



Лобырев Ф.С.

ОРДИНАЦИОННЫЕ МЕТОДЫ

Учебное пособие по курсу лекций
«Практическая биостатистика»

*Московский государственный университет им. М.В. Ломоносова
Биологический факультет
Кафедра ихтиологии*

Лобырев Ф.С.

ОРДИНАЦИОННЫЕ МЕТОДЫ

**Учебное пособие по курсу лекций
«Практическая биостатистика»**

Белгород, 2026

УДК 597.2/.5(075.8)
ББК 28.693.33с51я73
Л68

Лобырев, Федор Сергеевич.
Л68 Ординационные методы : учебное пособие по курсу лекций «Практическая биостатистика» / Ф. С. Лобырев ; Московский государственный университет им. М. В. Ломоносова, Биологический факультет, Кафедра ихтиологии. – Белгород : КОНСТАНТА, 2026. – 174 с. : ил.

УДК 597.2/.5(075.8)
ББК 28.693.33с51я73

ISBN 978-5-908148-10-8

ОГЛАВЛЕНИЕ

1. ВВЕДЕНИЕ	6
2. СТАНДАРТИЗАЦИЯ ДАННЫХ	8
3. АНАЛИЗ И ИНТЕПРЕТАЦИЯ ОРДИНАЦИОННЫХ ДИАГРАММ.....	10
Scatter Plot. Ординация объектов/проб	10
Biplot. Ординация объектов и переменных	12
Triplot. Ординация объектов, переменных и видов	14
4. ПАНЕЛЬ НАСТРОЕК ВИЗУАЛИЗАЦИИ ОРДИНАЦИОННЫХ ДИАГРАММ.....	18
5. МЕРЫ ДИСТАНЦИИ	20
6. КЛЮЧ ДЛЯ ВЫБОРА МЕТОДА ОРДИНАЦИИ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ.....	22
7. АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ (РСА).....	27
Введение	27
Методология	27
Характер данных	29
Практическая задача. Режим Disregard	33
Практическая задача. Режим Within-group	41
Практическая задача. Режим Between-group	44
Философское эхо	48
8. АНАЛИЗ ГЛАВНЫХ КООРДИНАТ (РС_oA).....	49
Введение	49
Методология	50
Характер данных	51
Практическая задача	52
Философское эхо	56
9. НЕМЕТРИЧЕСКОЕ МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ (nMDS).....	57
Введение	57

Методология	57
Характер данных	58
Пример анализа данных	60
Философское эхо	65
10. АНАЛИЗ СООТВЕТСТВИЙ (CA)	66
Введение	66
Методология	67
Характер данных	68
Пример анализа данных	69
Философское эхо	74
11. ДЕТРЕНДИРОВАННЫЙ АНАЛИЗ СООТВЕТСТВИЙ (DCA)	75
Введение	75
Методология	76
Характер данных	77
Пример анализа данных	79
Философское эхо	83
12. КАНОНИЧЕСКИЙ АНАЛИЗ СООТВЕТСТВИЙ (CCA)	84
Введение	84
Методология	85
Характер данных	86
Пример анализа данных	88
Философское эхо	93
13. АНАЛИЗ ИЗБЫТОЧНОСТИ (RDA)	95
Введение	95
Методология	96
Характер данных	97
Пример анализа данных	98
Философское эхо	105
14. ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ (LDA)	106
Введение	106
Методология	107
Характер данных	108

Пример анализа данных	110
Философское эхо	117
15. Calgary And Brown Factor Analysis (CABFAC).....	118
Введение	118
Методология	119
Характер данных	121
Пример анализа данных	122
Философское эхо	138
16. Метод ANOSIM	139
Введение	139
Методология	140
Характер данных	142
Пример анализа данных. One-way ANOSIM	144
Пример анализа данных. Two-way ANOSIM	148
Философское эхо	150
17. Метод PERMANOVA.....	151
Введение	151
Методология	153
Характер данных	154
Пример анализа данных. One-way PERMANOVA	155
Пример анализа данных. Two-way PERMANOVA	160
Философское эхо	162
18. Метод SIMPER.....	163
Введение	163
Методология	164
Характер данных	165
Пример анализа данных	167
Философское эхо	169
19. Сравнительный анализ ANOSIM, PERMANOVA и SIMPER.....	170
20. СПИСОК ЛИТЕРАТУРЫ.....	172
21. СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ.....	173

1. ВВЕДЕНИЕ

При изучении природных популяций исследователь часто сталкивается с необходимостью осмысленного анализа сложных многомерных данных. Стандартные подходы быстро оказываются несостоятельными при работе с десятками видов, сотнями проб и множеством параметров среды, образующих систему со сложными взаимосвязями. Подобные массивы характеризуются высокой размерностью и неочевидной внутренней организацией; их многомерная природа требует специальных методов, учитывающих совместное влияние всех компонентов системы. Возникает ключевой вопрос: как выявить в этом массиве скрытые структуры, обнаружить устойчивые закономерности и перейти от простого описания к содержательной интерпретации? Для решения подобных задач применяется совокупность статистических методов, объединяемых понятием многомерной ординации.

Ординационные методы представляют собой инструмент упорядочивания и визуализации данных, позволяющий проецировать многомерные отношения между объектами (пробами) и переменными (видами) в пространство меньшей размерности (обычно 2–3 измерения) (Benzécri, 1969). Это не просто технический приём, а целостный исследовательский подход, исходящий из предположения, что за кажущимся хаосом данных скрываются упорядоченные градиенты и структурные закономерности. Ординация позволяет математически выявлять и подтверждать подобные экологические закономерности (Gower, 1966).

Фундаментальная предпосылка интерпретации ординации заключается в следующем: структура сообщества, его видовой состав и обилие формируются под влиянием комплекса абиотических условий и биотических взаимодействий. Эти часто взаимосвязанные факторы создают многомерный градиент условий, определяющий распределение видов. Следовательно, анализируя изменчивость биотических данных, можно делать обоснованные предположения о скрытых экологических факторах, стоящих за наблюдаемыми закономерностями. Ординация служит инструментом, который позволяет не только выявить скрытую структуру, но и соотнести параметры среды со структурой сообщества, обеспечивая переход от описания к содержательной экологической интерпретации.

Все многообразие ординационных методов делят на два больших класса, соответствующих разным этапам исследования. Непрямые (неограниченные) методы, такие как анализ главных компонентов (PCA) или неметрическое многомерное шкалирование (nMDS), основаны преимущественно на биотической матрице «пробы × виды» и позволяют выявлять внутреннюю структуру данных: основные градиенты, группы и возможные выбросы. Они используются на стадии первичного, разведочного анализа, когда необходимо получить общее представление о структуре данных и сформулировать предварительные гипотезы. Прямые (ограниченные) методы, например, канонический анализ соответствий (CCA) и анализ избыточности (RDA), строят ординацию с учётом матрицы предикторов, что позволяет напрямую проверять гипотезы о факторах среды, связанных с распределением видов. С их помощью можно количественно оценить вклад отдельных факторов и статистически проверить значимость выявленных взаимосвязей. Для методов прямой ординации тщательный отбор предикторов является ключевым условием корректного анализа и интерпретации (Kruskal, 1964).

Настоящее методическое пособие посвящено практическому освоению ординационных методов в среде бесплатного программного обеспечения PAST. Выбор этой платформы обусловлен тем, что PAST является универсальным и доступным инструментом: использование программы не требует покупки лицензии и снижает барьер входа в многомерный анализ. Интерфейс PAST относительно прост по сравнению с другими коммерческими аналогами, но логичен и интуитивно понятен, что позволяет в короткий срок освоить работу в программе. Сочетание функциональности и доступности делает PAST удобной стартовой площадкой для перехода от теории к практической работе с многомерными данными. На рынке статистического программного обеспечения доступны различные альтернативы, такие как CANOCO или PRIMER, однако PAST уверенно занимает нишу «первого шага», позволяющего в деталях освоить ключевые этапы многомерного анализа — от подготовки данных и выбора метода до его реализации и интерпретации результатов, что облегчает переход к более сложным инструментам с сохранением понимания сути аналитических процедур.

Пособие снабжено дополнительными материалами и практическими задачами, доступными в репозитории, что позволяет отрабатывать ординационные методы на реальных наборах данных: <https://github.com/Feodor1974/Ordination>

2. СТАНДАРТИЗАЦИЯ ДАННЫХ

Успех ординационного анализа напрямую зависит от корректной предварительной обработки исходных данных. Поскольку ординационные методы чувствительны к масштабу и распределению переменных, ключевым этапом является их стандартизация, целью которой является устранение произвольного влияния единиц измерения и выравнивание вклада различных переменных в итоговые ординационные модели. Выбор стратегии стандартизации принципиально различается для биотических и абиотических данных.

Абиотические переменные (факторы среды), такие как содержание элементов в почве, pH или температура, зачастую измеряются в принципиально разных единицах и имеют несопоставимые масштабы и дисперсию. Без приведения их к единому масштабу переменная с большими численными значениями (например, содержание кальция) искусственно получит неоправданно большой вес в анализе по сравнению с переменной, значения которой малы (например, pH). Для нивелирования данного эффекта наиболее универсальным является Z-стандартизация, в результате чего все абиотические параметры приобретают нулевое среднее и единичную дисперсию, становясь сравнимыми между собой и внося равный вклад в построение градиентов (Legendre, Legendre, 2012). Данная процедура необходима для приведения всех исходных признаков к единому масштабу, что устраняет доминирование признаков с большей дисперсией и обеспечивает более корректную совместную обработку параметров среды в рамках одной ординационной модели.

Иную физическую природу имеют биотические данные, представленные матрицей обилия видов. Здесь основной проблемой является не разница в единицах измерения (все виды измеряются, например, в процентах покрытия или количестве особей), а специфическое распределение данных: наличие множества нулей (отсутствие вида в пробе), резкая асимметрия (большинство видов редки, немногие — массовы), а также ситуации, когда редкий вид в одной пробе может быть массовым в другой. В результате такие матрицы плохо соответствуют предпосылкам многих стандартных методов многомерного анализа, основанных на евклидовой метрике, и требуют предварительного преобразования. Для экологических данных наилучшим компромиссом является преобразование Хеллингера (Legendre, Legendre, 2012).

Данное преобразование ослабляет влияние доминирующих видов, снижает вес редких и случайных встреч, уменьшает проблему «двойных нулей» и приближает распределение к форме, более пригодной для последующего анализа, особенно при использовании методов, основанных на евклидовых расстояниях.

Важно подчеркнуть, что преобразование Хеллингера следует применять к матрице обилий перед проведением как непрямого, так и прямого линейного ординационного анализа, например, перед tb-PCA и tb-RDA (tb — transformation based), поскольку именно оно готовит биотические данные к корректному моделированию в евклидовом пространстве. В отличие от линейных методов, основанных на евклидовой метрике, такие методы как классический CCA, изначально опираются на χ^2 -дистанцию и трактуют исходные обилия иначе; необходимость и характер трансформаций для них обсуждаются отдельно (см. специализированные руководства). В программе PAST, к сожалению, отсутствуют встроенные опции для описанных видов стандартизации, однако эти процедуры легко реализуются в Excel на основании следующих формул.

Формула для преобразования Хеллингера имеет следующий вид:

$$y_{ij}^* = \sqrt{\frac{y_{ij}}{\sum_{k=1}^S y_{ik}}}$$

где y_{ij}^* — стандартизированное значение вида j в пробе i , y_{ij} — исходная численность/биомасса вида j в пробе i , $\sum_{k=1}^m y_{ik}$ — общая сумма всех видов в пробе i (сумма по строке), S — количество видов. Допустим, в наличии 4 столбца данных в Excel, и нужно преобразовать, например, параметр в 1-й строке 1-го столбца. Тогда, 1) вычисляется сумма четырёх параметров из 1-й строки, 2) параметр в 1-й строке 1-го столбца делится на эту сумму и 3) результат деления возводится в степень 0.5, что соответствует извлечению квадратного корня. Подобная процедура продлевается для каждого параметра из матрицы данных.

Z-преобразование производится по следующей формуле:

$$z = (x - \mu) / \sigma,$$

где z — стандартизованная величина параметра, x — исходное значение признака, μ — среднее арифметическое признака по столбцу, σ — стандартное отклонение признака по столбцу.

3. АНАЛИЗ И ИНТЕПРЕТАЦИЯ ОРДИНАЦИОННЫХ ДИАГРАММ

Корректная интерпретация результатов многомерного анализа невозможна без понимания визуального языка ординационных диаграмм. Диаграммы рассеяния — биплоты (*biplot*) и триплоты (*tripplot*) — представляют собой проекции многомерных данных на плоскость первых двух ординационных осей, где каждый графический элемент несёт строгий аналитический смысл. Ключевыми элементами для чтения таких графиков являются взаимное расположение точек объектов (проб), отражающее сходство сообществ, и направление векторов переменных, показывающее градиенты среды или распределение видов. Умение «читать» эти элементы позволяет перейти от абстрактной геометрии графика к содержательным экологическим выводам о структуре сообществ, факторах их сдвига и взаимоотношениях между видами и средой. Данный раздел объясняет, как интерпретировать расстояния и углы для формулирования научных гипотез.

Scatter Plot. Ординация объектов/проб

1. Различия в размерах облаков точек разных групп

Размер облака (эллипса) отражает внутригрупповую вариабельность.

а) *Большое облако*: высокая неоднородность условий внутри группы (например, объединены выборки из различных биотопов) или высокая изменчивость видового состава (Рис. 3.1, D).

б) *Маленькое, компактное облако*: группа однородна; сообщества в этих пробах стабильны и похожи друг на друга (Рис. 3.1, C).

2. Нахождение облаков точек в одном квадранте или разных квадрантах

Отражает качественное сходство или различие сообществ, если первые оси действительно описывают главные градиенты в данных.

а) *В одном квадранте*: сообщества в целом схожи по видовому составу и структуре; их различия носят преимущественно количественный характер (Рис. 3.1, A–B).

б) *В разных квадрантах*: сообщества качественно различны, в них присутствуют разные наборы видов индикаторов или виды с противоположными экологическими требованиями (Рис. 3.1, A–B vs. C, D).

3. Расстояние между облаками; перекрывание облаков

Отражает степень сходства между группами.

а) *Близко расположенные или перекрывающиеся облака*: группы слабо различимы; сообщества плавно переходят друг в друга (экологический континуум) — высокое сходство (Рис. 3.1, А–В).

б) *Далеко расположенные, неперекрывающиеся облака*: чёткие, дискретные различия между группами; сообщества формируются в принципиально разных условиях — возможная экологическая граница (Рис. 3.1, А–В — D).

4. Наложение облака точек на координатную ось

Если облако вытянуто вдоль одной из осей, изменчивость внутри этой группы определяется преимущественно градиентом, коррелирующим с данной осью (Рис. 3.1, D). Влияние фактора, соответствующего перпендикулярной оси, для этой группы вторично или относительно однородно.

5. Степень удаленности облака точек от координатных осей

Отражает степень отклонения сообществ от «средней» пробы в пространстве показанных осей.

а) *Облако далеко от центра*: факторы, определяющие оси, сильно связаны с формированием этих сообществ; пробы имеют выраженные характеристики по показанным градиентам (Рис. 3.1, C).

б) *Облако близко к центру*: данные сообщества близки к среднему состоянию по учтённым факторам; их специфика, как правило, определяется другими, не попавшими на график переменными или последующими осями (Рис. 3.1, А–В).

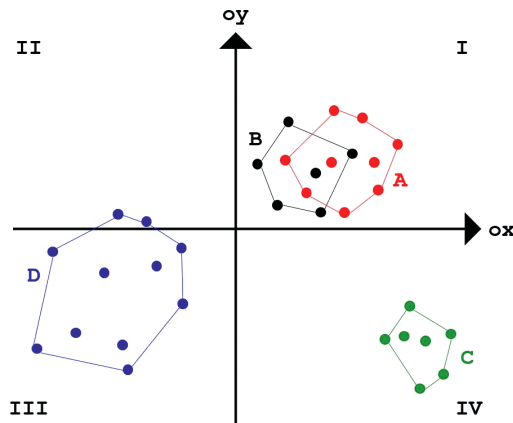


Рис. 3.1. Диаграмма рассеяния *Scatter plot*; I–IV — номера квадрантов

Biplot. Ординация объектов и переменных**1. Облако точек и вектор фактора среды в одном квадранте**

Прямая (положительная) связь. Пробы/сообщества, расположенные в этом квадранте, характеризуются высокими значениями данного абиотического параметра; чем дальше облако вдоль направления вектора, тем выше значение параметра (Рис. 3.2, облако А и факторы 1–2).

2. Облако точек и вектор фактора среды в смежных квадрантах

Слабая или опосредованная связь. Параметр имеет некоторое влияние на расположение этих проб, но не является определяющим; возможно, его эффект проявляется только во взаимодействии с другими факторами (Рис. 3.2, облако А и фактор 3).

3. Облако точек и вектор фактора в диаметрально противоположных квадрантах

Обратная (отрицательная) связь. Пробы/сообщества характеризуются низкими значениями данного параметра (Рис. 3.2, облако А и фактор 4).

4. Нахождение векторов фактора:

а) В одном квадранте и под острым углом: параметры положительно скоррелированы; высокие значения одного сопровождаются высокими значениями другого (Рис. 3.2, факторы 1 и 2).

б) В противоположных квадрантах (угол близок к 180°): параметры отрицательно скоррелированы (Рис. 3.2, факторы 2 и 4).

в) Под углом $\sim 90^\circ$: параметры практически не скоррелированы; их влияние на сообщества существенно независимо (Рис. 3.2, факторы 1 и 3).

5. Различная длина векторов

Длина вектора пропорциональна силе связи переменной с ординационными осями (её вкладу в объяснённую дисперсию в плоскости графика). Длинный вектор означает параметр, хорошо отражённый на диаграмме и сильно коррелирующий с осями (Рис. 3.2, параметр 1); короткий вектор соответствует параметру, слабо связанному с показанными осями или вносящему небольшой вклад в вариацию данных (Рис. 3.2, факторы 2 или 4).

6. Величина проекции векторов

Поскольку первая ось обычно объясняет большую долю вариации, чем вторая, проекция фактора на ось Х (первую) имеет больший статистический вес, чем равная по длине проекция на ось Y (вторую). Поэтому одинаковое по модулю смещение вдоль первой оси отражает

более существенные различия между объектами, чем такое же смещение вдоль второй (Рис. 3.2, проекция фактора 1 на ось ox , Nox vs. проекция фактора 3 на ось oy , Noy).

7. Углы между векторами (уточнение корреляции):

а) *Острый угол* ($<90^\circ$): положительная корреляция; чем острее угол, тем связь сильнее (Рис. 3.2, факторы 1 и 2).

б) *Тупой угол* ($>90^\circ$): отрицательная корреляция; рост одного фактора сопровождается снижением другого (Рис. 3.2, факторы 1 и 5).

в) *Угол $\sim 90^\circ$* : Параметры независимы (не скоррелированы) (Рис. 3.2, факторы 1 и 3).

8. Угол вектора с осями:

а) *Острый угол с осью абсцисс* (ox): параметр сильно коррелирован с первой (главной) осью ординации (Рис. 3.2, фактор 1).

б) *Острый угол с осью ординат* (oy): параметр сильно коррелирован со второй осью (Рис. 3.2, фактор 3).

в) *Угол $\sim 45^\circ$ к обеим осям* (сопоставимые проекции на X и Y): параметр в примерно равной степени связан с первой и второй осью; его вариабельность объясняется их комбинацией, а не только скрытой третьей компонентой. Если при этом вектор очень короткий, это указывает, что основная изменчивость параметра лежит вне плоскости первых двух осей и связана с последующими компонентами (Рис. 3.2, фактор 4).

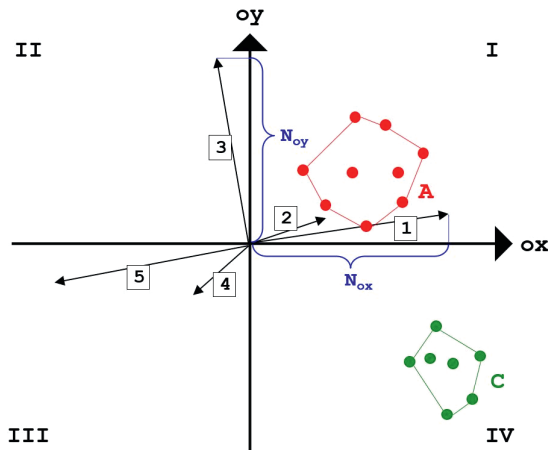


Рис. 3.2. Биплот *Biplot*, 1–5 — номера векторов факторов среды, I–IV — номера квадрантов

Triplot. Ординация объектов, переменных и видов

Интерпретация строится на взаимных проекциях: вектор параметра задаёт градиент среды, а точка вида показывает его оптимум (предпочитаемые условия). Положение вида на графике моделирует его экологическую нишу и показывает, в каких условиях среды этот вид достигает наибольшей относительной обильности по сравнению с другими исследованными условиями.

1. Вектор параметра, облако точек и точки видов в одном квадранте

Прямые экологические предпочтения. Сообщества находятся в условиях высоких значений параметра; вид является индикатором этих условий и достигает здесь высокой обильности (Рис. 3.3, квадрант I).

2. Вектор параметра в одном квадранте, а облако точек и точки видов в другом

Избегание условий. Сообщества и виды связаны с низкими значениями данного параметра; виды выступают индикаторами противоположных условий (Рис. 3.3, облако А и виды 1, 2 в I квадранте — факторы 4, 5 в III квадранте).

3. Вектор параметра и облако точек в одном квадранте, а точки видов в другом

Сложная взаимосвязь. Параметр важен для формирования сообщества в целом (пробы расположены вдоль его градиента), но конкретно этот вид не доминирует в данных условиях; возможно, его лимитируют другие факторы (Рис. 3.3, облако А и факторы 1, 2 в I квадранте — вид 4 в III квадранте).

4. Вектор параметра и точки видов в одном квадранте, а облако точек в другом

Локальное влияние. Параметр важен для данного вида (определяет его обилие), но не является главным фактором, структурирующим всё сообщество в этих пробах; структура сообщества в целом определяется другими градиентами (Рис. 3.3, виды 1, 2 и факторы 1, 2 в I квадранте — облако С в IV квадранте).

5. Степень удаления друг от друга объектов

Определяет силу связи.

а) *Вектор и точка вида близко, под острым углом:* вид сильно и положительно связан с этим параметром (Рис. 3.3, вид 1 — параметр 1; вид 4 — параметр 5).

б) *Точка вида и облако проб близко*: вид имеет высокую обильность в данном сообществе (Рис. 3.3, вид 2 — облако А).

в) *Точка вида внутри облака проб*: вид индикатор (характерный представитель) для данного типа сообществ (Рис. 3.3, вид 1 — облако А).

г) *Большие расстояния между всеми объектами* подчёркивают слабые экологические связи между сообществами и специфичность условий среды в каждой пробе; это, как правило, сопровождается существенными различиями в видовом составе.

6. Степень удаления точки данного вида от данной координатной оси

Отражает экологическую валентность вида.

а) *Далеко от центра*: вид со специализированными, чёткими требованиями к факторам, представленным осями (стенобионт); яркий индикатор конкретных условий (Рис. 3.3, вид 5).

б) *Близко к центру*: вид генералист (эврибионт); его обилие слабо зависит от показанных градиентов, он встречается в широком диапазоне условий (Рис. 3.3, вид 3).

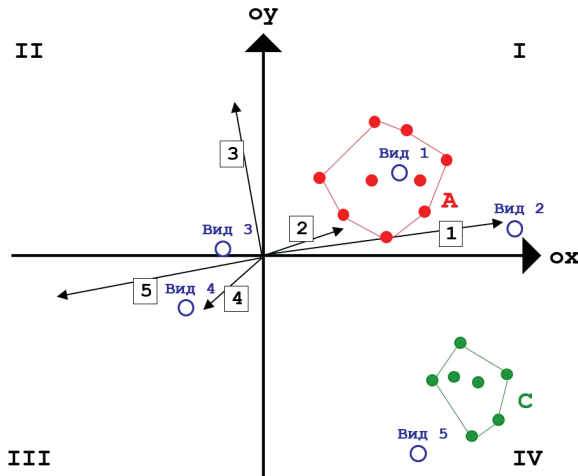


Рис. 3.3. Триплот Triplot, 1–5 — номера векторов факторов среды, I–IV — номера квадрантов.

Дополнение:

а) *Форма облака*: вытянутое вдоль оси/вектора облако указывает, что градиент этого фактора — главный источник изменчивости внутри группы; округлое облако может свидетельствовать о влиянии нескольких сопоставимых факторов или случайной вариабельности.

б) *Выбросы (outliers)*: отдельные точки, далеко отстоящие от своего облака или от всех остальных, могут быть как ошибками ввода/измерения, так и уникальными, редкими сообществами, представляющими особый интерес для анализа.

в) *Важность 3D-ординации*: если векторы переменных или точки видов на 2D графике короткие или расположены примерно под 90° к обеим осям, их интерпретация ненадёжна; необходимо учитывать долю дисперсии, объясняемую третьей осью, и по возможности анализировать трёхмерные проекции.

г) *Статистические тесты*: визуальную интерпретацию ординационных диаграмм следует подкреплять тестами значимости различий между группами (например, PERMANOVA, ANOSIM) и анализа вклада видов (SIMPER или аналогами).

Типичные ошибки интерпретации ординационных диаграмм

1. Интерпретация рисунка без учёта доли объяснённой дисперсии. Интерпретация рисунка опирается только на геометрию, без оценки того, какую долю вариации объясняют первая и вторая оси. Если эти оси описывают малую часть дисперсии, двухмерный график плохо отражает реальные отношения между пробами и переменными, и далеко идущие выводы по нему некорректны.

2. Отношение к биплоту как к обычной диаграмме рассеяния. Отношение к биплоту как к обычному двухмерному графику не учитывает, что точки и стрелки отображают разные объекты анализа (пробы и переменные). В результате углы, расстояния и длины векторов трактуются так, как если бы это был простой график «X–Y», без связи с методом и типом скейлинга.

3. Интерпретация углов и длины стрелок при неподходящем скейлинге. Интерпретация углов между векторами как корреляций и длины векторов как «силы влияния» проводится без проверки, какой вариант биплота построен (*distance biplot*, *correlation biplot* и др.). При ряде настроек расстояния между точками отражают диссимиляции, а углы между стрелками уже не являются прямой мерой корреляции между переменными.

4. Выводы о третьей и последующих осях по 2D-проекции.

