ЧАСТЬ 3. СТАТИСТИЧЕСКИЙ АНАЛИЗ В ГИДРОЭКОЛОГИИ: ЗАДАЧИ И РЕШЕНИЯ

> **Мем № 27**: *«Как ни странно, но задачи фитоиндикации, вероятностные по своей природе, до сих пор решаются в основном без использования каких-либо статистических методов»* В.И. Василевич [1969].

Эта часть книги посвящена описанию конкретных методов математической статистики, распознавания образов и алгоритмов искусственного интеллекта применительно к анализу результатов гидроэкологического мониторинга.

Деление излагаемого материала на главы выполнено не вполне традиционным образом: не по "генетическому" сродству отдельных методов, а в соответствии с общностью постановок конкретных задач гидроэкологии и схемы последующей обработки данных:

- *Глава 5* объединяет различные параметрические и непараметрические методы анализа, когда исследователь располагает одним или двумя вариационными рядами измерений, представленными в количественной шкале;
- *Глава 6* является, в определенном смысле, продолжением главы 5 применительно к данным, измеренным в порядковых шкалах (или сведенным к таковым различными методами "интервальной" математики);
- *Глава 7* объединяет методы статистического анализа "без учителя" и представляет интерес, если исследователь имеет двухмерную таблицу наблюдений, в которой явным образом отсутствует моделируемая величина $\boldsymbol{Y}$ (т.е. "отклик");
- В *главе 8* описаны методы анализа "с учителем", когда исследователь располагает некоторой обучающей выборкой – многомерной матрицей варьируемых переменных и сопряженным с нею вектором измерений моделируемой величины $\boldsymbol{Y}$;
- И, наконец, при комплектовании *главы 9* мы включили туда методы, которые по своему смыслу довольно мало отличаются от представленных в главе 8, но характеризуются относительной новизной и не слишком долговечной, на наш взгляд, рекламной меткой «***ИИ***» (искусственный интеллект).

Общее описание и сравнительный обзор всех этих методов выполнен в разделах 4-8 главы 2, где приведена также основная рекомендуемая библиография.

Каждый раздел представляемой части построен по единой стандартной схеме из трех следующих подразделов:

- «***Формулировка задачи***», содержащая общую экологическую и математическую постановку проблемы по принципу описания "мясорубки" (т.е. берем на входе объект «А» и желаем получить на выходе объект «Б»);
- ***«Математический лист»*** или краткое описание теоретических идей, используемых при построении статистических моделей и оценки их достоверности (в названии этого подраздела мы отдали должное Э.Т. Гофманну, впервые использовавшему *макулатурные листы* в "Житейских воззрениях кота Мурра"; читатель, не интересующийся математической "кухней", может их вполне пропустить, иногда, без большого ущерба для понимания существа дела);
- ***«Результаты расчетов»,*** содержащие более или менее развернутые примеры использования каждого метода на основе единого тестового массива гидробиологических показателей.

Расчеты, иллюстрирующие изложенные методы, были выполнены на основе одного и того же "сквозного" для всех глав набора исходных измерений: данных гидроэкологического мониторинга донных организмов на 40 малых реках, расположенных в степной и лесостепной зонах Среднего Поволжья (см. карту на рис. 1.8). Подробно структура и информационный состав базы данных описан в разделе 1.6.

Выборка, использовавшаяся при построении статистических моделей, характеризовалась следующими основными параметрами:

- количество водных объектов (малых рек Самарской области) – **34**;
- количество станций наблюдений (по выделенным створам рек) – **247**;
- количество гидробиологических проб и сопряженных с ними гидрохимических и гидрологических измерений – **571**;
- диапазон дат измерений – от 10.07.1985 до 31.07.2000 г.;
- сезонный диапазон дат измерений – с 1 мая по 1 ноября;
- количество видов макрозообентоса – **580**;
- количество значений численности и биомассы по видам гидробионтов, полученных в результате обработки всех 571 проб – **5937**;
- количество учитываемых качественных и количественных гидрологических показателей – **12**;
- количество учитываемых гидрохимических показателей – **18**;
- общее количество измерений гидрохимических показателей – **3102**.

На рисунках, в таблицах и уравнениях нами использовались следующие условные обозначения для измеренных и рассчитанных показателей применительно к каждой гидробиологической пробе:

S – число видов макрозообентоса в пробе;
N_i – численность i-го вида в пробе, экз/м^2;
N_S – суммарная численность всех видов в пробе;
B_i – биомасса i-го вида в пробе, мг/м^2;
B_S – суммарная биомасса всех видов в пробе;
$\sqrt{N_i \cdot B_i}$ или $(N_i \cdot B_i)^{0.5}$ – индекс плотности, рассчитанный для вида в пробе (см. раздел 2.3);
$(N_s \cdot B_s)^{0.5}$ – индекс плотности всех видов в пробе;
H – информационный индекс Шеннона (см. раздел 4.3);
V – биотический индекс Вудивисса (см. раздел 4.5);
P – олигохетный индекс Гуднайта-Уитлея-Пареле (см. раздел 4.2);
D – хирономидный индекс Балушкиной (см. раздел 4.2).

Для формирования обучающей выборки, использованной в главах 8 и 9, по каждому из 571 комплексов измерений нами оценивался класс качества воды по шестибальной шкале в соответствии с ГОСТ 17.1.3.07–82. Эта оценка выполнялась на основе гидрохимического индекса загрязнения воды ИЗВ (см. раздел 3.5), если имелся в наличии необходимый набор из 6 показателей, либо по методике Былинкиной и Драчева (см. раздел 3.6), если имеющихся гидрохимических данных было недостаточно.

При выполнении расчетов мы использовали следующие программные продукты:

- пакет прикладных программ для статистических расчетов STATISTICA v5.5A, методически наиболее полный на сегодняшний день, но, вследствие "странной стохастичности" видеоинтерфейса, далеко не всегда "дружественный" к пользователю (русскоязычные ссылки в Интернет http://www.statsoft.ru и http://www.exponenta.ru/soft/Statist);
- статистическую программу StatGraphics v5.0 для Windows, менее полную, но более гармоничную и "дружественную";
- программу статистического анализа на основе нейросетей Statistica Neural Networks v1.0;
- превосходную программу статистического анализа Matrix, разработанную А.А. Цыплаковым (Новосибирский госуниверситет) и распространяемую бесплатно на сайте Интернет www.nsu.ru/ef/tsy/ecmr/mtx;
- аналитический пакет Deductor, ориентированный на решение задач многомерного анализа и разработанный российской лабораторией BaseGroup Labs, специализирующейся в области искусственного интеллекта (сайт в Интернет http://www.basegroup.ru/labs).

Значительная часть расчетов, реализующих "нестандартные" математические методы и алгоритмы распознавания образов, была выполнена с использованием собственных программных модулей, разработанных авторами и входящих в состав компонентов базы гидробиологических данных по малым рекам Самарской области (СУБД Access 97, алгоритмический язык Visual Basic for Application).

Глава 5. Задачи о выборках: анализ распределений, сравнение, поиск зависимостей

5.1. Задача о законе распределения гидробиологических показателей

> **Мем № 28**: *«Случайность существует объективно и не зависит от того, знаем ли мы причины явления или не знаем…»* М.М. Розенталь [1952].

Формулировка задачи

Пусть имеется выборка из n значений $X_1, X_2, \ldots, X_n$ некоторого измеренного гидробиологического показателя. Необходимо:

- оценить закон вероятностного распределения натуральных величин X;
- подобрать такое функциональное преобразование $f(X)$ исходных значений X, которое позволило нам принять гипотезу о нормальном характере распределения преобразованных значений.

Эта задача внешне кажется вспомогательной в ряду предлагаемых задач, поскольку само по себе оценивание закона распределения не имеет большого практического смысла. Однако этот подготовительный этап носит обязательный и важный характер для последующего корректного применения большинства методов математической статистики.

Рекомендуемая литература: [Урбах, 1963; Смирнов, Дунин-Барковский, 1963; Кендалл, Стьюарт, 1966; Гмурман, 1972; Крамер, 1975; Закс, 1976; Джонсон, Лион, 1980, 1981; Гнеденко, 1988; Вентцель,1999; Калинина, Панкин, 2001; С.А. Прохоров, 2001а,б, 2002; Прохоров с соавт., 2003].

Математический лист

Случайная величина и ее характеристики

Одним из центральных понятий теории вероятностей является *случайная величина* – любая количественная характеристика, которая в результате случайного эксперимента может принять одно из некоторого множества значений.

Каждая случайная величина ξ полностью определяется своей функцией распределения:

$$F(x) = F_\xi(x) = P(\xi < x), \qquad (5.1)$$

где $P(\xi < x)$ – вероятность того, что случайная величина ξ принимает значение, меньшее x. Функция $F(x)$ монотонно возрастает на всей числовой оси, причем $F(-\infty) = 0$, $F(\infty) = 1$. Функция распределения является "паспортом" случайной величины: она содержит всю информация о ξ и поэтому изучение случайной величины заключается в исследовании ее функции распределения, которую часто называют просто *распределением*.

Если функция распределения $F_\xi(x)$ непрерывна, то случайная величина ξ называется *непрерывной* случайной величиной. Если функция распределения непрерывной случайной величины дифференцируема, то более наглядное представление о случайной величине дает *плотность вероятности* случайной величины $p_\xi(x)$, которая связана с функцией распределения $F_\xi(x)$ формулами

$$F_\xi = \int_{-\infty}^{x} p_\zeta(t)\cdot dt \quad \text{и} \quad p_\xi(x) = \frac{dF_\xi(x)}{dx}. \qquad (5.2)$$

Отсюда, в частности, следует, что для любой случайной величины : $\int_{-\infty}^{\infty} p_\zeta(x)\cdot dx = 1$.

При решении практических задач часто требуется найти значение x, при котором функция распределения $F_\xi(x)$ случайной величины ξ принимает заданное значение p, т.е. требуется решить уравнение $F_\xi(x) = p$. Решения такого уравнения (т.е. соответствующие значения x) в теории вероятностей называются *квантилями*. Квантилью x_p (p-квантилью, квантилью уровня p) случайной величины ξ, имеющей функцию распределения $F_\xi(x)$, называют решение x_p уравнения

$F_\xi(x) = p$, $p \in (0, 1)$. Для некоторых p уравнение $F_\xi(x) = p$ может иметь несколько решений, для некоторых – ни одного. Это означает, что для соответствующей случайной величины некоторые квантили определены неоднозначно, а некоторые квантили не существуют.

Квантили, наиболее часто встречающиеся в практических задачах, имеют свои названия: *медиана* (квантиль уровня 0,5); нижняя *квартиль (*квантиль уровня 0,25); верхняя квартиль (квантиль уровня 0,75); *децили (*квантили уровней 0,1, 0,2, …, 0,9); *процентили – (*квантили уровней 0,01, 0,02, …, 0,99).

Вероятность того, что значение непрерывной случайной величины $F_\xi(x)$ попадает в интервал (a, b), равная $P(a < \xi < b) = F_\xi(b) - F_\xi(a)$, вычисляется по формуле:

$$P(a < \xi < b) = \int_a^b p_\xi(x) \cdot dx = F(b) - F(a), \tag{5.3}$$

причем, если $a = -\infty$, то $P(a < \xi < b) = P(\xi < b) = F_\xi(b)$, а если $b = \infty$, то $P(a < \xi < b) = P(a < \xi) = 1 - P(\xi < a) = 1 - F_\xi(a)$.

Наиболее часто применяемыми числовыми характеристиками случайной величины ξ являются начальные и центральные моменты различного порядка. Для непрерывной случайной величины моменты порядка k определяются следующими формулами :

$$\alpha_k = \int_{-\infty}^{\infty} x^k f(x)dx, \mu_k = \int_{-\infty}^{\infty} (x - m_\xi)^k f(x)dx. \tag{5.4}$$

Чаще всего используется первый начальный момент $\alpha_1 = M_\xi$, называемый *математическим ожиданием* случайной величины ξ или центром распределения, и второй центральный момент $\alpha_2 = D_\xi$, называемый *дисперсией,* которая характеризует разброс случайной величины относительно центра распределения. Часто вместо дисперсии используют *среднее квадратичное отклонение* $\sigma_\xi = \sqrt{D_\xi}$.

Основные законы распределения

Перечислим наиболее распространенные распределения непрерывных случайных величин.

- *Равномерное распределение*. Непрерывная случайная величина x, принимающая значения на отрезке $[a, b]$, распределена равномерно, если ее плотность распределения $p_\xi(x)$, функция распределения $F_\xi(x)$ и моменты M_ξ и D_ξ имеют соответственно вид:

$$p_\xi(x) = \begin{cases} 0, & x \notin [a,b] \\ \dfrac{1}{b-a}, & x \in [a,b] \end{cases}; \quad F_\xi(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x \le b \\ 1, & x > b \end{cases}; \quad M_\xi = \frac{a+b}{2}; \quad D_\xi = \frac{(b-a)^2}{12}. \tag{5.5}$$

- *Нормальное распределение.* Случайная величина ξ нормально распределена с параметрами a и σ, $\sigma > 0$, если ее плотность распределения $p_\xi(x)$, функция распределения $F_\xi(x)$ и моменты M_ξ и D_ξ имеют соответственно вид:

$$p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}; \quad F_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{z-a}{\sigma}\right)^2} dz; \quad M_\xi = a \text{ и } D_\xi = \sigma^2. \tag{5.6}$$

Нормальное распределение играет исключительно важную роль в теории вероятностей и математической статистике.

- *Логнормальное распределение*. Случайная величина ξ имеет логарифмическое нормальное (логнормальное) распределение с параметрами a и σ, если случайная величина $\ln x$ имеет нормальное распределение с параметрами $a > 0$ и σ. Функция плотности вероятностей $p_\xi(x)$,

функция распределения $F_\xi(x)$ и моменты M_ξ, D_ξ логнормального распределения имеют соответственно вид:

$$p_\xi(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\ln x-\ln a}{2\sigma^2}};\quad F_\xi(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\ln x}e^{-\frac{z-\ln a}{2\sigma^2}}dz;\quad M_\xi=a\cdot e^{\frac{\sigma^2}{2}};\quad D_\xi=a^2\cdot e^{\sigma^2}\left(e^{\sigma^2}-1\right).$$

В ряде областей науки и техники нашли широкое применение такие одномерные распределения непрерывной случайной величины как экспоненциальное распределение, гамма-распределение, распределение Вейбулла и многие другие.

Основным предметом математической статистики является вычисление ***статистик*** (да простит нас читатель за тавтологию), являющихся критериями для оценки достоверности априорных предположений, гипотез или выводов по существу эмпирических данных: *«Статистики – это предписания, по которым из выборки рассчитывается некоторое число – значение статистики для данной выборки»* [Закс, 1976]. Выборочные среднее и дисперсия, отношение дисперсий двух выборок или любые другие функции от выборки могут рассматриваться как статистики[1].

Вильям ГОССЕТ -"СТЬЮДЕНТ"
(W.S.Hosset –"Student" 1876-1937)
известный английский статистик

Статистики также являются случайными переменными. Распределения статистик (тест-распределения) лежат в основе критериев, которые построены на этой статистике. Например, В.Госсет, работая на пивоварне Гиннеса и публикуясь под псевдонимом «Стьюдент», в 1908 г. доказал очень полезные свойства распределения отношения разности между выборочным средним и средним значением генеральной совокупности ($\bar{x}-\mu$) к стандартной ошибке среднего значения генеральной совокупности $\sigma/\sqrt{n}$, или t–статистики (*распределение Стьюдента*):

$$\frac{\{\text{Ошибка среднего значения}\}}{\{\text{Стандартная ошибка среднего}\}}=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}. \quad (5.7)$$

Распределение Стьюдента по форме при некоторых условиях приближается к нормальному.

Другими двумя важными распределениями выборочных статистик является χ^2*-распределение* и F*-распределение*, широко используемые в последующих разделах для проверки статистических гипотез.

Дадим определения и опишем основные свойства наиболее известных вероятностных распределений для дискретной случайной величины.

- <u>*Схема Бернулли и биномиальное распределение*</u>. Бернуллиевская модель (по имени Якоба Бернулли (1654–1705) – выдающегося швейцарского математика), является подходящей математической моделью для любого эксперимента с двумя исходами ("успех" - "неуспех"), т.е. простейшего статистического эксперимента.

  Случайная величина имеет распределение Бернулли с параметрами p ($0< p <1$), если она имеет лишь два значения, обозначаемые обычно 0 и 1, и при этом

  $$P\{X=1\}=p=1-P\{X=0\}$$

  В терминах плотности $f(x)$ это можно записать в следующем виде:

  $$f(x)=f(x\mid p)=p^x q^{1-x},\ x=\{0,1\},\ q=1-p. \quad (5.8)$$

  Пусть $\{X_1,\ldots,X_n\}$ – бернуллиевская последовательность с параметром p. Тогда сумма $X=X_1+\ldots+X_n$ имеет биномиальное распределение с параметрами n и p:

  $$P\{X=x\}=f(x\mid n,p)=C_n^x p^x q^{n-x},\quad x=0,1,\ldots,n \quad (5.9)$$

  Термин "биномиальное распределение" связан с тем, что вероятности P являются членами известного "бинома Ньютона": $\sum_{x=0}^{n}C_n^x p^x q^{n-x}=(p+q)^n=1$

[1] *Статистика* – классический пример представления «одним числом» сложного стохастического процесса.

Таким образом, биномиальная модель **Bi**(n, p) описывает распределение числа "успехов" в n испытаниях Бернулли с неизменной вероятностью "успеха" p.

Среднее и дисперсия такой случайной величины есть $M_\xi = n \cdot p$ и $D_\xi = n \cdot p \cdot q$.

- *Отрицательное биномиальное распределение*. С бесконечной последовательностью испытаний Бернулли $\{X_1, X_2, \ldots\}$ связано еще одно важное дискретное распределение, которое обозначается **Bi**(r, p) и называется отрицательным биномиальным распределением с параметрами r и p (здесь r – натуральное число). Это есть распределение числа "успехов" (1), предшествующих r-му "неуспеху" (0), и оно задается вероятностями

$$f(x \mid r, p) = C^{x}_{r+x-1} p^{x} q^{r}, \quad x = 0, 1, 2, \ldots \qquad (5.10)$$

Заметим, что выражение $f(x \mid r, p)$ совпадает с x-м членом разложения функции $q^r(1-p)^{-r}$ в ряд по степеням p; т.е. отрицательного бинома (отсюда происходит и название распределения). Если случайная величина имеет распределение **Bi**(r, p), то первые два центральных момента равны $M_\xi = \frac{r \cdot p}{q}$ и $D_\xi = \frac{r \cdot p}{q^2}$.

В частном случае при $r = 1$ распределение **Bi**(1, p) называется *геометрическим*: это есть распределение числа частиц, предшествующих первому нулю в бернуллиевской последовательности:

$$f(x \mid 1, p) = p^{x} q, \qquad x = 0, 1, 2, \ldots \qquad (5.11)$$

- *Распределение Пуассона*. Это одно из важнейших дискретных вероятностных распределений впервые было исследовано в 1837 г. С. Пуассоном (французский математик, механик и физик, 1781–1840 гг.) Распределение Пуассона обычно описывает схему редких событий, происшедших за фиксированный промежуток времени или в фиксированной области пространства, и дает хорошую аппроксимацию биномиального распределения для больших значений n и малых значений p. Случайная величина имеет распределение Пуассона с параметром λ ($\lambda > 0$), если

$$P\{X = x\} = f(x \mid \lambda) = e^{-\lambda} \frac{\lambda^{x}}{x!}, \quad x = 0, 1, 2, \ldots \qquad (5.12)$$

При этом $M_\xi = D_\xi = \lambda$.

Характер основных вероятностных распределений непрерывной и дискретной случайной величины представлены на рис. 5.1.

Проверка закона распределения эмпирического ряда

Обычно закон распределения случайной величины ξ неизвестен и его приближенно определяют (оценивают) опытным путем. С этой целью над величиной ξ проводят ряд независимых испытаний. Вся мыслимая (т.е. бесконечная) совокупность этих измерений называется *генеральной совокупностью*. А каждый конкретный ряд измерений $(x_1, x_2, \ldots, x_n)$ называют *простой случайной выборкой.*

Если простую выборку упорядочить по возрастанию, то ее называют *вариационным рядом.* Если для каждого неповторяющегося элемента вариационного ряда x_i указать относительную частоту его появления $p_i^* = \frac{m_i}{n}$, то такой вариационный ряд называют *статистическим рядом распределения случайной величины* ξ. Здесь m_i – число повторений x_i (абсолютная частота появления элемента), а n – общее число измерений, или *объем выборки*.

Имея вариационный ряд, легко построить *эмпирическую* (или статистическую) *функцию распределения* $F_n(x) = \frac{m_x}{n}$, где m_X – число членов вариационного ряда, лежащих левее от x, а m_X/n – частота попадания выборочного значения левее x. $F_n(x)$ представляет собой ступенчатую неубывающую функцию, заданную на всей числовой оси, со скачками в точках x_i, причем величина скачка равна частоте p_i^*. Заметим, что поскольку сумма абсолютных частот $\sum_{i=1}^{n} m_i = n$, то

сумма относительных частот $\sum_{i=1}^{n} p_i^* = 1$.

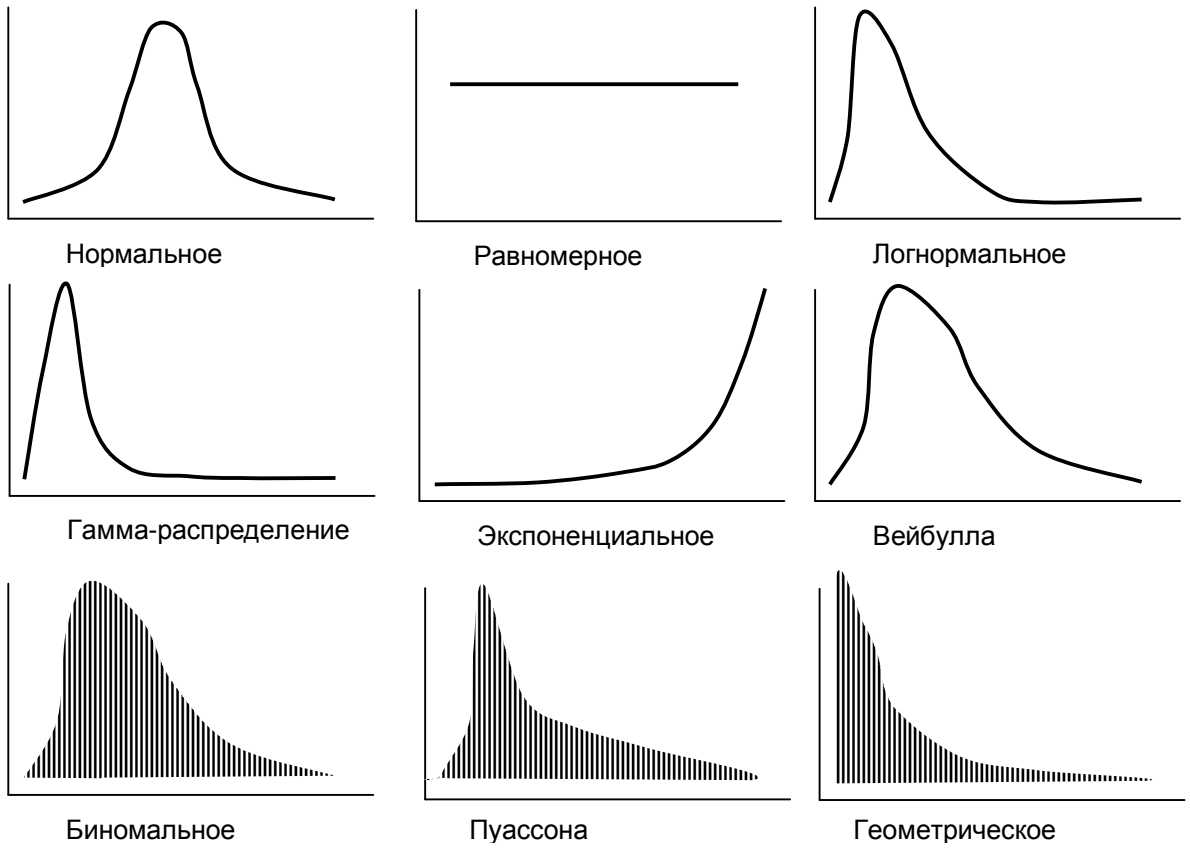


Рис. 5.1. Основные типы вероятностных распределений случайной величины

Согласно центральной теореме математической статистики – теоремы Гливенко - Кантелли о равномерной сходимости эмпирической функции распределения к истинной с ростом объема выборки, можно доказать, что $F_n(x) \xrightarrow{p} F(x)$ при $n \to \infty$. Отсюда ясно, что эмпирическую функцию распределения можно использовать как оценку теоретической функции распределения $F(x)$.

Совокупность разрядов и соответствующих частот статистического ряда геометрически изображают в виде *гистограммы*. По оси абсцисс откладывают интервалы и над каждым интервалом, как на основании, строят прямоугольник, высота которого равна значению плотности распределения для данного интервала $m_i/n \cdot h$. Таким образом, площадь каждого прямоугольника гистограммы равна его частоте, а общая площадь равна единице.

Большинство статистических вычислений сопровождается проверкой некоторых предположений или гипотез об источнике этих данных. Основное проверяемое предположение называется ***нулевой гипотезой*** и часто формулируется как отсутствие различий, незначительность влияния фактора, равенство нулю значений выборочных характеристик и т.п. Другое проверяемое предположение (не всегда строго противоположенное или обратное первому) называется конкурирующей или альтернативной гипотезой. Причина такого выделения нулевой гипотезы заключается в том, что она обычно рассматривается как утверждение, несостоятельность которого более бесспорно, чем истинность. Это основано на общем принципе, гласящем, что теория должна быть отвергнута, если есть противоречащий пример, но не обязательно должна быть принята, если такого примера не удалось найти.

Для проверки нулевой гипотезы задается ***статистический критерий*** (от греч. kriterion - средство для суждения; мерило оценки), согласно которому вычисляется (по заданному правилу или формуле) значение соответствующей статистики и уровень значимости α, который представляет собой вероятность ошибочно отвергнуть нулевую гипотезу тогда, когда она на самом деле верна (так называемая, ошибка 1-го рода).

Технически проверка нулевой гипотезы H_0 «нет статистически значимого различия» сводится к двум возможным операциям:

- выбирается критический уровень значимости $\alpha_{кр}$ из стандартной линейки типа 0,001; 0,01; 0,05 (в медико-биологических исследованиях часто принимают $\alpha_{кр} = 0.05$) и по таблицам находится пороговое значение критерия $K_{пор}$ для этого уровня значимости;
- с использованием программных средств или аппроксимационных формул находят точный уровень значимости α нулевой гипотезы (или p–значение, т.е. вероятность ошибочно отвергнуть гипотезу, когда она верна).

Нулевая гипотеза H_0 не отклоняется, если вычисленное значение статистики критерия $K_{рас}$ не превышает порогового $K_{пор}$ (первый случай), или если вычисленное значение α превышает критический уровень значимости $\alpha_{кр}$ (второй случай). В противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза H_1.

При подборе распределений возникает вопрос: а верна ли гипотеза о том, что функция распределения именно $F(x)$, а не какая-либо другая? Выражаясь более точно, не противоречит ли гипотеза о законе распределения $F(x)$ результатам эксперимента? Чтобы ответить на этот вопрос, пользуются *критериями согласия*. Под критерием согласия понимают некоторую величину $\Delta(F_n, F)$, которая отражает количественную меру расхождения гипотетического $F(x)$ и эмпирического $F_n(x)$ распределений. Величину разности между двумя распределениями можно выбрать многими способами и на ее основе имеются различные статистики для проверки интересующей нас гипотезы, например:

статистика Колмогорова
$$\Delta(F_n, F) = D_n = \sup\left|F_n(x) - F(x)\right|, \qquad (5.13)$$

статистика омега-квадрат Мизеса
$$\Delta(F_n, F) = \omega^2 = \int_{-\infty}^{+\infty}\left[F_n(x) - F(x)\right]^2 dF(x). \qquad (5.14)$$

Схема применения критерия согласия следующая. Возьмем $\alpha_{кр}$ из α**:** $(1 > \alpha > 0)$ настолько малым, чтобы осуществление события с вероятностью, не превышающей $\alpha_{кр}$ можно было считать практически невозможным в единичном опыте. Зная закон распределения случайной величины $\Delta = \Delta(F_n, F)$, найдем ее возможное значение Δ_0 из уравнения $P(\Delta > \Delta_0) = \alpha$. По данной выборке вычислим значение критерия согласия $\Delta_1 = \Delta(F_n, F)$. Если окажется, что $\Delta_1 > \Delta_0$, то это значит, что произошло практически невероятное событие. Следовательно, эксперимент опровергает нашу гипотезу, и она отбрасывается. При этом вероятность того, что мы ошибочно отбросили верную гипотезу, не превышает принятый уровень значимости $\alpha_{кр}$ критерия и равна α. Если $\Delta_1 < \Delta_0$, то гипотеза не противоречит эксперименту и должна быть принята.

А.Н. Колмогоров в 1933 г. нашел предельную функцию распределения величины $\lambda = \sqrt{n}\, D_n$, которую для больших n можно вычислить по формуле:

$$K(x) = \lim_{n \to \infty} P\left(\sqrt{n} D_n < x\right) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 x^2}, x > 0. \qquad (5.15)$$

Функцию $K(x)$ стали называть *критерием согласия Колмогорова*[2]. Чтобы воспользоваться критерием Колмогорова, нужно построить графики гипотетической и выборочной функций рас-

[2] Затем Н.В. Смирнов исследовал супремум (sup) и инфимум (inf) этого эмпирического процесса, поэтому нередко встречается название "критерий Колмогорова–Смирнова".

пределения, по графикам найти статистику D_n и вычислить величину $\lambda_1 = \sqrt{n}\, D_n$. Найти вероятность события $\sqrt{n}\, D_n > \lambda_1$ можно по формуле

$$P\left(\sqrt{n}\, D_n > \lambda_1\right) = 1 - K(\lambda_1) = -2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2\lambda_1^2} . \qquad (5.16)$$

Если эта вероятность меньше α, то гипотеза отвергается, если больше, то признается не противоречащей эксперименту.

Предположим теперь, что из некоторых соображений мы можем высказать гипотезу только о виде закона распределения, а параметры, входящие в него, нам неизвестны. В таких случаях часто используют *критерий согласия Пирсона*.

Всю числовую ось разобьем на r непересекающихся разрядов точками $-\infty = x_0 < x_1 < x_2 < \ldots < x_r = \infty$. Примем гипотезу о функции распределения, а неизвестные параметры, входящие в нее, заменим их оценками. Таким образом, гипотетическая функция распределения $F(x)$ будет полностью известна и можно будет найти вероятности $P_i = F(x_i) - F(x_{i-1})$ попадания случайной величины в i-й разряд. Возьмем статистику

$$t_n = \chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} , \qquad (5.17)$$

где n – объем выборки, r – число разрядов, m_i – число значений в i-м разряде.

За меру расхождения между гипотетической $F(x)$ и эмпирической $F_n(x)$ функциями распределения примем статистику $t_n = \Delta(F_n, F)$, определенную формулой (5.17). Р. Фишером с использованием метода максимального правдоподобия было доказано, что предельным законом распределения статистики t_n является распределение χ^2 с r-m-1 степенями свободы, где m – число параметров, входящих в гипотетическую функцию распределения. Доказано также, что при объеме выборки $n > 30$ с достаточной точностью можно пользоваться предельным законом распределения, если $np_i > 5$.

Схема применения критерия Пирсона следующая. По приведенным выше формулам вычисляют значение статистики $t_n = \Delta_0$ и вероятность

$$p(\Delta > \Delta_0) = \int_{\Delta_0}^{\infty} f(x)dx . \qquad (5.18)$$

Если эта вероятность меньше уровня значимости α, то гипотезу следует отбросить.

Специально для проверки нормальности распределения малых выборок, численностью от 3 до 50 вариант, разработан W *критерий Шапиро и Уилка* (Shapiro–Wilk), основанный на распределении порядковых статистик [Хан, Шапиро, 1969]. Этот критерий при наличии ограниченного объема данных является более мощным для проверки гипотезы нормальности, чем применяемые обычно критерии согласия [Лисенков, 1979].

Вычисления производятся по формулам:

$$W = b^2 / S^2, \qquad S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad b = \sum_{j=1}^{k} a_{n-j+1}\left(x_{n-j+1} - x_j\right), \qquad (5.19)$$

где x_i – ранжированный ряд, $i = 1,2,\ldots, n$;

$k = n/2$, если n – четное, или $k = (n-1)/2$, если n – нечетное;

a_{n-j+1} – константы ($j = 1,2,\ldots,k$), протабулированные для n от 3 до 50.

P-значение вычисляется по формуле $P = \Phi(z)$, где $\Phi(.)$ – функция стандартного нормального распределения, а величина $z = z(W)$ рассчитывается как $z = \gamma + \eta \ln \frac{W - \varepsilon}{1 - W}$, где греческими буквами обозначены табулированные константы.

Применение критериев согласия связано с определенными теоретическими и вычислительными сложностями. Поэтому для решения вопроса о возможности применения тех или иных параметрических тестов и методов дисперсионного анализа иногда считается удобным использовать группу критериев, которые позволяют оценить отклонение некоторых широко известных характеристик эмпирического ряда от нормального закона распределения. Например, в литературе [Айвазян с соавт., 1983] описан ***d-критерий Гири*** (Джири, Giri), вычисляемый по формулам:

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 \quad ; \quad b_2 = \frac{1}{ns^4}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 \quad ; \quad d = \frac{1}{ns}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| , \qquad (5.20)$$

использующим выборочную дисперсию s^2, асимметрию g_1 и эксцесс b_2.

Результаты расчетов

Распределение численности и биомассы организмов в пробах характеризуется определенными статистическими законами, обусловленными как совокупностью абиотических факторов, так и влиянием внутрипопуляционых, межпопуляционных и межвидовых отношений. Оценку законов распределения гидробиологических показателей осуществим на основе данных из базы по малым рекам Самарской области (см. преамбулу к настоящей части книги). Для этого сформируем выборку из численности N экземпляров отдельных видов зообентоса, приходящихся на 1 м^2 дна водоема, и биомассы B (г/м^2). Отдельно рассчитаем суммарные численности N_s и биомассы B_s по каждой взятой пробе, а также индекс Шеннона H.

Из полученных выборок сформируем дополнительные выборки на основе некоторого подмножества функциональных преобразований (логарифмирование, возведение в различную степень и т.д.) исходных вариационных рядов [Тьюки, 1981; Шитиков с соавт., 1985; Цейтлин, URL].

Основные описательные статистики по полученным выборкам представлены в табл. 5.1. Гистограммы распределения некоторых показателей приведены на рис. 5.2.

Представленные материалы позволяют сделать следующие выводы:

- выборки гидробиологических показателей, полученных как по отдельным видам, так и суммарных в пределах пробы, в натуральных шкалах измерения имеют очень сильные визуальные отличия от нормального закона распределения (см. фиг. «а» и «д» на рис.5.2); например, в один частотный диапазон 10-интервальной гистограммы попадает от 97.3 до 99.7% измерений;
- средние арифметические значения выборок по этим показателям очень сильно отличаются от моды и медианы; статистические моменты высших порядков (асимметрия и эксцесс) имеют крайне "неблагополучные" значения;
- различные логарифмические или степенные функциональные преобразования натуральных шкал существенно приближают законы распределения исходных выборок к нормальному (остальные фигуры рис.5.2); средние арифметические сближаются с медианами, асимметрия и эксцесс становятся достаточно малыми;
- интересным показателем является "индекс плотности населения" $(N*B)^{1/2}$ [Дедю, 1990], в комбинированной содержательной форме учитывающий оба показателя (численность и биомассу), которые, по мнению ряда авторов, "конкурируют" между собой (что мы подробно обсуждали в разделах 2.3 и 4.3).

Выполним проверку простой статистической гипотезы о нормальном законе распределения представленных выборок с использованием критериев согласия Колмогорова и χ^2 Пирсона, а также критерия эксцесса Гири d. Результаты расчетов для логарифмированных рядов суммарных численностей $\ln(N_s)$ и биомассы $\ln(B_s)$, индекса плотности населения $\ln((N_s*B_s)^{1/2})$ и индекса Шеннона H представлены в таб. 5.2.

Проверка нулевой гипотезы с помощью критериев согласия свидетельствует о том, что предположение о нормальном законе распределения может быть однозначно принято только для рядов $\ln((N_s*B_s)^{1/2})$ и $\ln(N_s)$. В разделе 4.3 было отмечено, что некоторые авторы ставят в заслугу информационной мере Шеннона именно нормальный характер распределения, что, в принципе, подтверждается и нашими расчетами: эта гипотеза отвергается только критерием χ^2. Но, как видно из табл. 5.2, простое логарифмирование численностей часто приводит к лучшему результату (сравните на рис. 5.2 фиг. «е» и «ж» с «з»).

Таблица 5.1

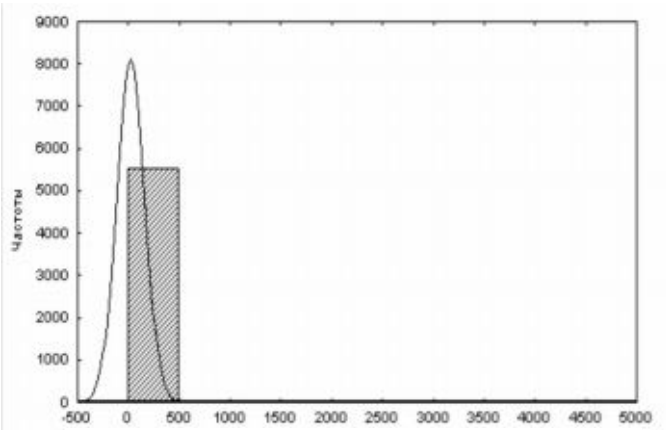
Основные описательные статистики по гидробиологическим показателям

Выборки	Минимум	Максимум	Медиана	Среднее	Стандарт-ное откло-нение	Асимметрия	Эксцесс
Численность по видам зообентоса (N) – 5788 измерений							
N	1	71680	80.00	427.33	1724.90	19.3	617
$N^{1/2}$	1	267.73	8.94	13.57	15.58	4.1	30
$N^{1/3}$	1	41.54	4.308	5.19	3.47	2.34	8.9
$\ln(N)$	0	11.18	4.38	4.43	1.69	0.358	0.051
$\ln(N+1)^{1/2}$	1	3.48	2.31	2.3	.37	-0.177	0.392
Биомасса по видам зообентоса (B) – 5578 измерений							
B	0.01	6000	0.07	11.48	152.59	24.6	749
$B^{1/2}$	0.10	77.46	0.264	0.84	3.2818	12	186
$B^{1/3}$	0.215	18.17	0.412	0.669	1.052	7.69	79.2
$\ln(B)$	-4.6	8.7	-2.66	-2.29	2.160	1.2	2.08
$\ln(B+10)^{1/2}$	2.3	4.3	2.71	2.75	0.37	0.82	0.58
Индекс плотности населения $(N*B)^{1/2}$ - 5573 измерений							
$(N*B)^{1/2}$	0.1	4174	2.47	21.35	136.7	19.4	463
$\ln((N*B)^{1/2})$	-2.3	8.33	0.91	1.12	1.68	0.68	0.23
Суммарная численность N_s и биомасса B_s в пробе – 514 измерений							
N_s	20	216800	2000	5039	11736	12.17	208.2
$\ln(N_s)$	2.93	12.3	7.6	7.47	1.57	-0.303	-.1781
B_s	0.01	12963.7	3.6	137.8	743	11.8	181
$\ln(B_s)$	-4.6	9.5	1.28	1.542	2.39	0.562	.5322
$(N_s* B_s)^{1/2}$	0.63	10608	84.4	380.4	947.6	6.19	49.3230
$\ln(N_s*B_s)^{1/2}$	-0.46	9.3	4.43	4.5	1.78	-0.06	-0.226
Информационный индекс Шеннона H - 507 измерений							
H	0.115	4.193	2.32	2.244	0.796	-0.284	-0.153

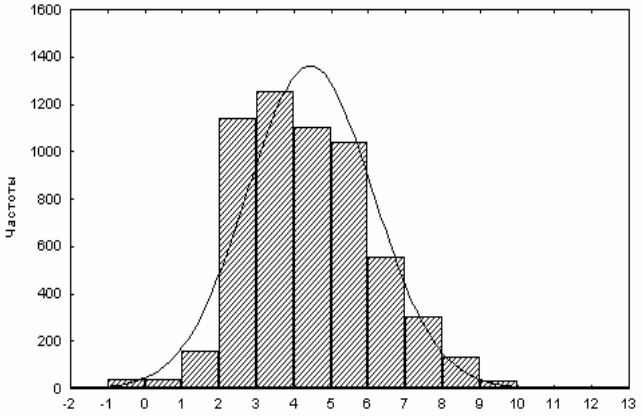
Таблица 5.2

Результаты проверки гипотезы о нормальности распределения некоторых функций от численности N_s и биомассы B_s зообентоса с помощью критериев согласия
(D_n – статистика Колмогорова, $\lambda = D_n \cdot n^{1/2}$, p_λ – вероятность, соответствующая λ, r – число степеней свободы, χ^2 – критерий Пирсона, $\chi^2_{0.05}$ – критическое значение критерия Пирсона при уровне значимости 0.05, p_χ – вероятность, соответствующая χ^2, p_d – вероятность, соответствующая критерию эксцесса Гири)

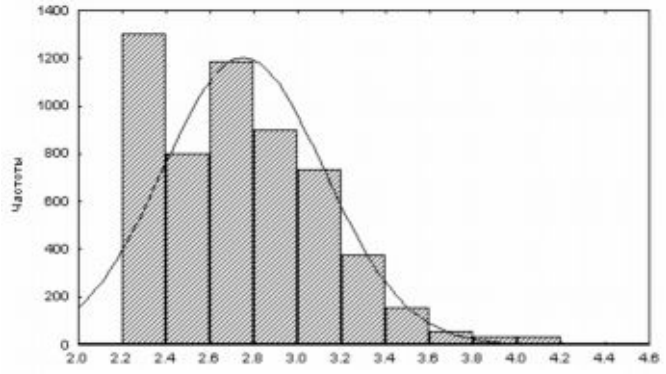
Выборка	D_n	λ	p_λ	r	χ^2	$\chi^2_{0.05}$	p_χ	p_d
$\ln(N_s)$	0.0367	0.833	0.492	8	15.32	15.5	0.053	0.48
$\ln(B_s)$	0.068	1.55	0.017	8	30.15	15.5	0.00019	~0
$(N_s*B_s)^{1/2}$	0.344	7.8	~0	8	1046	15.5	~0	~0
$\ln((N_s*B_s)^{1/2})$	0.0215	0.487	0.989	8	4.64	15.5	0.794	0.219
Индекс Шеннона H	0.0516	1.166	0.134	7	18.01	14.1	0.0119	0.202



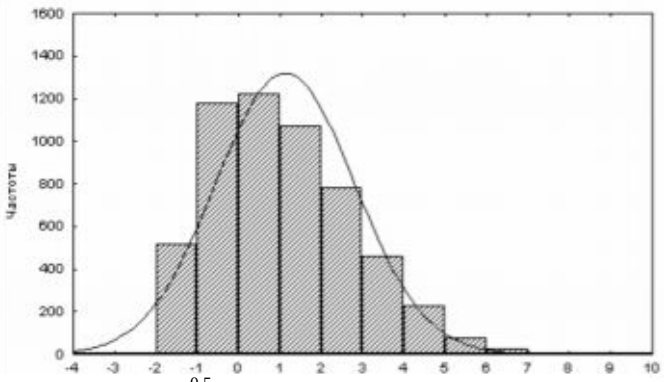
а) Численность экземпляров ***N***



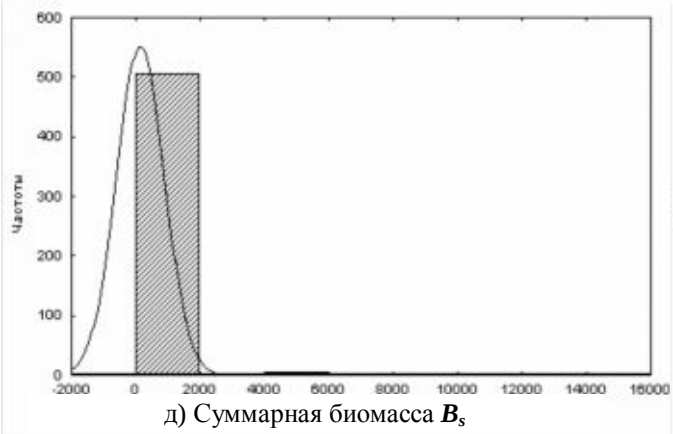
б) ln(***N***), ***N***- численность



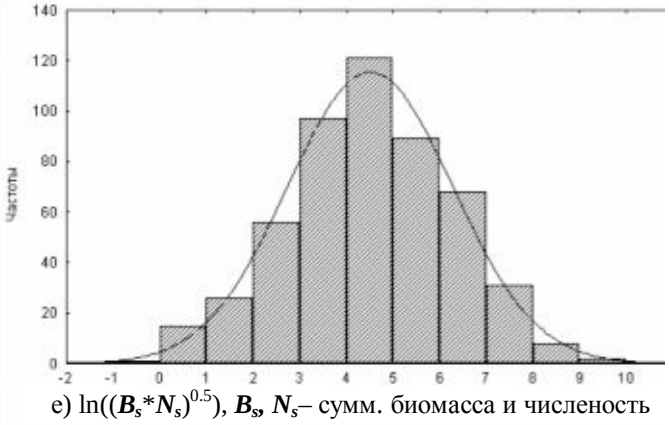
в) $(\ln(\boldsymbol{B}))^{0.5}$, ***B***- биомасса



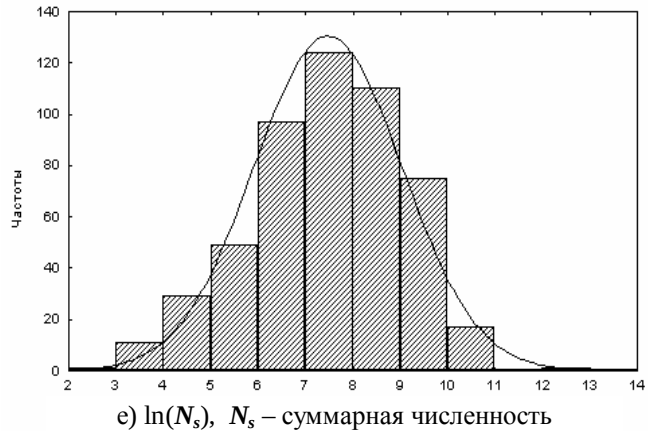
г) $\ln((\boldsymbol{B}*\boldsymbol{N})^{0.5})$, ***B*** – биомасса, ***N*** - численность



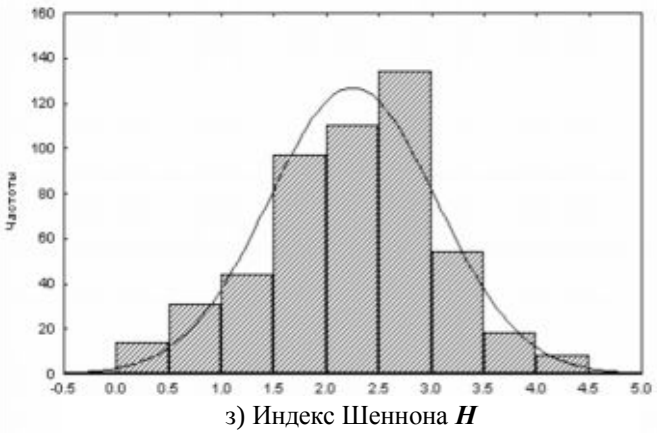
д) Суммарная биомасса $\boldsymbol{B_s}$



е) $\ln((\boldsymbol{B_s}*\boldsymbol{N_s})^{0.5})$, $\boldsymbol{B_s}$, $\boldsymbol{N_s}$– сумм. биомасса и численость



е) $\ln(\boldsymbol{N_s})$, $\boldsymbol{N_s}$ – суммарная численность



з) Индекс Шеннона ***H***

Рис. 5.2. Гистограммы распределения выборок из некоторых гидробиологических показателей и их функциональных преобразований

Для прочих рядов, не приведенных в табл. 5.2, гипотеза нормальности решительно отвергается, хотя тенденцию резкого снижения величин статистики Колмогорова $\lambda = D_n \cdot n^{1/2}$ при функциональном преобразовании численностей и биомассы отдельных видов бентоса легко проследить на рис. 5.3. Можно с большой уверенностью сказать, что эти ряды подчиняются логнормальному распределению: критерий Колмогорова для численности составляет D_n = 0.0048, для биомассы D_n = 0.0098, что позволяет с высоким уровнем значимости ($p \approx 1$) не отвергать гипотезу о логнормальном законе.

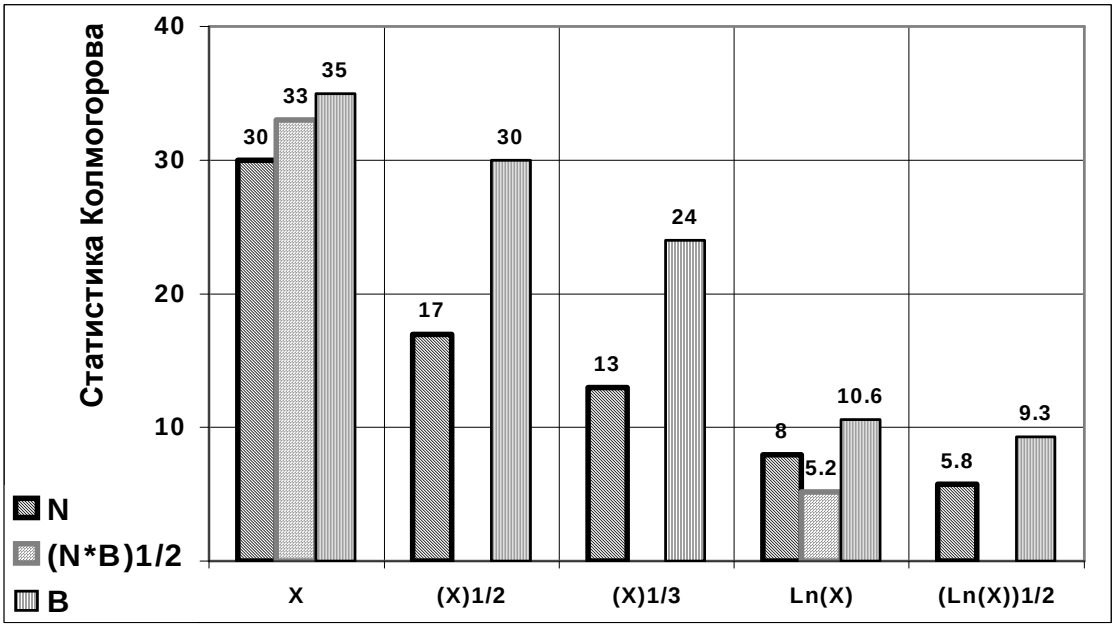
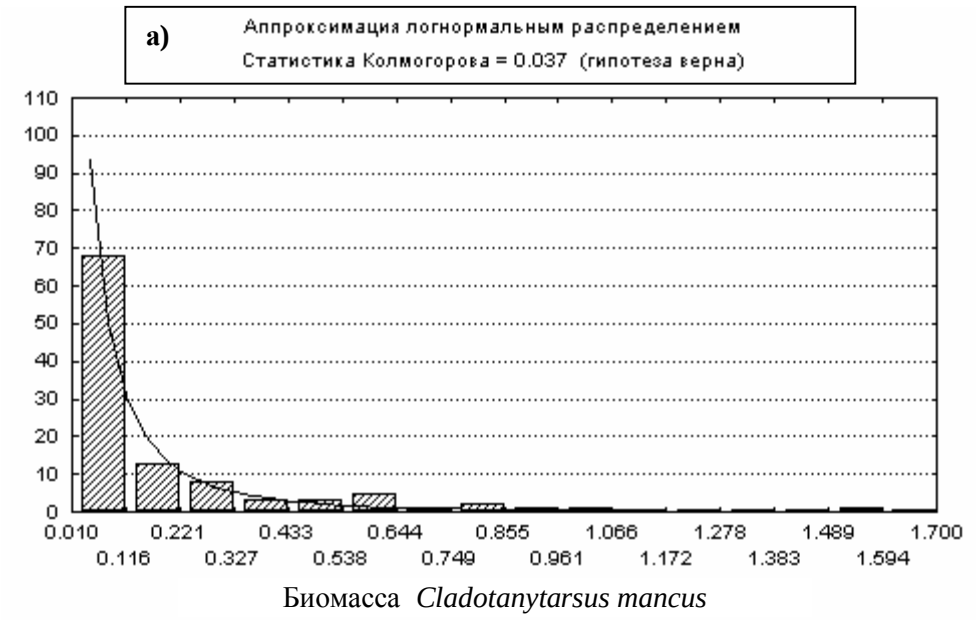


*Рис. 5.3. Снижение статистики Колмогорова $D_n \cdot n^{1/2}$ в результате функционального преобразования выборок численности **N** и биомассы **B** зообентоса*

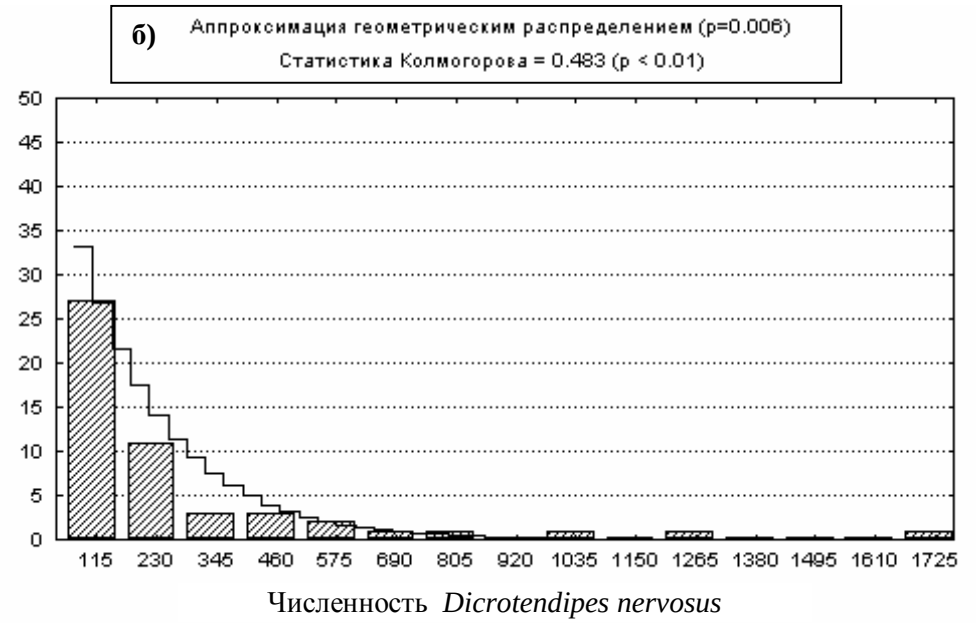
По литературным данным [Баканов, 2000а] для большинства видов бентосных организмов наиболее характерны отрицательное биномиальное или логнормальное распределения численности и биомассы, а при низком обилии – распределение Пуассона. Выполненные нами расчеты показали, что распределение численности и биомассы большинства видов с высоким уровнем значимости подчиняются логнормальному закону распределения (см. рис. 5.4 фиг. «а»). Проверка нулевой гипотезы при подборе остальных теоретических функций распределения дала однозначно отрицательные результаты. В частности, для численности *Dicrotendipes nervosus* оказалась неудачной попытка аппроксимации геометрическим распределением (см. рис. 5.4 фиг. «б»), в то время, как верна гипотеза о логнормальном распределении (статистика Колмогорова равна 0.029, $p \approx 0.85$).

В некоторых работах [Gray, 1981; Мокеева, Межов, 1986] делаются выводы, о том что, параметры статистических распределений меняются при изменении условий обитания животных. Например, вследствие загрязнения или ином ухудшении условий, асимметрия увеличивается, а кривая распределения имеет несколько пиков, в то время как при улучшении условий обитания кривая имеет более ровный характер, а полимодальность отсутствует.

Нам не удалось ни подтвердить, ни опровергнуть эти суждения, поскольку осталось непонятным, относительно какого закона распределения следует оценивать коэффициент асимметрии, а надежных критериев оценки полимодальности найти не удалось. Действительно, визуально можно обнаружить на гистограммах рис. 5.5 некоторые проявления полимодальности, однако, насколько имеющиеся пики статистически значимы, еще следует доказать, используя механизм проверки гипотез.



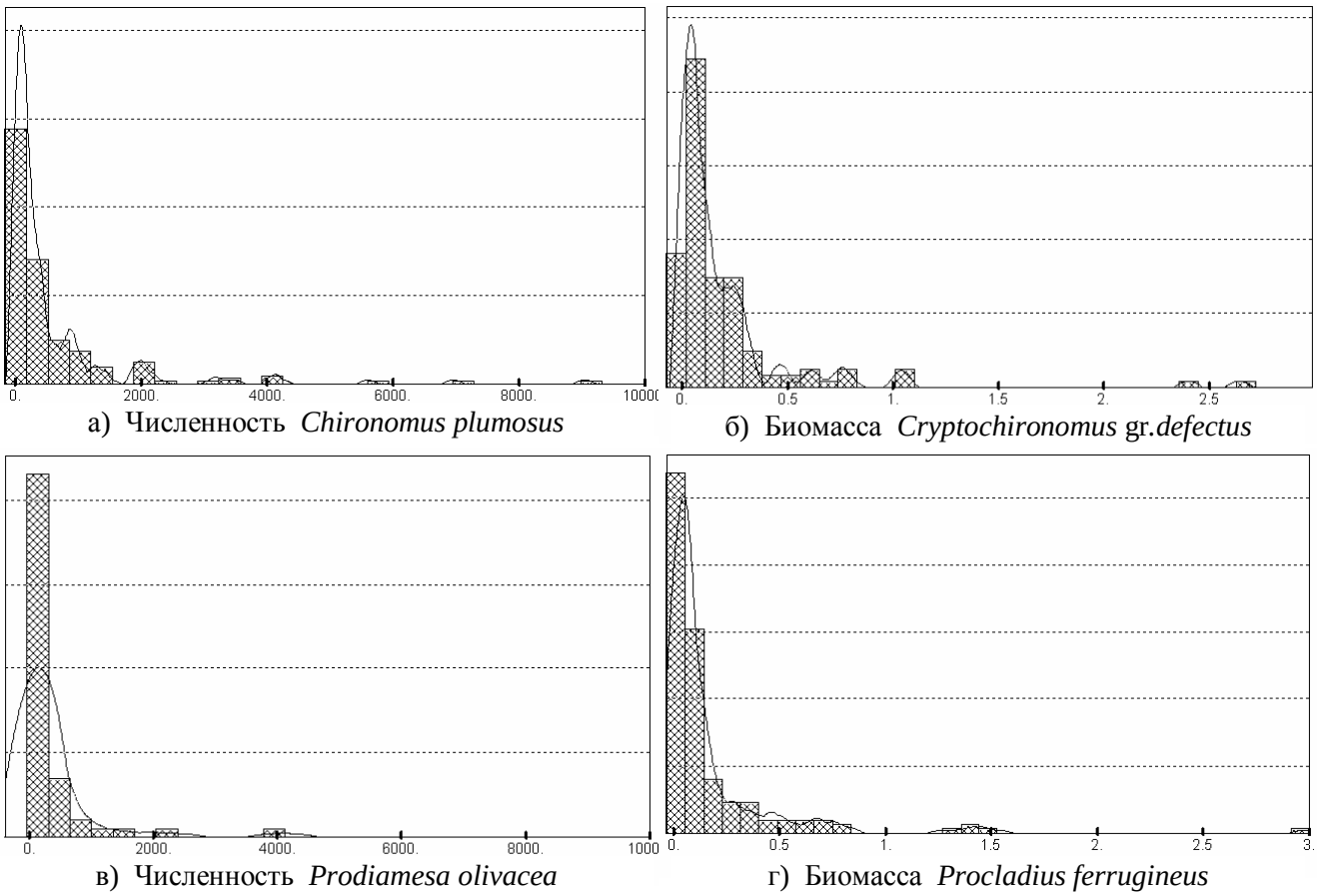
Биомасса *Cladotanytarsus mancus*



Численность *Dicrotendipes nervosus*

Рис 5.4. Подбор функции распределения показателей обилия некоторых видов бентоса

В то же время, с характером распределения можно, при желании, связать ряд выводов об экологии вида, глубине "экологической ниши" и проч. Например, сравнивая закономерности распределения численностей фиг. «а» и «в» на рис. 5.5, можно усмотреть, что экологический "спектр" *Chironomus plumosus* значительно шире, чем *Prodiamesa olivacea*, что объясняется широкой экологической валентностью первого вида и особенностями биологии хирономид [Зинченко, 2002]. Аналогично, из гистограммы распределения биомассы личинок хирономид можно заметить различия, связанные с особенностями питания хищных *Cryptochironomus* gr. *defectus* («б») и факультативных фитофагов *Procladius ferrugineus* («г»). Впрочем, кое-кому может показаться, что подобные упражнения сильно напоминают "гадания на кофейной гуще".



а) Численность *Chironomus plumosus*

б) Биомасса *Cryptochironomus* gr.*defectus*

в) Численность *Prodiamesa olivacea*

г) Биомасса *Procladius ferrugineus*

Рис 5.5. Формы кривых распределения обилия некоторых видов бентоса

5.2. Задача об однородности выборок: однофакторный дисперсионный анализ

Формулировка задачи

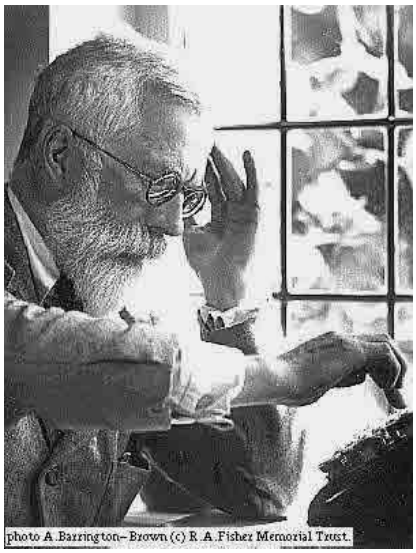
Пусть имеется выборка из n значений $X_1, X_2,\ldots,X_n$ некоторого измеренного количественного гидробиологического показателя, который будем называть ***откликом***. Каждому значению отклика поставлен в соответствие некоторый качественный признак (фактор), на основании которого общую выборку можно сгруппировать в частные независимые случайные выборки: если фактор имеет k уровней $A_1, A_2,..., A_k$, то при каждом уровне A_j фактора, $j = 1, ..., k$, имеется n_j измерений. Необходимо:

- проверить гипотезу о принадлежности всех выборок одной генеральной совокупности, т.е. нельзя ли объяснить расхождение наблюдаемых значений для разных значений фактора случайными обстоятельствами или влиянием неучтенных факторов;
- если нулевая гипотеза отвергается, то оценить степень влияния фактора;
- выделить все пары групп i - j, $i = 1,2,\ldots k$, $j = 1,2,\ldots k$, $i \neq j$, которые имеют между собой статистически значимые различия.

В рамках этой задачи изучается влияние, которое оказывает на количественный результат измерения тот или иной фактор, который имеет принципиально нечисловую природу и измерен в шкале наименований ("левый берег", "правый берег", "русло"), либо в порядковой шкале (месяцы или годы отбора проб).

Опыт применения статистики показывает, что условия формирования выборочных совокупностей сказываются, в первую очередь, на изменении положения центра распределения измеряемой характеристики на числовой оси – т.е. сдвиг медианы или среднего значения. При больших

различиях в условиях отбора проб наряду с дрейфом центра распределения изменяется и его разброс, т.е. дисперсия. Масштаб и форма распределения обычно остаются практически неизменными. Поэтому задача о степени однородности измеренной случайной величины обычно сводится к оценке статистической значимости различия между средними двух (или нескольких) групп.



Рональд Эйлмер ФИШЕР
(**R.A.Fisher** 1890-1962)
английский математик и генетик, основоположник современной прикладной статистики

В дисперсионном анализе ANOVA (англ. Analysis of variance), фундаментальная концепция которого была предложена Р. Фишером в 1920 г., процедура сравнения средних основывается на разложении суммы квадратов выборочной вариации. При этом принимается ряд следующих априорных допущений:

- все наблюдения принадлежат некоторому сдвиговому семейству распределений;
- эти распределения внутри каждой группы аппроксимируются нормальной кривой Лапласса – Гаусса (предположение о нормальности распределения);
- дисперсии и ковариации зависимой переменной в разных группах одинаковы (предположение об однородности дисперсий).

Рекомендуемая литература: [Урбах, 1963; Плохинский, 1970, 1982; Закс, 1976; Джонсон, Лион, 1980, 1981; Любищев, 1986; Дрейпер, Смит, 1986, 1987].

Математический лист

Пусть измеряемая величина x есть результат действия фактора A и некоторой случайной составляющей ε, от фактора не зависящей и отражающей внутреннюю изменчивость наблюдений: $x = f(A) + \varepsilon$.

Примем следующие предположения:

- значение отклика x_{ij} можно представить в виде суммы вкладов влияния уровней фактора, т.е. приемлема аддитивная модель:

$$x_{ij} = a_j + \varepsilon_{ij}, \qquad i = 1, ..., n_j, \tag{5.21}$$

  где $a_j = f(A_j)$ – неслучайные неизвестные величины, определяющие влияние каждой категории фактора j, $j = 1,2,\ldots k$. Если влияние фактора отсутствует, то все a_j равны между собой;
- случайная составляющая ε распределена нормально $N(0, \sigma^2)$ с дисперсией σ^2.

Основное соотношение дисперсионного анализа можно записать как разложение общей или "полной" суммы квадратов отклонений (Q) отдельных наблюдений от общей средней на две составляющие:

$$Q = Q_A + Q_e, \tag{5.22}$$

где Q_A – вариабельность за счет влияния исследуемого фактора или сумма квадратов отклонений "между группами";

Q_e – остаточное рассеивание, обусловленное случайными факторами, или сумма квадратов отклонений "внутри групп".

Долю объясненной суммы квадратов в полной вариации отклика, рекомендованную для биологических исследований Н.А. Плохинским, связывают с *корреляционным отношением Пирсона* (см. далее разделы 5.4 и 7.2):

$$\eta^2_A = Q_A / (Q_A + Q_e). \tag{5.23}$$

Чем ближе значение η^2_A к единице, тем большим принимается влияние группировочного фактора A на результативный признак X. Можно рассчитать корреляционное отношение и для остаточной суммы квадратов

$$\eta^2_e = Q_e / (Q_A + Q_e),$$

тогда очевидно, что $\eta^2_A + \eta^2_e = 1$. Корреляционные отношения иногда использовать проще, чем суммы квадратов, поскольку они нормированы и не зависят от шкалы измерений. Однако они не могут быть использованы для проверки гипотез, поскольку не включают число степеней свободы.

Проверка нулевой гипотезы.

Пусть имеется k выборок объемами $n_1, ..., n_k$, $\sum_{j=1}^{k} n_j = N$. По каждой из выборок методом наибольшего правдоподобия оценим групповые средние a_j и групповые дисперсии σ^2_j:

$$\bar{a}_j = \bar{x}_j \equiv \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}, \qquad s_j^2 = \frac{1}{n_j}\sum (x_{ij} - \bar{x}_{\bullet j})^2, \tag{5.24}$$

а затем оценим σ^2 по всем выборкам:

$$\sigma^{2*} = \frac{Q_e}{N-k} = \frac{1}{N-k}\sum_{j=1}^{k} n_j s_j^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{j=1}^{n_j} (x_{ij} - \bar{x}_{\bullet j})^2 . \tag{5.25}$$

Сформулируем гипотезу H об отсутствии влияния фактора A, которая справедлива, если групповые средние равны между собой: $a_1 = a_2 = ... = a_k$.

Статистика, рассчитанная по формуле (5.25) несмещенно оценивает общую дисперсию σ^2, независимо от того, верна или нет гипотеза H. Другую оценку для σ^2 построим, используя только значения групповых средних $\hat{a}_1, \hat{a}_2, ..., \hat{a}_k$:

$$\hat{a} = \frac{1}{N}\sum_{j=1}^{k} n_j \hat{a}_j , \qquad \sigma^{2**} = \frac{Q_A}{k-1} = \frac{1}{k-1}\sum_{j=1}^{k} n_j (\bar{a}_j - \bar{a})^2 . \tag{5.26}$$

Если H верна, то $M\bar{a}_j = a$, $D\bar{a}_j = \sigma^2 / n_j$, для всех $j = 1, 2, \ldots, k$.

Из теоремы о совместном распределении оценок среднего и дисперсии нормальной совокупности следует, что статистики $(N - k)\sigma^{2*}$ и $(k - 1)\sigma^{2**}$ независимы и распределены как $\sigma^2\chi^2_{N-k}$ и $\sigma^2\chi^2_{k-1}$, соответственно, и потому их отношение

$$F_H = \frac{\sigma^{2**}}{\sigma^{2*}} = \frac{\sigma^2\chi^2_{k-1}/(k-1)}{\sigma^2\chi^2_{N-k}/(N-k)} , \tag{5.27}$$

если гипотеза H верна, имеет F-распределение Фишера, зависящее только от двух степеней свободы $(k-1)$ и $(N - k)$.

Пусть $F_{крит} = F(1-\alpha; k-1, N-k)$ – квантиль уровня $(1-\alpha)$ F-распределения с $k-1$ и $N-k$ степенями свободы, где α – выбираемый уровень значимости. Если гипотеза H не верна, то σ^{2**} имеет тенденцию к увеличению за счет разброса средних a_j, приводящую к тому, что F_H принимает слишком большое значение, превышающее критическое

$$F_H > F_{крит} . \tag{5.28}$$

В этом случае гипотеза об отсутствии влияния фактора A отклоняется, и следует считать, что среди подмножества внутригрупповых средних $a_1, a_2, \ldots, a_k$ имеются хотя бы два не равных между собой.

Если же (5.28) не выполняется, то это означает, что наблюдения не противоречат гипотезе об отсутствии влияния фактора и условие (5.28) может быть записано иначе:

$$P\{F \geq F_H\} < \alpha , \tag{5.29}$$

где F – случайная величина, распределенная по закону Фишера.

Таким образом, влияние фактора признается достоверным, если средняя сумма квадратов отклонений между группами $S^2_A = Q_A/(k-1)$ значимо превышает средние квадраты внутри групп $S^2_e = Q_e/(N - k)$.

Оценка влияния фактора

Если гипотеза H_o оказалась несовместимой с наблюдениями, есть основания для обсуждения значений параметров $a_1, a_2, \ldots, a_j, \ldots, a_k$, оценками которых являются внутригрупповые средние. Отношение $\frac{\hat{a}_j - a_j}{\sigma^*}\sqrt{n_j}$ подчиняется распределению Стьюдента с $N-k$ степенями свободы, и если $t_{1-\alpha} = t(1-\alpha, N-k)$ – квантиль уровня (1 - α) этого распределения, то получаем доверительный интервал для a_j с уровнем доверия (1 - 2α):

$$\hat{a}_j \pm \frac{\sigma^*}{\sqrt{n_j}} t_{1-\alpha} \quad . \tag{5.30}$$

Для оценки влияния фактора используют ряд состоятельных статистик, предложенных различными авторами:

- *сила влияния фактора по Снедекору* [Гинзбург, 1969]

$$h_x^2 = \frac{S_A^2 - S_e^2}{S_A^2 + (\nu - 1)S_e^2} \; ; \qquad \nu = \frac{1}{k-1}\left(N - \frac{\sum_{j=1}^{k} n_j^2}{N}\right) \; ; \tag{5.31}$$

- *показатель Миллса-Лукомского* [Урбах,1963] $\mathrm{ML} = 1 - \frac{Q_e\,(N-1)}{(Q_e + Q_A)(N-k)}$; (5.32)

- *доля дисперсии, объяснимая фактором* $\eta = \frac{\frac{k}{N}\left(S_A^2 - S_e^2\right)}{S_e^2 + \frac{k}{N}\left(S_A^2 - S_e^2\right)}$. (5.33)

Аргументы "за" и "против" использования каждого из этих критериев были проанализированы нами ранее [Розенберг, Долотовский, 1988]. Здесь важно подчеркнуть другое – вычисление значений перечисленных статистик, в отличие от статистики Фишера, не дает ответа на главный вопрос: какой альтернативной гипотезе о влиянии показателя следует отдать предпочтение.

С точки зрения обоснованности применения метода дисперсионного анализа в целом (см. исходные допущения метода), наряду с тестами на нормальность распределения выборок, для оценки эффектов вариации наблюдений на разных уровнях факторов целесообразно использовать критериальные статистики проверки однородности групповых дисперсий, такие как критерии Бартлетта и Кохрена:

- ***M*-критерий *Бартлетта*** служит для проверки нулевой гипотезы о равенстве дисперсий нескольких генеральных совокупностей и вычисляется по формулам:

$$M = \frac{\ln 10}{c}\left[(N-k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2\right], \tag{5.34}$$

где $c = 1 + \frac{1}{3(k-1)}\left(\sum_{k=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)$; $\quad S^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\cdot s_i^2$;

- ***G*-критерий *Кохрена*** используется для тех же целей и вычисляется как

$$G = \frac{\max\limits_{1 \le i \le k} s_i^2}{\sum\limits_{i=1}^{k} s_i^2}, \tag{5.35}$$

где для обеих формул k – число выборок ($k > 2$); n_i – численность i-й выборки, $i = 1,2,\ldots k$; s_i^2 – выборочная дисперсия i-й выборки.

При больших объемах выборок M-статистика Бартлетта имеет асимптотическое χ^2 распределение с числом степеней свободы $(k - 1)$, а способ расчета критических значений G-статистики Кохрена по аппроксимационным формулам представлен в статистических справочниках. Заметим, однако, что критерии Кохрена и Бартлетта также весьма чувствительны к отклонению от предположения, что нормированные оценки дисперсии у сравниваемых выборок подчиняются распределению χ^2.

Проверка гипотез о равенстве групповых средних

Если гипотеза H о равенстве средних отклоняется, то следует определить, по каким именно уровням фактора групповые средние значимо различаются. В своем простейшем варианте эта задача сводится к попарному сравнению средних значений двух произвольных выборок.

Процедуру сравнения двух произвольных выборочных средних a_1 и a_2, основанных на выборках размерностью n_1 и n_2 и имеющих оценки дисперсий s^2_1 и s^2_2 соответственно, выполняют с помощью t-статистики, численно равной нормированной разности групповых средних. Известно, что это распределение близко к t-распределению Стьюдента (5.36) с числом степеней свободы ν, равным округленному до ближайшего целого выражению (5.36а):

$$t = \frac{|a_1 - a_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}; \qquad (5.36)$$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}. \qquad (5.36а)$$

Если для вычисленной t-статистики выполняется неравенство:

$$t > t_{1-\alpha, \nu}, \qquad (5.37)$$

где $t_{1-\alpha, \nu}$ – критическое значение распределения Стьюдента при ν степенях свободы, то нулевая гипотеза о равенстве средних отвергается с уровнем значимости α.

Однако простейший метод попарного сравнения групповых средних по Стьюденту не позволяет указать вероятность, с которой одновременно выполняются несколько неравенств типа (5.37), и поэтому считается неполным. Метод множественных сравнений Г. Шеффе позволяет получать совместные доверительные интервалы для контрастов.

Линейным контрастом L называется любая линейная комбинация выборочных средних a_j с произвольными коэффициентами $c_1, c_2, \ldots, c_k$,

$$L = \sum_{j=1}^{k} c_j a_j \; ; \quad \sum_{j=1}^{k} c_j = 0, \qquad (5.38)$$

вычисленная для k независимых выборок с неизвестными, но равными дисперсиями.

Пусть для некоторого контраста L: оценка среднего – $\tilde{L} = \sum_{j=1}^{k} c_j \hat{a}_j$, а оценка дисперсии – $S^2_L = \sigma^{2*} \sum_{j=1}^{k} \frac{c_j^2}{n_j}$. Зафиксируем произвольное число r контрастов $L^{(1)}, L^{(2)}, \ldots, L^{(r)}$. Можно показать, что одновременно для всех $\tilde{L}^{(1)}, \ldots, \tilde{L}^{(r)}$ с вероятностью $(1 - \alpha)$ выполняются соотношения:

$$\left| L^{(m)} - \tilde{L}^{(m)} \right| < S_L^{(m)} \sqrt{kF(1 - \alpha, k, N - k)}, \quad m = 1, \ldots, r. \qquad (5.39)$$

Неравенства (5.39), использующие критические значения F-распределения, позволяют сделать вывод о достоверности одновременного влияния любых интересующих нас комбинаций уровней факторов.

Вычисления статистики *критерия Шеффе* при проверке нулевой гипотезы $L = L_0$ производятся по следующей формуле:

$$t_S = \frac{\left|\sum_{i=1}^{k} c_i a_i - L_0\right|}{\sqrt{M \sum_{i=1}^{k} \frac{c_i^2}{n_i}}}, \qquad \text{где} \quad M = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - a_i)^2 \text{ – средний квадратичный остаток,}$$

N – общая численность, k – число выборок ($k > 2$); n_i – численность i-й выборки, $i = 1,2,\ldots k$; a_i – среднее i-й выборки. Уровень значимости вычисляется на основе функции плотности F-распределения с параметрами (k -1) и N.

Гипотезы о простых линейных контрастах могут быть также проверены с помощью двух других методов: с использованием *критерия множественных сравнений Дункана* t_D и *критерия достоверно значимой разности Тьюки* t_T, которые по своим предпосылкам применения и технике вычислений очень похожи на методику Шеффе. Вычисления этих критериев проводят по следующим формулам при тех же обозначениях , что и t_S:

$$t_D = \frac{\left|\sum_{i=1}^{k} c_i a_i - L_0\right|}{\sqrt{\frac{M}{2} \sum_{i=1}^{k} \frac{c_i^2}{n_i}}} \quad ; \qquad t_T = \frac{\left|\sum_{i=1}^{k} c_i a_i - L_0\right|}{0.5 \sum_{i=1}^{k} |c_i| \sqrt{M}} \sqrt{m} \quad , \tag{5.40}$$

а для нахождения p-значений критериев Дункана и Тьюки сформированы таблицы распределений "стьюдентизированного размаха".

Формирование однородных групп средних значений возможно также с использованием *LSD-критерия*. Для этого упорядочивают k средних значений выборочных групп по их возрастанию и проверяют, больше ли разность Δ между соседними величинами, чем наименьшая значимая разность (least significant difference, LSD):

$$LSD_{(a,b)} = t_{n-k,\alpha} \sqrt{s_e^2 \left(\frac{n_a + n_b}{n_a n_b} \right)} = \sqrt{s_e^2 \left(\frac{n_a + n_b}{n_a n_b} \right) F(1; N-k; \alpha)} \; . \tag{5.41}$$

При $\Delta < LSD$ или $\Delta_{(a, b)} < LSD_{(a, b)}$ нет оснований для отклонения нуль-гипотезы H_o о равенстве соседних средних значений.

Результаты расчетов

Рассмотрим влияние типа грунта на общие показатели обилия хирономид. На основе данных из базы по малым рекам Самарской области сформируем выборку из численности N экземпляров отдельных видов Chironomidae на м^2 и их биомассы B (г/м^2). Выполним группировку суммарных значений N_s и B_s для каждой пробы и вычислим дополнительно преобразованные выборки из прологарифмированных значений этих показателей обилия и индекса плотности населения $(N_s * B_s)^{1/2}$. Качественные оценки грунтов на дне водоемов в местах отбора проб представим в виде 6 категорий:

1- грунты преимущественно без растительных остатков (гравий, камни, песок);
2- заиленный песок;
3- серые илы и заиленная почва;
4- илы с почвой и растительными остатками;
5- глинистая почва;
6- черный ил и ил с запахом сероводорода.

Выполним проверку нулевой гипотезы об отсутствии влияния типа грунта на общее обилие хирономид, ориентируясь на p-значения, соответствующие вычисленным F-статистикам. Результаты, приведенные в табл. 5.3, свидетельствуют о том, что из всех анализируемых выборок значимые отличия (p-значения меньше 0.05) показателей обилия для различных грунтов имеют место только для $\ln(N_s)$. Весьма близка к этому выводу оказалась и выборка $\ln((N_s * B_s)^{1/2})$.

Таблица 5.3

Результаты проверки по критерию Фишера гипотезы о влиянии типа грунтов на обилие Chironomidae

Показатели для анализа	Источник вариации	Сумма квадратов	Степеней свободы	Средние квадраты	***F***-отношения	***p***-значения
1.Численность N_s	Между группами	$2.49*10^8$	5	$0.49*10^8$	0.41	0.84
	Внутри групп	$5.3*10^{10}$	439	$1.2*10^8$		
	Итого	$5.33*10^{10}$	444			
2. Биомасса B_s	Между группами	59200	5	11840	0.61	0.6914
	Внутри групп	8504210	439	19371		
	Итого	8563400	444			
3.ln (N_s)	Между группами	54.7	5	10.9	3.07	0.0099
	Внутри групп	1567.5	439	3.57		
	Итого	1622.2	444			
4. ln(B_s)	Между группами	22.96	5	4.59	1.65	0.144
	Внутри групп	1218.4	439	2.77		
	Итого	1241.4	444			
5. ln((N_s*B_s)$^{1/2}$)	Между группами	31	5	6.2	2.16	0.057
	Внутри групп	1263	439	2.87		
	Итого	1294	444			

Отличия в основных статистических выводах между параллельными выборками из натуральных и прологарифмированных значений могут быть вполне объяснимы, если вспомнить основное предположение дисперсионного анализа о нормальном характере распределения анализируемых выборок.

Кроме описанных в разделе 5.1 критериев согласия, для проверки нормальности закона распределения часто бывает достаточно визуальной оценки, основанной на удивительной способности человеческого глаза обнаруживать сходство геометрического образа с прямой линией. Для этого используют построение графиков на "*нормальной вероятностной бумаге*", где на оси абсцисс откладывают значения x_i, а на оси ординат – значения функции, обратной интегралу вероятностей $\Phi(z_i) = (x_i - a)/\sigma$. В случае нормального распределения полученные точки ложатся на прямую, как это имеет место для прологарифмированных значений и никак не характерно для натуральных значений (см. рис. 5.6)

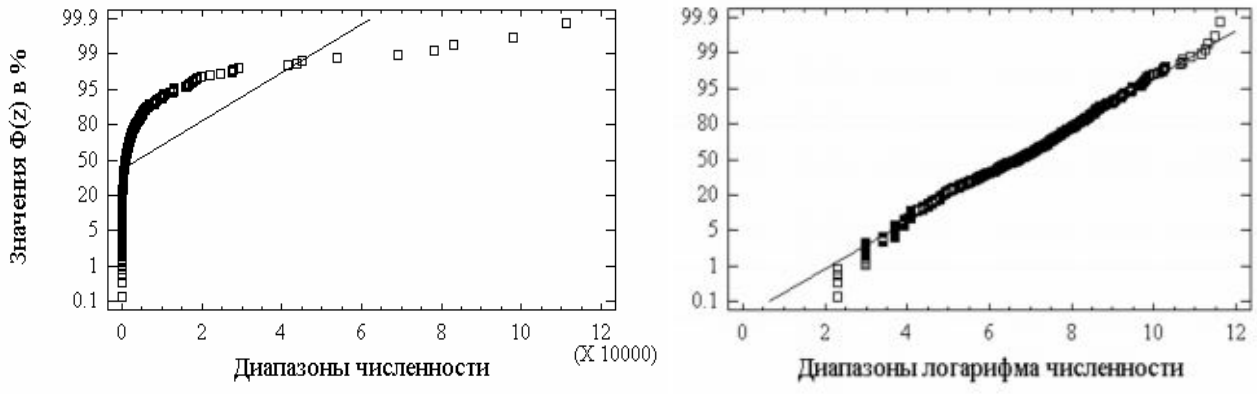


Рис.5.6. Эмпирические функции распределения общей численности хирономид на нормальной вероятностной бумаге

Как подтверждается результатами, приведенными в табл. 5.3, ***F***-критерий очень неустойчив к отклонению выборок от нормальности (подробнее см. [Lindman, 1974]). Если эксцесс существенно больше 0, что имеет место в случае с натуральными значениями численности и биомассы, то значение статистики F может стать очень маленьким. Нулевая гипотеза при этом не может

быть отвергнута, хотя она и не верна. Ситуация обычно меняется на противоположную, если эксцесс меньше 0.

Для проверки другого исходного предположения дисперсионного анализа об однородности групповых дисперсий воспользуемся критериями Кохрена и Бартлетта, представленными в табл. 5.4. Очевидно, что выборки 1 и 2 из натуральных значений показателей обилия не соответствуют требованиям анализа, в то время как для выборок из прологарифмированных значений гипотеза о равенстве групповых дисперсий не отвергается.

Таблица 5.4

Результаты проверки однородности выборочных дисперсий обилия Chironomidae *в зависимости от типа грунтов*

Выборки для анализа	Тест по методу Кохрена		Тест по методу Бартлетта	
	G-критерий	***P***-значение	***M***-критерий	***P***-значение
1. Численность N_s	0.406	0.0	1.381	0.0
2. Биомасса B_s	0.429	0.0	1.159	0.0
3.ln (N_s)	0.210	0.291	1.008	0.602
4. ln(B_s)	0.206	0.396	1.009	0.538
5. ln((N_s*B_s)$^{1/2}$)	0.200	0.571	1.010	0.492

Основываясь на этих рассуждениях, всем предпосылкам дисперсионного анализа отвечает вывод о том, что тип грунта оказывает существенное влияние на численность личинок хирономид в реках, но не влияет на их общую биомассу (при представлении показателей обилия в логарифмических шкалах).

Выполним анализ групповых средних, значения которых для каждой категории грунта представлены в табл. 5.5.

Таблица 5.5

Распределение выборочных значений показателей обилия хирономид для различных категорий грунта

Показатели для анализа	Категория грунта						Общее среднее
	1	**2**	**3**	**4**	**5**	**6**	
Количество проб	86	99	127	95	23	15	445
Численность N_s	3683	4912	2985	4376	3163	3455	3870
Биомасса B_s	2.274	2.941	1.954	2.333	2.033	2.708	2.346
ln (N_s)	5.923	6.788	6.776	6.812	6.225	6.760	6.593
ln(B_s)	-0.605	-0.141	-0.332	-0.132	-0.515	-1.080	-0.334
ln((N_s*B_s)$^{1/2}$)	2.659	3.323	3.222	3.340	2.855	2.840	3.129

Более детально параметры семейства сдвиговых распределений представляются на диаграммах размаха типа «прямоугольники и усы» (box-whisker), показанных на рис. 5.7. Вокруг медианы для каждой выборки рисуется прямоугольник, замыкающий 50% наблюдений; среднее значение отмечено знаком «+», а отрезки, концы которых расположены вне прямоугольника, определяют вариационный размах (но не более 1.5 межквартили с каждой стороны).

Представленные данные позволяют сделать следующие выводы о тенденциях изменения обилия хирономид в зависимости от типа грунтов (шкалы – логарифмические):

- численность и биомасса достигают своих максимальных значений на грунтах категории 2-4 (заиленный песок с растительными остатками) и существенно уменьшаются как на "чистых" грунтах категории 1, так и на вязких, глинистых илах категории 5;
- весьма неоднозначно поведение этих показателей на "черных" анаэробных илах категории 6, где при большой медиане общей численности, биомасса достигает своего минимального значения, что, возможно, объясняется развитием видов с небольшим индивидуальным весом особей;
- индекс плотности населения $(N_s*B_s)^{1/2}$ предоставляет более обобщенные выводы об основных закономерностях развития гидробионтов на грунтах различных категорий, чем отдельные показатели численности или биомассы, учитывая динамику их обоих.

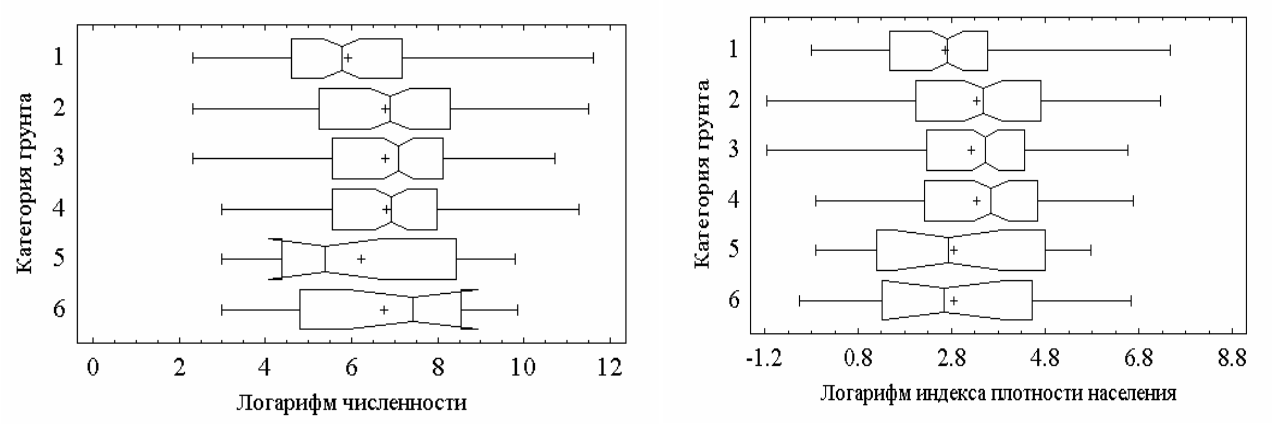


Рис 5.7. Диаграммы размаха типа «прямоугольники и усы»

Естественно, что суммарные показатели численности и биомассы хирономид позволяют оценить лишь самые общие тенденции влияния фактора, т.к. на нижних таксономических уровнях поведение различных групп организмов по отношению к типу грунтов будет далеко неоднозначно. Для оценки влияния фактора на различные трофические группы Chironomidae, разделим все 228 встретившихся видов на 6 групп в зависимости от способа и характера питания, сформируем дополнительные выборки из прологарифмированных значений суммарной численности по каждой группе для каждой пробы и выполним однофакторный дисперсионный анализ влияния категории грунта. Полученные результаты представлены в табл. 5.6.

Таблица 5.6

Однофакторный дисперсионный анализ влияния категории грунтов на логарифм численности трофических групп хирономид

Трофические группы	Количество видов	Общее среднее	Частота нулевого класса, %	Дисперсия фактора S^2_A	Остаточная дисперсия S^2_e	***F***-отношения	***p***-значения
Хищники хвататели	30	3.64	28.1	17.42	6.6	2.64	0.023
Всеядные собиратели +хвататели	4	2.50	50.5	72.39	6.8	10.51	0.0
Детритофаги собиратели	60	4.13	24.7	18.6	7.8	2.38	0.038
Сестонофаги+детритофаги фильтраторы	19	1.35	69.4	2.77	4.87	0.57	0.722
Детритофитофаги собиратели + фильтраторы	14	2.97	43.5	56.62	7.8	7.22	0.0
Фитодетритофаги собиратели	101	2.89	42.9	66.07	7.28	9.08	0.0

При реализации этих расчетов мы столкнулись с фундаментальной проблемой, не решенной за 100 лет количественной гидробиологии: следует ли учитывать в составе выборок нулевые значения, т.е. так называемую "частоту нулевого класса", свойственную по теории статистики любому статистическому распределению. Как видно из табл. 5.6, из 445 проб, в которых встречались хирономиды, виды выделенных трофических групп далеко не обязательно присутствовали в каждой пробе, и исходная анализируемая матрица размерностью 445x6 оказалась заполнена данными примерно наполовину, причем "пустые" клетки можно трактовать и как нули, и как отсутствующие значения.

В.Н. Максимов с соавторами [1999] считают, что *«…в действительности, значение численности, равное нулю, совсем не обязательно говорит об отсутствии вида в пробе. При существующих методиках измерения численности вполне вероятно, что ни один экземпляр вида с низким обилием просто не попадает в поле зрения исследователя»*. Иными словами, нельзя говорить об отсутствии вида в пробе, а только о чрезвычайно низкой, близкой к нулю удельной численности экземпляров.

К сожалению, такая статистически обоснованная концепция "нулевой" численности для видов, не встретившихся в пробе, приводит к серьезным практическим сложностям. Во-первых, становится неопределенным понятие размера выборки, ибо, если количество проб, в которых ***встретился*** некоторый вид всегда счетно и определено, то число проб, в которых он ***не встретился***, сильно зависит от точки зрения экспериментатора (вспомним, например, Александра Ширвиндта, пробующего перечислить роли, которые он ***не*** *сыграл*). Во-вторых, на понятии частоты встречаемости видов основаны индексы, являющиеся общеупотребительными в классической количественной гидробиологии, например, индекс доминирования Палия-Ковнацкого (см. раздел 4.3):

$$D_i = \frac{m_i \cdot \sum_k N_{ik}}{M \cdot \sum_k N_{sk}} 100\% \quad , \qquad (5.42)$$

где M – общее количество серий измерений, m_i – количество серий наблюдений, в которых встретился i-й вид, N_{ik} – численность вида i в k-м измерении, N_{sk} – сумма численностей всех видов в k-м измерении. Если принимается концепция "нулевой" численности, то частота встречаемости m_i/M автоматически теряет всякий смысл (поскольку вид встречается ***всегда***, принимая иногда значения $N_{ik} \to 0$) и, после исключения этого соотношения, индекс доминирования D_i гармонично превращается в традиционную вероятность p встретить i-й вид. Вряд ли многие гидробиологи будут с этим согласны.

Тем не менее, нами был избран принцип "значащих нулей" и при логарифмировании шести одинаковых по размеру выборок численностей трофических групп мы использовали прием добавления единичной численности $\ln(N_{ik} + 1)$, поскольку логарифм нуля математически не определен. В этих условиях расчета связь численности подгрупп хирономид с категорией грунта проявляется достаточно отчетливо: влияние фактора во всех случаях достоверно, за исключением группы сестоно-детритофагов, для которых влияние грунта незначимо. Из диаграммы сдвига групповых средних на рис. 5.8 можно, например, видеть, что рост численности хирономид на грунтах категории 6 объясняется развитием одной группы – детритофитофагов. В то же время низкая численность организмов на чистых твердых грунтах категории 1 обусловлена, в основном, всеядными собирателями и детритофитофагами. Напомним также, что на численность сестонофагов + детритофагов фильтраторов тип грунта заметного влияния не оказывает.

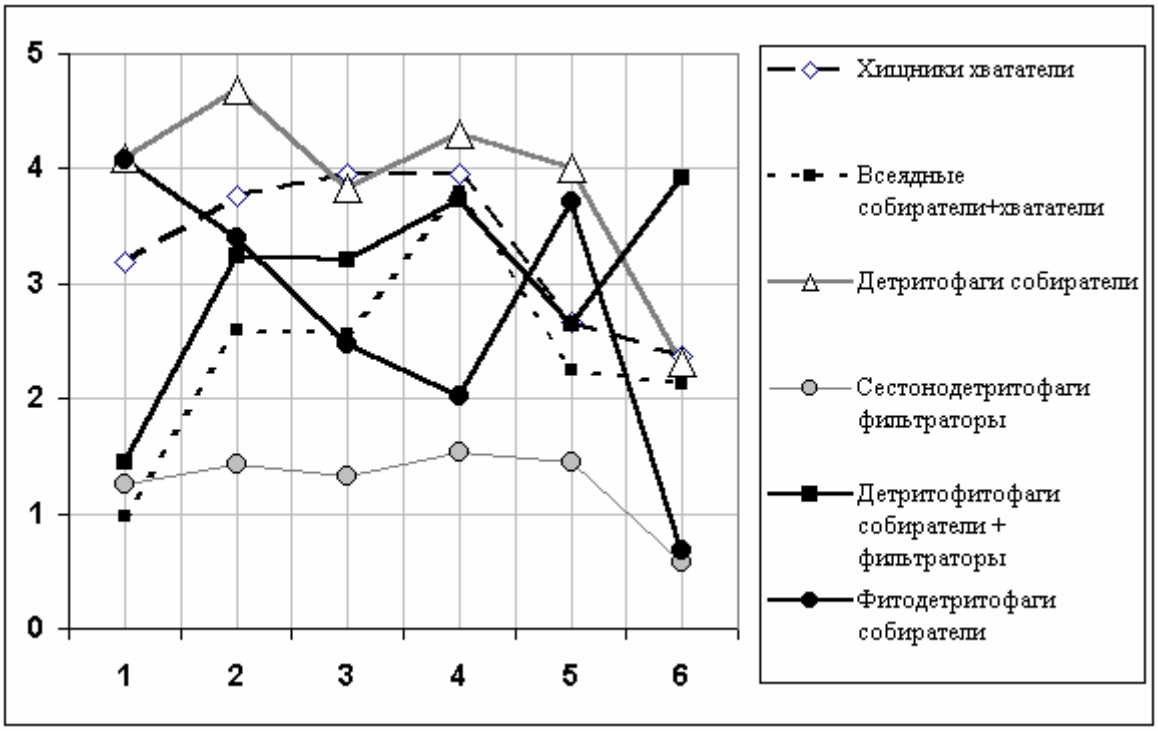


Рис 5.8. Зависимость средних значений логарифмов численности особей отдельных трофических групп (ось ординат) от категории грунтов

Поскольку результаты в табл. 5.6 свидетельствуют о высокой достоверности наличия факторного эффекта, то может быть выполнен анализ, по каким именно уровням фактора групповые средние значимо различаются. В табл. 5.7 приведена матрица парных сравнений Г. Шеффе, в которой для всех возможных пар категорий грунта {1}…{6} приведены вероятности предположения об отсутствии различий между средними значениями логарифма численности хищников-хватателей семейства Chironomidae. Для пар групп, выделенных жирным шрифтом, не отклоняется гипотеза об отсутствии различий.

Таблица 5.7

Вероятность отсутствия различий групповых средних логарифма численности хищников хватателей на разных категориях грунта, оцененная по методу Шеффе

Категории	{1}	{2}	{3}	{4}	{5}	{6}
Средние по группам	3.193	3.764	3.961	3.949	2.658	2.371
{1}	***	0.809	0.472	0.564	**0.977**	0.933
{2}	0.809	***	**0.997**	**0.998**	0.629	0.574
{3}	0.472	**0.997**	***	**1.000**	0.417	0.401
{4}	0.564	**0.998**	**1.000**	***	0.458	0.431
{5}	**0.977**	0.629	0.417	0.458	***	**0.999**
{6}	0.933	0.574	0.401	0.431	**0.999**	***

Как любую симметричную матрицу расстояний, таблицу парных сравнений удобно интерпретировать в виде графа пороговой связности. На рис. 5.9 представлен такой граф, узлами которого являются категории грунта, а все связи между узлами соответствуют принятым нулевым гипотезам об отсутствии значимых различий численности хирономид-хищников для различных категорий грунтов с использованием трех различных критериев: Шеффе, Дункана и Тьюки.

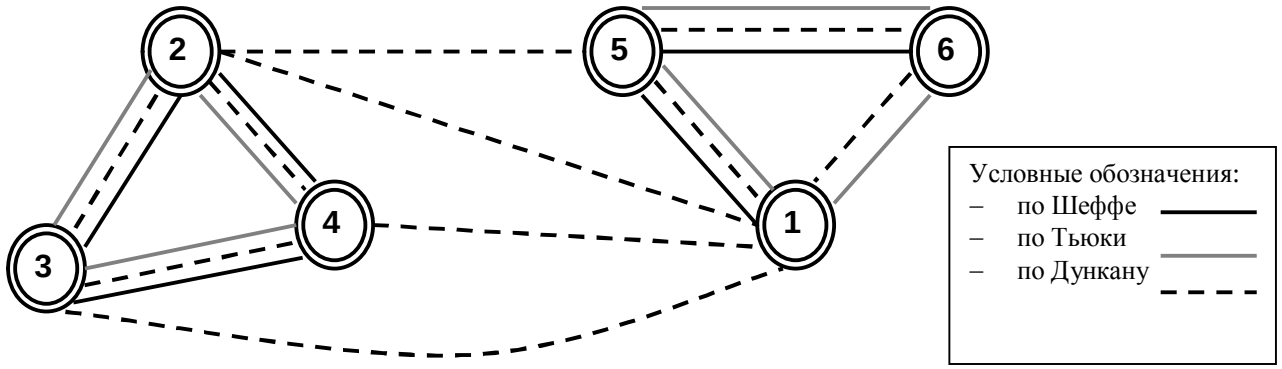


Рис 5.9. Граф связности категорий грунтов на основе равенства групповых средних численности хирономид-хищников (ребра графа соответствуют принятию нулевой гипотезы по критериям Шеффе, Тьюки и Дункана)

На графе отчетливо выделяются две характерные группы: грунты с большой численностью хищников (категории 2, 3, 4) и грунты, где эта численность относительно мала (категории 1, 5, 6). Можно также отметить, что выводы с использованием критериев Шеффе и Тьюки дали весьма сходные результаты, в то время как критерий Дункана на нашем исходном материале оказался склонен к сильной недооценке возможных различий между категорированными средними.

5.3. Теория и практика двухфакторного дисперсионного анализа

Формулировка задачи

Пусть имеется выборка из m значений $X_1, X_2, \ldots, X_m$ некоторого гидробиологического показателя, измеренного в количественной шкале. Изучается влияние, которое оказывают на отклик два качественных признака – фактор A, который имеет k уровней (градаций) $A_1, \ldots, A_k$, и фактор B, разбитый на n уровней $B_1, \ldots, B_n$

Необходимо:

- проверить гипотезу о принадлежности всех выборок одной генеральной совокупности, т.е. устанавливается, нельзя ли объяснить расхождение наблюдаемых значений для разных уровней факторов случайными обстоятельствами или влиянием неучтенных факторов;
- если нулевая гипотеза отвергается, то оценить степень влияния каждого фактора;
- выделить все пары групп $\{i - j\}$, $i = 1,2,\ldots,k \cdot n$, $j = 1,2,\ldots,k \cdot n$, $i \neq j$, которые имеют между собой статистически значимые различия.

В рамках этой задачи весьма типична ситуация, когда один из факторов (например, фактор B) является "мешающим": он включается в рассмотрение по той причине, что его влияние следует элиминировать, чтобы обнаружить и оценить индивидуальное влияние фактора A.

Общая методика анализа, как и в случае одного фактора, заключается в разложении общей вариации совокупности результатов наблюдения на частные вариации, обусловленные воздействием отдельных факторов и их комбинаций, и на остаточную вариацию, обусловленную случайными причинами. Оценка достоверности влияния факторов проводится по методу Фишера путем расчета отношения дисперсии, характеризующей статистическое колебание групповых средних по отдельным факторам, к дисперсии, характеризующей случайную вариацию.

Рекомендуемая литература та же, что и для раздела 5.2.

Математический лист

Предполагается, что измеряемая величина X есть результат действия факторов A и B и составляющей ε, которая является независимой, нормально $N(0,\sigma^2)$ распределенной случайной величиной: $X = f(A,B) + \varepsilon$.

Принимается аддитивная и независимая модель действия факторов:

$$f(A_j, B_i) = c + a_j + b_i \text{ , причем } \sum_{j=1}^{k} a_j = 0 \text{ , } \sum_{i=1}^{n} b_i = 0 \text{ .} \qquad (5.43)$$

Величины a_j и b_i называются вкладами факторов. Последние два условия всегда можно выполнить масштабированием величин a_j и b_i за счет изменения величины c.

Для каждого наблюдения из рассматриваемой совокупности справедливо уравнение:

$$x_{ij} = c + a_j + b_i + \varepsilon_{ij} \text{ , } \quad i = 1, \ldots, n; \quad j = 1, \ldots, k. \qquad (5.44)$$

Обычно наблюдения представляют структурной таблицей статистического комплекса. Приведем простейший двухфакторный комплекс, в которой каждому сочетанию (A_j, B_j) уровней (градаций) факторов, т.е. одной клетке таблицы, соответствует одно наблюдение (в таблице сочетание символов «( )^» обозначает статистическую оценку групповых средних):

Фактор B	Фактор A				Средние по строкам (оценки вкладов B)
	A_1	A_2	...	A_k	
B_1	x_{11}	x_{12}	...	x_{1k}	$x_{1\bullet}=(c+b_1)$^
B_2	x_{21}	x_{22}	...	x_{2k}	$x_{2\bullet}=(c+b_2)$^
...			...		...
B_n	x_{n1}	x_{n2}	...	x_{nk}	$x_{n\bullet}=(c+b_n)$^
Средние по столбцам (оценки вкладов A)	$x_{\bullet 1}=(c+a_1)$^	$x_{\bullet 2}=(c+a_2)$^		$x_{\bullet k}=(c+a_k)$^	$x_{\bullet\bullet}=c$^

Основное тождество дисперсионного анализа

Оценки c, b_i, a_j могут быть получены с помощью метода наименьших квадратов (МНК) минимизацией суммы

$$\sum_{i,j}(x_{ij}-c-b_i-a_j)^2 \Rightarrow \min \qquad (5.45)$$

Основываясь на МНК-оценках

$$\hat{c}=\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}x_{ij}\equiv x_{\bullet\bullet}\ ,\ \hat{b}_i=\frac{1}{k}\sum_{j=1}^{k}x_{ij}-\hat{c}\equiv x_{i\bullet}-x_{\bullet\bullet}\ ,\ \hat{a}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}-\hat{c}\equiv x_{\bullet j}-x_{\bullet\bullet}\ , \qquad (5.46)$$

введем следующие обозначения:

- для сумм квадратов отклонений под влиянием k уровней фактора A и n уровней фактора B:

$$Q_A=n\sum_{j=1}^{k}(x_{\bullet j}-x_{\bullet\bullet})^2=n\sum_{j=1}^{k}\hat{a}_j^2\ ,\qquad Q_B=k\sum_{i=1}^{n}(x_{i\bullet}-x_{\bullet\bullet})^2=k\sum_{i=1}^{n}\hat{b}_i^2\ ; \qquad (5.47)$$

- для остаточной суммы квадратов:

$$Q_e=\sum_i\sum_j(x_{ij}-\hat{c}-\hat{b}_i-\hat{a}_j)^2=\sum_i\sum_j(x_{ij}-x_{i\bullet}-x_{\bullet j}+x_{\bullet\bullet})^2=\sum_{ij}\hat{\varepsilon}_{ij}^2\ ; \qquad (5.48)$$

- для полной суммы квадратов наблюдений относительно общего среднего $\hat{c}=x_{\bullet\bullet}$:

$$Q=\sum_i\sum_j(x_{ij}-x_{\bullet\bullet})^2\ . \qquad (5.49)$$

Тогда справедливо следующее соотношение:

$$Q=Q_A+Q_B+Q_e=n\sum_{j=1}^{k}\hat{a}_j^2+k\sum_{i=1}^{n}\hat{b}_i^2+\sum_i\sum_j\hat{\varepsilon}_{ij}^2\ , \qquad (5.50)$$

т.е. полная сумма квадратов отклонений является суммой квадратов вкладов по факторам и квадратов случайных отклонений (или остатков $\hat{\varepsilon}_{ij}^2$). Другими словами, полное рассеяние есть сумма вариации под влиянием факторов и случайной составляющей.

Проверка нулевых гипотез

По имеющимся наблюдениям требуется проверить предположение об отсутствии влияния фактора A (или B) на результат измерения, т.е. проверить гипотезы

$$H_A:\ a_1=a_2=\ .\ .\ .=a_k=0$$

$$H_B:\ b_1=b_2=\ .\ .\ .=b_n=0\ .$$

Основой процедуры проверки гипотезы является сравнение двух статистически независимых оценок дисперсии σ^2.

Одна из них, σ^{2*} оценивает дисперсию вне зависимости от того, верна или нет гипотеза H_A (или H_B) и основана на сумме квадратов случайных отклонений:

$$\sigma^{2*}=\frac{Q_e}{(n-1)(n-k)}\ . \qquad (5.51)$$

Другая, σ^{2**} оценивает дисперсию, если H_A (или H_B) верна. Для гипотезы H_A эта дисперсия основана на сумме квадратов разностей “между столбцами”, т.е. по уровням фактора A:

$$\sigma_A^{2**}=\frac{Q_A}{k-1}\ . \qquad (5.52)$$

Если гипотеза H_A верна, то отношение

$$F_A=\frac{\sigma_A^{2**}}{\sigma^{2*}}=\frac{Q_A/(k-1)}{Q_e/r}=\frac{\sigma^2\chi_{k-1}^2/(k-1)}{\sigma^2\chi_r^2/r} \qquad (5.53)$$

имеет F-распределение Фишера с $(k-1)$ и r степенями свободы. Если

$$F_A\geq F_{1-\alpha}\ , \qquad (5.54)$$

где $F_{1-\alpha}$ – квантиль этого распределения порядка 1- α, α – выбранный уровень значимости, то гипотеза H_A отклоняется.

Вместо (5.54) можно использовать эквивалентную процедуру: гипотеза H_A отклоняется, если

$$P\{ F \geq F_A \} \leq \alpha, \qquad (5.55)$$

где $P\{ F \geq F_A \}$ – вероятность при справедливости H_A получить значение, большее, чем F_A; F – случайная величина, имеющая распределение Фишера.

Для проверки гипотезы H_B используют сумму квадратов разностей "между строками", то есть по уровням фактора B:

$$\sigma_B^{2**} = \frac{Q_B}{n-1} . \qquad (5.56)$$

Аналогичным образом, если отношение $F_B = \frac{\sigma_B^{2**}}{\sigma^{2*}}$ велико, то гипотеза H_B отклоняется.

Результаты расчетов

Формирование речных сообществ, их видовой состав и продуктивность гидробионтов находятся в постоянной динамике под воздействием большого количества экологических факторов. Будем оценивать временнỳю изменчивость сообществ, когда набор видов и количество особей претерпевают ежедневные, сезонные и многолетние колебания под воздействием температурного режима, мгновенной гидродинамики водотоков, сбросов токсикантов и т.д. Для прогнозирования динамики экологических систем методами статистического анализа существует обширный и специализированный математический аппарат, который остается за рамками настоящего изложения. Подробно теоретические и прикладные вопросы анализа хронологических трендов экосистем с использованием функциональных предикторов временных рядов рассматривались нами ранее [Розенберг с соавт., 1994]. Здесь мы ограничимся изложением частной методики дисперсионного анализа для оценки влияния многолетней и сезонной составляющих на общие показатели обилия зообентоса.

На основе данных из базы по малым рекам Самарской области сформируем выборку из суммарных значений численности N_s (экз./м^2) и биомассы B_s (г/м^2) зообентоса и индекса разнообразия Шеннона H для каждой пробы наблюдений. Данные по численности и биомассе предварительно прологарифмируем. Каждому значению варьируемой переменной поставим в соответствие три фактора: водоток, из которого взяты пробы, год исследования (с 1988-99 гг.) и порядковый номер месяца (с 5 по 9) отбора пробы.

Для оценки влияния регионального фактора выполним предварительно "разведывательный" однофакторный анализ по всем 33 градациям, соответствующим отдельным рекам (см. табл. 5.8).

Таблица 5.8

Результаты проверки гипотезы о влиянии места отбора проб на логарифм численности зообентоса по F-критерию

Источник вариации	Сумма квадратов	Степеней свободы	Средние квадраты	F-отношение	p-значение
Между группами	85.58	32	2.67	0.99	0.49 (фактор не значим)
Внутри групп	1357.12	500	2.71		
Итого	1442.71	532			

Оговоримся, что проблемы пространственной изменчивости, отражающей распределение показателей обилия в зависимости от географических координат поверхности, неизмеримо сложнее описываемого примера, поэтому в данном контексте речь идет не об оценке влияния места отбора пробы вообще, а о конкретном разбиении влияющего фактора на градации. Во-первых, проблематично само понятие «река», как средство обобщения гидробиологических данных, т.к. вариации данных между станциями одной реки, как правило, превышают межрегиональную вариацию (см. фиг. «а» рис. 5.10). Во-вторых, сама природа дисперсионного анализа предполагает тенденцию к недооценке влияния фактора при числе градаций больше 10. Однако выводы табл. 5.8 дают нам

формальные основания осуществить дисперсионный анализ в градациях остальных двух факторов: «год» – «месяц», основные результаты которого приведены в табл. 5.9.

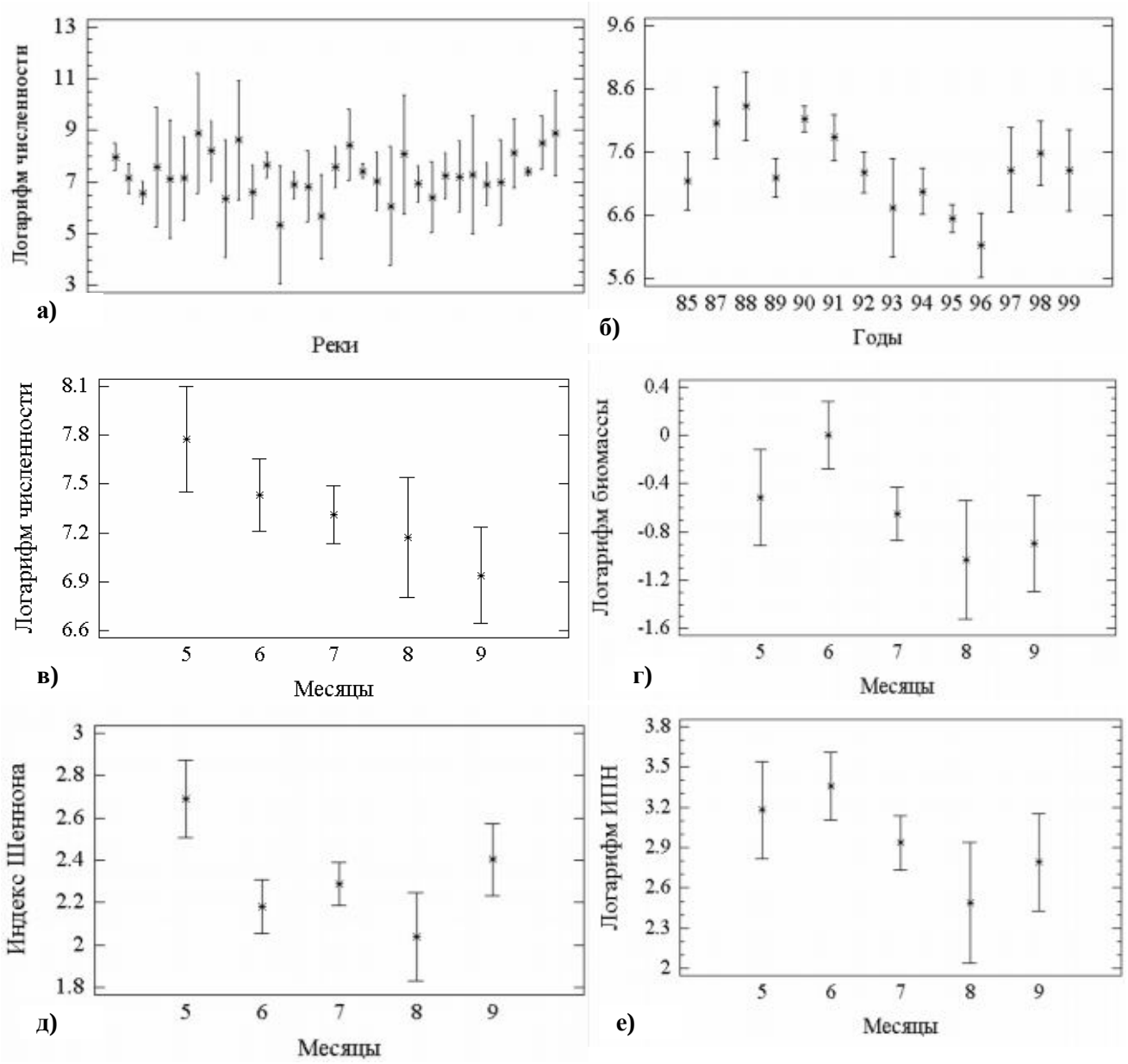


Рис 5.10 Диаграммы изменения групповых средних по результатам дисперсионного анализа показателей обилия и биоразнообразия зообентоса

По существу рассматриваемой проблемы выполненные расчеты влияния факторов позволяют сделать следующие выводы:

1. Во всех случаях (выборки 1-4, 6 в табл. 5.9) отчетливо проявляется влияние многолетнего фактора. Динамика изменения групповых средних по годам для численности зообентоса представлена на рис. 5.10, фиг. «б». Однако вряд ли следует понимать это влияние слишком буквально, как результат изменения гидрологических или радиационно-климатических условий. Можно, например, предложить следующую версию влияния многолетнего фактора. В 1987-91 гг. в коллектив исследователей пришла группа ученых, владеющих современными методами биологического контроля, что существенно расширило диапазон исследований и создало эффект роста численности определяемого материала. Начиная с 1992 г., наблюдалось резкое снижение уровня финансирования академических исследований, в результате чего сузилась широта географического охвата, и внимание исследователей сконцентрировалось на водоемах, под-

верженных экстремальным антропогенным воздействиям. Рост обилия гидробионтов с 1997 г. можно объяснить как расширением региона исследований на более "чистые" зоны, так и эффектом постепенного самоочищения водотоков после экономической стагнации 1991-97 гг. В любом случае, многолетний фактор является сложным конгломератом трудно идентифицируемых "человеческих", экономических и гидрометеорологических составляющих.

Таблица 5.9

Результаты двухфакторного дисперсионного анализа влияния временных факторов на показатели обилия зообентоса

Выборки для анализа	Источник вариации	Сумма квадратов	Степеней свободы	Средние квадраты	***F***-отношения	***p***-значения
1. Численность зообентоса в пробе (логарифм N_s)	Фактор «Месяц»	24.23	4	6.06	2.58	0.0365
	Фактор «Год»	208.12	13	16.01	6.82	0.0
	Остаточная	1208.84	515	2.35		
	Всего	1442.71	532			
2. Численность *Chironomidae* в пробе (логарифм N_s)	Фактор «Месяц»	25.65	4	6.41	2.23	0.0651
	Фактор «Год»	80.38	13	6.18	2.15	0.0108
	Остаточная	1341.48	446	2.87		
	Всего	1455.08	463			
3. Индекс Шеннона H (по численности зообентоса)	Фактор «Месяц»	11.93	4	2.98	3.95	0.0036
	Фактор «Год»	46.6	13	3.58	4.74	0.0
	Остаточная	388.6	514	0.756		
	Всего	442.16	531			
4. Биомасса зообентоса в пробе (логарифм B_s)	Фактор «Месяц»	40.08	4	10.02	3.09	0.0157
	Фактор «Год»	203.16	13	15.63	4.82	0.0
	Остаточная	1451.89	448	3.24		
	Всего	1694.25	465			
5. Индекс плотности населения $(N_s*B_s)^{1/2}$	Фактор «Месяц»	145189	4	36297	1.28	0.27
	Фактор «Год»	598507	13	46039	1.62	0.0762
	Остаточная	12727600	448	28409		
	Всего	13472500	465			
6. Логарифм индекса плотности населения $\ln((N_s*B_s)^{1/2})$	Фактор «Месяц»	22.49	4	5.62	2.09	0.081
	Фактор «Год»	119.1	13	9.16	3.40	0.0001
	Остаточная	1206.8	448	2.69		
	Всего	1356.6	465			

2. В очередной раз подтверждается необходимость обращать внимание на закон распределения зависимого показателя: влияние факторов на индекс плотности населения $(N_s*B_s)^{1/2}$ не было обнаружено, но проявилось после логарифмирования отклика.
3. С чисто формальных позиций, влияние сезонного фактора, выраженного календарным месяцем отбора пробы, является значимым для индекса Шеннона, численности и биомассы зообентоса (выборки 1, 3, 4), но незначимо для численности хирономид и индекса плотности населения (выборки 2, 6). Более детальные содержательные сведения можно получить из анализа сдвигов групповых средних, представленных на рис. 5.10. Месяц май характеризуют высокие значения численности N_s и биоразнообразия по Шеннону H, обусловленные развитием личинок амфибиотических насекомых весенней генерации. Начиная с июня, происходит постепенный захват экологических ниш видами-доминантами, в результате чего общая численность и, особенно резко, биоразнообразие H начинают падать. В то же время, для июня характерна максимальная продуктивность и энергообмен в трофических цепях, выраженные через биомассу. Результатом майско-июньской конкуренции между N_s и B_s явилось сглаживание контрастов индекса плотности населения (ИПН). Увеличение индекса Шеннона в сентябре обусловлено развитием осенних генераций амфибиотических насекомых в бентосе рек на фоне небольших показателей их обилия в конце лета.

Представляет интерес сравнить результаты двухфакторного и однофакторного анализа, например, для выборки 2 (численность хирономид). Если оценить по Фишеру локальное влияние сезонного фактора, то его вклад в общий статистический разброс численности следует трактовать как значимый (статистика Фишера $\boldsymbol{F}$ = 2.80, а соответствующее ему значение $\boldsymbol{p}$ = 0.0256). При переходе к двухфакторной схеме оказалось, что часть дисперсии, приписываемой в однофакторной схеме сезонному фактору, на самом деле объясняется многолетней составляющей (например, в разные годы экспедиции проводились во время разных месяцев).

Хотя у нас нет формальных оснований отвергнуть нулевую гипотезу о равенстве групповых дисперсий по тестам Кохрена и Бартлетта для данного примера, однако значения уровней значимости для обоих критериев находились в слишком опасной близости от порогового значения (α = 0.05):

- по тесту Кохрена: ***G***-критерий = 0.259, ***p***-значение = 0.07;
- по тесту Бартлетта: ***M***-критерий = 1.019, ***p***-значение = 0.06.

Поскольку в рамках схемы двухфакторного дисперсионного анализа оценка различий групповых средних представляет собой нетривиальную задачу, выполним на основе однофакторного анализа с использованием метода Шеффе выделение в течение сезона гомогенных групп значений фактора, т.е. комбинаций месяцев, между которыми нет значимых отличий в средних значениях (см. табл. 5.10). Например, для численности хирономид (выборка 2 табл. 5.9) можно составить три последовательности месяцев, отмеченных затененными клетками таблицы 5.10, внутри которых данные можно считать статистически однородными: а) с мая по июль, б) с июня по август, в) с июля по сентябрь. Таким образом, не отвергнув нулевую гипотезу об отсутствии влияния фактора сезонности для этого показателя, мы можем ставить вопрос об объединении в гомогенные группы некоторые подмножества уровней фактора.

Таблица 5.10

Результаты анализа множественных сравнений для среднемесячных значений показателей обилия зообентоса

Выборки для анализа	Месяцы					Пары месяцев с отличающимися средними
	5	6	7	8	9	
1. Численность зообентоса в пробе (ln $\boldsymbol{N_s}$)	7.77	7.43	7.31	7.17		{5-9}, {6-9}
		7.43	7.31	7.17	6.94	
2. Численность Chironomidae в пробе (ln $\boldsymbol{N_s}$)	6.90	6.68	6.39			{5-8}, {5-9}, {6-9}
		6.68	6.39	6.16		
			6.39	6.16	6.03	
3. Индекс Шеннона ***H***	2.68				2.40	{5-6}, {5-7}, {5-8}
		2.18	2.28	2.04	2.40	
4. Биомасса зообентоса в пробе (ln $\boldsymbol{B_s}$)	-0.518	-.0029				{6-7}, {6-8}, {6-9}
	-0.518		-0.65	-1.032	-0.895	
Число измерений в группах	58	135	213	50	76	

5.4. Непараметрические критерии для оценки однородности выборок

Формулировка задачи

В рамках постановки задачи однофакторного и двухфакторного дисперсионного анализа (см. разделы 5.2-5.3) выполнить проверку нулевой гипотезы об отсутствии различий между группами наблюдений в условиях неопределенности априорных предпосылок о нормальности распределений и однородности дисперсий.

Ранее было показано, что основные гидробиологические показатели, – численность и биомасса, – не удовлетворяют исходным предположениям о нормальности распределения и для их анализа не могут быть использованы параметрические методы и критерии. Выход был найден в функциональном преобразовании сравниваемых выборок; однако, подвергая деформированию

шкалу измерений, мы неизбежно должны оговариваться, что влияние, например, сезонного фактора сказывается в отношении логарифма численности, а не самой численности, т.к., по сути дела, формально это два разных показателя.

В условиях неопределенности исходных предпосылок дисперсионного анализа, проще опираться в своих выводах на отношениях "больше - меньше" между наблюдениями, т.к. эти отношения не зависят от формы распределения. Рассмотрим применение для проверки нулевой гипотезы семейства ранговых непараметрических критериев, опирающихся на ***рангах***, которые получают числа X_{ij} при упорядочении всей совокупности.

Рекомендуемая литература: [Большев, Смирнов, 1968; Гаек, Шидак, 1971; Гублер, Генкин, 1973; Кендалл, 1975; Рунион, 1982; Холлендер, Вулф, 1983].

Математический лист

Статистики Вилкоксона и Манна-Уитни для сравнения двух выборок

Пусть имеются две выборки $X_1, X_2, ..., X_m$ и $Y_1, Y_2, ..., Y_n$ из n и m числовых результатов наблюдений, функции распределения которых непрерывны и строго возрастают, причем эти случайные величины независимы в совокупности. Из непрерывности функций распределения следует, что с вероятностью 1 все $m + n$ результатов наблюдений различны. В реальных статистических данных иногда встречаются совпадения, и тогда теоретическая схема действует как приближенная, основанная на усреднении рангов, а надежность ее выводов снижается по мере увеличения числа совпадений. Рассмотрим критерии сравнения средних тенденций двух выборок.

Статистика W двух выборочного критерия, предложенного в 1945 г. Ф. Вилкоксоном (F.Wilcoxon), определяется следующим образом. Все элементы объединенной выборки $X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$ упорядочиваются в порядке возрастания. Последовательность рангов (т.е. порядковых номеров места, которое каждое значение занимает в общем упорядоченном ряду объема $m + n$) является некоторой перестановкой чисел 1, 2, …, $m + n$, а общее число возможных ранговых последовательностей равно $(m+n)!$ Пусть элементы первой выборки $X_1, X_2, ..., X_m$ занимают в общем вариационном ряду места с номерами $R_1, R_2, ..., R_m$, другими словами, имеют ранги $R_1, R_2, ..., R_m$. Тогда *критерий (статистика) Вилкоксона*

$$W = R_1 + R_2 + ... + R_m \tag{5.57}$$

будет уменьшаться по мере того, как наблюдения из первой группы будут оказываться чаще в левой части вариационого ряда чисел $1,2,\ldots,m+n+1$, т.е. наблюдения Y-ов получают тенденцию превосходить наблюдения X-ов. Вилкоксон показал, что при справедливости нулевой гипотезы, когда значения выборок хорошо перемешаны, статистику W можно считать приближенно нормально распределенной со средним $m\cdot(m+n+1)/2$ и дисперсией $m\cdot n\cdot(m+n+1)/12$. Это позволило разработать методики проверки справедливости нулевой гипотезы на заданном уровне значимости [Большев, Смирнов, 1983].

Статистика U Манна-Уитни (H.B. Mann, D.R. Whitney), предложенная в 1947 г., определяется как число пар (X_i, Y_j) таких, что $X_i < Y_j$, среди всех $m\cdot n$ пар, в которых первый элемент – из первой выборки, а второй – из второй. Иными словами, уславливаются, что $X_i < Y_j$ означает "успех", а всякое событие $X_i > Y_j$ – "неудачу". На всем множестве возможных парных сравнений $n\cdot m$ подсчитаем число "успехов" U и полученную случайную величину назовем статистикой Манна-Уитни. На "языке формул" имеем:

$$U = \sum_{\substack{i=1,2,...,m \\ j=1,2,...,n}} \varphi(x_i, y_j), \quad \text{где} \quad \varphi(x_i, y_j) = \begin{cases} 1, & \text{если } x_i \langle y_j \\ 1/2, & \text{если } x_i = y_j \\ 0, & \text{если } x_i \rangle y_j \end{cases} \tag{5.58}$$

Ясно, что U может принимать значения от 0 до $m\cdot n$.

Имеются доказательства [Оуэн, 1973], что статистики W и U эквивалентны в смысле связывающей их линейной функции

$$U = m\cdot n + m(m+1)/2 - W, \tag{5.59}$$

поэтому часто говорят об *общем критерии Вилкоксона–Манна–Уитни*.

Вернемся к задаче однофакторного дисперсионного анализа, оперирующей с таблицей $\mathbf{X} = \{x_{11}, x_{21}, \ldots, x_{n11}, x_{12}, \ldots, x_{nkk}\}$, где $j = 1,2,\ldots k$ – количество уровней группировки фактора, n_j – количество наблюдений на j-м уровне. Упорядочим величины x_{ij} (все равно как – от меньшего к большему, либо наоборот) и обозначим через r_{ij} ранг числа x_{ij} во всей совокупности. При выполнении гипотезы H_o любые возможные расположения рангов по местам дисперсионной таблицы равновероятны. Согласно стратегии анализа возникает вопрос: нельзя ли объяснить эмпирическое расположение рангов в таблице измерений чисто случайными причинами. Это соответствует проверке статистической гипотезы о том, что все k предъявленных выборок однородны, т.е. являются выборками из одного и того же закона распределения, для чего необходимо сконструировать ранговый статистический критерий, чувствительный к определенной совокупности альтернатив.

Критерий Краскела-Уоллиса: проверка однородности при одном факторе

Для каждого уровня фактора j рассчитывается средний ранг:

$$R_j = \frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij} \quad . \tag{5.60}$$

Если между столбцами нет систематических отличий, то средние групповые ранги R_j, j=1,2,…,k, не должны значительно отличаться от среднего ранга, рассчитанного по всей совокупности $\|r_{ij}\|$, который равен (N+1)/2 (N – общее количество измерений). Поэтому при справедливости H_o величины $\left(R_1 - \frac{N+1}{2}\right)^2, \ldots, \left(R_k - \frac{N+1}{2}\right)^2$ в совокупности не должны быть большими.

Исходя из этого утверждения, *статистика Краскела-Уоллеса* рассчитывается как

$$KU = \frac{12}{N(N+1)}\sum_{j=1}^{k} n_j\left(R_k - \frac{N+1}{2}\right)^2 \quad , \tag{5.61}$$

где множитель перед знаком суммы присутствует в качестве нормировочного для обеспечения асимптотической сходимости распределения KU к распределению χ^2 с числом степеней свободы (κ -1).

Критерий Джонкхиера для альтернатив с упорядочением

В рамках однофакторного дисперсионного анализа для каждой пары натуральных чисел u и v, соответствующих сравниваемым выборкам, где $k \geq v > u \geq 1$, составим *статистики Манна-Уитни*:

$$U_{u,v} = \sum_{\substack{i=1,2,\ldots m_u \\ j=1,2,\ldots n_v}} \varphi(x_{i_u}, y_{j_v}) \quad . \tag{5.62}$$

Определим *статистику Джонкхиера* (иное название – *критерий Джонкхиера-Терпстра*) как усложненный вариант критерия Манна-Уитни для случаев нескольких выборок:

$$I = \sum_{1 \leq u < v \leq k} U_{u,v} \quad . \tag{5.63}$$

Нетрудно заметить, что значение статистики зависит от порядка расположения градаций от 1 до k, поэтому I применяется в тех случаях, когда заранее предполагается, что имеющиеся группы результатов упорядочены по возрастанию влияния фактора. Например, первая группа наблюдений соответствует I классу качества воды в водоеме, 5 группа – V классу, а промежуточные номера соответствует последовательности классов качества. Свидетельством в пользу упорядоченности эффектов против гипотезы однородности служат большие значения статистики, полученные в эксперименте. Критерий Джонкхиера в этих условиях оказывается более чувствительным в оценке влияния фактора, чем критерий Краскела-Уоллиса.

При небольших объемах выборок распределение статистики I табулировано, а для больших выборок действует нормальная аппроксимация.

Критерий Фридмана : проверка однородности при двух факторах

Будем основываться на структурной таблице двухфакторного статистического комплекса (см. раздел 5.3). Обратим внимание на то, что критерий Фридмана можно применять только к данным, состоящим из равного числа наблюдений для каждой клетки из $k \cdot n$. Подобные планы эксперимента называются *сбалансированными*. Критерий основывается на тех же концепциях, что и статистика Краскела-Уоллеса, однако, в отличие от однофакторного анализа, ранжирование происходит не по всей совокупности величин x_{ij}, а по блочно, т.е. рассматривается каждая отдельная i-я строка и пересчет в ранги происходит для $j = 1,2,\ldots,k$. Тем самым устраняется влияние "мешающего" фактора B, значение которого для всех строк постоянно.

При гипотезе H_o: $a_1 = a_2 = \ldots = a_k = 0$ об отсутствии влияния фактора A каждая строка рангов будет представлять случайную перестановку чисел от 1 до k, причем все $k!$ перестановок равновероятны. Введем величину $R_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$, являющуюся средним значением рангов по столбцу j, и будем сравнивать ее с общим средним рангом всех элементов таблицы, равном $R_{cp} = (k + 1)/2$.

Статистика Фридмана для проверки гипотезы H_o имеет следующий вид:

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{k+1}{2} \right)^2 . \qquad (5.64)$$

При справедливости нулевой гипотезы (в силу равновероятности построчных перестановок рангов) значения $(R_j - R_{cp})^2$ достаточно малы для всех j и, следовательно, значение S сравнительно невелико. При нарушении H_o баланс суммы рангов по столбцам нарушается и статистика Фридмана возрастает. Для небольших величин n и k критические значения $S(\alpha,k,n)$ могут быть найдены по таблицам. При $n \to \infty$ статистика Фридмана асимптотически распределена как χ^2 с числом степеней свободы $(\kappa - 1)$.

Критерий Пейджа для альтернатив с упорядочением

Сформируем таблицу рангов так, как это было сделано при обосновании статистики Фридмана. Критерий Пейджа предназначен для проверки гипотезы H_o об отсутствии эффекта влияния фактора A (H_o: $a_1 = a_2 = \ldots = a_k = 0$) против альтернативы "влияние фактора A увеличивается (вариант: уменьшается) в определенном направлении изменения градаций" , т.е. гипотезы об упорядочении H_1: $a_1 \le a_2 \le \ldots \le a_k$, где хотя бы одно из неравенств строгое.

Введем величину $Rs_j = \sum_{i=1}^{n} r_{ij}$, являющейся суммой рангов по столбцу j. *Статистика Пейджа* по определению есть:

$$L = \sum_{j=1}^{k} jRs_j = Rs_1 + 2Rs_2 + \ldots + kRs_k . \qquad (5.65)$$

Методы проверки нулевой гипотезы с использованием L-критерия Пейджа изложены в литературе [Холлендер, Вульф, 1983; Ликеш, Ляга, 1985] и иллюстрируются ниже на примерах.

