Л.И. Дубровская, Г.Б. Князев

КОМПЬЮТЕРНАЯ ОБРАБОТКА ЕСТЕСТВЕННО-НАУЧНЫХ ДАННЫХ МЕТОДАМИ МНОГОМЕРНОЙ ПРИКЛАДНОЙ СТАТИСТИКИ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Л.И. Дубровская, Г.Б. Князев

КОМПЬЮТЕРНАЯ ОБРАБОТКА ЕСТЕСТВЕННО-НАУЧНЫХ ДАННЫХ МЕТОДАМИ МНОГОМЕРНОЙ ПРИКЛАДНОЙ СТАТИСТИКИ

Учебное пособие

Допущено УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению «Гидрометеорология»

ТОМСК
2011

УДК 519.237: (556.5+549)
ББК 22.172: (26.22+26.3)
Д79

Рецензенты:
О.Г. Савичев – д-р геогр. наук, проф. Томского государственного политехнического университета,
А.А. Ельцов – канд. техн. наук, проф. кафедры высшей математики Томского государственного университета систем управления и радиоэлектроники

Дубровская Л.И., Князев Г.Б.

Д79 Компьютерная обработка естественно-научных данных методами многомерной прикладной статистики: Учебное пособие. – Томск: ТМЛ-Пресс, 2011. – 120 с.

ISBN 5-91302-114-2

В учебном пособии изложены методы, наиболее часто используемые при обработке естественно-научных данных дискриминантный анализ, множественная регрессия, кластерный анализ, метод главных компонент, факторный анализ. Для иллюстрации особенностей компьютерной реализации каждого метода приведены примеры решения задач гидрометеорологического и геохимического содержания с использованием пакета Statistica или электронной таблицы Excel.

Данное пособие разработано для обеспечения дисциплины федерального компонента ДПМ 03 «Компьютерные технологии в гидрометеорологии» для магистров направления 020600 – «Гидрометеорология» (Образовательные магистерские программы: 020600.68.01 «Гидрология суши», 020600.68.10 «Экологическая климатология», 020600.68.05 «Метеорология»), также оно может быть полезно студентам, аспирантам и молодым научным сотрудникам других специальностей для приобретения навыков корректного использования статистических методов в научно-исследовательской работе

УДК 519.237: (556.5+549)
ББК 22.172: (26.22+26.3)

ISBN 5-91302-114-2

ПРЕДИСЛОВИЕ

Чистая математика делает то, что можно, так, как нужно;
прикладная – то, что нужно, так, как может.
Перефразировка известного выражения

В настоящей книге читатель найдёт не столько теоретическое описание количественных методов анализа в таких разделах науки о Земле, как гидрометеорология и геохимия, но, главным образом, обширное и подробное изложение особенностей и тонкостей применения методов многомерной прикладной статистики на многочисленных примерах из этих областей. Будет верно сказать, что изложение многомерного статистического анализа построено в данной книге в духе прикладной математики.

Применение статистического анализа в любой конкретной прикладной области обладает специфическими особенностями, как в выборе методов исследования, так и в оценках и интерпретации результатов, зачастую требующих профессиональной интуиции в данной предметной области. Поэтому даже при наличии большого выбора литературы по статистическим методам общего плана остается необходимость в учебниках, методических пособиях и узкопрофильных указаниях.

Ранее, до появления персональных компьютеров, анализ реальных данных был чрезвычайно сложным делом, требующим больших интеллектуальных усилий, и ни о каких технологиях не могло быть и речи. Это было дело небольшого круга профессионалов.

В последнее время в связи с развитием вычислительной техники, особенно с появлением высокопроизводительных персональных компьютеров, широкое распространение получили готовые пакеты (комплексы программ) для статистической обработки естественно-научной информации, как универсального, так и специализированного назначения. Благодаря таким системам, как, например, пакеты Statistica, Statgraphics, SPSS и др., открылся путь к новым технологиям анализа данных, максимально сокращающий рутинные процедуры и делающий анализ доступным для широкого круга пользователей.

Однако оборотной стороной массовости и доступности пакетов статистической обработки данных стало некритичное отношение пользовате-

лей к внутреннему содержанию анализа, границам его применимости, вследствие чего столь же массовыми стали правдоподобные ошибочные результаты.

Например, очень часто студентами полностью игнорируется этап проверки адекватности регрессионной модели, получаемой с целью прогноза или восстановления отсутствующих данных. В научных работах при использовании факторного анализа зачастую даже при невозможности получения простой структуры факторных нагрузок производится интерпретация структуры зависимостей и делаются выводы, хотя этот случай следовало бы классифицировать как неудавшееся исследование. Поэтому задачами данного учебного пособия являются: обучить читателя осознанно и правильно пользоваться стандартными программами, отдавая себе отчёт о достоверности полученных результатов; показать читателям, на какие результаты вообще можно рассчитывать, применяя данный метод, и как избежать явных ошибок.

Существуют прекрасные учебники по статистическим методам для гидрологов (Рождественский, Чеботарёв, 1974; Смирнов, Скляренко, 1986; Дружинин, Сикан, 2001; Шелутко, 1991; Румянцев, Бовыкин, 1994, 2009 и др.), метеорологов (Алексеев, 1971; Алёхин, 1963; Белоцерковский, 1993 и др.), геологов (Дэвис, в двух книгах, 1990; Каждан, Гуськов,1990; Миллер, Канн Дж., 1967) и т.д. Что же послужило в таком случае поводом для создания данного учебного пособия? Причин на этого несколько:

– во-первых, часть вышеперечисленных трудов, являющаяся на данный момент «классикой», опубликована 10–30 лет назад, не переиздавалась; многие из них перешли в разряд библиографических редкостей и имеются в единичных экземплярах в библиотеках и на кафедрах;

– во-вторых, написанные на высоком теоретическом уровне, они не имели целью рассмотрение прикладных аспектов реализации статистических методов в компьютерных пакетах обработки данных;

– в-третьих, за последнее десятилетие произошло стремительное обновление версий статистических пакетов, перевод их в Windows-интерфейс, что существенно упростило обработку данных и, соответственно, расширило контингент пользователей;

– в-четвёртых, развитие геоинформационных технологий, основным продуктом которых являются тематические карты с возможностью анализа, а также необходимость решения задач классификации и районирования, требуют формирования у выпускников естественных факультетов вузов квалифицированных знаний статистических методов и практиче-

ских навыков работы в специализированных и универсальных пакетах обработки данных.

Если в США статистической обработке данных начинают учить со школы, то у нас этому обучают далеко не на всех факультетах высших учебных заведений. Появление пакета Statistica 7.0 и его русифицированной версии значительно облегчило как преподавание, так и эффективность освоения статистических методов обработки информации, в том числе на геолого-географическом факультете Томского государственного университета.

Авторы старались, насколько возможно, соединить строгость математического изложения с доступностью интерпретации чисто практических моментов, вынуждающих к отступлениям от строгих требований теории. Этот подход вынесен в качестве эпиграфа в предисловие. Теоретически достаточно разработанные методы часто основаны на сильных ограничениях, которые обычно не выполнимы при исследовании гидрометеорологических, геохимических и т.п. данных.

К сожалению, пробы исследователи вынуждены брать только там, где это возможно. Гидрологи и метеорологи также работают большей частью в условиях недостаточности данных и, тем не менее, необходимо из них извлекать сведения, изучать недостатки этих сведений и выявлять имеющиеся тенденции.

Например, критерий Стьюдента для проверки однородности ряда по среднему, строго говоря, применим только к рядам, распределённым по нормальному закону. Тем не менее, он рекомендуется как основной в официальных документах (Пособие... 1984) при гидрологических исследованиях, несмотря на характерную для гидрометеорологических данных асимметричность.

Авторы не скрывают своего стремления подражать в методологическом плане и стиле изложения превосходной книге Дж.С. Дэвиса «Статистический анализ данных в геологии». И хотя настоящее учебное пособие имеет чёткую нацеленность преимущественно на компьютерную обработку данных, цель его та же, что и в вышеупомянутой книге. Как образно сформулировал Дж.С. Дэвис, «на практике большинство исследователей осваивают методы количественного анализа читая, спрашивая, учась на своих ошибках. Этот неудовлетворительный и даже опасный метод обучения, возможно, сравним с обучением врача в процессе работы. ...Эта книга может помочь организовать процесс самообучения, а именно даёт возможность сделать первые шаги к познанию описанных в ней алгоритмов».

Методически пособие написано по принципу дедукции – от общего к частному, сначала теоретическая часть метода, затем его реализация на примерах из гидрометеорологии и геохимии.

Теоретическая часть составлена на основании источников, список которых приведён в конце учебного пособия, оригинальными являются приложения методов к конкретным задачам, особенности их реализации в пакете Statistica, обсуждение и интерпретация результатов.

Предисловие, главы 1, 4 и 5 написаны доцентом, кандидатом физико-математических наук Л.И. Дубровской, параграф 1.4 и глава 3 – доцентом, кандидатом геолого-минералогических наук Г.Б. Князевым.

1. ПОНЯТИЕ О МНОГОМЕРНОМ СТАТИСТИЧЕСКОМ АНАЛИЗЕ

Многообразие свойств природных объектов и многофакторность природных процессов приводит исследователя к проблеме обработки огромной массы индивидуальных наблюдений. В многомерных данных каждый *объект* наблюдений характеризуется множеством *признаков* (переменных). Многомерные методы позволяют одновременно изучать изменение набора характеристик. Конечной целью большинства многомерных статистических методов является предсказание (прогнозирование) тех или иных свойств изучаемых объектов, будь то гидрометеорологические, экологические и т.д.

Можно привести много примеров гидрометеорологических и геохимических данных, к которым применимы методы многомерного анализа. Среди них можно назвать химические анализы, в которых переменные представляют собой содержание микро- или макро элементов в воде, почве, снеге. Примером многофакторного процесса может служить речной сток, являющийся результатом взаимодействия многих геофизических процессов (прямая и рассеянная радиация, осадки, температура воздуха и подстилающей поверхности, давление и влажность воздуха, скорость и направление ветра и т.д.) физико-географических условий бассейна (ландшафт, почвы, геологическое строение, растительность) и т.д.

Многомерные методы позволяют исследователю работать с большим числом переменных, объём которых невозможно обработать вручную без компьютера. Однако эти методы сложны как с методологической, так и с теоретической точки зрения. Статистические критерии и процедуры большей части этих методов разработаны лишь при очень сильных ограничениях, а поведение при решении реальных задач изучено слабо.

Некоторые процедуры многомерного анализа совсем не имеют теоретического обоснования, для них не созданы ещё критерии проверки соответствующих гипотез. Например, до сих пор не разработаны способы оценки адекватности результатов кластер-анализа. Тем не менее, эти методы «работают» и дают неплохие результаты при условии сочетания их с профессиональным опытом и интуицией исследователя в конкретной предметной области, то есть реализуется принцип ***«доказать нельзя, а использовать можно»***.

Есть два пути решения проблемы обработки многомерных данных: 1) отсечь часть малоинформативных характеристик и возвратиться к мало-

размерным классическим задачам; 2) объединить характеристики в группы (в дальнейшем – факторы) для уменьшения признакового пространства. Второй подход вылился в задачу обратного сведения множества характеристик к небольшому ряду обобщающих параметров, выражающих реально существующие закономерности в наборе данных, и соответственно родилось направление, получившее название «многомерный анализ» (Факторный, дискриминантный и кластерный анализ, 1989).

Развитие многомерного статистического анализа как науки началось с 1901–1904 гг. В это время появились статьи К. Пирсона и Ч. Спирмена, посвящённые теории факторного анализа. Методы многомерного статистического анализа базируются на представлении исходной информации в многомерном признаковом пространстве и позволяют определять неявные, но объективно существующие закономерности в данных и тенденциях развития изучаемых явлений и процессов.

Круг основных теоретических и практических задач, решаемых с помощью методов многомерной статистики, заключается в анализе и выявлении связей внутри комплекса исходных признаков; выделении групп случайных признаков, обладающих наиболее сильными связями; оценке вклада ведущих признаков и факторов (последние представляют комплекс генетически однородных характеристик) в общую дисперсию; в типизации (группировке) объектов в многомерном пространстве.

Постепенно в многомерном анализе образовались разделы, взаимодополняющие друг друга – кластерный анализ, таксономия, распознавание образов, метод главных компонент, факторный анализ (Харман, 1972; Тьюки, 1981).

Особо стоит сказать о задаче классификации, одной из важнейших в обработке естественно-научных данных. Под решением *задачи классификации* понимается установление правил отнесения объекта к одной или нескольким группам (категориям, классам) на основании некоторого числа его характеристик (признаков) и построение описаний классов. Само же отнесение объекта к тому или иному классу с известным описанием называется *идентификацией*. В данной работе по примеру многих изданий термин классификация используется в широком смысле, включая идентификацию.

Если совокупность объектов разбивается на группы (классы) на основании одного признака, то классификация называется *монотетической*. Если для построения классификации используется несколько признаков одновременно, то она называется *политетической*. Из наиболее известных методов ярким представителем политетического метода является

дискриминантный анализ (ДА). Число признаков, используемых в ДА, как правило, невелико.

Кроме того, задачу классификации можно решать с помощью методов *автоматической классификации*, которая учитывает все признаки объектов. Это методы кластерного анализа, численной таксономии и т.д. Они относятся к группе методов распознавания образов.

Разновидностью политетической классификации являются методы факторного анализа (включая метод главных компонент). В них классификация строится на основе нескольких обобщённых показателей, именуемых факторами, компонентами.

1.1. Характеристика методов многомерного статистического анализа

Для успешного изучения методов многомерной статистики необходимы знания в таких областях высшей математики как аналитическая геометрия, матричная алгебра, многомерный математический анализ. Практически все методы многомерной статистики условно можно разбить на две группы, представленные на схеме (рис. 1).

Рис. 1. Схема методов математической и прикладной статистики (Горелова, Кацко, 2005). ФПВ – функция плотности вероятностей случайной величины (СВ)

В данном учебном пособии рассматриваются только методы многомерной прикладной статистики. Эту группу методов ещё называют мно-

гомерный статистический анализ данных. Характерной особенностью этих методов является геометрическое представление данных. Наблюдаемые объекты можно изобразить как точки в n-мерном пространстве, соответствующем числу признаков, которыми они характеризуются. Если признаки разнородны, то их нормируют.

Назначение методов многомерного статистического анализа данных не является однозначным. Например, задачи группирования объектов по принципу сходства признаков можно осуществлять и кластерным и факторным анализом. У каждого метода есть свои сильные и слабые стороны. Ниже приводится краткая сравнительная характеристика основных методов многомерного статистического анализа данных, рассматриваемых в данном учебном пособии.

Множественная корреляция и множественный регрессионный анализ. Множественная корреляция используется для установления *степени тесноты связи* между признаками объекта, а множественный регрессионный анализ для определения *вида этой связи*. Конечная цель регрессионного анализа – построить по корреляционной матрице уравнение регрессии, по которому можно содержательно интерпретировать результаты наблюдений и осуществлять прогноз. Одним словом, множественная корреляция и множественный регрессионный анализ применяются для измерения и моделирования связей изучаемых признаков и объектов.

В гидрометеорологии очень широко используется метод множественной регрессии в двух направлениях – для восстановления по уравнению регрессии пропущенных данных и в целях прогноза. Например, по известным рядам наблюдений за расходами воды на реках-аналогах и частично на контрольной реке, можно получить уравнение регрессии и по нему рассчитать (восстановить) часть отсутствующих данных на контрольной реке при соблюдении, естественно, определённых требований, как к рекам-аналогам, так и к самому уравнению регрессии.

Имея многолетние данные по осеннему увлажнению бассейна, запасам снега, жидким осадкам за период снеготаяния и т.д., можно найти связь между объёмом половодья и перечисленными факторами формирования этого процесса, а затем по ней спрогнозировать с определённой заблаговременностью объём половодья в текущем году.

Дискриминантный анализ является мощным статистическим средством решения классификационных задач, т.е. разделения *(дискриминации)* многомерных нормально распределённых совокупностей на группы. На основании имеющихся данных (обучающая выборка, «учитель») формулируется правило, по которому новые единицы исследуемой совокуп-

ности относятся к одному из существующих классов, при этом новые классы не образуются. Таким образом, производится расклассификация новых объектов по известным «эталонным группам».

Чаще всего дискриминантный анализ используется просто для разделения совокупностей на два класса, например, отделения загрязнённых территорий от незагрязнённых по какому-либо химическому элементу.

В гидрометеорологии дискриминантный анализ применяется чаще в области улучшения качества гидрометеорологических прогнозов в сочетании с другими статистическими методами, например, с методом множественной регрессии. В большинстве случаев дискриминантному анализу подвергаются два класса объектов (например, совокупность ситуаций выше нормы и ниже нормы).

В дискриминантном анализе число классов (групп) задаётся заранее. Дискриминантный анализ называют классификацией с «учителем».

Кластерный анализ – это совокупность методов, предназначенных для разбиения множества объектов на однородные группы (кластеры). В большинстве методов кластерного анализа заранее неизвестно, сколько классов будет выделено в данной совокупности объектов. В отличие от дискриминантного анализа кластерный анализ называют классификацией без «учителя», потому что его методы не используют обучающую выборку.

Задачи кластерного анализа:

– классификация объектов с учётом признаков, определяющих их природу;

– проверка предположения о структуре данных;

– построение новых классов.

Например, можно на основании проб геохимических показателей в снежном покрове провести разбиение территории города на классы по уровню техногенной нагрузки. Имея таблицу с набором физико-химических показателей с разных месторождений такого полезного ископаемого как торф, можно разбить торфа на классы по тому или иному направлению использования в промышленности или сельском хозяйстве.

Факторный анализ применяется в основном для «сжатия» данных, т.е. сведения множества элементарных признаков к небольшому числу «обобщённых признаков», и выявления латентных (скрытых, неизмеряемых) факторов. Эта же задача может решаться не только относительно признаков, но также и объектов.

Факторный анализ после выявления обобщённых показателей можно использовать для целей классификации. Например, с его помощью можно решать задачу районирования территории по условиям формирования

стока или по данным химических анализов осуществлять иерархическое группирование территории по уровню превышения ПДК и т.д.

Результаты, полученные на основе статистических методов, зависят как от правильности выбора самих методов, так и точности исходных данных. К прикладной статистике полностью применимо замечание английского натуралиста Гексли: «*математику можно сравнить с мельницей превосходного устройства, которая перемалывает что угодно до любой тонкости. Тем не менее, то, что вы получите, зависит от того, что вы засыпаете. И как самая великолепная мельница в мире не доставит вам пшеничную крупчатку из лебеды, так и страницы формул не доставят вам определённого результата из сомнительных данных*».

1.2. Краткие сведения о пакетах статистической обработки данных

Рынок компьютерных программ анализа данных очень разнообразен, что является отражением многоплановости задач обработки экспериментальных данных в различных областях человеческой деятельности. Все пакеты, как зарубежные, так и отечественные делятся на три группы: профессиональные, универсальные и специализированные.

1. Профессиональные могут работать с очень большими базами данных и имеют узкоспециализированные методы (SAS, BMPD). Стоимость таких пакетов составляет от 2 до 10 тыс. долларов.

2. Универсальные (или пакеты общего назначения) близки к профессиональным, но имеют меньшие возможности и более доступны по цене. Из зарубежных универсальных наиболее известны Statistica, SPSS, Statgraphics, S-Plus; из отечественных – Stadia, Olimp и др.

3. Специализированные пакеты содержат несколько методов (1–2), например, анализ временных рядов (Эвриста, Мезозавр, СтатЭксперт), методы классификации (Класс-мастер), контроль качества продукции и т.д. Полный сравнительный анализ современных статистических пакетов можно посмотреть в аналитической статье С.А. Айвазяна, В.С. Степанова (см. ссылку на электронные ресурсы).

Средства статистической обработки данных часто включают в электронные табличные процессоры (например, Excel), но для большей достоверности лучше проводить анализ в специальных статистических пакетах.

Методы многомерного анализа данных с разной степенью полноты представлены во всех универсальных статистических пакетах. В данном учебном пособии рассматриваются приёмы работы с универсальным па-

кетом Statistica и параллельно, там, где это возможно, с более лёгкими в освоении статистическими функциями Excel. Следует отметить, что «тяжеловесные» пакеты SPSS и Statistica предназначены для пользователей, владеющих статистическими методами на профессиональном уровне. Однако многолетний опыт авторов по обучению статистическим методам на базе этого пакета позволяет говорить об эффективности его освоения студентами.

Интегрированная система комплексного статистического анализа и обработки данных Statistica занимает одно из первых мест в мире среди программ статистической обработки данных. Первая версия пакета была разработана фирмой Statsoft Inc. (США) в 1991 г., в последние годы появились версии 7.0rus и 8.0. Начиная с версии 6.0, пакет полностью подстроен под стандартный Windows-интерфейс.

Пакет состоит из 19 специализированных статистических модулей, обладает мощной графической системой визуализации данных и результатов, имеет специальный инструмент для создания отчетов, встроенные языки программирования SQL, Statistica Basic и макрокоманд, может обрабатывать очень большие массивы наблюдений (корреляционные матрицы размером 32 000×32 000). Поддерживает все стандарты: импорт из популярных электронных таблиц, публикация в Internet, мастер запросов к ODBC-базам данных.

1.3. Основные способы представления многомерных данных

Статистика имеет дело с совокупностями объектов, описываемых некоторыми свойствами (качественными или количественными характеристиками). Если каждый объект имеет одну характеристику, то принято говорить об *одномерных данных*. Если таких характеристик у каждого объекта две и более, то данные рассматриваются как *многомерные*. Изучение природных (генеральных) совокупностей заключается в измерении характеристик объектов и получении выборочных совокупностей или выборок. Генеральная совокупность характеристик природных объектов может рассматриваться как случайная величина соответствующей мерности, свойства которой оцениваются с помощью выборки.

Первичные наблюдения (*выборку* многомерной случайной величины) в науках о Земле обычно формируют в виде таблицы размером $n \times m$, где где n – число строк, соответствующее числу объектов, попавших в выборку; m – число столбцов, содержащих характеристики (измерения) ка-

ждого объекта. Статистика же оперирует матрицами. *Матрицей* называется прямоугольная таблица из чисел, содержащая некоторое количество строк (***n***) и некоторое количество столбцов (***m***). Если ***m*** = ***n***, матрица называется *квадратной*, а число ***m*** или ***n*** – её *порядком*. *Рангом* матрицы называется число линейно независимых строк (или столбцов) матрицы. Квадратными являются, например, ковариационная и корреляционная матрицы, которые используются во всех методах многомерной прикладной статистики. Матрицу будем обозначать заключённой в квадратные скобки, например [**D**] или [**D**(d_{ij})], где d_{ij} – элементы матрицы; i – номер строки; j – номер столбца.

Столбец или строку матрицы можно рассматривать как *вектор*. Будем для его обозначения также использовать квадратные скобки, например, [**X***j*] – j-й столбец матрицы [**X**]. Будем употреблять также полную форму записи для вектора, например, вектор-строка $[\mathbf{X}_i] = \{x_{i1}, x_{i2}, \ldots x_{im}\}$, где x_{i1}, x_{i2}, … x_{im} – компоненты (элементы, координаты) вектора. Векторы называют иногда точками. Например, вектор $[\mathbf{X}_i]$ можно геометрически представить как точку с координатами x_{1j}, x_{2j}, … x_{nj} в ***n***-мерном пространстве. Это свойство векторов используется в факторном, дискриминантном и кластерном анализах.

Собственные значения и собственные векторы. Исключительный практический интерес (особенно для факторного анализа) представляют собственные векторы и собственные значения квадратных матриц. *Собственным значением* (собственным числом, характеристическим числом) называется корень λ_i характеристического уравнения квадратной матрицы, которое, например, для матрицы S будет иметь вид:

$$\begin{bmatrix} s_{11}-\lambda & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22}-\lambda & \cdots & s_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ s_{m1} & s_{m2} & \cdots & s_{mm}-\lambda \end{bmatrix} = 0$$

Корней (собственных значений) у уравнения, которое получится, если вычислить определитель, будет столько, чему равен ранг матрицы S. Не вдаваясь в математические подробности и доказательства, приведём лишь способ использования теории собственных чисел и собственных векторов в методах многомерного прикладного анализа. Для каждого отдельного собственного числа λ_k находят *собственный вектор*, решая систему ***m*** однородных линейных уравнений вида:

$$(s_{11}-\lambda_k)x_1^k+s_{12}x_2^k+\cdots+s_{1m}x_m^k=0,$$
$$s_{21}x_1^k+(s_{12}-\lambda_k)x_2^k+\cdots+s_{1m}x_m^k=0,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$s_{m1}x_1^k+s_{m2}x_2^k+\cdots+(s_{mm}-\lambda_k)x_m^k=0,$$

где $x_1^k, x_2^k, \cdots x_m^k$ – решение системы и одновременно элементы собственного вектора $[\mathbf{X}_k]$. Собственные векторы обладают одним замечательным свойством, а именно, – собственные векторы, принадлежащие различным собственным значениям, линейно независимы (ортогональны). Это свойство используется в факторном анализе. Если [S] – корреляционная матрица, то собственные векторы этой матрицы являются искомыми факторами.

В электронных таблицах каждый элемент матрицы занимает соответствующую ячейку, располагающуюся на пересечении строки и столбца. Представление этого элемента (его формат) в принципе не зависит от представления другого элемента данных.

В *базах данных* выборочная совокупность может быть представлена как атрибутивная таблица с заданными свойствами соответствующих атрибутивных полей. Аналогичным образом многомерные выборки организованы в таблицах прикладных статистических программ типа Statistica.

С точки зрения программного кода многомерная выборка всегда представлена как массив индексированных переменных: ИМЯ (число строк, число столбцов).

Основные типы представления многомерных данных. Все статистические пакеты обрабатывают информацию, представленную в виде матрицы одного из трёх типов: матрицы данных, матрицы мер близостей (расстояний), матрицы связей (между переменными). Каждая из этих матриц может являться входной матрицей, либо возникать как результат работы программ обработки. Такое деление данных носит в первую очередь содержательный характер, файлы для хранения и ввода этих данных имеют одинаковую логическую организацию, отличаясь лишь содержимым.

Матрица данных. Эта матрица, задающая отношение «*объект – признак*», является одной из наиболее частых форм организации данных и представляет собой двумерную таблицу размера $n \times m$

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix},$$

где каждая строка – новый объект, значения в строке – характеристики свойств объекта (признаки) или, по математической терминологии, вектор-строка $[\mathbf{X}_i] = \{x_{i1}, x_{i2}, \ldots x_{im}\}$ является совокупностью значений m признаков, измеренных у i-го объекта. В дальнейшем наряду с термином *признак* будем употреблять также термин *переменная*. Например, в пакете «Statistica» *переменная* – это столбец с данными.

Матрица связи между признаками. Эта матрица задаёт отношение «*признак – признак*» и представляет собой двумерную *симметричную квадратную* матрицу размера $m \times m$:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ S_{m1} & S_{m2} & \cdots & S_{mm} \end{bmatrix},$$

где S_{ij} может быть либо коэффициентом корреляции (ковариации), между i-м и j-м признаками, либо какой-либо другой мерой связи между признаками. Наиболее часто в качестве матрицы связи «признак – признак» используется ковариационная матрица, которая наряду с векторами средних значений признаков является основным промежуточным звеном для методов сокращения размерности, регрессии, дискриминантного анализа, основанных на линейных моделях. Диагональный элемент этой матрицы представляет собой оценку дисперсии признака x_j, которая характеризует степень рассеивания значений этого признака относительно среднего.

Матрица близостей (удаленностей расстояний). Эта матрица, задающая отношение «*объект – объект*» представляет собой квадратную симметричную матрицу размера $n \times n$ с *неотрицательными* элементами:

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}.$$

Элемент d_{ij} является значением некоторой меры близости (удалённости) между объектами x_i и x_j. Диагональные элементы этой матрицы значения не имеют, так как в последующей обработке матрицы не используются. Если элемент матрицы вычисляется как расстояние между двумя

объектами, то матрица [D] отражает геометрическую конфигурацию точек (объектов) в m-мерном пространстве признаков. Матрица расстояний (близостей) между объектами, как правило, вычисляется по матрице [X] и применяется в процедурах кластер-анализа или метрического шкалирования.