Интерпретация коротких или «неудобно» расположенных векторов на 2D-графике как признака связи переменной исключительно с третьей осью не опирается на собственные значения и долю дисперсии. Короткий вектор может отражать как связь с последующими осями, так и низкую вариабельность переменной; это требует проверки по численным результатам ординации.

5. Подмена связи причинностью. Интерпретация близости точек и стрелок как доказательства причинно-следственных отношений выходит за пределы возможностей ординации. Диаграммы показывают паттерны ассоциаций; выводы о причинности требуют дополнительного анализа (эксперименты, моделирование, специальные статистические тесты, проверка альтернативных гипотез).

6. Игнорирование масштаба осей и искажение диаграммы.

Интерпретация углов и расстояний на графике, где шкалы по осям изменены (растянуты или сжаты для верстки), нарушает геометрический смысл диаграммы. Корректная интерпретация расстояний и углов возможна только при одинаковом масштабе осей.

7. Перегруженные диаграммы и «подгон» интерпретации.

Отображение слишком большого числа переменных и видов на одном графике приводит к «ёжикам», где легко «считать» желаемые связи из визуального шума. Для корректной интерпретации требуется отбор переменных по значимости и, при необходимости, построение отдельных диаграмм для разных наборов факторов.

8. Использование неподходящего метода ординации. Интерпретация диаграмм, полученных линейными методами (PCA, RDA) для данных с ярко выраженной unimodal структурой, или, наоборот, нелинейными методами (CA, CCA) при фактически линейных отношениях, приводит к искажённой геометрии. Выбор метода должен опираться на форму отклика и длину градиента (например, по DCA), иначе смысл расстояний и углов на диаграмме становится ненадёжным.

9. Отсутствие проверки визуальных выводов статистическими тестами. Интерпретация различий между облаками и связей с градиентами только «на глаз» без проверки PERMANOVA, ANOSIM, SIMPER или аналогичными процедурами повышает риск ошибочной интерпретации случайных эффектов. Визуальный анализ должен рассматриваться как предварительный шаг, а не как окончательное доказательство.

4. ПАНЕЛЬ НАСТРОЕК ВИЗУАЛИЗАЦИИ ОРДИНАЦИОННЫХ ДИАГРАММ

Интерпретация ординационных диаграмм в программе PAST является ключевым этапом анализа, позволяющим извлекать содержательные выводы из многомерных данных. Для облегчения этого процесса программа предоставляет набор инструментов визуализации, расположенных на панели настроек графика. Каждый элемент панели решает специфическую задачу: от выделения групповой структуры до отображения взаимосвязей между объектами, видами и факторами среды. Систематизация этих функций позволяет исследователю настраивать вывод результатов в соответствии со стандартами научных публикаций и избегать визуальных искажений при анализе многомерных экологических данных. Ниже приведены ключевые опции визуализации, составляющие неотъемлемую часть рабочего процесса при проведении ординационного анализа в PAST.

95% ellipses (95% эллипсы). Отображают доверительную область вокруг центроида каждой группы с заданным уровнем значимости (обычно 0,05). Эллипс показывает зону, в которой с заданной вероятностью располагается среднее положение группы в ординационном пространстве, что позволяет визуально оценить статистическую однородность выборок и степень перекрывания групп.

Convex hulls (Выпуклые оболочки). Минимальный выпуклый многоугольник, охватывающий все точки данной группы. В отличие от эллипсов, выпуклая оболочка демонстрирует фактические границы разброса объектов группы в ординационном пространстве, не опираясь на вероятностные допущения о форме распределения.

Min. spanning tree (Минимальное остовное дерево). Граф, соединяющий все объекты (пробы) линиями минимальной суммарной длины без циклов, построенный на основе выбранной матрицы расстояний. Визуализирует топологию сходства: показывает, какие пробы наиболее близки друг к другу и как формируется общая структура связей в выборке.

Row labels (Метки объектов). Выводят на график идентификаторы отдельных проб (номера, названия станций и т. п.) непосредственно рядом с соответствующими точками, обеспечивая точную привязку конкретных наблюдений к их положению в ординационном пространстве.

Group labels (Метки групп). Отображают названия категорий (например, «озеро», «река») обычно вблизи центроида соответствующего скопления точек. Упрощают чтение графика при большом числе перекрывающихся объектов и позволяют быстро идентифицировать принадлежность проб к заданным группам.

Eigenvalue scale (масштабирование по собственным значениям). Пропорционально масштабирует длину осей ординации в соответствии с их собственными числами (долей объяснённой дисперсии). Предотвращает визуальное искажение градиентов, возникающее при принудительном выравнивании масштаба осей, и позволяет корректно оценивать относительную важность первой и второй компонент.

Biplot (Биplot). Режим отображения, совмещающий на одном графике точки объектов и векторы переменных (факторов среды или, в специальных случаях, видов). Векторы указывают направление и относительную силу корреляции переменных с ординационными осями; угол между векторами отражает степень их взаимной корреляции, а проекция объекта на вектор характеризует значение переменной в данной пробе (при корректном типе биplotа и скейлинга).

Triplot (Триplot). Расширенный режим визуализации, одновременно отображающий три слоя данных: объекты (пробы), векторы переменных среды и точки видов, координаты которых рассчитываются как взвешенные средние. Применяется преимущественно в методах анализа соответствий (CA, CCA) для комплексной оценки связей «среда — сообщество — вид».

Column scores (WA) (оценки столбцов / взвешенные средние видов). Отображают позиции видов на диаграмме, вычисленные как взвешенные средние (Weighted Averages, WA) координат проб, в которых данные виды присутствуют. Позволяют интерпретировать экологические оптимумы видов: чем ближе точка вида к определённым пробам или векторам среды, тем более выражены его предпочтения к соответствующим условиям.

Similarity index (индекс сходства / мера расстояния). Выпадающее меню для выбора математической метрики, используемой при расчёте попарных различий между объектами. По умолчанию в PAST часто применяется евклидово расстояние, однако для матриц видового обилия оно редко подходит из-за чувствительности к нулям и абсолютному масштабу. Для экологических данных рекомендуется выбирать меры, устойчивые к двойным нулям и учитывающие относительное обилие (например, расстояние Брея–Кёртиса, хордовое расстояние, χ^2 -дистанция), в зависимости от типа данных и используемого метода ординации.

5. МЕРЫ ДИСТАНЦИИ

Выбор меры расстояния (метрики) является ключевым этапом ординационного анализа, поскольку именно она определяет, как вычисляется сходство или различие между объектами в многомерном пространстве признаков. Иначе говоря, метрика задаёт способ «сравнения» объектов и тем самым влияет на геометрию ординации и возможную интерпретацию результатов. Универсальной меры не существует: разные метрики предназначены для различных типов данных — непрерывных количественных признаков, матриц обилия видов, бинарных данных «присутствие–отсутствие» или сильно коррелированных переменных. Неправильный выбор метрики может привести к искажению структуры данных, маскировке реальных градиентов или появлению ложных закономерностей, обусловленных математическими свойствами самой меры расстояния. Поэтому выбор должен основываться на типе данных, исследовательской задаче и известных свойствах конкретной метрики.

Евклидова дистанция (*Euclidean*). Классическая мера «прямолинейного» расстояния между точками в многомерном пространстве. Подходит для непрерывных количественных данных (например, морфометрические или физиологические показатели, физико-химические параметры среды), измеренных в сопоставимых единицах. Чувствительна к масштабу переменных, поэтому требует предварительной стандартизации. Не рекомендуется для матриц видового обилия с большим числом нулей, поскольку совместное отсутствие вида в двух пробах трактуется как признак сходства (проблема «двойных нулей»), что в экологии сообществ часто ведёт к искажению структуры.

Дистанция Брея–Кертиса (*Bray–Curtis*). Одна из наиболее распространённых мер для анализа сообществ по данным об обилии видов (численность, биомасса, проективное покрытие). Оценивает различия, основываясь на суммах обилий в двух пробах и суммах попарных минимумов, игнорируя совместное отсутствие видов. Значения лежат в диапазоне от 0 (полное сходство) до 1 (полное различие). Является стандартом в экологии для выявления пространственных и временных градиентов, сукцессионных сдвигов и изменений доминирования, так как устойчива к обилию нулей и хорошо отражает различия в структуре сообществ.

Дистанция Жаккара (*Jaccard*). Мера сходства/различия, основанная исключительно на данных присутствия–отсутствия видов. Вычисляется как отношение числа видов, общих для двух проб, к общему числу видов в обеих пробах; обилие при этом не учитывается. Полезна при работе с видовыми списками, когда данные по численности отсутствуют или ненадёжны (например, исторические или музейные данные, палеоэкологические реконструкции). Чувствительна к редким видам, которые могут существенно увеличивать оценку различий; часто используется близкий по смыслу индекс Сёренсена, придающий больший вес общим видам.

Дистанция Махаланобиса (*Mahalanobis*). Метрика, учитывающая не только разброс признаков, но и ковариацию (взаимосвязь) между ними, что делает её инвариантной к масштабу и корреляциям переменных. Применяется преимущественно для сравнения объектов или групп по набору взаимосвязанных количественных признаков (например, морфометрические характеристики популяций, многомерные физиологические профили), а также в моделях экологических ниш. Требуется достаточного объёма данных для устойчивой оценки ковариационной матрицы и предполагает близость распределений к многомерному нормальному; в ординации сообществ используется редко, чаще — в дискриминантном анализе и распознавании образов.

Дистанция Хорна (*Horn*). Мера сходства сообществ, учитывающая как присутствие видов, так и их обилие, и основанная на соотношении обилий видов в сравниваемых пробах. Менее чувствительна к различиям в общем числе особей и видовом богатстве, чем расстояние Брея–Кёртиса, и более устойчива к различиям в объёме выборки. Рекомендуются для сравнения разнородных сообществ, где общая численность и усилие учёта сильно варьируют между пробами, поскольку даёт более сбалансированную оценку сходства.

Таблица 5.1. Выбор меры расстояния

Тип данных/задача	Рекомендуемая мера	Краткое обоснование
Непрерывные количественные признаки (морфометрия, физиология), переменные в сопоставимых единицах	<i>Euclidean</i> (после стандартизации)	Простая геометрическая интерпретация, работает в линейных методах (PCA, RDA).

Матрицы обилия видов (численность, биомасса, покрытие), много нулей	<i>Bray–Curtis</i>	Игнорирует двойные нули, хорошо отражает различия в составе и доминировании.
Матрицы обилия при сильных различиях в общем числе особей/усилии учёта	<i>Horn / Morisita–Horn</i>	Более устойчива к различиям в видовом богатстве и объёме выборки.
Видовых списков (присутствие–отсутствие) численности нет или ненадёжны	<i>Jaccard, Sørensen</i>	Учитывают только совместное присутствие видов, подходят для бинарных данных.
Сравнение объектов по набору коррелированных количественных признаков (морфометрия, физиология групп)	<i>Mahalanobis</i>	Учитывает ковариацию признаков, инвариантна к масштабу; требует больших выборок и близкой к нормальной структуры данных.

6. КЛЮЧ ДЛЯ ВЫБОРА МЕТОДА ОРДИНАЦИИ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Корректный выбор метода ординации является фундаментальным этапом эколого-статистического исследования, от которого напрямую зависят достоверность и интерпретируемость получаемых результатов. Многомерный характер биологических данных, наличие сложных нелинейных взаимосвязей и разнообразие статистических подходов создают закономерные методологические трудности даже для опытных исследователей. Такой системный выбор минимизирует риск методологических ошибок и обеспечивает обоснованность применения конкретного метода. Последовательная схема принятия решений, опирающаяся на ключевые критерии (тип матриц, характер градиентов и цель анализа), позволяет исследователю осознанно выбирать оптимальный метод. Такой системный подход не только экономит время исследователя, но и способствует более глубокому пониманию внутренней логики многомерного анализа, превращая фор-

мальное применение статистических тестов в содержательное научное моделирование. Использование данного ключа обеспечивает воспроизводимость результатов и укрепляет методологическую строгость экологических работ, что особенно важно в контексте современных требований к научным публикациям. Таким образом, этот алгоритм служит не только практическим руководством, но и важным элементом формирования методологической культуры в исследовательском сообществе.

Подготовительный этап — стандартизация данных: а) биотические параметры (обилие видов): рекомендуется преобразование Хеллингера (альтернативно — преобразование через квадратный корень или логарифм); б) абиотические параметры (факторы среды): рекомендуется Z-преобразование (вычитание среднего и деление на стандартное отклонение) для устранения влияния единиц измерения и выравнивания вклада переменных.

Раздел А. Методы прямого градиентного анализа (*Constrained ordination*)

Шаг А1. Наличие матрицы факторов среды. Имеется ли отдельная матрица данных по факторам внешней среды, которые должны выступать объясняющими переменными?

- Да → Шаг А2.
- Нет → перейти к Разделу В (Шаг В1, непрямой анализ).

Шаг А2. Тип предикторов (факторов среды). Объясняющие переменные преимущественно:

- Количественные (непрерывные) → Шаг А3.
- Категориальные или смешанные (значимая доля факторных переменных) → Шаг А7.

Шаг А3. Тип основной матрицы (отклика). Основная матрица — это:

- Данные по обилию видов (пробы × виды) → Шаг А4.
- Прочие количественные признаки (морфометрия, физиология, абиотические параметры) → Шаг А5.

Шаг А4. Длина градиента (для матрицы видов). Оценка длины первой оси DCA (*Detrended Correspondence Analysis*) в SD единицах:

- < 3.0 SD → градиенты короткие, отклик видов близок к линейному → Шаг А6 (линейные модели).
- ≥ 4.0 SD → градиенты длинные, отклик видов унимодальный → ССА (*Canonical Correspondence Analysis*) предпочтительна.

- 3.0–4.0 SD → переходная зона: рекомендуется выполнить **и RDA, и CCA**, затем сравнить долю объяснённой дисперсии и интерпретируемость диаграмм.

Шаг А5. Количественные признаки (не виды). Основная матрица состоит из непрерывных количественных признаков в сопоставимых единицах (или после стандартизации)?

- Да → **RDA** (*Redundancy Analysis*): линейный метод, основанный на евклидовом расстоянии, предполагает примерно линейные связи между откликом и предикторами.

- Нет → данные сильно разнородны по шкалам/типам → рассмотреть предварительное преобразование/кодирование; при невозможности — использовать непрямую ординацию (Раздел В) и интерпретировать результаты *post-hoc* по факторам среды.

Шаг А6. Линейные модели для коротких и умеренных градиентов (видовые данные). Цель анализа:

- Нужен отбор и ранжирование факторов среды (скрининг переменных).

- Использовать **RDA** с процедурой пошагового отбора (*forward selection*) и пермутационными тестами для оценки значимости и ранжирования предикторов.

- Нужен анализ комплексного влияния всех факторов без акцента на отбор.

- Использовать **RDA** как базовый метод прямой ординации, обязательно проверяя мультиколлинеарность предикторов (сильные корреляции) и при необходимости исключая или объединяя такие переменные.

Шаг А7. Категориальные и смешанные предикторы. При наличии категориальных или смешанных предикторов (типы местобитаний, классы нагрузок и т. п.):

1. Основная матрица — обилие видов:

- При коротких градиентах (< 3 SD) и возможности представить категориальные переменные в виде набора 0/1 признаков (отдельный столбец для каждого уровня) можно применять RDA с такими переменными.

- При длинных градиентах и ожидаемом унимодальном отклике видов предпочтительна **CCA**.

2. Основная матрица — другие количественные признаки:

- Закодировать категориальные переменные (*dumtmy*-переменные) и использовать **RDA**, либо воспользоваться непрямой ординацией и последующей интерпретацией с учётом категорий.

Раздел В. Методы непрямого градиентного анализа (*Unconstrained ordination*)

Шаг В1. Тип основной матрицы. Основная матрица:

- Обилие видов (пробы $\times$ виды) $\rightarrow$ Шаг В2.
- Количественные признаки (объекты $\times$ параметры; морфометрия, абиотика) $\rightarrow$ Шаг В6.

Шаг В2. Стратегия анализа для матрицы обилия видов. Предпочтительный подход:

- Анализ через матрицу расстояний (*distance based*) $\rightarrow$ Шаг В3.
- Анализ через градиенты в матрице видов (*gradient based*) $\rightarrow$ Шаг В5.

Шаг В3. Выбор меры расстояния (видовые данные). Тип данных:

- Учитывается **обилие видов** $\rightarrow$ стандартный выбор — *Bray–Curtis* (для устойчивости к двойным нулям). $\rightarrow$ Шаг В4.
- Используется информация **присутствие–отсутствие** $\rightarrow$ применяют индексы Жаккара и Серенсена $\rightarrow$ Шаг В4.

Шаг В4. Выбор метода на основе матрицы расстояний. (Применяется к любой матрице, полученной на шаге В3: *Bray–Curtis*, Жаккар, Сёренсен и др.). Что важнее сохранить в пространстве ординации?

- Абсолютные значения расстояний, линейную интерпретируемость осей и скорость расчёта $\rightarrow$ **PCoA** (*Principal Coordinates Analysis*) — анализ главных координат.
- Ранговый порядок расстояний, устойчивость к нелинейным градиентам, шуму и выбросам $\rightarrow$ **nMDS** (*non-metric multidimensional scaling*); качество решения оценивать по стрессу и стабильности конфигураций (критерии *stress*: <0.05 — отличное, <0.1 — хорошее, $0.1–0.2$ — приемлемое, >0.2 — интерпретация осей ненадёжна; рекомендуется несколько запусков с разными начальными конфигурациями).

Шаг В5. Анализ экологических градиентов в матрице видов. По результатам пробного DCA или CA:

- Есть длинные градиенты (длина первой оси DCA > 4 SD, выраженный «эффект подковы») $\rightarrow$ **DCA** (*Detrended Correspondence Analysis*) как метод для длинных *unimodal* градиентов.
- Градиенты короткие (< 3 SD), «арка» выражена слабо $\rightarrow$ **CA** (*Correspondence Analysis*) как базовый метод для видовых данных с умеренными градиентами.

При длине 3–4 SD рекомендуется сравнить CA и DCA и выбрать метод с более интерпретируемыми диаграммами.

Шаг В6. Матрица количественных признаков (не виды).

Матрица включает количественные переменные (морфометрия, физиология, физико химические параметры объектов). Переменные измерены в разных единицах → выполнить Z-стандартизацию. Далее:

- Основная цель — выделить несколько главных осей, объясняющих максимальную дисперсию, с понятной структурой нагрузок → **PCA** (*Principal Component Analysis*) — основной метод для линейных данных.

- Аналитический процесс строится вокруг матрицы расстояний (например, для PERMANOVA, кластеризации) и требуется ординация на её основе → **PCoA** с евклидовой метрикой: при евклидовой дистанции даёт решение, эквивалентное PCA, но в терминах расстояний.

Раздел С. Проверка статистической значимости и виды-индикаторы

Шаг D1. Проверка значимости группировок. Если заданы априорные группы проб (например, «загрязнённые» vs «контроль»):

- Для методов с ограничением (RDA, CCA) → **PERMANOVA** на соответствующей матрице расстояний для оценки вклада факторов среды.

- Для методов без ограничения (PCA, PCoA, nMDS, CA, DCA) → **PERMANOVA** или **ANOSIM** на матрице расстояний (чаще всего *Bray–Curtis* для видовых данных).

Для матриц видового обилия стандартным и надёжным выбором метрики является *Bray–Curtis*.

Шаг D2. Определение видов индикаторов. Для выявления видов, вносящих наибольший вклад в различия между группами:

- Используется **SIMPER** (*Similarity Percentages*) на основе выбранной меры расстояния (обычно *Bray–Curtis*).

- Виды с высоким средним вкладом в различия и низким отношением SD/Mean рассматриваются как стабильные индикаторы различий между группами.

7. АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ (PCA)

Введение

Разработка метода анализа главных компонент (PCA, *Principal Component Analysis*) относится к началу XX века. В 1901 году английский статистик Карл Пирсон опубликовал работу «О линиях и плоскостях, ближайших к системам точек в пространстве» (Pearson, 1901), где представил геометрическую интерпретацию метода как поиск осей и плоскостей, наилучшим образом аппроксимирующих облако исходных точек. В 1930-е годы американский статистик Гарольд Хотеллинг развил эти идеи в рамках статистической теории, ввёл термин «главные компоненты» и показал их связь с ковариационной матрицей данных (Hotelling, 1933), заложив современные основы PCA.

Метод PCA исходит из того, что многомерные данные часто содержат избыточность: многие признаки коррелируют между собой, и истинная структура может быть представлена в пространстве меньшей размерности. Задача PCA — найти новые ортогональные направления (главные компоненты), вдоль которых дисперсия данных максимальна. Первая главная компонента (PC1) объясняет наибольшую часть общей дисперсии, вторая (PC2) — наибольшую часть оставшейся при условии ортогональности первой, и так далее.

В результате анализа исследователь получает небольшой набор новых переменных (обычно 2–3 компоненты), которые суммарно объясняют значительную долю дисперсии исходных данных, что позволяет снизить размерность и визуализировать структуру в 2–3-мерном пространстве. Это способствует выявлению скрытых структур — кластеров объектов, градиентов изменчивости, выбросов. При этом PCA не «создаёт» структуру, а лишь выявляет её при условии, что она реально присутствует и измеренные признаки адекватно её отражают; поэтому метод требует тщательной предварительной обработки данных и критической интерпретации результатов с учётом его допущений и ограничений.

Методология

Во многих биологических задачах каждый организм описывается набором количественных признаков (длина, масса, метаболические показатели, экспрессия генов и т. п.), и каждый объект можно представить точкой в многомерном пространстве размерности, равной числу

признаков. При увеличении числа признаков визуализация и анализ таких данных затрудняются: корреляции между переменными создают избыточность, а шум маскирует базовые закономерности. Метод PCA предоставляет формальный аппарат для упрощения этой системы за счёт перехода к новой системе координат, основанной на собственной изменчивости данных и сохраняющей наиболее существенные их характеристики.

Математическая основа PCA опирается на понятия дисперсии и ковариации. Дисперсия количественно описывает разброс значений вдоль заданной оси, ковариация характеризует линейную взаимосвязь между двумя признаками. Алгоритм PCA начинается с центрирования данных (вычитания среднего) и, при необходимости, масштабирования, после чего вычисляется ковариационная (или корреляционная) матрица, содержащая дисперсии признаков на диагонали и их попарные ковариации вне диагонали. Эта матрица описывает форму и ориентацию многомерного облака точек.

Ключевой шаг метода — нахождение нового набора ортогональных осей, упорядоченных по убыванию объясняемой дисперсии. Первая главная компонента ориентирована в направлении максимальной дисперсии данных, вторая — перпендикулярна первой и объясняет максимум оставшейся дисперсии, последующие строятся аналогично. Направления этих осей являются собственными векторами ковариационной матрицы, а соответствующие собственные значения пропорциональны доле общей дисперсии, объясняемой каждой компонентой.

В результате линейного преобразования исходных переменных формируются новые переменные — значения главных компонент (*scores*) для каждого объекта. Они ортогональны (линейно независимы и некоррелированы) и упорядочены по убыванию дисперсии, что позволяет заменить десятки коррелированных признаков несколькими компонентами, сохраняющими основную структуру вариации. Это обеспечивает радикальное снижение размерности: анализ часто сводится к рассмотрению первых двух–трёх компонент, на плоскости которых становятся видимыми кластеры, градиенты или выбросы. Интерпретация нагрузок признаков на компоненты (*loadings*) позволяет связать математические результаты с биологическими процессами: высокие по модулю нагрузки указывают, какие признаки формируют тот или иной осевой градиент.

Характер данных

РСА является одним из ключевых инструментов снижения размерности и визуализации структуры данных в ихтиологии, однако его применимость и информативность напрямую зависят от свойств исходных данных и соблюдения основных допущений метода. Алгоритм поиска новых ортогональных осей наиболее эффективен, когда признаки заметно коррелированы: в этом случае их совместная изменчивость позволяет «упаковать» основную информацию в меньшее число компонент. Предполагается также приблизительная линейность взаимосвязей между переменными, а чувствительность метода к масштабу требует стандартизации при работе с признаками, измеренными в разных единицах (например, длина в мм и масса в граммах). Данные должны быть представлены непрерывными числовыми величинами или, по крайней мере, порядковыми шкалами, допускающими интерпретацию линейных комбинаций. Важно, чтобы выборка не содержала большого числа экстремальных выбросов, способных существенно исказить ориентацию главных компонент.

Метод РСА особенно продуктивен, когда анализ связан с многомерным описанием организмов и/или среды, где признаки естественным образом коррелированы. Характерный пример — морфометрический анализ формы тела рыб, где различные линейные размеры тела и плавников обычно тесно связаны. Первая компонента часто отражает общий размер особи (изометрический рост), а последующие — аллометрические изменения (пропорции туловища, относительная высота тела, длина хвостового стебля). Это позволяет наглядно выявлять различия между популяциями, возрастными группами, экологическими формами, а также обнаруживать гибридные или переходные варианты.

Другой классический пример — анализ стабильных изотопов в трофодинамике. Наборы значений $\delta^{13}\text{C}$, $\delta^{15}\text{N}$ (а иногда $\delta^{34}\text{S}$, $\delta^2\text{H}$) для разных видов рыб и их кормовых объектов образуют многомерное «изотопное пространство». РСА позволяет сжать эту информацию, выделив компоненты, связанные с основными источниками органического вещества (например, пелагические vs бентосные цепи питания, отражаемые $\delta^{13}\text{C}$) и трофическим уровнем ($\delta^{15}\text{N}$), и визуализировать структуру пищевой сети и степень перекрытия экологических ниш. Аналогично, в геометрической морфометрии координаты опорных точек (*landmarks*) образуют высокоразмерные и сильно коррелированные наборы данных; РСА позволяет выявить основные направления

вариации формы, связанные с гидродинамикой, способом питания или филогенетическими различиями. Наконец, анализ комплексов абиотических параметров среды (температура, глубина, скорость течения, мутность, кислород, ионы и др.) на станциях отбора также эффективно выполняется с помощью РСА: метод выделяет основные градиенты среды, которые затем сопоставляются с распределением видов или их функциональных групп.

В то же время существуют типы данных и исследовательские задачи, где применение РСА либо малоэффективно, либо может приводить к искажённым выводам. Прежде всего это категориальные данные: попытка применить РСА к номинальным признакам (видовые названия, тип питания, наличие/отсутствие паразитов) приводит к результатам, зависящим от способа числового кодирования уровней и не имеющим биологического смысла. Проблематичен также анализ данных с выраженными нелинейными зависимостями: поскольку РСА выявляет только линейные направления максимальной дисперсии, сложные криволинейные взаимосвязи оказываются искажёнными или распределёнными по нескольким компонентам.