Результаты расчетов

В предыдущем разделе был проведен анализ влияния сезонного фактора (т.е. номера месяца, в котором была отобрана проба) на общую численность зообентоса и личинок Chironomidae, в частности. Однозначных выводов, в конечном итоге, сделано не было, в первую очередь потому, что сезонная динамика различных видов водных организмов далеко не однозначна. Рассмотрим результаты проверки нулевой гипотезы об однородности выборок численности отдельных видов хирономид, взятых в различные месяцы отбора проб, с использованием методов непараметрической статистики. В таблице 5.11, кроме значений критерия Краскела-Уоллиса, приведены параллельно величины параметрических F-критериев Фишера, рассчитанных по натуральным и предварительно прологарифмированным выборкам.

Таблица 5.11

Результаты параметрического и непараметрического однофакторного дисперсионного анализа влияния фактора сезонности на численность отдельных видов зообентоса

Наименование видов	Объем выборки	По численности		По логарифму численности		По методу Краскела-Уоллиса	
		F-критерий	***p***-значение	***F***-критерий	***p***-значение	***KU***-критерий	***p***-значение
Chironomus plumosus	190	2.62	0.0367	0.39	0.815	2.064	0.723
Cladotanytarsus mancus	138	1.93	0.109	0.85	0.493	3.385	0.495
P.nubeculosum	183	0.90	0.463	0.73	0.574	3.279	0.512
Dicrotendipes nervosus	75	0.06	0.99	3.43	0.0127	13.16	0.010
Cryptochironomus gr. *defectus*	139	4.65	0.0015	2.07	0.088	10.718	0.029
Procladius ferrugineus	177	0.22	0.926	0.62	0.651	3.302	0.508
Prodiamesa olivacea	59	0.30	0.828	0.35	0.788	1.229	0.746

По результатам расчетов можно сделать следующие выводы:

- обращает на себя внимание факт близости оценок уровней значимости ***P***, рассчитанных по ***F***-критерию для прологарифмированных выборок и по статистике Краскела-Уоллиса, где несовпадение, имеющее принципиальное значение, зафиксировано только для *Cryptochironomus* gr. *defectus*;
- в очередной раз приходится констатировать, что использование дисперсионного анализа для выборок с сильной асимметрией распределения, какими являются ряды натуральных значений численности зообентоса, приводит к совершенно непредсказуемым результатам;
- из семи проанализированных видов зообентоса непротиворечивые выводы о статистически существенном влиянии сезонного фактора на численность зафиксированы только для *Dicrotendipes nervosus* (в отношении *Cryptochironomus* gr. *defectus* параметрические и непараметрические оценки разошлись).

Сезонная динамика значений среднего ранга численности некоторых видов представлена на рис. 5.11. Можно отметить, например, не обычные для жизненных циклов *P. nubeculosum* и *Cladotanytarsus mancus* сглаженные подъёмы численности их личинок.

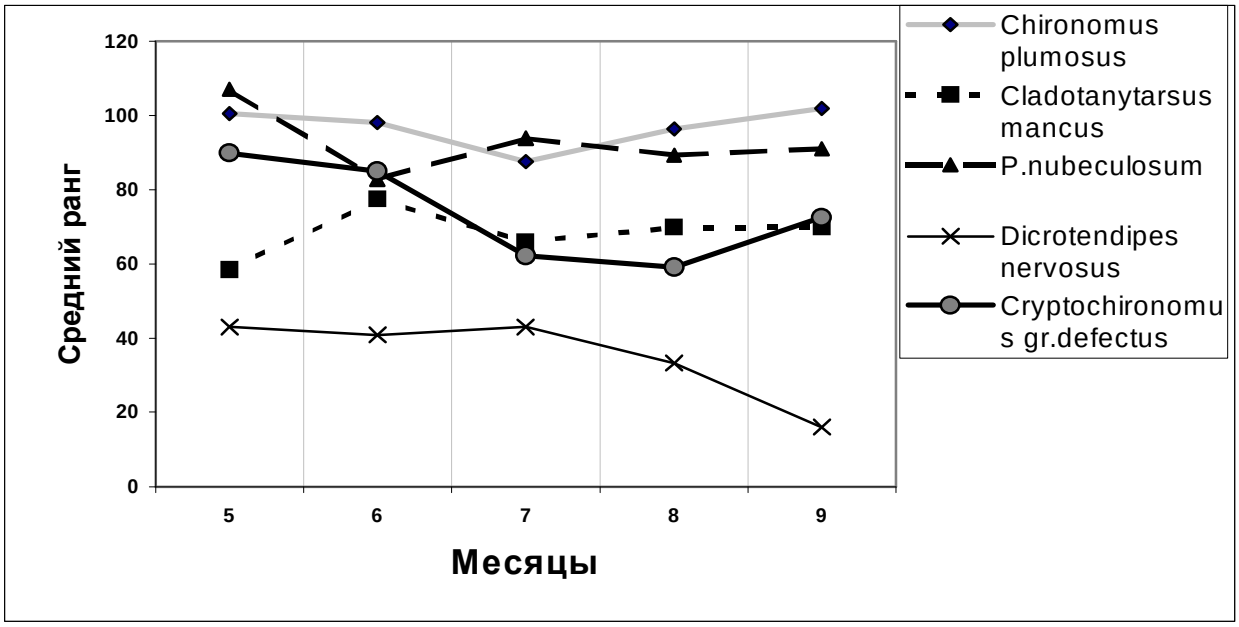


Рис. 5.11. Изменение среднемесячного ранга численности отдельных видов хирономид

Для окончательного уточнения сезонных изменений обилия *Cryptochironomus* gr. *defectus* проверим в качестве альтернативы к нулевой гипотезе предположение о монотонном снижении численности вида с мая по сентябрь с использованием статистики Джонкхиера. Для ее вычисления найдем значения критерия Манна-Уитни U для всех комбинаций индексов u и v, которые меняются от 5 до 9, причем $u < v$:

$U_{56} = 149$; $U_{67} = 621$; $U_{78} = 610.5$; $U_{89} = 145$
$U_{57} = 171$; $U_{68} = 198.5$; $U_{79} = 443$;
$U_{58} = 55.5$; $U_{69} = 228$;
$U_{59} = 71$

Откуда $I = \sum_{1 \le u < v \le k} U_{u,v} = 2692.5$.

Для нахождения минимального уровня значимости критерия Джонкхиера воспользуемся нормальной аппроксимацией: $I \sim N(M_I, D_I)$, где

$$M_I = \left(N^2 - \sum_{j=1}^{k} n_j^2 \right) / 4 = 3460.5; \qquad D_I = \left(N^2 (2N+3) - \sum_{j=1}^{k} n_j^2 (2n_j + 3) \right) / 72 = 67645 .$$

Величина $I^* = (I - M_I) / \sqrt{D_I}$ асимптотически имеет стандартное нормальное распределение и по таблицам для $I^* = -2.95$ имеем p = 0.00051, т.е. принимается гипотеза о монотонном снижении численности *Cryptochironomus* gr. *defectus* в течении сезона. Заметим, что критерий Джонкхиера оказался более чувствительным к выявлению тенденции, чем статистика Краскела-Уоллиса.

Рассмотрим теперь, как влияют на разнообразие видов зообентоса, оцененное по индексу Шеннона H, два фактора – место расположения точки отбора пробы и месяц, во время которого произведено наблюдение. Поскольку метод Фридмана использует сбалансированные планы, сформируем таблицу средних значений индексов Шеннона для р. Чапаевка, разбив всё течение реки от истоков до устья на 6 участков (см. табл. 5.12). Поскольку на участках 1-4 в августе измерения не проводились, в соответствующие пропущенные клетки таблицы были подставлены средние за август значения индекса на устьевых участках ($x_{ij} = 1.605$).

Таблица 5.12

Средние значения индекса Шеннона по численности зообентоса для различных участков р. Чапаевка и месяцев отбора проб наблюдений

Участки / месяцы	Май	Июнь	Июль	Август	Сентябрь	В среднем
1. Исток (ст. 1-2)	2.637	2.595	2.007	1.605	1.751	2.247
2. Верховья (ст. 3-4)	2.798	1.807	2.683	1.605	2.370	2.414
3. Средний участок (ст. 5-9)	2.756	2.589	2.335	1.605	1.889	2.392
4. Район Колывани (ст. 10-12)	2.692	2.484	2.413	1.605	2.346	2.484
5. Низовья (ст. 14-22)	0.501	1.219	1.414	0.875	1.373	1.077
6. Устье (ст. 23)	1.827	1.436	2.292	2.336	1.220	1.822
В среднем	2.202	2.022	2.191	1.605	1.825	2.073

Выполним предварительно стандартный двухфакторный дисперсионный анализ (см. таб. 5.13). Очевидно значимое влияние фактора **Ф1** расстояния от устья и отсутствие влияние фактора сезонности **Ф2**.

Сформируем в соответствии с методом Фридмана две таблицы рангов и рассчитаем для каждого столбца и каждой строки таблицы 5.11 суммы рангов и средние суммы рангов, поместив результаты в табл. 5.14.

Таблица 5.13

Дисперсионный анализ значений индекса Шеннона в зависимости от участков р. Чапаевка и месяцев отбора проб наблюдений

Источник вариации	Сумма квадратов	Степеней свободы	Средние квадраты	***F***-отношение	***p***-значение
Ф1 – участок	5.73	5	1.146	6.04	0.0017
Ф2 – месяц	1.69	4	0.4224	2.226	0.102
Остаточная	3.795	20	0.1897		
Всего	11.21	29	0.3867		

Таблица 5.14

Ранжирование значений индекса Шеннона по участкам р. Чапаевка и месяцам взятия проб

По строкам таблицы 5.11			По столбцам таблицы 5.11		
Участки	Средний ранг	Сумма рангов	Месяцы	Средний ранг	Сумма рангов
1. Исток (ст. 1-2)	3.5	17.5	Май	4.00	24
2. Верховья (ст.3-4)	4.9	24.5	Июнь	3.17	19
3. Средний участок (ст. 5-9)	4.3	21.5	Июль	3.67	22
4. Район Колывани (ст. 10-12)	4.3	21.5	Август	1.83	11
5. Низовья (ст.14-22)	1.2	6	Сентябрь	2.33	14
6. Устье (ст. 23)	2.8	14			

Рассчитав значение критерия Фридмана для левой части таблицы 5.14, мы получим $S = 13.67$. Этому значению соответствует статистика χ^2 с 5 степенями свободы и вероятностью $p = 0.0178$, т.е. с таким уровнем значимости нулевая гипотеза отвергается в пользу альтернативы и влияние места отбора пробы на индекс Шеннона можно считать установленным.

Аналогично для правой части таблицы имеем $S = 7.86$, $p = 0.096$, т.е. в представленных данных фактор сезонности не оказывает влияния на биоразнообразие. Нельзя не отметить практическое совпадение уровней значимости гипотез, оцененных по критериям Фишера и Фридмана.

Сделаем теперь предположение о том, что биоразнообразие зообентоса монотонно убывает от истоков к устью. Проверим эту гипотезу по критерию Пейджа:

$$L = \sum_{j=1}^{k} jRs_j = 14 + 2\cdot 6 + 3\cdot 21.5 + 4\cdot 21.5 + 5\cdot 24.5 + 6\cdot 17.5 = 404 .$$

Для нахождения приближенного значения минимального уровня значимости критерия Пейджа воспользуемся (аналогично расчетам уровня значимости для критерия Джонкхиера) нормальной аппроксимацией распределения следующей статистики:

$$L^* = \frac{L - nk(k+1)^2/4}{\left[n(k^3 - k^2)/144(k-1)\right]^{0.5}} = 14.4,$$

что очень близко к нулю по таблицам стандартного нормального распределения и нулевая гипотеза уверенно отвергается, т.е. наше предположение о монотонном снижении индекса Шеннона не противоречит данным.

Впрочем, для незначимого фактора **Ф2** сезонной изменчивости критерий Пейджа также достаточно велик и равен $L = 242$, чему соответствует $p = 0.011$. Это иллюстрирует тезис о том, что в случае упорядоченных альтернатив для критерия Пейджа характерна не всегда обоснованная гипердиагностика в пользу оценки влияния фактора.

5.5. Задача о законе влияния фактора: линейный регрессионный анализ

Мем № 29: *«Теоретическая кривая никогда не проходит в точности через все точки, полученные в результате измерений…»* В. Фукс [1975].

Формулировка задачи

Пусть имеется две группы числовых переменных $\boldsymbol{X} \equiv (x_1, ..., x_p)$ и $\boldsymbol{Y} \equiv (y_1, ..., y_m)$, причем предполагается, что $\boldsymbol{X}$ – независимая переменная (фактор) влияет на значения $\boldsymbol{Y}$ – зависимой переменной (отклик). В общем случае предполагается, что обе переменных измерены в количественных шкалах (интервальной, абсолютной или шкале отношений). Также постулируется независимость самих измерений, т.е. одни наблюдения не оказывают систематического влияния на другие.

Предположим, что из совокупности наблюдений $\boldsymbol{X}$-$\boldsymbol{Y}$ может быть укомплектована репрезентативная обучающая выборка $(\boldsymbol{X_i}, \boldsymbol{Y_i})$, $\boldsymbol{i} = 1, ..., \boldsymbol{n}$ сопряженных измерений, выполненных в идентичных пространственно-временных условиях. Необходимо по имеющейся обучающей выборке построить функцию $\boldsymbol{f}(\boldsymbol{X})$, которая приближенно описывала бы изменение $\boldsymbol{Y}$ при изменении $\boldsymbol{X}$: $\boldsymbol{Y} \approx \boldsymbol{f(X)}$.

Этой постановке соответствует широкий круг задач, связанных с построением (восстановлением) всевозможных зависимостей по имеющимся эмпирическим данным, т.е. открытие больших и малых законов, начиная от закона всемирного тяготения И. Ньютона и кончая законом Б.Л. Гутельмахера о линейной зависимости логарифма скорости экскреции биогенных элементов от массы тела зоопланктеров.

Статистика как метод исследования в принципе не интересуется причинно-следственной связью явлений. Однако рассмотренные в настоящем разделе методы аппроксимации близки идеям математического моделирования в смысле поиска наилучшей функциональной зависимости. Более того, уравнение кривой, полученной в результате обработки данных наблюдений, может подтолкнуть исследователя к пониманию внутренних взаимосвязей изучаемого явления.

Рекомендуемая литература: [Хальд, 1956; Андерсен, 1963; Себер, 1980; Дрейпер, Смит, 1986; Дюк, 1997; С.А. Прохоров, 2002; Прохоров с соавт. 2003].

Математический лист

Предполагается, что множество допустимых функций, из которого подбирается $\boldsymbol{f}(\boldsymbol{X})$, является параметрическим: $\boldsymbol{Y} \approx \boldsymbol{f}(\boldsymbol{X}, \theta)$, где θ – неизвестный параметр (вообще говоря, многомерный). Если имеет место равенство $\boldsymbol{f}(\boldsymbol{X}, \theta) = \boldsymbol{A}(\theta)\boldsymbol{X}$, где $\mathbf{A}(\theta)$ – некоторая матрица коэффициентов, то функция $\boldsymbol{f}(\boldsymbol{X}, \theta)$ линейно зависит от параметров θ и мы имеем дело с задачей *линейного регрессионного анализа*.

При построении аппроксимирующей функции будем считать, что

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}, \theta) + \varepsilon, \tag{5.66}$$

где первое слагаемое – закономерное изменение $\boldsymbol{Y}$ от $\boldsymbol{X}$, а второе ε – случайная составляющая с нулевым средним. Функция $\boldsymbol{f}(\boldsymbol{X}, \theta)$ является условным математическим ожиданием $\boldsymbol{Y}$ при условии известного $\boldsymbol{X}$ и называется ***регрессией Y по X.*** Слагаемое ε отражает как внутреннюю присущую отклику стохастическую изменчивость (ошибку измерения), так и влияние факторов, не учтенных в $\boldsymbol{f}(\boldsymbol{X}, \theta)$.

Классический регрессионный анализ предполагает, что статистическая природа случайной составляющей является неизменной для всех наблюдений, т.е. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ статистически независимы в опытах и одинаково распределены. Неоднородность дисперсий ошибок (*гетероскедастичность*, heteroscedasticity) в целом характерна для гидробиологических данных, поскольку обработке подвергаются показатели, полученные в разные моменты времени, по различным регионам, в различных экспедиционных условиях. Последствия гетероскедастичности могут привести к тому, что вычисленные значения $\boldsymbol{t}$- и $\boldsymbol{F}$-отношений уже нельзя рассматривать как наблюдаемые значения случайных величин, имеющих $\boldsymbol{t}$- и $\boldsymbol{F}$-распределения, что может приводить к ошибочным статистическим выводам в отношении гипотез о значениях коэффициентов линейной модели.

Простая линейная регрессия

Пусть X и Y одномерные величины, обозначенные как x и y, а функция $f(x, \theta)$ имеет вид $f(x, \theta) = A + bx$, где $\theta = (A, b)$. Относительно имеющихся наблюдений (x_i, y_i), $i = 1, ..., n$, полагаем, что

$$y_i = A + bx_i + \varepsilon_i \ , \tag{5.67}$$

где $\varepsilon_1, ..., \varepsilon_n$ – независимые (ненаблюдаемые) одинаково распределенные случайные величины.

Существует ряд методов подбирать "лучшую" прямую линию, из которых наиболее широко используется *метод наименьших квадратов* (МНК), который заключается в следующем. Построим оценку параметра $\theta = (A, b)$ так, чтобы величины

$$e_i = y_i - f(x_i, \theta) = y_i - A - bx_i \ , \tag{5.68}$$

называемые *остатками*, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - A - bx_i)^2 \ \rightarrow \ min \ \text{ по } (A, b) \ . \tag{5.69}$$

Решение $(\hat{a}, \hat{b})$, для которого сумма квадратов невязок оказывается наименьшей, может быть найдено с использованием выборочных средних $\bar{x}$ и $\bar{y}$ по следующим формулам:

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{n \sum_{i=1}^{n} y_i x_i - n \cdot \bar{y} \cdot \bar{x}}{\sum_{i=1}^{n} x_i^{\,2} - n \cdot \bar{x}^2} \ , \text{ где } \ \bar{y} = \sum_{i=1}^{n} y_i / n \ , \quad \bar{x} = \sum_{i=1}^{n} x_i / n \ . \tag{5.70}$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

Рассчитать уравнение регрессии не представляет никакой сложности, особенно при наличии компьютера с установленным табличным процессором Microsoft Excel, или любого другого пакета статистических программ. Основные методологические трудности возникают в ходе проверки выполнения априорных предпосылок регрессионного анализа и последующей оценки адекватности полученного уравнения.

Как и в случае дисперсионного анализа, оценка уравнений регрессии предполагает два основных подхода: непараметрический и гауссовый, которые различаются предположениями о характере распределения остатков $\varepsilon_1, \ldots, \varepsilon_n$. Гауссова модель простой линейной регрессии предполагает, что случайная величина ε распределена по нормальному закону $N(0, \sigma^2)$ с некоторой неизвестной дисперсией σ^2. Сам по себе метод наименьших квадратов в этих предположениях не нуждается, однако в гауссовой модели, во-первых, МНК обладает дополнительными свойствами оптимальности и, во-вторых, оценки неизвестных параметров приобретают ясные статистические свойства.

Разложение полной суммы квадратов и коэффициент детерминации

Нетрудно, как и для однофакторного дисперсионного анализа, рассчитать полную сумму квадратов вариации отклика $Q = \sum_{i=1}^{n} (y_i - \bar{y})^2$, и две ее составляющие: сумму квадратов объясненную моделью $Q_X = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, и остаточную сумму квадратов $Q_e = \sum_{i=1}^{n} (y_i - \hat{y})^2$. Все эти три суммы квадратов связаны соотношением

$$Q = Q_x + Q_e \ , \tag{5.71}$$

которое представляет собой разложение полной суммы квадратов.

Если вариация отклика, объясненная влиянием переменной X (точнее, уравнением регрессии) $Q_x \to 0$, то полная сумма квадратов соответствует значению Q_e. В этой ситуации привлечение информации о значениях переменной X не дает ничего нового для объяснения закономерности поведения Y в наблюдениях, а значит $\hat{b} = 0$ и "наилучшая" прямая имеет вид $y = \bar{y}$, отрицающий наличие линейной зависимости Y от X.

Аналогично, тенденция линейной связи между X и Y выражена в максимальной степени, если $Q_e = 0$: при этом, все точки (x_i, y_i), $i = 1, 2,..., n$, располагаются на одной прямой.

Рассмотрим традиционную для дисперсионного анализа F-статистику

$$F = \frac{Q_X/(k-1)}{Q_e/(n-k)} = \frac{(Q - Q_e)/(k-1)}{Q_e/(n-k)}, \tag{5.72}$$

где $k = 2$ – число степеней свободы простой линейной регрессии. Если остатки $\varepsilon_1,\ldots,\varepsilon_n$ распределены нормально $N(0,\sigma^2)$, то F-статистика, рассматриваемая как случайная величина, при справедливости гипотезы H_0 (т. е. когда действительно $\hat{b} = 0$) имеет стандартное распределение Фишера с $(k-1)$ и $(n-k)$ степенями свободы. В соответствии с этим, нулевая гипотеза отвергается при "слишком больших" значениях F, превышающих пороговое значение $F_{1-\alpha}(k-1, n-k)$. При этом, вероятность ошибочного отвержения гипотезы H_0 равна α. Большинство пакетов прикладных программ при анализе статистической значимости уравнения регрессии, наряду со значением F-статистики, приводят соответствующее ему P-значение (P-value), т. е. вероятность $P\{F(p-1, n-p) > F\}$.

Для оценки тесноты ("меры выраженности") линейной связи принято использовать долю объясненной суммы квадратов, не зависящую от шкал, в которых измерены значения переменных x и y и называемую часто *коэффициентом детерминации*:

$$R^2 = \frac{Q_x}{Q} = 1 - \frac{Q_e}{Q}. \tag{5.73}$$

Этот коэффициент изменяется в пределах от 0 (при $\hat{\beta} = 0$, т. е. $Q = Q_e$) до 1 (при $Q_e = 0$). Чем ближе к 1 значение R^2, тем лучше качество подгонки. При $R^2 \to 0$ можно сделать два вывода: либо фактор не оказывает никакого влияния на отклик, либо функция регрессии имеет существенно нелинейный характер.

Коэффициент детерминации тесно связан с другой мерой связи между переменными – выборочным *коэффициентом линейной корреляции* Пирсона. Пусть определены выборочные дисперсии, характеризующие степень разброса значений X и Y:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{5.74}$$

а также *выборочная ковариация*, характеризующая совместное распределение этих двух выборок в N-мерном евклидовом пространстве:

$$C(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{5.75}$$

Коэффициент линейной корреляции Пирсона между выборками Y и X определяется как

$$r_{xy} = \frac{C(x,y)}{\sqrt{S_x^2}\sqrt{S_y^2}}. \tag{5.76}$$

Нетрудно заметить его симметричность, т.е. $r_{xy} = r_{yx}$. Констатируем также без доказательств два известных факта:

- коэффициент детерминации и коэффициент корреляции между переменными x и $\hat{y}$ при гауссовских предположениях связаны простой формулой $r_{y\hat{y}}^2 = r_{xy}^2 = R^2$;
- коэффициент корреляции r_{xy} изменяется в диапазоне от -1 до 1, причем при $r_{xy} = 0$ линейная связь отсутствует, а при $r_{xy} < 0$ мы имеем дело с обратной линейной связью ($b < 0$).

Следует отметить, что традиционные мнения о возможности оценить тесноту линейной связи только по коэффициенту корреляции Пирсона r (либо его квадратам R^2 и η^2) не являются статистически корректными. Например, Я.Я. Вайну [1977] приводит следующую классификацию линейности по r: менее 0.2 – слабая, от 0.2 до 0.4 – слабее средней тесноты, от 0.4 до 0.6 – средняя и т.д. Нетрудно заметить, что коэффициенты корреляции, основанные на сумме квадратов, не учитывают числа степеней свободы, поэтому их сравнение правомочно только для выборок примерно одинаковой размерности. При увеличении числа обработанных измерений статистическая надежность уравнений регрессии по критерию Фишера начинает возрастать, когда как величина коэффициента корреляции чаще всего монотонно снижается.

Для удобства интерпретации параметра угла наклона b иногда пользуются *коэффициентом эластичности*

$$Э = \hat{b}\frac{\bar{x}}{\bar{y}} = \frac{\Delta y}{\Delta x}\cdot\frac{\bar{x}}{\bar{y}} = \frac{\Delta y}{\bar{y}}\Big/\frac{\Delta x}{\bar{x}}\ , \tag{5.77}$$

который показывает среднее изменение (в долях или %) зависимой переменной y при изменении фактора x: $\dfrac{\Delta y}{\bar{y}} = Э\cdot\dfrac{\Delta x}{x}$.

Доверительные интервалы для коэффициентов и проверка гипотез

Если принять дополнительные предположения о гауссовом характере распределения остатков $\varepsilon_1, ..., \varepsilon_n$, то:

- оценки $\hat{a}$ и $\hat{b}$ нормально распределены и независимы от дисперсии остатков σ^2;
- несмещенной оценкой дисперсии остатков является величина $s^2 = Q_e / (n - k)$, а среднеквадратичное отклонение остаточной ошибки модели регрессии может быть вычислено по формуле

$$s = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - A - bx_i\right)^2}\ , \tag{5.78}$$

  где $k = 2$ – число степеней свободы регрессии;
- отношения $\sqrt{n}\,\dfrac{\hat{a}-a}{s}$ и $\dfrac{\hat{b}-b}{s_b}$, где $s_b = s\Big/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, имеют распределение Стьюдента с $(n-2)$ степенями свободы;
- доверительные интервалы для a и b определяются как

$$\left|\hat{a}-a\right| \le t_p\frac{s}{\sqrt{n}} \quad \text{и} \quad \left|\hat{b}-b\right| \le t_p s_b\ , \tag{5.79}$$

  где t_p – квантиль уровня $(1 + P) / 2$ распределения Стьюдента с $n - 2$ степенями свободы, P – доверительная вероятность.

Статистическая значимость уравнения линейной регрессии проверяется отклонением нулевой гипотезы о равенстве нулю коэффициента b, отражающего угол наклона линии тренда. Гипотеза H_o: $b = 0$ соответствует предположению, что изменчивость y обусловлена только случайными составляющими ε_i. Существует два эквивалентных способа проверки нулевой гипотезы:

- если 0 не входит в доверительный интервал для b, т.е. $|b| / s_b > t_p$, то гипотезу H следует отклонить; уровень значимости при этом $\alpha = 1 - P$;
- вычисляется статистика, которая в случае, если H_o верна, распределена по закону Фишера F с числом степеней свободы 1 и $n - 2$:

$$F = \frac{\hat{b}^2 / D\hat{b}}{Q_e{}^2/(\sigma^2(n-2))} = \frac{\hat{b}^2}{s_b^2}\ ; \tag{5.80}$$

  если $F > F_{1-\alpha}$, где $F_{1-\alpha}$ – квантиль уровня $(1 - \alpha)$ распределения $F(1, n-2)$, то гипотеза H отклоняется с уровнем значимости α.

Тестирование правильности спецификации регрессионной модели

Необходимость изменить функциональную форму регрессии возникает, если неверна одна из следующих гипотез, выполнение которых требуется для того, чтобы метод наименьших квадратов в применении к регрессионной модели (5.67) давал хорошие результаты.

1. Ошибки имеют нулевое математическое ожидание, или, что то же самое, математическое ожидание зависимой переменной является линейной комбинацией регрессоров:

$$\mathbf{E}(\varepsilon_i) = 0, \qquad \mathbf{E}(\boldsymbol{Y}_i) = \boldsymbol{b}\,\boldsymbol{X}_i.$$

2. Ошибки гомоскедастичны, т. е. имеют одинаковую дисперсию для всех наблюдений:

$$\mathbf{V}(\varepsilon_i^2) = \mathbf{E}(\varepsilon_i^2) = \sigma^2.$$

Если перечисленные предположения неверны, то в ошибке осталась детерминированная (неслучайная) составляющая, которая может быть функцией входящих в модель регрессоров. Это означает, что имеет место неверная спецификация функциональной формы. Заметить ошибку спецификации можно на глаз с помощью графиков остатков. Формальный тест на «нормальность остатков» можно провести с помощью вспомогательной регрессии квадратов остатков по "подозрительным" переменным, значимость которой можно оценить по обычной методике, описанной выше. Похожим образом обнаруживается и гетероскедастичность (отсутствие гомоскедастичности), которая проявляется в том, что разброс остатков меняется в зависимости от значения переменной ***X***. Тест на гетероскедастичность может быть выполнен с помощью LM-статистики, которая равна половине объясненной суммы квадратов из регрессии по константе и ***X*** и распределена асимптотически как χ^2 [Цыплаков, URL].

При проверке правильности спецификации нулевая гипотеза всегда состоит в том, что модель специфицирована корректно, а альтернативная гипотеза - в том, что имеется ошибка соответствующей спецификации. Если статистика незначима (например, уровень значимости больше 5%), то следует принять гипотезу о правильности спецификации.

Обычно умеренные ошибки в спецификации не приводят к серьезным последствиям, таким как несостоятельность оценок. При отклонениях остатков от нормальности МНК может давать не очень точные оценки, а вычисляемые ***t***- и ***F***-статистики не распределены в конечных выборках точно как ***t*** и ***F*** (хотя эти статистики остаются состоятельными, то есть их использование оправдано асимптотической теорией).

Учет нелинейности связи факторов в моделях линейной регрессии

В разделе 2.6 мы подробно останавливались на необходимости учета объективной нелинейности реального мира и проблемах нахождения модели оптимальной сложности.

Лучшим инструментом оценки качества приближения экспериментальных точек к расчетной кривой является человеческий глаз. Для проверки гипотезы о линейном характере связи ***Y*** и ***X*** более педантичные исследователи, например, делят область рассеяния точек наблюдения на четыре равных квадранта и подсчитывают количество экспериментальных точек, попавших в каждый квадрант: для линейной зависимости характерен четкий дисбаланс между частотами по обеим диагоналям. Более строгие методы проверки гипотезы о линейности связаны с анализом выборочных корреляционных отношений и представлены в главе 6.

Перечислим в таблице 5.15 основные виды функциональных форм регрессионной модели с примерами нелинейных зависимостей. Более сложные нелинейные модели связаны с использованием сплайнов, алгоритмов МГУА и др. [Розенберг с соавт., 1994].

Слово «линейный» в названии "линейный регрессионный анализ" указывает на линейность модели относительно параметров $\boldsymbol{a_j}$, но не факторов $\boldsymbol{x_j}$. Это означает, что мы можем проделать с выборкой ***X*** любые функциональные преобразования и включить преобразованные факторы в линейное уравнение.

Таблица 5.15

Функциональные формы регрессионных моделей с примерами нелинейных зависимостей

Функциональная форма	Примеры
Полиномиальная	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \ldots + a_k X^k$
Гиперболическая	$Y = a_0 / (a_1 + X)$
Семейство обращенных полиномиальных функций	$Y = \mathrm{a} / (a_0 + a_1 X + a_2 X^2 + \ldots + a_k X^k)$; $Y = X / (a_0 + a_1 X + a_2 X^2 + \ldots + a_k X^k)$; $Y = (b_0 + b_1 X + b_2 X^2 + \ldots + \mathrm{b}_l X^l) / (a_0 + a_1 X + a_2 X^2 + \ldots + a_k X^k)$;
Логлинейная	$\ln Y = a_0 + a_1 \ln X$
Обобщенная логарифмическая	$Y = a_0 + a_1 \ln X + a_2 X$
Степенная	$Y = a_0 + a_1 X^c$, где c – любое вещественное число
Экспоненциальная	$Y = a\, \mathrm{e}^{b X}$
Функция Гомперца	$\ln Y = b - a\, \mathrm{e}^{-X}$
Логистическая	$Y = c /(1 + \mathrm{e}^{(a + b X)})$
Экспоненциально-степенная	$Y = a\, \mathrm{e}^{b X} X^c$
Обращенная экспоненциальная	$Y = 1/(a + b^* \mathrm{e}^{-X})$
Показательная	$Y = a^* b^X$
Тригонометрическая	$Y = \beta_o + \beta_1 \sin\omega x + \beta_2 \cos \omega x$

Линеаризации, например, могут легко подвергаться полиномы

$$y_i = a_o + a_1 x_i + a_2 x_i^2 + \ldots + a_k x_i^{\mathrm{k}} + \varepsilon_{\mathrm{i}}, \qquad (5.81)$$

поскольку для имеющейся обучающей выборки (x_i, y_i), $i = 1, \ldots, n$, можно записать матричную форму:

$$Y = X A + \varepsilon, \quad \text{где} \quad X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots x_1^k \\ 1 & x_2 & x_2^2 & \cdots x_2^k \\ & \cdots & & \\ 1 & x_n & x_n^2 & \cdots x_n^k \end{bmatrix}.$$

В этом и в ряде других случаев расчет коэффициентов a_j можно свести к стандартной задаче линейной регрессии, и потому все формулы МНК оказываются справедливыми и для нелинейной формы аппроксимирующего уравнения.

Выбор между альтернативными функциональными формами обычно осуществляют на основе некоторого критерия, оценивающего точность подбора, в качестве которого обычно используется коэффициент детерминации (R^2). Чтобы учитывать при селекции простоту модели, делают поправку на количество регрессоров k. Это дает коэффициент детерминации, скорректированный на количество степеней свободы:

$$R_0^2 = 1 - \frac{(1 - R^2) n}{n - k}. \qquad (5.82)$$

Оценки метода наименьших квадратов являются одновременно и оценками метода максимального правдоподобия [Дрейпер, Смит, 1986]. Поэтому предлагается сравнивать модели на основе максимума логарифмической функции максимального правдоподобия ($\hat{L}$). Если учесть

при этом количество наблюдений (n) и ввести "штраф" за большое количество регрессоров (k), то получится *информационный критерий Акаике* (Akaike information criterion):

$$\mathrm{AIC} = -2/n\,(\hat{L} - k)\,. \tag{5.83}$$

Чем меньше AIC, тем лучшей считается модель.

В некоторых случаях, приведенных в таблице 5.15, нельзя свести нелинейную функцию $f(X)$ к линейной форме. В такой ситуации оценки наименьших квадратов параметров a и b приходится получать с помощью итерационных вычислительных процедур, осуществляющих последовательное приближение к минимуму суммы квадратов

$$Q(a,b) = \sum_{i=1}^{n} \left(Y_i - f(a,b,x_i)\right)^2. \tag{5.84}$$

Выполнить это приближение можно стандартными методами нелинейной оптимизации – методами Ньютона-Рафсона, Нелдера-Мидда и проч. [Банди, 1988; Гайдышев, 2001], включаемыми в большинство пакетов статистических программ.

В ряде случаев линеаризация оказывается возможной, но методически неверным является "обратное" преобразование линеаризированных уравнений. Например, в разделе 4.8 отмечалось, что для расчета большинства компонентов материального баланса отдельной особи канонизировано употребление параболических (степенных) зависимостей:

$$Y = a\,X^k\,, \tag{5.85}$$

где X и Y – произвольные биоэнергетические или аллометрические показатели, a и k – коэффициенты уравнения регрессии. Подавляющее большинство ученых-экологов (например, [Тодераш, 1984; Балушкина, 1987; Голубков, 2000]) в своих работах рекомендуют применять не вполне корректную схему расчета коэффициентов a и k путем нахождения параметров линейной регрессии вида

$$\lg Y = \lg a + k \lg X\ . \tag{5.86}$$

Определенная методологическая неточность присутствует и у А.А. Умнова [1976], когда он при стандартных допущениях регрессионного анализа предлагает найти в выражении $Y = aX^k$ значения коэффициентов a и k, обеспечивающие минимальность квадратичной формы

$$Q_e(a,k) = \sum_{i=1}^{n} (\lg Y_i - \lg a - k \lg X_i) \Rightarrow \min\ . \tag{5.87}$$

Нетрудно видеть из (5.84), что аксиоматика метода наименьших квадратов требует нахождения экстремума совсем иного функционала:

$$Q_e^{'}(a,k) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - aX_i^k) \Rightarrow \min\ . \tag{5.88}$$

Согласно известным положениям регрессионного анализа [Дрейпер, Смит, 1986] нелинейная функция может быть приведена к линейной форме (т.е. линеаризирована) только в том случае, если ошибка обоих уравнений регрессии остается ***аддитивной***, то есть зависимая переменная является суммой своего математического ожидания и ошибки. Пусть, например, исследователь из своих теоретических соображений желает получить степенную зависимость модели наблюдений $Y = a \cdot X^k$. Выполнив логарифмирование данных и применив МНК, он получает линейную в логарифмах зависимость $\lg Y_i = \lg a + k \lg X_i + \varepsilon_i$. Однако, при обратном преобразовании логлинейного уравнения (5.86) в уравнение (5.85), будет имеет место ***мультипликативное*** вхождение ошибок ε_i в нелинейное уравнение отклика: $Y_i = a \cdot X_i^k \cdot \varepsilon_i$. Последняя форма записи далеко не идентична стандартной регрессионной модели с аддитивными ошибками $Y_i = a \cdot X_i^b + \varepsilon_i$. Иными словами,

выражения (5.85) и (5.86) представляют собой два *принципиально разных* уравнения регрессии, поэтому использование для модели $\boldsymbol{Y = a \cdot X^k}$ МНК-оценок $(\boldsymbol{a}, \boldsymbol{k})$ из линеаризированного уравнения $\lg \boldsymbol{Y_i} = \lg \boldsymbol{a} + \boldsymbol{k} \lg \boldsymbol{X_i} + \varepsilon_i$, неизбежно приведет к заведомо искаженным результатам.

Из такой ситуации существует два возможных выхода:

- оценки наименьших квадратов параметров $\boldsymbol{a}$ и $\boldsymbol{k}$ для (5.85) получать с использованием итерационных процедур нелинейной оптимизации;
- просто-напросто "забыть" о существовании степенного уравнения (5.85) и во всех случаях анализа данных и последующего расчета составляющих баланса оперировать с $\boldsymbol{a}$ и $\boldsymbol{k}$, как с коэффициентами вполне корректного и самодостаточного логлинейного уравнения (5.86).

Результаты расчетов

> **Мем № 30**: *«По отношению к сообществам донных животных значения индекса Шеннона тесно коррелируют со значениями биотического индекса Вудивисса, отражающего степень загрязнения вод. Существенно, что между разнообразием системы и биомассой организмов существует вполне определенная обратная зависимость, которая по отношению к планктонным сообществам была количественно выражена А.М. Гиляровым, и подтверждена мною на основании данных по нескольким латгальским озерам»* А.Ф. Алимов [1990].

Рассмотрим, насколько этот тезис подтверждается для зообентоса рек средней полосы России. Сформируем выборку из 533 сопряженных наблюдений, куда включим следующие показатели зообентоса: индекс Шеннона $(\boldsymbol{H})$, биотический индекс Вудивисса $(\boldsymbol{V})$, число видов $(\boldsymbol{S})$, суммарную численность $(\boldsymbol{N_S})$ и биомассу $(\boldsymbol{B_S})$ всех видов в пробе, а также обилие отдельных таксономических групп – хирономид $(\boldsymbol{B_{chi}})$, олигохет[3] $(\boldsymbol{N_{oli}}, \boldsymbol{B_{oli}})$ и хищников-хватателей $(\boldsymbol{N_h})$.

Для оценки связи этих показателей выполним расчет подмножества регрессионных уравнений, основные параметры которых представлены в табл. 5.16. По каждой диагностической статистике приведены уровни значимости использованных критериев. Если $\boldsymbol{P}$-значение, соответствующее величине статистики, больше 5%, то нет оснований отвергать исходную нулевую гипотезу: об отсутствии влияния фактора, равенстве 0 коэффициента угла наклона $\boldsymbol{b}$, отличии распределения остатков от нормального или влиянии фактора на остатки соответственно.

Некоторые графики полученных зависимостей и гистограмм распределений показаны на рис. 5.12.

По результатам расчетов можно сделать следующие выводы:

1. На используемом фактическом материале отсутствует сколько-нибудь достоверная статистическая связь между индексом Шеннона и натуральными значениями суммарной биомассы зообентоса в пробе (см. уравнение 1 в табл. 5.16). Однако, после логарифмирования биомассы между $\boldsymbol{H}$ и $\ln(\boldsymbol{B_S})$ появляется вполне значимая ***прямо*** пропорциональная зависимость (см. уравнение 2 табл. 5.16 и фиг. «а» на рис. 5.12), качественно противоположная установленным на латгальских озерах.
2. Как обсуждалось выше (см. раздел 4.3), статистическая вариация индекса Шеннона включает в себя две компоненты: чисто функциональную зависимость от числа видов $\boldsymbol{S}$ и вклад фактора эквитабельности (равномерности распределения численности экземпляров по видам). Уравнение 4 демонстрирует достаточно тесную связь логарифмов числа видов и биомассы, что выглядит откровенным трюизмом для биологов (чем больше видов в пробе, тем больше биомасса). Поскольку уравнение 5, где откликом является пронормированное на число видов значение индекса Шеннона, значительно уступает по статистическим критериям надежности уравнениям 2 и 4, это заставляет предположить что прямо пропорциональная связь между $\boldsymbol{H}$ и $\boldsymbol{B}$ определяется, в основном, числом видов, а не эквитабельностью.

[3] Только виды-детритофаги без учета фитофагов *Nais communis*, *Stylaria lacustris* и др.

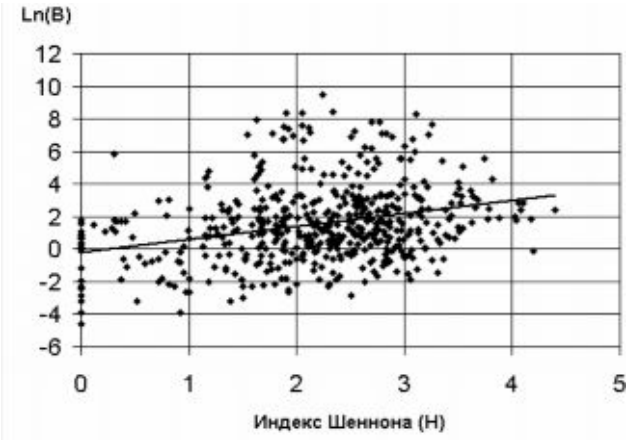
Таблица 5.16

Регрессионные зависимости между отдельными гидробиологическими показателями (обозначения по тексту)

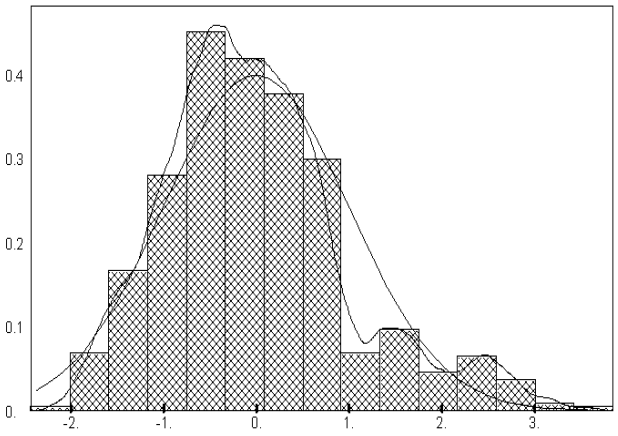
Уравнение регрессии	Коэффициент детерминации R^2, %	Значимость по F-критерию Фишера	Гипотеза b=0 по t-критерию Стьюдента	Распределение остатков	
				Нормальность по критерию χ^2	Гетероскедастичность по критерию χ^2
1. $H = 2.15 + 4.6\ 10^{-5} B_s$	-0.048	0.75 [0.38]	0.866 [0.38]	22.6 [0.0]	3.21 [0.07]
2. $H = 1.98 + 0.112 \ln(Bs)$	8.69	50.4 [0.0]	7.09 [0.0]	12.0 [0.0025]	6.99 [0.0082]
3. $\ln(B_s) = -0.159 + 0.79 H$	8.69	50.4 [0.0]	7.09 [0.0]	58.0 [0.0]	0.317 [0.57]
4. $\ln(S) = 1.935 + 0.136 \ln(B_s)$	16.89	106.48 [0.0]	10.31 [0.0]	54.0 [0.0]	6.56 [0.0104]
5. $H/\ln(S) = 0.873 + 0.0174 \ln(B_s)$	1.768	10.34 [0.0014]	3.216 [0.0014]	177.68 [0.0]	38.87 [0.0]
6. $H = 2.319 + 0.09 \ln(B_{Chi})$	4.35	22.49 [0.0]	4.74 [0.0]	10.88 [0.0043]	4.197 [0.0]
7. $H = 2.19 - 0.019 \ln(B_{Oli})$	~ 0	0.926 [0.3363]	0.96 [0.3363]	19.58 [0.0]	0.0053 [0.9418]
8. $H = 2.415 - 1.155 (N_{Oli}/N_s)$	16.53	103.8 [0.0]	-10.19 [0.0]	14.96 [0.0]	0.0587 [0.8084]
9. $H = 1.46 + 1.41 \ln(N_h/N_s)$	24.65	170.8 [0.0]	13.07 [0.0]	20.3 [0.0]	1.67 [0.195]
10. $H = 1.43 + 0.203 V$	22.66	153.0 [0.0]	12.37 [0.0]	4.45 [0.1078]	36.15 [0.0]

Примечание: в квадратных скобках приведены P-значения для соответствующей статистики.

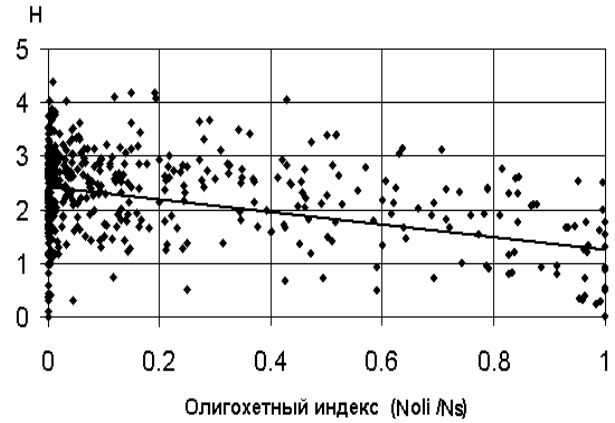
3. Если использовать в уравнениях биомассу отдельных таксономических групп бентоса, то их связь с индексом Шеннона для хирономид значительно слабее, а для олигохет-детритофагов вообще отсутствует (см. уравнения 6 и 7).
4. Уравнения 8 и 10 подтверждают распространенные выводы об обратной зависимости индекса Шеннона и индекса Гуднайта-Уитлея и Пареле и прямой зависимости H и биотического индекса Вудивисса (см. также фиг. «в» и «г» на рис. 5.11).
5. Наибольшая надежность модели регрессии была получена для прямо пропорциональной зависимости индекса Шеннона от доли логарифма численности видов хищников-хватателей (уравнение 9).
6. Выполненные расчеты показывают, что, на достаточно большой для гидробиологических данных обучающей выборке n, полученные коэффициенты корреляции от 0.3 до 0.5 при высоком F-критерии не следует оценивать как малые.
7. Практически для всех уравнений не отвергается гипотеза о нормальности распределения остатков, что связано, в первую очередь, с нормальностью распределения самого отклика. Этот вывод справедлив, если даже переставить местами отклик и фактор, как это сделано для уравнения 3 (см. фиг. «б» на рис. 5.12).
8. В пяти приведенных случаев из десяти гетероскедастичность остатков отсутствует. Однако, даже визуально на фиг. «в» рис. 5.12 можно увидеть значительную сконцентрированность вариации остатков в области малых значений олигохетного индекса. Если бы К.Г. Гуднайт, Л.С. Уитли и Э.А. Пареле использовали при расчете своего индекса прологарифмированные значения численностей, статистическая надежность уравнения 8 существенно бы повысилась.



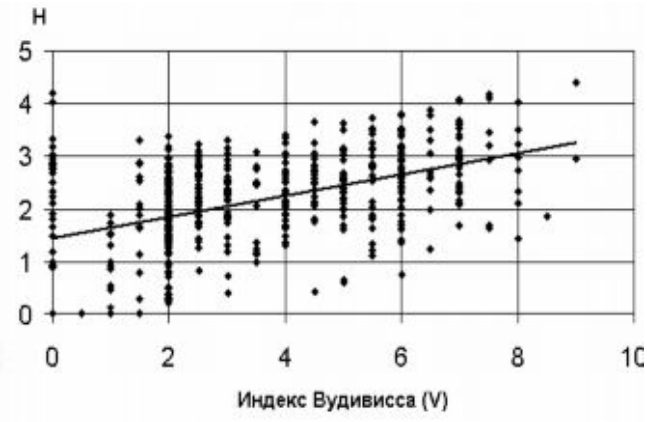
а) график зависимости 3 : $\ln(B_s)=a + bH$



б) гистограмма остатков зависимости 3.



в) график зависимости 8: $H = a + b\ (N_{oli}/N_s)$



г) график зависимости 10: $H = a + b\ V$

Рис. 5.12. Графики зависимостей между индексом Шеннона и некоторыми гидробиологическими показателями (уравнения приведены в таб. 5.15)

Мем № 31: *«Для практического использования при определении качества воды предлагаются следующие методы: система сапробности; биотический индекс Вудивисса в нашей модификации. Для углубленного изучения экологических условий водного объекта можно использовать индекс видового разнообразия Шеннона-Уивера»* В.А. Яковлев [1988].

Рассмотрим, подтверждаются ли распространенные утверждения о закономерном влиянии факторов гидрохимического загрязнения на такие показатели как индекс Шеннона ***H***, олигохетный индекс Гуднайта-Уитлея и Пареле ***P*** и биотический индекс Вудивисса ***V***, полученные для малых рек Самарской области. Рассчитанные уравнения линейной регрессии зависимости этих индексов от традиционных гидрохимических показателей представлены в табл. 5.17 и на рис. 5.13.

Выполненные расчеты показывают существенно более слабую корреляционную связь гидробиологических индексов с уровнем загрязнения, чем это обычно декларируется. Низкую статистическую значимость уравнений регрессии следует объяснить, видимо, двумя причинами: *объективной* и *технологической*.

Объективная составляющая заключается в том, что обобщенные индексы ***H***, ***P*** и ***V*** работоспособны, по нашему мнению, лишь в случае ощутимых различий в степени антропогенного воздействия на сравниваемые экосистемы. В изучаемом регионе уровень химического загрязнения воды относительно однороден, за редкими исключениями (устье р. Чапаевка), которые мало влияют на общую статистическую оценку. Поэтому вариабельность общих индексов определяется в значительной мере природно-географическими факторами и особенностями биотопов, нежели концентрациями химических ингредиентов. Влияние токсического загрязнения на экологию гид-

робионтов следует оценивать не в "общестатистическом болоте" данных, а по совокупности специально подобранных примеров, полярно разнесенных на шкале "патология – норма", и с учетом всех структурных изменений видовых комплексов биоценозов (о чем пойдет речь в последующих главах).

Таблица 5.17

*Регрессионные зависимости гидробиологических показателей (индекс Шеннона **H**, олигохетный индекс Гуднайта-Уитлея и Пареле **P** и биотический индекс Вудивисса **V**) от гидрохимических факторов*

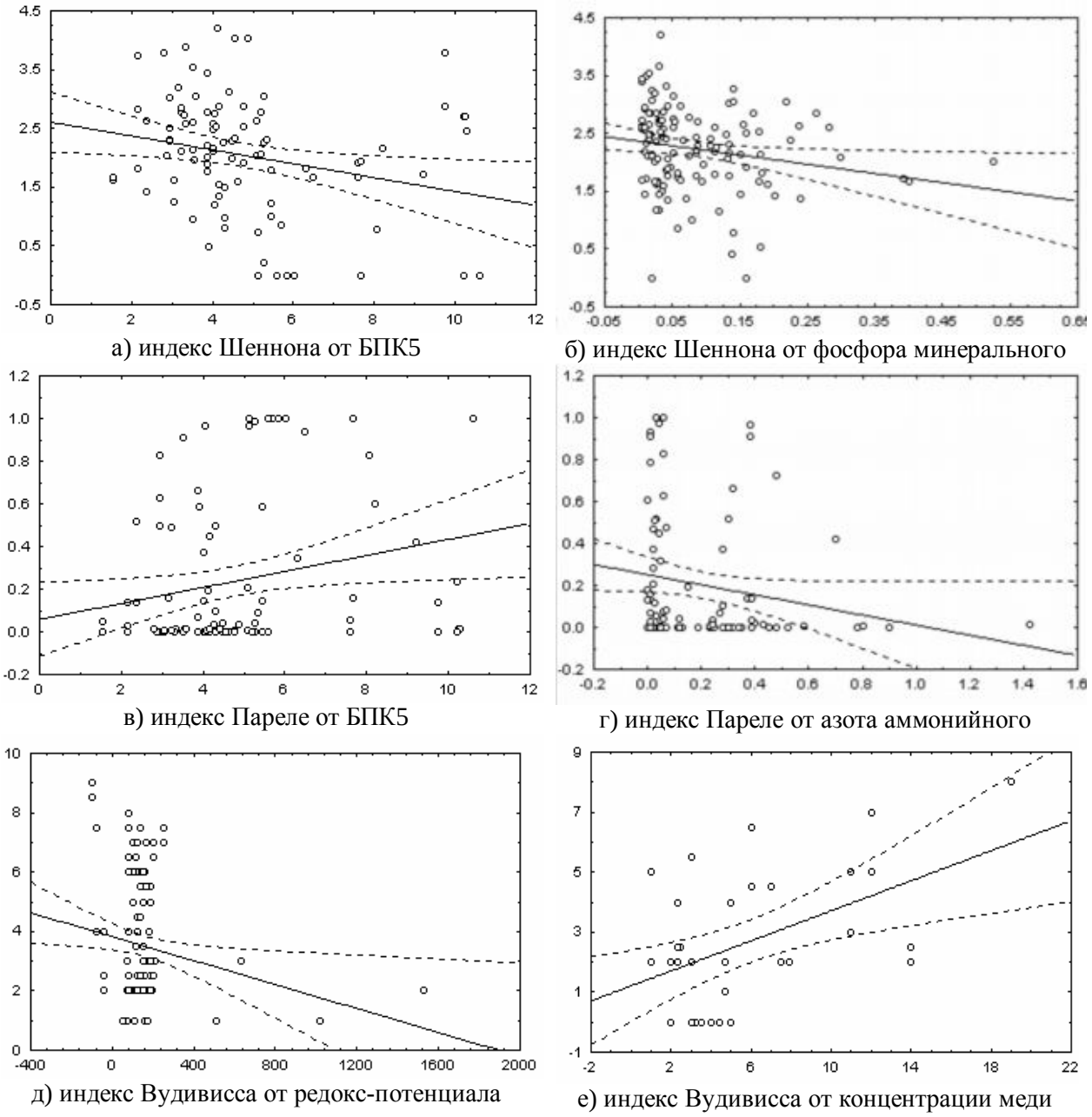
Наименование фактора, объем выборки и ее размах (*min-max*)	Уравнения регрессии (***C*** – гидрохимический фактор из столбца 1)	Коэффициент детерминации R^2, %	***F***-критерий Фишера и его ***p***-значение
pH в придонном слое, 331 изм., 5.3 ÷ 9.8	***H*** = 2.18 + 0.001****C***	0.303	10^{-4} [0.99]
	P = 0.464 - 0.031****C***	0.046	0.845 [0.358]
	V = 6.68 - 0.351****C***	0.368	2.22 [0.137]
Растворенный кислород (на дне), 306 изм., 6.5 ÷ 184 мг/л	***H*** = 2.49 - 0.002****C***	0.065	0.799 [0.37]
	P = 0.0785 + 0.0012****C***	0.467	2.43 [0.12]
	V = 4.35 - 0.005****C***	0.062	**0.81 [0.36]**
Редокс-потенциал (Eh) в поверхностном слое, 139 изм., 100÷1530 мВ	***H*** = 2.403 - 0.001****C***	4.78	**7.93 [0.0056]**
	P = 0.254 + 0.0003****C***	2.54	**4.6 [0.034]**
	V = 3.819 - 0.25****C***	3.11	**5.43 [0.02]**
Фосфор минеральный, 125 изм., 0.005 ÷ 0.525 мг/л	***H*** = 2.363 - 1.59****C***	2.99	**4.82 [0.029]**
	P = 0.203 + 0.598****C***	2.02	3.568 [0.061]
	V = 3.497 - 1.041****C***	0.537	0.336 [0.56]
Аммонийный азот, 90 изм., 0.001 ÷ 1.42 мг/л	***H*** = 2.196 - 0.172**C*	0.856	0.24 [0.62]
	P = 0.251 - 0.24****C***	2.85	**3.61 [0.06]**
	V = 3.7 - 0.8****C***	0.036	0.967 [0.32]
$БПК_5$, 99 изм., 1.55 ÷ 10.59, мг/л	***H*** = 2.6 - 0.118****C***	4.61	**5.74 [0.018]**
	P = 0.0575 + 0.0375****C***	3.87	**4.95 [0.028]**
	V = 3.89 - 0.8****C***	2.629	3.64 [0.059]
Железо, 71 изм., 0.001 ÷ 2.6 мг/л	***H*** = 2.35 - 0.477****C***	2.99	3.16 [0.08]
	P = 0.2 + 0.12****C***	0.512	1.36 [0.24]
	V = 3.89 - 0.8****C***	1.38	0.045 [0.83]
Медь, 33 изм., 1 ÷ 19 мкг/л	***H*** = 2.05 + 0.0323****C***	1.29	0.589 [0.44]
	O = 0.367 - 0.013****C***	0.497	0.84 [0.36]
	V = 1.202 + 0.25****C***	21.7	**9.9 [0.0036]**

Примечание: жирным шрифтом ***F***-критерия выделены статистически значимые уравнения регрессии.

Статистическая надежность уравнений регрессии зависит и от ***технологии*** работы с данными: тщательности подбора аппроксимирующей функции и выполнения исходных предпосылок анализа (нормальности и независимости остатков). Приведем несколько комментариев этому тезису:

1. Нетрудно заметить сильную асимметрию распределения олигохетного индекса Пареле в сторону тяжелого "нулевого хвоста" (т.е. в нижнюю часть графиков на фиг. «в» и «г» рис. 5.13). Нами уже отмечалось, что при расчете любых индексов, построенных на долях численности или биомассы отдельных таксономических групп, предпочтительнее использовать вместо натуральных сумм их прологарифмированные значения.
2. Можно усмотреть в корреляционном облаке точек определенный вид нелинейной зависимости, которая окажется предпочтительнее, чем линейная. Например, использование для фиг. «в» степенной функции ***P*** = -0.163 + 0.186 $(БПК_5)^{0.5}$ увеличивает коэффициент детерминации с 3.87 до 4.63%, а ***F***-отношение – с 4.95 до 5.76 (***p*** = 0.0183).
3. Наконец, очевидна асимметрия распределения большинства гидрохимических показателей и, связанная с этим гетероскедастичность остатков. Наилучший способ борьбы с этим явлением –

логарифмирование переменных. В частности, для фиг. «д» предпочтительнее уравнение $\boldsymbol{V} = 28.53 - 3.55 \ln(1000 + Eh)$, $\boldsymbol{R^2} = 4.24$, $\boldsymbol{F}(1,137) = 7.1$, $\boldsymbol{p} = 0.0086$.



а) индекс Шеннона от БПК5

б) индекс Шеннона от фосфора минерального

в) индекс Пареле от БПК5

г) индекс Пареле от азота аммонийного

д) индекс Вудивисса от редокс-потенциала

е) индекс Вудивисса от концентрации меди

Рис. 5.13. Графики зависимостей гидробиологических индексов по оси ординат от гидрохимических показателей (уравнения приведены в таб. 5.17)

Однако никакая методологическая тщательность обработки не гарантирует от артефактов в виде распространенного феномена "ложной" корреляции. Например, мы затрудняемся дать какое-либо осмысленное объяснение чрезвычайно достоверной прямо пропорциональной (!) связи индекса Вудивисса и концентрации ионов меди (см. фиг. «е» рис. 5.13).

Мем № 32: *«Несмотря на большое число исследований, рассматривающих зависимость скорости обмена от индивидуального веса, не было встречено случаев, показывающих необходимость или желательность применения для той же цели других функций. Впрочем, сделана попытка опровергнуть какое-то "реальное существование" (?) степенной зависимости R от W на основании того, что в отдельных случаях…данные аппроксимируются не только прямой… Нельзя принимать всерьез эти предельно наивные высказывания»* Г.Г. Винберг [1976].

В математической части была показана недостаточная обоснованность методов линеаризации при оценке параметров степенного уравнения регрессии. Оценим, насколько практически отличаются между собой коэффициенты степенного (5.85) и логлинейного (5.88) уравнений, оптимальные относительно сумм квадратов невязок (5.86) и (5.87), соответственно.

Воспользуемся для этого выборкой [Умнов, 1976], связывающей величину энергетического обмена с массой тела пескожила *Arenicola marina*. А.А. Умновым для этой серии наблюдений после логарифмирования было рассчитано линейное уравнение регрессии

$$\log \boldsymbol{Y} = 1.804 + 0.672\boldsymbol{X} .$$

Поскольку 26 лет назад методы нелинейного программирования были своего рода "вычислительной экзотикой", автор рекомендует использовать найденные коэффициенты для степенной модели, которая после обратного преобразования приобретает вид

$$\boldsymbol{Y} = 63.75\boldsymbol{X}^{0.672} . \qquad (5.89)$$

В настоящее время любой исследователь может легко вычислить истинные МНК-оценки нелинейной функции (5.84) и получить для данных, представленных в табл. 5.18, оптимальное степенное уравнение энергетического обмена пескожила

$$\boldsymbol{Y} = 53.36\boldsymbol{X}^{0.785} . \qquad (5.90)$$

Если принять во внимание, что по литературным данным (см. раздел 4.8) для различных групп зообентоса показатель степенного уравнения $\boldsymbol{k}$ варьируется от 0.75 до 0.895, то отличия уравнений (5.89) и (5.90) можно признать существенными. Достаточно значительным (на 17%) имело место и снижение остаточной суммы квадратов $\boldsymbol{Qe}$.

Из графика на рис. 5.14 легко увидеть, что уравнение (5.89), совпадая с (5.90) в средней части, плохо учитывает расположение экспериментальных точек на "хвостах" кривой.

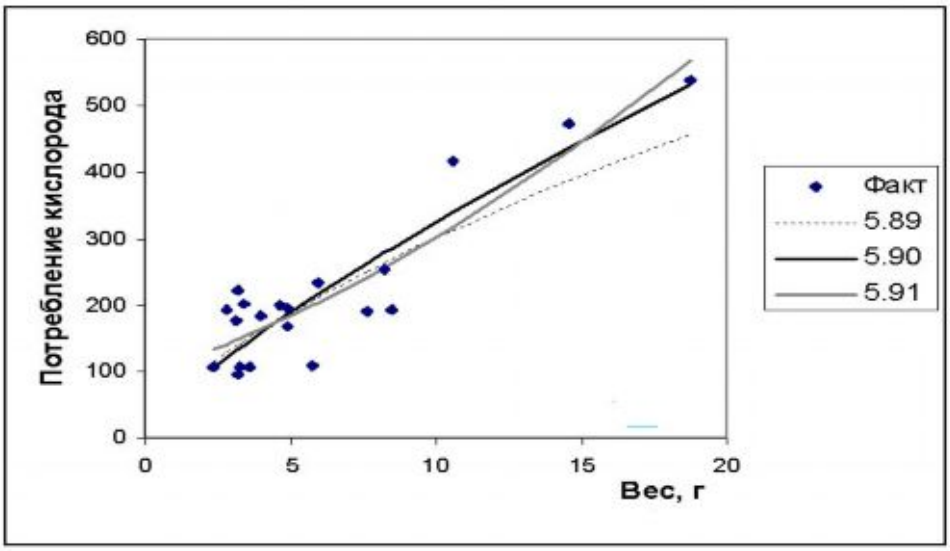


Рис. 5.14. Графики аппроксимирующих функций (номера по тексту) потребления кислорода от веса тела пескожила

Таблица 5.18

Расчет энергетического обмена в зависимости от индивидуального веса пескожила по различным эмпирическим уравнениям (цифровые обозначения по тексту)

Индивидуальный вес, г/экз	Энергетический обмен, мм^3 O_2/час экз				
	Получены в эксперименте	Рассчитаны по уравнениям регрессии			
		(5.89)	(5.90)	(5.91)	(5.92)
2.37	105.33	113.89	**105.03**	132.66	160.76
2.82	192.63	128.01	120.38	140.85	**156.20**
3.12	177.13	137.02	130.31	146.43	**154.51**
3.21	221.84	139.67	133.26	148.12	**154.20**
3.22	95.9	139.96	**133.58**	148.31	154.18
3.24	107	140.54	**134.23**	148.68	154.12
3.38	202.5	144.60	138.76	151.34	**153.88**
3.6	105.4	150.86	**145.80**	155.55	153.92
4	182.38	161.94	158.37	**163.33**	155.33
4.65	198.2	**179.19**	178.23	176.35	161.01
4.91	167.45	185.87	186.00	181.68	**164.39**
4.93	195	186.38	**186.60**	182.09	164.68
5.72	108.8	205.97	209.68	198.75	**178.60**
5.95	233.13	211.51	**216.27**	203.72	183.57
7.65	190.17	250.45	263.41	242.22	**230.90**
8.26	254.35	263.71	279.76	**256.78**	251.55
8.5	191.88	268.84	286.12	262.62	**260.07**
10.62	416.87	312.26	340.75	316.83	**341.50**
14.6	472.75	386.79	437.42	431.46	**485.71**
18.76	538.34	457.82	532.52	569.22	**537.81**
Остаточная сумма квадратов отклонений $\boldsymbol{Q_e}$		66656	56935	49973	42359
Скорректированный коэффициент детерминации $\boldsymbol{R^2_o}$, %		73.7	77.6	79.1	81.2

Примечание: жирным шрифтом выделены значения с минимальной ошибкой относительно экспериментальных данных

Мем, предваряющий этот блок расчетов, относится к любопытной категории – "указующим мемам". При этом Г.Г. Винберг [1976] признает: *«нет оснований считать, что когда данные могут быть переданы степенной функцией, это указывает на некоторый определенный механизм явления»*, т.е. на самом деле никакой биофизической закономерности уравнение (5.85) не отражает. Это означает, что с равным правом можно предложить целый ряд иных математических формул, которые будут описывать те же экспериментальные данные с большей степенью адекватности.

Мы бы сочли проявлением собственной "наивности", если бы лишили себя возможности проверить, нет ли иной формы зависимости, которая бы лучше аппроксимировала данные по энергетическому обмену пескожила, чем канонизированное степенное уравнение. Наши поиски не были слишком упорными – первые же тестируемые полиномиальные уравнения 2-й и 3-й степени оказались существенно более адекватными по отношению к экспериментальным данным (коэффициент детерминации в табл. 5.18 рассчитан по формуле (5.82) и учитывает количество регрессоров):

$$Y = 0.5298\boldsymbol{X}^{2} + 15.441\boldsymbol{X} + 93.093\ ; \qquad (5.91)$$

$$Y = -0.2492\boldsymbol{X}^{3} + 8.0469\boldsymbol{X}^{2} - 46.844\boldsymbol{X} + 229.9. \qquad (5.92)$$

Согласно *принципу множественности моделей* В.В. Налимова [1971], список подобных эмпирических зависимостей может быть существенно продолжен.

Окончательная селекция лучших из них целесообразна, вероятно, исходя из совокупных соображений "ошибка регрессии" + "удобство использования". Если под фактором "удобства" понимать визуальное сравнение коэффициентов, оцененных для разных таксономических групп животных, то выбор степенной зависимости в качестве единого стандарта интерпретации и обобщения следует признать обоснованным.

Однако, в случае разработки ряда имитационных моделей материально-энергетического баланса экосистем [Меншуткин, 1971], применение, например, полиномиальных уравнений значительно предпочтительней, поскольку схема расчета их коэффициентов методом линеаризации является статистически корректной и не требует громоздких итеративных алгоритмов нелинейной оптимизации.

5.6. Непараметрическая корреляция и регрессия

Формулировка задачи

Исходные условия те же, что и для регрессионного анализа: т.е. имеется две группы сопряженных наблюдений

$$\boldsymbol{X} \equiv (x_1, \ldots, x_m) \text{ и } \boldsymbol{Y} = (y_1, \ldots, y_m) .$$

Если есть сомнения в применимости гауссовой модели распределения данных (а они, в большинстве случаев, небезосновательны), то для оценки связи между переменными $\boldsymbol{Y}$ и $\boldsymbol{X}$ можно воспользоваться некоторыми альтернативами метода наименьших квадратов. Обсуждая в разделе 5.4 измерения в порядковых (ординальных) шкалах, мы убедились, что реальным содержанием этих измерений является тот порядок, в котором выстраиваются объекты по степени выраженности измеряемого признака. Порядковый номер числа в таком отсортированном списке называется его *рангом*.

Необходимо оценить степень влияния признака $\boldsymbol{X}$ на степень выраженности отклика $\boldsymbol{Y}$. Если такого влияния нет, то справедлива нулевая гипотеза $\boldsymbol{H}_o$ о независимости порядковых признаков. Решение этой задачи будем искать, основываясь на рангах измерений.

Рекомендуемая литература та же, что и для раздела 5.4.

Математический лист

Пусть каждому $\boldsymbol{i}$-му измерению приписана пара натуральных чисел (r_i, s_i), где r_i – ранг x_i среди чисел $(x_1, \ldots, x_m)$, а s_i – ранг y_i среди чисел $(y_1, \ldots, y_m)$. Будем при этом считать, что среди рядов чисел $\boldsymbol{X}$ и $\boldsymbol{Y}$ нет повторяющихся значений, так что переход к рангам вопросов не вызывает.

Если признаки $\boldsymbol{X}$ и $\boldsymbol{Y}$ взаимосвязаны, то последовательность рангов $r_1, r_2, \ldots, r_m$ влияет на ранговую последовательность $s_1, s_2, \ldots, s_m$; в противном случае порядок среди $\boldsymbol{Y}$ ***случаен*** по отношению порядка среди $\boldsymbol{X}$. Поэтому центральным моментом обсуждения гипотезы $\boldsymbol{H_o}$ является оценка, насколько являются ранги $s_1, s_2, \ldots, s_m$ равновозможными (т.е. равновероятными) при любом порядке чисел $r_1, r_2, \ldots, r_m$. Вторым важным моментом является выбор меры сходства двух наборов рангов.

Коэффициент ранговой корреляции, предложенный в 1900 г. знаменитым психологом Ч. Спирменом, основан на том, что близость этих двух рядов чисел отражает величина:

$$\boldsymbol{S} = \sum_{i=1}^{m} (r_i - s_i)^2 , \qquad (5.93)$$

которая варьируется от 0 , если последовательности полностью совпадают, до $(m^3 - m)/3$, когда последовательности рангов полностью противоположны.

Нормированный по своему максимальному значению, *коэффициент ранговой корреляции Спирмена*

$$\rho = 1 - \frac{6S}{m^3 - m} \tag{5.94}$$

варьирует от +1 до –1 и свои крайние значения принимает в случаях полной предсказуемости одной ранговой последовательности по другой. Заметим, что значение S не зависит ни от значения первого номера последовательности, ни от порядка сортировки.

Другой коэффициент ранговой корреляции, получивший популярность после работ М. Кендалла, в качестве меры сходства между двумя ранжировками использует минимальное число перестановок, которое надо осуществить между соседними объектами, чтобы одно упорядочение объектов превратить в другое.

Статистику Кендалла K подсчитывают следующим образом. Выстраивают сопряженные наблюдения в порядке возрастания признака X и для каждого значения y_i определяют его ранг s_i. На последовательности рангов $s_1, s_2, \ldots, s_m$ определяют количество *инверсий*, т.е. нарушений порядка следования. Например, при $m = 4$ и последовательности рангов $\{4, 3, 1, 2\}$ имеем количество инверсий (суть – статистику Кендалла) $K = 3 + 2 = 5$, где 3 – количество инверсий для числа 3 и 2 – количество инверсий для числа 3. Наименьшее возможное значение $K = 0$ получается при полном совпадении ранговых последовательностей, наибольшее $K = m(m-1)/2$ – при полной их противоположности.

Коэффициент ранговой корреляции по Кендаллу представляет собой статистику, нормированную по ее максимуму, и изменяется в тех же пределах, что и коэффициент корреляции Спирмена

$$\tau = 1 - \frac{4K}{m(m-1)}. \tag{5.95}$$

Статистика τ Кендалла эквивалентна ρ Спирмена как по мощности, так и по выполнению основных предположений. Обычно, однако, числовые значения ρ Спирмена и τ Кендалла различны, потому что они отличаются как своей внутренней логикой, так и способом вычисления. Более важно то, что статистики Кендалла и Спирмена имеют различную интерпретацию: если коэффициент корреляции Спирмена может рассматриваться как прямой аналог коэффициента корреляции r Пирсона, вычисленный по рангам, то статистика Кендалла скорее основана на подсчете *вероятностей* (выражаясь более точно, проверяется наличие различий между вероятностями порядка расположения наблюдаемых данных для двух величин).

Если в данных имеется много совпадающих значений, то предпочтительнее использовать третью ранговую статистику γ – *критерий*, который по своей интерпретации и вычислениям эквивалентен статистике Кендалла, за исключением того, что совпадения явно учитываются в нормировке. Выражаясь кратко, γ представляет собой разность между вероятностью того, что ранговый порядок двух переменных совпадает, минус вероятность того, что он не совпадает, деленную на единицу минус вероятность совпадений.

Для проверки предположения об отсутствии связи между признаками надо вычислить выборочное значение любого коэффициента ранговой корреляции и сравнить его с критическим значением для данного уровня значимости. Нулевую гипотезу H_o следует отвергнуть, если полученное в опыте значение коэффициентов ρ или τ по модулю превосходит критическое.

Критические значения ранговых критериев можно найти по таблицам, либо вычислить по приближенным формулам, которые основаны на том, что при H_o и с увеличением m распределение случайных величин $\rho\sqrt{m-1}$ и $\tau\sqrt{\frac{9m(m+1)}{2(2m+5)}}$ асимптотически приближается к стандартному нормальному закону $N(0,1)$.

Результаты расчетов

В результате гидробиологических наблюдений установлено, что при загрязнении водоемов происходит закономерное изменение соотношения численности личинок хирономид подсемейств Chironominae, Orthocladiinae и Tanypodinae. Ортокладиины обычно доминируют в чистых водах, таниподины – в загрязненных, что дало основания Е.В. Балушкиной предложить индекс, отражающий соотношение обилия представителей этих трех подсемейств и описанный в разделе 4.2.

Проверим справедливость этого предположения с использованием ранговых критериев ρ, τ и γ (см. табл. 5.19, где приведены статистики и соответствующие им значения вероятностей p). Расчет был выполнен по выборке из 88 наблюдений, а в качестве показателя загрязнения водоема использовались значения биологического потребления кислорода $БПК_5$.

Таблица 5.19

*Корреляция между $БПК_5$ и обилием различных групп личинок хирономид с использованием ранговых критериев (**N** – суммарная численность представителей группы в пробе, **A** – отношение численности группы к общей численности хирономид)*

Наименование показателя и подсемейства хирономид	Коэффициент Спирмена		Коэффициент Кендалла		Статистика γ	
	ρ	p	τ	p	γ	p
N Chironominae	-0.162	0.131	-0.108	0.135	-0.110	0.135
N Orthocladiinae	**-0.271**	**0.011**	**-0.196**	**0.007**	**-0.253**	**0.007**
N Tanypodinae	-0.205	0.055	-0.141	0.051	-0.161	0.051
A Chironominae	0.190	0.076	0.129	0.076	0.135	0.076
A Orthocladiinae	**-0.251**	**0.018**	**-0.185**	**0.011**	**-0.238**	**0.011**
A Tanypodinae	-0.159	0.139	-0.113	0.118	-0.128	0.118
Индекс Балушкиной	**0.251**	**0.018**	**0.185**	**0.011**	**0.238**	**0.011**

Примечание: жирным шрифтом выделены строки со значимым влиянием фактора

Проведенные расчеты позволяют сделать следующие выводы.

1. На представленном массиве гидробиологических данных выявлена достоверная обратно пропорциональная связь между $БПК_5$ и численностью представителей подсемейства Orthocladiinae.
2. На водотоках Самарской области не подтверждается вывод о влиянии органического загрязнения в диапазоне до 10 $мгO_2/л$ на абсолютную или относительную численность подсемейства Chironominae.
3. Нет веских оснований использовать, как это делает Е.В. Балушкина, в качестве анализируемого показателя относительную (в %) долю численности гидробионтов отдельных подсемейств вместо натурального значения общей численности. Например, если по абсолютной численности Tanypodinae достоверность нулевой гипотезы находится на уровне порога значимости, то для относительной численности гипотезу об отсутствии связи следует принять без колебаний.
4. Несмотря на то, что хирономидный индекс имеет значимую прямо пропорциональную связь с уровнем органического загрязнения, эта корреляция практически полностью основывается на удельной составляющей Orthocladiinae, поэтому прагматическая ценность конечного математического выражения для индекса Балушкиной в условиях рассматриваемого примера не определена.
5. Все три используемых ранговых коэффициентов корреляции продемонстрировали весьма близкие результаты оценки значимости нулевой гипотезы, что свидетельствует о надежности полученных выводов.

Представляемый пример демонстрирует также высокую технологичность ранговых коэффициентов корреляции по сравнению с коэффициентом корреляции Пирсона в условиях негауссовых распределений анализируемых выборок. Для доказательства этого выполним параллельный расчет уравнений линейной регрессии классическим методом наименьших квадратов (графики рассчитанных зависимостей представлены на рис. 5.15):

- для численности ортокладиин:
  $\boldsymbol{N_{Ort}}$ = 107.23 - 7.02 * $БПК_5$
  при коэффициенте корреляции $\boldsymbol{r}$ = 0.068, статистике Фишера $\boldsymbol{F}(1,86)$ = 0.41 и уровне значимости $\boldsymbol{p}$ = 0.526;
- для индекса Балушкиной
  $\boldsymbol{K_b}$ = 7.35 + 0.33 * $БПК_5$,
  при $\boldsymbol{r}$ = 0.179, $\boldsymbol{F}(1,86)$ = 2.85, $\boldsymbol{p}$ = 0.095 .

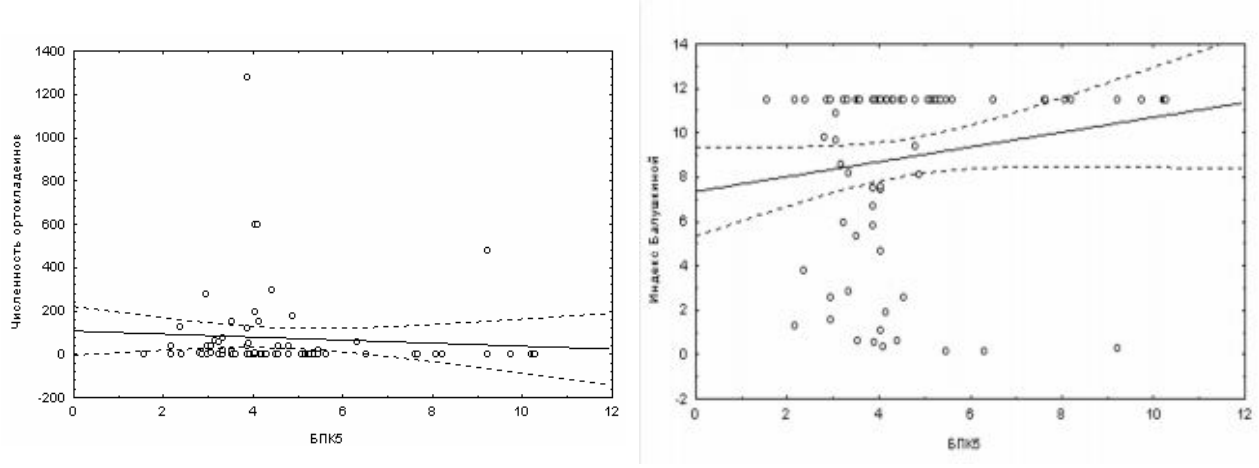


Рис. 5.15. Графики зависимостей численности ортокладиин и индекса Балушкиной от значения $БПК_5$

Поскольку основные предположения регрессионного анализа на этих выборках не выполняются, полученные уравнения и коэффициенты корреляции оказались недостоверными. В то же время, использование непараметрических критериев дало уверенное заключение о наличии такой связи.

Глава 6. Таблицы сопряженности и «интервальная» математика

6.1. Оценка зависимости признаков в таблицах сопряженности

Формулировка задачи

Пусть имеется ряд из m сопряженных наблюдений двух переменных $A \equiv (a_1, ..., a_m)$ и $B \equiv (b_1, ..., b_m)$, причем, предполагается, что A – независимая переменная (фактор) влияет на значения B – зависимой переменной (отклик). При этом типы данных, в которых представлены показатели, носят вполне определенный характер: они должны быть измерены в классификационных или порядковых шкалах, либо сведены к таковым в ходе предварительной обработки.

Предположим, что признак A имеет r градаций (или уровней) $A_1, A_2, \ldots, A_r$, а признак B подразделяется на s градаций $B_1, B_2, \ldots, B_s$. В "свернутом" виде результаты наблюдений можно представить *таблицей сопряженности*, состоящей из r строк и s столбцов, в ячейках которых проставлены частоты событий n_{ij}, т.е. количество объектов выборки, обладающих комбинацией уровней A_i и B_j.

Если между переменными A и B имеется взаимно однозначная прямая или обратная функциональная связь, то все частоты n_{ij} концентрируются по одной из диагоналей таблицы. При связи не столь сильной некоторое число наблюдений попадает и на недиагональные элементы. В этих условиях перед исследователем стоит задача выяснить, насколько точно можно предсказать значение одного признака по величине другого.

В отличие от регрессионного анализа, в данном случае нас интересует не сколько конкретный вид расчетного уравнения $B \approx f(A)$, а надежная и непротиворечивая оценка степени и характера влияния фактора на зависимую переменную. Говоря языком статистики, надо указать распределение вероятностей между возможными значениями второго признака при известном значении первого. Этой проблеме обычно предшествует более простая процедура: надо сначала проверить, существует ли вообще какая-либо связь между этими признаками, или же они ведут себя *независимо* друг от друга.

Рекомендуемая литература: [Елисеева, Рукавишников, 1977; Аптон, 1982; Енюков, 1986; Флейс, 1989].

Математический лист

Проверка нулевой гипотезы

Признаки A и B будут независимыми, если значение, принятое признаком A не влияет на вероятности возможных значений признака B:

$$P(B_j/A_i) = P(B_i) \quad \text{или} \quad P(A_i,B_j) = P(A_i)\,P(B_j) \tag{6.1}$$

Значения использованных вероятностей нам неизвестны, однако, по теореме Бернулли, при большом объеме выборки $(n \to \infty)$ частоты в ячейках таблицы сопряженности будут являться оценками этих вероятностей. При выполнении гипотезы о независимости признаков справедливо

$$p_{ij} = p_{i\bullet} \cdot p_{\bullet j}\,, \tag{6.2}$$

где следующие величины трактуются как ожидаемые частоты:

$$\frac{n_{ij}}{n} \to p_{ij}\,; \quad \frac{n_{i\bullet}}{n} \to p_{i\bullet}\,; \quad \frac{n_{\bullet j}}{n} \to p_{\bullet j}\,; \quad n = \sum_{i=1}^{r}\sum_{j=1}^{S} n_{ij}\,,$$

(замена индекса точкой означает результат суммирования по этому индексу). Тогда проверка нулевой гипотезы сводится к оценке, насколько близки значения фактических и ожидаемых частот, т.е.

$$n_{ij} \approx \frac{n_{i\bullet} n_{\bullet j}}{n}\,. \tag{6.3}$$

Методы сравнения эмпирических (H) и теоретических (T) частот по А. Брандту (A. Brandt) и Г. Снедекору (G. Snedecor) основываются на расчете критерия согласия χ^2, оценивающего меру близости по всем ячейкам таблицы сопряженности:

Карл ПИРСОН
(Carl Pearson, .1857-1936)
английский математик, биолог-евгеник, философ-позитивист

$$\chi^2 = \sum \frac{(H-T)^2}{T} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} . \qquad (6.4)$$

Если в конкретном опыте величина χ^2 оказывается чрезмерно большой, то приходится признать, что ожидаемые частоты слишком сильно отличаются от наблюдаемых. Ответ на естественный вопрос, о том, какие значения статистики следует считать чрезмерно большими, дает теорема К. Пирсона – Р. Фишера, из которой следует:

- для независимых признаков при неограниченном росте числа наблюдений распределение случайной величины χ^2 стремится к распределению «хи-квадрат»;
- гипотезу о независимости можно принять, если χ^2 не превосходит критического для заданного уровня табличного значения с $(r-1)(s-1)$ степенями свободы;
- для зависимых признаков χ^2 неограниченно возрастает с увеличением n.

В 1934 г. Ф. Иэйтс (F. Yates) предложил ввести в выражение для статистики χ^2 так называемую поправку на непрерывность, которая связана с тем, что непрерывные распределения хи-квадрат и, соответственно, нормальное распределение используются для представления дискретных выборочных частот. С учетом такой поправки данное выражение примет следующий вид:

$$Y^2 = 2\sum H \cdot \ln \frac{H}{T} = 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} \cdot \ln\left(\frac{n_{ij}n}{n_{i\bullet}n_{\bullet j}}\right) . \qquad (6.5)$$

Оценка силы связи

Как всегда в статистике, интерес исследователя не ограничивается принятием гипотезы, оценивающей величину риска предположения о существовании связи. Если признаки оказались взаимосвязаны (т.е. гипотеза об их независимости была проверена и отвергнута) представляет интерес оценка силы связи, которую хочется видеть в некотором привычном интервале величин, например, от –1 до +1 с нулевым значением при отсутствии связи. Сама по себе такая постановка проблемы определенным образом дискуссионна. Достаточно сказать, что нет единого мнения даже у соавторов настоящей книги: один из нас считает приоритетным при оценке силы связи уютный коэффициент корреляции Пирсона r (суть – долю факториальной вариации), а другой – статистики Фишера или χ^2 (то же, но только с учетом степеней свободы), напрямую связанные с фундаментальными для статистики уровнями значимости.

В случае таблиц сопряженности для измерения силы связи предложены десятки формул [Миркин, Розенберг, 1979; Миркин и др., 1989], которые можно свести к трем основным группам:

- традиционные коэффициенты связи, основанные на χ^2;
- меры и статистики, основанные на рангах;
- коэффициенты, измеряющие информационную связь между факторами.

Коэффициенты связи, основанные на χ^2, исходят из предпосылки о том, что, чем больше объем выборки m, тем легче получить статистически значимую величину критерия даже при очень слабой взаимосвязи переменных (т.е. при больших объемах выборки даже слабые связи будут статистически значимыми).

Чтобы элиминировать влияние объема выборки m, К. Пирсон предложил в качестве меры связи *среднеквадратическую сопряженность* (он же – *редуцированный коэффициент корреляции*)

$$\phi^2 = \frac{\chi^2}{m}, \tag{6.6}$$

который изменяется в диапазоне от 0 до $min(r-1, s-1)$.

Стремясь нормировать меру связи к единому диапазону, С. Крамер видоизменил формулу (6.6) для своего *коэффициента Крамера*:

$$V^2 = \frac{\chi^2}{m \cdot \min(r-1, s-1)}, \tag{6.7}$$

верхний предел которого единица.

А.А. Чупров нашел для похожей формулы более звучное название – *полихорический коэффициент сопряженности (коэффициент Чупрова)*:

$$T^2 = \frac{\chi^2}{m \cdot \sqrt{(r-1)(s-1)}}. \tag{6.8}$$

Нетрудно заметить, что T^2 и V^2 эквивалентны, когда число столбцов равно числу строк, в иных случаях V^2 всегда больше, чем T^2. Для таблицы 2x2 обе меры равны ϕ^2.

Наконец, можно упомянуть еще один коэффициент, связанный с именем К. Пирсона – *коэффициент контингенции*:

$$C^2 = \frac{\chi^2}{\chi^2 + m}. \tag{6.9}$$

Перечисленные коэффициенты, основанные на χ^2, остаются неизменными при перестановке местами строк или столбцов таблицы и всегда выражаются положительными числами, поэтому уяснение направления зависимости должно производиться только по виду таблицы сопряженности.

Коэффициенты, основанные на рангах, позволяют извлечь информацию о направлении связи между признаками, используя понятие коррелируемости на основе подсчета числа пар объектов с взаимно возрастающими, взаимно убывающими и равными значениями признаков.

Коэффициент τ Кендалла учитывает число пар с равными признаками и может достигать значений -1 и +1, отражающих высшую степень положительной или отрицательной корреляции между признаками. Обычно вычисляется два варианта статистики Кендалла: τ_b и τ_c, которые различаются только способом обработки совпадающих рангов.

Если в данных имеется много совпадающих значений, предпочтительнее *γ-статистика Гудмана-Кендалла,* которая представляет собой нормированную разность между вероятностью P того, что ранговый порядок двух переменных совпадает, и вероятностью Q того, что он не совпадает:

$$\gamma = (P - Q)/(P + Q).$$

Таким образом, γ-статистика в основном эквивалентна τ Кендалла, за исключением того, что совпадения явно учитываются в нормировке.

Коэффициент d Соммера аналогичен коэффициенту γ с дифференциальным учетом пар с равными значениями признаков. Вычисляются два значения коэффициента, учитывающих равенство первого $d(A|B)$, и второго $d(B|A)$ признака.

Информационный подход к оценке связи. Традиционные меры связи, основанные на χ^2, представляют собой сугубо эвристические конструкции, интерпретация и математико-статистическое обоснование которых оставляет желать много лучшего. Поэтому понятен интерес к оценкам, основанным на теоретико-информационном подходе к анализу таблиц частот.

Современная трактовка статистической связи между переменными A и B сводится к оценке количества информации

$$I(A, B) = H(B) - H(B|A), \tag{6.10}$$

которое устраняет неопределенность того, какое значение примет B, если известно значение A. Таким образом, мера сопряженности оценивается как доля сопряженного разнообразия величины

$\boldsymbol{B}$, включенной в систему $(\boldsymbol{A},\boldsymbol{B})$, по сравнению с разнообразием (энтропией) $\boldsymbol{B}$, рассматриваемой отдельно:

$$\boldsymbol{R}_{\boldsymbol{B}|\boldsymbol{A}} = \boldsymbol{I}(\boldsymbol{A},\boldsymbol{B})/\boldsymbol{H}(\boldsymbol{B}) . \qquad (6.11)$$

Практическая трудность построения мер, относящихся к этому семейству, заключается в трудности выбора конкретных дефиниций и формальных выражений понятий "неопределенность" и "информация" из множества возможных. Например, согласно принципа "пропорциональной предикции", высказанного Л. Гудменом и В. Крускалом (L. Goodman, V. Kruscal), мерой связи должно служить относительное уменьшение вероятности ошибки предикции (предсказания) значения зависимого признака по значению независимого. Если в рамках этой концепции производить оценку относительной ошибки, используя отношения правдоподобия и минимизацию числа ошибочных предсказаний, мы получаем меру связи, известную как λ-*коэффициент Гудмана*:

$$\lambda_{\boldsymbol{B}|\boldsymbol{A}} = \frac{\sum_{\boldsymbol{i}} \boldsymbol{p}_{\boldsymbol{i}} \cdot \max_{\boldsymbol{j}} \boldsymbol{p}_{\boldsymbol{j}|\boldsymbol{i}} - \max_{\boldsymbol{j}} \boldsymbol{p}_{\bullet \boldsymbol{j}}}{1 - \max_{\boldsymbol{j}} \boldsymbol{p}_{\bullet \boldsymbol{j}}} , \qquad (6.12)$$

где $\max_{\boldsymbol{j}} \boldsymbol{p}_{\bullet \boldsymbol{j}}$ – вероятность, соответствующая модальной категории маргинального распределения $\boldsymbol{B}$; $\max_{\boldsymbol{j}} \boldsymbol{p}_{\boldsymbol{j}|\boldsymbol{i}}$ – вероятность модального значения $\boldsymbol{B}$ при $\boldsymbol{i}$-м значении $\boldsymbol{A}$. Значения λ лежат в пределах от 0 до 1: 0 означает невозможность предсказания зависимого признака, а 1 означает, что по значению независимого признака можно уверено предсказать значение зависимого признака. Необходимо отметить, что нулевое значение λ эквивалентно не состоянию статистической независимости, а тождеству модальных категорий переменных $\boldsymbol{A}$ и $\boldsymbol{B}$.

Кроме несимметричных мер λ, требующих предварительного уточнения, какая из двух переменных является объясняемой, чешские авторы (J. Rehak, B. Rehakova) предложили *симметричный коэффициент Гудмана*:

$$\lambda_{\boldsymbol{A},\boldsymbol{B}} = \frac{\left[\sum_{\boldsymbol{i}} \boldsymbol{p}_{\boldsymbol{i}} \cdot \max_{\boldsymbol{j}} \boldsymbol{p}_{\boldsymbol{j}|\boldsymbol{i}} - \max_{\boldsymbol{j}} \boldsymbol{p}_{\bullet \boldsymbol{j}} + \sum_{\boldsymbol{j}} \boldsymbol{p}_{\boldsymbol{j}} \cdot \max_{\boldsymbol{i}} \boldsymbol{p}_{\boldsymbol{i}|\boldsymbol{j}} - \max_{\boldsymbol{i}} \boldsymbol{p}_{\boldsymbol{i}\bullet} \right] / 2}{1 - (\max_{\boldsymbol{i}} \boldsymbol{p}_{\boldsymbol{i}\bullet} + \max_{\boldsymbol{j}} \boldsymbol{p}_{\bullet \boldsymbol{j}}) / 2} . \qquad (6.13)$$

Кроме λ-статистики Гудмана, можно упомянуть еще ряд критериев, основанных на теоретико-информационном подходе к построению мер связи: коэффициент Валлиса, информационные коэффициенты корреляции Линфута, Райского и т.д. [Елисеева, Рукавишников, 1977].

Результаты расчетов

В разделах 2 и 5 главы 4 были описаны различные биотические индексы, предложенные Ф.Вудивиссом, Э.А. Пареле и Е.В. Балушкиной. Фиксированный диапазон значений этих индексов соответствует классам качества вод шестибалльной шкалы (см. табл. 4.7 главы 4). Сформируем выборку значений этих трех индексов, рассчитанных по результатам гидробиологического мониторина для различных створов рек Самарской области. Для тех же точек наблюдений оценим класс качества воды по химическим показателям на основе гидрохимического индекса загрязения воды ИЗВ, либо по методике Былинкиной и Драчева, если имеющихся гидрохимических данных было недостаточно (см. разделы 3.5-3.6).

Сравним, используя таблицы сопряженности (см. табл. 6.1), насколько соответствуют оценки по гидробиологическим индексам реальным классам качества воды по гидрохимическим показателям. Предварительный анализ легко выполнить визуально: при полном совпадении оценок классов все ненулевые значения должны сконцентрироваться на главной диагонали. Этого, разумеется, не произошло и совпадение прогноза класса качества по индексу Вудивисса составило 38% случаев, индексу Пареле – 21%, индексу Балушкиной – 33%.

Таблица 6.1

Таблицы сопряженности, основанные на классах качества воды по гидрохимическим показателям (столбцы) и биотическим индексам (строки)

Градации строк		Градации столбцов – классы качества по гидрохимическим показателям					Итог по строкам
		2	**3**	**4**	**5**	**6**	
Классы качества вод, оцененные по биотическому индексу Вудивисса	**2**	25	22	8	2	0	57
		4.32%	3.80%	1.38%	0.35%	0.00%	9.84%
	3	21	52	46	9	3	131
		3.63%	8.98%	7.94%	1.55%	0.52%	22.63%
	4	2	27	39	4	3	75
		0.35%	4.66%	6.74%	0.69%	0.52%	12.95%
	5	4	24	88	66	44	226
		0.69%	4.15%	15.20%	11.40%	7.60%	39.03%
	6	4	6	18	23	39	90
		0.69%	1.04%	3.11%	3.97%	6.74%	15.54%
Итоги по столбцам		56	131	199	104	89	579
		9.67%	22.63%	34.37%	17.96%	15.37%	100.00%
Классы качества вод, оцененные по индексу Пареле	**1**	23	38	39	10	6	116
		6.10%	10.08%	10.34%	2.65%	1.59%	30.77%
	2	7	14	26	12	9	68
		1.86%	3.71%	6.90%	3.18%	2.39%	18.04%
	3	6	11	18	9	4	48
		1.59%	2.92%	4.77%	2.39%	1.06%	12.73%
	4	3	8	29	10	4	54
		0.80%	2.12%	7.69%	2.65%	1.06%	14.32%
	5	1	5	17	24	6	53
		0.27%	1.33%	4.51%	6.37%	1.59%	14.06%
	6	1	8	9	12	8	38
		0.27%	2.12%	2.39%	3.18%	2.12%	10.08%
Итоги по столбцам		41	84	138	77	37	377
		10.88%	22.28%	36.60%	20.42%	9.81%	100.00%
Классы качества вод, оцененные по хирономидному индексу Балушкиной	**2**	31	32	31	5	11	110
		6.30%	6.50%	6.30%	1.02%	2.24%	22.36%
	3	17	54	76	19	8	174
		3.46%	10.98%	15.45%	3.86%	1.63%	35.37%
	4	5	35	66	41	30	177
		1.02%	7.11%	13.41%	8.33%	6.10%	35.98%
	5	1	8	10	10	2	31
		0.20%	1.63%	2.03%	2.03%	0.41%	6.30%
Итоги по столбцам		54	129	183	75	51	492
		10.98%	26.22%	37.20%	15.24%	10.37%	100.00%

Более объективный анализ связи между оценками качества воды в гидробиологических и гидрохимической шкалах можно сделать с использованием описанных выше статистических критериев. Как свидетельствуют расчеты, представленные в табл. 6.2, в соответствии с критериями χ^2 и τ_b Кендалла нет никаких оснований принимать гипотезу об отсутствии связи между классами качества, оцененными по биотическим индексам и по гидрохимическим показателям.

Сравнительный анализ значений коэффициентов связи позволяет сделать вывод о том, что на имеющемся экспериментальном материале ощутимое превосходство в прогностической силе имеет биотический индекс Вудивисса, тогда как индексы Пареле и Балушкиной дают значительно более скромные и приблизительно одинаковые по точности результаты.

Таблица 6.2

Анализ силы связи в таблице сопряженности (см. табл. 6.1) с использованием различных статистических критериев и коэффициентов

Наименование критерия или коэффициента	Класс по индексу Вудивисса	Класс по индексу Пареле	Класс по индексу Балушкиной
Критерий χ^2	288.45 (p=0.0)	69.68 (p=0.0)	88.6 (p=0.0)
V-статистика Крамера	0.3529	0.215	0.245
Коэффициент контингенции C	0.5766	0.395	0.3906
τ_b - статистика Кендалла	0.5114 (p=0.0)	0.2781 (p=0.0)	0.2778 (p=0.0)
τ_c - статистика Кендалла	0.483	0.2708	0.2662
γ - статистика	0.6567	0.3533	0.3803
Коэффициент d Соммера симметр.	0.5114	0.2779	0.2776
то же, по строкам	0.5048	0.2877	0.2673
то же, по столбцам	0.5181	0.2688	0.2887
λ Гудмана-Крускала симметр.	0.1269	0.054	0.0897
то же, по строкам	0.1388	0.0651	0.1746
то же, по столбцам	0.1158	0.0418	0.0032

Следует еще раз подчеркнуть, что сама по себе величина коэффициентов связи мало о чем говорит и имеет содержательный смысл только при сравнении между собой выборок, имеющих примерно одинаковую размерность и условия формирования. Например, коэффициент λ Гудмана, также как и коэффициент детерминации R^2, обычно имеет небольшие значения. Оценки этого коэффициента для наших таблиц не превышали 0.15, т.е. объяснено всего около 10% качественной дисперсии. В то же время, значимость критерия "хи-квадрат" свидетельствует о высоком уровне связи. Поэтому, на наш взгляд, не следует недооценивать влияние фактора, ориентируясь на небольшие величины коэффициентов детерминации как для количественных, так и для неколичественных переменных, а полагаться на содержательные результаты, подтвержденные значимостью связей при статистической проверке результатов.

Другой формой визуального анализа таблиц сопряженности является их графическое представление в виде различного рода диаграмм. На рис. 6.1 представлен вариант столбчатой диаграммы совместного распределения значений индекса Вудивисса и классов качества воды, оцененных по гидрохимическим показателям.

Несмотря на выявленный высокий уровень связи между этими признаками, можно отметить значительное снижение адекватности прогноза класса качества воды в области малых значений показателя V: величина индекса Вудивисса менее 3 далеко не всегда свидетельствует о реальном химическом загрязнении воды, а может определяться посторонними факторами (условиями отбора проб, сезонностью и проч.).

> **Мем № 33**: *«Об экологическом благополучии водного объекта можно судить по составу доминирующего комплекса донных организмов, соотношению численности личинок хирономид, относящихся к роду* Chironomus, *подсемейству* Orthocladiinae *и трибе* Tanytarsini *...и другим показателям донных сообществ»*
>
> В.А. Яковлев [1988].

Проанализируем это часто встречающееся в литературе утверждение, сформировав частотные таблицы сопряженности.

Поскольку алгоритмы анализа сопряженности связаны с признаками, измеренными в порядковых шкалах, предварительно выполним следующие преобразования:

- по каждому измерению и анализируемой таксономической группе рассчитаем логарифм индекса плотности населения $\ln((N_s * B_s)^{0.5})$, учитывающему в одном показателе как численность N_s, так и биомассу B_s и имеющему распределение, близкое к нормальному;

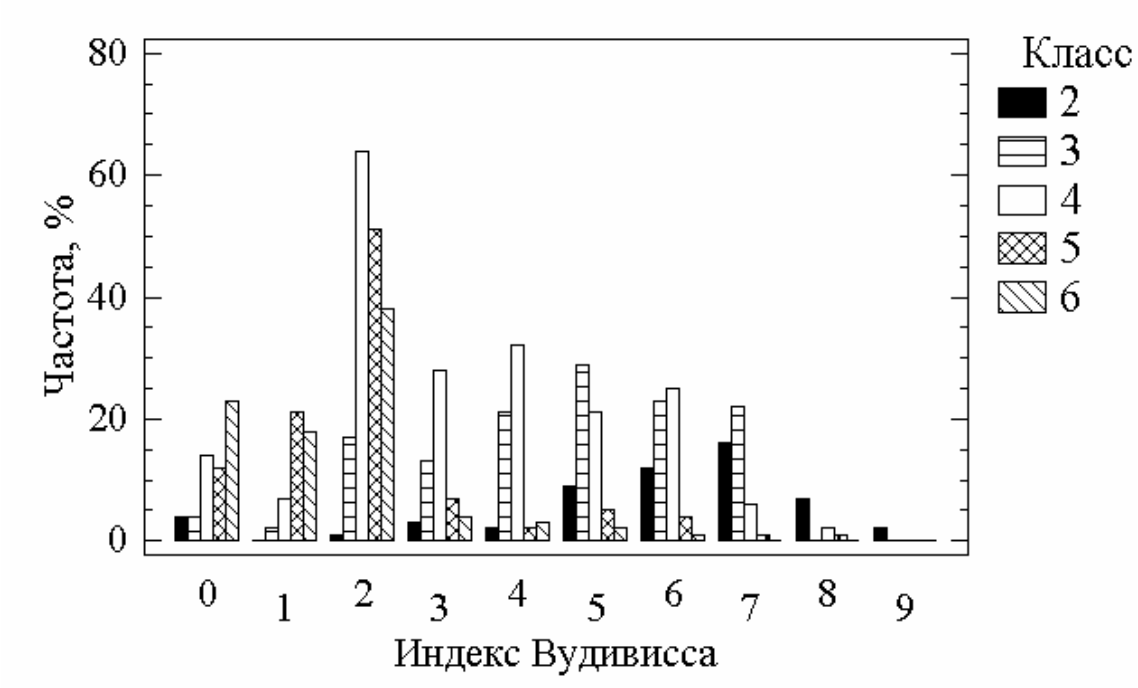


Рис. 6.1. Частотное распределение значений биотического индекса Вудивисса на станциях с разным классом качества воды

- пересчитаем полученные значения $x = \ln((N_s * B_s)^{0.5})$ в диапазон шкалы чисел $\{6 \geq x' \geq 1\}$ по известным формулам масштабирования

$$x' = (x'_{\min} - x_{\min}) + \frac{x'_{\max} - x'_{\min}}{x_{\max} - x_{\text{мin}}} x ; \qquad (6.14)$$

- округлим значения x' до ближайшего целого, установив значения ранга обилия равным нулю, если данная таксономическая группа в наблюдении отсутствовала (общая выборка наблюдений включала все пробы, в которых был найден хотя бы один вид хирономид).

Используя полученные градации, сформируем таблицы сопряженности обилия хирономид подсемейств Tanypodinae (см. пример в табл. 6.3), Orthocladiinae и Chironominae с двумя факторами среды – классом качества воды по гидрохимическим показателям и ландшафтно-географической категорией станции наблюдения.

Результаты анализа силы и достоверности связи по некоторым основным критериям представлены в табл. 6.4, что дает нам основания сделать следующие выводы:

1. В разделе 5.6 по результатам рангового корреляционного анализа остался открытым вопрос о влиянии органического загрязнения (по $БПК_5$) на численность хирономид подсемейства Tanypodinae. Анализ таблиц сопряженности дал более категорическое заключение – обилие таниподин определяется, в основном, типологическими особенностями водоемов (являясь вместе с тем показателем процесса их эвтрофирования) и мало зависит от химического загрязнения,;
2. Экологический диапазон видов подсемейств Orthocladiinae и Chironominae (трибы Tanytarsini и Chironomini) существенно более узкий, однако их обилие примерно в равной степени определяется как классом качества вод по гидрохимическим показателям, так и факторами, напрямую не связанными с антропогенным воздействием.

Таблица 6.3

Таблица сопряженности, основанная на градациях обилия личинок подсемейства Tanypodinae *(столбцы) и ландшафтно-географическим категориям станций (строки)*

Ландшафтно-географические категории	Градации обилия Tanypodinae							Итого по строкам
	0	1	2	3	4	5	6	
1. Ручьи и родники	8	3	4	4	1	1	0	21
	1.69%	0.64%	0.85%	0.85%	0.21%	0.21%	0.00%	4.45%
2. Малые реки возвышенностей	82	10	23	14	17	10	3	159
	17.37%	2.12%	4.87%	2.97%	3.60%	2.12%	0.64%	33.69%
3. Малые равнинные реки	6	4	2	4	2	3	1	22
	1.27%	0.85%	0.42%	0.85%	0.42%	0.64%	0.21%	4.66%
4. Средние равнинные реки	31	8	13	19	21	7	1	100
	6.57%	1.69%	2.75%	4.03%	4.45%	1.48%	0.21%	21.19%
5. Устья рек	22	1	11	12	7	2	0	55
	4.66%	0.21%	2.33%	2.54%	1.48%	0.42%	0.00%	11.65%
6. Пруды, озера, водохранилища	57	3	17	20	15	3	0	115
	12.08%	0.64%	3.60%	4.24%	3.18%	0.64%	0.00%	24.36%
Итого по столбцам	206	29	70	73	63	26	5	472
	43.64%	6.14%	14.83%	15.47%	13.35%	5.51%	1.06%	100 %

Таблица 6.4

Анализ силы связи в таблицах сопряженности между градациями обилия отдельных групп видов хирономид и факторами среды: классом качества воды по гидрохимическим показателям и ландшафтно-географические категориям станций

Факторы среды	Подсемейства/трибы хирономид	Статистика χ^2 «хи-квадрат»		V статистика Крамера	τ_b - статистика Кендалла		λ Гудмана симметр.
		критерий	p		критерий	p	
Класс качества воды (гидрохимический)	Tanypodinae	23.26	0.504	0.1110	-0.0222	0.558	0.0018
	Orthocladiinae	97.94	0	0.2270	-0.2919	0	0.0413
	Chironomini	67.2	0.0003	0.189	0.516	0	0.0431
	Tanytarsini	86.10	0	0.2135	-0.2780	0	0.0277
Ландшафтно-географические категории	Tanypodinae	47.54	0.022	0.1419	0.0311	0.409	0.0173
	Orthocladiinae	91.29	0	0.1967	-0.2216	0	0.0642
	Chironomini	143.3	0	0.2481	0.248	0	0.126
	Tanytarsini	93.91	0	0.1995	-0.1959	0	0.0625

6.2. Нелинейность отношений и «уиттекеровские» колокола

Формулировка задачи

Пусть имеется ряд из m сопряженных наблюдений двух переменных $X \equiv (x_1, ..., x_m)$ и $Y \equiv (y_1, ..., y_m)$, причем предполагается, что хотя бы одна из этих переменных (или обе вместе) измерены в количественных шкалах: интервальной, абсолютной или шкале отношений.

В разделе 5.5 было приведено выражение для расчета коэффициента корреляции Пирсона r_{XY}, который является мерой линейной связи между векторами X и Y. Однако в случае нелинейной зависимости между фактором и откликом значение r_{XY} теряет свою достоверность и может дать ошибочное представление о тесноте связи. Классический пример – параболическая зависи-

мость, для которой коэффициент линейной корреляции может быть близок к 0, как бы не были близки экспериментальные точки к расчетной кривой.

Необходимо в условиях предполагаемой нелинейности оценить наличие связи между X и Y, а также степень ее близости к линейной форме. Для этого определяются показатели, характеризующие концентрацию распределения (и, следовательно, тесноту связи) около кривых регрессии $\mathbf{Y}(x)$ и $\mathbf{X}(y)$.

Рекомендуемая литература: [Ван дер Варден, 1960; Смирнов, Дунин-Барковский, 1965; Плохинский, 1970; Батоян, 1983; Генкин, 1999; С.А. Прохоров, 2001б, 2002].

Математический лист

Определим метод расчета корреляционных отношений $\eta^2_{y|x}$ и $\eta^2_{x|y}$, введенных К. Пирсоном и являющихся наиболее общими мерами оценки нелинейных связей.

Для вычисления $\eta^2_{y|x}$ разобьем весь диапазон изменения фактора X на k поддиапазонов (интервалов). Пусть теперь $y_{1i}, y_{2i}, \ldots, y_{ni}$ – ординаты всех тех точек, абсциссы которых принадлежат i-му интервалу ($i = 1,2,\ldots,k$). Если m_i – количество точек, попавших в i-й интервал, то среднее значение ординат точек, попавших в этот интервал, определяется как

$$\bar{y}_i = \left(\sum_{j=1}^{m_i} y_{ij} \right) / m_i , \tag{6.15}$$

а общая средняя всех ординат – как

$$\bar{\bar{y}} = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{m_i} y_{ij} \quad . \tag{6.16}$$

Для нахождения меры нелинейной связи вычисляется также дисперсия всех ординат от общей средней (обычная дисперсия ординат)

$$s_y^2 = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{m_i} (y_{ij} - \bar{\bar{y}})^2 \tag{6.17}$$

и разброс интервальных средних $\bar{y}_i$ ($i = 1, 2, \ldots, k$) от общей средней:

$$s_{\bar{y}}^2 = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}})^2 . \tag{6.18}$$

Последняя дисперсия тем меньше, чем хуже зависимость y от x, а в случае полного хаоса $\bar{y}_i = \bar{\bar{y}}$ и $s_{\bar{y}}^2 = 0$. Квадрат *корреляционного отношения* $\eta^2_{y|x}$ определяется как отношение дисперсии интервальных средних к дисперсии всех ординат:

$$\eta^2_{y|x} = s_{\bar{y}}^2 / s_y^2 \quad . \tag{6.19}$$

Аналогично вычисляется квадрат корреляционного отношения x от y; при этом квантуется диапазон изменения признака y на оси ординат:

$$\eta^2_{x|y} = s_{\bar{x}}^2 / s_x^2 . \tag{6.20}$$

Между $\eta^2_{y|x}$ и $\eta^2_{x|y}$ нет какой либо простой зависимости: Y может быть не скоррелирована с X и $\eta^2_{y|x} = 0$, когда как другой показатель $\eta^2_{x|y} = 1$ (пример – та же парабола, или колоколовидная кривая по Р. Уиттекеру [1980]). Если $\eta^2_{y|x} = \eta^2_{x|y} = 1$, то функциональная зависимость $Y = f(X)$ обратима и Y представляет монотонную функцию от X.

Корреляционное отношение, как и коэффициент детерминации R^2, всегда положительно и изменяется от 0 до 1. Заметим еще, что во всех случаях $R^2 < \eta^2_{y|x}$ и $R^2 < \eta^2_{x|y}$, так что из равенства нулю любого их корреляционных отношений, коэффициент линейной корреляции r_{xy} также равен 0.

Вернемся к уже обсуждавшейся проблеме: как оценить наличие связи и степень ее близости к линейной форме.

Ошибка коэффициента линейной корреляции определяется по формуле:

$$\sigma_r = \frac{1 - r_{xy}^2}{\sqrt{m}} \ , \tag{6.21}$$

где m – число сравниваемых реализаций пар признаков X и Y. Тогда достоверность отличия коэффициента корреляции от нуля определяется по критерию Стьюдента:

$$t = \left| \frac{r_{xy}}{\sigma_r} \right| > t_\nu \ , \tag{6.22}$$

где число степеней свободы $\nu = N - 2$.

Ошибка квадрата корреляционного отношения задается следующей формулой:

$$\sigma_{\eta^2} = \frac{(1 - \eta^2)(k - 1)}{(m - k)} \ , \tag{6.23}$$

где k – число классов корреляционной решетки по соответствующему признаку.

Тогда критерий достоверности отличия корреляционных отношений $\eta^2_{y|x}$ и $\eta^2_{x|y}$ от нуля будет определяться с использованием критерия Фишера:

$$F = \frac{\eta^2}{\sigma_{\eta^2}} > F_{\nu_1,\nu_2} \ , \tag{6.24}$$

где $\nu_1 = k - 1$, $\nu_2 = m - k$.

Критерий криволинейности связи Φ позволяет установить границу достоверного различия линейного коэффициента корреляции и корреляционного отношения и выносить решение о существовании сугубо нелинейной связи:

$$\Phi = \frac{(\eta^2 - R^2)(m - k)}{(1 - \eta^2)(k - 2)} > \Phi_{\nu_1,\nu_2} \ . \tag{6.25}$$

Распределение этого критерия также в случае нулевой гипотезы асимптотически приближается к распределению Фишера $F(\nu_1 = k - 1, \ \nu_2 = m - k)$, что позволяет статистически проверить предположение о линейности.

Таким образом, схема *полного корреляционного анализа* выглядит так: первоначально определяется коэффициент линейной корреляции r_{xy} и его достоверность σ_r; далее – корреляционное отношение $\eta^2_{y|x}$ и его достоверность σ_η. Здесь возможны следующие варианты [Плохинский, 1970; Миркин, Розенберг, 1978]:

- если r_{xy} достоверен, а $\eta^2_{y|x}$ недостоверно, то необходим дополнительный анализ распределений;
- если r_{xy} недостоверен, а $\eta^2_{y|x}$ достоверно, то связь нелинейная;
- если r_{xy} недостоверен и $\eta^2_{y|x}$ недостоверно, то связь отсутствует;
- если r_{xy} достоверен и $\eta^2_{y|x}$ достоверно, то определяется критерий криволинейности Φ;
- если Φ достоверен, то связь нелинейная;
- если Φ недостоверен, то выбирается линейная зависимость как более простая.

Другой способ оценки нелинейности взаимосвязи $Y \approx f(X)$ заключается в проведении двух пересекающихся вспомогательных линий регрессии через точки, соответствующие интервальным средним $\overline{y_i}$ ($i = 1, 2, \ldots, k$) и $\overline{x_l}$ ($l = 1, 2, \ldots, n$). Угол θ между этими прямыми определяется по формуле:

$$\mathrm{tg}\theta = \frac{k_2 - k_1}{1 + k_1 k_2} \ , \tag{6.26}$$

где $k_1 = \eta^2_{y|x}$ и $k_2 = 1/\eta^2_{x|y}$ – угловые коэффициенты обеих прямых. Величина этого угла близка к 0 при наличии линейной корреляции.

В предыдущем разделе отмечалось, что с точки зрения теоретико-информационного подхода измерение статистической связи между переменными есть измерение сопряженного разнообразия:

$$(\boldsymbol{U}(\boldsymbol{X}) - \boldsymbol{U}(\boldsymbol{X}|\boldsymbol{Y}))/\boldsymbol{U}(\boldsymbol{Y}) , \tag{6.27}$$

где $\boldsymbol{U}(\boldsymbol{X})$ – неопределенность (или "энтропия") переменной $\boldsymbol{X}$, рассматриваемой отдельно, т.е. без знания $\boldsymbol{Y}$; $\boldsymbol{U}(\boldsymbol{X}|\boldsymbol{Y})$ – неопределенность $\boldsymbol{Y}$ при знании $\boldsymbol{X}$. Спецификация понятия «энтропия» в случае признаков, измеренных в интервальных и номинальных шкалах, наиболее явным образом связана с традиционными статистическими понятиями «дисперсии» и «вариации». Учитывая, что при наблюдениях на уровне интервальной шкалы мы получаем информацию не только о целевой категории объекта по $\boldsymbol{Y}$, но и количественно оцениваем его отличие от других объектов, естественно использовать следующие дефиниции неопределенностей:

$$\boldsymbol{U}(\boldsymbol{X}) = \boldsymbol{s}_{\boldsymbol{x}}^2 = \frac{1}{\boldsymbol{m}} \sum_{\boldsymbol{k}=1}^{\boldsymbol{L}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m_i}} (\boldsymbol{x_{kj}} - \bar{\bar{\boldsymbol{x}}})^2 \tag{6.28}$$

$$\text{и} \quad \boldsymbol{U}(\boldsymbol{X} \mid \boldsymbol{Y}) = \boldsymbol{s}_{\bar{\boldsymbol{x}}}^2 = \frac{1}{\boldsymbol{L}} \sum_{\boldsymbol{k}=1}^{\boldsymbol{L}} (\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \bar{\bar{\boldsymbol{x}}})^2 , \tag{6.29}$$

где $\bar{\boldsymbol{x}}_{\boldsymbol{k}} = \left(\sum_{\boldsymbol{j}=1}^{\boldsymbol{m_k}} \boldsymbol{x_{kj}} \right) / \boldsymbol{m_k}$ – среднее значение варьируемой переменной для примеров, попавших в $\boldsymbol{k}$-й класс, а $\bar{\bar{\boldsymbol{x}}} = \frac{1}{\boldsymbol{m}} \sum_{\boldsymbol{k}=1}^{\boldsymbol{L}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m_k}} \boldsymbol{x_{kj}}$ – общая средняя всех измерений. Тогда нормированная мера связи $(\mathrm{U}(\mathbf{X}) - \mathrm{U}(\mathbf{X}|\mathbf{Y}))/\mathrm{U}(\mathbf{Y})$ примет вид знакомого по (6.19) квадрата корреляционного отношения:

$$\eta_{x|y}^2 = \frac{\sum_{\boldsymbol{k}=1}^{\boldsymbol{L}} \boldsymbol{m_k} (\bar{\boldsymbol{x}}_{\boldsymbol{k}} - \bar{\bar{\boldsymbol{x}}})^2}{\sum_{\boldsymbol{k}=1}^{\boldsymbol{L}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m_k}} (\boldsymbol{x_{kj}} - \bar{\bar{\boldsymbol{x}}})^2} , \tag{6.30}$$

которое, таким образом, является частной реализацией информационного подхода к построению мер связи:

Как и корреляционные отношения, с идеей анализа тенденций интервальных средних связан *прямой градиентный анализ* – один из широко используемых методов ординации растительности, уходящий корнями в труды Л.Г. Раменского начала XX века и получивший развитие в работах Р. Уиттекера и Дж. Кертиса [Whittaker, 1952; Curtis, 1959; цит. по: Миркин, Розенберг, 1978]. Количественный прямой градиентный ***R***-анализ [McIntosh, 1973; Kershaw, 1974; Миркин, Наумова, 1983, 1998; Розенберг, 1984] складывается из следующих этапов:

- интересующий исследователя фактор $\boldsymbol{X}$ разбивается на классы (интервалы - $\boldsymbol{x_i}$, $\boldsymbol{i}$ = 1,2,…,$\boldsymbol{k}$);
- для анализируемого биологического показателя $\boldsymbol{Y}$ (встречаемость, обилие численности или биомассы для некоторого вида или ассоциации видов) традиционным образом рассчитываются интервальные средние и оценки дисперсий (6.15) – (6.18);
- проводится статистическая проверка гипотезы о влияния фактора с использованием классического однофакторного дисперсионного анализа (например, по ***F***-критерию Фишера);
- осуществляется построение графика эмпирического распределения интервальных средних $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ в зависимости от градаций фактора среды, ориентированных вдоль оси $\boldsymbol{X}$;
- визуально или с помощью эвристических приемов оценивается гипотеза об одновершинном характере этого распределения (модель "колоколовидного распределения" Либиха – Шелфорда – Уиттекера [модель предложена Р. Уиттекером, но отражает *принцип лимитирующих факторов Либиха – Шелфорда*]);
- в случае полимодальности кривой интервальных средних запускается процедура выравнивания распределения методом скользящей средней [Розенберг, 1984] до достоверно одновершинного распределения (эта процедура может осуществляться несколько раз);

- проводится определение средневзвешенной напряженности фактора и ее дисперсии:

$$\chi = \sum x_i * p_i, \quad \sigma^2 = \sum (x_i - \chi)^2 * p_i, \tag{6.31}$$

где x_i – значение фактора X для середины i-й градации; p_i – доля площади под выровненной кривой распределения (криволинейной трапеции), приходящаяся на i-й интервал.

Положение средневзвешенной напряженности для данного вида на оси фактора X будет свидетельствовать о "принадлежности" вида к минимальным или максимальным значениям исследуемого фактора, а величина дисперсии – о степени эвритопности (большая дисперсия) или стенотопности вида (маленькая дисперсия).

Результаты расчетов

Рассмотрим выборку сопряженных значений концентрации нитрат-ионов в воде C_{NO3} и индекса Шеннона H, состоящую из 68 измерений, и выполним полный корреляционный анализ.

Для расчета корреляционных отношений предварительно выполняется трудно формализуемая процедура выбора количества и граничных значений диапазонов, поскольку по эмпирическим соображениям в каждой ячейке совместной корреляционной решетки для двух показателей должно быть не менее 6-8 измерений.

Осуществим разбиение области варьирования переменных на интервалы, основываясь на классической стратегии их равной ширины в натуральной шкале. Если с индексом Шеннона все складывается относительно благополучно, то в случае с концентрацией нитратов, мы сталкиваемся с некоторым разочарованием: большинство гидрохимических показателей, как и подробно рассмотренные в разделе 5.1 гидробиологические показатели, характеризуются сильной асимметрией распределения. Поскольку в каждую клетку корреляционной решетки должно попасть хотя бы 1 значение (напомним, лучше – не менее 6), проводим логарифмирование значений C_{NO3} с добавлением "страховочной" единицы, что в значительной мере стабилизирует распределение – см. фиг. «а» и «б» на рис. 6.2. Но даже и после этого, достичь полного "заселения" решетки удается лишь "склеив" несколько крайних правых интервалов.

Расчет корреляционных отношений представлен в табл. 6.5.

Таблица 6.5

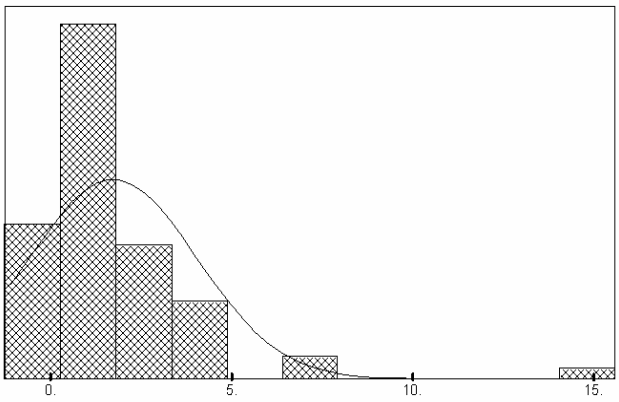
Расчет корреляционных отношений по выборке «индекс Шеннона (H) – концентрация нитратов (C_{NO3}), мг/л»

Интервалы $C = \ln(C_{NO3} + 1)$, общее среднее = 0.791				Интервалы H, общее среднее = 2.098			
Диапазоны	Измерений	Групповые средние		Диапазоны	Измерений	Групповые средние	
		C	H			H	C
До 0.5	23	0.177	2.363	До 1	8	0.584	0.965
0.5 - 1	23	0.718	2.121	1 - 2	20	1.576	0.767
1 -1.5	12	1.246	1.688	2 - 3	31	2.392	0.847
1.5 - 2.8	10	1.829	1.930	3 - 4.19	9	3.594	0.504
Корреляционное отношение $\eta^2_{y\|x}$			8.53%	Корреляционное отношение $\eta^2_{x\|y}$			7.89%
Статистика Фишера для $\eta^2_{y\|x} = 0$			1.99	Статистика Фишера для $\eta^2_{x\|y} = 0$			1.83
Критерий криволинейности связи Φ			1.58	Критерий криволинейности связи Φ			1.35

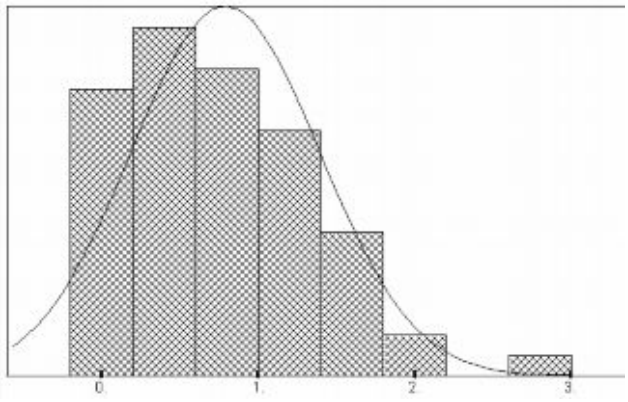
Классический линейный регрессионный анализ приводит к следующему уравнению (фиг. «в» на рис. 6.2):

$$H = 2.329 - 0.292 \ln(C_{NO3} + 1)$$

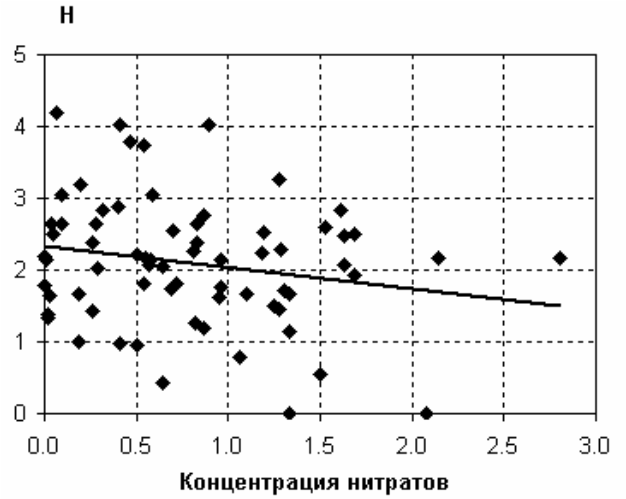
при коэффициенте детерминации $R^2 = 4.017\%$ и статистике Фишера $F(1,66) = 2.76$ [$p = 0.101$].



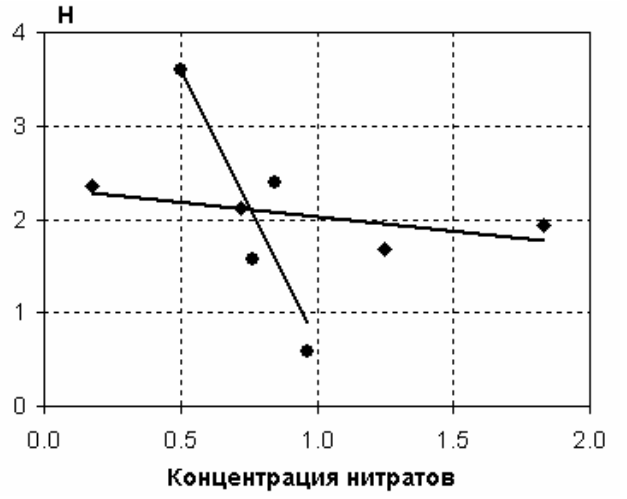
а) гистограмма натуральных значений концентраций нитратов, мг/л



б) гистограмма логарифмированных значений концентраций нитратов



в) корреляционное поле зависимости $\boldsymbol{H} = \boldsymbol{a} - \boldsymbol{b}\ln(\boldsymbol{C}_{NO3} + 1)$



г) прямые, построенные по интервальным средним $\boldsymbol{H}$ и $\ln(\boldsymbol{C}_{NO3} + 1)$

Рис. 6.2. Графики распределений и регрессионных зависимостей индекса Шеннона от концентрации нитратов

Вряд ли проведенный анализ корреляционных отношений добавил что-нибудь существенное к выводам регрессионного анализа о влиянии концентрации нитратов на индекс Шеннона. С одной стороны, на 5% уровне надежности, как уравнение регрессии, так и корреляционные отношения являются незначимыми, да и угол $\theta = 58^{\circ}$ между прямым $\mathbf{Y}(\boldsymbol{x})$ и $\mathbf{X}(\boldsymbol{y})$ на фиг. «г» трудно назвать небольшим. С другой стороны, при 10% пороге надежности, который не является слишком рискованным для биологических исследований, утверждения о линейной форме зависимости между этими переменными становятся непротиворечивыми, а, используя лексику градиентного анализа можно сказать, что *«средневзвешенная напряженность показателя на оси градиента имеет устойчивую обратно пропорциональную тенденцию для индекса Шеннона»*.

В качестве примера использования градиентного анализа рассмотрим влияние такого важнейшего показателя загрязнения воды, как степень насыщения кислородом в придонном слое ($\boldsymbol{C}_{O2}$) на индекс Шеннона $\boldsymbol{H}$, общую численность видов хищников-хватателей зообентоса ($\boldsymbol{N_h}$), общую биомассу видов Chironomidae ($\boldsymbol{B_{chi}}$) и долю хищников этого рода ($\boldsymbol{B_h}/\boldsymbol{B_{chi}}$).

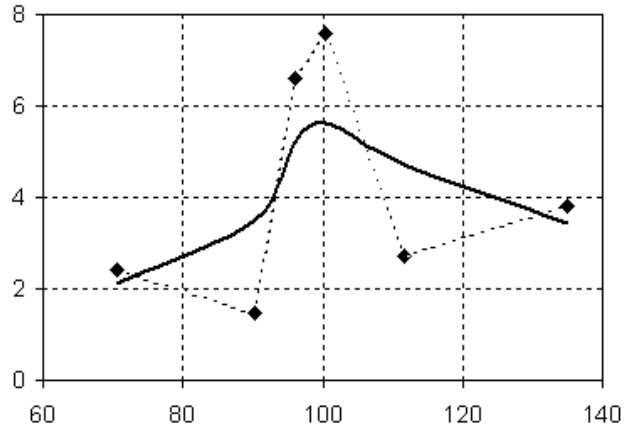
Отличительной особенностью распределения содержания кислорода является сильное сгущение точек измерений относительно среднего значения при наличии мощных "хвостов". Как было показано в разделе 5.5, это делает неустойчивыми большинство уравнений регрессионного анализа. Для выделения границ интервалов воспользуемся концепцией равной заселенности и представим результаты разбиения в табл. 6.6.

Графики распределения средних интервальных значений анализируемых показателей по оси градиента и оценка силы влияния фактора по статистике Фишера приведены на рис. 6.3.

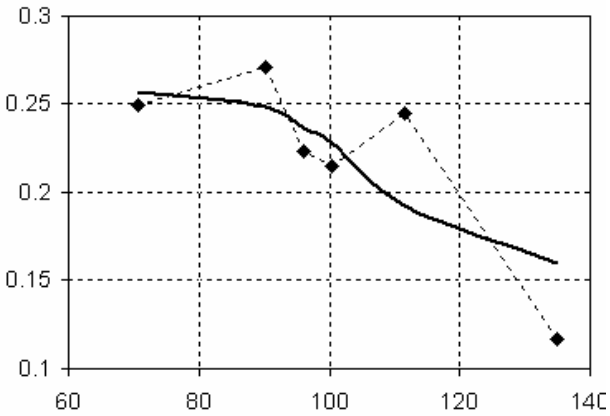
Таблица 6.6

Распределение по диапазонам значений содержания растворенного кислорода, % от степени насыщения

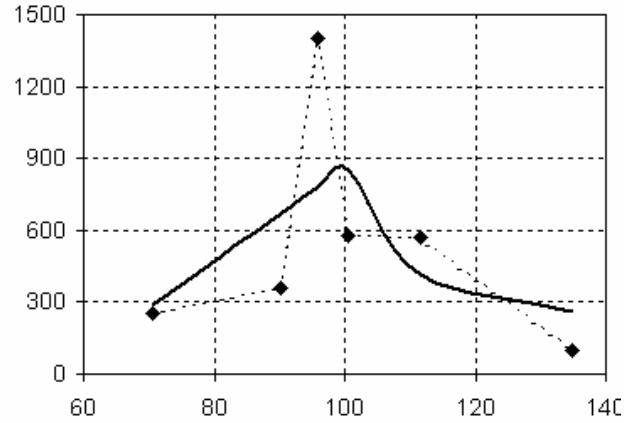
№№ интервалов	Количество точек	Минимум	Максимум	Среднее
1	52	6.5	84	70.68
2	49	85	93	90.2
3	53	94	98	95.99
4	50	99	103	100.42
5	50	105	116	111.6
6	52	118	184	134.90
Всего	306			100.65



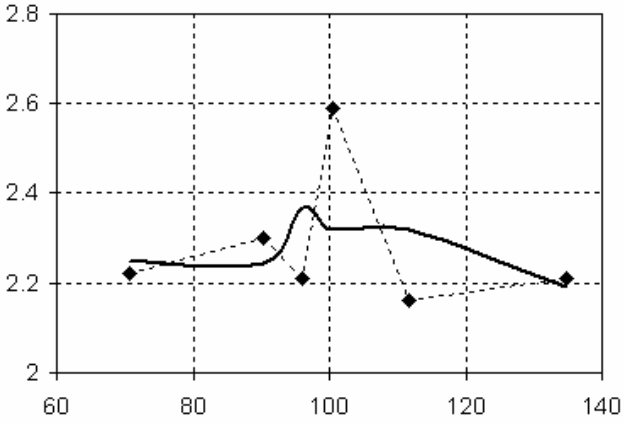
а) суммарная биомасса Chironomidae B_{chi}
(F = 2.14, p = 0.06)



б) доля видов хищников в биомассе Chironomidae
B_h/B_{ch} (F = 1.83, p = 0.106)



в) суммарная численность хищников N_h
(F = 2.54, p = 0.0286)



г) индекс Шеннона H
(F = 1.55, p = 0.175)

Рис. 6.3. Изменение некоторых гидробиологических показателей от содержания растворенного кислорода (по оси абцисс в %); пунктирная линия – до выравнивания, сплошная – после.