1.4. Математическое ожидание и дисперсия многомерной случайной величины

Многомерная случайная величина. Так как строки и столбцы матрицы данных можно представить как соответствующие векторы, то многомерная случайная величина может быть названа векторной. Двумерная случайная величина в простейшем варианте может быть изображена с помощью бивариантной гистограммы, имитирующей функцию распределения вероятностей (функцию плотности). Гистограмму можно легко построить, используя графические процедуры пакета Statistica: меню *Graphs* → *3D Sequential Graphs* → *Bivariate Histograms*. Представить с помощью обычных средств функцию распределения многомерной векторной случайной величины затруднительно, однако можно оценить функцию распределения каждой из m одномерных случайных величин.

Математическим ожиданием многомерной случайной величины [X], состоящей из m одномерных векторов $[X_1]$, $[X_2]$, ... $[X_m]$ является вектор $[M(X)] = \{M(X_1), M(X_2), \ldots M(X_m)\}$. Оценкой математического ожидания многомерной случайной величины является *выборочный вектор средних* $\bar{x} = \{\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m\}$, компонентами которого являются выборочные средние значения всех атрибутов.

Матрицу данных путём вычитания вектора средних можно преобразовать в матрицу вариаций [Y] размера $n \times m$ с элементами $y_{ij} = x_{ij} - \bar{x}_j$.

Изменчивость векторной случайной величины характеризуется дисперсионной матрицей, называемой также матрицей ковариаций или вариационно-ковариационной матрицей. Это квадратная матрица размером $m \times m$. Матрицу дисперсий и ковариаций можно представить как квадратную симметричную матрицу [V]. По диагонали этой матрицы располагаются суммы квадратов отклонений значений всех признаков от своих средних: $v_{ii} = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2$. Недиагональные элементы матри-

цы [V] представлены суммами смешанных произведений: $v_{ij} = \sum_{\kappa=1}^{n}(x_{\kappa j} - \bar{x}_j)(x_{\kappa i} - \bar{x}_i)$. При делении всех элементов матрицы на объём выборки получаем матрицу дисперсий и ковариаций [D]. Если каждый элемент матрицы [D] разделить на корень квадратный из произведения соответствующих дисперсий, то получим корреляционную матрицу [R] с элементами, называемыми парными коэффициентами корреляции и оценивающими линейные зависимости между свойствами объекта

$$r_{ij} = \frac{d_{ij}}{\sqrt{d_{ii} d_{jj}}}.$$

Рассмотрим процедуры, связанные со сравнением многомерных выборочных совокупностей, которое используется в дальнейшем в дискриминантном анализе.

Сравнение средних двух многомерных выборочных совокупностей. Пусть имеется две случайные векторные величины $\mathbf{X^{(1)}}$ и $\mathbf{X^{(2)}}$, охарактеризованные выборками объёма n_1 и n_2. Каждая случайная величина характеризует объект, описанный k признаками. Многомерные величины $X_t^{(1)}\{X_{1t}^{(1)}, X_{2t}^{(1)}, \ldots X_{kt}^{(1)}\}, (t = 1 \ldots n_1)$ и $X_t^{(2)}\{X_{1t}^{(2)}, X_{2t}^{(2)}, \ldots X_{kt}^{(2)}\}, (t = 1 \ldots n_2)$ нормально распределены с общей ковариационной матрицей и векторами средних $\bar{X}^{(1)}$ и $\bar{X}^{(2)}$.

Проверим гипотезу о равенстве математических ожиданий случайных величин H_0: $M\mathbf{X^{(1)}} = M\mathbf{X^{(2)}}$. Воспользуемся критерием Хоттелинга (T^2), являющимся многомерным аналогом критерия Стьюдента. Статистика T^2 вычисляется по достаточно громоздкой формуле:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}^{(1)} - \bar{x}^{(2)})'[D]^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}), \qquad (1)$$

где $[\mathbf{D}]^{-1}$ – матрица, обратная обобщённой эмпирической ковариационной матрице системы, умноженная слева на вектор строку, а справа на вектор столбец разности векторов средних двух выборок.

Обобщённая ковариационная матрица может быть получена путём объединения выборок. Таблицы распределения Хоттелинга не всегда доступны, поэтому рассчитаем F-критерий, связанный со статистикой T^2, имеющей F-распределение:

$$F = \frac{T^2(n_1 + n_2 - k - 1)}{(n_1 + n_2 - 2)k}. \qquad (2)$$

Если $F_{рассч} \geq F_{крит}$ для уровня значимости α и степеней свободы k и $(n_1 + n_2 - k - 1)$, то нулевая гипотеза о равенстве векторов средних отвергается. Основное допущение, на котором основан рассмотренный критерий, заключается в том, что выборки взяты из нормально распределённых совокупностей, имеющих одну и ту же или одинаковые ковариационные матрицы. Предположение о нормальности распределения и равенстве ковариационных матриц, как и при сравнении одномерных средних и дисперсий, в реальности часто нарушается.

Сравнение дисперсий двух многомерных выборочных совокупностей. Сравнение ковариационных матриц возможно с помощью *критерия обобщённых дисперсий,* являющегося многомерным аналогом F-критерия. Пусть имеется две группы наблюдений объёмом n_1 и n_2. Найдём для них ковариационные матрицы $[\mathbf{D}_1]$ и $|\mathbf{D}_2]$. Сформулируем нулевую гипотезу H_0: $|\mathbf{D}_1| = |\mathbf{D}_2|$ при альтернативе H_1: $|\mathbf{D}_1| \neq |\mathbf{D}_2|$. Объединим выборки и получим обобщённую оценку ковариационной матрицы **[D]**, предполагаемую общей для исследуемых генеральных совокупностей. Далее вычислим статистику M:

$$M = (n_1 + n_2 - 2)\ln|D| - \frac{1}{2}[(n_1 - 1)\ln|D_1| + (n_2 - 1)\ln|D_2|], \qquad (3)$$

представляющую собой разность между логарифмом определителя обобщённой ковариационной матрицы и средним значением логарифмов определителей выборочных ковариационных матриц. Критическое значение статистики M апроксимируется χ-квадрат распределением с числом степеней свободы равным $0{,}5p(p + 1)$, где p – мерность случайной величины:

$$\chi^2 = MC^{-1},$$

$$C^{-1} = 1 - \frac{2(2p^2 + 3p - 1)}{6(p + 1)}\left(\frac{1}{(n_1 + n_2 - 1)} - \frac{1}{(n_1 + n_2 - 2)}\right), \qquad (4)$$

χ^2 – апроксимация будет более точной, если количество наблюдений в каждой из выборок превышает 20.

Если полученное значение M превосходит критическое, то нулевая гипотеза о равенстве ковариационных матриц должна быть отвергнута в пользу альтернативной.

Пример. *Вычисление ковариационной и корреляционной матриц в Excel. Сравнение средних двух многомерных выборочных совокупностей.* Для иллюстрации рассмотренных понятий и процедур используем выборки результатов химического анализа, представляющие собой пятимерные векторные случайные величины (табл. 1–2). Общий объём двух выборок $N = 37$, $n_1 = 19$, $n_2 = 18$ при $p = 5$. Объединённая выборка представляет собой матрицу размерностью 37×5, а отдельные выборки, соответственно 19×5 и 18×5.

Ковариационные матрицы выборок найдём с помощью соответствующей процедуры пакета анализа электронного процессора Excel. Рассчитанные векторы средних представлены в табл. 1–5.

Таблица 1

Выборка № 1 результатов сокращённого химического анализа

	SiO_2	Al_2O_3	FeO	MgO	CaO
1	59,92	0	0,38	39,51	0,32
2	57,73	0,95	3,57	36,13	0,23
3	57,1	0,7	5,21	34,52	0,62
4	57,7	1,87	6,47	32,72	0,95
5	55,94	1,61	7,15	32,12	1,48
6	55,02	2,69	7,18	32,13	0,48
7	55,2	1,5	11,86	28,14	1,93
8	54,11	1,52	15,73	27,03	1,16
9	52,6	0,12	17,14	25,65	2,07
...	...	...	...	...	...
12	52,00	0,57	22,46	21,62	0,75
13	52,07	1,70	22,65	21,13	1,55
14	50,6	0,16	25,71	18,96	1,65
15	52,22	0,43	25,91	18,54	1,28
16	50,26	3,13	26,54	16,36	1,76
17	50,06	1,84	29,39	13,63	1,43
18	49,43	0,38	34,91	12,96	0,71
19	44,52	4,74	38,66	6,59	1,40
Среднее	53,30	1,44	17,69	24,46	1,29
Стандарт	3,61	1,25	10,98	8,79	0,66

Таблица 2

Выборка № 2 результатов сокращённого химического анализа

	SiO_2	Al_2O_3	FeO	MgO	CaO
1	53,95	4,24	8,4	30,17	0,57
2	55,12	1,25	11,96	29,11	0,4
3	53,37	1,86	12,96	27,33	1,85
4	53,17	2,46	13,06	24,74	4,19
5	53,2	1,15	21,64	22,50	0,82
6	46,91	8,26	19,88	20,02	0,34
7	50,34	3,14	22,53	19,52	0,6
8	49,44	2,21	28,06	16,61	0,23
9	50,08	1,23	27,85	15,78	1,44
...	...	...	...	...	...
12	48,29	2,83	33,67	10,77	2,29
13	48,21	1,37	36,9	10,45	0,43
14	47,55	1,9	39,37	8,68	1,23
15	48,7	1,04	42,35	6,88	0,85
16	46,36	0,29	44,93	5,09	1,64
17	46,36	0,23	48,10	3,7	0,71
18	45,95	0,9	41,65	3,49	1,43
Среднее	49,66	2,15	28,88	15,53	1,15
Стандарт	2,91	1,82	12,35	8,72	0,97

Таблица 3

Ковариационная матрица |D| объединённой выборки

	SiO_2	Al_2O_3	FeO	MgO	CaO
SiO_2	13,58	–1,54	–43,08	33,77	–0,20
Al_2O_3	–1,54	2,43	–1,01	–0,58	–0,07
FeO	–43,08	–1,01	160,09	–120,53	0,26
MgO	33,77	–0,58	–120,53	92,51	–0,70
CaO	–0,20	–0,07	0,26	–0,70	0,65
Вектор средних объединённой выборки					
$\|\bar{x}\|$	53,30	1,44	17,69	24,46	1,29

Таблица 4

Ковариационная матрица $|D_1|$ выборки № 1

	SiO_2	Al_2O_3	FeO	MgO	CaO
1	2	3	4	5	6
SiO_2	12,39	–1,83	–36,34	29,36	–0,89
Al_2O_3	–1,83	1,50	3,77	–3,84	0,11
FeO	–36,34	3,77	114,28	–90,84	2,40

Окончание табл 4

1	2	3	4	5	6
MgO	29,36	–3,84	–90,84	73,26	–2,21
CaO	–0,89	0,11	2,40	–2,21	0,41
Вектор средних					
$[\bar{x}_1]$	53,30	1,44	17,69	24,46	1,29

Таблица 5

Ковариационная матрица $|D_2|$ выборки № 2

	SiO_2	Al_2O_3	FeO	MgO	CaO
SiO_2	8,03	0,09	–29,25	21,73	0,27
Al_2O_3	0,09	3,15	–10,17	6,11	–0,21
FeO	–29,25	–10,17	144,11	–100,56	–1,21
MgO	21,73	6,11	–100,56	71,94	0,26
	0,27	–0,21	–1,21	0,26	0,89
Вектор средних					
$[\bar{x}_2]$	49,66	2,15	28,88	15,53	1,15

Предполагая принадлежность выборок к одной генеральной совокупности, сравним их ковариационные матрицы и векторы средних. Сформулируем нулевую гипотезу о равенстве ковариационных матриц H_0: $[\mathbf{D}_1] = [\mathbf{D}_2]$ при альтернативе H_1: $[\mathbf{D}_1] \neq [\mathbf{D}_2]$. Найдём значения определителей ковариационных матриц, используя функцию МОПР Excel:

$$|\mathbf{D}| = 75,53;\ |\mathbf{D}_1| = 9,51;\ [\mathbf{D}_2| = 68,05.$$

Далее по формуле (3) рассчитаем статистику M:

$$M = (18+19-2)\ln(75,53) - \frac{1}{2}[(18-1)\ln(9,51) + (19-1)\ln(68,06)] = 94,23.$$

Согласно выражениям (4) рассчитаем значения C^{-1} и χ^2:

$$C^{-1} = 1 - \frac{2\times(2\times 25^2 + 3\times 5 - 1)}{6\times(5+1)}\left(\frac{1}{(18+19-1)} - \frac{1}{(18+19-2)}\right) = 0,94,$$

$$\chi^2 = MC^{-1} = 94,23\times 0,944268 = 88,98.$$

Критическое значение χ-квадрат для уровня значимости 0,05 и степеней свободы 5 и 15 примерно равно 25. Значение M намного превосходит критическое и, следовательно, нулевая гипотеза о равенстве ковариационных матриц должна быть отвергнута в пользу альтернативной. Таким образом, рассматриваемые выборки существенно отличаются друг от друга, представляя собой различные генеральные совокупности.

Установив различие генеральных совокупностей продолжим, тем не менее, процедуру сравнения и сопоставим векторы средних. Найдём матрицу, обратную объединённой, и по формуле (1) рассчитаем статистику Хоттелинга: $T^2 = (18\times19\times[L]'\times[D]^{-1}\times[L])/(18 + 19) = 10,3$, где $[L]'$ и $[L]$ – вектор-строка и вектор-столбец разности векторов средних. По формуле (2) рассчитаем статистику $F{:}F = (10,3\times31)/(35\times5) = 1,82$. Значение $F_{крит} = 25$ для уровня значимости 0,05 и степеней свободы 5 и 31. При равенстве ковариационных матриц для такой ситуации мы не смогли бы отвергнуть нулевую гипотезу о равенстве средних векторов.

Вопросы для самопроверки

1. Являются ли синонимами понятия: свойство объекта, признак, переменная?
2. Совпадает ли порядок квадратной матрицы с её рангом?
3. Как рассчитывается и что характеризует ковариационная матрица?
4. Какие величины стоят на главной диагонали ковариационной матрицы?
5. Что понимают под корреляционной матрицей?
6. Что характеризует коэффициент парной корреляции?
7. В каких пределах изменяется коэффициент парной корреляции?
8. Как связаны между собой корреляционная и ковариационная матрицы?
9. Какие критерии сравнения параметров многомерных случайных величин Вам известны?
10. Поясните термин «собственное значение» квадратной матрицы.
11. Сколько различных собственных значений может иметь матрица?
12. Дайте определение собственного вектора матрицы?
13. Сколько собственных векторов может быть у ковариационной или корреляционной матрицы?
14. В каком методе прикладного статистического анализа используются собственные векторы и собственные значения корреляционной матрицы? Выберите правильный ответ:

а) дискриминантный анализ;

б) кластерный анализ;

в) корреляционный анализ;

г) факторный анализ.

15. В качестве оценки математического ожидания многомерной случайной величины принимают:

a) собственный вектор;

b) вектор средних;

c) среднее значение.

16. Что понимается под вектором средних?

17. Какой критерий используется для сравнения средних двух многомерных выборочных совокупностей?

18. Какие статистика и критерий используются для сравнения дисперсий двух многомерных выборочны совокупностей?

19. В каком методе прикладного статистического анализа используется критерий Хоттелинга сравнения средних двух многомерных выборочных совокупностей? Выберите правильный ответ:

а) дискриминантный анализ;

б) кластерный анализ;

в) корреляционный анализ;

г) факторный анализ.

20. В каком методе прикладного статистического анализа используется матрица близостей (расстояний) типа «объект – объект»?

2. АНАЛИЗ ЗАВИСИМОСТЕЙ В МНОГОМЕРНЫХ ДАННЫХ. РЕГРЕССИОННЫЙ АНАЛИЗ

Множественный корреляционный и регрессионный анализы относятся к числу немногих количественных методов, которые могут быть использованы для исследования взаимосвязей природных процессов. *Основная задача корреляционного анализа состоит в оценке степени тесноты линейной связи между переменными*, т.е. в расчёте корреляционной матрицы по выборкам и определении частных и множественных коэффициентов корреляции и детерминации.

Основное назначение регрессионного анализа заключается в установлении вида стохастических зависимостей между переменными. Он устанавливает форму зависимости между одной переменной (**Y**), рассматриваемой в качестве *зависимой*, и значениями одной или нескольких переменных величин из этого же набора данных, рассматриваемых как *независимые* ($\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_n$) и принимающих некоторые заданные значения. Зависимую переменную называют также *предиктантом*, а независимую – *предиктором*.

Полученное уравнение можно использовать для оценки одновременного влияния нескольких факторов на данный процесс с целью его прогнозов и расчётов. Кроме того, этот метод позволяет определять относительное влияние на прогноз каждого фактора и измерять полный эффект с помощью коэффициентов. Можно также оценить значимость связи между зависимой и каждой независимой переменной и получить «лучшее» расчётное уравнение.

2.1. Определение параметров уравнения регрессии

В простейшем случае двух переменных **X** и **Y** связь между случайными величинами в линейном регрессионном анализе представляется уравнением прямой, принимающей вид $\mathbf{Y} = a\mathbf{X} + b$. Все переменные, входящие в уравнение регрессии, должны быть случайными непрерывными величинами.

В случае, когда рассматривается зависимость **Y** от нескольких независимых переменных $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots \mathbf{X}_m$, говорят о множественной линейной регрессии. В качестве исходных данных будем использовать матрицу

наблюдений **X**, содержащую n строк (число наблюдений) и $m + 1$ столбцов ($\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots \mathbf{X}_m$), т.е. значения каждой переменной представлены в виде вектора-столбца. *Модель множественной линейной регрессии* – это уравнение вида

$$\mathbf{Y} = b_0 + b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \ldots + b_m\mathbf{X}_m, \qquad (5)$$

где **Y** – зависимая переменная (предиктант, отклик); b_j – коэффициенты уравнения; $\mathbf{X}_j$ – независимые переменные (предикторы, факторы), $j = 1, 2, \ldots m$; m – количество предикторов (слово «предиктор» произошло от *англ.* predict – предсказывать).

Процедуры множественной регрессии будут оценивать (вычислять) *параметры* уравнения, то есть коэффициенты $b_0, b_1, b_2, \ldots b_m$. Величины $b_0, b_1, b_2, \ldots b_m$ называются также регрессионными коэффициентами, где b_0 – свободный член уравнения.

Заметим, что модель (5) записана в векторном виде и может быть расписана для любого i-го элемента векторов **Y**, $\mathbf{X}_1, \ldots \mathbf{X}_m$, где $i = 1, 2, \ldots n$; n – число наблюдений.

Обозначим символом y_i значения зависимой случайной величины (элементы вектора **Y**), полученные в результате эксперимента, а y'_i – значения зависимой случайной величины, рассчитанные по уравнению регрессии. Для оценки коэффициентов уравнения регрессии чаще всего исходят из принципа наименьших квадратов, основанном на том, что при нормальном законе распределения оценка коэффициентов уравнения b_j считается наилучшей, если средний квадрат разности между фактическими и рассчитанными значениями является наименьшим, т.е. из условия минимизации суммы квадратов отклонений расчётных значений от экспериментальных:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1x + b_2x_{i2} + b_mx_{im}))^2 = \min_{b_0, b_1 \cdots b_m}.$$

Далее частные производные S^2 по неизвестным коэффициентам приравнивают к нулю. В результате получают систему линейных алгебраических уравнений (6) относительно неизвестных $b_0, b_1, b_2, \ldots b_m$. На практике все наблюдения заменяют их отклонениями от среднего значения $\Delta y_i = y_i - \overline{y}, \Delta x_{ij} = x_{ij} - \overline{x_j}$. Это преобразование позволяет уменьшить абсолютную величину переменных и приводит к переменным, имеющим об-

щее среднее значение, равное нулю. При этом преобразовании коэффициент b_0 обращается в нуль, так что порядок матрицы системы уравнений снижается на единицу.

$$\begin{aligned}
&\frac{\partial S^2}{\partial b_0} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im})) = 0;\\
&\frac{\partial S^2}{\partial b_1} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}))x_{i1} = 0;\\
&\frac{\partial S^2}{\partial b_2} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}))x_{i2} = 0; \qquad (6)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\frac{\partial S^2}{\partial b_m} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}))x_{im} = 0.
\end{aligned}$$

В отклонениях система уравнений (6) для ***k***-й строки будет иметь вид:

$$\frac{2}{n}\sum_{i=1}^{n}(\Delta y_i - \sum_{j=1}^{m} b_j \Delta x_{ij})(-\Delta x_{ik}) = 0.$$

Эту систему уравнений можно уже выразить через коэффициенты корреляции, используя формулу для вычисления коэффициента корреляции между независимыми переменными (r_{kj}) и между зависимой и каждой из независимых (r_{0j}):

$$r_{kj} = \frac{\sum_{i=1}^{n}\left(x_{ik} - \overline{x_k}\right)\left(x_{ij} - \overline{x_j}\right)}{n\sigma_k\sigma_y} = \frac{\sum_{i=1}^{n}\Delta x_{ik}\Delta x_{ij}}{n\sigma_k\sigma_j},$$

$$r_{0j} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)\left(y_i - \overline{y}\right)}{n\sigma_j\sigma_y} = \frac{\sum_{i=1}^{n}\Delta x_{ij}\Delta y_i}{n\sigma_j\sigma_y}.$$

где σ_j, σ_k – средние квадратические отклонения j-й и k-й независимых переменных; σ_0 – то же для зависимой переменной.

В новой системе уравнений с целью упрощения вводятся новые обозначения для коэффициентов при независимых переменных $\beta_j = (\sigma_j/\sigma_y)b_j$ – стандартизованные коэффициенты уравнения регрессии. Учитывая, что $r_{jj} = 1$, система уравнений приводится к нормальному виду:

$$r_{01} = \beta_1 + \beta_2 r_{12} + \ldots + \beta_m r_{1m};$$

$$r_{02} = \beta_1 r_{21} + \beta_2 + \ldots + \beta_m r_{2m};$$

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

$$r_{0m} = \beta_1 r_{m1} + \beta_2 r_{m2} + \ldots + \beta_m.$$

Решение системы уравнений относительно β_j может быть получено разными путями. В частности, если предикторов немного (2–3), то можно применить точные методы (Крамера, Гаусса), и тогда решение сводится к вычислению определителей. Если независимых переменных в уравнении много (Excel, например, допускает до 16), то используют приближённые методы, например, последовательных приближений (итераций) Зейделя.

Бета-коэффициенты имеют и самостоятельное значение при анализе модели. Они показывают, на сколько средних квадратических отклонений (σ_y) изменится величина y при увеличении x_j на одно среднее квадратическое отклонение (σ_j) при неизменности остальных независимых переменных, входящих в уравнение регрессии. Таким образом, сравнение бета-коэффициентов позволяет судить о большей или меньшей степени влияния соответствующих независимых переменных на величину y.

2.2. Оценка качества уравнения регрессии

Значимость уравнения регрессии проверяется по критерию Фишера [20]. Выдвигается нулевая гипотеза о равенстве всех коэффициентов уравнения регрессии нулю (H_0: $b_0 = b_1 = \ldots = b_m = 0$). Рассчитывается статистика F по формуле

$$F_{\text{набл}} = \frac{R^2(n-m-1)}{(1-R^2)m},$$

где R^2 – коэффициент детерминации, характеризующий долю дисперсии **Y**, объяснённую переменными $\mathbf{X}_i$ ($i = 1, 2, \ldots m$):

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_{\text{рассч},i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}.$$

По таблице F-распределения для заданных уровня значимости α и степеней свободы $\nu_1 = m$, $\nu_2 = n - m - 1$ находят критическое значение статистики $F_{кр}$. Гипотеза H_0 отклоняется на уровне значимости α, если $F_{набл} > F_{кр}$. В этом случае уравнение является значимым, т.е. хотя бы один из коэффициентов уравнения регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов, т.е. гипотезы H_0: $b_j = 0$, где $j = 1, 2, \ldots m$ используют t-критерий. Гипотеза H_0 отвергается на уровне значимости α, если $t_{набл} > t_{кр}$. При выполнении этого условия соответствующий коэффициент b_j значим. В противном случае коэффициент регрессии незначим, и соответствующая переменная в модель не включается. Тогда реализуется алгоритм пошагового регрессионного анализа, состоящий в том, что исключается одна из незначимых переменных, которой соответствует минимальное по абсолютной величине значение $t_{набл}$. Затем вновь проводят регрессионный анализ с числом предикторов, уменьшенным на единицу. Алгоритм заканчивается получением уравнения регрессии со значимыми коэффициентами.

Возможен и другой алгоритм пошагового регрессионного анализа – с последовательным включением предикторов. В пакете Statistica можно использовать на выбор любой из рассмотренных алгоритмов пошагового включения/исключения предикторов, а в электронных таблицах Excel их нет.

Величина стандартной ошибки уравнения регрессии или степень приближения рассчитанных величин к наблюдённым, находится по формуле

$$s = \sqrt{\frac{1}{(n-m)}\sum_{i=1}^{n}\left(y_i - y_{\text{рассч}}\right)^2}.$$

Здесь s – стандартная ошибка уравнения на зависимом материале, т.е. на данных, использовавшихся в расчётах коэффициентов уравнения; m – число степеней свободы, равное количеству параметров в уравнении рег-

рессии. Например, для уравнения $y = a_1x_1 + a_2x_2 + a_3$ число степеней свободы $m = 3$.

Для оценки качества модели (уравнения регрессии) используется отношение его стандартной ошибки (s) к среднему квадратическому отклонению (σ) ряда-предиктанта **Y**. Если число проверочных прогнозов (n) больше 25, методика считается удовлетворительной при $0,5 < s/\sigma \leq 0,8$ и хорошей при $s/\sigma \leq 0,5$. Если число членов ряда $n \leq 15$, то пороговое значение s/σ должно быть уменьшено на 0,1. При $15 < n \leq 25$ значение s/σ должно быть уменьшено на 0,05 (Бефани, Калинин, 1965).

Важную роль показателя качества модели или применимости данного набора предикторов для описания зависимой переменной играет величина R^2 – *коэффициент детерминации* (определенности), где R – множественный коэффициент корреляции. Коэффициент детерминации непосредственно интерпретируется следующим образом: например, если $R^2 = 0,4$, то только 40% исходной изменчивости могут быть объяснены предикторами $\mathbf{X}_j$, а 60% остаются необъясненными.

В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R^2 является индикатором степени подгонки модели к данным. Значение R^2, близкое к 1, показывает, что модель объясняет почти всю изменчивость соответствующей переменной. Нижняя допустимая граница для коэффициента детерминации зависит от конкретной области исследований, например, в гидрологии она составляет 0,5 (Определение основных расчётных гидрологических характеристик, 2004), что соответствует значению множественного коэффициента корреляции $R > 0,7$.

Перечисленные выше некоторые способы оценки качества регрессионного уравнения являются в большей степени предварительными и не заменяют основного метода оценки адекватности модели – *анализа остатков.*

2.3. Анализ остатков

Остатком называют разность наблюдённого значения и предсказанного по уравнению регрессии (рис. 2). Анализ остатков является одним из способов проверки качества модели или степени адекватности математической модели линейной регрессии.

Если остатки представляют собой временной ряд *случайных независимых величин, распределенных по нормальному закону*, это может слу-

жить обоснованием пригодности уравнения для прогноза. Достаточно информативным в этих целях является графическое представление зависимости остатков от x или y (Шелутко, 1991). На графике остатки должны вести себя достаточно хаотично, не должно быть резких выбросов, закономерностей в чередовании знаков.