Особые трудности возникают при анализе матриц видового состава с преобладанием редких видов. В таких таблицах содержится большое число нулевых значений (отсутствий), что нарушает предпосылку примерно линейных зависимостей между видами, усиливает влияние редких таксонов и маскирует закономерности в распределении массовых видов. Совместные нули могут искусственно сближать пробы в евклидовом пространстве, даже если их экологическое сходство невелико, а композиционный характер данных (работа с долями и процентами) приводит к дополнительным искажениям ковариационной структуры. В таких случаях предпочтительны специализированные методы ординации сообществ (CA, DCA, nMDS, PCoA с подходящей метрикой), а не прямое применение стандартного РСА.

Таким образом, успешное использование РСА требует осмысленного подхода к данным. Перед анализом необходимо ответить на ключевые вопросы: являются ли переменные непрерывными и коррелированными, допустимо ли предположение о линейности связей, отсутствуют ли сильные выбросы и избыток нулевых значений? Обязателен предварительный разведочный анализ: визуализация распределений, построение матриц корреляций, выявление выбросов и оценка необходимости стандартизации. Следует помнить, что РСА — это

инструмент поиска линейных комбинаций и направлений максимальной дисперсии, и его эффективность проявляется только при соответствии структуры данных этим принципам. В этом случае метод становится мощным средством выявления биологических закономерностей в многомерных данных.

Анализ данных

Анализ главных компонент (PCA) в PAST выполняется через главное меню *Multivariate* → *Ordination* → *PCA*. В открывшемся окне на вкладке *Summary* необходимо задать параметры анализа в выпадающих меню *Matrix* и *Groups*. В *Matrix* доступны два основных варианта: *Variance–Covariance* (ковариационная матрица), сохраняющая информацию об абсолютной изменчивости признаков и уместная, когда все переменные измерены в сопоставимых единицах, и *Correlation* (корреляционная матрица), автоматически выполняющая Z-стандартизацию и обязательная при работе с признаками в разных единицах или сильно различающейся дисперсии.

Ключевым является выбор режима анализа в меню *Groups*, определяющего учёт априорных групп: *Disregard* — классический PCA без учёта групп, когда все особи анализируются как единая совокупность и выявляется общая структура вариации; *Within-group* — PCA внутригрупповой изменчивости, при котором межгрупповые различия «вычитаются» и акцент смещается на вариацию внутри каждой группы; *Between-group* — PCA межгрупповых различий, где анализируются только средние значения признаков по группам, а внутригрупповая изменчивость игнорируется (табл. 7.1). Тем самым выбор режима задаёт фокус анализа: общая структура данных, внутрипопуляционная изменчивость или различия между группами (например, озе́рами).

В меню *Summary* содержатся, в частности, значения *Eigenvalue* и *% variance*, зависящие от выбранного режима *Disregard*, *Within-group* или *Between-group*. *Eigenvalue* (λ) главной компоненты количественно отражает величину дисперсии исходных данных, объясняемую данной компонентой, тогда как *% variance* показывает долю общей изменчивости данных (в процентах), приходящуюся на каждую главную компоненту; суммарный вклад первых двух–трёх компонент служит ориентиром пригодности низкоразмерного представления для интерпретации.

Таблица 7.1. Выбор режима анализа изменчивости в зависимости от структуры выборки

Режим	Особи из одного озера, групп нет	Особи из нескольких озёр, группы есть
<i>Disregard</i>	Задача. Анализ общей изменчивости популяции в озере	Не применим, т. к. смешивает внутри- и межгрупповую изменчивость, затрудняя интерпретацию
<i>Within-group</i>	Не применим, т. к. групп нет	Задача. Насколько разнообразны рыбы внутри каждого озера? Исключаем различие между озёрами
<i>Between-group</i>	Не применим, т. к. групп нет	Задача. Насколько сильно отличаются средние показатели между озёрами? Какие популяции/озера наиболее схожи?

Ключевым результатом РСА является ординационная диаграмма — *биplot* (*biplot*), совмещающий в одном пространстве проекции объектов (например, отдельных рыб) и векторы нагрузок исходных признаков. Выбор осей для отображения осуществляется на вкладке *Scatter plot* через меню *X axis* и *Y axis* (обычно PC1 и PC2). Биplot позволяет одновременно анализировать положение объектов и вклад переменных в формирование осей: первые компоненты (PC1 и PC2) обычно объясняют наибольшую долю дисперсии, поэтому интерпретация чаще всего ограничивается ими.

Анализ биplotа следует проводить совместно с таблицей *Loadings* (нагрузки). Нагрузки представляют собой коэффициенты линейных комбинаций исходных признаков, формирующие каждую главную компоненту, и показывают вклад и направление каждого признака в новую ось. При использовании корреляционной матрицы (стандартизированные данные) значения *loadings* численно равны коэффициентам корреляции Пирсона между исходными переменными и компонентами, что облегчает интерпретацию биологического смысла осей PC1 и PC2.

Практическая задача. Режим Disregard

Задача. Выполнены измерения следующих морфометрических параметров леща из одного озера: C — длина по Смитту, aD — антедорсальное расстояние, aA — расстояние между рылом и началом анального плавника, aV — антевентральное расстояние, ao — длина головы, H — максимальная высота тела, vz — расстояние между грудным и брюшным плавниками, hD — высота спинного плавника, ID — длина спинного плавника, IP — длина грудного плавника, IV — длина брюшного плавника, hA — длина анального плавника, IA — ширина основания анального плавника, pD — длина между проекцией конца спинного плавника и окончанием чешуйного покрова, kj — минимальная высота тела. Используя анализ главных компонент РСА, выявить основные закономерности изменчивости морфометрических признаков леща данного озера.

Перед загрузкой данных в PAST необходимо провести Z-стандартизацию переменных. Выбор этого метода предобработки обусловлен природой исходных данных: морфометрические параметры представляют собой непрерывные количественные показатели, измеренные в одинаковых единицах, но имеющие различный масштаб изменчивости. В отличие от преобразования Хеллингера, предназначенного для матриц видового обилия и долей, Z-стандартизация сохраняет метрические свойства исходных переменных, приводя их к нулевому среднему и единичной дисперсии. Это позволяет интерпретировать результаты РСА в терминах стандартных отклонений от среднего и предотвращает доминирование признаков с большими абсолютными значениями.

Метод главных компонент (РСА) является корректным инструментом для изучения внутрипопуляционной изменчивости, что особенно актуально при анализе единственной выборки, например, морфометрии рыб из одного водоёма. В этом случае фокус исследования смещается с поиска дискретных групп на выявление внутренней структуры вариабельности: РСА позволяет идентифицировать основные направления морфологического разнообразия — такие как общий размер тела (изометрия) или особенности пропорций (аллометрия) — и визуализировать непрерывный спектр изменчивости в виде облака точек в пространстве первых главных компонент. Таким образом, для отдельной популяции РСА применяется не для кластеризации, а для анализа структуры изменчивости, выявления выбросов и формирования гипотез о влиянии пола, возраста или условий среды.

При анализе единственной популяции в диалоговом окне PCA программы PAST целесообразно использовать режим *Disregard*, при котором вся совокупность особей рассматривается как единая выборка. Этот режим фокусируется на общей структуре ковариации признаков без искусственного разделения на подгруппы, что полностью соответствует задаче изучения внутренней морфометрической вариабельности внутри одной популяции.

РЕШЕНИЕ ЗАДАЧИ

Интерпретация *Loadings* и доли объясненной дисперсии

Первая главная компонента PC1 (объясняет 38,53% дисперсии; рис. 7.1, слева) демонстрирует положительные нагрузки по всем морфометрическим признакам (рис. 7.1, справа), что указывает на доминирование «компоненты общего размера» (*size component*). Наибольший вклад в PC1 вносят параметры *ao* (длина головы, 0,370), *kj* (минимальная высота тела, 0,369), *pD* (постдорсальное расстояние, 0,358) и *H* (максимальная высота тела, 0,326). Это свидетельствует о согласованном увеличении линейных размеров: особи с высокими значениями PC1 характеризуются крупными абсолютными размерами тела, тогда как особи с низкими значениями PC1 — меньшими.

Вторая главная компонента PC2 (объясняет 14,78% дисперсии; рис. 7.1, слева) отражает градиент формы тела (*shape component*), о чём свидетельствуют разнонаправленные нагрузки признаков. Положительные нагрузки характерны для *aD* (антедорсальное расстояние, 0,453) и *lP* (длина грудного плавника, 0,422), тогда как отрицательные — для *aA* (расстояние до анального плавника, -0,462) и *pD* (постдорсальное расстояние, -0,253). Такая противоположная направленность нагрузок указывает на аллометрические изменения пропорций: особи с высокими значениями PC2 характеризуются смещённым вперёд спинным плавником (удлинённая передняя часть тела), более развитыми грудными плавниками и относительно укороченной задней частью тела (меньшее расстояние *aA* и *pD*), тогда как особи с низкими значениями PC2 имеют относительно удлинённую заднюю часть тела и менее выраженный передний отдел. Эта вариабельность может отражать возрастные изменения (различия между молодыми и взрослыми особями) или индивидуальную пластичность формы тела в пределах популяции.

Суммарно первые две главные компоненты объясняют 53,31% общей дисперсии признаков, что позволяет считать двумерную проекцию (биplot) информативной для интерпретации основных закономерностей морфометрической изменчивости.

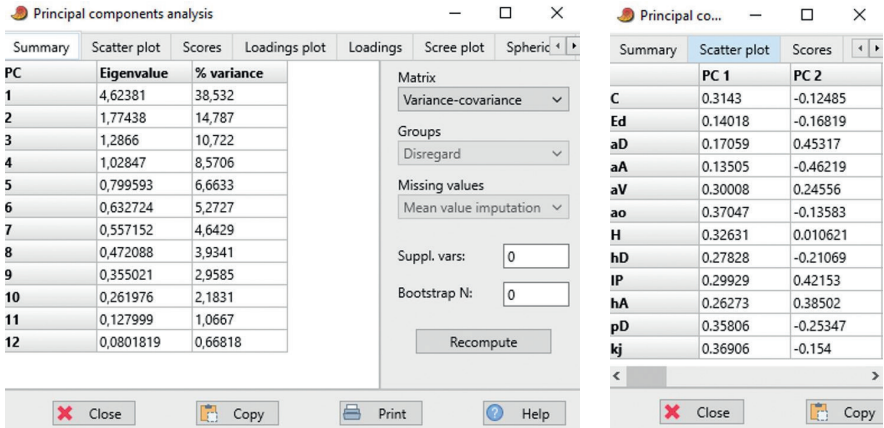


Рис. 7.1. Режим *Regard*. Значения *Eigenvalues* и процент объясненной дисперсии *%variance* (слева); значения нагрузок (*loadings*) для первых двух компонент PC1 и PC2 (справа)

График *loadings*

График нагрузок в плоскости PC1–PC2 визуализирует векторы признаков, где длина вектора отражает силу вклада переменной в компоненты, а направление — знак и относительное соотношение нагрузок (рис. 7.2). Векторы *ao*, *H*, *kj* и *aV* ориентированы преимущественно вдоль PC1, что подтверждает их определяющую роль в формировании компоненты общего размера: именно эти признаки в наибольшей степени задают градиент «крупные–мелкие» особи. Векторы *aD* и *IP* имеют выраженные положительные нагрузки по PC2, тогда как *aA* и *pD* — отрицательные (рис. 7.2, правый график). При сохранении положительных нагрузок всех признаков по PC1 это формирует тупые углы между векторами *aD*–*aA* (~140°) и близкие к прямым углы для пары *IP*–*pD* (~90°). Тупой угол *aD*–*aA* отражает устойчивый «морфологический компромисс»: увеличение длины передней части тела и смещение спинного плавника вперёд статистически сопряжены с относительным укорочением заднего отдела. Ортогональность пары *IP*–*pD* указывает на то, что изменения длины спинного

плавника и расстояния до хвостового стебля варьируют независимо друг от друга при фиксированном общем размере тела.

В данном анализе, выполненном на основе ковариационной матрицы (вкладка *Summary* → *Matrix* → *Variance-covariance*), в качестве нагрузок по умолчанию используются коэффициенты (вкладка *Loading plot* → радиокнопка *Coefficients*), а не корреляции (радиокнопка *Correlations*) (рис. 7.2). Такой подход методически оправдан, поскольку все морфометрические признаки измерены в сопоставимых единицах (мм), и коэффициенты сохраняют информацию об исходном масштабе данных и абсолютной величине дисперсии каждого признака; это важно, в частности, для интерпретации компоненты общего размера (PC1), отражающей согласованное изменение линейных размеров.

Использование опции *Correlations* нормирует нагрузки так, чтобы они отражали корреляции признаков с компонентами; при предварительной Z-стандартизации данные по сути становятся сопоставимыми, и общий размерный градиент сохраняется и в этом представлении. Однако в морфометрических задачах, где интересует не только направление, но и относительный вклад различных линейных размеров в общую дисперсию, работа с коэффициентами на ковариационной матрице делает интерпретацию size-компоненты более прозрачной. Поэтому для анализа абсолютных размеров тела и их совместной изменчивости в пределах популяции использование режима *Variance-covariance* в сочетании с *Coefficients* является предпочтительным.

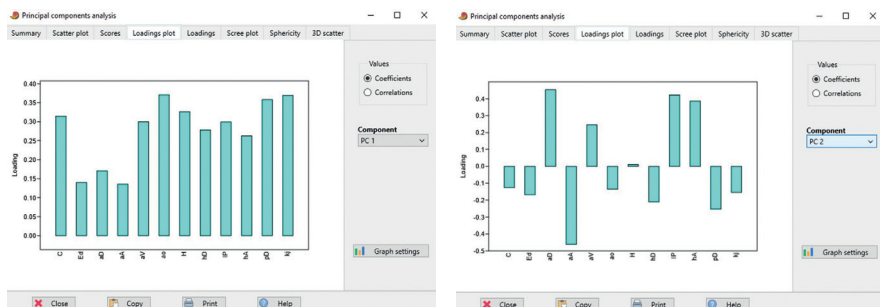


Рис. 7.2. Режим *Regard*. Графики *loadings* для первой (слева) и второй (справа) главных компонент

Визуализация (*Biplot*)

Биплот проекций особей в пространстве PC1–PC2 демонстрирует непрерывное облако точек без чётких кластеров, что характерно для единой популяции без явно обособленных групп по комплексу морфометрических признаков (рис. 7.3). Разброс объектов вдоль оси PC1 отражает вариацию абсолютного размера тела: особи с отрицательными значениями PC1 имеют меньшие линейные размеры, тогда как особи с положительными значениями — более крупные. Разброс вдоль PC2 выявляет полиморфизм формы: особи в положительной области PC2 отличаются относительно удлинённой передней частью тела и более развитыми грудными плавниками, тогда как особи в отрицательной области характеризуются относительно удлинённой задней частью тела. Отсутствие дискретных скоплений указывает на отсутствие выраженной внутренней субструктуры по анализируемым признакам, по крайней мере в пределах разрешения данного набора переменных.

Практическое чтение биплота основывается на проекции точек объектов на векторы нагрузок: особь, расположенная в направлении вектора, как правило, имеет значение соответствующего признака выше среднего; расположение в противоположную сторону указывает на пониженные значения, близость к началу координат — на значения, близкие к среднему. Наблюдаемый непрерывный градиент вдоль PC1, вероятнее всего, отражает возрастную структуру выборки, где меньшие особи соответствуют младшим возрастным группам, а крупные — более старшим. Вариация вдоль PC2 может интерпретироваться как проявление аллометрического роста или экологической пластичности формы тела. Крайние точки облака (потенциальные выбросы) могут указывать на аномальные особи или ошибки измерений и требуют отдельной проверки. Для подтверждения гипотез о возрастной динамике, аллометрии и роли среды необходимы независимые данные (оценка возраста, сведения об условиях обитания) и, при необходимости, специализированные методы анализа формы.

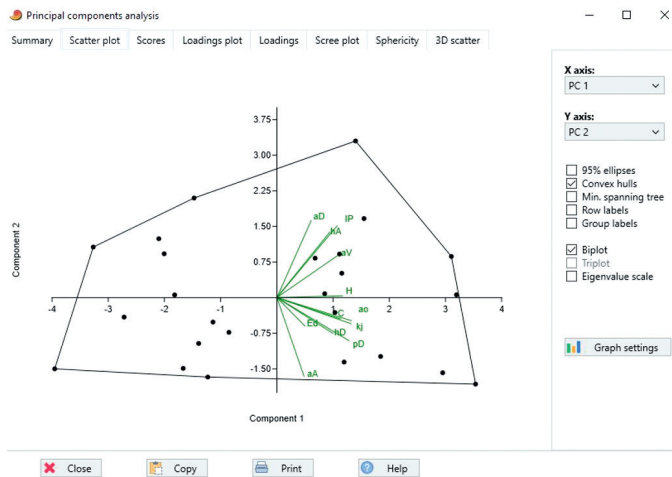


Рис. 7.3. Режим *Regard*. Ординационная диаграмма (биplot) морфометрической изменчивости леца из одной популяции

Sphericity

В программе PAST вкладка *Sphericity* предоставляет статистические тесты, позволяющие оценить обоснованность применения анализа главных компонент (PCA) к данным (рис. 7.4).

Тест сферичности Бартлетта (*Bartlett's test of sphericity*) проверяет нулевую гипотезу о том, что матрица корреляций является единичной, то есть все переменные статистически независимы и не коррелируют между собой. Альтернативная гипотеза предполагает наличие значимых корреляций между переменными, что делает применение PCA целесообразным. В данном анализе получены следующие результаты: определитель матрицы (*Determinant*) составляет $1,5466 \times 10^{-16}$; значение статистики $\chi^2 = 757,43$ при 77 степенях свободы; $p\text{-value} = 2,14 \times 10^{-112}$. Крайне малое значение определителя указывает на высокую степень коррелированности признаков (мультиколлинеарность), что для морфометрических данных ожидаемо, поскольку у морфометрических характеристик имеется сильная ковариация. Статистически значимый результат теста ($p \ll 0,05$) позволяет уверенно отвергнуть нулевую гипотезу и сделать вывод о наличии значимых корреляционных связей между переменными, что подтверждает применимость PCA для выявления латентной структуры данных.

Мера адекватности выборки Кайзера–Мейера–Олкина (*Kaiser–Meyer–Olkin*, КМО) оценивает пригодность данных для факторного анализа путём сравнения величин парных и частичных корреляций. Принята следующая шкала интерпретации: значения КМО $> 0,9$ считаются превосходными, $> 0,8$ — хорошими, $> 0,7$ — приемлемыми, $> 0,6$ — посредственными, тогда как значения $< 0,5$ указывают на непригодность данных для факторного анализа. В рассматриваемом примере расчёт КМО для ковариационной матрицы в PAST технически не реализован, что является ограничением программы. Однако, учитывая высокую значимость теста Бартлетта ($p < 0,001$) и биологически обоснованную коррелированность морфометрических признаков, можно заключить, что данные обладают достаточной структурной организацией для применения PCA. При этом интерпретация компонент с малыми собственными числами (PC3 и выше) требует особой осторожности, поскольку они с большей вероятностью отражают шум или локальные особенности выборки, а не устойчивые биологически значимые градиенты.

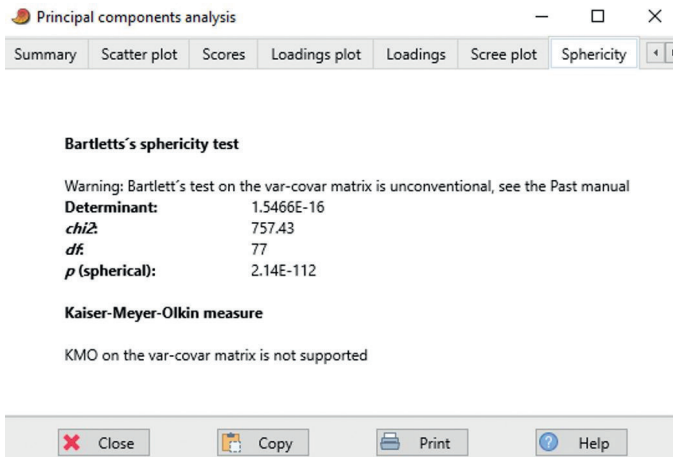


Рис. 7.4. Режим *Regard*. Сферичность

График «каменной осыпи» (*Scree plot*, вкладка *Scree plot*) визуализирует долю объяснённой дисперсии для каждой главной компоненты в порядке их убывания (рис. 7.5). По оси абсцисс откладываются номера компонент, по оси ординат — соответствующие им собственные значения (*Eigenvalues*), пропорциональные величине объяснённой

дисперсии. Основная цель графика — визуальное определение точки «перелома» («локтя»), после которой крутизна кривой заметно уменьшается. В данном случае выраженный излом («локоть») наблюдается после 1-й компоненты, что указывает на её доминирующую роль и целесообразность интерпретации как основного направления изменчивости данных. Компоненты, расположенные до этой точки, рассматриваются как информативные и подлежат интерпретации, тогда как компоненты после «локтя» в большинстве случаев отражают шум и, как правило, не интерпретируются самостоятельно.

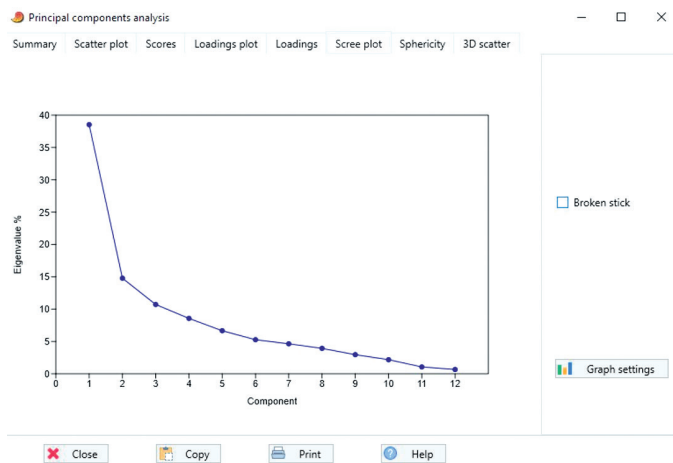


Рис. 7.5. График «каменной осыпи», *Scree plot*

Заключение

РСА выявил два основных направления изменчивости в популяции леща: компоненту общего размера тела (PC1) и компоненту пропорций, отражающих соотношение длины отделов тела и плавников (PC2). Совокупность этих осей описывает внутривидовое морфологическое разнообразие, вероятнее всего обусловленное онтогенетическими изменениями и, возможно, адаптивными различиями в форме тела.

Практическая задача. Режим *Within-group*

Режим *Within-group* РСА предназначен для анализа и визуализации внутренней структуры данных внутри предопределённых групп, что позволяет выявлять скрытые закономерности, оценивать популяционную изменчивость и обнаруживать потенциальные субпопуляции.

Задача. Выполнены измерения следующий морфометрических параметров гольца из трёх различных озёр: C — длина по Смитту, aD — антедорсальное расстояние, aA — расстояние между рылом и началом анального плавника, aV — антевентральное расстояние, ao — длина головы, H — максимальная высота тела, vz — расстояние между грудным и брюшным плавниками, hD — высота спинного плавника, ID — длина спинного плавника, IP — длина грудного плавника, IV — длина брюшного плавника, hA — длина анального плавника, IA — ширина основания анального плавника, pD — длина между проекцией конца спинного плавника и окончанием чешуйного покрова, kj — минимальная высота тела. Используя анализ главных компонент РСА, выявить основные закономерности изменчивости морфометрических признаков гольца из разных озёр.

РЕШЕНИЕ ЗАДАЧИ

Интерпретация *Loadings* и дисперсии

Проведённый анализ главных компонент в режиме *Within-group* выявил структуру внутрипопуляционной изменчивости морфометрических признаков гольца, устранив влияние различий между средними значениями групп (озёр). Первые две главные компоненты объясняют сравнительно небольшую, но типичную для морфометрических данных долю общей изменчивости: PC1 — 12,4%, PC2 — 12,2%. Низкая концентрация дисперсии в первых осях отражает сложную корреляционную структуру признаков, когда вариация «распределена» по множеству измерений.

Биологический смысл компонент раскрывается через анализ нагрузок. PC1 отражает контраст пропорций передней и задней частей тела. Положительный полюс нагружен признаками IV (длина брюшного плавника, 0,47), ao (длина головы, 0,37) и H (максимальная высота тела, 0,31), тогда как отрицательный формируют ID (длина спинного плавника, -0,41) и pD (постдорсальное расстояние, -0,30). Таким

образом, особи с высокими значениями PC1 характеризуются более крупной головой, более высоким телом и длинными брюшными плавниками при относительно укороченном спинном плавнике и задней части тела; отрицательные значения PC1 соответствуют противоположной конфигурации пропорций.

PC2 описывает вариацию плавников и основания анального плавника. Положительный полюс связан с lA (ширина основания анального плавника, 0,58), aV (антевентральное расстояние, 0,34) и lP (длина грудного плавника, 0,31), тогда как отрицательный определяют IV (длина брюшного плавника, -0,39) и kj (высота хвостового стебля, -0,37). Положительные значения PC2 описывают рыб с широким основанием анального плавника и длинными грудными плавниками при относительно укороченных брюшных плавниках и более тонком хвостовом стебле; отрицательные значения соответствуют рыбам с противоположным сочетанием этих признаков.

Визуализация *Biplot*

На биплоте (рис. 7.6), совмещающем положения особей и векторы признаков, выпуклые оболочки (*convex hulls*), ограничивающие выборки трёх озёр, заметно перекрываются. Это показывает, что внутривидовая изменчивость существенно превышает межвидовые различия и общая морфологическая структура популяций сходна. Тем не менее прослеживаются определённые популяционные тенденции: Озеро 1: точки смещены преимущественно в положительную область PC1; Озеро 2: особи группируются преимущественно в отрицательной области PC1; Озеро 3: демонстрирует наибольший разброс, занимая широкий диапазон значений по обоим осям, что может указывать на более высокую морфологическую пластичность или разнородность условий обитания в данном водоёме.

Распределение точек вдоль осей согласуется с направлением векторов нагрузок: особи из Озера 1, расположенные в области положительных значений PC1, тяготеют к признакам ao , H и IV , тогда как особи из Озера 2 (отрицательный PC1) — к признакам ID и pD .

