 2496/52 = 48 месяцев (4 года) и далее при увеличении длины ряда соответственно 96, 192, 384 месяца (8, 16, 32 года).

Из рис. 9.16 следует, что определенное значение имеют колебания с константой 52, которые соответственно порождают периоды колебаний с менее выраженной амплитудой:

первая — 1248/26 = 48 месяцев (4 года), вторая — 1248/78 = 16 месяцев, третья — 1248/130 = 9,6 месяца и т.д.

Наконец, можно выделить гармоники и с минимальной мощностью с константой 34. Она описывает самый длинный период колебаний, который можно достоверно выделить в этом ряду, составляющий 73 месяца (6,208 лет).

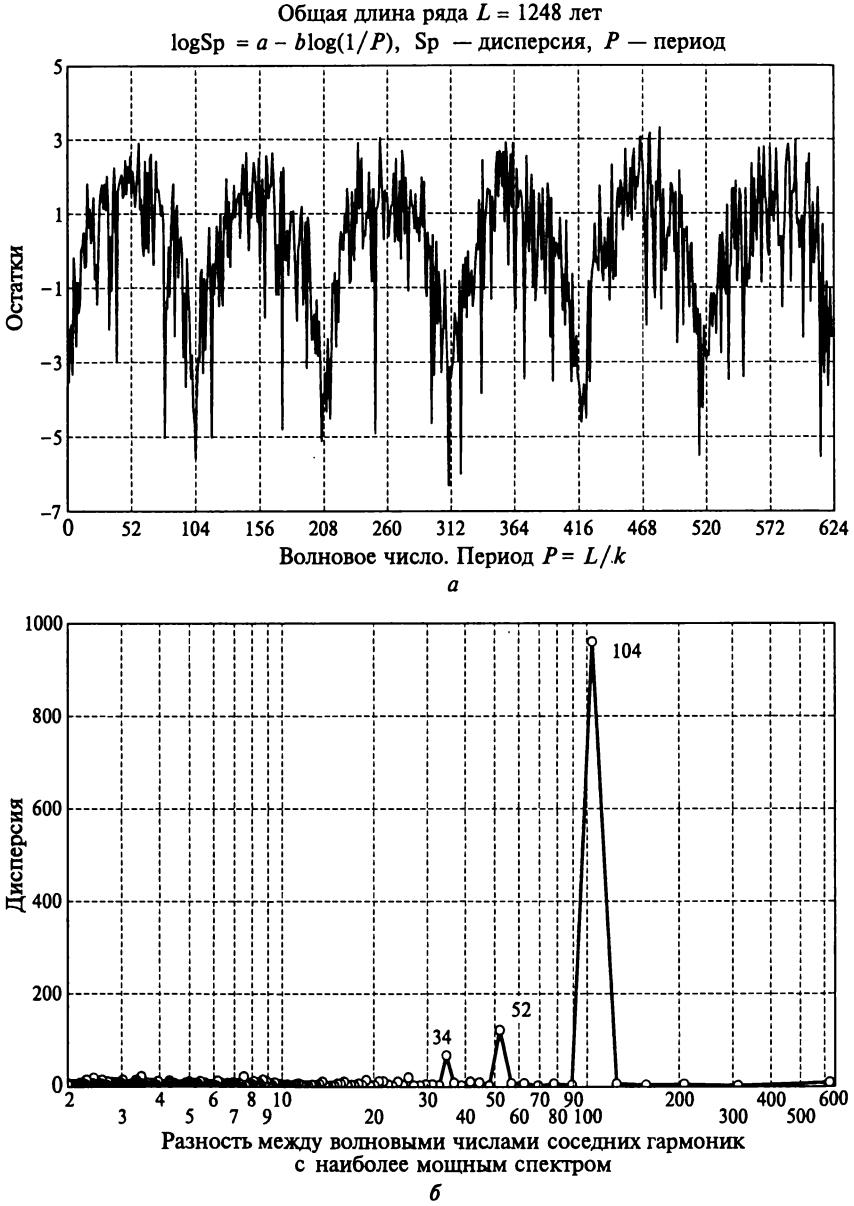


Рис. 9.16. Остатки от спектра сглаженного ряда осадков после удаления линейной зависимости «логарифм периодограммы — логарифм частоты» (*а*) и его периодограмма (*б*)

Итак, можно утверждать, что наряду с сезонной составляющей примерно 29 % варьирования ряда описывается квазигармоническими колебаниями, порождаемыми, скорее всего, автоколеба-

Таблица 9.6

Корреляционная матрица между осадками 12 соседних месяцев по всему ряду наблюдений

Переменная	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	Многомерный коэффициент детерминации R^2
M1	1,00	0,27	0,09	–0,01	–0,08	–0,15	–0,15	–0,17	–0,07	–0,04	0,13	0,22	0,141446
M2	0,27	1,00	0,27	0,09	–0,01	–0,08	–0,15	–0,15	–0,17	–0,07	–0,04	0,13	0,163638
M3	0,09	0,27	1,00	0,27	0,09	–0,01	–0,08	–0,15	–0,15	–0,17	–0,07	–0,04	0,157186
M4	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	–0,08	–0,15	–0,15	–0,17	–0,07	0,154810
M5	–0,08	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	–0,09	–0,15	–0,15	–0,17	0,157809
M6	–0,15	–0,08	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	–0,08	–0,14	–0,14	0,153258
M7	–0,15	–0,15	–0,08	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	–0,08	–0,14	0,153008
M8	–0,17	–0,15	–0,15	–0,08	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	–0,08	0,158443
M9	–0,07	–0,17	–0,15	–0,15	–0,09	–0,01	0,09	0,27	1,00	0,27	0,09	–0,01	0,157587
M10	–0,04	–0,07	–0,17	–0,15	–0,15	–0,08	–0,01	0,09	0,27	1,00	0,27	0,09	0,160731
M11	0,13	–0,04	–0,07	–0,17	–0,15	–0,14	–0,08	–0,01	0,09	0,27	1,00	0,27	0,167747
M12	0,22	0,13	–0,04	–0,07	–0,17	–0,14	–0,14	–0,08	–0,01	0,09	0,27	1,00	0,143211

Таблица 9.7

Нагрузки на компоненты

Коэффициент детерминации 12 компонентами всего ряда месячных сумм осадков $R^2 = 0,91$; F-критерий = 1030,0; средняя квадратическая ошибка 0,64351

Номер компоненты	Нагрузка	Процент от общей дисперсии	Накопленная нагрузка	Накопленный процент от общей дисперсии
1	2,059692	17,16410	2,05969	17,1641
2	2,042938	17,02448	4,10263	34,1886
3	1,065894	8,88245	5,16852	43,0710
4	0,991063	8,25886	6,15959	51,3299
5	0,988014	8,23345	7,14760	59,5633
6	0,803720	6,69767	7,95132	66,2610
7	0,789309	6,57757	8,74063	72,8386
8	0,730679	6,08899	9,47131	78,9276
9	0,697918	5,81598	10,16923	84,7436
10	0,692837	5,77364	10,86206	90,5172
11	0,598163	4,98469	11,46023	95,5019
12	0,539774	4,49811	12,00000	100,0000

ниями большой системы циркуляции атмосферы. Обсуждение этого механизма выходит за рамки настоящего пособия. Здесь важно отметить то, что существуют возможности выделения из ряда квазигармонических колебаний с очень сложной структурой.

Не останавливаясь на анализе спектров остальных составляющих, рассмотрим применение метода главных компонент для разложения временно́го (пространственного) ряда по ортогональному базису. Для этого исследуется матрица корреляций между n-рядами, смещенными относительно друг друга на один шаг (в данном случае шаг равен месяцу). Для временны́х рядов естественно рассматривать матрицу при $n = 12$ (24; 36) переменных. Чем больше переменных включается в анализ, тем полнее будет описываться ряд ортогональным базисом. В табл. 9.6 приведены коэффициенты между суммами месячных осадков, рассчитанные для всего ряда наблюдений. Соседние месяцы коррелируют с максимальной корреляцией всего 0,27. Осадки месяцев, отстающих друг от друга на три (например, январь и апрель), вообще независимы. Максималь-

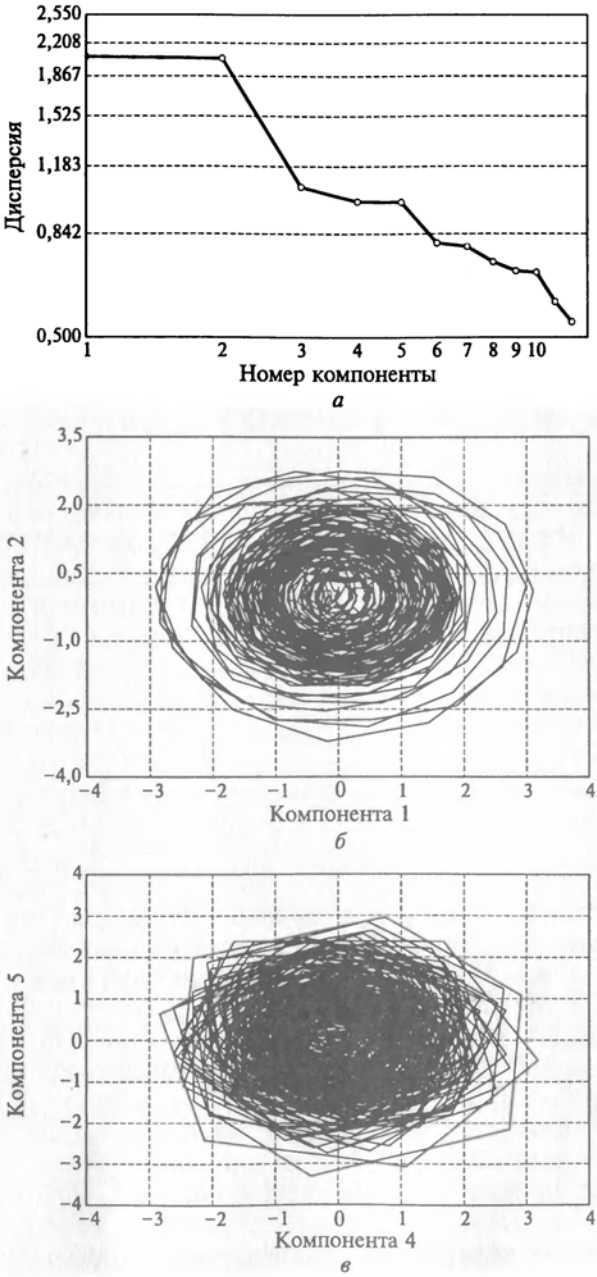
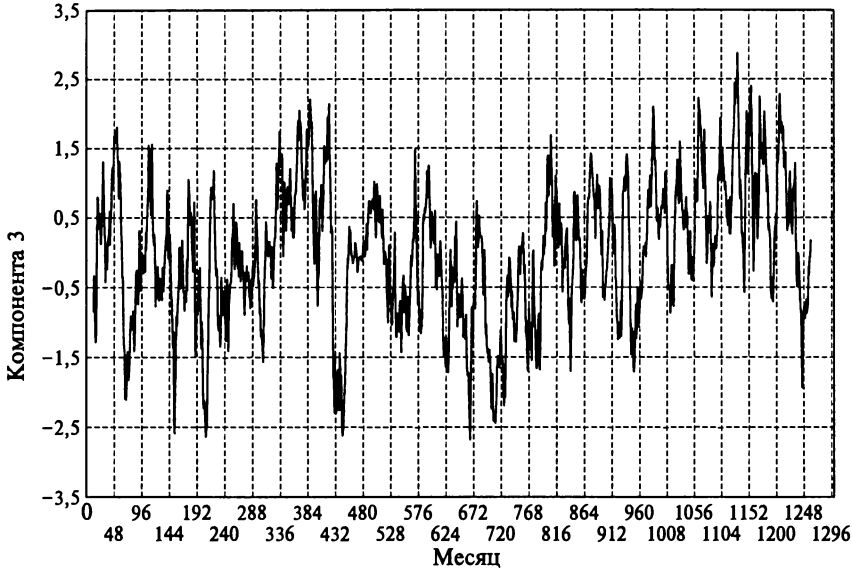
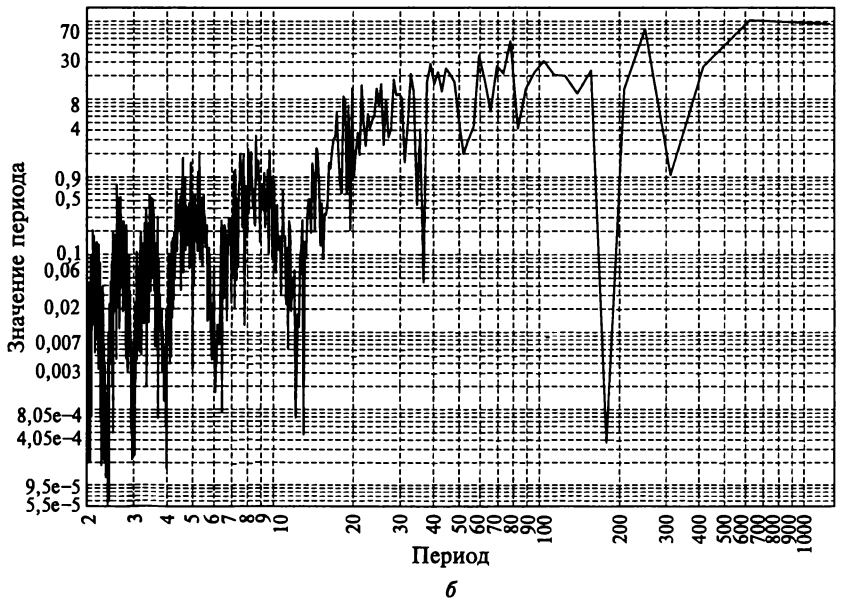


Рис. 9.17. Разложение ряда осадков по ортогональному базису методом главных компонент:

а — нагрузки на компоненты при разложении по 12-мерному базису; *б* — отображение 12-месячного периода в двух первых компонентах; *в* — отображение 6-месячного периода в компонентах 4 и 5



а



б

Рис. 9.18. Третья компонента ряда осадков (*а*) и ее периодограмма (*б*)

ный коэффициент детерминации (0,1677) характерен для ноября. Минимально от остальных месяцев зависят осадки в июле.