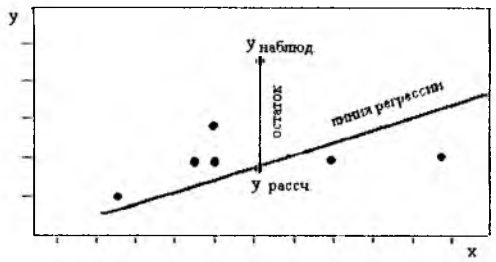

Рис. 2. Схема для иллюстрации парной линейной регрессии

Если остатки попадают в горизонтальную полосу с центром по оси абсцисс, то модель можно рассматривать как адекватную (рис. 3, *а*). Если полоса расширяется, то необходимо преобразование ряда **Y** (рис. 3, *б*). График, показывающий линейный тренд (рис. 3, *в*), даёт основание для введения в модель дополнительной переменной. График, представленный на рис. 3, *г*, показывает, что в модель должен быть добавлен квадратический член.

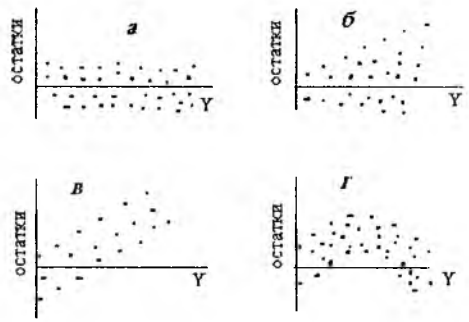

Рис. 3. Примеры графиков остатков (погрешностей расчёта): *а* – адекватная модель; *б* – необходимо преобразование ряда Y; *в* – не учтена линейная независимая переменная; *г* – не учтена квадратическая независимая переменная (Шелутко, 1991)

Для проверки независимости остатков можно использовать статистику Дарбина – Уотсона, являющуюся стандартным методом обнаружения их автокоррелированности, или автокорреляционную функцию остатков. Статистика Дарбина – Уотсона d используется для проверки гипотезы о том, что остатки построенной регрессионной модели не коррелированны (корреляции равны нулю), против альтернативы – остатки связаны автокорреляционной зависимостью.

Вычисленное значение статистики d надо сравнить с двумя критическими d: нижним DW1 и верхним DW2.

– Если $d < DW1$ и $4 - d < DW1$, то в остатках имеется автокорреляция на заданном уровне значимости.

– Если $d > DW2$ и $4 - d > DW2$, то автокорреляция отсутствует.

– Если $DW1 < d < DW2$, то случай сомнительный, нужны дополнительные исследования.

Критические точки (DW1, DW2) для данного числа наблюдений и числа предикторов находят в таблице, составленной для определенного уровня значимости α (см. Приложение). В пакете Statistica уровень значимости обозначается ***p***, а для проверки независимости остатков уравнения регрессии предлагается критерий Дарбина – Уотсона.

Соответствие ряда остатков нормальному закону распределения можно проверить по критериям Колмогорова – Смирнова, Лилиефорса, Шапиро – Уилкса. Если уровень значимости, вычисленный для ряда остатков, меньше критического, то гипотеза о соответствии ряда остатков нормальному закону распределения отвергается. Для визуальной оценки в этих целях можно использовать также гистограмму остатков, нормальный вероятностный график.

Гистограмма остатков помимо визуальной оценки нормальности распределения содержит также информацию о наличии грубых ошибок, проявляющихся на «хвостах» гистограммы, когда значительная доля остатков сосредотачивается в крайних интервалах.

2.4. Требования к исходным данным

Для получения удовлетворительных результатов при использовании модели множественной регрессии необходимо выполнение ряда требований к исходной информации, соблюдение которых зачастую вообще не проверяется, в то время как во многих случаях они не выполняются или выполняются не полностью (Шелутко, 1991).

Основные требования к рядам наблюдений, следующие из сути математической модели, заключаются в следующем:

1. Связи между всеми рядами должны быть линейными. Если нелинейность связи очевидна, то можно рассмотреть или преобразование переменных, или явно допустить включение нелинейных членов.

2. Исследуемые ряды должны подчиняться нормальному закону распределения. *Близость законов распределения выборок к нормальному является одним из главных показателей надёжности математических моделей, основанных на принципе метода наименьших квадратов.*

3. Корреляция между рядами-предикторами должна отсутствовать или быть незначительной. При наличии тесной связи между предикторами корреляционная матрица становится вырождающейся, её детерминант стремится к нулю, и возникают трудности в вычислении коэффициентов уравнения регрессии. Они становятся неустойчивыми. В этом случае надо *исключать дублирующие предикторы.*

4. Ряд-предиктант должен представлять собой выборку значений случайной величины, т.е. его значения должны быть некоррелирваны между собой. В применении ко многим рядам наблюдений за природными явлениями это требование не выполняется, так как для них характерно наличие внутрирядной связности.

5. Объём выборки должен в несколько раз превосходить число независимых переменных (рекомендуется в 2–3 раза).

Практика показывает, что при использовании одного предиктора длина рядов n должна быть не менее 10, при двух предикторах минимальная длина рядов должна составлять не менее 25–30, при четырёх – 50–60, при пяти – 100–120 и т.д. (Дружинин, Сикан, 2001). Только в этом случае можно получить более или менее надёжные оценки параметров уравнения регрессии.

В гидрометеорологии уравнение регрессии может использоваться для практических расчетов при выполнении условий $R^2 \geq 0{,}5$, $R/\sigma_r \geq 2$, $b_i/\sigma_i \geq \geq 2$, где σ_r – стандартная ошибка вычисления множественного коэффициента корреляции; σ_i – стандартная ошибка вычисления коэффициента уравнения (Дружинин, Сикан, 2001).

2.5. Использование множественной линейной регрессии при решении задач прикладных исследований

Пример 1. *Подбор рек-аналогов и восстановление расходов воды по уравнению множественной регрессии с использованием пакета «Statis-*

tica». Важной проблемой в гидрометеорологии является наличие пропусков в наблюдениях и необходимость восстановления данных. В этих целях широко используется метод множественной регрессии.

Рассмотрим задачу, целью которой является восстановление недостающих данных по уравнению множественной линейной регрессии подбором рек-аналогов.

Имеется восемь временных рядов со среднегодовыми расходами воды рек, водосборные бассейны которых расположены на однородной по физико-географическим условиям территории. Период наблюдений составляет 24 года. В одном из рядов имеются пропуски. Сформируем таблицу исходных данных в виде матрицы типа «объект – признак» размером 24×8. В гидрологии принято наблюдения по конкретному посту (объекту) размещать в столбце.

Для получения удовлетворительных результатов при использовании модели множественной регрессии необходимо выполнение ряда требований к исходной информации. Поэтому на первом шаге исследований необходимо проверить однородность и репрезентативность рядов наблюдений, а также линейность и тесноту связей между ними. После анализа корреляционной матрицы, которая легко строится в пакете Statistica, и удаления на её основе дублирующих рядов остались две независимые переменные, удовлетворяющие требованиям, предъявляемым к связям как предикторов друг с другом, так и их связям с предиктантом (табл. 6).

Таблица 6

Корреляционная матрица связей предикторов (Q_2, Q_3) и предиктанта (Q_1)

Data: Spreadsheet3* (3v by 3c)

	Коррел		
	1 Q2, м³/с	2 Q3, м³/с	3 Q1, м³/с
Q2, м³/с	1,00	-0,08	0,76
Q3, м³/с	-0,08	1,00	0,84
Q1, м³/с	0,76	0,84	1,00

Если уравнение множественной регрессии используется для восстановления пропущенных данных, то коэффициент корреляции между предиктором и предиктантом должен быть не менее 0,7–0,8 (Шелутко, 1991).

Независимые переменные Q_2 и Q_3 не коррелированны ($r = -0,08$) и характеризуются тесной связью с зависимой переменной Q_1 ($r = 0,76$ и $0,84$), поэтому с точки зрения математических требований их можно использовать в качестве рек-аналогов. Гидрологические требования к рекам-аналогам изложены в (Определение основных гидрологических характеристик... 2004).

Практика показывает, что при использовании двух предикторов минимальная длина рядов должна составлять не менее 25–30 (Дружинин, Сикан, 2001). Только в этом случае можно получить более или менее надёжные оценки параметров уравнения регрессии. Объём выборки в нашем случае ($n = 24$) намного превосходит число независимых переменных ($m = 2$), и можно рассчитывать на получение удовлетворительных результатов.

На диалоговом окне результатов (рис. 4) модуля *Множественная регрессия (Multiple Regression)* в первую очередь стоит обратить внимание на значение скорректированного коэффициента детерминации (adjusted $R^2 = 0,643$). Это высокое значение, которое означает, что 64,3% дисперсии предиктанта объясняют выбранные нами предикторы Q_1 и Q_2. Здесь же в информационной области окна приводятся результаты проверки значимости полученного уравнения регрессии по критерию Фишера: статистика $F = 24,4$ и уровень значимости $p = 0,000002$. Так как p много меньше критического уровня значимости равного 0,05, то гипотеза о равенстве нулю коэффициентов уравнения отвергается, т.е. уравнение значимо.

Ниже располагается информация о проверке значимости свободного члена (*Intercept*) по критерию Стьюдента.

Важную информацию несут *beta*-коэффициенты. Это стандартизованные значения коэффициентов уравнения регрессии. Преимущество *beta*-коэффициентов в том, что они позволяют сравнивать относительный вклад каждой независимой переменной в прогнозе. Их интерпретация подобна анализу частных коэффициентов корреляции. Например, согласно информации в окне результатов, предиктор Q_2 вносит в прогноз величины **Y** намного больший вклад, чем Q_3. Значимые *beta*-коэффициенты выделяются красным цветом. Предикторы, имеющие незначимые *beta*-коэффициенты, из уравнения регрессии удаляются как неинформативные. В нашем случае оба *beta*-коэффициенты значимы.

Для оценки адекватности модели осталось проверить ряд остатков на независимость и соответствие нормальному закону распределения. В модуле для проверки независимости остатков используется статистика Дарбина – Уотсона, являющаяся стандартным методом обнаружения их ав-

токоррелированности. Статистика Дарбина – Уотсона d используется для проверки гипотезы о том, что остатки построенной регрессионной модели некоррелированы (корреляции равны нулю), против альтернативы: остатки связаны авторегрессионной зависимостью.

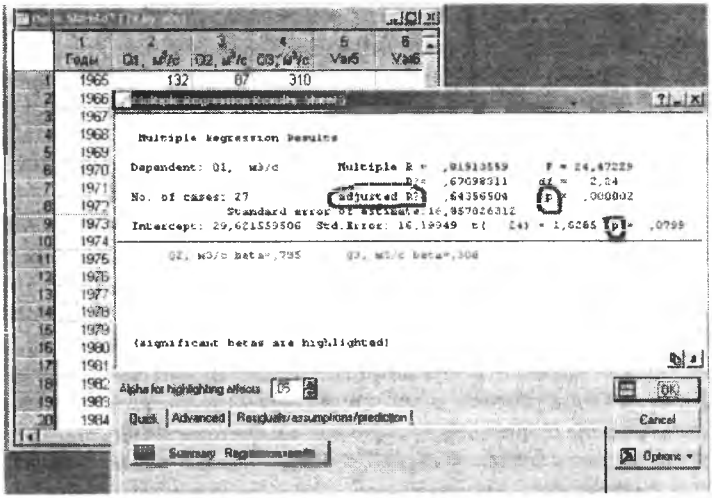

Рис. 4. Окно результатов уравнения множественной регрессии

Сравнение значений $d = 2{,}35$ (рис. 5) с DW1 и DW2 из таблицы критических точек статистики Дарбина – Уотсона при уровне значимости $p = 0{,}05$ позволяет сделать вывод об отсутствии значимых внутрирядных связей в ряде остатков.

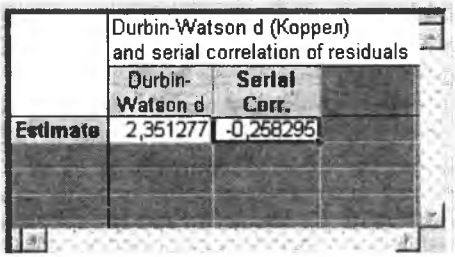

Рис. 5. Значения d-статистики критерия Дарбина – Уотсона

Стандартным способом проверяем соответствие распределения остатков нормальному закону: визуально по гистограмме, нормальному вероятностному графику и по теоретическим критериям Колмогорова – Смирнова, Лилиефорса и Шапиро – Уилкса.

Для осуществления прогноза введём понятие обучающей и контрольной выборки. *Обучающая выборка* – это просто матрица исходных данных, на основе которой вычисляются коэффициенты уравнения регрессии; *контрольная выборка* – это совокупность наблюдений, которые не использовались для получения регрессионных коэффициентов.

Вычисление Y по значениям предикторов из обучающей выборки называется прогнозом на зависимом материале, а по данным из контрольной выборки – на независимом. Последний является единственным надёжным способом оценки экстраполяционных свойств полученного уравнения регрессии.

Расчёт значений Q_1 как восстанавливаемых, так и прогнозируемых осуществляется по уравнению множественной регрессии с помощью вкладки *Residuals/Assumptions/Prediction (Остатки/предположения/предсказание)* (рис. 6). Прогноз осуществляется в поле, обведенном кругом.

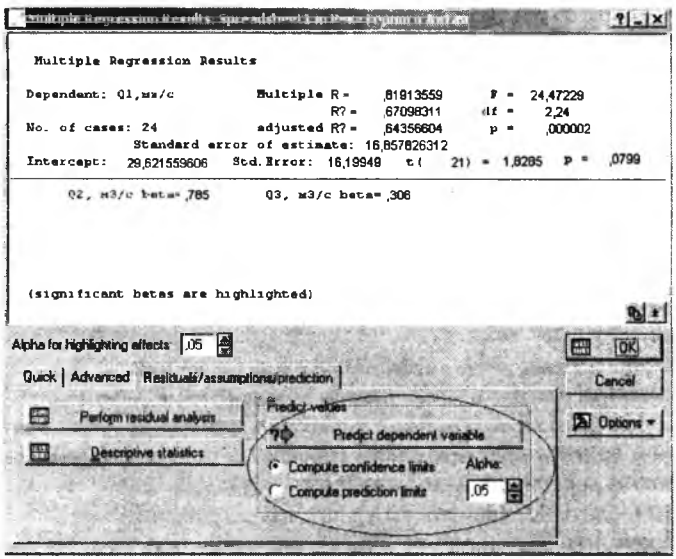

Рис. 6. Окно результатов модуля «Множественная регрессия»

Результаты прогноза выдаются в виде таблицы (рис. 7), из которой выпишем коэффициенты уравнения и значение свободного члена (Intercept). Окончательно модель прогноза примет вид линейного уравнения:

$$Q_1 = 0{,}613824\, Q_2 + 0{,}142025\, Q_3 + 29{,}6216.$$

Variable	Predicting Values for (Коррел) variable: Q1, м³/с B-Weight	Value	B-Weight * Value
Q2, м³/с	0,613824	87,0000	53,4027
Q3, м³/с	0,142025	310,0000	44,0279
Intercept			29,6216
Predicted			127,0521
-95,0%CL			110,1414
+95,0%CL			143,9629

Рис 7 Результаты прогноза на независимом материале

Для значений $Q_2 = 87$ и $Q_3 = 310$ получаем по уравнению регрессии значение $Q_1 = 127$ м^3/с, которое попадает в 95% доверительного интервала.

Пример 2. *Прогноз расходов воды по уравнению множественной регрессии с использованием электронной таблицы Excel.*

Для тех же данных, которые приведены в примере 1, требуется осуществить прогноза с заблаговременностью 2–3 шага в заданном масштабе времени и дать оценку результатов прогноза на зависимом и независимом материале. Принципиально эта задача с вычислительной точки зрения ничем не отличается от вышерассмотренной. Алгоритм ее решения будет состоять из следующих этапов:

1. Сформировать матрицу совместных наблюдений из всех имеющихся рядов.

2. Вычислить матрицу коэффициентов корреляции.

3. Оставить в списке предикторов (независимых переменных) только те ряды, которые характеризуются высоким парным коэффициентом корреляции с прогнозируемым рядом (предиктантом) и незначимыми коэффициентами корреляции между собой.

4. Вычислить коэффициенты уравнения регрессии.

5. Проверить адекватность модели анализом остатков. Чтобы можно было использовать полученное уравнение для прогноза, в остатках долж-

на отсутствовать внутрирядная зависимость, и они должны быть распределены по нормальному закону.

6. Вычислить по уравнению регрессии прогнозируемые значения для заданных моментов времени.

Будем решать эту задачу на примере электронного процессора Excel, подбирая уравнение множественной регрессии, в котором в качестве зависимой переменной Y будет выступать ряд Q_1 среднегодовых модулей стока, а независимых X_i – ряды Q_2–Q_8. Этапы 1–3 исследователю придется осуществлять самостоятельно, используя для получения матрицы коэффициентов корреляции процедуру *Корреляция* из *Пакет анализа* (меню *Сервис)*.

Электронный процессор Excel позволяет автоматизировать нахождение уравнения множественной линейной регрессии. Его *Пакет анализа* (меню *Сервис)* располагает для этого процедурой *Регрессия*.

Для реализации процедуры *Регрессия* необходимо вначале осуществить ряд установок, задающих условия задачи в окне *Регрессия*. Выполните команду *Сервис → Анализ данных* и в появившемся диалоговом окне выберите строку *Регрессия*. В следующем диалоговом окне (рис. 8) укажите входной интервал зависимой переменной Y путём ссылки на соответствующий диапазон таблицы данных.

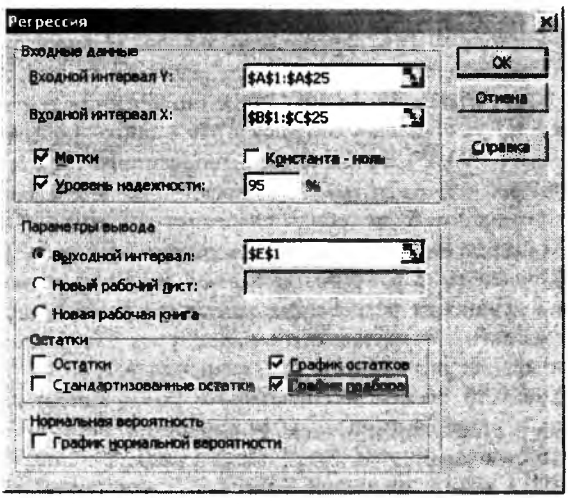

Рис. 8. Вид диалогового окна настроек регрессионного анализа

Далее, для следующего входного интервала укажите ссылку на диапазон таблицы данных, содержащий одну или несколько (не более 16) независимых переменных $\mathbf{X}_i$. Укажите выходной диапазон путём ссылки на верхнюю левую ячейку предполагаемой области вывода (или выберите новый рабочий лист).

Если используемые диапазоны данных содержат имена (метки), то установите флажок в поле «*Метки*». Для визуальной проверки отличия экспериментальных значений зависимой переменной от предсказанных по уравнению регрессии поставьте флажок в окно «*График подбора*». Если необходимо, можно поставить флажки в окошки графиков и остатков. При необходимости можно пометить и другие поля.

Результаты процедуры *Регрессия* показаны в табл. 7. В качестве первых результатов приводятся регрессионные статистики – коэффициент множественной корреляции R, оценивающий степень влияния независимых переменных на исследуемую зависимую переменную, и R^2 – коэффициент детерминации, оценивающий точность полученной регрессионной модели.

Если коэффициент детерминации (Нормированный R-квадрат в таблице итогов) больше 0,5, то обычно считается, что модель хорошо описывает данные (модель адекватна описываемому явлению). Если $R^2 < 0,5$, принято считать, что точность аппроксимации недостаточна и необходим пересмотр модели путём изменения типа зависимости или набора предикторов.

В середине таблицы (дисперсионный анализ) сравнивается доля изменчивости, описываемая регрессионным уравнением, и случайная изменчивость. Рассчитывается F-критерий и оценивается его уровень значимости. Если по умолчанию допускать 5% ошибку (критический уровень значимости $\alpha = 0,05$), то значение вычисленного уровня значимости (приводится в графе *Значимость F*) меньшее 0,05, как и высокий коэффициент детерминации, будет свидетельствовать о высокой достоверности уравнения регрессии (доверительная вероятность выше 0,95).

В нижней части табл. 7 приводятся значения коэффициентов и свободного члена уравнения регрессии с указанием доверительных границ и уровня значимости, обозначаемого в процедурах Excel символом *p*. Если вычисленный для коэффициентов уравнения регрессии уровень значимости больше 0,05, то можно считать этот коэффициент равным нулю и практически не влияющим на значение зависимой переменной.

Для условий нашей задачи коэффициент множественной корреляции составляет 0,81. Судя по значению R-квадрат, принятая линейная модель достаточно хорошо аппроксимирует зависимость Q_1 от Q_2 и Q_3. В целом

уравнение значимо, что видно из результатов дисперсионного анализа. Рассчитанный уровень значимости практически равен нулю ($p = 2E{-}6$ означает $p = 2 \cdot 10^{-6}$). Так как он меньше критического, то нулевая гипотеза о равенстве нулю коэффициентов уравнения может быть отвергнута.

Таблица 7

Итоги регрессионного анализа

ВЫВОД ИТОГОВ					
	Регрессионная статистика				
Множественный *R*	0,81				
R-квадрат	0,57				
Нормированный *R*-квадрат	0,64				
Стандартная ошибка	2,56				
Наблюдения	24				
	Дисперсионный анализ				
	df	*SS*	*MS*	*F*	*Значим F*
Регрессия	7	921,77	153,63	23,44	2,00E–6
Остаток	16	825,86	6,55		
Итого	23	1747,3			
	Коэфф	*Ст. ошиб.*	*t-стат.*	*p-знач.*	
Y-пересечение	29,622	5,94	–0,38	0,07	
Q2	0,614	0,13	237,17	0,04	
Q3	0,142	0,01	0,17	0,01	

На основании числовых данных, приведённых в нижней части табл. 7, можно составить уравнение регрессии. В него будут включены те ряды, у которых уровень значимости в столбце *p-значение* меньше критического 0,05. Без учёта незначимых коэффициентов уравнение регрессии может быть представлено следующим образом:

$$Q_1 = 0{,}614\ Q_2 + 0{,}142\ Q_3 + 29{,}622.$$

Наглядно уровень предсказания можно представить в виде графиков подбора, таблиц и графиков ошибок предсказания (остатков).

Вопросы для самопроверки

1. Каким образом можно количественно выразить степень тесноты связи между двумя переменными?

2. Позволяет ли установить «причинную» связь наличие корреляционной связи между двумя рядами наблюдений?

3. Как определить, является ли корреляционная связь между исследуемыми переменными статистически значимой?

4. Какие требования предъявляются к виду связей между рядами наблюдений?

5. Какие требования предъявляются к корреляции между предикторами?

6. Какие требования предъявляются к корреляции между предиктантом и предикторами?

7. В каких пределах изменяется множественный коэффициент корреляции R?

8. Может ли коэффициент множественной корреляции уменьшаться при добавлении ещё одного предиктора?

9. Дайте интерпретацию коэффициенту детерминации, равному 0,65.

10. Какой метод используется для оценки коэффициентов уравнения регрессии?

11. Какие критерии используются для оценки качества модели?

12. Какие величины называются остатками?

13. Какие способы проверки независимости остатков Вы знаете?

14. Как проверить согласие ряда остатков нормальному закону распределения?

15. Дайте определение обучающей выборке, контрольной выборке.

16. Поясните разницу между прогнозом на зависимом и независимом материале.

17. Каким способом проверяется независимость остатков в модуле «Множественная регрессия»?

18. Поясните физический смысл коэффициента детерминации.

19. Критерий качества прогноза S/σ получился равным 0,4. Является ли прогноз:

а) хорошим;

б) удовлетворительным;

в) неудовлетворительным.

20. Какие графические возможности имеются в пакете Statistica для отображения корреляционных связей?

3. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

3.1. Теоретические положения дискриминантного анализа

Дискриминантный анализ представляет собой расширение методов регрессионного анализа и находит широкое применение в науках о Земле. Некоторые исследователи рассматривают дискриминантный метод, как связывающий классическую элементарную одномерную статистику и многомерные статистические процедуры. Дискриминация трактуется как разделение и имеет много общего с классификацией. В обоих случаях происходит разделение объектов, однако в случае классификации количество объектов или кластеров нам априори не известно и может лишь предполагаться. При дискриминационных процедурах группы и их количество заданы заранее. Задача состоит в достаточно обоснованном отнесении объектов, в том числе и вновь обнаруженных, к уже известным группам.

Все методы дискриминантного анализа делятся на параметрические и непараметрические. Первые из них основаны на знании закона распределения в классах.

Идея дискриминантного анализа заключается в нахождении некоторого направления (линейной дискриминантной функции), вдоль которого изменчивость векторной случайной величины невелика, и точки-векторы разных групп объектов образуют в многомерном пространстве неперекрывающиеся области. Значение функции, располагающееся на границе областей, называется дискриминантным числом и может быть вычислено, а впоследствии использовано для разделения и отнесения объектов к соответствующим группам.

Дискриминантное преобразование даёт минимум отношения разности многомерных средних значений для некоторой пары групп к их многомерной дисперсии путём построения уравнения регрессии, в котором зависимыми переменными являются разности между многомерными средними этих групп. В матричном обозначении это уравнение может быть представлено как $[L] = [b] \times [D]$, где $[L]$ – вектор-столбец разности между векторами средних двух групп; $[b]$ – вектор коэффициентов в дискриминантной функции; $[D]$ – матрица дисперсий и ковариаций объединённой выборки. Уравнение решается с помощью операций обращения и умножения матриц

$$[b] = [D]^{-1} [L].$$

Значимость дискриминантной функции можно проверить с помощью уже известного нам критерия T^2, оценивая существенность различия между векторами средних групп наблюдений.

К сожалению, в пакете анализа Excel отсутствует процедура дискриминантного анализа, но можно воспользоваться уже известной нам процедурой расчета ковариационных матриц, регрессионного анализа и набором матричных функций.

Пусть имеются выборки двух объектов, предполагаемых разными, т.е. матрицы $\mathbf{X}^{(1)}$ и $\mathbf{X}^{(2)}$ размерностью соответственно $n_1 \times p$ и $n_2 \times p$. Обозначим объединённую выборку **X** объёмом $N \times p$, где $N = n_1 + n_2$. Алгоритм нахождения дискриминантной функции Excel состоит из следующих шагов. Найдём векторы средних, соответственно $\overline{X}^{(1)}$ и $\overline{X}^{(2)}$. Используя процедуру ковариации пакета анализа данных, рассчитаем матрицы дисперсий и ковариаций $[\mathbf{D}]^{(1)}$, $[\mathbf{D}]^{(2)}$ и [D]. Рассчитаем вектор [L] как разность векторов $\overline{X}^{(1)}$ и $\overline{X}^{(2)}$. Рассчитаем вектор [b] коэффициентов дискриминантной функции по приведённой выше формуле и запишем дискриминантное уравнение

$$S = b_1 F_1 + b_2 F_2 + \ldots + b_p F_p.$$

где $F_1, \ldots F_p$ – поименованные значения признаков; $b_1, \ldots b_p$ – коэффициенты уравнения регрессии.

Пусть для простоты дискриминантным параметрическим методом решается задача разделения двух классов объектов A_1 и A_2. Распознавание каждого вновь поступившего вектора (объекта) будет осуществляться по знаку разницы $S(\mathbf{F}) - S_0$, где S_0 – значение дискриминантной функции на границе между двумя классами. Если разность положительна, то согласно решающему правилу (Байеса) вектор **F** следует отнести к классу A_1, если отрицательна – то к A_2. Руководствуясь этим правилом, мы добиваемся минимума средних потерь, связанных с неправильным распознаванием.

Прибегая к геометрическим представлениям, можно сказать, что дискриминантная функция $S(\mathbf{F})$ описывает некоторую плоскость, которая наилучшим образом – в смысле средних потерь – разделяет объекты, принадлежащие разным классам.

3.2. Пример использования дискриминантного анализа для классификации речных вод по химическому составу

В качестве примера рассмотрим выборки содержаний микро- и макроэлементов в речных водах по результатам гидрохимического анализа проб, взятых из реки – водоприемника болотных вод (выборка Б) и реки, водосбор которой характеризуется малым коэффициентом заболоченности (выборка А). Продемонстрируем процедуры дискриминантного анализа с использованием ЭТ Excel на этом формальном примере.