По результатам расчетов можно отметить четко выраженный "экологический оптимум" для абсолютных значений численности хищников N_h, биомассы хирономид B_{chi} и, в меньшей мере, индекса Шеннона H в области содержания растворенного кислорода C_{O2} = 100%. В то же время, доля хищных видов B_h/B_{ch} в общей массе хирономидофауны заметно снижается по мере обогащения воды кислородом.

Приемы градиентного анализа могут быть использованы также в том случае, когда отклик $\boldsymbol{Y}$ измерен в порядковой шкале, состоящей из альтернатив (1 – признак присутствует и 0 – в противном случае). Средние интервальные значения отклика заменяются при этом распределением частот встречаемости по оси градиента фактора.

Основные теоретические изыскания градиентного анализа, отличающие его от стандартных процедур дисперсионного анализа, направлены на борьбу с "многовершинностью" частных средних, которая противоречит представлениям о существовании только одного экологического оптимума распределения показателя по градиенту среды.

Одним из простейших способов, позволяющих элиминировать случайные колебания эмпирического тренда и получить плавную "колоколообразную" теоретическую кривую, является метод выравнивания скользящей средней. Предлагается [Розенберг, 1984] определять новые выравненные интервальные средние $\boldsymbol{X_i'}$ по следующим формулам:

$$\boldsymbol{X_i'} = (\boldsymbol{X_{i-1}} + \boldsymbol{X_i} + \boldsymbol{X_{i+1}})/3 \text{ для } \boldsymbol{k} > \boldsymbol{i} > 1 \quad \text{и}$$

$$\boldsymbol{X_1'} = (2\boldsymbol{X_1} + \boldsymbol{X_2})/3 \ , \ \boldsymbol{X_k'} = (\boldsymbol{X_{k-1}} + 2\boldsymbol{X_k})/3 \quad \text{для концевых точек,}$$

где $\boldsymbol{X_i}$ – частные средние исходного дисперсионного комплекса, $\boldsymbol{i} = 1,2,\ldots,\boldsymbol{k}$. Выравнивание способом скользящей средней рекомендуется повторять многократно, до тех пор, пока кривая частных средних по градациям фактора становится достоверно одновершинной, что проверяется сравнением средних по критерию Стьюдента.

Изложенная методика не является "истиной в последней инстанции" (как, впрочем, и любой эмпирико-статистический приём) и требует ряда уточнений и пояснений.

1. Поскольку исследователь, как правило, имеет дело с некоторой ограниченной областью варьирования переменных и выделить в чистом виде роль конкретного фактора среды в ходе наблюдений бывает крайне трудно, "колоколообразная" кривая взаимного влияния в гидроэкологии является скорее исключением, чем правилом (нам пришлось немало потрудиться, чтобы найти подходящий пример). Можно постулировать скорее прямую или обратную линейную зависимость, либо различные фрагменты ***S***-образной кривой, когда в интервалах слева (или справа) влияние фактора вообще отсутствует.
2. Любая, а, тем более, многократная "стрижка неровностей" может утопить в "болоте" усреднения все специфичные "трещинки", которые могут иметь глубокий содержательный смысл. Например, "вздернутые" края эмпирического распределения биомассы хирономид (рис. 6.3 «а») в 1 и 6 интервалах (при небольшом разнообразии по Шеннону) могут быть связаны с закономерным доминированием видов, развивающихся в условиях "нестандартных" биотопов. После сглаживания скользящими средними этот эффект теряется.
3. Вряд ли является плодотворной идея сравнивать два эмпирических распределения по отдельным парам точек с использованием критерия Стьюдента. Эта процедура традиционно поручается в статистике критериям Колмогорова-Смирнова и другим критериям согласия, описанным в разделе 5.1.
4. Определенные сомнения вызывает и правомочность оценки статистической значимости влияния фактора по Фишеру с использованием выравненных значений интервальных средних. Любые суммы квадратов отклонений от некоторых субъективных "средних", являются *смещенными* оценками, поэтому в рассчитанных статистиках исследователь получает искусственно завышенную силу влияния фактора.

Однако, несмотря на достаточно острую критику отдельных работ в области прямого градиентного анализа [Василевич, 1972], он является, безусловно, одним из наиболее эффективных методов ординации, о чем свидетельствуют представленные примеры.

6.3. Интервальные и бинарные структуры

Мем № 34*: «…нередко случается, что ловкой обработкой одного и того же материала можно выжать из него при помощи этого приема прямо противоположные заключения»* (*о методах группировки данных по интервалам. – Ремарка наша*) А.А. Чупров [1960].

Формулировка задачи

Пусть имеется ряд наблюдений показателя $Y = (y_1, ..., y_n)$, принимающий значения на отрезке $[a, b]$. Необходимо найти такое разбиение δ шкалы Y на k интервалов, при котором наилучшим образом используются дифференциально-диагностические возможности признака Y для поиска закономерности его связи с заданным фактором X.

При градиентном анализе или расчете корреляционных отношений возникают нетривиальные вопросы: каков механизм выделения интервалов и на сколько поддиапазонов следует разбивать область варьирования переменных. Нетрудно заметить, что при выделении только одного интервала ($k = 1$) корреляционное отношение $\eta^2_{y|x} = 0$. Если же выбрать разбиение на N интервалов таким образом, чтобы в один поддиапазон попало бы ровно по одной точке, то на той же выборке данных $\eta^2_{y|x}$ становится уже равным 1. Конечно, оба этих разбиения противоречат, если не букве полного корреляционного анализа, то здравому смыслу, однако показывают, что величина корреляционного отношения сильно зависит не только от характера распределения зависимой переменной вдоль оси фактора, но и в значительной мере от способа группировки.

При количественном выражении взаимной корреляции признаков выбор числа групп и границ интервалов – центральная проблема, так как этим обуславливается объективность характеристик связи. Субъективным критерием правильности выбора числа классов k является верная передача типа распределения эмпирических частот данной совокупностью. Если выбрано слишком мало классов, можно потерять характерную картину связи Y с X. При слишком подробном делении на классы можно стушевать реальную картину распределения частот случайными отклонениями. С точки зрения последователей градиентного анализа, акцент делается на "удобство интерпретации результатов" и предлагается всю область варьирования фактора разбивать на пять равных частей [Миркин, Наумова, 1983].

«Если имеется система долей, например, распределение количественного показателя с относительными частотами (в долях), то общая энтропия вариационного ряда равна сумме частных энтропий по классам распределения» [Плохинский, 1982]. Таким образом, задачу выбора границ интервалов можно определить как, в некотором смысле, минимизацию совокупной энтропии, получающейся в результате группировки.

Дополнительный смысл постановка этой задачи приобретает, когда для нахождения разбиения на поддиапазоны δ некоторой локальной выборки используются свойства распределения другой выборки, сопряженной с анализируемой (т.е. ищется минимум энтропии, основанной на условных вероятностях многомерного распределения нескольких показателей). В этом случае деление на интервалы учитывает реально существующие статистические зависимости факторов и наилучшим образом использует дифференциально-диагностические возможности признака Y для поиска закономерностей его связи с другим признаком X.

Рекомендуемая литература: [Айвазян с соавт., 1983; Алгоритмы и программы.., 1984; Генкин, 1999].

Математический лист

Методы деления количественной шкалы на интервалы

К числу эмпирических способов вычисления числа классов k для выборок умеренных размеров m можно отнести *правило Стургеса (Sturgess)* [Зайцев, 1984]:

$$K = 3.32 \lg(m) + 1 = 1.44 \ln(m) + 1, \qquad (6.32)$$

т.е. от 5 до 9 на наших примерах. Другие авторы [Хан, Шапиро, 1969; Лакин, 1990] считают, что число классов k должно быть 12 ± 3, т.е. разброс мнений весьма велик.

Поскольку нет единых теоретических оснований для оценки качества группировки, то принципиально допустим любой формальный алгоритм, удовлетворяющий определенным требованиям. Можно выделить следующие основные стратегии разбиения на градации:

- *стратегия равной ширины диапазонов* (при этом граничным значениям присваивается ряд "аккуратных" чисел, например, 0, 20, 40, 60, 80, 100);
- *стратегия равной заселенности диапазонов* (граничные значения выбираются таким образом, чтобы в один интервал попало примерно одинаковое количество измерений);
- *сигмальная стратегия* (разбиваемый показатель имеет отчетливое нормальное распределение и выбор границ диапазонов осуществляется в долях дисперсии: одна сигма, две сигмы, три сигмы и т.д. [причем, три сигмы теоретически считаются статистической границей "норма-патология"]);
- *стратегия равных площадей под кривой* $\boldsymbol{X}$ *между границами диапазонов* (т. е. дискретный аналог определенного интеграла);

Разработаны и практически применяются более строгие способы различных аппроксимаций частотных распределений: оценки Парзена-Надарая [Горелик, Скрипкин, 1984; Фомин, Тарловский, 1986], сглаживание гистограмм [Ивашко, Кузнецов, 1989] и другие оптимизационные стратегии, когда граничные значения выбираются из условия экстремума некоторого критерия. Такие критерии оптимизации разбиения могут быть двух типов:

- *внутрение* (основанные только на свойствах распределения самой разбиваемой выборки);
- с использованием *внешнего дополнения* (например, использующие свойства распределения другой выборки, сопряженной с анализируемой).

Примером оптимального решения с использованием внутреннего критерия является минимизация функционала [Браверман, Мучник, 1983]:

$$\boldsymbol{F} = \frac{1}{\boldsymbol{m}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{k}} \boldsymbol{m_i} \sum_{\boldsymbol{j}=1}^{\boldsymbol{m_i}} (\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})^2 \Rightarrow \min \ , \tag{6.33}$$

где обозначения те же, что и при расчете корреляционных отношений (6.15)-(6.18).

Все методы, использующие внутренние критерии, предполагают локальный анализ закономерности частотного распределения признака раздельно для каждой выборки $\boldsymbol{X}$ или $\boldsymbol{Y}$ без учета их взаимной статистической обусловленности, что с точки зрения теории информации нельзя считать вполне адекватным.

Использование информационных мер для оптимизации разбиения

Предположим, что каждая величина $\boldsymbol{y_i}$, $\boldsymbol{i}$ = 1,2,…,$\boldsymbol{m}$, принимающая значения на отрезке [$\boldsymbol{a}$, $\boldsymbol{b}$], принадлежит к одному из $\boldsymbol{n}$ классов измерений $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, …, $\boldsymbol{D_n}$ (это могут быть, например, водоемы, в которых проводились измерения, сезонные признаки или классы качества вод).

Введем разбиение $\boldsymbol{\delta}$ диапазона [$\boldsymbol{a}$, $\boldsymbol{b}$] на заранее заданное количество интервалов $\boldsymbol{k}$, границы которых заранее не определены. Обозначим через $\boldsymbol{p_j}(\boldsymbol{y}|\boldsymbol{D_s})$ частоту попадания значения показателя $\boldsymbol{Y}$ из подмножества $\{\boldsymbol{y}\}_{\boldsymbol{Ds}}$ в $\boldsymbol{j}$-й диапазон.

Тогда для двух классов $\boldsymbol{D_s}$ и $\boldsymbol{D_l}$ в качестве наилучшего разбиения диапазона [$\boldsymbol{a}$, $\boldsymbol{b}$] выбирается такое, которое максимизирует значение меры дивергенции, введенной С. Кульбаком [1967]:

$$\boldsymbol{J}(\boldsymbol{D_s} : \boldsymbol{D_l}; \boldsymbol{y}) = \sum_{\boldsymbol{j}=1}^{\boldsymbol{k}} (\boldsymbol{p_j}(\boldsymbol{y} \mid \boldsymbol{D_s}) - \boldsymbol{p_j}(\boldsymbol{y} \mid \boldsymbol{D_l})) \cdot \ln \frac{\boldsymbol{p_j}(\boldsymbol{y} \mid \boldsymbol{D_s})}{\boldsymbol{p_j}(\boldsymbol{y} \mid \boldsymbol{D_l})} \Rightarrow \max \tag{6.34}$$

В общем случае $\boldsymbol{m}$ классов максимизируется величина:

$$\boldsymbol{J} = \sum_{\boldsymbol{s}=1}^{\boldsymbol{n}} \sum_{\boldsymbol{l}=1}^{\boldsymbol{s}} \boldsymbol{J}(\boldsymbol{D_s} : \boldsymbol{D_l}; \boldsymbol{y}) \ . \tag{6.35}$$

Получаемое таким образом разбиение вместе с вероятностями появления значений признака в соответствующих интервалах $\boldsymbol{p_j}(\boldsymbol{y}|\boldsymbol{D_s})$ называется *интервальной структурой* [Генкин, 1999].

Для двух признаков $\boldsymbol{Y_1}$ и $\boldsymbol{Y_2}$, зная разбиения $\boldsymbol{\delta_1}$ и $\boldsymbol{\delta_2}$ для каждого из них, естественным образом строятся оценки $\boldsymbol{p_l}(\boldsymbol{Y_1},\boldsymbol{Y_2}|\boldsymbol{D_s})$ – частоты попадания пары значений анализируемых признаков в прямоугольники со сторонами, равными интервалам соответствующих разбиений. Множест-

во прямоугольников, вместе с оценками вероятностей попадания в них пары значений признаков $p_l(Y_1,Y_2|D_s)$, называется *бинарной* (матричной) *структурой*.

Таким образом, описываемый методологический принцип анализа различий заключается в том, что сравнению подвергаются не сами наблюдения, а их нормированные частоты попадания в ячейки интервальных или бинарных структур.

Поскольку интервальные и бинарные структуры мало чувствительны к систематическим (а частично и к случайным) ошибкам наблюдений, эти методы нашли широкое применение при обработке клинико-лабораторных признаков. Большой вклад в развитие интервальных методов при решении медико-биологических задач внесли Е.В. Гублер и А.А. Генкин, воплотившие и развившие эти информационные структуры в среде Оболочки Медицинских Интеллектуальных систем [Генкин, 1999].

Рассмотрим принципы формирования интервальных структур и решающих правил при сравнении двух выборок. *Дивергенция Кульбака*, которая имеет смысл средней информационной меры различия двух эмпирических распределений, для этого случая может быть вычислена по формуле

$$J = \sum_{i=1}^{k} \left(\frac{f_i}{m_1} - \frac{q_i}{m_2}\right) \ln \frac{f_i m_2}{q_i m_1} \quad , \tag{6.36}$$

где f_i, q_i – частоты попадания в i-й интервал примеров сравниваемых выборок, m_1, m_2 – численность обеих выборок. С.Кульбаком [1967] было показано, что статистика

$$J'(x,1:2) = \frac{m_1 m_2}{m_1 + m_2} J \quad , \tag{6.37}$$

основанная на дивергенции J, имеет распределение χ^2 с $(k - 1, 1)$ степенями свободы, что позволяет использовать ее для проверки нулевых гипотез. Критерий различия двух выборок $J'(x, 1:2)$, использующий информационную меру Кульбака, предлагается [Генкин, 1999] назвать *J-критерием*, а соответствующий ему уровень значимости обозначать P_J. А.А. Генкиным приводятся формулы, распространяющие использование меры Кульбака на случай сравнения n выборок, однако практических примеров техники такого анализа нам найти не удалось.

В.Н. Вапником с соавторами [Алгоритмы и программы.., 1984] представлен более общий алгоритм нахождения наилучшего разбиения, основанный на минимизации шенноновской энтропии и определяющий как границы диапазонов, так и оптимальное число градаций k.

Результаты расчетов

Рассмотрим выборку значений общей численности хищников-хватателей N_h, причем каждой гидробиологической пробе поставлен в соответствие класс качества воды по шестибальной шкале [ГОСТ 17.1.3.07–82; Драчев, 1964]. Найденные границы интервалов разбиения на 5 интервалов с использованием различных стратегий и критериев представлены в табл. 6.7.

Таблица 6.7

Границы интервалов и количество наблюдений в каждой градации при различных стратегиях деления на 5 диапазонов выборки из численностей хищников-хватателей (описательные статистики: объем выборки $m = 540$, среднее $M = 409.08$, стандартное отклонение $\sigma = 1663$, максимум – 27820, медиана – 80, мода – 0)

Равная ширина интервалов		В долях среднеквадратического отклонения		Минимизация функционала Бравермана – Мучника (6.33)		Равное количество значений без учета повторов		Равномерная "заселенность" интервалов	
Градации	Кол-во	Градации	Кол-во	Градации	Кол-во	Градации	Кол-во	Градации	Кол-во
< 6000	537	от 0 до M	433	< 575	467	< 104	280	= 0	141
< 12000	1	< (M + σ)	63	< 2640	63	< 260	97	< 55	99
< 18000	0	< (M +2σ)	17	< 5900	7	< 560	81	< 160	104
< 24000	1	< (M +3σ)	8	< 18278	2	< 1280	53	< 480	104
< 30000	1	> (M +3σ)	19	< 27820	1	< 27820	29	< 27820	92

Как видно из представленных результатов, выделение границ интервалов с использованием выборок, типичных для гидробиологических данных, представляет собой далеко не тривиальную проблему. Использование традиционных стратегий равной ширины или долей сигмы, а также большинства формальных критериев, приводит к существенно асимметричному разбиению, которое не может продуктивно использоваться в последующем анализе.

Разделим все множество наблюдений из численностей хищников-хватателей N_h на две выборки: измерения на "чистых" станциях с классом качества вод 3 и менее и измерения на "грязных" станциях. Традиционное сравнение средних с использованием ***t***-критерия Стьюдента не выявляет статистических различий между этими выборками ($p = 0.45$). Осуществим такое разбиение всей области варьирования значений численности на 5 интервалов, которое обеспечивало бы максимальную расщепляющую способность обоих подмножеств, т.е. наибольшую суммарную разность частот (пересчитанную в доли ***J***) в ячейках таблицы сопряженности. Определение оптимального вектора границ диапазонов осуществлялось нами по алгоритму случайного поиска до тех пор, пока значение информационной меры Кульбака ***J*** не перестает возрастать (см. табл. 6.8).

В первом столбце табл. 6.8 – интервалы, найденные компьютером, наилучшим образом подчеркивающие различие вариабельности численности хищников в рассматриваемых группах. Во втором и третьем столбцах – частота (в скобках – относительная частота) наблюдений численности из соответствующих интервалов. Справа приводятся средние арифметические, не различающиеся по ***t***-критерию, тогда как мера Кульбака ***J'*** свидетельствует о значимом ($p_J < 0.00001$) изменении численности N_h в зависимости от уровня гидрохимического загрязнения на станциях наблюдения.

Таблица 6.8

*Интервальные структуры численности хищников-хватателей при относительно низком (группа **A** – класс качества воды < 4) и высоком (группа **B** – класс качества воды ≥ 4) уровнях химического загрязнения*

Градации численности N_h хищников, экз./м^2	Группа ***A*** (класс < 4) $N_1 = 186$	Группа ***B*** (класс ≥ 4) $N_2 = 142$	Вклад в информативность	Дивергенция и статистика Кульбака	Средние значения численности в группах ***A*** и ***B*** и их отличие по ***t***-критерию	
0 - 3	28 (15.1%)	118 (33.3%)	0.263		Гр. ***A***	Гр. ***B***
5 - 10	9 (4.8%)	4 (1.1%)	0.030	$J = 0.307$	406.1	390.3
12 - 120	62 (33.3%)	99 (28.0%)	-0.025	$J'(4,1)=37.4$	$t = 0.129$	
130 - 140	7 (3.8%)	1 (0.3%)	0.068	$P_J < 0.0001$	$t_{кр} = 1.65$	
≥ 149	80 (43.0%)	132 (37.3%)	-0.029		$\underline{p} = 0.45$	

Анализ соотношения составляющих дивергенции Кульбака (вклада в информативность) свидетельствует о том, что различия между группами на 85% обусловлены малыми значениями численности хищных видов зообентоса (диапазон от 0 до 3), которые в "грязных" условиях среды встречаются значительно чаще.

Рассмотрим еще один пример. На гистограммах рис. 6.4 представлено распределение численности некоторых подсемейств и триб хирономид по оптимальным диапазонам разбиения, контрастирующим различия групп наблюдений с разными классами качества воды (группирующий признак тот же, что и в предыдущем примере).

Из всех групп хирономид, которые на рис. 6.4 следуют в порядке убывания информативности, наилучшими индикаторами чистых вод явились виды подсемейства Orthocladiinae, в то время, как наименьшей информационная ценность приходится на подсемейство Tanypodinae. Отчетливая зависимость от уровня загрязнения просматривается и для видов трибы Tanytarsini. Интересной оказалась связь с классом качества для видов трибы Chironomini, частота появления больших численностей которых характерна именно для грязных вод.

С использованием ***J***-критерия, предложенного А.А. Генкиным, для всех проанализированных таксономических подмножеств видов были установлены статистически значимые ($p_j < 0.0003$) отличия между группами с разным уровнем загрязнения.

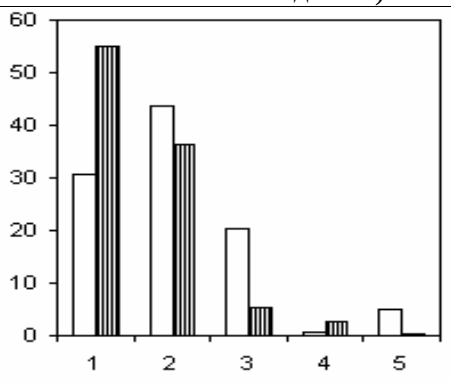
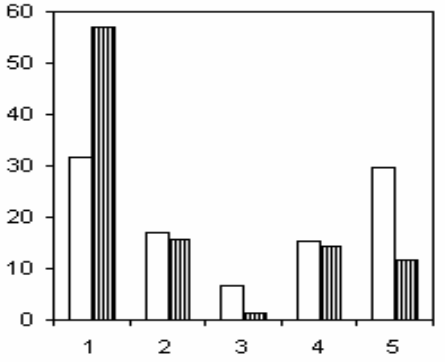
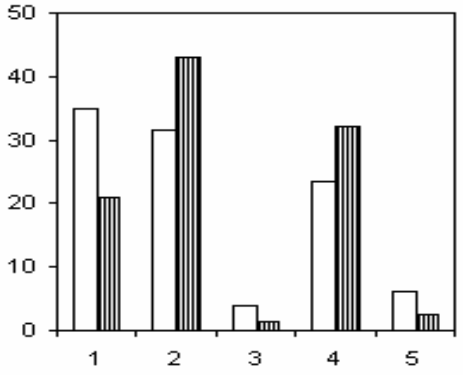
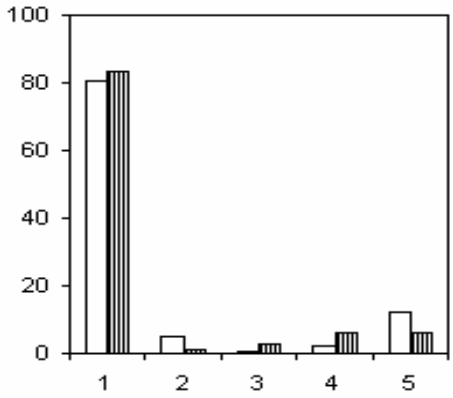
Таксономические группы	Градации диапазонов		Вклад в информативность	Распределение относительных частот по интервалам структуры, % (заштрихована частота группы с классом качества воды > 3)
	начало	конец		
Подсемейство Orthocladiinae, ***J*** = 0.507	0	10	0.145	
	13	334	0.013	
	340	1440	0.192	
	1480	3160	0.032	
	3300	22880	0.125	
Триба Tanytarsini, ***J*** = 0.404	0	0	0.15	
	1	40	0.001	
	48	72	0.086	
	74	230	0.001	
	240	14000	0.167	
Триба Chironomini, ***J*** = 0.189	0	35	0.072	
	40	560	0.034	
	570	640	0.028	
	660	6800	0.028	
	6880	24168	0.029	
Подсемейство Tanypodinae, ***J*** = 0.169	0	240	0.001	
	250	280	0.049	
	299	350	0.039	
	360	540	0.041	
	560	18278	0.04	

Рис. 6.4. Гистограммы распределения численности по интервалам для различных таксономических групп хирономид

Поскольку для видов Tanypodinae предыдущие методы ставили под сомнение эту гипотезу, можно предположить, что J-критерий (впрочем, как и χ^2) склонен к гипердиагностике вероятности различий между выборками, особенно, в случае большого их объема.

Таким образом, сопряженные таблицы в рамках изложенной технологии оказываются уже не просто набором независимых друг от друга частот появления значений признака в определенных интервалах, а *структурой*, элементы которой – экологически значимые интервалы проявления жизнедеятельности таксономических групп в различных условиях среды.

Многолетний опыт использования описанного метода обработки при анализе лабораторных и инструментальных признаков отчетливо выявил, по крайней мере, 6 различных типов интервальных структур. Приведем некоторые из них.

Самый простой тип – *линейные структуры*, когда частоты для одного условия монотонно возрастают (убывают), а для другого – монотонно убывают (возрастают), т.е изменяются разнонаправлено. Более сложны для интерпретации часто встречаемые *реципрокные структуры*, в которых нет монотонности, но частоты отрицательно коррелируют. Например, интервальные структуры, у которых частоты для минимальных и максимальных значений статистически неразличимы, а значения центральных частот реципрокны, называются реципрокными в центре (в фитоценологической ординации такие структуры носят названия «клинов»: топоклины, термоклины, галоклины, ценотопоклины и проч.) В некоторых типах структур реципрокность четко выражена только в двух диапазонах изменения признака, справа или слева.

6.4. Нахождение пороговых значений с использованием детерминационного анализа

Формулировка задачи

Пусть имеется две группы числовых переменных $\boldsymbol{X} \equiv (x_1, ..., x_p)$ и $\boldsymbol{Y} \equiv (y_1, ..., y_m)$, причем предполагается, что $\boldsymbol{X}$ – независимая переменная ("объясняющий признак") влияет на значения $\boldsymbol{Y}$ – зависимой переменной ("объясняемый признак"). В общем случае предполагается, что обе переменных измерены в количественных шкалах (интервальной, абсолютной или шкале отношений).

Необходимо найти такое значение $\boldsymbol{E} = \boldsymbol{x}_{крит}$, которое может трактоваться на используемом эмпирическом материале как некоторый *порог толерантности*. Синонимами термина $\boldsymbol{x}_{крит}$ являются "предельно допустимая или критическая нагрузка", "экологически допустимый уровень (ЭДУ) нарушающих воздействий" [Левич, 1994] и ряд других словоформ, которые подробно обсуждались в разделе 1.4.

На протяжении всей книги мы неоднократно обращались и будем обращаться к понятию «пороговости», а в этом разделе покажем, как значение $\boldsymbol{x}_{крит}$ может быть получено с использованием несложного алгоритма анализа таблиц сопряженности 2×2, который в социологических исследованиях именуется как «*детерминационный анализ*». Использованию этого метода для анализа антропогенных воздействий на пресноводные экосистемы посвящена серия статей коллектива авторов из Московского университета – В.Н. Максимова, Н.Г. Булгакова, А.П. Левича и др. [Замолодчиков с соавт., 1992; Булгаков с соавт., 1995; Левич, Терехин, 1997; Максимов с соавт., 1999, 2000а,б, 2001]. Эти публикации следует трактовать как первый серьезный (хотя и методологически не во всем бесспорный) вклад в разрешение сложных вычислительных проблем оценки границ между областями нормального и патологического функционирования природных объектов.

Математический лист

Основная концепция детерминационного анализа заключается в том, что роль функций традиционного регрессионного или факторного анализа выполняют *правила*, которые авторы метода возводят к силлогистике Аристотеля. *Правило – это особый математический объект, представляющий суждение вида «Если $\boldsymbol{A}$, то $\boldsymbol{B}$» (или сокращенно $\boldsymbol{A} \rightarrow \boldsymbol{B}$), где $\boldsymbol{A}$, $\boldsymbol{B}$ – соответственно, объясняющий и объясняемый признаки.*

Любое правило вида $\boldsymbol{A} \rightarrow \boldsymbol{B}$ всегда рассматривается вместе с двумя характеристиками:

- *точностью* $\boldsymbol{T} = \boldsymbol{N}(\boldsymbol{A},\boldsymbol{B})/\boldsymbol{N}(\boldsymbol{A})$, равной доле случаев $\boldsymbol{B}$ среди случаев $\boldsymbol{A}$ (т.е. доле случаев, в которых правило действует безошибочно);

- *полнотой* $P = N(A,B)/N(B)$, равной доле случаев A среди случаев B (т.е. доле случаев, объясняемых данным правилом),
  где $N(A,B)$ – количество случаев, в которых A и B встречаются вместе; $N(A)$ – количество случаев, когда имеет место A безотносительно к B; $N(B)$ – количество случаев, когда имеет место B безотносительно к A.

Собственно, любое логическое суждение становится правилом детерминационного анализа ("детерминацией") только тогда, если его точность и полнота удовлетворяют некоторым критериям, т.е. находятся в некоторых границах, причем показатель T оценивает надежность, а P – универсальность суждения.

Рассмотрим теперь применимость детерминаций для оценки порога толерантности. Пусть X – нормируемый фактор среды, а Y – оценка экологического состояния. Предположим, что нам известны два числа: $E = x_{крит}$ – значение пороговой нагрузки фактора, которое нам предстоит найти, и $F = y_{норм}$ – пограничное значение на шкале Y, за пределами которой экологическое состояние оценивается как неблагополучное. В этих условиях корреляционное поле рассеяния всех n значений Y-X можно свернуть до стандартной таблицы сопряженности 2×2, в ячейках которой находятся частоты (количества измерений), соответствующие приведенным условиям:

Условия для X и Y	$Y \geq F$	$Y < F$	Итого
$X < E$	a	c	$a + c$
$X \geq E$	b	d	$b + d$
Итого	$a + b$	$c + d$	$n = a + b + c + d$

В соответствии с этими обозначениями, точность детерминации «Из того, что $X < E$, следует $Y \geq F$» определяется по формуле

$$T = a / (a+c), \tag{6.38}$$

а ее полнота – по формуле

$$P = a / (a + b). \tag{6.39}$$

Приведенное правило соответствует обратно пропорциональной связи между признаками, хотя в определенных ситуациях (например, для концентрации растворенного в воде кислорода) детерминационное правило может быть сконструировано по альтернативному механизму связи.

В общем случае, граница области экологически допустимых значений фактора двухсторонняя (вспомним, *по ком звонят "уиттекеровские колокола"*); следовательно, необходимо ввести две пороговые переменные E_1 и E_2. При этом меняются только столбец условий для X, первое из которых будет выглядеть как «$X \geq E_2$ или $X \leq E_1$», а второе – как «$E_2 > X > E_1$», в то время как само правило и выражения для полноты и точности никаких изменений не претерпевают.

Естественно, что можно предложить достаточно большое (из соображений общности можно сказать "бесконечно большое") количество чисел-претендентов для пороговых значений E и F. Предлагается считать оптимальной такую величину E, которой соответствует максимальное значение полноты P при заданной нижней границе точности T. Процедура поиска оптимального решения в детерминационном анализе использует технику полного перебора и сводится к следующему алгоритму:

- рассматривается правило $A \rightarrow B$, где $A = \{ E_1 \leq x \leq E_2 \}$;
- точность $T(E_1, E_2)$ и полнота $P(E_1, E_2)$ правила $A \rightarrow B$ определяются как функции от нижней границы E_1 и верхней границы E_2 интервала X;
- задается число h, означающее нижнюю границу точности правила, которое должно получиться в результате проведения оптимизации;
- вычисляются функции $T(E_1, E_2)$, $P(E_1, E_2)$, т.е. определяются точность и полнота для всех допустимых значения границ E_1, E_2 интервала X;
- отбираются те правила, для которых $T(E_1, E_2) \geq h$; если таких правил нет, это значит, что задача оптимизации при заданной величине порога h не имеет решения;
- если такие правила существуют, решением задачи оптимизации служит пара чисел E_1, E_2, для которых $P(E_1, E_2) = max$.

В работах В.Н. Максимова с соавторами подробно описана также технология детерминационного анализа совокупного действия нескольких факторов среды в виде многофакторных (до 5) комплексов, поэтому читатель, заинтересованный в изучении проблем синергетики воздействий, может обратиться непосредственно к первоисточникам.

Подробно ознакомиться с концепцией детерминационного анализа можно на сайте разработчиков ДА-программы: http://www.context.ru, где также приводятся своеобразные "рецепты" стратегии и тактики статистической обработки. Вот, например, цитата с этого сайта, олицетворяющая тип "агрессивно-самоутверждающего" рекламного мема:

> **Мем № 35**: *«В настоящее время имеется около сотни различных методов измерять связь между признаками… Такое "разнообразие методов" не имеет научной ценности, поэтому без ущерба для дела может и должно быть упразднено. Мы сделали это в ДА-системе, предложив пользователям ясную концепцию связи, реализованную в понятии правила (детерминации). Аналогично обстоит дело и с "разнообразием методов" в решении задачи построения новых характеристик на основе заданных. Здесь "многочисленность методов" также имеет фиктивный характер…»* [http://www.context.ru].

Думается, что любой читатель, памятуя о *принципе множественности моделей сложных систем* (см. раздел 2.4), будет неприятно озадачен предложением "закрыть за ненадобностью" всю прикладную математику последнего полувека, заменив ее соотношениями полноты и точности двухпольных таблиц сопряженности…

В то же время, используемый в детерминационном анализе алгоритм выбора наилучших границ интервала $\boldsymbol{X}$ с позиций классических методов оптимизации нельзя признать *результативным*. Расчеты по ДА-программе предполагают большую неопределенность конечного результата при практически неконтролируемой свободе выбора параметра условия $\boldsymbol{h}$-порога точности $\boldsymbol{T}$.

Будем, например, искать экологически допустимый уровень (ЭДУ) концентрации минерального фосфора в условиях, описанных ниже в таб. 6.9, используя график полноты и точности для различных значений ЭДУ на рис. 6.5. Очевидно, что значение полноты всегда монотонно увеличивается по мере увеличения ширины диапазона $\boldsymbol{E_1}$-$\boldsymbol{E_2}$, причем естественный максимум полноты $\boldsymbol{P_{max}}$ = 100% достигается при полном охвате всей области варьирования. При этом можно выбрать широкое множество пороговых значений $\boldsymbol{h}$ для точности, которые субъективно могут показаться вполне приемлемыми:

- при $\boldsymbol{h} = \boldsymbol{T} = 60$ – получаем $\boldsymbol{P} = 35.4$ и ЭДУ = 0.023 мг/л;
- при $\boldsymbol{h} = \boldsymbol{T} = 53.1$ – получаем $\boldsymbol{P} = 54.2$ и ЭДУ = 0.036 мг/л;
- при $\boldsymbol{h} = \boldsymbol{T} = 50$ – получаем $\boldsymbol{P} = 68.8$ и ЭДУ = 0.054 мг/л и т.д.

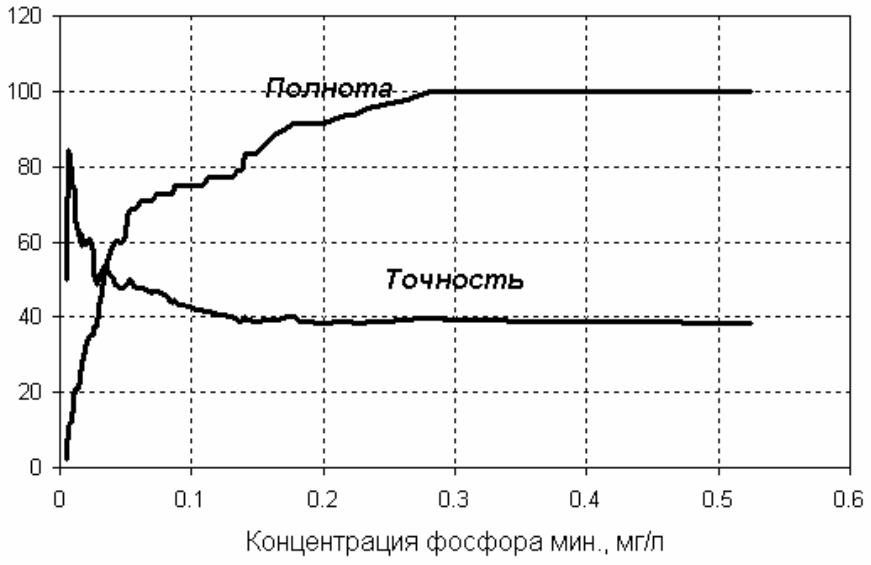


Рис. 6.5. Зависимость полноты и точности детерминационных правил от выбранного порога толерантности для концентрации фосфора

Поэтому, целесообразнее осуществить постановку задачи поиска наилучшего решения в иной интерпретации без параметра ***h***, сконструировав подходящий критерий безусловной оптимизации в виде функциональной комбинации ***P*** и ***T***. Например, в теории конструирования библиографических информационно-поисковых систем проблема соотношения полноты и точности (именно оттуда социология заимствовала эти понятия) исследовалась с начала 60-х годов [Аветисян, 1973; Селтон, 1973]. В частности, там широко используется [Попов, 1981, URL] *критерий релевантности* выдачи документов

$$\nu = (P*T)^{0.5}, \qquad (6.40)$$

который при определенных условиях сводится к тетрахорическому коэффициенту корреляции Чупрова – см. формулу (6.8) раздела 6.1.

Результаты расчетов

Сформируем по результатам мониторинга сообществ зообентоса на малых реках Самарской области несколько выборок, содержащих сопряженные значения наблюдений пар признаков, один из которых (гидробиологический) считается объясняемым, а другой (гидрохимический) – объясняющим.

Предпосылки детерминационного анализа требуют непременной трансформации объясняемого признака в шкалу двух градаций. Сама по себе эта процедура не является тривиальной, поскольку в разделе 6.3 перечислено не менее пяти различных стратегий такого разбиения. В.Н. Максимов с соавторами предлагают делить область варьирования на "два примерно равно заполненных класса" или использовать в качестве границы "среднемноголетнее значение". Впрочем, анализ устойчивости получаемой оптимальной величины порога толерантности в зависимости от стратегии трансформации гидробиологического показателя в шкалу двух градаций выходит за рамки настоящего изложения.

Сведения об использованных выборках, выбранные условия "благополучности" экосистем и результаты расчетов экологически допустимых уровней (ЭДУ) представим в форме стандартных таблиц сопряженности в табл. 6.9.

Таблица 6.9

*Результаты оценки экологически допустимого уровня трех гидрохимических показателей методом детерминационного анализа (затемнены ведущие ячейки таблиц сопряженности, по которым вычислялись критерии полноты **P** и точности **T**)*

Действующий фактор	Градации фактора относительно критического порога	Неблагополучное состояние	Благополучное состояние	Итого по строкам
$БПК_5$, $мгO_2/л$	$ПДК_{вр}$ – не более 3	Биотический индекс Вудивисса ***V***		
		Менее 3	3 и более	
	Менее 4.02	18	**24**	**42**
	Более 4.02	43	14	57
	Итого по столбцам	61	**38**	***m*** = 99
	Критерии:	***P*** =24/38 = 63.2%	***T***=24/42 = 57.1%	
Фосфор минеральный, мг/л	$ПДК_{вр}$ – не более (0.05 ÷ 0.2)	Индекс Шеннона ***H***		
		Менее 2.5	Более 2.5	
	Менее 0.036	23	**26**	**49**
	Более 0.036	54	22	76
	Итого по столбцам	77	**48**	***m*** = 125
	Критерии:	***P*** = 26/48 = 54.2%	***T***= 26/49 = 53.1%	
Кислород у дна, % от степени насыщенности	$ПДК_{вр}$ – не менее 75	Численность хищников-хватателей N_h		
		Менее 100 экз/м2	Более 100 экз/м2	
	Между 89 и 119	77	**117**	**194**
	Менее 89 или более 119	65	47	112
	Итого по столбцам	142	**164**	***m*** = 306
	Критерии:	***P*** =117/164=71.3%	***T***=117/194 =60.3%	

По каждой из трех выборок ведущей ячейкой в представленных таблицах сопряженности является «Благополучное состояние экосистемы» и «Действие фактора в диапазоне толерантности (ЭДУ)», а критериями оптимальности сформированной таблицы считаются доли частоты в ведущей ячейке к маргинальным значениям по строке (полнота ***P***) и столбцу (точность ***T***). Диапазоны ЭДУ были рассчитаны по алгоритму соотношения полноты и точности, используемому в программе ДА-анализа.

Найденные значения порогов толерантности по $БПК_5$ (менее 4.02 мг/л) и концентрации фосфатов (менее 0.036 мг/л) оказались весьма близки соответствующим значениям ПДК для водоемов рыбохозяйственного назначения (см. табл. 3.1). Что касается ЭДУ растворенного кислорода (между 89 и 119%), то можно предположить сильное влияние на его величину характера распределения использованных выборок.

Глава 7. Задачи о классификациях: отношения сходства и порядка для многомерных объектов

7.1. Техника выделения ассоциаций по Браун-Бланке

Мем № 36: *«…сообщества – это условности, объединенные в многомерный континуум… При создании классификаций используют* ***прагматический подход****. Любая классификация рассматривается как рабочее приближение, и если ее уровень достаточен для практического использования, то дальнейшее совершенствование классификации нецелесообразно»* Б.М. Миркин с соавт. [2000].

Формулировка задачи

Пусть имеется матрица наблюдений $\boldsymbol{X}$ размерностью $\boldsymbol{n \times m}$. Столбцами этой матрицы являются исследуемые объекты, соответствующие, в нашем случае, тем географическим точкам наблюдений $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{n}$, где были взяты гидробиологические пробы, а в строках, $\boldsymbol{j} = 1, 2, \ldots, \boldsymbol{m}$ будем размещать список видов изучаемых гидробионтов (или некоторых более крупных таксономических групп). Значениями матрицы наблюдений $\boldsymbol{x_{ij}}$ являются конкретные свойства $\boldsymbol{j}$ объекта $\boldsymbol{i}$, которые имеют смысл обилия (численность, биомасса, балльная оценка) $\boldsymbol{j}$-го вида и могут быть выражены в шкале произвольного характера.

Задачи классификации экологических сообществ (или задачи *синтаксономии*) сводятся к следующему: необходимо найти для столбцов или строк матрицы наблюдений статистически обоснованные и непротиворечивые структурные отношения сходства и упорядоченности, или иначе – установить классификационную типологию. Если интерпретировать экологические сообщества в традиционных терминах многомерной геометрии, то описания могут быть представлены своеобразными "галактиками" точек в многомерном пространстве осей, представляющих обилия видов [Goodall, 1963; Василевич, 1962, 1969; Миркин, Розенберг, 1978, 1979; Герасименко, Ипатов, 1980; Розенберг, 1984; Миркин, Наумова, 1998]. Задача синтаксономической классификации сводится в этом случае к исследованию закономерностей плотности распределения объектов: выделению наиболее "густых" зон скопления точек и игнорированию "разрежений".

Фундаментальным принципом, используемым для группировки нечетких множеств любой происхождения и имеющим солидное теоретическое обоснование (в том числе, в биоценотических исследованиях [Шмидт, 1964; Куприянова, 1977; Колодяжный, Пааль, 1985; Ястребов, 1991]), является процедура отбора наиболее информативных признаков (т.е. индикаторных видов), играющих ведущую роль в формировании однородных кластеров объектов.

Задача выделения ассоциаций в пространстве видов по Браун-Бланке в своей исходной постановке во многом совпадает с общей задачей распознавания образов "без учителя": на множестве имеющихся описаний необходимо сформировать такие комбинации видов (синтаксоны различных рангов), которые могут оказаться потенциально полезными для эколого-флористической диагностики произвольных сообществ.

Рекомендуемая литература: [Westhoff, van der Maarel, 1978; Миркин, 1989; Миркин, Наумова, 1998; Миркин с соавт., 2000].

Математический лист

Идея классифицировать экологические сообщества по признаку присутствия и отсутствия групп видов возникла независимо в ряде научных направлений, хотя наибольшее развитие она получила в работах Института SIGMA (Station Internationale de Geobotanique Mediterranean et Alpine) в Монпелье (Франция), созданного Жозья Браун-Бланке [J. Braun-Blanquet][1].

В понимании последователей этой школы выделяемые кластеры сообществ – это лишь центры полей рассеивания, образуемые подмножествами индикаторных видов или условными синтаксонами, в то время, как виды-эвритопы из классификационных правил исключаются. Конструкции синтаксонов, сформированных по характерным комбинациям диагностических видов, регули-

[1] История вопроса подробно рассмотрена в ряде публикаций Б.М. Миркина (см., например, [Миркин, Наумова, 1998]).

руются "Кодексом фитосоциологической номенклатуры" [Barkman et al., 1986; цит. по: Миркин, 1989], который регламентирует определенную последовательность и отличительные особенности синтаксономических рангов.

Синтаксоны эколого-флористической классификации устанавливаются на основе диагностических видов, среди которых различаются следующие типы:

- *константные виды*, встречаемые часто в разных ассоциациях, но не обязательно связанные своим экологическим диапазоном с условиями среды, где распространены ее сообщества;
- *дифференциальные виды*, которые заходят в ассоциацию лишь частью своего диапазона и могут входить в состав нескольких синтаксонов;
- *"верные"* или *характерные виды*, встречающиеся только в одном синтаксоне, экологическая амплитуда которых целиком укладывается в рамки заданных условий существования.

По верности виды подразделяют на пять ступеней: от "абсолютно верных" до "случайных спутников". Поскольку видам присуща экологическая индивидуальность, различают также главные, региональные и локальные "верные" виды: ареал первых совпадает с ареалом ассоциаций, вторые выходят за ее пределы, третьи занимают часть ареала. В современных геоботанических исследованиях характерные виды выступают именно в последнем качестве и чаще всего играют диагностическую роль только в отдельных районах (т.е. являются *локальными*).

Основной классификационной единицей эколого-флористической классификации растительности является *ассоциация* - растительное сообщество определенного флористического состава с единообразными условиями местообитания и единообразной физиономией. Ассоциация, тем не менее, занимает промежуточное положение в общей синтаксономической иерархии, которая имеет следующую последовательность ступеней:

класс $\Rightarrow$ *порядок* $\Rightarrow$ *союз* $\Rightarrow$ *ассоциация* $\Rightarrow$ *субассоциация* $\Rightarrow$ *вариант* $\Rightarrow$ *фация*.

На последних ступенях деления проявляются малосущественные видовые различия, как правило, находящие свое выражение в степени представленности диагностических групп видов.

Перечисленные классификационные положения общей концепции Браун-Бланке – Тюксена на определенном этапе сыграли важную роль в теоретическом обосновании общих конструкций синэкологии. Дальнейший опыт обработки массовых табличных данных показал, что далеко не всегда удается критерии "верных" видов или "эталонных ассоциаций" распространить на сколько-нибудь широкий круг биотопов или географических регионов [Василевич, 1969]. Однако, в последние 15-20 лет метод Браун-Бланке переживает определенный ренессанс [Миркин с соавт., 2000], связанный как с появлением эффективных компьютерных программ, так и с развитием новых методологических принципов (*дедуктивный метод классификации К. Копецки и С. Гейни*, *принцип множественности синтаксиномических решений* и т.д.; см. [Миркин и др., 1989]).

Отличительной особенностью классификационной техники Браун-Бланке является использование в качестве операционного поля непосредственно первичной матрицы наблюдений, тогда как большинство других алгоритмов кластеризации подвергают анализу вторичную матрицу – матрицу связи признаков или описаний.

Синтетический этап методики Браун-Бланке заключается в последовательном преобразовании описанной выше исходной "валовой" матрицы, где по столбцам расположены точки отбора проб (описания), а по строками – виды растений или организмов. Элементами матрицы являются значения численности или биомассы, которые необходимо пересчитать в баллы обилия от 1 до 9. Основная идея этого этапа метода чрезвычайно проста – необходимо целенаправленно переставлять строки и столбцы матрицы наблюдений до тех пор, пока таблица не приобретет, насколько это возможно, отчетливую блочную структуру. Процесс такой ***RQ***-диагонализации первоначальной таблицы приводит к тому, что по ее главной диагонали начинают просматриваться прямоугольные блоки клеток, соответствующие встрече групп близких по экологии видов в группах биоценозов сходных местообитаний.

Существует много версий алгоритмов такого перебора, основанных как на интуитивных, полуэмпирических стратегиях, так и на достаточно строгих формальных методах (например, на использовании математического аппарата задачи о назначениях). В качестве примера приведем краткое описание схемы обработки в модификации, предложенной Х. Элленбергом [Ellenberg, 1956, цит. по: Розенберг, 1984]:

- проводится ранжирование строк (т.е.видов), из которых укомплектовывается три блока: малоинформативные виды ***A*** высокой (свыше 70%) и ***C*** низкой (менее 10%) встречаемости, а также информативные виды ***B*** средней встречаемости;
- блоки ***A*** и ***C*** образуют пассивную часть таблицы, а блок ***B*** – ее активную часть;
- столбцы активной части таблицы ранжируются в соответствии со специально подбираемой системой коэффициентов, учитывающей соотношение разных диагностических групп и расстояние их центра тяжести от главной диагонали активного блока;
- столбцы активной и пассивной части переставляются в соответствии со значениями ранговых коэффициентов, после чего все блоки склеиваются в итоговую синтаксономическую таблицу.

Завершает автоматическую обработку синтаксономический этап, где рассматриваются новые варианты порядка строк и столбцов, при которых виды со сходной ролью в сходных биоценозах оказались бы рядом, с последующей оценкой сформированных ассоциаций. Здесь также приходится пользоваться "ножницами и клеем" и проводить "разрезание" таблицы по столбцам и строкам, только все эти операции выполняются уже "с ведома и по распоряжению" исследователя.

Значительная алгоритмическая завершенность аналитической части метода Браун-Бланке была достигнута в компьютерной программе TWINSPAN, составленной В. Ностом (V. Noest) в постановке М.О. Хилла [Hill, 1979] и дающей хорошие результаты обработки с возможностью гибкой настройки процесса классификации со стороны пользователя.

Результаты расчетов

Сформируем исходную таблицу для обработки по методу Браун-Бланке, используя численность ***N*** и биомассу ***B*** видов хирономид, встретившиеся на каждой из 22 станций наблюдений реки Чапаевка. Для сокращения объема публикационного материала из серии наблюдений, полученных для каждого створа реки, было случайным образом отобрано только по одной пробе, в результате чего сформировалась исходная матрица из 22 столбцов и 44 строк.

В качестве показателей обилия использовалось значения логарифмов от индекса плотности населения $\ln((N*B)^{0.5})$, округленные до целого и нормированные на интервале от 1 до 9 (подробное описание алгоритма расчетов см. в разделе 6.1). Результаты синтетической обработки по программе TWINSPAN представлены в табл. 7.1.

Визуальный анализ результирующей таблицы позволяет сделать, например, следующие предварительные выводы:

- приписать ассоциативность блокам таблицы {***I***6:***K***7}, (***R***23:***S***28}, {***R***39:***U***44},
- виды 5, 13, 15 трактовать как малоинформативные, в противоположность видам 7, 8, 44;
- ограничившись третьим уровнем разбиения, сформировать 4 группы станций по ассоциированности видового состава: {2, 3, 16}, {13, 15, 11, 20, 21, 8, 10, 23, 14, 17}, {1, 4, 9, 18} и {2, 3, 16}, выделив отдельно станции 7 и 19.

Дополнительная полуавтоматическая синтаксономическая обработка позволяет наполнить формально найденную ассоциативность содержательным экологическим смыслом на основе опыта исследователя-гидробиолога и существенно улучшить познавательную ценность итоговой таблицы.

Используя технику Браун-Бланке, можно находить взаимосвязи не только внутри ассоциаций видов, но и на более высоком трофико-таксономическом уровне. Сформируем исходную таблицу, выбрав в качестве столбцов 14 станций наблюдения на р. Сок и приняв за показатели обилия ранжированные значения $\ln((N_s*B_s)^{0.5})$, где N_s и B_s – средние по всем имеющимся наблюдениям суммарные значения численности и биомассы по отдельным группам водных организмов. В качестве основания для группировки было использовано сочетание системологического признака (подсемейство или триба) и принадлежность к той или иной трофической группе.

Например, из 8 проб, взятых на станции 1 р. Сок, виды трибы Tanytarsini подсемейства Chironominae, относящиеся к детритофагам собирателям, встретились 4 раза с суммарной численностью 10, 80, 210 и 20 экз./м^2 (т.е. в среднем 40 экз./м^2 при средней биомассе 0.156 г/м^2). Отсюда индекс обилия равен $\ln((40*0.156)^{0.5}) = 1.54$ или 2 после нормировки.

Использование показателей обилия видов или таксономических групп, осредненных по всем наблюдениям, полученным на анализируемом участке, вообще говоря, проблематично:

Таблица 7.1

Результаты обработки по методу Браун-Бланке матрицы видового состава хирономид по станциям наблюдений на р. Чапаевка

№№ пп	Названия видов хирономид	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***	***K***	***L***	***M***	***N***	***O***	***P***	***Q***	***R***	***S***	***T***	***U***	***V***	Уровни разбиения видов
		Номера станций наблюдений																						
		2	3	16	13	15	11	20	21	8	10	23	14	17	1	4	9	18	5	6	12	7	19	
1	*Cladotanytarsus mancus*	2	2	-	-	-	-	-	-	2	3	-	-	-	-	2	-	-	2	-	-	2	-	000
2	*M.pedellus*	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	2	-	000
3	*Paratanytarsus confusus*	-	-	2	-	-	-	-	-	2	2	-	-	-	1	2	-	-	3	-	-	2	-	000
4	*Stictochironomus crassifor*	-	-	-	-	-	-	-	-	2	3	-	-	-	-	-	2	-	-	-	2	2	-	000
5	*P.nubeculosum*	1	-	-	-	-	-	2	-	3	2	2	2	-	-	3	2	-	2	-	3	2	-	001
6	*Ablabesmyia* sp.	1	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	01000
7	*Dicrotendipes notatus*	-	-	-	-	-	-	-	-	2	3	-	2	-	-	-	-	-	-	-	-	-	-	01000
8	*G.glaucus*	-	-	-	-	-	-	-	-	2	2	-	2	-	-	-	-	-	-	-	-	-	-	01000
9	*Ablabesmyia mallochi*	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	01001
10	*Endochironomus albipennis*	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	01001
11	*Orthocladius annectens*	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	01001
12	*Psectrocladius fabricus*	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	01001
13	*Chironomus plumosus*	1	3	3	3	3	3	3	3	3	3	3	3	-	2	2	3	-	2	2	2	-	-	01010
14	*Cryptochironomus* gr. *defectus*	-	-	-	-	-	-	-	-	2	3	2	2	2	-	3	2	-	-	-	2	-	-	01010
15	*P.ferrugineus*	2	-	-	-	-	3	2	2	2	2	3	2	2	1	2	2	-	2	3	2	-	-	01010
16	*Tanypus vilipennis*	1	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	01010
17	*P.arcuatus*	-	-	-	-	3	-	-	-	-	-	-	2	-	-	2	-	2	-	-	-	-	-	01011
18	*Microchironomus tener*	1	-	-	-	-	-	-	-	2	2	2	-	-	-	-	2	2	2	-	2	-	-	0110
19	*Parachironomus varus*	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	0110
20	*Ablabesmyia monilis*	1	-	-	-	-	-	-	-	-	-	-	2	-	-	-	2	-	3	2	2	-	-	011100
21	*Cladopelma* gr. *lateralis*	-	-	-	-	-	-	-	-	2	-	2	-	-	-	2	2	-	2	3	2	-	-	011100
22	*C.sylvestris*	2	-	2	-	-	-	-	-	-	-	-	-	-	2	2	2	2	-	-	2	-	-	011100
23	*Hydrobaenus distylus*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	011101
24	*Nanocladius bicolor*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	011101
25	*Parakiefferiella bathophila*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	011101
26	*Paracladius conversus*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	011101
27	*Paratanytarsus lauterborni*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	011101
28	*Trissocladius* sp.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	011101
29	*P.bicrenatum*	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	2	-	3	-	-	-	-	011110
30	*Corynoneura scutellata*	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	011111
31	*Cricotopus algarum*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	011111
32	*Harnischia fuscimana*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	011111
33	*Paratendipes albimanus*	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	011111
34	*Polypedilum* sp.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	011111
35	*Procladius choreus*	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	2	3	-	2	2	-	-	011111
36	*Stictochironomus histrio*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	011111
37	*T.punctipennis*	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-	-	-	-	011111
38	*Tanytarsus* sp.	1	-	-	-	-	-	-	-	2	-	-	-	-	-	2	-	-	2	-	-	2	-	100
39	*Dicrotendipes nervosus*	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	2	2	3	2	2	3	1010
40	*Stictochironomus rosenscholdi*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	-	1010
41	*Thienemannimyia* sp.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	2	-	1010
42	*Cricotopus* sp.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	1011
43	*Endochironomus impar*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	1011
44	*Cricotopus bicinctus*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	2	2	-	11
Уровни разбиений станций наблюдений		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
		0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1			
		0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1			
					0	0	0	0	0	1	1	1	1	1										
					0	0	1	1	1	0	0	0	1	1										

- с одной стороны искусственно завышается реальная встречаемость биологических объектов, т.к. рано или поздно при большом количестве проб в каждой точке могут встретиться все виды, характерные для региона;
- с другой стороны, повышается статистическая надежность результатов по сравнению с выводами, сделанными по одной, случайно взятой пробе [2].

Тем не менее, результаты обработки программой TWINSPAN исходной матрицы из осредненных показателей обилия по станциям р. Сок предоставляют возможность сделать, по нашему мнению, более интересные выводы, чем это продемонстрировали расчеты по р. Чапаевка (см. табл. 7.2).

Таблица 7.2

Результаты обработки по методу Браун-Бланке матрицы встречаемости подсемейств и трофических групп хирономид по станциям наблюдений на р. Сок

№ № пп	Таксоны хирономид		***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***	***K***	***L***	***M***	***N***	Уровни разбиения таксонов
	Подсемейства/ трибы	Трофические группы	Номера станций наблюдений														
			6	7	10	11	8	9	12	13	14	4	5	1	2	3	
1	Chironomini	Хищ.	2	-	-	-	2	2	3	2	3	2	2	-	-	1	00
2	Chironomini	Сест	-	-	2	-	-	-	2	2	3	2	-	-	-	-	00
3	ORTHOCLADIINAE	Детр	2	-	-	-	-	2	1	2	-	-	-	-	2	-	010
4	Chironomini	Фит/дет	-	-	-	2	-	2	3	3	3	2	-	2	-	2	010
5	Chironomini	Детр	2	2	2	-	2	2	2	2	2	3	2	2	2	2	01100
6	ORTHOCLADIINAE	Фит/дет	2	2	2	2	2	2	2	2	2	2	2	2	3	2	01100
7	TANYPODINAE	Хищ.	2	-	-	-	2	2	2	2	3	2	2	2	2	2	01101
8	Tanytarsini	Детр	2	-	2	-	2	2	2	2	3	3	2	2	2	2	01101
9	Chironomini	Дет/филь	3	-	-	-	2	3	3	3	3	3	-	3	-	3	01101
10	Chironomini	Соб/вс	-	-	-	-	2	2	2	3	3	2	2	3	2	2	0111
11	ORTHOCLADIINAE	Соб/вс	-	-	-	-	-	2	-	2	2	2	2	2	-	2	0111
12	PRODIAMESINAE	Хищ.	-	-	-	-	-	-	-	-	-	2	3	2	3	3	10
13	DIAMESINAE	Фит/дет	-	-	-	-	-	-	-	-	-	3	-	2	2	2	10
14	PRODIAMESINAE	Фит/дет	2	-	-	-	-	-	-	-	-	2	2	-	3	2	10
15	Tanytarsini	Сест	-	-	-	-	-	-	-	-	-	2	-	3	2	2	110
16	DIAMESINAE	Хищ.	-	-	-	-	-	-	-	-	-	-	-	2	-	-	111
17	PRODIAMESINAE	Сест	-	-	-	-	-	-	-	-	-	-	-	2	3	2	111
18	Tanytarsini	Дет/филь	-	-	-	-	-	-	-	2	-	-	-	2	2	2	111
19	Tanytarsini	Фит/дет	-	-	-	-	-	-	-	-	-	-	-	-	2	-	111
Уровни разбиений станций наблюдений			0	0	0	0	0	0	0	0	0	0	0	1	1	1	
			0	0	0	0	0	0	0	0	0	1	1				
			0	0	0	0	1	1	1	1	1						
							0	0	0	0	1						

Примечание: В таблице использованы следующие условные обозначения трофических групп: «Хищ.» - хищники хвататели; «Соб/вс» - всеядные собиратели+хвататели; «Детр» - детритофаги собиратели; «Сест» - сестонофаги+детритофаги фильтраторы; «Дет/филь» - детритофитофаги собиратели + фильтраторы; «Фит/дет» - фитодетритофаги собиратели.

Из анализа табл. 7.2 можно четко выделить несколько пустых и заполненных блоков, например, A12:I19 , L12:N19 и т.д. По критерию общей встречаемости групп хирономид легко можно ранжировать четыре кластера станций: {1, 2, 3} с максимальным разноообразием трофических групп, {4, 5}, {8, 9, 12, 13, 14} и, наконец, {6, 7, 10, 11}, где трофическое разнообразие минимально. Аналогичную типизацию (визуально или с применением уровней разбиения) нетрудно проделать и с таксономическими группами хирономид, выделив, например, малоинформационные группы 5, 6, 8 и группы 12-19, имеющие существенно узкий экологический диапазон.

[2] Авторы сами мучительно размышляют над этой ключевой проблемой применения статистики в гидробиологии и гидроэкологии, но не рискуют давать никаких рекомендаций ни себе, ни, тем более, читателям.

7.2. Задача о статистической связи: корреляционный анализ признаков и объектов

Формулировка задачи

Пусть имеется матрица наблюдений X размерностью $n \times m$, строки i которой соответствуют гидробиологическим пробам, $i = 1,2,\ldots,n$, а столбцы j содержат конкретные гидробиологические показатели, $j = 1,2,\ldots,m$, полученные в точке наблюдения i и выраженые в шкале произвольного характера.

Параметры многомерного объекта чаще всего связаны между собой, причем эта связь более или менее тесная. В большинстве случаев она проявляется в виде тенденции, т.е. равномерному увеличению одного из параметров в среднем будет сопутствовать пропорциональное увеличение или уменьшение другого. Необходимо количественно оценить меру стохастической связи между признаками, т.е. уровень их взаимной корреляции.

Рекомендуемая литература: [Василевич, 1969; Кэндалл, Стьюарт, 1973; Миркин, Розенберг, 1978; Плохинский, 1982; Зайцев, 1984].

Математический лист

Корреляционным анализом называется совокупность методов обнаружения корреляционной зависимости между случайными величинами или признаками. В разделе 5.5 нами уже приводилось выражение для коэффициента парной корреляции Пирсона r_{xy} как меры детерминации y и x. В разделе 6.2 коэффициент Пирсона рассматривался в контексте проблемы нелинейности и в сравнении с другими статистиками – корреляционными отношениями. В настоящем разделе осуществим естественное обобщение корреляционного анализа на многомерные объекты.

Геометрическая интерпретация

Числовые значения экспериментальных данных, получаемых при исследовании какого-нибудь конкретного явления, всегда зависят от принятого начала отсчета каждого параметра и от масштаба, в котором этот параметр измеряется. Начало отсчета и масштаб никак не связаны с сущностью явлений, поэтому естественно стремятся к представлению данных в некотором стандартном виде, сопоставимом для различных параметров всего массива измерений. С этой целью принято приводить параметры к так называемой центрированной и нормированной форме.

Центрирование параметра равносильно переносу начала координат в точку, соответствующую арифметическому среднему его значений. *Нормирование параметра* имеет целью представить параметр не в абсолютных единицах (градусах или мг/м^2), а в некоторых безразмерных единицах, харктеризующих отосительное значение признака. Выполним нормирование произвольного измерения j-го признака на i-м объекте по следующей формуле:

$$X_{ij} = \frac{x_{ij} - (\sum_{k=1}^{n} x_{kj})/n}{\frac{1}{n}\sqrt{\sum_{k=1}^{n} (x_{kj})^2}}, \qquad (7.1)$$

где n – количество измерений признака j. Признаки, нормированые по среднеквадратическим значениям, можно сравнивать, несмотря на их возможную физическую неоднородность.

Введем в рассмотрение n-мерное пространство объектов, где каждый признак будет отображаться точкой в многомерном нормированном пространстве с координатами $X_{1j}, X_{2j}, \ldots, X_{nj}$. Каждой точке будет соответствовать вектор, направленный из начала координат в эту точку (по математической традиции между ними не делается особенных различий).

Зададимся вопросом о том, какие отношения между векторами в пространстве объектов будут соответствовать корреляционной связи между признаками. Во-первых, следует обратить внимание на то, что все векторы в нормированном пространстве имеют одинаковую длину (модуль), равную $\sqrt{n}$. Во-вторых, если имеется два вектора X_1 и X_2 в n-мерном пространстве и $x_{11}, x_{21}, \ldots, x_{n1}$ и $x_{12}, x_{22}, \ldots, x_{n2}$ – проекции векторов на координатные оси, то косинус угла между этими векторами равен:

$$\cos(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n} X_{i1} \cdot X_{i2} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1})^2}\sqrt{\sum_{i=1}^{n}(x_{i2})^2}} = r(x_1, x_2) \quad , \qquad (7.2)$$

т.е. коэффициенту корреляции между признаками x_1 и x_2.

Если векторы совпадают по направлению, соответствующие им параметры связаны между собой наиболее тесно: косинус угла между векторами и коэффициент корреляции между параметрами равны +1. По мере увеличения угла между векторами связь между параметрами уменьшается и исчезает, когда векторы направлены под прямым углом друг к другу: косинус угла и коэффициент корреляции равны нулю. При дальнейшем увеличении угла между векторами степень связи вновь увеличивается, хотя приращения параметров имеют противоположные знаки. Когда векторы направлены противоположно друг другу, косинус угла и коэффициент корреляции равны –1.

Совокупность коэффициентов корреляции между всеми возможными парами признаков принято представлять в виде корреляционной матрицы $r(x_j, x_k)$; $j = 1,2,\ldots,m$; $k = 1,2,\ldots,m$, которая служит основным "сырьем" для работы многих алгоритмов многомерной статистики (например, в факторном анализе). Матрица симметрична относительно главной диагонали, поскольку $r(x_j, x_k) = r(x_k, x_j)$, а члены матрицы, стоящие на этой диагонали, равны 1, так как $r(x_j, x_j) = 1$.

Как рассматривалось в разделе 6.2, коэффициент линейной корреляции адекватен, если график бинарного отношения двух признаков хорошо аппроксимируется прямой линией. С другой стороны, на количественную оценку уровня связи влияет также закон распределения переменых. По этому поводу существует ряд противоречивых утверждений. Например, М. Кэндалл и А. Стьюарт [1973], а впоследствии и Г. Крамер [1975], утверждают, что формула линейного коэффициента корреляции не зависит от вида распределения и является применимой для большинства случаев закономерностей варьирования данных. Однако сложившаяся практика экспериментальной фитоценологии [Василевич, 1969; Миркин, Розенберг, 1978; Зайцев, 1984] показывает, что использование коэффициента корреляции Пирсона в качестве меры связи оправдано лишь тогда, когда совместное распределение пары признаков нормально или приближено к нормальному и когда встречаемость сравниваемых признаков одинакова (элиминация "d-эффекта", где d – параметр таблицы сопряженности 2×2). Коэффициенты ранговой корреляции Спирмена и Кендалла (см. раздел 5.6) менее чувствительны к законам распределения и отдельным выбросам значений признаков. Поэтому при анализе корреляционых зависимостей обоснованный положительный вывод предпочтительнее делать при наличии значимых корреляций, установленных *всеми* мерами связи.

Формы коэффициента корреляции для разных шкал измерений

Аналогами коэффициента корреляции Пирсона для оценки связи между признаками, измеренными в номинальной шкале, являются многочисленные формулы для мер сходства, описанные в разделе 4.7: коэффициенты Жаккара, Съеренсена, Рассела-Рао и т.п.

Для признаков, измеренных в порядковых шкалах, кроме ранговых коэффициентов корреляции Спирмена и Кендалла, ориентировочная оценка корреляционной связи может быть получена с использованием любого из многочисленных коэффициентов оценки зависимости признаков в таблицах сопряженности, описанных ранее в разделе 6.1 (например, полихорического коэффициента Чупрова). Для полноты изложения приведем также некоторые меры сопряженности (см. [Миркин и др., 1989]), используемые специально для таблиц ассоциативности 2×2 и имеющие смысл коэффициентов корреляции:

- *коэффициент сопряженности Бравайса* (он же – *показатель подобия Чупрова*)

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \quad ; \qquad (7.3)$$

- *коэффициент ассоциации Юла*

$$Q = \frac{ad - bc}{ad + bc} \quad ; \qquad (7.4)$$

где a, b, c, d – значения в клетках таблицы сопряженности 2×2.

Коэффициент корреляции Фехнера, используя количественные признаки, сводит, тем не менее, анализ связи к подсчету совпавших отклонений от арифметического среднего:

$$R_s = (C - H) / (C + H), \qquad (7.5)$$

где C – число совпадений знаков отклонений вариант от соответствующих средних, H – число несовпадающих знаков; $H = n - C$.

Анализ частных и множественных корреляций.

Коэффициенты частной корреляции оценивают "чистую" связь между двумя признаками, элиминируя возможную зависимость между ними за счет других признаков. Если $r(x_1, x_2)$ – коэффициент парной корреляции Пирсона между признаками x_1 и x_2, а $r(x_1, x_3)$ и $r(x_2, x_3)$ – подобные коэффициенты между x_1 и x_3 и x_2 и x_3, то коэффициент частной корреляции x_1 и x_2 при исключении зависимости от x_3 вычисляется по формуле:

$$r(x_1 x_2 : x_3) = \frac{r(x_1, x_2) - r(x_1, x_3) r(x_2, x_3)}{\sqrt{r(x_1, x_3)^2} \cdot \sqrt{r(x_2, x_3)^2}} \quad . \qquad (7.6)$$

Поскольку термины «связь» и «зависимость» имеют разный философский и статистический смысл, корреляционный анализ в принципе не предназначен для исследования причинно-следственных отношений между явлениями. Однако расчет коэффициентов частной корреляции позволяет косвенно оценивать и направленность взаимного влияния признаков. Например, если значение $r(x_1 x_2 : x_3)$ существенно превышает $r(x_1, x_2)$, то можно предположить, что признак x_3 выполнет роль "статистического шума", заглушающего связь между признаками x_1 и x_2. Бывают и обратные ситуации, когда коэффициент частной корреляции показывает, что связь между признаками на самом деле обусловлена тем, что они оба зависят от третьего признака и при элиминации его влияния связь пропадает. С использованием похожих формул можно получать *парциальные корреляции*, выравнивая выборки не только по одному, а по двум и более признакам.

Представляет также значительный интерес при изучении структуры экосистем исследование множественных корреляций, т.е. корреляций более чем между двумя признаками. Например, коэффициент множественной корреляции признака x_1 с двумя другими признаками x_2 и x_3 будет иметь по В.И. Василевичу [1969] следующий вид:

$$r(x_1 - x_2 x_3) = \sqrt{\frac{r(x_1, x_2)^2 + r(x_1, x_3)^2 - 2 r(x_1, x_2) \cdot r(x_1, x_3) \cdot r(x_2, x_3)}{1 - r(x_2, x_3)^2}} \quad . \qquad (7.7)$$

В этой формуле перед корнем следует всегда брать знак «+», следовательно, коэффициент множественной корреляции может принимать значения от 0 до 1.

Методы анализа корреляционной матрицы

На основании значений матрицы $r(x_j, x_k)$ парных коэффициентов корреляций могут делаться разнообразные выводы о природе связей между явлениями, зависящие от направления исследования и изучаемой предметной области. Например, можно отобрать наиболее близкие между собой биотопы или объединить в группы виды гидробионтов, имеющих сходную экологию (точнее, сходное распределение по станциям наблюдения). Следует отметить принципиальную эквивалентость техники Q- и R-анализа, не зависящей о того, подвергаются анализу сами объекты или их признаки: путем простого транспонирования матрицы исходных данных и перемены местами строк и столбцов легко можно заменить тип анализа на противоположный.

Для анализа корреляционной матрицы используется ряд эвристик. Например, Гопкинс [Gopkins, 1957, цит. по: Василевич, 1969], используя в качестве переменных фитоценотические виды, опубликовал методику выделения групп положительно коррелируемых признаков, которые он назвал "основными единицами" (basic unit). Для этого в качестве центров групп выделяются виды, имеющие отрицательные сопряженности, а затем к этим центрам добавляются виды, положительно коррелируемые с ними. В дальнейшем проводится объединение двух или нескольких групп, содержащих общие виды.

На анализе знаков коэффициентов корреляции основана также технология "концептуального моделирования" COMOD [Качанова, Фомин, 1997]: если у любых трех значимо коррелируемых признаков имеется только одна отрицательная связь или все три связи отрицательны, то в

этом случае нарушен знаковый баланс и эмпирические данные образуют "треугольник противоречий", что говорит о нецелостости системы и возможности ее членения на части. При всей внешней привлекательности концепции треугольника непротиворечивых корреляций как простейшего элемента структуры, вряд ли знаковый баланс является универсальным свойством экосистем, а чрезвычайная лабильность статистических связей гидробиологических показателей не позволят сводить анализ поведения биоценозов исключительно к теугольникам противоречий.

Как отмечалось в разделе 2.6, основной способ первичного анализа и визуализации корреляционной матрицы сводится к построению специальных графиков – дендрограмм или дендритов (графов "максимального корреляционного пути").

Наиболее простыми способами построения графических интерпретаций подобного типа являются метод "корреляционных плеяд" П.В. Терентьева [1959; цит. по: Выханду, 1964; Мандель, 1988] и "вроцлавская таксономия", разработанная польскими учеными Вроцлавского математического института [Florek and oth.,1950, цит. по: Василевич, 1969; Ястребов, 1991].

Алгоритм Терентьева осуществляет выделение сильно связанных групп признаков ("корреляционных плеяд") и сводится к следующему. Задается пороговое значение коэффициента корреляции r_0, с помощью которого производится построение срезов корреляционного цилиндра, из которых формируется последовательность подграфов, принимаемых в качестве "плеяд". Узлами этих подграфов являются все рассматриваемые признаки, а ребрами – корреляционные связи по абсолютной величине больше r_0. При последовательном уменьшении критического уровня, количество ребер увеличивается, плеяды становятся крупнее и начинают сливаться друг с другом. Окончательно выбирается порог r_0, скорее отвечающий эстетическим вкусам исследователя, чем каким-то формальным правилам.

Результатом вроцлавской таксономии является полный незамкнутый корреляционный путь, который можно отобразить в виде оптимального дерева – дендрита. Он представляет собой графическую структуру, состоящую из m вершин, соединенных между собой $(m - 1)$ ребрами так, что каждая вершина соединена хотя бы с одной другой. Если длину каждого ребра ассоциировать с величиной коэффициента корреляции r_{ij} между вершинами i и j, то оптимальный дендрит имеет максимальную сумму длин соединяющих отрезков из всех возможных. Если принять во внимание, что корреляционная мера по своему смыслу обратна мере дистанции, то граф максимального корреляционного пути идентичен "минимальному дендриту", т.е. дереву минимальной протяженности (minimum spanning tree по [Gower, Ross, 1969]).

Построение полного дендрита начинается с выбора двух наиболее сопряженных признаков, для чего в матрице коэффициентов корреляции определяется максимальное значение r_{ij}, $i \neq j$; признаки i и j образуют две первые вершины графа. Далее в строках i и j находится следующий наиболее сопряженный признак (для определенности – r_{jk}, где $k \neq j, i$), образующий новую третью вершину, соединенную с вершиной j. Далее ищется максимальное значение связи в строках k и j, кроме тех, что уже выбраны. Эта процедура повтояется до тех пор, пока не будут задействованы все признаки. При желании, задавшись пороговым значением r_0, можно полученный полный граф максимального корреляционного пути разбить на подграфы (плеяды), проводя разрыв между теми признаками, которые вошли в первоначальный граф со значением сопряженности меньшем r_0. Алгоритм вроцлавской таксономии полностью соответствует известному в кластерном анализе методу ближайшей связи (он же – метод "одиночного сцепления" по терминологии Р. Сокала и П. Снита). Это правило строит "волокнистые" кластеры, т.е. кластеры, сцепленные вместе только отдельными элементами, случайно оказавшимися ближе остальных друг к другу.

Оба метода имеют достаточно хорошую иллюстративность при любой размерности m корреляционной матрицы. Метод Терентьева, выделяющий все возможные внутриплеядные связи, более чувствителен к величине порога разбиения r_0 и при его снижении возможно лавинообразное загромождение графа малоинформативными ребрами, число которых стремится к $0.5(m\text{-}1)^2$ при $r_0 \to 0$. Количество ребер дендрита в этих условиях никогда не превышает $(m - 1)$, однако структура полученных кластеров сильно зависит от случайных флуктуаций корреляционной матрицы: при незначительных изменениях величины хотя бы одного коэффициента r_0 может произойти коренная перестройка всего графа (своего рода "баттерфляй-эффект").

Ряд исследователей полагают недостатком описанных методов визуализации связей *«недостаточную адекватность самого коэффициента Пирсона для выделения сопряженных групп видов и игнорирование доверительных интервалов выборочных статистик»* [Василевич, 1969].

Но, если отделить "мух от котлет", то становится очевидным, что сами алгоритмы группировки не имеют никакого отношения ни к способу расчета, ни к степени достоверности корреляционной матрицы. По известному выражению Т. Гексли *«математика есть только жернов и, засыпав плохое зерно, мы не можем получить хорошей муки»*. Это дело исследователя - провести необходимый отбор измерений, рассчитать исходную матрицу любым методом, который ему покажется более надежным и адекватным его целям (например, используя в качестве меры связи критерий χ^2 или иной другой индекс) и принудительно обнулить все коэффициенты, которые ему, по каким-то соображениям, покажутся недостоверными.

Результаты расчетов

Сформируем исходную таблицу наблюдений, выбрав в качестве признаков сочетание систематической (подсемейство или триба) принадлежности и трофической группировки видов хирономид – всего $m = 19$ переменных. В строки таблицы по всем имеющимся $n = 453$ наблюдениям зообентоса поместим значения $\ln(N_s)$, где N_s – суммарные значения численностей видов по выделенным группам водных организмов. Включим также три столбца с дополнительными переменными, отражающими условия взятия каждой пробы: температуру воды в придонном слое, глубину и ширину водоема в точке наблюдения.

Матрица парных коэффициентов корреляции Пирсона, рассчитанная по формуле (7.2) и оценивающая тесноту взаимной связи между всеми 22 исходными переменными, представлена в табл. 7.3. На основе матрицы $r(x_j, x_k)$ сформируем граф корреляционных плеяд по П.В. Терентьеву, представленный на рис. 7.1. В качестве первого среза агрегации плеяд коррелируемых признаков примем порог $r_0 = 0.31$.

На этом уровне детализации легко выделяются два следующих обособленных хирономидных комплекса:

- обширная группа, включающая почти всех представителей трибы Chironomini, к которым примкнули некоторые Orthocladiinae и хищники Tanypodinae;
- компактная и обособленная группа из всех таксонов Prodiamesinae.

Остальные таксономические группы хирономид мало скоррелированы друг с другом. Снижение порога агрегирования до $r_0 = 0.25$ принципиальных изменений в характер классификации не вносят (на рис. 7.1 вновь обозначившиеся связи отмечены пунктиром).

В целом внутри систематических групп прослеживаются более тесные связи, чем между трофическими группировками. Характерен положительный знак практически всех значимых коэффициентов корреляции между численностями гидробионтов. Это свидетельствует о том, что явление конкурирования за пищевые ресурсы мало распространено среди групп хирономид (что не исключает конкуренцию или замещение на видовом уровне). Наибольшая отрицательная корреляция ($r = -0.24$) между фильтраторами Orthocladiinae и детритофитофагами Chironomini вряд ли обусловлена серьезными экологическими причинами.

Корреляция между численностью гидробионтов и физико-гидрологическими условиями биотопа находится приблизительно на том же или более низком уровне, что и между самими гидробиологическими показателями.

Характерно, что такой параметр, как ширина реки вообще не оказывает никакого влияния на обилие зообентоса. Отрицательная корреляция глубины отбора пробы с численностью большинства групп хирономид не противоречит сложившимся представлениям. Интересным оказалась связь численности с температурой воды в придонном слое: в диапазоне измерений от 7 до 30 ^oC обилие зообентоса снижается с ростом температуры, что в каждом конкретном случае объясняется особенностями биологии организмов в вегетационный период.

Не претендуя на содержательное истолкование полученных расчетов, отметим, что численность хищников Chironomini (признак 1), температура (признак 20) и глубина отбора проб (признак 21) образуют "треугольник противоречий" по терминологии концептуального моделирования COMOD при одной отрицательной корреляционной связи. В этом случае частные коэффициенты корреляции после поочередной элиминации гидрофизических показателей оказываются больше соответствующих обычных коэффициентов Пирсона:

$$r_{1,20\,:\,21} = 0.227 > r_{1,20} = 0.196\ ; \quad r_{1,21\,:\,20} = -0.135 > r_{1,21} = -0.068$$

Таблица 7.3

Матрица парных коэффициентов корреляции Пирсона, рассчитанная по численностями подсемейств и трофических групп хирономид (жирным шрифтом отмечены корреляции, значимые при $p < 0.05$)

	Признаки	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	22	23
1	Хищ. ChC	1.00																					
2	Хищ. Di	-0.09	1.00																				
3	Хищ. Or	-0.06	**0.10**	1.00																			
4	Хищ. Pr	**-0.17**	**0.20**	-0.03	1.00																		
5	Хищ. Tn	**0.21**	-0.03	-0.04	-0.02	1.00																	
6	Вс/Соб ChC	**0.30**	**-0.10**	-0.06	**-0.17**	**0.35**	1.00																
7	Вс/Соб Or	0.06	0.02	-0.03	**-0.10**	0.06	**0.13**	1.00															
8	Детр ChC	**0.38**	**-0.11**	-0.02	**-0.10**	**0.34**	**0.38**	0.06	1.00														
9	Детр ChT	0.06	0.04	-0.01	0.09	**0.31**	**0.23**	**0.13**	**0.32**	1.00													
10	Детр Or	-0.01	-0.02	-0.02	-0.05	**0.16**	**0.09**	-0.07	**0.17**	**0.19**	1.00												
11	Сест ChC	**0.15**	-0.07	0.01	**-0.16**	**0.13**	**0.31**	**0.34**	**0.16**	**0.13**	-0.03	1.00											
12	Сест ChT	-0.08	**0.19**	**0.24**	**0.18**	-0.02	**-0.10**	-0.05	-0.02	0.05	-0.03	-0.05	1.00										
13	Сест Pr	**-0.09**	0.04	-0.02	**0.37**	-0.04	**-0.10**	-0.04	**-0.11**	0.04	-0.03	**-0.11**	0.03	1.00									
14	деТ/филь ChC	**0.25**	**-0.12**	0.03	**-0.20**	**0.28**	**0.47**	**0.13**	**0.33**	**0.10**	0.06	**0.18**	**-0.13**	**-0.13**	1.00								
15	деТ/филь ChT	**-0.11**	**0.30**	-0.02	**0.23**	0.04	-0.08	0.02	-0.05	0.02	-0.03	-0.07	-0.01	**0.12**	-0.06	1.00							
16	Фит/дет ChC	**0.22**	-0.03	-0.04	-0.08	**0.15**	**0.18**	-0.06	**0.29**	**0.17**	**0.10**	0.09	-0.06	-0.05	**0.24**	-0.06	1.00						
17	Фит/дет Di	**-0.13**	**0.20**	-0.02	**0.15**	-0.00	**-0.12**	-0.02	**-0.11**	**0.14**	0.00	-0.08	-0.03	0.05	**-0.11**	**0.14**	-0.03	1.00					
18	Фит/дет Or	**-0.20**	**0.14**	-0.01	**0.23**	-0.04	**-0.14**	0.08	0.00	**0.40**	**0.16**	**-0.10**	**0.18**	**0.24**	**-0.24**	**0.14**	-0.04	0.22	1.00				
19	Фит/дет Pr	-0.07	-0.03	-0.02	**0.36**	0.01	**-0.11**	-0.09	0.06	0.09	0.08	-0.08	**0.16**	**0.20**	**-0.10**	0.04	-0.03	-0.02	0.16	1.00			
20	Темп. дна	**0.20**	**-0.22**	**-0.11**	**-0.32**	0.07	**0.27**	0.05	**0.27**	-0.07	0.05	**0.14**	**-0.19**	**-0.20**	**0.29**	**-0.14**	**0.21**	**-0.25**	**-0.40**	**-0.16**	1.00		
21	Глубина	-0.07	**-0.10**	0.01	**-0.18**	0.08	0.01	**-0.11**	-0.05	**-0.26**	-0.05	-0.04	**-0.09**	**-0.12**	**0.12**	-0.09	-0.08	**-0.14**	**-0.36**	-0.09	**0.30**	1.00	
22	Ширина	0.03	-0.08	-0.04	**-0.17**	-0.01	0.06	-0.01	-0.05	**-0.22**	-0.04	-0.05	-0.08	**-0.10**	0.06	-0.08	0.01	**-0.11**	**-0.22**	**-0.10**	**0.30**	**0.50**	1.00
		X ChC	X Di	X Or	X Pr	X Tn	B ChC	B Or	Д ChC	Д ChT	Д Or	C ChC	C ChT	C Pr	T ChC	T ChT	Ф ChC	Ф Di	Ф Or	Ф Pr	Темп.	Глуб.	Шир.

Примечание: В таблице использованы следующие условные обозначения трофических групп:
«Хищ.» («Х») - хищники хвататели; «Вс/Соб » («В») - всеядные собиратели+хвататели; «Детр» («Д») - детритофаги собиратели; «Сест» («С») - сестонофаги+детритофаги фильтраторы; «деТ/филь» («Т») - детритофитофаги собиратели + фильтраторы; «Фит/дет» («Ф») - фитодетритофаги собиратели.
Условные обозначения семейств и триб: Or - Orthocladiinae, Tn - Tanypodinae, Di - Diamesinae, Pr - Prodiamesinae, Ch – Chironominae (ChC - Chironomini, ChT – Tanytarsini)

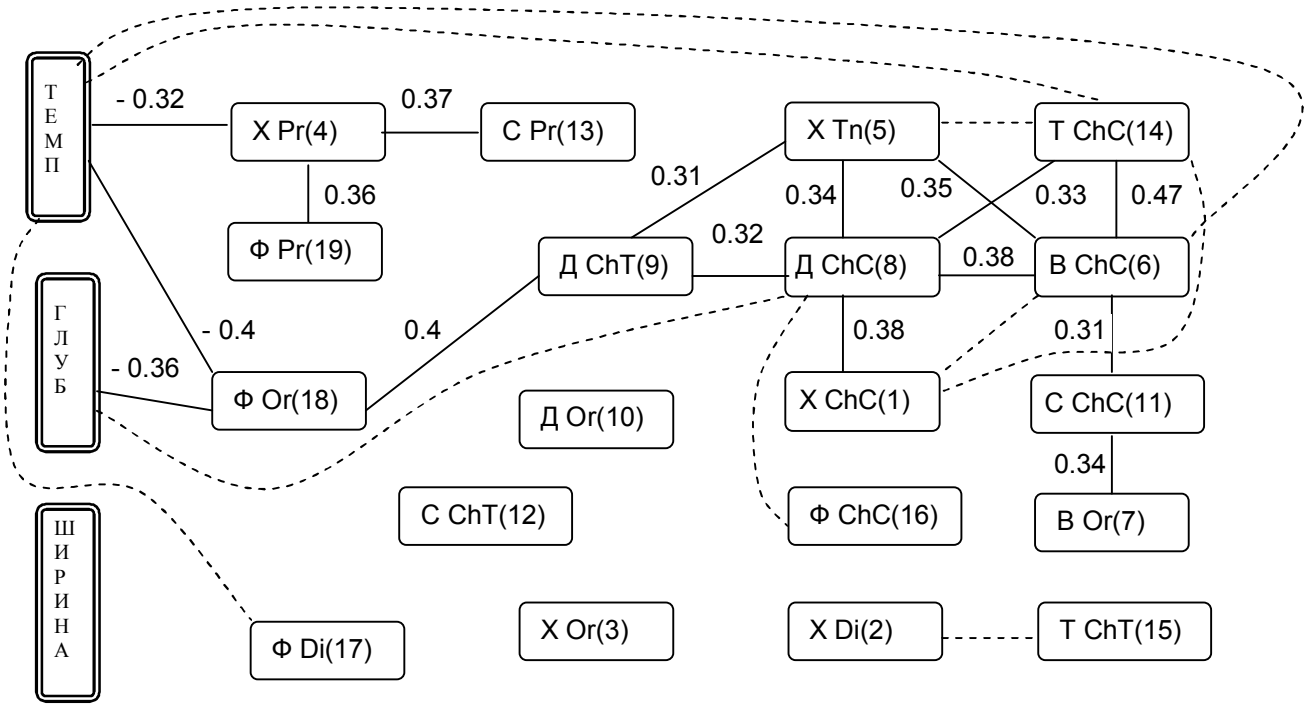


Рис. 7.1. Корреляционные плеяды Терентьева, выделенные из матрицы парных коэффициентов корреляции Пирсона (данные и условные обозначения – из табл. 7.3)

В аналогичной ситуации фитодетритофаги собиратели Orthocladiinae (признак 18) с теми же переменными образуют непротиворечивый треугольник с двумя отрицательными связями и парциальные корреляции оказываются меньше полных:

$$\boldsymbol{r}_{18,20\,:\,21} = -0.326 < \boldsymbol{r}_{18,20} = -0.397\ ; \qquad \boldsymbol{r}_{18,21\,:\,20} = -0.273 < r_{18,21} = -0.357.$$

Множественные коэффициенты корреляции этих гидробиологических признаков с обоими гидрофизическими факторами в обоих случаях увеличиваются:

$$\boldsymbol{r}_{1\,-\,20,21} = 0.236\ ; \qquad \boldsymbol{r}_{18\,-\,20,21} = = 0.469.$$

Результаты вроцлавской таксономии по той же корреляционной матрице представим в виде минимального дендрита – графа максимального корреляционного пути на рис. 7.2. Анализируя полученный граф, можно отметить, что при пороговом значении $\boldsymbol{r}_o = 0.29$ все множество таксономических единиц хирономид распадается на 4 индивидуальных элемента и 3 группы, качественно совпадающие с корреляционными плеядами Терентьева. Связи с коэффициентом корреляции менее 0.29 представлены на рис. 7.2 пунктиром.

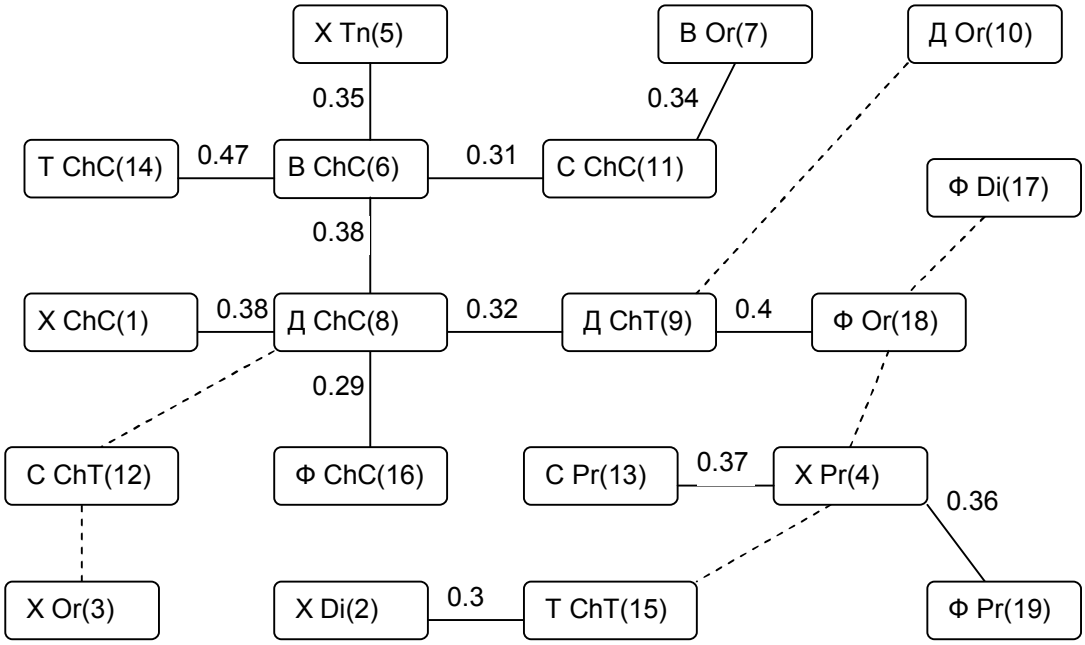


Рис. 7.2. Минимальный дендрит – граф максимального корреляционного пути (данные и обозначения – из табл. 7.3)

Поскольку ранее нами отмечалось, что дендрит и дендрограмма – визуальное отображение одной сущности, представим на рис. 7.3 те же результаты в виде дендрограммы. Т.к. для последней традиционно использование *мер расстояния*, а коэффициент корреляции Пирсона ***r*** имеет смысл *меры сходства*, то обычно проводят построение дендрограммы в инверсной шкале (1 –abs(***r***)), где значения ***r*** берутся по абсолютной величине.

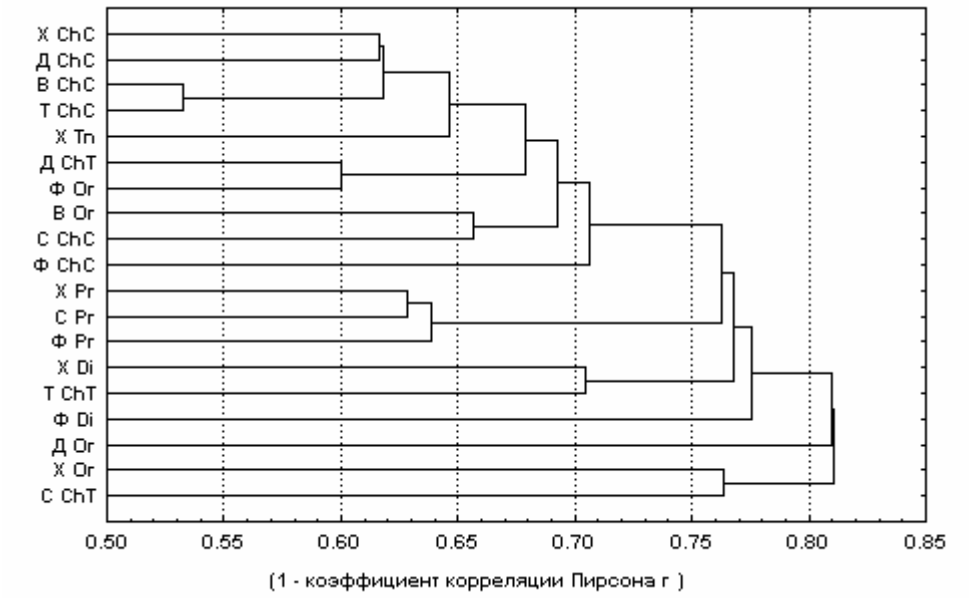


Рис. 7.3. Дендрограмма группировки признаков по методу вроцлавской таксономии (данные и обозначения – из табл. 7.3)

7.3. Задача о разбиении на группы: кластерный анализ

Формулировка задачи

Пусть имеется матрица наблюдений X размерностью $n \times m$, строки i которой соответствуют гидробиологическим пробам, $i = 1,2,\ldots,n$, а столбцы j содержат конкретные гидробиологические показатели, $j = 1,2,\ldots,m$, полученные в точке наблюдения i и выраженые в шкалах измерений произвольного характера.

Если эти данные понимать как точки в признаковом пространстве, то задача кластерного анализа формулируется как выделение "сгущений точек" и разбиение исходной совокупности на однородные подмножества объектов. Кластерный анализ можно рассматривать также как метод редукции (сжатия) некоторого множества данных в более компактную классификацию объектов.

Рекомендуемая литература: [Айвазян с соавт., 1974; Миркин, Розенберг, 1978; Дюран, Оделл, 1980; Классификация и кластер.., 1980; Жамбю, 1988; Мандель, 1988].

Математический лист

Кластер определяется, как совокупность точек, лежащих на расстоянии не больше, чем r от некоторого "центра тяжести" в m-мерном пространстве (внутри гиперсферы радиуса r или гиперкуба со сторонами $2r$).

В литературе описывается множество различных методов кластеризации, основанных на использовании матриц сходства, оценивании функций плотности статистического распределения, эвристических алгоритмах перебора, идеях математического программирования и др. Большая часть этих алгоритмов, при всей их несхожести, методически основаны на одной предпосылке – *гипотезе компактности*, т.е. *«в используемом пространстве признаков измерения, принадлежащие одному и тому же классу, близки между собой, а измерения, принадлежащие разным классам хорошо разделимы друг от друга»* [Кольцов, 1989].

Рассмотрим некоторые алгоритмы, основанные на использовании меры расстояния между объектами D. Введение метрики m-мерного пространства (т.е. способа оценки расстояний) является естественным приемом квантификации свойства схожести объектов: чем ближе между собой объекты в данной метрике, тем они более сходны и наоборот. Без этого само понятие «кластер» во многом теряет смысл, поэтому алгоритмы кластерного анализа часто формулируют в терминах дистанций.

Был предпринят ряд попыток разработать аксиоматический подход к введению метрических мер, согласно которым, например, расстоянием D называется двухместная действительная функция $D(x_1, x_2)$, обладающая следующими свойствами:

- $D(x_1, x_2) \geq 0$ – неотрицательная определенность расстояния (хотя тот же коэффициент корреляции Пирсона принимает и отрицательные значения);
- $D(x_1, x_2) = 0$ тогда и только тогда, когда $x_1 = x_2$ – неразличимость тождественных объектов;
- $D(x_1, x_2) = D(x_2, x_1)$ – симметричность расстояния (хотя в разделе 4.7 приводятся примеры несиметричных мер);
- $D(x_1, x_2) + D(x_2, x_3) \geq D(x_1, x_3)$ – неравенство треугольника (длина любой стороны треугольника не больше суммы длин двух оставшихся).

Более конкретная математическая формулировка не имеет однозначного смысла, поскольку разные субъекты вкладывают в эту аксиоматику неодинаковое содержание. Проблемы выбора конкретных выражений для мер близости или расстояния между объектами подробно обсуждались нами в разделе 4.7.

Пусть мы имеем симметричную матрицу расстояний D между объектами исходной матрицы наблюдений:

$$\mathbf{D} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \cdots & 0 \end{pmatrix}$$

Компоненты матрицы **D** могут быть рассчитаны с использованием любой из перечисленных выше концепций или формул, что не имеет значения для работы собственно алгоритмов таксономии.

Наиболее распространенную группу эвристических методов кластеризации составляют методы, основывающиеся на иерархической агломеративной процедуре (от латинского agglomero – присоединяю, накапливаю). Эти алгоритмы дают лишь условно-оптимальное решение в некотором подмножестве локальных разбиений (кластеров), однако достоинством этих методов является простота вычислений и интерпретации полученных результатов. Смысл иерархический агломеративной процедуры заключается в следующем. Перед началом кластеризации все объекты считаются отдельными кластерами, т.е. имеется $p = n$ кластеров, каждый из которых включает по одному элементу. На первом шаге алгоритма определяются два наиболее близких или сходных объекта, которые объединяются в один кластер, общее количество которых сокращается на 1 ($p \to p$ - 1). Итеративный процесс повторяется, пока на последнем (p - 1)-м шаге все классы не объединятся. На каждом последующем шаге агломеративной процедуры требуется пересчет лишь одной строки и одного столбца матрицы D, т.е. рассчитываются расстояния от образованного кластера до каждого из оставшихся кластеров.

Процедура иерархического кластерного анализа предусматривает возможность группировки как объектов (строк матрицы данных), так и переменных (столбцов). В последнем случае роль объектов кластеризации играют признаки исходной мадрицы:, например, виды гидробионтов.

Использовать построенную дендрограмму для выделения того или иного количества отдельных кластеров можно путем "разрезания" этой дендрограммы на определенном значении шкалы D. Фактически это означает, что мы проводим горизонтальную линию, рассекая дерево связей в том месте, где наблюдается максимальный скачок в изменении межкластерного расстояния.

Для определения расстояния между произвольной парой кластеров $\{X_i\}$, $i = 1,\ldots k_1$ и $\{Y_j\}$, $j = 1,\ldots k_2$ с использованием различных версий алгоритмов классификации были сформулированы следующие подходы:

- метод «*одиночной связи или минимального локального расстояния*» (single linkage), знакомый нам по "вроцлавской таксономии", когда для включения объекта в кластер требуется максимальное сходство всего лишь с одним членом кластера;
- метод «*полной связи или максимального локального расстояния*» (complete linkage), когда последовательность сцепления между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями");
- метод «*средней связи Кинга или попарного арифметического среднего*» (unweighted pair-group method using arithmetic averages), где мера сходства между "кандидатом" и членами кластера устанавливается как арифметическое среднее

$$D_3 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} d(X_i, Y_j)/k_1 k_2 \quad . \tag{7.8}$$

Выделяется также совокупность методов, использующих статистические расстояния между кластерами (*метод групповых средних*, *центроидный метод*, *метод Уорда* и т.д.), где предполагается объединение, приводящее к минимизации суммы квадратов отклонений между каждым объектом и центром кластера, содержащим этот объект. Например, в методе Уорда [Ward, 1963] используется мера:

$$D_4 = \frac{k_1 k_2}{k_1 + k_2}(\overline{X} - \overline{Y})^T(\overline{X} - \overline{Y}), \quad \text{где} \quad \overline{X} = \sum_{i=1}^{k_1} x_i /k_1; \quad \overline{Y} = \sum_{j=1}^{k_2} y_j / k_2 \quad . \tag{7.9}$$

Кроме иерархических методов классификации большое распространение получили также различные итерационные процедуры, которые пытаются найти наилучшее разбиение, ориентируясь на заданный критерий оптимизации, не строя при этом полного дерева. В начале последовательных итераций в качестве центра выбирается один из элементов и формируется кластер из элементов, удаленных от него не далее чем на r. Далее процедура повторяется для остальных элементов, причем в качестве очередного центра выбирается, например, "типическая" точка – лежа-

щая на минимальном расстоянии от центра оставшегося множества объектов. После выполнения очередного шага выясняется, достигнуто ли желательное разбиение. Существуют различные методы определения критерия остановки процедуры:

- получено определенное заранее количество кластеров;
- все кластеры содержат более определенного числа элементов;
- кластеры обладают требуемым соотношением внутренней однородности и разнородности между собой.

На первом условии основывается наиболее популярный алгоритм – *метод k-средних Мак-Кина* [Фрей, 1967], в котором сам пользователь должен задать искомое число конечных кластеров, обозначаемое k. Принцип классификации заключается в следующем:

- выбираются или назначаются k наблюдений, которые будут первичными центрами кластеров;
- остальные наблюдения приписываются к ближайшим заданным кластерным центрам;
- текущие координаты первичных кластерных центров заменяются на кластерные средние;
- предыдущие два шага повторяются до тех пор, пока изменения координат кластерных центров не станут минимальными.

Наиболее важным свойством, используемым при анализе, является плотность распределения объектов внутри кластеров. Это свойство дает нам возможность определить кластер в виде скопления точек в многомерном пространстве, относительно более плотного по сравнению с иными областями этого пространства, которые либо вообще не содержат точек, либо содержат малое количество наблюдений. Несмотря на достаточную очевидность этого свойства, однозначного способа вычисления такого показателя плотности не существует. Наиболее удачным показателем, характеризующим компактность "упаковки" многомерных наблюдений в данном подмножестве, является дисперсия расстояния от центра кластера до отдельных его точек.

Другим примером критерия однородности может быть, например, функция, описанная А.А. Дорофеюком [1971]:

$$I_m = \frac{1}{m}\sum_{i=1}^{m} D(p,p) - \frac{2}{m(m-1)}\sum_{q=1}^{m-1}\sum_{p>q} D(p,q) \quad \to \max , \tag{7.10}$$

значения которой подсчитываются для всех возможных вариантов разбиения исходного множества на m классов. В этой формуле $D(p, p)$ – среднее сходство между собой всех векторов, попавших в одну группу, а $D(p, q)$ – среднее сходство по всем парам векторов из разных групп p и q:

$$D(p,p) = \frac{2}{n_p(n_p-1)}\sum_{x=1}^{n_p-1}\sum_{y>x} D(x,y) \qquad D(p,q) - \frac{1}{n_p n_q}\sum_{x\in q}\sum_{y\in p} D(x,y) , \tag{7.11}$$

где n_p и n_q – число элементов в группах p и q.

Результаты расчетов

Основной задачей классификации в гидробиологических исследованиях является установление отношений сходства между станциями наблюдений, створами, участками рек и целыми водоемами. В общем случае, на нижнем уровне этой иерархии – станции – может быть сделано несколько проб, составляющих некоторый статистический "образ" водного объекта.

Выполним в терминах меры сходства анализ воспроизводимости повторяющихся измерений, сделанных в разное время в окрестности одной географической точки. Для каждой малой реки Самарской области рассчитаем по формулам (7.11) А.А. Дорофеюка два показателя:

- среднее внутригрупповое расстояние измерений, сделанных на одной станции $D(p, p)$;
- среднее межгрупповое расстояние проб, полученных на разных станциях наблюдения $D(p,q)$.

Воспользуемся следующей метрикой для оценки сходства между измерениями, которая, по нашему субъективному мнению, наилучшим образом отражает отношения сходства между измерениями в многомерном пространстве видов.:

- для численности N_i и биомассы B_i каждого вида в пробе рассчитаем значение

$$X_i = \ln((N_i * B_i)^{0.5}) ;$$

- пронормируем полученные величины на интервале от 0 до 1 по формуле

$$Y_i = (\max X_i - X_i) / (\max X_i - \min X_i) ;$$

- используем в качестве меры расстояния между пробами $D(x_i, x_k)$ манхэттенское расстояние, равное сумме разностей нормированных $\ln((N_i * B_i)^{0.5})$, вычисленных по всем видам – см. формулу (4.32) из раздела 4.7.

Результаты расчетов, представленные в табл. 7.4, показывают, что изменчивость видового состава и показателей обилия в измерениях, взятых внутри одного биотопа, весьма велика в зависимости от даты наблюдения, конкретной точки отбора проб и прочих факторов. При этом отсутствуют статистически значимые отличия в уровне внутригрупповой вариации манхеттенского расстояния между пробами одной и той же станции и вариацией этого показателя в пробах, взятых на разных станциях той же реки. Иными словами, все пробы, взятые в пределах одной реки из перечисленных в таблице 7.4, принадлежат к одной генеральной совокупности измерений.

Таблица 7.4

Среднее количество проб, взятых на одной станции наблюдения, количество пар сравниваемых между собой измерений N_D, среднее манхэттенское расстояние M_D и его доверительный интервал E_D (отдельно для проб, относящихся к одной и той же станции и к разным станциям)

Наименование реки	Проб на станции	Внутригрупповое среднее			Межгрупповое среднее		
		N_D	M_D	E_D	N_D	M_D	E_D
Чапаевка (верховья)	10.8	812	11.82	± 0.40	7573	13.63	± 0.14
Чапаевка (низовья)	12.9	647	5.73	± 0.31	4106	6.08	± 0.11
Сок	6.8	365	14.95	± 0.63	3913	15.95	± 0.20
Байтуган	3.9	47	18.93	± 1.96	388	16.96	± 0.72
Маза	2.6	20	10.87	± 1.56	170	11.34	± 0.56
Тайдаков	2.6	10	10.42	± 3.64	45	13.25	± 1.29
Муранка	2.5	20	11.20	± 2.93	151	11.35	± 0.98

К аналогичным выводам можно прийти, сравнив вариационный ряд дистанций между парами измерений, принадлежащих одной реке, с межгрупповыми мерами расстояния проб из разных рек. Матрицы исходных данных, представленные пробами из разных водных объектов, в нашем случае оказались очень трудно статистически разделимы.

В дополнение к этому, если ставится задача классифицирования рек или станций наблюдений, то следует принять во внимание следующие малоприятные обстоятельства, связанные со спецификой детерминационного кластерного анализа:

- невозможно использование полного объема повторяющихся данных, т.к. метод работает только с непересекающимися и необъединяемыми строками исходной матрицы наблюдений;
- поскольку отсутствуют методики адекватного "усреднения" проб одной станции, для проведения классификации объектов необходимо по каждому из них отобрать только одну пробу, которая, по субъективному мнению исследователя, наилучшим образом отражает гидробиологические особенности данного биоценоза;
- если для класифицируемого объекта (станции, створа, участка, реки и т.д.) сделано более, чем одно измерение, то остальные пробы будут игнорироваться.

Таким образом, на этапе подготовки данных для кластерного анализа исследователь оказывается в весьма затруднительном положении: какие измерения следует выбрать за "опорные", наилучшим образом характеризующие каждый классифицируемый объект, принимая во внимание, что от этого выбора будет сильно зависеть конечный результат группировки.

Выберем из 97 измерений, сделанных на 14 станциях наблюдения р. Сок, по одной пробе для каждой станции, отобрав их из общего множества примеров по критерию максимального биологического разнообразия (наибольшему количеству видов). Общее количество видов зообентоса, которое встретилось в этих пробах, составило 155.

Выполним кластерный анализ участков реки с использованием различных методов объединения и мер расстояния:

- евклидова расстояния в пространстве показателей обилия 155 видов, рассчитанных по формуле $\ln((\boldsymbol{N}*\boldsymbol{B})^{0.5})$;
- евклидова расстояния с использованием показателя обилия, пронормированного от 0 до 1 по максимальному размаху для каждого вида;
- меры сходства по Съеренсену.

Полученные дендрограммы, представленые на рис. 7.4, свидетельствуют о том, что при выполнении кластерного анализа исследователь находится в тяжелом комбинаторном положении, будучи поставлен перед необходимостью выбора не только комплекта исходных данных, но также метрики расстояния и алгоритма объединения. Например, для тех же 14 классифицируемых станций р. Сок можно использовать не менее 5 общеупотребимых формул для матрицы сходства и не менее 5 широко распространенных методов построения иерархической классификации. В результате для 14 объектов мы получаем 25 возможных вариантов разбиений, т.е. деревьев, в разной степени отличающихся друг от друга. В итоге неопределенность исходных данных подменяется другой, еще более туманной – неопределенностью результатов классификаций.

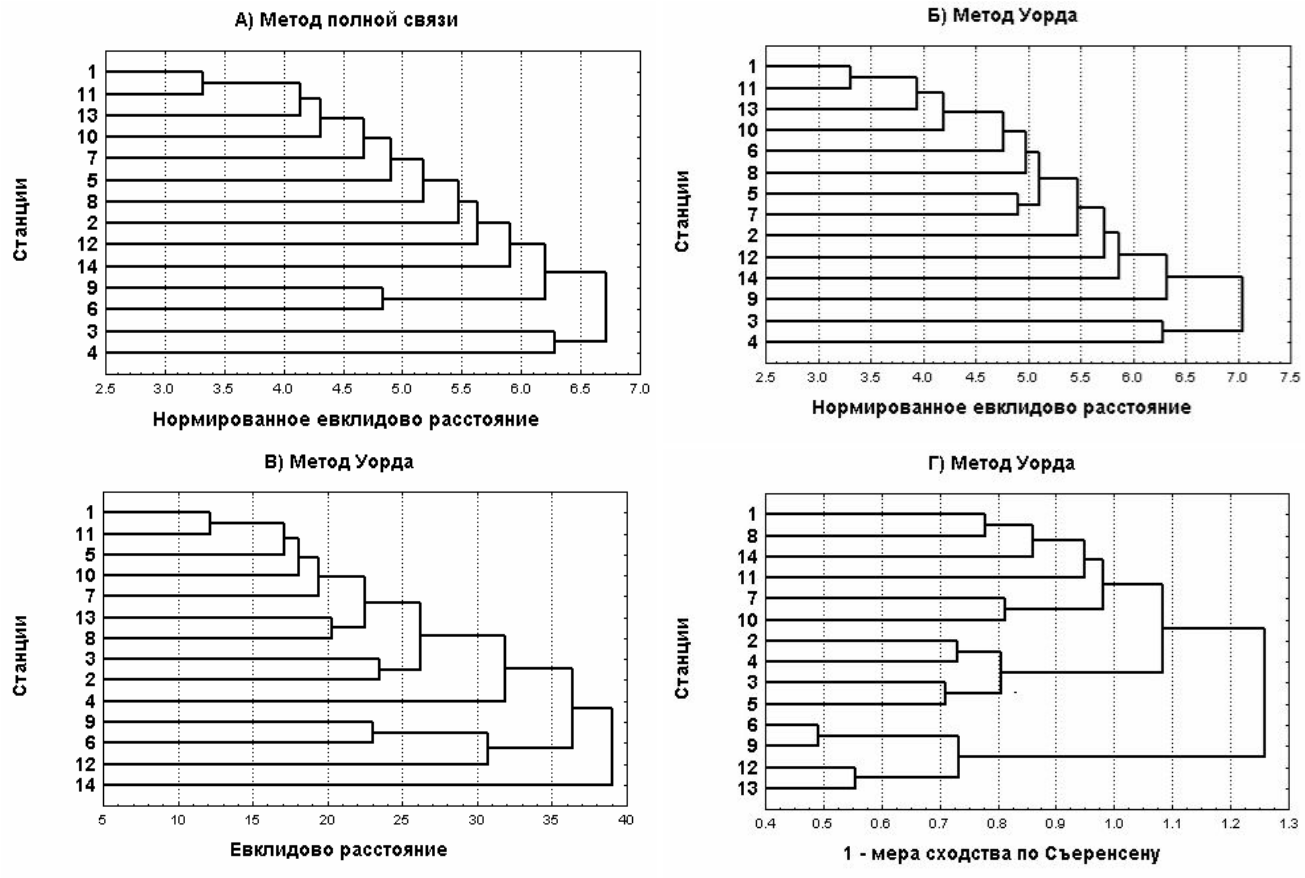


Рис. 7.4. Дендрограммы классификации станций наблюдения р. Сок по пробам зообентоса, выполненные с использованием различных методов и мер расстояний

Существуют некоторые, в разной степени логически размытые рекомендации по выбору технологии расчетов. Например, считается, что более изощренные центроидный метод или алгоритм Уорда приводят к хорошим результатам, хотя при сравнении фиг. «А» и «Б» рис. 7.4 серьезных изменений не обнаруживается.

Евклидово расстояние зависит только от нескольких "доминирующих" разностей, поскольку они возводятся в квадрат, и практически игнорирует длинный "хвост" небольших расхождений. Если это расстояние основывается на натуральных показателях обилия, то его величина определяется 2–3 видами с наибольшей численностью или биомассой. Если, однако, используются нормированные исходные данные, то на первый план в формировании расстояния выдвигаются

редкие виды вне зависимости от их абсолютной численности. Напомним, что мера сходства по Съеренсену вообще не учитывает обилия, а только факт наличия или отсутствия каждого вида бентоса.

Любой класс (таксон, кластер), состоящий из некоторого подмножества реальных объектов и полученный по технологии "без учителя" – всегда некоторая умозрительная теоретическая конструкция, созданная из субъективных предположений или на основе эвристических алгоритмов, качество которой принципиально невозможно точно измерить. Сказать, например, что классификация «Г» лучше, чем «В», может только коллектив компетентных экспертов, основывающийся, чаще всего, не на формальных критериях, а на интуитивном опыте. Правда, один из авторов [Розенберг, 1975] предлагал в качестве формализации такого сравнения вычисление *меры диссонанса* полученной классифицированной матрицы (напрмер, матрицы сходства, упорядоченной по последовательности «Г») от "случайно перемешанной" матрицы (т.е. матрицы, упорядоченной по случайной последовательности станций); чем больше мера диссонанса, тем лучше классификация. Этот эвристический прием был весьма эффективно использован при анализе мозаичности травянистых растительных сообществ на ценотическом уровне [Миркин, Розенберг, 1977а].

Нетривиальной является и задача оценки близости (предупорядоченности) двух произвольных деревьев, состоящих из одних и тех же "листьев". Чтобы дать точный ответ, значимо ли отличаются друг от друга варианты классификаций, следует провести, своего рода, кластерный анализ результатов кластерного анализа [Миркин, Черный, 1970; Наумова, 1979]. Например, В.Л. Андреев [1979а] рекомендует использовать в этом случае графо-аналитические методы, известные под общим названием "поиск лидера" и основанные на ранговой корреляции последовательностей агрегируемых элементов, однако, подробное рассмотрение этой задачи выходит за рамки нашей монографии.

7.4. Оценка различий многомерных комплексов наблюдений

Формулировка задачи

Пусть в матрице наблюдений $\boldsymbol{X}$ каждый $\boldsymbol{j}$-й столбец, $\boldsymbol{j} = 1, 2,\ldots, \boldsymbol{m}$, представлен гидробиологическим показателем, а каждая $\boldsymbol{i}$-я строка, $\boldsymbol{i} = 1,2,\ldots,\boldsymbol{n}$, описывает гидробиологическое измерение, выполненное в некотором пространственно-временном аспекте. Предположим, что каждой строке $\boldsymbol{i}$ поставлен в соответствие некоторый качественный признак (фактор), на основании которого общую многомерную выборку можно сгруппировать в частные независимые случайные выборки: если фактор имеет $\boldsymbol{p}$ уровней $\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_p}$, то при каждом уровне $\boldsymbol{A_k}$ фактора, $\boldsymbol{k} = 1,...,\boldsymbol{p}$, имеется $\boldsymbol{n_k}$ измерений. Необходимо проверить, насколько статистически значимо различаются между собой блоки матрицы $\boldsymbol{X}$, относящиеся к разным уровням фактора $\boldsymbol{A}$, и количественно оценить меру таких различий.

Как и в случае дисперсионного анализа, фактор, который оказывает влияние на количественные результаты измерений, имеет принципиально нечисловую природу и может соответствовать географическому объекту, где была взята проба, категории водоема или сезонному периоду.

Рекомендуемая литература: [Урбах, 1963; Дуда, Харт, 1978; Горелик, Скрипкин, 1984; Андреев, 1979а]

Математический лист

Статистические методы

Использование детерминистических методов кластеризации, описанных в разделе 7.3, оправдано в некоторых частных случаях, когда по условию задачи нужно получить простые решения или невозможно использовать вероятностные методы. Преемущества последних заключается в том, что они допускают наличие ошибок и неполноту знаний о сравниваемых объектах и оперируют с плотностью распределения вероятностей переменных.

В общем случае параметрические методы анализа многомерных наблюдений, принадлежащих к разным классам, основываются на формуле условных вероятностей, предложенной в конце XVIII в. малоизвестным священником и прекрасным английским математиком Томасом Байесом:

$$\boldsymbol{P}(\boldsymbol{x})\ \boldsymbol{P}(\boldsymbol{A_k}/\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{A_k})\ \boldsymbol{P}(\boldsymbol{x}/\boldsymbol{A_k}) \ , \qquad (7.11)$$

где $P(x)$ – функция распределения плотности вероятности всех данных в m-мерном пространстве независимо от того, к какому классу они принадлежат; $P(A_k)$ – вероятность наблюдения класса A_k; $P(A_k/x)$ – условная вероятность того, что вектор x принадлежит классу A_k; $P(x/A_k)$ – условная вероятность получения для класса A_k вектора данных x.

Формула Байеса позволяет вычислить вероятность справедливости некоторой гипотезы на основании принятых априорных вероятностей. Этот метод в строгом смысле оправдан, если альтернативные гипотезы (в данном случае принадлежность измерения к некоторому классу) основываются на функциях плотности вероятностей, т.е. известны законы распределения случайной величины, которые могут быть оценены по эмпирическим данным.

Как отмечалось выше, в принятой модели данных – таблице наблюдений X – совокупность n измерений может быть представлена "облаком" n точек в m-мерном пространстве, которое в целом можно охарактеризовать положением и степенью компактности. В большинстве методов многомерного анализа предполагается, что имеет место нормальное распределение случайной величины x_{ij}. Это связано с тем, что центральная предельная теорема для одной величины распространяется и на многомерный случай, т.е. последовательность сумм независимых одинаково распределенных случайных векторов сводится к многомерному нормальному распределению $N(\mu, \Sigma)$ с вектором $(m \times 1)$ из средних μ и матрицей $(m \times m)$ ковариаций Σ.

Для определенности рассмотрим задачу разделения двух множеств объектов (множества "Класс 1" и множества "Класс 2"). Очевидно, что два множества будут разделяться тем лучше, чем больше расстояние между их центрами. Кроме того, задача разделения множеств упрощается при условии сужения диаметров этих множеств, если фиксировано расстояние между оболочками и их центрами. Если к тому же многомерный закон распределения значений параметров является нормальным, то для любой пары признаков мы получим два эллипса, как это показано для двух координат x_j и x_k на рис. 7.5. Левая часть рисунка соответствует ситуации, когда разделимость двух областей оценивается, как расстояние между центром области "Класс 1" и всеми точками множества "Класс 2". Правая часть соответствует ситуации, когда разделимость двух областей оценивается как расстояние между центром области "Класс 2" и точками "Класс 1". Пунктирные эллипсы на рис. 7.5 соответствуют ситуации, когда одновременно оценивается дисперсия обоих разделяемых областей, что упрощает задачу их разделения при неизменности расстояний между центрами этих областей. Центрам обеих областей соответствуют векторы математических ожиданий значений каждого признака: $m(X_1)$ и $m(X_2)$ для множеств объектов "Класс 1" и "Класс 2", соответственно.

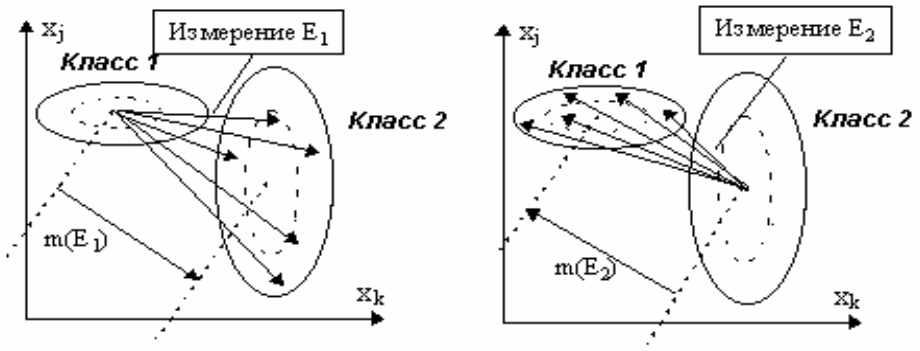


Рис. 7.5. Интерпретация расстояния Махаланобиса для объектов двух классов

Для измерения расстояния от центра области "Класс 1" до точек образа "Класс 2" целесообразно пользоваться квадратичной мерой – выборочным *расстоянием Махалонобиса,* которое при записи в векторной форме будет выглядеть следующим образом:

$$(D_1)^2=(X_2-m(X_1))^{\mathrm{T}}\,\mathbf{C}_1^{-1}\,(X_2-m(X_1)), \qquad (7.12)$$

где $(X_2-m(X_1))^{\mathrm{T}}$ – транспонированный вектор расстояний между каждой точкой множества "Класс 2" и центром области "Класс 1", $(X_2-m(X_1))$ – этот же вектор расстояний по выбранным координа-

там, но без его транспонирования, $\mathbf{C}_1^{-1}$ – обратная ковариационная матрица контролируемых параметров образов "Класс 1".

По аналогии, для измерения расстояния от центра области "Класс 2" до точек множества "Класс 1" мера Махалонобиса будет иметь вид:

$$(\boldsymbol{D_2})^2=(\boldsymbol{X_1}-\boldsymbol{m}(\boldsymbol{X_2}))^{\mathrm{T}}\,\mathbf{C}_2^{-1}\,(\boldsymbol{X_1}-\boldsymbol{m}(\boldsymbol{X_2})), \tag{7.13}$$

где $(\boldsymbol{X_1}-\boldsymbol{m}(\boldsymbol{X_2}))$ и $(\boldsymbol{X_1}-\boldsymbol{m}(\boldsymbol{X_2}))^{\mathrm{T}}$ – прямой и транспонированный векторы расстояний между конкретным примером "Класс 1" и центром области "Класс 2", $\mathbf{C}_2^{-1}$ – обратная ковариационная матрица переменных для множества объектов "Класс 2".

В многомерном случае элементы матрицы $\mathbf{C}$, которая является несмещенной оценкой *ковариационной матрицы* $\boldsymbol{\Sigma}$, вычисляются по следующей формуле:

$$\boldsymbol{c_{jl}} = \frac{1}{\boldsymbol{n}-1}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}(\boldsymbol{x_{ij}}-\boldsymbol{x_{\bullet j}})(\boldsymbol{x_{il}}-\boldsymbol{x_{\bullet l}}) \ , \tag{7.14}$$

где $\boldsymbol{j}$ и $\boldsymbol{l}$ – все возможные пары индексов измеряемых признаков, $\boldsymbol{j}$ = 1,2,…,$\boldsymbol{m}$, $\boldsymbol{l}$ = 1,2,…,$\boldsymbol{m}$. Выражения в скобках – отклонения значений переменных $\boldsymbol{x_{ij}}$ от соответствующего общего среднего $\boldsymbol{x_{\bullet j}}$. При $\boldsymbol{j} = \boldsymbol{l}$ по формуле (7.14) вычисляются среднеквадратичные отклонения, которые соответствует выборочным дисперсиям отдельных переменных, а при $\boldsymbol{j} \neq \boldsymbol{l}$ оцениваются ковариации между двумя переменными. Если каждый элемент ковариационной матрицы $\mathbf{C}$ разделить на квадратный корень из произведения соответствующих диагональных элементов $\boldsymbol{c_{jl}} \Big/ \sqrt{\boldsymbol{c_{jj}}\boldsymbol{c_{ll}}}$., то получается рассмотренная в разделе 7.2 *корреляционная матрица* $\mathbf{R}$.

Обратная матрица $\mathbf{C}^{-1}$ находится специальными методами линейной алгебры путем определения нетривиального решения матричного уравнения $\mathbf{C}\,\mathbf{C}^{-1} = \mathbf{I}$, где $\mathbf{I}$ – единичная матрица (т.е. матрица, состоящая из единиц, расположенных на главной диагонали). Следует обратить особое внимание на то, что вычисление ковариационных матриц $\mathbf{C}$ для векторов, состоящих из десятков и сотен переменных – это вполне реализуемая устойчивая техническая задача, имеющая квадратичную сложность. Однако, располагая вычисленными ковариационными матрицами $\mathbf{C}$, поиск обратных матриц (например, по алгоритму Гаусса) является неустойчивой задачей кубической сложности, поэтому реальным является обращение матриц не более 100-го порядка.

Структура приведенных выражений для расстояния Махалонобиса между вектором и множеством служит основой для построения *обобщеного расстояния Махалонобиса,* между образами "Класс 1" и "Класс 2" с векторами средних значений $\boldsymbol{m}(\boldsymbol{X_1})$ и $\boldsymbol{m}(\boldsymbol{X_2})$, соответственно:

$$\boldsymbol{D^2} = (\boldsymbol{m}(\boldsymbol{X_1})-\boldsymbol{m}(\boldsymbol{X_2}))^{\mathrm{T}}\,\mathbf{C}^{-1}\,(\boldsymbol{m}(\boldsymbol{X_1})-\boldsymbol{m}(\boldsymbol{X_2}))\ . \tag{7.15}$$

Приведенная статистическая мера удовлетворяет аксиомам расстояния только в случае равенства ковариационных матриц обоих классов $\mathbf{C}_1$ и $\mathbf{C}_2$. Поэтому под $\mathbf{C}^{-1}$ обычно понимают некоторую усредненную величину, например, объединенную выборочную ковариационную матрицу вида

$$\mathbf{C} = [(\boldsymbol{n_1} - 1)\,\mathbf{C}_1 + (\boldsymbol{n_2} - 1)\,\mathbf{C}_2] \,/\, (\boldsymbol{n_1} + \boldsymbol{n_2} - 2). \tag{7.16}$$

К основным преимуществам обобщенного расстояния Махалонобиса следует отнести учет коррелированности признаков между собой и инвариантность относительно невырожденных линейных преобразований, что избавляет от необходимости нормировки исходной матрицы наблюдений.

С использованием расстояния Махалонобиса возможна статистическая проверка гипотез о равенстве двух подмножеств векторов при неизвестной ковариационной матрице. Многомерным аналогом для двухвыборочной $\boldsymbol{t}$-статистики Стьюдента является *двухвыборочная* $\boldsymbol{T^2}$*-статистика Хоттелинга*:

$$\boldsymbol{T^2} = (\boldsymbol{m}(\mathbf{X_1})-\boldsymbol{m}(\mathbf{X_2}))^{\mathrm{T}}\,\mathbf{C}^{-1}\,(\boldsymbol{m}(\boldsymbol{X_1})-\boldsymbol{m}(\boldsymbol{X_2}))\ \boldsymbol{n_1 n_2}/(\boldsymbol{n_1}+\boldsymbol{n_2}) = \boldsymbol{D^2}\ \boldsymbol{n_1 n_2}/(\boldsymbol{n_1}+\boldsymbol{n_2})\ . \tag{7.17}$$

Если гипотеза $\boldsymbol{H_o}$: $\boldsymbol{m}(\boldsymbol{X_1}) = \boldsymbol{m}(\boldsymbol{X_2})$ верна, то величина

$$\boldsymbol{F} = \frac{\boldsymbol{n}_1 + \boldsymbol{n}_2 - \boldsymbol{m} - 1}{(\boldsymbol{n}_1 + \boldsymbol{n}_2 - 2)\boldsymbol{m}} \cdot \boldsymbol{T}^2 \tag{7.18}$$

имеет F-распределение с m и $(n_1 + n_2 - m - 1)$ степенями свободы, где m – число переменных.

Алгебраический метод

Описанные выше методы многомерной статистики корректно применимы при выполнении ряда условий: мультинормальность распределения значений измеряемых признаков, равенство ковариационных матриц и достаточно большой объем выборок, позволяющий получать хорошие оценки ковариаций. Каждое из этих условий является скорее исключением, нежели обычной ситуацией, с которой имеет дело биолог. Эти обстоятельство побудили исследователей к поиску иных методов решения задачи разграничения двух совокупностей, основанных на некоторых эвристических принципах. Опишем кратко один из таких методов, предложенный В.Н. Котовым и Н.Г. Терентьевой [1989] и использующий понятие «*биоквант*», который сами авторы определили как "алгебраический".

Общая формализация самого понятия разграничения казалось бы проста и естественна: две совокупности X_1 и X_2 считаются различными, если в некоторой метрике все внутригрупповые расстояния меньше межгрупповых. И действительно, случаи "идеального" разделения встречаются в практике. Однако объективная реальность имеет дело, как правило, с трангрессирующими совокупностями, где допускается существование общих или промежуточных форм, т.е. где торжествует принцип "неопределенности таксона", характерный для "размытых" множеств.

Более слабое понятие различия можно сформулировать следующим образом: две совокупности X_1 и X_2 считаются различными, если, используя некоторую метрику, в них можно выделить достаточно представительные по численности "скученности точек" X_1^* и X_2^*, различные в смысле предыдущего определения. Таким образом, *биоквантом* называются подмножества объектов X_1^* и X_2^*, для которых в заданном метрическом пространстве все внутригрупповые расстояния меньше всех межгрупповых. Пользуясь общепринятой терминологией, согласно которой максимальное расстояние между элементами множества называется диаметром этого множества, биокванты можно охарактеризовать следующим свойством: «*Расстояние между биоквантом X_1^* из совокупности X_1 и биоквантом X_2^* из совокупности X_2 больше диаметра каждого из этих биоквантов*»

Для нахождения биоквантов предложена [Котов, Терентьева, 1989] некоторая эвристическая процедура, основанная на преобразовании матрицы расстояний в стохастическую матрицу и расчете для каждого объекта X_1 и X_2 специальных оценок включения в биокванты, равных стационарным вероятностям перехода в марковской цепи. После выделения биоквантов принимается или отвергается гипотеза о наличии различий между группами измерений по следующему достаточно условному эвристическому правилу:

- если биокванты составляют не менее 60% от объема соответствующей совокупности, то подмножества хорошо разграничиваются,
- если хотя бы один из биоквантов составляет менее 30% количества измерений, то принимается гипотеза об отсутствии различий;
- в остальных случаях совокупности считаются "слабо разделимыми".

Результаты расчетов

Сформируем матрицу из 412 наблюдений, относящихся к 10 рекам Самарской области, по которым был проделан наиболее репрезентативный объем экспедиционных исследований. В составе р. Чапаевка, характеризующейся определенной региональной неоднородностью, выделим две группы станций, расположенных в верхнем течении (ст. 1-12) и ниже г. Чапаевска (ст. 13 – 23).

В качестве 8 контролируемых переменных будем использовать три группы признаков:

- показатели обилия – прологарифмированные значения общей численности N_s и биомассы B_s по всем видам бентоса;
- показатели биоразнообразия – количество видов S и информационный индекс Шеннона H, рассчитанный по численности зообентоса;
- показатели хирономидного комплекса – значения $\ln((N_s \cdot B_s)^{0.5})$ отдельно по подсемействам Orthocladiinae и Tanypodinae, а также трибам Tanytarsini и Chironomini.

Для каждой пары групп измерений, относящихся к разным водоемам, рассчитаем обобщенное расстояние по Махалонобису D^2, критерий Хоттелинга T^2 и соответствующие ему F-критерий и значение вероятности p (см. табл. 7.5).

Таблица 7.5

Матрица оценок различий между многомерными комплексами гидробиологических наблюдений на реках Самарской области

	Наименование рек	*m*	1	2	3	4	5	6	7	8	9	10	11
1	Уса	8		2.24	2.31	1.45	2.40	2.15	1.54	1.63	3.03	3.12	1.95
				27.23	30.50	11.77	24.44	24.62	17.72	19.61	57.43	43.13	27.40
2	Б.Черемшан	17	2.37		2.02	2.40	3.66	2.71	1.75	1.86	4.54	3.03	1.90
			0.0677		37.43	51.66	79.01	60.56	45.77	49.54	220.50	57.75	49.46
3	Б.Кинель	20	2.79	3.74		2.62	3.25	2.34	1.51	1.51	4.10	2.56	3.53
			0.0318	0.0043		66.86	65.58	48.52	39.25	37.14	199.18	43.65	193.91
4	Маза	19	1.06	5.13	6.78		1.36	1.07	1.71	1.41	3.13	2.92	1.66
			0.4319	0.0006	0.0001		11.29	9.87	48.02	31.29	112.36	55.93	41.45
5	Тайдаков	9	1.63	7.00	6.07	1.03		1.44	2.84	1.24	2.22	3.21	2.41
			0.2527	0.0004	0.0005	**0.447**		12.01	67.58	12.61	33.77	48.72	46.32
6	Муранка	16	2.10	5.86	4.82	0.97	1.04		1.87	1.12	2.63	1.26	1.47
			0.1028	0.0003	0.0009	**0.4787**	**0.4451**		49.35	17.05	71.44	9.80	28.15
7	Чапаевка (верховья)	124	2.10	5.43	4.66	5.70	8.00	5.86		1.45	3.45	1.84	1.68
			0.0411	0	0.0001	0	0	0		108.58	279.26	31.29	127.73
8	Сок	89	2.27	5.78	4.34	3.65	1.46	1.99	13.12		1.23	0.93	1.89
			0.0294	0	0.0002	0.0009	**0.1827**	**0.0562**	0		33.13	7.83	140.38
9	Байтуган	29	5.74	23.18	21.19	11.91	3.40	7.48	33.29	3.89		2.94	3.48
			0.0002	0	0	0	0.0071	0	0	0.0005		64.13	249.03
10	Самара	10	3.03	5.20	4.09	5.18	3.58	0.87	3.70	0.91	6.50		2.67
			0.0593	0.0018	0.0045	0.0013	0.0315	**0.5611**	0.0006	**0.5131**	0.0001		62.37
11	Чапаевка (низовья)	71	3.11	5.68	22.33	4.77	5.27	3.23	15.39	16.77	28.91	7.11	
			0.0046	0	0	0.0001	0	0.0032	0	0	0	0	

Примечание: Для каждой пары рек в клетках, расположенных выше главной диагонали - обобщенное расстояние по Махалонобису D^2 (вверху) и значения T^2-критерия Хоттелинга (внизу); в клеткак, расположенных ниже главной диагонали – F-критерий и соответствующая ему вероятность p.