Изменение значений нагрузок (табл. 9.7) на компоненты не является монотонной функцией. Первые две компоненты имеют практически одинаковую степень влияния, третья — существенно меньше, четвертая и пятая — вновь близки. В результате таких отношений график имеет ступенчатую форму (рис. 9.17, *а*).

Из рис. 9.17, *б* хорошо видно, что первые две компоненты отображают косинус и синус разложения ряда с периодом 12 месяцев, при этом форма кривых на рисунке по области высокого сгущения линий позволяет визуально выделить область типичных колебаний, а линии, далеко отстоящие от центра, характеризуют так называемый предельный цикл. Компоненты четвертая и пятая отображают синус-косинус разложения ряда с периодом 6 месяцев (рис. 9.17, *в*).

Третья компонента (рис. 9.18) отображает уже хорошо знакомые нелинейные колебания, практически те же, что были получены сглаживанием ряда методом скользящего среднего.

Если рассматривать следующие компоненты, то две их пары опишут циклические колебания с периодами 4 и 3 месяца и высокочастотные самостоятельные колебания, не дополняющие друг друга как косинусы и синусы одного периода. Таким образом, метод главных компонент, будучи строгой схемой разложения переменных по ортогональному базису для анализа временны́х (пространственных) рядов, дает вполне приемлемые результаты, имеющие определенное аналитическое содержание. В общем случае для временны́х (пространственных) рядов при их представлении через n-переменных, получаемых последовательным сдвигом ряда на один шаг относительно самого себя, применимы все методы многомерного анализа. Использование непараметрических методов оправдано в том случае, когда есть сильные подозрения о мультипликативном влиянии прошлых событий на настоящее.

9.3. Методы прогноза на основе временны́х рядов

Если ряд стационарен и имеет постоянную ковариацию, то совершенно очевидно, что она может быть использована для прогноза. При наличии в ряду сезонной составляющей возможность прогноза на ее основе также вполне естественна. Если в ряду существует устойчивый тренд, то, несмотря на стационарность, он также используется для прогноза. Можно полагать, что тренд отражает динамику с очень большим моментом инерции, и движение, определяемое такой динамикой, не может моментально изменить свое направление. И хотя ряд с трендом не является стационарным, прогноз на его основе все-таки возможен.

Таким образом, логическая основа прогноза по временно́му ряду достаточно прозрачна и фактически использует все уже известные методы многомерного анализа. Вместе с тем для удобства анализа временно́го ряда и построения на его основе прогноза все методы обычно объединяют в одну программу, называемую по имени ее автора Джексон-Бокс-анализ (ARIMA-Auto-Regressive Integrated Moving Average) или (АРПСС — Авторегрессионное проинтегрированное Скользящее Среднее).

Изложение метода в основном опирается на русский текст к пакету программ Statistica, в котором он представлен компактно и наиболее просто.

Прежде чем приступать к построению модели прогноза, желательно иметь максимально полное представление о структуре ряда, а именно:

- имеется ли в нем тренд;
- есть ли в нем сезонная составляющая;
- наличие регулярной составляющей.

В прогнозной модели рассматриваются два основных процесса.

Процесс авторегрессии. Если во временно́м ряду при небольшом лаге существует положительная или отрицательная автокорреляция, то значения в дат (элементы ряда) последовательно зависят друг от друга. Эту зависимость можно выразить следующим уравнением:

$$x_t = a + b_1 x_{t-1} + b_2 x_{t-2} + b_3 x_{t-3} + \ldots + \varepsilon,$$

где a — константа (свободный член); b_1, b_2, b_3 — параметры авторегрессии; ε — вклад «белого шума».

Значение переменной x_t в ряду X есть сумма случайной компоненты (случайное воздействие ε) и линейной комбинации предыдущих наблюдений. Важно обратить внимание на то, что процесс авторегрессии будет стационарным только в том случае, если его параметры лежат в определенном диапазоне. Например, при наличии только одного параметра он должен находиться в интервале $-1 < b < +1$. В противном случае, предыдущие значения будут накапливаться и значения последующих x_t могут неограниченно расти. Такой ряд уже не будет стационарным и возможности его прогноза весьма ограничены. При нескольких параметрах авторегрессии должны выполняться те же условия, т.е. соотношение параметров должно быть таким, чтобы ряд не уходил в бесконечность, а мог бы вернуться на достаточно большом интервале к некоторому среднему. Однако если иметь это условие в виду и если не удается получить стационарную модель, можно, выделив в структуре ряда порождающую нестационарность, осуществлять прогноз на очень локальный интервал времени.

Процесс скользящего среднего. В отличие от рассмотренного выше в этом процессе каждый элемент ряда подвержен суммарному воз-

действию ошибок или, лучше сказать, случайных флуктуаций. В общем виде это можно записать следующим образом:

$$x_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \theta_3\varepsilon_{t-3} - \ldots,$$

где μ — константа; θ_1, θ_2, θ_3 — параметры скользящего среднего.

Другими словами, текущее наблюдение ряда представляет собой сумму случайной компоненты (случайное воздействие ε в данный момент) и линейной комбинации случайных воздействий в предыдущие моменты времени. Следует обратить внимание на то, что скользящее среднее, о котором здесь идет речь, не является тем «скользящим средним», по которому осуществляется сглаживание ряда, позволяющее избавиться, скорее всего, от чисто случайных отклонений. Если в модели авторегрессии подразумевается функциональная зависимость значения x в момент t от предшествующего значения в момент $t-1$, то в модели скользящего среднего подразумевается его зависимость от чисто случайного процесса на «входе» в систему, который приводит к затухающей автокорреляционной функции. До некоторого обычно небольшого лага автокорреляционная функция или положительная или отрицательная, а при лаге (сдвиге) k она становится равной нулю и далее не выходит за границы доверительного интервала. Параметры модели скользящего среднего определяются в программах через значения автокорреляции. В частности, для процесса скользящего среднего первого порядка имеем:

$$x_t = \mu + \theta\varepsilon_{t-1}.$$

Коэффициент автокорреляции определяют по формуле

$$\rho_k = \begin{cases} \dfrac{-\theta_1}{1+\theta_1^2} & \text{при} \quad k = 1; \\ 0 & \text{при} \quad k = 0. \end{cases}$$

Модель авторегрессии и скользящего среднего. Общая модель включает как параметры авторегрессии, так и параметры скользящего среднего. Обычно рассматриваются три типа параметров модели: параметры авторегрессии (p), порядок разности (d), параметры скользящего среднего (q). В обозначениях Бокса и Джексона модель записывается как АРПСС (p, d, q). Например, модель (0, 1, 2) содержит 0 (нуль) параметров авторегрессии (p) и 2 параметра скользящего среднего (q), которые вычисляются для ряда после взятия разности с лагом 1. Модель может рассчитываться для разности первого, второго и большего порядков. Так как для модели АРПСС желательно, чтобы ряд был стационарным (среднее — постоянно, а выборочные дисперсия и автокорреляция не меняются во времени), часто необходимо брать разности ряда до тех пор, пока он не станет стационарным (часто

также для стабилизации дисперсии целесообразно применять логарифмическое преобразование). Число разностей, взятых для достижения стационарности, определяется параметром *d*. Для того чтобы найти необходимый порядок разности, нужно исследовать график ряда и автокоррелограмму. Сильные изменения *уровня* (скачки вверх или вниз) обычно требуют взятия несезонной разности первого порядка (лаг = 1); сильные изменения *наклона* — разности второго порядка. Для сезонной составляющей необходима соответствующая сезонная разность. Если имеется медленное убывание выборочных коэффициентов автокорреляции в зависимости от лага, обычно берут разность первого порядка. Однако во всех случаях желательно начать построение модели или на основе одной авторегрессии, или на основе одного скользящего среднего, или на основе их комбинации, а в случае неудачи переходить к разностной схеме.

Необходимо обратить внимание на то, что *чрезмерное количество взятых разностей* приводит к менее стабильным оценкам коэффициентов. На этом же этапе (который обычно называют *идентификацией* порядка модели) необходимо решить, как много параметров авторегрессии (p) и скользящего среднего (q) должно присутствовать в эффективной и экономной модели процесса. (*Экономность* модели означает, что она содержит наименьшее число параметров и наибольшее число степеней свободы среди всех моделей, которые подгоняются к данным). Иногда говорят, что модель не должна быть переопределена.

Константа в моделях АРПСС. Дополнительно модели АРПСС могут содержать константу, интерпретация которой зависит от подгоняемой модели. А именно, при отсутствии в модели параметров авторегрессии, константа μ есть среднее значение ряда, в противном случае константа представляет собой свободный член. Если бралась разность ряда, то константа есть среднее или свободный член преобразованного ряда. Например, если использовалась первая разность (разность первого порядка), а параметров авторегрессии в модели нет, то константа представляет собой среднее значение преобразованного ряда и, следовательно, *коэффициент наклона линейного тренда* от исходного.

Число оцениваемых параметров. До начала работ по оцениванию необходимо решить, какой тип модели будет подбираться к данным и какое количество параметров присутствует в модели, иными словами, нужно идентифицировать модель АРПСС. Основными инструментами идентификации порядка модели являются графики, автокорреляционная функция (АКФ), частная автокорреляционная функция (ЧАКФ). Это решение не является простым и требует основательной работы с альтернативными моделями. Тем не менее, большинство встречающихся на практике временны́х рядов можно с достаточной степенью точности аппроксимировать

одной из пяти основных моделей, которые можно идентифицировать по виду АКФ и ЧАКФ:

1) *один параметр авторегрессии* (*p*): АКФ экспоненциально убывает; ЧАКФ имеет резко выделяющееся значение для лага 1, нет корреляций на других лагах;

2) *два параметра авторегрессии* (*p*): АКФ имеет форму синусоиды или экспоненциально убывает; ЧАКФ имеет резко выделяющиеся значения на лагах 1, 2, нет корреляций на других лагах;

3) *один параметр скользящего среднего* (*q*): АКФ имеет резко выделяющееся значение на лаге 1, нет корреляций на других лагах. ЧАКФ экспоненциально убывает;

4) *два параметра скользящего среднего* (*q*): АКФ имеет резко выделяющиеся значения на лагах 1, 2, нет корреляций на других лагах. ЧАКФ имеет форму синусоиды или экспоненциально убывает;

5) *один параметр авторегрессии* (*p*) *и один параметр скользящего среднего* (*q*): АКФ и ЧАКФ экспоненциально убывают с лага 1.

Конечно, возможны и более сложные отношения и их идентификация требует как определенного опыта, так и более глубокого знания теории.

Сезонные модели. Мультипликативная сезонная АРПСС представляет естественное развитие и обобщение обычной модели АРПСС на ряды, в которых имеется периодическая сезонная компонента. В дополнение к несезонным в модель вводятся сезонные параметры для определенного лага (устанавливаемого на этапе идентификации порядка модели). Аналогично параметрам простой модели АРПСС эти параметры называются: сезонная авторегрессия (*ps*), сезонная разность (*ds*) и сезонное скользящее среднее (*qs*). Таким образом, полная сезонная АРПСС может быть записана как АРПСС (*p*,*D*,*q*)(*ps*,*ds*,*qs*). Например, модель (0,1,2) (0,1,1) включает 0 регулярных параметров авторегрессии, 2 регулярных параметра скользящего среднего и 1 параметр сезонного скользящего среднего. Эти параметры вычисляются для рядов, получаемых после взятия одной разности с лагом 1, и далее сезонной разности. Сезонный лаг, используемый для сезонных параметров, определяется на этапе идентификации порядка модели.

Общие рекомендации относительно выбора обычных параметров (с помощью АКФ и ЧАКФ) полностью применимы к сезонным моделям. Основное отличие состоит в том, что в сезонных рядах АКФ и ЧАКФ имеют существенные значения на лагах, кратных сезонному лагу (в дополнение к характерному поведению этих функций, описывающих регулярную (несезонную) компоненту АРПСС).

Оценивание параметров. Существуют различные методы оценивания параметров, которые дают очень похожие результаты, но для данной модели одни оценки могут быть более эффективны, а

другие — менее. В общем случае в процессе оценивания порядка модели используется так называемый квазиньютоновский алгоритм максимизации правдоподобия (вероятности) наблюдения значений ряда по значениям параметров, что требует вычисления (условных) сумм квадратов (*SS*) остатков модели. Имеются различные методы нахождения суммы квадратов остатков *SS*: 1) приближенный метод максимального правдоподобия МакЛеода и Сейлза; 2) приближенный метод максимального правдоподобия с итерациями назад; 3)точный метод максимального правдоподобия по Меларду.

Сравнение методов. В основном все методы дают очень похожие результаты. Также все методы показали примерно одинаковую эффективность на реальных данных. Однако метод 1 — самый быстрый, им можно пользоваться для исследования очень длинных рядов (например, содержащих более 30 000 наблюдений). Метод 3 может оказаться неэффективным, если оцениваются параметры сезонной модели с большим сезонным лагом (например, 365 дней). С другой стороны, можно вначале использовать приближенный метод максимального правдоподобия (для того чтобы найти прикидочные оценки параметров), а затем точный метод; обычно требуется только несколько итераций метода 3, для того чтобы получить окончательные оценки.

Стандартные ошибки оценок. Для всех оценок параметров рассчитываются так называемые асимптотические стандартные ошибки, для вычисления которых используется матрица частных производных второго порядка, аппроксимируемая конечными разностями.

Штраф. Процедура оценивания минимизирует (условно) сумму квадратов остатков модели. Если модель не является адекватной, то может случиться так, что оценки параметров на каком-то шаге станут неприемлемыми — очень большими (например, не удовлетворяют условию стационарности). В таком случае *SS* будет приписано очень большое (*штрафное*) значение. Обычно это «заставляет» итерационный процесс удалить параметры из недопустимой области. Однако в некоторых случаях и эта стратегия может оказаться неудачной, и будут получены очень большие значения *SS* на серии итераций. В таких случаях следует с осторожностью оценивать пригодность модели. Если модель содержит много параметров и, возможно, имеется переопределение, то следует несколько раз испытать процесс оценивания с различными начальными условиями и параметрами.