Целью анализа является дискриминация (разделение) гидрохимического состава речных вод на основе обучающей выборки с тем, чтобы проба из нового источника была отнесена с высокой долей вероятности к одной из этих групп.

Используя процедуру «*Ковариация*» пакета анализа Excel, рассчитываем ковариационную матрицу объединённой выборки. Процедура выдаёт лишь нижнюю левую половину симметричной квадратной матрицы. Для дальнейших расчетов элементы, лежащие под главной диагональю, нужно скопировать в верхнюю часть матрицы, соответствующим образом их трансформировав (табл. 8).

Таблица 8

Фрагмент ковариационной матрицы объединённой выборки |D|

	Ca^{2+}	$Na^{+} + K^{+}$	HCO_3	Mn	Zn
Ca^{2+}	3,52	–0,69	–10,61	1,24	8,27
$Na^{+} + K^{+}$	–0,69	1,61	4,85	–0,44	4,66
HCO_3	–10,61	–4,85	124,14	–1,94	–107,07
Mn	1,24	–0,44	1,94	4,76	–2,41
Zn	8,27	4,66	–107,07	–2,41	97,56

Далее, используя функцию МОБР, рассчитываем обратную матрицу (табл. 9).

Таблица 9

Фрагмент обратной матрицы $|D|^{-1}$

	Ca^{2+}	$Na^{+} + K^{+}$	HCO_3	Mn	Zn
Ca^{2+}	0,60	0,44	0,04	–0,11	–0,02
$Na^{+} + K^{+}$	0,44	1,05	0,05	–0,01	–0,03
HCO_3	0,04	0,05	0,46	0,44	0,51
Mn	–0,11	–0,01	0,44	0,68	0,51
Zn	–0,02	–0,03	0,51	0,51	0,59

Необходимо помнить, что при расчёте обратной матрицы и при операциях умножения матриц для получения результата недостаточно кнопки <*OK*> диалогового окна, результат получается одновременным нажатием трёх клавиш <*Ctrl*>+ <*Shift*>+ <*Enter*>.

Рассчитываем разность векторов средних значений для выборок А и Б, используя обычные возможности электронных таблиц. Умножаем вектор разности средних на обратную матрицу и получаем вектор коэффициентов дискриминантного уравнения (табл. 10).

Таблица 10

Фрагмент таблицы с коэффициентами уравнения регрессии

Ca^{2+}	$Na^+ + K^+$	HCO_3	Mn	Zn
0,047	–0,024	–0,137	–0,154	0,043

Записываем итоговое уравнение дискриминантной линейной функции:

$$S = 0{,}047315\ Ca^{2+} - 0{,}02462\ (Na^+ + K^+) - 0{,}13746\ HCO_3 - 0{,}15444\ Mn + \\ + 0{,}043063\ Zn + \ldots.$$

Подставляя значения признаков в полученную формулу, можем рассчитать дискриминантные метки всех реализаций векторных случайных величин. Разделяющее дискриминантное число можно получить, подставляя элементы вектора средних объединённой выборки. Для нашего случая оно равно $S_0 = -1{,}30269$. Дискриминантные метки можно рассчитать все сразу в исходной таблице данных, подставляя в ячейки дискриминантное уравнение.

В нашем случае разделение на две группы произошло достаточно удовлетворительно. Все дискриминантные метки выборки А оказались больше S_0, а выборки Б – меньше.

Для наглядного представления результатов дискриминации построим средствами Excel диаграмму, выбрав в качестве осей координат один из признаков и дискриминантные метки DM (рис. 9). Значимость дискриминантной функции проверяется сравнением средних векторов групп анализов с использованием критерия Хоттелинга.

Несколько иную процедуру дискриминантного анализа предлагает система Statistica. Вернёмся к выборке А. В меню *Статистика* (*Statistics*) выберем модуль *Многомерные методы исследования (Multivariate Exploratory Techniques)* → *Дискриминантный анализ (Discriminant Analysis)*.

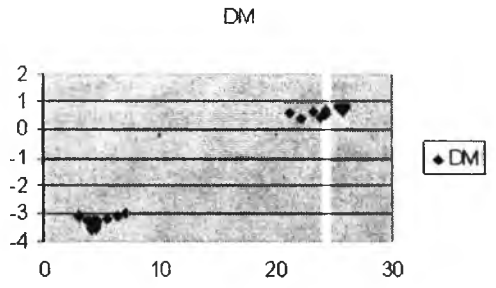

Рис. 9. Результаты дискриминации: нижняя группа – выборка Б, верхняя – выборка А, DM – дискриминантная метка

В качестве переменной с группирующими кодами *(Grouping)* назначим столбец с названием Min из таблицы с исходными данными, а в качестве кодов *(Codes for ...)* используем названия выборок – А и Б. Группирующая переменная – это вспомогательный столбец, с помощью которого объединенная выборка делится на две части.

Другие переменные, включённые в анализ, будем рассматривать как независимые *(Independent)*. В этом же диалоговом окне поставим флажок в окошко *Расширенные опции (пошаговый анализ) – Advanced options (stepwise analysis)*. В следующем диалоговом окне *Описание модели* выберем метод *вперёд пошагово*. В результате реализации метода появится диалоговое окно *Результаты дискриминантного анализа (Discriminant Function Analysis Results)* (рис. 10).

Верхняя информационная часть окна выдаёт обобщающую информацию:

а) использован пошаговый анализ;

б) в модель включено 8 переменных;

в) значение статистики λ Уилкса, близкое к нулю, свидетельствует в целом о хорошей дискриминации с уровнем значимости заметно меньшим 0,05.

Общие итоги пошагового анализа можно увидеть на вкладке *Расширенный (Advanced)* диалогового окна результатов анализа, где переменные расположены в порядке очерёдности в соответствии с уменьшением их вклада в дискриминантную функцию (табл. 11).

Уровень значимости *p-level* показан для отдельных переменных в шестой колонке.

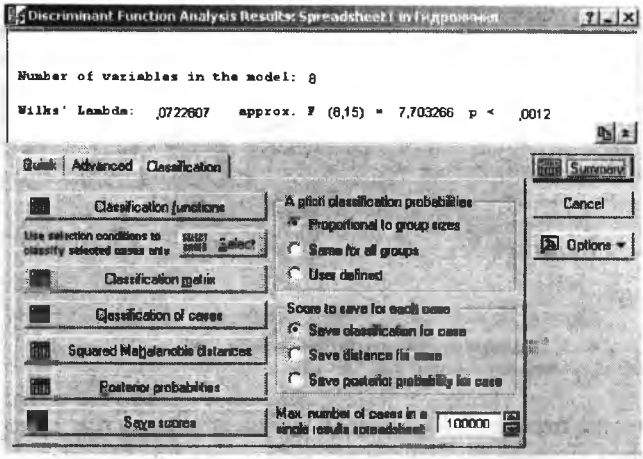

Рис. 10. Диалоговое окно результатов дискриминантного анализа

Таблица 11

Фрагмент результирующей таблицы дискриминантного анализа

	Шаг	*F*	*df* 1	*df* 2	*p-level*	*n*	Лямбда	*F-value*	*df* 1	*df* 2	*p-level*
Ca^{2+}	1	1043,72	1	20	0,00	1	0,018	1043,7	1	20	0,00
$Na^{+} + K^{+}$	2	3,62	1	19	0,07	2	0,015	592,2	2	19	0,00
HCO_3	3	16,24	1	18	0,00	3	0,008	717,0	3	18	0,00
Mn	4	5,28	1	17	0,03	4	0,006	667,0	4	17	0,00

Коэффициенты дискриминантной функции (табл. 12) можно найти на вкладке *Классификация (Classification)* диалогового окна *Результаты дискриминантного анализа,* нажав кнопку *Функции классификации (Classification Functions).*

Таблица 12

Коэффициенты функций классификации

	Выборка А	*Выборка Б*
Ca^{2+}	23,00	16,93
$Na^{+} + K^{+}$	13,85	33,43
HCO_3	10,21	32,44
Mn	20,94	12,67
...	...	...
Постоянная	–928,40	–1032,50

С помощью этих функций можно вычислить дискриминантные метки (классификационные значения) для вновь получаемых реализаций случайной величины. Функции для вычисления дискриминантных меток в рассматриваемом примере будут иметь вид:

$$\textit{Выборка А} =$$
$$= 23{,}001\ Ca^{2+} + 13{,}858\ (Na^{+} + K^{+}) + 10{,}217\ HCO_3 + \ldots + 928{,}402;$$

$$\textit{Выборка Б} =$$
$$= 16{,}93\ Ca^{2+} + 33{,}43\ (Na^{+} + K^{+}) + 32{,}44\ HCO_3 + 1 + \ldots + 1032{,}50.$$

Новая реализация случайной величины будет отнесена к той группе, для которой классификационное значение окажется максимальным.

Вопросы для самопроверки

1. Что понимается под термином дискриминация?
2. Что такое переменная группирования?
3. Как связаны между собой регрессионный и дискриминантный анализы?
4. Что такое дискриминантное число?
5. Какая случайная величина рассматривается как зависимая при дискриминантном анализе?
6. Назовите основные составляющие дискриминантной функции.
7. В чём сходство и различия дискриминантного и кластерного анализов?
8. Что такое дискриминантная метка и как она рассчитывается?
9. Какие независимые переменные включаются в дискриминантную функцию?
10. Как оценивается вклад независимых переменных в значение дискриминантной метки?
11. Какой смысл имеет термин «пошаговый дискриминантный анализ»?
12. В каком случае независимые переменные могут не учитываться в составе дискриминирующей функции?

4. КЛАСТЕРНЫЙ АНАЛИЗ

4.1. Основные положения кластерного анализа

Под термином «кластер-анализ» понимают набор методов, позволяющих выделять в исходных многомерных данных такие однородные подмножества, чтобы объекты внутри групп были похожи друг на друга, а объекты из разных групп – не похожи. Под «похожестью» понимается близость объектов в многомерном пространстве признаков. Тогда задача сводится к выделению в этом пространстве естественных скоплений объектов, которые и считаются однородными группами (Тюрин, Макаров, 1995). Полученные в результате разбиения группы обычно называются кластерами (таксонами, образами). Кластер (cluster) по-английски означает гроздь, пучок, скопление, группа элементов, характеризующихся каким-либо общим свойством. Это не очень строгое определение. Строго говоря, ***кластером называется такая группа объектов из рассматриваемого множества, для которой средний квадрат внутригруппового расстояния до центра группы меньше среднего квадрата расстояния до общего центра в исходной совокупности*** (Енюков, 1986).

Если данные представлены в виде матрицы **X** (объект – признак), то анализируемые объекты удобно интерпретировать геометрически как точки в многомерном признаковом пространстве. Если признаков всего три, то исследуемые объекты представляются в виде точек в нашем обычном трёхмерном евклидовом пространстве. Естественно предполагать, что геометрическая близость двух или нескольких точек в этом пространстве обозначает близость физических состояний соответствующих объектов, их однородность. Тогда проблема классификации состоит в разбиении анализируемой совокупности точек-наблюдений на сравнительно небольшое число классов, таких, что точки, принадлежащие к одному классу, «близки» друг к другу, а точки из разных классов «далеки друг от друга».

Кластеры должны удовлетворять требованиям:

1) кластер не может быть пустым множеством;

2) кластеры должны отличаться друг от друга;

3) все объекты должны быть расклассифицированы однозначно, т.е. кластеры не могут пересекаться, каждый объект должен принадлежать только одному кластеру.

Последнее требование очень жёсткое и трудновыполнимое в данных с «плохой структурой». Реальные системы зачастую обладают настолько «плохой» структурой с трудно различимыми между собой состояниями, что границу между ними установить нелегко. Современные алгоритмы кластер-анализа не дают успешных результатов в случае, когда имеют место «цепочки» или «мосты» между кластерами (рис. 11, *а*), когда кластеры несферической формы (рис. 11, *б–в*), когда они сильно различаются по объёму и плотности. Если третий пункт не выполняется, то говорят о «зашумлении» данных.

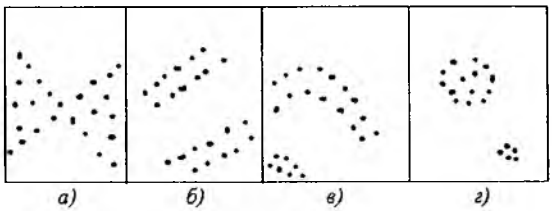

Рис. 11. Примеры формы кластеров для данных с «плохой» структурой (Елисеева, Рукавишников, 1977)

В отличие от других методов в кластер-анализе для группирования используются все признаки объектов, и отсутствует обучающая выборка (распознавание образов без «учителя»). Эти отличия являются достаточно существенными и определяют во многом своеобразие методологии применения кластерного анализа для решения задач классификации.

Центральное место в кластерном анализе принадлежит двум понятиям (мера близости объектов друг к другу и мера близости групп (классов объектов)), которые порождены двумя главными проблемами:

1) как вычислять расстояния между объектами;

2) по какому условию считать, достаточно ли мало расстояние между двумя объектами и их можно включить в один кластер, или оно «достаточно большое» и точки должны попасть в разные кластеры?

Наиболее трудным является первый пункт, связанный с понятием однородности объектов. В общем случае понятие однородности объектов задается введением метрики – правила вычисления расстояний между любой парой объектов исследуемого множества.

В зависимости от способа решения этих двух проблем и сами методы кластерного анализа можно разбить на виды. Традиционно их разделяют

на два вида: *иерархические* (древообразные) (рис. 12) и *неиерархические*. Последние ещё называют структурными, потому что в них реализуется идея образования кластеров по принципу выделения сгущений. Введём понятие *сгущения как группы объектов, для которых максимальный квадрат расстояния её точек до центра группы меньше среднего квадрата расстояния между объектами* (Енюков, 1986), т.е., проще говоря, сгущения – это места наибольшей концентрации точек в рассматриваемом пространстве. В качестве представителя структурной кластер-процедуры чаще всего в статистических пакетах выступает метод *k*-средних в различных модификациях.

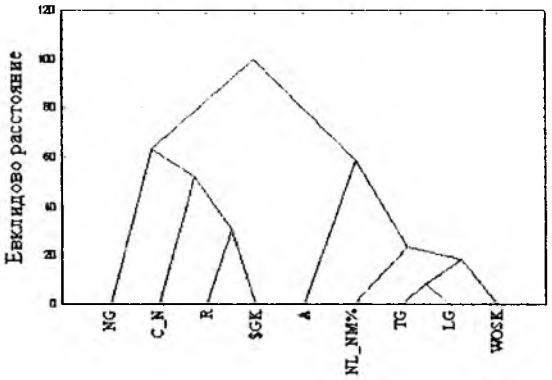

Рис. 12. Пример дендрограммы иерархической процедуры группирования объектов, построенной в пакете Statistica

По направлению кластеризации методы делятся на *агломеративные* (объединяющие) и итеративные *дивизивные* (разделяющие). В агломеративных вначале каждый объект образует кластер из одного себя. На каждом последующем шаге происходит объединение двух наиболее близких кластеров в один. В конечном итоге все попадают в один класс. В некоторых алгоритмах можно заказать момент остановки этого процесса указанием расстояния, при котором допустимо объединение. В результате выделяются формальные иерархические уровни «кластеризации» всех объектов (Факторный, дискриминантный и кластерный анализ, 1989). В дивизивных методах, наоборот, сначала все объекты представляют один класс, а затем производится их разбиение.

Результаты работы агломеративных иерархических методов изображаются в виде древовидного графа – *дендрограммы* – дерева объединения кластеров. Если данные имеют ясную «структуру», то в иерархическом дереве будут чётко видны различные ветви. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их (рис. 12).

По оси абсцисс дендрограммы располагаются символические обозначения объектов исследования, а по оси ординат – минимальные значения дистанционных коэффициентов, соответствующих каждому шагу классифицирующей процедуры. Таким образом, по оси ординат можно проследить иерархический уровень группирования.

Возможность визуального представления структуры данных в виде дендрограммы делает агломеративные иерархические методы привлекательными и удобными для анализа, поэтому в статистических пакетах они представлены наиболее полно.

Как методы объединяющего, так и разделяющего типов могут быть реализованы при помощи различных алгоритмов. Как правило, в статистических пакетах предпочтение отдаётся иерархическим методам кластер-анализа, например, в пакете Statistica предлагается шесть иерархических методов и один неиерархического типа (метод k-средних). Это объясняется тем, что в отличие от неиерархических, представляющих исследователю конечный результат группировки данных, иерархические процедуры позволяют проследить процесс выделения группировок и иллюстрируют соподчинённость кластеров, образующихся на разных этапах работы.

К недостаткам иерархических процедур следует отнести громоздкость вычислительной реализации на ЭВМ, немалые требования к памяти, большие затраты машинного времени. В связи с этим использование иерархических алгоритмов для задач с большим набором данных (сотни объектов) нецелесообразно или невозможно.

Следует подчеркнуть ещё одно различие иерархических и неиерархических алгоритмов: первые всегда построят дендрограмму, а задача естественного расслоения исходных данных на четко выраженные кластеры, решаемая вторыми, может и не иметь решения.

4.2. Способы вычисления расстояний между объектами

Как вычислять расстояния между объектами? Для количественной оценки сходства объектов, как уже говорилось выше, вводится метрика,

используемая для вычисления расстояния (удалённости) между объектами. Выбор метрики играет определяющую роль при решении задач кластер-анализа. Использование неадекватной функции расстояния может привести либо к плохому качеству разбиения, либо к разбиению, лишённому содержательного смысла с точки зрения исследователя даже в том случае, когда объекты потенциально могут быть сгруппированы.

Если входная информация задана в виде матрицы **X** (объект – признак), то возможные варианты вычисления типов расстояний между объектами будут иметь вид:

– *Евклидово расстояние.* Это наиболее общий тип расстояния. Оно попросту является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2},$$

где d_{ij} – расстояние между i-м и j-м объектами (точками); k – номер свойства (признака) этих объектов (k = 1, 2, … m). Если признаков всего три (k = 1, 2, 3), то это наше обычное трёхмерное пространство, а d – расстояние между двумя точками, измеренное, например, линейкой.

Евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным. Это обычный способ его вычисления, который имеет определенные преимущества. Например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться выбросом. Недостаток этой метрики заключается в неравноправности осей пространства. При ненормированных осях возможен случай, когда объекты, сходные по всем признакам, кроме одного, по которому они сильно разнятся, будут находиться далеко друг от друга в евклидовом пространстве.

Этот недостаток можно устранить подбором весов $\omega(k)$, приписывая более важным признакам больший вес, тогда мерой сходства будет *взвешенное евклидово расстояние*:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \omega(k) \left| x_{ik} - x_{jk} \right|^2},$$

$$\sum_{k} \omega(k) = 1.$$

Выбор весов связан с дополнительными исследованиями и полностью находится в руках исследователя.

Понятно, что в реальных данных признаки могут очень сильно отличаться по абсолютному значению, что также приводит к неравноправности их в вычислении расстояний, а, значит, и в оценке сходства между объектами. Можно устранить этот недостаток и по-другому, нормируя данные перед использованием кластер-анализа одним из следующих способов:

$z_{ik} = (x_{ik} - \overline{x}_i)/\sigma_i$ – обычная стандартизация, где $\overline{x}_i$ – среднее арифметическое значение, σ_i – среднее квадратическое отклонение ряда значений x_i;

$z_{ik} = x_{ik}/x_{\max i}$ – отнесение признаков объекта к максимальному значению в их наборе;

$z_{ik} = x_{ik}/x_i$ – отнесение признаков объекта к среднему значению в их наборе;

$z_{ik} = x_{ik}/x_{\min i}$ – отнесение признаков объекта к минимальному значению в их наборе;

$z_{ik} = (x_{ik} - x_{\min i})/(x_{\max i} - x_{\min i})$ – нормирование к единичному интервалу. Значения признаков будут заключены в *единичном* интервале

$$0 \le z_{ik} \le 1.$$

Евклидово расстояние рекомендуется применять в случаях, если:

а) признаки достаточно однородны по своему физическому смыслу, причём, установлено, что все они одинаково важны с точки зрения решения вопроса об отнесении объекта к тому или иному классу;

б) размерность признакового пространства (***m***) невелика.

– *Квадрат евклидова расстояния.* Иногда используют квадрат евклидова расстояния, чтобы придать большие веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

$$d_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2.$$

– *Расстояние городских кварталов (манхэттенское расстояние).* Это расстояние является просто средним разностей по координатам. В боль-

шинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат). Манхэттенское расстояние вычисляется по формуле:

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|.$$

– *Расстояние Чебышева.* Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением). Расстояние Чебышева вычисляется по формуле:

$$d_{ij} = \max_{k} \left| x_{ik} - x_{jk} \right|.$$

– *Степенное расстояние.* Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния. Например, надо увеличить вес того свойства объекта, которое более важно для данного исследования. Степенное расстояние вычисляется по формуле:

$$d_{ij} = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^{p} \right)^{1/r},$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r отвечает за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p – равны двум, то степенное расстояние совпадает с расстоянием Евклида.

– *Процент несогласия.* Эта мера используется в случаях, когда данные являются категориальными.

– *Коэффициент корреляции.* В качестве меры близости точек используется величин $1 - r$, где r – коэффициент корреляции Пирсона. Матрицы расстояний обычно применяются с бóльшим успехом, чем матрицы коэффициентов корреляции, потому что они меньше реагируют на замену алгоритма кластеризации.

Существует ещё насколько типов расстояний (Хемминга, Махаланобиса и т.д.), которые здесь не приводятся.

4.3. Правила объединения кластеров

Следующим важным этапом кластеризации данных является процедура объединения объектов в классы. По какому условию считать, достаточно ли мало расстояние, чтобы включить объект в данный кластер? По какому правилу объединять кластеры? Рассмотрим некоторые правила объединения объектов в кластер, реализованные в алгоритмах агломеративных иерархических методов (Боровиков, 2003; Электронные ресурсы).

Поскольку в агломеративных методах на первом шаге каждый объект понимается как кластер, то будем говорить о правилах объединения кластеров, подразумевая, что эти правила распространяются и на условия включения объекта в кластер.

Метод ближайшего соседа или одиночная связь. В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Отсюда и название метода одиночной связи, так как для объединения нужна только одна связь. Недостатком этого метода является образование слишком больших «продолговатых» кластеров – «цепочек».

Метод наиболее удаленных соседей (полная связь). Метод полных связей позволяет устранить недостаток предыдущего метода. В нём расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. «наиболее удаленными соседями»). Здесь мера сходства между кандидатом на включение в группу и членами группы не может быть меньше некоторого *порогового* значения. Этот метод обычно работает очень хорошо, когда объекты существенно различаются. Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является «цепочечным», то этот метод непригоден.

Невзвешенное попарное среднее. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные классы, однако он работает одинаково хорошо и в случаях протяженных («цепочного» типа) кластеров.

Взвешенное попарное среднее. Метод идентичен методу невзвешенного попарного среднего за исключением того, что при вычислениях

размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован, когда предполагаются неравные размеры кластеров.

Невзвешенный центроидный метод. В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

Взвешенный центроидный метод (медиана). Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Если подозреваются значительные отличия в размерах кластеров, то этот метод оказывается предпочтительнее предыдущего.

Метод Уорда. Этот метод отличается от всех других, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Так как алгоритмы объединения работают формально, то задачей исследователей является подбор правильного метода для учета особенностей физической природы своих данных.

Из *неиерархических методов* рассмотрим метод ***k***-средних, принадлежащий к группе итеративных методов эталонного типа. Он является частным случаем общего метода динамических сгущений, отличающихся друг от друга по способу определения центра кластера. В методе ***k***-средних за центр класса принимается центр тяжести, который вычисляется как среднее арифметическое объектов класса. Расстояние между объектом и классом понимается как евклидово расстояние между объектом и центром тяжести класса. Например, если в кластере 2 объекта, соединим их отрезком прямой и найдем середину. Эта точка и будет центром тяжести нашего кластера. Объект относится к тому классу, расстояние до которого минимально. В отличие от иерархических процедур метод ***k***-средних работает непосредственно с объектами, а не с матрицей сходства.

Принципиально метод ***k***-средних «работает» следующим образом:

1) для начала процедуры классификации должны быть заданы ***k*** случайно выбранных объектов, которые будут служить *эталонами*, т.е. центрами кластеров (число их определяется пользователем);

2) затем происходит перемещение точек: каждая точка перемещается в ближайший к ней кластер;

3) вычисляются центры тяжести новых кластеров;

4) шаги 2–3 повторяются, пока не будет найдено такое разбиение объектов на заранее заданное количество классов k, чтобы минимизировался функционал качества, вычисляющий сумму внутриклассовых расстояний между объектами, или число итераций не превысит заданное пользователем. Шаги 2–3 представляют собой одну итерацию, т.е. одно очередное по счёту приближение к решению.

Сущность итеративных методов заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров.

Например, выбранное случайным образом число кластеров может не только сильно увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок задать в качестве начального приближения k случайно выбранных объектов, которые будут служить *эталонами*, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие.

Как и в иерархическом кластерном анализе в итерационных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров. Но для окончательного решения вопроса о структуре изучаемой совокупности можно испробовать несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры. Тогда появляется возможность выбрать наилучшее разбиение по задаваемому критерию качества.

4.4. Функционалы качества разбиения на классы

Как оценить успешно или неуспешно решена задача кластеризации данных? Чётких разработок по проверке адекватности результатов разбиения данных на однородные группы пока нет, хотя имеются способы определения количественных критериев, следуя которым можно было бы

предпочесть одно разбиение другому. С этой целью в кластер-анализе вводится понятие так называемого функционала качества разбиения $Q(S)$, определённого на множестве всех разбиений. Функционал зависит от объёмов групп и расстояний между объектами, вошедшими в отдельные группы. Под наилучшим разбиением понимается то разбиение S^* из всех S, на котором достигается экстремум (минимум или максимум) выбранного функционала качества. Выбор того или иного функционала качества, как правило, также осуществляется произвольно и опирается скорее на профессионально-интуитивные соображения, чем на какие-либо строгие правила (Айвазян, Бежаева, Староверов, 1974).

Ниже приводятся примеры нескольких функционалов качества, когда в качестве $Q(S)$ используется:

1. Сумма квадратов расстояний до центров классов:

$$Q(S) = \sum_{l=1}^{k} \sum_{i \in S_l} d^2(X_i, \bar{X}_l),$$

где l – номер кластера (l = 1, 2, ... k); $\bar{X}$ – центр i-го кластера, $[X_i]$ – вектор значений переменных для i-го объекта, входящего в i-й кластер; $d(X_i, \bar{X}_l)$ – расстояние между i-м объектом и центром l-го кластера. При использовании этого критерия стремятся получить такое разбиение совокупности объектов на k кластеров, при котором значение $Q(S)$ было бы *минимальным*.

2. Сумма внутриклассовых расстояний между объектами:

$$Q(S) = \sum_{l=1}^{k} \sum_{i,j \in S_l} d_{ij}^2 .$$

В этом случае наилучшим следует считать такое разбиение, при котором достигается *минимальное значение* $Q(S)$, т.е. получены кластеры большой «плотности». Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

3. Суммарная внутриклассовая дисперсия:

$$Q(S) = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{lj}^2,$$

где σ_{ij}^2 – дисперсия j-й переменной в кластере S_l. В данном случае разбиение, при котором сумма внутриклассовых (внутригрупповых) дисперсий будет *минимальной*, следует считать оптимальным.

В последнее время чаще применяются критерии в виде отношений показателей «населённости» кластеров к расстоянию между ними. Это может быть, например, отношение суммы межклассовых расстояний к сумме внутриклассовых (между объектами) расстояний, или отношение общей дисперсии данных к сумме внутриклассовых дисперсий.

После распределения объектов по k кластерам сравнивают первоначальный состав этих кластеров с вновь полученным. Если обнаруживается несовпадение, работа алгоритма продолжается. Локальный экстремум достигается в том случае, если совпадают результаты последующей и предыдущей группировок.