Если величина p превышает выбранный уровень значимости, то нет оснований отвергать нулевую гипотезу о том, что измерения, выполненные на данной паре водоемов, принадлежат к одной генеральной совокупности. Например, можно считать статистически сходными по данному комплексу признаков реки Маза (4) и Тайдаков (5), Маза и Муранка (6), Тайдаков и Муранка – эти и другие пары вероятностей, превышающие 0.05, отмечены в табл. 7.5 жирным шрифтом.

Полученные значения обобщенной меры Махаланобиса, T^2 и F-критериев могут быть интерпретированы как матрицы расстояния между классифицируемыми объектами и обработаны описанными в предыдущем разделе алгоритмами кластерного анализа с целью построения дендрограмм.

На рис. 7.6 приведены результаты классификации рек Самарского региона по методу Уорда. Необходимо отметить, что выборочные меры Махалонобиса D^2 зависят от объема n_1 и n_2 сравниваемых подвыборок, в результате чего при большом количестве измерений, характерных, например, для р. Чапаевка, коэффициенты расстояния оказываются сильно заниженными, что нашло свое отражение в результатах кластеризации на дендрограмме фиг. «а». Значения критерия Хоттелинга T^2 и F-критерия в этих же условиях являются несмещенными оценками, не так сильно зависящими от количества измерений, поэтому дают, по нашему мнению, более адекватные результаты кластеризации (см. фиг. «б» рис. 7.6).

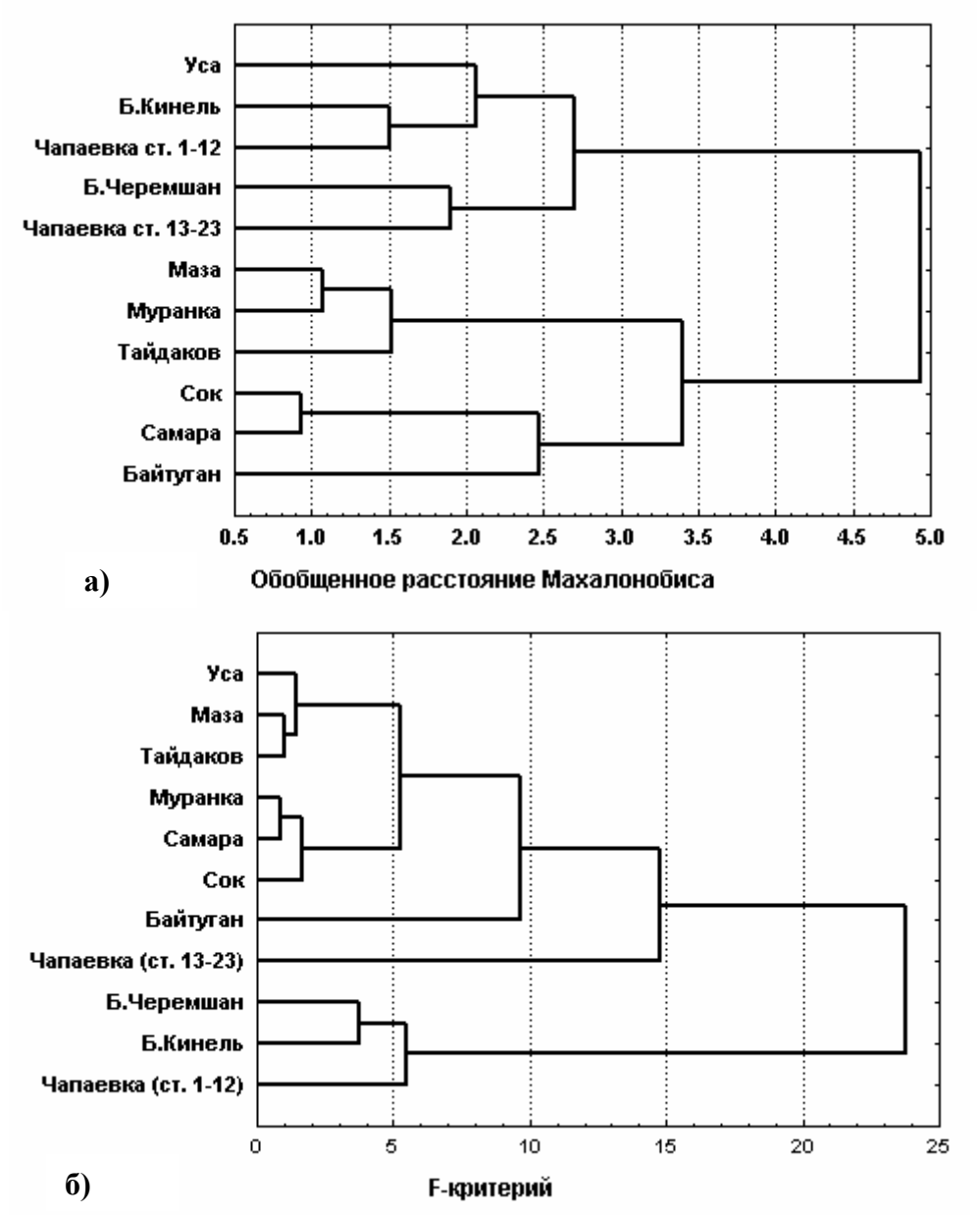


Рис. 7.6. Дендрограммы кластеризации рек Самарской области по методу Уорда с использованием обобщенного расстояния Махалонобиса и критерия Фишера

7.5. Задача о снижении размерности многомерного пространства: факторный анализ

Формулировка задачи

Исходными данными является знакомая по предыдущим разделам таблица наблюдений **X**, в столбцах которой представлены гидробиологические показатели, измеренные в количественной шкале ($j = 1,2,\ldots,m$), а строки содержат описания экологических объектов в некотором пространственно-временном аспекте ($i = 1,2,\ldots,n$).

Необходимо найти такое линейное преобразование исходной матрицы **X**, которое позволило бы получить сжатое (редуцированное) представление входных данных в виде матрицы **F** с меньшим числом переменных p ($m > p$) без существенной потери содержательной информации об экологических объектах.

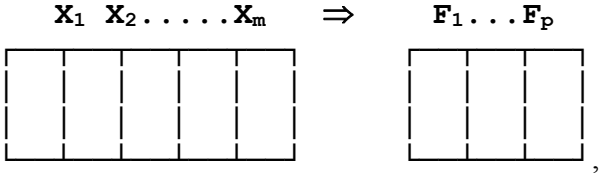
Под факторным анализом понимается совокупность статистических моделей, описывающих и объясняющих наблюдаемые данные с помощью небольшого числа скрытых (латентных) факторов, которые могут быть сконструированы с помощью определенных математических методов. Модели факторного анализа применяются при решении следующих задач:

- редукция данных или понижение размерности признакового пространства типа "объект – признак" за счет сведения многочисленных взаимозависимых наблюдаемых переменных к некоторым обобщенным ненаблюдаемым факторам;
- преобразование исходных переменных к виду, более удобному для визуализации или интерпретации; классификация объектов на основе сжатого признакового пространства;
- создание структурной теории исследования объектов и интерпретация косвенных факторов, не поддающихся непосредственному измерению.

В экологии факторный анализ можно считать одним из первых серьезных многомерных методов статистики, использованных, в частности, для ординации растительности [Goodall, 1954]. Имеется обширная литература как по применению факторного анализа вообще [Лоули, Максвелл, 1967; Харман, 1972; Айвазян с соавт., 1974; Дубров, 1978], так и по использованию его в экологических исследованиях [Василевич, 1969; Миркин, Розенберг, 1977б, 1978; Миркин с соавт., 1978а,б; Гелета, Крауклис, 1979; Нинбург, 1985; Колодяжный, 1985; Ястребов, 1991; Красовский, Воробьева, 1998; Сердюцкая, Каменева, 2000].

Математический лист

Под *моделью факторного анализа* понимают представление исходных переменных в виде линейной комбинации факторов $\boldsymbol{F}$



рассчитанных так, чтобы наилучшим способом (с минимальной погрешностью) представить $\boldsymbol{X}$:

$$X_j = \sum_{k=1}^{p} a_{jk} F_k + U_j \qquad (7.19)$$

В этой модели латентные переменные $\boldsymbol{F_k}$, $\boldsymbol{k}$ = 1,2,…,$\boldsymbol{p}$, называются *общими факторами*, а переменные $\boldsymbol{U_j}$, $\boldsymbol{j}$ = 1,2,…,$\boldsymbol{m}$, – *специфическими факторами* (англ. – unique factor). Значения $\boldsymbol{a_{jk}}$ называются *факторными нагрузками*.

Основное требование к исходным данным для факторного анализа – это то, что они должны подчиняться допущению о многомерном нормальном распределении в совокупности. Для проверки этой гипотезы используют *тест "сферичности" распределения данных Бартлетта*, где оценивается предположение о диагональности матрицы корреляций. Если эта гипотеза не отвергается (т.е. наблюдаемый уровень значимости превышает 5%) – нет смысла в факторном анализе, поскольку направления главных осей случайны. На практике предположение о многомерной нормальности проверить весьма трудно, поэтому факторный анализ чаще всего применяется без такой процедуры, тем более, что ряд исследователей [Лоули, Максвелл, 1967] считает это допущение излишним.

Однако если предполагается, что все признаки $\boldsymbol{X_j}$ стандартизованы ($\boldsymbol{\sigma_i}$=1, $\boldsymbol{m}(X_j)$ = 0), а факторы $\boldsymbol{F_1}$, $\boldsymbol{F_2}$, ..., $\boldsymbol{F_p}$ независимы и не связаны со специфическими факторами $\boldsymbol{U_j}$, то факторные нагрузки $\boldsymbol{a_{jk}}$ совпадают с коэффициентами корреляции между общими факторами и переменными $\boldsymbol{X_j}$. Общая дисперсия признака $\boldsymbol{X_j}$ раскладывается при этом на сумму квадратов факторных нагрузок $\boldsymbol{H_i^2}$, которая называется *общностью*, и дисперсию специфического фактора $\boldsymbol{S_{u_i}^2}$, или *специфичность*:

$$S_{x_i}^2 = H_i^2 + S_{u_i}^2 , \qquad \text{где} \quad H_i^2 = \sum_k a_{ik}^{\,2} . \qquad (7.20)$$

Другими словами, общность H_i^2 представляет собой часть дисперсии переменных, объясненную факторами, а специфичность $S_{u_i}^2$ – часть дисперсии, обусловленную случайными ошибками или переменными, неучтенными в модели. В соответствии с постановкой задачи, необходимо искать такие факторы, при которых суммарная общность максимальна, а специфичность – минимальна.

Основным объектом преобразований в факторном анализе является корреляционная матрица из коэффициентов корреляции Пирсона (иногда – дисперсионно-ковариационная матрица), полученная обычным путем обработки массива данных X. Выделение общих факторов и сжатие информации в ходе факторного анализа сводится к воспроизведению с той или иной степенью точности исходной корреляционной матрицы, т.е. предполагается, что редуцированная корреляционная матрица получена с использованием тех же объектов, но описанных меньшим числом переменных. Таким образом, следует уточнить, что фактически под сжатием информации в факторном анализе понимается уменьшение размерности корреляционной матрицы, а не самих данных, тем более что восстановить исходные данные по корреляционной матрице нельзя.

Поскольку коэффициенты, составляющие корреляционную матрицу, могут вычисляться разным способом, различают следующие техники факторного анализа:

- R – техника, когда коэффициенты корреляции вычисляются между переменными и исходная матрица X сжимается по столбцам, т.е. число признаков уменьшается с m до p;
- Q – техника, когда изучается корреляция между объектами (точнее, их состояниями, описываемыми векторами параметров) и их количество уменьшается с n до p;
- P – техника, предполагающая факторный анализ результатов экспериментальных исследований, выполненных на одном и том же объекте в различные промежутки времени.

Одним из наиболее распространенных приемов поиска факторов является *метод главных компонент*. Его основное различие от факторного анализа заключается в том, что главные компоненты F_k связаны с наблюдаемыми переменными X_j линейными функциями преобразования:

$$X_j = \sum_{k=1}^{p} a_{jk} F_k \quad \text{и} \quad F_k = \sum_{j=1}^{m} a_{jk} X_j \ . \tag{7.21}$$

Метод главных компонент более прост в расчетах и интерпретации, но одна из главных трудностей его использования – необходимость преобразования исходных данных, представленных в разных единицах измерения, в сопоставимые величины. Традиционным методом преобразования является нормирование по стандартным отклонениям, когда матрица Z стандартизованных исходных данных определяется по формуле $z_{ij} = \dfrac{x_{ij} - x_{\bullet j}}{S_j}$, где $x_{\bullet j}$ – среднее значение j-го признака; S_j – стандартное отклонение; $j = 1,2,\ldots,m$; $i = 1, 2,\ldots,n$.

Для вычисления корреляционной матрицы R размером $m \times m$, имеет место простое матричное соотношение:

$$R = \frac{1}{m-1} ZZ^{\mathrm{T}} \ , \tag{7.22}$$

где T – символ транспонирования.

Основная идея метода главных компонент основана на следующем предположении: чем выше дисперсия вдоль какой-нибудь оси, тем больше информации содержат значения проекций на эту ось. Поэтому вполне естественно предпринять попытку отыскать ось с максимальной дисперсией, которую можно было бы рассматривать как "ординационную" со всеми вытекающими отсюда последствиями. Такая ось называется *первой главной компонентой (фактором)*.

Поиск всей системы взаимно перпендикулярных осей по методу главных компонент сводится к последовательной процедуре: т.е. вначале ищется первый фактор, который объясняет наибольшую часть дисперсии, затем независимый от него второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т.д.

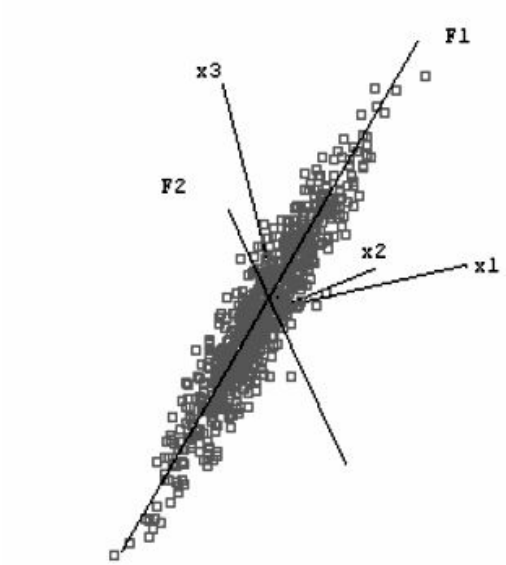


Рис. 7.7. Сжатие признакового пространства с применением факторного анализа

Геометрически это выглядит следующим образом (см. рис. 7.7). Для построения первого фактора F_1 берется прямая, проходящая через центр координат и облако рассеяния данных. При этом отыскивается такая ось, для которой сумма квадратов расстояний всех точек до перпендикуляра к этой прямой была бы максимальна. Это означает, что этой осью объясняется максимум дисперсии переменных. Найденная ось после нормировки используется в качестве первого фактора. Если "облако" данных вытянуто в виде эллипсоида (имеет форму "огурца"), фактор F_1 совпадет с направлением, в котором вытянуты объекты, и по нему с наибольшей точностью можно предсказать значения исходных переменных. Для поиска второго фактора F_2 ищется ось, перпендикулярная первому фактору, также объясняющая наибольшую часть дисперсии, не объясненной первой осью. После нормировки эта ось становится вторым фактором. Если данные представляют собой плоский элипсоид ("блин") в трехмерном пространстве, два первых фактора позволяют в точности описать эти данные. Максимально возможное число главных компонент равно количеству переменных.

Основная модель метода главных компонент записывается в матричном виде следующим образом:

$$\mathbf{Z} = \mathbf{AF}, \tag{7.23}$$

где $\mathbf{Z}$ – матрица $m \times n$ стандартизованных исходных данных; $\mathbf{A}$ – матрица $m \times p$ факторых нагрузок (факторное отображение); $\mathbf{F}$ – матрица $p \times n$ значений факторов; m – количество переменных и n – количество объектов исходной матрицы; p – количество выделенных факторов. Очевидно, что неизвестными являются матрицы $\mathbf{A}$ и $\mathbf{F}$.

Вычислительные аспекты метода главных компонент сводятся к следующим шагам:

1. Решается характеристическое матричное уравнение

$$\mathbf{R} = \Lambda\mathbf{V}\ , \tag{7.24}$$

которое в общем случае имеет m корней λ, называемых *собственными* или *характеристическими числами* (англ. – eigen value) корреляционной матрицы $\mathbf{R}$, каждому из которых соответствует вектор-столбец $\mathbf{V}$ базисных функций. Собственными значениями квадратной матрицы $\mathbf{R}$ порядка m называются такие значения λ_k, при которых система следующих m уравнений имеет нетривиальное решение:

$$\mathbf{RV_k} = \lambda_k\mathbf{V_k}\ , \tag{7.25}$$

где $\mathbf{V_k}$ – *собственные векторы* матрицы $\mathbf{R}$, соответствующие λ_k.; $k = 1,2,\ldots,m$.

2. Из последовательности собственных значений λ_k выбирается p максимальных. Матрица факторных нагрузок $\mathbf{A}$ каждой исходной переменной j на каждый выделенный фактор k, соответствующая коэффициентам линейных преобразований a_{jk}, вычисляется по формуле

$$a_{jk} = v_{ik} \cdot (\lambda_k)^{0.5}\ , \quad j = 1, 2, \ldots, m;\ k = 1, 2, \ldots, p\ . \tag{7.26}$$

3. Редуцированную матрицу факторов $\mathbf{F}$, соответствующую исходной таблице $\mathbf{X}$ наблюдений, в которой количество столбцов уменьшено с $\mathbf{m}$ до $\mathbf{p}$, рассчитывают по формуле:

$$f_{ik} = \sum_{j=1}^{m} a_{jk} z_{ij}\ . \tag{7.27}$$

Основная проблема расчетов состоит в оценке того, сколько главных компонент необходимо построить для оптимального представления анализируемых исходных факторов. Величина λ_k представляет не что иное, как часть суммарной дисперсии совокупности преобразованных данных, объясненную главной компонентой F_k. Если переменные стандартизованы, то $\lambda_1 > \lambda_2 > \lambda_3, \ldots$ и

необходимо иметь в виду, что первые несколько членов разложения дают основной вклад в объяснение вариации величин в исходных данных.

Решение о том, когда следует остановить процедуру выделения компонент, зависит главным образом от точки зрения на то, что считать малой долей дисперсии. Это решение достаточно произвольно, однако имеются два критерия: *критерий Кайзера* (Kaiser) и *критерий "каменистой осыпи" Кэттелла* (Cattell), которые в большинстве случаев позволяют рационально выбрать число компонент. Но анализ составляющих с малыми величинами собственных значений едва ли целесообразен еще и потому, что они могут оказаться статистически недостоверными из-за ошибок различного происхождения. Ввиду того, что иллюстративной целью факторного анализа часто является получение факторного отображения в графическом виде, обычно ограничивают $\boldsymbol{p} = 2$ и дают изображение пространства в двумерном срезе, поскольку выполнить это для трех и более выделенных факторов проблематично.

Для интерпретации факторов необходимо приписать каждому из них некоторый содержательный смысл, связанный с предметной областью. Чтобы понять, какая гидробиологическая реальность скрыта в найденных факторах, необходимо провести анализ корреляций факторных нагрузок с исходными переменными. Для повышения интерпретируемости факторов используют метод варимаксного вращения VARIMAX, позволяющий добиться большей "выразительности" матрицы факторных нагрузок [Харман, 1972]. Его суть состоит в изменении координатных осей, образуемых факторами, с целью получить более контрастные нагрузки так называемой *простой факторной структуры*. Новые факторы в результате вращения осей ищутся в виде специального вида линейной комбинации имеющихся факторов: $\hat{\boldsymbol{F}}_{\boldsymbol{i}} = \sum_{\boldsymbol{k}=1}^{\boldsymbol{m}} \boldsymbol{b}_{\boldsymbol{ik}} \boldsymbol{F}_{\boldsymbol{k}}$, максимизирующей "дисперсии" квадратов факторных нагрузок для переменных

$$\sum_{\boldsymbol{i}} \left[\sum_{\boldsymbol{k}} \boldsymbol{a}_{\boldsymbol{ik}}^{4} / \boldsymbol{m} - \left[\sum_{\boldsymbol{k}} \boldsymbol{a}_{\boldsymbol{ik}}^{2} / \boldsymbol{m} \right]^{2} \right] \rightarrow \max \quad . \tag{7.28}$$

Чем сильнее разойдутся квадраты факторных нагрузок к концам отрезка [0,1], тем больше будет значение целевой функции вращения и тем четче интерпретация факторов.

Результаты расчетов

Выполним факторный анализ методом главных компонент с использованием примера, приведенного в разделе 7.2. Матрица парных коэффициентов корреляции Пирсона, рассчитанная по 453 наблюдениям для 19 признаков, отражающих обилие различных групп хирономид, и 3 гидрофизических признаков, представлена в табл. 7.3.

Были проведены две серии расчетов: для общего набора признаков и с использованием только гидробиологических показателей. В первом случае фактор 1, нагрузки для которого представлены в табл. 7.6, достаточно отчетливо можно приписать условиям существования биотопа – температуре, глубине и ширине водоема, кампанию которым, по трудно объяснимым причинам, составили личинки хирономид фитодетритофагов-собирателей из подсемейства Orthocladiinae.

Рассмотрим более подробно факторный анализ структуры хирономидного комплекса (расчет 2 в табл. 7.6). Последовательность выделения главных факторов можно представить в виде графика "каменной осыпи" на рис. 7.8, на котором можно усмотреть два "обвала": при 2 и при 6 отбираемых факторов.

По отношению к водным экосистемам величины главных компонент и базисных векторов могут рассматриваться как показатели определенного типа взаимоотношений между совместно обитаемыми видами, или, иначе говоря, определенного типа ассоциирования. Например, первый фактор объединяет почти все трофические группы трибы Chronomini, к которым добавились хищники Tanypodinae. Второй фактор структурно обозначен фитодетритофагами-собирателями Orthocladiinae и Tanytarsini. Для остальных четырех выделенных нами факторов также можно достаточно адресно подобрать доминирующие группы. Все 6 факторов объясняют около 55% общего статистического разброса, причем на первые два фактора приходится 29%.

Таблица 7.6

Факторные нагрузки по главным компонентам, рассчитанные на основе численности различных подсемейств и трофических групп хирономид (жирным шрифтом отмечены нагрузки, превышающие 0.5)

Переменная	Факторы расчета 1			Факторы расчета 2					
	1	2	3	1	2	3	4	5	6
Собственные значения	3.53	2.56	1.47	3.14	2.15	1.42	1.28	1.22	1.21
Объясненная дисперсия, %	16.04	27.67	34.37	16.52	27.85	35.33	42.07	48.48	54.83
Хищ. ChC	-0.17	0.49	0.20	**0.60**	-0.20	-0.04	0.05	-0.04	-0.10
Хищ. Di	0.39	-0.19	0.02	-0.05	0.01	0.02	-0.30	-0.01	**0.73**
Хищ. Or	0.09	-0.11	0.08	-0.03	-0.04	0.00	**-0.74**	-0.15	0.03
Хищ. Pr	0.41	-0.07	**-0.54**	-0.11	0.04	0.09	-0.06	0.73	0.30
Хищ. Tn	0.02	**0.60**	0.00	**0.61**	0.20	-0.07	0.03	0.05	0.10
Вс/Соб ChC	-0.12	**0.63**	0.31	**0.67**	0.01	-0.27	0.06	-0.10	-0.06
Вс/Соб Or	0.25	0.06	**0.60**	0.00	0.07	**-0.82**	0.05	-0.06	0.04
Детр ChC	-0.04	**0.75**	0.04	**0.71**	0.22	-0.03	-0.06	0.03	-0.12
Детр ChT	0.49	**0.54**	0.01	0.32	**0.70**	-0.22	-0.05	0.12	0.08
Детр Or	0.04	0.36	-0.23	0.16	**0.58**	0.27	0.03	-0.11	-0.15
Сест ChC	0.09	0.25	**0.59**	0.24	-0.03	**-0.70**	-0.02	-0.10	-0.11
Сест ChT	0.26	-0.07	-0.19	-0.08	0.09	0.02	**-0.77**	0.20	0.02
Сест Pr	0.26	-0.03	-0.43	-0.11	0.01	-0.02	0.12	**0.67**	0.09
деТ/филь ChC	-0.25	**0.54**	0.26	**0.65**	-0.14	-0.12	0.03	-0.17	-0.03
деТ/филь ChT	0.31	-0.09	-0.09	0.02	-0.07	0.00	0.08	0.20	**0.71**
Фит/дет ChC	-0.09	**0.50**	-0.03	**0.51**	0.12	0.22	0.04	-0.08	0.00
Фит/дет Di	0.41	-0.13	0.00	-0.17	0.35	0.06	0.17	-0.09	**0.52**
Фит/дет Or	**0.68**	0.03	-0.21	-0.23	**0.71**	-0.11	-0.08	0.27	0.16
Фит/дет Pr	0.17	0.14	**-0.58**	0.01	0.12	0.11	-0.12	**0.70**	-0.19
Температура воды (дно)	**-0.61**	0.30	0.19						
Глубина реки	**-0.67**	-0.12	-0.07						
Ширина реки	**-0.60**	-0.08	-0.04						

Примечание: В таблице использованы следующие условные обозначения трофических групп: «Хищ.» - хищники хвататели; «Вс/Соб» - всеядные собиратели+хвататели; «Детр» - детритофаги собиратели; «Сест» - сестонофаги+детритофаги фильтраторы; «деТ/филь» - детритофитофаги собиратели + фильтраторы; «Фит/дет» - фитодетритофаги собиратели; подсемейств и триб: Or – Orthocladiinae, Tn – Tanypodinae, Di – Diamesinae, Pr – Prodiamesinae, Ch – Chironominae (ChC – Chironomini, ChT – Tanytarsini)

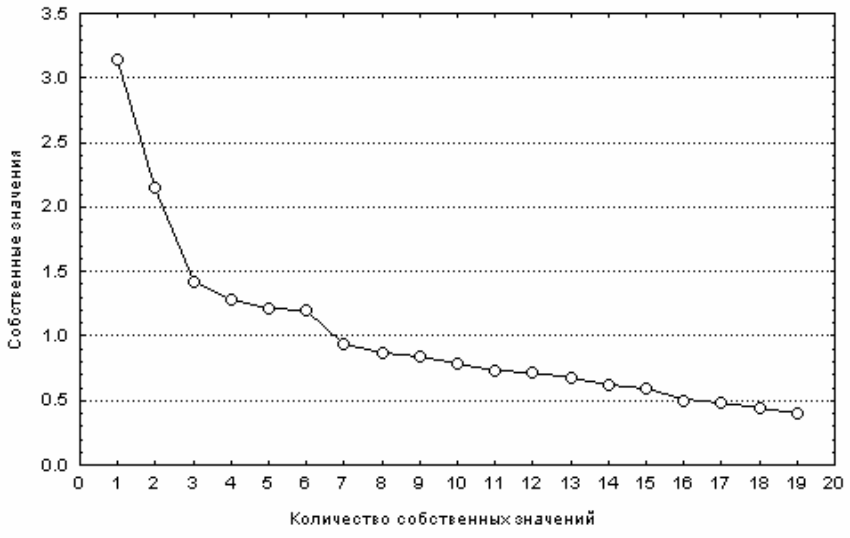


Рис. 7.8. График последовательности собственных значений при выделении главных факторов ("каменистая осыпь" Кэттелла)

Кроме табличной формы используются графические методы визуализации результатов факторного анализа в виде двухмерных срезов факторного пространства или трехмерных диаграмм (см. рис. 7.9 и 7.10, соответственно).

Используя расчитанные факторные нагрузки как коэффициенты линейного преобразования, можно сформировать редуцированную матрицу исходных данных, где столбцами являются новые факторизованные признаки. Выделение первых двух главных компонент дает возможность выполнить анализ двухмерной визуализации взаимного расположения объектов в свернутом пространстве факторов.

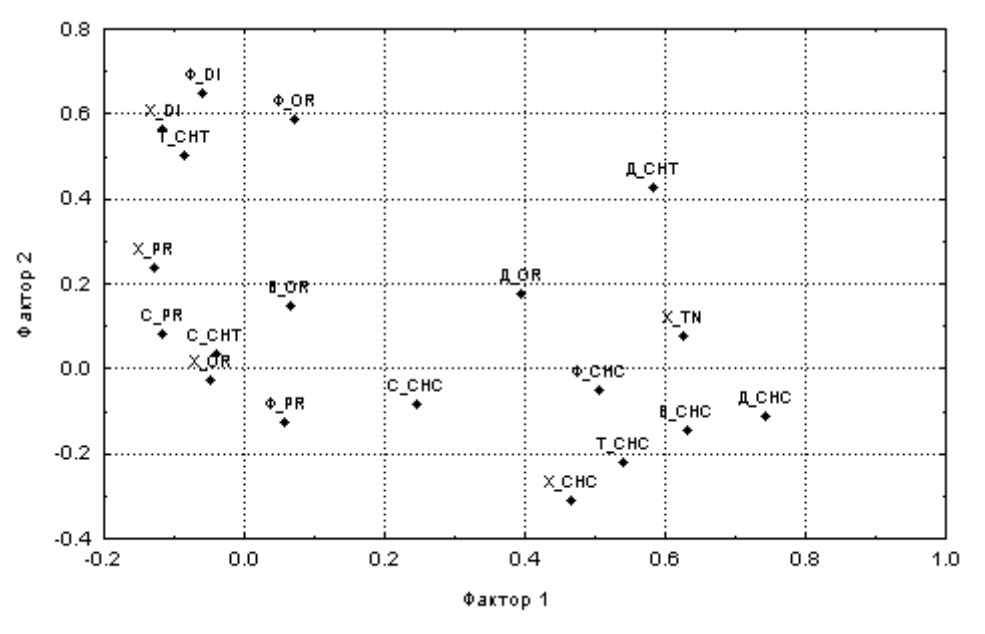


Рис. 7.9. Двухмерный график отображения факторных нагрузок

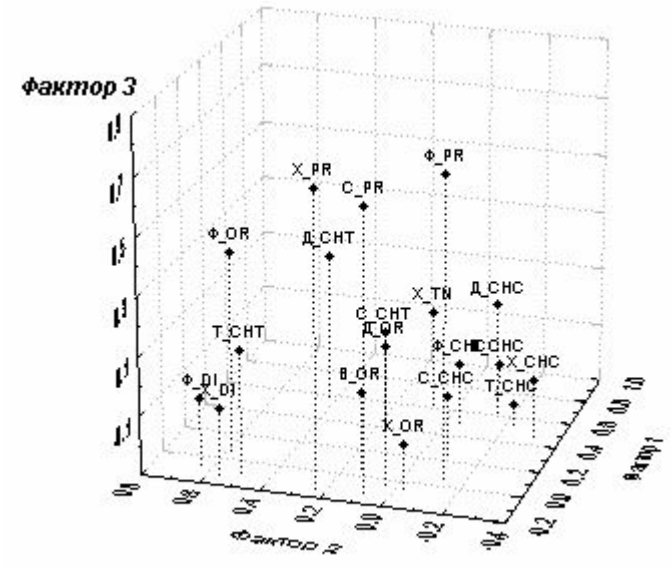


Рис. 7.10. Трехмерный график отображения факторных нагрузок

Поскольку выделение сгущений в облаке из 453 точек вызывает некоторые "изобразительные" трудности, представим на рис. 7.11 расчетные точки, соответствующие групповым средним значений факторов, вычисленным по множеству измерений для каждой малой реки. При таком способе отображения следует иметь в виду, что большинство точек, представленных на рис. 7.11, является в свою очередь трангрессирующими кластерами, диаметр которых может перекрывать значительную часть диапазона варьирования факторов.

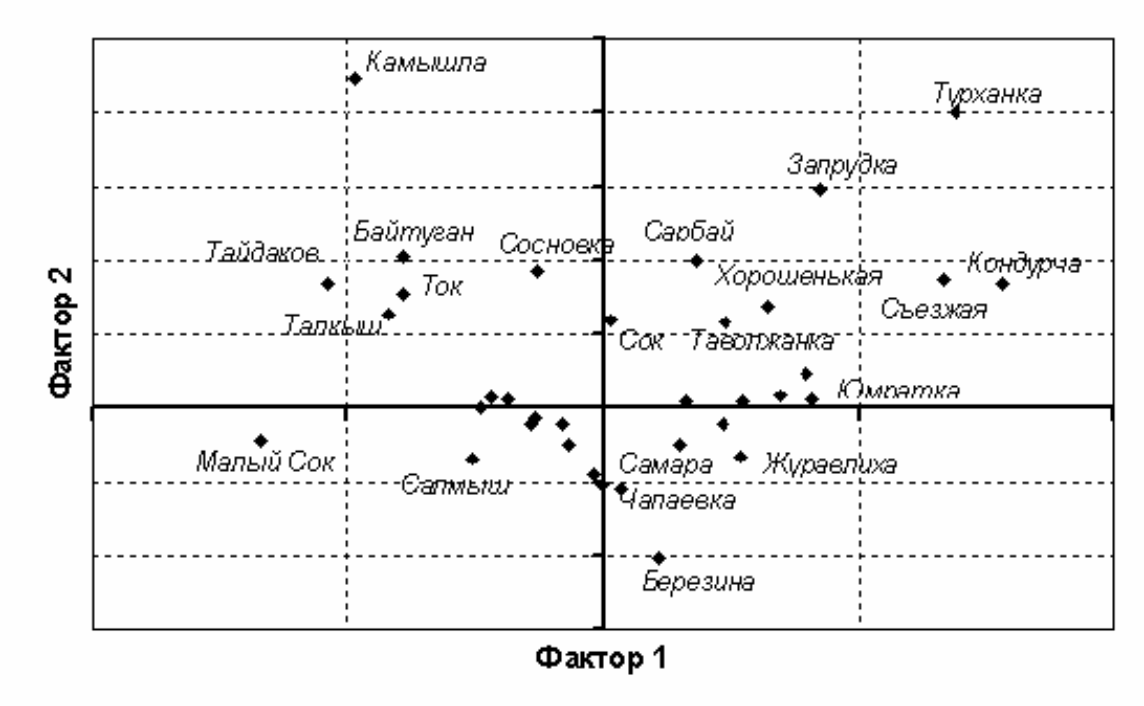


Рис. 7.11. Отображение малых рек Самарской области в пространстве двух главных факторов по результатам отбора проб хирономид

На рис. 7.12 представлено несколько рек из "срединной" части общего графика и для них пунктиром обозначен доверительный интервал значений факторов (который значительно уже минимаксного интервала, соответствующего диаметру подмножеств). Река Чапаевка при этом была разделена на две части, соответствующие верхнему и нижнему течениям.

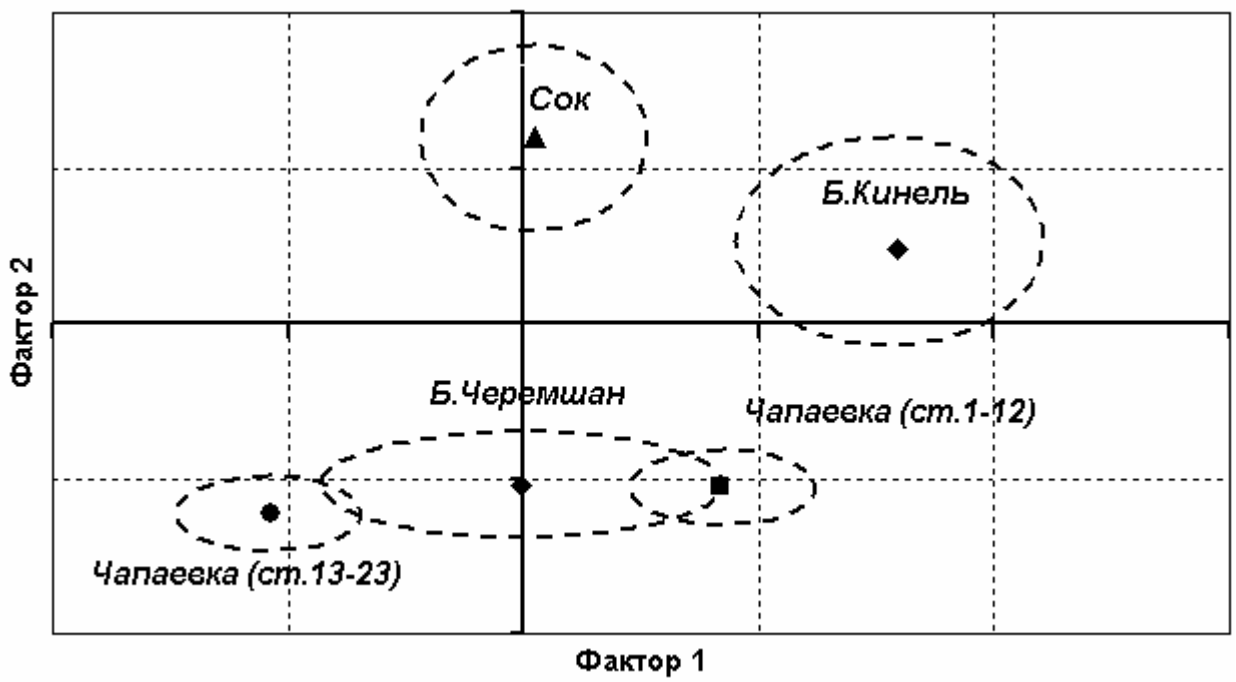


Рис. 7.12. Отображение четырех малых рек Самарской области в пространстве двух главных факторов по результатам отбора проб хирономид (пунктиром обозначена доверительная область варьирования факторов для каждой реки)

Верхний правый квадрант обоих графиков соответствует рекам, у которых велико как значение фактора 1 (который мы ранее связали с обилием ассоциации хищных хирономид из трибы Chronomini и подсемейства Tanypodinae), так и фактора 2 (фитодетритофаги-собиратели Orthocladiinae и Tanytarsini).

Нижний левый квадрант объединяет объекты с низким обилием всех этих групп. Можно, например, предположить, что хирономидный комплекс р. Сок отличается от р. Б.Черемшан высоким обилием видов, объединяемых фактором 2, а верхнее течение р. Чапаевка по сравнению с нижним течением той же реки характеризуется более высокой численностью видов хирономид при одинаковом относительно низком обилии фитофильных личинок ортокладеин.

7.6. Метод многомерного шкалирования

Формулировка задачи

Пусть имеется квадратная матрица $\boldsymbol{R}$ размерностью $\boldsymbol{n}\times\boldsymbol{n}$, каждый элемент которой на пересечении $\boldsymbol{i}$-й строки и $\boldsymbol{j}$-го столбца, содержит достаточно произвольные сведения о попарном сходстве анализируемых объектов $\boldsymbol{i}$ и $\boldsymbol{j}$. На выходе алгоритма многомерного шкалирования получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во "вспомогательных шкалах", связанных с латентными переменными).

Многомерное шкалирование (МШ) можно рассматривать как альтернативу факторному анализу, когда, кроме корреляционных матриц, в качестве исходных данных можно использовать произвольный тип матрицы сходства объектов. МШ – это не сколько определенная математическая процедура, а скорее способ наиболее эффективного размещения объектов, приближенно сохраняющий расстояния между ними в новом пространстве признаков, размерность которого существенно меньше исходного.

Рекомендуемая литература: [Терехина, 1986; Краскэл, 1986; Дэйвисон, 1988; Ципилева, 1989].

Математический лист

Методы многомерного шкалирования размещают объекты в пространстве заданной размерности и проверяют, насколько точно полученная конфигурация сохраняет расстояния между объектами. При этом использует алгоритм минимизации некоторой функции, оценивающей качество получаемых вариантов отображения.

Первая в этом направлении работа В. Торгерсона [Torgerson, 1952; цит. по: Терехина, 1986] была посвящена поиску оптимальных шкал и линейному преобразованию матрицы исходных расстояний $\mathbf{R}$, минимизирующему ошибку $\mathbf{E}$:

$$\mathbf{T} = \boldsymbol{L}\{\mathbf{R}\} = \mathbf{D}^2 + \mathbf{E}\,, \tag{7.29}$$

где $\boldsymbol{L}\{\mathbf{R}\}$ – линейное преобразование исходной матрицы расстояний, $\mathbf{T}$ – матрица расстояний, полученная на основе созданных шкал, $\mathbf{E}$ – матрица отклонений модели от исходных данных.

Р. Шепард и Дж. Краскэл [Shepard,1962; Kruscal,1964; цит. по: Терехина, 1986] совершили существенный прорыв, разработав *метод неметрического шкалирования* (МНШ), который имеет больше шансов получить действительно геометрическое пространство. Суть этого метода состоит в нелинейном (монотонном) $\boldsymbol{M}\{\mathbf{R}\}$ преобразовании исходной матрицы расстояний:

$$\mathbf{T} = \boldsymbol{M}\{\mathbf{R}\} = \mathbf{D}^2 + \mathbf{E}\,.$$

Пусть с помощью специальной итерационной процедуры мы определили $\boldsymbol{r}$ шкал $\boldsymbol{F^1}, \ldots, \boldsymbol{F^r}$. Тогда расстояние между парой объектов $\boldsymbol{i}$ и $\boldsymbol{j}$; $\boldsymbol{i} = 1,2,\ldots,\boldsymbol{n}$; $\boldsymbol{j} = 1,2,\ldots,\boldsymbol{n}$; определяется, например, формулой Евклида:

$$t_{ij} = \sqrt{\sum_{k=1}^{r} (F_i^k - F_j^k)^2}\,. \tag{7.30}$$

Для однозначности задания шкал предполагается, что $\sum_{i}^{n} F_i^k = 0$ и $\sum_{i}^{n}\sum_{k}^{r} (F_i^k)^2 = nr$. Кроме того, по аналогии с методом главных компонент, первая шкала выбирается с наибольшей диспер-

сией, вторая – имеет вторую наибольшую дисперсию и т.д. Кроме формулы Евклида могут быть использованы манхеттенское расстояние, метрика Минковского, формула Колмогорова и проч. (см. раздел 4.7)

В качестве критерия оптимизации итерационной процедуры выбора шкал используются различные похожие между собой показатели *стресса* (слово stress в английском языке имеет множество значений, одно из которых –"нагрузка"). Большинство из них сводится к вычислению суммы квадратов отклонений исходных r_{ij} и вычисленных шкалированием d_{ij} расстояний между объектами:

$$\varphi_{ij} = [t_{ij} - f(r_{ij})]^2. \qquad (7.31)$$

Здесь t_{ij} – воспроизведенные расстояния в пространстве заданной размерности, r_{ij} – исходное расстояние, а $f(r_{ij})$ обозначает функцию неметрического монотонного преобразования. Таким образом, МНШ воспроизводит не количественные меры сходств объектов, а лишь их относительный порядок.

Для измерения качества подгонки модели Дж. Такейном [Takane, 1977; цит. по Терехина, 1986] был предложен *нормированный показатель стресса:* $S = \left(\frac{\|E\|}{\|T\|} \right)^{1/2}$, (7.32)

где норма матрицы $||\ ||$ означает сумму квадратов элементов матрицы. Этот показатель изменяется от 0 до 1: равенство его нулю означает точную подгонку модели, единице – полную ее бессмысленность. Таким образом, чем меньше значение стресса, тем лучше матрица исходных расстояний согласуется с матрицей результирующих расстояний. Кроме стресса Такейна употребимы *нестандартизованный стресс*, *коэффициент стресса Краскела*, использующий простые разности, вместо их квадратов, и *коэффициент отчуждения*. Показателем качества модели является также квадрат коэффициента корреляции между матрицами **T** и **E**, который, как и в регрессионном анализе, может быть интерпретирован как доля дисперсии преобразованных расстояний **T**, объясненная матрицей расстояний **D**.

Результаты анализа обычно представляются на двух- и трехмерных диаграммах рассеяния в пространстве шкал с отмеченными точками-объектами, образуя зрительный образ "экологического пространства" наблюдений или свойств. Уровень согласия может быть оценен с помощью графика ступенчатой функции Шепарда, где оси ординат OY показываются воспроизведенные расстояния, а по оси OX откладываются истинные расстояния между измерениями. Если все воспроизведенные расстояния легли на эту ступенчатую линию, то ранги наблюдаемых расстояний были в точности воспроизведены полученным решением.

Несмотря на то, что имеется много идентичного в характере решения исследуемых проблем, методы МНШ и факторного анализа имеют ряд существенных отличий. Так, факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. Методы МНШ не накладывают таких ограничений: они применимы на любых данных, где сохраняет смысл порядок следования рангов сходств (например, матрица сходства может быть получена с использованием оценок экспертов). С точки зрения различий получаемых результатов, факторный анализ стремится разложить общую дисперсию на большее число факторов (координатных осей или латентных переменных) по сравнению с МНШ, в результате чего МНШ часто приводит к проще интерпретируемым решениям.

Результаты расчетов

В табл. 7.5 раздела 7.4 была представлена матрица расстояний Махалонобиса между 10 реками Самарской области, рассчитанная по гидробиологическим показателям. Выполним нелинейное преобразование этой матрицы и отображение координат водных объектов в пространстве двух шкал. Параллельно осуществим многомерное шкалирование матрицы из F-критериев Фишера, соответствующих T^2-критериям Хоттелинга из той же табл. 7.5.

Технически итеративный процесс шкалирования по описанному примеру дал хорошие результаты согласия между исходной и преобразованной функциями расстояния. Значения нормированного стресса по Такейну составили 0.1409 для меры Махалонобиса и 0.1063 для критерия Фишера. Диаграмма Шеппарда, представленная на рис. 7.13, показывает достаточно незначительные отклонения от графика ступенчатой функции, что свидетельствует о хорошем качестве подгонки модели. По крайней мере, можно утверждать, что преобразование исходной матрицы расстояний

размерностью 11×11 в матрицу координат объектов 11×2 произошло без существенной потери информации.

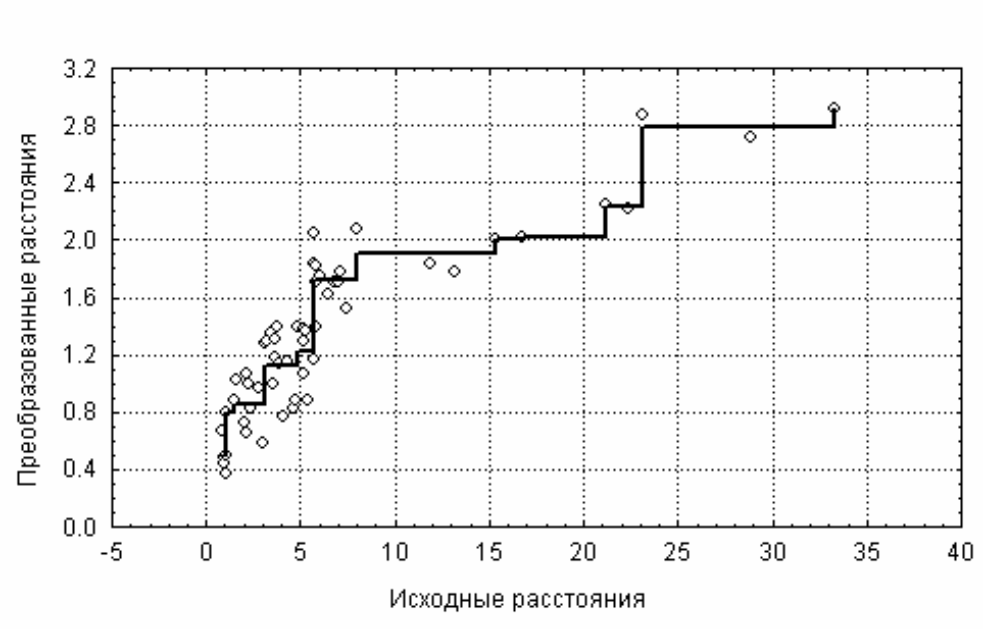


Рис. 7.13. Ступенчатая диаграмма Шеппарда для оценки качества многомерного шкалирования

Полученные двухмерные диаграммы рассеяния, представленные на рис. 7.14 («а», «б»), описывают другими средствами ту же предметную сущность, что и дендрограммы на рис. 7.6 («а», «б»). Однако, на наш взгляд, устойчивость и обоснованность решений, полученных в представленном примере методами МНШ, существенно выше, чем у методов кластерного анализа.

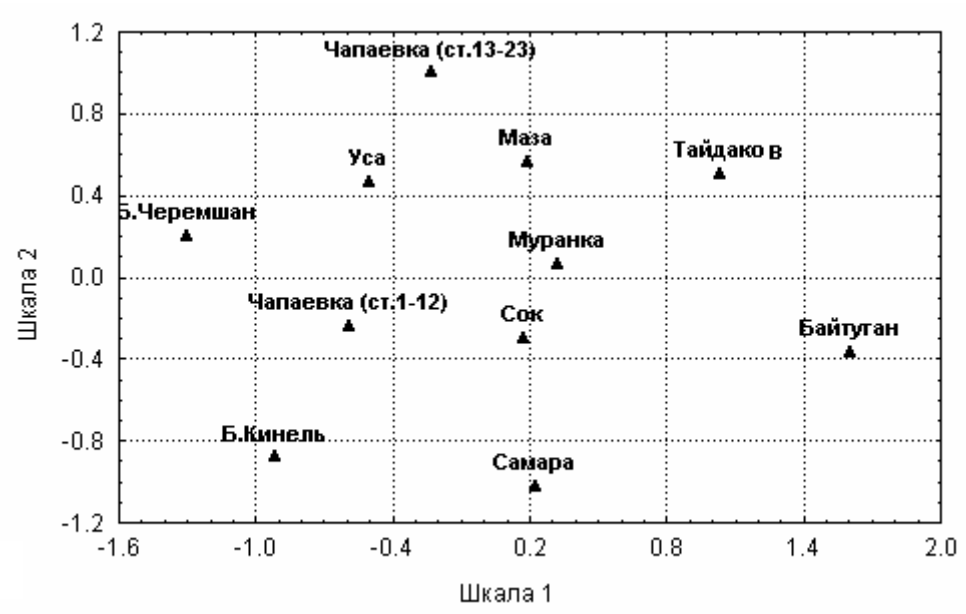
7.7. Общая методика анализа водных объектов по многомерным данным гидробиологического мониторинга (вместо резюме)

Обобщая изложенное в разделах 7.4-7.6, можно предложить следующую методику классификации групп многомерных гидроэкологических измерений, основные этапы которой рассмотрим на примере оценки сходства 14 станций наблюдений по течению р. Сок.

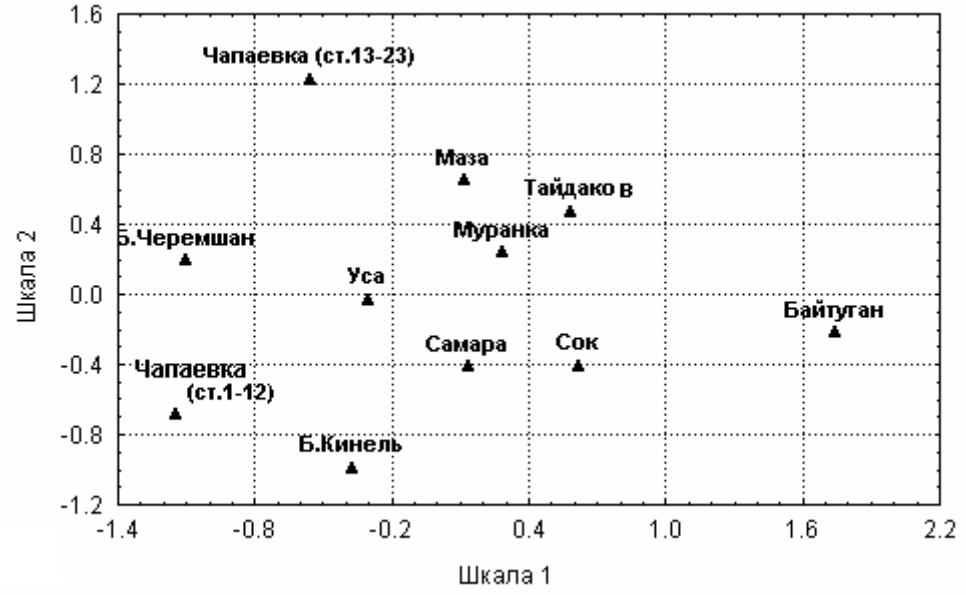
1. Формирование матрицы исходных данных

Методы анализа многомерных наблюдений связаны с определенными ограничениями, как по размеру обрабатываемых матриц, так и по характеру распределения значений признаков. В связи с этим, не представляется возможным использовать для расчетов детализацию гидробиологических переменных на уровне видов: нельзя рассчитать, например, достоверную корреляционную матрицу с использованием столбцов измерений, где встречаемость на всем массиве составляет только 1-2 раза, что для зообентоса далеко не редкость.

Альтернативой информативному пространству видов являются суммарные показатели обилия по систематическим или трофическим группам гидробионтов и другие обобщенные показатели. В качестве исходных данных по 88 наблюдениям, выполненным на р. Сок, будем использовать 11 признаков: логарифмы общих численностей и биомасс по всем видам бентоса, индексы Шеннона, Пареле и Вудивисса и 6 показателей обилия $\ln((\boldsymbol{N}_s*\boldsymbol{B}_s)^{0.5})$ по отдельным подсемействам и трибам хирономид. Объем полученной информации в исходном состоянии: 88∗11 = 968 значений.



а)



б)

Рис. 7.14. Отображение малых рек Самарской области в пространстве двух шкал, полученных на основе расстояний Махалонобиса («а») и критерия Фишера («б»)

2. Снижение признакового пространства путем выделения главных компонент

Сократим размерность переменных с 11 до 3 ведущих факторов, которые объясняют 58.2% статистического варьирования исходных признаков. Интерпретация выделенных главных компонент, нагрузки которых на исходные переменные представлены в табл. 7.7, достаточно очевидна: первый фактор связан с общим обилием зообентоса и наиболее массовыми видами хирономид, второй фактор трактуется как биоразнообразие, третий фактор объединяет более редкие группы хирономид. Можно лишний раз обратить внимание читателя на отрицательный вклад олигохетного индекса в формирование обобщенных переменных.

Объем информации после факторизации составил 88*3 = 264 значения.

Таблица 7.7

Значения факторных нагрузок по трем главным компонентам, рассчитанным по обобщенным индексам и показателям обилия группп хирономид на р. Сок (жирным шрифтом отмечены нагрузки, превышающие 0.5)

Обобщенные индексы и показатели обилия групп	Главные компоненты		
	1	2	3
Численность зообентоса N_s	**0.717**	0.133	0.453
Биомасса зообентоса B_s	**0.833**	0.074	-0.077
Индекс Шеннона H	0.225	**0.783**	-0.089
Индекс Пареле P	-0.212	**-0.598**	-0.106
Индекс Вудивисса V	-0.085	**0.714**	-0.081
Триба Chironomini	**0.840**	-0.040	-0.139
Триба Tanytarsini	0.377	0.346	**0.653**
Подсемейство Diamesinae	0.025	-0.261	**0.652**
Подсемейство Orthocladiinae	-0.119	**0.605**	**0.610**
Подсемейство Prodiamesinae	-0.090	-0.004	**0.574**
Подсемейство Tanypodinae	**0.574**	0.133	0.044

3. Дисперсионный анализ факторов

Значения трех главных факторов, вычисленных для каждого из 88 измерений, могут явиться основой для дисперсионного анализа, где проверяется нулевая гипотеза об отсутствии влияния группировки по станциям на характер выполненных наблюдений. Эта гипотеза не отвергается при $F(13,74) = 1.05$ и $p=0.41$ для фактора 2, т.к. на всех станциях зафиксирован приблизительно одинаковый уровень биоразнообразия, которое мы связали с этим фактором. Для остальных двух компонент факторной модели влияние распределения по станциям оказалось достоверным с высоким уровнем значимости и величиной F-критерия, равной 3.4 и 4.2 для фактора 1 и 3, соответственно.

Анализ пространственной динамики групповых средних значений факторов, представленной на рис. 7.15, дает возможность сделать вывод о существенных сдвигах в видовой структуре хирономид вдоль течения реки от истока к устью, где постепенно выпадают виды Tanytarsini, Diamesinae и Prodiamesinae, замещаемые видами Chironomini на фоне общего увеличения обилия зообентоса.

4. Формирование матрицы расстояний

Рассчитаем для каждой пары станций р. Сок статистики, отражающие расстояние между "центрами тяжести" соответствующих выборок (меры Махалонобиса и T-критерии Хотеллинга) и достоверность различий между ними (F-критерии и значения вероятностей p) в факторизованном пространстве.

Смысл преобразований в главные компоненты, выполненных на втором этапе, заключался не сколько в том, чтобы уменьшить объем вычислений, а чтобы обеспечить устойчивую невырожденность ковариационных матриц. Например, виды подсемейства Diamesinae встречались только на станциях 1-4, что означает полную невозможность расчета расстояния Махалонобиса в исходном пространстве признаков для пар станций от 5 до 14 из-за проблем с нахождением обратных матриц. Использование преобразованных факторов практически исключает эти проблемы, что позволило нам гладко вычислить уровни различий между парами всех 14 станций р. Сок, представленные в табл. 7.8.

Объем информации, заключенной в матрице расстояний с учетом ее симметричности: $14*(14-1)/2 = 91$ значение.

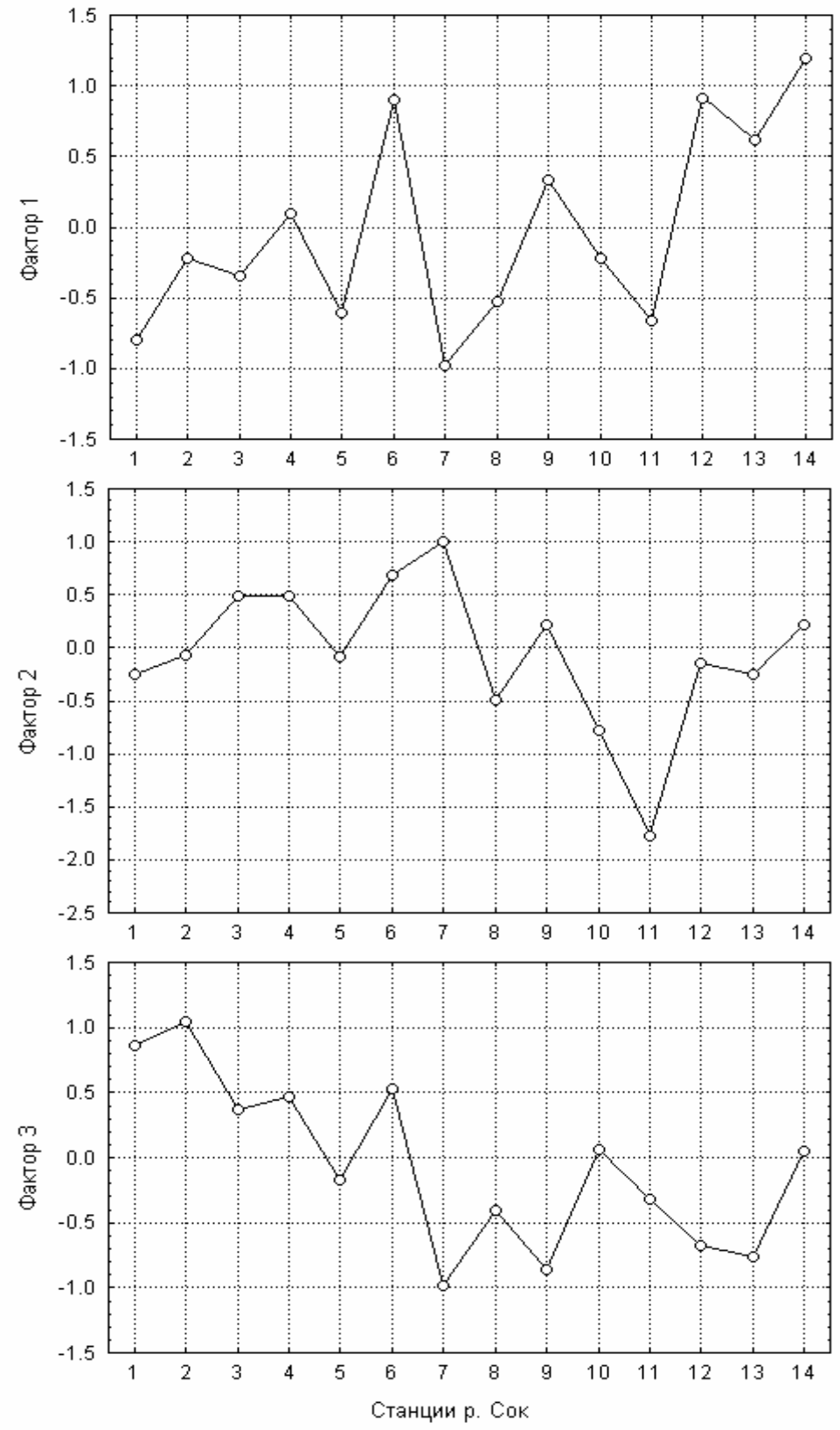


Рис. 7.15. Изменение групповых средних значений главных компонент по станциям от истока к устью р. Сок

Таблица 7.8

*Значения T^2-критерия Хоттелинга (выше главной диагонали) и вероятностей **p** оценки различий по **F**-критерию (ниже главной диагонали) между парами станций р. Сок по комплексу гидробиологических показателей*

№	***n***	Номера станций наблюдений на р. Сок													
ст.	изм.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	8		2.9	9.0	19.5	6.4	14.3	2.8	5.4	18.3	2.5	8.3	28.0	24.0	21.3
2	9	**0.497**		8.5	7.5	20.3	4.6	8.6	15.0	31.9	5.6	6.6	42.9	41.5	19.1
3	11	**0.086**	**0.096**		2.8	4.8	5.3	9.5	6.5	29.4	4.2	7.6	33.7	31.2	28.3
4	9	0.011	**0.135**	**0.492**		10.0	3.1	1.5	8.9	22.4	9.2	15.6	30.3	28.8	15.9
5	11	**0.175**	0.006	**0.264**	**0.064**		8.3	6.2	1.7	21.3	1.5	2.7	22.3	19.9	21.1
6	1	**0.110**	**0.401**	**0.309**	**0.551**	**0.165**		5.9	12.2	4.0	2.6	4.6	4.9	4.8	2.1
7	1	**0.611**	**0.194**	**0.130**	**0.773**	**0.253**	*		3.6	3.4	3.5	4.1	11.0	4.9	6.4
8	5	**0.285**	0.037	**0.191**	**0.121**	**0.697**	**0.346**	**0.676**		6.0	3.4	0.6	8.8	6.4	12.3
9	6	0.022	0.003	0.002	0.009	0.008	**0.570**	**0.617**	**0.285**		6.1	5.4	3.9	1.7	5.3
10	2	**0.630**	**0.307**	**0.383**	**0.154**	**0.757**	*	*	**0.618**	**0.377**		1.2	5.9	5.3	10.1
11	1	**0.235**	**0.276**	**0.189**	**0.074**	**0.563**	*	*	**0.955**	**0.474**	*		4.8	4.0	7.3
12	11	0.002	0.0	0.000	0.001	0.003	**0.338**	**0.099**	**0.107**	**0.371**	**0.256**	**0.344**		0.6	5.1
13	10	0.004	0.0	0.001	0.002	0.006	**0.360**	**0.359**	**0.207**	**0.702**	**0.310**	**0.435**	**0.909**		4.9
14	3	0.029	0.029	0.005	0.045	0.014	*	*	**0.178**	**0.383**	**0.586**	*	**0.295**	**0.325**	

Примечания: 1. Жирным шрифтом отмечены вероятности, где различия недостоверны. 2. Для клеток, отмеченных *, значения вероятностей не определены, а вместо расстояния Махаланобиса рассчитывались квадраты евклидовой дистанции.

5. Многомерное шкалирование

В качестве коэффициентов матрицы сходства между комплексами измерений на различных станциях р. Сок можно выбрать любую из четырех статистик, полученных на этапе 4, поскольку для метода МНШ важны не абсолютные значения расстояний, а их взаимная предупорядоченность. На этот раз выполним многомерное шкалирование с использованием матрицы расстояний, составленной из значений критерия Хоттелинга (см. табл. 7.8).

Полученное решение с достаточной степенью надежности воспроизводит взаимную упорядоченность объектов, заданную в исходной метрике: нормированный коэффициент стресса равен 0.163. Представленные визуально на рис. 7.16 координаты расположения станций в пространстве двух основных шкал позволяют сделать вывод о закономерном изменении гидробиологической обстановки вдоль течения р. Сок: номера станций почти монотонно возрастают вдоль шкалы 1.

Окончательное количество информации после трех этапов сжатия составило только $14*2 = 28$ значений, т.е. около 3% от исходного объема.

Если по той же матрице расстояний, составленной из Т-статистик Хотеллинга, выполнить иерархический кластерный анализ с использованием, например, алгоритма Уорда, то получим дендрограмму, изображенную на рис. 7.17. Обе формы представления одной и той же сущности в некотором смысле похожи, но далеко не идентичны, поскольку подходят к проблеме классификации объектов с различных концептуальных позиций.

Сравнение дендрограммы на рис. 7.17, полученной с использованием всего массива наблюдений, и аналогичных дендрограмм на рис. 7.4, где каждую станцию представляло только одно, случайным образом взятое наблюдение, свидетельствует о несомненном преемуществе статистических методов классификационных построений в гидробиологии.

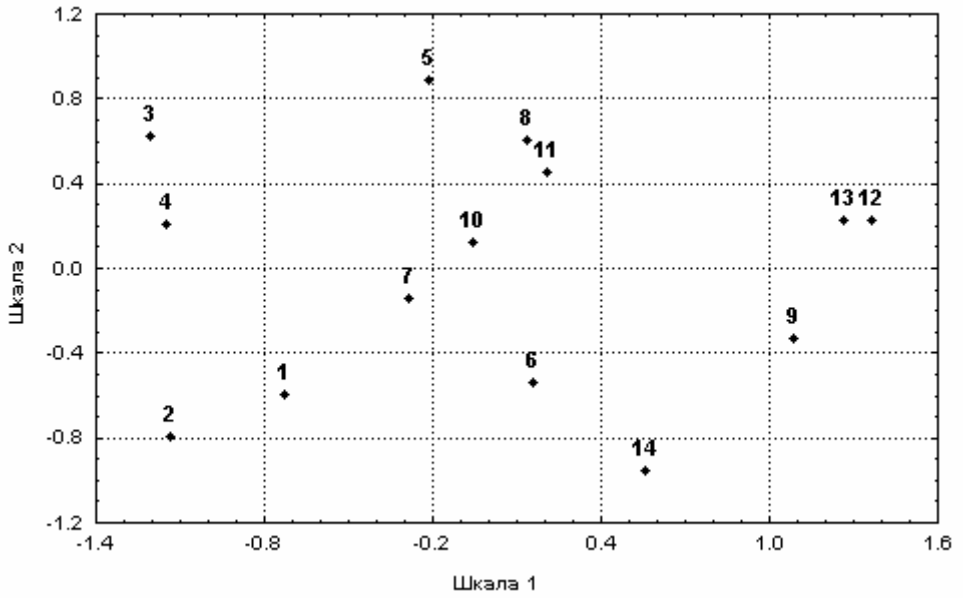


Рис. 7.16. Отображение станций р. Сок в пространстве двух шкал, полученных с использованием статистики Хотеллинга

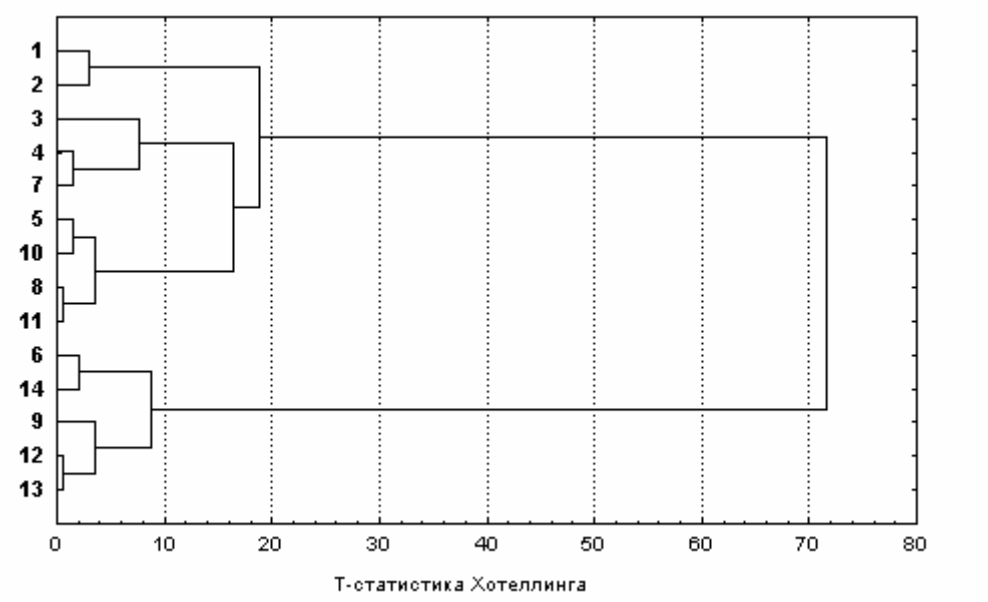


Рис. 7.17. Дендрограмма классификации станций наблюдения р. Сок по пробам зообентоса, выполненная с использованием метода Уорда и Т-критерия

Глава 8. Задача о классе качества вод: прогноз отклика по многомерным эмпирическим данным

8.1. Модель множественной регрессии

Формулировка задачи

Пусть задано пространство признаков X' размерностью $p > 1$, точками которого являются конкретные измерения $x = \{x_1, \ldots, x_j, \ldots, x_p\}$, где x_j – значение j-го гидробиологического показателя в пробе или некоторого параметра среды, сопутствующего наблюдению. Предположим, что в матрице X' один из столбцов считается объясняемой переменной или откликом y, а некоторое количество остальных признаков m, $m < p$, $m > 1$, определены нами как объясняющие или варьируемые переменные. Если массив наблюдений статистически репрезентативен, то можно сформировать обучающую выборку в виде матрицы независимых переменных $X \to x_{ij}$, $j = 1,2,\ldots,m$, и сопряженного с ней вектора-столбца $Y \to y_i$, где $i = 1,2,\ldots,n$ – количество строк измерений ($n > m$), для которых все значения численно определены. Если не оговорено противное, то матрица X дополняется столбцом с фиктивной переменной, тождественно равной единице, $x_{i1} \equiv 1$, что обеспечит вычисление свободного члена регрессии.

Необходимо сконструировать уравнение, выражающее закон изменения отклика Y в зависимости от конкретных значений независимых переменных $X \to x_{ij}$.

По аналогии с одномерной линейной регрессией (см. раздел 5.5) будем предполагать, что модель наблюдений имеет вид

$$y_i = \theta_1 x_{i1} + \ldots + \theta_m x_{im} + \varepsilon_i, \; i = 1,\ldots,n, \; n \geq m, \tag{8.1}$$

где y_i – значение объясняемой переменной в i-м наблюдении; x_{ij} – известное значение j-ой объясняющей переменной в i-м наблюдении; θ_j – неизвестный коэффициент при j-ой объясняющей переменной; ε_j – случайная составляющая ("ошибка") модели для i-го наблюдения.

Использование множественного регрессионного анализа имеет чрезвычайно широкие возможности для обработки таблиц гидробиологических наблюдений, содержащих, как правило, десятки и сотни потенциальных переменных. В разделе 2.7 была выполнена формальная постановка задачи оценки качества воды, где в "роли" отклика Y фигурировал некоторый показатель произвольной этиологии. Комбинируя факторы в различных сочетаниях, можно, дав простор фантазии, построить тысячи различных вариантов регрессионных моделей и доказать с их помощью любые, в том числе и диаметрально противоположные гипотезы о механизмах функционирования экосистем.

Рекомендуемая литература: [Хальд, 1956; Андерсен, 1963; Себер, 1980; Дрейпер, Смит, 1986; Дюк, 1997; Айвазян, Мхитарян, 1998; С.А. Прохоров, 2001а,б, 2002].

Математический лист

Предположения модели и оценивание по методу наименьших квадратов

Нормальная линейная модель множественной регрессии переменной y с m объясняющими переменными $x_1, \ldots, x_m$ основана на следующих предположениях:

- значения $\varepsilon_1,\ldots,\varepsilon_m$ в формуле (8.1) представляют собой случайные величины, независимые в совокупности, имеющие одинаковые нормальные распределения $N(0,\sigma_i^2)$ с нулевым математическим ожиданием и дисперсиями $\sigma_i > 0$;
- $y_1,\ldots,y_n$ являются наблюдаемыми значениями нормально распределенных случайных величин $Y_1,\ldots,Y_n$, которые независимы в совокупности, и для которых

$$E(Y_i) = \theta_1 x_{i1} + \ldots + \theta_m x_{im}, \qquad D(Y_i) = \sigma^2,$$

так что $Y_i \sim N(\theta_1 x_{i1} + \ldots + \theta_m x_{im}, \sigma^2),\ i = 1,\ldots,n.$;

- в отличие от $\varepsilon_1,\ldots,\varepsilon_n$, случайные величины $Y_1,\ldots,Y_n$ имеют распределения, отличающиеся сдвигами.

Термин «множественная» указывает на использование в правой части модели наблюдений двух и более объясняющих переменных, отличных от постоянного члена.

Оценивание неизвестных коэффициентов модели *методом наименьших квадратов* (МНК) состоит в минимизации по всем возможным значениям $\theta_1,\ldots,\theta_m$ суммы квадратов

$$Q(\theta_1,\ldots\theta_m) = \sum_{i=1}^{n} (y_i - \theta_1 x_{i1} - \ldots - \theta_m x_{i\,\mathrm{m}})^2 \to \min. \tag{8.2}$$

Для поиска значений коэффициентов $\hat{\theta}_1,\ldots,\hat{\theta}_m$, минимизирующих эту сумму, необходимо решить систему из m нормальных линейных уравнений с m неизвестными, которая в векторно-матричной форме имеет вид:

$$X^T X\hat{\theta} = X^T y\,, \tag{8.3}$$

где: $X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{n\,\mathrm{m}} \end{pmatrix}$ – матрица значений m объясняющих переменных в n наблюдениях; $\mathbf{X}^{\mathrm{T}}$ – та же матрица в транспонированном виде;

$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ и $\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_m \end{pmatrix}$ – соответственно, вектор-столбец значений объясняемой переменной в n наблюдениях и вектор-столбец оценок m неизвестных коэффициентов. Система нормальных уравнений (8.3) имеет единственное решение

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T y\ , \tag{8.4}$$

если матрица $X^{\mathrm{T}}X$ не вырождена, т.е. ее определитель отличен от нуля ($\det X^T X \neq 0$), что соответствует линейной независимости столбцов матрицы X.

Обозначив расчетные (т.е. подобранные – fitted) значения объясняющей переменной по оцененной линейной модели связи как $\hat{y}_i = \hat{\theta}_1 x_{i1} + \ldots + \hat{\theta}_m x_{im}$, $i = 1,\ldots,n,$ и остаток (residual) для i-го наблюдения как $e_i = y_i - \hat{y}_i$, получим остаточную сумму квадратов:

$$Q_e = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} e_i^2. \tag{8.5}$$

Если рассчитать полную сумму квадратов отклонений

$$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{8.6}$$

и объясненную моделью (explained) сумму квадратов

$$Q_X = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2\ , \tag{8.7}$$

то, как и в случае простой линейной регрессии с $m = 2$, все эти три суммы квадратов связаны соотношением

$$\boldsymbol{Q} = \boldsymbol{Q_x} + \boldsymbol{Q_e}\,, \tag{8.8}$$

которое представляет собой разложение полной суммы квадратов. Коэффициент детерминации модели регрессии $\boldsymbol{R^2}$ определяется как

$$\boldsymbol{R}^2 = 1 - \frac{\boldsymbol{Q_e}}{\boldsymbol{Q}} \tag{8.9}$$

и равен $\boldsymbol{R}^2 = r^2_{\boldsymbol{y},\hat{\boldsymbol{y}}}$, где $\boldsymbol{r_{y,\hat{y}}}$ – выборочный множественный коэффициент корреляции между переменными $\boldsymbol{y}$ и $\hat{\boldsymbol{y}}$. Значение $\boldsymbol{R^2}$ монотонно возрастает с ростом числа переменных (регрессоров) в регрессии, что зачастую не означает улучшения качества предсказания. Потому правильнее использовать скорректированный (adjusted) коэффициент детерминации, учитывающий число использованных регрессоров:

$$\boldsymbol{R^2_{adj}} = 1 - \frac{\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 / (\boldsymbol{n} - \boldsymbol{m} - 1)}{\left\|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\right\|^2 / (\boldsymbol{n} - 1)} = 1 - (1 - \boldsymbol{R}^2)\frac{(\boldsymbol{n} - 1)}{(\boldsymbol{n} - \boldsymbol{m} - 1)}\ . \tag{8.10}$$

Проверка адекватности модели

Определяющим для проверки статистической значимости уравнения является то обстоятельство, что в нормальной линейной модели с несколькими объясняющими переменными оценки $\hat{\theta}_1,\ldots,\hat{\theta}_{\boldsymbol{m}}$ коэффициентов $\theta_1,\ldots,\theta_{\boldsymbol{m}}$ как случайные величины имеют нормальные распределения. Это дает возможность проверить гипотезу $\boldsymbol{H}_0 : \theta_{\boldsymbol{j}} = 0$, которая соответствует предположению исследователя о том, что $\boldsymbol{j}$-я объясняющая переменная не имеет существенного значения для интерпретации изменчивости величины переменной $\boldsymbol{y}$ и может быть исключена из модели. Если гипотеза $\boldsymbol{H}_0 : \theta_{\boldsymbol{j}} = 0$ не отвергается, то отношение оценки $\hat{\theta}_{\boldsymbol{j}}$ к его стандартному отклонению $\boldsymbol{s}_{\hat{\theta}_{\boldsymbol{j}}}$ соответствует $\boldsymbol{t}$-распределению Стьюдента с $\boldsymbol{n - m}$ степенями свободы, а критическое множество для уровня значимости α имеет вид

$$\left| \frac{\hat{\theta}_{\boldsymbol{j}}}{\boldsymbol{s}_{\hat{\theta}_{\boldsymbol{j}}}} \right| > \boldsymbol{t}_{1-\frac{\alpha}{2}}(\boldsymbol{n} - \boldsymbol{m})\,. \tag{8.11}$$

В современных компьютерных программах кроме $\boldsymbol{t}$-статистики приводится также $\boldsymbol{P}$–значение – вероятность того, что случайная величина, имеющая распределение Стьюдента с ($\boldsymbol{n} - \boldsymbol{m}$) степенями свободы, примет значение, не меньшее по абсолютной величине, чем наблюденное значение $\left|\hat{\theta}_{\boldsymbol{j}} / \boldsymbol{s}_{\hat{\theta}_{\boldsymbol{j}}}\right|$. Если указываемое $\boldsymbol{P}$-значение меньше выбранного уровня значимости α, то это равносильно тому, что значение $\boldsymbol{t}$-статистики $\hat{\theta}_{\boldsymbol{j}} / \boldsymbol{s}_{\hat{\theta}_{\boldsymbol{j}}}$ попало в область отвержения гипотезы $\boldsymbol{H}_0$, т.е. $\left|\hat{\theta}_{\boldsymbol{j}} / \boldsymbol{s}_{\hat{\theta}_{\boldsymbol{j}}}\right| > \boldsymbol{t}_{1-\frac{\alpha}{2}}(\boldsymbol{n} - \boldsymbol{m})$. В этом случае параметр $\theta_{\boldsymbol{j}}$ *статистически значим* и наличие $\boldsymbol{j}$-й объясняющей переменной в правой части модели существенно для описания наблюдаемой изменчивости объясняемой переменной.

Кроме того, полезно проверить гипотезу об информационной способности всей модели в целом (или, другими словами, *гипотезу об общей значимости регрессии* в рамках нормальной линейной модели $\boldsymbol{H}_0 : \theta_2 = \theta_3 = \ldots = \theta_{\boldsymbol{m}} = 0$) с использованием $\boldsymbol{F}$-статистики, которая основана на отношении регрессионной суммы квадратов $\boldsymbol{Q_x}$ к остаточной сумме квадратов $\boldsymbol{Q_e}$:

$$\boldsymbol{F} = \frac{\boldsymbol{Q}_X / (\boldsymbol{m} - 1)}{\boldsymbol{Q_e} / (\boldsymbol{n} - \boldsymbol{m})}. \tag{8.12}$$

Действительно, чем больше отношение $\boldsymbol{Q_x} / \boldsymbol{Q_e}$, тем больше есть оснований говорить о том, что совокупность переменных $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\boldsymbol{m}}$ действительно объясняет изменчивость отклика

Y. Если выполняются перечисленные выше предположения, то F-статистика, рассматриваемая как случайная величина, имеет при гипотезе H_0 стандартное F-распределение Фишера с $(m - 1)$ и $(n - m)$ степенями свободы. В соответствии с этим, гипотеза $H_0 : \theta_2 = \theta_3 = \ldots = \theta_m = 0$ отвергается при "слишком больших" значениях F, превышающих пороговое значение при заданном уровне α значимости:

$$F = \frac{Q_X/(m-1)}{Q_e/(n-m)} > F_{1-\alpha}(m-1, n-m). \qquad (8.13)$$

При этом вероятность ошибочного отвержения гипотезы H_0 равна α. Статистические пакеты, выполняющие регрессионный анализ, приводят кроме F -статистики соответствующее ей P-значение, т.е. оценивается вероятность $P\{F(p-1, n-p) > F\}$. Если P-значение меньше заданного уровня значимости (равного, например, $\alpha = 0.05$), то уравнение регрессии считается информативным или "значимым в целом". Можно также отметить такой неслучайный факт, что при анализе модели простой (парной) линейной регрессии ($p = 2$) вычисленные P-значения F-статистик совпадают с P-значениями t-статистик, используемых для проверки гипотезы $\theta_2 = 0$.

Переменные, включаемые в модели

"Фиктивные переменные" используются как противоположность "значащим переменным", показывающим уровень количественного показателя, принимающего значения из непрерывного интервала. Как правило, фиктивная переменная – это индикаторная переменная, отражающая некоторую качественную характеристику. Например, сезонные фиктивные переменные принимают разные значения в зависимости от того, какому месяцу или кварталу года или какому дню недели соответствует наблюдение. Часто применяются бинарные фиктивные переменные, принимающие два значения, 0 и 1, в зависимости от определенного условия. Например, в результате моделирования 0 может означать, что наблюдение принадлежит к "грязным" водоемам, а 1 – к "чистым". Фиктивные переменные, будучи экзогенными, не создают каких-либо трудностей при применении МНК и являются эффективным инструментом построения регрессионных моделей и проверки гипотез.

НАЛИМОВ
Василий Васильевич
(1910-1997),
математик, крупный специалист по наукометрии и теории эксперимента

В линейной регрессионной модели математическое ожидание зависимой переменной – это линейная комбинация регрессоров с неизвестными коэффициентами, которые и являются оцениваемыми параметрами модели. Такая модель является *линейной по виду* и в матричной форме ее можно записать как $Y = X\theta + \varepsilon$. Однако не обязательно, чтобы влияющие на Y факторы входили в модель линейно – регрессорами могут быть любые точно заданные (не содержащие неизвестных параметров) функции исходных факторов – это не меняет свойств МНК:

$$y_i = \theta_0 + \sum_{j=1}^{k} \theta_j \varphi_j(x^i) + \varepsilon_i, \quad i = 1, \ldots, n, \qquad (8.14)$$

где $\varphi_j(\,)$, $j = 1, \ldots, m$ – система некоторых функций.

Для применения метода наименьших квадратов важно, чтобы выполнялись два условия:

- каждую из функций $\varphi_j(\,)$ можно переобозначить как новую переменную, т.е. преобразованием уравнения регрессии ее можно привести к *внутренне* линейному виду;
- ошибка уравнения регрессии оставалась *аддитивной*, то есть, чтобы зависимая переменная являлась суммой своего математического ожидания и ошибки.

Линеаризация не должна быть самоцелью – может случиться, что "истинная" модель бывает настолько нелинейной, что приходится пожертвовать удобствами общего МНК и использовать нелинейные методы оценивания параметров.

Основное преимущество учета нелинейности для простых моделей связано с компенсацией гомоскедастичности при расчете параметров уравнения регрессии. В методе наименьших квадратов все наблюдения выступают в одинаковых "весовых категориях", и поэтому в оценках непропорционально мало используется информация от признаков с меньшей дисперсией. В то же время, многие гидробиологические переменные таковы, что размер отклонений, связанных с ними, пропорционально зависит от величины самих переменных и возникающая при этом гетероскедастичность снижает эффективность оценок параметров. Например, логарифмирование численности или биомассы существенно может улучшить адекватность расчетного уравнения.

Расширение переменных уравнения регрессии за счет использования различных функциональных генераций исходных признаков позволяет также уменьшить недоопределенность модели, когда сложность структуры аппроксимирующей функции недостаточна для отображения сложности изучаемого динамического процесса (*«Время простых моделей прошло»* – У.Р. Эшби). Например, в разделе 2.7 было показано, что огромное большинство процессов в природе может быть описано в виде полиномов высокой степени, являющихся частным случаем обобщенного полинома Колмогорова – Габора (1.13):

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_i a_j a_k x_i x_j x_k + \ldots \quad .$$

Поэтому основная задача моделирования сложных систем с использованием регрессионных уравнений заключается в том, чтобы "обнулить" (вычеркнуть) в полиноме Колмогорова – Габора подмножество "лишних" или незначимых коэффициентов и сохранить оптимальное сочетание объясняющих членов (выражаясь образным языком – *«отсечь все лишнее от глыбы мрамора, превратив ее в статую»*).

Кроме натуральных степеней исходных переменных можно использовать и другие функции от них: $\ln X$, $\sqrt{X}$, $1/X$, $e^{\alpha X}$, тригонометрические преобразования, логистическую функцию $1/(1+e^{-X})$, преобразование Бокса-Кокса $\frac{X^{\alpha}-1}{\alpha}$ и т.д. В качестве примера использования такого подхода в гидробиологии можно привести работу В.А. Тереховой с соавторами [1998].

Методы структурной идентификации моделей.

Обычно исследователь обладает достаточной свободой при выборе функциональной формы модели. Важно лишь, чтобы при этом не нарушались те условия, которые необходимы для хорошей работы применяемых методов оценивания. Но при этом нужно не забывать проводить проверку правильности спецификации модели и исправлять уравнение, когда получена плохая диагностика (например, исключать или добавлять одночлены более высоких степеней в полиномиальную модель).

Принцип множественности моделей утверждает, что для сложных систем по экспериментальным данным нельзя ограничиваться одной единственной моделью. Для каждого объекта, рассматриваемого как некоторый черный ящик, можно найти бесконечное множество уравнений, имеющих одинаковые или почти одинаковые внешние проявления. Однако логика научных исследований требует селекции одной или нескольких моделей регрессии оптимальной или субоптимальной структуры. Для решения этой задачи генерируются определенные наборы уравнений различной сложности и отбираются лучшие из них по некоторому целесообразно заданному критерию регуляризации.

Большинство таких критериев стремится найти компромисс между сложностью и лаконизмом. Иными словами, в уравнение регрессии включается только то минимальное подмножество входных информативных переменных x, которое без существенной потери информации позволяет объяснить имеющийся статистический разброс.

Стандартная пошаговая процедура "включений с исключениями", впервые описанная в работе М.А. Эфроимсона [Efroimson, 1960; Афифи, Эйзен, 1982; Дрейпер, Смит, 1986], и базирующаяся на общей идее метода наименьших квадратов, позволяет с заданной надежностью выбрать из полной матрицы стандартизированных нормальных уравнений наилучшую невырожденную подматрицу, т.е. выбрать модель наиболее оптимальной структуры. Включение и исключение

переменных в модель осуществляется с использованием некоторой статистики – ***t***-критерия для проверки равенства нулю частного коэффициента корреляции. Квадрат этого критерия имеет ***F***-распределение и поэтому называется *последовательным* (или *частным*) ***F-критерием Фишера для включения*** (либо *исключения*).

Выбор первой переменной для включения в модель осуществляется для признака x_l, который имеет наибольший по абсолютной величине коэффициент парной корреляции с откликом r_{ql}. При этом процедура включения выполняется, если справедливо неравенство для последовательного ***F***-критерия: $F > F_o$, где F_o – наперед заданное исследователем пороговое значение. Процесс расширения количества переменных модели повторяется многократно, пока статистическая значимость включения очередного признака по ***F***-критерию на каждом шаге превышает заданный порог F_o. После очередного расширения модели анализируется взаимная коррелированность отобранных переменных и, если их взаимосвязь существенна, то лишние факторы, вносящие наименьший вклад, из модели исключаются. Более точно, исключению подлежат те переменные, для которых вычисленное значение частного ***F***-критерия меньше F_o. Вычисления прекращаются, если не осталось ни одной переменной, для которой вычисленное значение последовательного ***F***-критерия превысило бы заданный порог.

Робастные методы регрессии

Если распределение ошибок в регрессии отличается от нормального, то это не приводит к таким серьезным последствиям, как несостоятельность оценок. Все же на нормальность рекомендуется обращать внимание, т.к. если распределение ошибок имеет "толстые хвосты" или сильно асимметрично, то метод наименьших квадратов может давать не очень точные оценки. Кроме того, отсутствие нормальности означает, что вычисляемые ***t-*** и ***F***-статистики не распределены в конечных выборках точно как ***t*** и ***F***. Хотя эти статистики остаются состоятельными, но при сильном отклонении от нормальности асимптотическое приближение может быть очень неточным, особенно, если размер выборки мал. Использование, так называемых, *робастных* методов оценивания позволяет повысить эффективность регрессионного анализа.

В *медианной регрессии* оценки получаются минимизацией суммы модулей отклонений, а не суммы квадратов отклонений, как в методе наименьших квадратов, что делает расчеты более устойчивыми к "аномальным выбросам" измерений.

В отличие от обычной регрессии, *квантильная регрессия* оценивает не математическое ожидание зависимой переменной, а одну из квантилей.

Метод *инструментальных переменных* применяется в случае, когда ошибка в регрессии может быть скоррелирована с некоторыми из регрессоров. Чаще всего его используют для оценивания отдельного уравнения из системы одновременных уравнений. В этом контексте он известен как *двухшаговый метод наименьших квадратов*.

В линейной регрессии с *мультипликативной гетероскедастичностью* дисперсия ошибки равна $e^{(Z(i)\omega)}$, где Z – матрица, состоящая из переменных, от которых зависит дисперсия, ω – вектор параметров гетероскедастичности.

Тобит (цензурированная регрессия) – это регрессионная модель, в которой зависимая переменная является *цензурированной*, т.е. зависимая переменная преобразовывается, если она меньше (или больше) некоторой границы. Типичным примером является модель с левым цензурированием в нуле, когда вместо наблюдаемого отклика y^*_i принимается переменная y_i, которая получает значения $y_i = 0$, если $y^*_i < 0$ и $y_i = y^*_i$, если $y^*_i \geq 0$. В отличии от тобита, в модели у*сеченной регрессии* наблюдение целиком исключается, если отклик меньше (или больше) некоторой границы.

Результаты расчетов:

Приведем несколько характерных примеров расчета уравнений регрессии, но прежде необходимо еще раз обратить внимание на то, что эти результаты мы трактуем как "истину" в некоторой ограниченной "области справедливости", а именно – все полученные выводы характерны для изучаемого нами объекта (малых рек Самарской области) и вытекают из конкретного собранного материала со всеми его неточностями и условностями.

Моделирование среднего веса особи

Одной из важнейших характеристик сообществ зообентоса является его размерная структура, выраженная в изменчивости *«средних индивидуальных масс особей* $W_{ср}$*»* (термин из работы В.А. Яковлева [2001] не вполне корректен, т.к. "средний" показатель не может быть одновременно "индивидуальным"). Показатель $W_{ср}$ предлагается [Яковлев, 2001] считать одним из «*фундаментальных свойств водных экосистем»*. Автором найдена основополагающая прямо пропорциональная зависимость между разнообразием (индексом Шеннона H) и $W_{ср}$, «*которая сохраняется даже в антропогенно-нарушенных условиях*». Высказываются гипотезы, что средний вес особи закономерно возрастает, например, от горных ландшафтов к лесистым, от бессточных озер к проточным, от глубоководных участков к мелководьям и т.д.

Будем считать средней массой особи $W_{ср}$ частное от деления суммарной биомассы (мг) на суммарную численность для некоторого подмножества организмов, наблюдаемых в конкретной гидробиологической пробе. Эта масса может быть рассчитана как средняя для вида, трибы, семейства, трофической группы или всего зообентоса в целом. Такое усреднение может показаться неправомочным, как не имеющее гидробиологического смысла, однако можно вспомнить, что весьма представительная теория термодинамики идеальных газов основывается на аналогичном показателе – *кажущейся* (или приведенной) *молекулярной массе* компонентов, входящих в газовую смесь, которую никто не считает неестественной. Разумеется, речь в данном контексте будет идти уже не о реальном весе какого-то конкретного организма, а о некотором интегральном показателе гидробиологического сообщества, напрямую зависящем от текущего соотношения видов.

Сформируем пять выборок, содержащих значения средней массы особи W для всего зообентоса в целом по результатам 540 проб наблюдений, а также отдельно для видов хищников-хватателей (384 измерения), семейств Oligochaeta (418 изм.), Chironomidae (473 изм.) и трибы Chironomini (337 изм.), встретившихся в тех же пробах. Будем искать регрессионную зависимость среднего веса особи от следующих восьми переменных:

- показателей обилия: X_{NB} – логарифма индекса плотности населения $\ln((N_s*B_s)^{1/2})$, где N_s и B_s – суммарные численность и биомасса, и X_S – общего числа S видов зообентоса в пробе;
- традиционных индексов: X_H – информационного индекса Шеннона, X_V – биотического индекса Вудивисса и X_P – олигохетного индекса Пареле;
- фиктивных переменных, измеренных в порядковых шкалах: X_{MS} – сезонной составляющей ($X_{MS} = d / 30$, где d – количество дней с начала года до даты проведения наблюдения, $X_{MS} = 5 \div 10$), X_{KK} – класса качества вод на станции, оцененного по гидрохимическим показателям, $X_{KK} = 2 \div 6$, и X_{TW} – типа водоема в точке отбора пробы, $X_{TW} = 1 \div 6$.

Тип водоема X_{TW} был специфицирован по следующей последовательности категорий: 1 – ручьи и родники, 2 – малые реки возвышенностей, 3 – малые равнинные реки, 4 – средние равнинные реки, 5 – устья, 6 – озера и водохранилища.

По результатам расчетов, представленным в таблице 8.1, можно сделать следующие выводы:

1. Анализ характера распределения средней массы особей зообентоса выявил большую асимметрию $A = 9.2$ и эксцесс $Э = 107$. Предварительное логарифмирование значений W позволяет существенно приблизить закон распределения отклика $\ln(W)$ к нормальному и одновременно улучшить параметры модели – коэффициент детерминации R^2 и оценку значимости регрессионного уравнения по F-критерию.
2. Для большинства полученных уравнений существует значимая прямо пропорциональная зависимость среднего веса особи от показателя обилия X_{NB}. На этом факте можно было бы развить некоторую содержательную гипотезу, объясняющую продукционные механизмы формирования размерной структуры сообществ. Однако здесь нельзя не принять во внимание возможность чисто статистического эффекта: частное от деления двух показателей всегда в определенной степени зависит от одного из них, особенно, если "вариационная эластичность" переменных различается.
3. Отсутствует достоверная связь W с индексом Шеннона X_H, отражающим эквитабельность численности по видам. В то же время, средний вес особей значимо уменьшается с увеличением числа видов в пробе X_S, представляющего биоразнообразие экосистемы в явном виде. Это значит, например, что гипотезу об увеличении W в условиях отсутствия трофической конкуренции будет чрезвычайно сложно опровергнуть.

Таблица 8.1

*Параметры уравнений регрессии, связывающих средний вес особи (**W**) с комплексом варьируемых переменных (обозначения см. по тексту).*
*В шапке таблицы: **M ± m** – среднее и доверительный интервал, **R^2** - коэффициент детерминации,%; **F** – статистика Фишера для оценки значимости регрессии*

Группа	***M* ± *m***	Уравнение регрессии	***R^2 / F***
Все виды зообентоса	53.8 ± 22.6	W = - **270.6** + **58.1·X_{NB}** + 34.2·X_H - **13.87·X_S** - 2.87·X_V - **126.8·X_P** + 4.77·X_{MS} - **30.65·X_{Kk}** + **21.45·X_{TW}**	14.15 11.69
		ln(W) = - **3.86**+ **0.671·X_{NB}** + **0.38·X_H** - **0.126·X_S** - 0.006·X_V - **0.885·X_P** + 0.019·X_{MS} - 0.13·X_{KK} + **0.189·X_{TW}**	33.86 34.22
Хищники хвататели	4.66 ± 1.08	ln(W) = 1 + 0.044·X_{NB} + 0.018·X_H - 0.006·X_S + **0.071·X_V** + 0.164·X_P - 0.059·X_{MS} - 0.0003·X_{KK} - **0.198·X_{TW}**	6.96 4.58
Семейство Oligohaetae	2.95 ± 3.26	ln(W) = -0.55 + **0.139·X_{NB}** + 0.056·X_H - **0.044·X_S** + 0.03·X_V - 0.006·X_P - **0.158·X_{MS}** + 0.067·X_{KK} + 0.029·X_{TW}	8.15 5.63
Семейство Chironomidae	2.01 ± 0.33	ln(W) = -0.93 + **0.148·X_{NB}** + 0.109·X_H - **0.046·X_S** - **0.06·X_V** + **0.375·X_P** - 0.036·X_{MS} + 0.065·X_{KK} + 0.033·X_{TW}	15.53 11.85
Триба Chironomini	1.52 ± 0.18	ln(W) = **1.1** + 0.014·X_{NB} + 0.028·X_H - 0.02·X_S - 0.003·X_V + 0.16·X_P - **0.113·X_{MS}** + 0.066·X_{KK} - **0.149·X_{TW}**	5.26 3.33

Примечание: Жирным шрифтом отмечены статистически значимые коэффициенты по t-критерию.

4. Влияние сезонного фактора X_{MS} значимо проявляется и становится определяющим для динамики среднего веса, вычисленного для отдельных таксономических групп – семейства Oligochaeta и трибы Chironomini. Этот факт имеет вполне объяснимый гидробиологический смысл: в большой группе организмов, таких, как весь бентос или семейство Chironomidae, всегда найдутся подмножества видов, заполняющих экологическую нишу в любой момент вегетационного периода.
5. Зависимость среднего веса особи от категории водоема неоднозначна для различных таксономических групп. Например, хищники и особи видов трибы Chironomini устойчиво крупнее в небольших проточных водоемах. Однако в целом для всего бентоса эта зависимость становится противоположной, что объясняется вкладом моллюсков Unionidae, Gastropoda, Dreissenidae с очень высоким индивидуальным весом особей, которые преобладают в малопроточных глубоководных участках рек.
6. Влияние на $W_{ср}$ уровня загрязнения водоемов, оцениваемого явно классом качества X_{KK} и неявно индексами X_V и X_P, носит достаточно эпизодический и нестационарный характер. Можно усмотреть некоторое уменьшение массы организмов зообентоса в целом в грязных водоемах (по X_P и X_{KK}) и аналогичное увеличение веса хирономид в тех же условиях (по X_V). Для остальных моделей факторы загрязнения оказались незначимыми. В определенной мере такой результат объясняется ощутимой взаимной коррелированностью переменных X_{KK}, X_V и X_P: например, коэффициент парной корреляции между индексом Вудивисса V и классом качества составил -0.539. Произошло своего рода "распыление объясняющей способности" между тремя конкурирующими показателями, что могло негативно сказаться на выводах модели.

Уместно на этом примере заметить, что полезность признака, как регрессора не всегда соответствует величине ***t***-критерия Стьюдента, поскольку последний исходит из предположения о взаимной независимости переменных и не учитывает их взаимную коррелированность.

Пошаговая регрессия для оценки связи гидрохимических и гидробиологических показателей

Искусство регрессионного анализа заключается не в прямолинейной технике расчета коэффициентов уравнений, а в тщательной селекции наиболее существенных регрессоров и учете нелинейного характера их связи с откликом, что объединяется в понятие «***структурной идентификации модели***». Рассмотрим использование пошаговой процедуры в расширенном пространстве переменных на примере анализа связи между гидрохимическими и гидробиологическими показателями.

Сформируем исходный набор признаков из следующих 7 показателей: X_H – информационного индекса Шеннона, X_V – биотического индекса Вудивисса, X_P – олигохетного индекса Пареле, X_{CI} – хирономидного индекса Балушкиной, числа видов X_S, логарифмов суммарной численности X_N и биомассы X_B зообентоса в пробе. Добавим в таблицу признаков столбцы вторичных переменных, которые будем получать за счет всех возможных парных произведений и различных математических функций от всех 7 исходных переменных: X_H^2, $X_H \cdot X_V$, $X_H \cdot X_P$, $X_H \cdot X_P X_{CI}$,…, $1/X_H$, $X_H^{0.5}$, $1/X_H^{0.5}$ и т.д. Общее число варьируемых переменных после преобразования матрицы данных и включения базисных функций увеличивается с 7 до 55.

Зададимся, согласно рекомендациям А. Афифи и С. Эйзена [1982], пороговым значением для частного F-критерия = 3.5 и, используя пошаговый метод включений с исключениями Эфроимсона, получим следующие модели регрессии для различных гидрохимических показателей:

- для концентрации минерального фосфора, мг/л (n = 117):

$$Y_p = 0.0917 - 0.0337 \cdot X_S^{0.5} + 0.012 \cdot X_B ,$$

  т.е. увеличение содержания фосфора сопровождается уменьшением числа видов и увеличением биомассы зообентоса, что характерно для процессов эвтрофирования;

- для концентрации аммонийного азота, мг/л (n = 86):

$$Y_{NH4} = 0.00619 + 0.69 / X_S + 0.00107 \cdot X_S \cdot X_N + 0.074 \cdot X_P \cdot X_N - 0.239 \cdot X_H \cdot X_P;$$

- для биохимического потребления кислорода, мгО^2/л (n = 87):

$$Y_O = 1.096 + 6.58 / X_S + 0.132 \cdot X_H \cdot X_{CI} - 9.22 \cdot X_P + 1.32 / (X_{CI})^{0.5}$$

- для содержания ионов железа, мг/л (n = 67):

$$Y_{Fe} = 0.13 - 0.087 \cdot X_P \cdot X_S + 1.39 \cdot X_P^2 + 0.0039 \cdot X_N \cdot X_B .$$

Преимущество пошаговой процедуры, выполнившей подбор информативной комбинации из 55 признаков, заключается не только в компактности получаемых уравнений, но и в существенном повышении уровня их адекватности. Все представленные модели являются достоверными с высоким уровнем значимости, когда как практически все коэффициенты регрессии аналогичных полных линейных уравнений на основе тех же 7 признаков статистически недостоверны по t-критерию. Читатель может сравнить в табл. 8.2 значения коэффициентов детерминации и F – статистик Фишера для оценки значимости двух типов регрессий: полной линейной и нелинейной с селекцией переменных для моделирования одного и того же отклика.

Таблица 8.2

Сравнительные характеристики регрессионных моделей, связывающих гидрохимические и гидробиологические показатели (R^2 – коэффициент детерминации, %, r – коэффициент множественной корреляции, F – статистика Фишера для оценки значимости регрессии, p - вероятность, соответствующая F-критерию)

Наименование отклика (гидрохимического показателя)	Типы регрессионных моделей					
	На основе индексов и обобщенных показателей				На основе относительной численности таксономических групп	
	Полная линейная модель		Нелинейная модель с селекцией переменных			
	R^2 / r	F / p	R^2 / r	F / p	R^2 / r	F / p
Фосфор минеральный	7.35 0.271	1.775 0.099	9.75 0.312	6.154 ~0.0	32.08 0.567	6.91 ~0.0
Азот аммонийный	4.185 0.205	0.804 0.587	17.49 0.418	4.293 ~0.0	47.65 0.69	8.09 ~0.0
$БПК_5$	5.02 0.224	1.123 0.357	29.87 0.547	6.899 ~0.0	43.51 0.66	10.12 ~0.0
Железо	11.6 0.341	1.675 0.133	28.26 0.532	8.27 ~0.0	43.75 0.661	6.03 ~0.0

Определяющими параметрами в уравнениях, полученных пошаговой процедурой, являются "натуральные" показатели X_S, X_N и X_B, а столь популярные в гидробиологических работах индексы в нашем примере играют роль очевидных статистов. Например, биотический индекс Вудивисса вообще оказался не связанным ни с одним из гидрохимических показателей.

Связь гидрохимических показателей с обилием таксономических групп

Сформируем матрицу варьируемых переменных из относительных долей таксономических групп зообентоса:

$$x_j = \ln(N^s_j) / \sum_{j=1}^{m} \ln(N^s_j) \quad , \tag{8.15}$$

где N^s_j – общая численность особей j-й группы, $j = 1,2,\ldots,m$; m – число таксономических групп, встретившихся в каждой пробе наблюдений.

Группы, как мы это уже ранее делали неоднократно, выделим по двум параллельным градациям – по систематике и трофическому признаку. Генерировать вторичные переменные в данном случае не будем.

Применим пошаговый метод включений с исключениями Эфроимсона для расчета моделей регрессии для тех гидрохимических показателей, что и в предыдущем примере. Ориентируясь на знак коэффициентов рассчитанных уравнений, представленных в табл. 8.3, можно сформулировать гипотезы об увеличении (знак «+») или уменьшении (знак «–») удельной доли таксономических групп в сообществе при возрастании соответствующего гидрохимического показателя. Относительную достоверность каждого предположения можно оценить по величине частных F-критериев, с которыми отдельные признаки включались в модель (пороговое значение в этом случае было равном 2.5). Коэффициенты множественной корреляции и оценки значимости полученных уравнений с использованием статистики Фишера приведены в табл. 8.2.

Выполненные расчеты со всей очевидностью свидетельствуют о том, что комплекс признаков, составленный из абсолютных или относительных значений обилия и представляющий все таксономические группы гидробионтов, значительно адекватнее связан с факторами среды, чем суммарные гидробиологические показатели или отдельные субъективные индексы. Впрочем, этот практический вывод лишь подтверждает естественное умозаключение: таблица численности и биомассы особей, суммированных по таксономическим группам, уже содержит в полном, хоть и неявном виде, всю информацию, содержащуюся в любом из обобщенных индексов. При этом, обратная информационная трансформация является невозможной, поскольку после расчета любого индекса значительная часть данных о структуре биоценоза становится безвозвратно потерянной.

8.2. Регрессия с качественной зависимой переменной

Формулировка задачи

Пусть в рамках задачи множественной регрессии зависимая переменная Y принимает фиксированные значения из некоторого заранее предопределенного набора, т.е. моделируемому объекту приписывается выбор между двумя и более возможными альтернативами. В частности, модель с бинарной переменной включает отклик, принимающий два значения (обычно 0 и 1), а также регрессоры X, которые содержат факторы, определяющие альтернативный выбор.

Эта задача возникает, как правило, если моделируемый показатель измерен в порядковой шкале, которая принципиально не может быть преобразована в непрерывную числовую последовательность. Пусть, например, рассматривается оценка пола особи: мужской (0) или женский (1). Тогда построенная обычная линейная регрессия будет предсказывать абсурдные значения Y – дробные, отрицательные и больше единицы. Может быть, это как-то и интерпретируется с медицинской точки зрения, но в практике гидробиологических исследований такое будет едва ли возможно.

Для случая с качественной зависимой переменной требуется найти модель, которая порождала бы дискретное распределение $E(Y \mid X)$, зависящее от X и хорошо описывающие исходные данные. Классическая модель регрессии не подходит для описания этой ситуации, поскольку предполагает, что зависимая переменная имеет непрерывное распределение.

Таблица 8.3

Коэффициенты регрессионных моделей, связывающих гидрохимические показатели и относительную численность групп зообентоса
*(**n** – количество измерений, **m** – количество таксономических групп)*

Гидрохимические показатели	Коэффициенты уравнения	Наименования таксономических групп зообентоса, относительные численности которых включены в модель	Частный ***F***-критерий
Фосфор минеральный (***n*** = 126, ***m*** = 66)	0.143	Свободный член	-
	0.7	Сестоно-детритофаги фильтраторы / Unionidae	21.39
	- 0.434	Хищники хвататели / Tanypodinae	17.66
	- 0.599	Хищники хвататели / Limoniidae	6.66
	- 0.3	Всеядные собиратели+хвататели / Chironomini	8.24
	- 0.233	Сестоно-детритофаги фильтраторы / Bivalvia	6.71
	- 0.846	Сестоно-детритофаги фильтраторы / Gastropoda	2.92
	-1.38	Сестоно-детритофаги фильтраторы / Simuliidae	3.67
	- 0.77	Фитодетритофаги собиратели / Psychodidae	2.93
Азот аммонийный (***n*** = 90, ***m*** = 61)	0.276	Свободный член	
	11.422	Хищники хвататели / Megaloptera	31.63
	0.362	Детритофаги собиратели / Chironomini	5.21
	- 0. 974	Всеядные собиратели+хвататели / Nematoda	6.56
	0.461	Детрито-фитофаги / Tanytarsini	4.68
	0. 445	Детрито-фитофаги / Chironomini	3.85
	- 0. 797	Хищники хвататели / Tanypodinae	6.28
	-1. 277	Фитодетритофаги собиратели / Ephydridae	5.84
	- 0.251	Детритофаги собиратели / Oligochaeta	4.72
	- 0.445	Хищники хвататели / Prodiamesinae	2.54
$БПК_5$ (***n*** = 100, ***m*** = 46)	3.446	Свободный член	
	37.21	Хищники хвататели / Megaloptera	20.46
	2.845	Детритофаги собиратели / Oligochaeta	17.25
	17.63	Всеядные собиратели+хвататели / Nematoda	14.36
	10.25	Сестоно-детритофаги фильтраторы / Chironomini	8.77
	3.844	Детрито-фитофаги / Chironomini	10.16
	- 4.532	Всеядные собиратели+хвататели / Chironomini	4.32
	27.46	Сестоно-детритофаги фильтраторы / Gastropoda	3.91
Железо (***n*** = 71, ***m*** = 58)	0.054	Свободный член	
	4.1	Детритофаги собиратели / Polychaeta	15.35
	1.3	Хищники хвататели / Prodiamesinae	7.24
	2.52	Хищники хвататели / Diamesinae	7.39
	1.25	Детрито-фитофаги / Chironomini	7.48
	2.67	Хищники хвататели / Limoniidae	5.63
	4.01	Хищники хвататели / Homoptera	5.23
	0.461	Детритофаги собиратели / Oligochaeta	5.73
	0.87	Хищники хвататели / Tanypodinae	2.64

С этой целью рассматривается логистическая регрессия, которая выражает статистическую связь в виде зависимости $\boldsymbol{P}\{\boldsymbol{Y}=1|\boldsymbol{X}\}=\boldsymbol{f}(\boldsymbol{X})$, т.е. прогнозируется вероятность события $\{\boldsymbol{Y}=1\}$, обусловленная значениями независимых переменных $\boldsymbol{X}^1,\ldots,\boldsymbol{X}^p$. Геометрически суть задачи состоит в том, чтобы найти одну из возможных гиперплоскостей, которая бы в определенном смысле наилучшим образом разделяла бы две группы наблюдений (соответствующие 0 и 1) в пространстве регрессоров.

Рекомендуемая литература: [Бикел, Доксам , 1983; Справочник по прикладной.., 1989].

Математический лист

Логистическая регрессия выражает модель связи между откликом и переменными в виде формулы

$$P\{Y=1|X_1,...,X_p\}=\frac{e^{\hat{Y}}}{1+e^{\hat{Y}}}, \tag{8.16}$$

где переменная $\hat{Y} = \theta_0+\theta_1X_1+...+\theta_pX_p$ называется *логитом*. Такая модель с бинарной зависимой переменной, по сути, является функцией логистического закона распределения

$$F(x)=\frac{e^{(x-a)/k}}{1-e^{(x-a)/k}}, \tag{8.17}$$

в которой в качестве аргумента используется линейная комбинация независимых переменных.

Наряду с моделью, имеющей логистически распределенное отклонение, используют также близкую ей модель *пробит* с нормально распределенным отклонением (см. рис 8.1).

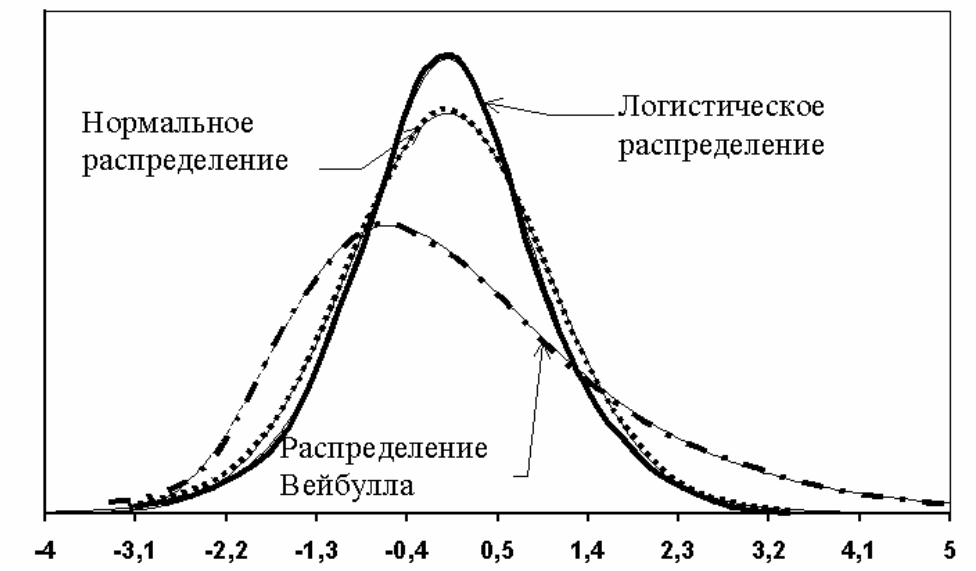


Рис 8.1. Виды распределений, используемых в логистической регрессии

Различить, когда следует применять логит, а когда – пробит, в случае малых выборок невозможно, поскольку оценки коэффициентов θ отличаются множителем, который практически постоянен.

Предлагается два вида моделей выбора, которые могли бы порождать интересующие нас распределения зависимой переменной: *пороговая* модель и модель, основанная на *полезности альтернатив*.

Пороговая модель предполагает, что прогнозирование отклика основывается на ненаблюдаемой непрерывной переменной $\hat{Y}$, математическое ожидание которой является линейной комбинацией набора регрессоров X: $\hat{Y} = X\hat{\theta} + \varepsilon$. Отклик Y, являющийся дискретной величиной, связан с $\hat{Y}$ следующим образом: если $\hat{Y}$ больше некоторой пороговой величины C, то $Y = 1$, если меньше, то $Y = 0$. Как обычно предполагается, что ошибки ε_i имеют нулевое математическое ожидание, одинаково распределены и независимы. Величину C обычно принимают равной 0.5. Пробит- или логит-уравнение $\hat{Y} = X\hat{\theta}$ задает задает в этом случае гиперплоскость, которой разделяются две группы точек: $\hat{Y}_i = X\hat{\theta} < 0.5 \Rightarrow 0$ и $\hat{Y}_i = X\hat{\theta} > 0.5 \Rightarrow 1$.

О качестве модели можно судить по графику оценки $E(Y)$ по $\hat{Y}$, который в случае "хорошей" модели должен быть "крутой" в нуле. На двух графиках, представленных на рис. 8.2, слева внизу и справа вверху расположены правильно предсказанные точки, а слева вверху и справа внизу — неправильно.

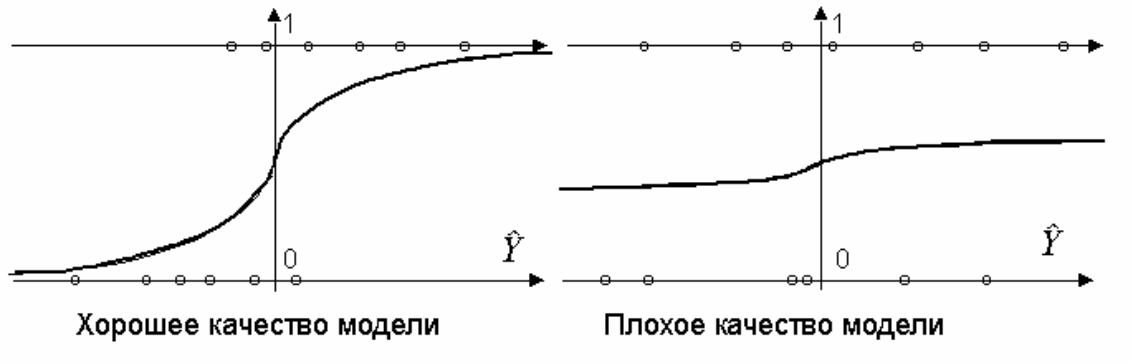


Рис. 8.2. График оценки $E(Y)$ *по* $\hat{Y}$ *для различных моделей с бинарной зависимой переменной*

Множественный логит- или *пробит-анализ* являются естественным продолжением бинарного и возникают, когда рассматривается выбор между более, чем двумя альтернативами. Упорядоченный логит, развивающий пороговую модель, имеет дело с альтернативами, которые можно расположить в определенном порядке. Например, это могут быть шкала оценок класса качества вод, зоны сапробности и т.д.

Будем предполагать, что альтернативы пронумерованы от 0 до S, а переменная Y принимает значение s, если выбрана альтернатива s. Как и в бинарной модели, в основе выбора лежит ненаблюдаемая величина $\hat{Y} = X\theta + \varepsilon$, для ранжирования которой рассчитывается s пороговых значений $\gamma_1, \gamma_2, \ldots, \gamma_s$. Предполагается, что $Y=0$, если $\hat{Y}$ меньше нижнего (первого) порогового значения, $Y = 1$, если $\hat{Y}$ попадает в промежуток от первого до второго порогового значения и т. д.; $Y = S$ выбирается, если $\hat{Y}$ превышает верхнее пороговое значение, т.е.:

$$Y_i = \begin{cases} 0, \text{ если } \hat{Y} \le \gamma_1 \\ 1, \text{ если } \gamma_1 < \hat{Y} \le \gamma_2 \\ \ldots \\ S, \text{ если } \hat{Y} > \gamma_S \end{cases} \qquad (8.18)$$

Коэффициенты пробита и логита обычно оценивают методом максимального правдоподобия, рассмотрение теоретических аспектов которого выходит за рамки нашего изложения (подробно см. [Цыплаков, URL]). Статистика отношения правдоподобия, которая распределена асимптотически как χ^2 с $m-1$ степенями свободы, где m – количество параметров в исходной модели, используется для построения показателя качества модели, аналогичного F-статистике для линейной регрессии, т.е. для проверки гипотезы о том, что коэффициенты при всех регрессорах, кроме константы, равны одновременно нулю. Для моделей с бинарной зависимой переменной можно сконструировать и некий аналог коэффициента детерминации — псевдо-R^2: Однако для логистической регрессии, на наш взгляд, существует наиболее естественный критерий качества – вероятность ошибки при оценке прогнозируемых альтернатив. Понятно, что "хорошая" модель должна давать высокий процент правильных предсказаний.

Если категории прогнозируемого отклика не упорядочены (например, сравниваются наблюдения для различных водоемов), то предполагается, что выбор делается на основе функции полезности альтернатив $u(Y, X)$. Для бинарной модели, если $u(1, X) > u(0, X)$, то выбираем 1, а если $u(0, X) < u(1, X)$, то выбираем 0. Для множественного логита Y_i выбирается равным s, если $u_s(Z_i) > u_t(Z_i)\ \forall\ s \neq t$. При выборе вида функции полезности обычно делают одно из двух упрощающих допущений:

- регрессоры для всех альтернатив одни и те же: $u_s = Z\beta_s + \varepsilon_s$;
- функция имеет один и тот же вид, а меняются только факторы, определяющие выбор, т.е.

$$u_s = Z_s\beta + \varepsilon_s.$$

При этом также принимается, что ошибки ε_s имеют стандартное распределение Вейбулла $F(X) = e^{-e^{-X}}$ (см. рис.8.1).

Результаты расчетов

Бинарная логистическая регрессия на основе показателей обилия групп

Сформируем выборку для построения бинарной логистической модели следующим образом:

- в качестве отклика $\boldsymbol{Y}$ примем альтернативу 0 («чисто»), куда отнесем измерения, сделанные на станциях с классом качества вод 3 и менее, и 1 («грязно»), соответствующую классу 4 и выше;
- в качестве варьируемых переменных примем общее число видов $\boldsymbol{X_1} = \boldsymbol{S}$ и показатели обилия отдельных семейств зообентоса (для хирономид -подсемейств и триб); $\boldsymbol{X_j} = \ln((\boldsymbol{N_{sj}} \cdot \boldsymbol{B_{sj}})^{0.5})$, $\boldsymbol{N_{sj}}$ и $\boldsymbol{B_{sj}}$ – суммарные по видам численность и биомасса $\boldsymbol{j}$-й таксономической группы в пробе, $\boldsymbol{j} = 2,3,\ldots,51$.

Полученное уравнение логистической регрессии оказалось значимым с позиций статистики отношения правдоподобия – $\boldsymbol{\chi^2}(50) = 225.9$, $\boldsymbol{p} \approx 0.0$, а коэффициент детерминации псевдо-$\boldsymbol{R^2}$ составил 33.6%. Верхняя часть списка коэффициентов регрессии, отсортированного по убыванию $\boldsymbol{t}$-статистики, представлена в табл. 8.4. Большинство рассчитанных коэффициентов имеют отрицательный знак, т.е. чем меньше обилие гидробионтов этих семейств, тем больше шанс, что проба взята из водоема «грязной» категории. Достоверная связь обратного характера наблюдается только для организмов Oligochaeta и Odonata.

Таблица 8.4

Коэффициенты логистической регрессионной модели, связывающей категорию качества вод и обилия таксономических групп зообентоса

Наименования таксономических групп зообентоса	Коэффициенты логита	Стандартная ошибка	$\boldsymbol{t}$-статистика	$\boldsymbol{P}$-значение
Ephemeroptera	-0.291	0.066	-4.409	**0.000**
Oligochaeta	0.147	0.047	3.120	**0.002**
Триба Tanytarsini	-0.154	0.065	-2.373	**0.018**
Amphipoda	-0.490	0.208	-2.352	**0.019**
Coleoptera	-0.222	0.098	-2.267	**0.024**
Crustacea	-0.274	0.124	-2.208	**0.028**
Odonata	0.213	0.108	1.974	**0.049**
Gastropoda	-0.093	0.061	-1.525	0.128
Diptera	-0.329	0.216	-1.523	0.128
Подсемейство Diamesinae	0.141	0.100	1.416	0.158
Dreissenidae	-0.094	0.068	-1.380	0.168
Limoniidae	-0.125	0.094	-1.335	0.183
Megaloptera	0.142	0.108	1.318	0.188
Psychodidae	-0.255	0.199	-1.282	0.201
Nematoda	0.113	0.091	1.248	0.213
Simuliidae	-0.135	0.113	-1.197	0.232
Hemiptera	-0.131	0.110	-1.195	0.233
Hidracarina	0.215	0.192	1.119	0.264
Rhagionidae	-0.201	0.180	-1.118	0.264
Число видов в пробе $\boldsymbol{S}$	-0.047	0.044	-1.068	0.286
Триба Chironomini	0.059	0.055	1.060	0.290

Гистограмма распределения примеров выборки по шкале прогнозируемой вероятности класса 1 («грязно») представлена на рис. 8.3. Если принять в качестве порогового значения $\boldsymbol{P} = 0.5$, то к классу «грязных» объектов относят значения $\boldsymbol{X}_i$, для которых $\hat{\boldsymbol{P}}_{\boldsymbol{i}} = \boldsymbol{P}\{\boldsymbol{Y} = 1 \big| \boldsymbol{X}_{\boldsymbol{i}}^{1},\ldots,\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{m}}\} > 0.5$, и тогда общая ошибка предсказания по уравнению регрессии составит менее 20%. Это правило оптимально с точки зрения минимизации числа ошибок, но не всегда верно с точки зрения исследования связи и поэтому порог зачастую сдвигают в сторону класса с минимальной априорной вероятностью встречаемости.

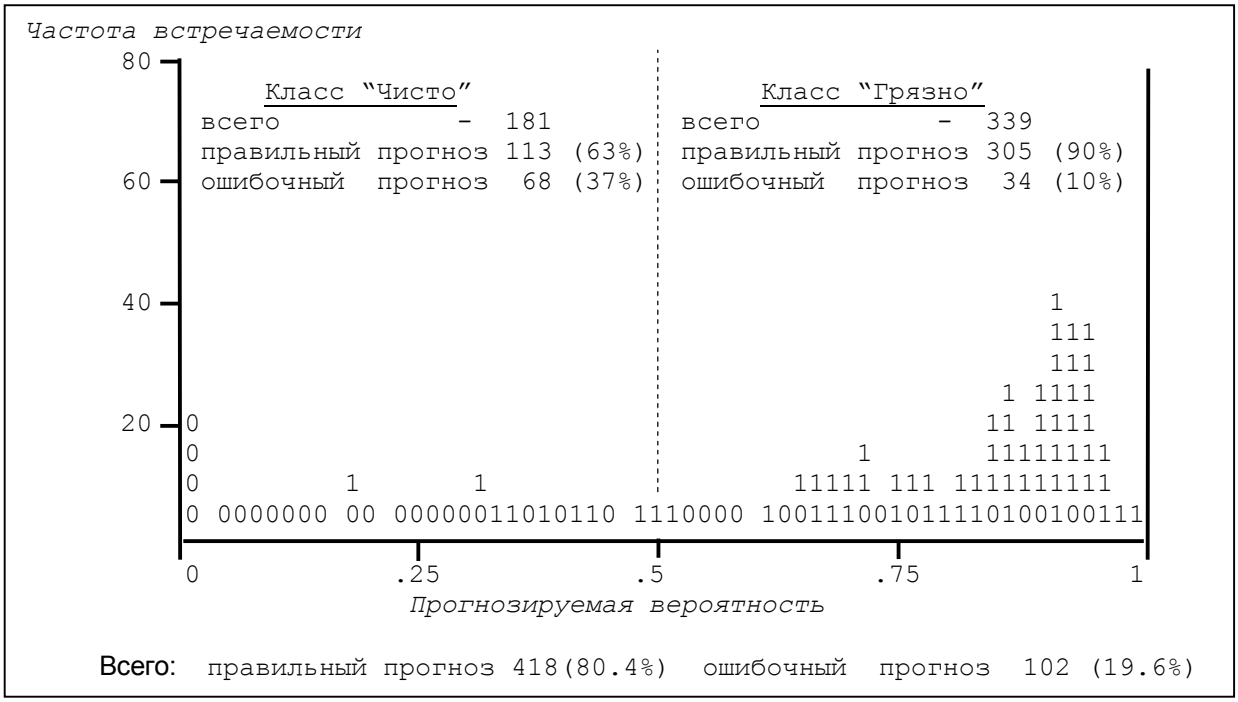


Рис. 8.3. Гистограмма распределения измерений по шкале прогнозируемой вероятности класса 1 - «грязно» (каждому символу на гистограмме соответствует около 5 объектов исходной выборки)

Множественный пробит-анализ по обобщенным показателям

Используем упорядоченный пробит-анализ для непосредственной оценки значения класса качества водоемов в виде числа от 1 до 6. Сформируем выборку из тех же 520 измерений, но в качестве девяти варьируемых переменных будем использовать различные обобщенные гидробиологические показатели и традиционные "интегральные" индексы, перечисленные в табл. 8.5.

С чисто статистической точки зрения было рассчитано вполне благополучное уравнение упорядоченного пробита: критерий $\chi^2(9)$ для статистики отношения правдоподобия составил 257.1 при $p \cong 0.0$, коэффициент детерминации псевдо-R^2 равен 42.3%.

Таблица 8.5

Коэффициенты модели упорядоченного пробита, связывающей класс качества вод и обобщенные показатели зообентоса

Наименования индексов и обобщенных показателей	Коэффициенты пробита	Стандартная ошибка	*t*-статистика	*P*-значение
1 Константа уравнения	2.42	0.276	8.76	0.0
2 Индекс Шеннона	0.1117	0.064	1.73	0.08
3 Число видов в пробе	- 0.0348	0.0125	-2.78	**0.0056**
4 Общая численность (ln экз/м^2)	0.011	0.044	0.249	0.803
5 Общая ,биомасса (ln мг/м^2)	0.0056	0.0246	0.228	0.812
6 Доля хищных видов (численность)	0.00108	0.00455	0.238	0.81
7 Доля хищных видов (биомасса)	-0.00459	0.00348	-1.318	0.189
8 Биотический индекс Вудивисса	- 0.285	0.0318	-8.958	**0.0**
9 Олигохетный индекс Пареле	0.7679	0.0894	8.584	**0.0**
10 Хирономидный индекс Балушкиной	0.0289	0.0166	1.738	0.083

Однако анализ уровня значимости коэффициентов пробит-уравнения, представленных в таблице 8.5, показывает, что вполне достоверно связаны с классом качества лишь число видов в пробе, биотический индекс Вудивисса (обратная зависимость) и олигохетный индекс (прямая зависимость).

Упорядоченный пробит представляет собой вероятностную модель, согласно которой попадание в группу, соответствующую каждому классу качества водоема представляет собой случайное событие. Вероятность $P(k|x_{i1},x_{i2},\ldots,x_{ip})$ принадлежности i-го измерения к k-му классу (вернее, вероятность попадания в интервал между границами классов) вычисляется по усеченному нормальному распределению, и при этом предполагается, что к расчетному значению $\hat{Y} = X\hat{\theta}$ прибавляется стандартная нормальная случайная величина. Графики этих вероятностей и граничные значения классов качества вод для полученного нами уравнения упорядоченного пробита изображены на рис. 8.4. В качестве прогноза для каждого наблюдения берется та группа, вероятность $P(k|x_{i1},x_{i2},\ldots,x_{ip})$ для которой наибольшая.

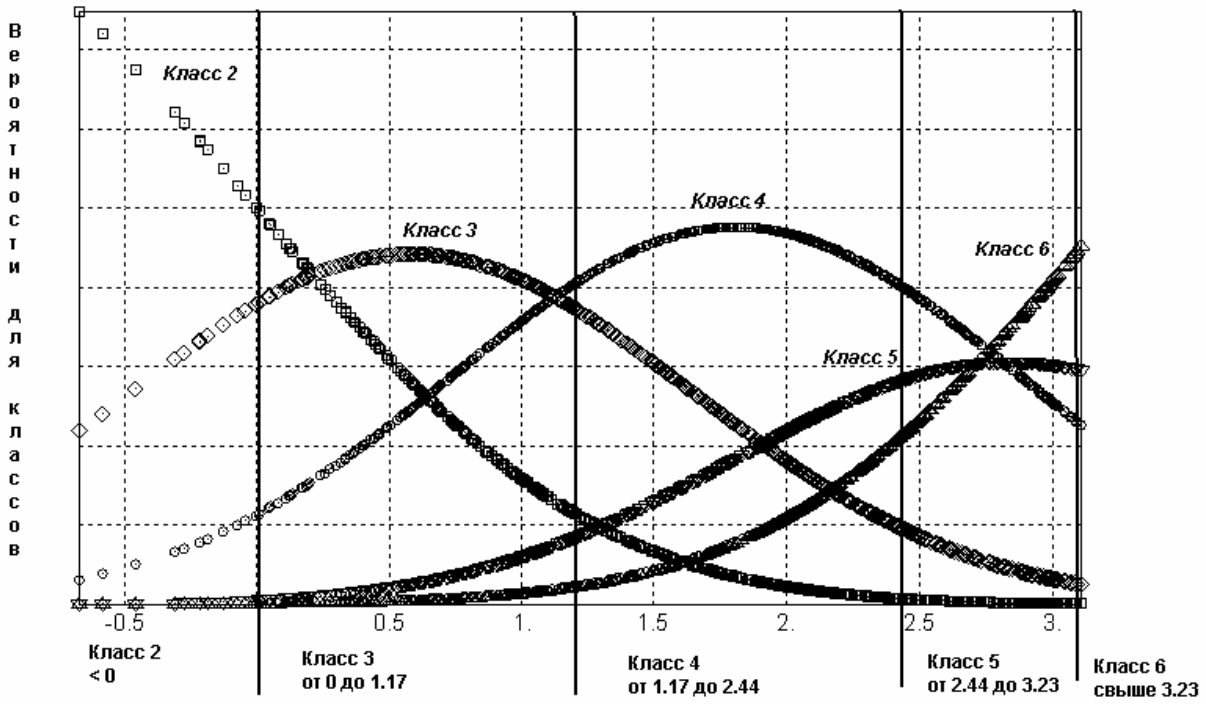


Рис. 8.4. График вероятностей прогнозирования класса качества вод и граничные значения для уравнения упорядоченного пробита

Результаты достоверности оценки классов качества представим в виде таблицы сопряженности «Факт – Прогноз» (табл. 8.6), где по главной диагонали проставлены частоты правильной оценки групп измерений, а в остальных клетках – имеющиеся ошибки оценки.

Таблица 8.6

Результаты прогнозирования класса качества вод по модели упорядоченного пробита

Классы качества вод		Фактические					Итого прогноз	Правильный прогноз, %	Ошибка на два и более класса, %
		2	***3***	***4***	***5***	***6***			
Прогнозируемые	***2***	**15**	10	2	0	0	27	55.56	7.41
	3	26	**53**	38	5	0	122	43.44	4.1
	4	11	61	**139**	56	45	312	44.55	17.95
	5	0	0	0	**0**	0	0	-	-
	6	1	4	8	26	**20**	59	33.90	22.03
Итого факт		53	128	187	87	65	520	43.65	14.6

Качество выполненного прогноза по модели упорядоченного пробита нельзя назвать вполне удовлетворительным, особенно, в области классов загрязненных вод 5 и 6. В частности, оценка класса 5 вообще не реализуема по представленным данным наблюдений. Действительно, ни для одного из измерений кривая вероятностей, соответствующая на рис. 8.4 классу 5, не проходит выше кривых остальных вероятностей.