Оценивание модели. Если значения вычисляемой *t* статистики не значимы, то соответствующие параметры в большинстве случаев удаляются из модели без ущерба подгонки. Другой обычной оценкой надежности модели является сравнение прогноза, построенного по урезанному ряду с «известными (исходными) данными».

Однако качественная модель должна не только давать достаточно точный прогноз, но быть экономной и иметь независимые остатки, содержащие только шум без систематических компонент (в частности, АКФ остатков не должна иметь какой-либо периодичности). Процедура оценивания предполагает, что остатки некоррелированны и нормально распределены. Хорошей проверкой модели являются: 1) график остатков и изучение их трендов; 2) проверка АКФ остатков (на графике АКФ обычно отчетливо видна периодичность). Если остатки систематически распределены (например, отрицательны в первой части ряда и примерно равны нулю — во второй) или включают некоторую периодическую компоненту, то это свидетельствует о неадекватности модели. Анализ остатков чрезвычайно важен и необходим при анализе временны́х рядов.

Полученные оценки параметров используются на последнем этапе (этапе-прогноза) для того, чтобы вычислить новые значения ряда и построить доверительный интервал для прогноза.

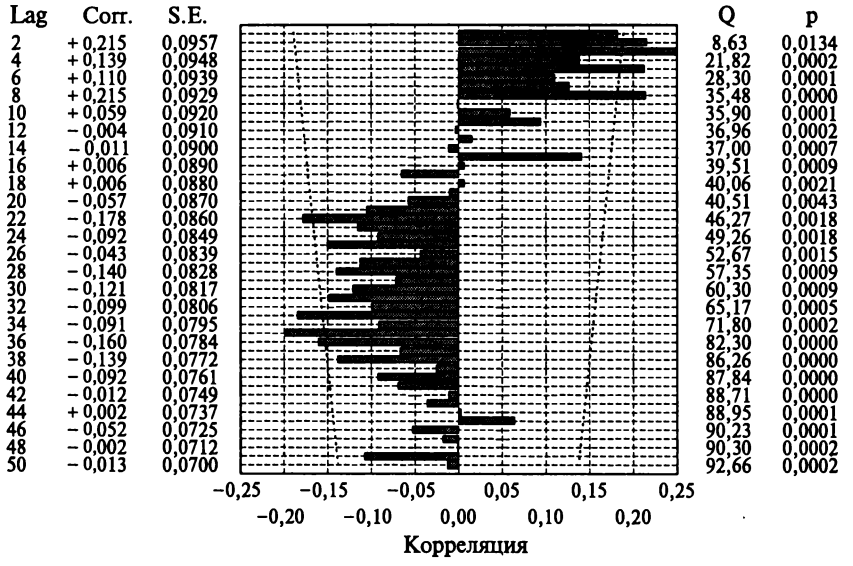
Такова самая общая схема построения прогностической модели на основе анализа временно́го ряда. Анализ временны́х рядов аккумулирует в себе по существу всю мощь параметрических методов статистики и в некотором смысле может рассматриваться как вершина всего ее стройного здания. Это исключительно творческий процесс, не допускающий механистического подхода. Должные навыки анализа приобретаются только с опытом и постоянным углублением в теоретические основы богатого арсенала его методов. В настоящем пособии даны только самые общие основания. В аннотированном списке рекомендуемой литературы приведены монографии, тщательный разбор содержания которых, параллельно с анализом реальных данных, позволит исследователю действительно освоить это важное направление.

Сочетая исследования реальных данных с базовыми моделями теории линейных и нелинейных колебаний и динамики систем, можно существенно продвинуться в область физической интерпретации наблюдаемых процессов. Следует отметить, что программные средства анализа временны́х рядов организованы во многих специальных, но трудно доступных пакетах программ, адаптированных к экономическим и геофизическим исследованиям. В России в 90-х гг. XX в. был очень популярен пакет программ анализа временны́х рядов «Мезозавр», разрабатывавшийся под эгидой фирмы «ДИАЛОГ».

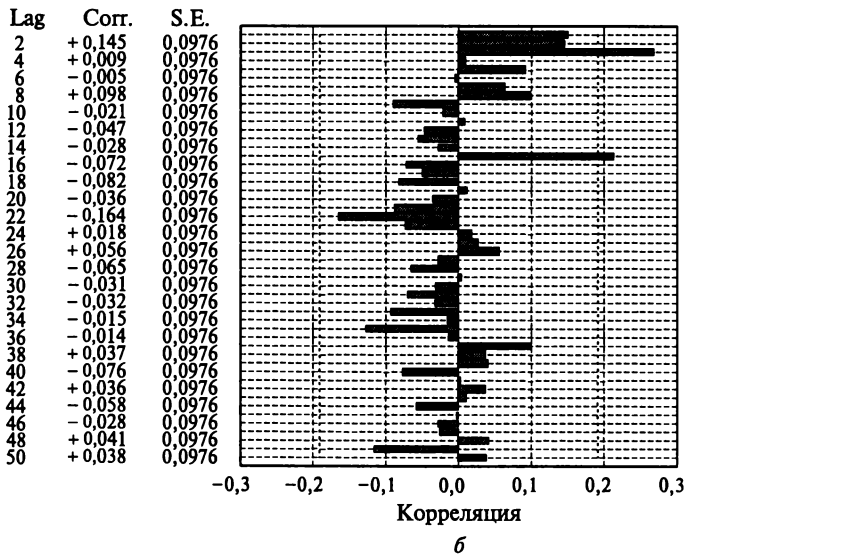
В настоящее время наиболее мощный пакет программ анализа временны́х рядов представлен в Statistica и Statgraf (в остальных пакетах возможности анализа временны́х рядов весьма ограничены).

В качестве примера рассмотрим применение методов прогноза временно́го ряда для сумм зимних осадков на метеостанции «Рязань».

На рис. 9.19 показана АКФ и ЧАКФ сумм осадков за январь. АКФ постепенно экспоненциально уменьшается (рис. 9.19, *а*), а ЧАКФ имеет достоверный локальный максимум на лагах 1, 2, 3 (рис. 9.19, *б*). Эти отношения точно не соответствуют описанным выше критериям вида модели, но ближе всего ко второй модели с



а



б

Рис. 9.19. АКФ (*а*) и ЧАКФ (*б*) сумм осадков за январь (по данным метеостанции «Рязань»). Стандартная ошибка — оценка границы «белого шума»

Таблица 9.8

Модель для сумм осадков в январе с трансформацией ln(x) – ARIMA (2,0,0)

Средняя квадратическая ошибка остатков MS Residual = 0,39205

Переменная	Параметры	Асимптотическая стандартная ошибка	t-критерий (102)	Уровень значимости p-level	Доверительный 95%-й интервал	
					нижний	верхний
Константа	3,216739	0,096042	33,49302	0,000000	3,026240	3,407238
p(1)	0,143590	0,098538	1,45720	0,148132	–0,051860	0,339040
p(2)	0,222434	0,098775	2,25191	0,026471	0,026513	0,418354

двумя параметрами авторегрессии. Так как осадки подчиняются гамма-распределению, установим автоматическое логарифмирование данных и проверим модель авторегрессии с лагом 2. В результате получаем следующие оценки параметров (рис. 9.20, табл. 9.8).

Из табл. 9.8 следует, что модель переопределена, так как второй параметр p(1) статистически не значим. Более того, функция автокорреляции остатков имеет значения, выходящие за границу «белого шума», т.е. в остатках существуют автокорреляции и они не могут рассматриваться как чисто случайные шумовые колебания. Следовательно, модель нельзя считать адекватной реальному процессу. Если признать, что ЧАКФ ряда экспоненциально убывает, то можно использовать пятую модель (1,0,1), табл. 9.9.

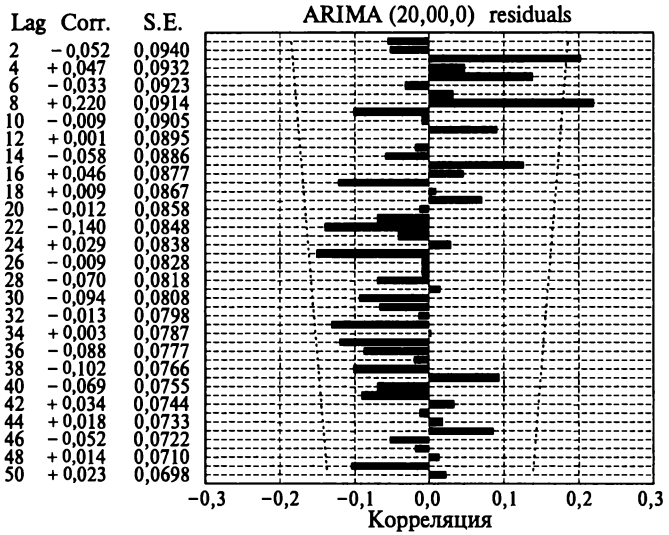
Таблица 9.9

Модель для сумм осадков в январе с трансформацией ln(x) – ARIMA (1,0,1); средняя квадратическая ошибка остатков MS Residual = 0,3744

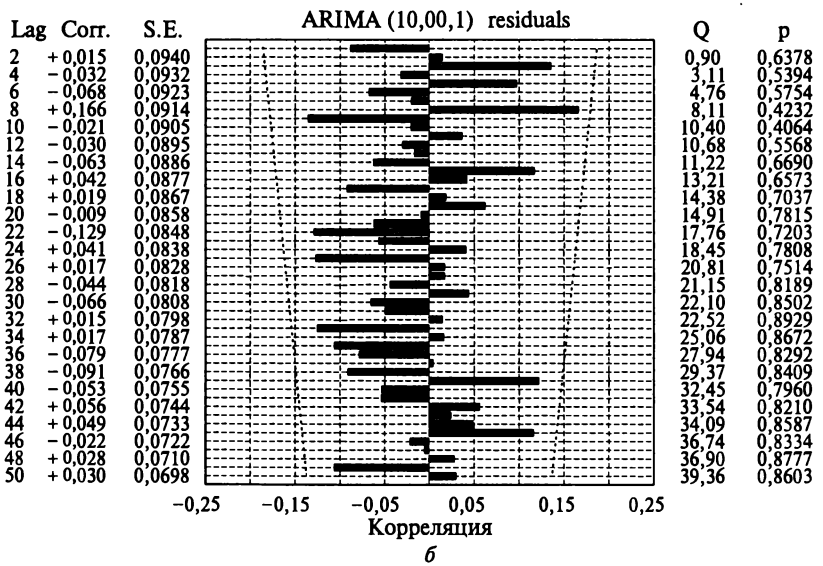
Model (1,0,1) MS Residual = 0,37440

Переменная	Параметры	Асимптотическая стандартная ошибка	t-критерий (102)	Уровень значимости p-level	Доверительный 95%-й интервал	
					нижний	верхний
Константа	3,212161	0,149824	21,43949	0,000000	2,914984	3,509337
p(1)	0,902490	0,071937	12,54559	0,000000	0,759803	1,045176
q(1)	0,739982	0,101565	7,28577	0,000000	0,538528	0,941437

Эта модель авторегрессии первого порядка и скользящего среднего по формальным критериям статистически значима и не переопределена. АКФ остатков показывает, что все корреляции не превышают уровня шума. Таким образом, по формальным критериям



а



б

Рис. 9.20. АКФ (*а*) и ЧАКФ (*б*) остатков сумм осадков за январь от модели (2,0,0) — по данным метеостанции «Рязань». Стандартная ошибка — оценка границы «белого шума»

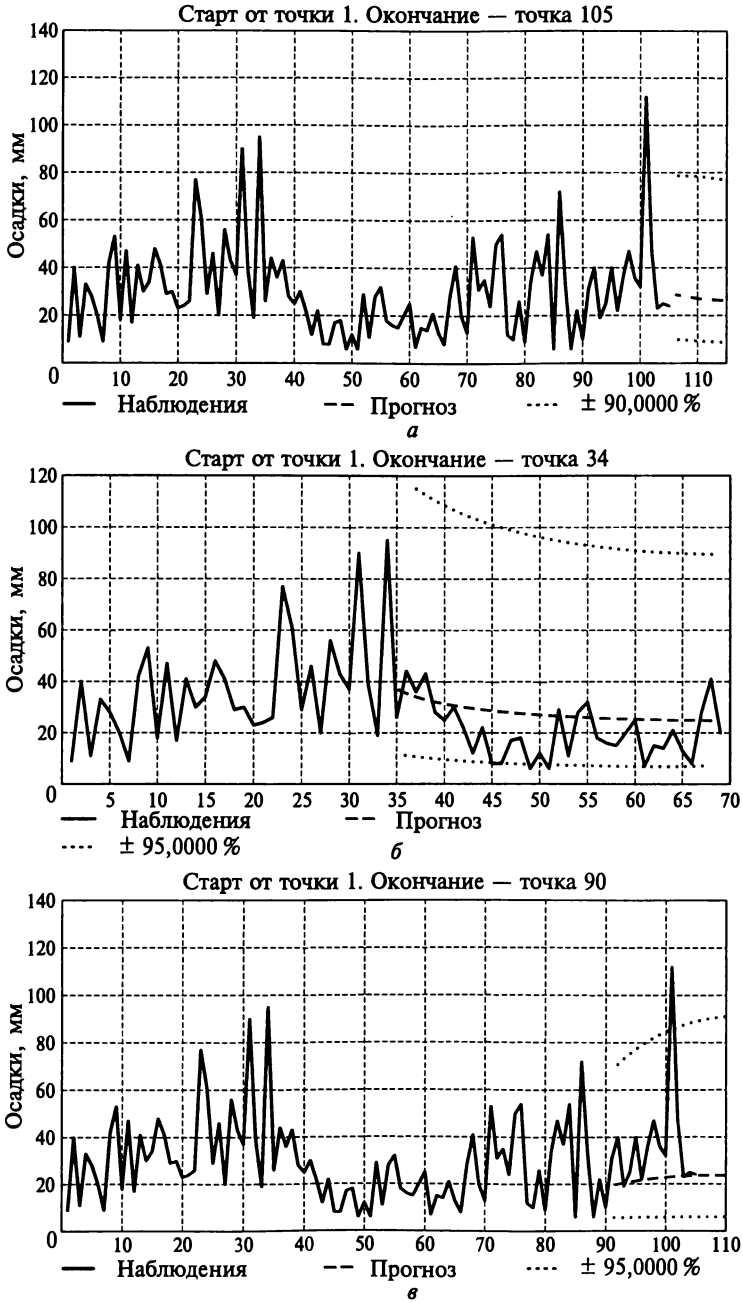


Рис. 9.21. Прогноз изменения осадков на 12 лет по модели (1,0,1) для трех вариантов:

а — на будущее; *б*, *в* — для интервалов с тенденцией к уменьшению (*б*) и увеличению (*в*) осадков

можно признать модель хорошей. На рис. 9.21 оцениваются прогностические возможности модели для трех вариантов. В первом варианте осуществляется прогноз от последнего года наблюдения (1989) на 12 лет вперед (рис. 9.21, *а*). Модель показывает возможность слабого отрицательного тренда осадков в январе. Далее демонстрируются прогнозные возможности модели от 1917 г. в сравнении с тем, что наблюдалось в реальности. В целом модель предсказывает тенденцию к снижению осадков (рис. 9.21, *б*). В третьем варианте при прогнозе от 1973 г. модель предсказывает тенденцию увеличения осадков (рис. 9.21, *в*). Таким образом, модель неплохо описывает тенденции, а ее построение по логарифмическому основанию позволяет предсказывать возможные экстремальные значения. Если вспомнить задачу прогноза вероятности получения экстремальных значений осадков по всему ряду, рассмотренную в рамках одномерного анализа, то очевидно, что эта модель дает некоторые уточненные оценки возможных экстремумов для временно́го лага около 10 лет.