В большинстве случаев критерии качества связаны с определёнными алгоритмами классификации, т.е. определенный алгоритм обеспечивает получение экстремального значения соответствующего функционала качества. Например, на принципе минимизации внутрикластерной дисперсии основаны алгоритмы метода k-средних и метода Уорда.

К настоящему времени предложено много различных способов как вычисления расстояний между объектами, так и целевых функций, поэтому кластерный анализ – это не один метод, а набор (более 200 схем).

4.5. Сравнительный анализ различных методов кластерного анализа

Столкнувшись с множеством методов, каждый из которых даёт несколько отличающийся результат, исследователь вправе спросить, какой из них лучше? Зависимость результатов кластеризации от методов тем сильнее, чем менее явно изучаемая совокупность данных разделяется на группы объектов. Существенное влияние на характеристики кластерной структуры оказывают: во-первых, набор признаков, по которым осуществляется классификация; во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например, метод k-средних, то задаваемые начальные усло-

вия разбиения в значительной степени определяют конечный результат разбиения.

Имеются некоторые рекомендации практического плана по выбору алгоритма кластеризации, полученные на основе экспериментального сравнения различных методов (Мандель, Чёрный, 1988):

– Результаты классификации в целом будут лучше, если признаков немного, а число классов большое.

– Близкие результаты наблюдаются только у методов группового среднего и центроидного. Особенно близкие результаты из неиерархических методов дают вариант Мак-Кина и вариант Бола и Холла метода ***k***-средних.

– Наиболее устойчивы к «зашумлению» данных метод ***k***-средних (вариант Болла и Холла), метод Уорда; наименее – медианный метод и особенно метод ближайшего соседа.

– Для алгоритмов типа ***k***-средних очень важен выбор начального разбиения. Случайный выбор центров классов и выбор по центроиду не рекомендуется.

– При малом числе признаков и большом числе классов хорошо работают (восстанавливают истинную структуру) метод Уорда, метод группового среднего, метод дальнего соседа. Плохо работают – метод ближайшего соседа, метод ***k***-средних (варианты Мак-Кина и Хартигана).

– При большом числе признаков и малом числе классов хорошо работают метод Уорда, метод ***k***-средних (вариант Мак-Кина), плохо – метод ближайшего соседа, медианный метод, центроидный метод.

– Во всех случаях хорошо работают метод дальнего соседа и особенно метод Уорда.

– Во всех случаях плохо работает медианный метод и особенно метод ближайшего соседа.

– С учётом чувствительности к «зашумлению» и способности к восстановлению структуры данных наилучшим является алгоритм Уорда, наихудшим – метод ближайшего соседа (Гублер, 1978).

– По исследованиям канадских учёных, помехоустойчив алгоритм метода «средняя связь», а центроидный и медианный зависят от размеров групп.

– Если число объектов (образцов) велико (1 000 и более), а число признаков мало (5 или 10), метод ***k***-средних предпочтительнее иерархических, так как процесс получения решения происходит почти на два порядка быстрее (Дэвис, 1990).

Подчеркнём, что результаты классификации тем успешнее, чем больше объём выборки ***n*** и меньше соотношение ***m/n*** (число признаков долж-

но быть в несколько раз меньше числа объектов). Для получения содержательно-осмысленной классификации (если она вообще потенциально возможна) полезны следующие методические приёмы:

– применять к данным несколько алгоритмов классификации;

– применять для анализа данных несколько типов метрик (расстояний) с последующим сравнением результатов;

– выбирать результат, наиболее устойчивый к изменениям.

Успех применения кластерного анализа во многом зависит от информации, которой обладает исследователь относительно ожидаемого разделения на классы. Все вышеперечисленные рекомендации носят формальный характер и могут оказать лишь косвенную помощь исследователю. **Оценка качества и разумности полученного разбиения должна носить содержательный характер, поэтому при выборе окончательного варианта очень важны интуиция и знания исследователя в данной предметной области.**

Несмотря на отсутствие надёжных статистических способов оценки адекватности результатов кластерного анализа истинной структуре данных, его методы широко используются в науках о Земле. Например, методы кластерного анализа могут быть полезны как вспомогательный этап в районировании территории по условиям формирования стока, в задачах районировании территории по гидрохимическим показателям загрязнения поверхностных и подземных вод и т. д.

4.6. Примеры использования кластерного анализа при решении задач прикладных исследований

Основные сведения об устройстве модуля Cluster Analysis из пакета Statistica 6.0. В пакете предлагается *семь* способов вычисления расстояния между объектами и *семь* методов объединения их в кластеры на основе рассчитанных расстояний. Из них шесть алгоритмов представляют агломеративные иерархические методы и один неиерархический (*k*-средних).

Методы кластеризации, используемые в пакете Statistica:

– Single linkage – метод одиночной связи («ближайшего соседа»).

– Complete linkage – метод полной связи («дальнего соседа»).

– Unweighted pair group average – невзвешенный метод «средней связи».

– Weighted pair group average – взвешенный метод «средней связи».

– Unweighted centroid pair group – невзвешенный центроидный метод.

– Weighted centroid pair group (median) – взвешенный центроидный метод.

– Ward method – метод Уорда.

В пакете доступны следующие меры расстояния между объектами:

– Squared Euclidian distances – квадрат евклидова расстояния.

– Euclidian distances – евклидова мера.

– City-block (Manhattan) - манхэттенское расстояние (или «расстояние городских кварталов»).

– Chebyshev distance metric – метрика Чебышева.

– Power: SUM(ABS(x – y)p)1/r – степенное расстояние.

– Percent disagreement – процент несогласия (для бинарных данных, например, результатов социологического опроса).

– 1 – Pearson *r* – пирсоновский коэффициент корреляции (точнее, 1 – *r*).

Как любая интеллектуальная автоматизированная система, модуль устроен так, что перед началом работы программа запрашивает у пользователя ряд уточняющих сведений, конкретизирующих постановку задачи.

Для определённости будем считать, что таблица с данными представляет собой матрицу из *n* строк и *m* столбцов. Пусть объекты расположены в столбцах, а каждая строка содержит значения какого-либо одного признака. В действительности данные в матрице могут располагаться и наоборот, что непринципиально.

На первом шаге надо определиться с типом анализа: иерархический *(Joining)*, ***k*-средних** *(K-means clustering)* или двухсторонний (по объектам и их признакам одновременно) – *Two-way joining*. Последний тип анализа считается перспективным, но используется крайне редко из-за недостаточности теоретических разработок, поэтому будем рассматривать только два первых (рис. 13).

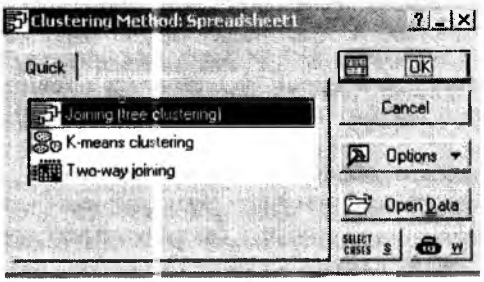

Рис. 13. Диалоговое окно для выбора типа анализа

Второй шаг – формулирование условий решения задачи. Если был выбран анализ иерархическими методами *(Joining)*, то, естественно, необходимо будет уточнить, какой именно метод желательно использовать.

Список для выбора предлагается в поле под названием *Amalgamation (linkage) rule* (рис. 14). Кроме того, потребуется выбрать тип расстояния из списка поля *Distance measure* и, конечно, указать программе, из каких столбцов таблицы брать данные для анализа.

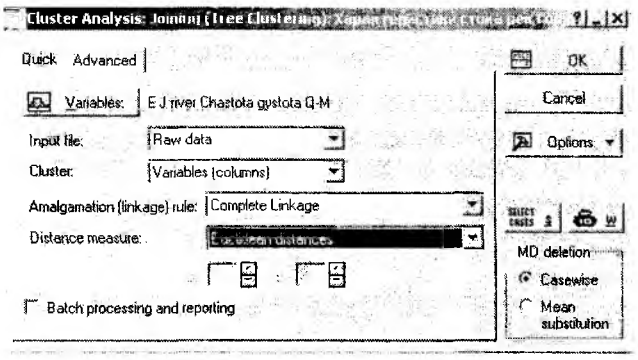

Рис. 14. Диалоговое окно для настроек кластер-анализа одним из иерархических методов

Диалоговое окно для выбора данных вызывается щелчком на *Variables*. Процедура может обрабатывать как матрицу с исходными данными *(Raw data)*, так и заранее подготовленную матрицу сходства (расстояний) типа признак-признак, поэтому следует также указать тип матрицы в поле *Input file*.

Если на предыдущем шаге был выбран неиерархический метод ***k***-средних, то набор установок для него другой, более обширный (рис. 15). В меню настройки анализа методом ***k***-средних необходимо указать, что является объектами – сроки или столбцы (в поле *Cluster*). Указать желаемое количество кластеров в поле *Количество групп (Number of clusters)* (по умолчанию 2); число итераций *(Number of iterations)* (по умолчанию 10). Необходимо также задать начальное положение центров тяжести кластеров (эталонов), которое затем будет уточняться на каждой последующей итерации.

В разделе *Начальные центры кластеров (Initial cluster centers)* для выбора эталонов предлагается три способа:

– *Выберите измеренные (Choose observations to maximize initial between cluster distances)* – Выбрать в качестве эталонов объекты, отстоящие друг друга от друга на максимальном расстоянии.

– *Отсортируйте расстояния... (Sort distances and take observations at constant intervals)* – Отсортировать расстояния между объектами и назначить эталонами точки, расположенные через равные интервалы (стоит по умолчанию).

– *Выберите первые номера... (Choose the first N (Number of clusters) observations)* – Назначить в качестве эталонов первые *N* наблюдений (*N* – число кластеров).

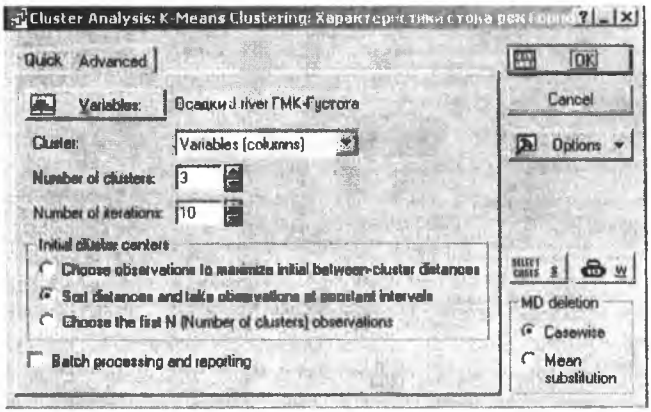

Рис. 15. Диалоговое окно для настроек метода *k*-средних

Допустим, Вы заказали разбиение на 4 класса, тогда при выборе этого алгоритма первые четыре объекта будут назначены эталонами.

Третий шаг – обсуждение результатов. Это наиболее творческий и мучительный этап для исследователя. К нему по праву можно отнести образное выражение: «*Математика может избавить нас от мучительной необходимости размышлять, но мы должны платить за эту привилегию, испытывая муки раздумий как до того, как математика вступает в действие, так и после*» (Каплан, 1964). Задачей исследователя на этом этапе является максимальное использование для анализа графических и табличных результатов работы модуля.

Диалоговое окно результатов состоит из двух частей: верхней информационной и нижней функциональной. В информационной части высве-

чиваются сведения о параметрах задания, а в нижней функциональные кнопки для анализа результатов (рис. 16).

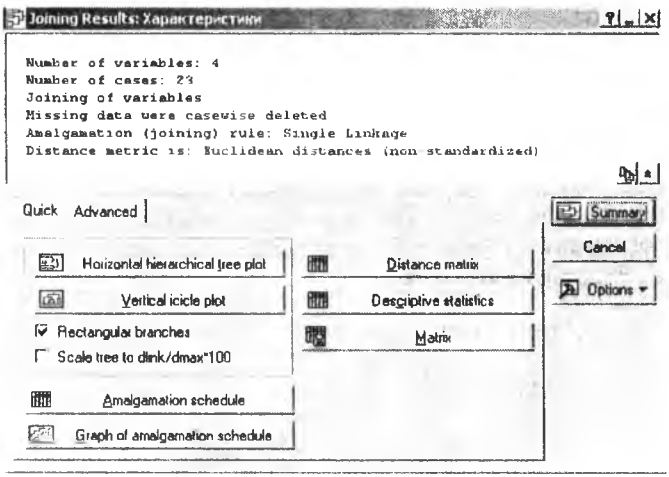

Рис. 16. Диалоговое окно результатов

Интерпретацию результатов кластерного анализа лучше рассматривать на примерах.

Пример 1. *Классификация рек Хакасии по внутригодовому распределению стока.* Рассмотрим пример разбиения рек Хакасии на однородные группы по водному режиму. Сочетание анализа физико-географических условий и группирование с помощью кластер-анализа, очевидно, позволит в дальнейшем более уверенно проводить районирование гидрологических объектов (бассейнов) по водному режиму.

Предварительную классификацию исследуемых бассейнов можно проводить по признаку однородности внутри классов. Задача такой классификации сводится к объединению бассейнов в группы с максимальной однородностью внутри них и минимальной – между ними.

Для классификации и районирования рек можно использовать внутригодовое распределение стока, которое отражает характерные черты режима стока, наиболее часто повторяющиеся в группе рек данного гидрологического района. Однако при этом теряются некоторые частные особенности. Поэтому мы будем иметь дело с исходными временными среднемесячными рядами расходов воды.

Исходной информацией для классификации и районирования послужили данные наблюдений за стоком, а именно, доли среднемесячных расходов воды от годового для 23 рек. Таким образом, имеем матрицу данных размером 23×12, в качестве объекта кластеризации выбирается река-пост. Так как все признаки представляли собой отношение расхода к среднегодовому, то в нормализации данных перед началом работы не было необходимости.

В первую очередь были построены дендрограммы объединения рек в однородные группы, полученные разными иерархическими методами кластерного анализа (рис. 17–18), что даёт возможность сравнить полученные результаты и с большей уверенностью выделять реки с общими условиями формирования стока.

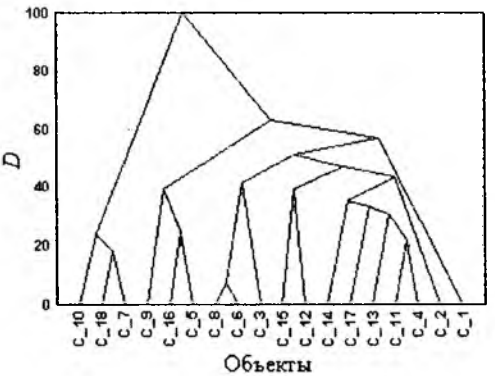

Рис. 17. Классификация рек по внутригодовому распределению стока методом «наиболее удалённого соседа» (*D* – евклидово расстояние)

Выбор типа выводимого графика дендрограммы (горизонтальный или вертикальный) целиком зависит от автора исследования. Для этого достаточно в диалоговом окне результатов (рис. 16) выбрать *Horizontal...* или *Vertical...* Кроме того, существуют на выбор исследователя две модификации дендрограммы: прямоугольная *(Rectangular branches)* и более точная *(Scale tree...)*. Разбиения постов на однородные группы методом «удалённого соседа » и методом Уорда по многим позициям совпадают. Например. по обоим методам в одну группу попали р. Джебаш – с. Джебаш (объект С9), р. Она – п. Малый Анзас (объект С16), р. Большой Он – п. Большой Он (объект С5) и т.д.

В иерархических методах на дендрограммах, как правило, хорошо видно прекращение объединения в классы по довольно резко увеличившемуся расстоянию по оси y.

Исследователь может выбрать *пороговое значение* расстояния, на котором следует ограничить объединение. На рис. 18 можно ограничиться 3–4 шагами. Каждый узел на ветви – это результат очередного шага объединения.

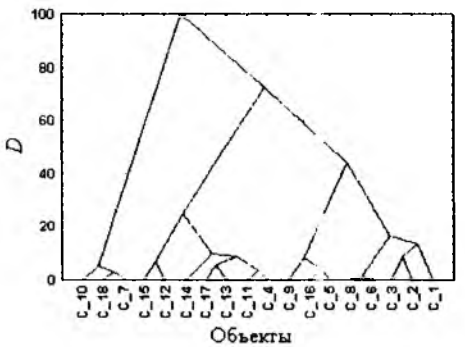

Рис. 18. Классификация рек по внутригодовому распределению стока методом Уорда (*D* – евклидово расстояние)

При применении метода Уорда уже при расстоянии 4 (рис. 19) образовалось 4 кластера. Посмотрим списочный состав кластеров на каждом шаге объединения в таблице (рис. 19).

Вызывается таблица из диалогового окна результатов кнопкой *Amalgamation schedule*.

В столбце *linkage distance* указаны расстояния, при которых происходит объединение объектов с символическими именами c_1, c_2... что означает «объект на строке1» и т.д. Из выделенных кластеров чёткую интерпретацию можно дать двум группам: реки первой относятся к степной зоне – это реки Тесь, Ерба, Уйбат, Бся, Табат; объекты второй – к району северо-западного склона хребта Западных Саян – реки Джебаш, Она, Большой Он. Данные две группы образуются при использовании всех вышеперечисленных методов. Реки выделенных групп не только близки по гидрологическому режиму, но и находятся в одной природной зоне. Последнее дает основания считать, что реки образуют районы с общими условиями формирования речного стока.

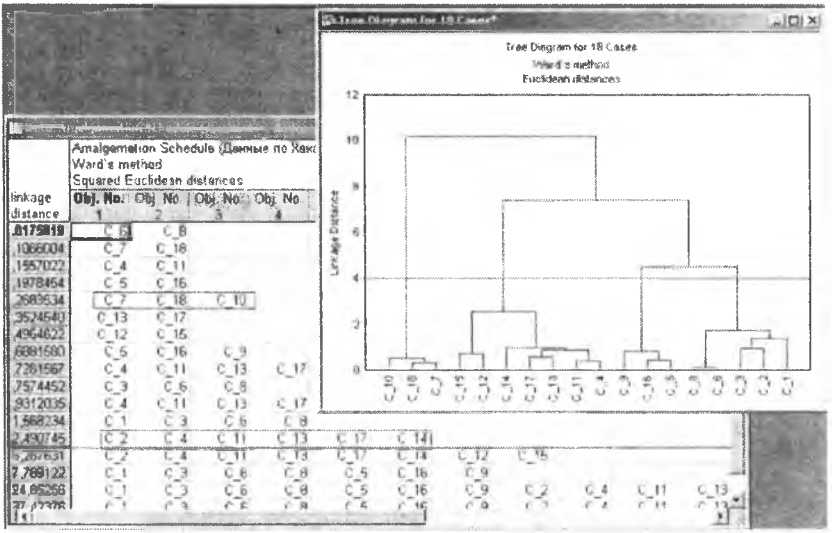

Рис. 19. Таблица со списком объединяемых объектов

При рассмотрении оставшихся кластеров возникают затруднения из-за несовпадения их составов для горных районов Кузнецкого Алатау и Абаканского хребта. Поэтому для содержательного анализа и принятия окончательного решения необходимо более детально рассмотреть данную территорию с учётом физико-географических характеристик, используя гипсометрическую карту, карту растительности, распределение увлажнённости по территории, геологическое строение, гидрологические показатели и т.д.

Пример 2. *Классификация рек Хакасии по внутригодовому распределению стока методом k-средних.* Решим задачу разбиения рек Хакассии на однородные группы по водному режиму методом ***k***-средних. Матрица данных та же, что и в предыдущем случае. В качестве объектов кластеризации назначаем строки. Рассмотрим только возможности, предоставляемые модулем для анализа результатов. Диалоговое окно результатов в методе ***k***-средних имеет несколько иной вид (рис. 20).

В функциональной части окна можно посмотреть расстояния от объектов до центра тяжести каждого кластера и матрицу расстояний между их центрами *(Cluster means & Euclidian distances)*, дисперсию признаков *(Analysis of variance)*, построить графики средних значений признаков для

каждого кластера *(Graph of means)*, получить описательную статистику для каждого кластера *(Descriptive statistics of each cluster)*, просмотреть состав и расстояния от объектов до центра тяжести каждого кластера *(Members of each cluster and distances)*.

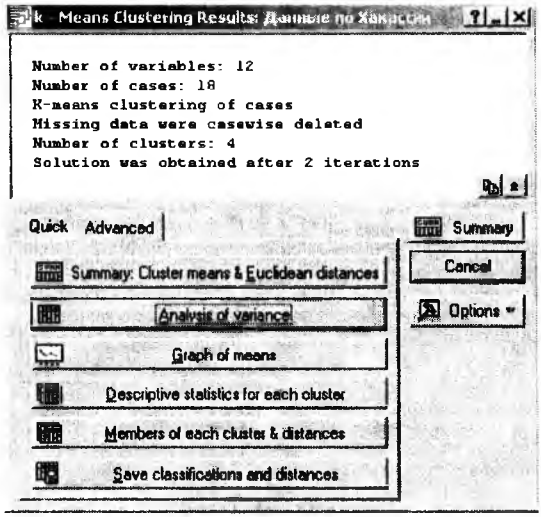

Рис. 20. Диалоговое окно результатов кластерного анализа методом *k*-средних

По графику средних визуально можно судить о качестве «расслоения» данных при заданном числе классов и соответственно этому корректировать решение о количестве кластеров. Иллюстрацией этого приёма служат приведённые на рис. 21 графики средних.

На формальном уровне этот пример очень показателен. Легко видеть, что кластеры *4* и *5* на верхнем графике синхронны, и их можно объединить в один. Аналогичная картина наблюдается у кластеров *2* и *3*. В результате деления на *3* кластера расслоение данных на группы чётче (нижний график).

Перенесём расстояния от объектов до центра тяжести каждого из трёх кластеров из таблицы, выдаваемой нажатием кнопки *Cluster means & Euclidian distances,* в ЭТ Excel и построим лепестковую диаграмму (рис. 22). У объектов первого кластера (ряд 1 на диаграмме) половодье

наблюдается в апреле, у второго кластера (ряд 2) – в мае–июне (в высокогорных бассейнах), у третьего кластера (ряд 3) – в мае.

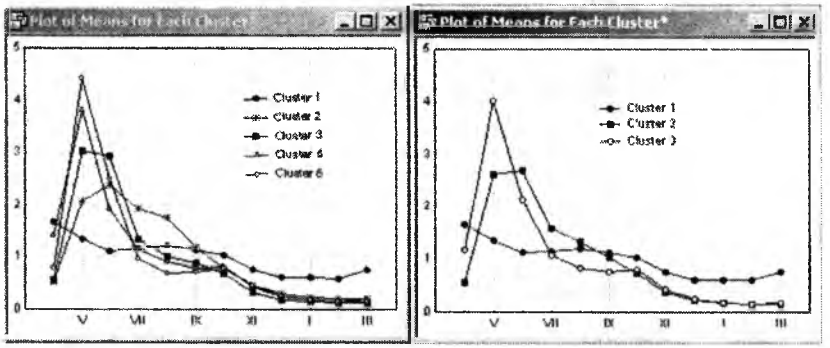

Рис. 21. Графики средних значений признаков в кластерах

Рис. 22. Лепестковая диаграмма среднемесячных расходов по каждому кластеру

Меню *Members of each cluster & distances* является обязательным для просмотра. Здесь содержится информация о метрических расстояниях внутри классов и само разбиение на классы.

Сравнение состава кластеров (рис. 23), полученных методами ***k***-средних и методом Уорда, позволяет сделать вывод о достаточно хорошем

совпадении результатов кластеризации. Полностью совпадают составы 2-х групп с очень тесными связями:

1) реки Тесь – с. Боград, Ерба – с. Ерба, Туи (С7, С10, С18) с малыми расстояниями 0,05–0,07;

2) реки Бея – с. Бея, Табат – с. Табат (С12, С15) с расстояниями 0,101.

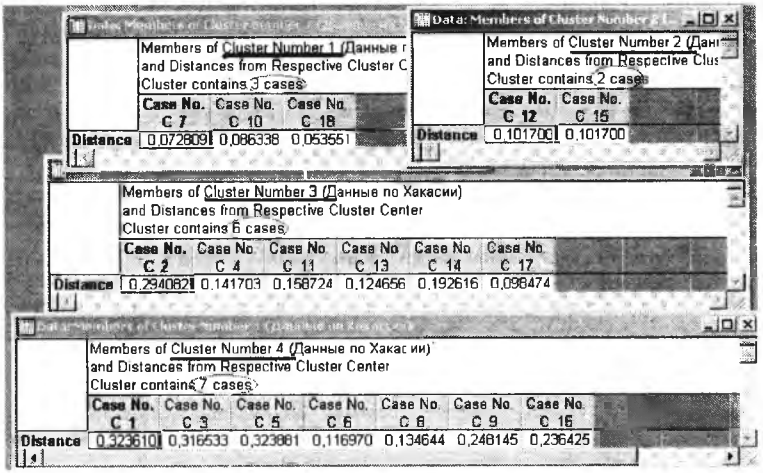

Рис. 23. Вид таблиц с информацией о кластерах: номер кластера (подчёркнут), символические имена объектов (С1 ...) и их расстояния до центра кластера (Distance)

Чем меньше расстояния, тем более компактную математическую область занимает класс и тем он статистически более устойчив.

В кластере *3* «лишним» является объект С2, в методе Уорда он попадал в кластер *4*. Действительно, из всех объектов третьего кластера он наиболее удалён от центра (расстояние 0,29), поэтому, возможно, лучше перенести С2 в кластер *4*. Однако такое решение должно быть подкреплено анализом физико-географических особенностей бассейна.

Естественно полагать, что степень компактности объектов в кластере отражает уровень близости объектов в признаковом пространстве.

В ряде задач удобно использовать в качестве меры близости объектов коэффициент корреляции, а точнее величину $1 - \text{Pearson}\ r$. Процедура применения кластерного анализа в этом случае имеет особенности в более удобном представлении дендрограммы. Таким образом, кластерный анализ обеспечивает относительно простой и прямой путь классифика-

ции объектов и позволяет представить результаты в удобном для понимания виде.

Пример 3. *Уточнение сезонов гидрологического года методом k-средних.* Рассмотрим задачу уточнения сезонов внутри гидрологического года по водному режиму рек Хакасии. Кластерным анализом также можно выделять сезоны внутри года (рис. 24), если в качестве объектов кластеризации выбрать признаки, расположенные в столбцах матрицы данных, т.е. характеристики стока. Иерархические методы хорошо выделяют все фазы водного режима, что немаловажно, так как отпадает необходимость в построении для этих целей гидрографов стока.

Сравнение результатов кластерного анализа с сезонами, выделенными на типовом гидрографе для рек равнинной Минусинской котловины, показало полное совпадение.

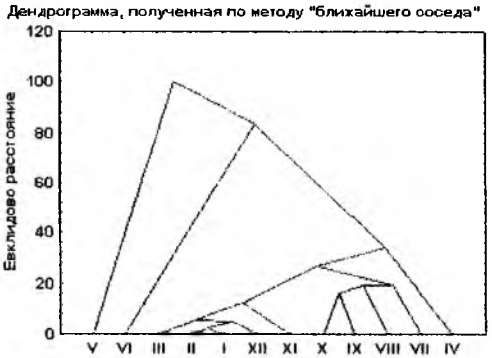

Рис. 24. Выделение сезонов водного режима

Выделены зимний (XI, XII, I, II, III месяцы) и летне-осенний (VII, VIII, IX, X) сезоны. Наиболее тесная связь наблюдается у зимних месяцев, менее связаны величины стока за весенние месяцы (IV, V, VI).

Вопросы для самопроверки

1. Дают ли различные алгоритмы кластерного анализа сходные результаты?
2. Дайте строгое определение понятию «кластер».
3. Как вы понимаете признаковое пространство?

4. Что собой представляет объект исследования?

5. Чем отличаются агломеративные методы кластерного анализа от дивизивных?

6. Поясните принцип работы итеративных методов кластерного анализа?

7. Как ещё называют неиерархические кластер-процедуры?

8. Объясните смысл термина «зашумление данных».

9. Какой алгоритм кластеризации наиболее устойчив к зашумлению данных?

10. Для чего проводится процедура нормирования исходных данных?

11. Что такое метрика?

12. Какие метрические расстояния Вы можете назвать?