8.3. Дискриминантные функции для классификации многомерных объектов

Формулировка задачи

Пусть задано пространство признаков $\boldsymbol{X}$ размерностью $\boldsymbol{m} > 1$, точками которого являются конкретные гидробиологические измерения $\boldsymbol{x} = \{\boldsymbol{x}_1,\ldots, \boldsymbol{x}_i,\ldots,\boldsymbol{x}_m\}$, где $\boldsymbol{x}_i$ – значения численности или биомассы $\boldsymbol{i}$-й таксономической группы гидробионтов в пробе, либо некоторые обобщенные индексы. Исходная таблица наблюдений разбита на $\boldsymbol{P}$ непересекающихся подмножеств строк, где каждой строке $\boldsymbol{x}$ поставлен в соответствие некоторый класс качества $\boldsymbol{y}_k$, $\boldsymbol{k} = 1,2,\ldots, \boldsymbol{p}$, причем любому из $\boldsymbol{P}$ классов принадлежит не менее одного объекта. Содержательный смысл задаваемой системы классификации $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p\}$ применительно к гидробиологическим исследованиям может иметь вполне произвольное толкование (например, любые градации сапробности, токсобности, классов качества вод, типов водоемов, природно-климатических зон и т.д.).

Необходимо определить набор формальных решающих правил, позволяющих для произвольного измерения $\boldsymbol{x}$ из $\boldsymbol{X}$ указать класс качества $\boldsymbol{y}_k$, к которому оно принадлежит.

С одним из методов классификации мы уже познакомились в предыдущем разделе, рассматривая логит- и пробит-анализ, как частный случай множественной регрессии. Другие решающие правила, формируемые на основе вероятностных методов, могут быть получены:

- в виде вероятности диагноза при заданном комплексе симптомов с использованием формулы Байеса или иных стратегий, основанных на минимаксном критерии или критерии Неймана-Пирсона;
- в виде простых классифицирующих функций, как это сделано в линейном дискриминантном анализе Фишера;
- в виде дискриминантных функций, являющихся результатом канонического дискриминантного анализа;
- в виде таких характеристик, как групповая ковариационная матрица, групповой вектор средних и определитель ковариационной матрицы, как это сделано в линейном дискриминантном анализе.

Рекомендуемая литература: [Урбах, 1964; Дуда, Харт, 1978; Кравцов, Милютин, 1981; Айвазян с соавт., 1989; Ким с соавт., 1989].

Математический лист

Байесовская схема принятия решений

Параметрические методы распознавания для поиска оптимальных дискриминационных функций используют аппроксимацию функции вероятностного распределения исходных данных и сводятся к определению отношения правдоподобия в различных областях многомерного пространства признаков. Классификатор может быть непосредственно построен из формулы условных вероятностей Байеса, описанной в разделе 7.4 и апеллирующей к априорным вероятностям принадлежности объектов к тому или иному распознаваемому классу и условным плотностям распределения значений вектора признаков.

Если априорные вероятности появления каждого класса равны, то вероятность того, что вектор $\boldsymbol{x}$ принадлежит классу $\boldsymbol{y}_i$ равна:

$$\boldsymbol{P}_i = \frac{\boldsymbol{P}(\mathbf{x} / \boldsymbol{y}_i)}{\sum_{\boldsymbol{k}=1}^{\boldsymbol{p}} \boldsymbol{P}(\mathbf{x} / \boldsymbol{y}_k)}. \qquad (8.19)$$

Очевидно, что наибольшая из величин $P(x/y_i)$ и будет обеспечивать наименьшую вероятность неправильной классификации или наименьший средний риск. Решающее правило можно сформулировать следующим образом: вектор измерений x принадлежит классу y_i, если

$$P(x/y_i) > P(x/y_j) \qquad \forall\, i \neq j\,. \tag{8.20}$$

Предположим, например, что каждый класс измерений описывается нормальным распределением и ковариационные матрицы C всех классов одинаковы. Тогда дискриминантная функция имеет следующий вид [Стьюпер с соавт., 1982]:

$$\frac{P(\mathbf{x}/y_i)}{P(\mathbf{x}/y_j)} = e^{[\mathbf{X'C}^{-1}(m_i - m_j) - 0.5(m_i + m_j)\mathbf{X'C}^{-1}(m_i - m_j)]}, \tag{8.21}$$

где m_i, m_j – математические ожидания векторов классов i и j. Для того, чтобы классифицировать произвольный вектор x, нужно рассчитать значения функции для всех возможных пар i и j при $i \neq j$ и отнести измерение к тому классу, для которого отношение условных вероятностей имеет наибольшее значение.

Если ковариационные матрицы классов неодинаковы, то добавляется некоторая функция потерь или платежная матрица, элементами которой R_{ij} являются значения штрафов за неправильную классификацию, когда объект x относят к классу j, когда как он принадлежит классу i. Чаще всего используют платежную матрицу R стандартного вида: ее элементы равны 0, если решающее правило правильно отнесло измерение к своему классу, и 1, если имела место ошибочная классификация. Нетрудно видеть, что при этом функционал среднего байесовского риска превращается в вероятность ошибочной классификации.

В простейшем случае для одной переменной и при двух классах процесс разделения можно представить графически на рис. 8.5. Если выборки признака X, относящиеся к обоим классам, подчинены нормальным законам распределения с дисперсией σ и средними m_1 и m_2, то пороговая величина x_o позволяет оптимальным образом разделить признаковое пространство на две области:

$$x_0 = \frac{m_1 + m_2}{2} + \frac{\sigma^2}{m_2 - m_1} \ln \lambda_0\,, \tag{8.22}$$

где λ_0 – критическое значение коэффициента правдоподобия, который зависит от платежных коэффициентов и априорных вероятностей появления объектов первого и второго класса. Если $r_{11} = r_{22} = 0$, $r_{12} = r_{21}$ и априорные вероятности равны, то $\lambda_0 = 0$ и линия x_o проходит посередине между средними обоих классов.

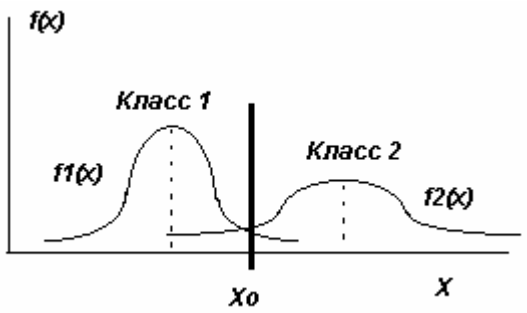


Рис. 8.5. Распределение двух совокупностей 1 и 2 по признаку X

Формула Байеса и оптимальные параметрические решающие правила могут быть использованы, если возможна достаточно точная аппроксимация функции плотности распределения данных. Если эта аппроксимация на основе обучающей выборки недостаточно точна, то и решающая функция будет далека от оптимальной. Сложность расчетов по восстановлению условных функций распределения $F(x/y_i)$ или ее плотности $p(x/y_i)$, $i = 1,2,\ldots, l$, является самым большим препятствием к использованию параметрических методов в многочисленных приложениях.

Однако, когда вид кривой плотности распределения неизвестен и нельзя сделать вообще никаких предположений о ее характере, то все равно общую стратегию Байеса можно обобщить на любой непараметрический метод расчета с участием двух матриц – платежной матрицы $\boldsymbol{R}$ и диагностической матрицы $\boldsymbol{P}$, содержащей некоторые оценки условных вероятностей отнесения объекта к каждому классу, если объект имеет определенную комбинацию признаков. Существует значительное множество различных алгоритмов формирования диагностической матрицы $\boldsymbol{P}$, использующих разные эвристические предположения их авторов. Расчет оценок может быть основан, например, на использовании *метода многомерных гистограмм* (частот встречаемости в различных классах объектов обучающей выборки, содержащих тот или иной признак [Гублер, 1978]), средних мер близости для компактных подмножеств объектов [Журавлев, 1978], нормированных разностей между внутригрупповыми средними и общим средним значением признака, эвристиках Е. Парзена [Parzen, 1962] и Э.А. Надарая [1964] и т.д.

Методы линейного дискриминантного анализа

Основной целью дискриминации является нахождение такой линейной комбинации переменных, которая бы оптимально разделила рассматриваемые группы. Линейная функция

$$\boldsymbol{d_{ik} = a_{ok} + a_{1k} \cdot x_{i1k} + a_{2k} \cdot x_{i2k} + \ldots + a_{jk} \cdot x_{ijk} + \ldots + a_{mk} \cdot x_{imk}}, \qquad (8.23)$$

при $\boldsymbol{i} = 1, 2,\ldots, \boldsymbol{n_k}$; $\boldsymbol{k} = 1, 2,\ldots, \boldsymbol{p}$; называется *дискриминантной функцией* с неизвестными коэффициентами $\boldsymbol{a_{jk}}$. Здесь $\boldsymbol{d_{ik}}$ – расчетное значение функции для $\boldsymbol{i}$-го объекта из группы $\boldsymbol{k}$, состоящей из совокупности $\boldsymbol{n_k}$ измерений; $\boldsymbol{x_{ijk}}$ – значение $\boldsymbol{j}$-й дискриминантной переменной, $\boldsymbol{j} = 1,2,\ldots,\boldsymbol{m}$ – столбцы матрицы наблюдений.

В общем случае необходимо рассчитать $\boldsymbol{p}$ линейных дискриминантных функций, равное количеству анализируемых популяций, после чего с использованием коэффициентов $\boldsymbol{a_{jk}}$ и постоянной $\boldsymbol{a_{0k}}$ можно провести классификацию любого произвольного наблюдения. Для этого необходимо подставить значения переменных вектора $\boldsymbol{x}$ в дискриминантные уравнения для каждой $\boldsymbol{k}$-й группы и рассчитать значения оценок отклика, $\boldsymbol{k} = 1, 2,\ldots, \boldsymbol{p}$. Вектор $\boldsymbol{x}$ классифицируется как принадлежащий тому классу (группе измерений, популяции) $\boldsymbol{k}$, для которого величина $\boldsymbol{d_k}$ имеет максимальное значение.

Для расчета коэффициентов дискриминантных функций нужен статистический критерий, оценивающий различия между группами. Очевидно, что классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Один из методов поиска наилучшей дискриминации данных заключается в нахождении таких дискриминантных функций $\boldsymbol{d_k}$, которые были бы основаны на максимуме отношения межгрупповой вариации к внутригрупповой.

Многомерное нормальное распределение случайной величины $\boldsymbol{x_{ijk}}$ характеризуется следующими статистическими компонентами:

- вектор из $\boldsymbol{m}$ средних значений $\boldsymbol{X_{j\bullet\bullet}}$ переменной $\boldsymbol{j}$ по всем классам (общее средние значения используемых признаков);
- матрица $\boldsymbol{m \times m}$ сумм квадратов и по парных произведений $\boldsymbol{T}$, которая показывает степень различий между признаками. Элементы матрицы $\boldsymbol{T}$ задаются соотношением:

$$\boldsymbol{t_{jl} = \sum_{k=1}^{P} \sum_{i=1}^{n_k} (x_{ijk} - x_{j\bullet\bullet})(x_{ilk} - x_{l\bullet\bullet})}, \qquad (8.24)$$

  где выражения в скобках – отклонения значений переменных от общего среднего. При $\boldsymbol{j = l}$ получается среднеквадратичное отклонение, которое показывает вариацию наблюдений по отдельной переменной. При $\boldsymbol{j \neq l}$ получается корреляция (ковариация) между двумя переменными. Если разделить каждый элемент матрицы **T** на $(\boldsymbol{n}$ - 1) получается ковариационная матрица $\boldsymbol{C}$. Для получения корреляционной матрицы $\boldsymbol{R}$ следует каждый элемент матрицы **T** разделить на квадратный корень произведения соответствующих диагональных элементов

$$\boldsymbol{t_{jl} \Big/ \sqrt{t_{jj} t_{ll}}};$$

- матрица $\boldsymbol{m \times p}$ из средних значений $\boldsymbol{X_{jk\bullet}}$ переменной $\boldsymbol{j}$ для измерений $\boldsymbol{k}$-го класса (групповые средние);

- матрица W, которая используется для определения степени разброса внутри классов и отличается от T тем, что при ее вычислении используются средние для отдельных классов, а не общие средние:

$$w_{jl} = \sum_{k=1}^{P} \sum_{i=1}^{n_k} (x_{ijk} - x_{jk\bullet})(x_{ilk} - x_{lk\bullet}) \,. \tag{8.25}$$

Если элементы матрицы $\mathbf{W}$ разделить на $(n - p)$, то получится *внутригрупповая ковариационная матрица* S.

Если расположение центров классов различается между собой, то степень вариации наблюдений внутри классов будет меньше общего статистического разброса: $w_{jl} < t_{jl}$, причем, чем больше расхождение этих величин, тем ощутимее влияния фактора группировки. Введем матрицу разницы этих двух матриц B, которая представляет собой межгрупповую сумму квадратов отклонений и попарных произведений $B = T - W$ (т.е. $b_{jl} = t_{jl} - w_{jl}$). Величины элементов B по отношению к величинам элементов W дают меру различия между группами.

Коэффициенты $a_{ok}, a_{1k}, a_{2k},\ldots, a_{mk}$ разделяющих функций могут быть найдены по методу дискриминантного анализа Фишера как элементы матрицы, обратной к W, что соответствует общей вычислительной процедуре множественной линейной регрессии. Более сложным в математическом отношении является канонический дискриминантный анализ, где ищутся независимые или ортогональные функции, вклады которых в разделение совокупностей не будут перекрываться. С вычислительной точки зрения здесь проводится анализ *канонических корреляций*, в котором определяются последовательные канонические корни и вектора.

Для нахождения p наборов коэффициентов канонических дискриминантных функций необходимо решить систему уравнений:

$$\sum b_{1i}\nu_i = \lambda \sum w_{1i}\nu_i \,, \sum b_{1i}\nu_i = \lambda \sum w_{1i}\nu_i \,,\ldots, \sum b_{1i}\nu_i = \lambda \sum w_{1i}\nu_i \,, \quad \sum \nu_i^2 = 1 \,. \tag{8.26}$$

Как известно из курса линейной алгебры, собственными значениями квадратной матрицы B порядка m называются такие значения λ_j, при которых система следующих m уравнений имеет нетривиальное решение:

$$B\nu_j = \lambda_j \nu_j \,, \quad j = 1, 2,\ldots, m, \tag{8.27}$$

где ν_j собственные векторы матрицы B, соответствующие λ_j. Нетривиальное решение системы уравнений: $B\nu_j = \lambda_j W\nu_j$, $j = 1, 2,\ldots, m$, где B и W – симметрические положительно определенные матрицы, относится к обобщенной проблеме собственных значений и может быть получено путем замены переменных, используя разложение по Холецкому.

Используя компоненты собственных векторов ν_j для описанных выше ковариационных матриц B и W, находят путем нормировки p наборов нормированных коэффициентов канонических дискриминантных функций $a_{jk} = \nu_{jk}\,(n - p)^{0.5}$. С геометрической точки зрения, полученные дискриминантные функции определяют гиперповерхности в m-мерном пространстве. В частном случае при $m = 2$ они являются прямыми, а при $m = 3$ – плоскостями. В этих обозначениях функция расстояния Махалонобиса, описанная в разделе 7.4 и учитывающая расстояние между центроидами двух классов k и r, будет равна:

$$D_{kr}{}^2 = \sum_{j=1}^{m} a_j (x_{jk\bullet} - x_{jr\bullet}) \,. \tag{8.28}$$

Заключительный этап дискриминантного анализа включает методы интерпретации межгрупповых различий и методы классификации наблюдений по группам. При интерпретации нужно ответить на вопросы: возможно ли, используя данный набор переменных, отличить одну группу от другой, насколько хорошо эти переменные помогают провести дискриминацию и какие из них наиболее информативны? Детальный анализ проводится с использованием объединенной матрицы ковариации T и ковариационных матриц для отдельных групп W_k, $k = 1, 2,\ldots, p$.

Напомним еще раз основные предположения дискриминантного анализа. Во-первых, считается, что анализируемые переменные представляют выборку из многомерного нормального распределения. Отметим, однако, что пренебрежение условием нормальности обычно не является "фатальным" в смысле доверия к результатам расчетов. Более важно второе предположение о статистическом равенстве внутригрупповых матриц дисперсий и корреляций. При искусственном объявлении ковариационных матриц W_k статистически неразличимыми могут оказаться отброшенными наиболее важные индивидуальные черты, имеющие большое значение для хорошей дискриминации. Критерии, используемые для проверки этих предположений, обсуждались нами в разделе 5.2.

Так как дискриминантные функции находятся по выборочным данным, они нуждаются в проверке статистической значимости. Определяющим для дискриминантного анализа является проверка гипотезы об отсутствии различий между групповыми средними H_o : $m_1 = m_2 = \ldots m_p$, что является многомерным аналогом однофакторного дисперсионного анализа. Для этого может быть использовано *обобщенное расстояние Махалонобиса*, которое в матричном виде можно записать как

$$D^2 = \sum_{k=1}^{p} n_k (\mathbf{x}_{k\bullet} - \mathbf{x}_{\bullet\bullet})\, \mathbf{C}^{-1} (\mathbf{x}_{k\bullet} - \mathbf{x}_{\bullet\bullet}) \quad . \tag{8.29}$$

На содержательном уровне его можно интерпретировать как взвешенную сумму расстояний от вектора средних каждой группы $X_{k\bullet}$ до общего вектора средних $X_{\bullet\bullet}$. Если гипотеза H_o верна, а объем выборки стремится к ∞, то D^2 может быть аппроксимирована F-распределением. Другим, в некоторых случаях более точным способом проверки гипотезы H_o является использование U-статистики Уилкса (она же – лямбда Вилкса), которая вычисляется как отношение детерминантов (*det*) матрицы внутригрупповой ковариации W и полной ковариационной матрицы T:

$$U = det(\mathbf{W}) / det(\mathbf{T}).$$

Аппроксимация статистики U-Уилкса с помощью F-распределения была выполнена К. Рао.

Наиболее общим принципом применения дискриминантного анализа является включение в исследование по возможности большего числа переменных с целью определения тех из них, которые наилучшим образом разделяют выборки между собой. Для этой цели, как и в случае множественного регрессионного анализа, используется пошаговая процедура, в которой на каждом шаге построения модели дискриминации просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная включается в модель на текущем шаге и происходит переход к следующему шагу. Можно также двигаться в обратном направлении и все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться признаки, вносящие наименьший вклад в предсказание. Пошаговая процедура дискриминантного анализа для отбора переменных основывается на F-критериях однофакторного дисперсионного анализа: «F-включения» и «F-исключения». Значение F-статистики для переменной указывает на ее статистическую значимость при дискриминации между совокупностями и является мерой вклада признака в предсказание членства в группах. Тогда в качестве результата успешного анализа можно сохранить только наиболее информативные переменные модели, то есть те переменные, чей вклад в дискриминацию больше остальных.

Другим полезным критерием, используемым для селекции признаков, является коэффициент множественной корреляции R^2 для соответствующей переменной со всеми другими переменными в текущей модели. При значениях R^2, близких к 1, анализируемый признак полностью определяется комбинацией других признаков и является избыточным. При сильной взаимной коррелированности переменных матрица задачи становится плохо обусловленной, что резко сказывается на погрешностях расчетов. И, наконец, дискриминантные функции представляются аналогами главных компонент, поэтому для нахождения оптимального числа переменных можно воспользоваться критериями, оценивающими остаточную дискриминантную способность, под которой понимается способность различать группы с помощью переменных, не включенных в модель. Это может быть, например, Λ-статистика, вычисляемая по формуле:

$$\Lambda = \prod_{l=k+1}^{p} \frac{1}{1+\lambda_l} \quad . \tag{8.30}$$

Если остаточная дискриминация мала, то выполненный анализ достиг своей цели.

Кроме задачи "объяснения", другой главной целью применения дискриминантного анализа является задача "прогнозирования". Как только модель установлена и получены дискриминирующие функции, возникает вопрос о том, как хорошо они могут *предсказывать*, к какой совокупности принадлежит конкретное измерение? Обычно классификация объектов осуществляется с использованием одного из следующих методов:

- произвольный вектор наблюдений $\boldsymbol{x}$ относится к классу $\boldsymbol{k}$, для которого значение отклика дискриминантных функций $\boldsymbol{d_k}$ является максимальным среди всех $\boldsymbol{k} = 1, 2,\ldots, \boldsymbol{p}$;
- вычисляется расстояние Махаланобиса от анализируемого наблюдения до «центра тяжести» каждой группы и наблюдение признается принадлежащим к той совокупности, к которой оно ближе в смысле минимума этого расстояния (этот метод классифицирования считается не вполне точным, так как предполагает нормальный закон распределения относительно среднего для каждой группы);
- используются оценки *апостериорных* вероятностей принадлежности вектора $\mathbf{x}$: $\boldsymbol{P(k} \mid \mathbf{x}) = \dfrac{e^{d_k}}{\sum_{k=1}^{p} e^{d_k}}$, вычисленные для каждого класса $\boldsymbol{k} = 1, 2,\ldots, \boldsymbol{p}$.

На точность классификации может сильно влиять способ спецификации *априорных* вероятностей наблюдений в различных совокупностях. Если неодинаковая заселенность классов в первоначальной выборке является отражением истинного распределения в популяции, то необходимо положить априорные вероятности пропорциональными объемам совокупностей в выборке. Если это только случайный результат процедуры формирования обучающей выборки, то априорные вероятности принимаются одинаковыми для каждой группы.

Результаты достоверности оценки классов анализируются с использованием таблицы сопряженности «Факт – Прогноз». С помощью этой таблицы можно оценить вероятность ошибочной классификации каждого класса, которая является смещенной.

Результаты расчетов

В качестве примеров используем те же выборки, что были получены в разделе 8.2 при рассмотрении логистической регрессии.

В первом примере разделим 520 гидробиологических измерений на 2 группы: 181 пробы на «чистых» станциях наблюдения с классом качества вод 3 и менее, и 339 проб на станциях, соответствующих классу 4 и выше. В качестве варьируемых переменных примем общее число видов $\boldsymbol{X_1} = \boldsymbol{S}$ и показатели обилия отдельных семейств зообентоса (для хирономид – подсемейств и триб); $\boldsymbol{X_j} = \ln((\boldsymbol{N_{sj}B_{sj}})^{0.5})$, $\boldsymbol{N_{sj}}$ и $\boldsymbol{B_{sj}}$ – суммарные по видам численность и биомасса $\boldsymbol{j}$-й таксономической группы в пробе, $\boldsymbol{j} = 2, 3,\ldots 51$.

Используем прямую пошаговую процедуру нахождения информативных дискриминантных функций при пороге $\boldsymbol{F}$-включения, равном 3, в результате которой было отберем 13 дискриминантных переменных из 50 при статистически незначимой остаточной дискриминации. Как статистике Уилкса ($\boldsymbol{U} = 0.68$), так и обобщенному расстоянию Махалонобиса между группами ($\boldsymbol{D^2} = 2.07$) соответствует один и тот же аппроксимированный критерий Фишера $\boldsymbol{F}(13,506) = 18.3$, что позволяет отвергнуть нулевую гипотезу о равенстве групповых средних на высоком уровне значимости.

По знаку и абсолютной величине рассчитанных коэффициентов дискриминантных функций, представленных в табл. 8.7, можно качественно оценить удельную индикаторную ценность отдельных семейств зообентоса для прогнозирования степени загрязнения вод. Естественно, например, предположить, что обилие групп Ephemeroptera, Amphipoda или Ptychopteridae и многих других тем выше, чем чище водоем, в то время как обратная зависимость имеется для Odonata и Dermaptera. Однако прямое использование коэффициентов уравнений в роли индексов «чис-

то/грязно» не вполне правомочно, поскольку механизм их формирования учитывает целый ряд дополнительных статистических аспектов, таких как взаимная коррелированность признаков и проч.

Таблица 8.7

Коэффициенты дискриминантных функций, оценивающих категорию качества вод по обилию таксономических групп зообентоса
*(**p** – априорная вероятность класса на обучающей выборке,*
R^2 – коэффициент множественной корреляции)

Таксономические группы зообентоса	Класс «Чисто» $p = 0.348$	Класс «Грязно» $p = 0.652$	F-критерий исключения	R^2
Свободный член	-4.51	-2.68		
Ephemeroptera	0.312	-0.050	35.4	0.315
Limoniidae	0.424	0.235	6.0	0.102
Amphipoda	0.380	-0.154	12.8	0.031
Oligochaeta	0.387	0.545	14.9	0.165
Число видов в пробе ***S***	0.199	0.160	2.3	0.532
Rhagionidae	0.311	0.060	3.2	0.052
Simuliidae	0.245	0.055	3.9	0.122
Dreissenidae	0.230	0.077	5.8	0.069
Ptychopteridae	1.363	-0.242	4.6	0.013
Триба Tanytarsini	0.167	0.049	4.1	0.322
Dermaptera	-0.904	0.069	4.3	0.037
Coleoptera	0.034	-0.157	4.3	0.145
Odonata	-0.287	-0.113	3.4	0.109

Использование полученных дискриминантных функций для прогноза категории водоема по шкале («чисто» / «грязно») на примерах обучающей выборки дал более скромные результаты, чем в случае логистической регрессии: всего было правильно опознано 382 измерения или 78.3% от общего числа. Особенно неудачный прогноз имел место для категории «чисто», где было правильно идентифицировано только 103 пробы (56.9%).

Во втором примере выборку из тех же 520 измерений разобьем на 5 групп в соответствии со значением класса качества водоемов, откуда бралась проба. В качестве девяти варьируемых переменных будем использовать различные обобщенные гидробиологические показатели и традиционные «интегральные» индексы. В ходе пошаговой процедуры дискриминантного анализа два признака, связанные с численностью, были исключены как неинформативные, в результате чего получены уравнения, представленные в табл. 8.8.

Таблица 8.8

Коэффициенты дискриминантных функций, оценивающих класс качества вод по обобщенным показателям зообентоса
(отсутствуют коэффициенты для переменных, признанных неинформативными)

Наименования индексов и обобщенных показателей	Классы качества вод					F-критерий исключения
	2	3	4	5	6	
Константа уравнения	-14.61	-14.10	-12.49	-13.89	-10.90	
Биотический индекс Вудивисса	1.78	1.46	0.99	0.78	0.66	23.33
Олигохетный индекс Пареле	3.88	4.04	5.07	7.38	6.20	11.58
Общая биомасса (ln ***B***) мг/м^2	1.29	1.55	1.47	1.55	1.40	3.27
Индекс Шеннона	2.44	2.61	3.15	3.28	2.66	3.54
Хирономидный индекс Балушкиной	0.04	0.22	0.22	0.29	0.21	3.58
Доля хищных видов (по биомассе)	0.07	0.06	0.07	0.06	0.05	3.27
Число видов в пробе	-0.27	-0.33	-0.35	-0.45	-0.40	3.26
Общая численность (ln ***N***), экз./м^2	-	-	-	-	-	1.48
Доля хищных видов (по численности)	-	-	-	-	-	0.99

Основываясь на статистике Уилкса $U = 0.51$, значению которой соответствуют аппроксимация критерия Фишера $F(28, 1836) = 13.4$ и вероятность $p \approx 0.0$, гипотеза о равенстве групповых средних для различных классов качества вод в целом должна быть решительно отвергнута.

Детальный анализ матрицы обобщенных расстояний Махалонобиса для всех возможных пар классов, представленный в табл. 8.9, также свидетельствует о значимых статистических различиях между группами измерений (в наименьшей мере это относится для классов качества 5 и 6, значение вероятности $p = 0.0034$ для которых значительно меньше критического).

Таблица 8.9

Обобщенные расстояния Махалонобиса D^2 (выше главной диагонали) между центроидами измерений на станциях с различным классом качества воды и значения F-статистики Фишера для оценки различий между классами (ниже главной диагонали)

		Классы качества вод				
		2	3	4	5	6
Классы качества вод	2	0	1.39	2.84	7.19	7.09
	3	5.61	0	0.81	3.29	3.81
	4	12.63	6.69	0	1.50	1.91
	5	25.49	18.45	9.66	0	0.68
	6	22.23	17.74	9.96	2.72	0

Не углубляясь в технику расчетов, которые разумно поручить компьютеру, рассмотрим, как с использованием дискриминантного анализа осуществляется классификация конкретных измерений. Пусть для двух гидробиологических проб, взятых на р. Чапаевка в русловой части станции 12, отнесенной по гидрохимическим показателям к классу качества 5, рассчитаны индексы и обобщенные показатели, перечисленные в табл. 8.8. Используя коэффициенты дискриминантных уравнений, вычислим для каждого из этих примеров значения апостериорных вероятностей и расстояний Махалонобиса относительно каждого класса качества вод (см. табл. 8.10). Для измерения от 13 июня 1990 г. наименьшее расстояние Махалонобиса соответствует 5 классу качества вод, в то время, как по максимуму оценок апостериорных вероятностей эта проба соответствует классу качества 4. В то же время, обе статистики, вычисленные по результатам наблюдения от 27 июля 1990 г., относят эту станцию к 5 классу.

Таблица 8.10

Расстояния Махалонобиса и апостериорные вероятности для оценки класса качества вод двух измерений, сделанных на ст.12 р. Чапаевка

Дата измерения	Статистика для классификации	Классы качества вод				
		2	3	4	5	6
13.06.90	Расстояние Махалонобиса	9.499	5.177	2.408	**2.391**	3.149
	Апостериорная вероятность	0.004	0.091	**0.529**	0.248	0.127
27.07.90	Расстояние Махалонобиса	11.164	6.929	4.502	**1.845**	2.235
	Апостериорная вероятность	0.003	0.050	0.247	**0.434**	0.267

Результаты достоверности оценки классов качества для всех примеров обучающей выборки представим в виде таблицы сопряженности «Факт – Прогноз» (табл. 8.11), где по главной диагонали проставлены частоты правильной оценки групп измерений, а в остальных клетках – имеющиеся ошибки прогноза.

Остается обратить внимание читателя на то, что общая эффективность прогнозирования класса качества вод с использованием дискриминантного анализа оказалась существенно выше, чем по модели упорядоченного пробита.

Таблица 8.11

Результаты прогнозирования класса качества вод с использованием дискриминантного анализа

Классы качества вод		Фактические					Итого прогноз	Правильный прогноз, %	Ошибка на два и более класса, %
		2	***3***	***4***	***5***	***6***			
Прогнозируемые	***2***	**26**	19	6	0	0	51	51.0	0
	3	12	**50**	33	8	1	104	48.1	8.6
	4	13	46	**129**	31	28	247	52.2	16.5
	5	0	11	11	**39**	15	76	51.3	14.4
	6	2	2	8	9	**21**	42	50.0	28.5
Итого факт		53	128	187	87	65	520	51.1	14.0

8.4. Задача о «классобности» видов: алгоритм распознавания, основанный на вычислении биоиндикационых индексов

Формулировка задачи

Пусть задано пространство признаков $\boldsymbol{X}$ размерностью $\boldsymbol{m} > 1$, соответствующее некоторому списку видов гидробионтов. Точками этого пространства являются конкретные гидробиологические измерения $\boldsymbol{x} = \{\boldsymbol{x_1},\ldots, \boldsymbol{x_i},\ldots,\boldsymbol{x_m}\}$, где $\boldsymbol{x_i}$ – значение обилия $\boldsymbol{i}$-го вида в пробе.

Как и в случае дискриминантного анализа, определены $\boldsymbol{P}$ классов $\{\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_p}\}$, к которым могут относиться всевозможные измерения $\boldsymbol{x}$ из $\boldsymbol{X}$, причем содержательный смысл используемой системы классификации имеет произвольный характер, включающий любые градации сапробности, токсобности, классов качества вод, типов водоемов, природно-климатических зон и т.д.

Необходимо на основе обучающей выборки определить набор формальных решающих правил, позволяющих для произвольного измерения $\boldsymbol{x}$ из $\boldsymbol{X}$ указать класс качества вод $\boldsymbol{y_k}$, к которому оно принадлежит.

Математический лист

Сформулируем метод классифицирования измерений $\boldsymbol{x}$, основанный на следующих эвристических процедурах:

- на этапе обучения рассчитывается матрица оценок $\boldsymbol{R}$, элементы которой $\boldsymbol{R_{ik}}$ являются тем больше, чем выше индикаторная ценность вида $\boldsymbol{i}$ для класса $\boldsymbol{k}$ (т.е. «сапробность», «токсобность» или, попросту выражаясь, некоторая «классобность»);
- на этапе экзамена подбирается такой алгоритм распознавания, т.е. последовательность формул преобразования матрицы $\boldsymbol{R}$ в вектор результирующих оценок классов $\boldsymbol{t}$:

$$\boldsymbol{F(R, x_r) \Rightarrow t}, \tag{8.31}$$

которая обеспечивает минимум ошибок классификации для широкого набора тестовых примеров $\boldsymbol{x_r}$.

Элементы матрицы $\boldsymbol{R}$ будем называть *индикаторными валентностями* (value of indication) вида $\boldsymbol{i}$ в классе $\boldsymbol{k}$. Рассмотрим три версии расчета индикаторных валентностей и последующей классификации, основанные на следующих различных концепциях:

- использование формулы Зелинки – Марвана вычисления сапробных валентностей;
- вероятностные оценки на основе частот встречаемости видов в водоемах разной категории;
- сбалансированные индикаторные валентности, полученные с использованием методов нелинейной оптимизации.

Алгоритм 1, использующий формулы вычисления сапробных валентностей.

Как было описано в разделе 4.4, оценка зон сапробности по показательным организмам методом Зелинки и Марвана основывается на следующих предпосылках:

- для каждого i-го вида гидробионтов устанавливаются значения сапробных валентностей a_{ik}, которые теоретически совпадают с оценками распределения вероятности встречаемости вида в каждой k-й ступени сапробности и выражаются одной или несколькими цифрами, сумма которых равна 10;
- вводится шкала индикаторного веса J_i, выраженного в баллах от 1 до 5 и оценивающего роль (дискриминирующую важность) i-го вида при оценке степени загрязнения;
- для произвольной гидробиологической пробы, в которой измерены значения обилия видов N_i, с использованием величин a_{ik} и J_i, рассчитывается средневзвешенная сапробность сообщества t_k, которая является оценкой принадлежности пробы к k-й зоне сапробности.

Поскольку метод классификации водоемов по сапробности является единственным практическим способом количественной водной биоиндикации, понятно стремление исследователей-гидробиологов, накопивших значительный массив экспедиционных данных, провести самостоятельные расчеты индикаторных валентностей с учетом региональных особенностей бентофауны, характера загрязнений и типологии водоемов. В разделе 4.4 была приведена формула расчета сапробных валентностей с использованием численности N и встречаемости D_i гидробионтов в основных зонах сапробности (о-р), предложенная П.А. Цимдинем [1979]. Если сделать естественное предположение, что численность вида-индикатора представляет собой среднее значение численности для всех j измерений в k-й зоне, $j = 1, 2,\ldots, n_{ik}$, то формула приобретает тривиальный вид отношения групповых средних к глобальной средней:

$$a_{ik} = \frac{N_{ik} \cdot D_{ik}}{\sum_{k=p}^{o} N_{ik} \cdot D_{ik}} \cdot 10 = \frac{\frac{n_{ik}}{n}\left(\sum_{j=1}^{n_{ik}} N_{ijk}\right)/n_{ik}}{\sum_{k=p}^{o} \frac{n_{ik}}{n}\left(\sum_{j=1}^{n_{ik}} N_{ijk}\right)/n_{ik}} \cdot 10 = \frac{\sum_{j=1}^{m_{jk}} N_{ijk}}{\sum_{k=p}^{o}\sum_{j=1}^{m_{jk}} N_{ijk}} \cdot 10 \quad , \tag{8.32}$$

где n_{ik} – встречаемость i-го вида в k-й зоне, $D_{ik} = n_{ik} / n$, n – общее число измерений.

Индикаторные веса J_i оцениваются по Сладечеку, ориентируясь на характер распределения сапробных валентностей по зонам. Например, индикаторный вес $J = 5$ присваивается "хорошим" индикаторам, если все 10 баллов сапробной валентности распределены в одной зоне сапробности. Если валентности равномерно распределяются по ступеням, то такие виды считаются индифферентными или "плохими" индикаторами и получают небольшой балл.

Сформированные индикаторные валентности a_{ik} используются на втором этапе для прогнозирования класса водоема по комбинации видов, встретившихся в произвольно взятой пробе. Для этого, в соответствии с методом Зелинки и Марвана, для каждой k-й зоны ($k = 1, 2,\ldots, p$) вычисляются средневзвешенные валентности t_k по численностям видов N_i ($i = 1,\ldots, n_p$) экзаменуемого измерения с учетом индикаторных весов J_i:

$$t_k = \frac{\sum_{i=1}^{m_r} N_i \cdot 0.1 \cdot a_{ik} \cdot J_i}{\sum_{i=1}^{m_t} N_i \cdot J_i} \quad , \tag{8.33}$$

где m_r – количество видов в тестируемой пробе.

В предыдущих разделах нами было показано, что для вариационных рядов обилия бентосных организмов наиболее характерно логнормальное распределение, в связи с чем логарифмирование численности особей в пробе существенно нормализует распределение значений N_i и обеспечивает корректное вычисление статистических характеристик: средней, дисперсии и проч. Поэтому альтернативными вариантами формул для a_{ik} и t_k являются:

$$a_{ik} = \frac{\sum_{j=1}^{n_{jk}} \ln(N_{jk}+1)}{\sum_{k=p}^{o}\sum_{j=1}^{n_{jk}} \ln(N_{jk}+1)} \cdot 10 \; ; \qquad t_k = \frac{\sum_{i=1}^{m_t} \ln(N_i+1) \cdot 0.1 \cdot a_{ik} \cdot J_i}{\sum_{i=1}^{m_t} \ln(N_i+1) \cdot J_i} \tag{8.34}$$

Рассчитав для тестируемого примера средневзвешенные валентности t_k, имеющие смысл оценок принадлежности пробы к классам качества вод, мы можем избрать одну из двух стратегий классифицирования.

Первая стратегия – наиболее простая и традиционно рекомендуемая математической теорией распознавания образов – предлагает отнести объект к классу, набравшему максимальную оценку, что соответствует минимальному среднему риску ошибки (см., например, формулу 8.20).

Вторая стратегия основывается на аналогии с расчетом индекса сапробности Пантле-Букка. И.К. Тодерашем [1984] предложена формула пересчета средневзвешенных сапробных валентностей $\{t_x, t_o, t_\beta, t_\alpha, t_p\}$ в индекс сапробности по Ротшайну (см. раздел 4.4), которая имеет следующий вид:

$$S_R = 0 \cdot t_x + 1 \cdot t_o + 2 \cdot t_\beta + 3 \cdot t_\alpha + 4 \cdot t_p \; . \tag{8.35}$$

По величине индекса S_R судят о принадлежности к зоне сапробности.

Алгоритм 2 – оценка индикаторных индексов по частоте встречаемости

Традиционным для гидробиологии является анализ видовой встречаемости, когда для исследователя имеет значение только факт наличия вида в пробе. Такой подход, например, широко используется в кластерном анализе с использованием коэффициентов сходства по Жаккару, Съёренсену и т.п. Поэтому представляется целесообразным выполнить расчет индикаторных индексов для отдельных видов на основе частот их встречаемости в пробах различных классов.

Сформируем матрицу гидробиологических наблюдений X в альтернативной шкале измерений, положив все ее элементы равными 1, если значение численности i-го вида в j-й пробе не меньше некоторого заданного порога ($N_{ij} \geq \theta$), и 0 в противном случае. Эта операция позволяет нам абстрагироваться от абсолютных величин обилия видов, нестационарных по своей природе, и использовать в качестве байесовских оценок условных вероятностей более устойчивые и унифицированные значения частот встречаемости.

На первом этапе с использованием примеров обучающей выборки, для которых известны значения классов качества вод $y_1, y_2, \ldots, y_p$, сформируем:

- матрицу P оценок условных вероятностей класса k для вида i: $p_{ik} = n_{ik} / n_i$; (8.36)
- вектор-столбец оценок априорной вероятности вида i : $p_i = n_i / n$; (8.37)
- вектор-строку оценок априорной вероятности класса k : $p_k = n_k / n$, (8.38)

где n – общее количество примеров обучающей выборки; n_i – количество измерений, содержащих вид k; n_k – количество объектов, принадлежащих классу k; n_{ik} – количество измерений класса k, содержащих вид i.

Сформированные оценки вероятностей используются на втором этапе для прогнозирования или экзамена – определения класса водоема по комбинации видов, встретившихся в произвольно взятой пробе.

Решающее правило, как и в предыдущем случае, представляет собой функциональный преобразователь, на выходе которого по априорным вероятностям и измерению x_r вычисляется вектор t результативных оценок $t_1, t_2, \ldots, t_p$. Суждение о принадлежности пробы x_r к водоему k-й категории может быть вынесено, например, если t_k – наибольшая оценка из $t_1, t_2, \ldots, t_p$. Простейшим вариантом расчета оценок t_k является усреднение значений условных вероятностей для всех m_r видов, встретившихся в тестируемом измерении:

$$t_k = \frac{1}{m_r} \sum_{i=1}^{m_r} p_{ik} \quad , \tag{8.39}$$

т.е. вероятность принадлежности пробы к k-му классу есть средняя вероятность класса k для всех видов, найденных в пробе.

Формула "простых средних вероятностей", традиционная для многих работ в этом направлении [Джурс, Айзенауэр, 1977; Авидон с соавт., 1978], часто дает вполне удовлетворительную точность. Однако ряд теоретико-вероятностных предположений заставляют усомниться в конечной оптимальности аппроксимации первого порядка. Поэтому другим вариантом усреднения вероятностей p_{ik} является использование известного углового преобразования Фишера [Урбах, 1964], при котором частотные оценки вероятностей имеют ошибку, почти не зависящую от самих вероятностей. Используемая при этом функция $\arcsin(2P - 1)$ ведет себя почти так же, как используемая в байесовских подходах функция $\ln(P/(1-P))$, но, в то же время, при P, близких к 0 или 1, не вырождается, устремляясь в бесконечность, а ограничена $\pm\pi/2$. На этих принципах основана работа компьютерной системы PASS [Poroikov et al., 2000], прогнозирующей спектр биологической активности химических соединений по их структурным формулам (адрес Интернет-версии: http://www.ibmh.msk.su/PASS). Тогда результирующие оценки принадлежности к классам некоторого тестируемого измерения x_p выражаются через условные вероятности p_{ik} и априорные вероятности классов p_k следующим образом:

$$t_k = \frac{1}{m_r} \sum_{i=1}^{m_r} J_i' * [\arcsin(2p_{ik} - 1) - \arcsin(2p_k - 1)] , \qquad (8.40)$$

где J_i' – индикаторный вес вида i, который мы интерпретировали как величину, обратную "шенноновской" энтропии распределения вероятностей по классам:

$$J'_i = 1/[1 - \sum_{k=1}^{p} p_{ik} \log_2(p_{ik})] . \qquad (8.41)$$

По нашему мнению, именно энтропия H, а не субъективно назначаемые баллы Сладечека, в явном виде характеризует равномерность распределения индикаторных валентностей по классам.

Алгоритм 3 – расчет валентностей с использованием методов оптимизации

Сделаем основные исходные предположения относительно природы индикаторных валентностей R_{ik}, оценивающих сродство (т.е. резистентность или экологическую значимость) i-го вида гидробионтов в k-м классе качества вод. Будем считать, что их величина является некоторой сложной математической функцией F от следующих факторов:

$$R_{ik} = F (n_i , v_k, J_i, k, a_{ik}) , \qquad (8.42)$$

где:

- n_i – количество измерений i-го вида, на основе которых вычислена R_{ik}. С позиций классической статистики, чем больше n_i, тем выше надежность и устойчивость рассчитываемых оценок. С другой стороны, гидробиологический опыт свидетельствует о том, что целостная картина наиболее характерных черт биоценоза определяется, в первую очередь, редкими видами, в то время, как массовые виды-эврибионты составляют некоторый "размытый фон";
- v_k – относительная вероятность оцениваемого класса. Поскольку частота встречаемости наблюдений различных классов неодинакова, то разумно предположить, что оценки R_{ik} для "редких" классов должны обладать некоторыми преимуществами при их сопоставлении. Примем в качестве конкретной дефиниции этой оценки класса величину n/n_k, обратную априорной вероятности k-го класса;
- J_i – индикаторная значимость вида i, которая, как было показано выше, формально может быть интерпретирована как величина, обратная к "шенноновской" энтропии распределения оценок по классам: $h_i = \sum_{k=1}^{p} a_{ik} \log_2(a_{ik})$; (8.43)
- k – порядковый номер класса. Поскольку биоразнообразие и обилие гидробионтов существенно уменьшается с увеличением номеров градаций качества воды, разумно предположить, что некоторым приоритетом должны обладать оценки классов, близких к p;
- a_{ik} – собственно индикаторные валентности или относительный вклад вида i в развитие биоценозов, характерных для каждого k-го класса качества, которые мы будем рассчитывать по

несколько видоизмененной формуле (8.32), в которой учтем как численность, так и биомассу видов (в мг/м^2), преобразовав их произведение в логарифмическую шкалу:

$$a_{ik} = \frac{\sum_{j=1}^{n_{jk}} \ln(\sqrt{B_{jk} N_{jk}})}{\sum_{k=1}^{6} \sum_{j=1}^{n_{jk}} \ln(\sqrt{B_{jk} N_{jk}})} . \qquad (8.44)$$

По аналогии с функцией желательности, примем мультипликативную модель обобщенных оценок, в которой будут учтены все пять перечисленных факторов:

$$R_{ik} = a_{ik} (n_i)^{\alpha} (n/n_k)^{\beta} [3.33/(h_i + 1)]^{\gamma} k^{\lambda}, \qquad (8.45)$$

где α, β, γ и λ – параметры модели, т.е. некоторые специально подобранные коэффициенты, 3.33 – константа, равная максимально возможному значению энтропии Шеннона при распределении валентностей по p классам ($p = 5$).

Для экзамена (т.е. определения класса водоема по комбинации видов, встретившихся в произвольно взятой пробе), как и в предыдущих версиях алгоритма, будем рассчитывать оценки принадлежности к классам по следующей формуле:

$$t_k = \frac{\sum_{i=1}^{m_t} \ln(\sqrt{N_i \cdot B_i}) \cdot R_{ik}}{\sum_{i=1}^{m_t} \ln(\sqrt{N_i \cdot B_i})}, \qquad (8.46)$$

и относить экзаменуемый вектор наблюдений к тому классу, которому соответствует максимальная оценка t_k из всех вычисленных.

Очевидно, что значения компонентов матрицы валентностей R, а, следовательно, и достоверность процесса распознавания, зависит от величин настроечных коэффициентов α, β, γ и λ, которые регулируют долю участия каждого из пяти перечисленных факторов в формировании оценок R_{ik}. Поэтому естественна постановка следующей оптимизационной задачи: *«необходимо найти такие α, β, γ и λ, которые сводят к минимуму число ошибок классификации при экзамене всех m примеров обучающей выборки в режиме скользящего контроля:*

$$D^2 = \sum_{j=1}^{m} (Y_j^{\phi} - Y_j^{p})^2 \Rightarrow \min \quad , \qquad (8.47)$$

где Y_j^{ϕ} и Y_j^{p} – расчетные и фактические значения класса качества вод».

Решение этой экстремальной задачи не может быть представлено в аналитическом виде или, по крайней мере, сведено к системе конечных линейных уравнений. Поэтому минимизация функционала D^2 может быть выполнена только специальными численными методами нелинейного программирования. Для нахождения оптимальных значений коэффициентов α, β, γ и λ воспользуемся модифицированным симплексным методом Нелдера-Мида [Банди, 1988; Гайдышев, 2001]. Этот метод, обеспечивающий достаточно быструю сходимость к экстремуму, не использует производные, что выгодно отличает его от других градиентных методов оптимизации. Основная идея симплекс-метода заключается в том, что по известным значениям целевой функции D^2 в вершинах выпуклого многогранника (симплекса) вычисляется вектор градиента, в направлении которого требуется сделать следующий шаг, чтобы получить наибольшее уменьшение критерия D^2. По новым точкам строится следующий симплекс и т.д., то есть, образно выражаясь, многогранник постепенно "перекатывается, двигаясь под горку" и, при этом, каждый раз находится наилучшее направление движения. Модификация Нелдера-Мида предполагает автоматическое изменение размеров ребер симплекса, что обеспечивает ему эффективное преодоление "оврагов" и быстрое движение по пологим спускам.

Результаты расчетов

Сформируем матрицу $\boldsymbol{X}$ ($\boldsymbol{n}$ = 542 , $\boldsymbol{m}$ = 546), содержащую информацию о 542 гидробиологических пробах, взятых на малых реках Самарской области. Элементами матрицы наблюдений являются значения численностей особей по 546 видам макрозообентоса, причем, диапазон частот встречаемости различных видов варьируется от 1 до 226 ($\boldsymbol{p_j} = \boldsymbol{n_j} / \boldsymbol{n}$ = 0.00184 ÷ 0.417). Каждое измерение отнесем к одному из классов качества вод по гидрохимическим показателям от 2 до 6 ($\boldsymbol{p}$ = 5). Априорные вероятности классов $\boldsymbol{p_k} = \boldsymbol{n_k} / \boldsymbol{n}$ варьируются от 0.103 для класса 2 до 0.356 для класса 4.

Алгоритм 1. Используем формулу П.А. Цимдиня (8.32) и методы анализа сапробных валентностей для прогнозирования класса качества воды. Индикаторные веса будем вычислять по приближенной формуле, связывающей $\boldsymbol{J_i}$ с мерой энтропии по Шеннону $\boldsymbol{H}$:

$$\boldsymbol{J_i} = 0.75\boldsymbol{H^3} - 2.8\boldsymbol{H^2} + 0.63\boldsymbol{H} + 5\,, \quad \text{где} \quad \boldsymbol{H} = \sum_{\boldsymbol{k}=1}^{6} \boldsymbol{a_{ik}} \log_2(\boldsymbol{a_{ik}}/10)/10 \qquad (8.48)$$

и дающей хорошую аппроксимацию весов по Сладечеку (коэффициент детерминации $\boldsymbol{R^2}$ = 0.943).

Для некоторых видов зообентоса рассчитанные значения индикаторных весов $\boldsymbol{J_i}$ и индикаторных валентностей $\boldsymbol{a_{ik}}$ для классов качества воды ($\boldsymbol{k}$ = 2, 3, 4, 5, 6) представлены в табл. 8.12. Для первых пяти видов дана более подробная информация, позволяющая провести детальный анализ механизма расчета.

Таблица 8.12

Значения индикаторных валентностей ($\boldsymbol{a_k}$) для классов качества воды и индикаторные веса $\boldsymbol{J}$, рассчитанные для некоторых видов зообентоса

($\Sigma\, \boldsymbol{N_{jk}}$ - сумма численностей вида, попавших в $\boldsymbol{k}$-й класс; $\boldsymbol{n_{ik}}$ – количество проб $\boldsymbol{i}$ вида, попавших в $\boldsymbol{k}$-й класс; $\boldsymbol{n_k}$ – общее число измерений $\boldsymbol{k}$-го класса)

Названия видов		Классы качества воды					По всем классам	Индикат. веса $\boldsymbol{J}$
		2	**3**	**4**	**5**	**6**		
1. Parachironomus varus	$\Sigma\boldsymbol{N_{jk}}$	63	107	490	603	1080	2343	1.4
	$\boldsymbol{n_{ik}}$	2	2	5	4	4	17	
	$\boldsymbol{n_{ik}}/\boldsymbol{n_k}$, %	3.6	1.5	2.6	4.3	5.8	3.14	
	$\boldsymbol{a_{ik}}$	0.3	0.5	2.1	2.6	4.6		
2. Parametriocnemus lundbecki	$\Sigma\boldsymbol{N_{jk}}$	628	316	40			984	3.2
	$\boldsymbol{n_{ik}}$	8	4	3			15	
	$\boldsymbol{n_{ik}}/\boldsymbol{n_k}$, %	14.3	3.1	1.6			2.76	
	$\boldsymbol{a_{ik}}$	6.4	3.2	0.4				
3. Polypedilum scalaenum	$\Sigma\boldsymbol{N_{jk}}$	60	3239	21697	290	320	25606	4.2
	$\boldsymbol{n_{ik}}$	3	15	22	4	2	46	
	$\boldsymbol{n_{ik}}/\boldsymbol{n_k}$, %	5.4	11.5	11.4	4.3	2.9	8.5	
	$\boldsymbol{a_{ik}}$		1.3	8.5	0.1	0.1		
4. Eukiefferiella minor	$\Sigma\boldsymbol{N_{jk}}$	10					10	5.0
	$\boldsymbol{n_{ik}}$	1					1	
	$\boldsymbol{n_{ik}}/\boldsymbol{n_k}$, %	1.8					0.2	
	$\boldsymbol{a_{ik}}$	10.0						
5. Eukiefferiella gr. *claripennis*		3.0	0.3	4.2	2.6			1.7
6. Cricotopus bicinctus		3.3	4.9	1.6	0.1			2.0
7. Cricotopus gr. *sylvestris*		2.7	0.0	2.9	0.0	4.4		2.0
8. Baetis rhodani		8.0	0.4	1.6				3.9
9. Ephemeroptera (прочие)		6.4	1.1	2.2	0.3			2.4
10. Simulium sp.		1.0	1.7	7.2	0.1			3.1
Количество измерений по классам $\boldsymbol{n_k}$		56	131	193	93	69	542	

Можно отметить следующие свойства оценок a_{ik}, вычисленных по формуле (8.32), не комментируя степень их позитивности или негативности:

- индикаторные валентности зависят только от соотношения групповых средних численности и никак не связаны с абсолютными значениями обилия: вид, встречающийся в единичных экземплярах, будет получать такие же оценки, что и вид, численность которого насчитывает тысячи особей;
- как и любые оценки, основанные на средних, валентности сильно зависят от характера распределения данных, наличия аномальных «выбросов» данных и проч.; так для вида *Polypedilum scalaenum* большое значение валентности 4-го класса ($a_4 = 8.5$) во многом определяется одним измерением с $N = 17840$ экз/м^2;
- наибольшим преимуществом в формировании оценок обладают редкие виды (в том числе, виды, встретившиеся только в одной пробе, например, *Eukiefferiella minor* – они автоматически приобретают максимальный вес 5 и максимальную валентность).

Рассмотрим методику классификации на примере данных, полученных на ст. 5 р. Сок (см. табл. 8.13). Индикаторные валентности и индикаторные веса для всех 7 видов зообентоса, найденных в этой пробе, указаны в табл. 8.12. Согласно данным гидрохимического анализа эта станция отнесена к 2 классу качества вод.

Таблица 8.13

Расчет по формуле (8.33) обобщенных индикаторных валентностей t_k на примере классификации данных экспедиционных наблюдений на ст. 5 р. Сок (30 июля 1999 г.)

Наименования видов	Численность экз/м^2	Для класса 2: Na_2J/10	Для класса 3: Na_3J/10	Для класса 4: Na_4J/10	Для класса 5: Na_5J/10	Для класса 6: Na_6J/10	Знаменатель NJ
Eukiefferiella minor	10	50					50
Cricotopus gr. *sylvestris*	150	82.91		88.44		135.98	307.33
Cricotopus bicinctus	10	6.64	9.75	3.29	0.21	0.09	19.98
Eukiefferiella gr. *claripennis*	10	5.03	0.44	7.03	4.40		16.90
Simulium sp.	10	3.09	5.32	22.04	0.30		30.75
Ephemeroptera (прочие)	10	15.48	2.79	5.33	0.74		24.33
Baetis rhodani	70	219.93	10.25	45.35			275.53
Итого		383.08	28.55	171.48	5.65	136.07	724.83
Средневзвешенные валентности		0.53	0.04	0.24	0.01	0.19	

Если ориентироваться на максимальные значения обобщенных индикаторных валентностей t_k, представленные в табл. 8.13, то при экзамене *безошибочно* выбираются $t_2 = 0.53$ и 2 класс качества.

Чтобы использовать формулу индекса сапробности в редакции И.К. Тодераша, необходимо предварительно обозначить параллель между зонами сапробности и классами качества вод по ГОСТ 17.1.3.07–82, для чего воспользуемся табл. 4.7 раздела 4.6, где это соответствие установлено: $\cdot t_2 = \cdot t_o$ для олигосапробной, $t_3 = t_\beta$ для β-мезосапробной, $t_4 = t_\alpha$ для α-мезосапробной и $\cdot t_5 = t_p$ для полисапробной зон. Подставив рассчитанные средневзвешенные сапробные валентности в формулу (8.35), получим для станции 5 р. Сок индекс сапробности:

$$S_R = 1\cdot 0.53 + 2\cdot 0.04 + 3\cdot 0.24 + 4\cdot 0.01 + 5\cdot 0.19 = 2.25,$$

согласно которому также уверенно, но уже *ошибочно* (!), тестируемый водоем относится к классу 3 (т.е. β-мезосапробной зоне с диапазоном индекса сапробности S_R от 1.5 до 2.5).

Для проверки работоспособности метода и отдельных его модификаций проведем оценку класса качества воды для всех 542 измерений зообентоса, использованных для обучения. Расчеты проведем тремя различными методами классифицирования, приведенными в п.п. 1.1-1.3 табл. 8.14.

Таблица 8.14

Сравнительный анализ адекватности различных методов вычисления оценок индикаторных валентностей и техник распознавания

Метод вычисления оценок принадлежности к классам	На всей выборке			Правильный прогноз без учета проб из 6 класса	Критерий оптимизации D^2
	Правильно распознано ед. / %	Ошибочно распознано ед. / %	Ошибка на 2 класса и более		
1.1. По натуральным численностям и с использованием индексов сапробности	260 48.0 %	282 52.0 %	90 16.6 %	259 54.8 %	691
1.2. По натуральным численностям и по максимумам оценок	315 58.1 %	227 41.9 %	88 16.2 %	307 64.9 %	607
1.3. По логарифмам численностей и по максимумам оценок	341 62.9 %	201 37.1 %	84 15.5 %	338 71.5 %	558
2.1. Простое усреднение оценок вероятности	319 58.9 %	223 41.1 %	86 15.9 %	317 67.0 %	571
2.2. Усреднение вероятностей с угловым преобразованием Фишера	366 67.5 %	176 32.5 %	81 14.9 %	350 74.0 %	580
3.1. Полная черехпараметрическая модель оптимальных оценок	391 74.9 %	131 25.1 %	41 7.9 %	359 78.6 %	274
3.2. Трехпараметрическая модель ($\alpha = 0,\ \beta = 0.41,\ \gamma = 1.21,\ \lambda = 0.71$)	370 70.9 %	152 29.1 %	54 10.3 %	345 75.5 %	345
3.3. Трехпараметрическая модель ($\lambda = 0,\ \alpha = -0.53,\ \beta = 0.39,\ \gamma = 1.189$)	354 67.8 %	168 32.2 %	77 14.8 %	343 75.1 %	499
3.4. Двухпараметрическая модель ($\lambda = 0,\ \alpha = 0,\ \beta = 0.35,\ \gamma = 1.225$)	345 66.1 %	177 33.9 %	76 14.6 %	339 74.2 %	500
3.5. Двухпараметрическая модель ($\lambda = 0,\ \beta = 0,\ \alpha = -0.47,\ \gamma = 1.185$)	345 66.1 %	177 33.9 %	79 15.1 %	340 74.4 %	514
3.6. Однопараметрическая модель ($\lambda = 0,\ \alpha = 0,\ \beta = 0,\ \gamma = 1.2$)	329 63.0 %	193 37.0 %	82 15.7 %	325 71.1 %	534

Примечание: правильный прогноз соответствует верной оценке нужного класса качества из 5 возможных альтернатив, т.е. случайный процесс угадывания соответствует вероятности 20%.

По результатам расчета можно сделать следующие выводы:

1. Подтверждается предположение о принципиальной математической несостоятельности подхода Пантле-Букка и всей техники вычисления индексов сапробности. Попытка выразить через "сапробный центр тяжести" некоторую функцию распределения вероятностей принадлежности измерения к 4 зонам сапробности (или 6 классам качества) неизбежно приводит к усреднению факторов видовой структуры биоценоза, а, следовательно, к сильному смещению прогноза в диапазон индекса сапробности от 2 до 3. В частности, по методу 1.1 (табл. 8.14) 371 наблюдение из 542 было отнесено к "среднему" классу 4 (т.е. α-мезосапробной зоне), причем в 178 случаях это было сделано ошибочно.
2. Вектор значений валентностей Зелинки-Марвана $\boldsymbol{T}$ более полно отражает индикаторное значение видов в сообществе. Принадлежность пробы $\boldsymbol{x_p}$ к водоему $\boldsymbol{k}$-й категории целесообразней осуществлять по $\max \boldsymbol{t_k}$ – наибольшей оценке средневзвешенные валентности из $\{\boldsymbol{t}_x, \boldsymbol{t}_o, \boldsymbol{t}_\beta, \boldsymbol{t}_\alpha, \boldsymbol{t}_p\}$.
3. Выбросы обучающей выборки, т.е. аномально высокие значения численностей по некоторым видам в отдельных пробах, могут сильно сказываться на устойчивости расчета индикаторных валентностей. Этим можно объяснить относительно большое для метода 1.2 количество "грубых" ошибок прогноза на два класса и более. При использовании в методе 1.3 предварительно логарифмированных значений численностей результаты можно назвать вполне стабилизировавшимися. Для сравнения приведем значения индикаторных валентностей, полученных по методу 1.3 для вида *Polypedilum scalaenum* (Sch.) , которые сильно отличаются от представленных в табл. 8.12: $\boldsymbol{A} = \{0.5;\ 3.2;\ 5.1;\ 0.8;\ 0.5\}$ при $\boldsymbol{J} = 1.6$.
4. Весьма неудовлетворительными оказались результаты экзамена для проб, взятых в загрязненных водоемах класса 5 и, особенно, 6, на которые пришлась преимущественная доля ошибок.

Сказывается, с одной стороны, некоторая жесткость санитарно-гигиенического подхода к оценке класса качества по гидрохимическим показателям, когда по некоторому лимитирующему фактору, например, концентрации поверхностно-активных веществ, водоем относится к 6 классу, что не мешает развиваться структурно деформированному, но количественно полноценному сообществу гидробионтов. С другой стороны, ошибки определяются и чисто статистическими эффектами: то небольшое количество видов-эврибионтов, обычно характерное для грязных водоемов, недостаточно для точной его идентификации, поскольку эти виды имеют невысокие индикаторные веса J и валентности, вносящие равномерный вклад в средневзвешенные оценки t_k.

Алгоритм 2. Используем в качестве индикаторных индексов оценки вероятности встречаемости видов. Результаты экзамена примеров обучающей выборки для двух вариантов: по формуле "простых средних вероятностей" и с использованием углового преобразования Фишера приведены в пунктах 2.1 и 2.2 табл. 8.14.

Расчеты показывают, что вероятностные оценки индикаторных индексов, полученные на основе частот встречаемости видов и не учитывающие абсолютные значения численностей организмов, оказались существенно лучше оценок индикаторных валентностей, рассчитанных с использованием формулы, включающей натуральные значения численностей (пункты 1.1 и 1.2 табл. 8.14). Это – лишнее подтверждение известного тезиса о некорректности сопоставления средних значений без учета закона статистического распределения выборок.

Алгоритм 3. Выполним расчет индикаторных валентностей, обеспечивающих минимум ошибок классификации. В качестве начальных приближений для полной модели

$$R_{ik} = a_{ik} (n_i)^{\alpha} (n/n_k)^{\beta} (3.33/(h_i+1))^{\gamma} k^{\lambda}$$

выберем, по априорным соображениям, наиболее ожидаемые значения $\alpha = 0.2$, $\beta = 0$, $\gamma = 1$ и $\lambda = 0.5$. В ходе симплекс-процедуры по методу Нелдера-Мида было просчитано 160 вариантов матрицы R размерностью 546×5 и для каждого варианта найдено число ошибок классификации D^2, пока, наконец, не было найдено оптимальное решение: $\alpha = -0.52$, $\beta = 0.41$, $\gamma = 1.21$, $\lambda = 0.68$, т.е. расчетная формула приобрела вид:

$$R_{ik} = \{a_{ik} (n/n_k)^{0.41} [3.33/(h_i+1)]^{1.21} k^{0.68}\} / n_i^{0.52} .$$

Значения полученных коэффициентов λ, β и γ полностью подтверждают наши исходные предпосылки, в то же время как отрицательная величина коэффициента α приводит к нетрадиционному для статистики, но важному для биологии выводу о том, что основными индикаторами состояния экосистем являются редкие виды.

Поскольку переусложнение расчетной модели так же вредно, как и ее недоусложнение, оценим степень вклада каждого из перечисленных выше параметров. Для этого выполним серию расчетов матрицы оптимальных оценок R по моделям разной степени сложности: по полной четырехпараметрической модели и по упрощенным моделям, в которых поочередно элиминировались один или несколько факторов, т.е. значения параметров α, β или λ априори принимался равным нулю.

Полученные результаты представлены в табл. 8.14 строками 3.1-3.6. Серия расчетов с исключением из общей формулы для валентностей R отдельных факторов дает уверенное основание полагать, что все компоненты модели (8.45) являются информативными для прогноза класса качества вод, поскольку снижение эффективности распознавания при элиминации любого из коэффициентов оказывается весьма существенным. Однопараметрическая модель оценок 3.6, учитывающая только такой фактор, как уровень доминантности индикаторных валентностей (т.е. индикаторный вес J_i в понимании Зелинки-Марвана), методически полностью соответствует модели 1.3 и отличается от нее лишь использованием комбинированного показателя обилия $(N\cdot B)^{0.5}$ вместо численности экземпляров зообентоса N.

Принципиальное отличие рассчитанных индикаторных валентностей от традиционных валентностей Зелинки-Марвана заключается в том, что для последних вводится жесткое условие нормировки – сумма сапробных валентностей должна быть равна 10. Нам представляется, что введение этого условия не связано с каким-то другим содержательным смыслом, кроме стремления непременно пересчитывать оценку по валентностям во вторичные по идеологии и нерезультатив-

ные по технике индексы сапробности Пантле-Букка. Достаточно отказаться от этого условия, чтобы получить другую, гораздо более привлекательную возможность – сравнивать между собой отдельные виды по их индикаторной значимости в данной зоне сапробности или водоемах определенного класса качества. Если сопоставить для отдельных видов характер распределения вычисленных нами значений R_{ik} по классам и сапробных валентностей, взятых из литературных источников (см. табл. 8.15) то можно усмотреть неплохое соответствие оценок для чистых водоемов, в то время как виды, объявленные классическими полисапробами, в наших региональных условиях оказались эврибионтами и не набрали высокой индикаторной значимости.

Не имея технической возможности представить полный список всех 546 видов зообентоса, приведем в табл. 8.16 индикаторные валентности для некоторых видов хирономид, сгруппировав их по отводимой им роли в отдельных классах качества вод. Нетрудно прийти к выводу, что наибольшей индикаторной значимостью обладают редко встречающиеся виды.

Доля правильной классификации качества вод, достигнутая описанными эвристическими методами распознавания с использованием видов зообентоса в качестве алфавита признаков, оказалась существенно выше, чем при аналогичных расчетах методами упорядоченного пробита и дискриминантного анализа на основе обобщенных гидробиологических индексов (см. разделы 8.2 и 8.3). В дополнение к этому, полученные результаты могут быть серьезно улучшены за счет чисто математических операций.

Во-первых, для повышения достоверности и эффективности вычисляемых индикаторных индексов необходимо предварительно провести тщательный ручной или автоматизированный отбор гидробиологических измерений для включения в обучающую выборку. Расчет индикаторных валентностей предпочтительней проводить не на всем массиве наблюдений, а на некотором "опорном" подмножестве надежных примеров, где сведены до минимума случайные ошибки измерений и влияние таких посторонних факторов, как сезонность, неудачный выбор места отбора пробы и т.д. Мы этого в нашей работе не сделали из принципиальных соображений, желая оценить общий уровень и степень влияния "шума", неизбежно сопровождающего гидробиологические измерения. Зато проведенный тщательный анализ ошибок экзамена показал, что не менее 100 проб из 542 вообще не могут быть правильно классифицированны ни компьютером, ни человеком.

Во-вторых, качество любой процедуры классификации определяет выбор информативного комплекса видов-индикаторов. Ввиду сложности гидробиологических объектов часто возникает тенденция непременно учесть в анализе все наблюдаемые виды, независимо от их реальной индикаторной значимости. Оптимальные классификационные процедуры всегда преследуют иную цель – добиться наибольшего эффекта наименьшим числом признаков, для чего большое внимание уделяется выбору наиболее информативного пространства переменных. Среди существующих способов предварительного отбора признаков можно упомянуть метод минимизации дивергенции, аппроксимацию функции распределения видового обилия в разных классах качества вод, метод максимизации кластеризуемости и т.д., о чем частично шла речь в главе 7.

В заключение затронем такую важную проблему как интерпретация и способ использования рассчитанных оценок индикаторных валентностей. По укоренившейся в гидробиологии традиции, полученные индексы рекомендуется использовать "As is" (т.е. "как они есть") для практического определения классов качества вод исследователям из других регионов. Мы полагаем, что оценки, полученные нами, работоспособны только для определенного типа водоемов (малых и средних равнинных рек) и в определенных ландшафтно-географических условиях. Неэффективность концепции "всемирных" сапробных валентностей, пригодных для водоемов любого типа или местоположения, стала очевидной еще 30 лет назад и было бы иллюзией надеяться на ее возрождение в ином информационном качестве. В то же время, развитие вычислительной техники, корпоративных баз данных и Интернет-технологий создают предпосылки для динамического формирования обучающей выборки и расчета "контекстно-зависимых" валентностей, оптимальных в условиях тестируемого комплекса измерений. В этом разделе мы ставили своей целью показать, какими математическими средствами и приемами может быть решена эта задача.

Наконец, как отмечалось выше, любая естественно-научная теория должна выполнять, как минимум, две функции: "объяснения" и "прогнозирования" наблюдаемых феноменов [Розенберг, 1988; Розенберг с соавт., 1999], причем для сложных систем объединение в одной модели этих функций невозможно.

Таблица 8.15

Вычисленные нами оценки валентностей для классов качества вод и сапробные валентности для некоторых видов зообентоса

Наименование видов зообентоса	Оценки классов качества					Сапробные валентности по зонам					
	2	**3**	**4**	**5**	**6**	**k**	**o**	**β-m**	**α-m**	**p**	***J***
Sphaeriastrum rivicola		0.1	0.2	0.2	0.1			3	5	2	2
Chaoborus sp.			2.3	0.8		-1	2	4	2	2	1
Ablabesmyia monilis		0.3	0.3	0.1	0.1		2	5	3		2
Ablabesmyia phatta	0.6	0.6	1.0				2	5	3		2
Eukiefferiella bavarica	1.9	1.4				1	6	3			3
Eukiefferiella coerulescens		10.0				1	6	3			3
Eukiefferiella longicalcar		10.0					7	3			4
Eukiefferiella longipes	10.0					1	5	4			3
Psectrotanypus varius		0.5	2.2					2	7	1	3
Paratanytarsus austriacus	1.8	1.5					4	6			3
Stempellinella minor	1.5		2.1				5	5			3
T.pallidicornis	0.1	0.3	0.5				2	6	2		3
Asellus aquaticus		0.7	0.4		0.3			2	8	-1	4
Astacus astacus		6.7				3	4	3			2
Dreissena polymorpha		0.2	0.1	0.3	0.5		4	6			3
B.bioculatus		2.1	1.1				1	6	3		3
Baetis rhodani	0.8	0.2	0.5			3	3	3	1		1
B.vernus	2.0	1.4					2	5	3		2
Centroptilum luteolum	0.2	0.2	0.3	0.2			2	7	1		3
Cloeon dipterum	0.2	0.3	0.3				3	4	3		2
Caenis macrura		0.2	0.2			4	4	2			2
Ecdyonurus sp.	10.0						4	5	1		2
Ephemera danica			9.4			1	4	4	1		1
Ephemerella ignita	0.5	0.4	0.2	0.2		1	3	3	3		1
Sipholonurus linneanus	1.9	1.5					3	4	3		2
Hydrobates fluviatilis	1.6	1.8					4	6	-1		3
Glossiphonia complanata	0.1	0.3	0.5	0.4				6	4		3
Herpobdella octoculata	0.4	0.2	0.3				-1	2	6	2	3
Limnodrilus hoffmeisteri			0.1	0.1	0.1				4	6	3
Tubifex tubifex		0.1	0.2	0.2	0.1			-1	2	8	4
Leuctra fusca	5.8						2	5	3		2
Leuctra sp.	2.2	1.3					2	5	3		2
Nemoura cinerea	1.6	0.6					4	4	2		2
Nemoura sp.	0.8	0.4	0.7				4	4	2		2
Atherix ibis	10.0					2	5	3	-1		2
Atherix sp.	0.4	0.5	0.2			2	5	3	-1		2
Simulium sp.	0.2	0.2	0.1	0.1		3	3	2	2		1
Hydropsyche ornatula	0.2	0.2	0.9					6	4		3
Hydropsyche sp.	1.0	0.5	0.6			1	2	4	3		1
Hydroptila sp.	5.8						3	6	1		3
Lepidostoma hirtum	1.2		2.6				3	7			4
Neureclipsis bimaculata	10.0					-1	6	4			3
Polycentropus flavomaculatus		10.0				1	3	4	2		1
Rhyacophila sp.	1.4	0.0	1.4				5	5			3

Таблица 8.16

Индикаторные валентности некоторых видов хирономид, характерные для отдельных классов качества воды малых и средних рек Самарской области

Виды хирономид	Индикаторные валентности классов качества					Встре-чаемость
	2	**3**	**4**	**5**	**6**	
Индикаторы класса 6						
Sergentia gr. *longiventris*					9.3	1
Stenochironomus sp.					9.3	1
Glyptotendipes paripes			0.9		1.5	2
Cricotopus gr. *intersectus*	0.6		0.4		0.6	3
Индикаторы класса 5						
Lipiniella agrayloides				8.3		1
Endochironomus sp.				5.8		2
E. donatoris			0.9	1.4		2
Ablabesmyia longistyla			0.9	1.3		2
Индикаторы класса 4						
Lymnophyes sp.			6.3			1
Brillia sp.			4.3			2
Monopelopia sp.			4.3			2
Polypedilum scalaenum						
Polypedilum sordens			4.3			2
Rheopelopia sp.			4.3			2
Pagastia sp.			4.3			2
Glyptotendipes barbipes			3.5			3
Microtendipes gr. *rydalensis*	1.1		1.3			2
Индикаторы класса 3						
Psectrocladius gr. *dilatatus*		7.3				1
Harnischia sp.		5.0				2
Hydrobaenus distylus		5.0				2
Metriocnemus gr. *hydropetricus*		5.0				2
Psectrocladius simulans		5.0				2
Conchapelopia melanops		5.0				2
Chironomus anthracinus		1.6	0.1			9
Diamesa heterodentata	1.1	1.5				2
Tanytarsus usmaensis		1.5	0.7			2
Индикаторы класса 2						
Paratrichocladius rufiventris	10.0					1
Parorthocladius sp.	7.0					2
Paramerina sp.	7.0					2
Cricotopus albiforceps	7.0					2
Orthocladius oliveri	7.0					2
Mesocricotopus sp.	7.0					2
Tvetenia discoloripes	7.0					2
Cricotopus gr. *cylindraceus*	5.7					3
Rheocricotopus effusus	5.7					3
Pseudodiamesa nivosa	4.9					4
Paratanytarsus austriacus	1.8	0.9				2
Cryptotendipes sp.	1.7	1.0				2
Telopelopia sp.	1.7	0.2				7
Плохие индикаторы						
Polypedilum nubeculosum	0.01	0.03	0.04	0.03	0.02	187
Chironomus plumosus		0.02	0.05	0.03	0.03	195
Procladius ferrugineus	0.01	0.04	0.05	0.01	0.02	177

Tanytarsus sp.	0.03	0.04	0.04	0.01	0.01	184
Cryptochironomus gr. *defectus*	0.01	0.03	0.06	0.03	0.02	143
Cladotanytarsus mancus	0.02	0.05	0.05	0.01	0.02	139
Cricotopus sylvestris	0.02	0.03	0.06	0.03	0.03	100
Cricotopus bicinctus	0.05	0.06	0.05	0.01		124
Dicrotendipes nervosus		0.06	0.04	0.04	0.07	75
Paratanytarsus confusus	0.04	0.08	0.07	0.03	0.01	64
Microchironomus tener	0.01	0.06	0.09	0.05	0.03	58
Cladopelma gr. *lateralis*	0.01	0.05	0.11	0.04	0.03	59
Procladius choreus	0.05	0.06	0.11	0.03	0.02	50
Prodiamesa olivacea	0.06	0.07	0.11	0.01		61
Paracladius conversus	0.05	0.11	0.08	0.02		55

Индикаторные валентности, выборочно приведенные в табл. 8.16, были рассчитаны нами исключительно для выполнения конкретной задачи – обеспечить минимум ошибок прогноза класса качества воды и изначально не предназначались для "объяснения" (например, формирования каких-либо научных гипотез о роли того или иного вида в общей системе классификации водоемов).

8.5. Задача о двух классах и разделяющей гиперплоскости: метод «обобщенного портрета»

Формулировка задачи

Пусть зависимая переменная $\boldsymbol{Y}$, отражающая качество вод, принимает одно из двух значений. Класс 1 может, например, трактоваться как "Чисто, "Норма" или "Относительно удовлетворительная ситуация", а 0 – как "Грязно", "Патология", "Чрезвычайная экологическая ситуация" и т.п. В соответствии с этой классификацией, таблица гидробиологических наблюдений, соответствующая обучающей выборке, разделяется на два конечных подмножества векторов: $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_\mathbf{a}$ и $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{\boldsymbol{b}}$. Элементы $\boldsymbol{x_i}$, $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{m}$, векторов $\boldsymbol{X}$ и $\tilde{\boldsymbol{X}}$ соответствуют варьируемым переменным, в состав которых могут входить значения обилия $\boldsymbol{i}$-го вида в пробе, а также всевозможные обобщенные индексы и гидробиологические показатели.

Необходимо найти такое уравнение оптимальной гиперплоскости в $\boldsymbol{m}$-мерном пространстве признаков

$$\boldsymbol{X}\varphi_0 - \mathbf{c}_0 = 0\,, \tag{8.49}$$

которая разделяет точки множеств $\boldsymbol{X}$ и $\tilde{\boldsymbol{X}}$ и, в то же время, наиболее удалена от выделяемых областей – выпуклых оболочек каждого из этих множеств. Геометрическая интерпретация задачи разделения двух подмножеств точек плоскостью (точнее, прямой линией) для случая двух переменных представлена на рис. 8.6.

Рекомендуемая литература: [Вапник, Червоненкис, 1974; Алгоритмы и программы.., 1984].

Математический лист

Пусть в некотором пространстве существует две (или несколько) областей, не имеющих общих точек, и что измерения – точки из этих областей. Каждой такой области можно приписать наименование, т. е. дать название, соответствующее образу. Сами эти области заранее не определены, т. е. нет каких-либо сведений о расположении их границ или правил определения принадлежности точки к той или иной области.

Процесс обучения распознаванию образов состоит в том, что предъявляются точки, случайно выбранные из этих областей. Необходимо построить поверхность, которая разделяла бы не только показанные в процессе обучения точки, но и все остальные точки, принадлежащие этим областям. Иначе говоря, цель обучения состоит в построении таких функций от векторов-измерений, которые были бы, например, положительны на всех точках одного и отрицательны на всех точках

другого образа. Если предъявляемые измерения принадлежат не двум, а большему числу образов, то задача состоит в построении поверхностей, разделяющей все области друг от друга.

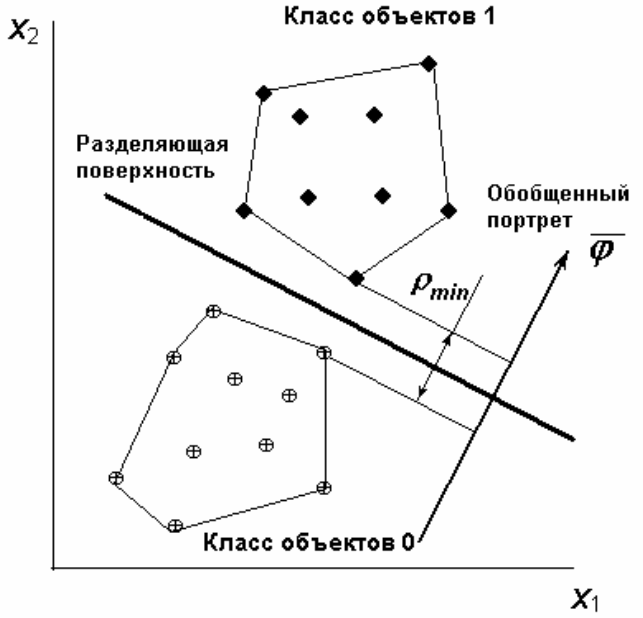


Рис. 8.6. Геометрическая интерпретация метода обобщенного портрета

Класс ***R***-моделей распознавания образов, использующих принцип разделения, основан на гипотезе полимодальности плотности распределения объектов в исходном признаковом пространстве. Иными словами, объекты группируются в "сгустки" точек, которые могут быть разделены друг от друга некоторой гиперповерхностью и при этом достигается приемлемая ошибка классификации.

Конкретные алгоритмы этого типа построены по единой схеме: из класса решающих правил подходящей емкости выбирается правило, минимизирующее количество неправильных опознаний на обучающей выборке. Во многих случаях класс решающих функций задается параметрически, т.е. считается, что вид функции известен с точностью до значения конечного числа параметров (коэффициентов модели). Самыми распространенными являются представления разделяющих функций в виде линейных уравнений, обобщенных нелинейных полиномов, эволюционных моделей и моделей метода группового учета аргументов (МГУА). Близки этой концепции и ***П***-модели, реализующие метод потенциальных функций, основанный на используемой в физике идее потенциала – всюду положительной и монотонно убывающей функции расстояния.

При использовании методов, основанных на предположениях о виде решающих функций, исследователь, прежде всего, обращается к линейным моделям. Это обусловлено высокой размерностью пространства признаков, характерной для реальных задач, вследствие чего при повышении степени полиноминальной решающей функции имеет место огромный рост числа ее членов при проблематичном сопутствующем повышении качества распознавания. Как отмечалось ранее, свойства линейных диагностических моделей, в которых изучаемый показатель представлен взвешенной суммой исходных признаков, хорошо изучены. Результаты этих моделей при соответствующем нормировании легко интерпретируются как расстояния от исследуемых объектов до некоторой гиперплоскости в пространстве признаков или, что эквивалентно, как проекции объектов на некоторую прямую линию в данном пространстве.

Рассмотрим метод нахождения разделяющей гиперплоскости, т.е. функции из класса линейных по параметру решающих правил. Наиболее конструктивным алгоритмом в этой области является *метод обобщенного портрета*, строгое математическое обоснование которого выполнено В.Н. Вапником и А.Я. Червоненкисом [1974]. Построение разделяющей поверхности идет здесь следующим образом. Ищется такое направление φ_0 в полном пространстве признаков, чтобы проекции выпуклых оболочек точек обучающей выборки первого и второго класса на это направление были максимально удалены друг от друга:

$$\varphi_0 = \max_{\varphi}\left[\min_{x_i \in \mathbf{X}} x_i\varphi - \max_{\tilde{x}_i \in \tilde{\mathbf{X}}} \tilde{x}_i\varphi\right] \tag{8.50}$$

Как показано на рис. 8.6, оптимальная разделяющая плоскость проводится перпендикулярно выбранному направлению φ_0 через середину отрезка, соединяющего проекции разделяемых областей:

$$c_0 = \frac{\min\limits_{x_i \in X} x_i\varphi + \max\limits_{\tilde{x}_i \in \tilde{X}} \tilde{x}_i\varphi}{2}. \tag{8.51}$$

Эта разделяющая гиперплоскость отделяет точки множества X, для которых $x\varphi_0 > c_0$, от точек множества $\tilde{X}$, для которых $x\varphi_0 < c_0$, а ее направляющий вектор φ_0 и называется, собственно, «*обобщенным портретом*».

Основным достоинством методов, основанных на предположениях о классе решающих функций, является ясность математической постановки задачи распознавания, как задачи поиска экстремума. В частности, нахождение оптимальной разделяющей гиперплоскости по методу обобщенного портрета представляет собой задачу квадратичного программирования, решение которой опирается на теорему Куна-Таккера. При этом ищется точка α_0, соответствующая положительному максимуму квадратичной формы

$$W(\alpha) = \sum_{i=1}^{a}\sum_{j=1}^{b}\alpha_{ij} - \psi^t\psi/2, \quad \text{где} \quad \psi = \sum_{i=1}^{a}\sum_{j=1}^{b}\alpha_{ij}(x_i - \tilde{x}_j). \tag{8.52}$$

Поиск экстремума достигается с помощью достаточно хорошо изученных специальных алгоритмов перцептронного типа (модификация метода Гаусса-Зайделя) или стандартных градиентных методов, к которым относится, в частности, метод сопряженных градиентов [Алгоритмы и программы.., 1984].

Особая ситуация возникает, когда безошибочное разделение векторов невозможно. Это бывает, если классы трангрессируют, а проекции их выпуклых оболочек перекрываются, либо "зазор" между классами ρ_{min} меньше некоторого заданного значения ρ_0. В этом случае из обучающей выборки исключается вектор, наиболее препятствующий успешному разделению. Затем, если разделение все еще невозможно, из оставшегося множества удаляется следующий малоинформативный элемент. Поиск продолжается до тех пор, пока либо задача не будет решена, либо число исключенных точек не превзойдет заданную долю общего числа векторов в обучающей выборке. В случае успешного нахождения обобщенного портрета, оставшаяся совокупность векторов называется информативной.

Важнейшей проблемой является проверка адекватности моделей прогнозирования, т.е. оценка достоверности полученных решающих правил. Эффективность работы различных алгоритмов распознавания образов оценивается с использованием критериев качества, которыми, в общем случае, являются либо собственно вероятности ошибочной классификации, либо связанные с ними некоторые функции потерь. При этом различают условную вероятность ошибочной классификации, ожидаемую ошибку алгоритма классификации на выборке заданного объема и асимптотическую ожидаемую ошибку классификации. Для оценки выбранного показателя качества того или иного алгоритма применяется три основных экспериментальных способа:

- выборка используется одновременно как обучающая и контрольная;
- выборка разбивается на две части - обучающую и контрольную;
- используется оценка "скользящего контроля".

Первый способ соответствует *критерию внутренней непротиворечивости модели*, выраженному в частотах ошибок первого и второго рода на обучающей выборке (или процентах несовпадений в классификациях "учителя" и машины). Способ дает завышенную оценку качества распознавания по сравнению с той же оценкой на независимых от обучения данных. Второй способ оценивает адекватность на *внешнем дополнении* (т.е. на экзаменуемых примерах, не участвовавших в построении самой модели). Метод является самым простым и убедительным, и им широко пользуются, если экспериментальных данных достаточно.

Оценка *скользящего контроля* формируется при работе алгоритма, который состоит в том, что из обучающей выборки поочередно удаляются описания одного объекта, на оставшемся материале строится решающее правило и по нему классифицируется исключенный объект. Такая процедура повторяется $(n - 1)$ раз. Доля правильных опознаний при скользящем контроле является несмещенной оценкой вероятности ошибки на всей генеральной совокупности и, следовательно, наиболее репрезентативной оценкой качества модели. Однако этот метод является и самым трудоемким, так как требует многократного построения правила распознавания.

Основной задачей построения моделей является структурная идентификация решающего правила, связанная с выбором оптимального признакового пространства, дающего наилучшее качество прогнозирования. Кроме чисто технических проблем, связанных с математической обработкой плохо обусловленных матриц большой размерности, методологически верно максимально учитывать известное предостережение Уильяма Оккама: *«частностей должно быть не больше, чем их необходимо»*. Действительно, сложная модель прогнозирования, основанная на бессистемном привлечении множества малоинформативных переменных, далеко не всегда оказывается оптимальной: добившись высокого качества предсказания на обучающей выборке, вероятность ошибки классификации векторов генеральной совокупности для такой модели часто возрастает.

Одним из путей борьбы с "проклятием размерности" является использование специальных пошаговых алгоритмов (см. также разделы 8.1 и 8.3), предназначенных для формального выделения в исходном пространстве переменных такого оптимального подпространства, в котором построенная гиперплоскость доставляет минимум критерию P, учитывающему как вероятность ошибок классификации, так и число использованных признаков. При прямой пошаговой процедуре фиксированное подпространство переменных наращивается добавлением на каждом шаге одного признака, в наибольшей степени улучшающего критерий P. Обратная пошаговая процедура заключается в исключении признаков до тех пор, пока критерий P не достигнет минимума. Специальный *метод последовательных включений с исключениями* представляет собой комбинацию прямой и обратной пошаговых процедур.

В отличие от параметрических методов распознавания, успешность применения метода обобщенного портрета мало зависит от рассогласования теоретических представлений о законах распределения объектов в пространстве признаков с эмпирической реальностью. Метод позволяет использовать любой способ представления информации – непрерывный, когда координаты вектора x могут принимать любые значения, и дискретный, которым удобно кодировать признаки, измеренные в порядковой шкале. Использованный нами программный комплекс *FOP* [Алгоритмы и программы.., 1984] предполагает предварительное преобразование всего исходного пространства признаков в бинарную форму, в результате чего область изменения каждого признака x_i разбивается на k интервалов с произвольным числом градаций $(9 \geq k \geq 2)$. Далее признаку x_i ставится в соответствие k бинарных признаков z_{ij} по традиционному правилу:

- $z_{ij} = 1$, если значение x_i лежит в j–м интервале;
- $z_{ij} = 0$, в противном случае.

Переход в пространство бинарных признаков реализует целый ряд преимуществ:

- все исходные признаки, независимо от разброса их индивидуальных значений, нормируются в единую (или, по крайней мере, сопоставимую) шкалу;
- в обобщенном портрете легко учитывается нелинейное и немонотонное влияние признака на результат классификации (разделяющая поверхность в исходном пространстве признаков является по своей сути нелинейной);
- в моделях прогнозирования корректно используется весь комплекс как количественных, так и качественных признаков (наличие или отсутствие видов).

Тогда уравнение разделяющей гиперплоскости в пространстве бинарных признаков имеет вид

$$\sum_{i=1}^{m}\sum_{j=1}^{k(i)}\lambda_{ij} - \beta = 0 , \qquad (8.53)$$

где m – число исходных признаков, λ_{ij} – настраиваемые коэффициенты, связанные кусочно-постоянными функциями с элементами вектора обобщенного портрета φ, β – свободный член

уравнения. С использованием этого уравнения *решающее правило* заключается в расчете расстояния η в многомерном пространстве от тестируемого объекта z' до разделяющей гиперплоскости и классификации наблюдения по следующей простой процедуре: если $\sum_{i=1}^{m}\sum_{j=1}^{k(i)}\lambda_{ij} - \beta \geq 0$, то экзаменуемый вектор z' относится к классу 1, в противном случае – к классу 0.

Следует отметить, что алгоритм обобщенного портрета применим и для распознавания объектов, относящихся более чем к двум классам, путем использования одного из методов: последовательной дихотомии или "один против всех". Например, можно сформировать обучающую выборку с априорными оценками трех классов: 2 – "стабильный", 1 – "кризисный" и 0 – "экологическое бедствие". Применима следующая стратегия обработки данных:

- рассчитывается гиперплоскость, разделяющая класс 2 (стабильный) от 0 + 1 (все остальные менее благополучные объекты);
- рассчитывается гиперплоскость, разделяющая класс 0 (бедствие) от 1 + 2 (все остальные более благополучные объекты);
- два полученных решающих правила позволяют непротиворечиво распознать одну из трех градаций для любого из экзаменуемых объектов.

В работе Б.А. Курляндского с соавторами [1988] описан пример прогнозирования количественных значений гигиенических нормативов и ПДК методом построения обобщенного портрета.

Линейные модели распознавания адекватны только простым геометрическим конфигурациям областей пространства признаков, в которые отображаются объекты разных классов, выделенных "учителем". При более сложных распределениях, порождающих геометрическую неоднородность объектов, эти модели принципиально не могут отражать многие особенности структуры экспериментальных данных, которые способны нести ценную диагностическую информацию.

С помощью методов, основанных на дополнительных предположениях о классе решающих функций (эволюционных, МГУА, нейросетевых и т.д. – см. главу 9), можно строить диагностические модели высокой сложности и получать практически приемлемые результаты. В то же время, достижению практических целей в этом случае не всегда сопутствует извлечение новых знаний о природе распознаваемых объектов. Возможность извлечения этих знаний, в частности, об экологических механизмах взаимодействия видов (признаков), принципиально ограничена емкостью класса решающих функций. Поэтому максимально, что можно сказать после построения той или иной диагностической модели высокой сложности – это перечислить комбинации признаков и сами признаки, вошедшие в результирующую модель. Но содержательный смысл явлений, отражающих внутреннюю природу и структуру исследуемых объектов, в рамках нелинейного подхода часто остается нераскрытым. Более простые линейные модели, используемые для целей "объяснения", имеют в этом смысле несомненные преимущества.

Результаты расчетов

В отличие от примеров предыдущих разделов, уравнения обобщенного портрета рассчитаем на основе опорного множества специально подобранных примеров с заранее известным откликом. Для формирования обучающей выборки из общего массива наблюдений малых рек Самарской области выделим: зону (т.е. область *образа*) с чрезвычайной экологической ситуацией (43 станции класса 0) и зону относительно экологической стабильности (43 станции класса 1).

При отборе и классификации точек наблюдения будем учитывать весь комплекс абиотических показателей, территориально-статистические и экспертные оценки степени антропогенной нагрузки, а также степень изученности и непротиворечивости данных. Всем измерениям, включаемым в обучающую выборку, присвоим признак класса 0 с *чрезвычайной экологической ситуацией* (ЧЭС) или 1 *относительной экологической стабильности* (ОУС) в соответствии с предварительной классификацией станций. Гидробиологические пробы на остальных 160 станциях наблюдения будем использовать лишь для экзамена.

Сформированная таким образом компактная обучающая выборка содержит $n = 160$ векторов x с исходными данными, из которых 70 было отнесено к классу ОУС и 90 к классу ЧЭС. На примере этой обучающей выборки выполним три варианта расчетов.

Вариант 1. Рассчитаем уравнение оптимальной разделяющей гиперплоскости с использованием следующих 5 основных индексов и обобщенных количественных показателей зообентоса: числа видов S в каждой пробе, суммарной численности N_s (экз./м^2) и биомассы B_s (г/м^2), информационного индекса Шеннона H и биотического индекса Вудивисса V. Область значений каждого из этих признаков разобьем на диапазоны (градации), причем границы диапазонов выберем оптимальным образом в соответствии с критерием равной заселенности диапазонов (см. раздел 6.3). Получим уравнение разделяющей гиперплоскости со свободным членом $\beta = 2.6$ и коэффициентами λ_j которые представлены в таблице 8.17.

Таблица 8.17

Коэффициенты λ_j модели распознавания состояния экосистемы по набору индексов и обобщенных количественных показателей зообентоса

№№ пп	Наименование показателей		Градации разбиения показателей j			
			1	**2**	**3**	**4**
1	Количество видов S	Диапазон	1 - 3	4-8	9-16	17-41
		λ_j	-69.53	-77.66	23.97	123.22
2	Численность N	Диапазон	1 - 400	401-1300	1301–3700	>3700
		λ_j	00.00	21.73	42.19	-63.92
3	Биомасса B	Диапазон	До 0.35	0.35-2.5	2.5–10	>10
		λ_j	15.06	6.68	-16.35	-5.38
4	Индекс Шеннона H	Диапазон	До 1.36	1.36-1.9	1.9-2.9	2.9-4.4
		λ_j	-72.35	17.26	26.05	29.03
5	Индекс Вудивисса V	Диапазон	0 - 1.2	равно 2	2.3-6	6.7-9
		λ_j	-22.15	-29.56	-30.62	82.33

Приведем примеры использования решающего правила.

На ст. 3 р. Байтуган в пробе от 17 июля 1991 г. было определено 20 видов бентоса (N = 3360 экз./м^2, B = 3.3 г/м^2, H = 3.29, V = 8). Используя коэффициенты табл. 8.17, имеем:

$$\eta = 2.6 + 123.22 + 42.19 - 16.35 + 29.03 + 82.33 = 263, \text{ т.е. } > 0 ,$$

что позволяет уверенно классифицировать экосистему этого участка реки как экологически стабильную (класс ОУС). Аналогично, на ст. 8 (в пределах с. Байтуган) той же реки в составе бентофауны зарегистрировано 6 видов (N = 490 экз./м^2, B = 1.06 г/м^2, H = 1.71, V = 2), что соответствует оценке

$$\eta = 2.6 - 77.66 + 21.73 + 6.68 + 17.26 - 29.56 = -58.9, \text{ т.е. } < 0 ,$$

дающей основание отнести станцию к зоне экологического кризиса (класс ЧЭС).

Сравнительный анализ коэффициентов уравнения разделяющей гиперплоскости по их абсолютной величине дает возможность утверждать, что основными показателями, определяющими качество экосистем, остается видовое разнообразие, выраженное через обилие (число видов S) и равномерность распределения (индекс H). В то же время, значения биомассы B вносят несущественный вклад в оценку качества экосистемы, а тенденция монотонного возрастания весовых коэффициентов λ_j по мере увеличения численности организмов при $N > 3700$ резко меняет свой знак на противоположный (т.е. в сторону экокризисных явлений), что является признаком серьезных структурных нарушений в водоеме. Имеет свои гносеологические корни и феномен явной нелинейности вклада обоих индексов – Шеннона H (значения до 1.36) и Вудивисса V (значения свыше 6).

Вариант № 2. Альтернативной моделью распознавания классов ОУС и ЧЭС является уравнение гиперплоскости, основанное на видовом составе зообентоса. При ее расчете используем следующие необязательные допущения:

- показатели обилия преобразуем в бинарные векторы, компоненты которых $z_{ij} = 1$, если j-й вид зафиксирован в i-м наблюдении, и $z_{ij} = 0$, в противном случае (напомним, что сам метод позволяет для каждого вида использовать до 9 градаций численности);

- список видов объектов обучающей выборки ограничим 166 видами семейства хирономид (отряд Diptera, сем. Chironomidae), в связи с отличными биоиндикационными свойствами последних и надежностью полученных нами натурных данных.

В ходе моделирования используем *метод последовательных включений с исключениями*, и в результате работы пошаговых процедур из всего списка хирономид отберем 68 видов, обладающих наилучшей классификационной значимостью (предполагается, что остальные 98 видов либо в равной мере присутствуют в водоемах обеих категорий, либо низкая частота их встречаемости не дает оснований для включения в решающее правило). В табл. 8.18 приведены верхняя и нижняя часть списка видов, отобранных в результате алгоритма селекции и отсортированных по убыванию значений коэффициентов λ, а позиции 20-49 с λ, близкими к 0, опущены.

Численные значения λ, по нашему мнению, имеют вполне обоснованный смысл индикаторных индексов "экологического благополучия" (при $\lambda>0$) или "экологического кризиса" (при $\lambda<0$). Например, наличие в составе донных сообществ видов *Cladotanytarsus mancus* (λ= +114), *Cricotopus bicinctus* (λ= +99) и других видов левой части табл. 8.18 свидетельствует о значительной вероятности отнесения водоема к "чистому" классу ОУС, а вида *Procladius choreus* (λ= -133.8) и других, представленных в табл. 8.18 справа, – к "грязному" классу ЧЭС.

С использованием уравнения обобщенного портрета для любого тестируемого наблюдения легко рассчитать его расстояние η от разделяющей гиперплоскости. Достаточно просто сложить рассчитанные коэффициенты для тех видов, которые встретились в пробе и добавить к ним свободный член уравнения. Если эта мера больше нуля, то это – класс ОУС, при отрицательной величине – класс ЧЭС (см. табл. 8.19.)

Таблица 8.19

Пример использования модели распознавания по видовому составу хирономид для двух произвольных станций наблюдения

№№ пп	Наименование видов	Коэффициенты обобщенного портрета λ,	
		Ст.6, р. Сок (15.07.1998)	Ст.5, р. Черновка (16.07.1987)
1	*Cladotanytarsus mancus*	114.88	114.88
2	*Harnischia fuscimana*	47.18	Отс.
3	*Tanypus punctipennis*	42.64	Отс.
4	*Chironomus obtusidens*	38.3	Отс.
5	*Microchironomus tener*	-2	Отс.
6	*Prodiamesa olivacea*	-5.36	-5.36
7	*Polypedilum scalaenum*	-16.56	Отс.
8	*Micropsectra* sp.	Отс.	-38.3
9	*Polypedilum nubeculosum*	Отс.	-38.3
10	*Cladopelma* gr. *lateralis*	Отс.	-38.3
11	*Cryptochironomus gr.defectus*	-106.42	Отс.
$\sum \sum \lambda_{ij} x_i^j$		112.66	-5.38
$\eta = \sum \sum \lambda_{ij} x_i^j - \beta$ (β = 47.04)		65.62	-52.42
Результат классификации		Относительно удовлетворительная ситуация (ОУС)	Чрезвычайная экологическая ситуация (ЧЭС)

Вариант № 3. Комбинированную модель, основанную на полном признаковом пространстве, получим с использованием всего набора переменных: из 5 индексов и обобщенных показателей обилия зообентоса и 166 признаков видовой структуры.

Сравнительная оценка достоверности полученных моделей-претендентов, представленная в табл. 8.20, осуществлялась по двум критериям:

- надежность распознавания на обучающей выборке;
- проверка на внешнем дополнении с использованием процедуры скользящего контроля.

Таблица 8.18

Коэффициенты уравнения разделяющей гиперплоскости для распознавания состояния экосистемы по видовому составу хирономид

NN пп	Наименования видов, характерных для класса 1 ("чисто")	Встречаемость в выборке	Коэффициенты λ модели	NN пп	Наименования видов, характерных для класса 0 ("грязно")	Встречаемость в выборке	Коэффициенты λ модели
1	*Microtendipes pedellus* (De Geer)	10	129.98	50	*Procladius* sp.	49	-5.36
2	*Cladotanytarsus mancus* (Walk.)	26	114.88	51	*Rheotanytarsus curtistylus* (Goetgh.)	3	-9.16
3	*Ablabesmyia monilis* (L.)	8	105.06	52	*Polypedilum scalaenum* Schrank	13	-16.56
4	*Paracladopelma camptolabis* K.	10	99.7	53	*Polypedilum* sp.	4	-22.14
5	*Cricotopus bicinctus* (Mg.)	35	99.32	54	*Chironomus plumosus* (L.).	1	-27.8
6	*Brillia* gr. *modesta* (Mg.)	5	99.18	55	*Chironomus muratensis* Ryser.	1	-31.9
7	*Corynoneura* sp.	4	97.76	56	*Parachironomus varus* Goetgh.	3	-33.74
8	*Clinotanypus nervosus* (Mg.)	6	93.06	57	*Dicrotendipes notatus* (Mg.)	5	-33.74
9	*Psectrocladius* gr. *sordidellus* (Zett.)	3	74.66	58	*Micropsectra* gr. *praecox* (Mg.)	8	-36.9
10	*Paratanytarsus confusus* Pal.	19	71.06	59	*Eukiefferiella* sp.	1	-37.14
11	*Cricotopus* sp.	2	65.42	60	*Polypedilum nubeculosum* (Mg.)	3	-38.3
12	*Prodiamesa olivacea* (Mg.)	2	58.76	61	*Cladopelma* gr. *lateralis* (G.)	5	-38.3
13	*Teleopelopia* sp.	5	51.58	62	*Micropsectra* sp.	1	-38.3
14	*Glyptotendipes gripekoveni K.*	5	47.18	63	*Stictochironomus crassiforceps* (K.)	5	-40.22
15	*Harnischia fuscimana* K.	8	47.18	64	*Tanytarsus pallidicornis* Walk.	6	-45.54
16	*Monodiamesa bathyphila* K.	6	43.78	65	*Eukiefferiella similis* Goetgh.	1	-65.7
17	*Tanypus punctipennis* (Mg.)	8	42.64	66	*Glyptotendipes glaucus* (Mg.)	3	-95.44
18	*Parametriocnemus* sp.	4	39.52	67	*Cryptochironomus* gr. *defectus* (K.)	29	-106.42
19	*Chironomus obtusidens* G.	7	38.3	68	*Procladius ferrugineus* (K.) *choreus*	1	-133.82

Таблица 8.20

Сравнительный анализ адекватности полученных моделей

Результат классификации	По обучающей выборке			При скользящем контроле		
	Число наблюдений	Число правильно распознанных, %	Число ошибочно распознанных, %)	Число наблюдений	Число правильно распознанных, %	Число ошибочно распознанных, %
1. Модель, основанная на структурных показателях зообентоса ***S, N, B, H, V***						
Класс ОУС	70	64(91.4)	6(8.6)	64	62(88.3)	2(11.7)
Класс ЧЭС	90	82(91.1)	8(8.9)	82	80(88.7)	2(11.3)
Всего	160	146(91.3)	14(8.8)	146	142(88.5)	4(11.5)
2. Модель, основанная на видовом составе хирономид						
Класс ОУС	68	68(100.0)	0(0.0)	68	51(75.0)	17(25.0)
Класс ЧЭС	89	89(100.0)	0(0.0)	89	77(86.5)	12(13.5)
Всего	157	157(100.0)	0(0.0)	157	128(81.5)	29(18.5)
3. Модель, учитывающая оба класса признаков						
Класс ОУС	70	70(100.0)	0(0.0)	70	59(84.3)	11(15.7)
Класс ЧЭС	90	90(100.0)	0(0.0)	90	80(88.9)	10(11.1)
Всего	160	160(100.0)	0(0.0)	160	139(86.9)	21(13.1)

По внутреннему критерию вне конкуренции оказались модели 2 и 3, включающие видовой состав и обеспечивающие почти 100% разделение гиперплоскостью классы "ЧЭС" и "ОУС". Однако экстраполяционные свойства этих моделей, оцененные по результатам скользящего контроля, несколько уступили по надежности модели, основанной на использовании обобщенных структурных показателей. Этот результат вполне соответствует сложившимся представлениям о недостаточной устойчивости моделей, построенных на основе обширных, разреженных матриц.

Метод обобщенного портрета дает надежность правильного распознавания экологической ситуации в пределах 85-90%. Ошибки классификации объясняются в основном вариабельностью проб под влиянием сезонной или многолетней динамики, а также определенной неоднозначностью критериев исходного деления на классы.

Таким образом, метод обобщенного портрета дает нам возможность:

- по произвольным гидробиологическим пробам проводить распознавание качества воды в водоеме;
- количественно ранжировать биоиндикационную ценность отдельных видов на множестве наблюдений, полученных непосредственно в изучаемом регионе;
- рассчитывать "модельно обоснованные" индексы экологического благополучия для отдельных точек наблюдения, а в комплексе – для пространственных гидроэкосистем

Действительно, для любого примера, к которому применимо найденное решающее правило, можно рассчитать его расстояние η от разделяющей плоскости (см. рис. 8.7), численно равное правой части уравнения (8.53) для решающего правила. Чем это расстояние больше, тем больше благополучие экосистемы (при $\eta > 0$) или глубже ее кризисность (при $\eta < 0$), что дает нам основания интерпретировать η как некоторую меру на шкале нормирования качества вод или очередной *индекс экологического благополучия* [Шитиков с соавт., 2001]. Нулевому значению η соответствует пограничное состояние, которое можно классифицировать как "Напряженное или критическое". Если оценить это расстояние η в многомерном пространстве видов от найденной гиперплоскости до каждой пробы, взятой на 23 станциях р. Чапаевка от истока до устья (см. рис. 8.7), получим сложную пространственную динамику благополучия экосистемы (с точки зрения развития зообентоса) с выделением трех зон:

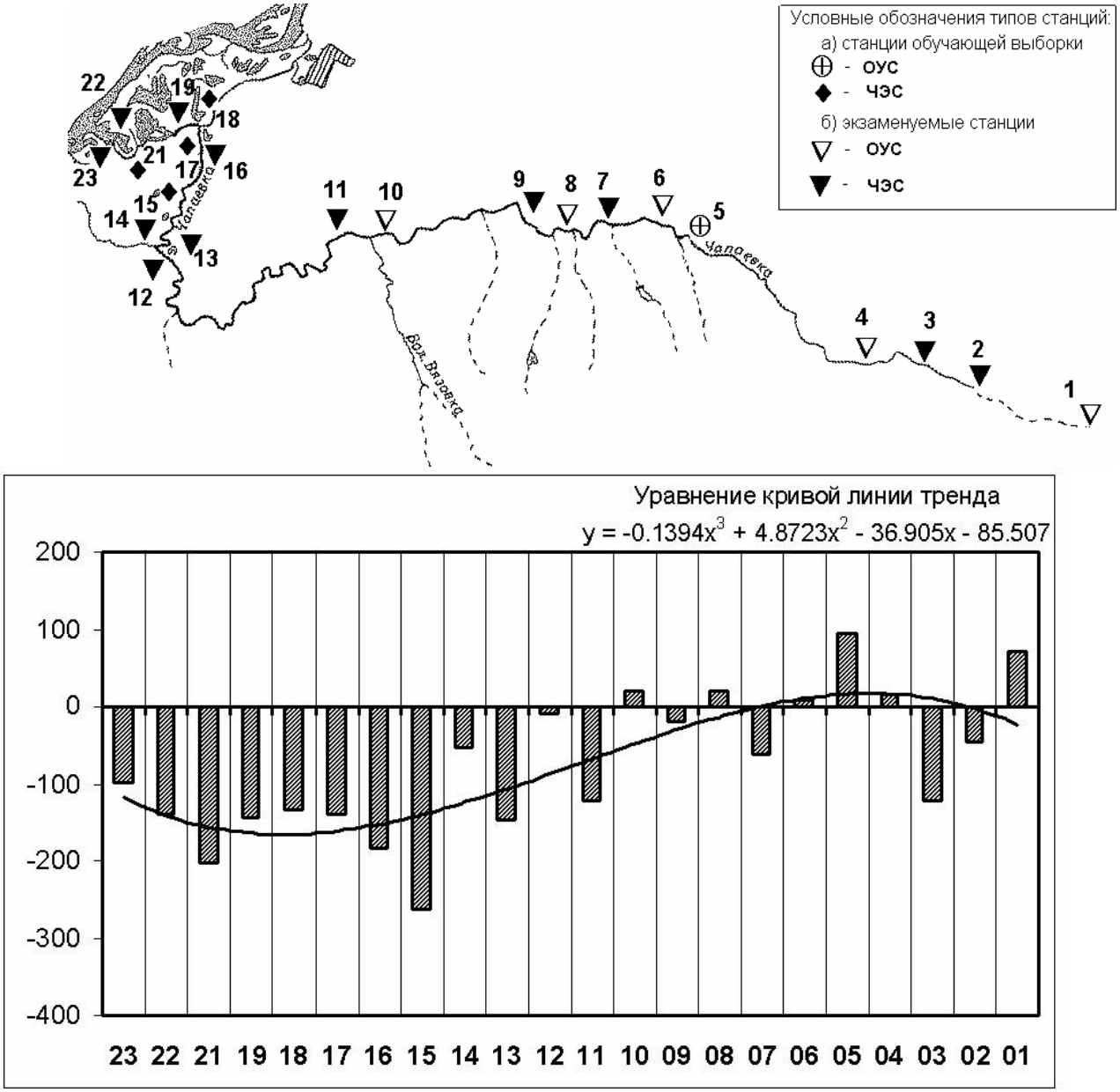


Рис. 8.7. Схема расположения станций наблюдений на р. Чапаевка и диаграмма распределения индекса экологического благополучия η по руслу реки
(на диаграмме по оси абсцисс – станции р. Чапаевка, по оси ординат – значения η)
Показана аппроксимирующая кривая тренда, соответствующая полиному третьего порядка.

- состояние в верхнем течении реки (ст.1–10), где при отсутствии промышленного загрязнения наблюдаются процессы евтрофирования под влиянием сельскохозяйственной нагрузки, может быть охарактеризовано как переходное от "стабильного" к "критическому";
- в районе г. Чапаевска (ст.13 и ниже) в зоне сброса промышленных сточных вод наблюдается состояние отчетливого экологического кризиса;
- в устье реки (ст. 23) кризисное состояние несколько стабилизируется за счет разбавления относительно чистыми водами Саратовского водохранилища.

Подробно результаты исследования состояние экосистемы р. Чапаевка в условиях антропогенного воздействия изложены нами ранее [Экологическое состояние.., 1997] и полученные выводы по представленным моделям распознавания вполне соответствуют данным комплексных многолетних наблюдений. Следует также отметить, что из 23 станций на рис. 8.7, только 5 были

использованы в обучающей выборке, что свидетельствует о вполне приемлемых экстраполяционных свойствах решающих правил, полученных по методу обобщенного портрета.

Как уже отмечалось выше, разделяющая гиперплоскость делит совокупность измерений только на 2 класса. Однако не представляет методологических трудностей осуществить пересчет диапазона значений расстояния η в любую из широкоупотребительных шкал оценки качества вод: 6-разрядную шкалу по Былинкиной и Драчеву, 9-разрядную систему по Оксиюк и Жукинскому или любую другую (см. раздел 3.4). Для этого достаточно составить репрезентативную выборку из примеров обучающей последовательности, каждому из которых будет поставлен в соответствие нормативный класс качества по выбранной шкале, и определить коэффициенты уравнения регрессии такого пересчета.

8.6. Задача об ассоциативности видов: алгоритм формирования логических высказываний

Формулировка задачи

Пусть пространство признаков $\boldsymbol{X}$ размерностью $\boldsymbol{m} > 1$, соответствующее некоторому списку видов гидробионтов, предварительно преобразовано к бинарному виду, т.е. $\boldsymbol{x_i} = 1$, если значение обилия $\boldsymbol{i}$-го вида в пробе превышает некоторый заданный порог, и $\boldsymbol{x_i} = 0$ в противном случае, $\boldsymbol{i} = 1,2,\ldots,\boldsymbol{m}$. Пусть также обучающая последовательность, в которую включаются специально отобранные измерения $\boldsymbol{x}$ из $\boldsymbol{X}$, распадается на два подмножества векторов – векторы первого класса $\boldsymbol{X}$ и векторы второго класса $\overline{\boldsymbol{X}}$.

Необходимо осуществить поиск по обучающей выборке непротиворечивых логических закономерностей и сформировать некоторую систему логических решающих правил, каждое из которых содержит информацию, заключенную не только в отдельных признаках, но и в различных сочетаниях значений признаков.

Биологический смысл формируемых логических конструкций заключается в попытке выделить в составе заданных биоценозов эволюционно сложившиеся совокупности взаимозависимых, контактирующих организмов, связанных общностью судьбы (консортивные группы, ассоциации, синузии и проч.).

Рекомендуемая литература: [Бонгард, 1967; Вайнцвайг, 1973; Голендер, Розенблит, 1978].

Математический лист

Рассмотрим логический метод распознавания, известный под названием алгоритм «Кора», который широко используется в геологоразведочных работах и скрининге лекарственных препаратов.

Задается множество характеристических логических функций $\Psi(\boldsymbol{x},\tau)$, которые называются логическими высказываниями и представляют собой некоторую комбинацию исходных переменных $\boldsymbol{x}_{ij}$, связанных между собой операцией коньюнкции, т.е. знаком логического произведения «И» (AND, $\wedge$): $\boldsymbol{x_{i1}} \wedge \boldsymbol{x_{j2}} \wedge \ldots \wedge \boldsymbol{x_{kl}}$, где $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k} \in \boldsymbol{M}$ – индексы исходного словаря признаков, $\boldsymbol{l} = 1 \div 5$ – количество элементов в логическом высказывании или его ранг. Каждый из сомножителей может трактоваться при этом на языке булевой алгебры как «Истина» (TRUE при $\boldsymbol{x_{ij}} = 1$) или ее отрицание, т.е. «Ложь» (NOT TRUE или FALSE при $\boldsymbol{x_{ij}} = 0$).

Алгоритм «Кора» многократно просматривает обучающую выборку, предварительно разделенную на два класса (0 и 1), и, с использованием операций алгебры логики, выделяет из множества высказываний так называемые непротиворечивые логические высказывания $\Psi(\boldsymbol{x},\boldsymbol{\tau}^*)$, покрывающие все множество примеров. Непротиворечивым высказыванием для каждого класса считается конъюнкция, которая встречается некоторое количество раз только в одном классе и ни разу не встречается в другом. При генерации логических высказываний алгоритм руководствуется рядом следующих правил:

1. Конъюнкции сортируются по продуктивности или мощности, оцениваемой числом наблюдений, для которых это высказывание справедливо. Чем больше продуктивность конъюнкции, тем выше прогностическая ценность выделенной комбинации признаков. В случае детерминист-

ской задачи распознавания в конечное решающее правило включаются конъюнкции, продуктивность которых превышает некоторый порог Δ.

2. Из генерируемого списка исключаются подчиненные (или дочерние) конъюнкции, полностью содержащие более короткие претенденты. После такой операции поглощения ликвидируется избыточность решающего правила, в котором остаются конъюнкции минимального ранга, содержащие выделенные закономерности в концентрированной форме.
3. Исключаются высказывания, которые по определенным критериям считаются "предрассудками". К ним относятся конъюнкции, не связанные с объективным правилом классификации, но в силу ограниченности выборки получившие хорошие оценки на обучении. Для выделения признаков, склонных к предрассудкам, выполняют bootstrep-процедуру, в которой обучающая выборка многократно случайным образом разбивается на классы. Разбитая таким образом выборка является тестом, позволяющим присвоить каждому исходному признаку штрафные баллы за склонность к предрассудкам. Конъюнкции, набравшие свернормативное количество штрафных очков, из решающего правила исключаются.

Описанием каждого класса является логическая сумма (дизъюнкция) некоторого количества непротиворечивых и продуктивных конъюнкций, прошедших описанные выше этапы отбора. Комбинация этих логических высказываний представляет собой своеобразную мозаично-фрагментарную разделяющую поверхность специального типа (в отличии от линейной поверхности "обобщенного портрета"). Существует возможность использовать сгенерированные конъюнкции для экзамена тестируемых примеров по принципу голосования (конъюнкции как бы используются в качестве "электората"). Однако, мы полагаем, что специальная ценность использования алгоритма «Кора» состоит в извлечении из моря исходных данных нетрадиционных и непротиворечивых гипотез об экологии видов – феноменах их взаимной обусловленности или конкурентности.

Алгоритм «Кора», как и другие логические методы распознавания образов, является достаточно трудоемким, поскольку при отборе конъюнкций необходим полный или частично направленный перебор. Поэтому при применении логических методов предъявляются высокие требования к эффективной организации вычислительного процесса, и эти методы хорошо работают при сравнительно небольших размерностях пространства признаков и емкости задаваемого класса характеристических функций.

Практический пример

Используем алгоритм «Кора» для логико-структурного анализа той же обучающей выборки, что и в алгоритме 2 раздела 8.5. Напомним, что алфавит признаков соответствовал 166 видам хирономид, класс $\boldsymbol{X}$ включал 70 проб на станциях с относительно удовлетворительной экологической ситуацией (ОУС), а класс $\overline{\boldsymbol{X}}$ – 90 измерений в зоне с чрезвычайной экологической ситуацией (ЧЭС).

Небольшая часть полученных высказываний приведена в табл. 8.21. Например, вид *Microtendipes pedellus* встретился в 9 наблюдениях класса ОУС и ни в одном наблюдении класса ЧЭС. Сочетание видов *Cladotanytarsus mancus* и *Cricotopus bicinctus* при непременном отсутствии *Cryptochironomus* gr. *defectus* достигли того же 8 раз. Логические высказывания класса ЧЭС несколько беднее и часто оказываются, хотя бы фрагментарно, зеркальным отражением коньюнкций класса ЧЭС.

Таблица 8.21

Фрагменты логических решающих правил, характеризующих категории качества вод по видовому составу хирономид
(затенены сомножители, являющиеся признаком отсутствия вида)

Класс ОУС		**Класс ЧЭС**	
Мощ-ность	Логическое высказывание	Мощ-ность	Логическое высказывание
5	***Brillia* gr. *modesta***	4	***Dicrotendipes notatus*** ∨ ***Harnischia fuscimana*** ∨ ***Tanytarsus pallidicornis***
6	***Clinotanypus nervosus***		
8	***Harnischia fuscimana***		
9	***Microtendipes pedellus***	4	***Ablabesmyia monilis*** ∧ ***Cladotanytarsus mancus*** ∧ ***Harnischia fuscimana*** ∧ ***Microchironomus tener***
9	***Cricotopus bicinctus*** ∧ ***Paratanytarsus confusus***		
8	***Cladotanytarsus mancus*** ∧ ***Cricotopus bicinctus*** ∧ ***Cryptochironomus* gr. *defectus***		
8	***Harnischia fuscimana*** ∧ ***Polypedilum nubeculosum*** ∧ ***Procladius ferrugineus***	5	***Cladotanytarsus mancus*** ∧ ***Endochironomus impar*** ∧ ***Paracladopelma camptolabis*** ∧ ***Polypedilum scalaenum***
9	***Microtendipes pedellus*** ∧ ***Micropsectra* gr.*praecox*** ∧ ***Dicrotendipes notatus***		

Глава 9. На пути к интеллектуальным биоиндикационным системам

9.1. Классификация наблюдений с использованием иерархических деревьев решений

Формулировка задачи

Пусть в таблице произвольных гидробиологических наблюдений $\boldsymbol{X}$ размерностью $\boldsymbol{m}$ >1 один из признаков, измеренный в порядковой шкале, определяет класс объекта и может принимать значения из некоторого фиксированного набора $\{y_1, y_2, \ldots, y_k, \ldots, y_p\}$. Необходимо на основе обучающей выборки сформировать дерево классификации (дерево решений), содержащее совокупность логических условий, позволяющих для произвольного измерения $\boldsymbol{x}$ из $\boldsymbol{X}$ указать класс качества y_k, к которому оно может принадлежать.

Еще в XVII столетии великий ученый Готфрид Лейбниц ("Новые опыты о человеческом разуме", 1704 г.) пытался раскрыть тайну "*Всеобщего Искусства Открытия*". Он утверждал, что одной из двух частей этого искусства является *комбинаторика* – перебор постепенно усложняющихся комбинаций исходных данных. Второй частью является *эвристика* – свойство догадки человека. На языке нашего времени эта часть соответствует модели мышления человека, включающей в себя процессы генерации эвристик (догадок, изобретений, открытий).

Универсальным методом поиска решений является *метод полного перебора* [1] и, обладай мы бесконечным запасом времени и ресурсов, то можно найти решение любой задачи. Здесь имеется в виду не конструирование нового знания, а, прежде всего, "выбор" наиболее правдоподобных вариантов. Можно отметить другой универсальный метод ускорения полного перебора — быстрое отсечение ложных (или вероятно ложных) ветвей перебора на основе использования алгебр логики.

Простейшие (одномерные) логические правила типа "если $\boldsymbol{A}$, то $\boldsymbol{B}$" мы рассматривали в разделе 6.4, когда описывали детерминационный анализ. Более широкие возможности предоставляют системы анализа на основе *деревьев решений* (Tree Analyzer), которые позволяют свести исходную матрицу данных $\boldsymbol{X}$ к набору простых правил, представленных в виде иерархической структуры – дерева. Этот метод моделирования сочетает мощный аналитический аппарат генерации решений с простотой использования технологии и интуитивно понятными конечными результатами.

Рекомендуемая литература: [Breiman et al., 1984; Коршунов, 1995; Loh, Shih, 1997; Деревья классификации.., URL]. Интересные методические материалы и Интернет-конференция по теме находятся также на сайте лаборатории BaseGroup Labs – http://www.basegroup.ru/labs.

Математический лист

Деревья решений – один из методов автоматического анализа данных, основные идеи которого восходят к работам П. Ховленда (P. Hoveland) и Е. Ханта (E. Hunt) конца 50-х годов XX в. Их итогом явилась основополагающая монография [Hunt et al., 1966], давшая импульс развитию этого направления.

Построение деревьев классификации – один из наиболее важных приемов, используемых при проведении "добычи данных и разведывательного анализа" (Data Mining), реализуемый как совокупность методов аналитической обработки больших массивов информации с целью выявить в них значимые закономерности и/или систематические связи между предикторными переменными, которые затем можно применить к новым совокупностям измерений.

Деревья решений представляют собой последовательные иерархические структуры, состоящие из узлов, которые содержат правила, т.е. логические конструкции вида "если … то …".

---

[1] Один из принципов системологии – принцип осуществимости моделирования сложных систем Б.С. Флейшмана [1982] делает невозможной процедуру полного перебора для большинства реальных экосистем. Но если понимать под «системой» совокупность только тех элементов и связей, которые необходимы для решения конкретной задачи (для достижения определенной цели), то процедура перебора может быть вполне реализуемой.

Конечными узлами дерева являются "листья", соответствующие найденным решениям и объединяющие некоторое количество объектов классифицируемой выборки. Это похоже на то, как положение листа на дереве можно задать, указав ведущую к нему последовательность ветвей, начиная от корня и кончая самой последней веточкой, на которой лист растет.

Есть целый ряд причин, делающих деревья классификации более гибким средством, чем традиционные методы анализа:

- схема *одномерного ветвления*, которая позволяет изучать эффект влияния отдельных предикторных переменных и проводить последовательный анализ их вклада;
- возможность одновременно работать с переменными различных типов, измеренных в непрерывных и порядковых шкалах, либо осуществлять любое монотонное преобразование признаков;
- отсутствие предварительных предположений о законах распределения данных.

Область применения деревья решений в настоящее время широка, но все задачи, решаемые этим методом, могут быть объединены в три следующие группы:

- *Описание данных*: деревья решений позволяют хранить информацию о данных в компактной форме, т.е. вместо обширных таблиц данных мы можем хранить дерево решений, которое содержит в концентрированной форме точное описание объектов;
- *Классификация*: деревья решений отлично справляются с задачами классификации, т.е. отнесения объектов к одному из заранее известных классов; при этом целевая переменная должна быть измерена в порядковой шкале;
- *Регрессия*: если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых (входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

На сегодняшний день существует значительное число алгоритмов, реализующих построение деревьев решений, из которых наибольшее распространение и популярность получили следующие:

- ***CART*** (Classification and Regression Tree), разработанный Л. Брейманом с соавторами [Breiman et al., 1984], представляет собой алгоритм построения бинарного дерева решений – дихотомической классификационной модели; каждый узел дерева при разбиении имеет только двух потомков; как видно из его названия, алгоритм решает задачи как классификации, так и регрессии;
- ***C4.5*** – алгоритм построения дерева решений с неограниченным количеством потомков у узла, разработанный Р. Куинленом [Quinlan, 1993]; не умеет работать с непрерывным целевым полем, поэтому решает только задачи классификации;
- ***QUEST*** (Quick, Unbiased, Efficient Statistical Trees) – программа, разработанная В. Ло и И. Ши [Loh, Shih, 1997], в которой используются улучшенные варианты метода рекурсивного квадратичного дискриминантного анализа, позволяющие реализовать *многомерное ветвление по линейным комбинациям порядковых предикторов*; содержит ряд новых средств для повышения надежности и эффективности индуцируемых деревьев классификации.

Основная идея построения деревьев решений из некоторого обучающего множества $\boldsymbol{X}$, сформулированная в интерпретации Р. Куинлена, состоит в следующем.

Пусть в некотором узле дерева сконцентрировано некоторое множество примеров $\boldsymbol{X^*}$, $\boldsymbol{X^* \subset X}$. Тогда существуют три возможные ситуации.

1. Множество $\boldsymbol{X^*}$ содержит один или более примеров, относящихся к одному классу $\boldsymbol{y_k}$. Тогда дерево решений для $\boldsymbol{X^*}$ – это "лист", определяющий класс $\boldsymbol{y_k}$.
2. ·Множество $\boldsymbol{X^*}$ не содержит ни одного примера, т.е. представляет пустое множество. Тогда это снова "лист", и класс, ассоциированный с "листом", выбирается из другого множества, отличного от $\boldsymbol{X^*}$ (скажем, из множества, ассоциированного с родителем).
3. Множество $\boldsymbol{X^*}$ содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество $\boldsymbol{X^*}$ на некоторые подмножества. Для этого выбирается один из признаков $\boldsymbol{j}$, имеющий два и более отличных друг от друга значений и $\boldsymbol{X^*}$ разбивается на новые подмножества, где каждое подмножество содержит все примеры, имеющие определенный диапазон значений выбранного признака. Это процедура будет рекурсивно продолжаться до тех пор, по-

ка любое подмножество $\mathbf{X^*}$ не будет состоять из примеров, относящихся к одному и тому же классу.

Описанная процедура построения дерева решений сверху вниз, называемая *схемой "разделения и захвата"* (divide and conquer), лежит в основе многих современных методов построения деревьев решений. Процесс обучения также называют *индуктивным обучением* или *индукцией деревьев* (tree induction).

При построении алгоритмов индукции деревьев решений особое внимание уделяется следующим ключевым вопросам:

- какой принять вид критерия для селекции признака, по которому пойдет разбиение;
- как выбрать момент, когда следует прекратить дальнейшие ветвления;
- каков механизм отсечения ветвей.

Правило разбиения: каким образом следует выбрать признак?

Для построения дерева с одномерным ветвлением, находясь на каждом внутреннем узле, необходимо найти такое условие проверки, связанное с одной из переменных $\boldsymbol{j}$, которое бы разбивало множество, ассоциированное с этим узлом на подмножества. Общее правило для выбора опорного признака можно сформулировать следующим образом: *«выбранный признак должен разбить множество* $\boldsymbol{X^*}$ *так, чтобы получаемые в итоге подмножества* $\boldsymbol{X^*_k}$, $\boldsymbol{k} = 1, 2, \ldots, \boldsymbol{p}$, *состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, т.е. количество чужеродных объектов из других классов в каждом из этих множеств было как можно меньше»*.

Были разработаны различные критерии, например, теоретико-информационный критерий, предложенный Р. Куинленом:

$$\boldsymbol{T}(\boldsymbol{j}) = \boldsymbol{H}(\boldsymbol{X}^*) - \sum_{\boldsymbol{k}=1}^{\boldsymbol{p}} \frac{|\boldsymbol{X}_{\boldsymbol{k}}^*|}{|\boldsymbol{X}^*|} \cdot \boldsymbol{H}(\boldsymbol{X}_{\boldsymbol{k}}^*) \Rightarrow \max \quad \forall \boldsymbol{j} = 1,2,\ldots,\boldsymbol{m} \ , \tag{9.1}$$

где $\boldsymbol{H}(\boldsymbol{X^*})$ и $\boldsymbol{H}(\boldsymbol{X^*_k})$ – энтропия подмножеств, разбитых на классы, рассчитанная по формуле Шеннона.

Алгоритм CART использует, так называемый, *индекс Джини* (в честь итальянского экономиста Corrado Gini), который оценивает "расстояние" между распределениями классов

$$\boldsymbol{G}(\boldsymbol{c}) = 1 - \sum_{\boldsymbol{j}=1}^{\boldsymbol{n}} \boldsymbol{p}_{\boldsymbol{j}} \ , \tag{9.2}$$

где $\boldsymbol{c}$ – текущий узел, а $\boldsymbol{p_j}$ – вероятность класса $\boldsymbol{j}$ в узле $\boldsymbol{c}$.

Большинство из известных алгоритмов являются "жадными алгоритмами": если один раз был выбран атрибут и по нему произведено разбиение на подмножества, то алгоритм не может вернуться назад и выбрать другой атрибут, который дал бы лучшее разбиение. И поэтому на этапе построения дерева нельзя сказать даст ли выбранный атрибут, в конечном итоге, оптимальное разбиение.

Правило остановки: разбивать дальше узел или отметить его как лист?

В дополнение к основному методу построения деревьев решений были предложены следующие правила:

- использование статистических методов для оценки целесообразности дальнейшего разбиения или так называемой "ранней остановки" (prepruning); в конечном счете, "ранняя остановка" процесса построения привлекательна в плане экономии времени обучения, но здесь уместно сделать одно важное предостережение: этот подход строит менее точные классификационные модели и поэтому ранняя остановка крайне нежелательна – признанные авторитеты в этой области Л. Брейман и Р. Куинлен советуют буквально следующее: «*Вместо остановки используйте отсечение*».
- ограничение глубины дерева; можно остановить дальнейшее построение, если разбиение ведет к дереву с глубиной превышающей заданное значение;
- контроль нетривиальности разбиения, т.е. получившиеся в результате узлы должны содержать количество примеров, не менее заданного порога.

Этот список эвристических правил можно продолжить, но на сегодняшний день не существует таких, которые имели бы глобальную практическую ценность. Многие из них применимы в каких-то частных случаях и, поэтому, к их использованию следует подходить с достаточной осторожностью.

Правило отсечения: каким образом ветви дерева должны отсекаться?

Очень часто алгоритмы построения деревьев решений дают сложные деревья, которые имеют много узлов и ветвей. Такие "ветвистые" деревья очень трудно понять, а ценность правила, справедливого скажем для 1-3 объектов, крайне низка и в целях анализа данных практически непригодно. Гораздо предпочтительнее иметь дерево, состоящее из малого количества узлов, не вполне идеально классифицирующее обучающую выборку, но обладающее способностью столь же хорошо прогнозировать результат для *тестовой выборки*.

К сожалению, достаточно непросто конкретно определить, что же такое дерево классификации "подходящего размера", кроме общего тезиса о том, что оно должно уметь использовать ту информацию, которая улучшает точность прогноза, и игнорировать ту информацию, которая прогноза не улучшает. Для решения вышеописанной проблемы часто применяется так называемое "отсечение ветвей" (pruning), которое происходит снизу вверх, двигаясь с листьев дерева, отмечая узлы как листья, либо заменяя их поддеревом. Если под точностью дерева решений понимается отношение правильно классифицированных объектов, то нужно отсечь или заменить поддеревом те ветви, которые не приведут к возрастанию ошибки.

Классификация новых примеров

После индукции дерева решений его можно использовать для распознавания класса нового объекта. Обход дерева решений начинается с корня дерева. На каждом внутреннем узле проверяется значение объекта $\boldsymbol{X_m}$ по атрибуту, который соответствует алгоритму проверки в данном узле, и, в зависимости от полученного ответа, находится соответствующее ветвление, и по этой дуге осуществляется движение к узлу, находящему на уровень ниже и т.д. Обход дерева заканчивается, как только встретится узел решения, который и дает название класса объекта $\boldsymbol{X_m}$.

Результаты расчетов

Сформируем обучающую выборку, состоящую из 117 наблюдений, признаками которой являются значения индексов Вудивисса, Шеннона, Пареле, Балушкиной, а также число видов в пробе, средние значения численности и биомассы, концентрация минерального фосфора и тип водоема (значения от 1 до 5 в зависимости от ширины русла и скорости течения). В качестве целевой переменной будем использовать класс качества вод от 2 до 6.