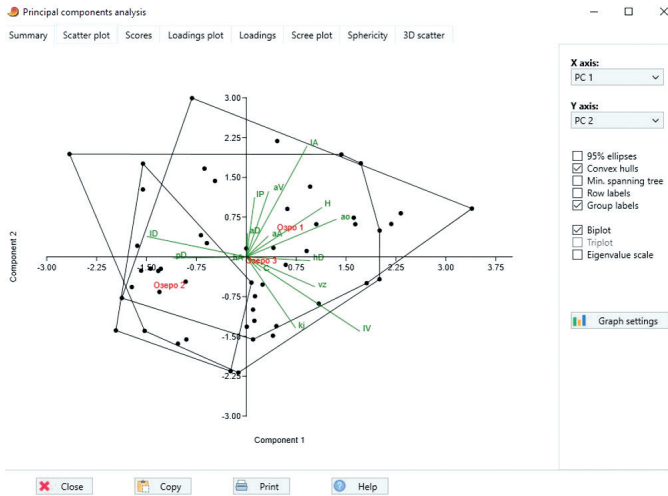


Рис. 7.6. Режим *Within-group*. Ординационная диаграмма (биplot) морфометрической изменчивости гольца из трёх озёр.

Важно отметить, что режим *Within-group* устраняет различия между средними значениями групп, центрируя данные внутри каждой популяции. Тот факт, что после такой центровки выборки из трёх озёр всё ещё демонстрируют частичное пространственное разделение, указывает на то, что характер индивидуальной изменчивости (ковариация признаков) не идентичен в разных популяциях. Иными словами, «форма» морфологического разнообразия внутри каждого озера имеет свои особенности. В то же время существенное перекрытие облаков точек показывает, что эти различия носят градиентный, а не дискретный характер, и индивидуальная вариабельность внутри популяций доминирует над межпопуляционными расхождениями.

Sphericity

Результаты теста сферичности Бартлетта подтверждают пригодность данных для применения PCA. Нулевая гипотеза о том, что матрица корреляций не отличается от единичной (отсутствуют значимые связи между признаками), уверенно отвергается ($p\text{-value} \approx 0$; $\chi^2 = 1989,6$ при 119 степенях свободы), что свидетельствует о выраженной коррелированности морфометрических переменных и структурированности данных. Это обосновывает использование анализа

главных компонент для выявления скрытых градиентов изменчивости и указывает на то, что обнаруженные паттерны морфологической вариации с высокой вероятностью отражают реальные, биологически осмысленные взаимосвязи между признаками, а не случайный шум.

Заключение

Проведённый внутригрупповой PCA выявил основные паттерны морфологической изменчивости гольца, общие для трёх исследованных озёр. Главные оси вариации описывают согласованное изменение комплекса признаков, связанных с размером головы, высотой тела и длиной плавников, то есть характеризуют сочетание размерных и формных градиентов, типичных для морфометрических данных. Визуальный анализ распределения особей на биплотах показывает значительное перекрытие групп при наличии выраженных популяционных тенденций, что указывает на сохранение общего плана строения при наличии тонких, но статистически значимых различий в пропорциях тела между популяциями. Выявленные морфологические градиенты представляют интерес для дальнейших исследований, направленных на сопоставление этих различий с экологическими условиями, структурой возрастных групп и другими факторами, влияющими на онтогенез и пластичность формы.

Практическая задача. Режим *Between-group*

Режим *Between-group* PCA предназначен для максимизации различий между заранее определёнными группами. Он позволяет выявить признаки, в наибольшей степени ответственные за дифференциацию популяций, и визуализировать основные тенденции межгрупповой изменчивости. Набор данных идентичен предыдущему примеру (голец из трёх озёр), но фокус анализа смещён с индивидуальной вариации на различия между средними значениями групп.

Интерпретация *Loadings* и дисперсии

Анализ выполнен на основе ковариационной матрицы (*Variance-covariance matrix*), поскольку все морфометрические признаки измерены в сопоставимых единицах (мм), что позволяет сохранить информацию об абсолютном масштабе различий. Проведённый *Between-group* PCA выявил выраженные различия между средними значениями морфометрических признаков гольца из трёх озёр: первые две главные

компоненты объясняют подавляющую часть межгрупповой дисперсии (PC1 — 62,66%, PC2 — 34,80%, суммарно 97,46%). Столь высокая доля объяснённой дисперсии типична для режима *Between-group*, так как внутригрупповая вариация не учитывается, а анализ концентрируется только на различиях средних.

Анализ нагрузок показывает структуру этих различий. PC1 наиболее сильно положительно нагружена признаками *ID* (длина спинного плавника, 0,63) и *kj* (минимальная высота тела, 0,46), то есть разделяет популяции по длине спинного плавника и высоте хвостового стебля. PC2 демонстрирует наибольшие положительные нагрузки по признакам *hD* (высота спинного плавника, 0,50), *ao* (длина головы, 0,41) и *H* (максимальная высота тела, 0,37), что интерпретируется как градиент общих размеров головы и высоты тела. В совокупности PC1 и PC2 описывают два основных морфологических градиента: специфические пропорции плавников (PC1) и общую «высотность» и крупность тела (PC2).

График *Loadings*

График нагрузок (векторов признаков) наглядно демонстрирует вклад каждого параметра в формирование осей межгрупповой изменчивости. Признаки *hD*, *ao* и *H* формируют компактную группу векторов, направленных вверх и вправо (область положительных PC1 и PC2), что указывает на их согласованную роль в дифференциации популяций по размеру головы и высоте тела. Вектор *ID* имеет практически чистую проекцию на ось PC1 (направлен преимущественно вправо), подчёркивая его доминирующую роль в разделении популяций по первой компоненте. Признаки *C*, *aA* и *IV* также ориентированы вправо, что указывает на их корреляцию с градиентом PC1 и участие в разделении групп по длине тела и передне-задним пропорциям. Совместное расположение векторов позволяет выделить два главных градиента: градиент общего размера и высоты тела (связанный с *hD*, *ao*, *H*) и градиент пропорций плавников и задней части тела (связанный с *ID*, *kj* и сопряжёнными признаками).

Визуализация (*Biplot*)

Диаграмма (рис. 7.7), отображающая положения групповых центроидов (средних значений по озёрам) в пространстве PC1–PC2, выявляет чёткое пространственное разделение всех трёх популяций. Популяция Озера 1 занимает обособленное положение в правом верхнем

квадранте, характеризуясь максимальными положительными значениями по обоим осям. Популяция Озера 2 концентрируется в левом нижнем квадранте, демонстрируя выраженные отрицательные значения по PC1 и PC2. Популяция Озера 3 располагается вблизи начала координат, занимая промежуточное положение по комплексу признаков с небольшим смещением в отрицательную область PC1.

Взаимное расположение векторов и точек групп раскрывает природу межпопуляционных различий. Векторы *lD* и *kj*, направленные вправо, указывают на то, что популяция Озера 1 отличается наибольшей длиной спинного плавника и высотой хвостового стебля по сравнению с общим средним, тогда как популяция Озера 2, расположенная в противоположной стороне, характеризуется укороченным спинным плавником и более низким хвостовым стеблем. Векторы *hD*, *ao* и *H*, направленные вверх, демонстрируют, что Озеро 1 объединяет наиболее крупнотелых рыб с высокой головой и телом, тогда как Озеро 2 — наиболее мелких и низкотелых особей. Популяция Озера 3, находящаяся ближе к центру, представляет собой промежуточный морфотип, чьи средние значения по ключевым признакам близки к общему среднему по набору. Такое распределение свидетельствует о наличии устойчивых морфологических различий между тремя популяциями: наибольший контраст наблюдается между Озером 1 (крупные, длинноплавниковые рыбы) и Озером 2 (мелкие, короткоплавниковые), тогда как Озеро 3 занимает промежуточное положение.

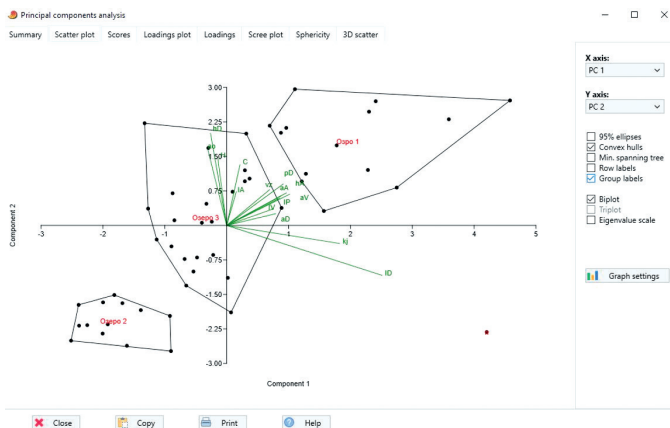


Рис. 7.7. Режим *Between-group*. Ординационная диаграмма (биplot) морфометрической изменчивости гольца из трёх озёр

Sphericity

Для *Between-group* PCA тесты сферичности Бартлетта и меры адекватности выборки КМО неприменимы, поскольку они оценивают корреляционную структуру *исходной* матрицы данных, тогда как *Between-group* метод оперирует вырожденной матрицей групповых центроидов. В такой матрице число ненулевых собственных значений всегда равно $k-1$ (где k — число групп), что автоматически приводит к нулевому определителю и формально «плохим» значениям КМО, не отражающим реальную пригодность данных для *Between-group* PCA.

Заключение

Проведённый *Between-group* PCA выявил статистически значимые морфологические различия между популяциями гольца из трёх озёр. Главные оси межгрупповой вариации демонстрируют высокий уровень объяснённой дисперсии, что указывает на выраженную дифференциацию средних значений признаков между популяциями. Наибольший вклад в разделение групп вносят признаки, связанные с размером и формой плавников, а также общими пропорциями тела; чёткое пространственное разделение групповых центроидов в пространстве PC1–PC2 свидетельствует об устойчивом характере выявленных различий.

Популяция из Озера 1 демонстрирует наиболее выраженные отличия по комплексу исследованных признаков, тогда как популяция из Озера 3 занимает промежуточное положение между Озером 1 и Озером 2. Полученные результаты согласуются с гипотезой об адаптивной дифференциации исследованных популяций в ответ на специфические условия обитания в разных водоёмах, однако для её окончательной проверки необходимы дополнительные экологические и генетические данные.

Сравнительный анализ режимов *Within-group* и *Between-group*

Режимы *Within-group* и *Between-group* PCA представляют собой взаимодополняющие подходы к анализу морфометрической изменчивости. *Within-group* анализ фокусируется на внутрипопуляционной вариации, показывая широкое перекрытие особей из разных озёр после устранения межгрупповых различий и тем самым описывая «форму» индивидуального морфологического разнообразия внутри популяций. *Between-group* анализ, напротив, выделяет различия между средними

значениями популяций и демонстрирует чёткое разделение групповых центроидов, акцентируя устойчивые межпопуляционные сдвиги по комплексу признаков. Ключевое отличие состоит в интерпретации результатов: *Within-group* отвечает на вопрос «Как варьируют особи внутри каждой популяции?», тогда как *Between-group* — на вопрос «Насколько и по каким признакам различаются средние характеристики популяций?». Совместное применение этих режимов даёт более полное представление о структуре морфологической изменчивости, позволяя одновременно описать индивидуальную вариабельность и межпопуляционные различия.

Философское эхо

Метод РСА — проектировщик многомерных конструкций, где исходные признаки сплетены, как арматурный каркас, а истинная геометрия объекта скрыта за лесами избыточных измерений. Его задача — не пересчитать каждую стяжку, а выявить главные несущие направления, вдоль которых конструкция воспринимает нагрузку изменчивости. Последовательно определяя ортогональные оси, он выстраивает систему независимых опор: первая (PC1) ложится в вектор максимальной жёсткости, принимая на себя основной вес информации; вторая (PC2) — строго перпендикулярно, фиксирует остаточные деформации в следующем измерении, и так далее, пока в облаке исходных замеров не проступит чёткая силовая схема сходства. Важно помнить: РСА не даёт детализировки каждого узла, а чертит каркас доминантных нагрузок, где видны траектории главных усилий. Эффективность расчёта зависит от приведения данных к единому масштабу — стандартизации, подобной выверке нулевых отметок фундамента: без неё элемент с крупным диапазоном значений перетянет на себя усилия, исказив картину второстепенных связей. Метод остро реагирует на локальные дефекты — выбросы: одна аномальная точка, словно скрытая раковина в бетоне или критический прогиб балки, способна сместить расчётные оси и обрушить устойчивость интерпретации. В руках исследователя, понимающего механику данных, РСА превращает отдельные замеры в рабочий проект скрытых структур, где каждая ось — направление силового потока, связывающее систему с её средой. Но когда признаки связаны нелинейными деформациями или данные разнородны по природе, расчёт теряет точность, подобно статической модели, применённой к пластичному

грунту: его линейная логика работает лишь в пределах упругой работы материала и при корректной подготовке исходной базы. Он даёт ясную картину несущего остова, отсеивая фоновый шум, но требует признания: главные направления распределения нагрузок — это и есть несущая конструкция понимания, на котором возводится здание интерпретации.

8. АНАЛИЗ ГЛАВНЫХ КООРДИНАТ (РСОА)

Введение

Анализ главных координат РСоА (*Principal Coordinate Analysis*), предложенный Джоном Гауэром в 1966 году, занимает особое место среди методов многомерной ординации. Этот метод был разработан для преодоления ключевого ограничения анализа главных компонент (РСА), связанного с требованиями к типу исходных данных. В то время как РСА работает непосредственно с матрицей признаков и предполагает количественные переменные, измеренные в сопоставимых шкалах и связанные преимущественно линейно, РСоА оперирует матрицей попарных расстояний между объектами. Это принципиально расширяет область его применения: становится возможным использование любой биологически осмысленной меры различия, включая неевклидовы метрики (например, Брея–Кёртиса, Жаккара). Фундаментальная задача РСоА заключается в реконструкции конфигурации объектов в пространстве низкой размерности (обычно 2D или 3D) таким образом, чтобы евклидовы расстояния между точками на ординационной диаграмме максимально точно воспроизводили исходные дистанции. Тем самым РСоА преобразует абстрактную матрицу несходств в геометрически наглядную «карту» структуры данных.

Благодаря своей универсальности РСоА широко используется в экологии, популяционной генетике и морфометрии для анализа структуры сообществ, генетических дистанций и морфологической изменчивости. На ординационной карте группы сходных объектов формируют кластеры, а оси отражают главные направления различий. При этом оси РСоА не имеют автоматической экологической интерпретации: их осмысление требует сопостав-

ления с внешними переменными (например, факторами среды) и, при необходимости, применения методов прямой ординации (CCA, RDA) для наложения векторов среды на полученную конфигурацию.

Методология

Метод анализа главных координат (РСоА) позволяет перейти от матрицы попарных различий между объектами к их координатам в пространстве меньшей размерности. В отличие от РСА, использующего исходные признаки, входными данными для РСоА служит симметричная матрица дистанций, элементы которой количественно описывают степень несходства между каждой парой объектов, рассчитанную по выбранной метрике (евклидова, Брея–Кёртиса, Жаккара, Махаланобиса и др.). Задача метода — найти такое расположение точек в двумерном или трёхмерном евклидовом пространстве, при котором расстояния между ними максимально точно аппроксимируют исходную матрицу различий.

Ключевым этапом алгоритма является преобразование матрицы квадратов дистанций в матрицу скалярных произведений, допускающую геометрическую интерпретацию. Сначала элементы матрицы расстояний возводятся в квадрат. Затем выполняется процедура двойного центрирования: из каждого элемента вычитаются средние значения по соответствующим строке и столбцу и добавляется общее среднее по всей матрице. Эта операция переводит описание системы с уровня «абсолютных» дистанций на уровень взаимного расположения объектов, подготавливая данные к спектральному разложению.

Полученная матрица скалярных произведений подвергается спектральному разложению: вычисляются собственные значения и собственные векторы. Собственные векторы задают направления координатных осей, а собственные значения отражают вклад соответствующих осей в общую вариацию. Координаты объектов вдоль каждой оси получают умножением собственных векторов на квадратные корни из соответствующих положительных собственных значений; первые оси, объясняющие наибольшую долю вариации, используют для построения двумерных или трёхмерных диаграмм.

Важной особенностью метода является возможность появления отрицательных собственных значений. Их наличие не свидетельствует об ошибке вычислений, а указывает на неевклидовость исходной

матрицы дистанций, то есть невозможность точного вложения этих расстояний в евклидово пространство без искажений. Для осей с отрицательными собственными значениями корректные координаты в вещественном пространстве задать нельзя, поэтому для интерпретации используют только оси с положительными значениями; сумма положительных собственных значений показывает, какая доля структурированной вариации дистанций может быть корректно отображена. Величина и суммарный вклад отрицательных собственных значений характеризуют степень неевклидовости метрики и потенциальный объём искажений. В практике используют специальные корректировки (например, преобразование *Cailliez* или *Lingoes*), позволяющие привести матрицу расстояний к евклидовому виду и улучшить интерпретируемость ординации.

Характер данных

Анализ главных координат (РСоА) применяют в тех случаях, когда исходной информацией служит не матрица измеренных признаков, а матрица различий между объектами. Метод работает с симметричной матрицей попарных дистанций, где каждое значение отражает степень несходства между парой объектов. Исходные переменные могут быть количественными, порядковыми или категориальными — при условии, что выбранная мера различия биологически осмысленна и согласована с типом данных. Если в качестве дистанции используется евклидова метрика по стандартизированным признакам, результат РСоА будет математически эквивалентен РСА, но получен через матрицу расстояний. Однако при сильной неевклидовости данных эффективность метода снижается: появляется заметная доля отрицательных собственных значений, и часть информации теряется при проекции.

РСоА особенно полезен там, где ключевой интерес связан со структурой различий. В популяционной генетике его используют для визуализации генетических дистанций между популяциями. В геометрической морфометрии к РСоА прибегают при анализе матриц прокрустовых расстояний, чтобы выявить главные направления изменчивости формы тела; прокрустово расстояние в этом случае выступает как мера различия форм, вычисляемая как корень из суммы квадратов расстояний между одноимёнными точками после их оптимального совмещения. В анализе сообществ РСоА широко применяют после

преобразования данных методом Хеллингера и вычисления евклидовых расстояний или при использовании метрик Брея–Кёртиса и Жаккара для отображения градиентов среды через относительное расположение проб.

Вместе с тем существуют ситуации, когда применение РСоА ограничено или малоинформативно. Использование евклидова расстояния по произвольно закодированным категориальным признакам лишено биологического смысла и приводит к артефактам. Матрицы с большим числом нулей и резко различной численностью видов могут давать сильную неевклидовость (значительные отрицательные собственные значения) и искажения ординации. Кроме того, РСоА оптимизирован для сохранения общей геометрии расстояний и лучше передаёт плавные градиенты, чем жёстко дискретные кластеры; при ярко выраженной кластерной структуре требуется осторожная интерпретация или использование дополнительных методов (например, pMDS).

Таким образом, корректное применение РСоА требует осознанного выбора меры различия и оценки её соответствия структуре данных. Когда дистанция биологически интерпретируема и достаточно близка к евклидовой (доля отрицательных собственных значений мала), метод даёт наглядное и информативное представление о структуре различий между объектами. В противном случае ординационные диаграммы следует рассматривать как приблизительную визуализацию и интерпретировать с особой осторожностью.

Практическая задача

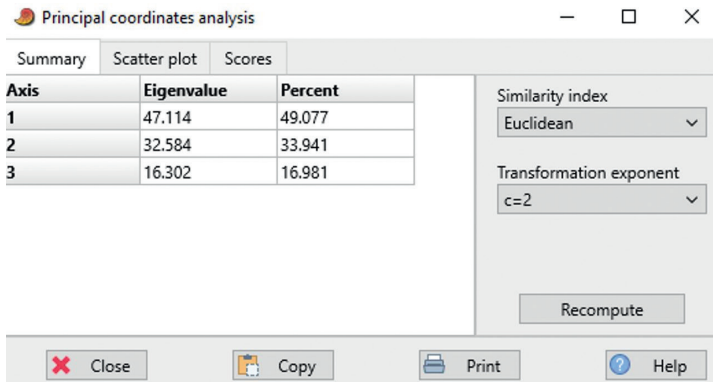
У плотвы из пяти озёр Самарской области оценили концентрацию Fe, Zn и Cu, мг/кг, в мышечной ткани. Используя метод РСоА, определить, какие популяции наиболее чётко разделяются по содержанию тяжелых металлов в мышечной ткани, и какие металлы вносят наибольший вклад в это разделение.

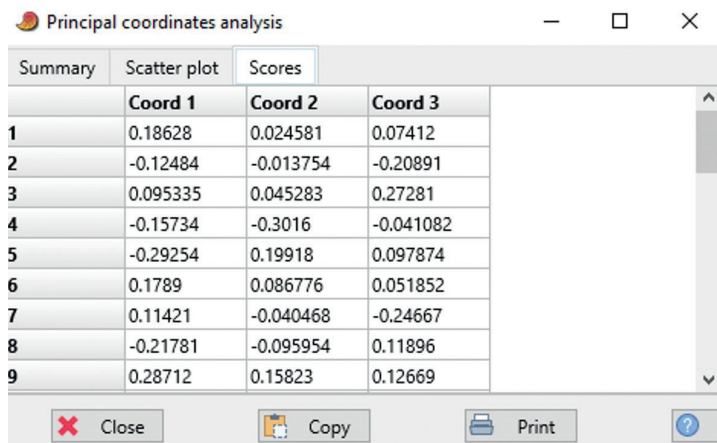
РЕШЕНИЕ ЗАДАЧИ

Перед загрузкой данных в PAST необходимо выполнить Z-стандартизацию переменных. Это позволяет привести концентрации металлов к единому масштабу (нулевое среднее, единичная диспер-

сия), чтобы ни один из металлов не доминировал в расчёте расстояний за счёт больших абсолютных значений. Анализ главных координат (PCoA) в PAST запускается через меню *Multivariate* → *Ordination* → *PCoA*. В меню *Summary* (рис. 8.1) в выпадающем списке *Similarity index* следует выбрать метрику *Euclidean*. В условиях Z-стандартизации евклидово расстояние является естественным выбором, так как оно напрямую отражает различия между образцами в стандартизованном многомерном пространстве и корректно работает с отрицательными значениями.

Метрика Брея–Кёртиса (*Bray–Curtis*) в данном случае неприменима: после Z-преобразования часть значений становится отрицательной, тогда как Брея–Кёртиса определена только для неотрицательных величин и трактуется как мера различия по относительным вкладом признаков. При наличии отрицательных значений суммирование модулей и попарных минимумов теряет интерпретируемость, и биологический смысл дистанций размывается. После нажатия кнопки *Recompute* в окне результатов появляется таблица с собственными числами (*Eigenvalue*) и долей объяснённой вариации (*Percent*) для каждой оси. По значениям *Percent* можно оценить, какая часть различий между популяциями по содержанию Fe, Zn и Cu адекватно отображена на первых осях PCoA.





	Coord 1	Coord 2	Coord 3
1	0.18628	0.024581	0.07412
2	-0.12484	-0.013754	-0.20891
3	0.095335	0.045283	0.27281
4	-0.15734	-0.3016	-0.041082
5	-0.29254	0.19918	0.097874
6	0.1789	0.086776	0.051852
7	0.11421	-0.040468	-0.24667
8	-0.21781	-0.095954	0.11896
9	0.28712	0.15823	0.12669

Рис. 8.1. Окно анализа главных координат (PCoA) в PAST, меню *Summary* (стр. 53) и меню *Scores* (стр. 54)

Визуализация

Метод PCoA выявил три ординационные оси, что математически соответствует числу исходных переменных (Fe, Zn, Cu). Совокупно они объясняют 100% дисперсии: ось 1 — 49,08%, ось 2 — 33,94%, ось 3 — 16,98%. Доля объяснённой дисперсии на первых двух осях (~83%) указывает на достаточную репрезентативность двумерной проекции для визуального анализа основных закономерностей.

На ординационной диаграмме (рис. 8.2) наблюдается значительное перекрытие выпуклых оболочек (*convex hulls*), соответствующих пяти озёрам. Точки-пробы концентрируются вблизи начала координат в диапазоне значений *Coord 1* $\approx [-0,3; +0,3]$ и *Coord 2* $\approx [-0,4; +0,4]$, что указывает на отсутствие резкой дифференциации популяций по профилю накопления металлов. Отсутствие обособленных кластеров и компактное расположение точек свидетельствуют о высокой степени сходства уровней биоаккумуляции Fe, Zn и Cu в мышечной ткани плотвы из всех исследованных водоёмов.

Важно отметить методологический нюанс: в PCoA — в отличие от PCA-биplota — экологическая интерпретация осей не является автоматической. Без наложения векторов нагрузок или корреляционного анализа с исходными признаками нельзя однозначно указать, какой именно металл преимущественно определяет ту или иную ось.

Тем не менее доминирование первой оси (~49%) указывает на наличие основного градиента изменчивости, тогда как вторая ось (~34%) отражает дополнительный, ортогональный компонент вариации; их возможная связь с общим уровнем накопления и пропорциями элементов должна проверяться отдельно с использованием векторов нагрузок или *envfit*-анализа.

Третья ось, объясняющая оставшиеся ~17% дисперсии, математически необходима при анализе трёх переменных и дополняет картину изменчивости, выходя за рамки двумерной проекции. Визуальная интерпретация обычно фокусируется на первых двух осях, однако при необходимости ось 3 может быть использована для трёхмерного отображения или анализа остаточной вариации.

Таким образом, результаты РСОА подтверждают высокую степень сходства популяций плотвы по изученному комплексу тяжёлых металлов. Для выявления более тонких различий или связи структуры ординации с факторами среды целесообразно дополнить анализ методами прямого градиентного анализа (RDA) либо корреляционным тестированием координат с экологическими переменными.