13. В нижеприведённом списке укажите алгоритмы кластерного анализа, которые рекомендуется использовать в условиях малого числа признаков и большого количества классов:

– метод k-средних;

– метод ближнего соседа;

– метод дальнего соседа;

– центроидный метод;

– метод Уорда.

14. Какие графические возможности имеются в статистических пакетах для отображения результатов кластерного анализа?

15. Можно ли заранее задать количество кластеров?

16. Какие кластер-процедуры называются структурными – иерархические или неиерархические?

17. Поясните смысл термина «размерность признакового пространства».

18. Какой физический смысл имеет евклидово расстояние, если объекты характеризуются всего тремя признаками?

19. Верно ли утверждение «Кластер-анализ использует при разбиении на группы только несколько наиболее важных свойств объектов»?

20. Используют ли методы кластер-анализа, как и дискриминантный анализ, обучающую выборку?

21. Какие величины откладываются по осям дендрограммы?

22. К какому типу методов принадлежит метод k-средних?

23. Что такое эталон в методе k-средних?

24. Как можно оценить плотность класса?

25. Что такое функционал качества?

26. Какой должна быть дисперсия полученного класса, чтобы признать качество разбиения хорошим?

27. Каким должно быть расстояние между классами, чтобы признать качество разбиения хорошим?

28. Что такое пороговое значение?

29. Назовите алгоритмы, дающие сходные разбиения.

5. ФАКТОРНЫЙ АНАЛИЗ

Фактор (от *лат.* factor – делающий, производящий) – причина, движущая сила какого-либо процесса, явления, определяющая его характер или отдельные его черты.
Советский энциклопедический словарь

5.1. Общие положения факторного анализа

Факторный анализ (в широком смысле) – это совокупность моделей и методов, ориентированных на выявление и анализ внутренних причин (факторов), формирующих процесс, на основании информации об их внешних проявлениях в виде измеренных характеристик (признаков).

В основе различных моделей факторного анализа лежит следующая гипотеза: наблюдаемые или измеряемые свойства являются лишь косвенными характеристиками изучаемого объекта или явления, на самом же деле существуют внутренние (скрытые, неподдающиеся прямому измерению) параметры или свойства, число которых мало и которые определяют значения наблюдаемых признаков. Эти внутренние параметры принято называть факторами. Задача классического факторного анализа – представить наблюдаемые параметры в виде линейных комбинаций факторов и, может быть, некоторых дополнительных, «несущественных» величин – «помех». Таким образом, хотя сами факторы неизвестны, такое разложение может быть получено и, более того, такие факторы могут быть определены, т.е. для каждого объекта могут быть указаны значения каждого фактора (Девис, 1990).

Рассмотрим очень показательный пример (Благуш, 1989). Он взят из минералогии в силу очень удачной иллюстративности латентных переменных. Месторождение руды, содержащей олово и цинк, с высокой вероятностью определяется совместным действием *трёх* факторов: температурой, силой деформации подстилающего пласта и проницаемостью горной породы. Известно, что месторождение руды, о котором идёт речь, может образоваться там, где значения этих трёх показателей или в иных терминах – латентных (скрытых) переменных – комбинируются на определённых уровнях. В соответствующих единицах эти уровни характеризуются значением около 80 для температуры, 40 – для деформации и 50 – для проницаемости. Латентные переменные были потенциально измери-

мыми в далёком прошлом, ныне же они не подаются прямому измерению. Однако латентную температуру мы можем косвенно оценить по трём непосредственно измеряемым переменным, значения которых устанавливаются с помощью химического анализа, а именно по количеству Mg в кальците, Fe в сфалерите и Na в мусковите. О силе деформации подстилающего слоя можем косвенно судить по образованию оолитов, разломам и складчатости. Проницаемость горной породы можно оценить по количеству жил или обломков на 1 м^2. Всего, таким образом, набирается $n = 8$ непосредственно наблюдаемых переменных x_j.

При исследовании большого числа N образцов (например, $N > 20$) можно вычислить корреляционную матрицу размером 8×8 и с помощью факторного анализа установить существование $m = 3$ латентных переменных – общих факторов f_1, f_2, f_3, а их значения оценить по непосредственно измеряемым переменным. В тех случаях, когда все три полученных значения достаточно мало отличаются от установленных уровней $f_1 = 80$, $f_2 = 40$, $f_3 = 50$, можно с высокой вероятностью ожидать, что поиск рудного месторождения увенчается успехом. Обратите внимание – измеряемых признаков 8, а факторов – 3. По терминологии прикладной статистики произошло «сжатие» (снижение) размерности признакового пространства с 8 до 3.

Пример применения факторного анализа в метеорологии. В метеорологии существует огромное множество переменных, используемых для предсказания погоды, и они коррелированны между собой, что затрудняет применение регрессионного анализа и приводит к накоплению ошибок. Ортогонализация с помощью компонентного анализа (раньше факторный анализ считали частным случаем компонентного) позволяет в значительной мере обойти эти указанные проблемы.

Факторный анализ применяется также в химии, гидрологии, экологии и других областях знаний, хотя своим возникновением он обязан психологии и долгое время применялся и развивался для решения проблем психологии, педагогики, социологии и т.д. Рождение факторного анализа связано с появлением работ математика и философа К. Пирсона и психолога Ч. Спирмена. Карлу Пирсону принадлежит первая работа, посвящённая методу главных компонент (1901), который был разработан им для решения проблемы измерения общих факторов в антропометрии. Чарльз Спирмен в своей статье (1904) доказывал на примере тестов для обследования способностей учеников английских школ, что все способности человека объясняются одним главным фактором с неподдающимся точному определению свойством, которое пытались измерить английские

психологи – фактором интеллекта. Впоследствии в теории личности было доказано, что это не совсем так, и процесс становления личности многофакторный. В дальнейшем развитие факторного анализа связано с именами Л. Гутмана, Г. Хоттелинга, Л. Тэрстоуна, Г. Хармана и др.

В общих чертах можно выделить следующие *цели*, или направления *применения* факторного анализа:

а) понижение числа переменных;

б) классификация переменных (таксономия), обычно сочетаемая с введением *общих вторичных* переменных на основе укрупнения первичных;

в) косвенное оценивание переменных, не поддающихся непосредственному измерению;

г) исследование структуры исследуемой области;

д) преобразование исходных переменных к более удобному для интерпретации виду, осуществляемому обычно путём ортогонализации исходных переменных.

Все модели с *латентными* (ненаблюдаемыми, скрытыми, не поддающимися измерению) переменными можно разделить на 3 группы:

1) классический линейный факторный анализ;

2) нелинейный факторный анализ;

3) анализ латентной структуры.

В каждой группе имеется ряд моделей. В данном учебном пособии речь будет идти только о классическом (линейном) факторном анализе.

Следует особо отметить присущие факторному анализу в определённой степени трудности восприятия терминологии и путаницу в терминах, в значительной мере обусловленные его междисциплинарным происхождением, профессиональным жаргоном и англоязычными образцами. Эта особенность отмечается во многих серьезных изданиях по факторному анализу. В связи с этим ниже при введении нового термина будут даваться его синонимы. В факторном анализе часто даже сам термин «фактор» приводит к неверным интерпретациям. *Фактор* в смысле факторного анализа означает, прежде всего, *математически сконструированную переменную*, удовлетворяющую аксиомам факторной модели (Благуш,1989).

5.2. Модель классического линейного факторного анализа

Определим цель наших исследований как желание лаконично объяснить природу анализируемой многомерной структуры. Что понимается под термином «многомерная структура»? Речь идёт о множестве стати-

стически обследованных объектах Q_1, Q_2, ... Q_n, представленных в виде матриц вида «объект – признак», «признак – признак» или «объект – объект». Последние две могут быть матрицами корреляций, ковариаций или других мер сходства. Будем считать, что анализируемые данные представлены матрицей вида «объект – свойство» из n строк и m столбцов $[\mathbf{X}] = \{\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_m\}$, где $[\mathbf{X}_j] = \{x_{1j}, x_{2j}, \ldots x_{nj}\}$ – вектор-столбец значений j-го признака (свойства, переменой).

Линейная зависимость между переменными является полезным упрощением, облегчающим их исследование. Стремление «объяснить» *зависимую переменную* (в нашем случае вектор-столбец $[\mathbf{X}_j]$) с помощью комбинации нескольких слагаемых, т.е. *независимых переменных* $[f_1]$, $[f_2]$, ... $[f_k]$ приводит к многофакторным регрессионным уравнениям. Удобнее работать не с исходными (например, измеренными) переменными, а с их отклонениями от средних значений $x_{i,j} - \overline{x_j}$, или в векторной форме вектор-столбец $\left[\mathrm{X}_j\right] - \overline{\mathrm{X}}_j$, что значительно упрощает математические выражения. Кроме того, удобно предположить, что каждая независимая переменная f_k нормирована таким образом, при котором её дисперсия равна 1. Обозначим через $[\mathbf{Y}_j']$ регрессионную оценку зависимой переменной в отклонениях, тогда линейная зависимость, записанная в векторной форме, будет иметь вид:

$$[\mathbf{Y}_j'] = a_{1j}[f_1] + a_{2j}[f_2] + \ldots + a_{kj}[f_k],$$

где k – число факторов, причём, k много меньше m ($k << m$). Естественно, что регрессионная оценка получена с ошибками, т.е. $[\mathbf{y}_j] = [\mathbf{y}_j'] + [e_j]$, где e_j – называется отклонением от регрессии или остатком. Предполагается, что остатки распределены по нормальному закону с единичной дисперсией.

В общем виде модель линейного факторного анализа выглядит как обычное регрессионное уравнение (далее везде квадратные скобки для обозначения векторов и матриц опущены):

$$y_j = \sum_{i=1}^{k} a_{ij} f_i + e_j; \quad j = \overline{1,m}; \quad k << m.$$

Если в уравнении учитываются все факторы ($k = m$), то имеет место метод главных компонент или компонентный анализ. В этом случае каждый из наблюдаемых параметров $[\mathbf{y}_j]$ линейно зависит от m некоррелиро-

ванных между собой новых компонент (факторов) $f_1, f_2, \ldots f_m$. Каждая очередная компонента даёт возможно максимальный вклад в суммарную дисперсию параметров. В факторном анализе, в отличие от метода главных компонент, оставляют небольшое количество факторов ($k < m$), чтобы на их долю приходилось 70–80% суммарной дисперсии. Однако для точной аппроксимации корреляций между параметрами необходимы все компоненты, поэтому факторный анализ не применяется для восстановления или прогноза y_j.

Коэффициенты a_{ij} указывают «вес» или важность i-го фактора в определении j-й переменной и называются *факторной нагрузкой* i-го фактора на j-ю переменную. Другое название – *факторные коэффициенты*. Латентные независимые переменные образуют вектор f, который называется *общим фактором модели*. Точное название матрицы коэффициентов регрессии a_{ij} – матрица общих факторных коэффициентов, но на практике используют более короткое – *матрица нагрузок*. Сумму квадратов нагрузок $\sum_{i=1}^{k} a_{ij}^2$ называют *общностью*, соответствующей переменной y_j, и чем больше это значение, тем лучше описывается y_j факторами f_i. В свою очередь остатки e_j^2 показывают, какая часть дисперсии остаётся необъяснённой при используемом наборе факторов, и данную величину называют *характерностью (специфичностью)* переменной y_j. Таким образом, *дисперсия = общность + характерность* и вычисляется через факторные нагрузки следующим образом:

$$D_j = \sum_{i=1}^{k} a_{ij}^2 + e_j \quad j = 1, \ldots m.$$

Слагаемые в правой части этой формулы представляют собой вклады различных факторов в общую дисперсию j-й переменной.

5.3. Вычисление матрицы факторных нагрузок

Процесс нахождения факторов начинается с преобразования исходной матрицы данных в квадратную симметричную матрицу, которая выражает либо степень взаимосвязей между переменными, либо между объектами. Это делается путём умножения слева или справа матрицы исходных данных на транспонированную к ней.

Если матрица данных **X** состояла их n строк наблюдений (объектов) и m столбцов переменных (свойств объектов), то умножение слева на матрицу, транспонированную к ней, приводит к квадратной матрице $\boldsymbol{R}$ порядка $m \times m$: $\boldsymbol{R} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. Символом $\mathbf{X}^{\mathrm{T}}$ обозначена транспонированная матрица. Если данные были стандартизованы, то $\boldsymbol{R}$ будет корреляционной матрицей m переменных. Своим именем ($\boldsymbol{R}$) эта матрица обязана общепринятому обозначению коэффициента корреляции r.

Процедура стандартизации исходных данных очень важна как способ выравнивания влияния переменных, дисперсия которых мала. Например, если длины объектов заданы в сантиметрах, а ширина в миллиметрах, то переменная, представляющая ширину, будет оказывать в 10 раз большее влияние на результат, чем переменная, представляющая длину. Иногда предпочтительнее работать с оригинальными переменными и вследствие этого с ковариационной матрицей.

Если теперь матрицу данных **X** умножить справа на транспонированную к ней матрицу $\mathbf{X}^{\mathrm{T}}$, то получим квадратную симметричную матрицу $\boldsymbol{Q}$, которая имеет n строк и n столбцов: $\boldsymbol{Q} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. Если данные были стандартизованы, то $\boldsymbol{Q}$ будет ковариационной матрицей n объектов наблюдений. Обе матрицы характеризуют статистические свойства исходной матрицы **X**.

Далее находят собственные значения и собственные векторы корреляционной матрицы (см. раздел 1.3).

В факторном анализе используется следующее свойство собственных векторов матрицы – собственные векторы, принадлежащие различным собственным значениям, линейно независимы (ортогональны). Если $[\boldsymbol{S}]$ – корреляционная матрица, то собственные векторы этой матрицы являются искомыми факторами.

Все методы факторного анализа основаны на выделении собственных значений и собственных векторов из корреляционных матриц. *Собственные векторы корреляционной матрицы, умноженные на корень квадратный из соответствующего ему собственного значения, и образуют вектор f – искомые взаимно некоррелированные (т.е. ортогональные) факторы.* Собственные значения характеризуют относительный вклад каждого собственного вектора в общую дисперсию. Длина собственного вектора, представляющего фактор, равна собственному значению (вернее, корню квадратному из него), поэтому факторы отражают также дисперсию (точнее, стандартные отклонения).

Существует много методов выделения факторов, которые можно разбить на 2 группы в зависимости от выбранного критерия оптимальности.

Если в качестве критерия оптимальности используют минимум расхождения между ковариационной матрицей исходных признаков и той, которая получится после оценивания нагрузок, то приходят к методу главных компонент.

Если критерием оптимальности является максимальная близость исходных корреляций признаков к тем, которые получены в модели после оценивания нагрузок, то возникает факторный анализ.

В пакете Statistica, например, предлагаются метод главных компонент *(Principal components)* и 5 методов из группы анализа главных факторов (факторный анализ):

– *Communalities = multiple* R^2 – Общности равны квадрату коэффициента множественной корреляции (классический метод).

– *Iterated communalities MINRES* – Метод итеративных общностей (минимальных остатков).

– *Maximum likelihood factors* – Метод максимального правдоподобия.

– *Centroid method* – Центроидный метод.

– *Principal axis method* – Метод главных осей.

Математический аппарат этих методов достаточно сложен и поэтому здесь не рассматривается. О методе максимального правдоподобия достаточно просто на примерах можно прочитать в книге (Лоули, Максвелл, 1967).

Какой из этих методов лучше? Приведём высказывание на этот счёт одного из основоположников современного факторного анализа Г. Хармана «*Ни в одной из работ не было показано, что какой-либо один метод приближается к „истинным" значениям общностей лучше, чем другие методы. ...Выбор среди группы методов „наилучшего" производится в основном с точки зрения вычислительных удобств, а также склонностей и привязанностей исследователя, которому тот или иной метод казался более адекватным его представлениям об общностях*» (Харман, 1972).

Обратим внимание читателя на следующий важный момент – *число учитываемых в дальнейшем факторов исследователь задаёт сам* (кроме метода главных компонент, использующего все факторы). Предполагается, что исследователь до проведения анализа в состоянии предсказать число факторов, от которых зависит изучаемая модель, исходя из некоторых предварительных рассуждений. Теоретически эта проблема до сих пор не решена, хотя существует ряд рекомендаций:

а) оставлять только факторы, собственные значения которых больше единицы;

б) оставлять только те факторы, общий вклад которых в суммарную дисперсию исследуемых рядов достигает заранее заданной величины, например, 60–80% и т.д.

В дальнейшем будем называть эти факторы эффективными. В пакете Statistica в рамках первой рекомендации предлагается вспомогательный чрезвычайно наглядный график «каменистой осыпи» (рис. 30), на котором в порядке убывания изображены собственные значения корреляционной матрицы. Собственные значения малой величины, как «обломки» горных пород у подошвы горы, не участвуют в дальнейших исследованиях.

Однозначного решения системы уравнений (7) не существует, так как число объектов не равно числу признаков ($n \neq m$), поэтому представление корреляционной матрицы факторами (факторизацию) можно произвести бесконечным числом способов. Если получено решение одним из вышеперечисленных методов, то любое ортогональное преобразование *(ортогональное вращение)* матрицы нагрузок приведёт к той же факторизации.

5.4. Интерпретация матрицы факторных нагрузок. Требования простой структуры

Следующей проблемой факторного анализа является содержательная интерпретация полученной матрицы факторных нагрузок. Крайне редко решение, полученное вышеуказанными методами, позволяет сразу содержательно охарактеризовать факторы. Поэтому прибегают к процедуре вращения факторных осей в многомерном пространстве для получения «простой структуры» факторной матрицы.

В настоящее время существует довольно много процедур вращения факторных осей, но все они нацелены на выполнение основных критериев «*простой структуры*» Л. Тэрстоуна, суть которых сводится к следующим положениям (Структура и динамика... 1987):

1) для каждой строки матрицы факторных нагрузок хотя бы одно значение должно быть равным или близким к нулю;

2) столбцы факторной матрицы должны содержать не менее k близких нулю элементов, где k – учитываемые факторы;

3) в любой паре факторов должны существовать признаки, факторные нагрузки которых имеют максимальные нагрузки на один фактор и близкие к нулю на другой фактор, а также признаки, имеющие малые нагрузки на оба фактора;

4) в любой паре факторов должно иметься небольшое число признаков, значимо отличающихся от нуля.

5.5. Вращение осей

Используя различные процедуры вращения, довольно трудно добиться полностью выполнения вышеуказанных критериев, особенно при работе с сильно коррелированной совокупностью признаков, но этот этап построения факторной моделей представляется весьма полезным, так как существенно облегчает интерпретацию результатов. В идеале при вращении желательно получить такой вид, чтобы точки располагались на концах факторных осей (рис. 25) с небольшим количеством точек около нуля.

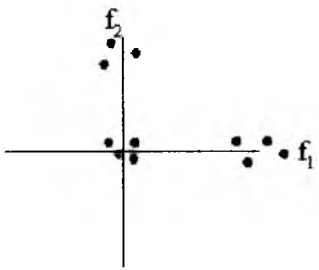

Рис. 25. Идеальное расположение факторных нагрузок для интерпретации в пространстве факторов f_1 и f_2, удовлетворяющее принципам «простой структуры»

Дадим краткую характеристику методам вращения для осмысленного их применения в конкретных задачах. Вращение методом *варимакс* Кайзера ставит целью упростить столбцы факторной матрицы, сводя все значения к 1 или 0. Вращение методом *квартимакс* ставит целью аналогичное упрощение, но по отношению к строкам факторной матрицы. *Эквимакс* занимает промежуточное положение – при вращении факторов по этому методу одновременно делается попытка упростить и столбцы, и строки.

У каждого из рассмотренных методов вращения осей есть нормализованный вариант *(normalized)*, в котором факторные нагрузки нормализуются, т.е. делятся на корень квадратный из соответствующей общности. Вращение осей не всегда облегчает анализ и в некоторых случаях даже

может приводить к дальнейшему ухудшению результатов. Последнее указывает на непригодность выбранной факторной модели.

Рассмотренные методы вращения относятся к ортогональным вращениям, т.е. в результате получаются некоррелированные факторы. Существуют ещё методы косоугольного вращения, приводящие к коррелированным между собой факторам. Эти методы есть в пакете SPSS, а в пакет Statistica они не включены.

5.6. Понятие о *R*- и *Q*-факторном анализе

В большинстве исследований число объектов значительно больше числа измеренных свойств, так что матрица $\boldsymbol{Q}$ значительно больше по размерам, чем матрица $\boldsymbol{R}$, несмотря на то, что они построены по одной исходной матрице **X** (рис. 26).

Соответственно получаемым матрицам методы факторного анализа делятся на два больших класса, называемых *R*- и *Q*-факторным анализом (Дэвис, 1990; Благуш, 1989). Первый связан с исследованием соотношений между переменными, второй – с исследованием отношений между объектами и часто используется при анализе внутренней структуры данных для представления в многомерном пространстве. В большинстве случаев факторный анализ проводится на корреляционной матрице $\boldsymbol{R}$.

При решении задач районирования (*Q*-факторный анализ) важны не сами значения факторов, а лишь взаимное расположение исследуемых рядов в k-мерном пространстве факторов.

Если взять сумму произведений факторных нагрузок на стандартизованные значения признаков по каждому наблюдению (объекту), то можно получить значение проекций наблюдений на факторы. Анализ расположения точек проекций наблюдений в плоскостях первых факторов позволяет выявить группировки, которые могут быть положены в основу районирования. Однородным классам многомерных наблюдений будут соответствовать относительно изолированные компактные группы точек в факторном пространстве. Степень компактности точек будет проявляться тем ярче, чем большая часть общей дисперсии признакового пространства будет приходиться на первые факторы (Структура и динамика... 1987).

Например, в задаче районирования рек бассейна р. Катуни по условиям формирования поверхностного стока проекции многомерных наблюдений на факторные оси изображены на рис. 27.

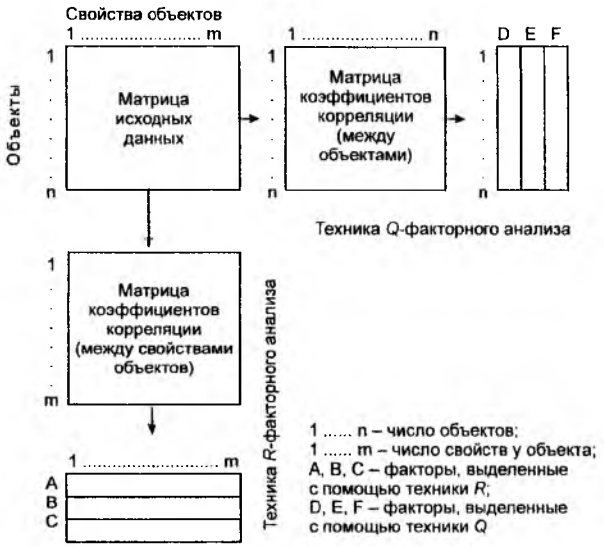

Рис. 26. Схема техник *R*- и *Q*-факторного анализа (Иберла, 1980)

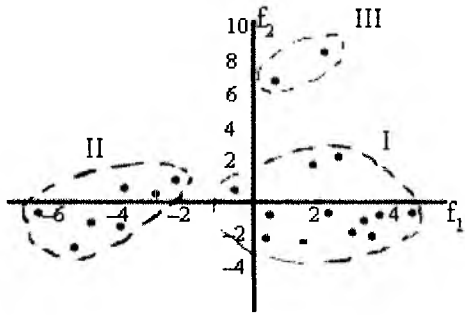

Рис. 27. Проекции многомерных наблюдений в плоскости первых двух факторов

Вся совокупность точек (объекты – бассейны) разбилась на три группы, которые по территориальным принадлежностям соответствуют различным ландшафтным зонам. Полученное разбиение можно рассматри-

вать как основу для последующего районирования территории по условиям формирования поверхностного стока.

5.7. Этапы факторного анализа

Можно выделить 5 этапов в осуществлении процедуры факторного анализа (Благуш, 1989).

1. *Формирование цели.* Можно выделить цели двух типов:

а) исследовательские, ориентированные на выявление и анализ латентных переменных, определяющих наблюдаемые признаки;

б) прикладные, связанные с построением агрегированных характеристик, необходимых для дальнейшего использования в задачах прогнозирования и управления.

В первом случае неизвестны ни количество факторов, ни структура их связи с признаками – они-то и определяются в процессе исследования. Во втором случае, исходя из результатов, полученных на других данных, другими исследователями, или из содержательных соображений, исследователь формирует гипотезу о количестве факторов и их связи с признаками, а потом проверяет её на реальных данных.

2. *Выбор объектов и признаков.* Это один из самых ответственных этапов, в значительной степени влияющий на результат. Следует тщательно проанализировать и обосновать выбор совокупности признаков и представительность множества объектов. Как правило, это удаётся сделать достаточно корректно только на основе многократной обработки данных.

3. *Получение матрицы факторных нагрузок.* Очень популярный некогда центроидный метод в настоящее время используется редко. Наиболее распространёнными являются алгоритмы аппроксимационного подхода. Метод главных компонент – аппроксимационный метод оценки параметров модели с помощью последовательного «отщепления» факторов, в соответствии с принципом: «снимаемая» на данном шаге компонента должна учитывать как можно большую долю «необъяснённого» разброса данных.

В определённой степени альтернативным предыдущему является метод максимального правдоподобия, разработанный Дж. Лоули в 1940 г. В этом методе можно избежать априорного задания числа факторов и, кроме того, факторы оказываются свободными от смещения, присущего более простым методам.

4. *Вращение факторной структуры.* Вращение факторной структуры осуществляется главным образом в связи с реализацией исследователь-

ских целей. При проверочном анализе критерием качества вращения является соответствие факторной структуры той, которая задана исследователем. При разведочном анализе требуется достижения «простой структуры», когда нагрузки максимально поляризованы между нулём и единицей. В целом решение проблемы вращении находится в начале своего развития. Обоснованные критерии ещё только предстоит разработать.

5. *Интерпретация и использование решения.* Рекомендаций общего плана здесь немного. В основном исследователь получает их на примерах решения задач в интересующей его конкретной предметной области.

5.8. Примеры использования факторного анализа при решении задач прикладных исследований

Пример 1. *Факторная модель процессов формирования поверхностного стока (на примере р. Катуни).* В основе построения факторной модели формирования поверхностного стока лежат следующие основные теоретические предпосылки:

1) процесс формирования поверхностного стока представляет совокупность различных состояний природной обстановки, но определяется относительно небольшим числом существенных характеристик (признаков);

2) признаки, оказывающие наибольшее влияние на формирование поверхностного стока, так же изменчивы во времени и в пространстве, как и сам поверхностный сток;

3) число признаков должно быть значительно больше количества факторов; признаки распадаются на группы, связанные с факторами (Структура и динамика... 1987).

В приводимом примере будут рассмотрены как R-, так и Q-факторный анализ. Результатом проведения первого будет установление иерархии в условиях формирования поверхностного стока, а второго – разбиение бассейнов рек на группы в качестве основы для последующего районирования.

Для построения факторной модели использовались 23 речных бассейна с разнообразными условиями формирования поверхностного стока, в различных ландшафтных зонах бассейна р. Катуни. В факторную модель было введено 14 различных характеристик речных бассейнов, которые можно условно разделить на несколько групп: геоморфологические (гидроморфологический коэффициент (ГМК), км/(м^3/с); геоморфологический фактор стока; частота потоков; густота речной сети, км/км^2), мор-

фометрические (площадь водосбора, км^2; средняя высота бассейна, м; средний уклон водосбора, ‰; средний уклон русла, ‰; расстояние от наиболее удаленной точки бассейна до замыкающего створа S, км), гидрометеорологические (осадки, мм; испарение, мм; средний многолетний расход воды Q, м^3/с).

Гидроморфологический коэффициент – это отношение суммарной длины речной сети к среднему многолетнему расходу воды: $\gamma_k = \sum l_i / \overline{Q}$. Он характеризует длину речной сети, необходимую для формирования среднего многолетнего расхода воды в 1 м^3/с (Горошков, 1979). Коэффициент ГМК предложен И.Н. Гарцманом, чаще используется в индикационных исследованиях.