Заключая краткое изложение анализа временны́х рядов, выделим две связанные основные сюжетные линии:

1) исследование ряда для выявления типов определяющих его процессов и выделение их в форме частных рядов для формулировки гипотез о возможных механизмах, определяющих динамику изучаемой системы;

2) прогноз возможных изменений в будущем на основе свойств самого временно́го ряда.

Следует отметить, что методы анализа временны́х рядов и исследования динамики сложных природных систем находятся в постоянном развитии. В данном случае рассмотрены лишь самые общепринятые. Читатель, интересы которого связаны с этой областью знаний, должен постоянно следить за текущей литературой и обогащать арсенал используемых технологий анализа.

9.4. Анализ пространственных рядов

Теоретические основания анализа временны́х и пространственных рядов тождественны. Основное отличие состоит в том, что пространственные ряды при полном отображении отношений двухмерны (для поверхности Земли: широта и долгота). Проблема сама по себе очень широка и включает, в том числе анализ любых изображений с целью максимально компактной упаковки информации при минимальных потерях качества восстановления исходного изображения. В частности, стандартные форматы типа *jpg* используют для упаковки информации алгоритмы анализа пространственной структуры изображения. Географов и экологов проблемы упаковки информации изображения для ее передачи по каналам

связи с последующим восстановлением приемником интересуют, в первую очередь, не как задачи оптимального кодирования, а как системы анализа структуры, т.е. правил организации изображения или в общем случае поверхности. В науках о земле эта тема разрабатывается в рамках так называемого пространственного анализа. Хотя логика выделения однородных по внутренней структуре форм рельефа и контуров, однородных по текстуре, на снимках достаточно прозрачна, строгая алгоритмитизация этих процедур остается актуальной задачей.

Под *пространственным анализом* (Spetial Analysis) в самом общем случае понимается набор методов исследования, в которых, по крайней мере, две из переменных связаны с пространственным местоположением (X и Y координатами), третья Z характеризует некоторый признак, изменяющийся в пространстве (обычно высоту или глубину). Пространственный анализ первоначально был ограничен, главным образом, качественным анализом карт. В настоящее же время он включает методы анализа полей и полигонов из прикладной математики и статистики. Таким образом, современный пространственный анализ строится на количественном анализе картографических данных и включает в себя в том числе и процессы моделирования и интерпретации. Имеется четыре традиционных метода пространственного анализа: 1) топологический оверлей; 2) анализ соседства; 3) анализ поверхности; 4) анализ линий и растра. В самом общем плане пространственный анализ можно определить как количественное изучение явлений, распределенных в пространстве, и как способ управления пространственными данными, представленными в различной форме с извлечением из этого управления дополнительного содержания, отражаемого в результатах.

Происхождение пространственного анализа связывается с развитием в начале 60-х гг. XX в. количественной географии. Его начальную стадию характеризовало использование количественных (главным образом статистических) процедур и методов для анализа размещения пунктов (точек), линий, областей и поверхностей, изображенных на картах или определенных координатами в двух- или трехмерном пространстве. Позже акцент был смещен на исследование пространственных отношений между явлениями для конкретных географических объектов, на пространственные процессы при исследовании пространственно-временного развития сложных систем. Как показали три десятилетия развития, пространственный анализ по содержанию больше, чем пространственная статистика. В целом же в пространственном анализе выделяется область статистического пространственного анализа, развивающего методологию и методы, обеспечивающие отображение широкого диапазона пространственных эффектов и пространственные модели изучаемых процессов, включая широкий диапазон

различных моделей географических, экономических и социальных явлений.

Пространственный анализ опирается на следующие технические средства и стандартные методические приемы: географические информационные системы, глобальную систему позиционирования (GPS), спутниковую информацию, пространственную статистику; элементарные технические приемы кластеризации и измерения соседства, буферизации, расчета времени движения, расчета доступности, пространственных моделей связей и зависимостей, а также более сложные технологии: пространственные автокорреляции, системную и ландшафтную метризацию, метод главных компонент и т. п.

На совещании службы лесов США, уделяющей особое внимание этому направлению, в 1998 г. были определены проблемы и приоритетные направления развития пространственного анализа. В первую очередь обращается внимание на необходимость развития пространственно-временны́х технологий анализа, ориентированных на анализ очень больших объемов многомерных данных как в локальном, региональном, так и глобальном масштабах. Рекомендуется развивать методы классификации, геостатистики, применять методы нейронных сетей, размытых множеств, анализа импульсов (wavelets), методы искусственного интеллекта, ориентированные на непосредственный анализ текстов и, напротив, раскрытия содержания многомерной географической информации.

Особое внимание уделено анализу влияния масштаба, в котором изучается явление, на результат и на связь результатов исследования отношений, выполненных в различных масштабах, а также на методы выделения особых ситуаций, аномалий, особых отношений — «горячих точек». В конечном итоге пространственный анализ рассматривается как путь интеграции, интегрального анализа и моделирования данных о состоянии природы и общества в различных пространственно-временны́х масштабах для повышения эффективности принятия решений, обеспечивающих адекватность действий человека в изменяющихся условиях среды и социума. Из этого описания совершенно очевидно, что пространственный анализ требует отдельного развернутого изложения и не может быть полно рассмотрен в настоящем пособии.

Поэтому ограничимся иллюстрацией подхода на примере одномерного пространственного анализа, содержание которого не выходит за рамки приведенных выше подходов к анализу временны́х рядов.

Объектом исследования будет прямолинейный нивелировочный ход (трансект) с постоянным шагом измерения превышения 10 м (рис. 9.22), проложенный через гряду московской морены на территории Центрально-лесного природного биосферного заповед-

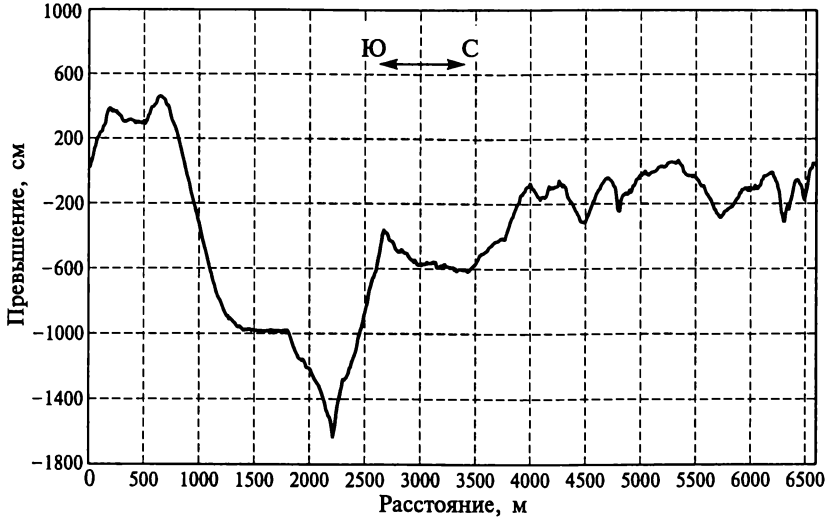


Рис. 9.22. Трансект по просеке 96/97 (длина 6580 м) при шаге нивелирования 10 м

ника. Анализ выполнен студенткой кафедры физической географии и ландшафтоведения МГУ К. А. Мерекаловой.

На рис. 9.23, *а* представлен спектр трансекта в логарифмических координатах. Линейная зависимость логарифма спектра от логарифма частоты (1/период) указывает на фрактальность структуры рельефа. Фрактальная размерность определяется по формуле

$$D = (5 - b_1)/2,$$

где b_1 — параметр в уравнении регрессии $\log(\text{спектр}) = b_0 + b_1 \times \log(1/\text{период})$. Для данного трансекта D = 1,25, что близко к «черному шуму». Однако в низкопериодической (высокочастотной) части спектра визуально выделяется интервал линейных размеров территориальных структур (приблизительно до 30 м), плохо описываемый этим уравнением регрессии. Это позволяет предположить, что рельеф формировался при взаимодействии, по крайней мере, двух генетически различных факторов, действовавших на интервалах меньше и больше 30 м.

Строгое доказательство многофакторности системы можно получить, исследуя отклонения от линии регрессии. Если полученные остатки содержат достоверный полиномиальный тренд, то его перегибы (точки, в которых первая производная равна нулю) соответствуют границам воздействия разных факторов. На рис. 9.23, *б* показаны остатки, полиномиальный тренд и график производной для рассматриваемого спектра. Очевидно, выделяются три интервала периодов, возможно отражающие действие трех различных

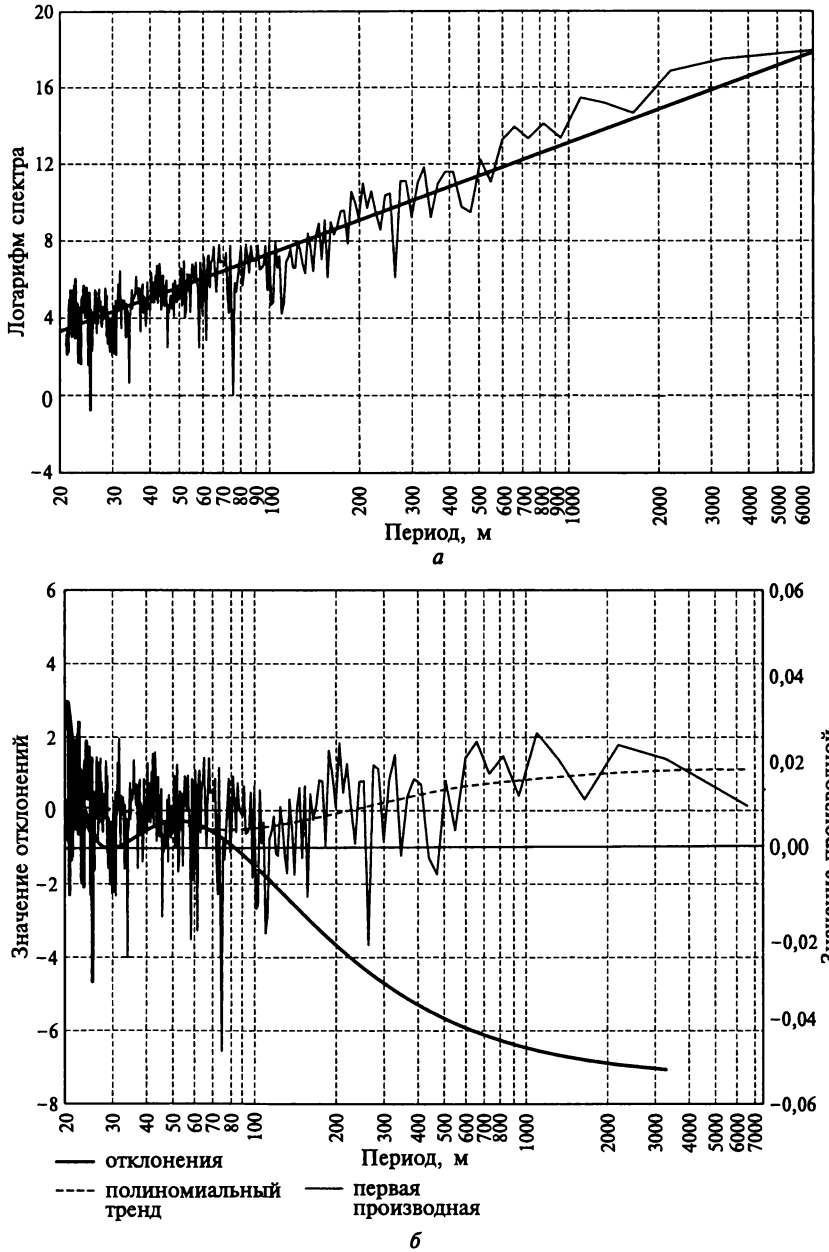


Рис. 9.23. Спектр трансекта по просеке 96/97 с шагом 10 м и линия регрессии log(спектр) = b_0 + $b_1$log(1/период) (*а*) и отклонения от линии регрессии, содержащийся в них полиномиальный тренд и первая производная тренда (*б*)

Таблица 9.10

Параметры спектра для трансекта с различными интервалами частот

Шаг нивелирования	Период, м	B1	Ошибка	Значимость	Фрактальная размерность
10 м	Все значения	–2,49979	0,070843	0,000000	1,250105
	<30	0,602359	0,996308	0,546719	2,1988205
	30—85	–2,12536	0,336027	0,000000	1,43732
	>85	–3,07319	0,146192	0,000000	0,963405

факторов: 1) < 30 м; 2) 30 до 85 м и 3) >85 м. При аппроксимации выделенных интервалов уравнениями регрессий можно оценить фрактальную размерность для каждого из них. Результаты такой операции представлены в табл. 9.10 и на рис. 9.24.