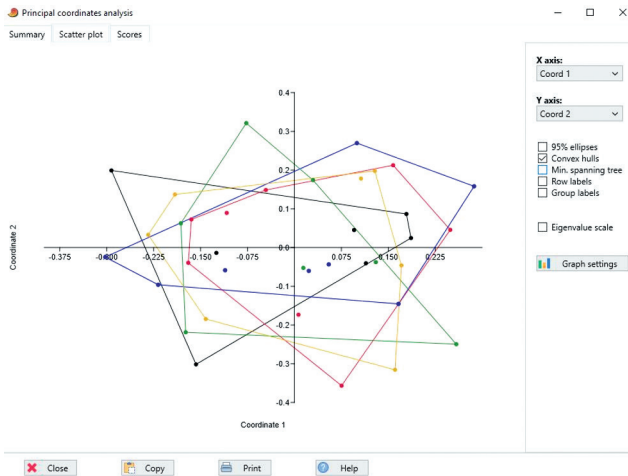


Рис. 8.2. Распределение проб рыб по концентрациям металлов (Fe, Zn, Cu) в пространстве главных координат (РСОА)

Философское эхо

Метод РСoA — хирург многомерного пространства, обнажающий скрытую анатомию сходства. Его инструмент — матрица дистанций, подобная послойному скану, выявляющему структурные сдвиги в тканях данных, где каждый параметр фиксируется как клинический признак системного состояния. Сила метода — в процедуре двойного центрирования и последующем спектральном разложении, которое работает как послойные надрезы, обнажающие главные оси изменчивости и демонстрирующие внутреннюю геометрию объекта. Однако если выбранная мера расстояния не укладывается в евклидову геометрию, отрицательные собственные значения проявляются как артефакты сканирования, порождающие фантомные контуры и нарушающие естественные анатомические пропорции среза, что сигнализирует о метрической несовместимости исходных данных. Метод формирует не плоский снимок, а объёмную диагностическую проекцию: скопления точек становятся зонами структурного сродства, а координатные расстояния — клинической мерой различия между состояниями. Исход процедуры зависит от выбранного протокола: некорректная метрика, подобная несовместимости тканей, вызывает методологическое отторжение и ставит под угрозу точность диагноза. В руках исследователя, способного пальпировать цифры и распознавать за ними реальные биологические силы — генетический дрейф, морфологическую пластичность, экологические градиенты, — РСoA превращает таблицу симптомов в связную историю болезни системы. Каждая точка здесь — не абстрактный маркер, а диагностический срез состояния, а оси — направляющие изменчивости, формирующие топографию разнообразия на аналитическом снимке. Но, как и после любой сложной процедуры, результат требует клинической верификации: без вдумчивой интерпретации даже точный рентгеновский снимок останется лишь стерильным изображением, где геометрическая чёткость так и не обретёт биологического смысла.

9. НЕМЕТРИЧЕСКОЕ МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ (NMDS)

Введение

Метод неметрического многомерного шкалирования nMDS (*non-metric Multidimensional Scaling*), разработанный Джозефом Крускалом в 1964 году, предназначен для анализа сложных данных с неевклидовой природой, высоким уровнем шума или выраженной нелинейностью. В отличие от анализа главных координат (РCoA), ориентированного на воспроизведение абсолютных значений дистанций в пространстве низкой размерности, цель nMDS заключается в сохранении рангового порядка различий между объектами. Такой подход позволяет визуализировать структуру сходства, опираясь на относительную иерархию несходства, что особенно важно в биологических исследованиях, где относительное положение объектов зачастую информативнее точных количественных оценок дистанций.

Метод nMDS широко применяется в экологии сообществ, генетике и морфометрии, где биологическое сходство часто описывается нелинейными или полуметрическими мерами дистанции. Алгоритм является итеративным и основан на минимизации функции стресса (*Stress*), которая количественно оценивает степень соответствия между ранговым порядком исходных дистанций и расстояний на ординационной диаграмме. Низкие значения стресса (например, менее 0,1) указывают на надёжность конфигурации, тогда как значения выше 0,2 свидетельствуют о существенных искажениях и требуют либо увеличения размерности, либо особо осторожной интерпретации. Таким образом, nMDS обеспечивает топологически корректное представление структуры данных, жертвуя метрической точностью ради сохранения относительного порядка различий.

Методология

Метод nMDS принципиально отличается от метрических методов (PCA, РCoA) своей философией: он стремится сохранить не величины расстояний, а их порядок. На ординационной диаграмме корректно отражается, какие объекты расположены ближе или дальше друг от друга, но сами численные значения дистанций в пространстве низкой размерности не обязаны совпадать с исходными.

Алгоритм реализуется в виде итеративной процедуры минимизации функции стресса: 1) формируется начальная конфигурация точек в заданном пространстве (обычно 2D или 3D); 2) для этой конфигурации вычисляются евклидовы расстояния между точками; 3) сравнивается ранговый порядок этих расстояний с ранговым порядком исходной матрицы дистанций; 4) степень рассогласования между ними вычисляется в виде стресса; 5) положение точек последовательно корректируется методами численной оптимизации (например, градиентного типа) до тех пор, пока стресс перестаёт заметно уменьшаться.

В итоговой конфигурации сохраняется относительный порядок различий: если в исходных данных дистанция А–В меньше, чем А–С, то на диаграмме точка В будет ближе к А, чем С. Ключевое свойство nMDS — инвариантность к монотонным преобразованиям дистанций: поскольку метод оперирует рангами, любые монотонные преобразования (логарифмирование, возведение в степень, извлечение корня) не изменяют ранговую структуру и, следовательно, не влияют на конфигурацию точек. Это делает nMDS устойчивым к масштабу дистанций и форме их распределения.

В то же время метод имеет ограничения. Итеративный процесс может сходиться к локальным минимумам функции стресса, поэтому на практике рекомендуется выполнять несколько запусков с разными начальными конфигурациями и выбирать решение с минимальным стрессом. Кроме того, расстояния на диаграмме не имеют строгого метрического смысла и отражают только порядок различий, что необходимо учитывать при интерпретации: нельзя напрямую трактовать «кратность» различий по длине отрезков в ординационном пространстве.

Характер данных

Метод nMDS основан на анализе ранговой структуры различий между объектами и требует наличия биологически осмысленной меры дистанции, даже если она не является евклидовой. Он не предъявляет строгих требований к линейности зависимостей, нормальности распределений или сопоставимости шкал признаков и работает с любой симметричной матрицей попарных различий. Метод демонстрирует относительную устойчивость к шуму и отклонениям от классических предпосылок, однако сильные выбросы могут изменять ранговую структуру дистанций и тем самым влиять на конфигурацию.

nMDS особенно эффективен в ряде типичных задач. При анализе сообществ с контрастной экологией (например, ихтиоценозов водоёмов разного трофического статуса) использование меры Брея–Кёртиса позволяет выявлять чёткие кластеры, соответствующие типам сообществ; при стрессе $<0,1$ такая конфигурация обычно считается адекватной. В популяционной генетике метод способен отражать сложную структуру гибридных популяций, где объекты формируют характерные геометрические фигуры (например, «треугольник» гибридизации, с вершинами — родительскими формами и промежуточными положениями гибридов разной степени интрогрессии). При анализе динамики сообществ после нарушений nMDS хорошо передаёт нелинейные траектории изменений (например, по данным обилия видов после токсического сброса, преобразованным через расстояния Хеллингера), выявляя значительное удаление между исходным и конечным состояниями и тем самым подчёркивая необратимость сдвигов. В отличие от строго метрических методов, дающих более линейную проекцию, nMDS лучше сохраняет сложность нелинейных изменений.

Существуют и принципиальные ограничения. При анализе плавных экологических градиентов nMDS, как и другие ординационные методы, может порождать искривлённые конфигурации (например, «подковы»), визуально искажая линейный градиент. В таких случаях методы, сохраняющие метрические свойства и осевую интерпретацию (PCA, PCoA), часто дают более прямое и интерпретируемое представление. nMDS не сохраняет информацию об абсолютных величинах различий, что ограничивает его применение в задачах, где важны сами значения дистанций и их изменение во времени (например, количественная оценка скорости генетической дивергенции). Метод также не предназначен для отображения реальных географических координат, поскольку формирует топологически корректную, но геометрически искажённую проекцию.

Таким образом, использование nMDS должно быть обосновано природой данных и исследовательскими задачами. Метод наиболее эффективен при анализе сложных, нелинейных структур различий и кластерной организации сообществ, но менее информативен в задачах с выраженными линейными градиентами, требованием точной метрической интерпретации или прямой географической визуализации.

Пример анализа данных

Задача. На четырёх участках реки с различной степенью антропогенного воздействия (фоновая, сельхозугодья, городская, промзона) получены уловы шести видов рыб (плотва, карась, окунь, пескарь, лещ, ерш) стандартным промысловым усилием с учётом селективности каждого вида по отношению к орудию лова. Используя ординационный тест nMDS, оценить различия в структуре ихтиоценозов между участками.

АНАЛИЗ ДАННЫХ

Перед загрузкой данных в PAST рекомендуется выполнить преобразование Хеллингера, что позволяет корректно работать с матрицами обилия видов и снижает влияние доминирующих таксонов при использовании евклидовой геометрии на расстояниях. При использовании меры Брея–Кёртиса возможно также работать с необработанными значениями обилия, поскольку эта метрика устойчива к нулям и опирается на относительные различия в структуре сообществ. Неметрическое многомерное шкалирование (nMDS) в PAST вызывается через меню *Multivariate* → *Ordination* → *nMDS*; ключевым для анализа является окно *Scatter plot*, где настраивается визуализация ординации.

Параметр *Similarity index* задаёт меру расстояния между пробами. Для анализа видового состава и обилия наибольшее распространение имеет мера Брея–Кёртиса, которая игнорирует совместные отсутствия видов и чувствительна к изменениям в доминировании. После выбора метрики необходимо нажать кнопку *Compute* для запуска расчёта. Опция *Column scores (WA)* в разделе настроек графики используется для наложения векторов видов: в поле задаётся число отображаемых векторов, отбираемых по силе корреляции с ординационной конфигурацией. Каждый вектор отражает направление максимального увеличения обилия вида в пространстве nMDS, а его длина пропорциональна степени согласованности распределения вида с конфигурацией проб; длинные векторы соответствуют видам-индикаторам, короткие — генералистам с размытым распределением.

Переключатели 2D и 3D определяют размерность пространства проекции. Двумерное решение (2D) предпочтительно для первичной интерпретации и иллюстраций, но может сопровождаться несколько более высоким значением стресса; переход к 3D даёт дополнительную

степень свободы и часто снижает стресс, однако усложняет визуальное чтение графика. При смене размерности в PAST ординация пересчитывается с новой случайной инициализации, из-за чего абсолютные координаты точек могут изменяться, но общая топология (взаимное расположение кластеров) при устойчивом сигнале в данных сохраняется. Кнопка *Compute* запускает итеративную оптимизацию: каждый запуск начинается со случайного расположения точек, и из-за наличия локальных минимумов разные прогоны могут давать несколько отличающиеся по ориентации конфигурации (повороты, отражения, масштаб), что является нормой для nMDS. Рекомендуется выполнить несколько запусков и ориентироваться на решение с минимальным значением стресса.

Визуализация

Анализ диаграммы рассеяния (рис. 9.1) демонстрирует выраженную пространственную сегрегацию ихтиоценозов, чётко привязанную к типу антропогенного воздействия. Первая координата выступает главным фактором разделения: отрицательная область объединяет участки с локальной или диффузной нагрузкой, тогда как положительная полуось выделяет территории с интенсивным промышленным влиянием. Кластер «Городская» образует компактную группу в верхнем левом квадранте, что указывает на формирование устойчивого, но трансформированного сообщества, адаптированного к хронической урбанистической нагрузке. Кластер «Фоновая» сосредоточен в нижнем левом квадранте, демонстрируя умеренную внутреннюю вариабельность, характерную для естественных условий русла. В противоположной части графика, в правой полуплоскости, располагается облако точек «Промзона», которое вытянуто вдоль первой оси и занимает отрицательные значения по второй координате; такая конфигурация отражает высокую неоднородность сообществ под воздействием стохастических промышленных факторов. Кластер «Сельхозугодья» отличается наибольшим разбросом и занимает центральную и правую части ординационного пространства, частично перекрываясь с фоновой зоной (одна из проб попадает в центральный диапазон), что свидетельствует о мозаичности условий и переходном характере агрогенной нагрузки. Перекрывание выпуклых оболочек между «Промзоной» и «Городской» с остальными участками минимально, подтверждая качественное различие ихтиоценозов в зонах точечного интенсивного воздействия по сравнению с диффузным или естественным фоном.

Интерпретация градиентов среды

Оси nMDS (*Axis 1*, *Axis 2*) являются математически произвольными и не имеют прямой физической или экологической размерности; их задача — обеспечить сохранение топологической структуры сходства, а не представлять конкретные факторы среды. Экологическая интерпретация выполняется через анализ взаимного расположения кластеров и наложение видовых векторов. В рассматриваемом примере первая координата выступает основным фактором разделения, чётко противопоставляя урбанизированные участки (отрицательные значения) всем остальным типам воздействия. Вторая координата дополнительно дифференцирует оставшиеся группы, выделяя промышленные зоны (положительные значения) в отдельный кластер, тогда как фоновые и агрогенные участки концентрируются в центральной и отрицательной области второй оси, формируя зону умеренного воздействия.

Анализ видовых векторов

Векторы видов (*Column scores*, *WA*) позволяют связать структуру ординации с эколого-физиологическими особенностями таксонов. Векторы окуня и пескаря ориентированы вдоль отрицательной части первой оси и пересекают облако точек «Фоновая», причём большая часть их длины приходится именно на эту зону. Такая геометрия указывает на то, что эти виды являются ядром ненарушенного сообщества, но их векторная траектория, выходящая за пределы выпуклой оболочки, допускает присутствие в переходных условиях с умеренным уровнем воздействия. Вектор карася примерно на две трети своей длины расположен внутри кластера «Сельхозугодья», что подчёркивает не просто толерантность вида, а его активное доминирование и оптимальное развитие именно в условиях диффузной агрогенной нагрузки, эвтрофикации и нестабильности гидрохимического режима.

Особого внимания заслуживает положение вектора плотвы, который целиком находится внутри облака «Сельхозугодья», не выходя за его пределы. Это свидетельствует о том, что в данном исследовании плотва выступает не как фоновый вид, а как специфичный индикатор трансформированных агроландшафтов, где она находит оптимальные условия и формирует устойчивые популяции. На диаграмме представлены векторы лишь четырёх из шести учтённых видов; отсутствие стрелок леща и ерша обусловлено алгоритмом векторной проекции, который отображает только таксоны, чьё распределение статистически

значимо коррелирует с осями ординации. Короткие или центрированные векторы леща и ерша были автоматически отфильтрованы программой, что указывает на их эврибионтность: эти виды встречаются во всех типах сообществ и не вносят значимого вклада в разделение кластеров.

Длина и точное расположение отображённых векторов характеризуют силу и специфику вклада вида в структурирование выборки. Чёткая приуроченность векторов окуня и пескаря к фоновой зоне, плотвы и карася — к сельскохозяйственной, а также обособленность кластера «Промзона» (где доминируют иные, не отображённые на векторной диаграмме механизмы формирования сообщества) подтверждают, что пространственная дифференциация ихтиоценозов определяется сменой комплекса доминантов вдоль градиента антропогенного воздействия. Для статистического подтверждения выявленной сегрегации групп целесообразно дополнить визуальный анализ методом непараметрической многомерной дисперсионной оценки (PERMANOVA), который количественно оценит значимость различий между кластерами. При необходимости детализации вклада отдельных таксонов в наблюдаемые различия рекомендуется использовать анализ SIMPER (*Similarity Percentages*), позволяющий ранжировать виды по степени их участия в формировании межгрупповой несхожести.

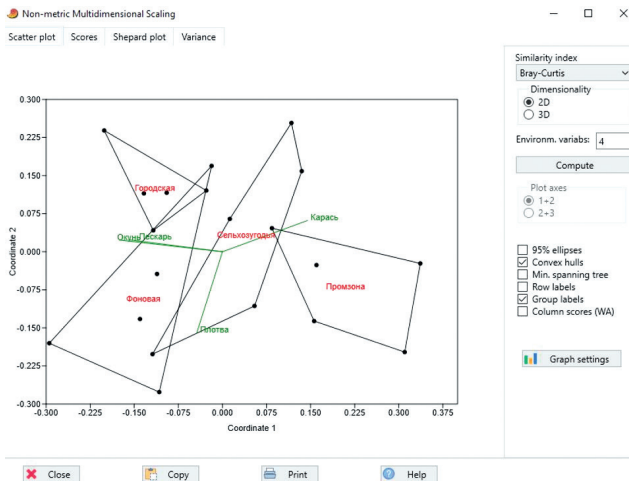


Рис. 9.1. Графи nMDS-ординации уловов рыб по четырём точкам отбора

Вкладка *Stress*

Stress является статистическим показателем, количественно характеризующим степень несоответствия между исходной матрицей ранговых различий и её геометрическим представлением в пространстве nMDS; его значения варьируют от 0 (идеальное совпадение) до 1 (полное расхождение). Чем ближе *Stress* к 0, тем точнее ординация сохраняет топологию сходства между объектами. Полученное значение $Stress = 0.103$ (рис. 9.2) является ключевой метрикой качества решения. Согласно общепринятым критериям (Clarke & Warwick, 2001), значения в интервале 0.1–0.2 интерпретируются как приемлемые/удовлетворительные, тогда как диапазон 0.05–0.1 считается хорошим, а >0.2 — указывающим на ненадёжность двумерной проекции. В данном случае стресс находится на верхней границе «хорошего» диапазона, что подтверждает адекватность визуализации общей структуры различий между ихтиоценозами.

Наблюдаемые на ординационной диаграмме паттерны — обособленность кластера «Промзона», широкий разброс точек «Сельхозугодий» и частичное перекрытие «Городских» и «Фоновых» проб — с высокой вероятностью отражают реальные экологические различия, а не артефакты алгоритма. При этом значение стресса около 0.10 требует осторожной интерпретации мелкомасштабных деталей: nMDS оптимизирует сохранение рангового порядка дистанций, а не их абсолютных величин, поэтому расстояния между близко расположенными точками внутри одного кластера могут быть локально искажены. Надёжнее опираться на глобальную конфигурацию (наличие, форму и взаимное расположение групп), чем на точные координаты отдельных проб; это особенно актуально при оценке направления и длины видовых векторов, которые могут незначительно варьировать между устойчивыми решениями.

Полученный уровень стресса согласуется с внутренней структурой данных. Суммарная доля объяснённой ранговой дисперсии двумя осями составляет ~84% (*Axis 1*: 0.497, *Axis 2*: 0.346), что дополнительно подтверждает достаточность двумерного представления. Высокая внутригрупповая вариабельность на участке «Сельхозугодья» проявляется в вытянутой форме облака точек, что усложняет точное воспроизведение всех многомерных дистанций в двух измерениях и вносит вклад в повышение стресса. Тем не менее ординация корректно передаёт ключевые биологические градиенты — выраженную

трансформацию сообществ вдоль оси антропогенного пресса и уникальность ихтиоценоза «Промзоны» — что делает полученную конфигурацию надёжной основой для содержательной экологической интерпретации.

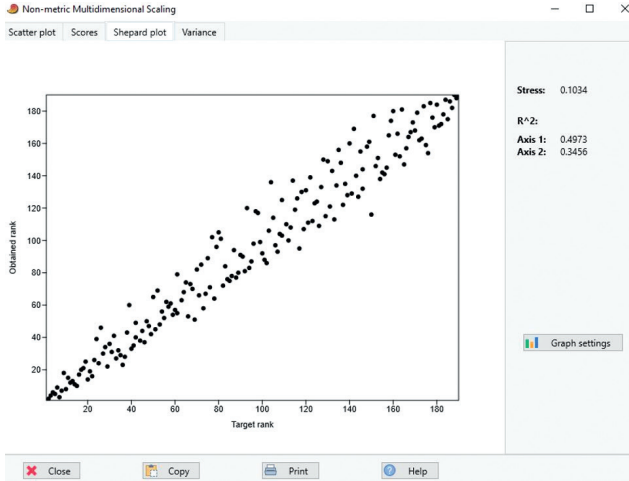


Рис. 9.2. Распределение проб в пространстве nMDS (Stress = 0.10); значение стресса ниже критического порога 0.2, что указывает на высокую достоверность полученной двумерной проекции и возможность её содержательной интерпретации

Философское эхо

Метод nMDS — систематик, работающий не с линейными замерами, а с порядком признаков, исходя из принципа, что в сложных сообществах ранговые отношения сходства часто надёжнее мнимой точности абсолютных расстояний. Он становится определительным ключом там, где данные зашумлены, нелинейны и не укладываются в жёсткие рамки евклидовой геометрии. Его логика основана на поиске конфигурации точек, минимизирующей «стресс» — меру рассогласования между ранговым порядком различий и их пространственным отображением, подобно тому, как таксономист последовательно исключает противоречивые диагностические признаки, чтобы ключ не уводил к неверному таксону. Метод не измеряет темп изменений и не выстраивает строгие градиенты; он чертит схему относительного сродства, где

важно лишь взаимное расположение объектов, как и в ранговой последовательности, где каждый шаг зависит от порядка, а не от величины отклонения. Его точность зависит от итеративной шлифовки: исследователь, подобно систематику, пересматривающему набор признаков, многократно запускает модель, чтобы избежать «ловушки локального минимума» — конфигурации, которая фиксирует случайное сходство вместо истинной структуры. В руках вдумчивого систематика-эколога nMDS превращает матрицу различий в живую схему классификации: каждая точка обретает место в пространстве сродства, а значение стресса служит индикатором надёжности определения. Но если исходные данные требуют строгой метрики или отражают непрерывный морфоклин, метод теряет остроту, подобно тому, как систематик пытается описать плавный переход формы дискретными ступенями. Без интерпретатора, способного перевести геометрию рангового напряжения в язык биологического смысла, диаграмма остаётся аккуратной, но немой таблицей признаков.

10. АНАЛИЗ СООТВЕТСТВИЙ (СА)

Введение

Анализ соответствий СА (*Correspondence Analysis*), разработанный Жан-Полем Бензери в 1960-х годах, был предложен как новый подход к выявлению латентных структур в категориальных данных. Метод возник как ответ на необходимость визуализировать скрытые взаимосвязи, закодированные в таблицах сопряжённости (частоты, присутствие/отсутствие), характерных для экологии, социологии, лингвистики и смежных дисциплин. Исходной идеей Бензери было рассматривать структуру данных не как набор свойств отдельных объектов или признаков, а как систему их взаимных соответствий.

Новаторство СА состоит в геометрическом моделировании таблиц сопряжённости. Матрица интерпретируется как два взаимосвязанных облака точек (строки и столбцы), проецируемых в единое пространство пониженной размерности. Центральным понятием является инерция (*Inertia*) — мера статистической значимости отклонения наблюдаемой совместной встречаемости категорий от ожидаемой при условии стохастической независимости; суммарная инерция пропор-

циональна статистике хи-квадрат, нормированной на объём выборки. Первая ось отражает направление максимальной инерции, то есть наиболее выраженную структурную закономерность в системе «строки–столбцы». Важнейшая особенность СА — симметричное представление: элементы обоих множеств отображаются в одном пространстве, и их взаимная близость используется для интерпретации ассоциаций. Вместе с тем метод подвержен специфическим артефактам, в частности «эффекту арки» (*arch effect*) при наличии сильных градиентов.

Концептуально СА отличается от PCA, PCoA и nMDS как целевой областью, так и базовой идеей. PCA работает с непрерывными количественными переменными и максимизирует дисперсию через линейные комбинации признаков; PCoA и nMDS используют заранее вычисленные расстояния между объектами и стремятся сохранить их (метрически или по рангам). СА же специализирован на таблицах сопряжённости и моделирует отклонение от независимости между строками и столбцами, представляя два разных множества сущностей (например, виды и местообитания) в общем факторном пространстве. Тем самым он ориентирован не столько на различия между объектами, сколько на структуру соответствий между классами категорий.

Методология

Анализ соответствий (СА) предназначен для исследования структуры таблиц сопряжённости, содержащих неотрицательные частоты, обилия или другие меры связи между двумя категориальными переменными (например, «станции × виды»). Исходные данные трактуются как система профилей: каждая строка описывается относительным распределением по столбцам (строковый профиль), а каждый столбец — по строкам (столбцовый профиль). Мерой различия между профилями служит расстояние хи-квадрат (*chi-square distance*), придающее больший вес отклонениям в редких категориях, поскольку остатки масштабируются ожидаемыми частотами.

Ключевой шаг СА — анализ отклонений от модели независимости. Для каждой ячейки вычисляется ожидаемое значение при предположении независимости строк и столбцов; далее формируются стандартизованные остатки (разность между наблюдаемой и ожидаемой частотой, делённая на корень из ожидаемой), образующие матрицу, к которой применяется сингулярное разложение (SVD). В результате выделяются ортогональные факторы (оси), упорядоченные по убыванию

их вклада в общую инерцию. Первая ось отражает наиболее сильный паттерн ассоциации между строками и столбцами, последующие оси описывают дополнительные независимые направления зависимости.

Результатом является биplot, где строки и столбцы одновременно отображаются в пространстве низкой размерности (обычно 2D). Близость точки строки и точки столбца интерпретируется как положительная ассоциация: соответствующая категория столбца относительно «перепредставлена» в этой строке по сравнению с моделью независимости. Близость точек внутри множества строк или столбцов указывает на сходство их профилей распределения, а удалённость от центра масс отражает специфичность профиля (чем дальше точка от центра, тем более она отклоняется от «среднего» профиля). При этом геометрическая интерпретация расстояний между строками и столбцами зависит от выбранного типа масштабирования (*scaling*), и их нельзя трактовать как обычную евклидову метрику без оговорок. Метод чувствителен к эффекту арки: при наличии выраженного основного градиента точки могут образовывать подковообразную конфигурацию, что требует осторожности либо использования детрендированного анализа соответствий (DCA).

Характер данных

Анализ соответствий (СА) ориентирован на данные, представленные в виде матриц сопряжённости с неотрицательными значениями, отражающими частоту или интенсивность связи между категориями. Типичными примерами являются экологические матрицы «участки × виды» с численностью, биомассой или проективным покрытием или генетические таблицы «популяции × аллели». Во всех этих случаях СА позволяет выявлять ассоциации между строками и столбцами и визуализировать их в общем факторном пространстве, фокусируясь на относительных соотношениях частот, а не на их абсолютном масштабе.

Метод неприменим к матрицам расстояний или сходств, рассчитанным заранее, а также к таблицам с отрицательными значениями или произвольным масштабированием, нарушающим предпосылку общего масштаба частот. СА не предназначен для анализа сырых непрерывных количественных переменных, которые требуют методов, основанных на ковариации или корреляции (например, PCA). Применение Z-стандартизации или других трансформаций, нарушающих интерпретацию частот и модели хи-квадрат, перед СА методически некорректно, поскольку искажает смысл инерции и расстояния.

Кроме того, крайне разреженные таблицы с очень высоким процентом нулей могут приводить к неустойчивым решениям и усиливать влияние редких категорий, что усложняет интерпретацию.