Геоморфологический фактор стока Ф характеризует крутизну и расчлененность рельефа на водосборе и определяется по формуле

$$\Phi = l / \sqrt{I},$$

где l – длина склонов, м; I – уклон склонов, ‰. Коэффициент предложен А.Н. Бефани (Горошков, 1979).

Такой подход к выбору данных обеспечивает достаточную степень объективности использования метода факторного анализа.

На первом этапе исследований необходимо сформировать матрицу исходных данных. В случае, когда исследуется таблица многомерных данных, может иметь место наличие дублирующих и малоинформативных признаков. При прочих равных условиях, чем больше взято признаков для анализа, тем хуже, ибо каждый признак, не несущий полезной информации, только создает помехи для анализа – создает «шум», в котором теряется полезный сигнал. Для отбора независимых признаков исследовалась матрица коэффициентов корреляции и, принимая во внимание, что мерой относительной информативности признака могут служить показатели его вариабельности, оставлялись признаки с бóльшими значениями коэффициента вариации C_v.

Окончательно был выявлен комплекс из 11 наиболее представительных признаков. Характеристики бассейнов – средний уклон водосбора, расстояние от наиболее удаленной точки бассейна до замыкающего створа и геоморфологический фактор стока были удалены из матрицы, как дублирующие.

Коэффициенты при общих факторах позволяют учесть бóльшую часть суммарной дисперсии параметров, если:

а) показатели распределены по нормальному закону,

б) показатели линейны относительно друг друга.

Эти достаточно жёсткие требования полностью выполняются редко, но проверка их является обязательной, так как уже по её результатам можно предвидеть степень успешности модели. Требование линейности негласно принято смягчать монотонностью. Первый этап исследований можно проводить и в ЭТ Excel, хотя в пакете Statistica заложены все вышеперечисленные подготовительные операции над данными.

Так как признаки разнородны, то исходные данные были стандартизованы. В пакете Statistica этой процедуре можно подвергнуть достаточно просто всю матрицу сразу в меню Правка *(Edit → Fill/Standardize Block → Standardize Columns/Rows)*.

На втором этапе построения факторной модели в пакете Statistica необходимо сформировать условия анализа (рис. 28):

1) Выбрать ***R***- или ***Q***-модель. Выбор зависит от вида таблицы с данными. В пакете Statistica под *Variables* всегда подразумеваются столбцы матрицы. Если в исходной матрице наблюдений *(Raw Data)* в столбцах размещены свойства объектов наблюдений, то выбрана ***R***-модель, в противном случае – ***Q***-модель. По матрице исходных данных будет рассчитываться корреляционная матрица. Можно исходные данные сразу подать модулю в форме корреляционной матрицы соответствующего размера.

2) Указать столбцы матрицы, участвующие в анализе *(Variables)*.

3) Указать число учитываемых факторов *(Max. no. of factor)*, если оно априори известно, или выбрать условие ограничения числа учитываемых факторов. Обычно накладывают условие на величину собственных значений корреляционной матрицы – учитывается в модели столько факторов, сколько собственных значений превышают единицу (задаётся в поле *Minieigenvalue*).

4) Выбрать метод вычисления факторных нагрузок *(Extraction method)* (рис. 28).

В зависимости от критерия оптимальности (дисперсии или корреляции) возможен либо анализ методом главных компонент, либо одним из методов:

1) главных осей;

2) центроидный метод;

3) минимальных остатков;

4) общности равны квадрату коэффициента множественной корреляции;

5) метод максимального правдоподобия.

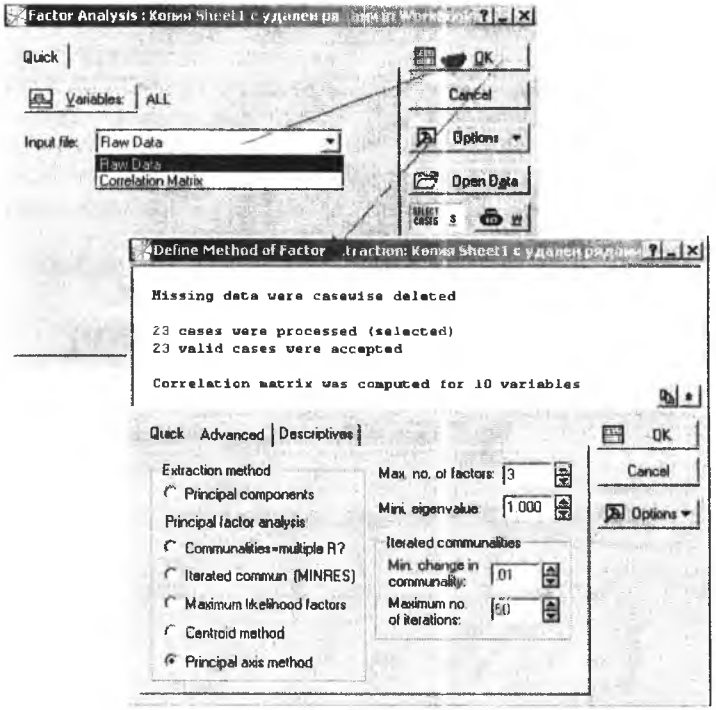

Рис. 28. Выбор настроек для модели факторного анализа

В рассматриваемом примере факторная модель была построена методом главных осей, применялось вращение осей *Quartimax*.

На третьем этапе анализируются результаты полученной модели. По таблице с собственными значениями определяется суммарный процент дисперсии исходных рядов. В рассматриваемом примере три первых фактора объясняют 64,26% дисперсии исходных рядов, как видно из таблицы собственных значений *(Eigenvalues)* на рис. 29.

По графику «каменистой осыпи» (рис. 30) можно уточнить число учитываемых факторов и повторить анализ. По графику видно, что можно выделить три главных фактора с собственными значениями, превышающими единицу. Построить этот график можно, выбрав кнопку *Screen plot* в диалоговом окне результатов (рис. 29).

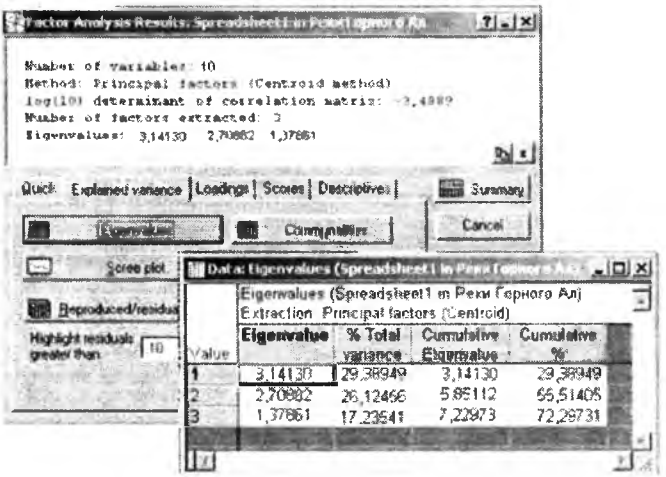

Рис. 29. Таблица собственных значений корреляционной матрицы

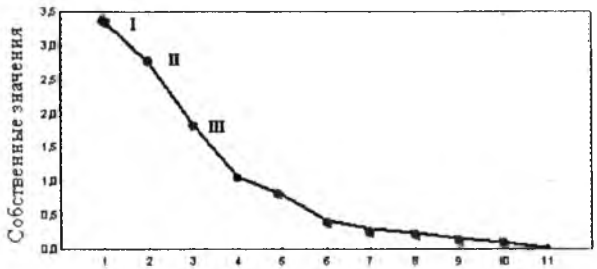

Рис. 30. График «каменистой осыпи» для выбора числа эффективных факторов I–III

Однако самым важным моментом этого этапа является анализ матрицы факторных нагрузок *(Loadings)*. Обычно рекомендуется интерпретировать нагрузки, превышающие 0,5. Дадим интерпретацию факторам по таблице нагрузок (табл. 13). Факторная модель описывает общую дисперсию полей рассчитанных параметров с помощью трёх факторов. Вклад каждого из них в общую дисперсию составляет: первого – 29%, второго – 26%, третьего – 17%, в сумме – 72%. Относительно равномер-

ное распределение нагрузок свидетельствует о многофакторности процесса формирования речного стока.

На первый фактор приходится большая часть дисперсии. В его формировании участвует значительная часть изменяющихся параметров. Наибольшую нагрузку на него имеют гидрометеорологические параметры: испарение (r = 0,92), осадки (r = 0,86). Менее значимые нагрузки несут морфометрические характеристики фактора: уклон склонов водосбора (r = –0,69), средняя высота водосбора (r = –0,57).

Таким образом, исходя из анализа факторной матрицы, расчета процентной доли параметров, наполняющих фактор, а также из учета коэффициентов корреляции между параметрами, первый фактор можно рассматривать как фактор комплексного воздействия характеристик, отражающих условия питания (увлажнение, испарение), движения (морфометрические параметры) речных вод.

Таблица 13

Матрица факторных нагрузок, полученная методом главных осей

Признак	I фактор	2 фактор	3 фактор
Средний многолетний расход $Q_{ср}$, м^3/с	0,27	**0,84**	–0,04
Площадь водосбора F, км^2	–0,66	–0,15	–0,25
Осадки X, мм	**0,86**	0,31	0,03
Испарение E, мм	**0,92**	0,27	0,13
Лесистость, %	0,49	–0,34	–0,46
Средняя высота водосбора h, м	–0,57	**0,69**	–0,13
Уклон речных русел I, ‰	–0,09	**0,82**	0,06
Уклон склонов водосбора I, ‰	–0,69	0,30	0,15
Частота речной сети	0,20	0,46	0,57
ГМК (гидроморфол. коэф-т)	–0,06	–0,61	**0,69**
Густота речной сети ρ, км/км^2	0,16	0,09	**0,86**
Вклад факторов в общую дисперсию	29%	26%	17%

Примечание. Факторный анализ применялся к стандартизованным данным

К изложенному следует добавить, что нагрузки метеорологических параметров входят в первый фактор с положительным знаком, а морфометрические – с отрицательным, т.е. связаны между собой обратной зависимостью.

Параметрами, имеющими наибольшие нагрузки на второй фактор, являются средний многолетний расход (r = 0,84) и морфологическая характеристика – уклон речных русел (r = 0,82).

В третьем факторе максимальные нагрузки имеют гидроморфологический коэффициент ГМК (r = 0,69), густота речной сети (r = 0,86). Интерпретировать его можно как влияние геологических условий.

Заключительным этапом анализа является проверка адекватности модели. Факторный анализ в отличие от кластерного имеет средства для оценки адекватности модели с помощью воспроизведенной матрицы корреляций и матрицы остаточных корреляций. Последняя получается поэлементным вычитанием воспроизведенной матрицы из исходной.

Стандартным образом проводится оценка значимости остаточных корреляций. Отсутствие значимых остаточных корреляций может рассматриваться как признак адекватности модели. Как видно на рис. 31, «остаточные» коэффициенты корреляции малы и не превышают 0,29. Красным отмечены коэффициенты корреляции, превышающие 0,1. По диагонали матрицы расположены средние квадратические отклонения соответствующих рядов.

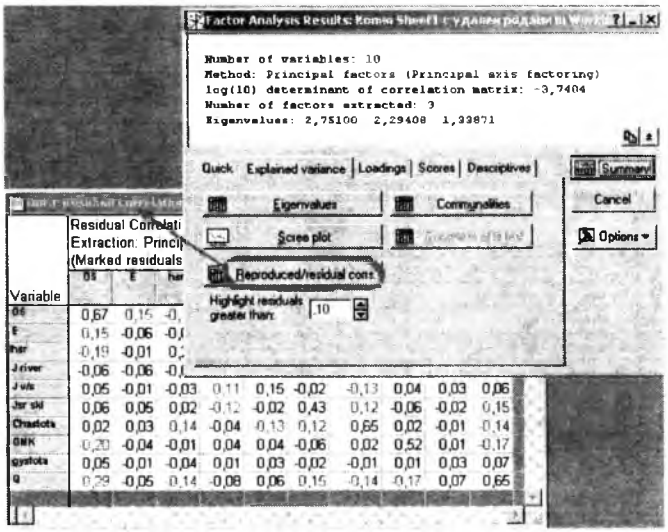

Рис. 31. Вызов матрицы остаточных корреляций (*Residual corelation*)

При анализе факторных и корреляционных матриц, в соответствии с рекомендациями Крамбейна и Грейбилла, связи считаются сильными, если коэффициенты корреляции r превышают 0,7, умеренными – при r от 0,4 до 0,7. Значения $r < 0,4$ говорят о слабых связях или их отсутствии.

Недостатки модели:

1) не всегда выполняется линейная зависимость между признаками;

2) не всегда выполняется нормальный закон распределения;

3) не учитывались средние летние температуры;

4) не было данных по подземному стоку.

Применение факторного анализа для построения модели формирования стока позволило оценить структуру взаимосвязей в общем процессе формирования стока.

По анализу корреляционной и факторной матриц можно судить о том, что факторная модель в целом отражает реальные связи между отдельными признаками в общей картине условий формирования речного стока.

Пример 2. *Факторная модель состава торфов олиготрофных и эвтрофных болот Западной Сибири.* Применим метод факторного анализа к исследованию природы торфов Западной Сибири как многомерных структур для выявления их внутренних связей.

Влияние болотных биогеоценозов и их компонентов на общую физико-географическую обстановку Западной Сибири огромно. Отметим, в частности, качественную сторону гидрологической роли болот, являющихся сложными комплексными ландшафтно-геохимическими барьерами, способствующими понижению pH, жесткости и концентрации главных ионов в речных водах, стекающих с заболоченных водосборов, увеличению содержания органического вещества.

Геохимические и климатические особенности территории, несомненно, накладывают отпечаток на состав и структуру органического вещества западносибирских торфов вследствие особенностей растительного покрова, а также замедления биологических процессов, обусловленного суровостью климатических условий.

Обработке подвергалось 140 образцов торфов по 35 параметрам, характеризующим фракционно-групповой состав углерода, азота, содержание подвижных форм основных питательных элементов и поглощающий комплекс торфа. В случаях, когда исследуется таблица многомерных данных, полученных экспериментально или агрегированием возможно наличие дублирующих и малоинформативных признаков.

Для отбора независимых параметров исследовалась матрица коэффициентов корреляции и из относительно однородных групп параметров, несущих информацию об одном свойстве исследуемого объекта, оставлялось по одному представителю, обладающему большей вариабельностью, оцениваемой коэффициентом вариации C_v.

Выбраковка нестандартных наблюдений, «отскакивающих» по терминологии Дж. Тьюки (1981) осуществлялась при превышении ими $2h$, где h – межквартильное расстояние выборки. Для успешного применения

факторного анализа должны выполняться два основных предположения относительно данных: независимость (или линейность взаимных связей с небольшими коэффициентами корреляции) и соответствие нормальному закону распределения. Поскольку эти требования накладывает сильное ограничение на круг решаемых задач, то на практике выработалось соглашение о допустимости применения факторного анализа для данных с монотонно изменяющимися связями и распределением, не очень «заметно» отклоняющимся от нормального. Нормальность можно проверить визуально по графику на нормальной вероятностной бумаге (рис. 32, *а*) и по критериям Колмогорова – Смирнова и Шапиро – Уилкса, монотонность по матричному графику рассеяний (рис. 32, *б*).

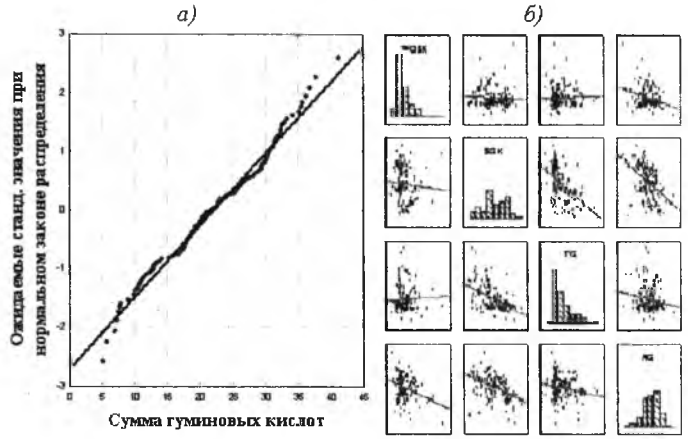

Рис. 32. Нормальный вероятностный график для суммы гуминовых кислот (*а*) и фрагмент матричного графика корреляций (*б*)

В результате такой процедуры для включения в дальнейший анализ было отобрано 8 относительно независимых параметров из 35: сумма гуминовых кислот (∑ГК), содержание углеводов (ЛГ и ТГ), негидролизуемый остаток (НГ), отношение углерода к азоту (C/N), сумма минерального и легкогидролизуемого азота (Нм + Нлг), степень разложения (*R*), зольность (*A*), содержание липидов (*W*). Первые пять параметров характеризуют с разных сторон биохимическую устойчивость торфов. Степень разложения и зольность торфа отражают конечный результат торфообразования, а содержание липидов рассматривается как признак, на-

следуемый торфами от растений торфообразователей и предположительно влияющий на состояние органического вещества торфа.

Таблица 14

Корреляционная матрица свойств торфов

	W	∑ГК	ТГ	НГ	C/N	*R*	*A*	Nлг + Nм
W	1,00	–0,05	0,10	–0,35	0,45	–0,17	–0,039	0,02
∑ГК	–0,5	1,00	–0,54	–0,50	–0,39	0,41	0,013	0,04
ТГ	0,10	–0,54	1,00	–0,18	0,37	–0,40	–0,17	0,17
НГ	–0,35	–0,50	–0,18	1,00	–0,02	–0,05	0,10	–0,21
C/N	0,45	–0,39	0,37	–0,02	1,00	–0,41	–0,48	0,36
R	–0,17	0,41	–0,40	–0,05	–0,41	1,00	0,46	–0,13
A	–0,39	0,13	–0,17	0,10	–0,48	0,46	1,00	–0,11
Nлг + Nм	0,02	0,04	0,17	–0,21	0,36	–0,13	–0,11	1,00

Будем решать эту задачу методом главных компонент (рис. 33). Главные компоненты – это собственные векторы ковариационной матрицы.

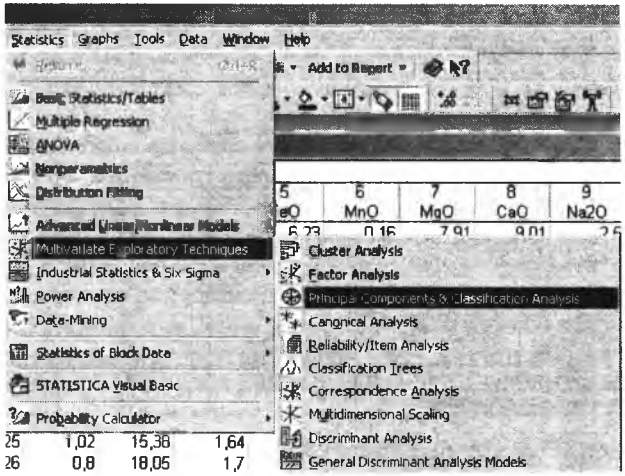

Рис. 33. Вызов модуля, реализующего метод главных компонент

Диалоговое окно результатов (рис. 34) в информационной части сообщит, что в системе из 8 переменных и 140 наблюдений выявлены три главные компоненты с собственными значениями, превышающими единицу (3,64 и 2,51; 1,78), и описывающие 69,3% изменчивости исходных данных.

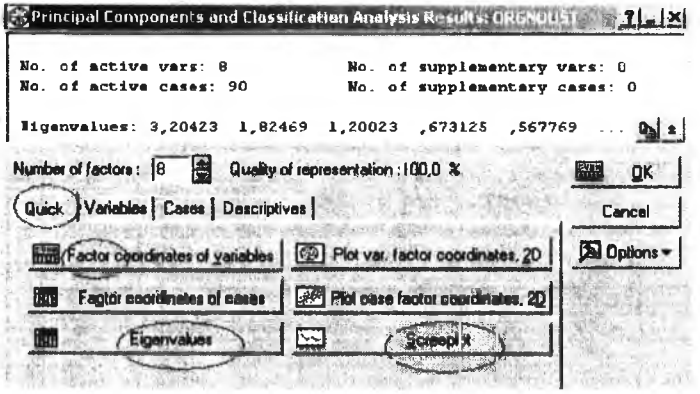

Рис. 34. Диалоговое окно метода главных компонент

Матрицу факторных нагрузок для 8 факторов (табл. 15), можно вызвать на экран кнопкой *Factor coordinates of Variables*, собственные значения – кнопкой *Eigenvalues*.

Таблица 15

Фрагмент матрицы факторных нагрузок

Параметры	Фактор 1	Фактор 2	Фактор 3
W	**–0,82**	0,18	–0,03
∑ГК	0,11	**0,91**	0,04
ТГ	–0,18	–0,50	0,57
НГ	0,32	–0,67	–0,45
C/N	–0,64	–0,36	0,37
R	0,48	0,53	–0,2
A	**0,79**	0,14	0,00
Nлг + Nм	–0,02	0,07	**0,83**
Вклад в суммарную дисперсию, %	26,4	25,1	17,8

Кроме того, можно построить *графики разброса переменных в виде точек в пространстве двух факторов* (кнопка *Plot var. factor coordinates*), выбор которых предоставляется исследователю. Конечно, наибольший интерес представляет разброс переменных в координатных осях первых двух факторов, так как они объясняют большую часть дисперсии исходных данных. Аналогичные графики можно построить и для объектов (кнопка *Plot case factor coordinates* в окне результатов).

Обратимся к интерпретации матрицы нагрузок главных компонент (табл. 15) и таблице собственных значений (табл. 16). Первая главная компонента определяет 26,4% изменчивости в исследуемой системе. Высокие нагрузки на первую компоненту имеют зольность (0,79) содержание липидов (–0,82).

Таблица 16

Собственные значения главных компонент

	Собственные значения	Суммарная дисперсия, %	Накопленная дисперсия, %
1	2,6455809	**26,45580**	26,4558
2	2,5161047	**25,16104**	51,6168
3	1,789080	17,89080	69,5076
4	0,881254	8,81254	78,3201
5	0,796214	7,96214	86,2823
6	0,785071	7,85071	94,1330
7	0,292253	2,92253	97,3555
8	0,264451	2,64451	100,0000

Вторая главная компонента менее значима, но она содержит высокую положительную нагрузку суммы гуминовых кислот (0,91) – самую высокую в таблице.

В третьей главной компоненте большую нагрузку имеет общее содержание легкогидролизуемого и минерального азота (0,83). Три первых фактора обеспечивают 69,3% суммарной дисперсии. Пороговая величина значимой факторной нагрузки в пакете «Statistica» установлена равной 0,7. В связи с этим, анализируя нагрузки второго фактора, можно сделать вывод, что при незначительной разнице в величине объясняемой дисперсии (25,1 против 26,4%), параметр ∑GK можно считать наиболее информативным (факторная нагрузка – 0,91). Третий фактор даёт небольшой вклад в суммарную дисперсию (17,8%), но лучше первых двух удовлетворяет принципу простой структуры.

Таким образом, можно установить иерархию в структуре анализируемых данных по степени их информативности:

1) сумма гуминовых кислот;

2) липиды, зольность;

3) сумма минерального и легкогидролизуемого азота.

Так как большинство параметров, включённых в анализ, имеют отношение к биохимической устойчивости торфов, то параметр ∑ГК может быть выбран в качестве базового для классификации.

Оценка адекватности данной модели может быть выполнена по признаку отсутствия значимых остаточных корреляций, полученных в результате вычитания воспроизведённой матрицы корреляций из исходной, так как это было изложено в предыдущем примере.

Вопросы для самопроверки

1. Как называются новые переменные, которые позволяет выявить факторный анализ?
2. Какие методы выделения факторов реализованы в пакете Statistica?
3. Является ли метод главных компонент одним из методов факторного анализа?
4. Можно ли отнести факторный анализ к методам сокращения признакового пространства?
5. Факторный анализ относится к группе методов (выберите правильный ответ):
а) монотетической классификации;
б) политетической классификации;
в) автоматической классификации.
6. Каким может быть количество выделенных факторов при проведении факторного анализа?
7. Начальным этапом факторного анализа является (выберите один правильный ответ из предложенных):
а) вычисление мер центральной тенденции;
б) вычисление показателей асимметрии и эксцесса;
в) построение корреляционной матрицы;
г) вычисление коэффициентов взаимной сопряжённости.
8. Объясните значение термина «простая структура».
9. Перечислите свойства «простой структуры».
10. Что Вы знаете о графике «каменистая осыпь» (вид и назначение)?
11. Дайте определение собственного вектора матрицы.
12. Зачем в факторном анализе вычисляют собственные векторы и собственные значения?
13. Какую цель преследует процедура вращения факторных осей?
14. Какие методы вращения факторной матрицы вам известны?
15. Какие типы исходных данных возможны в факторном анализе?
16. При проведении факторного анализа количество выделенных факторов может быть любым?
17. Что означает термин «матрица воспроизведённых корреляций»?
18. Что такое «матрица остаточных корреляций» и как она вычисляется?
19. Какую величину в факторном анализе называют «общностью»?
20. Какую величину в факторном анализе называют «характерностью»?

ЛИТЕРАТУРА

Основная литература

1. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. М.: Статистика, 1974. 200 с.

2. *Алексеев Г.А.* Объективные методы выравнивания и нормализации корреляционных связей. Л.: Гидрометеоиздат, 1971. 362 с.

3. *Алёхин Ю.М.* Статистические прогнозы в геофизике. Л.: Изд-во ЛГУ, 1963. 86 с.

4. *Белоцерковский А.В.* Спектральный анализ в гидрометеорологии: Учеб. пособие. СПб.: Изд-во РГМИ, 1993. 64 с.

5. *Берестнева О.Г., Муратова Е.А., Уразаев А.М.* Компьютерный анализ данных. Томск: Изд-во ТПУ, 2003. 204 с.

6. *Бефани Н.Ф., Калинин Г.П.* Упражнения и методические разработки по гидрологическим расчетам. Л.: Гидрометеоиздат, 1965. 427 с.

7. *Благуш П.* Факторный анализ с обобщениями. М.: Финансы и статистика, 1989. 248 с.

8. *Бокс Дж., Дженкинс Г.* Анализ временных рядов. Прогноз и управление / Пер. с англ. М.: Мир, 1974. 406 с.

9. *Бондаренко В.Н.* Статистические решения некоторых задач геологии. М.: Недра, 1970.

10. *Боровко Н.Н.* Статистический анализ пространственных геологических закономерностей. М.: Недра, 1971.

11. *Боровиков В.П.* Statistica. Искусство анализа данных на компьютере: Для профессионалов. СПб.: Питер, 2003. 688 с.

12. *Боровиков В.П., Боровиков И.П.* Statistica. Статистический анализ и обработка данных в среде Windows. М.: Инф.-изд. дом «Филинъ», 1997. 608 с.

13. *Горелова Г.В., Кацко И.А.* Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: Учеб. пособие для вузов. Ростов н/Д: Феникс, 2005. 480 с.

14. *Горошков И.Ф.* Гидрологические расчеты. Л: Гидрометеоиздат, 1979. 431 с.

15. *Гублер Е.В.* Алгоритм оценки расхождения распределений признаков в медицинских автоматизированных системах // Проблемы системотехники и автоматизированные системы управления. Л.: Медицина, 1978. 230 с.

16. *Дружинин В.С., Сикан А.В.* Методы статистической обработки гидрометеорологической информации. СПб.: Изд-во РГМУ, 2001. 168 с.

17. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы: Учеб. М.: Финансы и статистика, 1998. 352 с.

18. *Дэвис Дж.С.* Статистический анализ данных в геологии / Пер. с англ. В.А. Голубевой; Под ред. Д.А. Родионова. М.: Недра, 1990. Кн. 1. 319 с.

19. *Дэвис Дж.С.* Статистический анализ данных в геологии / Пер. с англ. В.А. Голубевой; Под ред. Д.А. Родионова. М.: Недра, 1990. Кн. 2. 427 с.

20. *Дюк В.* Обработка данных на ПК в примерах. СПб.: Питер, 1997. 240 с.