Получим два дерева решений: "компактное" дерево с использованием жестких процедур отсечения ветвей и упрощения правил и "полное" дерево, где единственным условием было концентрация в одном узле не менее 2 примеров обучающей выборки. "Компактное" дерево, представленное на рис. 9.1, состоит только из 8 узлов, в основе логических правил которых лежит всего 3 исходных признаков из 9.

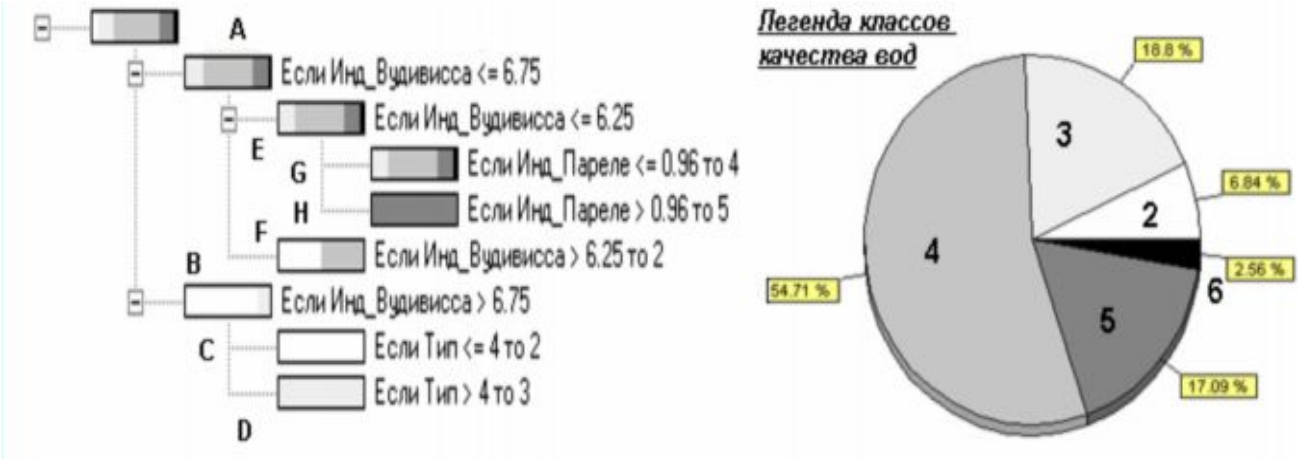


Рис. 9.1. Дерево решений, построенное для классификации качества вод с использованием процедур отсечения ветвей и упрощения.

В частности, двигаясь от корня, мы, вместе с 7 измерениями попадаем в узел **B**, если индекс Вудивисса больше 6.75, либо, в противном случае, с остальными измерениями поступаем в узел **A**. Из узла **B**, используя дополнительное условие по типу водоема, осуществляется переход на два листа **C** и **D**. Двигаясь же по основной ветви от узла **A** мы, в конце концов, приходим к узлу **G**, где сосредотачиваются все плохо распознаваемые примеры в количестве 63 измерений. Всем этим объектам присваивается класс 4, причем в 44 случаях это было выполнено ошибочно. Таким образом, исполнив свою объясняющую роль интерпретации существенных факторов, "компактное" дерево показало посредственные результаты в прогнозировании (38.4% ошибок).

Полное дерево, построенное без применения излишне жестких мер по обрезанию ветвей, использует все 9 предикторных переменных и гораздо сложнее в интерпретации, поскольку состоит из 54 узлов, из которых 28 являются "листьями" (см. табл. 9.1). Однако громоздкость сформированных правил компенсируется великолепными интерполяционными свойствами: на обучающей выборке зафиксировано только 9 ошибок классификации, что составляет 7.6%.

9.2. Генетический алгоритм селекции информативных переменных

Формулировка задачи

Пусть для решения произвольной задачи регрессии или классификации имеется множество измерений в пространстве варьируемых переменных размерностью $\boldsymbol{m} > 1$. Необходимо заданный набор признаков разбить на две категории: информативные переменные, существенные для решения поставленной задачи, и незначимые переменные, несущие мало дополнительной информации для нахождения искомой зависимости.

Важным условием применения любых статистических методов является объективно существующая связь между известными входными значениями и неизвестным откликом. Эта связь может носить случайный характер, искажена шумом, но она должна существовать. Известный афоризм «*garbage in, garbage out*» («*мусор на входе – мусор на выходе*») нигде не справедлив в такой степени, как при использовании методов самоорганизации и нейросетевого моделирования. Это объясняется, во-первых, тем, что итерационные алгоритмы направленного перебора комбинаций параметров нейросети оказываются весьма эффективными и очень быстрыми лишь при хорошем качестве исходных данных. Однако, если это условие не соблюдается, число итераций быстро растет и вычислительная сложность оказывается сопоставимой с экспоненциальной сложностью алгоритмов полного перебора возможных состояний. Во-вторых, сеть склонна обучаться, прежде всего, тому, чему проще всего обучиться, а, в условиях сильной неопределенности и зашумленности признаков, это – прежде всего артефакты и явления "ложной корреляции".

Отбор информативных переменных в традиционной регрессии и таксономии осуществляют путем "взвешивания" признаков с использованием различных статистических критериев. Так в главе 8 нами были описаны пошаговые процедуры, основанные, в той или иной форме, на анализе коэффициентов частных корреляций или ковариаций. Однако трудность проблемы формирования наиболее информативного подмножества признаков обусловлена тем, что после отбрасывания одного признака соотношение значимостей остальных анализируемых переменных в общем случае изменяется. Прямой путь решения этой задачи заключается в полном переборе всех $\boldsymbol{C}_{\mathbf{m}}^{\mathbf{m-p}}$ сочетаний переменных, что требует гигантского объема вычислений. Поэтому для этих целей используют различные секвенциальные (последовательные) процедуры, не всегда приводящие к результату, достаточно близкому к оптимальному. Элегантный автоматизированный подход к выбору значимых входных переменных может быть реализован с использованием *генетического алгоритма*, который можно считать "интеллектуальной" формой метода проб и ошибок.

Рекомендуемая литература: [Goldberg, 1989; Скурихин, 1995; Васильев, Ильясов, 1999].

Таблица 9.1

Дерево решений, построенное для классификации качества вод без использования процедур отсечения ветвей и упрощения

*(условные обозначения в шапке: **M** – количество примеров обучающей выборки, ассоциированных с узлом, **f** – число ошибок, **k** – лист класса качества вод)*

Правило узла	***M***	***f***	***k***
if Инд_Вудивисса <= 6.5 then	110		
| if Инд_Вудивисса <= 6.25 then	108		
| | if Инд_Вудивисса <= 0.75 then Класс 3	2	0	3
| | if Инд_Вудивисса > 0.75 then	106		
| | | if Инд_Пареле <= 0.96 then	105		
| | | | if P_мин <= 0.346 then	103		
| | | | | if Инд_Балушкиной <= 0.165 then Класс 4	2	1	4
| | | | | if Инд_Балушкиной > 0.165 then	101		
| | | | | | if Инд_Вудивисса <= 1.5 then	8		
| | | | | | | if P_мин <= 0.0905 then Класс 5	6	0	5
| | | | | | | if P_мин > 0.0905 then Класс 4	2	0	4
| | | | | | if Инд_Вудивисса > 1.5 then	93		
| | | | | | | if Инд_Шеннона <= 1.075 then Класс 3	2	0	3
| | | | | | | if Инд_Шеннона > 1.075 then	91		
| | | | | | | | if P_мин <= 0.0075 then Класс 4	5	1	4
| | | | | | | | if P_мин > 0.0075 then	86		
| | | | | | | | | if Инд_Шеннона <= 1.52 then	8		
| | | | | | | | | | if Тип <= 3.5 then Класс 2	2	1	2
| | | | | | | | | | if Тип > 3.5 then Класс 4	6	0	4
| | | | | | | | | if Инд_Шеннона > 1.52 then	78		
| | | | | | | | | | if Инд_Балушкиной <= 6.43 then	41		
| | | | | | | | | | | if N_видов <= 6.5 then	6		
| | | | | | | | | | | | if P_мин <= 0.1335 then Класс 3	4	0	3
| | | | | | | | | | | | if P_мин > 0.1335 then Класс 4	2	1	4
| | | | | | | | | | | if N_видов > 6.5 then	35		
| | | | | | | | | | | | if Инд_Вудивисса <= 4.75 then	27		
| | | | | | | | | | | | | if Ср_Биомасса <= 171.305 then Класс 4	23	0	4
| | | | | | | | | | | | | if Ср_Биомасса > 171.305 then	4		
| | | | | | | | | | | | | | if P_мин <= 0.0915 then Класс 4	2	0	4
| | | | | | | | | | | | | | if P_мин > 0.0915 then Класс 3	2	0	3
| | | | | | | | | | | | | if Инд_Вудивисса > 4.75 then	8		
| | | | | | | | | | | | | | if Ср_Численность <= 2590 then Класс 3	5	0	3
| | | | | | | | | | | | | | if Ср_Численность > 2590 then Класс 4	3	0	4
| | | | | | | | | | | | if Инд_Балушкиной > 6.43 then	37		
| | | | | | | | | | | | if Ср_Биомасса <= 225.245 then	34		
| | | | | | | | | | | | | if Тип <= 5 then	30		
| | | | | | | | | | | | | | if Инд_Пареле <= 0.625 then	19		
| | | | | | | | | | | | | | | if Ср_Численность <= 5380 then Класс 4	15	1	4
| | | | | | | | | | | | | | | if Ср_Численность > 5380 then	4	0	
| | | | | | | | | | | | | | | | if Тип <= 2 then Класс 4	2		4
| | | | | | | | | | | | | | | | if Тип > 2 then Класс 5	2	1	5
| | | | | | | | | | | | | | if Инд_Пареле > 0.625 then	11		
| | | | | | | | | | | | | | | if Тип <= 3.5 then Класс 5	8	0	5
| | | | | | | | | | | | | | | if Тип > 3.5 then Класс 4	3	0	4
| | | | | | | | | | | | | if Тип > 5 then	4		
| | | | | | | | | | | | | | if P_мин <= 0.074 then Класс 5	3	0	5
| | | | | | | | | | | | | | if P_мин > 0.074 then Класс 3	1	0	3
| | | | | | | | | | | | | if Ср_Биомасса > 225.245 then Класс 3	3	1	3
| | | | if P_мин > 0.346 then Класс 4	2	1	4
| | | if Инд_Пареле > 0.96 then Класс 5	1	0	5
| if Инд_Вудивисса > 6.25 then Класс 2	2	1	2
if Инд_Вудивисса > 6.5 then	7		
| if Тип <= 4 then Класс 2	6		2
| if Тип > 4 then Класс 3	1		3

Математический лист

Генетический алгоритм, позаимствованный у природных аналогов, является наиболее ярким представителем *эволюционных методов* (см. раздел 2.8) и представляет собой мощное поисковое средство, эффективное в различных проблемных областях.

Принципы эволюционной теории, заложенные Чарльзом Дарвиным в работе "Происхождение видов", сводятся к двум основным выводам:

- естественный отбор как движущая и направляющая сила эволюции, что предполагает некоторый механизм выделения самых сильных и полезных экземпляров (решений, структур, особей, алгоритмов);
- наличие некоторых степеней свободы эволюционного процесса в виде изменчивости объектов, т.е. наличие шага генерации новых структур искомых объектов (перечисление то же) в виде непрекращающейся последовательности "проб и ошибок".

Именно эти принципы отбора наилучших объектов являются ключевой эвристикой всех эволюционных математических методов, позволяющих зачастую уменьшить время поиска решения на несколько порядков по сравнению со случайным поиском. Механизм естественного отбора связывается здесь с некоторой функцией оптимальности $f(x)$, определяющей сравнительную ценность произвольного варианта, а изменчивость привносится путем специальных модификаций фрагментов бинарного кода.

Генетический алгоритм был разработан Дж. Холландом (J. Holland) в 1975 г. в Мичиганском университете. В дальнейшем Д. Голдберг (D. Goldberg) выдвинул ряд гипотез и теорий, помогающих глубже понять природу эволюции, а К. ДеДжонг (C. DeJong) предложил оптимальный вариант подбора параметров алгоритма для повышения общей эффективности работы.

Канонический генетический алгоритм характеризуется следующими особенностями.

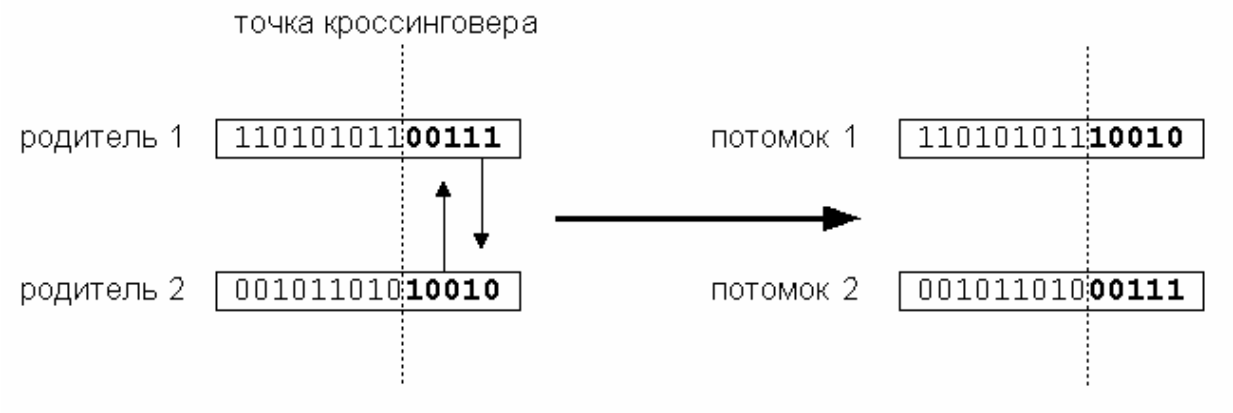
1. Задается функция оптимальности $f(x)$, определяющая эффективность каждой найденной комбинации признаков. Формируемое решение кодируется как вектор x, который называется "*хромосомой*" и соответствует битовой маске, т.е. двоичному представлению набора исходных переменных. В *хромосоме* выделяются части вектора – "*гены*", изменяющие свои значения в определенных позициях – "*аллелях*".
2. В соответствии с определенными ограничениями инициализируется исходная "*популяция*" $P^0(x_1^0..x_\lambda^0)$ потенциальных решений – совокупность решений на конкретной итерации, состоящая из некоторого количества *хромосом* λ, число которых задаётся изначально и в процессе перебора обычно не изменяется.
3. Каждая *хромосома* x_i, $i = 1,..., \lambda$ в *популяции* декодируется в форму, необходимую для последующей оценки, и ей присваивается значение эффективности $\mu(x_i)$ в соответствии с вычисленной функцией оптимальности. Кроме того, каждой *хромосоме* присваивается вероятность воспроизведения $P(x_i)$, $i = 1,..., \lambda$, которая зависит от эффективности данной *хромосомы*. Существуют различные схемы отбора, самая популярная из них – пропорциональный отбор:

$$p(x_i^t) = \frac{f(x_i^t)}{\sum_{j=1}^{\lambda} f(x_j^t)}. \qquad (9.3)$$

4. В соответствии с вероятностями воспроизведения $P(x_i)$ создается новая популяция *хромосом*, причем с большей вероятностью воспроизводятся наиболее эффективные элементы. *Хромосомы* производят потомков, используя операции рекомбинации: кроссинговера и мутации.
5. Процесс останавливается, если получено удовлетворительное решение, либо если исчерпано отведенное на эволюцию время. Если процесс не окончен, то вновь повторяются процессы оценки и воспроизведения новой *популяции*.

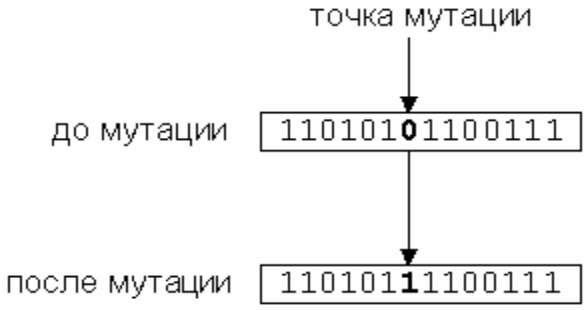
Операция воспроизведения на шаге 4 служит для создания следующей *популяции* на основе предыдущей при помощи операторов кроссинговера и мутации, которые имеют случайный характер. Каждой *хромосоме* промежуточной *популяции* X_i^t в случае необходимости подбирается партнёр и созданная *хромосома* помещается в *результирующую популяцию*.

Оператор кроссинговера производит скрещивание *хромосом* и обмен генетическим материалом между родителями для получения потомков. Этот оператор служит для исследования новых областей пространства и улучшения существующих (*эволюционное приспособление*). Простейший одноточечный кроссинговер производит обмен частями, на которые *хромосома* разбивается точкой кроссинговера, выбираемой случайно.



Двухточечный кроссинговер обменивает кусок строки, попавшей между двумя точками. Предельным случаем является равномерный кроссинговер, в результате которого все биты *хромосом* обмениваются с некоторой вероятностью.

Оператор мутации применяется к каждому биту *хромосомы* с небольшой вероятностью ($p_i \approx 0.001$), в результате чего бит (*аллель*) изменяет значение на противоположный.



Мутация нужна для расширения пространства поиска ("*эволюционное исследование*") и предотвращения невосстановимой потери бит в *аллелях*.

Существует также оператор воспроизведения, называемый «*инверсией*», который заключается в реверсировании *аллелей* между двумя случайными позициями, однако для большинства задач он не имеет практического смысла и поэтому мало эффективен.

Для исследования эффективности генетических алгоритмов используется понятие «*шаблона*». Шаблоны представляют собой гиперплоскости различной размерности в l-мерном пространстве и определяются с помощью элементов множества $\{0,1,*\}^l$, где l – длина *хромосомы* в битах, * – в данной позиции может быть любой бит. Генетический алгоритм обрабатывает шаблоны, и производит выборку значительного числа гиперплоскостей из областей с высокой приспособленностью, причем в течение одного поколения *популяции* оценивается $O(\lambda^3)$ структур. Это и отличает эволюционные процессы от различных эвристических и случайно-поисковых методов, в которых единственное решение развивается само по себе, а предыдущий опыт не используется.

Решение, получаемое при помощи генетического алгоритма, по своему характеру субоптимально, но это не мешает применять метод для поиска глобальных экстремумов в широком классе задач оптимизации многоэкстремальных функций.

Результаты расчетов

Используем для последующего анализа выборку из 520 измерений, где в качестве варьируемых переменных представлены общее число видов $X_1 = N$ и показатели обилия отдельных таксонов зообентоса (для хирономид – подсемейств и триб); $X_j = \ln((N_{sj}B_{sj})^{0.5})$, N_{sj} и B_{sj} – суммарные

по видам численность и биомасса j-й таксономической группы в пробе, j = 2, 3, …, 51. В предыдущей главе эта выборка уже использовалась нами при описании логистической регрессии и дискриминантного анализа (см. разделы 8.2 и 8.3). В качестве прогнозируемого отклика была взята категория качества воды в альтернативной шкале «Чисто» (2 и 3 класс) / «Грязно» (4 – 6 класс).

Функцию оптимальности получаемых решений $f(x)$ определим как статистику, связанную с качеством прогноза отклика на примерах обучающей выборки и полученную с помощью моделей вероятностных и обобщенно-регрессионных нейронных сетей. Сети этих типов выбраны потому, что для них общее время обучения и оценки относительно мало. Кроме того, эти сети очень сильно страдают от присутствия неинформативных входных переменных, а потому являются хорошим детекторами их обнаружения.

Выполним экзамен большого числа комбинаций входных переменных с помощью функции описанного типа. Каждый возможный вариант набора входных переменных можно представить в виде битовой маски по одному биту на каждую переменную (см. табл. 9.2). Ноль в соответствующей позиции означает, что эта исходная переменная не включена во входной набор, а единицы соответствуют базовому набору признаков. Начальная *популяция хромосом* на итерации 1 была сформирована нами случайным образом, хотя для этого возможно использование специальных простых правил.

Таблица 9.2

Результаты эволюционного процесса формирования информативного набора переменных с использованием генетического алгоритма
($\mu(x_i)$ – оценка эффективности хромосомы)

№ итераций	$\mu(x_i)$	Битовые маски хромосом
1	2.0001	10000011011001011010100111010010110100010001011110
2	2.0093	11101010010110010100000101100001110010110010011010
3	2.0049	10010101100100010011110111001111111110001001101011
4	2.0124	11100000110011000100011101010010110000010001011110
5	2.0154	11000111001101101101101111010010110101010001010110
6	2.0076	00011010010110010100010101100001110010111011000111
7	2.0078	00011010010111010100010101100001110010010010000111
…	…	…
97	2.0063	10010011101100010001101011110110111000001101100011
98	2.0070	10111001110101000100101001001001000101001101110010
99	2.0149	10111001101100000100101001001001000101101001000010
100	2.0042	01111000111100000110001101000010111101101001000010
Наилучшее найденное решение:		
	2.0450	11011010111111100010001001110000110101101110101010

Емкость *популяции* индивидуумов в нашем случае была установлена λ = 100, а процесс эволюции продолжался на протяжении 100 поколений (т.е. цикл отбор-порождение-оценка был повторен 100 раз и при этом построено и оценено 10000 версий нейронных сетей). Генетический алгоритм в поисках оптимального набора *генов* следил за *популяциями хромосом*, оценивая эффективность $\mu(x_i)$ каждой из них с помощью стандартного отклонения от регрессии при обучении обобщенно-регрессионной нейронной сети одинаковой конфигурации с коэффициентом сглаживания, равном 0.3. По значениям ошибки производился отбор лучших вариантов масок, которые комбинировали друг с другом с помощью искусственных генетических операций скрещивания и мутаций, интенсивностью которых можно было управлять (скорость мутации была выбрана 1, а скорость скрещивания – 0.3).

В соответствии с найденным наилучшим решением, все подмножеств видов зообентоса можно разбить на две примерно равные категории: 27 групп информативных индикаторов качества воды (первые три уровня в табл. 9.3) и 22 группы организмов, малозначимых для нейросетевого моделирования (в той же таблице внизу). Следует еще раз подчеркнуть, что отнесение какого-либо класса или семейства к малоинформативной категории вовсе не означает отсутствие фактической

биологической зависимости обилия организмов этого таксона от уровня загрязнения. Это может быть, например, объяснено его сильной корреляцией с какой-либо другой значащей группой, вследствие чего пришлось пожертвовать одним из таксонов, объявив его малоценным для статистической обработки.

Иногда бывает полезно уменьшить размерность задачи даже ценой некоторой потери точности, поскольку это повышает иллюстративность и улучшает способность нейросетевой модели к обобщению. Можно создать дополнительный стимул к исключению лишних переменных, назначив специальный штраф за элемент. Это число будет умножаться на количество элементов и результат будет прибавляться к уровню ошибки при оценке качества сети. Назначим штраф за элемент, равный 0.002, и повторим эволюционный процесс. В этих условиях из 50 исходных переменных будет отобрано в качестве информативных только четыре признака, представленных первыми двумя уровнями в табл. 9.3: обилие семейств Ephemeroptera, Oligochaeta, а также подсемейства Orthocladiinae и трибы Tanytarsini семейства Chironomidae. Если же увеличить штраф за элемент до 0.005, то выбирается только один значимый признак – обилие организмов семейства Ephemeroptera. Таким образом, варьируя параметрами генетического алгоритма можно эшелонировать весь список переменных по уровню их связи с целевой функцией, что и нашло свое отражение в табл. 9.3.

Таблица 9.3

Результаты селекции информативного набора переменных с использованием генетического алгоритма

Уровень информативности	Таксономические группы зообентоса
Наилучшие биоиндикаторы	Ephemeroptera
Хорошие биоиндикаторы	Oligochaeta, Chironomidae (подсемейство Orthocladiinae и триба Tanytarsini)
Информативные группы	Amphipoda, Bivalvia, Chaoboridae, Ceratopogonidae, Coleoptera, Dermaptera, Dreissenidae, Gastropoda, Hidracarina, Limoniidae, Megaloptera, Nematoda, Odonata, Plecoptera, Polychaeta, Psychodidae, Ptychopteridae, Simuliidae, Tabanidae, Trichoptera, Chironomidae (подсемейства Diamesinae, Prodiamesinae, Tanypodinae и триба Chironomini)
Группы, незначимые для прогноза качества воды	Arachnoidae, Collembola, Crustacea, Culicidae, Cylindrotomidae, Diptera, Dolichopodidae, Diplura, Dixidae, Ephydridae, Hemiptera, Hirudinea, Homoptera, Hydrida, Muscidae, Nematocera, Rhagionidae, Stratiomyidae, Tipulidae, Unionidae

9.3. Многорядный алгоритм МГУА для оценки качества вод

Формулировка задачи

Предположим, что имеется набор исходных данных в виде матрицы $\boldsymbol{X}$ из $\boldsymbol{n}$ наблюдений в пространстве варьируемых переменных размерностью $\boldsymbol{m} > 1$, характерный для стандартной задачи множественной регрессии. Пусть сформирована обучающая последовательность примеров, в которой каждой строке матрицы $\boldsymbol{X}$ поставлено в соответствие известное значение отклика $\boldsymbol{Y}$, измеренное в количественной шкале.

Необходимо, используя методы самоорганизации, получить модель, выражающую закон изменения отклика $\boldsymbol{Y}$ в зависимости от конкретных значений независимых переменных $\boldsymbol{X}$.

Рекомендуемая литература: [Ивахненко, 1969, 1982; Ивахненко, Лапа, 1971; Ивахненко с соавт., 1976; Брусиловский, 1987; Ивахненко, Юрачковский, 1987; Розенберг с соавт., 1994].

Математический лист

Ранее, в разделе 2.8 были рассмотрены основные принципы самоорганизации моделей, лежащие в основе такого направления в математическом анализе данных как *метод группового учёта аргументов* – МГУА (Group Method of Data Handling, GMDH). Модели самоорганизации МГУА можно рассматривать как своеобразное связующее звено, объединяющее различные методологические концепции, представленные, в том числе, и разделами настоящей книги.

С одной стороны, МГУА считается, своего рода, интеллектуальным обобщением регрессионного анализа, понимаемого в наиболее широком смысле. От классической множественной регрессии МГУА отличается лишь использованием специфических квадратичных критериев внешнего или внутреннего типа, а также многорядными итерационными процедурами нахождения оптимального решения задачи.

С точки зрения организации вычислений метод группового учета аргументов можно представить как следующий итеративный цикл:

- задается некоторое множество $\mathfrak{R}$ достаточно простых функций от исходных аргументов, которые называются *предикторами* или *частными структурами* модели, и формируется первый слой модели;
- из частных структур текущего слоя генерируется по определенным правилам новый слой предикторов, которые теперь сами становятся последним слоем;
- из частных описаний последнего слоя отбираются L лучших, где L – ширина отбора (*селекции*);
- если не выполняется условие прекращения селекции (например, продолжает возрастать критерий качества модели), осуществляется генерация нового слоя;
- самый лучший набор частных структур последнего слоя объявляется искомым оптимальным решением задачи.

В этом описании налицо все признаки эволюционного алгоритма – отбор (селекция) и генерация нового поколения.

Наконец, форма многорядного представления моделей МГУА, где в каждом слое локализуются достаточно простые функции (полиномы не более 2 порядка от двух переменных), но общая целостная модель представляет чрезвычайно сложную конструкцию, содержит много общего с описываемыми далее моделями искусственных нейронных сетей.

В рекомендованной литературе представлены различные схемы процесса самоорганизации при синтезе моделей МГУА: комбинаторные, многорядные, гибридизации, основанные на конечных стохастических автоматах и т.д.

Остановимся на общей схеме *многорядного алгоритма МГУА*, которая воспроизводит схему массовой селекции, аналогичную задаче нахождения оптимальной структуры перцептрона. В многорядной полиномиальной модели "полное" описание (т.е. регрессионная модель от m факторов)

$$y = F(x_1, x_2, ..., x_m) \tag{9.4}$$

заменяется последовательностью рядов "частных" описаний:

- первый ряд селекции – $y_1 = f(x_1, x_2)$, $y_2 = f(x_1, x_3)$, …, $y_s = f(x_{m-1}, x_m)$, где $s = C_m^2$;
- второй ряд селекции – $z_1 = f(y_1, y_2)$, $z_2 = f(y_1, y_3)$, …, $z_p = f(y_{s-1}, y_s)$, где $p = C_s^2$ и

т.д.

Общая результирующая сложность модели (9. 4) зависит, таким образом, от двух факторов – вида частного описания f и количества рядов селекции.

Каждое частное описание является функцией только двух переменных. Поэтому коэффициенты такого регрессионного уравнения могут быть легко определены даже по небольшому числу наблюдений обучающей последовательности методом наименьших квадратов. Различные модификации многорядного алгоритма отличаются друг от друга по виду опорной функции f. В алгоритме с линейными полиномами используются частные описания вида

$$Y_k = a_0 + a_1 \cdot x_i + a_2 \cdot x_j, \quad 0 < i < m, \quad 0 < j < m. \tag{9.5}$$

Усложнение модели в этом случае происходит только за счет увеличения числа учитываемых аргументов: на первом ряду селекции синтезируются модели, содержащие по 2 аргумента, на втором - по 3 или 4, на третьем - до 8 аргументов и т.д.

Многорядные алгоритмы при использовании нелинейных опорных функций, например:

$$Y_k = a_0 + a_1 \cdot x_i + a_2 \cdot x_j + a_3 \cdot x_i \cdot x_j ;$$

$$Y_k = a_0 + a_1 \cdot x_i + a_2 \cdot x_j + a_3 \cdot x_i \cdot x_j + a_4 \cdot x_i^2 + a_5 \cdot x_j^2 ; \qquad (9.6)$$

позволяют получить модели практически любой сложности, так как на каждом ряду селекции степень полинома удваивается. При этом число коэффициентов модели может исчисляться уже миллионами, хотя минимум критерия селекции обычно достигается достаточно быстро.

Чтобы обеспечить несмещенность получаемого решения, исходную выборку предварительно разделяют случайным образом на две статистически однородные части: *обучающую* и *проверочную (контрольную) последовательности*. Для этого все имеющиеся экспериментальные точки ранжируются, т.е. располагаются в ряд по величине дисперсии

$$D^2 = \frac{1}{n} \sum_{i=1}^{n} [(y_i - \bar{y}) / \bar{y}]^2 \quad , \qquad (9.7)$$

где $\bar{y}$ – среднее значение отклика, и делятся на две части. Точки с четными номерами образуют первую последовательность, а точки с нечетными номерами – вторую последовательность.

Обучающая последовательность используется для нахождения обычным методом наименьших квадратов коэффициентов a_0 - a_5 частных описаний (9.5)-(9.6), связывающих отклик Y с любыми двумя аргументами – исходными признаками, либо выходными переменными частных описаний предыдущего ряда. Проверочная последовательность, которая в этих расчетах участия не принимает, служит в качестве модельно-независимого порогового фильтра селекции, играющего роль внешнего дополнения к обучающей выборке.

Из одного ряда селекции в другой на каждом шаге самоорганизации пропускаются не все частные описания, полученные путем полного перебора пар факторов (s, p и т.д.), а только небольшая их часть, например, m уравнений, которые являются "наилучшими" в смысле заданного критерия регулярности, определяемого по частным описаниям на проверочной последовательности. В качестве конкретных математических выражений, используемых для регуляризации, обычно используют одну из следующих статистик:

- *абсолютной среднеквадратичной ошибки*
$$\delta^2 = \frac{1}{n_{пр}} \sum_{i=1}^{n_{пр}} (y_i - y_i^*)^2 \qquad (9.8)$$

- *относительной среднеквадратической ошибки*
$$\Delta^2 = \frac{n_{пр} \delta^2}{\sum_{i=1}^{n_{пр}} y_i^2} \qquad (9.9)$$

- *коэффициента корреляции* выходной переменной Y с аргументом x_k
$$K_{yx_k} = \frac{\sum_{i=1}^{n_{пр}} y_i x_{ik}}{\sqrt{\sum_i y_i^2 \sum_i x_{ik}^2}}, \qquad (9.10)$$

где $n_{пр}$ – количество точек проверочной выборки, Y и Y^* – фактическое и расчетные значения прогнозируемой переменной.

Поскольку при использовании нелинейных опорных функций отмечается опасность потери существенного аргумента, то предпочтительнее использовать алгоритмы, оптимизирующие на каждом шагу длину частного описания (например, выбирающие вид частного описания с максимумом коэффициента корреляции на проверочной последовательности [Справочник по типовым.., 1980]).

Количество рядов селекции обычно рекомендуется наращивать до $s = (m - 1)$, хотя в литературе описан случай, когда самая несмещенная линейная модель в примере с 5 аргументами получилась на 30-м ряду селекции. На практике усложнение модели прекращают, когда дальнейшее улучшение критерия селекции не будет превышать некоторого числа ε (параметр алгоритма).

Результаты расчетов

Рассмотрим использование многорядного алгоритма МГУА на примере анализа связи между гидрохимическими и гидробиологическими показателями. Сформируем исходный набор признаков из следующих 7 показателей: X_H – информационного индекса Шеннона, X_V – биотического индекса Вудивисса, X_P – олигохетного индекса Пареле, X_{CI} – хирономидного индекса Балушкиной, числа видов X_S, логарифмов суммарной численности X_N и биомассы X_B зообентоса в пробе. Как и в разделе 8.1 будем искать зависимость этих показателей от концентраций различных химических ингредиентов: аммонийного азота, минерального фосфора, ионов железа и БПК.

Выполним предварительное нормирование переменных от 0 до 1 по вариационному размаху и разобьем исходные выборки на обучающую и проверочную в примерном соотношении 2.5:1. Используем многорядный алгоритм МГУА, ограничившись при этом линейным частным описанием (9.5). Наращивание рядов селекции будем прекращать, если на очередной итерации прирост максимальной величины коэффициента корреляции (9.10) оказывался по абсолютной величине меньше ε=0.0001.

Модели, полученные для каждого гидрохимического показателя и представленные в табл. 9.4, оказались достаточно лаконичными – количество рядов селекции не превысило 3, что обычно характерно для простых, умеренно зашумленных зависимостей. На каждом шаге итерации, в том числе, на завершающем, было отобрано по 7 возможных моделей-претендентов. Структурные матрицы в нижней части таблицы показывают, из каких конкретно исходных переменных состоят те или иные модели. Нетрудно сделать вывод, что в результате селекции отбирались для включения в частные описания три основных индекса – Шеннона, Вудивисса и Пареле. Остальные переменные попадали в модели эпизодически.

Наилучшая модель № 2 для прогноза концентрации аммонийного азота, оцененная по максимуму коэффициента корреляции $K_{кор}$ на проверочной последовательности, была получена на 3-м ряду селекции и основывалась на 3 исходных аргументах из 7.

Оптимальная модель (M_2) имела вид:

$$Y_{NH4} = -0.0489 + 0.939 \cdot U_2 + 0.794 \cdot U_3,$$

где промежуточные переменные U_2 и U_3 могут быть вычислены по частным описаниям 2-го ряда селекции:

$$U_2 = -0.0998 + 0.797 \cdot Z_2 + 0.843 \cdot Z_5;$$
$$U_3 = -0.0173 + 0.345 \cdot Z_3 + 0.766 \cdot Z_4.$$

Таблица 9.4

Основные характеристики многорядных моделей МГУА, полученных для прогнозирования гидрохимических показателей ($n_{обуч}$ и $n_{пров}$ – размерность обучающей и проверочной последовательностей, критерий регулярности по формуле (9.10))

Прогнозируемая переменная	Аммонийный азот							Минеральный фосфор							Ион железа							$БПК_5$						
$n_{обуч}$ / $n_{пров}$	53 / 33							79 / 38							87 / 42							62 / 25						
Рядов селекции	3							3							3							2						
№ лучшей модели	2							3							3							2						
$K_{кор\ max}$	0.718							0.766							0.805							0.845						
Факторы/№модели	1	**2**	3	4	5	6	7	1	2	**3**	4	5	6	7	1	2	**3**	4	5	6	7	1	**2**	3	4	5	6	7
Индекс Шеннона	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Индекс Вудивисса	*	*	*	*	*	*	*		*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*
Индекс Пареле	*	*	*	*	*	*	*		*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*
Индекс Балушкиной		*		*	*	*	*				*	*	*	*			*	*	*	*	*		*	*	*	*	*	*
Число видов		*	*	*	*	*	*	*	*	*	*	*	*	*				*	*	*	*				*	*	*	*
Биомасса						*	*						*	*						*	*						*	*
Численность			*	*	*	*	*		*			*	*	*	*	*	*	*	*	*	*			*		*	*	*

В свою очередь, промежуточные переменные Z_2, Z_3, Z_4 и Z_5 вычисляются на первом ряде селекции уже с использованием нормированных исходных переменных:

$$Z_2 = 0.1983 - 0.0138 \cdot X_V + 0.00073 \cdot X_S;$$
$$Z_3 = 0.1868 + 0.0522 \cdot X_P - 0.0117 \cdot X_V;$$
$$Z_4 = 0.1687 + 0.0071 \cdot X_{CI} - 0.0113 \cdot X_V;$$
$$Z_5 = 0.2222 + 0.0059 \cdot X_S - 0.0571 \cdot X_H;$$

где X_H – информационный индекса Шеннона, X_V – биотический индекс Вудивисса, X_P – олигохетный индекс Пареле, X_{CI} – хирономидный индекс Балушкиной, X_S – число видов в гидробиологической пробе.

Необходимо отметить, что отдельные модели последнего ряда селекции весьма незначительно отличаются между собой по критериям качества: коэффициент корреляции колеблется от 0.7117 до 0.717, а стандартное отклонение на проверочной выборке – от 0.247 до 0.2463. Можно также напомнить, что в аналогичном примере множественной регрессии (см. раздел 8.1), версия этой же модели вообще не содержала индекса Вудивисса, столь популярного в оптимальной модели МГУА M_2. Эти примеры, иллюстрирующие, как примерно одного и того же результата можно достичь совершенно разными способами, служат еще одним убедительным доказательством *принципа множественности моделей В.В. Налимова.*

9.4. Нейросетевое моделирование: многослойный персептрон

Формулировка задачи

Пусть в таблице произвольных гидробиологических наблюдений X размерностью $m > 1$ один из признаков, измеренный в порядковой шкале, определяет класс объекта и может принимать значения из некоторого фиксированного набора $\{y_1, y_2, \ldots, y_k, \ldots, y_p\}$.

Необходимо, используя таблицу X, как обучающую выборку, решить общую задачу распознавания образов, т.е. разработать модель предсказания для произвольного объекта значения целевого признака, выраженного в порядковой шкале или шкале наименований.

В предыдущих главах мы неоднократно обращались к поставленной задаче, решая ее методами регрессионного и дискриминантного анализа (разделы 8.2 - 8.3), с использованием эвристических и линейных алгоритмов распознавания (разделы 8.4 - 8.5), а также на основе иерархических деревьев решений (раздел 9.1). При этом неоднократно отмечалось, что традиционная параметрическая статистика требует для адекватного прогноза либо очень большого объема известных данных, либо очень сильных предположений о виде функций распределения. Поэтому естественным является желание иметь единую парадигму построения различных эмпирических моделей, решающих задачи предсказания и классификации, которая бы удовлетворяла следующим условиям:

- достаточно высокая вычислительная эффективность;
- отсутствие определяющих ограничений на функцию распределения данных;
- обеспечение возможности обработки разнотипных экспериментальных данных (без сведения всех признаков к одной шкале) и инвариантность к допустимым преобразованиям шкал признаков;
- простота получения результата и отсутствие привязки к конкретной проблемной области;
- работа в случае, если число измеренных признаков превышает число объектов, и число объектов достаточно мало;
- работа при наличии пропусков в таблице.

Как отмечалось в разделе 2.8, большинству перечисленных требований удовлетворяет подход, основанный на применении искусственных нейронных сетей (ИНС).

Рекомендуемая литература: [Горбань, 1990, 1998а,б; Уоссермен, 1992; Горбань, Россиев, 1996; Васильев с соавт., 1997; Царегородцев, Погребная, 1998; Дьяконов, Круглов, 2001; Нейронные сети.., 2001].

Математический лист

Структура искусственного нейрона

Многие современные НС сконструированы из *формальных нейронов*, отдалённо напоминающих свой биологический прототип. Структура нейрона имеет вид, представленный на рис. 9.2, при следующих обозначениях:

- $x_1,...,x_n$ – значения, поступающие на входы (синапсы) нейрона;
- $w_1,..,w_n$ – ***веса*** синапсов, которые могут быть как тормозящими, так и усиливающими;
- S – взвешенная сумма входных сигналов:

$$S = \sum_{i=1}^{n} w_i x_i - T \; ; \tag{9.11}$$

- T – ***порог*** нейрона (во многих моделях обходятся без него);
- F – функция *активации* нейрона, преобразующая взвешенную сумму в выходной сигнал

$$y = F(S) \; . \tag{9.12}$$

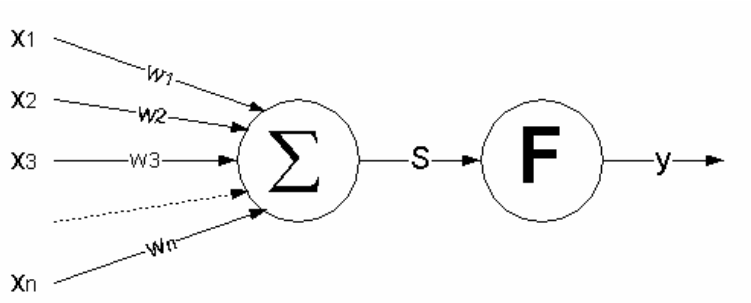


Рис. 9.2 Структура искусственного нейрона

Вид функции активации может иметь различное математическое выражение, выбор которого определяется характером решаемых задач. Например, функция активации может быть:

- *линейная*

$$y = k \cdot S \; ; \tag{9.13}$$

- *пороговая бинарная*

$$y = \operatorname{sign}(S) = \begin{cases} 1, S > 0; \\ 0, S \le 0. \end{cases} \quad \text{или биполярная} \quad y = \begin{cases} 1, S > 0; \\ -1, S \le 0. \end{cases} ; \tag{9.14}$$

- *линейная ограниченная*

$$y = \begin{cases} p, S > \alpha; \\ -p, S < -\alpha; \\ S, -\alpha \le S \le \alpha. \end{cases} ; \tag{9.15}$$

- *сигмоидная*

$$y = \frac{1}{1 + e^{-CS}} \; , \tag{9.16}$$

где $C > 0$ – коэффициент ширины сигмоиды по оси абсцисс;

- функция *гиперболического тангенса*

$$y = th(cS) = \frac{e^{cS} - e^{-cS}}{e^{cS} + e^{-cS}} \; ; \tag{9.17}$$

- *логарифмическая*

$$y = \ln(S + \sqrt{S^2 + 1}) \tag{9.18}$$

(данная функция характерна тем, что имеет диапазон $[-\infty;+\infty]$ с точкой перегиба в начале координат, усиливает очень слабые сигналы и ослабляет очень сильные);

- *радиально-базисная*, имеющая вид *функции Гаусса*

$$y = e^{-\frac{S^2}{2\sigma^2}}, \qquad (9.19)$$

где σ – среднеквадратичное отклонение нормального распределения, характеризующее ширину функции, S в этом случае определяется как расстояние между входным и весовым вектором: $S^2 = \sum_i (x_i - w_i)^2$.

Вид некоторых функций активации представлен на рис. 9.3.

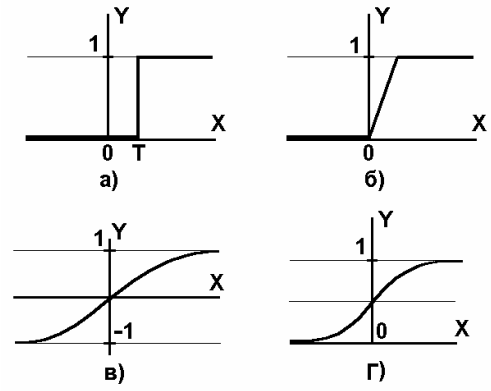


Рис. 9.3 Вид функций активации:
а) функция единичного скачка;
б) линейный порог (гистерезис);
в) гиперболический тангенс;
г) сигмоид

Одной из наиболее распространенных функций активации является нелинейная функция с насыщением - так называемая логистическая функция или сигмоид (9.16). При уменьшении коэффициента C сигмоид становится более пологим, вырождаясь в пределе при $C = 0$ в горизонтальную линию на уровне 0.5. При увеличении C сигмоид приближается по внешнему виду к функции единичного скачка с порогом T в точке $x = 0$.

Из выражения для сигмоида очевидно, что выходное значение нейрона лежит в диапазоне [0,1].

Популярность сигмоидной функции определяют следующие ее ценные свойства:

- способность усиливать слабые сигналы лучше, чем большие, и сопротивляться "насыщению" от мощных воздействий;
- монотонность и дифференцируемость на всей оси абсцисс;
- простое выражение для ее производной:

$$f'(x) = C \cdot f(x) \cdot (1 - f(x)), \qquad (9.20)$$

что дает возможность использовать широкий спектр оптимизационных алгоритмов.

Следует также отметить существование других типов формальных нейронов (сигма-пи нейроны и стохастические нейроны, нейроны с поправочным блоком), учитывающих другие свойства биологического прототипа. Наиболее интересны голографические нейроны, используемые в голографической парадигме ИНС.

Структура многослойного персептрона

Теоретическое обоснование нейросетевого моделирования базируется на теореме А.Н.Колмогорова, доказавшего, что любую непрерывную многомерную функцию на единичном отрезке $[0;1]$ можно представить в виде конечного числа одномерных [Головко, 1999]:

$$f(x_1, x_2, ..., x_n) = \sum_{p=1}^{2n+1} g\left(\sum_{i=1}^{n} \lambda_i \varphi_p(x_i)\right), \qquad (9.21)$$

где g и φ_p непрерывные и одномерные функции, $\lambda_i = \text{const}$. Отсюда с помощью любой многослойной ИНС всего с двумя перерабатывающими слоями можно с любой точностью аппроксимировать любую многомерную функцию на единичном отрезке.

На сегодняшний день существует много разновидностей нейронных сетей, характерных для различных типов задач, которые можно классифицировать по следующим признакам:

- тип входной информации (аналоговые или двоичные данные);
- характер обучения (с учителем и без);

КОЛМОГОРОВ Андрей Николаевич
(1903-1987)
выдающийся отечественный математик

- характер настройки синапсов (фиксированные или динамические связи);
- метод обучения (обратное распространение, конкурентная настройка, использование правила Хебба, гибридные сети и т.д.);
- характер связей (прямое и прямое-обратное распространение информации);
- архитектура (персептроны, самоорганизующиеся ИНС Кохонена, сети адаптивного резонанса, рециркуляционные, рекуррентные, встречного распространения, ИНС с обратными связями Хэмминга и Хопфилда, с двунаправленной ассоциативной памятью, с радиально-базисной функцией активации и т.п.).

Рассмотрим возможности обучения ИНС на примере многослойного персептрона с прямым распространением информации. Характер связей сети для нашего случая будет выглядеть, как показано на рис 9.4:

Нейроны регулярным образом организованы в слои, причем элементы некоторого слоя связаны только с нейронами предыдущего слоя, и информация распространяется от предыдущих слоёв к последующим. Входной слой, состоящий из чувствительных (сенсорных) ***S***-элементов, на который поступают входные сигналы X_i, никакой обработки информации не совершает и выполняет лишь распределительные функции. Каждый ***S***-элемент связан с совокупностью ассоциативных элементов (***A***-элементов) первого промежуточного слоя, а ***A***-элементы последнего слоя соединены с реагирующими элементами (***R***-элементами).

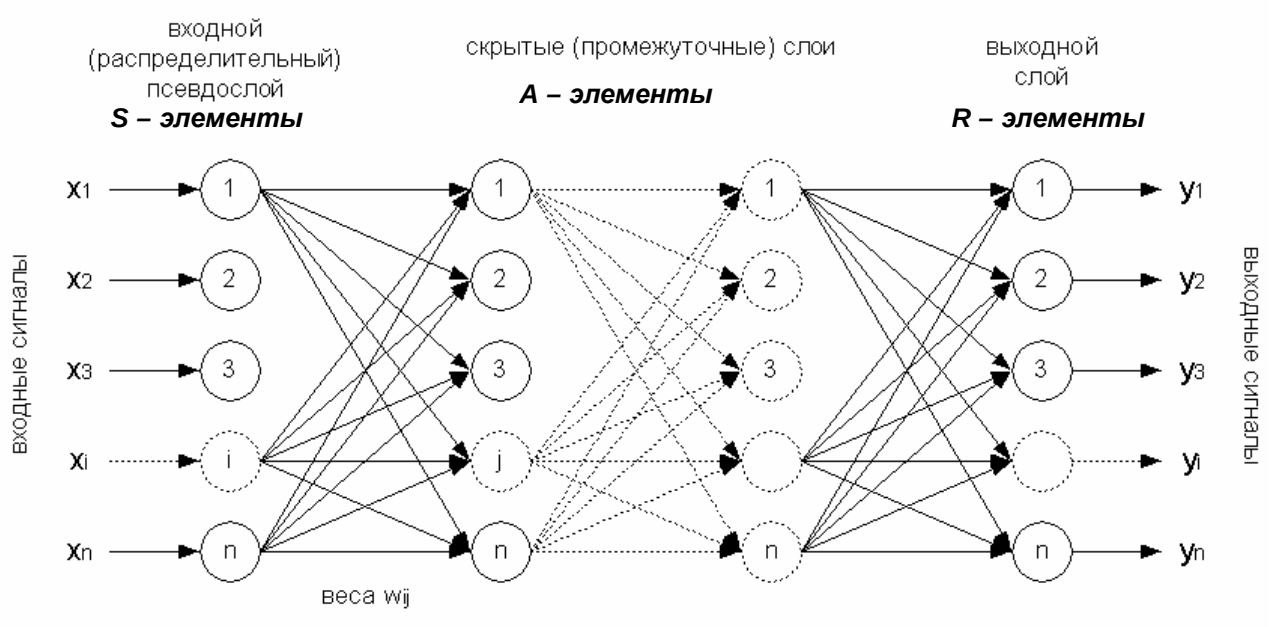


Рис. 9.4. Структура многослойного персептрона

Взвешенные комбинации выходов ***R***-элементов составляют реакцию системы, которая указывает на принадлежность распознаваемого объекта определенному образу. Если распознаются только два образа, то в персептроне устанавливается один ***R***-элемент, который обладает двумя реакциями – положительной и отрицательной. Если образов больше двух, то для каждого образа ус-

танавливают свой ***R***-элемент, а выход каждого такого элемента представляет линейную комбинацию выходов ***A***-элементов.

Выбор количества слоёв и типа активационной функции влияет на способность сети решать те или иные задачи. *Однослойная сеть* способна формировать линейные разделяющие поверхности и легко сводится к рассмотренному нами в разделе 8.5 методу обобщенного портрета. Линейная модель является хорошей точкой отсчета для сравнения эффективности различных *многослойных сетей* с нелинейными функциями активации, позволяющими, как показано выше, аппроксимировать любые выпуклые многомерные функциональные зависимости.

Уточним механизм моделирования разделяющих поверхностей многослойным персептроном. Уровень активации каждого нейрона представляет собой простую линейную функцию ***X***-ов, т.е. берется взвешенная сумма подающихся на вход сигналов с добавлением к ней порогового значения. Эта активация затем преобразуется с помощью заданной функции $y = F(S)$. Если в качестве функции активации использовать сигмоидную кривую, имеющую ***S***-образную форму, то комбинация линейной функции нескольких переменных и скалярной логистической функции $y = \frac{1}{1+e^{-cs}}$ приводит к характерному профилю "сигмоидного склона", который выдает элемент первого промежуточного слоя.

На приведенном рис. 9.5а соответствующая поверхность изображена в виде функции двух входных переменных. При изменении весов $w_1 \ldots w_n$ и порогов T может меняться как ориентация всей поверхности отклика, так и крутизна склона (большим значениям весов соответствует более крутой склон). Элемент с большим числом входов выдает многомерный аналог представленной поверхности.

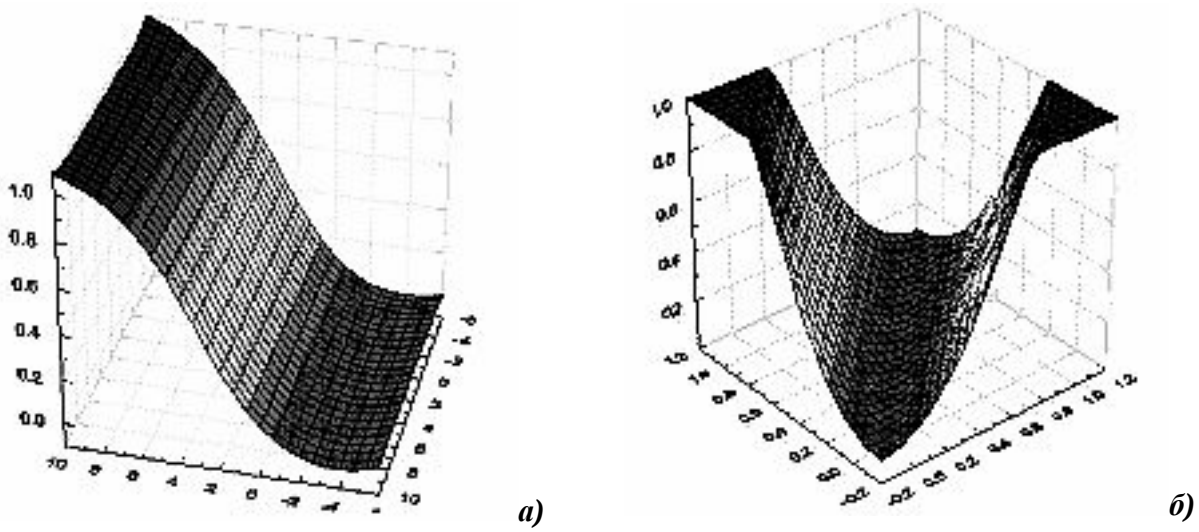


Рис. 9.5. Вид простейшей разделяющей поверхности сигмоидного типа для одного нейрона (фиг. а) и персептрона с одним промежуточным слоем (фиг. б)

В многослойной сети подобные функции отклика комбинируются друг с другом с помощью последовательного взятия их линейных комбинаций и применения нелинейных функций активации. На рисунке 9.5б изображена типичная поверхность отклика для сети с одним промежуточным слоем, состоящим из двух элементов, и одним выходным элементом. Две разных сигмоидных поверхности объединены в одну поверхность, имеющую форму буквы «**U**».

Обучение многослойного персептрона

После того, как определено число слоев и число элементов в каждом из них, нужно найти значения весовых коэффициентов $w_1,\ldots,w_n$ и порогов для каждого составляющего нейрона, которые бы минимизировали ошибку прогноза, выдаваемого сетью. Для того, чтобы оценить, насколько верное решение сделала ИНС, определяется среднеквадратичная ошибка сети как разница между эталонным и выходным вектором:

$$E = \frac{1}{n}\sum_j (y_j - y_j^*)^2, \qquad (9.22)$$

где Y^*_j – известное выходное значение для примеров обучающей выборки. Процесс обучения сети заключается в многократной подгонке коэффициентов $w_1,...,w_n$ к имеющимся выборочным данным с использованием различных алгоритмов нелинейной оптимизации.

Здесь оказывается очень полезным понятие поверхности ошибок. Каждому из N свободных параметров модели (весов и порогов) соответствует одно измерение в многомерном пространстве. Пусть (N+1)-е измерение соответствует ошибке сети E. Для всевозможных сочетаний значений весов соответствующие им значения ошибок представляют собой множество точек, образующих поверхность ошибок. Тогда цель обучения нейронной сети состоит в том, чтобы найти на этой многомерной поверхности самую низкую точку. В случае линейной модели с суммой квадратов в качестве функции ошибок поверхность ошибок представляет собой параболоид – гладкую поверхность, похожую на часть поверхности сферы, с единственным минимумом, который достаточно легко локализовать. В случае нейронной сети поверхность ошибок имеет гораздо более сложное строение и имеет локальные минимумы, плоские участки, седловые точки и длинные узкие овраги (см. рис. 9.6).

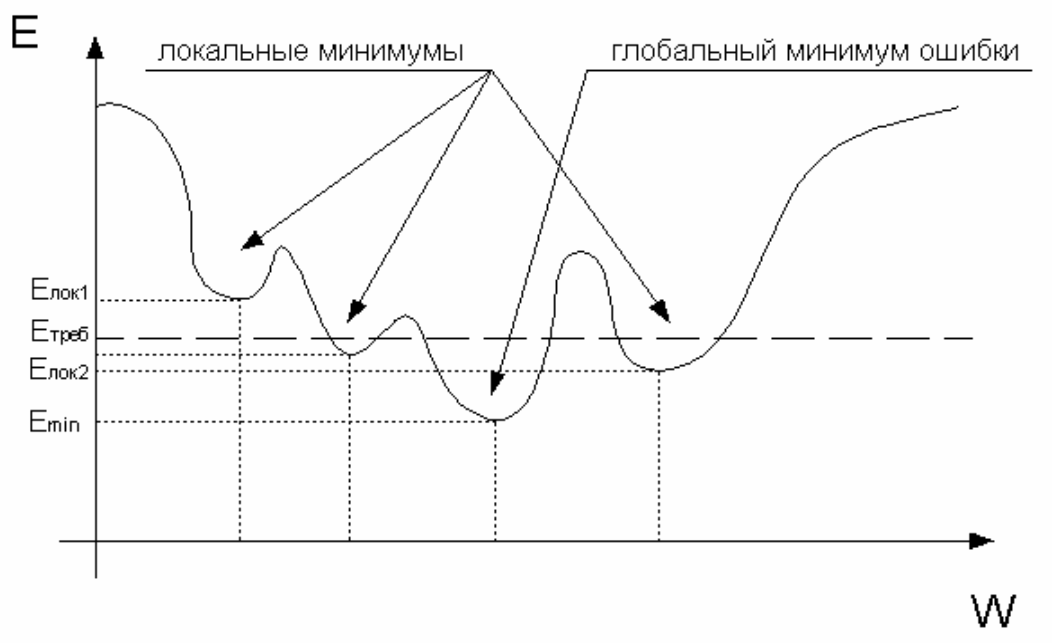


Рис. 9.6. Интерпретация поверхности ошибок нейронной сети

Аналитическими средствами найти в таких условиях глобальный минимум невозможно, поэтому процесс обучения ИНС, по сути дела, заключается в исследовании этой поверхности. Отталкиваясь от случайной начальной конфигурации весов и порогов, *алгоритм оптимизации* начинает передвижение к минимуму, используя на каждом шаге вектор градиента (т.е. направления кратчайшего спуска). Можно сказать, что процесс ведет себя как слепой кенгуру – каждый раз прыгает в направлении, которое ему кажется наиболее привлекательным. В общем случае результат обучения ИНС может соответствовать субоптимальному решению, т.е. не глобальному оптимум, а решению, которое нас устроит. В самом деле, если $E_{лок2} < E_{требуемое}$, то такое решение вполне приемлемо. Другое дело, когда веса попадают в область локального минимума (такого как $E_{лок1}$) и при этом $E_{лок} > E_{требуемое}$, а величины шага обучения не хватает чтобы выйти оттуда.

Самый известный алгоритм обучения сети – так называемый, *алгоритм обратного распространения* (back propagation) [Нейронные сети.., 2001]. В основу метода легли прямой ход вычисления выходных значений, вычисление ошибки последнего слоя и рекурсивное обратное распространение.

Для выходного слоя определение ошибки δ_j нейрона j тривиально:

$$\delta_j = (y_j - y_j^*)F'(S_j)$$

и напоминает систему поощрений-наказаний, используемую при обучении однослойных сетей. Для каждого предыдущего слоя ошибка определяется рекурсивно через ошибку следующего слоя:

$$\delta_j = (\sum_{i=1}^{m} \delta_i w_{ij}) \cdot F'(S_j),$$

т.е. для каждого j-го нейрона ошибки следующего слоя как бы распространяются к нему обратно сквозь соответствующие веса. Этот механизм обратного распространения дополнен традиционными для многих градиентных методов оптимизации процедурами оценки вектора кратчайшего спуска, изменения величины шага пропорционально крутизне склона и проч. Существуют также современные алгоритмы обучения сетей второго порядка, такие как *метод сопряженных градиентов* и *метод Левенберга-Маркара*, которые во многих задачах работают на порядок быстрее [Bishop, 1995].

Переобучение и обобщение

Основная задача моделирования состоит в подборе сети, адекватно прогнозирующей совершенно новые наблюдения. С. Бишоп [Bishop, 1995] пишет: «*Минимизация ошибок на обучающем множестве, которое не бывает ни идеальным, ни бесконечно большим, - это совсем не то же самое, что минимизация «настоящей» ошибки заранее неизвестной модели явления*». Сеть с небольшим количеством весов может оказаться недостаточно гибкой, чтобы смоделировать имеющуюся зависимость. Однако, слишком большое количество нейронов и слоев позволяет моделировать очень сложные функции, но это часто приводит к переобучению сети, когда модель будет давать совершенно правильные ответы, но только на тех примерах, которым её обучили.

Выбор нейросети «правильной» сложности сводится к двум рецептам: использование контрольных выборок и экспериментирование. Механизм контрольной кросс-проверки заключается в том, что некоторая часть обучающих наблюдений резервируется, т.е. подгонка коэффициентов модели и поиск минимума ошибки сети по ним не осуществляется. Эти измерения, как и в алгоритмах МГУА, используются только для независимого контроля результата и называются контрольной выборкой. Если разбиение на обучающее и контрольное множества было выполнено однородно, то, по мере того как сеть обучается, ошибка обучения и ошибка на контрольном множестве будут одновременно уменьшаться. Если же контрольная ошибка перестала убывать или даже стала расти, это указывает на то, что сеть стала чересчур точно аппроксимировать данные и наступает фаза переобучения. В этом случае следует уменьшить число A-элементов или слоев сети.

Описанные проблемы с локальными минимумами и выбором размера сети приводят к тому, что в практической работе с ИНС приходится экспериментировать с большим числом различных вариантов конфигурации.

Понижение размерности входных переменных

Цель задачи заключается в таком преобразовании входных данных, чтобы та же информация была бы записана с помощью меньшего числа переменных. Мы уже обсуждали эту проблему, когда знакомились с методом главных компонент (см. раздел 7.5). Следует подчеркнуть, что один из основных недостатков традиционного факторного анализа состоит в том, что это преобразование является линейным и может улавливать только линейные направления максимальной вариации. Поэтому рассмотрим другой подход к решению этой проблемы: *нелинейный вариант метода главных компонент*, основанный на применении автоассоциативных сетей.

Автоассоциативная сеть – это сеть, предназначенная для воспроизведения на выходе своих же сигналов. У такой сети число выходов совпадает с числом входов, а все нейроны имеют особое свойство. Если число элементов промежуточного слоя сделать меньше числа входов/выходов, то это заставляет сеть "сжимать" информацию, представляя ее в меньшей размерности. Трехслойная автоассоциативная сеть сначала линейно преобразует входные данные в меньшую размерность промежуточного слоя, а затем снова линейно разворачивает их в выходном слое. Можно показать, что такая сеть на самом деле реализует стандартный алгоритм анализа главных компонент. Для того, чтобы выполнить нелинейное понижение размерности, нужно использовать

пятислойную сеть (см. рис. 9.7). Ее средний слой служит для уменьшения размерности, а соседние с ним слои, отделяющие его от входного и выходного слоев, выполняют нелинейные преобразования.

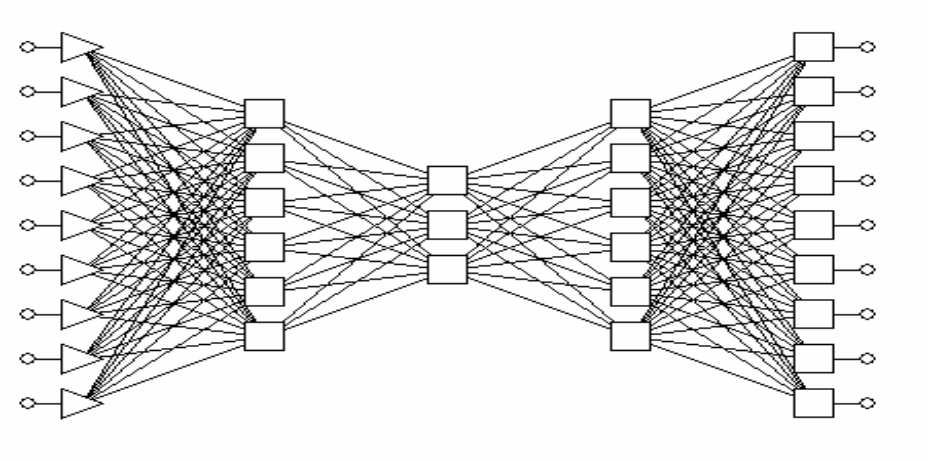


Рис. 9.7. Автоассоциативная сеть для понижения размерности признакового пространства с девяти до трех

Если из автоассоциативной сети удалить два последних слоя, то получается сеть для препроцессирования, с помощью которой генерируется версия входных данных в уменьшенной размерности.

Результаты расчетов

Прогнозирование альтернативной переменной

Построим нейронную сеть для классификации двух категорий качества вод "Чисто" / "Грязно" на базе многослойного персептрона с использованием обучающей выборки из 520 примеров, описанной в разделах 8.2, 8.3 и 9.2. Входами сети являются значения обилия 27 таксономических групп зообентоса, отобранных в результате генетического алгоритма (см. раздел 9.2), а также количество видов в пробе. По условиям работы всех алгоритмов построения сетей данные подвергаются *препроцессингу* – предварительному масштабированию исходных значений входов в единую шкалу: в нашем случае используем известную формулу минимакса, в результате чего каждый преобразованный признак варьируется на интервале [0,1].

Примем стандартную модель персептрона с 28 входами, одним выходом и одним промежуточным слоем, состоящем из 14 ***A***-нейронов, т.е. полусумме числа входных и выходных элементов (см. рис. 9.8). Выполним обучение сети в трех вариантах:

- с использованием линейной функции активации;
- с использованием логистической функции активации и оценкой качества сети по среднеквадратичной ошибке;
- с использованием логистической функции активации и оценкой качества сети по информационному критерию (энтропии).

Для каждого из вариантов обучение проведем в двух режимах: без использования контрольного множества и с применением кросс-проверки на контрольном множестве. В первом случае сеть обучается на всех 520 примерах, а во втором случае 120 примеров выделяются в качестве внешнего дополнения и используются для независимого контроля качества сети.

Построенная сеть в нашем случае для каждого произвольного вектора $\boldsymbol{X}$ выдает на выходе некоторое значение апостериорной вероятности на интервале от 0 до 1. Для конкретной классификации примеров используются два задаваемых доверительных уровня: *порог принятия* (т.е. минимальное значение выхода, при котором наблюдение будет считаться принадлежащим классу 1) и *порог отвержения* (т.е. максимальное значение выхода, относящее измерение к классу 0). В нашем примере мы задали эти пороги равными 0.55 и 0.45 соответственно. Все вектора, предъявленные сети и имеющие отклик внутри диапазона доверительных уровней, классифицируются как неопределенные.

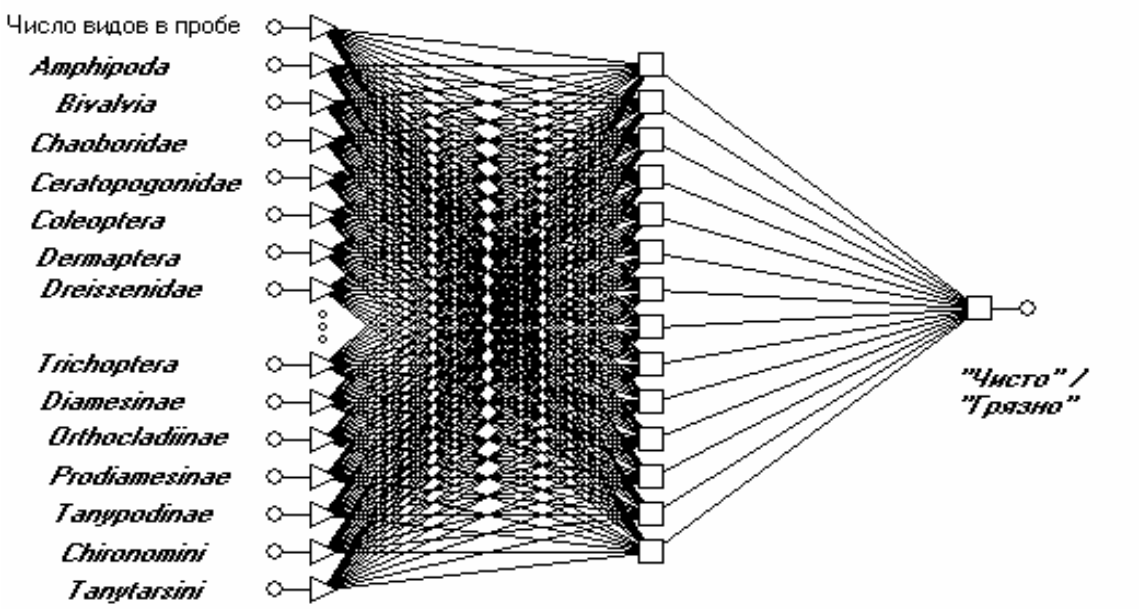


Рис. 9.8. Схема многослойного персептрона для оценки категории качества воды по обилию групп зообентоса

Результаты моделирования по традиции для всех задач классификации оценим как долю ошибочных опознаний на обеих выборках (см. табл. 9.5, где приведены данные только для режима с кросс-проверкой).

Таблица 9.5

Результаты использования модели персептрона с одним промежуточным слоем для оценки категории качества воды

Вид функции активации нейронов	Оценка качества сети	Результат прогноза	Обучающая выборка			Контрольное множество		
			Класс "Чисто"	Класс "Грязно"	Правильный прогноз, %	Класс "Чисто"	Класс "Грязно"	Правильный прогноз, %
Линейная функция	Среднеквадратическая ошибка	"Чисто"	136	5	94.4	28	7	75.6
		"Грязно"	7	271	98.1	7	53	84.1
		Отказ	1	0		2	3	
		Итого	144	276	**96.9**	37	63	**81.0**
Логистическая (сигмоидная) функция	Среднеквадратическая ошибка	"Чисто"	140	0	97.2	33	7	89.2
		"Грязно"	3	276	100	3	55	87.3
		Отказ	1	0		1	1	
		Итого	144	276	**99.0**	37	63	**88.0**
Логистическая (сигмоидная) функция	Простая энтропия	"Чисто"	141	0	97.9	34	4	91.9
		"Грязно"	2	276	100	2	58	92.1
		Отказ	1	0		1	1	
		Итого	144	276	**99.3**	37	63	**92.0**

На основе выполненных расчетов можно сделать следующие очевидные выводы:

- использование методов нейросетевого моделирования для этого примера дало весьма ощутимый выигрыш по сравнению с логистической регрессией (раздел 9.2) и линейным дискриминантным анализом (раздел 9.3);

- вариант построения нелинейной разделяющей поверхности с использованием сигмоидной функции активации более эффективен по сравнению с вариантом линейных разделяющих гиперплоскостей;
- использование энтропии в качестве функции ошибок сети с альтернативным параметром выхода имеет некоторые преимущества по сравнению со стандартным среднеквадратичным отклонением, которое теоретически больше подходит для задач регрессии;
- кросс-проверка на контрольном множестве, включенная при обучении, в двух случаях из трех привела к уменьшению ошибки на самом обучающем множестве и дала возможность провести независимую оценку качества на внешнем дополнении;
- ошибка на контрольной выборке оказалась несколько выше, чем при обучении, однако, в нашем случае этот факт вряд ли можно объяснить феноменом переобучения.

Нейронную сеть вследствие ее принципиальной многомерности обычно интерпретируют как "черный ящик", поскольку визуальный или статистический анализ коэффициентов усиления и порогов, как это делается, например, в множественной регрессии, для всех структурных связей графа представляет собой весьма утомительную процедуру, не гарантирующую достоверных выводов. Тем не менее, при желании, можно заглянуть "внутрь" этого ящика и попытаться выяснить роль каждого скрытого элемента, создаваемый ими потенциал активации и прочие нейронные характеристики (в рамках нашего изложения мы этого желания не испытали).

Сети для множественной классификации

Построим нейронную сеть для непосредственной оценки значения класса качества водоемов в виде числа от 1 до 6. Сформируем выборку из тех же 520 измерений, но в качестве девяти варьируемых переменных будем использовать различные обобщенные гидробиологические показатели и традиционные "интегральные" индексы зообентоса.

Архитектура выбранного персептрона с одним промежуточным слоем представлена на рис. 9.9. Сеть имеет 5 выходов, соответствующих каждому из присутствующих в обучающей выборке классов качества воды. При правильном распознавании на одном из выходных элементов появляется высокое значение активации при незначительной ее величине на остальных четырех.

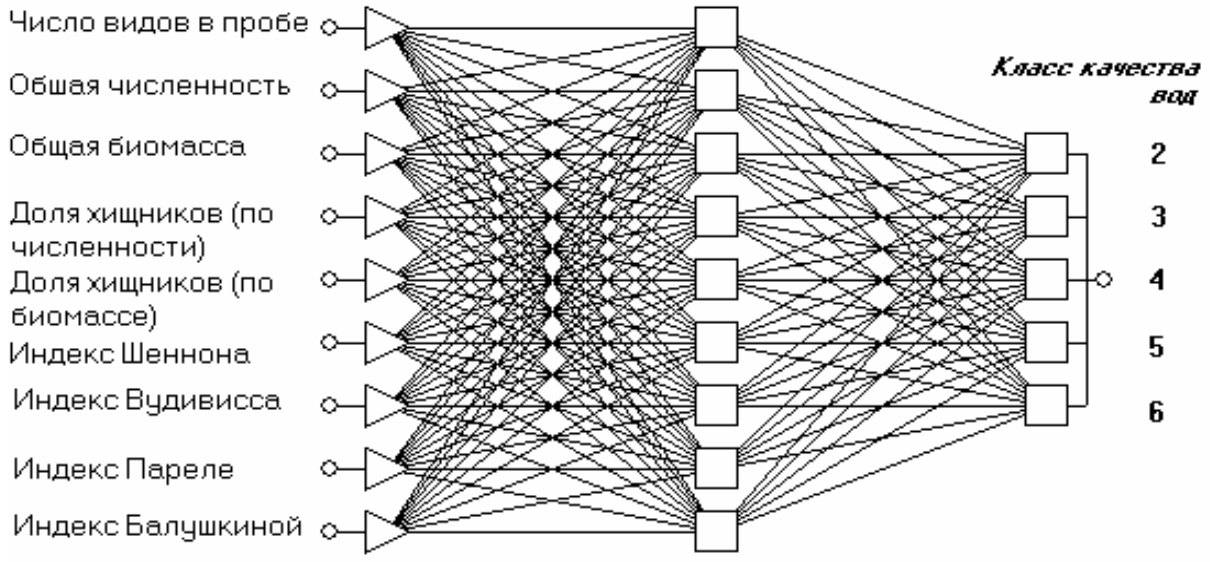


Рис. 9.9. Нейронная сеть для прогнозирования пяти классов качества воды

Нами использовалась в выходном слое специальная функция активации (Softmax), представляющая собой взвешенную и нормированную на единицу сумму экспонент. Можно показать, что если входные данные представляют собой выборку из какого-либо экспоненциального распределения, то выходы софтмакс-элементов можно трактовать как вероятности (напомним, что самым известным примером экспоненциального распределения является нормальное). Например, если

для измерения на ст. 1 р. Байтуган активации выходных нейронов сети оказались равными {0.314, 0.503, 0.142, 0.028, 0.018 }, то с вероятностью 0.503 можно предположить, что это измерение было взято из водоема 3 класса качества, а с вероятностью 0.817 – из водоема 2-3 классов.

Хотя анализ весовых коэффициентов и активностей отдельных фрагментов сети представляется неблагодарной деятельностью, методы нейросетевого моделирования позволяют оценить относительную значимость влияния отдельных исходных переменных на отклик, т.е. выполнить анализ чувствительности сети по ее входам. Этот анализ дает возможность идентифицировать неинформативные переменные с низкой чувствительностью, которые могут благополучно игнорироваться в последующих просчетах. Разумеется, подобные заключения должны приниматься со всей осторожностью, поскольку никакой анализ чувствительности не гарантирует абсолютную точность и надежность оценок "полноценности" переменных в модели. Например, в литературе приводится много примеров взаимного влияния и обусловленности исходных факторов, когда ни одна из переменных по отдельности не несет существенную информацию (классы будут выглядеть совершенно перемешанными), но, учитывая всю совокупность признаков, классы можно легко разделить.

Чувствительность каждой входной переменной измеряется в терминах ошибки сети: т.е. вычисляется прирост общей погрешности предсказания, если бы анализируемая переменная была бы исключена на входе. Более удобно для анализа *отношение двух ошибок*: ошибки сети без использования ***j***-й переменной к аналогичной ошибке при ее наличии. Если это отношение равно или менее 1, то анализируемая переменная либо не участвует в работе нейросетевой модели вообще, либо даже мешает ее работе, снижая общую эффективность. Иногда стоит удалять переменные даже с чувствительностью несколько более чем 1.0, и переквалифицировать сеть с целью достижения ее большей компактности и надежности. Значения прироста ошибки δ_j и отношения чувствительности η_j для переменных, использованных в рассматриваемом примере, представлены в табл. 9.6. Подробные выводы легко может сделать сам читатель; мы же выскажем только предположение, что странная "рокировка" ролей общей биомассы и индекса Шеннона при переходе от стадии обучения к контролю объясняется не только случайными причинами.

Таблица 9.6

Анализ чувствительности входных переменных модели многослойного персептрона для прогнозирования класса качества вод

Анализируемые входные переменные	Обучающая выборка			Контрольная выборка		
	Ранг	δ_j	η_j	Ранг	δ_j	η_j
Биотический индекс Вудивисса	1	0.377	1.095	1	0.381	1.078
Олигохетный индекс Пареле	2	0.355	1.030	2	0.369	1.045
Общая биомасса (ln), мг/м^3	3	0.351	1.018	7	0.354	1.003
Число видов в пробе	4	0.350	1.016	3	0.357	1.012
Хирономидный индекс Балушкиной	5	0.348	1.010	8	0.354	1.003
Общая численность (ln), экз/м^3	6	0.348	1.009	5	0.356	1.007
Доля хищников (по численности)	7	0.347	1.007	9	0.353	0.999
Доля хищников (по биомассе)	8	0.347	1.006	6	0.356	1.007
Индекс Шеннона	9	0.346	1.003	4	0.356	1.008

Результаты достоверности оценки классов качества представим в виде таблицы сопряженности "Факт/Прогноз" (табл. 9.7), где по главной диагонали проставлены частоты правильной оценки групп измерений, а в остальных клетках – имеющиеся ошибки прогноза отдельно для обучающей выборки и контрольной последовательности.

Несмотря на всю математическую мощь нейронных сетей, эффективность распознавания класса качества вод с использованием персептрона лишь незначительно превысила достоверность дискриминантного анализа (см. раздел 8.3) и уступила методу, основанному на видовых индикаторных индексах (раздел 8.4). Очевидно, что здесь был достигнут порог насыщения для уровня информативности "интегральных" показателей зообентоса, превысить который невозможно никакими интеллектуальными ухищрениями. Модель на данном наборе признаков сделала все, что

могла и дальнейшие резервы повышения эффективности – в расширении признакового пространства с привлечением данных о видовой структуре. Это подтверждает и тот уникальный для методов распознавания факт, когда ошибка классификации на контрольной выборке оказалась меньше, чем на обучающей последовательности.

Таблица 9.7

Результаты прогнозирования класса качества вод по модели многослойного персептрона

Классы качества вод		Фактические					Итого прогноз	Правильный прогноз, %	Ошибка на два и более класса, %
		2	***3***	***4***	***5***	***6***			
Прогноз по обучающей выборке	***2***	**18**	10	3	0	0	31	58.06	9.67
	3	13	**52**	24	6	0	95	54.47	6.31
	4	11	25	**77**	15	17	145	53.10	19.31
	5	0	10	14	**36**	15	75	48.0	13.3
	6	1	2	6	6	**9**	24	37.5	37.5
Итого факт		43	99	124	63	41	370	51.89	13.5
Прогноз по контрольной выборке	***2***	**7**	3	2	0	0	12	58.33	16.6
	3	2	**15**	14	3	1	35	42.86	11.42
	4	1	8	**40**	6	7	62	64.52	12.9
	5	0	2	5	**13**	11	31	41.94	6.45
	6	0	1	2	2	**5**	10	50.00	30.0
Итого факт		10	29	63	24	24	150	53.33	12.66

Задача о понижении размерности

Чтобы осуществить нелинейное понижение размерности исходной матрицы наблюдений для прогнозирования класса качества вод, используемой в предыдущем примере, выполним следующие действия:

- построим автоассоциативную сеть – персептрон с пятью слоями, как показано на рис. 9.7;
- обучим автоассоциативную сеть на имеющейся обучающей выборке 9x520 с использованием любого итеративного алгоритма (для определенности используем метод сопряженных градиентов);
- удалим два последних слоя автоассоциативной сети и получим сеть для пре-процессирования, понижающую размерность данных (см. рис. 9.10);
- с помощью пре-процессирующей сети получим версию входных данных в уменьшенной размерности: те же 520 строк исходной таблицы наблюдений с метками класса качества вод, но количество варьируемых признаков сведено от 9 к 3 без существенной потери информации;
- построим обрабатывающую сеть в виде второго трехслойного персептрона и обучим ее на модифицированном обучающем множестве, полученном на предыдущем шаге;
- объединив обе сети (см. рис. 9.10), получим единую сеть, которая последовательно пересчитывает исходные данные в пространство трех главных компонент, после чего обрабатывает уже преобразованные факторы, прогнозируя значение отклика.

В нашем конкретном случае достоверность распознавания по полученной двухуровневой сетевой модели несколько снизилась по сравнению с обычным персептроном, представленным на рис. 9.9. Однако общий смысл перехода в пространство небольшой размерности состоит не столько в том, чтобы повысить эффективность прогноза, сколько в попытке дать какое-то разумное объяснение имеющимся внутренним механизмам анализируемых явлений. Один из способов это сделать – проанализировать двух- или трехмерную визуализацию классифицируемых объектов в осях главных факторов (см. рис. 9.11). Изучив на полученной "картинке" сильно трангрессирующее облако точек, принадлежащих разным классам качества воды, мы, по крайней мере, поняли, какая трудновыполнимая задача была поставлена перед методами распознавания.

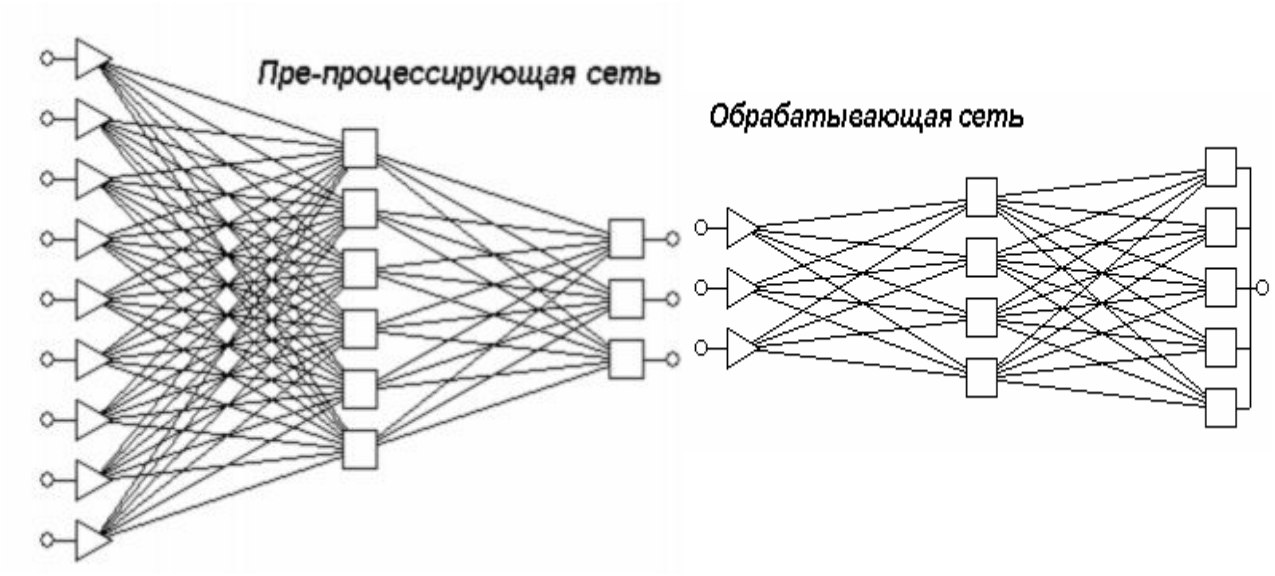


Рис. 9.10. Последовательность двух сетей для понижения размерности признакового пространства

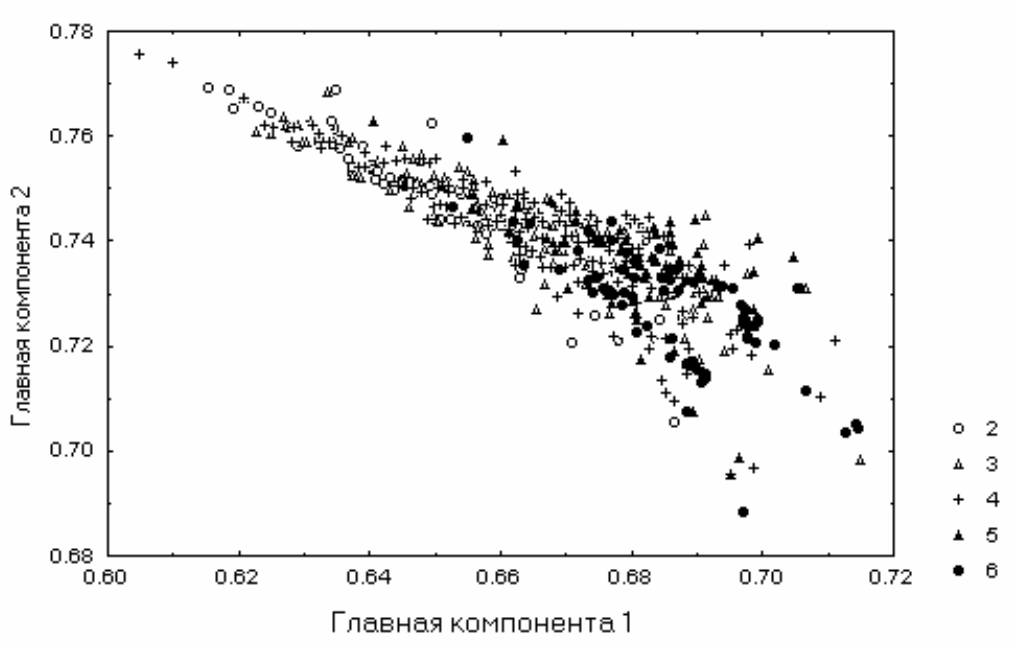


Рис. 9.11. Изображение в пространстве двух главных компонент облака точек, соответствующих различным классам качества воды

9.5. Решение задачи регрессии с помощью нейросетей различной архитектуры

Формулировка задачи

Пусть в таблице произвольных гидробиологических наблюдений $\boldsymbol{X}$ размерностью $\boldsymbol{m}$ >1 откликом $\boldsymbol{Y}$ является один из любых признаков, измеренных в количественной шкале.

Необходимо решить задачу регрессии, целью которой является оценка по матрице входных переменных параметров функции выходной переменной, принимающей непрерывный диапазон числовых значений.

Математический лист (продолжение раздела 9.4)

Нейронные сети наиболее часто используемых архитектур выдают выходные значения в некотором определенном диапазоне (например, на отрезке от 0 до 1 в случае логистической функции активации). Для задач классификации это не создает никаких трудностей. Однако для задач регрессии особую важность имеет масштаб и диапазон существования выходных значений, поскольку на передний план выходят проблемы, связанные с эффектом экстраполяции.

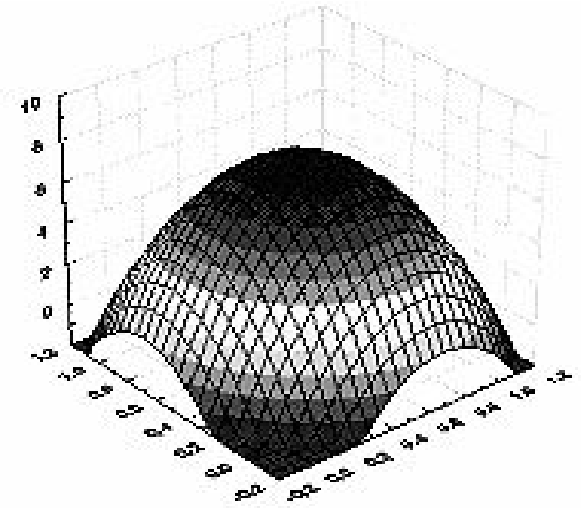


Рис. 9.12. Вид функции радиального элемента

Как показано в главе 1.5, простейшей из масштабирующих функций, сводящей переменные сети к "приемлемому" диапазону, является минимаксная функция: она находит минимальное и максимальное значение переменной по обучающему множеству и выполняет линейное преобразование так, чтобы значения лежали в нужном диапазоне, как правило, на отрезке [0,1]. Если эти действия применяются только к измерениям обучающей выборки, то есть гарантия, что результаты преобразования попадут в область возможных выходных значений сети. Сеть может быть обучена, но выход сети будет находиться в определенных границах, пересечение которые будет пресекаться.

Это обстоятельство можно считать достоинством, если бы не проблема экстраполяции: если продолжать кривую вправо по числовой оси, то выход ее за лимитируемые пределы неизбежен, даже если мы еще достаточно близко отошли от диапазона обучающих векторов. Чтобы избежать этого, сужают целевой диапазон минимаксной масштабирующей функции, например, делают его от 0.25 до 0.75, создавая некоторый запас. Интересно заметить, что на среднем участке сигмоидная кривая "почти линейна", поэтому другой путь для учета экстраполяции - использование линейного выходного слоя.

Задачи регрессии методами нейросетевого моделирования можно решать с помощью сетей различных типов: многослойного персептрона, линейной сети, радиальной базисной функции и обобщенной регрессионной сети. Линейная модель по сути ничем не отличается от обычной линейной регрессии, но на языке нейронных сетей представляется сетью без промежуточных слоев, которая в выходном слое содержит только линейные элементы (то есть элементы с линейной функцией активации). Обучить линейную сеть можно с помощью стандартного алгоритма линейной оптимизации.

В предыдущем разделе было описано, как многослойный персептрон моделирует функцию отклика с помощью функций "сигмоидных склонов". Столь же естественным является подход, основанный на разбиении пространства окружностями или, в общем случае, гиперсферами, которые задаются своим центром и радиусом. Поверхность отклика такого радиального элемента представляет собой гауссову функцию колоколообразной формы, с вершиной в центре и понижением к краям (см. рис. 9.12). Наклон гауссова радиального элемента можно менять подобно тому, как можно менять наклон сигмоидной кривой в персептроне.

Сеть, построенная на *радиальных базисных функциях* (*RBF*), имеет промежуточный слой из радиальных элементов, каждый из которых воспроизводит гауссову поверхность отклика. Поскольку эти функции нелинейны, то для моделирования любой произвольной функции отклика нет необходимости использовать более одного промежуточного слоя – достаточно лишь взять оптимальное число радиальных элементов. RBF-сети имеют как ряд достоинств (компактность, быстрая обучаемость), так и недостатков. Например, с "групповым" представлением пространства модели связано неумение сетей RBF экстраполировать свои выводы за область известных данных: при удалении от обучающего множества значение функции отклика быстро падает до нуля.

В предыдущем разделе, говоря о задачах классификации, мы упомянули о том, что выходы сети можно интерпретировать как оценки вероятности того, что элемент принадлежит некоторому классу, и сеть, фактически, "учится" оценивать функцию плотности вероятности. Аналогичная интерпретация может иметь место и в задачах регрессии – выход сети рассматривается как ожидаемое значение модели в данной точке пространства входов, связанное с плотностью вероятности совместного распределения входных и выходных данных.

Задача оценки плотности вероятности имеет давнюю историю в математике и относится к области байесовой статистики. Возможный подход к оценке плотности вероятности основан на *ядерных оценках Парзена* [Parzen, 1962], связывающих ансамбли близко лежащих точек с некоторым доверием к уровню плотности, которое по мере отдаления убывает и стремится к нулю. В методе ядерных оценок в точке, соответствующей каждому наблюдению, помещается некоторая простая функция (например, гауссова функция), затем все они складываются, и в результате получается оценка для общей плотности вероятности. Если обучающих примеров достаточное количество, то такой метод дает достаточно хорошее приближение к истинной плотности вероятности.

Аппроксимация плотности вероятности с помощью ядерных функций является методологической основой для *вероятностных* (*PNN*) и *обобщенно-регрессионных* (*GRNN*) *нейронных сетей*. В этих сетях в точку расположения каждого обучающего наблюдения помещается гауссова ядерная функция. Окончательная выходная оценка сети получается как взвешенное среднее выходов по всем обучающим наблюдениям, где величины весов отражают расстояние от этих наблюдений до той точки, в которой производится оценивание. Таким образом, более близкие точки вносят больший вклад в оценку.

Первый промежуточный слой сети GRNN состоит из радиальных элементов, а второй промежуточный слой содержит элементы, которые помогают оценить взвешенное среднее и состоит из двух нейронов. Обобщенно-регрессионная сеть обучается почти мгновенно, но может получиться большой и медленной. Как и сеть RBF, сеть GRNN не обладает способностью эффективно экстраполировать данные.

Результаты расчетов

<u>*Моделирование индивидуального веса особи*</u>

Выполним синтез различных нейронных сетей для прогнозирования зависимости среднего веса особей семейства хирономид от восьми различных переменных, подробно описанных в разделе 8.1:

1. логарифма индекса плотности населения $\ln((\boldsymbol{N_s}*\boldsymbol{B_s})^{1/2})$, где $\boldsymbol{N_s}$ и $\boldsymbol{B_s}$ – суммарные численность и биомасса;
2. общего числа видов зообентоса в пробе;
3. информационного индекса Шеннона;
4. биотического индекса Вудивисса;
5. олигохетного индекса Пареле;
6. сезонной составляющей, учитывающей дату проведения наблюдения;
7. класса качества вод по гидрохимическим показателям в точке отбора пробы;
8. типа водоема.

В составе общей таблицы наблюдений, состоящей из 473 измерений, выделим обучающую выборку из 400 объектов, а остальные примеры используем для контроля. Средние значения выходной переменной составили для обучающей и контрольной выборок, соответственно, 1.78 и 3.24 мг, стандартные отклонения – 2.70 и 6.71.

С помощью "интеллектуального генератора" построим 50 нейронных сетей различного типа и архитектуры, из которых отберем наилучшие модели, т.е. имеющие наименьшую ошибку на контрольном множестве. Будем попутно исключать при этом из числа входных переменных признаки, имеющие низкую чувствительность. Основные результаты расчетов представлены в табл. 9.8.

Эти результаты позволяют сделать два достаточно общих вывода:

- еще раз подтверждается тезис "множественности моделей" В.В. Налимова (в частности, исключение из модели от 2 до 4 различных переменных не приводит к серьезному ухудшению их качества);

- результаты свидетельствуют о хороших интерполяционных и плохих экстраполяционных качествах обобщенно-регрессионных сетей, антиподом которых в этом отношении являются линейные сети.

Однако, как основной вывод, следует признать, что между средним весом особей и представленным комплексом переменных существует достаточно слабая статистическая зависимость, которую *не удалось существенно улучшить* с использованием различных методов нейросетевого моделирования. Например, на обучающей выборке лишь GRNN-модель оказалась чуть лучше (коэффициент корреляции 0.464 против 0.394) обычной линейной регрессионной модели, представленной в таблице 8.1 раздела 8.1.

Доля общей вариации индивидуального веса, объясненная любой из протестированных сетей табл. 9.8, составляет от 3 до 5%. В то же время, если на той же выборке в качестве выходной переменной принять, например, индекс Шеннона, то доля объясненной дисперсии даже при моделировании трехслойным персептроном составит от 42.3 % на обучающей выборке до 38.6 % на контрольной последовательности. Коэффициент корреляции Пирсона в тех же условиях изменяется от 0.817 до 0.790 соответственно, что следует оценить как весьма высокий показатель применительно к гидробиологическим моделям.

Таблица 9.8

Результаты прогнозирования среднего веса особи с использованием нейронных сетей различного типа (в числителе – показатель для обучающей выборки, в знаменателе – для контрольной последовательности)

Наименование показателей	Трехслойный персептрон	Линейная модель	Радиальные базисные функции ***RBF***	Обобщенно-регрессионная модель ***GRNN***
Характеристики сети: количество элементов входного и скрытого слоев	Входных – 6 Скрытых – 6	Входных – 4	Входных – 6 Скрытых – 2	Входных – 8 Скрытых – 400
Перечень входных переменных, исключенных из модели	6 , 8	1, 3, 5, 7	4, 6	-
Средняя абсолютная разность расчетных и фактических значений	3.08 4.02	1.49 2.80	1.48 2.80	1.42 2.80
Стандартное отклонение ошибки сети	2.58 6.51	2.65 6.50	2.62 6.67	2.49 6.59
Отношение стандартного отклонения ошибки сети к стандартному отклонению отклика	0.955 0.970	0.980 0.968	0.970 0.994	0.922 0.982
Коэффициент корреляции расчетных и фактических значений	0.297 0.328	0.199 0.402	0.242 0.111	0.464 0.264

Анализ связи между гидрохимическими и гидробиологическими показателями

Ранее нами были приведены примеры построения регрессионных моделей по методу "включений с исключениями" Эфроимсона (раздел 8.1) и моделей МГУА на основе алгоритма многорядной селекции (раздел 9.3). В этих примерах оценивалась взаимосвязь между некоторыми гидрохимическими параметрами качества воды (концентрациями ионов железа, аммонийного азота, минерального фосфора и БПК) и 7 основными гидробиологическими показателями обилия и индексами: $\boldsymbol{X_H}$ – информационного индекса Шеннона, $\boldsymbol{X_V}$ – биотического индекса Вудивисса, $\boldsymbol{X_P}$ – олигохетного индекса Пареле, $\boldsymbol{X_{CI}}$ – хирономидного индекса Балушкиной, числа видов $\boldsymbol{X_S}$, логарифмов суммарной численности $\boldsymbol{X_N}$ и биомассы $\boldsymbol{X_B}$ зообентоса в пробе.

Для прогнозирования каждого из указанных гидрохимических показателей в ходе перебора многих вариантов была построена наилучшая нейросетевая модель со следующими параметрами:

- для ионов железа – трехслойный персептрон (MLP-сеть) с 5 элементами промежуточного слоя на основе 5 входных переменных, в состав которых не вошли индекс Шеннона $\boldsymbol{X_H}$ и число видов $\boldsymbol{X_S}$;

- для аммонийного азота – RBF-сеть на основе радиально-базисных функций с 6 входными переменными (исключен индекс Пареле $\boldsymbol{X_P}$) и 9 элементами промежуточного слоя;
- для минерального фосфора – RBF-сеть с 6 входными переменными (исключена биомасса $\boldsymbol{X_B}$) и 27 элементами промежуточного слоя;
- для БПК – MLP-сеть с 9 элементами промежуточного слоя на основе 6 входных переменных, в состав которых не вошел индекс Вудивисса.

Несмотря на известную теоретическую проблематичность корректного сравнения математических моделей, имеющих разную параметричность, количество степеней свободы и проч. [Брусиловский, Розенберг, 1981; Розенберг, 1989], мы сочли возможным предложить читателю некоторый анализ эффективности моделирования всеми протестированными методами. Как и в другой нашей работе [Розенберг с соавт., 1994], сопоставление моделей-претендентов проведем по системе критериев, представленных в табл. 9.9.

В целом, определенные преимущества в "соревновании" методов прогнозирования оказались на стороне моделей ИНС. Однако даже в ходе перебора более четырехсот версий различных сложных нейросетей, не удалось найти модель для прогноза концентрации аммонийного азота, лучшую, чем простенькое уравнение регрессии. В некоторых случаях с позиций различных критериев эти оценки могут разойтись: например, для прогноза БПК, если принять во внимание коэффициент корреляции, лучшей оказалась модель ИНС, но меньшую среднюю ошибку доставляет уравнение регрессии.

Таблица 9.9

Оценка эффективности различных моделей-претендентов прогнозирования гидрохимических показателей по совокупности критериев (сокращения: МШР – модель пошаговой регрессии, МГУА – модели самоорганизации, ИНС – нейросетевые модели)

Химический компонент	Тип модели	Среднеквадратическая ошибка	Средний модуль ошибки	Максимальный модуль ошибки	Критерий регулярности	Коэффициент корреляции	Критерий Дарбина-Уотсона
Железо	МШР	0.317	0.242	1.09	0.847	0.532	1.83
	МГУА	0.322	0.237	1.08	0.840	0.440	2.04
	ИНС	**0.293**	0.205	1.27	0.781	**0.628**	2.18
Аммонийный азот	МШР	**0.230**	0.169	1.04	0.908	**0.418**	1.37
	МГУА	0.237	0.177	1.09	0.907	0.364	1.147
	ИНС	0.233	0.171	1.11	0.920	0.395	1.22
Минеральный фосфор	МШР	0.0756	0.0575	0.269	0.950	0.312	0.981
	МГУА	0.0730	0.0567	0.241	0.932	0.372	1.072
	ИНС	**0.0665**	0.0485	0.234	0.836	**0.551**	1.33
$БПК_5$	МШР	**1.70**	**1.19**	5.96	0.837	0.546	1.11
	МГУА	2.00	1.44	5.46	0.969	0.247	0.925
	ИНС	1.71	1.28	4.95	0.828	**0.561**	1.22

Примечание. Жирным шрифтом выделены "наилучшие" с точки зрения эффективности значения критериев.

Резерв повышения надежности моделей прогнозирования гидроэкологических показателей видится нам в объединении отдельных моделей в "коллектив", суммарная эффективность которого практически всегда оказывается значительно выше любого из его членов. Структурные связи в коллективе выбираются таким образом, чтобы положительные свойства той или иной модели (метода) дополняли друг друга, а отрицательные – компенсировались [Растригин, Эренштейн, 1981]. Системы коллективного распознавания и прогнозирования значительно более устойчивы к не вполне объяснимым "провалам", которые свойственны отдельным индивидуальным методам (см., например, модель МГУА для $БПК_5$ в табл. 9.9). Конкретные методы и примеры объединения отдельных прогнозов в работоспособный коллектив были описаны ранее [Брусиловский, Розенберг, 1983; Брусиловский, 1987; Розенберг с соавт., 1994].

9.6. Обучение без учителя: нейросети Кохонена

Формулировка задачи

Пусть имеется таблица произвольных гидробиологических наблюдений $\boldsymbol{X}$ размерностью $\boldsymbol{m} > 1$.

Рассмотренный в предыдущих разделах алгоритм обучения нейронной сети с помощью процедуры обратного распространения подразумевал наличие некоего внешнего классификатора (как правило, человека), предоставляющего сети как входные, так и целевые выходные образы. Алгоритмы, пользующиеся подобной концепцией, называются *алгоритмами обучения с учителем*.

В то же время вся глава 7 была посвящена задаче кластерного анализа – *алгоритмам без учителя* – разбиению множества объектов на заданное или неизвестное число классов на основании некоторого априорного математического критерия качества классификации, отражающего в той или иной мере следующие неформальные требования:

- внутри групп объекты должны быть тесно связаны между собой;
- объекты разных групп должны быть далеки друг от друга;
- при прочих равных условиях распределение объектов по группам должно быть равномерным.

Главная черта, делающая привлекательным обучение без учителя, – это его "самостоятельность", а возможная область применения состоит в обнаружении новых явлений. Естественным оказалось проникновение нейросетевых методов и в эту область моделирования.

Рекомендуемая литература: [Кохонен, 1982; Kohonen, 1982; Уоссермен, 1992; Нейронные сети.., 2001]

Математический лист

Самоорганизующиеся карты (Self Organizing Maps – SOM), разработанные Т. Кохоненом [Kohonen, 1982], представляют собой мощный аналитический инструмент, объединяющий в себе две основные парадигмы анализа – кластеризациию и проецирование, т.е. визуализацию многомерных данных на плоскости. Сеть Кохонена распознает кластеры в многомерных обучающих данных и относит все данные к тем или иным кластерам, используя алгоритм проецирования с сохранением топологического подобия. При этом те элементы выборки, которые находятся в относительной близости в исходном многомерном пространстве, оказываются рядом и в пространстве с более низкой размерностью.

Разумеется, при любой попытке свернуть информацию об объекте из $\boldsymbol{m}$-мерного пространства в точку на плоскости могут быть потеряны некоторые детали, однако, такой прием часто бывает полезен, так как он позволяет визуализировать данные, которые никаким иным способом проанализировать невозможно. Если, например, сеть встретится с набором данных, не похожим ни на один из известных образцов, то она не сможет классифицировать такое наблюдение и тем самым выявит его новизну.

Сеть Кохонена имеет всего два слоя: *входной* и *выходной*, составленный из радиальных нейронов упорядоченной структуры (выходной слой называют также *слоем топологической карты*). Нейроны выходного слоя располагаются в узлах двумерной сетки с прямоугольными или шестиугольными ячейками. Количество нейронов в сетке определяет степень детализации результата работы алгоритма, и, в конечном счете, от этого зависит точность обобщающей способности карты.

Процесс обучения, как и в случае обучения с учителем, заключается в подстраивании весов синапсов методом последовательных приближений на основании их значений от предыдущей итерации. Обучение по алгоритму Кохонена сводится к минимизации разницы между входными сигналами нейрона, поступающими с выходов нейронов предыдущего слоя $\boldsymbol{y_i^{(n-1)}}$, и весовыми коэффициентами его синапсов:

$$\boldsymbol{w_{ij}(t) = w_{ij}(t-1) + \alpha \cdot \left[y_i^{(n-1)} - w_{ij}(t-1) \right]} \ , \qquad (9.23)$$

где $\boldsymbol{t}$ обозначает номер эпохи (итерации).

Полный алгоритм обучения будет выглядеть так.

1. На стадии инициализации всем весовым коэффициентам присваиваются небольшие случайные значения.
2. На входы сети подается входной образ, и сигналы возбуждения распространяются в выходном слое согласно принципам классических прямопоточных сетей, то есть для каждого нейрона рассчитывается взвешенная сумма его входов, к которой затем применяется активационная (передаточная) функция нейрона, в результате чего получается его выходное значение $y_i^{(n)}$, i=0,...,M_i-1, где M_i – число нейронов в слое i; n=0,...,N - 1, а N – число слоев в сети.
3. Из всего выходного слоя выбирается нейрон, значения синапсов которого максимально походят входному образу, и для него осуществляется подстройка весов синапсов с применением формулы (9.23). Эта, так называемая, аккредитация может сопровождаться "затормаживанием" всех остальных нейронов слоя и введением выбранного нейрона в насыщение. Иными словами, в ходе обучения модифицируется не только нейрон-"победитель", но, в меньшей степени, и его соседи.
4. Цикл повторяется с шага 2, где попеременно предъявляются все образы из входного набора пока выходные значения сети не будут стабилизированы с заданной точностью.

Оценка выигравшего нейрона на шаге 3 может осуществляться с использованием любого алгоритма k-ближайших соседей (например, путем расчета скалярных произведений векторов весовых коэффициентов с вектором входных значений и максимальное произведение будет указывать на выигравший нейрон).

В результате итеративной процедуры обучения сеть организуется таким образом, что каждому входному измерению, заданному в m-мерном пространстве исходных признаков, будет соответствовать ячейка-"победитель" на двумерной решетке топологического слоя сети. Для визуализации структуры кластеров, полученных в результате обучения карты, применяется унифицированная матрица расстояний. Элементы матрицы определяют расстояние между весовыми коэффициентами каждого нейрона и его ближайшими соседями. Большое значение расстояния говорит о том, что данный нейрон сильно отличается от окружающих и относится к другому классу.

Основная трудность применения сетей Кохонена, как и в случае факторного анализа, заключается в смысловой интерпретации топологической карты и увязывании ее отдельных участков с некоторыми конкретными обобщениями из предметной области.

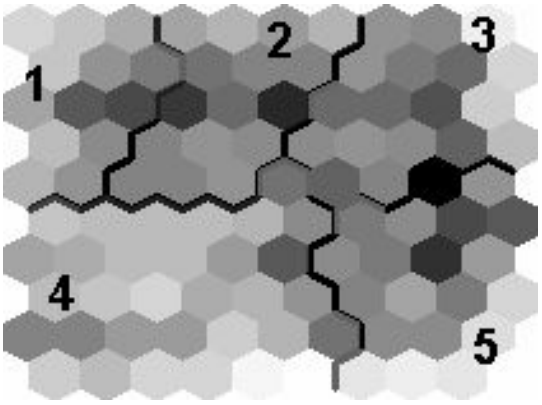
Результаты расчетов

Сформируем выборку из 88 наблюдений, выполненных на 15 станциях р. Сок. В качестве конкретных признаков, описывающих эти измерения, используем показатели обилия по 6 основным таксономическим группам хирономид (отдельно по подсемействам Orthocladiinae, Tanypodinae, Diamesinae, Prodiamesinae и трибам Chironomini и Tanytarsini), а также индексы Шеннона, Вудивисса и Пареле.

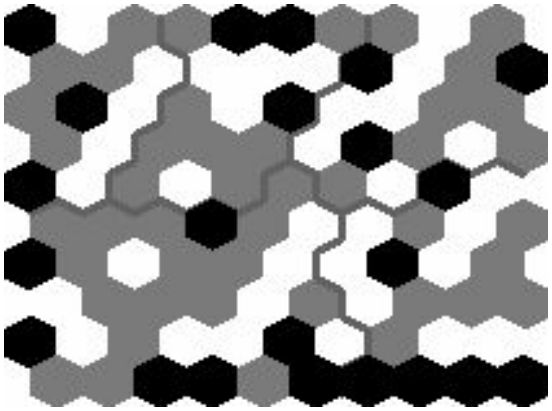
Выполним обучение самоорганизующейся сети с выходным топологическим слоем 10x10 ячеек и представим на рис. 9.13 серию карт Кохонена для рассматриваемого примера. Каждая карта представляет собой отображение выходного слоя нейронов, расположенных в узлах двумерной координатной сетки с прямоугольными или шестиугольными ячейками (шестиугольники дают более корректные результаты, т.к. расстояние между центрами ячеек ближе к евклидову, чем между центрами прямоугольников). Для визуализации карт будем применять градации серого цвета, т.е. чем больше значение отображаемого показателя, тем темнее прорисовывается связанный с ним узел. Полученный набор раскрасок может использоваться для анализа закономерностей, имеющихся между компонентами набора данных.

Представленные на рис. 9.13 карты Кохонена могут быть интерпретированы следующим образом.

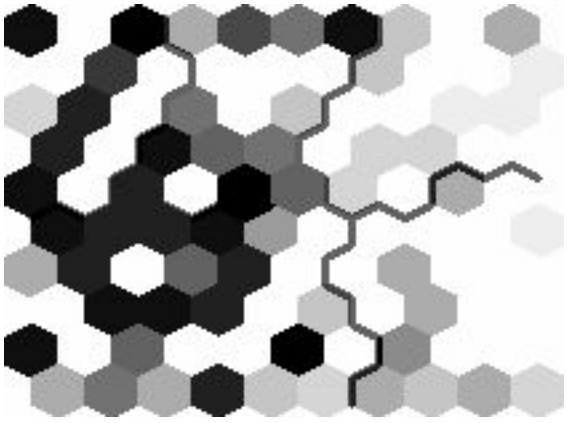
1. Карта а) описывает унифицированную матрицу расстояний между каждым нейроном и его ближайшими соседями. Узлам, резко контрастирующим со своей окрестностью, соответствует черный цвет, а участкам, носящих характер "сглаженного плато", – белый. Группу ячеек, расстояние между которыми внутри этой группы меньше, чем расстояние до соседних групп, определим как кластер. В качестве примера используем разбиение топологической карты на 5 кластеров.



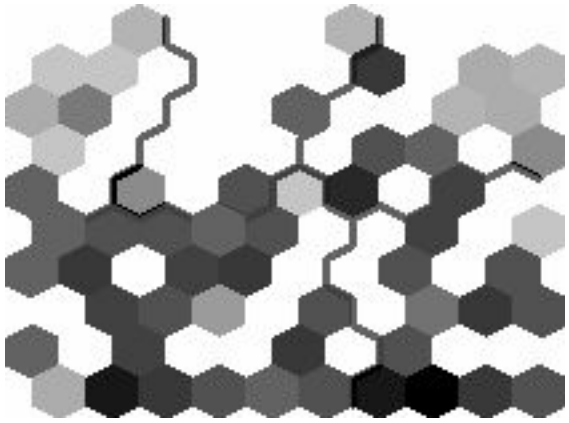
Карта а). Области кластеров



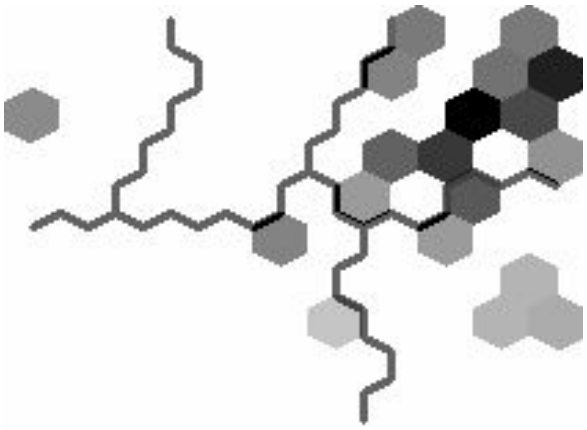
Карта б). Частоты выигрышей



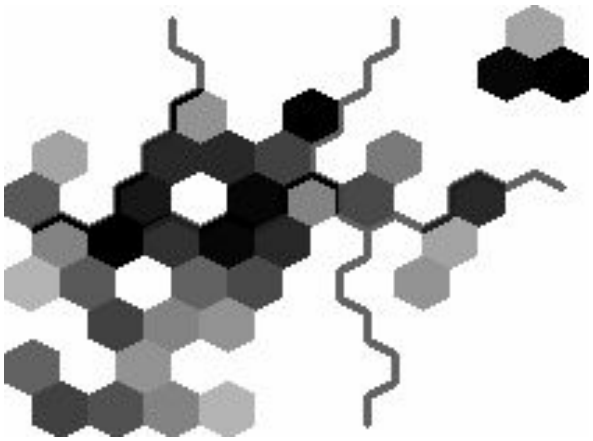
Карта в). По станциям



Карта г). Индекс Вудивисса



Карта д). Обилие Prodiamesinae



Карта е). Обилие Tanypodinae

Рис. 9.13. Топологические карты Кохонена для комплекса наблюдений, сделанных на р. Сок

2. С каждым произвольным измерением обучающей или контрольной выборок связывается "нейрон-победитель", т.е. нейрон выходного слоя, имеющий максимальную близость в смысле некоторой функции расстояния. На карте б) представлена матрица частот выигрышей, которая показывает, сколько раз каждый элемент выиграл (т.е. оказывался ближайшим к обрабатываемому наблюдению) после тестирования всех 88 примеров обучающей выборки. Если узел выигрывал два раза, то он окрашен в черный цвет. Ввиду того, что количество измерений меньше числа элементов решетки сети, то некоторые узлы оказались незадействованными и окрашены в белый цвет. Большие значения частот выигрышей традиционно указывают на центры кластеров топологической карты.
3. Можно выбрать любую переменную исходной таблицы и отобразить ее в виде карты. Эти карты представляют собой проекции матрицы расстояний на соответствующую компоненту (признак в таблице $\boldsymbol{X}$). На карте в) более темным цветом окрашены узлы, связанные с наблюдениями на станциях, расположенных ближе к устью. Можно отчетливо увидеть, что измерения, сделанные в верховьях, попали в кластеры 3 и 5, в то время, как станции, расположенные ниже по течению, сконцентрировались в кластерах 1 и 4.
4. Темным цветом на карте г) представлены измерения с наибольшим значением индекса Вудивисса. Хотя распределение этого показателя существенно размыто, определенное его превышение можно усмотреть в кластерах 4 и 5.
5. На картах д) и е) можно ясно оценить как изменяется видовой состав индикаторной части хирономидного комплекса: если в кластере 3, который мы связываем с верхним течением, сравнительно велико обилие реофильных представителей подсемейства Prodiamesina, то для 4-го кластера, где больше устьевых измерений, отчетливо увеличивается удельный вес пелофильных видов подсемейства Tanypodinae.

ЗАКЛЮЧЕНИЕ

"Что такое счастье – это каждый понимал по-своему..."[1]

Работая над этой книгой, мы не хотели создавать некий императив, обозначающий законченное руководство к действию, и, тем более, не собирались покушаться на чужое понимание «счастья».

«*Наука есть способ описания окружающей действительности на основе аксиоматических моделей*»[2]. И все научные знания, непротиворечивые в рамках границ объявленных исходных предпосылок, представляют собой абсолютную истину. Таковым может быть любой индекс A, показывающий убедительные результаты на подмножестве рек B региона R.

Но любые выдвинутые аксиомы всегда имеют ограниченную область справедливости, вне которой они ошибочны. Неокончательность аксиоматики в науке есть обязательное для нее явление и ошибочность, связанная с применением утверждений за рамками обозначенных границ, является, наряду с ленью, "двигателем прогресса". Все развитие науки сводится к последовательному приближению путем замены действующих аксиом на более общие.

Поэтому, принимая индексы, предложенные Ф. Вудивиссом, Э.А. Пареле, Е.В. Балушкиной (да и собственные – Т.Д. Зинченко с соавторами) и многими другими, как некоторую самодостаточную данность, мы своей работой лишь хотели обратить внимание наших коллег на то, что идентичные (и, рискуем предположить, более простые и обоснованные) результаты могут быть получены путем *простой подстановки* исходных данных в обычные регрессионные или дискриминантные уравнения (см. главу 8).

«*Частностей должно быть не больше, чем их необходимо*», но оценить, какую дополнительную информационную нагрузку несет новая сформулированная закономерность или индекс, можно только обладая некоторым объемом знаний и приемов, в том числе, в области математической статистики.

"Сотри случайные черты и, ты увидишь, мир прекрасен…"[3]

Мы не хотели в очередной раз подчеркивать принципиальную сложность окружающего экологического мира – многочисленные сетования по этому поводу успели основательно надоесть. Но вряд ли очень парадоксальным будет иное утверждение – модели экологических систем, как средство отображения для наблюдателя свойств окружающей действительности, должны быть *принципиально просты*.

Любое зеркало, отражающее реальный объект, всегда ограничено (например, лишено объемности), но стремится выделить ту конкретную сущность, которая нужна в конкретный момент наблюдателю (например, блеск губной помады модницы). Вспомним хотя бы слова Гамлета: «*Я зеркало поставлю перед Вами, где Вы насквозь увидите себя*». Поэтому при моделировании действительности (суть – конструировании экологической системы) нужно уметь выделить только тот процесс или свойства, которые необходимы для решения ***нужной задачи***.

Приведем не совсем экологический пример – процесс падения шляпы со шкафа, который, при желании, можно трактовать как очень сложный. Например, с позиций квантовой механики это падение – проявление корпускулярно-волнового дуализма материи: любой "частице" с энергией E и импульсом p соответствует волна де Бройля с длиной $\lambda = h/p$ и частотой $\nu = E/h$, где h – постоянная Планка. С позиций гидродинамики это падение подчиняется уравнению Навье-Стокса и аксиоматике механики сплошных сред. Астроном начнет вычислять траекторию падения с учетом мгновенного расположения всех основных космических объектов и конфигурации общего гравитационного поля. И только ученик 8 класса уверенно подставит массу шляпы в простейшее уравнение для свободного падения.

---

[1] Заключительная фраза из повести Аркадия Гайдара "Чук и Гек".

[2] Хазен А. О лженауке, ее последствиях и об ошибках в науке // Наука и жизнь, 2002, № 10. – С. 103-110.

[3] Иоганн В. Гёте "Фауст".

Конечная цель конструирования экологической теории всегда проста и должна быть выражена "одним числом" – исследуемую экосистему нужно отнести к одной из немногих категорий качества ("хорошая", "плохая", "грязная", "неустойчивая", "депрессивная" и т.д.), которая является итогом поставленной задачи. Поэтому, анализируя процесс смены парадигм в экологии [Розенберг, Смелянский, 1997], можно предполагать в текущем столетии обратное движение "маятника" (точнее, новый виток "диалектического штопора") от диатропической познавательной модели С.В. Мейена [1990; Чайковский, 1992] назад к прагматической ньютоновской модели, оснащенной, однако, мощным интеллектуальным придатком отсева "лишних" (малоинформативных в каждом конкретном случае) частностей. Основная задача науки – разработать методы адекватного представления *«простыми средствами сложных свойств экосистем»* (точнее говоря, сложных реалий окружающего мира). Перефразируя Ю. Одума [1975], *«…принцип сводимости свойств целого к сумме свойств его частей должен служить первой рабочей заповедью экологов»*, нужно только выбрать физиономически верные свойства "частей" и выполнить над ними целеобусловленные функциональные преобразования.

Конечная простота модели представления экосистемы предполагает одновременно и ее необходимую внутреннюю сложность, внешне скрытую, по возможности, от внимания исследователя. Прообразом такого решающего автомата является нейросетевая модель (см. главу 9), внутри которой, при ограниченном наборе входных и целевых переменных, заложен сложный, многоуровневый и самонастраивающийся механизм преобразований.

ПРИЛОЖЕНИЕ 1

Л.Ф.Сердюцкая, И.П.Каменева

Модели водных экосистем и их фазовые портреты (на примере модели круговорота азота в Куйбышевском водохранилище)

Мем № 37: *«В свое время прославленный французский писатель Бернарден де Сент-Пьер, "бессмертный" член Академии наук Франции, пожаловался Бонапарту, что в Академии к нему относятся без должного уважения. Бонапарт осведомился, знает ли знаменитый автор "Поля и Вирджинии" дифференциальное исчисление. Получив отрицательный ответ, Бонапарт дал понять, что академик, не знающий дифференциалов, не заслуживает уважения. К концу XX века подобная требовательность никому бы не пришла в голову».* В.В. Дружинин и Д.С. Конторов [1976].

Круговорот биогенных элементов представляет собой ключевой механизм формирования качества воды, однако сложность процессов круговорота приводит к созданию сложных многокомпонентных имитационных моделей, которые оказываются недостоверными из-за неопределенности входящих в модели параметров и возможности получения качественно различных прогнозов при практически одних и тех же параметрах. Отсюда следует крайняя актуальность решения задач достоверного формального описания процессов в экосистемах, в том числе, процессов круговорота азота и фосфора, как одних из основных биогенных элементов, определяющих продуктивность и качество воды водных объектов.

Модель круговорота азота в Куйбышевском водохранилище создана с целью выявления основных причин цветения водоема, а также для прогноза продукционной способности водохранилища, исходя из существующих антропогенных нагрузок.

Для моделей круговорота азота разработана одна из методик верификации – методика связности, существенно уменьшающая неопределенность параметров с помощью выявления связей, наложенных на параметры, из условий сохранения устойчивости нетривиальных особых точек последовательно для нескольких структур круговорота.

Суть этой процедуры состоит в следующем. На начальном этапе составляется система небольшого количества (два–три) уравнения, в которых априори заключена особенность имитируемого процесса и которые легко поддаются теоретическому анализу, в частности, удовлетворяют условиям положительности стационарной точки и устойчивости. Однако эта исходная модель не является достаточно конструктивной, и задача состоит в ее расширении с последовательным введением необходимых фазовых координат при сохранении свойств динамики процесса, заключенных в исходную систему.

Отметим, что верификацией модели нельзя найти однозначно все параметры – некоторые из них всегда будут свободными. Поэтому последние необходимо идентифицировать (назначить численные значения) по реальным данным. Здесь часть параметров моделей идентифицирована по гидрохимическим и гидробиологическим характеристикам Куйбышевского водохранилища, а часть – по другим водоемам. В последнем случае идентифицировались параметры, которые приблизительно равны для круговорота азота в разных условиях. Они выбирались так, чтобы при их варьировании в достаточно широкой области сохранялись устойчивость и качественная картина.

Реальным объектом моделирования является Куйбышевское водохранилище. Методика связности применяется для моделей с постоянными коэффициентами (среднегодовой ход), однако демонстрируется переход и к моделям с переменными параметрами (температура воды, освещенность, поступления органического и неорганического азота в водоем), которые отражают сезонную динамику.

В дальнейшем, если это не будет оговорено, то средняя температура воды (T) водохранилища принимается равной 14 ˚С. Все расчеты ведутся в азотных единицах (тонны азота). Литературные ссылки, обосновывающие основные исходные предположения и численные значения величин, использованных при построении моделей, приведены в [Сердюцкая, Каменева, 2000].

П.1.1. Среднегодовые модели с постоянными коэффициентами

М о д е л и р у е м ы е п р о ц е с с ы

В последующих моделях в качестве динамических переменных используются следующие величины: $\boldsymbol{P}$ – количество фитопланктона в водохранилище; $\boldsymbol{B}$ – количество бактериопланктона; $\boldsymbol{Z}$ – количество зоопланктона; $\boldsymbol{N_0}$ – количество растворенного и детритного органического азота; $\boldsymbol{N_1}$ – количество минерального азота в водохранилище. Последний в свою очередь подразделяется на такие компоненты: NH_4 ($\boldsymbol{N_1}$) – количество аммонийного азота; NO_3 ($\boldsymbol{N_2}$) – количество нитратного азота в водохранилище. Все перечисленные выше величины рассчитаны в тоннах азота.

Предполагается, что весь азот в водохранилище можно представить как неорганический (где $\boldsymbol{N_1}$ – переменная) и органический (где $\boldsymbol{P, B, Z, N_0}$ – переменные). Органический азот, в свою очередь подразделяется на живой (где $\boldsymbol{P, B, Z}$ – переменные) и неживой (где $\boldsymbol{N_0}$ – переменная).

Теперь попытаемся описать процессы, происходящие между этими динамическими переменными.

Физические процессы

Сбросы в водоем органического и неорганического азота в сутки (равны $\boldsymbol{C}$, где $\boldsymbol{C}$ – некоторая константа).

Сток азота через плотину для органического и минерального азота в сутки. Данный процесс линейно зависит от величины переменной.

Биологические процессы

Фиксация свободного азота. Запас связанного азота в водоеме может пополниться путем усвоения молекулярного азота воздуха свободноживущими фиксаторами азота (разными видами *Azotobacter, Clostridium pasteurianum*) и некоторыми сине-зелеными водорослями. Поскольку последние в водоемах распределены очень широко, то фиксация азота из воздуха за счет их развития может достигать в природных водах очень большой величины.

Здесь рассматриваются процессы азотфиксации фито- и бактериопланктона. Для наших модельных условий запишем:

- – азотфиксация фитопланктоном за сутки равна $\boldsymbol{kP}$[1];
- – азотфиксация бактериопланктоном за сутки равна $\boldsymbol{kB}$.

Предполагается также, что эти два процесса линейно зависят от величины переменной.

Усвоение минерального азота. Минеральный азот в воде обычно встречается в очень небольших концентрациях, не превышающих долей миллиграмма на литр. В основном он используется водорослями, в частности, планктонными.

Так как рост фитопланктона зависит от количества питания (минеральных соединений), запишем, пользуясь законом взаимодействия масс, что потребление фитопланктоном неорганического азота за сутки равно $\boldsymbol{kPN_1}$.

Минерализация органического азота. Образовавшиеся в водоеме белковые вещества растений и животных в дальнейшем, после отмирания организмов, подвергаются минерализации с помощью бактерий. Участие бактерий в азотном цикле является необходимым условием круговорота. Азот распадающихся белковых веществ выделяется в первую очередь в виде аммиака. Таким образом, в результате этого процесса водоем пополняется минеральными соединениями и обедняется органическими. Заметим, что потребность фитопланктона в биогенных элементах обеспечивается за счет регенерации в умеренных широтах и составляет 18–25 %. При моделировании этот процесс разбивается на две стадии:

- минерализация органического азота бактериопланктоном за сутки равна $\boldsymbol{kBN_0}$ (по закону взаимодействующих масс);
- поступление неорганических соединений азота в водоем за счет минерализации за сутки равно $\boldsymbol{kB}$ (линейная зависимость).

Денитрификация и другие процессы, связанные с потерей азота в водоеме. Переход азота в газообразную форму из нитратов, т.е. $NO_3^- \rightarrow NO_2^- \rightarrow N_2$, называется денитрификацией. Этот процесс при наличии денитрифицирующих бактерий протекает достаточно энергично лишь тогда, когда в наличии имеются нитраты и достаточное количество легкоусвояемого органического вещества, а также анаэробные условия.

Если в окружающей среде одновременно с NO_2^- присутствуют аммонийные соли или аминокислоты, то свободный азот может в этом случае выделяться за счет их химического взаимодействия:

$$2KNO_2 + (NH_4)_2SO_4 \rightarrow 2NH_4NO_2 + K_2SO_4,$$

$$NH_4NO_2 \rightarrow N_2 + 2H_2O \ .$$

Этот процесс называется *косвенной динитрификацией* .

---

[1] Символ $\boldsymbol{k}$ здесь и в дальнейшем обозначает константу.

Потери азота возможны также в результате деаммонификации, $NH_4 \rightarrow N_2$, а также при образовании летучих окислов азота с участием мочевины, некоторых аминокислот и разнообразных нестойких азотсодержащих соединений.

Будем считать, что на нашем объекте все перечисленные выше процессы в той или иной степени имеют место. Хотя чаще всего денитрификация наблюдается в грунтах, а не в толще воды. Для краткости объединим их в один процесс и будем называть просто денитрификацией, подразумевая наличие и других процессов, связанных с потерей азота в водоеме.

Разделим и этот процесс на две стадии:

- потери неорганического азота в водоеме за счет денитрификации за сутки равны $\boldsymbol{kBN_1}$ (по закону взаимодействующих масс);
- переход азота в газообразную форму (выход из системы) в сутки равен $\boldsymbol{kB}$ (линейная зависимость).

Выделения зоопланктона. Роль зоопланктона в круговороте азота и других биогенных элементов подчеркивается во многих работах.

Экспериментальным путем показано, что взятые для опыта животные-детритофаги (пресноводные планктонные рачки-фильтраты, дафнии, пресноводные моллюски лужанки и насекомые жучки-кожееды) обеспечивают достаточно высокий уровень поступления биогенных элементов, доступных для усвоения звеном первичных продуцентов. Так, в эвтрофных водоемах (чем является и наш объект) выделения аммиака зоопланктоном почти на 56 % обеспечивают ежегодные потребности фитопланктона в азоте. Все это свидетельствует о необходимости учета этого процесса в круговороте азота.

Таким образом, для модели выделения зооплактона за сутки равны $\boldsymbol{kZ}$ (линейная зависимость).

Смертность. При моделировании учитывается также и смертность живых компонентов ($\boldsymbol{P, Z, B}$), т.е. переход в мертвую органику ($\boldsymbol{N_0}$):

- отмирание фитоплактона за сутки равно $\boldsymbol{kP}$ (линейная зависимость);
- отмирание бактериопланктона за сутки равно $\boldsymbol{kB}$ (линейная зависимость);
- отмирание зоопланктона за сутки равно $\boldsymbol{kZ + k_1Z}$ (линейная зависимость), где под $\boldsymbol{k_1Z}$ подразумевается выедание зооплактона хищниками более высокого трофического уровня, другими словами, выход из системы.

Питание зоопланктона. Считаем, что зоопланктон питается фито-, бактериопланктоном и детритным азотом. Отсюда, учитывая закон взаимодействующих масс, получим:

- выедание зоопланктоном фитопланктона за сутки равно $\boldsymbol{kZP}$;
- выедание зоопланктоном бактериопланктона за сутки равно $\boldsymbol{kZB}$;
- потребление зоопланктоном детритного азота за сутки равно $\boldsymbol{kZN_0}$.

Такие процессы, как обмен биогенами толщи воды с дном, фотохимическое окисление при моделировании круговорота азота не учитывались ввиду отсутствия натурных данных по конкретному водоему.

Модель из трех уравнений

Рассмотрим простейший случай круговорота биогенного элемента в водоеме, где в качестве фазовых переменных приняты следующие величины: $\boldsymbol{P}$ – количество фитопланктона; $\boldsymbol{N_0}$ – количество органического азота; $\boldsymbol{N_1}$ – количество неорганического азота в водоеме.

Блок-схема круговорота азота в Куйбышевском водохранилище (три фазовые переменные) представлена на рис. П.1. На ее основе составлена модель, состоящая из трех обыкновенных дифференциальных уравнений:

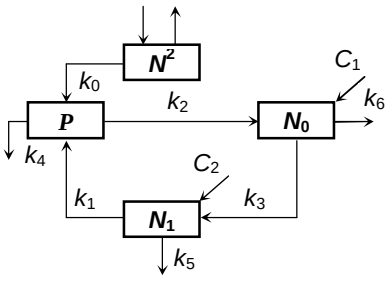


Рис. П.1

$$\begin{aligned} \dot{\boldsymbol{P}} &= \boldsymbol{k_0 P} + \boldsymbol{k_1 T N_1 P} - \boldsymbol{k_2 P} - \boldsymbol{k_4 P}, \\ \dot{\boldsymbol{N}}_0 &= \boldsymbol{C_2} + \boldsymbol{k_2 P} - \boldsymbol{k_3 T N_0} - \boldsymbol{k_6 N_0}, \\ \dot{\boldsymbol{N}}_1 &= \boldsymbol{C_1} + \boldsymbol{k_3 T N_0} - \boldsymbol{k_1 T N_1 P} - \boldsymbol{k_5 N_1}, \end{aligned} \qquad (\text{П.1})$$

где $\boldsymbol{k_0}$ – коэффициент азотфиксации фитопланктона;
$\boldsymbol{k_1}$ – коэффициент потребления фитопланктоном неорганического азота;
$\boldsymbol{k_2}$ – смертность фитопланкона;
$\boldsymbol{k_3}$ – коэффициент трансформации органического азота;
$\boldsymbol{k_4}$ – выедание фитопланкона зоопланктоном;
$\boldsymbol{k_5}$ – сток неорганического азота через плотину;
$\boldsymbol{k_6}$ – сток органического азота через плотину; $\boldsymbol{C_1}$ – поступление неорганического азота в водоем; $\boldsymbol{C_2}$ – поступление органического азота в водоем.