Таким образом, анализ соответствий является специализированным инструментом для исследования категориальных данных в форме частотных таблиц. При корректном выборе типа данных и учёте ограничений метод позволяет выявлять содержательные структурные зависимости между строками и столбцами и наглядно представлять их в пространстве главных осей, дополняя арсенал исследователя при изучении систем взаимосвязей.

Пример анализа данных

Задача. Получен количественно-видовой состав фоновых видов (плотва, окунь, густера, красноперка, лещ, карась) ихтиоценозов четырёх озёр Тверской области через стандартные промысловые усилия. Используя ординационный тест СА, оценить различия в структуре ихтиоценозов между озёрами.

АНАЛИЗ ДАННЫХ

Отсутствие стандартизации. При подготовке данных к анализу в СА и DCA важно избегать предварительной стандартизации, в том числе Z -преобразования. Эти методы опираются на χ^2 -расстояние для измерения различий между образцами, а данная метрика внутренне выполняет собственную стандартизацию: 1) преобразует абсолютные обилия видов в каждом образце в относительные доли (профили строк), 2) автоматически присваивает веса видам, обратно пропорциональные корню из их общей частоты встречаемости. Эта двойная стандартизация — по строкам (образцам) и столбцам (видам) — является частью математического аппарата СА/DCA и обеспечивает их чувствительность к различиям в видовых профилях. Поэтому внешняя Z -стандартизация нарушает внутреннюю логику χ^2 -расстояния, искажая встроенную систему весов видов и изменяя относительные соотношения обилий, на которые метод рассчитан.

Оценка градиентов. Перед проведением ординационного анализа в программе PAST принципиально важно оценить длину экологического градиента анализируемых данных. Этот показатель определя-

ет тип отклика видов на изменения среды и, соответственно, выбор метода: при коротких градиентах (<3 SD) отклик видов обычно близок к линейному и предпочтительны линейные методы (PCA, RDA), при длинных градиентах (>4 SD) отклик становится унимодальным (колоколообразным), и более уместны унимодальные методы (CA, CCA, DCA). Поскольку PAST не имеет прямой опции для этой оценки, используют DCA (*Detrended Correspondence Analysis*) как диагностический тест-индикатор. Выбор DCA, а не CA, связан с тем, что DCA устраняет артефакт «дуги», характерный для CA, и обеспечивает метрическую интерпретацию осей. Для оценки градиента нужно загрузить матрицу «образцы $\times$ виды» в PAST, выбрать *Multivariate* $\rightarrow$ *Ordination* $\rightarrow$ *Detrended Correspondence Analysis* (DCA) и открыть вкладку *Raw scores* для *Axis 1*. Значения *raw scores* следует скопировать в Excel и разделить на 100, так как PAST для удобства визуализации масштабирует их, умножая на 100; биологическая интерпретация требует возврата к исходному масштабу. Непосредственно длину градиента в стандартных отклонениях (SD) вычисляют как разность между максимальным и минимальным значениями *raw scores*, делённую на 100:

$$SD = \frac{\text{MAX (raw score)} - \text{MIN (raw score)}}{100}$$

Эта процедура занимает 1–2 минуты, но позволяет избежать серьёзных ошибок в анализе. Ключевое свойство DCA состоит в том, что единица длины вдоль оси 1 соответствует одному стандартному отклонению (SD) смены видового состава; это позволяет напрямую использовать разброс значений *Axis 1* как меру длины градиента: расстояние между двумя точками на оси отражает степень экологической дистанции в единицах SD. Например, если два образца разнесены на 4 единицы по оси 1 (после деления на 100), их видовые составы различаются примерно на 4 SD. Интерпретация результата строится по правилу: 1) <3 SD — короткий градиент, предпочтительны линейные методы (PCA, RDA); 2) >4 SD — длинный градиент, уместны унимодальные методы (CA, CCA, DCA); 3) 3–4 SD — «серая зона», в которой допустимы оба подхода в зависимости от целей исследования. Длину экологического градиента оценивают по оси с наибольшей объяснённой инерцией, которой обычно является *Axis 1*.

Анализ соответствий (CA) в PAST вызывается через главное меню *Multivariate* → *Ordination* → *CA*. Ключевой для интерпретации является вкладка *Scatter plot*, где отображается диаграмма рассеяния. На панели доступны режимы масштабирования *Row principal* и *Column principal*, определяющие, для какого множества — строк (образцов) или столбцов (видов) — оптимизируется отображение расстояний; обе опции используют одни и те же ординационные оси и сохраняют ключевые взаимосвязи, но смещают визуальный акцент. Оценка экологического градиента в рассматриваемом примере дала величину порядка 4 SD, что указывает на достаточно длинный градиент и методически обосновывает применение CA как унимодального метода ординации.

Визуализация (*Biplot*)

Основными результатами анализа являются диаграмма рассеяния (*biplot*) (рис. 10.1), таблица координат видов (*Column scores*) и сводная матрица собственных значений. Первая ось (*Axis 1*) отражает главный градиент межозёрных различий, объясняя 72,8% общей инерции (*Eigenvalue* = 0,181). Вторая ось (*Axis 2*) описывает дополнительный, менее выраженный градиент и вносит 12,1% в общую структуру (*Eigenvalue* = 0,030); суммарно первые две оси охватывают 84,8% изменчивости данных, что подтверждает высокую информативность двумерной проекции и её пригодность для содержательной интерпретации пространственной организации ихтиоценозов.

По оси 1 озёра 1 и 4 занимают противоположные положения и формируют чётко разделённые кластеры, свидетельствующие о фундаментальных различиях в видовом составе. Озеро 4 смещено в область положительных значений *Axis 1* и ассоциировано с таксонами, имеющими высокие положительные координаты: плотвой (0,546), краснопёркой (0,276) и лещом (0,165); это позволяет охарактеризовать его как водоём с выраженным «плотвично-лещовым» типом сообщества. Озеро 1, напротив, расположено в отрицательной части оси и связано с видами, имеющими отрицательные координаты: окунём (−0,489) и карасём (−0,155), представляя типичный «окунёво-карасиный» комплекс. Озёра 2 и 3 занимают промежуточные позиции в отрицательной области оси 1, указывая на структурную близость их ихтиофауны к «окунёвому» полю, однако с менее доминантной ролью окуня и более сложной видовой мозаикой.

Разделение по оси 2 уточняет внутреннюю дифференциацию озёр 2 и 3, близких по первому градиенту. Озеро 2 смещено в положительную область Axis 2 и коррелирует с видами, имеющими положительные координаты по этой оси, прежде всего с плотвой (0,148) и карасём (0,125), формируя смешанную структуру с элементами реофильного и эврибионтного комплексов. Озеро 3, напротив, расположено в отрицательной зоне оси 2 и характеризуется повышенной представленностью густеры (–0,381), красноперки (–0,172) и леща (–0,145); при общей близости к «окунёвому» кластеру по первой оси оно выделяется акцентом на придонно-бентофагную группировку. Положение видов на диаграмме определяется их координатами и векторной направленностью относительно начала координат: густера, обладающая максимальной отрицательной нагрузкой по оси 2, выступает строгим индикатором условий озера 3, тогда как плотва и окунь маркируют крайние положения градиента. В совокупности анализ выявляет два независимых экологических континуума: первый отражает переход от сообществ с доминированием плотвы и леща к окунёво-карасёвым ассоциациям, а второй описывает перераспределение вклада придонных и пелагических видов в переходных водоёмах в зависимости от локальных лимнологических факторов.

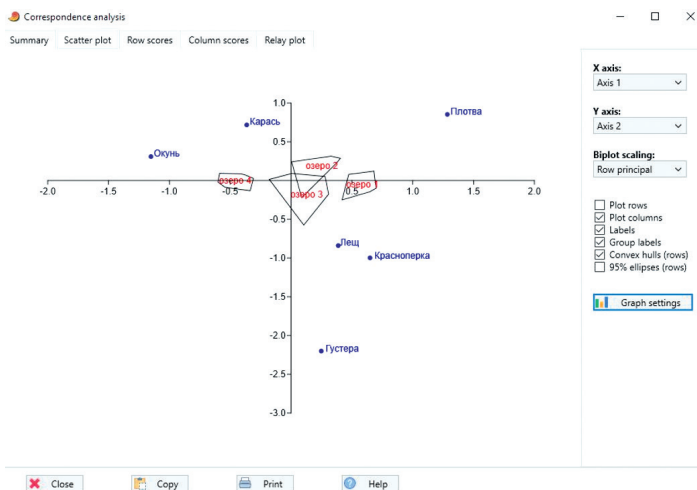


Рис. 10.1. График ординации анализа соответствий (СА)

 Correspondence analysis

Summary	Scatter plot	Row scores	Column scores	Relay plot	
	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Плотва	0.546262	0.148037	-0.0658407	0.0180565	-0.0143549
Окунь	-0.489921	0.0536377	-0.0317839	-0.000102388	-0.036388
Густера	0.10563	-0.381261	-0.164387	0.0365612	0.0567637
Красноперка	0.276058	-0.172828	0.193361	-0.320337	-0.0692775
Лещ	0.164996	-0.145396	0.319156	0.306358	-0.0524189
Карась	-0.154999	0.124539	0.137515	-0.046085	0.224195

 Correspondence analysis

Summary	Scatter plot	Row scores	Column scores	Relay plot	
Axis	Eigenvalue	% of total	Cumulative		
1	0.180937	72.762	72.762		
2	0.0300319	12.077	84.839		
3	0.0165294	6.6472	91.487		
4	0.0152498	6.1326	97.619		
5	0.00592015	2.3807	100		

Рис. 10.2. Количественные параметры ординации СА: 1) значения *column scores* (вверху), 2) собственные значения и % объясненной дисперсии (внизу)

Вкладка *Relay plot*

Данный график визуализирует распределение видов рыб вдоль первой оси ординации, интерпретируемой как главный градиент структуры данных (рис. 10.3). По вертикальной оси *Relay index* (от 0 до ~105) отложено относительное положение проб вдоль *Axis 1* анализа соответствий; шкала представляет собой нормированное ранжирование координат проб, приблизительно приведённое к диапазону 0–100, и не выражается напрямую в единицах стандартного отклонения, как в DCA.

Каждый вид представлен вертикальной чёрной линией рядом с его названием; эта линия центрирована в точке его оптимума по главной оси ординации. Горизонтальные отрезки («леденцы»), отходящие от вертикальной линии вправо, соответствуют отдельным пробам, в которых данный вид был обнаружен. Вертикальное положение

«леденца» показывает, где по градиенту расположена соответствующая проба (её координата по *Axis 1*), а длина отрезка пропорциональна относительному обилию вида в этой пробе.

Распределение «леденцов» вдоль вертикальной линии одного вида отражает его экологическую амплитуду: узкий вертикальный диапазон указывает на стенобионтность и приуроченность к определённому диапазону условий, тогда как широкий разброс свидетельствует об эврибионтности и способности вида обитать в широком спектре состояний среды. В рассматриваемом примере визуальный диапазон *Relay index* составляет около 105 единиц ($105/100 \approx 1.05$ в пересчёте на исходную шкалу оси), что указывает на умеренную длину градиента в пределах отображённой части ординационного пространства.



Рис. 10.3. График *Relay plot*

Философское эхо

Метод СА — прецизионный эхолот, посылающий локационные импульсы в глубины таблицы частот и распознающий скрытые паттерны по их резонансному отклику. Его сила — в умении улавливать слабые сигналы ассоциаций, подобные отражениям от редких выступов на абиссальной равнине сквозь акустический шум случайных совпадений. Он не просто фиксирует форму рельефа, но и перенастраивает чувствительность приёмника, усиливая значимость редких, но информативных сигналов и приглушая фоновые доминирующие

шумы, тем самым меняя акустическую картину среды. Он преобразует сухую матрицу данных в карту подводного рельефа: станции и виды фиксируют позиции, а их взаимное расположение формирует изрезанный профиль дна сообщества, где расстояния не измеряют силу связи напрямую, а задают направление и плотность соответствий. Однако когда данные, подобно илистым осадкам, теряют чёткую частотную структуру или растворяются в непрерывных градиентах, сигнал рассеивается и искажается, рождая призрачные отражения. Его проникающая способность раскрывается лишь в прозрачной среде таблиц сопряжённости, где становятся различимы даже небольшие перепады глубин; в иных условиях метод может создавать иллюзорный рельеф, принимая шум за форму. Особенно коварен эффект «подковы», подобно термоклину сворачивающий градиент и искажающий геометрию пространства. В руках опытного гидроакустика СА становится инструментом точной навигации, прокладывающим курс сквозь многомерную толщу данных, где каждая станция ложится на чарт как опорная точка, каждый вид звучит акустическим репером, а их взаимное положение вместе вычерчивает батиметрический план скрытых сообществ.

11. ДЕТРЕНДРОВАННЫЙ АНАЛИЗ СООТВЕТСТВИЙ (DCA)

Введение

Во второй половине XX века экология сообществ столкнулась с быстрым ростом объёмов данных о видовом составе и распределении организмов, что потребовало методов визуализации многомерных структур и выявления скрытых экологических градиентов. Анализ соответствий (СА) стал одним из ключевых инструментов, позволяя проецировать сложные взаимосвязи между видами и пробами в пространство пониженной размерности. Однако применение СА сопровождалось характерным артефактом — «эффектом дуги» (*arch effect*): ординационные диаграммы приобретали подковообразную форму, в которой объекты на противоположных концах градиента визуально сближались, а центральная часть казалась

растянутой, что искажало реальные экологические расстояния и затрудняло интерпретацию осей в терминах факторов среды.

Для устранения этого искажения Марк Хилл и Хью Гаух предложили детрендированный анализ соответствий DCA (*Detrended Correspondence Analysis*) (Hill & Gauch, 1980). DCA представляет собой модификацию СА, основанную на процедуре детрендинга — устранения систематической кривизны, индуцированной первой осью на последующие, и последующем рескейлинге конфигурации. В результате ординационное пространство становится более «выпрямленным», а оси — лучше соотносимыми с реальными градиентами среды (температура, солёность, трофность, уровень нарушений и др.). Важным нововведением DCA стала калибровка осей в единицах стандартных отклонений (SD) смены видового состава: значения <2 SD интерпретируются как слабый градиент, $2-4$ SD — как умеренный, >4 SD — как длинный градиент с почти полной сменой видов. Таким образом, DCA предоставляет не только визуальное представление, но и стандартизированную метрику для количественной оценки и сравнения экологических градиентов. Несмотря на критику отдельных аспектов (искусственности детрендинга и выбора параметров), метод занял устойчивое место в экологическом анализе и широко применяется для изучения структуры сообществ вдоль абиотических и биотических градиентов.

Методология

Детрендированный анализ соответствий (DCA) опирается на математический аппарат анализа соответствий (СА) и реализуется как его поэтапная модификация. На первом шаге исходная таблица «виды $\times$ пробы» анализируется методом СА: в метрике χ^2 -расстояния выделяется последовательность ортогональных осей, каждая из которых объясняет максимально возможную часть остаточной инерции, а χ^2 -метрика усиливает вклад редких видов и нормирует различия между пробами с разной суммарной численностью. При унимодальных ответах видов на градиенты среды сочетание χ^2 -метрики и нелинейных откликов приводит к тому, что одномерный градиент «загибается» на второй оси, формируя подковообразную конфигурацию — классический эффект дуги.

DCA не заменяет СА, а добавляет к нему корректирующий этап. Ключевая операция детрендинга состоит в удалении тренда координат по следующей оси относительно предыдущей. Диапазон значений

первой оси СА разбивается на сегменты; внутри каждого сегмента для всех проб вычисляется среднее значение координат по второй оси, после чего координаты по второй оси локально центрируются так, чтобы внутри сегмента среднее стало равно нулю. Эта процедура повторяется вдоль всей первой оси и устраняет систематическую кривизну второй оси; затем может применяться сглаживание, чтобы избежать резких переходов между сегментами. Аналогичные операции могут применяться к последующим осям: координаты по третьей оси корректируются относительно положения точек в пространстве первых двух осей (возможен регрессионный подход с вычитанием предсказанных значений), чтобы каждая новая ось описывала вариацию, максимально независимую от уже выделенных направлений.

Важным следствием детрендинга и последующего рескейлинга является изменение интерпретации масштаба осей. В отличие от СА, где *eigenvalues* напрямую связаны с долей объяснённой инерции, в DCA оси калибруются в единицах стандартных отклонений смены видового состава: расстояние в 1 SD вдоль оси примерно соответствует такому изменению в видовом составе, при котором один «средний» вид заменяется другим. Градиент длиной порядка 4 SD рассматривается как почти полная смена видов, тогда как 1 SD указывает на слабую дифференциацию. Эта метризация осей — одно из ключевых практических достоинств DCA: оси можно трактовать как приближённо линейные меры экологических расстояний и использовать длину градиента для сопоставления различных исследований. В то же время DCA остаётся постобработкой СА: его операции сегментации, локального центрирования и перекалибровки носят характер методологических допущений, а не строго выведенных статистических процедур, что служит источником критики. Тем не менее именно эта целенаправленная корректировка эффекта дуги сделала DCA одним из наиболее употребляемых методов непрямой ординации для протяжённых градиентов.

Характер данных

Эффективность применения DCA в экологии сообществ во многом определяется соответствием структуры данных его ключевым предпосылкам. Как развитие СА, DCA ориентирован на матрицы «виды × пробы» с неотрицательными значениями обилия (численность, биомасса, покров), где интерпретация χ^2 -метрики и унимодальных откликов видов вдоль градиентов среды уместна. Метод наследует

чувствительность СА к редким видам, которые могут непропорционально влиять на результаты ординации, поэтому на практике часто используют предварительную фильтрацию редких таксонов или мягкие преобразования (например, логарифмирование или преобразование Хеллингера) для стабилизации структуры данных.

Наибольшую информативность DCA демонстрирует при анализе данных, отражающих протяжённые, приблизительно одномерные экологические градиенты, вдоль которых происходит последовательная, унимодальная смена видового состава. Классическими примерами являются эстуарные системы с градиентом солёности, вертикальная стратификация водоёмов (глубина, температура, освещённость, кислород), крупные речные системы с континуумом от истока к устью (изменение скорости течения, типа субстрата, трофности, антропогенной нагрузки), а также градиенты загрязнения, где наблюдается закономерная замена чувствительных видов толерантными. В таких условиях DCA устраняет дугообразную конфигурацию СА, выявляет основной тренд и позволяет количественно оценить протяжённость градиента в единицах SD.

Ограничения DCA проявляются там, где данные не формируют выраженного линейного (или квазILINEЙНОГО) градиента. В системах изолированных озёр, не образующих экологического континуума, ординация с помощью DCA, как и СА, может не выявить интерпретируемой структуры или навязать искусственный «градиент», отражающий лишь комбинацию дискретных различий. В мозаичных местообитаниях, таких как коралловые рифы или сложные донные ландшафты, где распределение видов определяется наложением множества локальных факторов без доминирующего направления, DCA может сгладить важные локальные паттерны и привести к излишне упрощённой картине. Дополнительные трудности возникают при анализе циклических или резко нелинейных процессов (например, сезонной динамики планктонных сообществ), где истинная структура ближе к замкнутым траекториям, чем к протяжённым градиентам: попытка «выпрямить» такие данные с помощью DCA способна исказить циклы и приводить к неверным выводам о направлении и масштабе изменений. Поэтому корректное применение DCA требует предварительной оценки длины градиента и осмысления экологической природы данных: при коротких градиентах (<3 SD) предпочтительны линейные методы (PCA, RDA), тогда как для длинных градиентов (>4 SD) DCA остаётся одним из наиболее интерпретируемых инструментов непрямой ординации.

Пример анализа данных

Задача. В пяти озёрах с разной степенью антропогенного загрязнения получены оценки количественно-видового состава ихтиофауны — пескарь, карась, окунь, плотва, лещ, ротан, укля — на основании стандартных уловов с учётом селективности орудия по отношению к разным видам. Используя тест DCA, оценить различия структуры ихтиоценозов в разных озёрах.

АНАЛИЗ ДАННЫХ

Детрендрированный анализ соответствия (DCA) в PAST расположен в главном меню *Multivariate* → *Ordination* → *DCA*. Тест DCA представляет собой дальнейшее развитие классического анализа соответствий (CA), сохраняя его базовые принципы ординации и графического представления данных. Поэтому фундаментальные концепции и подходы к интерпретации результатов, применимые к CA, в большинстве случаев актуальны и для DCA. Функционал для DCA в PAST версии 4.17с несколько беднее, чем для CA, и представлен лишь *raw scores* и *column scores*, которые здесь для наглядности даны в масштабе 1:100; соответственно, для оценки величины градиента в единицах SD необходимо вернуться к истинному масштабу, разделив значения координат на 100 (см. главу «CA»). Оценку величины экологического градиента, как и в случае методов непрямого градиентного анализа, требуется проводить с использованием ординационного теста DCA (см. раздел «CA»). В данном случае оценка градиента по оси 1 составляет ~3.3 SD, что соответствует переходной зоне (3–4 SD) и при ожидаемой нелинейности видовых откликов делает применение DCA методически оправданным.

Основной функционал теста размещён на вкладке *Scatter plot* и представлен: 1) *Eigenvalues*, величины которых характеризуют относительную важность осей и приближённо пропорциональны их вкладу в инерцию; 2) опцией *Segments* (см. ниже); и 3) «флажком» с надписью *Detrending*, со снятием которого DCA (рис. 11.1) фактически превращается в CA (рис. 11.2), что даёт возможность применять к одному набору данных два теста и сравнивать их результаты, не выходя из меню DCA. К сожалению, в функционал DCA не входит автоматический расчёт доли объяснённой инерции, но её можно приблизительно оценить из величин *Eigenvalues*: например, для оси 1 эта доля составит

$0.606 / (0.606 + 0.057 + 0.036 + 0.019) \approx 84.4\%$, а для оси 2 — $0.057 / (0.606 + 0.057 + 0.036 + 0.019) \approx 7.9\%$, что позволяет судить о доминирующей роли первой оси в структуре данных.

Параметр *Segments* определяет, на какое количество равных частей разбивается первая ось DCA (*Axis 1*) при выполнении процедуры детрендинга; границы этих сегментов одновременно отображаются в виде фоновой сетки на графике ординации. Основная цель такого разбиения заключается в локальном выравнивании тренда второй оси внутри каждого сегмента, что устраняет эффект дуги и в то же время помогает визуализировать плотность распределения объектов на плоскости анализа. Менять значение *Segments* стоит тогда, когда текущая визуализация структуры на графике DCA недостаточно информативна: например, точки образуют сплошное «месиво» без чётких кластеров или картина выглядит излишне «пятнистой». Изменением величины данного параметра добиваются подбора оптимального масштаба детрендинга и отображения плотности данных, что является аналогом настройки фокуса камеры и делает структуру данных максимально ясной, не затрагивая при этом сами результаты расчёта ординационных осей.

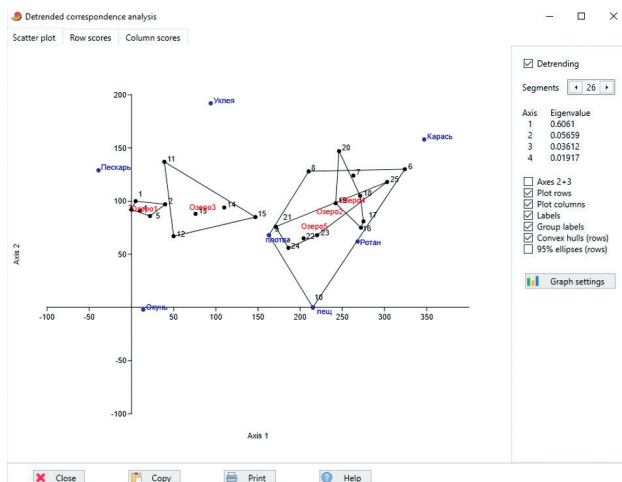


Рис. 11.1. Ординационная диаграмма детрендированного анализа соответствий (DCA)

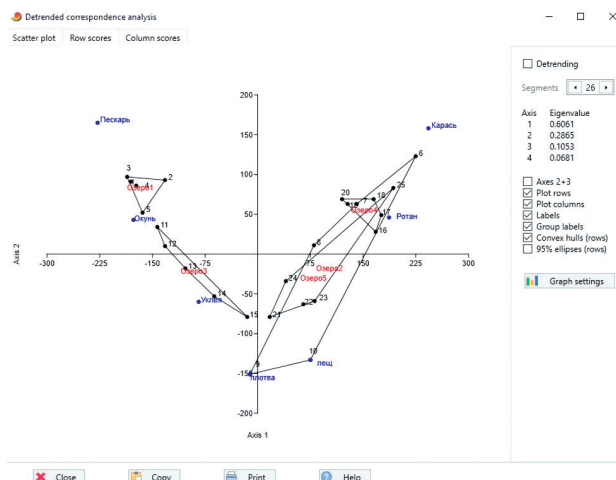


Рис. 11.2. Эффект «дуги» при использовании анализа соответствий (CA) путём снятия «флажка» *Detrending* в правом верхнем углу графика; устраняется использованием DCA путем активации «флажка» *Detrending* (см. рисунок выше)

Интерпретация ординационной диаграммы

На ординационной диаграмме DCA (рис. 11.1) чётко визуализируется непрерывный экологический градиент вдоль первой оси, аккумулирующей основную долю вариации видового состава (~84%). Облака проб выстраиваются в линейную последовательность, отражающую нарастание антропогенной трансформации: от изолированного положения Озера 1 в крайней левой части (фоновые, олиготрофные условия) через переходную зону Озера 3 к слитному массиву Озёр 5, 2 и 4 в правой полуплоскости. Резкая обособленность Озера 1 и отсутствие перекрытия с остальными локациями подтверждает сохранение эталонной структуры ихтиоценоза с доминированием стенобионтных видов (пескарь, окунь). Остальные водоёмы формируют единый континуум, где выпуклые оболочки Озёр 2, 4 и 5 демонстрируют значительное геометрическое слияние, свидетельствующее о конвергентной перестройке сообществ под давлением хронического антропогенного пресса. Распределение видов вдоль этой оси строго соответствует их экологической толерантности: чувствительные таксоны маркируют левый край, умеренно толерантная плотва и уклейка занимают центральные позиции, а рези-

стентные лещ, ротан и карась концентрируются в зоне максимальных значений, формируя ядро трансформированных экосистем.