Видно, что крупнопериодические колебания (>85 м) с учетом среднеквадратической ошибки описываются типичным «черным шумом» (фрактальная размерность около 1), колебания с перио-

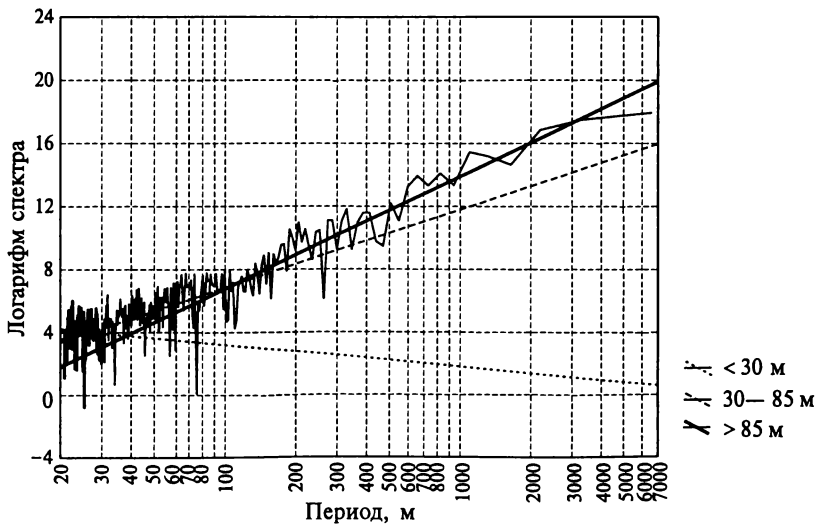


Рис. 9.24. Три интервала (<30; 30...85; >85 м) спектра трансекта 96/97 (шаг 10 м), ассоциируемые с различными факторами генезиса, и линии регрессии

дом от 30 до 85 м — «бурым шумом», а с периодом менее 30 м — «розовым шумом». Последнее значение является, однако, недостоверным (уровень значимости — 0,546719), что возможно связано с недостаточным описанием десятиметровым шагом опробования таких высокочастотных колебаний.

На рис. 9.25, *а* можно видеть, что остатки от линии регрессии «логарифм спектра — логарифм частоты» содержат периодические составляющие и не являются «белым шумом». Следовательно, поверхность подчиняется правилам иерархической организации, описываемым квазипериодическими разномасштабными колебаниями.

Осредняя периодограмму спектром (рис. 9.25, *б*), можно выделить три основных независимых фактора, определяющих строение поверхности:

1) фактор, описываемый полиномиальной формой;

2) фактор, описывающий волны с периодами $P_i = L/(k_0 + 27k_i)$ (k_0 — минимальное волновое число волны с номером $i = 0$), волновая константа — 21,8—29,8 пикселей;

3) фактор, описывающий волны с периодами $P_i = L/(k_0 + 3{,}3k_i)$, волновая константа — 3,3 пикселя.

Статистически значимый полиномиальный тренд 7-й степени имеет константу по волновому числу 164, т.е. в ряду длиной $L = 658$ укладывается три иерархически соподчиненные волны, для которых волновое число определяют по формуле

$$k_i = k_0 + 164k_i,\ i = 0,\ 1,\ 2,\ 3,\ \ldots,\ n,$$

а период соответственно равен

$$P_i = L/(k_0 + 164k_i).$$

Константу k_0 необходимо найти по графику полинома:

$$k_0 = \frac{L}{P_{\max}},$$

где $P_{\max}$ — наибольший период по полиному ($P_{\max} = 47$) и соответственно $k_0 = 14$.

Таким образом, этот первый независимый фактор, определяющий иерархическую организацию рельефа, порождает пространственные формы поверхности с периодами

$P_0 = 658/14 = 47{,}0$ пикселей (470 м); $P_1 = 3{,}1$ (31 м); $P_2 = 1{,}9$ (20 м).

Построим функции линейных размеров пространственных волн, порождаемых этим фактором, вне зависимости от длины наблюдаемого ряда. Для этого перенумеруем волны так, чтобы волна с наименьшим периодом имела бы номер 1 (первый иерархический уровень) и будем обозначать номера этих волн, как номера иерар-

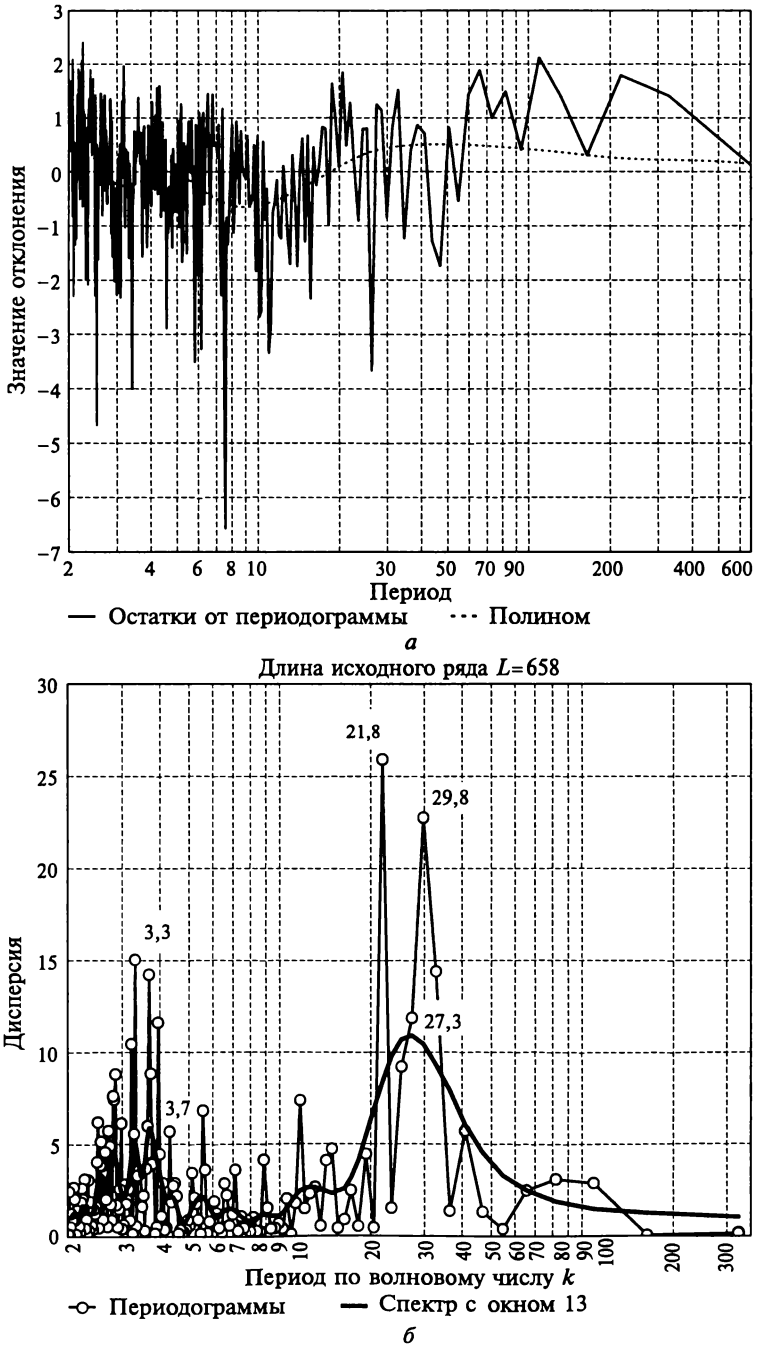


Рис. 9.25. Остатки от уравнения регрессии «логарифм спектра — логарифм частоты» (*а*) и их периодограмма и спектр (*б*)

хических уровней, буквой S. Тогда функция, описывающая связь номера волны с периодом, будет определяться по формуле

$$P_s = (P_1 + a)S^b,$$

где P_1 — константа ($P_1 \approx 20$ м).

Определить константу можно с помощью метода оценки параметров нелинейных уравнений.

В результате аппроксимации получаем

$$P_s = 1000(20 + 0{,}01941S^{9{,}1}) \text{ км}.$$

Эти наиболее низкочастотные волны порождают иерархические уровни, кроме указанных выше, с линейными размерами (рис. 9.26): период 6,1; 47,2; 250 и 1024 км. Естественно, что это чисто прогностические оценки и они реалистичны при условии, что на всем этом интервале сохраняется действие рассматриваемого фактора.

Теперь проанализируем иерархические структуры, порождаемые вторым фактором с волновой константой, лежащей в интервале 21,8—29,8 шагов.

Для определения k_0 аппроксимируем остатки уравнением «косинус — синус»:

Остатки = $a\sin(2\pi(329/29{,}8)k) + b\cos(2\pi(329/29{,}8)k)$ с частотой, равной половине длины ряда, деленной на волновую константу. Подбор в ряде волновых чисел от 21,8 до 29,8 с помощью множественной регресси и позволяет выбрать оптимальную константу, равной 29,8. На рис. 9.27, *а* показано, как описываются остатки от волновых чисел моделью с этой константой. Соответственно на наблюдаемом интервале в 6,59 км рассматриваемый фактор порождает 11 полных иерархических уровней с $k_0 = 5$.

Рис. 9.26. Расчет возможных иерархических уровней организации рельефа за пределами длины ряда наблюдений

Таким образом, для данного иерархического уровня имеем

$$P_i = 658/(5 + 29{,}8i).$$

В этой форме записи наиболее высокий иерархический уровень, с которым связаны поверхности с наибольшими линейными размерами, соответствует $i = 0$. Для лучшей наглядности примем этот уровень в качестве

максимального и присвоим ему индекс 11. Соответственно запишем отношение в ином виде:

$$P_i = 658/[5 + 29{,}8(12 - i)].$$

При этой форме записи первый уровень, отражающий пространственные уровни с минимальной длиной пространственной волны, будет иметь номер 1, а наибольший уровень — номер 12 (рис. 9.27, *б*).

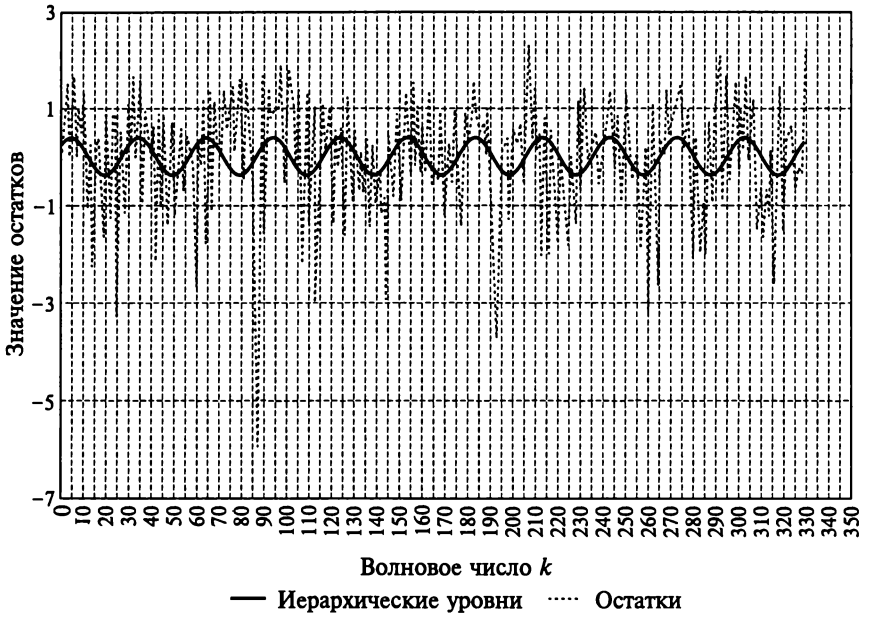
Не рассматривая действие третьего высокочастотного фактора, обратим внимание на то, что сложный спектральный образ рельефа с множеством волн вероятнее всего порождается случайными фрактальными процессами с элементами нелинейных колебаний, вызываемых несколькими независимыми факторами.

Первый фактор, описываемый полиномом, с большой вероятностью можно ассоциировать со структурой геологического фундамента. Второй, скорее всего, отражает правила формирования четвертичных отложений, сформированных московской моренной. Третий, наиболее высокочастотный, фактор, возможно, отражает формирование маломощного покрова валдайского оледенения. Однако эти высказывания можно рассматривать лишь как предположения. Следует отметить, что изучение правил организации поверхности только начинаются, и в этой важной для географии области исследования пока не существует достаточно общепринятых решений.

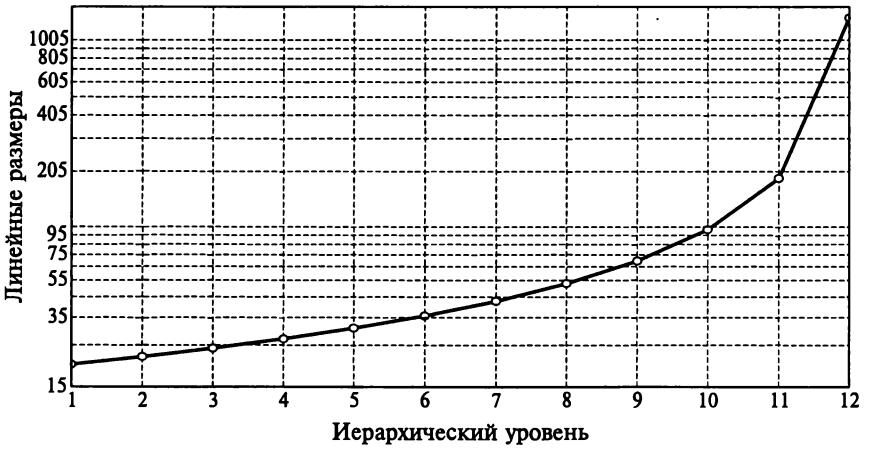
С практической точки зрения полезно выделить поверхности различного иерархического уровня. Эту операцию можно осуществлять с помощью выделения части изображения в определенной полосе частот или на основе факторного анализа. Продемонстрируем второй метод, так как, кроме выделения форм рельефа различного иерархического уровня, он дает независимое представление о его возможной факторной основе.