21. *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. М.: Статистика, 1977. 143 с.

22. *Енюков И.*С. Методы и алгоритмы многомерного статистического анализа. М.: Финансы и статистика, 1986. 232 с.

23. *Иберла К.* Факторный анализ / Пер. с нем. В.М. Ивановой. М.: Статистика, 1974. 398 с

24. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика: Учеб. пособие для втузов. М.: Высшая школа, 1992. 304 с.

25. *Каждан А.Б., Гуськов О.И.* Математические методы в геологии. М.: Недра, 1990. 250 с.

26. *Кендалл М.Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976. 736 с.

27. *Крамбейн У., Грейбилл Ф.* Статистические модели в геологии. М.: Мир, 1969.

28. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. М.: Мир, 1967. 144 с.

29. *Никитин А.А.* Теоретические основы обработки геофизической информации. М.: Недра, 1986.

30. *Окунь Я.* Факторный анализ. М.: Статистика, 1974. 198 с.

31. *Определение* основных расчётных гидрологических характеристик: СП 33-101-2003. М., 2004.

32. *Рождественский А.В., Чеботарев А.И.* Статистические методы в гидрологии. Л.: Гидрометеоиздат, 1974. 424 с.

33. *Румянцев В.А., Бовыкин И.В.* Математико-статистические основы совместного анализа временных гидрологических рядов. СПб.: Наука, 2009.

34. *Румянцев В.А., Бовыкин И.В.* Многомерные распределения вероятностей и их применение в гидрологии. СПб.: Гидрометеоиздат, 1994.

35. SPSS: искусство обработки информации. Анализ статистических данных и выявление скрытых закономерностей / Ахим Бююль, Петер Цёфель. СПб.: ООО «ДиаСофтЮП», 2002. 608 с.

36. *Статистические* методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. М.: Наука, 1986. 464 с.

37. *Структура* и динамика речного стока горных регионов / Под ред. В.И. Верболова. Новосибирск: Наука, 1987. 160 с.

38. *Тьюки Дж.* Анализ результатов наблюдений / Пер. с анг. А.Ф. Кушнира. М.: Мир, 1981. 693 с.

39. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 1998. 528 с.

40. *Факторный*, дискриминантный и кластерный анализ / Под ред. И.С. Енюкова. М.: Статистика, 1972. 486 с.

41. *Факторный*, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. 215 с.

42. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983.

43. *Харман Г.* Современный факторный анализ. М.: Статистика, 1972. 486 с.

44. *Шелутко В.А.* Численные методы в гидрологии: Учеб. Л.: Гидрометеоиздат, 1991. 238 с.

45. *Jöreskog K.G., Klovan J.E., Reyment.* Geological factor analysis. Amsterdam: Elsiver, 1976. P. 260.

Электронные ресурсы

1. *Электронный* учебник по промышленной статистике [Электронный ресурс]. М.: Statsoft Inc., 2001. Режим доступа: http://www.statsoft.ru/home/portal/textbook_ind/default.htm, свободный.

2. http://tora-centre.ru/ – сайт Тора-центра. Приводится описание большого числа систем интеллектуального анализа данных.

3. http://is1.cemi.rssi.ru/ruswin/index.html – содержит детальный обзор пакетов для статистического анализа.

4. *Программное* обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка / С.А. Айвазян, В.С. Степанов [Электронный ресурс]. Режим доступа: http://is1.cemi.rssi.ru/ruswin/publication/ep9700lt.html, свободный.

5. *Анализ* и прогнозирование гидрометеорологических данных в пакете «Statistica» / Л.И. Дубровская, И.В. Кужевская [Электронный ресурс]. Томск, 2007.Режим доступа: http://geo.tsu.ru, свободный.

Справочная литература

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.

2. *Болшев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1982.

3. *Вероятность* и математическая статистика. М.: Большая Российская Энциклопедия, 1999.

4. *Иванов В.В.* Методы вычислений на ЭВМ: Справочное пособие. Киев: Наукова думка, 1986. 584 с.

5. *Справочник* по прикладной статистике / Э. Ллойд, У. Ледерман. М.: Финансы и статистика, 1989. Т. 1. 510 с.

6. *Справочник* по прикладной статистике / Э. Ллойд, У. Ледерман. М.: Финансы и статистика, 1990. Т. 2. 526 с.

7. *Поллард Дж.* Справочник по вычислительным методам статистики / Пер. с англ. В.С. Занадворова. М.: Финансы и статистика, 1982. 344 с.

ГЛОССАРИЙ

Альтернативная гипотеза – гипотеза о значимости различий.

Асимметрия (коэффициент асимметрии) – число, характеризующее степень симметричности рассеяния данных относительно математического ожидания генеральной совокупности (эмпирический коэффициент асимметрии – относительно среднего значения выборки).

Антропометрия – область науки о размерах человеческого тела.

Вариация – рассеяние, разброс, неоднородность или изменчивость значений выборки.

Выборка – отобранная тем или иным способом часть генеральной совокупности.

Генеральная совокупность – множество относительно однородных, но индивидуально различимых объектов (наблюдений, измерений, описаний), объединённых для совместного изучения.

Гипотеза – это предположение, которое вызывает сомнение.

Гистограмма – графическое изображение частоты попадания элементов выборки в соответствующий интервал группировки.

Главные компоненты – собственные вектора ковариационной матрицы.

Дендрограмма (иерархическое дерево) – график, иллюстрирующий соподчинённость кластеров.

Диаграмма рассеивания – графическое отображение связи между двумя переменными.

Дивизивный метод кластерного анализа – разновидность иерархического алгоритма, в котором вначале все объекты представляют один класс, а затем производится их разбиение на группы.

Дисперсия – мера варьирования числовых значений признака вокруг его среднего значения.

Зашумление данных – термин, означающий, что объекты не могут быть расклассифицированы однозначно, т.е. кластеры пересекаются. Такая ситуация встречается в данных с «плохой» структурой, с трудно различимыми состояниями.

Иерархические алгоритмы – группа методов кластерного анализа, результатом работы которых является иерархический ряд (иерархическое дерево) из кластеров. Иерархические процедуры позволяют проследить процесс выделения группировок и иллюстрируют соподчинённость кластеров, образующихся на разных этапах работы.

Итеративный метод – тип вычислительной процедуры, когда решение находится методом последовательных приближений; требует задания примерных значений решения в качестве нулевого приближения.

Итерация – отдельные повторяющиеся шаги вычислительного цикла в итеративной процедуре, дающие последовательные приближения к искомому результату.

Категориальные данные – данные номинальной шкалы измерений (например, данные социологических опросов).

Квантиль – значение случайной величины X_p, соответствующее заданной вероятности *непревышения* p. В гидрометеорологической практике по аналогии с квантилями используются ординаты кривой обеспеченности $X_{p'}$, соответствующие вероятности *превышения* p' ($p = 1 - p'$).

Кластер – группа (скопление) элементов выборки, характеризующихся каким-либо общим свойством.

Кластеризация – выделение различных однородных групп данных.

Ковариация (корреляция) – мера связи между двумя исследуемыми признаками.

Компонентный анализ (метод главных компонент) – отличается от факторного анализа тем, что учитываются все факторы. О соподчиненности этих двух методов нет устоявшегося мнения.

Контрольная выборка – выборка, на которой проверяется точность уравнения регрессии.

Коэффициент вариации – относительный показатель изменчивости данных. Представляет собой среднее квадратическое отклонение, выраженное в процентах или в долях единицы от среднего значения.

Коэффициент детерминации – коэффициент, равный квадрату коэффициента множественной корреляции, т.е. R^2. Имеет смысл доли дисперсии исходного ряда y, объясняемой независимыми переменными x_1, x_2, ... x_n.

Коэффициент ковариации – число, характеризующее меру тесноты связи двух случайных величин. Рассчитывается по формуле

$$\mathrm{cov}(x, y) = \sum_{i=1}^{n} (x_i - x_{sr})(y_i - y_{sr}).$$

Коэффициент корреляции – безразмерная величина, характеризующая меру тесноты **линейной** связи двух случайных величин. Рассчитывается по формуле

$$r = \sum_{i=1}^{n}(x_i - x_{sr})(y_i - y_{sr}) / \sqrt{\sum_{i=1}^{n}(x_i - x_{sr})^2 \sum_{i=1}^{n}(y_i - y_{sr})^2};$$

принимает значения из интервала [–1, +1].

Критерий Стьюдента – параметрический критерий проверки однородности по среднему двух рядов наблюдений.

Критерий Хоттелинга (T^2) – критерий для проверки гипотезы о равенстве математических ожиданий двух многомерных случайных величин (многомерный аналог критерия Стьюдента).

Критерий оптимальности – условие, выбираемое исследователем для оценки расхождения модели с исходными данными.

Лаг (сдвиг) – число, означающее сдвиг элементов ряда на несколько значений вперёд или назад относительно исходного ряда (шаг запаздывания).

Латентная (скрытая) переменная – переменная, которая не поддаётся прямому измерению.

Матрица воспроизведённых корреляций – корреляционная матрица для переменных, рассчитанных по уравнению факторной модели.

Матрица остаточных корреляций – разность двух корреляционных матриц: воспроизведённой и исходных данных.

Медиана – значение, расположенное в середине выборки, *упорядоченной* по возрастанию или убыванию. Таким образом, одна половина значений выборки оказывается меньше медианы, а другая – больше.

Методы распознавания образов – *см.* таксономия.

Метод главных компонент (компонентный анализ) – один из методов теории латентных переменных, в котором, в отличие от факторного анализа, учитываются все компоненты (факторы). Устоявшегося мнения об иерархии факторного анализа и метода главных компонент в настоящее время не существует.

Метрика – правило вычисления расстояний между любой парой объектов исследуемого множества.

Мода – значение, встречающееся в выборке наиболее часто.

Неиерархические (структурные) алгоритмы – группа методов кластерного анализа, в которых реализуется идея образования кластеров по принципу выделения сгущений.

Непараметрический критерий – статистический критерий, не включающий в формулу расчёта параметры распределения и основанный на оперировании частотами и рангами.

Непрерывная переменная – переменная, принимающая любое значение внутри некоторой области числовых значений.

Нормировка – вычислительный приём приведения ряда к безразмерному виду (*см.* п. 4.2).

Нулевая гипотеза – это гипотеза об отсутствии различий.

Обеспеченность – вероятность превышения заданного значения случайной величины.

Объём выборки – количество элементов в выборке.

Обучающая выборка – выборка, по которой оценивались коэффициенты уравнения регрессии.

Общностью, соответствующей переменной y_j, в факторном анализе называют сумму квадратов нагрузок $\sum_{i=1}^{m} a_{ij}^2$. Чем больше это значение, тем лучше описывается y_j факторами f_i.

Ошибка I-го рода – ошибка, состоящая в отклонении нулевой гипотезы, в то время как она верна.

Ошибка II-го рода – ошибка, состоящая в принятии нулевой гипотезы, в то время как она неверна.

Ортогонализация – математическое преобразование множества линейно независимых функций с сохранением их линейной независимости.

Остаток – разность наблюдённого данного и предсказанного по уравнению регрессии.

Параметрический критерий – это критерий, включающий в формулу расчёта параметры распределения, например, среднее и дисперсию для нормального закона распределения.

Переменная – изучаемая характеристика объекта исследования.

Порядок матрицы – числа n (число строк) и m (число столбцов).

Признак – изучаемая характеристика объекта исследования, свойство объекта.

Простая структура – термин, означающий вид, к которому приводится в результате вращения матрица факторных нагрузок. Простую структуру легче интерпретировать.

Предиктор – зависимая переменная (Y) в уравнении регрессии.

Предиктант – независимая переменная (X) в уравнении регрессии.

Порог (пороговое значение) – число, равное расстоянию между двумя кластерами, при достижении которого следует прекратить процедуру объединения объектов в классы. Может быть как постоянным, так изменяющимся по какому-либо правилу. Используется в кластерном анализе.

Размах – это разность максимального и минимального значений в выборке.

Размерность признакового пространства – число, равное количеству измеренных признаков объекта.

Ранг матрицы – число линейно независимых строк (или столбцов) матрицы.

Ранг числа – это номер позиции числа в выборке, упорядоченной по убыванию или возрастанию.

Регрессионный анализ – статистический метод исследования зависимости случайной величины y от независимых переменных x_i. Конечная цель регрессионного анализа – построить по корреляционной матрице уравнение регрессии, по которому можно содержательно интерпретировать результаты наблюдений и осуществлять прогноз.

Репрезентативная выборка – выборка, адекватно представляющая пропорции генеральной совокупности.

Сгущения – это места наибольшей концентрации точек в рассматриваемом пространстве. Понятие используется в кластерном анализе.

Сдвиг (лаг) – это число, означающее, на сколько элементов исследуемый ряд смещён вперёд или назад относительно исходного ряда (шаг запаздывания).

Сжатие данных – сокращение размерности признакового пространства или, по-другому, уменьшение числа признаков, характеризующих объект, без потери информации.

Среднее значение – число, вокруг которого группируется большинство значений в выборке.

Среднее квадратическое отклонение – положительное значение квадратного корня из дисперсии.

Среднее отклонение – отклонение каждого значения от среднего значения в выборке.

Стандартизация – вычислительный приём приведения совокупности данных к безразмерному виду вычитанием из каждого члена ряда среднего значения и делением на среднее квадратическое отклонение. Среднее стандартизованного ряда равно нулю, а дисперсия – единице.

Статистическая гипотеза – это предположение о свойстве генеральной совокупности, которое мы хотим проверить по имеющимся данным.

Статистическая закономерность – форма проявления причинной связи, выражающаяся в последовательности, регулярности, повторяемости событий с достаточно высокой степенью вероятности, если причины, порождающие события, не изменяются или изменяются незначительно.

Статистическая совокупность – множество единиц, обладающих массовостью, однородностью, определённой целостностью, взаимозависимостью состояний отдельных единиц и наличием вариации.

Статистический критерий – решающее правило, обеспечивающее принятие истинной и отклонение ложной гипотезы с высокой вероятностью.

Стохастическая зависимость – вид неоднозначной зависимости между двумя случайными величинами, когда одному значению X может соответствовать несколько наблюдений Y.

Таксон – систематизированная группа любой категории.

Таксономия – численная таксономия, распознавание образов с самообучением – методы нахождения таксонов.

Транспонированная матрица – матрица, полученная поворотом исходной на 90°.

Уровень значимости – вероятность отвергнуть правильную гипотезу, т.е. совершить ошибку первого рода. В анализе естественнонаучных данных обычно назначают уровень значимости, равный 5%, или в долях единицы – 0,05.

Фактор – в факторном анализе это гипотетический, непосредственно не измеряемый (латентный, скрытый) показатель, в той или иной мере связанный с измеряемыми характеристиками, или, по-другому, *обобщённый* показатель свойств комплекса генетически однородных характеристик.

Факторные нагрузки – компоненты (составляющие) нормализованного собственного вектора ковариационной (в методе главных компонент) или корреляционной (в факторном анализе) матриц, умноженные на корень квадратный из соответствующего собственного значения.

Факторизация – вычисление матрицы нагрузок в факторном анализе.

Функционал качества разбиения – числовая функция, определённая на некотором множестве (например, множество разбиений на группы в кластерном анализе).

Характерность – *син.* специфичность – остатки e_j^2, которые показывают, какая часть дисперсии остаётся необъяснённой при используемом наборе факторов.

Эталон (эталонная точка) – точка в группе, которая по какому-либо правилу может быть выбрана в качестве представителя этой группы. В кластерном анализе эталоном считают центр тяжести группы объектов.

Эталонная группа – группа с заранее известными свойствами, с которой соотносят объект для принятия решения о принадлежности к ней. Термин используется в дискриминантном анализе.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1

Основные положения пользовательского интерфейса системы комплексного статистического анализа и обработки данных Statistica в приложении к задачам гидрометеорологического содержания

Цель расчётов – не числа, а понимание.
Р.В. Хемминг

Применение статистического анализа в любой конкретной прикладной области обладает специфическими особенностями, как в выборе методов исследования, так и в оценках и интерпретации результатов. Вышесказанное в полной мере относится к гидрометеорологии, в которой, например, нормативно принято вместо интегральной функции вероятностей использовать кривую обеспеченности. Поэтому даже при наличии большого выбора литературы по статистическим методам общего плана остается необходимость в учебниках, методических пособиях и указаниях узкопрофильных.

Средства пакета Statistica могут быть использованы в гидрометеорологии для первичного анализа данных средствами показателей описательной статистики, проверки статистических гипотез, расчёта ординат кривой обеспеченности и побора к ней теоретического закона распределения, создания сложных комплексных графиков, анализа и прогнозирования временных рядов, определения циклических колебаний в наблюдениях средствами спектрального анализа, кластерного и факторного анализа для исследования структуры рядов наблюдений и как вспомогательного средства для районирования территории по какому-либо признаку. Для прогнозирования можно использовать методы авторегрессии и скользящего среднего (ARIMA), множественной регрессии, нейронные сети.

Рынок компьютерных программ анализа данных очень разнообразен, что является отображением многоплановости задач анализа экспериментальных данных в различных областях человеческой деятельности. Все пакеты, как зарубежные, так и отечественные, делятся на две большие группы: универсальные и специализированные.

Из зарубежных универсальных наиболее известные Statistica, SPSS, Statgraphics, S-Plus, из отечественных – Stadia, Olimp и др. Наибольший интерес для задач гидрометеорологического содержания представляют

методы анализа временных рядов. Методы анализа временных рядов широко представлены во многих универсальных статистических пакетах STADIA, STATGRAPHICS, SPSS, STATISTICA. Но анализ временных рядов – очень специфическая область статистики, отличающаяся по кругу задач и методов их решения, а также по составу пользователей, применяющих эти методы. Поэтому для анализа временных рядов имеются также и специализированные статистические пакеты – ЭВРИСТА, МЕЗОЗАВР, FOREGAST EXPERT, СТАТИСТИК-КОНСУЛЬТАНТ и т.д.

МЕЗОЗАВР разработан специально для обработки гидрометеорологической информации. Пакет ЭВРИСТА является одним из лучших отечественных специализированных пакетов для анализа временных рядов (Тюрин, Макаров, 1998). Его функциональные возможности значительно шире стандартных процедур анализа временных рядов универсальных статистических пакетов. Например, в трудном вопросе подбора порядка модели авторегрессии ЭВРИСТА проводит экспертную оценку с помощью различных критериев (Парзена, Акаике, Хеннана – Куина) и выдает их результаты пользователю в качестве «подсказки». В пакете Statistica эту проблему приходиться решать «вручную», строя и анализируя графики автокорреляционной (АКФ) и частной автокорреляционной (ЧАКФ) функций.

Лидирующее положение на рынке статистического программного обеспечения в настоящее время занимает интегрированная система комплексного статистического анализа и обработки данных Statistica.

Пакет состоит из 19 специализированных статистических модулей, обладает мощной графической системой визуализации данных и результатов, имеет специальный инструмент для создания отчетов, встроенные языки программирования SQL и Statistica Basic, может обрабатывать очень большие массивы наблюдений (корреляционные матрицы размером 32 000×32 000).

Наибольший интерес для решения анализа гидрометеорологической информации представляют 6 следующих модулей:

1. **Basic Statistics/Tables (Основные статистики и таблицы).** Выдает описательные статистики, таблицы частот и матрицу коэффициентов корреляции, позволяет провести анализ однородности рядов наблюдений параметрическими методами, содержит калькулятор для расчёта вероятностей и обеспеченностей.

2. **Multiple Regresion (Множественная регресия).** При изучении многих гидрометеорологических процессов требуется установить вид линейной зависимости между несколькими переменными. Для решения этой задачи привлекается метод множественной регрессии. На основе материа-

лов наблюдений за величиной y и определяющими её величинами x_1, x_2, ... x_n методом наименьших квадратов подбираются коэффициенты уравнения регрессии.

3. **Nonparametrics (Непараметрическая статистика).** В этом модуле осуществляется проверка различных гипотез о характере распределения данных. Модуль содержит большой набор непараметрических критериев согласия: Колмогорова – Смирнова, Манна - Уитни, Вальда – Вольфовица, Уилкоксона и др.

4. **Distribution Fitting (Подбор распределения).** Предназначается для сравнения различных видов распределения рядов наблюдений с теоретическими («подгонка» теоретического распределения к эмпирическому).

5. **Time Series/Foregasting (Временные ряды/Прогнозирование).** Общее назначение модуля – построить простую модель, описывающую ряд, сгладить его, спрогнозировать будущие значения временного ряда на основе наблюдаемых до данного момента, построить регрессионные зависимости одного ряда от другого, провести спектральный анализ Фурье или Фурье-анализ и т.д.

6. **Cluster Analysis (Кластерный анализ).** Позволяет разбивать данные на классы по различным признакам близости или, наоборот, по отдалённости свойств. В гидрометеорологии привлекается для районирования территории по какому-либо признаку, для всевозможных классификационных построений.

7. **Spectral (Fourier) analysis (Спектральный (Фурье) анализ).** Один из важнейших видов анализа гидрометеорологических наблюдений, заключающийся в разложении исследуемого ряда на различные частотные составляющие (спектр), показывающий вклад колебаний с разными частотами в общую энергию процесса. Используется для предсказания погоды, для обнаружения в спектре пульсаций морского волнения или ветрового потока опасных частотных составляющих для морских и воздушных судов, анализа колебаний речного стока и т.д. Входит как составная часть в модуль анализа временных рядов Time Series/Foregasting (Временные ряды/Прогнозирование).

Основные положения пользовательского интерфейса пакета Statistica

Пакет Statistica 6.0 является сложной многофункциональной системой и, чтобы в нем ориентироваться, необходимо иметь представление об

особенностях его функционирования. Ниже приводятся некоторые наиболее важные положения об устройстве и интерфейсе пакета.

– Данные для обработки должны быть помещены в электронную таблицу данных – *Spreadsheet.* Таких таблиц может быть несколько. Но только одна из них является активной (её имя высвечивается в красной рамочке). Данные для обработки доступны только из активной таблицы данных. Сделать активной можно любую из таблиц.

– Результаты выдаются в виде графиков и таблиц (Scrollsheet), каждый результат в своем окне.

– Для ведения протокола текущего сеанса можно заказать документ под названием рабочая книга (Workbook), в которой будут храниться в хронологическом порядке сведения обо всех использовавшихся модулях и результаты их работы.

– Для подготовки и печати отчета в пакете существует еще один тип документа – отчет (Report). В отличие от предыдущих документов его можно сохранить в формате *.rtf (расширенный текстовый формат) для просмотра и редактирования в Word. Однако возможности редактирования отчета в Word значительно уступают таковым в среде Statistica.

– В соответствии со стандартами среды Windows каждый из пяти вышеперечисленных типов документов выводится на экран в своем собственном окне в рабочей области пакета. Каждый из них может быть сохранен в отдельном файле в специальном формате: таблица с данными – *.sta, *.xls, *.dbf, *.txt, *.htm; таблица с результатами – *.sta, *.xls, *.dbf, *.txt, *.htm (в версии 5.0 – *.scr); график – *.stg, *.wmf, *.bmp; рабочая книга (Workbook) – *.stw; отчет (Report) – *.rtf, *.str, *.txt, *.htm, *.html.

– Как только окно документа становится активным, изменяется панель инструментов и меню. В них появляются команды и кнопки, предназначенные для обработки данного типа документа.

– Столбцы электронной таблицы называются Variables (переменными), а строки Cases (случаями).

– Строки могут иметь наряду с нумерацией текстовые имена. Переключение осуществляется кнопкой Shows/Hide Cases Name (Показать/ Спрятать имена случаев).

– В соответствии с соглашением о двойной записи текстовые значения могут иметь числовые метки, что позволяет вместо текста вводить числа. Переключение осуществляется кнопкой в виде замка.

– Рекомендуется перед началом работы выбрать способ вывода результатов на экран с помощью Output Manager (Диспетчер вывода), кото-

рый предоставляет на выбор несколько вариантов размещения окон с результатами вычислений:

– Каждый результат выдается в виде индивидуального окна в рабочей области пакета.

– Окно с результатом автоматически помещается только в рабочую книгу Workbook.

– Окно с результатом автоматически помещается только в отчет Report.

– Окно с результатом автоматически помещается и в рабочую книгу и в отчет.

Диалоговое окно Output Manager (Мастер вывода) можно вызвать из меню File.

– Любой график, таблицу можно «изъять» из рабочей книги или отчёта и преобразовать в индивидуальное окно в рабочей области пакета путём буксировки мышью или командой Extract из меню Workbook.

– Любое индивидуальное окно с графиком, таблицей, а также электронную таблицу с данными можно перенести в рабочую книгу или отчёт. Кнопки добавления Add to Workbook и Add to Report доступны только при активном индивидуальном окне.

– Формула задается сразу для всех ячеек столбца, а не для отдельной ячейки, как принято в электронной таблице Excel. Формула рассматривается как длинное имя столбца (Long Name) и задается в окне его спецификаций, вызов которого осуществляется двойным щелчком на имени столбца.

– По умолчанию выбранная пользователем операция применяется ко всему столбцу. Если обработке подлежит часть значений столбца, то диапазон ячеек надо указать с помощью кнопки Select Case. Эта кнопка имеется в диалоговых окнах у всех модулей.

– В пакете имеется набор быстрых блоковых операций, которые применяются к предварительно выделенному блоку данных. Во всех остальных случаях построения графиков предварительное выделение области с данными бесполезно. Всё равно придется в диалоговом окне повторно указывать имена обрабатываемых столбцов.

– Если график построен по данным из таблицы Spreadsheet, то в любой момент можно вызвать окно с этими данными и отредактировать. Изменения в данных немедленно отразятся на графике, причем, исходную таблицу Spreadsheet они не затрагивают. Осуществляет эту процедуру Graph Data Editor (Редактор данных графика), который можно вызвать из меню View или Format при активном окне графика.

– Полный список модулей представлен в меню Statistics, а графиков – в меню Graphs.

ПРИЛОЖЕНИЕ 2

Таблица критических точек статистики Дарбина – Уотсона
(n – число наблюдений, к – число предикторов, $p = 0{,}05$)

	к = 1		к = 2		к = 3		к = 4		к = 5	
n	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,96	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,03	1,38	1,02	1,54	1,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,98	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,56	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,25	1,56	1,18	1,65	1,10	1,76	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,25	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,05	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,78
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,46	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,79
50	1,50	1,58	1,46	1,63	1,42	1,67	1,36	1,72	1,34	1,77
55	1,51	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78

ОГЛАВЛЕНИЕ

Учебное издание

Лариса Ивановна Дубровская
Георгий Борисович Князев

КОМПЬЮТЕРНАЯ ОБРАБОТКА ЕСТЕСТВЕННО-НАУЧНЫХ ДАННЫХ МЕТОДАМИ МНОГОМЕРНОЙ ПРИКЛАДНОЙ СТАТИСТИКИ

Издание подготовлено в авторской редакции

Оригинал-макет О.П. Шершневой
Дизайн обложки А.В. Бабенко

Подписано к печати 11.01.2011 г. Формат 64×80/16.
Ризография. Бумага писчая. Гарнитура Times.
Усл. печ. л. 6,9. Тираж 250 экз. Заказ № 3682.

ООО «Издательство „ТМЛ-Пресс“»
634050, г. Томск, ул. Гагарина, 31. Оф. 49.

ООО «Графика»
634050, г. Томск, ул. Беленца, 17. Оф. 58.

Дубровская Лариса Ивановна

Кандидат физико-математических наук, доцент кафедры гидрологии Национального исследовательского Томского государственного университета.

Основные научные интересы связаны с моделированием природных процессов, статистическими методами в гидрологии, геоинформационными системами.

Автор более ста научных статей, учебного пособия по компьютерным технологиям в гидрометеорологии.

Князев Георгий Борисович

Кандидат геолого-минералогических наук, доцент кафедры минералогии и кристаллографии Национального исследовательского Томского государственного университета.

Основные научные интересы связаны с геологией месторождений полезных ископаемых Алтае-Саянской складчатой области.

Автор около двухсот научных статей, четырех учебных пособий по кристаллографии, теории вероятностей и математической статистике, геоинформатике.