Геометрия выпуклых оболочек детализирует внутреннюю структуру правого кластера: Озёра 2, 4 и 5 образуют плотный функциональный массив, при этом координаты проб Озера 4 полностью вложены в диапазоны Озёр 2 и 5, что указывает на их глубокую экосистемную гомогенизацию. Векторные привязки, подтверждённые таблицами координат строк и столбцов (рис. 11.3), показывают, что видовое ядро этой группы формируют плотва и лещ, тяготеющие к периферии Озера 5, тогда как ротан и карась смещены к правому флангу, маркируя наиболее деградированные условия Озера 4. Вертикальная ось (*Axis* 2, ~7,9% инерции) выступает независимым стратификационным градиентом, разделяющим сообщества по предпочитаемым горизонтам обитания: верхняя область графика занята пелагофильными формами (уклейка, карась, пескарь), а нижняя — бентофагами и придонными видами (лещ, ротан, окунь). Тем не менее доминирующим сигналом ординации остаётся горизонтальное разделение по уровню нарушения среды. Слияние Озёр 2, 4 и 5 в единый кластер наглядно демонстрирует, что при превышении порога устойчивости разнородные по генезису нагрузки водоёма теряют исходную уникальность и переходят в унифицированное состояние, управляемое общим набором абиотических лимитаторов и доминированием эврибионтов.

Представленная DCA-конфигурация реконструирует траекторию сукцессионного замещения чувствительных специалистов устойчивыми генералистами вдоль градиента антропогенного пресса. Значительное перекрытие облаков Озёр 2, 4 и 5 отражает фундаментальный экологический принцип: при достижении критического уровня стресса разнородные экосистемы нивелируют исходные различия, выравнивая ключевые физико-химические параметры и формируя унифицированные «суперниши». В таких условиях адаптивный потенциал видов-универсалов подавляет локальное биоразнообразие, а ихтиоценозы приобретают конвергентные черты независимо от типа стоков (промышленные, городские или агрогенные). Таким образом, ординация не только диагностирует текущий статус водоёмов, но и служит прогностическим инструментом, визуализируя необратимость структурных изменений при сохранении существующего режима нагрузки.

Detrended correspondence analysis

Scatter plot	Row scores	Column scores	
	Axis 1	Axis 2	Axis 3
Озеро1	5	100	11
Озеро1	40	97	25
Озеро1	0	92	41
Озеро1	10	91	60
Озеро1	22	86	86
Озеро2	324	130	96
Озеро2	263	124	96
Озеро2	210	128	59
Озеро2	163	68	61
Озеро2	215	0	0
Озеро3	39	137	40
Озеро3	50	67	113
Озеро3	76	88	97
Озеро3	110	94	82
Озеро3	147	85	63
Озеро4	272	75	32
Озеро4	275	81	30
Озеро4	271	105	63
Озеро4	242	98	40
Озеро4	246	147	101
Озеро5	171	76	51
Озеро5	204	65	36
Озеро5	220	68	46
Озеро5	186	56	57
Озеро5	303	118	88

Detrended correspondence analysis

Scatter plot	Row scores	Column scores	
	Axis 1	Axis 2	Axis 3
Пескарь	-39	129	-41
Карась	347	158	134
Окунь	14	-2	211
плотва	163	68	61
лещ	215	0	0
Ротан	268	62	-12
Уклея	94	192	92

Рис. 11.3. Таблицы координат DCA: положения проб (*Row scores*) (слева), видовые координаты (*Column scores*), определяющие экологическую приуроченность таксонов (справа)

Философское эхо

DCA преподносит экологу дар точных мер: метод переводит абстрактные оси СА в рабочий эталон экологического расстояния — шкалу в единицах стандартного отклонения. Ось DCA перестаёт быть просто указателем направления и становится отградуированной шкалой, позволяющей соизмерять протяжённость средовых градиентов в единой метрике и сопоставлять их в различных анализах. Подобно своему предшественнику — СА, — DCA остаётся живым инструментом с характером. Его преобразующая сила выпрямления геометрии данных проявляется наиболее полно там, где отклики образуют мощный, непрерывный поток, организованный вдоль экологического градиента: здесь он действует как калибровщик измерительных шкал, приводя сложную геометрию облаков точек к линейной и измеримой форме. Однако в мозаике изолированных озёр, в хаотичном узоре рифовых сообществ или в цикличности экологических процессов попытка натянуть всё на единую мерную линейку DCA неизбежно искажает естественный рисунок распределений, сглаживая локальные неоднородности и навязывая им искусственную линейность. В таких условиях

детрендинг перестаёт быть аккуратной калибровкой и становится источником дополнительной деформации, перераспределяя реальные различия и придавая им несуществующую форму. Метод аккуратно выпрямляет дугу «подковы» СА, снимая геометрическое искажение и возвращая структуре данных ясность, однако не выводит абсолютную линейку экологических дистанций — лишь калибрует рабочий эталон последовательной смены видов. Выбирая DCA, исследователь работает не с окончательной точностью, а с тщательно откалиброванной мерой, с помощью которой можно сравнивать, насколько далеко друг от друга расположены сообщества вдоль единого градиента. В руках биолога-метролога DCA превращается в инструмент тонкой калибровки экологического пространства, позволяющий проследить, как виды последовательно сменяют друг друга вдоль выпрямленной оси, формируя упорядоченный ряд. Это не статическая карта, а динамическая картина потока, в котором виды входят в градиент и покидают его, а каждое стандартное отклонение становится надёжным шагом на пути к количественному пониманию смены доминант.

12. КАНОНИЧЕСКИЙ АНАЛИЗ СООТВЕТСТВИЙ (ССА)

Введение

Канонический анализ соответствий ССА (*Canonical Correspondence Analysis*) является ключевым методом прямого ограниченного (*constrained*) градиентного анализа в экологии сообществ, предназначенным для выявления и количественной оценки связи между структурой сообществ и факторами среды. Принцип «ограниченности» заключается в том, что ординация проб выполняется не в пространстве всей вариации данных, а строго в подпространстве, заданном линейными комбинациями выбранных исследователем переменных среды. Таким образом, метод фокусируется на той части изменчивости видового состава, которая объясняется заданными предикторами, и отсекает влияние неучтённых факторов.

В отличие от базового анализа соответствий (СА), который описывает лишь внутреннюю структуру матрицы «пробы–виды», ССА

напрямую связывает вариацию сообществ с вариацией измеренных параметров среды, что придаёт результатам направленный и прикладной характер. Метод был разработан голландским экологом математиком К. Й. Ф. тер Брааком (Cajo J.F. ter Braak) и представлен в 1986 году как развитие СА для экологических приложений (ter Braak, 1986). Термин «канонический» указывает на то, что анализ проводится в жёстко заданном пространстве объясняющих переменных: оси ординации конструируются как линейные комбинации экологических факторов и поэтому интерпретируются как градиенты среды.

Таким образом, ССА служит важным инструментом в ихтиологии для выявления того, какие экологические факторы (глубина, температура, скорость течения, тип субстрата, уровень загрязнения) в наибольшей степени детерминируют распределение видов рыб. Метод позволяет визуализировать специфические реакции видов на градиенты условий и идентифицировать зоны со специфическими типами сообществ, что особенно ценно для оценки антропогенных воздействий и мониторинга качества водных экосистем.

Методология

В основе ССА лежит задача согласованного анализа двух матриц: матрицы обилия видов («пробы × виды») и матрицы факторов среды («пробы × переменные»). Если стандартный СА выявляет главные направления внутренней изменчивости сообщества, то ССА целенаправленно ищет те направления, которые одновременно отражают вариацию сообщества и максимально связаны с изменениями внешних факторов.

Методологически ССА формирует синтетические канонические оси как взвешенные линейные комбинации исходных переменных среды. Параллельно строятся соответствующие канонические оси для сообщества, отражающие вариацию видового состава в ограниченном пространстве, заданном этими комбинациями. Пары осей среды и сообщества подбираются по строгому критерию: максимизации зависимости (канонической связи) между синтетической осью среды и синтетической осью сообщества. Первая пара осей строится так, чтобы значения вдоль канонической оси среды наилучшим образом объясняли вариацию вдоль оси сообщества; последующие пары ищутся по тому же принципу, но при обязательной ортогональности к предыдущим, описывая новые, независимые аспекты влияния среды.

Технически ССА можно рассматривать как проекцию облака точек СА в подпространство, определяемое факторами среды: из всей вариации сообщества выделяется только та часть, которая может быть объяснена заданными переменными, тогда как остаточная (необъяснённая) вариация остаётся вне канонического пространства. Проще говоря, ССА «фильтрует» многомерное пространство данных, выделяя градиенты, вдоль которых распределение видов сильнее всего обусловлено измеренными условиями среды, и игнорируя изменчивость, не связанную с ними.

Графически результат представляется в виде ординационной диаграммы (биплота или триплота), визуально похожей на график СА, но с принципиально иной логикой интерпретации. На ней отображаются: 1) точки проб — положение сообществ в пространстве канонических осей; 2) точки видов — их оптимумы и центры тяжести относительно градиентов среды; 3) векторы факторов среды — направление увеличения переменных и относительная сила их влияния. Близость вида к направлению вектора фактора указывает на положительную связь его обилия с увеличением этого фактора; виды, расположенные в противоположную сторону, ассоциируются с низкими значениями переменной. Длина вектора отражает относительную силу влияния фактора на структуру сообщества вдоль осей, а углы между векторами интерпретируются как корреляции между самими переменными (острый угол — положительная, тупой — отрицательная, близкий к прямому — слабая или отсутствующая). При интерпретации важно учитывать выбранный тип масштабирования (*scaling*) и помнить, что главный смысл несут направления и относительное положение объектов, а не абсолютные расстояния.

Характер данных

Канонический анализ соответствий (ССА) применяется к данным, представленным двумя матрицами: «пробы–виды» и «пробы–факторы среды». Первая содержит показатели обилия (численность, биомасса, встречаемость), вторая — экологически значимые переменные: физико-химические (температура, кислород, солёность, pH), гидрологические (глубина, скорость течения, тип субстрата), морфометрические, трофические и антропогенные индикаторы (загрязняющие вещества, степень зарегулированности реки). Факторы могут быть количественными или категориальными (тип биотопа, сезон),

но их набор должен быть осмысленно подобран, чтобы минимизировать статистический шум и лишнюю коррелированность.

ССА наиболее эффективен при анализе данных, отражающих градиенты обилия видов вдоль изменений условий среды. Метод устойчив к умеренному количеству нулей, но чрезмерная разреженность (много редких видов и пустых проб) может скрывать структуру и усиливать влияние случайных событий. Как и СА, ССА чувствителен к доминирующим видам и редким таксонам: сверхмассовые виды могут «перетягивать» ординацию, а единичные появления — исказить расположение проб. Поэтому часто используют предварительные преобразования обилий (логарифмирование, преобразование Хеллингера) и фильтрацию ультраредких видов для повышения интерпретируемости результатов.

Ключевое допущение ССА связано с унимодальной моделью отклика: предполагается, что виды демонстрируют колоколообразные (унимодальные) отклики вдоль скрытых градиентных осей, а сами оси задаются как линейные комбинации факторов среды. Это делает метод особенно подходящим для анализа данных с длинными градиентами (обычно >4 SD по DCA), где виды успевают пройти через зоны оптимума и спада. Важным ограничением является мультиколлинеарность: если факторы среды сильно коррелируют между собой (например, глубина и температура в стратифицированном озере), ССА корректно выделит общий градиент, но не позволит однозначно разделить индивидуальные вклады переменных. В таких случаях рекомендуется предварительный отбор предикторов или использование процедур пошагового включения (*forward selection*).

ССА демонстрирует высокую эффективность в системах с выраженными градиентами. В водотоке с нарастающим загрязнением от верховьев к устью метод выявит градиент, связывающий уменьшение численности чувствительных видов и увеличение доли толерантных с ростом концентраций тяжёлых металлов, органики или снижением кислорода, позволяя картировать зоны экологического неблагополучия. В цепочке озёр от олиготрофных до эвтрофных ССА покажет смену доминирующих комплексов и свяжет её с градиентами прозрачности, содержания фосфора и кислородного режима у дна. Анализ траловых съёмок на континентальном склоне с помощью ССА продемонстрирует, как с увеличением глубины и снижением температуры меняется количественно-видовой состав ихтиоценозов.

В то же время есть ситуации, в которых ССА работает хуже или даёт труднопонятные результаты. Пороговые эффекты (например, резкая граница пресной и солоноватой воды в эстуарии) приводят к ступенчатой смене сообществ; линейные комбинации факторов в ССА не всегда адекватно описывают такие резкие переходы, и структура может выглядеть «смазанной». Сильно разреженные данные с большим числом редких видов делают метод гиперчувствительным к шуму. Мощная мультиколлинеарность (например, включение множества тесно коррелированных показателей загрязнения) может породить статистически значимый, но экологически неинтерпретируемый результат, не позволяющий выделить конкретные ключевые токсиканты.

Таким образом, ССА является специализированным методом объяснительной ординации. Его сила заключается в выявлении и визуализации унимодальных/монотонных связей между комплексом факторов среды и структурой сообществ в континуальных градиентах (реки, эстуарии, батиметрические разрезы). Он превосходит СА по объяснительной способности и наглядности связей «среда–сообщество», но требует от исследователя критической оценки структуры данных, проверки предпосылок (длина градиента, коллинеарность, редкие виды) и готовности применять альтернативные методы (например, RDA для более коротких градиентов), когда допущения ССА нарушены.

Пример анализа данных

Задача. Применить канонический анализ соответствий (ССА) для плотностных параметров видов рыб, экз./м² (окунь, плотва, лещ, густера, ротан, карась, жерех, язь, щука, судак), полученных из пяти озёр, для которых оценены следующие абиотические параметры: площадь (км²), средняя глубина (м), прозрачность по диску Секки (м), температура (°C), растворенный кислород у дна (мг/л). Указать экологические закономерности, отражающие выявленные взаимоотношения, и оценить статистическую значимость осей и долю объяснённой дисперсии.

АНАЛИЗ ДАННЫХ

Подготовка данных к анализу начинается со стандартизации. Данные для методов прямого градиентного анализа, к которым относится ССА, включают как биотические, так и абиотические параметры. Для абиотических переменных обязательна Z-стандартизация (приведение

к нулевому среднему и единичному стандартному отклонению), чтобы переменные с разными единицами измерения и масштабами вносили сравнимый вклад в модель. Для биотических характеристик факультативно применяется преобразование Хеллингера: оно особенно полезно, когда в матрице видов много редких таксонов и нулей, так как снижает влияние доминирующих видов, придаёт редким таксонам умеренный вес и делает оси ССА более чувствительными к экологическим градиентам, не искажая структуру двойных нулей.

Канонический анализ соответствий (ССА) в PAST расположен в главном меню *Multivariate* → *Ordination* → *CCA*. После копирования данных в рабочий лист PAST и выбора теста ССА появляется окно, где требуется вручную указать количество абиотических факторов; в текущей реализации PAST их число не может превышать шести. Базовой является вкладка *Scatter plot*, отображающая ключевой результат ССА — триплот, то есть график, на котором в единой системе координат показаны: 1) позиции отдельных озёр (проб), 2) позиции видов (переменных отклика, в данном случае — плотностные характеристики), и 3) векторы факторов среды. Основными результатами анализа являются диаграмма рассеяния (триплот) на вкладке *Scatter plot* (рис. 12.2), вкладка *Eigenvalues* с процентом объяснённой вариации (% of total) по каноническим осям (рис. 12.1, слева) и вкладка *Permutation*, содержащая результаты пермутационных тестов значимости отдельных осей (рис. 12.1, справа).

Проверку длины экологического градиента, как и в случае методов непрямой ординации, следует проводить с использованием DCA (см. раздел «СА»). Для анализируемых данных длина градиента по первой оси DCA составляет ~4.2 SD, что указывает на протяжённый градиент и делает применение ССА (как унимодального, ограниченного метода) методически обоснованным. Важно подчеркнуть, что длина экологического градиента оценивается по биотическим данным (матрице видов), а не по самим факторам среды: именно структура видового состава несёт информацию о том, насколько «длинным» является градиент условий, и какой класс ординационных методов подходит лучше.

Объясненная дисперсия и значимость осей

Первые две канонические оси ССА объясняют подавляющую часть значимой изменчивости в данных. Ось 1 — абсолютный доминант: она аккумулирует 77.8% от общей инерции (изменчивости) системы; Ось 2 также статистически значима, объясняя дополнительно 15.6%

общей инерции (Рис. 12.1,верху) (*Пояснение.* В ССА инерция является аналогом дисперсии, но учитывает «вес» — массу — каждой точки, что заимствовано из концепции момента инерции в физике). В сумме эти две оси охватывают 93.4% всей структурной изменчивости, что свидетельствует об исключительной адекватности двумерной модели и надёжности интерпретации пространства. Общая значимость модели подтверждается значением $Trace = 0.2838$ с $p = 0.001$ (вкладка *Permutation*, верхняя правая часть окна) (*Пояснение.* $Trace$ представляет собой суммарную инерцию всех канонических осей — интегральную меру объяснённой связи между всем комплексом факторов среды и структурой сообщества). Статистическая значимость отдельных осей также подтверждена перестановочным тестом (*permutation test*, $p = 0.001$ для каждой оси — вкладка *Permutation*), что свидетельствует о том, что все выделенные оси достоверно объясняют связь между видами и факторами среды и исключает случайность выявленных закономерностей (Рис. 12.1, внизу). Третья и последующие оси объясняют лишь незначительные доли инерции (менее 3% каждая) и для визуального анализа могут быть проигнорированы.

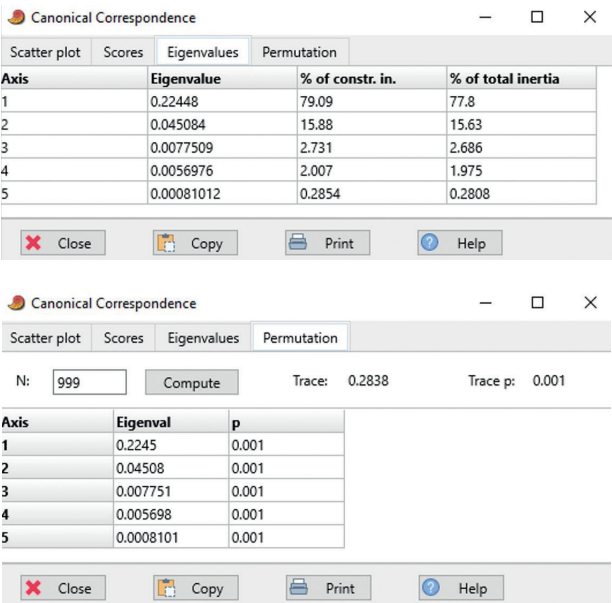


Рис. 12.1. Собственные числа и доля объяснённой дисперсии (инерции) (вверху); результаты перестановочного теста (внизу)

Интерпретация осей и общая структура графика

Проведенный анализ позволил выявить пространственные закономерности во взаимосвязи между составом ихтиофауны в ряде водоёмов и ключевыми морфометрическими и гидрофизическими параметрами этих озёр. Анализ триплота (Рис. 12.2), основанный на координатах объектов (озёр), видов (рыб) и факторов среды, раскрывает структуру этих взаимосвязей. Расположение всех объектов на плоскости первых двух осей выявило чёткую экологическую структуру. Ось 1 (77.8% инерции) интерпретируется как главный градиент «температура–глубина/прозрачность/кислород». Положительные значения ассоциированы с хорошо прогреваемыми, вероятно, более эвтрофными водоёмами (озёра 3, 4, 5), характеризующимися высокой температурой воды (координата вектора +0.96). Отрицательные значения оси — с большими, глубокими, холодными, прозрачными и богатыми кислородом водоёмами (озёра 1, 2), что типично для олиготрофных условий (векторы глубины -0.96 , прозрачности -0.99 и кислорода -0.99 направлены влево). Ось 2 (15.63% инерции) выступает как второстепенный градиент, дополнительно разделяющий озёра по их морфометрии. Вектор площади озера (Площадь_км²) имеет координаты (0.36; -0.87) и направлен вниз, что позволяет интерпретировать эту ось как градиент размера водоёма: положительные значения (Озеро 1) соответствуют малым акваториям, а отрицательные (Озёра 2, 3, 4) — крупным.

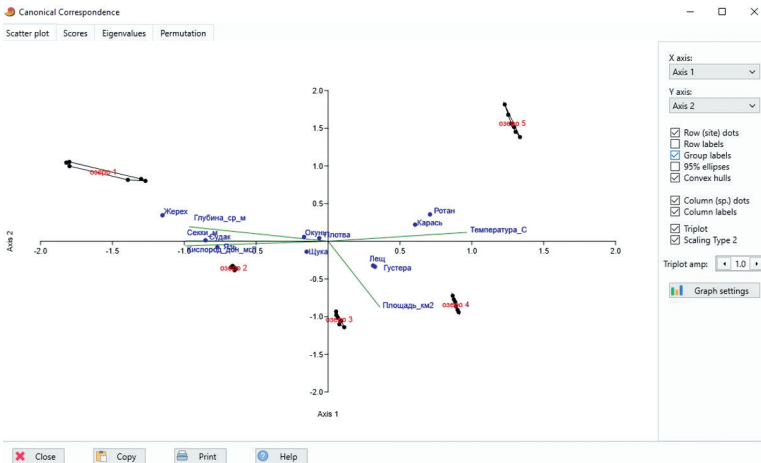


Рис. 12.2. Ординационная диаграмма — триплот — канонического анализа соответствий (CCA)

В правом верхнем квадранте (Квадрант I) расположено Озеро 5, образующее компактный, сильно обособленный кластер в области высоких положительных значений Оси 1 (~ 1.3) и Оси 2 (~ 1.5). Его уникальная позиция резко контрастирует со всеми другими озёрами. Позицию этого озера определяет самый сильный драйвер в системе — температура воды. Это чётко указывает на то, что Озеро 5 является самым теплым из всех исследованных и, вероятно, мелководным, так как находится в противоположной стороне от вектора глубины. Данный водоём населяют ротан (0.71; 0.36) и карась (0.61; 0.22) — виды, толерантные к высоким температурам и характерные для малых, сильно прогреваемых, часто эвтрофных водоёмов. Их тесная группировка с вектором температуры подтверждает сильную положительную связь с этим фактором.

В левом верхнем Квадранте II Озеро 1 формирует чёткий кластер в области максимально отрицательных значений Оси 1 (~ -1.6) и положительных значений Оси 2 (~ 0.9). Позиция этого озера определяется комплексом взаимосвязанных параметров, указывающих на значительную глубину, высокую прозрачность и кислородный режим. Вектор глубины имеет очень острый угол с отрицательной частью Оси 1, что означает его максимальную важность для объяснения вариативности по этой оси. Вектор прозрачности (по Секки) почти коллинеарен оси X (направлен влево, -0.99), являясь почти идеальным индикатором градиента Оси 1. К данному типу озёр (глубоким, прозрачным, но малым по площади) приурочены жерех (-1.15 ; 0.34) и судак (-0.85 ; 0.01) — типичные пелагические хищники, требующие высокого содержания кислорода. Их удаленность от центра указывает на высокую специфичность предпочтений. Окунь (-0.17 ; 0.06) и плотва (-0.06 ; 0.04), будучи более эвритопными видами, расположены ближе к центру графика, что говорит об их высокой экологической валентности, но все же с тяготением к данному типу условий.

В левом нижнем Квадранте III Озеро 2 занимает область отрицательных значений по Оси 1 (~ -0.65) и отрицательных значений по Оси 2 (~ -0.35). Важным уточнением является то, что, находясь в отрицательной зоне Оси 1, это озеро является глубоким (вдоль вектора глубины), а не мелководным. Отрицательное значение по Оси 2 указывает на его большую площадь (вдоль вектора площади). С этим озером пространственно ассоциируется язь (-0.77 ; -0.08) и щука (-0.15 ; -0.14). Щука расположена близко к центру координат, что указывает

на её широкую экологическую валентность, однако её смещение в отрицательную область Оси 1 и 2 связывает её с крупными, глубокими водоёмами, богатыми растительностью (прибрежный хищник). Язь же четче привязан к отрицательным значениям Оси 1, разделяя предпочтения среды обитания с судаком.

В правом нижнем квадранте (Квадрант IV) Озера 3 и 4 образуют тесно перекрывающийся кластер в квадранте с положительными значениями по Оси 1 и отрицательными по Оси 2. Озеро 4 (0.88; -0.86) смещено правее, чем Озеро 3 (0.05; -1.05). Ключевым фактором для этой группы является площадь озера, вектор которого направлен вниз. Это указывает на то, что эти озёра, будучи по трофности и температуре (Ось 1) ближе к озеру 5, отличаются от него большими размерами, что модифицирует их термический режим. Данный биотоп населяют лещ (0.31; -0.32) и густера (0.33; -0.34) — типичные бентофаги, предпочитающие водоёмы с развитой зоной глубин и илистыми грунтами, что часто коррелирует с достаточно большой площадью озера и высокой продуктивностью.

Сводная интерпретация сочетаний подтверждает выявленную структуру: Квадрант I (Озеро 5 + Температура + Ротан/Карась) демонстрирует, что высокая температура является ключевым драйвером в малом водоёме. Квадрант II (Озеро 1 + Глубина/Прозрачность + Жерех/Судак) показывает, что комплекс параметров «большая глубина и прозрачность» в условиях малой площади определяет наличие специализированных пелагических хищников. Квадрант III (Озеро 2 + Глубина/Площадь + Язь/Щука) характеризует среду обитания для щуки и язя в крупных, глубоких водоёмах. Квадрант IV (Озера 3/4 + Площадь + Лещ/Густера) подчеркивает, что большая площадь водоёма, сочетающаяся с признаками продуктивности (положительная ось 1), является важным фактором для донных видов-бентофагов.