Так же как и при факторном анализе климатических рядов, будем смещать ряд относительно самого себя. Имея в виду, что наибольший вклад в регулярную составляющую вносит процесс с константой 29,7, будем рассматривать 15 сдвинутых относительно друг друга рядов. Это означает, что каждая точка трансекта описывается 15 переменными.

Решаем задачу методом главных компонент. Все переменные описывают друг друга с коэффициентом детерминации $R^2 > 0{,}99$. Шесть компонент отображают варьирование высоты. На рис. 9.28 показана связь нагрузки компоненты с ее порядковым номером и график монотонной функции, описывающей эту зависимость. Наибольшие отклонения значений нагрузок от этой монотонной функции лежат в интервале четвертой и восьмой компоненты. Точки перегиба графика «нагрузка — компонента» приходятся на восьмую и шестую компонент. Чтобы не усложнять анализ, ограничимся



а



б

Рис. 9.27. Иерархические уровни организации для фактора с константой 29,8 (*а*) и соотношение линейных размеров пространственных структур и номера иерархического уровня (*б*)

Таблица 9.11

Нагрузки на первые шесть компонент при разложении рельефа по ортогональному базису методом главных компонент

Компонента	Дисперсия	Процент описываемой дисперсии	Накопленная дисперсия	Процент накопленной дисперсии
1	14,77989	98,53259	14,77989	98,53259
2	0,19650	1,30997	14,97638	99,84256
3	0,01557	0,10378	14,99195	99,94633
4	0,00387	0,02583	14,99582	99,97217
5	0,00156	0,01041	14,99739	99,98258
6	0,00087	0,00581	14,99826	99,98838

разложением рельефа по шестимерному ортогональному базису. В табл. 9.11 приведены нагрузки на компоненты в шестимерном пространстве.

Из табл. 9.11 следует, что шесть компонент описывают варьирование рельефа почти на 100 % (не охвачено только 0,012 % его варьирования).

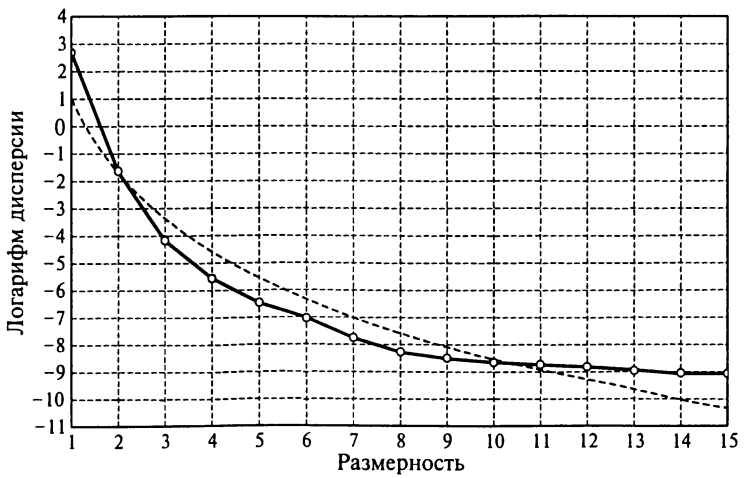
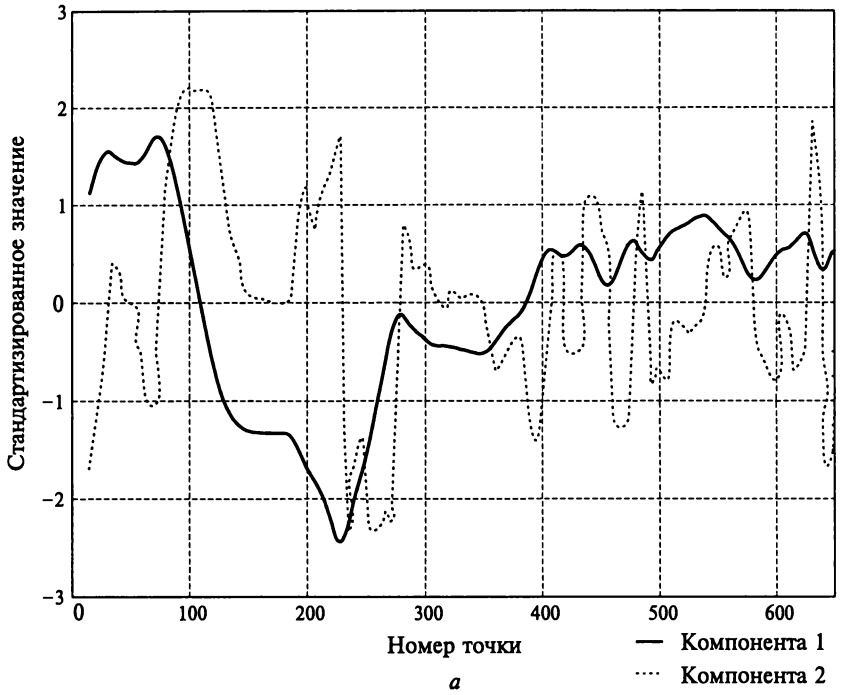
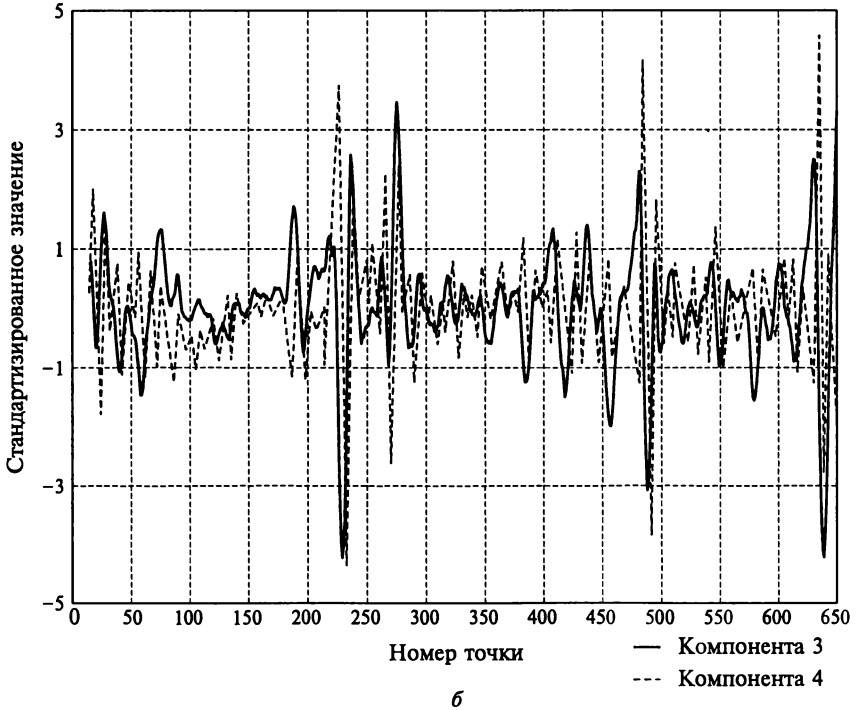


Рис. 9.28. Нагрузки на факторы при разложении ряда высот по трансекту методом главных компонент


Стандартизированное значение
3
2
1
0
−1
−2
−3
0
100
200
300
400
500
600
Номер точки
Компонента 1
Компонента 2

а


Стандартизированное значение
5
3
1
−1
−3
−5
0
50
100
150
200
250
300
350
400
450
500
550
600
650
Номер точки
Компонента 3
Компонента 4

б

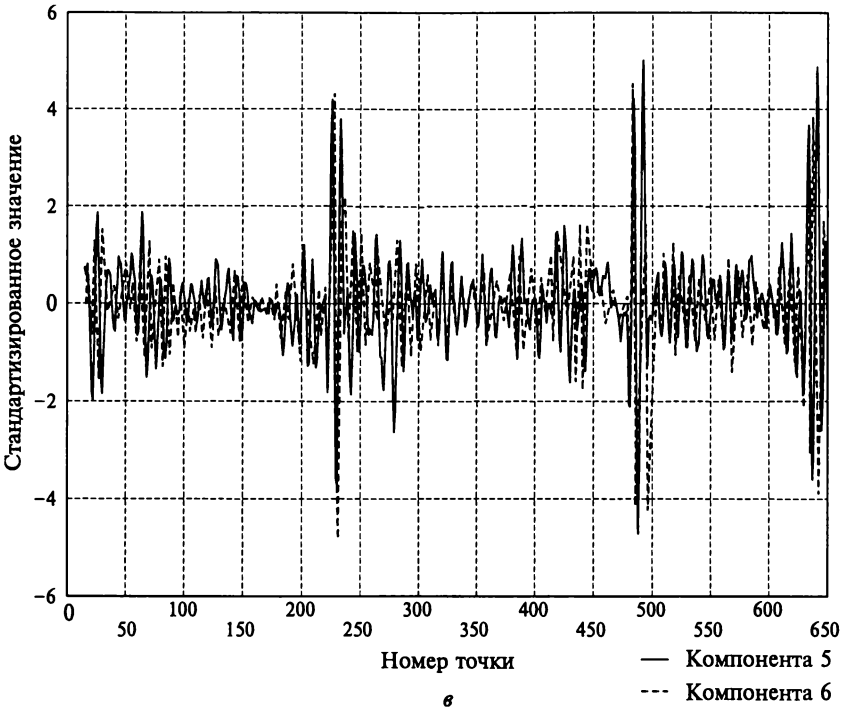


Рис. 9.29. Шесть главных компонент из разложения ряда высот по ортогональному базису:

а — компоненты 1 и 2; *б* — компоненты 3 и 4; *в* — компоненты 5 и 6

Первая компонента, очевидно, отображает макроформы рельефа, вторая — мезоформы и т.д. (рис. 9.29, *а* — *в*). Используя спектральный анализ, можно достаточно однозначно определить основные иерархические уровни организации, которые отображает каждая отдельно взятая компонента (рис. 9.30). Спектр показывает, что каждая компонента имеет свой максимум дисперсии, соответствующий некоторому диапазону длин волн. Первая компонента отображает тренд или формы рельефа с периодами более 100 точек трансекта или 1 км. Вторая компонента имеет максимум дисперсии при периоде около 108 точек, т.е. практически точно для структур с линейными размерами в 1 км. Третья компонента отображает формы рельефа с линейными размерами от 23 до 30 точек (230—300 м), четвертая — от 10 до 14 (100—140 м), пятая — от 7 до 9 (70—90 м), шестая — 6,5 (65 м) со вторым максимумом 3 точки (30 м). Весьма характерно, что график спектра каждой компоненты, хотя и имеет свой максимум, но содержит некоторую информацию и о соседних иерархических уровнях. Это выражается в существовании совпадающих у разных компонент локальных мак-

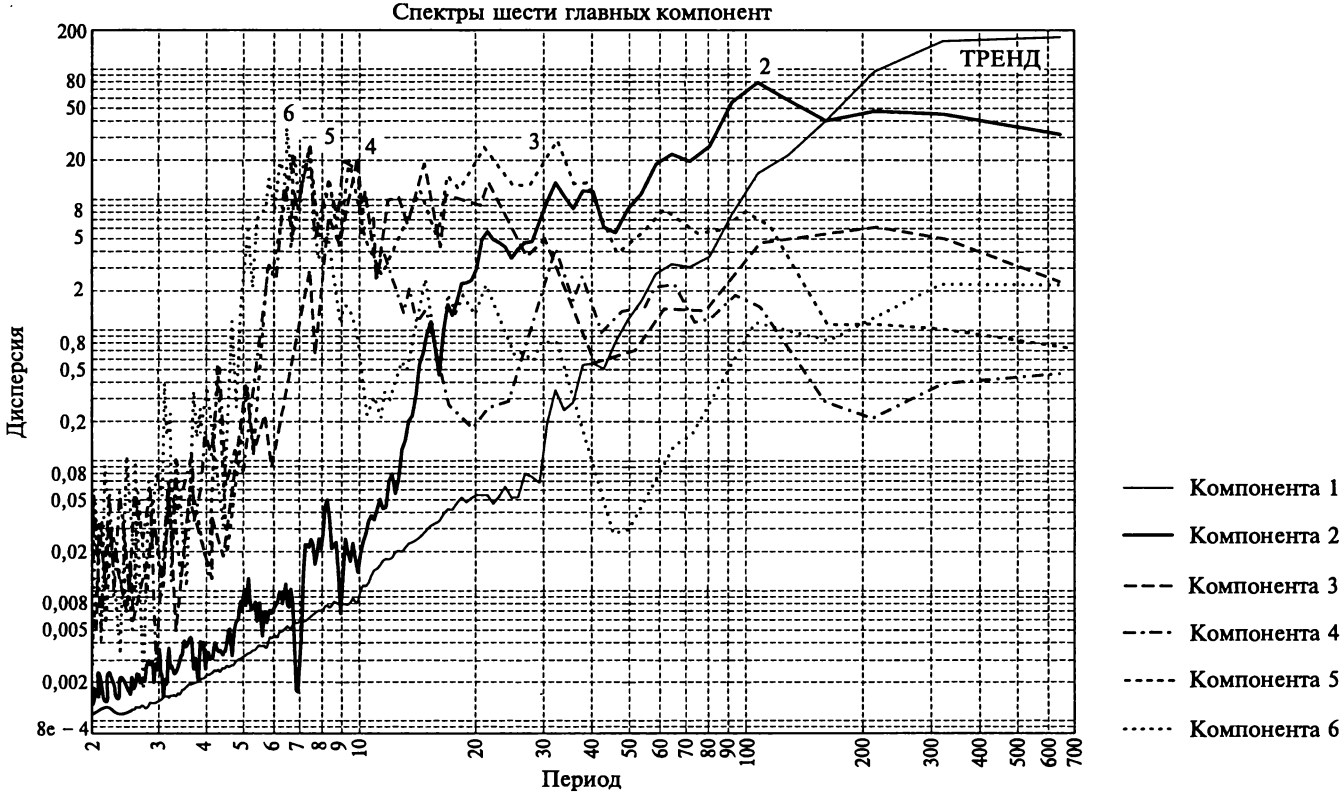


Рис. 9.30. Периодограммы шести главных компонент ряда высот по трансеку

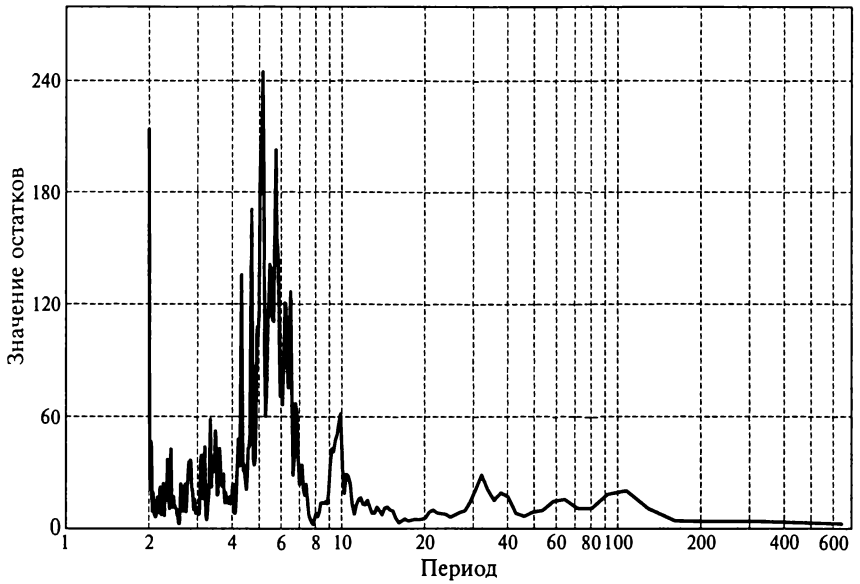


Рис. 9.31. Спектр части варьирования высот, не описываемых шестью главными компонентами (остатки)