Нахождение реальной стационарной точки модели. Объем Куйбышевского водохранилища равен 58 км^3. Средняя концентрация неорганического азота ($\boldsymbol{N_1}$) ~ 0,4 мг/л или в переводе на все водохрани-

лище $\boldsymbol{N_1}$ = 23200 т. Для органического азота средняя концентрация ~ 0,657 мг/л, т.е. в пересчете на весь водоем $\boldsymbol{N_0}$ = 38106 т.

Теперь переведем биомассу фитопланктона $\boldsymbol{P}$ в азотные единицы. Средняя концентрация фитопланктона ~ 9.334 мг/л. В переводе на сухой вес эта цифра принимает значение 2,3335 мг/л. Общий азот составляет в среднем 9 % от сухого веса, т.е. 0,21 мг/л (азота), или для всего водохранилища P = 12181 т.

Таким образом, получаем стационарную точку модели (П.1) (в тоннах азота):

$$\boldsymbol{P}^0 = 12181; \quad \boldsymbol{N}_0^0 = 38106; \quad \boldsymbol{N}_1^0 = 23200 \qquad \text{(П.2)}$$

Поиск коэффициентов. За год в Куйбышевское водохранилище поступает 426900 т общего азота, т.е.

$$C_1 = 1185{,}8 - C_2 \text{ (т/сут).} \qquad \text{(П.3)}$$

За год сток общего азота – 385900 т, т.е. $\boldsymbol{k}_6\boldsymbol{N}_0^0 + \boldsymbol{k}_5\boldsymbol{N}_1^0 = 385900$ т или $\boldsymbol{k}_6\boldsymbol{N}_0^0 + \boldsymbol{k}_5\boldsymbol{N}_1^0 =$ 1071,944 т/сут, отсюда

$$\boldsymbol{k_5} = 0{,}046 - 1{,}6425\boldsymbol{k}_6 \text{ (сут}^{-1}\text{).} \qquad \text{(П.4)}$$

В Куйбышевском водохранилище за год аккумулируется 41000 т азота, т.е. 113 т/сут. Следовательно, $\boldsymbol{k}_1\boldsymbol{T}\boldsymbol{N}_1^0\boldsymbol{P} = 113$, отсюда

$$\boldsymbol{k_1} = 2{,}856 \cdot 10^{-8} \text{ (С·т·сут)}^{-1}. \qquad \text{(П.5)}$$

Коэффициент прижизненных выделений и смертность фитопланктона составляет

$$\boldsymbol{k_2} = 0{,}35/360 = 0{,}972 \cdot 10^{-3} \text{ сут}^{-1}.. \qquad \text{(П.6)}$$

Коэффициент азотфиксации фитопланктона $\boldsymbol{k_0}$ = 0,9458 год$^{-1}$ или $2{,}63 \cdot 10^{-3}$ сут$^{-1}$.

Оставшиеся коэффициенты $\boldsymbol{C_2}$, $\boldsymbol{k_6}$, $\boldsymbol{k_4}$, $\boldsymbol{k_3}$ верифицируем по условию стационарности (П.2) модели (П.1), используя соотношениями (П.3)–(П.6):

$$\boldsymbol{k_4} = 0{,}0109 \text{ сут}^{-1}, \qquad \text{(П.7)}$$

$$\boldsymbol{k_3} = \frac{\boldsymbol{C}_2 + 11{,}84 - 38106\boldsymbol{k}_6}{533484}. \qquad \text{(П.8)}$$

Относительно двух коэффициентов $\boldsymbol{k_6}$ и $\boldsymbol{C_2}$ сделаем два предположения: $\boldsymbol{k_6} = \boldsymbol{k_4} - \boldsymbol{k_0} = 8{,}304 \cdot 10^{-3}$ сут$^{-1}$, тогда из (П.4) получим k_5 = 0,0324 сут$^{-1}$; $C_1 = C_2$, тогда C_2 = 592,9 т/сут, а по (П.8) получим $k_3 = 5{,}404 \cdot 10^{-4}$ (С·т·сут)$^{-1}$.

Из третьего уравнения системы (П.1) (стационарный вариант), уточняя коэффициент C_1 и используя (П.5)–(П.8), получаем C_1 = 576,4 т/сут.

Список коэффициентов:

$\boldsymbol{C_1}$ = 576,4 т/сут;
$\boldsymbol{C_2}$ = 592,9 т/сут;
$\boldsymbol{k_0} = 2{,}63 \cdot 10^{-3}$ сут$^{-1}$;
$\boldsymbol{k_1} = 2{,}856 \cdot 10^{-8}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_2} = 0{,}972 \cdot 10^{-3}$ сут$^{-1}$;
$\boldsymbol{k_3} = 5{,}404 \cdot 10^{-4}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_4}$ = 0,0109 сут$^{-1}$;
$\boldsymbol{k_5}$ = 0,0324 сут$^{-1}$;
$\boldsymbol{k_6} = 8{,}304 \cdot 10^{-3}$ сут$^{-1}$.

Исследование на устойчивость. Исследуя на устойчивость стационарную точку модели (П.1) (приравнивая к нулю ее правые части), получаем следующие необходимые условия устойчивости:

$$\begin{aligned}
&1)\ \boldsymbol{k}_4 > \boldsymbol{k}_0;\\
&2)\ \boldsymbol{M} = (\boldsymbol{C}_1 + \boldsymbol{C}_2)\boldsymbol{k}_1\boldsymbol{k}_2\boldsymbol{T} + \boldsymbol{k}_1(\boldsymbol{k}_4 - \boldsymbol{k}_0)\boldsymbol{T}\boldsymbol{C}_2 - \boldsymbol{k}_2\boldsymbol{k}_5(\boldsymbol{k}_2 + \boldsymbol{k}_4 - \boldsymbol{k}_0) > 0;\\
&3)\ (\boldsymbol{k}_1\boldsymbol{T} + \boldsymbol{k}_6 + \boldsymbol{k}_1\boldsymbol{T}\boldsymbol{P}^0 + \boldsymbol{k}_5)[(\boldsymbol{k}_1\boldsymbol{T}\boldsymbol{P}^0 + \boldsymbol{k}_5) + (\boldsymbol{k}_1\boldsymbol{T})^2\boldsymbol{P}^0\boldsymbol{N}_1^0] >\\
&\quad > (\boldsymbol{k}_1\boldsymbol{T})^2\boldsymbol{P}^0\boldsymbol{N}_1^0(\boldsymbol{k}_3\boldsymbol{T} + \boldsymbol{k}_6) - \boldsymbol{k}_1\boldsymbol{k}_2\boldsymbol{k}_3(\boldsymbol{T})^2\boldsymbol{P}^0;\\
&4)\ (\boldsymbol{k}_3\boldsymbol{T} + \boldsymbol{k}_6)\boldsymbol{M} > \boldsymbol{C}_2\boldsymbol{k}_1\boldsymbol{T}(\boldsymbol{k}_4\boldsymbol{k}_3\boldsymbol{T} + (\boldsymbol{k}_4 - \boldsymbol{k}_0)\boldsymbol{k}_6 + \boldsymbol{k}_2\boldsymbol{k}_6).
\end{aligned} \qquad \text{(П.9)}$$

Условия (П.9) для численных значений коэффициентов выполняются.

Система (П.1), а также все последующие модели были исследованы на устойчивость численным образом с помощью определителей Гурвица.

Модель из четырех уравнений

Расширим модель (П.1), добавив фазовую координату Z – количество зоопланктона в водоеме, при этом численные значения тех коэффициентов новой модели, которые отражают те же процессы, что и в предыдущей, остаются прежними.

На рис.П.2 представлена блок-схема процесса круговорота азота в Куйбышевском водохранилище (четыре фазовые переменные). На ее основе составлена модель, состоящая из четырех обыкновенных дифференциальных уравнений:

$$\begin{aligned}
\dot{\boldsymbol{P}} &= \boldsymbol{k}_0\boldsymbol{P} + \boldsymbol{k}_1\boldsymbol{T}\boldsymbol{N}_1\boldsymbol{P} - \boldsymbol{k}_2\boldsymbol{P} - \tilde{\boldsymbol{k}}_4\boldsymbol{T}\boldsymbol{P}\boldsymbol{Z},\\
\dot{\boldsymbol{Z}} &= \tilde{\boldsymbol{k}}_4\boldsymbol{T}\boldsymbol{P}\boldsymbol{Z} + \boldsymbol{k}_3\boldsymbol{T}\boldsymbol{N}_0\boldsymbol{Z} - \boldsymbol{k}_7\boldsymbol{Z} - \boldsymbol{k}_{10}\boldsymbol{Z},\\
\dot{\boldsymbol{N}}_0 &= \boldsymbol{C}_2 + \boldsymbol{k}_2\boldsymbol{P} - \boldsymbol{k}_3\boldsymbol{T}\boldsymbol{N}_0 - \boldsymbol{k}_6\boldsymbol{N}_0 - \boldsymbol{k}_9\boldsymbol{T}\boldsymbol{N}_0\boldsymbol{Z} + \boldsymbol{k}_7\boldsymbol{Z},\\
\dot{\boldsymbol{N}}_1 &= \boldsymbol{C}_1 + \boldsymbol{k}_3\boldsymbol{T}\boldsymbol{N}_0 - \boldsymbol{k}_1\boldsymbol{T}\boldsymbol{N}_1\boldsymbol{P} - \boldsymbol{k}_5\boldsymbol{N}_1,
\end{aligned} \qquad \text{П.10}$$

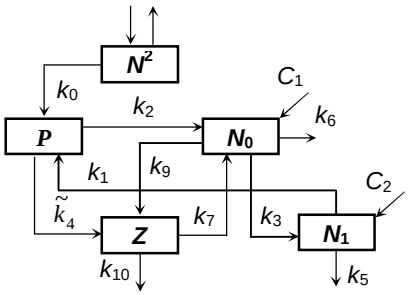


Рис. П.2

где: k_7 – смертность зоопланктона; k_9 – коэффициент потребления зоопланктоном органического азота; k_{10} – коэффициент выедания зоопланктона хищниками более высокого трофического уровня; $\tilde{k}_4$ – измененный коэффициент выедания фитопланктона зоопланктоном.

Нахождение реальной стационарной точки модели. Стационарное значение для зоопланктона обозначим через Z^0. Средняя биомасса зоопланктона составляет 448 мг/м^3. Предположим, что процентное соотношение по азоту в зоопланктоне то же, что и в бактериях, тогда значение средней концентрации зоопластона в азотных единицах – 22,4 мг/м^3. Для всего водохранилища – 1299,2 т.

Таким образом, получаем стационарную точку модели (П.10) (в тоннах азота).:

$$P^0 = 12181;\ Z^0 = 1299{,}2;\ N_0^0 = 38106;\ N_1^0 = 23200 \qquad (П.11)$$

Поиск коэффициентов. Приравнивая к нулю правые части системы (П.9), из первого уравнения, используя (П.11), получаем: $\tilde{k}_4 = 6{,}012\ 10^{-7}$ (С·т·сут)$^{-1}$. В эвтрофном водоеме пища зоопланктона содержит 5–15 % фитопланктона и 10–20 % детрита, причем $k_9 = \tilde{k}_4$. Из третьего уравнения системы получим k_7 = 0,32 сут$^{-1}$. Из второго уравнения – k_{10} = 0,743 сут$^{-1}$.

Список коэффициентов:

C_1 = 576,4 т/сут;
$k_0 = 2{,}63 \cdot 10^{-3}$ сут$^{-1}$;
$k_2 = 0{,}972 \cdot 10^{-3}$ сут$^{-1}$;
$\tilde{k}_4 = 6{,}012 \cdot 10^{-7}$ (ºС·т·сут)$^{-1}$;
$k_6 = 8{,}304 \cdot 10^{-3}$ сут$^{-1}$;
$k_9 = 6{,}012 \cdot 10^{-7}$ (ºС·т·сут)$^{-1}$;
C_2 = 592,9 т/сут;
$k_1 = 2{,}856 \cdot 10^{-8}$ (С·т·сут)$^{-1}$;
$k_3 = 5{,}404 \cdot 10^{-4}$ (С·т·сут)$^{-1}$;
k_5 = 0,0324 сут$^{-1}$;
k_7 = 0,32 сут$^{-1}$;
k_{10} = 0,103 сут$^{-1}$.

Модель из пяти уравнений

Расширим модель (П.10), добавив фазовую координату B – количество бактериопланктона в водоеме, при этом численные значения коэффициентов новой модели, которые отражают те же процессы, что и в предыдущей, остаются прежними.

На рис. П.3 представлена блок-схема процесса круговорота азота в Куйбышевском водохранилище при наличии пяти фазовых переменных, на основе которой составлена модель, состоящая из пяти обыкновенных дифференциальных уравнений:

$$\begin{aligned}
\dot{P} &= (k_0 - k_2)P + k_1 T N_1 P - \tilde{k}_4 TPZ,\\
\dot{Z} &= \tilde{k}_4 TPZ + k_9 T N_0 Z + k_{11} TBZ - (k_{10} + k_{12} + k_7)Z,\\
\dot{B} &= \tilde{k}_3 T N_0 B - k_{11} TBZ + k_{16} T N_1 B - (k_8 + k_{15} + k_{14} - k_{13})B,\\
\dot{N}_0 &= C_2 + k_2 P + k_{15} B + k_7 Z - \tilde{k}_3 T N_0 B - k_9 T N_0 Z - k_6 N_0,\\
\dot{N}_1 &= C_1 + k_{12} Z + k_8 B - k_{16} T N_1 B - k_1 T N_1 P - k_5 N_1,
\end{aligned} \qquad (П.12)$$

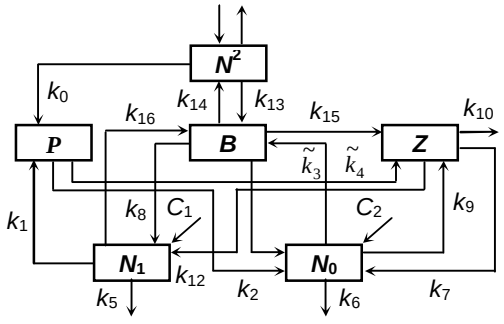


Рис. П.3.

где k_8 – коэффициент выделения бактериопланктоном минерального азота; k_{11} – коэффициент потребления зоопланктоном бактериопланктона; k_{13} – коэффициент азотфиксации бактериопланктона; k_{12} – коэффициент выделения бактериопланктоном молекулярного азота; k_{15} – смертность бактериопланктона; k_{16} – коэффициент денитрификации; $\tilde{k}_3$ – измененный коэффициент трансформации органического азота.

альной стационарной точки модели. Найдем стационарное значение для бактериопланктона, средняя биомасса которого составляет 1,9 мг/л. 10% от сухого веса составляет азот, а 75% – вода, т.е. получим 0,471 мг/л. Для всего водохранилища B = 2731,8 т азота.

Таким образом, получаем (в тоннах азота) стационарную точку модели (П.12):

$\boldsymbol{P}^0 = 12181;\ \boldsymbol{Z}^0 = 1299{,}2;\ \boldsymbol{B} = 2731;\ \boldsymbol{N}_0^0 = 38106;\ \boldsymbol{N}_1^0 = 23200.$

Поиск коэффициентов. Приравняем к нулю правые части системы (П.13). Из второго уравнения получим $\boldsymbol{k_{12}} = 0{,}115$ сут$^{-1}$. По результатам исследований $\boldsymbol{k_{13}} = 0{,}01 \cdot \boldsymbol{k_0} \approx 3{,}63 \cdot 10^{-4}$ сут$^{-1}$. Предположим, $\boldsymbol{k_{15}} = \boldsymbol{k_7} = 0{,}3$ сут$^{-1}$, где $\boldsymbol{k_{15}}$ – смертность бактериопланктона, а $\boldsymbol{k_7}$ – смертность зоопланктона, т.е. $\boldsymbol{k_{15}} = 0{,}3$ сут$^{-1}$. Из четвертого уравнения получим $\boldsymbol{k_{13}} = 7{,}595 \cdot 10^{-7}$ (С·т·сут)$^{-1}$.

Для процесса денитрификации примем $\boldsymbol{k_{16}} = \boldsymbol{k_1} \cdot 10^{-2}$, отсюда $\boldsymbol{k_{16}} = 2{,}856\ 10^{-10}$ (С·т·сут)$^{-1}$. Из пятого уравнения системы (П.12) получим $\boldsymbol{k_8} = 0{,}0509$ сут$^{-1}$. Из третьего уравнения следует: $\boldsymbol{k_{14}} = 6{,}132 \cdot 10^{-5}$ сут$^{-1}$.

Список коэффициентов:

$\boldsymbol{C_1} = 576{,}4$ т/сут;
$\boldsymbol{k_0} = 2{,}63 \cdot 10^{-3}$ сут$^{-1}$;
$\boldsymbol{k_2} = 0{,}972 \cdot 10^{-3}$ сут$^{-1}$;
$\boldsymbol{k_4} = 6{,}012 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_6} = 8{,}304 \cdot 10^{-3}$ сут$^{-1}$;
$\boldsymbol{k_9} = 6{,}012 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_{11}} = 3{,}006 \cdot 10^{-6}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_{14}} = 6{,}132 \cdot 10^{-5}$ сут$^{-1}$;

$\boldsymbol{C_2} = 592{,}9$ т/сут;
$\boldsymbol{k_1} = 2{,}856 \cdot 10^{-8}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_3} = 7{,}595 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$\boldsymbol{k_5} = 0{,}0324$ сут$^{-1}$;
$\boldsymbol{k_7} = 0{,}32$ сут$^{-1}$;
$\boldsymbol{k_{10}} = 0{,}103$ сут$^{-1}$;
$\boldsymbol{k_{12}} = 0{,}115$ сут$^{-1}$;
$\boldsymbol{k_{16}} = 2{,}856 \cdot 10^{-10}$ (С·т·сут)$^{-1}$.

М о д е л ь и з ш е с т и у р а в н е н и й

Расширим модель (П.12), разбив компоненту $\boldsymbol{N_1}$ (минеральный азот) на две составляющие: N_1 – количество аммонийного азота в водоеме и $\boldsymbol{N_2}$ – количество нитратного азота в водохранилище. Численные значения коэффициентов новой модели, которые отражают те же процессы, что и в предыдущей, остаются прежними.

Блок-схема круговорота азота в Куйбышевском водохранилище уже из расчета шести фазовых переменных изображена на рис. П.4. На основе блок-схемы составлена модель, состоящая из шести обыкновенных дифференциальных уравнений:

$$
\begin{aligned}
\dot{\boldsymbol{P}} &= \boldsymbol{k}_0\boldsymbol{P} + \boldsymbol{k}_1^1\boldsymbol{TPN}_1 + \boldsymbol{k}_1^2\boldsymbol{TPN}_2 - \boldsymbol{k}_2\boldsymbol{P} - \tilde{\boldsymbol{k}}_4\boldsymbol{TPZ},\\
\dot{\boldsymbol{Z}} &= \tilde{\boldsymbol{k}}_4\boldsymbol{TPZ} + \boldsymbol{k}_9\boldsymbol{TN}_0\boldsymbol{Z} + \boldsymbol{k}_{11}\boldsymbol{TBZ} - (\boldsymbol{k}_{10} + \boldsymbol{k}_7 + \boldsymbol{k}_{12}^1 + \boldsymbol{k}_{12}^2)\boldsymbol{Z},\\
\dot{\boldsymbol{B}} &= \boldsymbol{k}_{13}\boldsymbol{B} + \boldsymbol{k}_{17}\boldsymbol{TN}_1\boldsymbol{B} + \tilde{\boldsymbol{k}}_{16}\boldsymbol{TN}_2\boldsymbol{B} + \tilde{\boldsymbol{k}}_3\boldsymbol{TN}_0\boldsymbol{B} - \boldsymbol{k}_{11}\boldsymbol{TBZ} - (\tilde{\boldsymbol{k}}_8 + \boldsymbol{k}_{18} + \boldsymbol{k}_{15} + \boldsymbol{k}_{14})\boldsymbol{B},\\
\dot{\boldsymbol{N}}_0 &= \boldsymbol{C}_2 + \boldsymbol{k}_2\boldsymbol{P} - \boldsymbol{k}_{15}\boldsymbol{B} - \boldsymbol{k}_7\boldsymbol{Z} - \tilde{\boldsymbol{k}}_3\boldsymbol{TN}_0\boldsymbol{B} - \boldsymbol{k}_9\boldsymbol{TN}_0\boldsymbol{Z} - \boldsymbol{k}_6\boldsymbol{N}_0,\\
\dot{\boldsymbol{N}}_1 &= \boldsymbol{C}_1^1 + \boldsymbol{k}_{12}^1\boldsymbol{Z} + \tilde{\boldsymbol{k}}_8\boldsymbol{B} - \boldsymbol{k}_{17}\boldsymbol{TN}_1\boldsymbol{B} - \boldsymbol{k}_1^1\boldsymbol{TPN}_1 - \boldsymbol{k}_5^1\boldsymbol{N}_1,\\
\dot{\boldsymbol{N}}_2 &= \boldsymbol{C}_1^2 + \boldsymbol{k}_{12}^2\boldsymbol{Z} + \boldsymbol{k}_{18}\boldsymbol{B} - \boldsymbol{k}_{16}\boldsymbol{TN}_2\boldsymbol{B} - \boldsymbol{k}_1^2\boldsymbol{TPN}_2 - \boldsymbol{k}_5^2\boldsymbol{N}_2,
\end{aligned}
\qquad (П.13)
$$

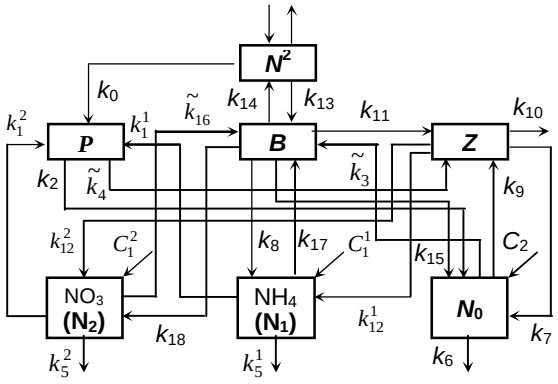


Рис. П.4.

где $\boldsymbol{k}_1^1$ – коэффициент потребления фитопланктоном аммонийного азота; $\boldsymbol{k}_1^2$ – коэффициент потребления фитопланктоном нитратного азота; $\boldsymbol{k}_{12}^1$ – коэффициент выделения зоопланктоном аммонийного азота; $\boldsymbol{k}_{12}^2$ – коэффициент выделения зоопланктоном нитратного азота; $\boldsymbol{k}_5^1$ – коэффициент стока через плотину аммонийного азота; $\boldsymbol{k}_5^2$ – коэффициент стока через плотину нитратного азота; $\boldsymbol{k}_{17}$ – коэффициент нитрификации; $\boldsymbol{k}_{18}$ – коэффициент выделения бактериопланктоном нитратного азота;

$\tilde{\boldsymbol{k}}_8$ – видоизмененный коэффициент выделения бактериопланктоном аммонийного азота; $\tilde{\boldsymbol{k}}_{16}$ – видоизмененный коэффициент денитрификации; $\boldsymbol{C}_1^1$ – поступления в водоем аммонийного азота; $\boldsymbol{C}_1^2$ – поступления в водоем нитратного азота.

Нахождение реальной стационарной точки модели. Найдем стационарное значение для аммонийного и нитратного азота. Средняя концентрация аммонийного азота ~ 0,226 мг/л, а в пересчете на все во-

дохранилище $N_1^0 = 13108$ т. Тогда $N_2^0 = 10092$ т. Таким образом, получаем (в тоннах азота) стационарную точку модели (П.13): $P^0 =$ = 12181; Z^0 = 1299,2; B^0 = 2731,8; N_0^0 = 38106; $N_1^0 = 13108$; $N_2^0 = 10092$.

Поиск коэффициентов. Приравняем к нулю правую часть системы (П.13).

По условию $k_1^1 T N_1^0 P^0 + k_1^2 T N_2^0 P^0 = k_1 T \tilde{N}_1^0 P^0$.

Отсюда $k_1^1 = \dfrac{k_1 \tilde{N}_1^0 - k_2 N_2^0}{N_1}$, где $\tilde{N}_1$ – общий минеральный азот.

Пусть $k_1^2 = \dfrac{1}{2} k_1 = 1{,}428 \cdot 10^{-8}$, тогда $k_1^1 = 3{,}955 \cdot 10^{-8}$ (С·т·сут)$^{-1}$.

Должны выполняться следующие условия:

1. $\tilde{k}_{16} T N_2^0 B^0 = k_{16} T N_1^0 B^0$, откуда $\tilde{k}_{16} = \dfrac{k_{16} N_1^0}{N_2^0} = 6{,}566 \cdot 10^{-10}$ (С·т·сут)$^{-1}$.
2. $C_1^1 + C_1^2 = C_1$.
3. $k_5^1 N_1^0 + k_5^2 N_2^0 = k_5 \tilde{N}_1^0$. Пусть $k_5^1 N_1^0 = k_5^2 N_2^0 = \dfrac{1}{2} k_5 \tilde{N}_1^0$. Тогда $k_5^1 = 0{,}0287$ сут$^{-1}$ и $k_5^2 = 0{,}0372$ сут$^{-1}$.
4. $k_{12}^1 Z^0 + k_{12}^2 Z^0 = k_{12} Z^0$ или $k_{12}^1 + k_{12}^2 = k_{12}$. Предположим: $k_{17} = \tilde{k}_3 \approx 7{,}97 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;

$k_{12}^1 = k_{12}^2 = \dfrac{1}{2} k_{12} = 0{,}0575$ сут$^{-1}$; $C_1^1 = C_1^2 = \dfrac{1}{2} C_1 = 288{,}2$ т/сут (по условию 2);.

Из шестого уравнения системы (П.13) находим k_{16} = 0,0138 сут$^{-1}$. Из третьего уравнения получим $\tilde{k}_8 = 0{,}0138$ сут$^{-1}$. В пятом уравнении уточним $C_1^1 = 289{,}54$ сут$^{-1}$.

Список коэффициентов:

$C_1^1 = 289{,}54$ т/сут;
$C_1^2 = 288{,}2$ т/сут;
$C_2 = 592{,}9$ т/сут;
$k_0 = 2{,}63 \cdot 10^{-3}$ сут$^{-1}$;
$k_1^1 = 3{,}955 \cdot 10^{-8}$ (С·т·сут)$^{-1}$;
$k_1^2 = 1{,}428 \cdot 10^{-8}$ (С·т·сут)$^{-1}$;
$k_2 = 0{,}972 \cdot 10^{-3}$ сут$^{-1}$;
$\tilde{k}_3 = 7{,}595 \cdot 10^{-3}$ (С·т·сут)$^{-1}$;
$\tilde{k}_4 = 6{,}012 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$k_5^1 = 0{,}0287$ сут$^{-1}$;
$k_5^2 = 0{,}0372$ сут$^{-1}$;
$k_6 = 8{,}304 \cdot 10^{-3}$ сут$^{-1}$;
$k_7 = 0{,}32$ сут$^{-1}$;
$\tilde{k}_8 = 0{,}0138$ сут$^{-1}$;
$k_9 = 6{,}012 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$k_{10} = 0{,}103$ сут$^{-1}$;
$k_{11} = 3{,}006 \cdot 10^{-6}$ (С·т·сут)$^{-1}$;
$k_{12} = 0{,}115$ сут$^{-1}$;
$k_{12}^1 = 0{,}0575$ сут$^{-1}$;
$k_{12}^2 = 0{,}0575$ сут$^{-1}$;
$k_{13} = 3{,}63 \cdot 10^{-4}$ сут$^{-1}$;
$k_{15} = 0{,}3$ сут$^{-1}$;
$k_{14} = 6{,}132 \cdot 10^{-5}$ сут$^{-1}$;
$k_{17} = 7{,}97 \cdot 10^{-7}$ (С·т·сут)$^{-1}$;
$\tilde{k}_{16} = 6{,}566 \cdot 10^{-10}$ (С·т·сут)$^{-1}$;
$k_{18} = 0{,}0138$ сут$^{-1}$.

П.1.2. Сезонные модели с переменными коэффициентами

По своей структуре сезонные модели не отличаются от моделей с постоянными коэффициентами: они описывают те же процессы и используют те же численные значения коэффициентов. Существенным отличием является только введение в ряд параметров (в явном виде) времени как аргумента.

В качестве переменных величин в этой части будут рассматриваться следующие параметры: температура воды, освещенность и сбросы в водоем азота. Все остальные коэффициенты будут такими же как и в соответствующих моделях с постоянными коэффициентами.

Модель из трех уравнений

Перепишем модель (П.1) уже при условии переменных параметров:

$$\begin{aligned} \dot{P} &= k_0 I(t) P + k_1 I(t) T(t) N_1 P - (k_2 + k_4) P, \\ \dot{N}_0 &= C_2(t) + k_2 P - k_3 T(t) N_0 - k_6 N_0, \\ \dot{N}_1 &= C_1(t) + k_3 T(t) N_0 - k_1 I(t) T(t) N_1 P - k_5 N_1, \end{aligned} \qquad \text{(П.14)}$$

где $C_1(t)$ и $C_2(t)$ – сбросы в водоем минерального и органического азота (рис. П.5), определены согласно поступлениям азота в Куйбышевское водохранилище, $T(t)$ – температура воды и $I(t)$ – освещенность воды Куйбышевского водохранилища определены соответственно по рис. П.6.

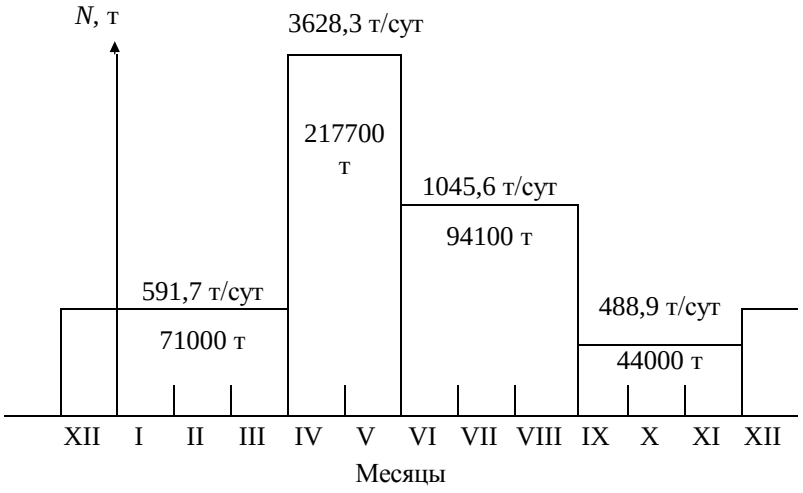
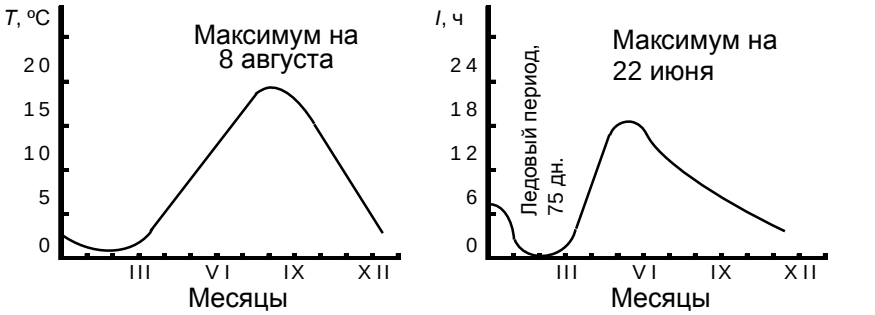


Рис. П.5. Поступление азота в Куйбышевское водохранилище



А) Температура воды Б) Освещенность зеркала

Рис. П.6. Температура и освещенность Куйбышевского водохранилища

Во всех имитационных расчетах временной интервал брался равным 360 дней, а временной шаг – 5 дней.

Аналогичным образом перепишем остальные уравнения (П.10), (П.11) и (П.12):

Модель из четырех уравнений

$$\begin{aligned}
\dot{P} &= k_0 I(t) P + k_1 I(t) T(t) N_1 P - k_2 P - k_4 PZ,\\
\dot{Z} &= k_4 PZ + k_9 N_0 Z - (k_7 + k_{10}) Z,\\
\dot{N}_0 &= C_2(t) + k_2 P - k_3 T(t) N_0 - k_6 N_0 - k_9 N_0 Z + k_7 Z,\\
\dot{N}_1 &= C_1(t) + k_3 T(t) N_0 - k_1 I(t) T(t) N_1 P - k_5 N_1,
\end{aligned} \qquad (П.15)$$

Модель из пяти уравнений

$$\begin{aligned}
\dot{P} &= k_0 I(t) P + k_1 I(t) T(t) N_1 P - k_2 P - k_4 PZ,\\
\dot{Z} &= k_4 PZ + k_9 N_0 Z + k_{11} BZ - (k_{10} + k_{12} + k_7) Z,\\
\dot{B} &= k_{13} I(t) B + k_3 T(t) N_0 B - k_{11} BZ + k_{16} N_1 B - (k_8 + k_{15} + k_{14}) B,\\
\dot{N}_0 &= C_2(t) + k_2 P + k_{15} B - k_3 T(t) N_0 B + k_7 Z - k_9 N_0 Z - k_6 N_0,\\
\dot{N}_1 &= C_1(t) + k_{12} Z + k_8 B - k_{16} N_1 B - k_1 I(t) T(t) N_1 P - k_5 N_1,
\end{aligned} \qquad (П.16)$$

Модель из шести уравнений

$$\begin{aligned}
\dot{P} &= k_0 I(t)P + k_1^1 I(t)T(t)PN_1 + k_1^2 I(t)T(t)PN_2 - k_2 P - k_4 PZ,\\
\dot{Z} &= k_4 PZ + k_9 N_0 Z + k_{11} BZ - (k_{10} + k_7 + k_{12}^1 + k_{12}^2)Z,\\
\dot{B} &= k_{13} I(t)B + k_{17} BN_1 + \tilde{k}_{16} N_2 B + \tilde{k}_3 T(t) N_0 B - k_{11} BZ - (\tilde{k}_8 + k_{18} + k_{15} + k_{14})B,\\
\dot{N}_0 &= C_2(t) + k_2 P + k_{15} B + k_7 Z - \tilde{k}_3 T(t) N_0 B - k_9 N_0 Z - k_6 N_0,\\
\dot{N}_1 &= C_1^1(t) + k_{12}^1 Z + \tilde{k}_8 B - k_{17} N_1 B - k_1^1 I(t)T(t)PN_1 - k_5^1 N_1,\\
\dot{N}_2 &= C_1^2(t) + k_{12}^2 Z + k_{18} B - \tilde{k}_{16} N_2 B - k_1^2 I(t)T(t)PN_2 - k_5^2 N_2.
\end{aligned} \qquad (П.17)$$

Исследование чувствительности моделей и их фазовые портреты

Изучение чувствительности моделей состоит, главным образом, в исследовании особенностей динамики экосистем внутри областей, ограниченных бифуркационными поверхностями, и является дополнительным к качественному исследованию моделей экосистем, в частности, к проблеме устойчивости.

Методы теории чувствительности могут быть использованы для решения двух задач, возникающих при построении динамических многокомпонентных имитационных моделей:

1. на этапе выбора структуры модели и для настройки параметров, обеспечивающих согласование на некотором интервале времени экспериментальных наблюдений и выхода модели. Обычно этот этап построения имитационных моделей сопряжен с большими трудностями, вызванными неопределенностью параметров моделей и связей между компонентами моделей. Как известно, в моделировании экосистем строгие методы для уточнения структуры имитационных моделей отсутствуют. Методы теории чувствительности дают формальный и в определенном смысле оптимальный метод определения поправки, которую надо добавить к предварительно назначенным параметрам модели, а также дают метод поиска существенных параметров модели;
2. на этапе исследования модели уже установленной структуры для отыскания чувствительных компонентов водной экосистемы, определения критических значений параметров при моделировании токсического эффекта, антропогенного эвтрофирования, а также при моделировании синергизма различного типа загрязнений.

Функции чувствительности выявляют те особенности динамики, которые не усматриваются при обычном переборе параметров, а качественные свойства моделей, обнаруженные функциями чувствительности, существенно уменьшают неопределенность моделей обычного типа, обусловленную неточностью исходных параметров (основная трудность имитационного моделирования).

Аппарат теории чувствительности оказывается адекватным сути задачи моделирования антропогенного воздействия на экосистему. Прогноз антропогенного воздействия на экосистему, собственно говоря, и заключается в выявлении тенденции изменения состояния экосистемы.

Если водная экосистема описывается уравнениями состояния

$$\dot{x} = f(x, P, t) \qquad (П.18)$$

при $x(t_0) = \alpha$, где x – вектор состояний; P – вектор параметров размерности m; t – время; $x(t_0)$ – вектор начальных состояний, то дифференциальные уравнения чувствительности первого порядка получаются путем дифференцирования уравнений модели по компонентам P_i:

$$\frac{d}{dt}\left(\frac{\partial x_j}{\partial P_i}\right) = \frac{\partial f_i}{\partial x_k}\frac{\partial x_k}{\partial P_i} + \frac{\partial f_j}{\partial P_i} \qquad (П.19)$$

и по компонентам векторов начальных состояний:

$$\frac{d}{dt}\left(\frac{\partial x_j}{\partial \alpha_i}\right) = \frac{\partial f_j}{\partial x_k}\frac{\partial x_k}{\partial \alpha_i}. \qquad (П.20)$$

Начальные условия для уравнений (П.19) и (П.20) имеют вид (П.16) и (П.17) соответственно:

$$\frac{\partial x_j}{\partial P_i}(t_0) = 0, \qquad (П.21)$$

$$\frac{\partial x_j}{\partial \alpha_i}(t_0) = \delta_{ji}, \quad \delta_{ji} = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases}. \qquad (П.22)$$

Уравнения (П.19) и (П.20) решаются совместно с уравнениями состояния (П.13), а полученные функции чувствительности $\frac{\partial x_j}{\partial P_i}$ и $\frac{\partial x_j}{\partial \alpha_i}$ применяются для изучения влияния антропогенных факторов.

Дифференцирование по параметрам и начальным условиям уравнений (П.14) приводит к дифференциальным уравнениям чувствительности 2-го порядка:

$$\frac{d}{dt}\left(\frac{\partial^2 x_i}{\partial \mu_j \partial \mu_k}\right) = \frac{\partial f_i}{\partial x_a}\frac{\partial^2 x_a}{\partial \mu_j \partial \mu_k} + \left(\frac{\partial^2 f_i}{\partial x_a \partial \mu_k} + \frac{\partial^2 f_i}{\partial x_a \partial x_b}\frac{\partial x_b}{\partial \mu_k}\right)\cdot\frac{\partial x_a}{\partial \mu_j} + \tag{П.23}$$

$$+\frac{\partial^2 f_i}{\partial x_a \partial \mu_j}\frac{\partial x_a}{\partial \mu_k} + \frac{\partial^2 f_i}{\partial \mu_j \partial \mu_k}.$$

При начальном условии $\frac{\partial^2 x_i}{\partial \mu_j \partial \mu_k}(t_0) = 0, \quad \left[\mu = \left(\frac{P}{\alpha}\right)\right].$ (П.24)

Функции чувствительности 2-го порядка $\frac{\partial^2 x_j}{\partial \mu_j \partial \mu_k}$ (П.23), (П.24) используются для изучения синергизма антропогенных воздействий.

Таким образом, функции чувствительности показывают, какие из фазовых координат и в какой момент времени наиболее сильно реагируют на изменения того или иного параметра, а также какова реакция системы при воздействии на несколько параметров одновременно.

На рис. П.7 представлена графическая реализация модели (П.17). Отметим, что динамика общих переменных этой модели совпадает как в количественном, так и в качественном отношении с моделью (П.16).

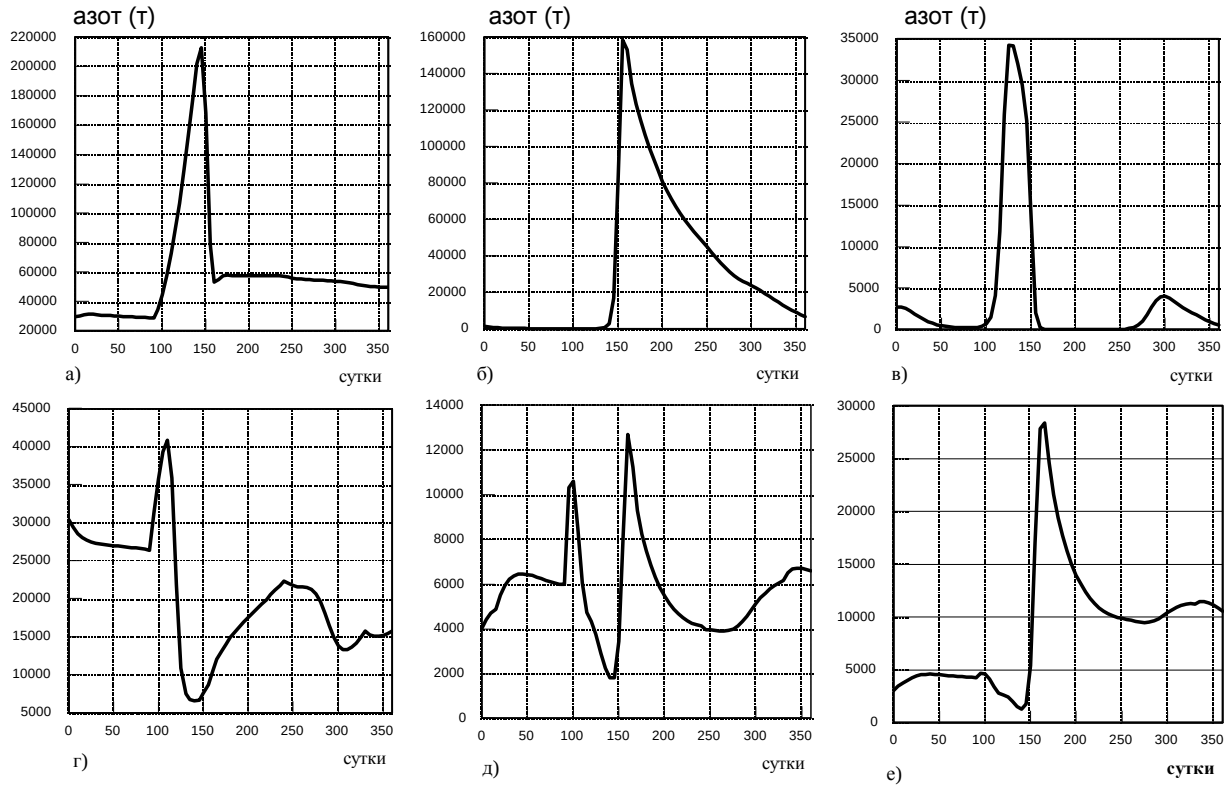


Рис. П.7. Годовая динамика фазовых переменных модели круговорота азота в Куйбышевском водохранилище (6 переменных): а) фитопланктон; б) зоопланктон; в) бактериопланктон; г) органический азот; д) аммонийный азот; е) нитратный азот

На рис. П.8 изображены функции чувствительности 1-го порядка по коэффициенту потребления минерального азота фитопланктоном для системы (П.16). Следует отметить наибольшую реакцию системы на изменение параметра потребления фитопланктоном нитратного азота.

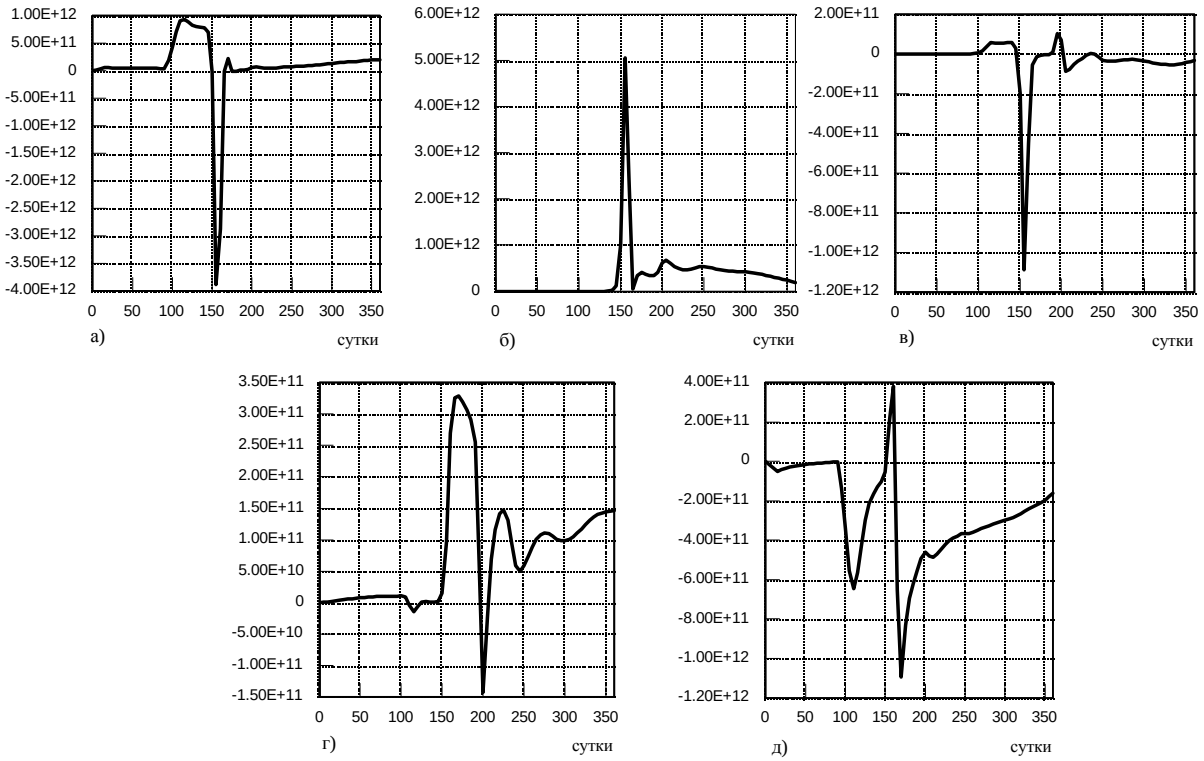


Рис. П.8. Функции чувствительности 1-го порядка фазовых переменных модели круговорота азота в Куйбышевском водохранилище (5 переменных): а) фитопланктон; б) зоопланктон; в) бактериопланктон; г) органический азот; д) минеральный азот

Построенная и синтезированная в результате применения метода связности имитационная модель круговорота азота с переменными параметрами дает возможность исследовать разные аспекты динамики биогенных в водохранилище.

Проведем теперь качественное исследование модели (П.17).

Компонентный анализ, примененный к численным значениям моделируемых переменных показал, что две первые главные компоненты содержат 85 % накопленной дисперсии. В табл.П.1 представлены факторные нагрузки после вращения осей главных компонент по варимаксному критерию.

Таблица П.1

Переменная	*Фактор 1*	*Фактор 2*
Фитопланктон	0,20868	–0,90646*
Бактериопланктон	0,27355	0,88999*
Зоопланктон	–0,91284*	0,06036
Органический азот	0,36445	–0,80868*
Аммонийный азот	–0,75069*	–0,48141
Нитратный азот	–0,97176*	–0,11180

Примечание. Звездочкой (*) помечены значения > 0,7

Интерпретируя факторы, представленные в таблице П.2, заметим, что фактор 1 связан в основном с минеральными формами азота, а фактор 2 – с органическими формами азота и продуцентами. График проекций для 1 года (360 суток, шаг 5 суток) по первым двум главным факторам представлен на рис. П.9.

Годовой фазовый портрет модели (П.17) отражает почти всю информацию, которую несут в себе сложные графики на рис. П.7 – П.8. Так, с марта по июнь возрастает роль фактора 2, четко наблюдается два пика – весенний (больший) и осенний (меньший). "Прогонка" модели (П.17) на 5 лет и последующее приме-

нение вышеизложенного многомерного структурного подхода к качественному исследованию результатов моделирования дает возможность установить устойчивый годовой цикл колебаний обобщенных фазовых координат.

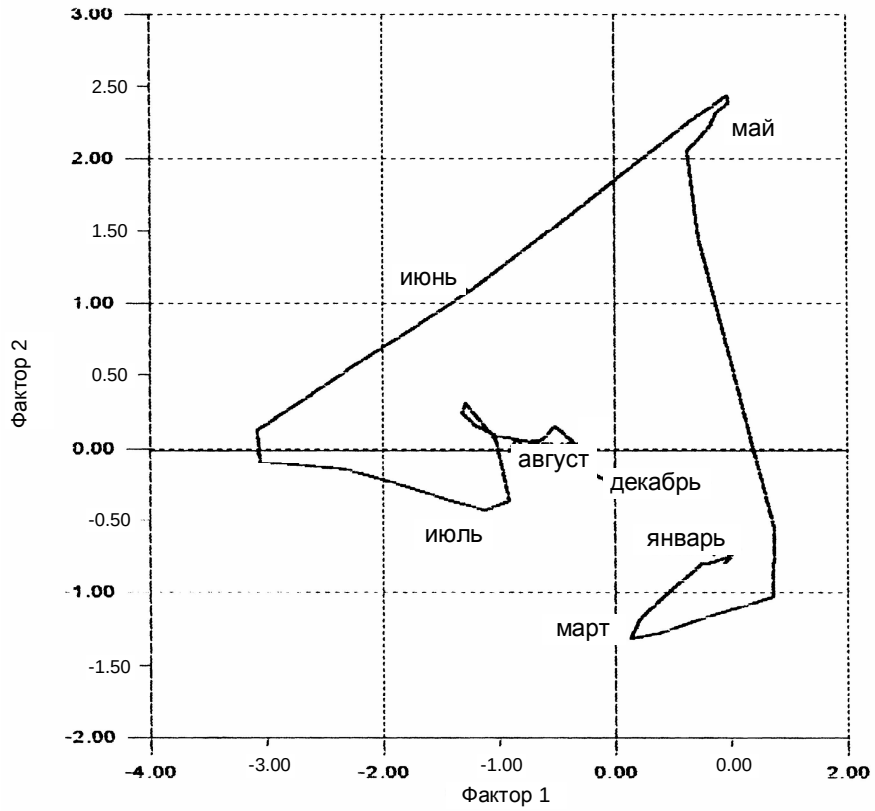


Рис. П.9. Годовой фазовый портрет шестипараметрической модели круговорота азота в пространстве двух главных факторов

ПРИЛОЖЕНИЕ 2

С.В. Крестин

Модель трансформации азотосодержащих веществ Куйбышевского водохранилища

Считается, что самым распространенным биогенным элементом в пресноводных водоемах является фосфор, на втором месте находится азот. Поэтому концентрации азота и фосфора в воде часто являются решающими при прогнозировании "цветения воды". Во многих работах (см., например, [Сердюцкая, 1984; Леонов, 1989; Леонов и др., 1991; Леонов, Цхай, 1995]) рассматриваются математические модели трансформации соединений азота, фосфора и кислорода в водной среде, а так же их взаимодействие с гидробионтами.

Ниже рассматривается еще одна оригинальная модель, опирающаяся на модели А.В. Леонова и описывающая трансформацию соединений азота в водной среде.

Модель включает в себя 17 переменных состояния водной среды C_i (i=1-17): соединения азота (биологические – в биомассах гетеротрофных бактерий ***BN***, фитопланктона ***FN***, простейших ***PRN***, зоопланктона ***ZON***; химические – в форме растворенного органического ***DON***, аммонийного NH_4, нитритного NO_2, нитратного NO_3, детритного ***ND*** [i=1-9, соответственно]), соединения фосфора (биологические – в биомассах гетеротрофных бактерий ***BP***, фитопланктона ***FP***, простейших ***PRP***, зоопланктона ***ZOP***; химические – в форме растворенного органического ***DOP*** и минерального ***DIP***; детритного ***PD*** [i=10-16, соответственно]) и содержание растворенного в воде кислорода O_2 (i=17).

Для оценки условий трансформации биогенных соединений в экосистеме водохранилища используется, так называемое, нульмерное приближение (или приближение полного перемешивания). При этом заведомо пренебрегают пространственной неоднородностью экологических процессов в водохранилище.

Модель экосистемы водохранилища задается уравнением общего вида:

$$d(C_i W)/dt = WR_i + \sum_j Q_j q_{ji} - Q C_i + Y_i , \qquad (П.25)$$

где R_i – скорость биохимической трансформации соответствующего соединения C_i, г/(м^3 сут); W – объем водохранилища, м^3; t – время, сут; Q_j и q_{ji} – расходы реки и боковых притоков и концентрации компонентов в них, м^3/сут и г/м^3, соответственно; Q – расход пропуска из водохранилища, м^3/сут; Y_i – скорости выноса биогенных веществ в водоем со дна и с водосбора (без учета притоков; для O_2 эта переменная описывает реаэрацию и поглощение донными отложениями, г/сут.).

Изменение биомассы гидробионтов характеризуется выражением

$$R_i = (U_i - L_i - S_i) C_i - G_i ,$$

где i = 1-4 и 10-13 для компонентов N и P, соответственно; U_i, L_i, S_i – удельные скорости потребления веществ, выделения продуктов метаболизма и отмирания гидробионтов, соответственно, сут$^{-1}$; G_i – скорость выедания гидробионтов согласно схеме их трофических взаимодействий, сут$^{-1}$ (см., например, [Уморин, 1983].

Для каждой группы гидробионтов C_i задается ассортимент $Pool_i$ взаимозаменяемых соединений N и P:

$$Pool_i = \sum d_{ij} C_j ,$$

где d_{ij} – коэффициенты предпочтения в потреблении субстратов, $\sum d_{ij} = 1$, $0 \le d_{ij} \le 1$.

Общее выражение для удельных скоростей потребления задается уравнениями:

$$U_i = \frac{K_i}{1 + (C_i / Pool_i) + (C_{i+9} / Pool_{i+9})} \quad \text{для } i = 1 - 4,$$

$$U_i = \frac{K_{i-9}}{1 + (C_i / Pool_i) + (C_{i-9} / Pool_{i-9})} \quad \text{для } i = 10 - 13.$$

где K_i – максимальная скорость потребления субстратов сут$^{-1}$, скорректированная по температуре и освещенности.

Скорости потребления U_{ij} гидробионтами C_i отдельных соединений C_j задаются:

$$U_{ij} = \frac{K_i\, d_{ij}\, C_j}{C_i + Pool_i\,(1 + C_{i+9} / Pool_{i+9})} \quad \text{для} \quad i = 1 - 4,$$

$$U_{ij} = \frac{K_{i+9}\, d_{ij}\, C_j}{C_i + Pool_i\,(1 + C_{i-9} / Pool_{i-9})} \quad \text{для} \quad i = 10 - 13.$$

Нетрудно видеть, что удельные скорости потребления U_i представляют собой суммы скоростей потребления отдельных U_{ij} веществ:

$$U_i = \sum_j U_{ij} \ .$$

Скорости метаболических выделений гидробионтов L_i и их смертности S_i составляют определенную долю удельных скоростей потребления и задаются уравнениями:

$$L_i = R_i\, U_i \ ,$$

$$R_i = \alpha_{i1} U_i / (1 + \alpha_{i2} U_i) + 1 - \alpha_{i1} / \alpha_{i2} \ ,$$

$$S_i = V_{i1} + V_{i2}\, C_i / U_i \ ,$$

где α_{i1}, α_{i2}, V_{i1}, V_{i2} – внутренние параметры модели.

Выражения для скоростей трансформации остальных соединений C_i, где i=5-9 и 14-17, имеют вид:

$$R_5 = K_5\, C_9 - U_1\, C_1 + C_3(L_3 - U_{3,5}) + L_2\, C_2 + L_4\, C_4 \ ,$$

$$R_6 = L_3\, C_3 - U_{2,6}\, C_2 - K_7\, C_6 \ ,$$

$$R_7 = K_7\, C_6 - U_{2,7}\, C_2 - K_8\, C_7 \ , \qquad R_8 = K_8\, C_7 - U_{2,8}\, C_2 \ ,$$

$$R_9 = C_3(S_3 - U_{3,9}) + S_2\, C_2 + S_1\, C_1 - K_5\, C_9 + C_4(S_4 - U_{4,9}) \ ,$$

$$R_{14} = K_6\, C_{16} + L_{11}\, C_{11} + L_{13}\, C_{13} - K_9\, C_{14} - U_{10}\, C_{10} + C_{12}(L_{12} - U_{12,14}) \ ,$$

$$R_{15} = K_9\, C_{14} - U_{11}\, C_{11} + L_{10}\, C_{10} \ ,$$

$$R_{16} = C_{12}\,(S_{12} - U_{12,16}) + S_{11}\, C_{11} + S_{10}\, C_{10} - K_6\, C_{16} + C_{13}(S_{13} - U_{13,16}) \ ,$$

$$R_{17} = p_1 U_1 C_1 / (1 + p_2 U_1) - H_1\, K_7\, C_6 - H_2\, K_8\, C_7 - H_3(K_1\, C_1 + L_2\, C_2 + L_3\, C_3 + L_4\, C_4) \ ,$$

где K_i – константы трансформации веществ, H_i – стехиометрические коэффициенты, p_i – константа фотосинтетического процесса.

При описании кислородного режима учитываются температурные зависимости скоростей потребления O_2 при нитрификации и его выделения при фотосинтезе. В расчетах оцениваются затраты на окисление продуктов метаболизма гидробионтов и поглощение O_2 донными отложениями. Использована зависимость коэффициента реаэрации водной среды от скорости ветра, предложенная для водохранилищ [Bahks, 1975, цит. по: Бреховских, 1988] .

Значения максимальной скорости потребления веществ гидробионтами K_{i0} корректируются по температуре и освещенности:

$$K_i = K_{i0}\, R_{Ti}\, R_{Ii} \ ,$$

где i – индекс рассматриваемого гидробионта (изменяется от 1 до 4); K_{i0} – оптимальное значение скорости потребления веществ; R_{Ti} и R_{Ii} – коэффициенты коррекции по температуре и освещенности соответственно.

Для описания температурной коррекции максимальных скоростей потребления веществ использована экспоненциальная функция:

$$R_{Ti} = A_{i0} + \frac{A_{i1}(\exp(A_{i2}T) - 1}{1 + A_{i3}\exp(A_{i2}T)} + \frac{A_{i4}(\exp(A_{i5}T) - 1}{1 + A_{i6}\exp(A_{i5}T)} \ ,$$

где T – температура воды, ^{0}C; A_{ik} – константы.

Температурная коррекция скорости трансформации детритных компонентов N и P, соответственно в DON и DOP, задается в модели линейной зависимостью:

$$K_5 = K_{50}\, T \ , \qquad K_6 = K_{60}\, T \ ,$$

где K_{50} и K_{60} – значения параметров при T = 1 ^{0}C.

Температурная коррекция скоростей трансформации неорганических химических соединений осуществлялась в соответствии с зависимостью

$$K_i = K_{i0}\,\Xi^{(T-20)}\ ,$$

где i = 7-10 при описании химической трансформации NH_4 до NO_2, NO_2 до NO_3, DOP до DIP; Ξ – температурный коэффициент, равный 1.05; K_{i0} – значения параметров при 20 ^{0}C.

Зависимость коэффициента коррекции R_{I2} от условий освещенности имеет вид [Леонов, Цхай, 1995]:

$$R_{I2} = (e / K_l h_0)\ [\ \exp(-r_x) - \exp(-r_1)\ ]\ ,$$

$$r_1 = I / I_0\ , \qquad r_x = r_1\ [\ \exp(-K_l h_0)\ ]\ ,$$

$$K_1 = K_a + K_b C_4\ ,$$

$$I = I_a\ [\ 1 + \cos(2\pi\,(t - t_p) / f\ ]\ ,$$

где t – время суток, I_a – среднесуточная освещенность, f – фотопериод, t_p=12 ч, I_0=350 кал/(см^2 сут); K_a, K_b и h_0 – внутренние параметры модели.

Для бактерий, простейших и зоопланктона коррекция соответствующих коэффициентов по освещенности не проводилась и $R_{I1} = R_{I3} = R_{I4} = 1$.

Величина нагрузки Y_i в общем случае может быть представлена в виде:

$$Y_i = J_i\,\Omega + \omega\,E_i\ ,$$

где ω – длина (в м), Ω – площадь зеркала водохранилища (м^2); J_i – поток вещества через межфазные поверхности, г/(м^2 сут); E_i – путевая нагрузка, связанная с поступлением с берегов, кроме учитываемых притоков, г/(м сут).

Для описания потока оседающего детрита используется следующее выражение:

$$J_i^{out} = -k\,C_i\ , \qquad i=9 \text{ и } 16\ ,$$

где k – скорость осаждения (гидравлическая крупность), м/сут.

В след за А.В. Леоновым [1989] считается, что величина потока взвешенного материала со дна в водную среду прямо пропорциональна скорости ветра V, м/с, и обратно пропорциональна глубине водохранилища H, м:

$$J_i^{in} = b_i\,V / H, \qquad i=9 \text{ и } 16\ ,$$

Значение k = 2.0, $b_9 = 6.4{*}10^{-3}$, b_{16} = 0.1 определены при параметризации модели.

Поток O_2 через свободную поверхность водохранилища представлен в виде

$$J_i^{in} = k_r\,(C_{0i} - C_i)\ , \qquad i=17\ .$$

Здесь коэффициент переноса через границу воздух-вода, имеет вид

$$k_r = 0.728\,V^{1/2} - 0.317V + 0.037V^2\ ,$$

а C_{017} – содержание растворенного кислорода в воде при насыщении, рассчитываемое в зависимости от температуры [Леонов, 1989]:

$$C_{017} = 14.61996 - 0.4042T + 0.00842T^2 - 0.00009T^3.$$

В практике моделирования кислородного режима водных объектов процесс поглощения O_2 донными отложениями обычно воспроизводится в виде реакции нулевого порядка, так как в широком диапазоне концентраций скорость поглощения не зависит от содержания O_2 в воде [Бреховских, 1988; Мизандронцев, 1991]. Поэтому учитывается лишь температурная зависимость соответствующего потока O_2:

$$J_i^{out} = J_{0i}\,(1.09)^{T-2}\ , \qquad i=17\ .$$

Здесь J_{017} – поток O_2 при T=20 ^{0}C, который соответствует конкретному типу донных отложений и характеру окислительно-восстановительных реакций вблизи границы вода-дно. Величина этого потока принимается неизменной и определяется при параметризации модели.

Описанная выше модель достаточно громоздка. Поэтому для иллюстрации проанализируем меньшую по объему модель трансформации только азотосодержащих соединений.

В работах [Леонов, 1989; Леонов, Цхай, 1995] изучаются органический, аммонийный, нитритный, нитратный и дитритный азот, а так же взаимодействие этих форм азота с бактериями, фитопланктоном, простейшими и зоопланктоном.

В нашей модели ради простоты ограничимся рассмотрением взаимодействий только трех видов азотосодержащих соединений (NH_4, NO_2, NO_3) и фитопланктона. Здесь применен камерный подход, состоящий в том, что в водоеме исключается пространственная неоднородность и все компоненты считаются идеально перемешанными. Учтены следующие процессы взаимодействия. Аммоний солевой частично поглощается фитопланктоном, частично перерабатывается в нитриты. Некоторое количество нитритов преобразуется в нитраты. Нитраты поглощаются фитопланктоном. Кроме взаимодействия их друг с другом учитывается

отток и приток воды и связанный с этим перенос компонентов. Все выше сказанное позволяет записать следующую систему дифференциальных уравнений:

$$N'_1 = Q_{ПР} N_1^{ПР}/W - Q_{ОТ} N_1/W + (S + M - (1 - \delta)\mu)\nu B - K_1 N_1$$
$$N'_2 = Q_{ПР} N_2^{ПР}/W - Q_{ОТ} N_2/W + K_1 N_1 - K_2 N_2 \qquad (П.26)$$
$$N'_3 = Q_{ПР} N_3^{ПР}/W - Q_{ОТ} N_3/W + K_2 N_2 - \delta\mu\nu B \ ,$$

где N_1, N_2, N_3 – концентрации, соответственно NH_4, NO_2, NO_3; K_1, K_2 – коэффициенты не консервативности для аммонийного и нитритного азота; S, M, μ – скорости осаждения, смертности и роста фитопланктона, ν – стехиометрический коэффициент, характеризующий содержание азота в единице сухого веса биомассы фитопланктона; δ – доля нитратного азота в потреблении азота фитопланктоном; $Q_{ПР}$, $Q_{ОТ}$ – количество втекающей и вытекающей из камеры воды, W – объём камеры, B – концентрация фитопланктона (см. [Рекомендации по прогнозированию.., 1984]):

$$B = -b/a + (B_0 + b/a)\exp(at/\varepsilon_1) \ ,$$

где:

$$a = \varepsilon_1(\rho+S+M+Q/W) \ , \qquad b = \mu/H - \varepsilon_0(\rho+S+M+Q_{ОТ}/W)+Q_{ПР}\varepsilon_1 B_{ПРФ}/W \ ,$$

ρ, S, M – скорость респирации, осаждения и гибели фитопланктона, соответственно; ε_0 – коэффициент поглощения света водой при отсутствии фитопланктона; ε_1 – коэффициент поглощения света в результате развития фитопланктона; B_0 – начальная концентрация фитопланктона; $B_{ПРФ}$ – концентрация фитопланктона в поступающей в камеру воде; H – средняя глубина камеры; t – время.

Введем обозначения:

$$\alpha = \frac{\mu}{\varepsilon_1 H(\rho S+M+Q_{ОТ}/W)} + \frac{Q_{ПР} B_{ПРФ}}{W(\rho+S+M+Q/W)} \ ,$$
$$\beta = B_0 - \alpha \qquad \gamma = (\rho+S+M+Q/W) \ .$$

Тогда:

$$b/a=-\alpha \ , \qquad B_0 - b/a = B_0 - \alpha = \beta \ , \qquad a/\varepsilon_1 = -\gamma \quad \text{и}$$
$$B = \alpha + \beta\exp(-\gamma t) \ . \qquad (П.27)$$

Решая уравнения системы (П.26) с учетом (П.27), после несложных выкладок получаем:

$$N_1 = A_1 \exp(-a_1 t) + E_1 + F_1 \exp(-\gamma t) \ ,$$

где: $a_1 = K_1 + Q_{ОТ}/W$,

$$A_1 = N^0_1 \frac{Q_{ПР} N_1^{ПР}+W(S+M(1-\delta)\mu)\nu\alpha}{a_1 W} - \frac{(S+M-(1-\delta)\mu)\nu\beta}{a_1 - \gamma} \ ,$$

$$E_1 = \frac{Q_{ПР} N_1^{ПР}+W(S+M(1-\delta)\mu)\nu\alpha}{a_1 W} \ , \qquad F_1 = \frac{(S+M-(1-\delta)\mu)\nu\beta}{a_1 - \gamma} \ ,$$

N^0_1 – начальная концентрация аммонийного азота.

$$N_2 = A_2 \exp(-a_2 t) + E_2 + D_2 \exp(-a_1 t) + F_2 \exp(-\gamma t) \ ,$$

где: $a_2 = K_2 + Q_{ОТ}/W$,

$$A_2 = N_2^0 \frac{Q_{ПР} N_2^{ПР}+K_1 E_1 W}{a_2 W} - \frac{K_1 A_1}{K_2 - K_1} - \frac{K_1 F_1}{K+Q/W-\gamma} \ ,$$

$$E_2 = \frac{Q_{ПР} N_2^{ПР}+K_1 E_1 W}{a_2 W} \ , \qquad D_2 = \frac{K_1 A_1}{K_2 - K_1} \ , \qquad F_2 = \frac{K_1 F_1}{K_2+Q_{ОТ}/W-\gamma} \ ,$$

N_2^0 – начальная концентрация NO_2.

$$N_3 = A_3 \exp(-a_3 t) + E_3 + D_3 \exp(-a_1 t) - A_2 \exp(-a_2 t) + F_3 \exp(-\gamma t) \ ,$$

где: $a_3 = Q_{ОТ}/W$,

$$A_3 = N^0_3 - \frac{Q_{ПР}N_3^{ПР} - K_2WE + \delta\mu\nu\alpha W}{Q_{ОТ}} + A_2 + \frac{K_2D_2}{K_1}\,\frac{K_2F_2 - \delta\mu\nu\beta}{Q_{ОТ}/W - \gamma},$$

$$D_3 = -\frac{K_2D_2}{K_1}, \qquad E_3 = \frac{Q_{ПР}N_3^{ПР} + K_2WE_2 - \delta\mu\nu\alpha W}{Q_{ПР}}, \qquad F_3 = \frac{K_2F_2 - \delta\mu\nu\beta}{Q_{ПР}/W - \gamma},$$

N^0_3 – начальная концентрация NO_3.

Расчеты производились по однокамерной модели с шагом в 10 дней на общем промежутке времени в 1 год. Коэффициенты, при которых производились вычисления, приведены в следующей таблице:

Таблица П..2.

Коэффициенты, при которых производились расчеты

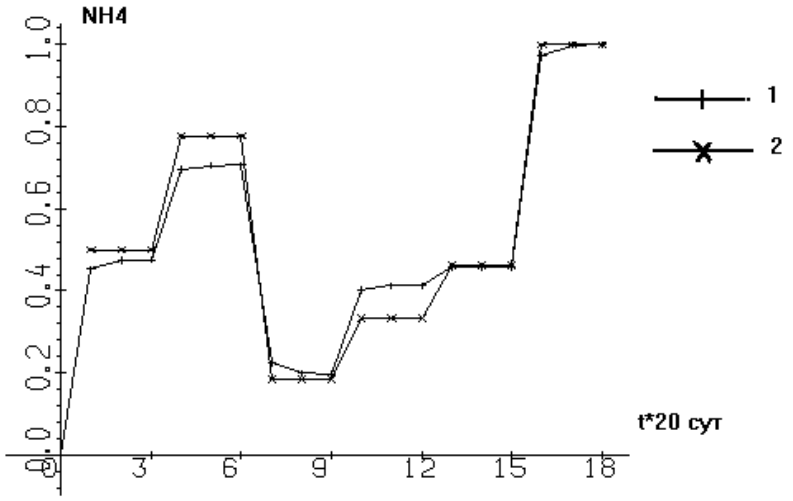
K_1	0.207	H	30.0
K_2	10.0	B_O	0.000305733
P	0.5	N^0_1	0.001
M	0.5	N^0_2	0.00002
S	0.6	N^0_3	0.14
μ	0.02	$Q_{ОТ}$	0.503
ν*	0.08	$Q_{ПР}$	0.503
ε_1	9.0	W	6.642
δ	0.1		

Параметры, отмеченные *, взяты из работы [Рекомендации по прогнозированию.., 1984], остальные получены экспериментально в ходе расчетов.

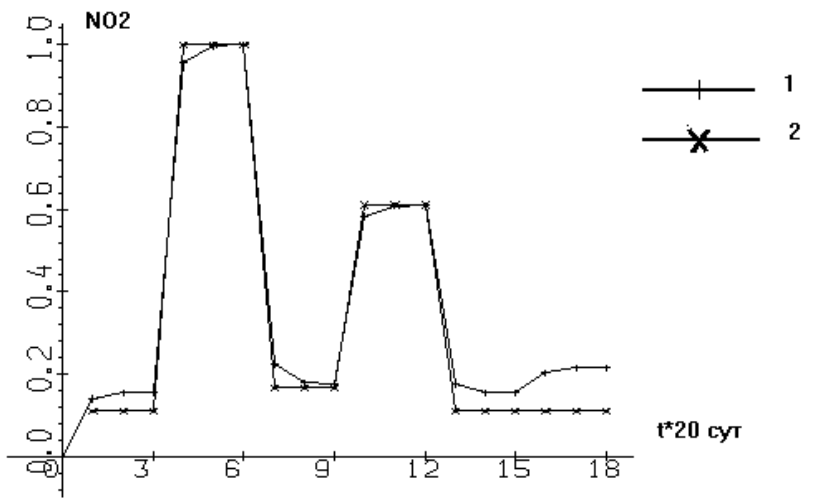
На вход камеры подавались концентрации, взятые из наблюдений, произведенных в 1988 г. на Приплотинном плёсе Куйбышевского водохранилища [Паутова, Номоконова, 1994]. В ходе расчетов выявилось хорошее совпадение данных, взятых из наблюдений с расчетными (см. рис. П.10-П.13). На каждом рисунке оба графика нормированы на единицу, то есть максимальное значение как экспериментальных, так и теоретических данных равно единице.

Графики на этих рисунках состоят из достаточно сложных сочетаний "впадин" и "поднятий", как для кривых наблюдений, так и для расчетных кривых. Сам характер совпадений поведения экспериментальных и теоретических кривых свидетельствует о высокой степени количественной адекватности рассмотренной модели. Сезонная динамика концентрации фитопланктона (рис. П.13), кроме того, демонстрирует качественную картину трех пиков "цветения": весеннего (незначительный), летнего и осеннего.

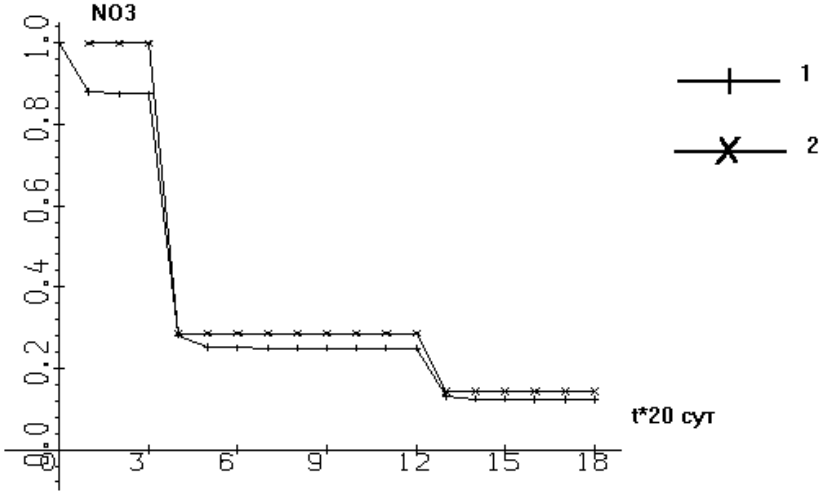
Дальнейшее совершенствование модели может идти, например, в направлении "корректировки" превышения расчетных значений над экспериментальными "поднятиями" и "впадинами".



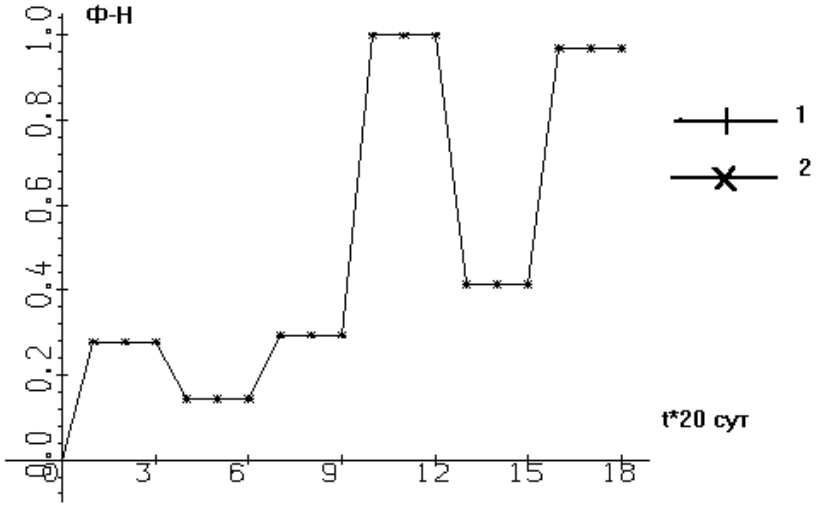
а) Концентрация NH_4



б) Концентрация NO_2



в) Концентрация NO_3



г) Концентрация фитопланктона

Рис. П10. Зависимость нормированных концентраций ингредиентов от времени ***t*** *(1 – результаты моделирования, 2 – наблюдения)*

ВИРТУАЛЬНЫЙ РАЗГОВОР АВТОРОВ С РЕЦЕНЗЕНТАМИ

Мем № 38: *«Поэт: …Ты рьяный чтец, но критик дикий.*
Так я, по-твоему, - великий, повыше Пушкина поэт?
Скажи, пожалуйста?!
Гражданин: Ну, нет!
Твои поэмы бестолковы, твои элегии не новы,
Сатиры чужды красоты, неблагозвучны и обидны,
Твой стих тягуч…» Н.А. Некрасов "Поэт и гражданин"

В процессе подготовки монографии мы ознакомили с ее текстом своих коллег–математиков и биологов, разослав им электронные вложения от Томска до германского Геттингена. Они взяли на себя труд внимательно прочитать рукопись, и при этом высказали немало ценных советов и замечаний. Часть из них мы учли в окончательном варианте, который читатель держит в руках. Но наша дискуссия по отдельным пунктам рецензий показалась нам столь интересной, что мы сочли возможным представить ее в форме виртуальной беседы, "стенографический" отчет о которой и представлен ниже.

Действующие лица:

ВШ, ГР, ТЗ – Владимир Шитиков, к.т.н., Геннадий Розенберг, чл.-корр. РАН, Татьяна Зинченко, к.б.н. (ИЭВБ РАН) – авторы;

НЦ – др. Натан Цейтлин, Институт Макса Планка (Геттинген, Германия);

ВЛ – Василий Леонов, доцент кафедры прикладной информатики Томского государственного университета, редактор электронного журнала "Биометрика";

ДФ – Дмитрий Филимонов, к.ф.-м.н., вед. науч. сотр. лаборатории структурно-функционального конструирования лекарств Института биомедицинской химии РАМН;

NN – анонимные рецензенты журналов эколого-биологического профиля.

ВЛ: Поздравляю вас с хорошей работой, которую по достоинству оценят многие читатели. Ваша книга меня очень порадовала своей глубиной, авторской увлеченностью и незаурядным кругозором авторов! Что в наше время встречается не часто. Однако я не вполне уверен в том, что вы сами ее сейчас оцениваете адекватно. Таких книг сейчас практически нет. Если вы издаете ее малым тиражом и в малоизвестном издательстве, книгу читатель не узнает. Поэтому настоятельно рекомендую не выкладывать ее в свободный доступ в интернете, а искать хорошее издательство. Книгу купят многие. Я обещаю сделать ей хорошую рекламу на своем сайте. Только если будете издавать в хорошем издательстве, почти всю графику надо переделывать. Графика бледная…

ВШ, ГР и ТЗ: Спасибо, Василий, за столь благоприятный отзыв на наш скромный труд.

ВЛ: О слове «интернет». Согласно современным доводам филологов его надо писать с маленькой буквы…

ВШ: Неисповедимы пути русского языка. Скоро, наверное, будут писать с маленькой буквы и windows и билл гейтс.

ВЛ: Говоря «А», вы не всегда развиваете ключевые для существа дела мысли. Например, на стр. 28 вы цитируете Е.П. Воробейника с соавторами [1994], который предлагает критические значения оценивать с использованием толерантных множителей, используемых в математической теории надежности. Определеннее было бы вам рассказать о коэффициенте ***k***, ссылаясь на теорию толерантных интервалов в статистике. Это не то же самое, что доверительные интервалы. Можете посмотреть краткое изложение толерантных интервалов в книге П. Мюллера с соавторами [1982].

ВШ: Мы бы с удовольствием рассказали еще о многом и объяли все необъятные просторы статистики, но помним предостережение Козьмы Пруткова.

НЦ: К списку основополагающей литературы, упомянутому во введении, я бы добавил и другие великолепные книги, которые приятно читать не только биологу, но и любому математику и исследователю: В.Ю. Урбаха [1964], В.В. Александрова и В.С Шнейдерова [1984], Е.В.Гублера и А.А. Генкина [1973].

ВШ: Но тогда нам нужно добавить сюда и Вашу Интернет-монографию "Из опыта аналитического статистика", помещенную на сайте http://matstat.gmxhome.de. Особенную ценность представляет там раздел "Рекомендуемая литература по математической теории эксперимента", содержащий проиндексированный список нескольких сотен книг на руссом языке с Вашими комментариями. Но вспомните эпиграф к этому разделу: *«Не надо читать много книг»* (Мао Дзэ Дун). Я бы вспомнил и подкорректированную фразу из Экклезиаста *«И далее предупреждаю тебя, сын мой, в делании множества книг нет конца и многие их изучения – суть утомления плоти»*. Но, если серьезно, – авторам было очень трудно исчерпывающе охватить всю литературу, небезынтересную для углубления рассматриваемых проблем, как, впрочем, и список самих этих проблем.

ВЛ: Мне очень понравился раздел 2.4. "Математические модели в экологии"**.** Рекомендую его дополнить работами В.В. Налимова, в частности, его книгой "В поисках иных смыслов", в которой он весьма остро ставит вопрос о том, в какой мере современная наука является "научной"? Немало интересного по этой теме и в его последней книге "Разбрасываю мысли". Спасибо за цитирование отдельных фрагментов моих статей. Но, чтобы читателю удобнее было найти первоисточник, лучше бы после каждого такого фрагмента давать точную ссылку.

ВШ: Признаемся, грешны. Если где-то встречали нетривиальную мысль или малоизвестный интересный факт, то старались повзаимствовать. И если где-то опускали точный адрес первоисточника, то только из стремления не перегружать текст "именами, адресами, явками, паролями…". Читателю не всегда интересно, кто первым обратил внимание на тот или иной факт. Кстати, приведу цифры цитируемости фамилий отдельных персоналий по тексту книги: Г.С. Розенберг – 64, Р. Фишер –59, А.Н. Колмогоров – 30, А.Ф. Алимов – 25, В.В. Налимов – 14, В.П. Леонов – 11, В.К. Шитиков – 10. Вы, Василий, находитесь во вполне приличной компании.

ГР: Готов констатировать, что это не плагиат, а хорошая память…

ВЛ: Далее, вы пишете: *«Основной задачей регрессионного анализа является идентификация вида восстанавливаемой функциональной зависимости* $\boldsymbol{Y \approx f(X)}$*»*. Но регрессионное уравнение является статистической или, вернее, стохастической зависимостью. И в этом случае некорректно говорить о функциональной зависимости. Это антиподы – функция и вероятностная зависимость.

ВШ: Признаться, я не знаю, что такое в природе "функциональная зависимость" без предлога "вероятностная". В реальном мире детерминированных функций просто не бывает – даже движение Земли вокруг Солнца подвергается массе возмущений.

ВЛ: Владимир, Вы просто "перегибаете палку"! Вспомните школьную алгебру, вузовскую аналитическую геометрию и т.д. Одно дело говорить о моделях реальных процессов, и совсем иное – в принципе отвергать существование такого понятия как функция. На днях я на сайте "Биометрика" открыл новый раздел, посвященный А.Н. Колмогорову в связи со столетием со дня его рождения. Так вот, его теория вероятностей построена полностью на функциях.

ВШ: Ох, мой дорогой доцент! Не знаю, насколько имеет смысл продолжать "функциональную" дискуссию, но что-то зацепило. Попытаемся *«отделить мух от котлет»*. В анфиладе миров, с которыми сталкивается человек, можно выделить два обособленных, но взаимно влияющих мира:

- *Абстрактный мир математики*, включающий и математический анализ в его детерминационной ипостаси и методы исследования стохастических процессов (теория вероятностей, математическая статистика и пр.). Как любой абстрактный мир он абсолютно точен и однозначен. Например, окружность - вполне однозначное математическое понятие и способ вычисления ее длины достаточно детерминирован (в смысле точности определения величины π);
- *Реальный мир физических, химических, биологических и прочих процессов*. Все эти процессы принципиально многофакторны, нестационарны и стохастичны. В них просто нет ни функций, ни детерминизма. В некотором приближении детерминированными процессами можно назвать лишь некоторые внутримолекулярные явления.

НЦ: Нельзя! Понятия «в некотором приближении» и «детерминированные» логически не совместимы.

ВШ: Например, длина волны одной спектральной линии криптона почти в точности равна $6.0378 \cdot 10^{-7}$ м. Процесс атомного излучения изотопа Kr^{86} можно назвать "в некотором приближении" детерминированным процессом, что и было использовано в качестве эталона меры длины вместо известной платиновой болванки. Аналогично принят и эталон времени. Но, дело не в этом. Изучение процессов реального мира заключается в приспособлении ("восстановлении" по В.Н.Вапнику) некоторой приглянувшейся абстрактной функции из математического мира для описания имеющихся эмпирических данных. Это проделывали и И. Ньютон, и Г. Ом, и А.Майкельсон (для вычисления скорости света). Возникали "законы", имеющие внешне детерминированную внешность, но вероятностные (как любая модель вообще) по своей природе,

поскольку процесс реального мира и приспособленная для его описания математическая "функция" принципиально неидентичные.

НЦ: Переход "закономерности" в "закон" – это переход количества в качество. Но, тем не менее, закономерность и закон – разные вещи. Закон (природы, в частности) не знает исключений – иначе это не закон, а закономерность или правило.

ВШ: Например, Вы вышли на футбольное поле исследовать реальный центральный круг. Измерили 20 раз длину окружности и радиус и оценили по уравнению регрессии число π. Оно может варьироваться от 3.05 до 3.3, потому что Ваш круг может являться и не кругом вовсе, а эллипсом и нужно вести расчеты по иной формуле. Или Вы неточно находили центр окружности, радиус и проч. "Закон" в виде формулы или модели и реальный процесс никогда не совпадают.

НЦ: Но мне кажется, нам тут нечего рассуждать. В любом учебнике по математическому моделированию эти вопросы толково разобраны. Могу открыть и процитировать.

ГР: И опять я готов констатировать, что эта дискуссия о соотношении случайности и детерминированности не конструктивна. Мне представляется, что должно существовать понимание того, что это – *«две стороны одной медали»* и очень многое определяют цели исследования (объяснение, прогноз, измерение и пр.).

ВШ: Отнюдь, нет: мы потихоньку уткнулись в ключевой вопрос системологии: "что есть СИСТЕМА?". Если СИСТЕМА – это и есть сам реальный мир, то это представление имеет право на жизнь, но в практических целях оно малопродуктивно. Рациональнее предложить другое понятие «системы» (а значит и ЗАКОНА системы) – "это ОТОБРАЖЕНИЕ в мозгу исследователя некоторых ***наблюдаемых свойств*** реального мира, необходимых для решения поставленной задачи". Отсюда следствие: "система не может быть СЛОЖНОЙ". Утверждая так, исследователь либо неверно обозначил цели, либо привлек незначимые факторы, либо не нашел пути вычленения главного индикаторного показателя *(«нет фригидных женщин, а есть только неумелые любовники»*). И.Ньютон, например, поставил своей целью оценить силу физического взаимодействия между телами. Он сконструировал специально для этой цели СИСТЕМУ взаимно влияющих материальных тел, нашел (вероятно, статистически!) закономерность в виде регрессионного уравнения и назвал ее ЗАКОНОМ всемирного тяготения. Так где та грань между законом и закономерностью? Понятно, что отличие между ними – в мере всеобщности. Но тогда надо разработать принципы измерения всеобщности. Одна моя шефиня, готовя докторскую диссертацию по экономике, открыла двадцать с лишним новых экономических законов рационального природопользования при социализме, которые вскорости благополучно загнулись…

НЦ: Вот Вам и примеры закономерностей… Все-таки я не во всём согласен как с Марксом, так и с Каутским! Моделируемые процессы имеют принципиально разную внутреннюю сложность, а сложность модели должна быть адекватна сложности процесса.

ГР: Совершенно верно. Взаимодействие двух шариков может быть описано по Ньютону простой формулой, взаимодействие двух популяций из хищников и их жертв – одним дифференциальным уравнением. Совершенно другое дело – модель экосистемы даже небольшого озера. В противоположность оценке вещественно-энергетических параметров простых физических систем, исследование экологических систем связано с изучением сложных морфофункциональных параметров. Описания экосистем нельзя "вогнать" в простые уравнения, поскольку аналитические модели отдельных качеств не адекватны принципам поведения всей системы в целом.

ВШ: Все зависит от поставленной задачи. Однако предлагаю закрыть тему и двинуться дальше.

НЦ: В главе 3 вы приводите хороший критический анализ показателей химического загрязнения воды. Убедительно, но одна загвоздка. Все показатели суть – случайные величины! Значит их сравнение с эталоном – предмет теории проверки статистических гипотез!

ВШ: Именно эта причина и побудила нас дополнить главу разделом 3.7.

НЦ: Я не совсем удовлетворен формой его изложения. В теории статистики приняты определённые нормы языка и обозначений. Хотелось бы их придерживаться, а вы этим грешите на протяжении всей книги. В частности, статистические параметры (константы) принято обозначать греческими буквами, а их статистические оценки – латинскими (или греческими же с "крышечкой"). Наконец, трудно понять, что к чему, из-за отсутствия примеров.

ВШ: Замечание по "нормам языка" справедливо. А пример мы не дали по следующим причинам.
Во-первых, мы не очень согласны с Вашим тезисом о "независимости" измерений концентраций отдельных ингредиентов. Все данные многокомпонентного химического анализа в пространственно-временном аспекте являются, как правило, СОПРЯЖЕННЫМИ. Тут правильно говорить о статистических оценках варьирования самого показателя μ, как суммы нормированных на ПДК концентраций n ингредиентов, измеренных в одной точке в одно и то

же время. Эти статистики должны быть рассчитаны по единым формулам, т.е. как и для остальных независимых столбцов.

Во-вторых, нам видится целая "матрешка" задач от простейшей к наиболее сложной:

а). Есть один сформулированный критерий качества воды μ и альтернатива гипотез: больше или меньше μ некоторого порогового значения ρ (в Вашей интерпретации $\rho = 1$);

б). Далее, перейдем от двух классов "чисто - грязно" к шести гостированным классам качества воды. Тогда нужно пять раз выполнить проверку гипотезы (а) для разных пороговых значений ρ (например, $\rho = 0.5, 0.8, 1, 2.5$ и 5). Получается вектор из пяти значений α-вероятностей.

в). В условиях задач (а) и (б) критерий качества воды не один, а несколько, например, "Концентрация химических примесей", "Показатель биологического разнообразия фауны", "Показатель рекреационного и эстетического благополучия водоема", причем граничные значения для всех классов определены. В итоге вычислений как в разделе 3.7. получается матрица из 3х5 α-вероятностей. А вот что с ней делать дальше? В этом и суть проблемы.

НЦ: Верно ли я понял, что μ может быть равно, скажем, 2.6 и мы отнесём эту грязищу к классу качества 4? Тогда и так ясно, что об остальных классах качества можно не беспокоиться и рассматривать только сравнение μ с ближайшей пороговой величиной 2.5!

ВШ: Альтернативная оценка "чисто-грязно" безобразно груба – для любой реки всегда можно найти такую пробу, где какой-нибудь компонент превысит ПДК. А нужная граница между классами находится как вероятность максимально приближенная к критической. Но для этого надо вычислить весь вектор вероятностей.

НЦ: Со временем эту задачу можно подробно разобрать, построить графики, номограммы, всё будет легко считать. Только, насколько я понимаю, в каждой пробе воды мы имеем дихотомизм: или она чистая, или грязная! А "среднее" (это тот самый индекс ИЗВ, который меня *ИЗВ-одит!*) – есть, по-моему, "средняя температура по больнице". Кстати, Вы, наверное, будете смеяться, но "средняя температура тела людей, заболевших одной болезнью" – вполне легитимная статистика! Только есть один нюанс: она не интерпретируемая! Просто она имеет размерность температуры и является неким эмпирическим коэффициентом в формуле для вероятностных расчётов. Её можно, например, использовать для расчёта вполне интерпретируемой *области допустимых значений* (см. Л.Н. Большев и Н.В. Смирнов [1968]) *температуры* гриппозных больных в критическом периоде их заболевания. То же можно сказать и об ИЗВ.

ВШ: ИЗВ не столь уж плох, как Вы его ругаете. Во-первых, это – единственная утвержденная на уровне министерства методика оценки загрязнения в шести-балльной шкале. Во-вторых, усреднение идет всегда строго по шести компонентам и, если компонентов измерено менее шести, то расчет просто не производится. В этих условиях сумма Σ и среднее M функционально идентичны $M=\Sigma/6$ или $\Sigma=6M$.

НЦ: Извините! Ваша река Сок – объект с распределёнными (по её длине и времени) параметрами физических свойств. В одних пробах $\mu<1$ (чистая), в других – $\mu>1$ (грязная). А как брать пробы на объекте – это не дело предметной науки (гидрологии, в вашем случае), а методов рандомизации и ГОСТов. И считать надо не ИЗВ: как я уже писал, это не интерпретируемая статистика, хоть трижды утверждённая бюрократами! Нужно оценивать долю "чистых" проб по отношению ко всем пробам, причем интерпретация результатов должно быть с распределёнными статистическими параметрами, как и сам объект.

ВШ: Вы, Натан, давно не были в России и забыли, что не параметры, а деньги здесь так мастерски распределяются, что их никто не найдет, тем более, на тщательные гидрохимические исследования…

Что касается Вашего подхода, то, по моему посконному мнению, в этом объективном мире есть только два вида измерений: простая выборка (или вектор измерений) и многомерная матрица СОПРЯЖЕННЫХ наблюдений (т.е. n различных показателей измерены при условии единства места и времени измерения). И БОЛЬШЕ НИЧЕГО. Есть некоторые попытки некоторых субъектов назвать многомерным пространством измерений некоторую совокупность несвязанных между собой простых выборок, но я, как честный программист и гидробиолог, с такими данными ничего общего иметь не хотел бы. А что же все-таки нам делать с несколькими частными критериями качества вод?

НЦ: Попробуйте использовать методику многокритериального экспертного оценивания (МЭО), подробно представленную в моей книге.

ВШ: Да, эксперты стоят дешевле (и потому легче покупаются). Из того, что я понял в МЭО, все сводится к одной формуле. Имеется n частных критериев качества K_i и общий критерий качества K_{cp} рассчитывается как средневзвешенное из этих значений: $K_{cp} = \Sigma B_i * K_i / \Sigma B_i$, где B_i – некоторые

веса. Остальное – сплошные частности, которые заключаются в многоярусном способе расчета B_i. В гидробиологии такой подход использовали многие: и Е.В. Балушкина, и В.И. Баканов, и наш соавтор Т.Д. Зинченко.

Кстати, Ваша *«статистическая оценка обобщенного среднего взвешенных частных критериев качества – ОСВ СО ЧКК»* (ух, правильно?) мало съедобна без какой-либо оценки ошибки или доверительного интервала. Т.е. является такой же не интерпретируемой статистикой, как и ИЗВ.

Честно говоря, я с большим сомнением отношусь к средним (или средневзвешенным) величинам, особенно, когда надо определить класс или категорию (в нашей книге это постоянно подчеркивается). Для убедительности по этому поводу приведу следующий пример. Пусть в некой Государственной Думе образовался комитет, состоящий из 12 чел, в том числе:

- Правые партии – 5 чел.
- Центр ("партия власти") – 1 чел.
- Левые партии (коммунисты) – 6 чел.

Нужно выбрать председателя, но мнения разделились. Обратились к специалисту по экспертному оцениванию г-ну *Z.* Он предлагает следующую методику:

*«Вводим для каждой партии "балл левизны" от 1 (Немцов) до 9 (Зюганов). Партии центра дадим балл 6. Тогда средний критерий левизны комитета: (1*5 +5*1+9*6)/12=6.1. Т.к. ближе всего к этому критерию "партия власти", то следует избрать председателем ее представителя».* Через некоторое время Президент вручает в Кремле г-ну *Z* премию за выдающиеся заслуги в области науки и техники. Вы спросите, "причем тут Президент?" А я Вам скажу, что именно таким образом оценивают класс качества воды по индексам сапробности, используя формулу Пантле-Букка.

Далее, веса B_i можно оценивать тремя методами:

- собрать шеренгу экспертов, как это делаете Вы, и заставить их пробаллировать ситуацию (собрать то можно, да где их взять и чем кормить?);
- придумать для ранжирования важности каждого частного критерия какое-нибудь аналитическое выражение (например, каждой пробе присвоить ценность, увеличивающуюся по мере приближения даты наблюдения к 1 июля, когда гидробиологические тесты имеют наименьшую погрешность);
- использовать для расчета значений весов методы оптимизации, доставляющие, например, минимум ошибки распознавания.

Третий метод с использованием коллектива предикторов наиболее перспективен, наименее субъективен и описывается в нашей предыдущей книге "Экологическое прогнозирование" [Розенберг и др., 1994а]. К сожалению, для него необходима обучающая выборка с реками, для которых точно известен объективный класс качества. Но тут брезжит очень занятный подход, связанный со следующей вычислительной процедурой:

а) выбирается 1-й частный критерий, он объявляется основанием для оценки классов, после чего по процедуре Бейтса-Гренджера или алгоритму Дикинсона-Ершова рассчитываются оптимальные (с точки зрения критерия 1) значения весов остальных критериев B_i;

б) процедура повторяется для всех частных критериев и получается матрица B весовых коэффициентов размерностью nxn (что с ней делать, я пока не придумал; надо ее как-то рекурсировать или свернуть в вектор).

ДФ: Вы собираетесь в неявном виде использовать метод стохастической аппроксимации, или, если вульгарно, "бутстреп". На сайте "Высокие статистические технологии" А.И. Орлова об этом в достаточном объеме написано.

НЦ: Однако, перейдем к части 3. На рис. 5.1 основные типы вероятностных распределений случайной величины представлены плохо. Координаты не названы; надо изобразить параметры (M, σ и др.)!

ВЛ: Рекомендую задать в EXCEL выражение для плотности, сгенерировать порядка 100-300 наблюдений и далее построить в ППП "STATISTICA" эти кривые распределения.

ВШ: А может дать возможность попрактиковаться самостоятельно нашим читателям?

НЦ: Описание распределения Стьюдента (стр. 208) мне не понравилось. Там всё не то и не так! Начинать надо с χ^2-распределения и т. д. Можно переписать у Н. Джонсона и Ф. Лиона [1981] или у других.

ВЛ: Было бы нелишним более четко акцентировать на необходимости проверки обязательных условий применения критерия Стьюдента при сравнении выборок.

ВШ: Но, помилуйте, мы писали не учебник по математической статистике, а практический справочник для гидробиологов, использующих те или иные статистические приемы в своей предметной области. И "математические листы" мы привели, скорее, не для обоснования самого подхода (об этом написаны груды специальной литературы), а чтобы привести конкретные расчетные формулы критериев и статистик, которые могут встретиться по ходу использования отдельных методов. Например, используя ППП "STATISTICA" для проверки гипотезы о законе распределения,

гидробиолог может столкнуться с возможностью рассчитать, скажем, критерий Шапиро-Уилка, о котором он не имеет ни малейшего представления. И тогда он открывает нашу книгу на стр. 212…, после чего остаются удовлетворенными и авторы и читатель. А более "продвинутый" пользователь в дальнейшем может обратиться к первоисточникам. Что касается критерия Стьюдента, то, в отличие, например, от медицины, его использование в гидробиологии почему-то весьма непопулярно и трудно сказать, насколько это явление обоснованно.

НЦ: Я перечитал много статистических работ (статей, книг, учебников). Многие авторы переписывают друг у друга и у "классиков", мягко говоря, не критически. Вот и появляются в новейших учебниках голые точечные оценки, некорректные сравнения популяций по средним, небрежно указанные (или не окаймленные вовсе) области определения регрессионной модели, неокругленные статистических оценки, неоправданно большие (и иногда неясно, зачем нужные) корреляционные матрицы, громоздкие математические модели с незначимыми факторами и многое другое из "классического наследия". Это относится и к давно раскритикованным критическим пятипроцентным уровням значимости. Вот и у вас: «*выбирается критический уровень значимости* $a_{кр}$ *из стандартной линейки типа 0,001; 0,01; 0,05 (например,* $a_{кр} = 0.05$*)*». Всё не то и не так!

ВШ: Вы в своих "Записках…" [Цейтлин, URL] предлагаете экспертно оценивать уровень ответственности в баллах, по которому уже и рассчитывается уровень значимости. Лично мне эта "двухуровенность" кажется малопонятной: эксперту все равно, что назначать – балл от 0 до 10 или непосредственно $a_{кр}$.

НЦ: Вы даете определение основной задачи математической статистики как *«вычисление статистик, являющихся критериями для оценки достоверности априорных предположений, гипотез или выводов по существу эмпирических данных»*. Мне это не нравится. Скажите, какие у Вас есть книги, и я скажу, откуда лучше переписать, что есть основной предмет математической статистики. Я считаю, что: *«Основным предметом математической статистики является сбор и математическая обработка информации, полученной в результате наблюдений реально существующего объекта (или явления), подлежащего экспериментальному изучению и математическому моделированию»*.

ВШ: Напишем сокращенно, обрезав информационные излишки: «*сбор и математическая обработка*». Совершенно с Вами не согласен – это ужасное гипертрофирование. "Сбором" занимается гидробиология, химия, гидродинамика и пр., но никак не статистика. "Математическая обработка" – это не есть БРЕНД статистики, т.к. это слишком общее определение, куда подходит масса других разделов математики или точных наук от топологии до прогнозирования маразмов. *Любое определение только тогда имеет смысл, когда оно подчеркивает основную характерную черту определяемой сущности.* Наше определение – гораздо конкретнее, и потому симпатичнее. А чем еще занимается статистика?

НЦ: Статистика занимается еще методологией сбора! В.В. Налимов приводит коллекцию из 165-и определений понятия «математическая статистика», многие из которых не только лишены чёткости, но и противоречат друг другу. Ваше будет 166-м. Но наиболее глубокое представление об изучаемом предмете можно приобрести только на собственном опыте.
На стр. 210 вы пишете: «*Причина такого выделения нулевой гипотезы заключается в том, что она обычно рассматривается как утверждение, несостоятельность которого более бесспорно, чем истинность.*» Это неверно. Обычно сравнивают величины ущербов от ошибочного отклонения гипотез. Далее, написано: «*Нулевая гипотеза* H_0 *не отклоняется, если вычисленное значение статистики критерия* $K_{рас}$ *не превышает порогового* $K_{пор}$» Просто неверно! Бывает и наоборот. Посмотрите формулировку статистических гипотез в разделе 1.1. моей книги.

ВШ: Вы упоминаете некоторые наиболее удачные разделы и формулировки из Вашей прекрасной работы. Я нисколько не собираюсь Вам возражать. Но что нам прикажете делать – выбрасывать куски из своей книги и заменять их на Ваши? Это вряд ли разумно, несмотря на позитивный конечный результат. «*Каждый пишет, как он дышит*». Но мы постараемся в нужных местах вставить рекомендации обратиться к Вашей книге – кому надо, тот прочтет, сравнит и поумнеет.

НЦ: Написано (с. 213): «*Применение критериев согласия связано с определенными теоретическими и вычислительными сложностями*». Дело не в этом – как теоретические, так и вычислительные (с появлением ЭВМ) сложности преодолены. Суть в том, что желательно использовать сразу несколько критериев! Особенно, если Вы заранее предполагаете, что распределение нормальное. Кстати, в моей книге описана более простая редакция ***d***-критерия Гири. Получилось ***F***-распределение Фишера, а расчёты элементарные!
Вы также приводите много гистограмм распределения выборок гидробиологических показателей и их функциональных преобразований. Я бы гистограммы вообще не рисовал – мне они не нравятся, поскольку сильно зависят от разбиения данных на интервалы. Обработку данных лучше представлять сразу в виде ЭФР (*эмпирической функции распределения*) в доверительных границах

для заданного (нормального, раз вы хотите) распределения! ЭФР не зависит от разбиения данных на интервалы (!), а если хотите посмотреть ЭФПВ, то сгладьте и продифференцируйте её численно – получится красиво!

На мой взгляд, вы слишком часто приводите значения критериев согласия, Стьюдента, Фишера и пр.. Эти значения ни уму ни сердцу ничего не дают, тем более, в итоговых таблицах (мало ли какие промежуточные величины вычисляются)! Зато не всегда приводите соответствующий им уровень значимости. Необходимо приводить всегда и только α-критерии (***P***-значения)!

ВШ: Не думаю! Народ должен привыкать к "натуральным" значениям статистик и, при некотором навыке, исследователю достаточно будет просто посмотреть на значение F-критерия, чтобы сделать исчерпывающий вывод.

ГР: А я здесь готов согласиться с Вашим замечанием, Натан.

ВЛ: Стоит подчеркнуть, что описываемые вами в гл. 5 классический дисперсионный анализ (ANOVA) является параметрическим методом. Однако есть и многочисленные методы непараметрического дисперсионного анализа. Кроме того, не понимаю, почему при изложении двухфакторной модели Вы исключили из рассмотрения эффект взаимодействия? Здесь было бы уместно перейти от 2-х факторного анализа к многофакторному, указав на другие виды эффектов факторных взаимодействий.

НЦ: Слишком подробно описан дисперсионный анализ. Я бы лучше потратил время на описание многомерного регрессионного анализа и просто указал, что дисперсионный анализ – это его частный случай!

ГР: Дисперсионный анализ родился в недрах биологии (Р. Фишер), тщательно прорабатывался биологами (Н.А. Плохинский) и считается основной математической доктриной биологии. Во многих случаях важно (с такой же необходимостью как клистир, прописываемый всем в "Бравом солдате Швейке") проверить гипотезу о влиянии на одномерную выборку того или иного качественного фактора, не прибегая к построению регрессии. Такова *традиция*, поэтому однофакторный ANOVA и все его детали должны быть растолкованы гидробиологу. Во всех пакетах прикладных программ дисперсионному анализу также уделяется достаточно важное место.

НЦ: Добавлю только: "*гнетущая традиция*"! (Со мной также согласны Н. Дрейпер и Г. Смит [1986]). У меня студенты мехмата не могли толком понять этот "клистир", пока я не начал его излагать в одном ключе с регрессионным анализом. Сразу всё стало просто и ясно.

По тексту раздела 5.2 я не во всем согласен с вашими выводами относительно статистик Бартлетта и Кокрена (мне они просто не нравятся, так как формулируются без альтернативной гипотезы). Лучше уж критерий Хартли – Дэвид (см. Дж. Себер [1980]), а ещё лучше (уж извините!) мой критерий, который я назвал критерием Фишера – Бонферрони. Не верно также мнение об ошибочности простого попарного сравнения групповых средних по Стьюденту (эта, так называемая "проблема Беренса – Фишера" давно решена). Но тут важно другое. Чтобы уйти от субъективного взгляда на частные методы и критерии, нужно каждую гипотезу проверять с помощью НЕСКОЛЬКИХ критериев, в том числе и непараметрических. Если ВСЕ критерии дружно дают малые ***P***-значения, значит разность безусловно достоверна! А наши внутренние статистические разборки – какой критерий мощней иного – только запутывают гидробиологов!

Далее, вы много раз повторяете формулировку: «*Проверка гипотез о равенстве групповых средних*». Это принципиально неверно! Параметрические гипотезы проверяют относительно параметров (констант)! А средние - это случайные оценки параметров. Корректнее говорить «*Проверка гипотез относительно центров распределения случайных величин*» или «*Проверка гипотез относительно математических ожиданий распределения случайных величин*».

ВШ: Несмотря на смешанные чувства от Вашей критики, она весьма полезна. Особенно по части необходимости точности формулировок. С точки зрения статистика Вы, безусловно, правы. Но, с позиций практического биолога, выражение «*Проверка гипотез о равенстве выборочных средних*», как-то ближе к телу. А при реализации конкретных вычислений эти реверансы в сторону ритуальных статистических схем вообще кажутся бессмысленными. Например, Вы подходите к компьютеру с традиционной цепочкой предположений типа «*Я хочу сравнить два центра распределения... Правда, я ничего не знаю о законе распределения генеральных совокупностей и не знаю как найти математические ожидания. Поэтому буду сравнивать их оценки в виде выборочных средних*». Компьютер пропустит Вашу "предобеденную молитву" и начнет сравнивать ***выборочные средние***. Хотя, повторяю, Вы совершенно правы.

НЦ: Теперь замечания общего характера: на рисунках надо по возможности приводить доверительные интервалы. Популярно говоря, мы по случайно разбросанным точкам пытаемся воссоздать как бы один или два холма. Склоны холмов можно изобразить так, как изображают склоны гор на географической карте, т.е. в виде линий равного уровня значимости: 80%, 90%, 95% и 99-%! С

компьютером это сделать просто. А эти «*прямоугольники и усы*» только путают некомпетентного читателя (а таких большинство)!

Далее, некоторые диаграммы плохо понятны. Не нужно рисовать логарифмы! Надо рисовать натуральные значения, но на полулогарифмической бумаге! И, наконец, за что мне нравятся медики – они любое утверждение подкрепляют ***P***-значением, а вы это делаете не всегда! И приводимые значения статистик округляйте до 2-3 значащих цифр, т.к. большая точность выглядит эдаким неуместным кокетством.

ВШ: Ваши замечания верны и исполнимы, но отложим их для будущих редакций. Хотя логарифмическая бумага выглядит некоторым анахронизмом. По-моему, все равно, что деформировать, оси координат или корреляционное поле…

Что касается излишней точности, то знаете, почему уважаемые фирмы продают, допустим, надувные матрацы за 1399.99 долларов? Учитывают психологию потребителя! А чем мы хуже?

НЦ: Вы сомневаетесь (стр. 227): «…*следует ли учитывать в составе выборок нулевые значения, т.е. так называемую "частоту нулевого класса"*…». Насколько я помню, Е. Маркова лет 25 назад докладывала решение. У неё есть свой сайт. Можете проконсультироваться...

Граф связности категорий грунтов на рис. 5.9 мне не понравился. Если вы написали пособие для гидрологов, то этими критериями и линиями вы их совсем запутываете. Вместо того, чтобы просто, достоверно и однозначно ответить на конкретные вопросы.

ВШ: Этот граф мы привели как некоторую прелюдию к разговору в главе 7 о трансформации матрицы сходства в граф минимального пути, подготавливая читателя к "будущим забегам".

НЦ: В разделах 5.3-5.4 вы анализируете влияние сезонного фактора, выраженного календарным месяцем отбора пробы... Не могу не высказать сожаления, что вы не пользуетесь регрессионным анализом вместо дисперсионного. Тогда вместо календарного месяца можно было бы просто записать несколько базисных функций ряда Фурье, и ЭВМ сама бы выделила значимые члены, дала бы (количественную) регрессионную модель и не надо было бы приводить лишних таблиц. А вы получили бы удовольствие от количественной интерпретации регрессионной модели...

ВШ: Совершенно с Вами согласен. Я сам вздохнул с огромным облегчением, дописав разделы 5.1-5.4. Но примеры носили чисто методический характер для иллюстрации именно этих методов.

НЦ: Мне совершенно не нравится ваша постановка задачи регрессионного анализа (раздел 5.5). Почему речь идет о законе, а не о закономерности? Почему обе переменных должны быть измерены в количественных шкалах (это не обязательно)? Вы собираетесь «*построить функцию* ***f(X)****, которая приближенно описывала...*». Зачем приближенно? Надо требовать адекватность!

ВШ: Регрессия – «*как много в этом звуке*…» и столько разногласий! Нам близки воззрения М.М. Бонгарда, А.Г. Ивахненко и В.Н. Вапника. Есть объективный закон природного явления, который необходимо восстановить по эмпирическим данным. "Выборочная закономерность" – как-то слишком мелко и не стоит возни.

Регрессия выполнима только для количественных шкал, поскольку оперирует со средними, дисперсиями и проч. Вы, конечно, можете какую-нибудь порядковую или бинарную шкалу интерпретировать как количественную – воля и ответственность Ваша. А как Вы найдете среднее из величин, которые принимают, например, только два значения 0 и 1?

Помните, у Окуджавы: «*Две спутницы верных: любовь и разлука, проходят сквозь сердце мое*». У регрессии – две стороны медали: адекватность как мера близости и вероятностность как мера удаленности. И та и другая выражаются суммой квадратов отклонений от линии регрессии. Здесь под словом «приближенно», в отличие от детерминированного подхода, мы хотели подчеркнуть *вероятностную суть любой регрессии*.

НЦ: Вы часто употребляете термин «*уравнение регрессии*». Так многие пишут, но это неверно. Следует писать «…*в виде модели регрессии*». Попробуем разобраться не спеша.

Имеем некоторую случайную величину ***Y***, которая распределена нормально с центром ν_y и стандартным отклонением σ_y, то есть $\boldsymbol{Y} \sim \boldsymbol{N}(\nu_y, \sigma_y)$. Существует некоторый фактор ***X*** – независимая переменная, которая может принимать определенные значения $\boldsymbol{X} = \boldsymbol{x}$. Если хотя бы один из параметров распределения ***Y***, например, центр ν_y, является функцией ***x***: $\nu_y = \eta(x)$, то *имеет место регрессия* ***Y*** *по* ***X***. Функция $\eta(x)$ называется *истинной функцией регрессии* (ее вид и свойства часто неизвестны). Задачей регрессионного анализа (РА) является восстановление по экспериментальным данным наиболее вероятного вида этой функции, которая после восстановления будет уже называться в отчете э*мпирической функцией регрессии (ЭФР)*.

Уравнением регрессионная функция бывает временно, когда она записывается в виде системы уравнений и решается относительно коэффициентов. В результате расчёта коэффициентов

уравнение регрессии становится эмпирической функцией регрессии (ЭФР)! Такова корректная терминология.

ВШ: В рамках своей книги мы используем, казалось бы, вполне общепринятые понятия:

- *Модель* – некоторая система взаимосвязанных уравнений произвольного типа, воспроизводящих суть наблюдаемых процессов или явлений;
- *Уравнение* – форма представления задачи о поиске аргументов, при которых значения двух функций равны;
- *Функция* – некоторый математический оператор, описывающий преобразования аргументов.

Обычно функция – одноместное понятие, не содержащее символ «=», и представляет собой правую часть уравнения регрессии. Обратимся, в частности, к «Большой Советской энциклопедии»: «*Цель регрессионного анализа состоит в определении общего вида* ***уравнения регрессии****, построении оценок неизвестных параметров, входящих в* ***уравнение регрессии****, и проверке статистических гипотез о регрессии*» (статья со ссылкой на Г. Крамера [1975], Н. Дрейпера и Г. Смита [1986]). Т.е. БСЭ ведет речь исключительно о "*регрессионных зависимостях*" и ничего не говорит о "*функциях регрессии*". Вряд ли можно предположить, что БСЭ не права – статьи там писались высочайшими профессионалами и тщательно редактировались.

Существующие традиции статистики требуют некоторой чисто ритуальной преамбулы типа «*Пусть где-то (непонятно где) существует некоторая генеральная совокупность данных или истинная функция регрессии. Мы о них ничего не знаем, поэтому*…». Но вряд ли этот процесс умозаключений очень интересен практикам. Им хочется просто взять быка за рога и на основе набора эмпирических данных рассчитать коэффициенты приближенного, но адекватного *уравнения регрессии*, которое можно было бы использовать для предсказания или управления.

ВЛ: Очень важно более четко сказать об основных предпосылках регрессионного анализа и о степени влияния отклонения от этих требований: Привожу кратко все семь "заповедей от регрессии" со ссылкой на книгу И. Вучкова с соавторами [1987]:

1. Ошибки (невязки) есть случайные величины.
2. Ошибки имеют нулевое математическое ожидание.
3. Все значения ошибок не коррелируют между собой и имеют одинаковые дисперсии (условие гомоскедастичности; это предположение в гидрологии, в связи с сезонностью, может нарушаться, особенно в данных, представляющих собой временные ряды).
4. Ошибки имеют нормальное распределение.
5. Матрица регрессоров **F** не случайна, т.е. ее элементы - известные числа, точно заданные исследователем.
6. На параметры уравнения регрессии априорно не накладывается никаких ограничений.
7. Матрица регрессоров **F** имеет ранг равный числу коэффициентов в уравнении регрессии.

Если все эти предположения выполняются, то имеет место случай классического РА независимо от метода оценки коэффициентов.

ДФ: Это стандартное заблуждение, что для метода наименьших квадратов (МНК) необходима нормальность распределения случайных ошибок зависимой переменной ***Y***, то есть разностей ***Y***-(***a***+***bX***). У В.Н. Вапника несколько десятков нудных страниц посвящено доказательству теорем, главным итогом которых является то, что среди распределений с одинаковой дисперсией нормальное является экстремальным (в определённом смысле, самым плохим) и, используя его, мы можем только перестраховаться, занизить значимость результата. Требование нормальности всего лишь удобно для легкого доказательства некоторых статистических теорем, но линейная регрессия по МНК может излагаться и с позиций функционального анализа, и вообще сама по себе. Получаемые результаты не зависят от доказательства нужных для остепенения теорем. В век быстрых компьютеров можно использовать метод стохастической аппроксимации, который, по сути, бóльшую часть классической статистики со всякими распределениями, параметрическими оценками и т.п. отправляет просто "в корзину", оставляя лишь фундаментальные понятия и теоремы.

NN: (*Из рецензии "Журнала общей биологии" на нашу статью, указывающую на некорректность сведения степенных уравнений энергетического обмена к линейной форме – см. пример к разделу 5.5*)*:* Теоретически проблема некорректности линеаризаций, возможно, и существует. Но при практическом применении обоих видов анализа к большим массивам данных, оказывается, что различия в оценках параметров находятся в пределах статистических ошибок.

ВШ: Возможно, Вы правы, но лишь только потому, что стандартные ошибки коэффициентов энергетических уравнений весьма высоки (хотя вопрос о том, насколько существенны различия, всегда достаточно субъективен). Но в нашей работе речь идет не о том, насколько "неверный" результат близок к "верному", а о том, как методологически правильно выполнить расчет. А выводы пусть делают гидробиологи, имеющие необходимые экспериментальные данные.

NN: Уравнение, описывающее эмпирические данные по интенсивности дыхания ракообразных от массы тела, Г.Г. Винберг предлагал в середине 50-х годов. Тогда это было блестящее исследование. Надо бороться с неряшливой математикой, но надо ли ворошить прах усопших, тем более, когда они были действительно выдающимися учеными.

ВШ: Такие глубоко почитаемые нами ученые как Г.Г. Винберг и А.А. Умнов, заложили основы научного мировоззрения, связанного с расчетом энергетических балансов. Но они не могли не ошибаться в некоторых второстепенных частностях (особенно, в методах статистической обработки, которые лимитировались уровнем технического обеспечения того времени). Мы считаем моральным долгом перед их памятью и проявлением бережного отношения к научному наследию (а не желанием "ворошить прах усопших") обсудить с научным обществом эти неточности, чтобы использовать в дальнейшем более корректные подходы в соответствие с духом времени.

NN: То, что полином "лучше" опишет любую зависимость – очевидный и наперёд известный результат (у полинома больше параметров). Но "лучше", не всегда "верно". Что же касается цитаты из статьи Г.Г. Винберга, то, по крайней мере, в использовании степенного уравнения для описания зависимости метаболизма от массы тела он оказался полностью прав. Достаточно посмотреть последние публикации за 3 года, чтобы увидеть большое число моделей, предсказывающих не только степенной вид зависимости, но и значения коэффициентов (см., например, [West G.B. et al., 1999] и др.). Вряд ли стоит опубликовывать свои взгляды на эту "проблему", апеллируя к состоянию науки 26 лет назад.

ВШ: Совершенно непонятно, что Вы понимаете в статистике под словом «***верно***». Обычно верной моделью из двух возможных претендентов считается та, что более ***точна*** (адекватна) по отношении к эмпирическим данным (разумеется, с учетом параметричности расчетного уравнения, для чего существуют специальные статистические критерии). И уж никогда массовость какого-либо предположения не была свидетельством его верности. Вместе с тем, традиционно (см. главу 2) принято различать модели для ***объяснения*** (относительно простые модели, к которым не предъявляются требования оптимальности, но, в силу их широкой распространенности, пригодные для поиска общих гипотез) и модели для ***прогнозирования*** (более сложные и точные модели). Не имея ничего против степенного уравнения как модели для объяснения, мы показываем, что не только ***верными***, но и более ***удобными*** моделями для целей прогнозирования могут являться и иные структурные версии (причем, не обязательно полиномы).

ГР: Добавлю от себя. В связи с различением в разных моделях объяснения и прогнозирования для сложных систем можно предложить такую дихотомическую схему [Брусиловский, Розенберг, 1981б; Розенберг, 1984, 1989]: различать *праксеологичность* (количественная "точность") и *собственно адекватность* (качественная "точность"). Тогда, регрессионные модели служат, в основном, целям прогнозирования и могут вообще не отражать ни структуру, ни механизм функционирования моделируемой сложной экосистемы и будут хотя и полезны для практики в силу количественной точности своих прогнозов, но не адекватны реальным объектам в гносеологическом смысле. Для аналитических моделей (типа Вольтерра), напротив, нельзя говорить о праксеологичности (так как при их построении исследователь идет на ряд существенных упрощений, модели основаны только на априорной информации и пр.), но имеет смысл говорить о гносеологической адекватности. Думается, именно здесь лежит непонимание ***NN***, который пытается в модель Г.Г. Винберга "вложить" и адекватность, и праксеологичность, что для сложных систем недостижимо.

Критерии оценки адекватности и праксеологичности моделей достаточно многочисленны, но и их также следует разделить на два основных класса: *внутренние* и *внешние*. Если для оценок праксеологичности моделей такое разделение достаточно очевидно (внутренние критерии основаны на той же информации, по которой строилась модель [МНК], а внешние – на новой [МГУА]), то для оценки собственно адекватности моделей такое разделение дать сложнее. Можно считать, например, что внутренними критериями являются теоретические предпосылки самой экологии или гидробиологии (численность популяции не должна быть отрицательной…). Тогда внешние критерии следует искать в области математики и математического анализа моделей экосистем.

НЦ: Методы структурной идентификации – это колоссальная наука, даже искусство! Сперва следует подобрать и обосновать структуру регрессии, включающую все подозреваемые факторы и соответствующие базисные функции, а только потом выбирать наилучшее в каком-то смысле подмножество базисных функций. А у вас он сведен только к выбору наилучшего подмножества базисных функций; в каком смысле – не ясно. Например, теория подобия даёт структуру мультипликативных функций регрессии при описании процессов в движущейся среде, в частности – гидродинамики процессов массопереноса. Можно подобрать дифференциальные уравнения этих процессов, интегралы которых и дают ОБОСНОВАННУЮ теорией структуру регрессионной функции, известную с точностью до параметров, которые, в свою очередь, и определяются по

экспериментальным данным. Что-то похожее есть, вероятно, и в гидробиологии. Там же следует упомянуть ортогональные полиномы Чебышева, сплайн-функции, функции Фурье и многие другие.

ВШ: Массопередача в сплошных средах – явление на несколько порядков более простое и предсказуемое, нежели взаимодействие популяций между собой и со средой. Единственный путь статистического моделирования экосистем – самоорганизация моделей. А для этого нужно сформулировать разумные критерии селекции и предоставить необходимые степени свободы выбора базисных функций, в том числе, и "экзотических". Об этом подробно – в главе 9.

НЦ: Вы пишете, что «*принципиально нельзя ограничиваться одной регрессионной модел*ью»? Иногда (и, даже, почти всегда) можно! Той, что адекватна (по ошибке воспроизводимости), имеет значимые коэффициенты регрессии и базисные функции, наиболее близкие к ортогональной системе. Иными словами, лучшей из эмпирических адекватных функций регрессии является менее сложная.

ВШ: Стоп!!! Качество любой модели регрессии определяют две конкурирующие величины: сумма квадратов отклонений от регрессии (СКО) и число степеней свободы, вернее, количество регрессоров (КР). Первый показатель отражает адекватность, второй – сложность. Увеличивая (КР), вы снижаете (СКО). Налицо – задача нахождения экстремума или модели оптимальной сложности.

НЦ: СКО – не интерпретируемая статистика. Вот если СКО поделить на число степеней свободы, получим интерпретируемую статистику – остаточную дисперсию. А если мы оценим дисперсию воспроизводимости (ДВО) отклика – получим оценку ошибки отклика. Модель регрессии, остаточная дисперсия погрешности которой не превосходит статистически дисперсию ошибки воспроизводимости отклика, и будет адекватной (см. литературу).

ВШ: Для существа задачи оптимизации различные функциональные формы статистик СКО, СКО/КР или (ДВО - СКО/КР) являются эквивалентными.

НЦ: ДВО не определяется как СКО/КР!

ВШ: Давайте четко различать, какую задачу мы перед собой ставим. Пусть есть 1 этап – оценка адекватности каждой из L рассчитанных моделей в отдельности – тут есть стандартная "кухня" интерпретируемых статистик: критерии Стьюдента, Фишера и проч. Но мы говорим о втором этапе – выборе "наилучшей" модели из L различных реализаций, каждая из которых уже прошла на 1-м этапе тест на адекватность. Ни F-критерий, ни остаточная дисперсия тут не могут являться тестом на "наилучшесть", поскольку МОНОТОННО уменьшаются с увеличением сложности.

НЦ: **О**статочная дисперсия в процессе расчётов методом включения обычно уменьшается с увеличением сложности, то есть, с возрастанием числа значимых регрессоров, но затем, при появлении мало значимых или вовсе незначимых регрессоров, проходит через минимум и возрастает.

ВШ: Такое далеко не всегда. Гораздо чаще СКО стремится к нулю: например, увеличивая степень полинома, вы придете к тому, что кривая точнехонько объедет все экспериментальные точки. Но подобная причудливая модель ни у кого не вызовет интереса. Нужно придумать что-то иное, основанное на той же "сладкой парочке" (КР и СКО). Интересные критерии, например, предлагает В.Н. Вапник [Алгоритмы и программы.., 1984].
Впрочем, мы начинаем повторяться или длиннотно обсасывать терминологические изыски.

НЦ: Это мы зря. Надо идти вперёд – к новым длиннотам и изыскам! Например, вы не вполне точно понимаете «адекватность». Все время в тексте, проверяя по Фишеру гипотезу об информационной способности модели регрессии, вы говорите об адекватности. Это же разные вещи!

ВШ: В общенаучном смысле «*адекватность (от лат. adaequatus – равный) – термин, который служит для обозначения верного воспроизведения объективных связей в представлениях, понятиях и суждениях*» [БСЭ]. В статистическом смысле один из тестов на адекватность – проверка на однородность двух дисперсий: воспроизводимости и остатков. Мы имеем полное право предположить, что дисперсия воспроизводимости отклика оценивается как его среднеквадратичное отклонение. Тогда, в определенном смысле, упомянутые Вами гипотезы – одинаковые вещи!

НЦ: **Д**исперсия воспроизводимости отклика оценивается как квадрат среднеквадратичного отклонения ошибки воспроизводимости отклика.

ВШ: Самое сложное в биологии – это понять, что есть «воспроизводимость» и его ошибка. Что, например, есть воспроизводимость роста человека, как не все множество его возможных значений?

НЦ: «*С чего начинается Родина*», т.е. регрессионная модель (РМ)? С области определения факторного пространства и с области значений для отклика. А в ваших примерах их нет! Надо оконтурить область E определения РМ хотя бы через минимумы и максимумы факторов! Правда при пассивном эксперименте область определения E, как правило, не образует прямоугольную фигуру (точнее, гиперпараллелепипед). Например, при корреляции двух факторов областью E может быть эллипс, сильно вытянутый вдоль случайно ориентированной главной оси. Задавая в таком случае область определения E в виде прямоугольника, мы ошибемся, если включим в нее подобласти, на

самом деле не принадлежащие ***E***. Эту задачу рекомендуется решать просто и правильно, чтобы полученные регрессионные функции не приводили к существенным ошибкам в периферийных точках факторного пространства. Я не знаю, где еще пишут о сложной задаче окаймления области определения РМ, поэтому можно сослаться на раздел 6 в моей монографии.

ДФ: Вы пишете, что в регрессионном анализе «*размерность признакового пространства практически не должна превышать 150-200, иначе возникают трудности вычислительного характера при матричных преобразованиях*». В книге Дж. Себера [1980], едва ли не исчерпывающей эту тему, описан модифицированный алгоритм Грама-Шмидта, успешно применяемый мною к матрицам из 1000 признаков и около 40000 объектов. Не надо с матрицами манипулировать, надо лишь решать свою задачу. Кстати, спектр возможных базисных функций гораздо шире, чем рекламируемый вами полином Колмогорова-Габора. Хотя многие из них, действительно, могут быть приближены полиномами. Но для одних – это очевидно, а для других – не нужно.

ВШ: Спасибо, коллеги. Наш диалог о регрессии – это ловкий слалом между вехами истин и заблуждений. Но давайте пойдем дальше.

ВЛ: Вы совершенно справедливо подчеркиваете, что результат, полученный при использовании кластерного анализа, является одним из возможных. Любое найденное разбиение необходимо сравнить с аналогичными результатами, полученными с применением других комбинаций метрик, алгоритмов объединения и т.д., а также с результатами использования других методов анализа данных. Следует убедиться самому и суметь убедить в этом своих оппонентов, что полученная классификация является оптимальной.

ГР: Нельзя забывать и о целях классификации. Свою библиотеку я могу "расклассифицировать" в соответствии с авторским или предметным указателями, или по цвету корешков книг…

ВШ: Гидробиологические данные – элементы размытых множеств с отсутствием четких границ и поэтому результат их кластеризации неустойчив к стратегии агломерации и задаваемой метрике. При этом нельзя сделать никаких априорных предположений ни о распределениях данных, ни о том, что должно получиться в итоге.

Пусть, например, мы имеем 25 классифицируемых объектов. Кроме того, мы имеем обычно не менее 5 общеупотребительных метрик и не менее 5 алгоритмов построения иерархической классификации. В результате для 25 объектов мы получаем 25 возможных вариантов разбиений (т.е. деревьев). Не подменяем ли мы неопределенность данных неопределенностью деревьев, еще более туманной и приводящей в ужас не только осла Буридана?

И мне видятся тут такие проблемы, требующие решения:

- раз уж мы получили 25 различных деревьев, то надо хотя бы оценить, значимо ли они отличаются друг от друга. Провести, своего рода, кластерный анализ результатов кластерного анализа. Каким методом можно оценить близость (предупорядоченность) двух произвольных деревьев из одних и тех же "листьев"?
- если нет метода сравнить два дерева, то как вообще можно сравнить две возможные версии кластеризации (например, оценить близость двух исходных матриц расстояний)?
- на основании какого критерия можно оценить, насколько одна классификация лучше другой?

ВЛ: Ничего не поделаешь - весь жизненный опыт человечества показывает, что в конечном счете применение новых, более совершенные методов анализа увеличивает количество информации об одном и том же объекте. Это, так сказать, общеметодологический аспект.

В нашем же случае (25 объектов, 5 метрик и 5 алгоритмов) ответ еще более очевиден. Во-первых, выбор метрик и алгоритмов не произволен, а имеет вполне целенаправленный характер, учитывающий априорную информацию о природе кластеризуемых объектов. Во-вторых, верифицируемость решения следует проверять по его предсказательной способности. А в этом случае, согласно теории байесовских оценок, мы получаем более точную оценку наших вероятностей (имеется в виду вероятности выбора решения).

ВШ: Не хотелось бы комментировать упоминание о мистической "предсказывающей способности" при кластеризации. Любой класс, агрегат, кластер, полученный по технологии "без учителя" и состоящий из некоторого подмножества реальных объектов, – всегда умозрительная теоретическая конструкция, которую принципиально невозможно точно измерить, а, следовательно, оценить качество предсказания. Например, В.И. Ленин разделил все человечество на эксплуататоров и эксплуатируемых, последних – на пролетариат и крестьянство, а последних – на кулаков, середняков и бедняков… В результате его "кластеризации" и "прогностической" операции одинаково перестреляли и тех, и иных. Но, к счастью (?), он имел большую обучающую выборку…

ВЛ: Уважаемый коллега Владимир Кириллович! Вряд ли Ваш аргумент достаточно убедителен. Не думаю, что стрельба началась от того, что "вожди всех времен и народов" поделили все

человечество на группы. В этом убеждает вся история человечества еще до появления Ленина. «*После того, не значит вследствие того*».

Далее, аргумент относительно предсказательности следует понимать в контексте всей остальной информации о свойствах объектов. Т.е. наличие группировок при анализе одного набора признаков означает высокую вероятность аналогичных группировок и при использовании других наборов.

ВШ: Давайте поближе к реальности. Допустим, что пять друзей-гидробиологов вышли на берег Женевского озера и в 25 его точках отобрали пробы зообентоса. Получили прямоугольную матрицу наблюдений и никаких особенных априорных вероятностей при этом не испытали. И решили выполнить кластеризацию объектов. Один исследователь с традиционным образованием сформировал матрицу коэффициентов сходства 25x25, используя классическую в гидробиологии меру сходства Съеренсена, учитывающую только факт присутствия или отсутствия каждого вида бентоса. Другой, прочитавший на пару книг больше, рассчитал ту же матрицу, применив традиционное для программы STATISTICA евклидово расстояние численностей видов. Третий – перевел численности видов в некоторые баллы обилия и избрал манхэттенское расстояние. Четвертый вычислил коэффициенты парной корреляции рангов по Спирмену. Пятый – используя метод главных компонент, свернул пространство видов до 5 главных факторов и вычислил евклидово расстояние в пространстве факторов. Получилось пять матриц сходства с некоторой вариабельностью предупорядоченности изученных объектов и, следовательно, пять различных деревьев иерархической классификации одних и тех же станций (даже расклассифицированных одним методом!). Предположим, они начали спорить, отстаивая каждый свою точку зрения и вызывая тень отца Байеса в надежде получить от него априорные вероятности. Но их дискуссия оказалась тщетной, поскольку не нашлось критерия, который бы оценивал, какой "результат является оптимальным (см. В. Леонова)", а тень отца Байеса безмолвствовала.

Выход я вижу один – считать все пять деревьев результатом работы 5 экспертов и искать синтез некоторого коллективного разбиения. Но, для этого необходимо научиться оценивать, насколько близки друг другу два результата классификации, т.е. два дерева. Это дает возможность некоторому обработчику результатов сказать «*ребята, вы напрасно спорите, все ваши деревья можно считать эквивалентными*» или «*вот вы трое получили очень близкие деревья, а приверженцы, предположим, Съеренсена и Спирмена могут пока отдохнуть*».

ГР: Действительно, получить картину "порядка" (априорные вероятности) зачастую невозможно, а вот создать "беспорядок" кажется более простой задачей. Задача автоматической классификации всегда может быть сведена к задаче диагонализации матрицы связи классифицируемых объектов. Тогда в матрице "порядка" вдоль диагонали стоят выделенные "блоки" сходных объектов. Можно представить процедуру диагонализации исходной матрицы как определение такой последовательности объектов (или признаков), которая бы минимизировала некоторую меру уклонения (*меру диссонанса*, [Лефевр, 1973; Розенберг, 1975]) этой матрицы от эталонной. Вот здесь, в качестве эталона и появляется "беспорядок" – исходная матрица, все строки и столбцы которой перепутаны случайным образом. Лучшей можно будет считать ту классификацию, для которой мера диссонанса между организованной матрицей и эталонной матрицей "беспорядка" (или средней из нескольких матриц "беспорядка") будет больше. Ясно, что предлагаемая процедура – эвристична, но ничего другого мне не известно…

ВЛ: Я не исключаю, что в теории графов возможно и есть такие варианты сравнения. Лично мне они также не известны. Однако я не могу согласиться с тем, что выход только один или два. При анализе данных нами выработан своеобразный стандартный набор приемов. Например, лично я в аналогичных ситуациях очень часто использую непараметрический дискриминантный анализ с локальной оценкой плотности распределения. Среди этих приемов есть и обязательный анализ таблиц сопряженности при кластеризации по количественным признакам и по дискретным качественным. Так что Вы зря иронизируете относительно предказательности!