симумов дисперсии для общих у всех компонент периодов. Спектр остатков (рис. 9.31) показывает, что неописанными остались в основном формы микрорельефа с периодом 3 точки (30 м) и самое главное — с периодом в 2 точки (20 м). Обратим внимание на то, что значения дисперсий для остатков существенно выше, чем значения для компонент. Это определяется тем, что спектр компонент рассчитывался по стандартизированным значениям, а спектр остатков — по реальным значениям высот, измеренных с точностью до 1 см.

Какие амплитуды высот описывает каждая компонента, легко определить по коэффициентам множественной регрессии (табл. 9.12). Из таблицы следует, что вторая и третья компоненты отображают рельеф с обратным знаком. Первая компонента описывает высоту со средним варьированием 4,6 м с учетом двух стандартных отклонений (±3) с амплитудой около 27 м; вторая компонента, описывающая мезорельеф, отражает амплитуду варьирования около 4,9 м, третья компонента — 1,3 м, четвертая компонента — 66 см, пятая компонента — 35,28 см и шестая компонента — 26 см.

Таким образом, разложение рельефа по ортогональному базису методом главных компонент позволяет выделить из очень сложно варьирующей поверхности основные иерархические уровни организации, которые ни в коем случае не противоречат уровням, выделяемым по спектру. Эти иерархические уровни есть, скорее

Таблица 9.12

Основные параметры описания рельефа шестью компонентами

R^2 = 0,999; стандартная ошибка 4,0869 см

Переменная	Амплитуда *b* (см)	Критерий t(638)
Константа	1267,054	7873,722
1	458,378	2846,242
2	–82,522	–512,409
3	–23,873	–148,237
4	11,056	68,651
5	5,884	36,535
6	4,341	26,956

всего, результат наложения и совместного действия неизвестных нам физических факторов, каждый из которых порождает свой спектр квазиволновых пространственных структур. Возможность достаточно строго и однозначно выделить иерархические уровни пространственной организации рельефа позволяет более обоснованно выбирать уровни его картографического отображения, строить модели переноса влаги, использовать полученные формы рельефа как возможные факторы, определяющие состояние других компонентов ландшафта.

Возникает естественный вопрос: можно ли исследовать иерархическую пространственную организацию какого-либо другого компонента, например, растительности? При очень небольшой фантазии легко найти положительное решение. Обратим внимание на то, что при нивелировании рельефа мы определяем превышение одной точки над другой, т. е. измеряем при постоянном шаге нивелирования производную для рельефа или дистанции между соседними точками по высоте. Таким образом, нужно изыскать способ измерения аналогов превышения для любого компонента. Это можно сделать, если измерить дистанцию или различия между состояниями, допустим растительности, в соседних точках, используя любую из возможных метрик. Так как большинство метрик не имеют отрицательных значений, их нужно стандартизировать по среднему квадратическому отклонению. Тогда они могут рассматриваться как аналоги превышения. Теперь, как и при расчете рельефа, возьмем для ряда приращений накопленную сумму и в результате получаем, то, что можно назвать поверхностью растительности, почвы или животного населения. Далее применимы все рассмотренные выше методы анализа.

В заключение еще раз отметим, что, изучая проблему исследования пространственной организации по сути любого явления природы, мы попадаем в очень слабо разработанную, но весьма перспективную и актуальную область исследования.

Контрольные вопросы

1. Что такое стационарный ряд? Опишите основные параметры временно́го (пространственного) ряда.
2. Постройте модель сложного ряда, включающего случайную составляющую, гармонические колебания в нескольких полосах частот и тренд, и исследуйте его структуры методами анализа временны́х рядов.
3. Постройте прогноз для этого ряда.

ЗАКЛЮЧЕНИЕ

В заключение изложения и разбора статистических методов анализа данных обратим внимание, в первую очередь, на тот факт, что полное рассмотрение практически любого раздела заслуживает объема добротной монографии. Читатель должен иметь это в виду и ни в коем случае не считать содержание настоящего пособия исчерпывающим. Вместе с тем, перечень приведенных методов достаточен для решения относительно простых задач анализа данных, наиболее типичных в работе эколога и географа.

Необходимо помнить о том, что методы статистики находятся в постоянном развитии и совершенствовании. Это, в первую очередь, относится к анализу данных, отражающих отношения в нелинейных системах с большим вкладом нестационарной и неравновесной составляющих. В настоящее время следить за новациями в области анализа данных в мировой науке несоизмеримо легче, чем десять лет назад. Записав соответствующее ключевое слово в любой поисковой системе Интернета, вы выйдите на сайты, содержащие интересующую вас информацию. Правда, чтобы не запутаться в сетях Интернета, нужно действовать в соответствии с некоторым простыми правилами:

- чтобы сузить область поиска, к основному ключевому слову на английском языке через знак плюс добавьте слово «ecology», «plant» и т.п.;
- сайты на первых двух-трех страницах поисковой системы обычно в наибольшей степени отвечают содержанию запроса, однако интересная информация бывает и на последних страницах;
- внимательно просматривайте аннотацию каждого сайта, позволяющую дать оценку его содержания и полноты;
- обратите внимание на источник: организация, которой принадлежит сайт, его назначение и т.п.;
- сайты с расширением *pdf* скачивайте через *download*, продолжая работать в формате *htm*;
- используйте разные поисковые системы, так как каждая из них реализует несколько отличные схемы поиска. Первые страницы трех-четырех поисковых систем с высокой вероятностью дадут наиболее полное представление;
- повторяйте свои изыскания через один-два месяца.

Следует иметь в виду, что современные темпы развития науки таковы, что требуют от специалиста постоянного самообучения.

В настоящее время, кроме базовых знаний в своей области и в области науки в целом, специалист должен знать, где найти новые знания и уметь их взять. Под последним понимается умение увидеть содержательную ценность новой информации, вектор научного знания, который она развивает, область ее приложения к собственным текущим или будущим исследованиям. Очень важно научиться быстро осуществлять селекцию новых материалов, отбрасывая информацию, которую можно определить как «шум», сохранять и разбираться в наиболее содержательной информации. В этой операции всегда существует риск ошибки, но вместе с тем можно предложить некоторые критерии. К «шуму» можно отнести уже известные вам материалы или те, которые многократно повторяются в различных сайтах, программные средства, написанные для DOS, материалы с очень частными примерами. Ценная информация включает:

1. Материалы с неизвестными вам понятиями и терминами (по ним целесообразно сделать специальный поиск).

2. Материалы с высокой полнотой изложения вопроса и с хорошим разбором логического и алгебраического содержания теории или метода.

3. Глоссарии и терминологические словари.

4. Практические реализации метода для репрезентативных данных.

Несколько слов необходимо сказать о пакетах статистических программ. В целом их очень много. Некоторые программные продукты можно найти в свободном доступе в Интернете, но обычно такие продукты имеют относительно частный характер, хотя некоторые из них весьма полезны.

Можно рекомендовать следующую последовательность приобретения программных продуктов:

1. Первый пакет, который вы должны приобрести — это, безусловно, Statistica. Он содержит практически все необходимые методы, очень хорошую графику, связанную со средствами оперативного анализа данных, относительно приемлемую стоимость и неплохой русский перевод, объясняющий сущность конкретных методов. Недостатки пакета связаны с ограниченным набором методов кластер-анализа и ограниченным числом переменных при многомерном шкалировании. К недостаткам можно отнести необходимость специальной подготовки матриц дистанций на основе корреляции Спирмена и Кендала.

2. Если вам необходимо решать сложные задачи классификации и многомерного шкалирования, то лучше ориентироваться на пакет NCSS.

В общем, в большинстве случаев можно удовлетвориться этими двумя пакетами, однако для полноценной работы желательно иметь пакеты SPSS и SYSTAT. Каждый из них содержит то, чего нет в других пакетах. Наконец, наиболее мощным пакетом является пакет SAS. Из отечественных пакетов можно рекомендовать Stadia 6.

АННОТИРОВАННЫЙ СПИСОК ЛИТЕРАТУРЫ

Анализ данных — столь обширная область знания и творчества, что изложение всех ее методов и нюансов невозможно в какой-либо одной книге. Одни из них переносят акцент на теорию, другие — на процедуры вычисления, третьи — на тонкости решения конкретных задач и применения конкретных методов. Одним авторам удалось доступно изложить одни методы, другим — другие. В результате все монографии и учебники по теории и практике статистики фактически в той или иной области дополняют друг друга. Как невозможно написать единый, универсальный учебник статистики, так и невозможно один раз и навсегда освоить весь арсенал ее методов. Это — та область знаний, которой нужно учиться всю жизнь, постепенно поднимаясь на более высокие ступени. Впрочем умение самообучаться и постоянно учиться есть необходимое условие работы современного специалиста. Именно для содействия старту этому перманентному процессу в заключение приведем список базовой литературы и некоторых сайтов в Интернет, число которых прогрессивно растет.

Ниже приведем список литературы на русском языке, который включает наиболее важные публикации последней четверти прошлого столетия. В настоящее время многие издания 70—80-х годов XX в., которые заложили основы статистических методов анализа, довольно легко приобрести у букинистов.

Абросов Н. С., Боголюбов А. Г. Экологические и генетические закономерности сосуществования и коэволюции видов. — Новосибирск: Наука, 1988.

Изложены очень интересные и общие модели организации сообществ, которые можно использовать как основу для формулировки гипотез при организации полевых экологических исследований.

Абросов Н. С., Ковров Б. Г., Черепанов О. А. Экологические механизмы сосуществования и видовой регуляции. — Новосибирск: Наука, 1982.

Айвазян С. А. и др. Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989.

Очень полное изложение методов многомерного анализа, включая подробный разбор задач классификации с обучением и без обучения, обсуждение проблем выбора метрик, метода главных компонент, факторного анализа, многомерного шкалирования, разведочного статистического анализа. Можно рекомендовать для читателей, углубленно работающих в первую очередь над задачами классификации.

Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. — М.: Издательское объединение «ЮНИТИ», 1998.

Учебник, охватывающий все проблемы статистического анализа данных, с большим акцентом на многомерный параметрический и непараметрический анализ. Отдельные главы доступны для начинающего исследователя, другие — для достаточно подготовленного читателя.

Афанасьев В. Н., Юзбашев М. М. Анализ временных рядов и прогнозирование. — М.: Финансы и статистика, 2001.

Очень доступный для понимания учебник с рядом оригинальных и полезных оценок параметров рядов.

Баврин И. И. Высшая математика. — М.: Академия, 2001.

Учебник математики, отчасти адаптированный для студентов естественных факультетов. Может рассматриваться как введение в высшую математику. Имеется очень короткий раздел теории вероятностей и математической статистики.

Бендат Дж., Пирсол А. Прикладной анализ случайных данных. — М.: Мир, 1989.

Полное и последовательное изложение теории и методов анализа временных рядов для достаточно подготовленного читателя.

Большов Л. Н., Смирнов Н. В. Таблицы математической статистики. — М.: Наука, 1983.

Классическое изложение методов параметрического и непараметрического оценивания с очень полным, но в целом доступным изложением их теоретических оснований. В настоящее время, когда таблицы распределений имеют относительно небольшую ценность и заменены программными средствами, интерес к этому справочному изданию определяется очень полным и четким изложением теоретических основ критериальных распределений.

Боровиков А. А. Курс теории вероятностей. — М.: Наука, 1972.

Лаконичное изложение теории вероятностей с акцентом на теорию статистики. В частности рассматриваются элементы количественной теории информации и статистическое понятие энтропии. Книга весьма полезна как справочник.

Боровиков А. А. Математическая статистика. Оценка параметров, проверка гипотез. — М.: Наука, 1984.

Теоретические основания статистики. Книга рассчитана на хорошо подготовленного читателя.

Боровиков В. П. STATISTICA: искусство анализа данных на компьютере.

В книге изложена концепция и технология современного анализа данных на компьютере. На основе элементарных понятий описываются углубленные методы анализа в системе STATISTICA с иллюстрацией многочисленными примерами из экономики, маркетинга, рекламы, бизнеса, медицины, промышленности и других областей знаний. Описываются классические и современные методы анализа данных, позволяющие получить всестороннее описание, провести классификацию, найти закономерности и зависимости в отношениях между переменными. Предлагаемая книга адресована самому широкому кругу читателей, желающих стать профессионалами в анализе данных на STATISTICA в бизнесе, финансах, управлении, экономике, промышленности, страховании, медицине и других приложениях. Книга дополнена компакт-диском, на котором представлена последняя, существенно дополненная версия знаменитого электронного учебника StatSoft по анализу данных, а также учебником по промышленной статистике, материалы обучающих курсов, демо-версии систем STATISTICA и STATISTICA *Neural Networks*, огромным количеством данных для обучения и проведения самостоятельных исследований в STATISTICA и SNN. http://www.statsoft.ru/home/registration/order_books.asp.

Боровиков В. П. Программа STATISTICA для студентов и инженеров.

Книга является первым шагом к знакомству с программой STATISTICA для статистического анализа данных в среде Windows, а также с нейронными сетями пакета SNN, дополняющими классические методы анализа данных. На простых, доступных каждому пользователю, примерах (описательная статистика, регрессия, дискриминантный анализ, кластерный анализ, анализ выживаемости и др.), взятых из различных сфер жизни, показаны возможности системы по анализу данных. В приложении даны краткие материалы по панели инструментов, языку STATISTICA BASIC и др. Второе издание книги существенно дополнено новыми материалами по анализу выживаемости, интенсивно применяемому в современных медицинских исследованиях и страховании, а также введением в нейронные сети, реализованные в пакете SNN фирмы StatSoft. Получившая прекрасные отзывы и написанная легким, изящным языком книга адресована самому широкому кругу читателей, желающих сделать самостоятельные шаги в анализе данных.

Боровиков В. П., Ивченко Г. И. Прогнозирование в системе STATISTICA® в среде Windows (основы теории и интенсивная практика на компьютере). — М.: Финансы и статистика, 1998.

Книга содержит описание практических методов прогнозирования в системе STATISTICA вместе с изложением необходимых теоретических основ. Состоит из двух взаимосвязанных частей. В первой части описано прогнозирование временны́х рядов с помощью признанной во всем мире системы статистической обработки данных STATISTICA®. Во второй части в сжатом виде изложены математические основы статистического прогнозирования. Эта часть необходима для углубленного понимания первой, практической части книги. В основу книги положен курс, читаемый в Московском государственном институте электроники и математики. Книга рассчитана на студентов, научных работников, аналитиков и специалистов, использующих методы прогнозирования в своей повседневной деятельности.

Боровиков В. П., Боровиков И. П. STATISTICA. Статистический анализ и обработка данных в среде Windows. — М.: Изд-во «Филинъ», 1997.

Подробное практическое руководство по работе с наиболее популярным и действительно наиболее эффективным пакетом статистических программ.

Гантмахер Ф. Р. Теория матриц. — М.: Наука, 1967.

Одно из лучших руководств по теории матриц, доступное в базовых разделах для слабо подготовленного читателя.

Громыко Г. Л. Теория статистики. — М.: Изд-во ИНФРА-М, 2000.

Общедоступный учебник по основам статистики, включающий задачи одномерного анализа с элементами анализа временных рядов. Учебник адаптирован для студентов экономических факультетов, что, однако, не снижает его общности.

Дэйвесон М. Многомерное шкалирование. — М.: Финансы и статистика, 1988.

Практически полное изложение теории и методов многомерного шкалирования. Лучшая литература по проблеме на русском языке.

Дженкинс Г., Ваттс Д. Спектральный анализ и его приложение. Вып. 1, 2. — М.: Мир, 1971.

Бокс Дж., Дженкинс Г. Анализ временных рядов, прогноз и управление. Т. 1, 2. — М.: Мир, 1974.

В обеих книгах, дополняющих друг друга, систематически изложены теория и методы анализа временных рядов в задачах исследования и теории управления. Это базовые книги, послужившие основой для современных методов анализа временных рядов и прогноза будущего на основе знания прошлого. Материал изложен в форме, доступной для достаточно подготовленного читателя. Это литература, к которой приходится обращаться многократно по мере возникновения новых проблем в решении конкретных задач и постепенного углубления в сущность методов.

Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. — М.: Наука, 1970.

Один из лучших, если не самый лучший справочник практически по всем проблемам математики, отличающийся исключительной лаконичностью и точностью изложения тем. Разделы теории множеств, теории вероятностей, статистики, с одной стороны, изложены с необходимой полнотой, а с другой — с максимальной четкостью и ясностью. Книга выдержала несколько изданий и, безусловно, не устарела и в настоящее время.

Кроновер Ричард М. Фракталы и хаос в динамических системах. — М.: ПОСТМАРКЕТ, 2000.

Последовательное изложение представлений о фракталах и областей их применения.

Кулаичев А. П. Компьютерный контроль процессов и анализ сигналов. — М.: НПО «Информатика и компьютеры», 1999.

Книга посвящена современным технологиям анализа временных рядов и фактически является подробным приложением к пакету программ *CONAN-m*. Стоимость этой системы, хотя и вполне реалистична, но превышает финансовые возможности экологов и географов. Однако сама книга содержит важные интересные идеи и реализации и очень полезна для активно работающего исследователя.

Лившиц Н.А., Пугачев В.Н. Вероятностный анализ систем автоматического управления. — М.: Советское радио, 1963.

Книга посвящена теории управления, однако по существу является очень полным и глубоким изложением теории случайных процессов, их статистического анализа. Изложение методов статистики осуществляется в прямой связи с разными типами моделей систем. В результате статистика становится средством описания поведения систем и их исследования. Основные результаты связаны с проблемами исследования и описания поведения во времени, что делает ее очень полезной при углубленном освоении теории и методов анализа временных рядов.

Лоев М. Теория вероятностей. — М.: Иностранная литература, 1962.

Классическая монография по теории вероятностей.

Мандельброт Б. Фрактальная геометрия природы. — М.: Изд-во Ин-та компьютерных исследований, 2002.

Перевод классической основополагающей монографии, автора термина «фрактал» и всей концепции фрактальных множеств. Особый, романтический стиль изложения создает неповторимое удовольствие при ее чтении, дающем в конечном итоге максимально полное представление проблемы.

Марпл-мл. С.Л. Цифровой спектральный анализ и его приложения. — М.: Мир, 1990.

Одна из наиболее полных монографий по проблеме анализа временных рядов. Описание методов двухмерного спектрального анализа и двухмерной авторегрессии позволяет использовать это руководство для анализа структуры пространства в географии. Книга доступна для достаточно подготовленного читателя.

Могилев А.В., Н.И.Пак, Е.К.Хеннер. Информатика. — М.: Академия, 2002.

Учебное пособие, охватывающее очень широкий круг проблем, которые авторы объединяют понятием «Информатика». Книга включает классические представления теории информации, как науки о передаче и преобразовании «сообщений» в самом общем понимании этого слова, так и обширные сведения по программному обеспечению современных персональных ЭВМ при различных аспектах их использования, языки и методы программирования, сведения о вычислительной технике, компьютерных сетях, телекоммуникации, информационных системах, под которыми имеются в виду базы информации и данных, автоматизированные информационные системы, экспертные и обучающие системы, а также математическое моделирование в различных областях науки и практики. Обширность предмета, несмотря на большой объем книги, позволила авторам в большинстве случаев сформулировать и объяснить основные проблемы по каждому направлению, что ставит эту книгу в ряд очень полезных справочников, опираясь на который можно начать профессиональное освоение самых различных аспектов работы с информацией. Книга весьма удобна для самообразования и может быть рекомендована начинающим исследователям и инженерам.

Количественные методы в почвенной зоологии / М.С.Гиляров, Б.Р.Стриганова. — М.: Наука, 1987.

Одно из немногих методических руководств, в котором изложены методы анализа полевых данных.

Нейман Ю. Вводный курс теории вероятностей и математической статистики. — М.: Наука, 1963.

Классическое руководство с изложением основ теории вероятностей и математической статистики в форме, максимально адаптированной для неподготовленного читателя. Автор пытается донести до читателя глубокий внутренний смысл теории вероятностей и статистики. В книге почти нет формул, в первую очередь автор стремится объяснить существо дела. В конечном итоге эта книга может быть определена как объяснение теории вероятностей и статистики на естественном языке. Работа с этой книгой весьма полезна, так как показывает предмет в необычном, но очень важном ракурсе.

Плохиниский Н.А. Биометрия. — М.: Изд-во МГУ, 1970.

Первый учебник статистики для биологов. В основном изложены методы одномерного анализа без каких-либо теоретических оснований. Приведено много примеров расчетов конкретных параметров. В настоящее время имеет больше историческое, чем практическое значение.

Рабинович М.И., Турбин Д.И. Введение в теорию колебаний и волн. — М.; Ижевск: Изд-во R&D. Dinamics, 2000.

Изложение базовых положений теории колебаний и динамических моделей. Книга важна как база поиска моделей для интерпретации результатов статистического анализа временных рядов.

Савельев Л. Я. Комбинаторика и вероятность. — Новосибирск: Наука, 1975.

Одна из немногих книг, в которой на очень высоком уровне рассматривается связь комбинаторики и теории вероятностей. Вообще представления о комбинаторике являются базовыми для понимания фундаментальных процессов, связанных с поведением ансамблей частиц. Без представлений комбинаторики невозможно понять смысл непараметрических критериев статистики. В книге очень просто и наглядно рассматриваются теоретико-множественные основания комбинаторики и теории вероятностей. Теория вероятностей вводится на комбинаторной основе, что после аксиоматизации ее теории делается не часто. Вместе с тем, представление о вероятности через конечные модели часто более наглядно и более доступно для понимания.

Свирежев Ю. М. Нелинейные волны, диссипативные структуры и катастрофы в экологии. — М.: Наука, 1987.

Развитие теории нелинейных колебаний для популяций и сообществ. Книга важна как база для поиска моделей, для постановки и интерпретации результатов полевых исследований.

Свирежев Ю. М., Логофет Д. О. Устойчивость биологических сообществ. — М.: Наука, 1976.

Модель экологических систем от уровня популяций до уровня сообществ. Основа для формулировки гипотез, проверяемых в полевых исследованиях.

Сошникова Л. А. и др. Многомерный статистический анализ в экономике. — М.: Изд-во UNITY, 1999.

В книге, которую можно рассматривать как справочное пособие, в достаточно популярной форме изложены основные методы многомерного анализа с некоторой адаптацией в область экономических исследований. Однако адаптация сводится фактически только к демонстрациям возможностей методов. Это делает книгу пригодной для работы специалистов во всех областях науки и практики, где приходится иметь дело с анализом данных. В книге наряду с изложением собственно методов многомерного анализа (многомерная регрессия, факторный анализ, многомерное шкалирование, кластерный анализ, дискриминантный анализ, каноническая корреляция) изложены основные положения общей теории статистики. Книга может быть рекомендована для начинающего исследователя.

Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Релстона, Г. С. Уилфа. — М.: Наука, 1986.

Книга представляет сборник тематически объединенных статей широко известных авторов, таких как Г. О. Хартли, Р. Р. Хогинг, Дж. Б. Краскелл и др. В каждом разделе тщательно разбираются математические основания методов и приводятся блок-схемы программ и сами программы на Фортране. Сочетание теории и изложения программ весьма способствует пониманию сути самих методов и основ логики программирования. В книге рассматриваются методы многомерного анализа, регрессионный и дискриминантный анализ, метод главных компонент и факторный анализ, кластерный анализ, многомерная пошаговая регрессия и временные ряды.

Особенно доступно изложен метод многомерного непараметрического шкалирования и дискриминантный пошаговый анализ. Книга может быть рекомендована специалистам, заинтересованным в более глубоком понимании сути используемых ими методов.

STATISTICA — Краткое руководство пользователя.

В книге изложены основные принципы работы с системой, рассматриваются панели инструментов, пользовательский интерфейс, файлы данных, практические примеры использования пакета. Отдельная глава посвящена настройке системы. Также книга содержит исчерпывающий справочник, который представляет собой краткие сведения о наиболее часто используемых соглашениях, функциях и возможностях системы STATISTICA, и предметный указатель.

Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. — М.: Инфра, 1998.

Очень хорошее руководство для освоения базовых положений и методов статистики, к сожалению, ориентированное на популярный всего несколько лет назад пакет Statgrafics и русскоязычный пакет STADIA. Оба пакета по разным причинам потеряли свою популярность. Пакет Stargrafics по возможностям и организации интерфейса был наилучшим для среды DOS, но не смог создать удобный интерфейс для среды Windows и потерял все свои преимущества, став мало удобным для пользователя. Однако полезные свойства руководства, построенного на основе этого пакета, сохранились. Важной особенностью этого учебника нового поколения является сочетание базовых положений теории и руководства по использованию пакета программ. По форме и содержанию это очень хороший учебник для начинающих.

Яглом А. М., Яглом И. М. Вероятность и информация. — М.: Наука, 1973.

Подробное изложение соотношения теории вероятностей и теории информации. Модели теории информации. Представления об энтропии. Для эколога книга полезна как основа общей теории разнообразия.

ОГЛАВЛЕНИЕ

Учебное издание

Пузаченко Юрий Георгиевич

Математические методы в экологических и географических исследованиях

Учебное пособие

Редактор *Л. В. Честная*
Технический редактор *Н. И. Горбачева*
Компьютерная верстка: *В. Н. Канивец*
Корректоры *В. А. Жилкина*, *Н. В. Савельева*

Диапозитивы предоставлены издательством.

Изд. № А-821-I. Подписано в печать 18.03.2004. Формат 60×90/16.
Гарнитура «Таймс». Бумага тип. № 2. Печать офсетная. Усл. печ. л. 26,0.
Тираж 5100 экз. Заказ №12981.

Лицензия ИД № 02025 от 13.06.2000. Издательский центр «Академия».
Санитарно-эпидемиологическое заключение № 77.99.02.953.Д.003903.06.03 от 05.06.2003.
117342, Москва, ул. Бутлерова, 17-Б, к. 223. Тел./факс: (095)330-1092, 334-8337.

Отпечатано на Саратовском полиграфическом комбинате.
410004, г. Саратов, ул. Чернышевского, 59.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЭКОЛОГИЧЕСКИХ И ГЕОГРАФИЧЕСКИХ ИССЛЕДОВАНИЯХ

ISBN 5-7695-1348-9

Издательский центр «Академия»