ВШ: Насколько я понимаю, частотные таблицы сопряженности могут быть использованы, когда (и только когда) на входе имеется матрица наблюдений из двух признаков, измеренных в порядковых шкалах не слишком большой дискретности ($n \times m < 100$, где n и m – число градаций признаков). Как их можно использовать для сравнения классификаций, мне абсолютно непонятно.

ВЛ: Я имел в виду многомерные таблицы сопряженности. Обозначаем принадлежность отдельных наблюдений отдельным ветвям и узлам каждого из 5 полученных деревьев. Вот Вам и 5-мерная таблица сопряженности. Далее вполне возможно в этом случае использовать, например, анализ соответствий Бензекри-Гуттмана.

ВШ: Нет ли более или менее полной публикации, описывающей технологию таких расчетов?

ВЛ: Нет, такой публикации нет. Каждая работа над отдельным массивом – это всегда творческая работа и комбинации методов, как правило, отличаются друг от друга. Различаются как исходные массивы

и задачи исследования, так и доступный набор статистических методов (пакетов). Спектр последних со временем расширяется – например, сейчас я активно использую через telnet статистические пакеты, установленные в США, в частности в Лос-Анджелесе. И хотя за это удовольствие приходится платить в USD, но зато имеешь доступ к мощным средствам.

Надо сказать, что анализ таблиц сопряженности вообще недооценивают большинство исследователей. А, между тем, он очень многое может идентифицировать! Анализ таблиц сопряженности можно (и нужно!) использовать для любых видов дискретных признаков. Сам анализ с виду прост и несложен, но его нюансы видны только профессионалу, глубоко знающему теорию метода. Прелести этого анализа – в переходе от 2-х признаков к большему их числу. Вычислительные сложности здесь огромнейшие, а программных реализаций отбора наиболее оптимальных вариантов нет. И хотя теория известна уже более 100 лет, большинство пакетов дальше самых элементарных операций этой теории не идут. Я сейчас пытаюсь сделать некоторые надстройки для более глубокого анализа таких таблиц. Но дело это очень трудоемкое и необходимо много времени потратить на программирование и отладку этих алгоритмов.

ГР: Здесь следует остановиться еще на одном аспекте количественных методов (не только в гидроэкологии) – то, что я называю «индексологией». Во второй главе мы говорим о том, что «*весь смысл математической обработки многомерных таблиц наблюдений заключается… в редукции данных или понижении размерности признакового пространства типа "объект-признак"*». Такое "схлопывание информации" по В.В. Налимову приводит к применению и в научных исследованиях, и в действующих методиках биологического мониторинга множества разнообразных индексов, за которыми пытаются закрепить либо новую информационную сущность (индекс Шеннона – хотя, это не так), либо обобщение квалифицированных экспертов (индекс Вудивисса), либо "нечто этакое…" (подавляющее большинство "придуманных" индексов). И здесь в мнениях не сходятся даже соавторы. Так, Владимир Кириллович ратует за использование, в частности, индекса $(\boldsymbol{N}*\boldsymbol{B})^{0,5}$, где $\boldsymbol{N}$ – численность, а $\boldsymbol{B}$ – биомасса организмов (см. стр. 66). В разделе 2.1 мы отмечали, что сложные системы обладают простыми (аддитивными) и сложными (неаддитивными) свойствами. В рамках "индексологии" иногда удается описать тем или иным эвристическим способом именно простые свойства сложных систем. Самый наглядный пример – введение Г. Омом без колебаний и особенных теоретических обоснований показателя «сопротивление» путем деления напряжения (в вольтах) электрической цепи на силу тока (в амперах), что позволяет нам в течение 175 лет успешно пользоваться этой никем не измеренной и имеющей (в данном контексте) сомнительный "реальный" смысл величиной. Этот показатель (как и величина индекса $(\boldsymbol{N}*\boldsymbol{B})^{0,5}$) были бы более "реальными", если бы они "вытекали" из некоторой оптимизационной модели. Например, $(\boldsymbol{N}*\boldsymbol{B})^{0,5}$ является максимумом функции $\mathbf{2/3}(\boldsymbol{N}*\boldsymbol{B})^{3/2}$. К сожалению (к счастью?) я не могу придумать модель, которая приводила бы именно к такому виду "взаимодействия" численности и биомассы…

Нельзя забывать и о том, что различные индексы сопряженности и сходства отражают различные "аспекты связи". В книге мы говорим об этом, но еще раз поясню. Например, коэффициент линейной корреляции (в показателях 4-хпольной таблицы; см. формулу 7.3) существенным образом зависит от, так называемого, $\boldsymbol{d}$-эффекта, т.е. от числа описаний, в которых оба сравниваемых вида отсутствуют. Причем, этот эффект достигается либо когда виды действительно "не хотят" встречаться по каким-то экологическим соображениям, либо у них существенно различается встречаемость (пальма и берёза в Тольятти). Таким образом, если мы определяем корреляцию между длиной хвоста и шириной ушей у кошек, мы должны быть уверены, что у всех наших объектов есть и уши, и хвост. К сожалению (для коэффициента корреляции), виды в природе имеют разную встречаемость и достичь значимой сопряженности можно увеличивая значение клетки $\boldsymbol{d}$.

ВШ: В разделе 8.4 мы предлагаем некоторый новый алгоритм оценки индикаторной значимости видов на основе матрицы $\boldsymbol{P}$ частот встречаемости $\boldsymbol{i}$-го признака в $\boldsymbol{k}$-м классе. Я не знаю никаких литературных ссылок, где бы описывался сколько-нибудь похожий алгоритм классификации, кроме нескольких малоизвестных работ и реально действующей системы PASS. К какому разделу распознавания образов или иной математики относится этот метод – нельзя ли помочь с терминологическими тонкостями?

ДФ: Раскладывание по полочкам – это, скорее, в корпоративных интересах узких специалистов. Опыт показывает, что одну и ту же, абсолютно одинаковую математику разные кланы от наук называют, записывают и интерпретируют по-разному, хотя практические результаты получаются одинаковыми. Например, есть частотная вероятность по Мозесу или аксиоматическая по Колмогорову, а есть ещё, кажется, Шметтерер, который вообще всю статистику выводит из определения средней величины, употребляя слово «вероятность» только, разве что, в ругательном смысле. А квантовая механика по Шрёдингеру, по Гейзенбергу, с интегралами по траекториям Фейнмана? Где даже наши доблестные философы немало копий наломали, причем и по судьбам ученых! Но потом математики напряглись и доказали абсолютную тождественность всех версий, так

что квантовая механика теперь у нас одна. Таких примеров не счесть! Да и этих "полкораспределителей" сколь угодно – чего стоит один многотомник Айвазяна "Прикладная статистика", где каждое третье слово непонятно, а в каждой второй формуле – ошибка.
Все ваши алгоритмы вытекают непосредственно из байесовского подхода (именно, подхода, а не формулы, т.к. подход – это целая идеология, включающая и формулу Байеса). На самом деле, мне очень трудно придумать случай, не сводимый к байесовскому подходу, потому что надлежащим выбором платёжной матрицы и априорной вероятности можно добиться почти всего. Например, можно изложить в свете этой идеологии метод максимума правдоподобия.

NN: (*Из рецензии журнала "Известия Академии наук", сер. биол. на отклоненную статью по материалам раздела 8.4*). Все же: чем предлагаемый метод "обобщенного портрета" лучше, чем хорошо и давно известный метод дискриминантного анализа, который содержится во всех пакетах статистической обработки данных (STATISTICA, SPSS, STATGRAPHICS и т.п.). У меня создается впечатление, что определенными преимуществами обладает именно дискриминантный анализ.

ВШ: Чтобы расширить спектр впечатлений, попробуйте на досуге выполнить дискриминантный анализ сильно разреженной матрицы размерностью, скажем, 250x300, которая на 95% заполнена нулями (а именно таковы, чаще всего, гидробиологические, фитоценологические и пр. данные). Негативные эмоции Вам гарантированы.
Кроме того, нельзя ограничивать круг разрешаемых к использованию математических методов только пакетами общего назначения, иначе остановится прогресс (громко сказано, но верно по сути). Кстати, индекс Шеннона или меру сходства по Сьеренсену тоже нельзя рассчитать ни в пакете STATISTICA, ни в других, но они почему-то широко используются в биологии.

NN: Судя по вашему тексту, метод обобщенного портрета, кажется, вообще игнорирует существование теории вероятностей и статистики и даже метода наименьших квадратов.

ВШ: Действительно, он основан на другой теоретико-вероятностной теореме – теореме Гливенко о равномерной сходимости частот появления событий к их вероятностям, которую математическая наука считает не менее глубокой, чем, например, принцип максимального правдоподобия. Строгому обоснованию метода обобщенного портрета посвящено несколько сот страниц [Вапник, Червоненкис, 1974], в том числе, вполне корректной проработке теорем о вероятности ошибки коэффициентов при построении оптимальной разделяющей гиперплоскости. Полагая, что ученым-экологам будут не слишком интересны чисто математические аспекты, мы сделали акцент исключительно на практической реализации метода, отослав желающих к специальной литературе (к сожалению, таких желающих не оказалось даже среди рецензентов).

ВЛ: Кстати, сейчас большое развитие в области распознавания образов получают *методы ресэмпинга* для получения асимптотических оценок. Кроме МНК есть и другие методы оценок параметров уравнений – наименьших модулей, ортогональная регрессия и пр. Математика и статистика продолжают бурно развиваться!..

NN: Фраза на с. 10 (*в рецензируемой статье. – Ремарка наша*): «*Модель смогла построить решающее правило, только предварительно исключив из рассмотрения 14 объектов обучающей выборки*» – вызывает, по меньшей мере, изумление. Это все равно, что при регрессионном анализе выбрасывать все точки, не укладывающиеся на аппроксимируемую кривую!

ВШ: Это исключение так же естественно, как выход некоторых экспериментальных точек за пределы доверительного интервала ("трубы", а не аппроксимирующей кривой) относительно линии регрессии. Не правда ли, Натан?

НЦ: Выход экспериментальных точек за пределы доверительного интервала, построенного вокруг линии регрессии, совершенно правомерен. Доверительный интервал строится не для экспериментальных точек, а для условного математического ожидания отклика. Большинство (скажем, 99%) экспериментальных точек лежат в пределах области допустимых значений отклика (с заданной вероятностью, скажем, 95%). Но и то некоторые точки «имеют право» выпрыгивать за границы этой области. И только 1 – 2 «выброса» могут вызвать у нас подозрение в том, что они не принадлежат к изучаемой совокупности. Их наличие проверяют по критериям «выброса» (например, Граббса).

ВШ: Так и тут, невозможно разделить плоскостью взаимопроникающие "облака" точек двух классов, не отбросив в ходе оптимизирующих операций несколько (~ 5%) точек, препятствующих разделению.

NN: Приведенный пример для р. Чапаевка (бывшая Мóча) выглядит не слишком убедительным: на протяжении всего нескольких десятков километров по течению реки на станциях с 6 по 9 "экологическая стабильность" дважды сменяется на "экологический кризис" и обратно и при этом авторы равнодушно замечают, что на этом участке реки «*...происходит естественное и стабильное развитие зообентоса, несмотря на отмеченные там процессы антропогенного эвтрофирования*».
Одномоментная съемка не может в принципе дать оценку стабильности, если иметь в виду относительное постоянство структурных и функциональных характеристик системы *во времени*.

Представляется также крайне сомнительным, что о "развитии зообентоса" можно судить только по видовому составу личинок хирономид. И почему появление в зообентосе, скажем, *Glyptotendipes glaucus* и *Procladius ferrugineus* – это признак неестественного развития зообентоса ("экологический кризис"), а обнаружение в пробах *Ablabesmia monilis* и *Cricotopus bicinctus* – это признак благополучия?

ТЗ: Хотелось бы в этот хор статистиков добавить и слово практика-гидроэколога.

В рамках нашей книги мы не стремились (да и не смогли бы) убедить педантичного читателя ни в отменных биоиндикационных свойствах отдельных видов личинок хирономид (хотя это – установленный в биологии факт), ни в возможности изменения видовой структуры зообентоса по течению реки среднего размера (хотя абсолютно не понятно, кто способен сомневаться в том, что существует не только временна́я динамика, но и пространственная динамика [неоднородность] экосистемы от истока к устью).

А если уж конкретно по четырем названным видам, то первые два индицируют процесс эвтрофирования участков р. Чапаевки, а вторые, реафильные виды характерны для водоемов с проточной и чистой водой. По-настоящему заинтересованный читатель может обратиться к:

- обширной литературе по проблемам и методам биоиндикации и оценке качества вод по зообентосу (в том числе и по хирономидам – например, [Зинченко, 2002]);
- цитируемым первоисточникам по реке Чапаевка (Мóча), чтобы по подробнее познакомиться с изучаемым объектом и объемом проведенных исследований (см., например, [Экологическое состояние.., 1994]).

NN: Авторы так и не определили, что именно они имеют в виду под терминами «качество экосистемы», «стабильность экосистемы» и «кризисность экосистемы». В отсутствие этих определений остается совершенно непонятым, какое отношение к этим терминам, допускающим далеко не однозначное толкование, имеет сугубо математическая конструкция – расстояние от точки в пространстве признаков до разделяющей гиперплоскости.

ВШ: А какое, кстати, имеет отношение стрелка пружинного барометра к такому ёмкому и неоднозначному термину, как «погода»? Если бы мы могли точно сформулировать термины «качество», «стабильность» и «кризисность экосистемы» и предложить фундаментальные способы их количественной оценки, экологическую науку можно было бы закрывать за неимением проблемных тем. Эти термины мы использовали не в рамках "мирозданья вообще", а исключительно в рамках *предположений нашей модели*: чем больше величина отклика модели, тем выше качество экосистемы.

ТЗ: А все-таки, нельзя ли определеннее сказать, как мы относимся к "индексам": на протяжении всей книги мы их то ругаем, то (тут же) признаем их прогностическую силу…

ВШ: Как отдельный феномен, каждый индекс (и любой другой частный показатель) ограничен и малоэффективен, но в рамках многофакторных моделей их роль весьма позитивна.

ВЛ: Уверен, что раздумья над противоречиями и сопоставление аргументов и контраргументов – лучший способ найти некую новую систему взглядов на любую проблему, в том числе и эту.

СПИСОК ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ И ИНТЕРНЕТ-ССЫЛОК

Абакумов В.А. Продукционные аспекты биомониторинга пресноводных экосистем // Продукционно-гидробиологические исследования водных экосистем. – Л.: Наука, 1987. С. 51-61.

Абакумов В.А., Максимов В.Н., Ганьшина Л.А. Экологические модуляции как показатель изменения качества воды // Научные основы контроля качества вод по гидробиологическим показателям: Тр. Всес. конф. – Л., 1981. С. 117-136.

Абакумов В.А., Сущеня Л.М. Гидробиологический мониторинг пресных вод и пути его совершенствования // Экологические модификации и критерии экологического нормирования: Тр. Междунар. симпоз. – Л.: Гидрометеоиздат, 1991. С. 41-51.

Абросов Н.С., Боголюбов А.Г. Экологические и генетические закономерности сосуществования и коэволюции видов. – Новосибирск: Наука, 1988. – 333 с.

Абросов Н.С., Ковров Б.Г., Черепанов О.А. Экологические механизмы сосуществования и видовой регуляции. – Новосибирск: Наука, 1982. – 301 с.

Авакян А.Б., Салтанкин В.П., Шарапов В.А. Водохранилища. – М.: Мысль, 1987. – 325 с.

Аверкин А.Н., Батыршин И.З., Блишун А.Ф. и др. Нечеткие множества в моделях управления и искусственного интеллекта // Под ред. Д.А. Поспелова. – М.: Наука, 1986. – 312 с.

Аветисян Д.О. Проблемы информационного поиска: (Эффективность, автоматическое кодирование, поисковые стратегии) - М.: Финансы и статистика, 1981. - 207 с.

Авидон В.В., Аролович В.С. , Козлов С.П., Пирузян Л.А. Применение подструктурного анализа для скрининга биологически активных соединений // Химико-фармацевтический журнал. 1978. № 5. С. 88-94.

Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. – М.: Наука, 1976. – 279 с.

Азгальдов Г.Г. Общие сведения о методологии квалиметрии // Стандарты и качество. 1994. №.11. С. 24-35.

Азгальдов Г.Г., Райхман Э. П. О квалиметрии. – М.: -Стандарты , 1973.

Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с.

Айвазян С.А., Буштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерностей. – М.: Финансы и статистика, 1989. – 607 с.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. Справочное издание. – М.: Финансы и статистика, 1983. – 472 с.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей: Справочник. – М.: Финансы и статистика, 1985. – 182с.

Айвазян С.А. , Мхитарян В.С. Прикладная статистика и основы эконометрики. – М. Юнити, 1998. – 1024 с.

Айвазян С.А, Степанов В.С. Инструменты статистического анализа данных // Мир ПК. 1997. № 8. Расширенная Интернет-версия по адресу http://www.tvp.ru/prog/progrfr.htm.

Айвазян С.А, Степанов В.С. Программное обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка // ТВП (Теория Вероятностей и ее Применения). Адрес в Интернет http://www.cemi.rssi.ru/rus/publicat/e-pubs/ep97001t.htm

Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. – М.: Наука, 1970. – 384 с.

Алгоритмы и программы восстановления зависимостей. – М.: Наука, 1984. – 816 с.

Алекин О.А. К вопросу о химической классификации природных вод // Вопросы геохимии. 1946. С. 14-35.

Алекин О.А. Основы гидрохимии. – Л.: Гидрометеоиздат, 1970. – 296 с.

Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. – Л.: Наука, 1983. – 208 с.

Александров В.В., Шнейдеров В.С. Обработка медико-биологических данных на ЭВМ. – Л.: Медицина, 1984. – 160 с.

Алексахин С.В. и др. Прикладной статистический анализ данных. Теория. Компьютерная обработка. Области применения. В 2-х томах. – М.: ПРИОР, 2002. – 688 с.

Алексеев В.А. Основы биоиндикации качества вод на уровне организмов // Водн. ресурсы. 1984а. № 2. С. 107-121.

Алексеев В.А. Система токсобности и ее место в унифицированной системе качества вод СССР // Водн. ресурсы. 1984б. № 5. С. 76-87.

Алимов А.Ф. Интенсивность обмена у пресноводных двустворчатых моллюсков // Экология. 1975. № 1. С. 10-20.

Алимов А.Ф. Структурно-функциональный подход к изучению сообществ водных животных // Экология. 1982. № 3. С. 45-51.

Алимов А.Ф. Исследования биотических балансов экосистем пресноводных водоемов в СССР // Гидробиол. журн. 1987. Т. 23. № 6. С. 7-10.

Алимов А.Ф. Общие основы учения биологической продуктивности водоемов // Гидробиол. журн. 1988. Т. 24. № 3. С. 40-51.

Алимов А.Ф. Введение в продукционную гидробиологию. – Л.: Наука, 1989. – 152 с.

Алимов А.Ф. Основные положения теории функционирования водных экосистем // Гидробиол. журн. 1990. Т. 26. № 6. С. 7-13.

Алимов А.Ф. Элементы теории функционирования экосистем. – СПб.: ЗИН РАН, 2000. – 147 с.

Алимов А.Ф., Балушкина Е.В., Умнов А.А. Подходы к оценке состояния водных экосистем // Экологическая экспертиза и критерии экологического нормирования. – СПб.: СПбНЦ РАН, 1996а. С. 37 - 47.

Алимов А.Ф., Бульон В.В., Гутельмахер Б.Л., Иванова М.Б. Методы изучения участия гидробионтов в процессах самоочищения водоемов // Роль гидробионтов в очистке сточных вод. – Фрунзе: Илим, 1977. С. 3-42.

Алимов А.Ф., Голубков С.М., Панов В.Е. Закономерности функционирования и стратегия управления экосистемами эстуария реки Невы // Экологическое состояние водоемов и водотоков бассейна реки Невы. – СПб.: СПбНЦ РАН, 1996б. С. 187-203.

Алимов А.Ф., Финогенова Н.П. Количественная оценка роли сообщества донных животных в процессах самоочищения пресноводных водоемов // Гидробиологические основы самоочищения вод. – Л.: ЗИН АН СССР, 1976. С. 5-14.

Андерсен Т. Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с.

Андреев В.Л. Системы классификации в биогеографии и систематике (детерминистские методы) // Иерархические классификационные построения в географической экологии и систематике. – Владивосток: ДВНЦ АН СССР, 1979а. С. 3-59.

Андреев В.Л. Статистические методы классификационных построений в биогеографии и систематике // Иерархические классификационные построения в географической экологии и систематике. – Владивосток: ДВНЦ АН СССР, 1979б. С. 60-96.

Аптон Г. Анализ таблиц сопряженности. – М.: Финансы и статистика, 1982. – 144 с.

Аркадьев А.Г., Браверман Э.М. Обучение машины классификации объектов. – М.; Наука, 1971. – 192 с.

Арманд А.Д. Информационные модели природных комплексов. – М.: Наука, 1975. – 126 с.

Арнольд В.И. Теория катастроф. – М.: Наука, 1990. – 128 с.

Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982. – 488 с.

Ащепкова Л.Я. Математические модели водных экосистем (обзор) // Математическое моделирование водных экологических систем. – Иркутск: ИГУ, 1978. С. 6-46.

Баевский Р.М. Прогнозирование состояния на грани нормы и патологии. – М.: Медицина, 1979. – 157 с.

База эколого-экономических данных крупного региона (Методическое пособие). – Тольятти: ИЭВБ РАН, 1991. – 62 с.

Базыкин А.Д. Математическая биофизика взаимодействующих популяций. – М.: Наука, 1985. – 180 с.

Баканов А.И. Обзор существующих подходов к районированию водохранилищ // Экологическое районирование пресноводных водоемов. Труды ИБВВ АН СССР, вып. 62(65). – Рыбинск: ИБВВ АН СССР, 1990. С. 3-16.

Баканов А.И. Способ ранжирования гидробиологических данных в зависимости от экологической обстановки в водоеме // Биол. внутр. вод. 1997. № 1. С. 53-58.

Баканов А.И. Использование комбинированных индексов для мониторинга пресноводных водоемов по зообентосу // Водн. ресурсы. 1999. Т. 26. № 1. С. 108-111.

Баканов А.И. Использование зообентоса для мониторинга пресноводных водоемов // Биол. внутр. вод. 2000а. № 1. С. 68-82.

Баканов А.И. О некоторых методологических вопросах применения системного подхода для изучения структур водных экосистем // Биол. внутр. вод. 2000б. № 2. С. 5-18.

Балушкина Е.В. Функциональное значение личинок хирономид / Тр. Зоол. ин-та АН СССР. Т. 142. – Л.: Наука, 1987. – 179 с.

Балушкина Е.В. Применение интегрального показателя для оценки качества вод по структурным характеристикам донных сообществ // Реакция озерных экосистем на изменение внешних условий. – СПб.: ЗИН РАН, 1997. С. 266-292.

Балушкина Е.В. Структура сообществ донных животных и оценка экологического состояния р. Ижоры: оценка качества вод р. Ижоры по структурным характеристикам донных животных в разные годы // Биол. внутр. вод. 2002. № 4. С. 61-68.

Балушкина Е.В., Финогенова Н.П. Структурные характеристики зообентоса как основа оценки состояния экосистем Невской губы и восточной части Финского залива // Структурно-функциональная организация пресноводных экосистем разного типа. Труды ЗИН РАН. Т. 279. – СПб.: Наука, 1999. – С. 269 - 292.

Балушкина Е.В., Финогенова Н.П., Слепухина Т.Д. Изменение характеристик зообентоса в системе Ладога – р. Нева – Невская губа – восточная часть Финского залива // Экологическое состояние водоемов и водотоков бассейна р. Невы. – СПб.: ЗИН РАН, 1996. С. 91-100.

Банди Б. Методы оптимизации. – М.: Радио и связь, 1988. – 178 с.

Барабаш Ю.Л. и др. Вопросы статистической теории распознавания / Под ред. Б.В. Варского. – М.: Советское радио, 1967. – 400с.

Батоян В.В. Решение задач геохимии ландшафтов и почвоведения с применением математических методов. – М.: МГУ, 1983. – 120 с.

Башкин В.Н. Оценка степени риска при критических нагрузках загрязняющих веществ на экосистемы // География и природные ресурсы. 1999. № 1. С. 35-39.

Бейм А.М., Красовский Г.И., Сутокская И.В, Васюкович Л.Я. Эколого-гигиенические подходы к биоиндикации качества воды // Самоочищение воды и миграции загрязнений по трофической цепи. – М.: Наука, 1984. С. 16-22.

Беляев В.И. Теория сложных геосистем. – Киев: Наук. думка, 1978. – 155 с.

Беляев В.И., Ивахненко А.Г., Флейшман Б.С. Имитация, самоорганизация и потенциальная эффективность // Автоматика. 1979. № 6. С. 9-17.

Беляев М.П. Справочник предельно допустимых концентраций вредных веществ в пищевых продуктах и среде обитания. – М.: Госсанэпиднадзор, 1993. – 141 с.

Берман Д.И., Виленкин Б.Я. Некоторые принципы исследования сообществ // Количественные методы в экологии и биоценологии животных суши. – Л.: Наука, 1975. С. 10-12.

Бершадский А., Новик К., Новицкий Д. и др. Азбука многомерного анализа (в картинках) / Научный руководитель проекта Л.Н. Столяров. – М.: Студенческая Лаборатория "Computer Science" МФТИ, ПрограмБанк, 1999. – Адрес в Интернет http://www.cslab.mipt.ru

Беспамятнов Г.П., Кротов Ю.А. Предельно допустимые концентрации химических веществ в окружающей среде. – Л.: Химия, 1985. – 528 с.

Бешелев С.Д., Гурович Ф.Г. Математико-статистические методы экспертных оценок. – М.: Статистика, 1980. – 263 с.

Бигон М., Харпер Дж., Таунсенд К. Экология: Особи, популяции, сообщества: В 2-х т. – М.: Мир, 1989. Т. 1. – 667 с.; Т. 2. – 477 с.

Бикел П., Доксам К. Математическая статистика. – М.: Финансы и статистика, 1983. Вып. 1. - 278 с.; Вып. 2. – 254 с.

Биоиндикация природных водоемов (телекоммуникационный проект) / ГУ ЯО "Центр телекоммуникаций и информационных систем в образовании", 2002. Адрес в Интернет www-windows-1251.edu.yar.ru/russian/misc/eco_page /bioind

Биоиндикация: теория, методы, приложения / Под ред. Г.С. Розенберга. – Тольятти: Изд-во Интер-Волга, 1994. – 266 с.

Боголюбов А.Г. Столетие биометрии в России // Изв. СамНЦ РАН. 2002. Т. 4. № 2. С. 189-196.

Богословский Б.Б. Озероведение. – М.: МГУ, 1960. – 335 с.

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: ВЦ АН СССР, 1968 (2-е изд.). - 474 с.

Бонгард М.М. Проблема узнавания. – М.: Наука, 1967. – 320 с.

Бондаренко В.В. 48 часов с Nostalgie. – Самара: Изд. Дом "Бахрах-М", 2002. – 240 с.

Борисов А.Н., Алексеев А.В., Крумберг О.А. и др. Модели принятия решений на основе лингвистической переменной. – Рига: Зинатне, 1982. – 256 с.

Борисов А.Н., Алексеев А.В., Меркурьева Г.В. и др. Обработка нечеткой информации в системах принятия решений. – М: Радио и связь. 1989. – 304 с.

Боровиков В.П. Statistica для студентов и инженеров. – М.: КомпьютерПресс, 2001. – 301 с.

Браверман Э.М., Мучник Ч.Б. Структурные методы обработки эмпирических данных. – М.: Наука, 1983. – 464 с.

Брагинский Л.П. Экологические подходы к исследованию механизмов действия токсикантов в водной среде. // Формирование и контроль качества поверхностных вод. Вып. 1. – Киев: Наук. думка, 1975. С. 5—15.

Брагинский Л.П. Теоретические аспекты проблемы нормы и патологии в водной экотоксикологии. // Теоретические вопросы водной токсикологии. – Л.: Наука, 1981. С. 29-40.

Брагинский Л.П. Некоторые принципы классификации пресноводных экосистем по уровням токсической загрязненности // Гидробиол. журн. 1985. Т. 21. № 6. С. 65-74.

Брагинский Л.П., Комаровский Ф.Я, Мережко А.И. Персистентные пестициды в экологии пресных вод. – Киев: Наук. думка, 1979. – 140 с.

Браунли К. А. Статистическая теория и методология в науке и технике. - М.: Наука, 1977. - 408 с.

Бреховских В.Ф. Гидрофизические факторы формирования кислородного режима водоемов. – М.: Наука, 1988. – 168 с.

Брусиловский П.М. Становление математической биологии. – М.: Знание, 1985. – 62 с.

Брусиловский П.М. Коллективы предикторов в экологическом прогнозировании. – Саратов: Изд-во Сарат. ун-та, 1987. – 104 с.

Брусиловский П.М., Розенберг Г.С. О возможности построения модели, удовлетворительно описывающей колебания в одной реальной системе хищник-жертва // Динамика эколого-экономических систем. – Новосибирск: Наука, 1981а. С. 84-91.

Брусиловский П.М., Розенберг Г.С. Проверка неадекватности имитационных моделей динамической системы с помощью алгоритмов МГУА // Автоматика. 1981б. № 6. С. 43-48.

Брусиловский П.М., Розенберг Г.С. Модельный штурм при исследовании экологических систем // Журн. общ. биол. 1983. Т. 44, № 2. С. 266-274.

БСЭ – Большая Советская Энциклопедия / РУБРИКОН - информационно-энциклопедический проект компании «Русс. портал». Адрес в Интернет http://www.rubricon.ru/bse_1.asp

Будилова Е.В., Дрогалина Ж.А., Терехин А.Т. Основные направления современной экологии и ее математический аппарат: анализ публикаций // Журн. общ. биол. 1995. Т. 56. № 2. С. 179-189.

Букатова И.Л. Эволюционное моделирование и его приложения. – М.: Наука, 1979. – 232 с.

Букатова И.Л., Михасев Ю.И., Шаров А.М. Экоинформатика. Теория и практика эволюционного моделирования. – М.: Наука, 1991. – 206 с.

Булгаков Г.П. Принципы оценки качества текучих вод Узбекистана с помощью МБМ // Тр. Среднеазиат. регион. н.-и. гидрометеорол. ин-та. 1989. № 135. С. 13-21.

Булгаков Н.Г., Дубинина В.Г., Левич А.П., Терехин А.Т. Метод поиска сопряженностей между гидробиологическими показателями и абиотическими факторами среды (на примере уловов и урожайности промысловых рыб) // Изв. РАН. Сер. биол. 1995. № 2. С. 218-225.

Бульон В.В. Первичная продукция планктона внутренних водоемов. – Л.: Наука, 1983. – 151 с.

Бульон В.В. Закономерности первичной продукции в лимнических экосистемах. - СПб.: Наука, 1994. – 222 с.

Бурдин К.С. Основы биологического мониторинга. – М.: МГУ, 1985. – 158 с.

Былинкина А.А., Драчев С.М., Ицкова А.И. О приемах графического изображения аналитических данных о состоянии водоема // Материалы 16-го совещ. Гидрохим. ин-та АН СССР. – Новочеркасск: АН СССР, 1962.

Вайну Я.Я.-Ф. Корреляция рядов динамики. – М.: Статистика, 1977. – 119 с.

Вайнцвайг М.Н. Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов / Под ред. В.Н. Вапника. – М.: Сов. радио, 1973. С. 110-116.

Вайнштейн Б.А. Об оценке сходства между биоценозами // Биология, морфология и систематика водных организмов. – Л.: ЗИН АН СССР, 1976. С. 156-164.

Ван дер Варден Б.Л. Математическая статистика. – М.: Изд-во иностр. лит., 1960. – 302 с.

Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – М.: Наука, 1974. – 487 с.

Василевич В.И. О количественной мере сходства между фитоценозами // Проблемы ботаники. Т. 6. – М.; Л.: Наука, 1962. С. 83-94

Василевич В.И. Статистические методы в геоботанике. – Л.: Наука, 1969. – 232 с.

Василевич В.И. Количественные методы изучения структуры растительности // Итоги науки и техники. Ботаника. Т. 1. – М.: Изд-во ВИНИТИ, 1972. С. 7-83.

Василевич В.И. Очерки теоретической фитоценологии. – Л.: Наука, 1983. – 248 с.

Василевич Н.Г. Системный подход и методологические проблемы экологии // Методология частных наук. – Иркутск, 1979. С. 104-112.

Васильев В.И. Распознающие системы: Справочник. – К.: Наук. думка, 1983. – 230с.

Васильев В.И., Ильясов Б.Г. Интеллектуальные системы управления с использованием нечеткой логики. Учеб. пособие. - Уфа: УГАТУ, 1995. - 80 с.

Васильев В.И., Ильясов Б.Г. Интеллектуальные системы управления с использованием генетических алгоритмов. Учебное пособие. – Уфа: УГАТУ, 1999. – 105 с.

Васильев В.И., Ильясов Б.Г., Валеев С.В., Жернавков С.В. Интеллектуальные системы управления с использованием нейронных сетей. Учебное пособие. – Уфа: УГАТУ, 1997. – 92 с.

Васильев В.И., Коноваленко В.В., Горелов Ю.И. Имитационное управление неопределенными объектами. – К.: Наукова думка, 1989. – 216 с.

Васильева Е.А., Виниченко В.Н, Гусева Т.В. и др. Как организовать общественный экологический мониторинг / Под ред. М.В. Хотулевой. – М.: СоЭС - Методический центр «Эколайн», 1998. Адрес в Интернет www.cci.glasnet.ru\mc\books\monitor

Васюкович Л.Я. Норма и патология // Теоретические проблемы водной токсикологии. – М.: 1983. С. 164-165.

Ватанабе С. Разложение Карунена–Лоэва и факторный анализ. Теория и приложения // Автоматический анализ сложных изображений. – М.: Мир, 1969. С. 239-253.

Вентцель Е.С. Теория вероятностей. – М.: Высш. шк., 1999. – 576 с.

Верхаген К., Дейн Р., Грун Ф и др. Распознавание образов: состояние и перспективы. – М.: Радио и связь, 1985. – 104 с.

Виленкин Б.Я. Взаимодействующие популяции // Математическое моделирование в экологии. – М.: Наука, 1978. С. 5-16.

Вильсон А.Дж. Энтропийные методы моделирования сложных систем. – М.: Наука, 1978. – 248 с.

Винберг Г.Г. Некоторые общие вопросы продуктивности озер // Зоол. журн. 1936. Т. 15. Вып. 4. С. 587-603.

Винберг Г.Г. Интенсивность обмена и пищевые потребности рыб. – Минск: Изд. Белорус. Ун-та, 1956. – 254 с.

Винберг Г. Г. Первичная продукция водоемов. – Минск: Изд-во АН БССР, 1960. – 328 с.

Винберг Г.Г. Зависимость энергетического обмена от массы тела у водных пойкилотермных животных // Журн. общ. биол. 1976. Т. 37. Вып. 1. С. 56-70.

Винберг Г. Г. Общие основы изучения водных экосистем. – Л.: Наука, 1979а. – 273 с.

Винберг Г.Г. Опыт применения разных систем биологической информации загрязнения вод в СССР // Влияние загрязняющих веществ на гидробионтов и экосистемы водоемов. – Л.: Наука, 1979б. С. 43-51.

Винберг Г.Г. Опыт применения разных систем биологической индикации загрязнения вод в СССР // Влияние загрязняющих веществ на гидробионтов и экосистемы водоемов. – Л., Наука, 1979в. С. 285-292.

Винберг Г.Г. Температурный коэффициент Вант-Гоффа и уравнение Аррениуса в биологии // Журн. общ. биол. 1983. Т. 44. № 1. С. 3-42.

Винберг Г.Г., Анисимов С.И. Математическая модель водной экосистемы // Фотосинтезирующие системы высокой продуктивности. – М.: Наука, 1966. С. 213-223.

Виттих В.А. Системообразующая функция интеграции знаний в организациях. Препринт № 2. – Самара: ИПУСС РАН, 1998. 20 с.

Виттих В.А. Инженерная эпистемология // Проблемы управления и моделирования в сложных системах: Труды III Международной конференции. – Самара: СамНЦ РАН , 2001. С. 92-100.

Водный кодекс Российской Федерации. Принят Государственной Думой 18 октября 1995 года. Адрес в Интернет http://www.medinfo.ru/ecolog/ek3.shtml

Волков И.В., Заличева И.Н., Ганина В.С. и др. О принципах регламентирования антропогенной нагрузки на водные экосистемы // Водн. ресурсы. 1993. Т. 20. № 6. С. 707-713.

Волков И.В., Заличева И.Н., Шустова Н.К. Ильмаст Т.Б. Есть ли экологический смысл у общефедеральных рыбохозяйственных ПДК? // Экология. 1996. № 5. С. 350-354.

Воробейник Е.Л., Садыков О.Ф., Фарафонтов М.Г. Экологическое нормирование техногенных загрязнений наземных экосистем. – Екатеринбург: УИФ Наука, 1994. – 380 с.

Временная типовая методика определения экономической эффективности осуществления природоохранных мероприятий и оценки экономического ущерба, причиняемого народному хозяйству загрязнением окружающей среды. – М.: Экономика, 1986. – 96 с.

Временные методические указания по комплексной оценке качества поверхностных и морских вод. Утв. Госкомгидрометом СССР 22.09.1986 г. № 250-1163. – М.: 1986. – 5 с.

Вудивисс Ф.С. Биотический индекс р. Трент. Макробеспозвоночные и биологическое обследование // Научные основы контроля качества поверхностных вод по гидробиологическим показателям. - Л.: Гидрометеоиздат, 1977. С. 132-161.

Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 239 с.

Выханду Л.К. Об исследовании многопризнаковых биологических систем // Применение математических методов в биологии. Т. III. – Л.: ЛГУ, 1964. С. 9-12.

Выхристюк Л.А., Зинченко Т.Д., Шитиков В.К. Комплексная оценка экологического состояния равнинных рек в условиях антропогенных воздействий // Научные аспекты экологических проблем России. – СПб.: Гидрометеоиздат, 2001. С. 70.

Гаазе-Рапопорт М.Г., Поспелов Д.А. От амебы до робота: модели поведения. – М.: Наука, 1987. – 285 с.

Габор Д. Перспективы планирования // Автоматика. 1972. № 2. С. 16-22.

Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука, 1971. – 376 с.

Гайдышев И.П. Анализ и обработка данных: специальный справочник. - СПб.: Питер, 2001. - 752 с.

Гегель Г. Энциклопедия философских наук. – М.: Наука, 1974. С. 240.

Гелашвили Д.Б., Безруков М.Е., Белышева О.И., Черников А.А. Количественные методы оценки кумулятивного и комбинированного действия ксенобиотиков // Экологический мониторинг. Ч. III: Методы биологического и физико-химического мониторинга. – Н. Новгород: ННГУ, 1998. С. 175-217.

Гелашвили Д.Б., Зинченко Т.Д., Выхристюк Л.А., Карандашова А.А. Интегральная оценка экологического состояния водных объектов по гидрохимическим и гидробиологическим показателям // Изв. СамНЦ РАН. 2002. № 2. С. 270-275.

Гелета И.Ф., Крауклис А.А. Применение метода главных компонент при анализе структуры и динамики наземного покрова таежных фаций // Моделирование и прогноз динамики геосистем. – Иркутск: Ин-т геогр. Сибири и Дал. Востока СО АН СССР, 1979. С. 111-132.

Генкин Я.Я. Новая информационная технология анализа медицинских данных. – СПб: Политехника, 1999. – 191 с.

Георгиевский В.Б. Идентификация и верификация моделей водных экосистем // Проблемы сохранения, защиты и улучшения качества природных вод. – М.: Наука, 1982. С. 156-163.

Герасименко Г.Г., Ипатов В.С. Анализ распределения обилия видов как метод классификации растительности // Бот. журн. 1980. Т. 65. № 5. С. 717-724.

Герасимов И.П. Научные основы современного мониторинга окружающей среды // Изв. АН СССР. Сер. геогр. 1975. № 3. С. 13-25.

Герасимов И.П. Мониторинг окружающей среды // Общие проблемы географии и моделирование геосистем: Тр. XXII Международного географического конгресса. – М., Наука, 1976. С. 15-34.

Гидробиологический режим малых рек в условиях антропогенного воздействия / Под ред. Андрушайтиса Г.П., Качаловой О.Л. – Рига: Зинатне, 1981. – 166 с.

Гильманов Т.Г. Математическое моделирование биогеохимических циклов в травяных экосистемах. – М.: МГУ, 1978. – 169 с.

Гиляров А.М. Соотношение биомассы и видового разнообразия в планктонном сообществе // Зоол. журн. 1969. Т. 48. № 4. С. 485-493.

Гинзбург Э.Х. Сравнение оценок показателя силы влияния // Генетика. 1969. Т. 5. № 4. С. 150-160.

Гладышев М.И. Основы экологической биофизики водных систем. – Новосибирск: Наука, 1999. – 113 с.

Глазовская М.А. Технобиогеомы – исходные физико-географические объекты ландшафтно-геохимического прогноза // Вест. МГУ. Сер. геогр. 1988. № 4. С.54-59.

Гмурман В.С. Теория вероятностей и математическая статистика. – М.: Высш. шк., 1972. – 368 с.

Гнеденко Б.В. Курс теории вероятностей: Учебник. – М.: Наука, 1988. – 380 с.

Голендер В.Е., Розенблит А.Е. Вычислительные методы конструирования лекарств. – Рига: Зинатне, 1978. – 179 с.

Головко В.А. Нейроинтеллект: Теория и применения. Книга 1 Организация и обучение нейронных сетей с прямыми и обратными связями. – Брест: БПИ, 1999. – 260 с. Книга 2 Самоорганизация, отказоустойчивость и применение нейронных сетей – Брест: БПИ, 1999. – 228 с.

Голубева Г.В. Использование хирономид в индикации качества воды малых рек Нечерноземной зоны РСФСР // Биоценология рек и озер Волжского бассейна. – Ярославль: ИБВВ АН СССР, 1985. С. 34-61.

Голубков С.М. Функциональная экология личинок амфибиотических насекомых // Тр. Зоол. ин-та АН СССР. Т. 284. – СПб.: Наука, 2000. – 294 с.

Горбань А.Н. Обучение нейронных сетей. – М.: Изд. СССР-США СП "ParaGraph", 1990. – 160 с.

Горбань А.Н. Нейроинформатика и ее приложения // Открытые системы. 1998а. № 4. Адрес в Интернет http://www.osp.ru/os/1998/04/05.htm

Горбань А.Н. Функции многих переменных и нейронные сети // Соросовский образовательный журнал. 1998б. № 12. С. 105-112.

Горбань А.Н., Дунин-Барковский В.Л., Кирдин А.Н. и др. Нейроинформатика. – Новосибирск: Наука. Сибирское предприятие РАН, 1998. – 296 с.

Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. – Новосибирск: Наука, 1996. – 276 с.

Горелик А.Л., Гуревич И.Б., Скрипкин В.А. Современное состояние проблемы распознавания. – М.: Радио и связь, 1985. – 160 с.

Горелик А.Л., Скрипкин В.А. Методы распознавания. – М.: Высш. шк., 1984. – 219 с.

Горстко А.Б., Эпштейн Л.В. Имитационная система «Азовское море» – инструмент анализа и прогнозирования // Математическое моделирование водных экологических систем. – Иркутск, ИГУ, 1978. С. 47-58.

ГОСТ 12.1.007-76. Система стандартов безопасности труда. Вредные вещества. Классификация и общие требования безопасности. – М.: Гос. комитет СССР по стандартам, 1976.

ГОСТ 14.413-80. Банки данных технологического назначения. Общие требования. – М.: Гос. ком. СССР по стандартам, 1980. – 3 с.

ГОСТ 17.1.1.01-77. Охрана природы. Гидросфера. Использование и охрана вод. Основные термины и определения. М: Гос. ком. СССР по стандартам, 1977.

ГОСТ 17.1.1.02-77. Охрана природы. Гидросфера. Классификация водных объектов. – М.: Гос. ком. СССР по стандартам, 1977.

ГОСТ 17.1.1.03-86 (СТ СЭВ 5182-85). Охрана природы. Гидросфера. Классификация водопользований. - М.: Гос. ком. СССР по стандартам, 1986.

ГОСТ 17.1.2.04-77. Охрана природы. Гидросфера. Показатели состояния и правила таксации рыбохозяйственных водных объектов. – М.: Гос. ком. СССР по стандартам, 1977.

ГОСТ 17.1.3.07-82. Охрана природы. Гидросфера. Правила контроля качества воды в водоемах и водотоках. – М.: Гос. ком. СССР по стандартам, 1982.

ГОСТ 17.1.5.02-80. Охрана природы. Гидросфера. Гигиенические требования к зонам рекреации водных объектов. – М.: Гос. ком. СССР по стандартам, 1980.

Государственный доклад "О состоянии окружающей природной среды Российской Федерации в 1995 г.". – М.: Центр междунар. проектов, 1996. – 458 с.

Гренандер У. Лекции по теории образов. – М.: Мир. Т. 1. Синтез образов. 1979. – 384 с.; Т. 2. Анализ образов. 1981. – 448 с.; Т. 3. Регулярные структуры. 1983. – 432 с.

Гродзинский М.Д. Эмпирические и формально-статистические методы определения областей допустимых и нормальных состояний // Научные подходы к определению норм нагрузок на ландшафты. – М.:Изд-во МГУ, 1988. С. 215-224.

Гублер Е.В. Вычислительные методы анализа и распознавания патологических процессов. – Л.: Медицина, 1978. – 296 с.

Гублер Е.В., Генкин А.А. Применение непараметрических критериев статистики в медико-биологических исследованиях. – Л.: Медицина, 1973. – 142 с.

Гусев А.Г. Охрана рыбохозяйственных водоемов от загрязнения. – М.: Пищ. пром-ть, 1975. – 179 с.

Гусева Т.В., Молчанова Я.П., Заика Е.А. и др. Гидрохимические показатели состояния окружающей среды: справочные материалы. – М.: СоЭС, Метод. центр «Эколайн», 2000. Адрес в Интернет www.cci.glasnet.ru\mc\refbooks\hydrochem

Гутельмахер БЛ. Метаболизм планктона как единого целого. – Л.: Наука, 1986. – 155 с.

Даддингтон К. Эволюционная ботаника. – М.: Мир, 1972. – 308 с.

Дедю И.И. Экологический энциклопедический словарь. – Кишинев: Гл. ред. Молдав. Сов. Энциклопедии, 1990. – 408 с.

Дейт К. Введение в системы баз данных. – Киев: Диалектика, 1998. – 784 с.

Деревья классификации / Электронный учебник по статистике. Адрес в Интернет http://www.statsoft.ru/home/

Джефферс Дж. Введение в системный анализ: применение в экологии. - М.: Мир, 1981. - 252 с.

Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. – М.: Мир. Т. 1. 1980. – 610 с.; Т. 2. 1981. – 520 с.

Джурс П., Айзенауэр Т. Распознавание образов в химии. – М.: Мир, 1977. – 230 с.

Дзюбан Н.А., Кузнецова С.П. О гидробиологическом контроле качества вод по зоопланктону // Научные основы контроля качества вод по гидробиологическим показателям: Тр. Всес. конф. – Л.: Наука, 1981. С. 117-136.

Дидэ Э. и др. Методы анализа данных: Подход, основанный на методе динамических сгущений. – М.: Финансы и статистика, 1985. – 357с.

Динамическая теория биологических популяций / Под ред. Р.А. Полуэктова. – М.: Наука, 1974. – 456 с.

Дмитриев В.В. Экологическое нормирование состояния и антропогенных воздействий на природные экосистемы // Вестн. СПб. ун-та. 1994. Сер. 7. Вып. 2(14). С. 60-70.

Дмитриев М.Т., Казнина Н.И., Пинигина И.А. Санитарно-химический анализ загрязняющих веществ в окружающей среде. – М.: Химия, 1991. – 544 с.

Долгов Г.И., Никитинский Я.Я. Гидробиологические методы // Стандартные методы исследования питьевых и сточных вод. – М.: Мосполиграф, 1927. С.142-217.

Домбровский Ю.А. Модель биотического круговорота Таганрогского залива // Изв. СКНЦ ВШ. Естеств. Науки. 1977. № 2. С. 94-102.

Дорофеюк А.А. Алгоритмы автоматической классификации // Проблемы расширения возможностей автоматов (Труды Ин-та пробл. управ. АН СССР). Вып 1. – М.: ИПУ АН СССР, 1971. С. 5-41.

Драбкова В.Г., Сорокин И.Н. Озеро и его водосбор – единая природная система. – Л.: Наука, 1979. – 196 с.

Драчев С.М. Борьба с загрязнением рек, озер и водохранилищ промышленными и бытовыми стоками. – М.-Л.: АН СССР, 1964. – 274 с.

Дрейпер Н., Смит Г. Прикладной регрессионный анализ: В 2-х кн.– М.: Финансы и статистика. Кн. 1. 1986. – 366 с.; Кн. 2. 1987. – 351 с.

Дружинин В.В., Конторов Д.С. Проблемы системологии (проблемы теории сложных систем). – М.: Сов. радио, 1976. – 296 с.

Дубницкий В.Ю., Цейтлин Н.А. Оценка качества материала по содержанию вредных примесей с учетом интересов производителя и потребителя // Модели и системы. Труды. Вып. 1. – Харьков: Харьк. военный ун-т, 1999. С. 30-32.

Дубров А.М. Обработка статистических данных методом главных компонент. – М.: Финансы и статистика, 1978. – 135 с.

Дуда Р., Харт П. Распознавание образов и анализ сцен. – М.: Мир, 1978. – 510 с.

Дуплаков С.М. Материалы к изучению перифитона // Труды Лимнологической станции в Косине. 1933. Т. 16. С. 9-136.

Дьяконов В., Круглов В. Математические пакеты расширения МАТЛАБ. Специальный справочник. – СПб.: Питер, 2001. – 480 с.

Дэйвисон М. Многомерное шкалирование. Методы наглядного представления данных. – М.: Финансы и статистика, 1988. – 348 с.

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

Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128 с.

Елисеева И.И., Юзбашев М.М. Общая теория статистики. – М.: Финансы и статистика, 1995. – 368 с.

Елисеева Л.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. – М.: Статистика, 1977. – 144 с.

Емельянов В.В., Ясиновский С.И. Введение в интеллектуальное имитационное моделирование сложных дискретных систем и процессов. Язык РДО. – М.: Изд-во АНВИК, 1998. – 427 с.

Емельянова В.П., Данилова Г.Н., Зенин А.А. Метод комплексной оценки загрязнения воды // Оценка и классификация качества поверхностных вод для водопользования. – Харьков, 1979. С. 126-128.

Емельянова В.П., Данилова Г.Н., Родзиллер И.Д. Способ обобщения показателей для оценки качества поверхностных вод // Гидрохим. материалы. 1980. Т. 77. С. 88-96.

Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. – М.: Финансы и статистика, 1986.

Жадин В. И. Донные биоценозы реки Оки и их изменения за 35 лет. // Загрязнение и самоочищение р. Оки. – М.; Л.: Наука, 1964. С. 226-287.

Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 342 с.

Жолдакова З.И., Синицына О.О., Харчевникова Н.В., Зайцев Н.А. Проблема единого эколого-гигиени-ческого нормирования химических веществ в окружающей среде // Гигиена и санит. 1998. № 4. С. 57-60.

Жукинский В.Н., Оксиюк О.П., Олейник Г.Н, Кошелева С.И. Проект системы комплексной оценки качества поверхностных пресных вод // Водн. ресурсы. 1978. № 3. С. 83-93.

Жукинский В.Н., Оксиюк О.П., Олейник Г.Н., Кошелева С.И. Принципы и опыт построения экологической классификации качества поверхностных вод суши // Гидробиол. журн. 1981. Т. 17. № 2. С. 38-50.

Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Вып. 33. – М.: Наука, 1978. С. 5-68.

Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок. // Кибернетика. 1971. № 3. С. 1-11.

Загоруйко Н.Г. Методы распознавания и их применение. – М.: Сов. радио, 1972. – 308 с.

Заде Л. Основы нового подхода к анализу сложных систем и процессов принятия решений // Математика сегодня. – М.: Знание, 1974. С. 5-49.

Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 165 с.

Заика В.Е. Сравнительная продуктивность гидробионтов. – Киев: Наук. думка, 1983. – 206 с.

Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. - М.: Наука, 1984. - 424 с.

Закс Л. Статистическое оценивание. – М.: Статистика, 1976. – 598 с.

Замолодчиков Д.Г. Оценка экологически допустимых уровней антропогенного воздействия на пресноводные экосистемы // Проблемы экологического мониторинга и моделирования экосистем. – СПб: Гидрометеоиздат, 1993. Т. 15. С. 214-233.

Замолодчиков Д.Г., Булгаков Н.Г., Гурский А.Г. и др. К методике применения детерминационного анализа для обработки экологических данных // Науч. докл. высш. шк. Биол. науки. 1992. № 7. С. 116-133.

Зенкевич Л.А. Продуктивность морских водоемов СССР // Тр. фаунист. конф. Зоологического инта АН СССР. – Л.: Секция гидробиол., 1934. С. 11-18.

Зенкевич Л.А. Фауна и биологическая продуктивность моря. Т. 1. – М.: Сов. наука, 1951. – 505 с.

Зернов С.А. Общая гидробиология. Изд. 2. – М.; Л.: Изд-во АН СССР, 1949. – 587 с.

Зинченко Т.Д. Хирономиды поверхностных вод бассейна Средней и Нижней Волги (Самарская область). Эколого-фаунистический обзор. – Самара: ИЭВБ РАН, 2002. – 174 с.

Зинченко Т.Д., Выхристюк Л.А, Шитиков В.К. Методологический подход к оценке экологического состояния речных систем по гидрохимическим и гидробиологическим показателям // Изв. СамНЦ РАН. 2000. Т. 2. № 2. С. 233-243.

Зинченко Т.Д., Молодых Н.В. Закономерности многолетних изменений хирономид в бентосе Куйбышевского водохранилища // Экологические проблемы бассейнов крупных рек. – Тольятти: ИЭВБ РАН, 1993. С. 78-79.

Зинченко Т.Д., Шитиков В.К. Гидробиологический мониторинг как основа типологии малых рек Самарской области // Изв. СамНЦ РАН. 1999. Т. 1. № 1. С. 118-127.

Зоммер Е.А. Теоретические и экспериментальные предпосылки к разработке экологически обоснованных региональных ПДК // Вторая Всесоюзная конференция по рыбохозяйственной токсикологии: Тез. докл. – СПб., 1991. Т. 1. С. 224-226.

Зорин Н.А О неправильном употреблении термина "достоверность" в российских научных психиатрических и общемедицинских статьях // Адрес в Интернет http://www.biometrica.tomsk.ru:8101/lib/naukoved/let1.htm

Ивахненко А.Г. Самообучающиеся системы распознавания и автоматического управления. – Киев: Техника, 1969. – 392 с.

Ивахненко А.Г. Долгосрочное прогнозирование и управление сложными системами. – Киев: Техника, 1975. – 311 с.

Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наук. думка, 1982. – 296 с.

Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. Принятие решений на основе самоорганизации. – М.: Сов. радио, 1976. – 280 с.

Ивахненко А.Г., Кротов Г.И., Чеберкус В.И. Многорядный алгоритм самоорганизации долгосрочных прогнозов (на примере экологической системы оз. Байкал) // Автоматика. 1980. № 4. С. 28-47.

Ивахненко А.Г., Лапа В.Г. Предсказание случайных процессов. – Киев: Наук. думка, 1971. – 416 с.

Ивахненко А.Г., Юрачковский Ю.П. Моделирование сложных систем по экспериментальным данным. – М.: Радио и связь, 1987. – 118 с.

Ивашко В.Г., Кузнецов С.О. Оценки правдоподобия в продукционных экспертных системах / Экспертные системы: состояние и перспективы. – М.: Наука, 1989. – 152 с.

Ивлев В.С. О превращении энергии при росте беспозвоночных // Бюлл. МОИП. Отд. биол. 1938. Т. 47, № 4. С. 267-277.

Ивлев В.С. О структурных особенностях биоценозов // Изв. АН Латв. ССР. 1954. Т. 10 (87). С. 53-68.

Ивлев В.С. Экспериментальная экология питания рыб. – М.: Пищепромиздат, 1955. – 252 с.

Израэль Ю.А. Глобальная система наблюдений. Прогноз и оценка изменений состояния окружающей природной среды. Основы мониторинга // Метеорология и гидрология. 1974. № 7. С. 3-8.

Израэль Ю.А. Концепция мониторинга состояния биосферы // Мониторинг состояния окружающей природной среды. Тр. 1 советско-английского симпозиума. – Л.: Гидрометеоиздат, 1977. С. 10-25.

Израэль Ю.А. Экология и контроль состояния природной среды. – М.: Гидрометеоиздат, 1984. – 560 с.

Израэль Ю.А., Абакумов В.А. Об экологическом состоянии поверхностных вод СССР и критериях экологического нормирования // Экологические модификации и критерии экологического нормирования. – Л.: Гидрометеоиздат, 1991. С. 7-18.

Исаков Ю.А., Казанская Н.С., Тишков А.А. Зональные закономерности динамики экосистем. – М.: Наука, 1986. – 150 с.

Исаченко А.Г. Ландшафтоведение и физико-географическое районирование. - М.: Высш. шк., 1991. – 336 с.

Каган Ю.С. О вероятностном подходе к установлению пороговых и неэффективных доз при действии химических веществ // Гигиена и санит. 1978. № 8. С. 278.

Каган Ю.С., Станкевич В.В. Коэффициент кумуляции как количественный критерий // Актуальные вопросы гигиены труда, промышленной токсикологии и профпатологии в нефтяной и нефтехимической промышленности. – Уфа, 1964. С. 48-49.

Казначеев В.П., Спирин Е.А. Космопланетарный феномен человека: проблемы комплексного изучения. – Новосибирск: Наука, 1991. – 304 с.

Калинина В.Н., Панкин В.Ф. Математическая статистика. – М.: Высш. шк., 2001. – 336 с.

Калужский государственный педагогический университет им. К.Э. Циолковского. Лаборатория биоиндикации. Адрес в Интернет http://kspu.kaluga.ru/biomon/index.htm

Каменев А.Г. Продукция сообществ макрозообентоса водотоков // Биопродуктивность и биоиндикация водотоков правобережного Средневолжья: Макрозообентос. - Саранск: 1993. С. 136-182.

Каминский В.С. Состав и качество поверхностных вод: Понятие «качество» воды // Основы прогнозирования качества поверхностных вод. – М.: Наука, 1982. С. 6-22.

Камлюк Л.В. Энергетический обмен у свободноживущих плоских и кольчатых червей и факторы, его определяющие // Журн. общ. биол. 1974. Т. 35. Вып. 6. С. 874-885.

Каплин В.М. Превращение органических веществ в природных водах. Автореф. докт. дисс. – Иркутск: Иркутск. гос. ун-т, 1973. – 36 с.

Качанова Т.Л., Фомин Б.Ф. Реконструктивный анализ поведения сложных систем по эмпирическим данным. Препринт № 1. – СПб: Изд-во СПбГЭТУ, 1997.

Кашкаров Д.Н. Среда и сообщество (Основы синэкологии). – М.: Медгиз, 1933. – 383 с.

Кендалл (Кендэлл) М. Ранговые корреляции. – М.: Статистика, 1975. – 212 с.

Кендалл М., Стьюарт А. Теория распределений. – М.: Наука, 1966. – 566 с.

Кендалл М., Стьюарт А. Статистические выводы и связи. – М .: Наука, 1973. – 899 с.

Ким Дж.О., Мьюллер Ч.У, Клекка У.Р. и др. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.

Кимстач В.А. Классификация качества поверхностных вод в странах Европейского экономического сообщества. – СПб.: Гидрометеоиздат, 1993. – 48 с.

Китаев С.П. Термические классификации озер мира // Водн. ресурсы. 1978. № 1. С. 87-103.

Китаев С.П. Экологические основы биопродуктивности озер разных природных зон. – М.: Наука, 1984. – 207 с.

Классификация и кластер / Под ред. Дж. Вэн-Райзина. – М.: Мир, 1980. – 390 с.

Кожова О.М. Применение методов экосистемного анализа к оценке качества вод (на примере Ангары и Байкала). // Научные основы контроля качества поверхностных вод по гидробиологическим показателям. Тр. Советско-английского семинара. – Л.: Гидрометеоиздат, 1977. С. 16-29.

Кожова О.М. Прогноз состояния водных экосистем и приемы экологической оценки действия антропогенных факторов // Прогнозирование экологических процессов. – Новосибирск: Наука, 1986. С. 27-34.

Кожова О.М., Ащепкова Л.Я., Загоренко Г.Ф. Исследование некоторых методов биологического контроля рек // Гидробиологические исследования в Восточной Сибири. Чтения памяти проф. М.М. Кожова. - Иркутск: Изд. ИГУ. 1979. Вып. 3. С. 67-82.

Кожова О.М., Павлов Б.К. Организация биомониторинга Байкала // Приемы прогнозирования экологических систем. – Новосибирск: Наука, 1985. С. 4-8.

Кожова О.М., Павлов Б.К. Популяционные аспекты исследования зоопланктона оз. Байкал // Прогнозирование экологических процессов. – Новосибирск: Наука, 1986. С. 132-138.

Колмогоров А.Н. Избранные труды: Математика и механика. – М.: Наука, 1985. С. 136-138.

Колодяжный С.Ф. Проблемы изучения взаимосвязей в биометрических исследованиях // Исследование биологических систем математическими методами. – Л.: 1985. С. 18-34.

Колодяжный С.Ф., Пааль Я.Д. О некоторых проблемах теории и практики количественной классификации растительности // Исследования биологических систем математическими методами. – Л.: 1985. С. 126-138.

Коломыц Э.Г. Модели цепных реакций в лесных экосистемах как основа локального мониторинга глобальных изменений // Теоретические проблемы экологии и эволюции (Третьи Любищевские чтения). – Тольятти: ИЭВБ РАН, 2000. С. 73-76.

Кольцов П.П. Математические модели теории распознавания образов // Компьютер и задачи выбора. – М.: Наука, 1989. С. 89-119.

Комплексные оценки качества поверхностных вод / Под ред. А.М. Никанорова. – Л.: Гидрометеоиздат, 1984. – 139 с.

Константинов А.С. Основные принципы оценки мощности системы биологического самоочищения водоемов, некоторые показатели ее работы в Волгоградском водохранилище и пути установления ПДК в естественных условиях // Водн. ресурсы. 1973. № 4. С. 149-153.

Константинов А.С. Общая гидробиология. – М.: Высш. шк., 1979. – 480 с.

Константинов А.С. Оценка и индикация состояния экосистем в условиях антропогенного воздействия // Научные основы контроля качества поверхностных вод по гидробиологическим показателям. – Л.: Гидрометеоиздат, 1981. С. 75-89.

Коршунов Ю.М. Математические основы кибернетики. – М.: УРСС, 1995. – 325 с.

Котов В.Н. Применение теории измерений в биологических исследованиях. – Киев: Наук. думка, 1985. – 98 с.

Котов В.Н., Терентьева Н.Г. Разграничение двух совокупностей биологических объектов на примере планктонных видов рода *Anabaena Bory*. Препринт. – Киев: Институт ботан. им. Н.Г. Холодного АН УкрССР, 1989. – 52 с.

Кохонен Т. Ассоциативные запоминающие устройства. – М.: Мир, 1982. – 383 с.

Кравцов Б.А., Милютин Л.И. Возможности применения многомерной классификации при изучении популяций древесных растений // Пространственно-временная структура лесных биогеоценозов. - Новосибирск: Наука, 1981. С. 47-65.

Крамер Г. Математические методы статистики. – М.: Мир, 1975. – 648 с.

Крапивин В.Ф. О теории живучести сложных систем. – М: Наука, 1978. – 248 с.

Крапивин В.Ф., Свирежев Ю.М., Тарко А.М. Математическое моделирование глобальных биосферных процессов. – М.: Наука, 1982. – 270 с.

Краскел Дж.Б. Многомерное шкалирование и другие методы поиска структуры // Статистические методы для ЭВМ. – М.: Наука, 1986. С. 301-347.

Краснощеков Г.П., Розенберг Г.С. Естественно-исторические аспекты формирования территории Волжского бассейна // Изв. СамНЦ РАН. 1999. Т. 1. № 1. С. 108-117.

Красовский Г.Н., ***Воробьева Л.В.*** Эколого-гигиеническая оценка водной среды на основе факторного анализа // Гигиена и санит. 1998. № 4. С. 19-23.

Красовский Г.Н., ***Егорова Н.А.*** Гигиенические и экологические критерии вредности в области охраны водных объектов // Гигиена и санит. 2000. № 6. С. 14-16.

Красовский Г.Я. Аэрокосмический мониторинг поверхностных вод (практические аспекты). – М.: Науч. совет по космич. исслед. для народн. хозва МКС АН СССР. – 231 с.

Крестин С.В., Розенберг Г.С. Об одном механизме "цветения воды" в водохранилище равнинного типа // Биофизика. 1996. Т. 41. Вып. 3. С. 650-654.

Крестин С.В., Розенберг Г.С. Двухмерная модель "цветения воды" в водохранилище равнинного типа // Изв. СамНЦ РАН. 2002. Т. 4. № 2. С. 276-279.

Криволуцкий Д.А., Степанов А.М., Тихомиров Ф.А., Федоров Е.А. Экологическое нормирование на примере радиоактивного и химического загрязнения экосистем // Методы биоиндикации окружающей среды в районах АЭС. – М.: Наука, 1988. С. 4-16.

Критерии оценки экологической обстановки территорий для выявления зон чрезвычайной экологической ситуации и зон экологического бедствия. Утверждено Приказом Минприроды РФ от 30 ноября 1992 г. – 51 с. (Опубликовано в газете "Зеленый мир". 1994. № 11).

Крогиус Ф.В., Крохин Е.М., Куренков И.И., Меншуткин В.В. Модель экологической системы озера Дальнего // Гидробиол. журн. 1969. Т.5. № 5. С. 14-22.

Кузин Л.Т. Основы кибернетики. (Основы кибернетических моделей). Т. 2. – М.: Энергия, 1979. – 584 с.

Кулагин Ю.З. К теории экологического прогнозирования // Экология. 1980. № 5. С. 36-41.

Кульбак С. Теория информации и статистика. – М.: Наука, 1967. – 408 с.

Кун Т. Структура научных революций. – М.: Прогресс, 1977. – 300 с.

Куприянова Т.П. Принципы и методы физико-географического районирования с применением ЭВМ. – М.: Наука, 1977. – 124 с.

Куркин К.А. Системный подход в экологических исследованиях // Системные исследования. – М.: Наука, 1977. С. 195-211.

Курляндский Б.А., Шитиков В.К., Тихонов В.Н., Ковалев А.Ф. Прогнозирование гигиенических нормативов методом построения "обобщенного портрета" // Гигиена труда и профессиональные заболевания. 1988. № 6. С. 33-37.

Кустов В.В., Тиунов Л.А., Васильев Г.А. Комбинированное действие промышленных ядов. – М.: Медицина, 1975. – 184 с.

Кутикова Л.А. Коловратки речного планктона как показатели качества вод // Методы биологического анализа пресных вод. – Л.: ЗИН АН СССР, 1976. С. 80-90.

Лазарев Н.В. Основы промышленной токсикологии. – Л.: Медгиз, 1938. – 388 с.

Лакин Г.Ф. Биометрия. - М.: Высш. шк., 1990. - 352 с.

Лапач С.Н., Чуйсенко А.В., Бабич П.И. Статистические методы в медико-биологических исследованиях с использованием Excel. – М.: Изд-во "Морион Лтд", 2000. – 320 с.

Лапко А.В., Цугленок Н.В., Цугленок Г.И. Имитационные модели пространственно распределенных экологических систем. – Новосибирск: Наука, 1999. – 190 с.

Ларичев О.И., Мечитов А.И., Мошкович Е.М., Фуремс Е.М. Выявление экспертных знаний (процедуры и реализации). - М.: Наука, 1989. - 128 с.

Лбов Г.С. Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука, 1981. – 157 с.

Левич А.П. Структура экологических сообществ. – М.: МГУ, 1980. – 181 с.

Левич А.П. Биотическая концепция контроля природной среды // Докл. Академии наук. 1994. Т. 337 № 2. С. 257-259.

Левич А.П., Терехин А.Т. Метод расчета экологически допустимых уровней воздействия на пресноводные экосистемы (метод ЭДУ) // Водн. ресурсы. 1997. Т. 24. № 3. С. 328-335.

Левушкин С.И. Понятие вида и экология // Материалы к II Всесоюзному совещанию "Вид и его продуктивность в ареале". – Вильнюс: Лит. АН, 1976. С. 83-85.

Лем С. Сумма технологии. – М.: Мир, 1968. – 608 с. Адрес в Интернет http://lib.ru//LEM/ summa.txt

Леман Э. Проверка статистических гипотез. – М.: Наука, 1964. – 498 с.

Леме Ж. Основы биогеографии. – М.: Прогресс, 1976. – 309 с.

Леонов А.В. Математическая модель совместной трансформации соединений азота, фосфора и кислорода в водной среде: ее применение для анализа динамики компонентов в евтрофном озере // Водн. ресурсы. 1989. № 2. С. 105-123.

Леонов А.В., Осташенко М.М., Лаптева Е.Н. Математическое моделирование процессов трансформации органического вещества и соединений биогенных элементов в водной среде: предварительный анализ условий функционирования экосистемы Ладожского озера // Водн. ресурсы. 1991. № 1. С. 51-72.

Леонов А.В., Цхай А.А. Прогноз качества воды проектируемого водохранилища на основе модели трансформации соединений азота и фосфора // Водн. ресурсы. 1995. № 3. С. 261-272.

Леонов В.П. Три "Почему ..." и пять принципов описания статистики в биомедицинских публикациях // "Биометрика". а. Адрес в Интернет http://www.biometrica.tomsk.ru

Леонов В.П. Долгое прощание с лысенковщиной // "Биометрика". б. Адрес в Интернет http://www.biometrica.tomsk.ru

Леонов В.П. Искушение знанием или «Сколько стоит биометрика?» // "Биометрика". в. Адрес в Интернет http://www.biometrica.tomsk.ru

Лефевр В.А. Конфликтующие структуры. – М.: Сов. радио, 1973. – 158 с.

Ли Н. Экологическая экспертиза. Учебное руководство. – М.: Экопрос, 1995. – 187 с.

Ликеш И., Ляга И. Основные таблицы математической статистики. – М.: Финансы и статистика, 1985. – 356 с.

Лисенков А.Н. Математические методы планирования многофакторных медико-биологических экспериментов. – М.: Медицина, 1979. – 344 с.

Литвак Б.Г. Экспертная информация. Методы получения и анализа. – М.: Радио и связь, 1982. – 184 с.

Логвиненко Н.В. Петрография осадочных пород. – М.: Высш. шк., 1974. – 399 с.

Логофет Д.О., Свирежев Ю.М. Концепция устойчивости биологических систем // Проблемы экологического мониторинга и моделирования экосистем. – Л.: Гидрометеоиздат, 1983, Т. 6. С. 159-171.

Лорьер Ж.-Л. Системы искусственного интеллекта. – М.: Мир, 1991. – 342 с.

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

Лукьяненко В.И. Общая ихтиотоксикология. – М.: Легкая и пищ. пром-ть, 1983. – 320 с.

Лукьяненко В.И. Экология водоемов. Охрана и рациональное использование рыбных запасов бассейна Волги. Концепции, цели, задачи. – Н.Новгород: ННГУ, 1992. – 32 с.

Лукьяненко В.И. Экологические ПДК и комплексный экологический мониторинг качества вод // Розенберг Г.С., Краснощеков Г.П. Волжский бассейн: экологическая ситуация и пути рационального природопользования - Тольятти: ИЭВБ РАН, 1996. С. 218-219.

Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (На примере универсальной автоматизированной системы распознавания образов "ЭЙДОС-5.1"). – Краснодар: КЮИ МВД РФ, 1996. – 278 с. Адрес в Интернет http://lc.narod.ru/aidos/ aidos96/aidos96.htm

Луценко Е.В. Автоматизированный системно-когнитивный анализ в управлении активными объектами (системная теория информации и ее применение в исследовании экономических, социально-психологических, технологических и организационно-технических систем). – Краснодар: КубГАУ. 2002. – 605 с. Адрес в Интернет http://lc.narod.ru/aidos/ aidos02/index.htm

Любищев А.А. Об ошибках в применении математики в биологии. I. Ошибки от недостатка осведомленности // Журн. общ. биол. 1969а. Т. 3. № 5. С. 572-584.

Любищев Л.Л. Об ошибках в применении математики в биологии. II. Ошибки, связанные с избытком энтузиазма // Журн. общ. биол. 1969б. Т. 3. № 6. С. 715-723.

Любищев А.А. Дисперсионный анализ в биологии. – М.: МГУ, 1986. – 200с.

Ляпунов А.А. Проблемы теоретической и прикладной кибернетики. – М.: Наука, 1980. – 335 с.

Мазуров В.Д. Плохо формализуемые задачи планирования технико-экономических систем. – Свердловск: Средне-Урал. кн. изд-во, 1982. – 64 с.

Майстренко В.Н., Хамитов Р.З., Будников Г.К. Эколого-аналитический мониторинг суперэкотоксикантов. – М.: Химия, 1996. – 319 с.

Макарова Н.П. Балансовый анализ уравнений биоэнергетики // Биология моря (Киев). 1975. Вып. 33. С. 49-55.

Макрушин А.В. Биологический анализ качества вод. – Л.: ЗИН АН СССР, 1974а. – 60 с.

Макрушин А.В. Библиографический указатель по теме "Биологический анализ качества вод" с приложением списка организмов-индикаторов загрязнения. – Л.: ЗИН АН СССР, 1974б. – 53 с.

Максимов В.Н. Об одном способе оценки качества природных вод // Самоочищение и биоиндикация природных вод. - М.: Наука, 1980. С. 212-219.

Максимов В.Н., Булгаков Н.Г., Левич А.П., Терехин А.Т. Методика применения детерминационного анализа данных мониторинга для целей экологического контроля природной среды // Успехи соврем. биол. 2001. Т. 121. № 2. С. 131-143.

Максимов В.Н., Булгаков Н.Г., Милованова Г.Ф. Детерминационный анализ связей между различными компонентами экосистем. Сравнение с методами традиционной статистики // Изв. РАН. Сер. биол. 1999. № 4. С. 469-477.

Максимов В.Н., Булгаков Н.Г., Милованова Г.Ф., Левич А.П. Детерминационный анализ в экосистемах: сопряженности для биотических и абиотических компонент // Изв. РАН. Сер. биол. 2000а. № 24. С. 482-491.

Максимов В.Н., Милованова Н.Г., Булгаков Н.Г., Левич А.П. Индикация состояния экосистем методами детерминационного анализа // Теоретические проблемы экологии и эволюции. – Тольятти: ИЭВБ РАН, 2000б. С. 113-120.

Малиновский А.А. Общие вопросы строения систем и их значение для биологии // Вопросы методологии системного исследования. – М.: Наука, 1970. С. 146-183.

Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.

Маргалеф Р. Облик биосферы. – М.: Наука, 1992. – 214 с.

Масалович А.И. Этот нечеткий, нечеткий, нечеткий мир // PC Week/RE. 1995. № 16. С. 15-18.

Масалович А.И. Нечеткая логика: на гребне "Третьей Волны" // Адрес в Интернет http://www.tora-centre.ru/library/fuzzy/ct_fuz. htm

Мейен С.В. Нетривиальная биология (Заметки о…) // Журн. общ. биол. 1990. Т. 51. № 1. С. 4 -14.

Мейер М. Теория реляционных баз данных. – М.: Мир, 1987. – 608 с.

Мейндональд Дж. Вычислительные алгоритмы в прикладной статистике. – М.: Финансы и статистика, 1988. – 350 с.

Меншуткин В.В. Математическое моделирование популяций и сообществ водных животных. – Л.: Наука, 1971. – 196 с.

Меншуткин В.В., Кожова О.М., Ащепкова Л.Я. Модель сезонной динамики экосистемы озера Байкал // Модели природных систем. – Новосибирск, Наука, 1978. С. 57-64.

Меншуткин В.В., Кожова О.М., Ащепкова Л.Я., Кротова В.А. Камерная модель динамики экосистемы озера Байкал с учетом трехмерной циркуляции вод // Математическое моделирование водных экосистем. – Л.: Гидрометеоиздат, 1981. С. 288-298.

Меншуткин В.В., Умнов А.А. Математическая модель простейшей водной экологической системы // Гидробиол. журн. 1970. Т. 6. Вып. 2. С. 28-35.

Меркурьев В.В., Молдавский М.А. Семейство сверток векторного критерия для нахождения точек множества Парето // Автоматика и телемеханика. 1979. № 1. С. 110-121.

Методика изучения биогеоценозов внутренних водоемов. – М.: Наука, 1975. – 240 с.

Методика расчета предельно-допустимых сбросов (ПДС) веществ в водные объекты со сточными водами – Харьков: ВНИИВО, 1990. – 115 с.

Методические рекомендации по применению биотестирования для оценки качества воды в системах хозяйственно-питьевого водоснабжения. МР № ЦОС ПВ Р 005-95. – М.: 1995. – 50 с.

Методические рекомендации по установлению ПДК загрязняющих веществ для воды рыбохозяйственных водоемов. – М.: 1986. – 36 с.

Методические указания к экспериментальному изучению процессов трансформации химических веществ при их гигиеническом регламентировании в воде. № 2966-84. – М.: Минздрав СССР, 1984. – 24 с.

Методические указания по изучению мутагенной активности химических веществ при обосновании их ПДК в воде. № 4110-86. – М.: Минздрав СССР, 1986. – 23 с.

Методические указания по обоснованию гигиенических нормативов химических веществ в воде водных объектов хозяйственно-питьевого и культурно-бытового водопользования. МУ 2.1.5.720-98. – М.: 1998. – 44 с.

Методические указания по определению концентраций химических веществ в воде централизованного хозяйственно-питьевого водоснабжения: Сборник методических указаний. – М.: 1997. – 112 с.

Методические указания по разработке и научному обоснованию предельно допустимых концентраций вредных веществ в воде водоемов. № 1296-75. – М.: Минздрав СССР, 1976. – 78 с.

Методология оценки состояния экосистем: учебное пособие / Под ред. О.М. Кожовой и В.В. Воробьева. – Ростов-на-Дону: Изд-во ООО «ЦВВР», 2000. – 128 с

Методы определения токсичности и опасности химических веществ (токсикометрия) / Под ред. И.В. Саноцкого. – М.: Медицина, 1970. – 343 с.

Мизандронцев И.Б. Химические процессы в донных отложениях водоемов. – Новосибирск: Наука, 1991. – 176 с.

Минский М., Пейперт С. Персептроны. – М.: Мир, 1971. – 252 с.

Миркин Б.Г., Черный Л.Б. Об измерении близости между различными разбиениями конечного множества объектов // Автоматика и телемеханика. 1970. № 5. С. 6-18.

Миркин Б.М. О парадигме в фитоценологии // Журн. общ. биол. 1984. Т. 45. № 6. С. 749-788.

Миркин Б.М. Теоретические основы современной фитоценологии. – М.: Наука, 1985. – 136 с.

Миркин Б.М. Современное состояние и тенденции развития классификации растительности методом Браун-Бланке // Итоги науки и техники. Ботаника. Т. 9. – М.: ВИНИТИ, 1989. – 126 с.

Миркин Б.М., Наумова Л.Г. Градиентный анализ растительности // Успехи совр. биол. 1983. Т. 95. Вып. 2. С. 304-318.

Миркин Б.М., Наумова Л.Г. Наука о растительности (история и современное состояние основных концепций). – Уфа: Гилем, 1998. – 413 с.

Миркин Б.М., Наумова Л.Г., Соломещ А.И. Современная наука о растительности. – Уфа: Логос, 2000. – 264 с.

Миркин Б.М., Розенберг Г.С. Анализ мозаичности травянистых растительных сообществ. 2. Ценотический уровень // Биол. науки. 1977а. № 2. С. 121-126.

Миркин Б.М., Розенберг Г.С. Факторный анализ в фитоценологии. 1. Общая характеристика модели // Биол. науки. 1977б. № 12. С. 121-126.

Миркин Б.М., Розенберг Г.С. Фитоценология. Принципы и методы. – М.: Наука, 1978. – 212 с.

Миркин Б.М., Розенберг Г.С. Количественные методы классификации, ординации и геоботанической индикации // Итоги науки и техники. Ботаника. Т. 3. – М.: ВИНИТИ, 1979. С. 71-137.

Миркин Б.М., Розенберг Г.С., Бурцева Е.А. Факторный анализ в фитоценологии. 2. Сравнение влияния ведущих факторов // Биол. науки. 1978а. № 4. С. 136-143.

Миркин Б.М., Розенберг Г.С., Гареева Л.М., Яценко О.В. Факторный анализ в фитоценологии. 3. Оценка экологичности классификаций // Биол. науки. 1978б. № 12. С. 130-135.

Миркин Б.М., Розенберг Г.С., Наумова Л.Г. Новый коэффициент межвидовой сопряженности, удобный для классификации пойменных лугов (трансформированный коэффициент Дайса, ТКД) // Растительность речных пойм, методы ее изучения и вопросы рационального использования: Тез. докл. Всесоюз. конф. – Уфа: БФАН СССР, 1972. С. 85-86.

Миркин Б.М., Розенберг Г.С., Наумова Л.Г. Словарь понятий и терминов современной фитоценологии. – М.: Наука, 1989. – 223 с.

Мозговой Д.П., Розенберг Г.С., Владимирова Э.Д. Информационные поля и поведение млекопитающих. – Самара: Изд-во "Сам. ун-т", 1998. – 92 с.

Моисеев Н.Н. Человек и ноосфера. – М.: Мол. гвардия, 1990. – 352 с.

Моисеенко Т.И. Методические подходы к нормированию антропогенных нагрузок на водоемы Субарктики (на примере Кольского севера) // Проблемы химического и биологического мониторинга экологического состояния водных объектов Кольского севера. – Аппатиты: Кольский науч. центр , 1995. С. 7-23.

Моисеенко Т.И. Экотоксикологический подход к нормированию антропогенных нагрузок на водоемы Севера // Экология. 1998. № 6. С. 452-461.

Моисеенкова Т.А., Шитиков В.К. Принципы организации регионального банка эколого-экономической информации // Моделирование процессов экологического развития (М., ВНИИСИ АН СССР). 1989. № 7. С. 110-117.

Мокеева Н.П., Межов Б.В. Численность зообентоса как показатель изменений в морских донных сообществах // Гидробиол. журн. 1986. Т. 22. № 4. С. 28-31.

Мокров И.В., Гелашвили Д.Б. Оценка качества городской среды по стабильности развития березы повислой (*Betula pendula* Roth) // Экологические и метеорологические проблемы больших городов и промышленных зон. Тез. докл. – СПб.: РГГМУ, 1999. С. 43-44.

Молчанов А.М. Математические модели в экологии: Роль критических режимов // Математическое моделирование в биологии. – М.: Наука, 1975. С. 133-141.

Мордухай-Болтовской Ф.Д. Особенности программы и методики биогеоценотических исследований внутренних водоемов // Программа и методика биогеоценотических исследований. – М.: Наука, 1974. – 76 с.

Морозов В.Г. Эволюционное моделирование рядов произвольной вариабильности: необходимость и методика прогнозирования // Изв. СамНЦ РАН. 2000. Т. 2. № 2. С. 206-215.

Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. - М.: Финансы и статистика, 1982. - 30 с.

Муравейский С.Д. Роль географических факторов в формировании географических комплексов // Вопр. географии. 1948. С. 9-36.

Мюллер П., Нейман П., Шторм Р. Таблицы по математической статистике: - М.: Финансы и статистика, 1982. – 272 с.

Нагель Э., Ньюмен Д. Теорема Гёделя. – М.: Знание, 1970. – 62 с.

Надарая Э.А. Об оценке плотности распределения случайных величин // Сообщ. АН ГССР. 1964. Т. 34. № 2. С. 277-280.

Налимов В.В. Применение математической статистики при анализе вещества. – М.: Физматгиз, 1960. – 430 с.

Налимов В.В. Теория эксперимента. – М.: Наука, 1971. – 207 с.

Налимов В.В. Теоретическая биология? Ее все еще нет... // Знание – сила. 1979. № 7. С. 9-11. Адрес в Интернет http://www.biometrica.tomsk.ru:8101/lib/naukoved/\nalimov1.htm

Налимов В.В. Вероятностная модель языка. – М.: Наука, 1979. – 176 с.

Налимов В.В. Спонтанность сознания. Вероятностная теория смыслов и смысловая архитектоника личности. – М.: Изд-во "Прометей" МГПИ им. Ленина, 1989. – 112 с.

Наумов Н.П. Экология животных. 2-е изд. – М.: Высш. шк., 1963. – 618 с.

Наумов Н.П. Биологические (сигнальные) поля и их значение в жизни млекопитающих // Успехи современной териологии. – М.: Наука, 1977. С. 93-108.

Наумова В.В. Некоторые приложения теории графов к классификации // Иерархические классификационные построения в географической экологии и систематике. – Владивосток: ДВНЦ АН СССР, 1979. С. 113-120.

Нейман Ю. Вводный курс теории вероятностей и математической статистики. – М.: Наука, 1968. – 448 с.

Нейронные сети. Statistica Neural Networks. – М.: Горячая линия – Телеком, 2001. – 182 с.

Нетрадиционные модели и системы с нечеткими знаниями /Под ред. А.Ф. Блишуна. – М.: Энергоатомиздат, 1991. – 144 с.

Никитин Я.Ю. Асимптотическая эффективность непараметрических критериев. – М.: Наука, 1995. – 240 с.

Николаевский В.С. Биомониторинг, его значение и роль в системе экологического мониторинга и охране окружающей среды // Методологические и философские проблемы биологии. – Новосибирск. Наука, 1981. С. 341-354.

Николис Дж. Динамика иерархических систем: Эволюционное представление. – М.: Мир, 1989. – 448 с.

Никулина В.Н. Опыт использования разных методов оценки степени загрязнения вод по альгофлоре // Методы биологического анализа пресных вод. – Л.: ЗИН АН СССР, 1976. С 38-58.

Нильсен Н. Д. Искусственный интеллект. Методы поиска решений.- М.: Мир, 1973. – 298 с.

Нинбург Е.А. Разграничение морских бентосных сообществ на основе данных факторного анализа (метод главных компонент) // Исследование биологических систем математическими методами. – Л.: 1985. С. 111-126.

Новиков С.М., Поройков В.В., Тертичников С.Н. и др. Анализ тенденций в развитии информационных технологий и обоснование концепции разработки банка токсикологических данных SARETbase // Гигиена и санит. 1995. № 1. С. 29-33.

Новиков С.М., Фурсова Т.Н. Метод количественной оценки кумулятивных свойств вредных веществ // Гигиена и санит. 1987. № 10. С. 52-55.

Новиков С. М., Шашина Т. А., Скворцова Н. С. Критерии оценки риска при кратковременных воздействиях химических веществ // Гигиена и санит. 2001. № 5. С 87-89.

Новиков Ю.В., Плитман С.И., Ласточкина К.С. и др. Оценка качества воды по комплексным показателям // Гигиена и санит. 1987. № 10. С. 7-11.

Новиков Ю.В., Ласточкина К.С., Болдина З.Н. Методы исследования качества воды водоемов. – М.: Медицина, 1990. – 376 с.

Ноулер Л., Хауэлл Дж., Голд Б. и др. Статистические методы контроля качества продукции. – М.: Изд-во стандартов, 1989. – 96с.

Одум Ю. Основы экологии. - М.: Мир, 1975. - 740 с.

Одум Ю. Экология: В 2-х т. – М.: Мир, 1986. Т. 1. – 328 с.; Т. 2. – 376 с.

Оксиюк О.П., Жукинский В.Н., Брагинский Л.П. и др. Комплексная экологическая классификация качества поверхностных вод суши // Гидробиол. журн. 1993. Т. 29. № 4. С. 62-77.

Оксиюк О.П., Олейник Г.Н., Юрченко В.В., Шевцова Л.В. Гидробиологический режим и процессы формирования качества воды в каналах // Проблемы гидробиологии каналов. – Киев: Наук. думка, 1978. С. 176-198.

Орлов А.И. Прикладная теория измерений // Прикладной многомерный статистический анализ. – М.: Наука, 1978. С. 68-138.

Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. – 296 с.

Орлов А.И. Задачи оптимизации и нечеткие переменные. – М.: Знание, 1980. – 53 с.

Орлов А.И. Репрезентативная теория измерений и ее применение // «Высокие статистические технологии» а. / Адрес в Интернет http://antorlov.chat.ru/reprteor.htm

Орлов А.И. Современная прикладная статистика // «Высокие статистические технологии» б. / Адрес в Интернет http://antorlov.chat.ru/statistc.htm

Орловский С.А. Проблемы принятия решений при расплывчатой информации. – М.: Наука, 1981. – 370 с.

Основы общей промышленной токсикологии / Под ред. Н.А. Толоконцева и В.А. Филова. – Л.: Медицина, 1976. – 304 с.

Оуэн Д.Б. Сборник статистических таблиц. Изд. 2-е, испр. – М.: ВЦ АН СССР, 1973. – 586 с.

Охапкин А.Г., Кузьмин Г.В. Фитопланктон как индикатор сапробности воды Саратовского водохранилища // Водн. ресурсы. 1978. № 3. С. 193-196.

Палий В.Ф. О количественных показателях при обработке фаунистических материалов // Зоол. журн. 1961. Т. 60. Вып. 1. С. 3-12.

Пастухова Е.В. Изменение фауны и флоры Москва-реки под влиянием антропогенных факторов // Растительное и животное население Москвы и Подмосковья. – М.: 1978. С. 112-114.

Патин С.А. Влияние загрязнения на биологические ресурсы и продуктивность мирового океана. – М.: Пищепромиздат, 1979. – 304 с.

Паутова В.Н., Номоконова В.И. Продуктивность фитопланктона Куйбышевского водохранилища. – Тольятти: ИЭВБ РАН, 1994. – 188 с.

Пашутин С.Б. Феномены биологической эволюции. Адрес в Интернет http://www.pereplet.ru/pops

Перечень рыбохозяйственных нормативов: предельно допустимых концентраций (ПДК) и ориентировочно безопасных уровней воздействия (ОБУВ) вредных веществ для воды водных объектов, имеющих рыбохозяйственное значение. – М: ВНИИРО. 1999. – 304 с.

Песенко Ю.А. Принципы и методы количественного анализа в фаунистических исследованиях. – М.: Наука, 1982. – 287 с.

Пикунов В.С., Цапук Д.А. Устойчивое развитие территорий: картографо-геоинформационное обеспечение. – М.; Смоленск: Изд-во Смоленского гос. ун-та, 1999. – 176 с.

Пинигин М.А. Гигиенические основы оценки суммарного загрязнения воздуха населенных мест // Гигиена и санит. 1985а. № 1. С. 66-69.

Пинигин М.А. Оценка комбинированного действия атмосферных загрязнений методом подбора его коэффициентов // Гигиена и санит. 1985б. № 9. С. 74-76.

Пинигин М.А. Теория и практика оценки комбинированного действия химического загрязнения атмосферного воздуха // Гигиена и санит. 2001. № 1. С. 9-12.

Пиотровски Е. Использование кинетики метаболизма и выведения токсических веществ в решении проблемы промышленной токсикологии. – М.: Мир, 1976. – 279 с.

Плохинский Н.А. Биометрия. – М.: МГУ, 1970. – 367 с.

Плохинский Н.А. Биометрический анализ в биологии. – М.: МГУ, 1982. – 157 с.

Плошко Б.Г., Елисеева И.И. История статистики: Учебное пособие. – М.: Финансы и статистика, 1990. – 295 с.

Подиновский В.В., Ногин В.Д. Парето-оптимальные решения многокритериальных задач. – М.: Наука, 1982. – 254 с.

Покровский А.А. Метаболитические аспекты фармакологии и токсикологии пищи. – М.: Медицина, 1979. – 180 с.

Попечителев Е.И., Романов С.В. Анализ числовых таблиц в биотехнических системах обработки экспериментальных данных. – Л.: Наука, 1985. – 148 с.

Попов И.И. Некоторые модели оценки и оптимизации информационных систем: математический аппарат моделирования // НТИ. 1981. Сер. 2. № 3. С. 10-16.

Попов И.И. Моделирование и оптимизация документальных информационных систем. / Адрес в Интернет http://www.rsuh.ru/e-library/ popov

Популярная медицинская энциклопедия / Под ред. Б.В. Петровского. – М.: Советская энциклопедия, 1988. – 513 с.

Попченко В.И. Закономерности изменений сообществ олигохет в условиях загрязнения водоемов // Водные малощетинковые черви: Матер. 6 Всесоюз. симп. – Рига: 1987. С. 117-122.

Попченко В.И. Экологические модификации сообществ зообентоса в условиях загрязнения водных экосистем // Экологические модификации и критерии экологического нормирования. – Л.: Гидрометеоиздат, 1991. С. 144-151.

Попченко В.И., Булгаков Г.П. Методы изучения зообентоса для оценки качества вод // Руководство по гидробиологическому мониторингу пресноводных экосистем. – СПб.: Гидрометеоиздат, 1992. С. 78-92.

Правила охраны поверхностных вод. – М.: Госкомитет СССР по охране природы, 1991. – 34 с.

Предельно допустимые концентрации (ПДК) химических веществ в воде водных объектов хозяйственно-питьевого и культурно-бытового водопользования: ГН 2.1.5.689-98. Ориентировочные допустимые уровни (ОДУ) химических веществ в воде водных объектов хозяйственно-питьевого и культурно-бытового водопользования: ГН 2.1.5.690-98. – М.: Миннздрав РФ, 1998.

Пригожин И., Стенгерс И. Порядок из хаоса: Новый диалог человека с природой. – М.: Прогресс, 1986. – 431 с.

Приемы прогнозирования экологических систем / Ред. О.М. Кожова и Л.Я. Ащепкова. – Новосибирск: Наука, 1985. – 126 с.

Прикладная статистика. Методы обработки данных. Основные требования и характеристики. – М.: ВНИИСтандартизации, 1987. – 64 с.

Проблема пороговости в токсикологии / Под ред. Г.Н. Красовского. – М.: Минздрав СССР, АМН СССР, 1979. – 103 с.

Прохоров С.А. Аппроксимативный анализ случайных процессов. 2-е изд. – Самара: СамНЦ РАН, 2001а. – 380 с.

Прохоров С.А. Прикладной анализ неэквидистантных временных рядов. – Уральск: СГАУ, 2001б. – 375 с.

Прохоров С.А. Моделирование и анализ случайных процессов: Лабораторный практикум. – Самара: СГАУ, 2002. – 191 с.

Прохоров С.А., Иващенко А.В., Графкин А.В. Автоматизированная система корреляционно-спектрального анализа случайных процессов. – Самара: СамНЦ РАН, 2003. – 286 с.

Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия. – М.: Большая Российская Энциклопедия, 2002. – 910 с.

Пузаченко Ю.Г., Скулкин В.С. Структура растительности лесной зоны СССР. Системный анализ. – М.: Наука, 1981. – 275 с.

Пушников А.Ю. Введение в системы управления базами данных. Часть 1. Реляционная модель данных: Часть 2. Нормальные формы отношений и транзакции. – Уфа: БГУ, 1999. Часть 1 – 108 с.; часть 2 – 138 с.

Пфанцагль И. Теория измерений. – М.: Мир, 1976. – 249 с.

Пых Ю.А., Малкина-Пых И.Г. Об оценке состояния окружающей среды. Подходы к проблеме // Экология. 1996. № 5. С. 323-329.

Райс Дж. (Rice J.R.) Будущее систем программного обеспечения для научных исследований // Computer-weekly (Computer-week Moskow). 1998. № 10. С. 25-26. (www.infoart.ru/it/press/cwm/10_98/sistem.htm)

Растригин Л.А., Марков В.А. Кибернетические модели познания: Вопросы методологии. – Рига: Зинатне, 1976. – 286 с.

Растригин Л.А., Эренштейн Р.Х. Метод коллективного распознавания. – М.: Энергоатомиздат, 1981. – 80 с.

Рашевски Н. Некоторые медицинские аспекты математической биологии. - М.: Мир, 1966. - 310 с.

Реймерс Н.Ф. Природопользование: Словарь-справочник. – М.: Мысль, 1990. – 637 с.

Рекомендации по прогнозированию качества поверхностных вод. – М.: ВНИИ траспорт. строит., 1984. – 44 с.

Розенберг Г.С. О сравнении различных методов автоматической классификации // Автоматика и телемеханика. 1975. № 9. С. 145-148.

Розенберг Г.С. Обзор методов статистической геоботаники. 3. Методы автоматической классификации. – М., 1977. 38 с. Деп. В ВИНИТИ 11.04.1977. № 1321-77.

Розенберг Г.С. Сравнение различных методов экологического прогнозирования. Прогноз динамики экосистем. // Экология. 1981. № 1. С. 12-18.

Розенберг Г.С. Модели в фитоценологии. – М.: Наука, 1984. – 256 с.

Розенберг Г.С. О системной экологии // Журн. общ. биологии. 1988. Т. 49. № 5. С. 580-591.

Розенберг Г.С. Адекватность математического моделирования экологических систем // Экология. 1989. № 6. С. 8-14.

Розенберг Г.С. К построению системы концепций современной экологии // Журн. общ. биологии. 1991. Т. 52. № 3. С. 422-440.

Розенберг Г.С., Беспалый В.Г., Голуб В.Б. и др. Экспертная система для оценки экологического состояния крупного региона (на примере Куйбышевской области) // Теоретические проблемы эволюции и экологии. – Тольятти.: ИЭВБ АН СССР, 1991, С. 183-192.

Розенберг Г.С., Долотовский И.М. Еще раз о показателях силы влияния // Биол. науки. 1988. № 9. С. 105-110.

Розенберг Г.С., Дунин Д.П., Костина Н.В. и др. Информационные технологии для оценки экологического состояния крупного региона (на примере Волжского бассейна и Самарской области) // Проблемы региональной экологии. Вып. 8. – Томск: СО РАН, 2000. С. 213-216.

Розенберг Г.С., Краснощеков Г.П. Волжский бассейн: экологическая ситуация и пути рационального природопользования - Тольятти: ИЭВБ РАН, 1996. – 249 с.

Розенберг Г.С., Краснощеков Г.П., Шитиков В.К. К созданию пространственно-распределенной базы эколого-экономических данных бассейна крупной реки (на примере Волжского бассейна) // Вопросы экологии и охраны природы в лесостепной и степной зонах: Межвед. сб. науч. тр. – Самара: Изд-во "Сам. ун-т", 1995. С. 8-15.

Розенберг Г.С., Мозговой Д.П., Гелашвили Д.Б. Экология. Элементы теоретических конструкций современной экологии. – Самара: СамНЦ РАН, 1999. – 396 с.

Розенберг Г.С., Смелянский И.Э. Экологический маятник (смена парадигм в современной экологии) // Журн. общ. биол. 1997. Т. 58. № 4. С. 5-19.

Розенберг Г.С., Феклистов П.А. Прогнозирование годичного прироста деревьев методами самоорганизации // Экология. 1982. № 4. С. 43-50.

Розенберг Г.С., Шитиков В.К., Брусиловский П.М. Экологическое прогнозирование (Функциональные предикторы временных рядов). – Тольятти: ИЭВБ РАН, 1994а. – 182 с.

Розенберг Г.С., Шитиков В.К., Мозговой Д.П. Экологическая информатика: Учебное пособие. – Самара: Изд-во Самар. ун-та, 1993. – 151 с.

Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга. – М.: Мир, 1965. – 480 с.

Розенталь М.М. Марксистский диалектический метод – М.: Госполитиздат, 1952.

Романенко В.Д., Оксиюк О.А., Жукинский В.Н. и др. Экологическая оценка воздействия гидротехнического строительства на водные объекты. – Киев: Наук. думка, 1990. – 256 с.

Россолимо Л.Л. Основы типизации озер и лимнологического районирования // Накопление вещества в озерах. – М.: 1964. С. 5-46.

Руководство по гидробиологическому мониторингу пресноводных экосистем. – СПб.: Гидрометеоиздат, 1992. – 318 с.

Руководство по методам гидробиологического анализа поверхностных вод и донных отложений / Под ред. В.А. Абакумова. – Л.: Гидрометеоиздат, 1983. – 239 с.

Рунион Р. Справочник по непараметрической статистике. - М.: Финансы и статистика, 1982. - 198 с.

Рябинин Н.А. Анализ фауны панцирных клещей темнохвойно-широколиственных лесов юга Дальнего Востока на основе методов теории множеств // Известия Академии наук, серия биологическая. 1993, вып. 2. С. 271-279.

Самарский А.А. Что такое вычислительный эксперимент? // Наука и жизнь. 1979. № 3. С. 27-33.

Саноцкий И.В. Индивидуальная реактивность и вероятность изменения здоровья человека при химических воздействиях (полемика по принципиальным вопросам) // Медицина труда и промышленная экология. 1993. № 3-4. С. 9-12.

Саноцкий И.В., Уланова И.П. Критерии вредности в гигиене и токсикологии при оценке опасности химических соединений. – М.: Медицина, 1975. – 328 с.

СанПиН 4630-88. Санитарные правила и нормы. Охрана поверхностных вод от загрязнения. – М.: Минздрав СССР, 1988. – 69 с.

СанПиН 2.1.5.980-00. Санитарные правила и нормы. Гигиенические требования к охране поверхностных вод. – М.: Минздрав России, 2000. – 24 с.

СанПиН 2.1.4.1074-01. Санитарные правила и нормы. Питьевая вода. Гигиенические требования к качеству воды централизованных систем питьевого водоснабжения. Контроль качества. – М.: Минздрав России, 2002. – 103 с.

Сафонов В.О. Экспертные системы – интеллектуальные помощники специалистов. – СПб: Санкт-Петербургская организация общества "Знания" России, 1992. – 196 с.

Сборник санитарно-гигиенических нормативов и методов контроля вредных веществ в объектах окружающей среды. – М.: Медицина, 1991. 134 с.

Свирежев Ю.М. Вито Вольтерра и современная математическая экология // Вольтерра В. Математическая теория борьбы за существование. – М.: Наука, 1976. С. 245-286.

Свирежев Ю.М. Нелинейные волны, диссипативные структуры и катастрофы в экологии. – М.: Наука, 1987. – 368 с.

Свирежев Ю.М., Логофет Д.О. Устойчивость биологических сообществ. - М.: Наука, 1978. - 350 с.

Себер Дж. Линейный регрессионный анализ. – М., Мир: 1980. – 286 с.

Северцов С.А. Динамика населения и приспособительная эволюция животных. – М.; Л.: Изд-во АН СССР, 1941. – 316 с.

Селтон Дж. Автоматическая обработка, хранение и поиск информации. – М.: Сов. радио, 1973. – 168 с.

Сёмкин Б.И. Эквивалентность мер близости и иерархическая классификация многомерных данных // Иерархические классификационные построения в географической экологии и систематике. – Владивосток: ДВНЦ АН СССР, 1979. С. 97-112.

Сёмкин Б.И., Двойченков В.И. Об эквивалентности мер сходства и различия // Исследование систем. – Владивосток: ДВНЦ АН СССР, 1973. С. 18-43.

Сердюцкая Л.Ф. Верификация экологических моделей круговорота азота (на примере Кременчугского водохранилища): Дисс. … канд. техн. наук. – Харьков., 1984. – 134 с.

Сердюцкая Л.Ф., Каменева И.П. Системный анализ и математическое моделирование медико-экологических последствий аварии на ЧАЭС и других техногенных воздействий. – Киев: "Медэкол" МНИЦ БИО-ЭКОС МЧС и НАН Украины, 2000. – 173 с.

Сидельников Ю.В. Теория и организация экспертного прогнозирования. – М.: ИМЭМО АН СССР, 1990. – 196 с.

Сильвестров Д.С. Программное обеспечение прикладной статистики. Обзор состояния. Тенденции развития. – М.: Финансы и статист., 1988. – 240 с.

Скурихин А.Н. Генетические алгоритмы // Новости искусственного интеллекта. 1995. № 4. С. 6-46.

Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. Изд. 2-е, испр. и доп. – М.: Наука, 1965. – 511 с.

Смит Дж. Математические идеи биологии. – М.: Мир, 1970. – 180 с.

Смолл Р.Д (Small R.D.) Интеллектуальный анализ данных: мифы и факты // Computer-weekly (Computer-Week Moskow). 1997. № 22-23. С. 38-39. Адрес в Интернет www.infoart.ru/it/press/cwm/22_97/data.htm

Соловьев В.Н., Фирсов А.А., Филов В.А. Фармакокинетика. – М.: Медицина, 1980. – 424 с.

Сотник С.Л. Основы проектирования систем с искусственным интеллектом (курс лекций). 2000. Адрес в Интернет http://www.alicetele.com/~sergei/index_e.htm

Сотник С.Л. Идентификация колебательного звена методом группового учета аргументов и искусственной нейронной сетью с генетическим алгоритмом обучения // Адрес в Интернет http://www.alicetele.com/~sergei/index_e.htm

Справочник по типовым программам моделирования / Под ред. А.Г. Ивахненко – Киев: Техника, 1980. – 184 с.

Справочник по прикладной статистике. В 2-х т. / Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, 1989. Т. 1. – 510 с.; 1990. Т. 2. – 526 с.

Старобогатов Я.И. Системный подход в экологии // Системные исследования. – М.: Наука, 1970. С. 114-118.

Статистический словарь. – М.: Финансы и статистика, 1989. – 623 с.

Стивенс С.С. Экспериментальная психология. Т.1. – М.: ИЛ, 1960. С. 5-78.

Стрельцов А.Б., Шпынов А.В., Гаркунов М.И. Организация биомониторинга в г. Калуге. // Антропогенные воздействия и здоровье человека. Вып. 1. – Калуга: 1995. С. 10-23.

Стьюпер Э., Брюггер У., Джурс П. Машинный анализ связи химической структуры и биологической активности. – М.: Мир, 1982. – 235 с.

Сукачев В.Н. Соотношение понятий биогеоценоз, экосистема и фация // Почвоведение. 1960. № 6. С. 4-10.

Сукачев В.Н. Основные понятия о биогеоценозах и общее направление их изучения // Программа и методика биогеоценотических исследований. – М.: Наука, 1966. С. 12-50.

Сурков Ф.А. и др. Моделирование абиотических факторов экосистемы Азовского моря // Изв. СКНЦ ВШ. Естеств. наук. 1977. № 2. С. 21-49.

Сущеня Л.М. Интенсивность дыхания ракообразных. – Киев: Наук. думка, 1972. – 195 с.

Тарасов А.Г. Биотический индекс дельты рек Волги и Северного Каспия: Автореф. дис. ... канд. биол. наук. – М.: МГУ, 1993. – 23 с.

Таунсенд К., Фохт Д. Проектирование и программная реализация экспертных систем на персональных ЭВМ. – М.: Финансы и статистика, 1990. – 346 с.

Ташкер И.Д. О статистических критериях и их использовании при гигиеническом контроле // Гигиена и санит. 1991. № 12. С. 85-87.

Ташкер И.Д. К проблеме установления безопасных уровней токсических веществ во внешней среде // Адрес в Интернет http://www.medved.kiev.ua/arhiv_mg/st_2000/00_2_9.htm

Телитченко М.М., Кокин К.А. Санитарная гидробиология. – М.: Изд. МГУ, 1968. – 103 с.

Теоретические вопросы классификации озер / Отв. ред. Н.П. Смирнов – СПб.: Наука, 1993. – 185 с.

Терентьев П.В. Метод корреляционных плеяд // Вест. Ленингр. ун-та. Серия биол. 1959. № 9. Вып. 2. С. 137-141.

Терехина А.Ю. Анализ данных методами многомерного шкалирования. - М.: Наука, 1986. - 168 с.

Терехова В.А., Шитиков В.К., Семенова Т.А. Микромицеты Куйбышевского водохранилища. IV. Взаимодействие с абиотическими и биотическими факторами // Микология и фитопатология, 1998. Т. 32. Вып 1. С. 44-48.

Тимофеев-Ресовский Н.В., Яблоков А.В., Глотов Н.В. Очерк учения о популяции. – М.: Наука, 1973. – 277 с.

Тихонов В.Н., Шитиков В.К. Анализ вариабельности признаков и биологическая норма // Гигиена и санит. 1984а. № 7. С. 63-64.

Тихонов В.Н., Шитиков В.К. К вопросу прогнозирования кумулятивных свойств веществ в токсикологических исследованиях// Гигиена и санит. 1984б. № 4. С. 79-80.

Тихонов В.Н., Шитиков В.К., Кондратов С.А. О математическом моделировании интоксикации при повторяющихся воздействиях // Гигиена и санит. 1987. № 7. С. 62-65.

Тодераш И.К. Функциональное значение хирономид в экосистемах водоемов Молдавии. – Кишинев: Штинница, 1984. – 172 с.

Толковый словарь по охране природы / Под ред. В.В. Снакина. – М.: Экология, 1995. – 191 с.

Трахтенберг И.М. Приоритетные аспекты проблем медицинской экологии в Украине (взгляд токсиколога) // Совр. проб. токсикологии. 1998. № 1. С. 5-8.

Трахтенберг И.М., Сова Р.Е., Шефтель В.О., Оникиенко Ф.А. Проблемы нормы в токсикологии. – М.: Медицина, 1991. – 208 с.

Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 410 с.

Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 693 с.

Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Финансы и статистика, 1995. – 384 с.

Уинстон П. Искусственный интеллект. – М.: Мир, 1980. – 520 с.

Уиттекер Р. Сообщества и экосистемы. – М.: Прогресс, 1980. – 328 с.

Умнов А.А. Математическая модель биотического круговорота в озерной системе // Гидробиол. журн. 1972. Т. 8. Вып. 5. С. 5-13.

Умнов А.А. Математическая модель биотического круговорота вещества и энергии, происходящего в загрязненной реке // Биологические процессы и самоочищение на загрязненном участке реки (на примере Верхнего Днепра). – Минск: БГУ, 1973. С 157-190.

Умнов А.А. Применение статистических методов для оценки параметров эмпирических уравнений, описывающих взаимосвязь между энергетическим обменом и массой тела животного // Журн. общ. биол. 1976. Т. 37. № 1. С. 71-86.

Умнов А.А. Использование математических моделей для оценки экологического состояния водоемов (на примере экосистемы Невской губы) // Экологическое состояние водоемов и водотоков бассейна реки Невы. – СПб.: СПбНЦ РАН, 1996. С. 155-186.

Уморин П.П. Оценка интенсивности выедания инфузориями водорослей и бактерий // Зоол. журн. 1983. № 3. С. 325-330.

Уникальные экосистемы солоноватых карстовых озер Среднего Поволжья. – Казань: Изд-во Казан. Ун-та, 2001. – 256 с.

Унифицированные методы исследования качества вод. Ч. 3. Методы биологического анализа вод. – М.: Изд-во СЭВ, 1977. – 175 с.

Уоссермен Ф. Нейрокомпьютерная техника. – М.: Мир, 1992. – 184 с.

Уотермен Д. Руководство по экспертным системам. – М.: Мир, 1980. – 384 с.

Урбах В.Ю. Математическая статистика для биологов и медиков. – М.: Изд-во АН СССР, 1963. – 323 с.

Урбах В.Ю. Дискриминантный анализ и его применение в биологической систематике и медицинской диагностике // Применение математических методов в биологии. Вып. 3. – Л.: 1964. С. 67-87.

Урманцев Ю.А. Что может дать биологу представление объекта как системы в системе объектов того же рода? // Журн. общ. биол. 1978. Т. 39. № 5. С. 699-718.

Федеральный закон от 10.01.2002 г. № 7-ФЗ "Об охране окружающей среды" / Принят Государственной Думой 20 декабря 2001 года, одобрен Советом Федерации 26 декабря 2001 года. - 31 с.

Федоров В.Д., Устойчивость экологических систем и ее измерение // Изв. АН СССР. Сер. биол. 1974. № 3. С. 402-415.

Федоров В.Д. Заметки о парадигме вообще и экологической парадигме в частности // Вести. МГУ. Сер. биол. 1977а. № 3. С. 8-22.

Федоров В.Д. Проблемы оценки нормы и патологии состояния экосистем // Научные основы контроля качества поверхностных вод по гидробиологическим показателям. - Л.: 1977б. С. 6-12.

Федоров В.Д., Гильманов Т.Г. Экология. – М.: МГУ, 1980. – 464 с.

Федоров В.Д., Максимов В.Н., Сахаров В.Б. Количественный способ оценки внешних воздействий на экологические системы // Человек и биосфера. – М.: МГУ, 1980. Вып. 5. С. 12-23.

Федоров В.Д., Сахаров В.Б., Левич А.П. Количественные подходы к проблеме оценки нормы и патологии экосистем // Человек и биосфера. – М.: МГУ, 1982. Вып. 6. С. 3-42.

Финогенова Н.П. Значение олигохет как индикаторов загрязненных вод // Методы биологического анализа пресных вод. – Л.: ЗИН АН СССР, 1976. С. 51-59.

Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989. – 319 с.

Флейшман Б.С. Системные методы в экологии // Статистические5 методы анализа почв, растительности и их связи. – Уфа: БФАН СССР, 1978. С. 7-28.

Флейшман Б.С. Основы системологии. – М.: Радио и связь, 1982. – 368 с.

Флейшман Б.С. Системология, системотехника и инженерная экология // Кибернетика и ноосфера. – М.: Наука, 1986. С. 97-110.

Флейшман Б.С., Брусиловский П.М., Розенберг Г.С. О методах математического моделирования сложных систем // Системные исследования. Ежегодник. – М.: Наука, 1982. С. 65-79.

Фогель Л., Оуэнс А., Уолш М. Искусственный интеллект и эволюционное моделирование. – М.: Мир, 1969. – 230 с.

Фомин Г.С. Вода. Контроль химической, бактериальной и радиационной безопасности по международным стандартам. Энциклопедический справочник. – М.: Протектор, 1995. – 624 с.

Фомин Я.А., Тарловский Г.Р. Статистическая теория распознавания образов. – М.: Радио и связь, 1986. – 264 с.

Форощук В.П. Водоохранная деятельность и экологическое нормирование качества водной среды // Гидробиол. журн. 1989. Т. 25. № 1. С. 36-41.

Форрестер Дж. Антиинтуитивное поведение сложных систем // Современные проблемы кибернетики. – М.: Знание, 1977. С. 9-25.

Форрестер Дж. Мировая динамика. – М.: Наука, 1978. – 167 с.

Фрей Т.Э.-А. О математико-фитоценотических методах классификации растительности: Автореф. дис. … докт. биол. наук. – Тарту: ТартГУ, 1967. – 32 с.

Фрумин Г.Т. Экологически допустимые уровни воздействия металлами на водные экосистемы // Биол. внутр. вод. 2000. № 1. С. 125-134.

Фрумин Г.Т., Баркан Л.В. Комплексная оценка загрязненности Ладожского озера по гидрохимическим показателям // Водн. ресурсы. 1997. Т. 24. № 3. С. 315-319.

Фу К. Структурные методы в распознавании образов. – М.: Мир, 1977. – 320 с.

Фукс В. По всем правилам искусства. Точные методы в исследованиях литературы, музыки и изобразительного искусства // Моль А., Фукс В., Касслер М. Искусство и ЭВМ. – М.: Мир, 1975. – 556 с.

Хайлов К.М. Системный подход в экологии // Системные исследования. – М.: Наука, 1970. С. 118-122.

Хальд А. Математическая статистика с техническими приложениями. – М.: Изд. иностр. лит., 1956. – 664 с.

Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969. – 396 с.

Хант Э. Искусственный интеллект. – М.: Мир, 1978. – 558 с.

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

Хатчинсон Д. Лимнология. – М.: Прогресс, 1969. – 592 с.

Хлебович Т.В. Значение инфузорий в оценке степени загрязнения вод // Методы биологического анализа пресных вод. – Л.: ЗИН АН СССР, 1976. С. 59-68.

Хованов Н.В. Математические основы теории шкал измерения качества. – Л.: ЛГУ, 1982. – 185 с.

Хокс Х.А. Биологический контроль качества речной воды // Научные основы контроля качества поверхностных вод по гидробиологическим показателям. Тр. Советско-английского семинара. – Л.: Гидрометеоиздат, 1977. С. 172-178.

Холлендер М., Вулф Д.А. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.

Хьюбер П. Робастность в статистике. – М.: Мир, 1984. – 304 с.

Хэтч Т. Значение допустимых пределов воздействия опасных веществ, присутствующих в воздухе рабочей зоны, рассматриваемое в аспекте предупреждения профессиональных заболеваний // Бюлл. ВОЗ. 1973. Т. 47. № 2. С. 153-161.

Царегородцев В.Г., Погребная Н.А. Нейросетевые методы обработки информации в задачах прогноза климатических характеристик и лесорастительных свойств ландшафтных зон // Методы нейроинформатики (сборник научных трудов). – Красноярск: КрГТУ, 1998. С. 65-110.

Цейтлин Н.А. Применение методов математической теории эксперимента в содовой промышленности. – М.: НИИТЭХИМ, 1984. – 37 с.

Цейтлин Н. Из опыта аналитического статистика. / Адрес в Интернет http://people.freenet.de/nzarchiv/buecher/ или http://matstat.gmxhome. de/

Цимдинь П.А. Коловратки как биоиндикаторы сапробности // Гидробиол. журн. 1979. Т. 15. № 4. С. 63-67.

Ципилева Т.А. Методы автоматической классификации в сжатии экологической информации // Алгоритмическое и информационное обеспечение систем экоинформации. – Томск: СО АН СССР, 1989. С. 23-61.

Цыпкин Я.З. Адаптация и обучение в автоматических системах. – М.: Наука, 1968. – 399 с.

Цыпкин Я.З. Основы информационной теории идентификации. – М.: Наука, 1984. – 520 с.

Цыплаков А.А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии, Новосибирск: Изд-во НГУ, 1997. 129 с. / Адрес в Интернет http://www.nsu.ru/ef/tsy/ecmr/study.htm

Чайковский Ю.В. Познавательные модели, плюрализм и выживание // Путь. 1992. № 1. С. 62-108.

Человек. Медико-биологические данные. – М.: Знание, 1977. – С. 36-40.

Чернов Ю.И. Понятие "животное население" и принципы геозоологических исследований // Журн. общ. биологии. 1971. Т. 32. № 4. С. 425 -438.

Чернов Ю.И. Природная зональность и животный мир суши. – М.: Мысль, 1975. – 222 с.

Черп О.М., Виниченко В.Н., Хотулёва М.В. и др. Экологическая оценка и экологическая экспертиза – М.: СоЭС – Методический центр «Эколайн», 2000. Адрес в Интернет www.cci.glasnet.ru\mc\ books\eiabook

Чертопруд М.В. Модификация метода Пантле-Букка для оценки загрязнения водотоков по качественным показателям макробентоса // Водн. ресурсы. 2002. Т. 29. № 3. С. 337-342.

Чупров А.А. Основные проблемы теории корреляции. – М.: Госиздат, 1960. – 176 с.

Шайтура С.В. Геоинформационные системы и методы их создания. – М.: 1998. – 252 с.

Шакин В.В. Выбор критериев классификации в методе собственных подпространств // Математическая обработка медико-биологической информации. – М.: Наука, 1976. С. 103-115.

Шакин В.В. Биосистемы в экстремальных условиях // Журн. общ. биол. 1991. Т. 52. № 6. С. 784-792.

Шафаревич И.Р. Доклад на собрании Японского математического общества от 28 сентября 1993 г. Адрес в Интернет http://www.doktor.u/doctor/biometr/naukoved/mathem.htm

Шварц С.С. Популяционная структура биоценоза // Изв. АН СССР. Сер. биол. 1971. № 4. С. 485-493.

Шеннон К. Математическая теория связи / Работы по теории информации и кибернетике. – М.: Ин. литер., 1963. С. 243-332.

Шестакова Г.А., Стрельцов А.Б., Логинов А.А. и др. Система регионального биологического мониторинга (на примере Калужской области) // Вопросы географии и геоэкологии. – Калуга: 1998. Вып. 2. С. 75-88.

Шитиков В.К., Зинченко Т.Д., Головатюк Л.В. Математические аспекты оценки патологии экосистем на примере зообентоса малых рек Самарской области // Малые реки: Современное экологическое состояние, актуальные проблемы. (Тез. Межд. науч. конф. Россия, г.Тольятти, 23-27 апреля 2001г.). – Тольятти: ИЭВБ РАН, 2001. С. 230.

Шитиков В.К., Зинченко Т.Д., Головатюк Л.В. Нейросетевые методы оценки качества поверхностных вод по гидробиологическим показателям // Изв. СамНЦ РАН. 2002. Т. 4. № 2. С. 280-289.

Шитиков В.К., Коппа Ю.В., Курляндский Б.А., Тихонов В.Н. Прогнозирование токсикологических показателей химических веществ методом самоорганизации моделей // Автоматика. 1986. № 4. С. 85-87.

Шитиков В.К., Тихонов В.Н., Быков С.Т., Ковалев А.Ф. Статистический анализ и нормальность распределения выборок в токсиколого-гигиенических исследованиях // Гигиена и санит. 1985. № 3. С. 61-62.

Шмальгаузен И.И. Рост и дифференцировка // Рост животных. – М.: АН СССР, 1935. С. 74-84.

Шмальгаузен И.И. Кибернетические вопросы биологии. – Новосибирск: Наука, 1968. – 396 с.

Шмидт В.М. Биометрический метод в ботанической систематике // Ботан. журн. 1964. Т. 49. № 1. С. 315-324.

Шноль С.Э. Герои, злодеи, конформисты российской науки. – М.: КРОН-ПРЕСС, 2001. – 875 с.

Штабский Б.М. Об оценке кумулятивных свойств вредных веществ по В.Н. Тихонову и В.К. Шитикову // Гигиена и санит. 1985. № 1. С. 70-73.

Штабский Б.М. О концепции дифференцированных во времени ПДК вредных веществ в связи с проблемой санитарной охраны водоемов // Гигиена и санит. 1986. № 11. С. 57-60.

Штабский Б.М., Каган Ю.С. К оценке кумулятивных свойств химических веществ по индексу и стандартизованному коэффициенту кумуляции // Гигиена и санит. 1974. № 3. С. 65-67.

Шуберт Р. Основные принципы методов биоиндикации // Изучение загрязнения окружающей природной среды и его влияния на биосферу: Матер. 3 заседания Междунар. рабочей группы по проекту № 14 МАБ ЮНЕСКО. – Л., 1986. С. 112-122.

Шустер Г. Детерминированный хаос: Введение. – М.: Мир, 1988. – 240 с.

Щукарев С.А. Попытка общего обзора грузинских вод с геохимической точки зрения // Труды Гос. ин-та курортологии. – М.: Медиздат, 1934. Т.4.

Щукин И.С. Общая геоморфология. – М.: МГУ, 1964. Т. 2. – 564 с.

Экологическая оценка. Пособие для преподавателей (UNEP Environmental Impact Assessment Training Resource Manual) / Программа ООН по окружающей среде (UNEP) . Пер. РОО "Эколайн" и Центра подготовки и реализации инвестиционных проектов (ЦПРП) Адрес в Интернет http://cci.glasnet.ru/mc/books/eiamanual/index.html

Экологический мониторинг. Методы биомониторинга / Под ред. Д.Б. Гелашвили. – Н. Новгород: ННГУ, 1995. Вып. 1. – 190 с.

Экологическое районирование пресноводных водоемов. – Рыбинск: ИБВВ АН СССР, 1990. Вып. 62(65). – 176 с.

Экологическое состояние бассейна реки Чапаевка в условиях антропогенного воздействия (Биологическая индикация). - Тольятти: ИЭВБ РАН, 1997. – 337 с.

Элти Д., Кумбс М. Экспертные системы: концепции и примеры. – М.: Финансы и статистика, 1987. – 191 с.

Эрроусмит Д., Плейс К. Обыкновенные дифференциальные уравнения. Качественная теория с приложениями. – М.: Мир, 1986. – 243 с.

Эшби У.Р. Введение в кибернетику. – М.: Ин. литер., 1959. – 432 с.

Эшби У.Р. Несколько замечаний // Общая теория систем. – М.: Мир: 1966. С. 171-178.

Эшби У.Р. Математические модели и анализ на вычислительных машинах функций центральной нервной системы // Автоматика. 1967. № 1. С. 57.

Яглом А.М., Яглом И.М. Вероятность и информация. – М.: Наука, 1973. – 512 с.

Яковлев В.А. Методы оценки качества вод по зообентосу озера Имандра // Мониторинг природной среды Кольского Севера. – Апатиты, 1984. С. 39-50.

Яковлев В.А. Оценка качества поверхностных вод Кольского Севера по гидробиологическим показателям и данным биотестирования (практические рекомендации). – Апатиты, 1988. – 27 с.

Яковлев В.А. Оценка степени закисления поверхностных вод северо-восточной части Фенноскандии по зообентосу // Водн. ресурсы. 1998. Т. 25. № 2. С. 244-251.

Яковлев В.А. Формирование размерной структуры сообществ зообентоса северо-восточной части Фенноскандии в зависимости от природных условий и антропогенных факторов // Тезисы докладов 8 съезда Гидробиологического общества РАН. Том 1. – Калининград, 2001. С. 320-321.

Ястребов А.Б. Методы изучения мозаичности растительного покрова с применением ЭВМ. – Л.: ЛГУ, 1991. – 200 с.

Andrewartha H.G., Birch L.C. The Distribution and Abundance of Animals. – Chicago: Univ. Press, 1954. - 782 p.

Bertalanffi L. fon. Basic concepts in quantitative biology of metabolism // Helgol and Wiss. Meeresuntersuch, 1964. V. 9. № 1. P. 5-34.

Bishop C. Neural Networks for Pattern Recognition. – Oxford: University Press, 1995. – 432 p.

Boysen-Yensen P. Valiation of the Limfyiord // I.-Rep. Dan bion. sta. 1919. V. 26. № 1. P. 1-44.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and regression trees. – Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. – 358 p.

Canter L.W. Environmental Impact Assessment. 2nd Ed. – NY.: McGraw-Hill, 1996. – 587 p.

Constandse-Westermann T.S. Coefficients of Biological Distance. – N.Y.: Humanities Press, 1972. – 142 p.

Critical loads for sulphur and nitrogen (Report from a Workshop held at Stokhoster, Sweden, March 19-24, 1988) / Miljo rapport, 1988 – Copenhagen: Nordic Council of Ministers, 1988. Vol. 15.

Czekanowcki J. Objectiv kriterien in der ethologie // Korrespondenz-blatt der Deutschen Gesellschaft fur Antropologie, Ethnologie, und Urgeschichte, 1911. 42. S. 1-5.

Dawkins R. The Selfish Gene. – NY.: Oxford University Press, 1976. – 112 p.

Dubois D.M. A model of patchines for prey-predator plankton populations // Ecol. Modeling. 1975. № 1. P. 67-80.

Efroimson M.A. Multiple regression analysis // Mathematical Methods for Digital Computers. 1960. V. 1. P. 191-203.

Goldberg D.E. Genetic Algorithms. – Reading, MA: Addison Wesley, 1989. – 236 p.

Goodall D.W. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis // Austral. J. Bot. 1954. V. 2. № 3. P. 304-324.

Goodall D.W. The continuum and individualistic association // Vegatatio. 1963. V. 11. № 5-6. P. 297-316.

Goodall D.W. Numerical classification // Handbook of Vegetation Science. Pt. 5. - The Haque: Dr. W.Junk, 1973. P. 105-156.

Gower J.C., Ross G.J. Minimum spanning trees and single linkage cluster analysis // Appl. Stat. 1969. V. 18. № 1. P. 54-64.

Gray J. Detecting pollution induced changes in communities using the lognormal distribution of individuals among species // Mar. Pollut. Bull. 1981. V. 12. № 5. P. 173-176.

Guidelines for drinking-water quality. – Geneva: World Health Organization.Vol. 1 – Recommendations. 1983; Vol. 2 – Health Criteria and Other Supporting Information. 1984.

Heatwole H. The concept of the econe, a fundamental ecological unit // Trop. Ecol. 1989. V. 30. № 1. P. 13-19.

Hill M.O. DECORANA and TWINSPAN, for ordination and classification of multivariate species data: a new edition, together with supporting programs, in FORTRAN 77. – Huntingdon: Inst. Of Terrestrial Ecology, 1979. – 58 p.

Hill M.O., Bunce R., Show M. Indicator species analysis, a divisive polithetie method of classification, and its application to a survey of native pinewoods in Scotland // J. Ecol. 1975. V. 63. № 9. P. 721-727.

Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities // Proc. Nat. Acad. Sci. USA. 1982. Vol. 79. P.2554-2558. Адрес в Интернет http://www-windows-1251.edu.yar.ru/russian/misc/eco_page/bioind/index.html

Hunt E.B., Marin J., Stone P.J. Experiments in Induction. – NY.: Academic Press, 1966. – 247 p.

Kershaw K.A. Quantitative and Dynamic Plant Ecology. Ed. 2. – London, 1974. – 308 p.

Kohonen T. Self-organized formation of topologically correct feature maps // Biological Cybernetics. 1982. № 43. P. 59-69.

Kolkwitz R., Marsson M. Grundsatze fur die biologishe Beurtheilung des Wassers hach seinerFlora und Fauna // Mitteil. aus der konigl. Prufungang fur Wasserbesorg. und Abwasserbes. 1902. H. 1. S. 33.

Kownacki A. Taxocens of Chironomidae in streams of the Polish Hight Tatra, Mts // Acta Hydrobiol. 1971. V. 13. № 2. P. 439-463.

Lim R.K., Runk K.G., Glass H.G., Soaje-Echague F. A method for the evolution of cumulative by the determination of acute and subchronic median effective doses // Arch. Inter. Pharm. Ther. 1961. V. 130. P. 336-352.

Lindeman R.L. The trophic-dynamic aspect of ecology // Ecology. 1942. V. 23. № 4. P. 399-418.

Lindman H. R. Analysis of variance in complex experimental designs. – San Francisco: W. H. Freeman & Co., 1974.

Loh W.-Y., Shih Y.-S. Split selection methods for classification trees // Stat. Sinica, 1997. V. 7. P. 815-840.

MacArtur R.H. On the relative abundance of bird species // Proc. Natl. Acad. Sci. USA. 1957. V. 45. P 293-295.

Machine Learning, Neural and Statistical Classification / Ed. D. Mitchie et al. – Ellis Horwood, Chichester, UK, 1994. - 304 p.

Manahan S.E. Environmental Chemistry. – NY.: Lewis Publishers, 1994. – 789 p.

Margalef R. Temporal succession and spatial heterogeneity in phytoplankton // Perspectives in Marine Biology. – Berkeley: Univ. of California Press, 1958. P. 323-347.

Margalef R. Perspectives in Ecological Theory. – Chicago: Univ. Press, 1968. – 123 p.

McIntosh R.P. Matrix and plexus techniques // Handbook of Vegetation Science. Pt. 5. – The Hague: Dr. W. Junk, 1973. P. 157-221.

Murthy S. Automatic construction of decision trees from data: A Multi-disciplinary survey // Data Mining and Knowledge Discovery (Kluwer Academic Publishers, USA). 1998. V. 2. № 4. P. 345 - 389 p.

Parzen E. On the estimation of probability density function and the mode // Ann. Math. Stat. 1962. V. 33. P. 1065-1076.

Pielou E.C. Shannon's formula as a measure of species diversity: its use and misure // Amer. Natur. 1966. V. 100. P. 463-465.

Pielou E.C. Ecological Diversity. – NY.: Gordon & Breach Sci. Publ., 1975. – 165 p.

Poroikov V.V., Filimonov D.A., Borodina Yu.V. et al. Robustness of Biological Activity Spectra Predicting by Computer Programm PASS for Noncongenetic Sets of Chemical Compounds // J. Chem. Inform. Comput. Sci. 2000. V. 40. № 6. P. 1349.

Quinlan J.R. C4.5: Programs for Machine learning. – San Mateo: Morgan Kaufmann Publishers, 1993. – 302 p.

Rosenberg D.M. Freshwater biomonitoring and Chironomidae // Neth. J. Aquat. Ecol. 1993. V. 26. № 2-4. P. 101-122.

Rozenberg G.S. Expert systems "REGION" and "RESERVOIR" as instruments of simulation of diffuse polution of large-scale ecosystems and reservoirs // Proceeding of the Second International IAWQ Specialized Conference on Diffuse Pollution. – Brno: Prague (Czech Repub.), 1995. Part 1. P. 72-77.

Rozenberg G.S., Krestin S.V. System of analytical models of processes of eutrophication in the reservoir (block approach) // Programme and Abstracts. 3rd International Conference on Reservoir Limnology and Water Quality. – Ceske Budejovice (Czech Republic), 1997. P. 151.

Rozenberg G.S., Shitikov V.K. Expert systems "REGION" as instruments of simulation of large-scale ecosystems and reservoirs // Экологические проблемы бассейнов крупных рек: Тез. докл. Междунар. конф. – Тольятти: ИЭАВБ РАН, 1993. С. 264.

Shannon C.B., Weaver W. The Mathematical Theory of Communication. – Urbana (Illinois): Univ. of Illinois Press, 1963. – 345 p.

Simpson E.H. Measurement of diversity // Nature (London). 1949. V. 163. № 4148. P. 668.

Sládeček V. The future of the saprobity system // Hydrobiologia. 1965. V. 25. № 3-4.

Sladeček V. System of water quality from the biological point of view. // Arch. Hydrobiol., Beihefz., Ergebnisse der Limnol. 1973. Bd 7. S. 1-218.

Sokal R., Sneath P. Principles of Numerical Taxonomy. - San Francisco: W.H. Freeman, 1963. - 573 p.

Steele J.H. The structure of marine ecosystems. – Cambrige (Massachusets): Harv. Univ. Press, 1974. – 110 p.

Tansley A.G. The use and anuse of vegetation concepts and terms // Ecology. 1935. V. 16. P. 248-307.

Tryon R.C. Cluster Analysis. – NY.: McGraw-Hill, 1939.

Vittikh V.A. Engineering theories as a basis for integrating deep engineering knowledge // Artifical Intel. in Engin. 1997. V. 1. P. 25-30.

Ward J.H. Hierarchical grouping to optimize an objective function // J. Amer. Statist. Assoc. 1963. V. 58. № 301. P. 236-244.

West G.B., Brown J.H., Enquist B.J. The fourth dimension of life: fractal geometry and allometric scaling of organisms // Science. 1999. V. 284. P. 1677-1679.

Westhoff V., van der Maarel E. The Braun-Blanquet approach // Classification of Plant Communities. – The Hague: Dr. W.Junk, 1978. P. 287-399.

Zadeh L. Fuzzy Sets // Information and Control. 1965. V. 8. № 3. P. 338-353.

Zinchenko T.D. Long-term (30 years) dynamics of chironomidae (*Diptera*) fauna in the Kuibyshev water reservoir associated with eutrophication processes // Netherlands J. of Aquatic Ecology. 1992. V. 26. № 2-4. P. 533-542.