Философское эхо

Метод ССА выступает штатным медиатором на саммите двух держав данных, где одна сторона диктует условия среды, а другая, представленная делегациями видов, выстраивает свои позиции в строго очерченных границах допустимого. Экологические факторы задают правила дипломатической игры, формируя переговорный коридор и пространство возможных компромиссов; каждый вид занимает нишу, соразмерную степени его соответствия повестке, выстраивая

непрерывный спектр от полного согласия до скрытого отторжения. Суть метода — в прокладывании магистралей согласования, вдоль которых архитектура сообщества наилучшим образом отзывается на диктат условий, превращая разрозненные реакции в читаемую геометрию договорённостей. Первая каноническая ось очерчивает главный переговорный трек — доминирующий градиент, вдоль которого происходит перекройка альянсов и перераспределение видовых квот, тогда как последующие оси вводят ортогональные линии влияния, уточняя расстановку сил и выявляя автономные векторы, где каждый новый уровень отражает самостоятельный механизм параллельного взаимодействия. ССА не фиксирует каждую отдельную реакцию, а чертит карту зон влияния, где решающими становятся направление векторов среды и проекция видов на эти линии притяжения; это не стенограмма переговоров, а динамическая расстановка позиций, на которой ясно видны центры притяжения, буферные зоны нейтралитета и области неизбежного расхождения интересов. Однако устойчивость урегулирования зависит от чистоты входящих условий: если в повестку вводятся перекрывающиеся или избыточные факторы, начинается скрытая конкуренция за рычаги влияния, рождающая иллюзию сложного многостороннего пакта там, где на деле правит один латентный мотив. Метод не выносит дипломатического хаоса — когда реакции видов становятся фрагментарными, а доминирующий градиент размывается, переговорная логика теряет опору, и карта взаимодействий начинает проецировать не реальную архитектуру связей, а лишь их бледный, упрощённый силуэт. Но в руках исследователя, чувствующего ритм экологических трендов, ССА превращает сырые массивы наблюдений в отлаженную систему договорных обязательств, где условия среды задают границы диалога, а биологические сообщества, принимая эти рамки, выстраивают многослойную, но предсказуемую структуру взаимного сосуществования, переводя шум полевых данных в ясную, проверяемую и устойчивую карту градиентных взаимодействий.

13. АНАЛИЗ ИЗБЫТОЧНОСТИ (RDA)

Введение

Анализ избыточности RDA (*Redundancy Analysis*) представляет собой метод ограниченной (*constrained*) ординации, специально разработанный для выявления и количественной оценки связи между структурой биологических сообществ и факторами окружающей среды. Его ключевая особенность заключается в том, что метод фокусируется не на всей наблюдаемой изменчивости, а исключительно на той ее части, которая может быть линейно объяснена заданными исследователем внешними переменными, такими как физико-химические параметры среды или характеристики местообитания. Основная задача RDA, таким образом, сводится к проверке причинно-следственного влияния конкретных факторов на состав и обилие видов, выходя за рамки простого описания данных.

Теоретические основы метода были заложены в работах статистика Чарльза Рао (C. R. Rao) в 1960-х годах. Однако широкую известность и практическую значимость в экологии, особенно в гидробиологии и ихтиологии, RDA приобрел позднее, в 1980–1990-е годы. Существенный вклад в его развитие, популяризацию и адаптацию для экологических задач внесли, прежде всего, Пьер Лежандр (Pierre Legendre) и Луи Лежандр (Louis Legendre). Именно тогда возникла острая потребность экологов в инструментах, способных целенаправленно проверять гипотезы о влиянии измеряемых факторов среды. Будучи методом ограниченной ординации, RDA позволяет оценить, насколько вариации предикторов (температура, глубина, течение, концентрация веществ) определяют закономерности в распределении и обилии видов, обеспечивая формализованный подход к анализу причинно-следственных связей.

Термин «избыточность» в названии метода отражает фундаментальное свойство экологических данных: виды часто демонстрируют согласованные реакции на изменения среды, что приводит к частичному дублированию информации в матрице сообщества. Эта ковариация создаёт избыточность, а RDA количественно выделяет именно ту ее часть, которая объясняется заданными факторами и формирует предсказуемую компоненту общей вариации сообщества. Понимание этой предсказуемой части позволяет не только объяснить текущие закономерности, но и прогнозировать возможные изменения сообщества

при варьировании ключевых факторов. Следовательно, метод служит мощным инструментом для выявления и ранжирования значимости конкретных абиотических или антропогенных факторов в формировании наблюдаемых биотических паттернов, что делает RDA особенно ценным в широком спектре экологических исследований.

Методология

Анализ избыточности (RDA) методологически основывается на анализе главных компонент (PCA). В многомерном пространстве, где каждая точка соответствует пробе с определённым видовым составом, PCA выявляет направления максимальной вариации данных через синтез новых ортогональных осей, представляющих линейные комбинации обилий видов. Этот метод описывает внутреннюю структуру сообществ без учёта внешних факторов. В отличие от PCA, RDA фокусируется на той части общей вариации, которая может быть предсказана с помощью набора внешних переменных — измеренных факторов среды. Математическая основа RDA заключается в последовательности двух статистических операций. На первом этапе выполняется многомерная линейная регрессия, где обилия видов рассматриваются как отклики, а факторы среды — как предикторы. Модель предсказывает ожидаемое обилие каждого вида в каждой точке исключительно на основе значений факторов среды.

Полученные предсказанные значения обилия видов формируют новое облако точек, отражающее только ту часть вариации, которая линейно связана с заданными факторами среды. На втором этапе к этому регрессионному облаку точек применяется процедура ординации, аналогичная PCA. Найденные новые оси, называемые каноническими осями RDA, являются линейными комбинациями факторов среды и объясняют максимально возможную долю вариации в предсказанных данных. Графически результат представляется в виде биплота: точки проб отражают их положение в пространстве предсказанных значений, стрелки факторов среды указывают направление и относительную силу влияния, а точки видов показывают их ассоциацию с определёнными комбинациями условий.

Ключевое математическое отличие RDA от PCA заключается в последовательности анализа и наложении ограничений. RDA сначала трансформирует исходное облако данных через призму линейной зависимости от факторов среды, а затем ординирует получившееся

подмножество данных. В этом смысле RDA можно рассматривать как PCA, выполненный на части данных, объясненной предикторами. Статистическая значимость модели RDA и отдельных канонических осей оценивается с помощью перестановочного тестирования, что позволяет проверить, является ли наблюдаемая структура следствием случайности. Интерпретация осей RDA требует учёта совокупного влияния факторов, которые могут действовать согласованно или противоположно, формируя сложные градиенты условий. RDA выделяет и визуализирует эти градиенты, позволяя количественно оценить и статистически проверить гипотезы о факторах, определяющих структуру сообщества организмов.

Характер данных

Успешное применение анализа избыточности (RDA) требует чёткого понимания его математических предпосылок и природы анализируемых данных. Метод предполагает наличие линейной или близкой к линейной связи между матрицей предикторов (факторы среды) и матрицей отклика (обилие видов). Ключевое допущение заключается в том, что изменения факторов среды вызывают приблизительно пропорциональные изменения в видовых откликах, что наиболее корректно при коротких или умеренных экологических градиентах. Важным условием является также отсутствие выраженной гетероскедастичности, то есть существенного изменения дисперсии ошибок по диапазону предсказанных значений. В отличие от описательных методов, таких как PCA, RDA направлен на проверку гипотез о связи между средой и сообществом, а не только на описание структуры данных.

Типичными примерами эффективного применения RDA в ихтиологии являются задачи, где предполагается монотонное, близкое к линейному влияние среды. К ним относится анализ влияния физико-химических параметров воды (глубина, скорость течения, температура, мутность) на структуру речных ихтиоценозов в пределах бассейна, оценка воздействия точечного источника загрязнения, позволяющая выявить зависимость между концентрациями загрязнителей и деградацией сообщества, а также изучение пространственного распределения молоди лососевых в эстуариях в зависимости от градиентов солёности и температуры. Во всех случаях применение RDA оправдано, когда исходные гипотезы предполагают прямолинейные или почти линейные связи, а данные охватывают умеренные средовые градиенты без выраженных пороговых эффектов.

Существуют, однако, сценарии, где применять RDA не рекомендуется. Метод плохо подходит для анализа данных с резко нелинейными откликами или сильными пороговыми эффектами, когда виды реагируют на среду нелинейно. Изучение трофических взаимодействий (хищничество, конкуренция) также плохо описывается линейной моделью из-за сложных опосредованных связей и поведенческих механизмов. Кроме того, RDA демонстрирует низкую эффективность при анализе сообществ с большим количеством нулевых значений или редких видов, где встречаемость часто носит нелинейный, пороговый характер. Для работы с категориальными предикторами (например, тип местообитания) RDA может быть использован после их преобразования в фиктивные переменные (*dummy variables*) в рамках линейной модели.

Близкой альтернативой RDA выступает канонический анализ соответствий (ССА), разработанный для моделирования унимодальных (колоколообразных) отношений между видами и средой. Выбор между этими методами определяется природой данных: RDA предполагает линейную модель отклика, тогда как ССА — унимодальную. На практике длина градиента среды, оцениваемая, например, с помощью DCA, является ключевым критерием: длина первой оси < 3 SD указывает на предпочтительность линейных методов, таких как RDA, длина > 4 SD — на целесообразность использования унимодальных методов, в том числе ССА; при значениях между 3 и 4 SD допустимы оба подхода. Таким образом, применение RDA требует тщательной проверки его допущений, анализа остатков и готовности использовать альтернативные методы, если данные не соответствуют линейной парадигме.

Пример анализа данных

Задача. В предустьевых зонах пяти рек бассейна Каспийского моря, рр. Урал, Самур, Терек, Кума и Сулак, на пяти станциях для каждой реки на одном и том же расстоянии от берега измерены величины абиотических факторов (солёность, температура, взвесь, кислород у дна и кислотность) и биомасса (г/м^2) 10-ти видов рыб (вобла, лещ, сазан, судак, сом, жерех, кутум, шемая, чехонь, окунь). Используя метод RDA, построить ординационную диаграмму и выявить связь плотностных характеристик разных видов с величинами факторов среды.

АНАЛИЗ ДАННЫХ

Анализ избыточности в PAST расположен в главном меню *Multivariate* → *Ordination* → *RDA*. Как и в случае ССА, при копировании данных необходимо вручную указать количество факторов среды (в текущей реализации PAST — до 6). Подготовительный этап включает обязательную стандартизацию абиотических переменных (Z-преобразование) и, при необходимости, преобразование биотических данных (например, методом Хеллингера), что снижает влияние доминирующих и редких видов, не затрагивая двойные нули. Длина экологического градиента составляет $\sim 2.4 \text{ SD}$ (< 3), что обосновывает выбор линейного метода RDA вместо унимодального ССА. В этих условиях предположение о линейном отклике видов на изменения среды является допустимым, а преобразование Хеллингера дополнительно усиливает чувствительность канонических осей к измеряемым градиентам среды.

Интерпретация осей и общая структура графика

Ключевым результатом RDA является триплот (вкладка *Scatter plot*), на котором одновременно отображаются точки проб, векторы видов и векторы факторов среды (Рис. 13.1). Для интерпретации рекомендуется использовать опцию *Fitted site scores* (последний чек-бокс на панели настроек), которая отображает положения проб, предсказанные моделью на основе факторов среды, а не исходных видовых данных. Это позволяет визуализировать именно объяснённую (*constrained*) компоненту вариации и оценить, насколько хорошо факторы среды воспроизводят структуру сообщества. Такое представление можно сравнить с переходом от «фотографического» изображения, фиксирующего все эмпирические детали, к схематическому чертежу, отражающему лишь те структурные взаимосвязи, которые согласуются с принятой моделью. Проведённый анализ избыточности позволил выявить и визуализировать чёткие зависимости между градиентами абиотических факторов и составом ихтиофауны в предустьевых пространствах рек, впадающих в Каспийское море; триплот, объединяющий проекции факторов среды, видов рыб и точек отбора проб (рек), демонстрирует согласованную структуру этих взаимосвязей.

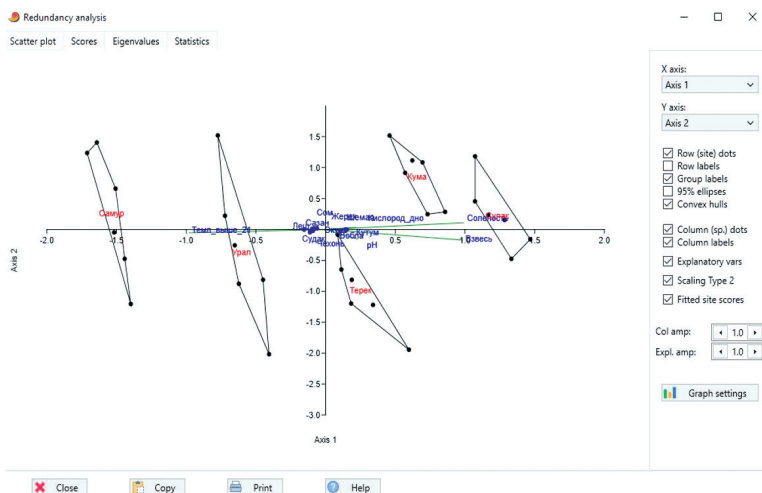
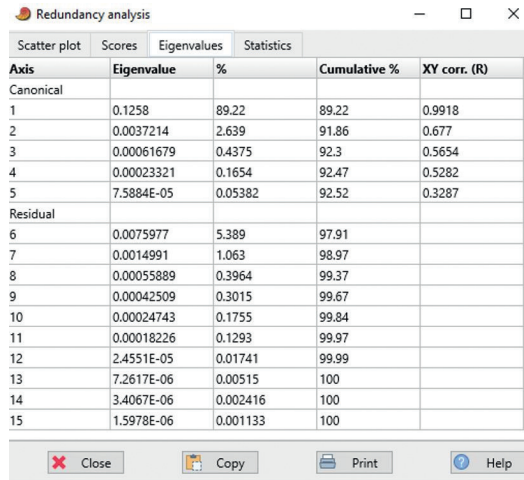


Рис. 13.1. Ординационная диаграмма — триплет — для результатов анализа избыточности (RDA)

Расположение объектов в пространстве первых двух осей выявляет выраженную экологическую структуру, обусловленную градиентом солёности. Первая каноническая ось аккумулирует 89,22 % объяснённой дисперсии ($Eigenvalue = 0,1258$) и интерпретируется как главный градиент минерализации (Рис. 13.2). Это подтверждается направлением и длиной вектора «Солёность», ориентированного практически параллельно оси 1 и демонстрирующего максимальную корреляционную связь ($XY \text{ corr. } R = 0,992$). Отрицательный полюс оси соответствует условиям максимального опреснения (реки Самур, частично Урал), тогда как положительный — зонам повышенной солёности (реки Сулак, Кума); река Терек занимает промежуточное положение вблизи начала координат, отражая переходные гидрохимические условия. Таким образом, первая ось формирует чёткий экологический континуум «пресная вода → солоноватая вода», определяющий основную структуру сообществ.

Вторая ось объясняет лишь 2,64 % вариации ($Eigenvalue = 0,0037$) и характеризуется умеренной корреляционной связью с матрицей факторов ($R = 0,677$), что указывает на её второстепенную роль. Низкая доля объяснённой дисперсии и преимущественная ориентация основных экологических векторов вдоль первой оси свидетельствуют о

том, что вторая компонента отражает слабые остаточные вариации, не связанные с доминирующим градиентом солёности, и может учитывать локальные стохастические факторы, microhabitat-различия или неучтённые параметры среды, не вносящие существенного вклада в общую структуру ихтиоценозов.



Axis	Eigenvalue	%	Cumulative %	XY corr. (R)
Canonical				
1	0.1258	89.22	89.22	0.9918
2	0.0037214	2.639	91.86	0.677
3	0.00061679	0.4375	92.3	0.5654
4	0.00023321	0.1654	92.47	0.5282
5	7.5884E-05	0.05382	92.52	0.3287
Residual				
6	0.0075977	5.389	97.91	
7	0.0014991	1.063	98.97	
8	0.00055889	0.3964	99.37	
9	0.00042509	0.3015	99.67	
10	0.00024743	0.1755	99.84	
11	0.00018226	0.1293	99.97	
12	2.4551E-05	0.01741	99.99	
13	7.2617E-06	0.00515	100	
14	3.4067E-06	0.002416	100	
15	1.5978E-06	0.001133	100	

Рис. 13.2. Собственные значения, доли дисперсии и канонические корреляции (RDA)

Вкладка *Statistics*

Проведённый анализ избыточности (RDA) демонстрирует очень сильную и статистически значимую связь между абиотическими параметрами и структурой ихтиоценоза в предустьевых зонах рек Каспийского бассейна (Рис. 13.3). Модель объясняет 92.5% вариации биотических данных ($R^2 = 0.925$), при скорректированном значении 91.0% ($R^2_{adj} = 0.909$), что указывает на высокую объяснительную способность включённых факторов среды и отсутствие выраженного переобучения модели. Статистическая надёжность полученных результатов подтверждается значением F-критерия ($F = 59.37$) и итогами перестановочного теста (*Permutation test*). При 999 случайных перестановках матрицы данных достигнутое значение значимости $p = 0.001$ указывает на то, что вероятность получения столь сильной структуры связей случайно практически отсутствует.

Совокупность этих данных позволяет заключить, что существенная часть вариабельности в распределении биомассы десяти видов рыб статистически связана с измеренными градиентами среды и объясняется включёнными в модель факторами, а выявленные паттерны отражают реальные экологические закономерности, а не стохастический шум. (Пояснение. Перестановочный тест (*Monte Carlo permutation test*) — непараметрическая процедура проверки статистических гипотез, основанная на многократном случайном перемешивании строк матрицы данных и построении эмпирического распределения F-статистики; такой подход позволяет корректно оценивать значимость, не опираясь на предположение о нормальном распределении данных, что особенно важно для эколого-биологических выборок).

Overall test

R² :	0.9252
R²_{adj} :	0.9096
F:	59.37
Perm p, n=999:	0.001

Рис. 13.3. Параметры статистического отчета RDA

Интерпретация облаков точек (проб) и их взаимного расположения

Размер, форма и положение облаков точек отражают степень внутренней однородности ихтиоценозов каждой реки и их экологическую дифференциацию вдоль главного градиента.

р. Самур. Кластер занимает крайнее левое положение на оси RDA1 и демонстрирует максимальный разброс по вертикали (RDA2). Это указывает на наиболее выраженные пресноводные условия и высокую неоднородность между сообществами отдельных рек, вероятно, обусловленную мозаичностью биотопов или неравномерностью гидрологического режима в предустьевой зоне.

р. Урал. Облако также смещено в отрицательную область RDA1, но расположено ближе к началу координат, чем у Самура. Это свидетельствует о преобладании пресноводных условий, однако с меньшим контрастом по главному градиенту. Выраженная вытянутость точек

вдоль оси RDA2 подтверждает значительную внутреннюю вариабельность состава проб, что может быть связано с протяженностью зоны смешения речных и морских вод.

р. Терек. Компактный кластер, смещенный в слабоположительную область RDA1. Его положение вблизи центра графика характеризует Терек как переходный водоём с умеренными значениями солёности и температуры, где сообщество отличается относительно высокой стабильностью и отсутствием резких внутренних контрастов.

р. Кума. Плотный кластер в правой части графика (положительные значения RDA1). Компактность облака указывает на стабильные гидрохимические условия и сформировавшееся однородное сообщество, адаптированное к повышенной минерализации воды.

р. Сулак. Крайнее правое положение на оси RDA1. Это облако соответствует участкам с максимальной солёностью и наиболее низкой температурой среди обследованных рек. Чёткая обособленность подтверждает уникальность экологических условий и высокую специализацию местного ихтиоценоза.

Пространственная сегрегация кластеров доказывает, что различия между реками (гидрология, солёность, температурный режим) существенно превышают изменчивость внутри рек. Принадлежность пробы к конкретной реке выступает жёстким экологическим фильтром, детерминирующим структуру сообщества. Такая чёткая дискретность валидирует адекватность модели и позволяет проводить уверенную экологическую интерпретацию для каждого водоёма.

Интерпретация расположения видов и векторов среды

Координаты векторов подтверждают, что ось RDA1 является доминирующим интегральным градиентом, вдоль которого согласованно, но разнонаправленно меняются ключевые абиотические параметры. Векторы «Солёность» (0,99) и «Взвесь» (0,98) практически коллинеарны и направлены в положительную область оси, что отражает типичную для эстуариев ситуацию: рост минерализации сопровождается увеличением концентрации взвешенных частиц морского или донного происхождения. Вектор «Температура» имеет сильную отрицательную нагрузку (–0,98) и направлен в противоположную сторону, указывая на выраженную обратную корреляцию с солёностью. Это закономерно для летнего периода: опреснённые речные воды (Самур, Урал) на мелководьях прогреваются быстрее, тогда как более солёные

морские массы (Сулак, Кума) обладают большей тепловой инерцией и остаются холоднее. Векторы «рН» и «Кислород_дно» крайне коротки и сгруппированы у начала координат, что свидетельствует об их статистической и экологической незначимости: вариация этих параметров не вносит существенного вклада в объяснение структуры сообществ на фоне мощного градиента солёности–температуры.

Распределение видовых координат чётко отражает их экологическую валентность. Большинство таксонов (сазан, сом, жерех, окунь) имеют небольшие отрицательные значения по RDA1, что указывает на их приуроченность к пресноводно-тёплому типу условий, но без строгой специализации — это эвритопные виды с широкой толерантностью. Лещ (–0,15) демонстрирует наиболее выраженную отрицательную нагрузку среди всех видов, выступая строгим индикатором опреснённых участков. Судак (–0,11) и чехонь (–0,09) также подтверждают свою приуроченность к пресным водам. Исключение составляют кутум (0,15) и шемай (0,14), чьи положительные координаты однозначно связывают их с солоновато-водным, более холодным комплексом. Вобла (0,10) тяготеет к положительным значениям оси, что указывает на её ассоциацию с солоноватыми водами рек Сулак и Кума, а не на эврибионтность.

Взаимное расположение видов и облаков проб

Проекция видов на конфигурацию проб подтверждает экологическую интерпретацию осей. Таксоны, тяготеющие к положительным значениям RDA1 (кутум, шемай, вобла), пространственно совпадают с кластерами рек Сулак и Кума, где зафиксированы максимальные значения солёности и минимальные температуры. Виды с отрицательными нагрузками (лещ, судак, чехонь) проецируются на область рек Самур и Урал, соответствующую пресноводным и более прогретым условиям, причём лещ выступает наиболее чётким индикатором этой зоны. Терек занимает промежуточное положение, что согласуется со смешанным характером его ихтиоценоза. Близость большинства видовых точек к центру координат (сазан, сом, жерех, окунь) указывает на их присутствие во всех обследованных реках, однако их количественное доминирование закономерно смещается вдоль градиента: в пресных зонах преобладают карповые и окуневые, в солоноватых — каспийские эндемики и полупроходные виды.

Экологические закономерности и выводы

Структура ихтиоценозов в предустьевых зонах рек западного Каспия на 89,22% детерминирована первым каноническим градиентом, отражающим сопряжённую вариацию солёности и температуры. Выявлена чёткая бинарная экологическая дифференциация: 1) пресноводно-тепловой комплекс (рр. Самур, Урал), характеризующийся доминированием леща, судака и чехони; 2) солоновато-водный, более холодный комплекс (рр. Кума, Сулак), где ключевыми видами выступают кутум, шемая и вобла. Река Терек демонстрирует переходный, сглаженный тип сообщества. Подавляющее число видов (сазан, сом, жерех, окунь) проявляют высокую пластичность, населяя весь диапазон условий, что типично для эстуарных экосистем. Параметры рН и донного кислорода не оказали значимого структурирующего влияния, уступив ведущую роль градиенту минерализации. Полученные данные подтверждают, что солёность, модулируемая температурным режимом, является главным абиотическим драйвером формирования рыбных сообществ в изученном регионе.

Философское эхо

Метод RDA выступает в роли мерчендайзера торгового зала, выстраивающего архитектуру ассортимента на перекрестье управляемых воздействий. Перед ним живая мозаика товарных потоков, где каждая единица товара имеет свою цену, сезонность и промо-активность, а задача специалиста — отделить управляемый рисунок спроса от фонового шума случайных колебаний продаж. Пропуская массив продаж через множественную регрессию, он вычисляет идеальную траекторию распределения: как должны располагаться товары, если бы их динамика полностью определялась заданными факторами. На этой отфильтрованной матрице предсказанного строятся оси главных компонент, задающие канонические направления — те осевые направления выкладки, вдоль которых структура спроса — влияние предикторов — становится измеримой. Разрозненные чеки складываются в читаемую геометрию коммерческих закономерностей, а неучтённая вариация остаётся за пределами планограммы, признавая импульсный спрос фактором, не поддающейся строгому прогнозу. Метод требует тщательной калибровки: без приведения факторов к единой шкале громкий ценовой дисконт заглушит визуальные маркеры оформления, искажая пропорции расчёта. При этом алгоритм остро реагирует на

выбросы — единичная аномальная транзакция работает как чужеродный элемент, способный развернуть оси в ложном направлении, что вызывает необходимость отсечь фоновый шум импульсных покупок. Визуализация разворачивается как рабочая схема зон влияния: векторы факторов расходятся из нулевой точки, показывая направление и силу влияния каждого параметра на спрос. Но линейная природа метода задаёт чёткие пределы. Если кривая спроса изгибается дугой, достигая пика в среднем сегменте, или ассортимент разбит на изолированные категории, оси теряют точность. Это напоминает попытку описать траекторию движения покупателя по торговому залу прямой линией: там, где связь нелинейна или дискретна, модель рисует лишь схематический контур, не улавливая реальную геометрию потока. Но там, где влияние факторов остаётся стабильным и непрерывным, RDA даёт прозрачную и проверяемую структуру, превращая многомерный хаос торговых данных в чёткий, управляемый коммерческий ландшафт.

14. ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ (LDA).

Введение

Линейный дискриминантный анализ LDA (*Linear Discriminant Analysis*) представляет собой статистический метод, предназначенный для анализа и классификации объектов с заранее известной групповой принадлежностью на основе множества измеренных признаков путём нахождения линейных комбинаций переменных, максимально разделяющих эти группы. Ключевая особенность подхода заключается в том, что для каждого наблюдения априорно известен класс, к которому оно относится. Цель LDA состоит в построении таких линейных функций исходных признаков, которые обеспечивают наибольшую разделимость заданных классов в многомерном пространстве и тем самым позволяют проецировать данные в подпространство меньшей размерности с максимально выраженными групповыми контрастами.

Метод был разработан Рональдом Фишером и представлен в работе «The Use of Multiple Measurements in Taxonomic Problems» (Fisher, 1936). Изначально он применялся для решения таксономических задач,

таких как разделение биологических видов или популяций по комплексу морфологических и экологических признаков. Фишер сформулировал критерий, направленный на максимизацию отношения межгрупповой изменчивости к внутригрупповой, что позволяет определить дискриминантные функции — направления в пространстве признаков, вдоль которых группы оказываются наилучшим образом разделены. LDA обладает рядом важных свойств. Во-первых, метод не только определяет направления максимального межгруппового различия, но и ранжирует их по значимости: первая дискриминантная функция объясняет наибольшую долю дисперсии, последующие — остаточную изменчивость при условии ортогональности. Во-вторых, при выполнении предположений о нормальности распределений и гомогенности ковариационных матриц LDA позволяет строить близкие к оптимальным классифицирующие правила. Таким образом, метод ориентирован на максимизацию различий между известными группами, что отличает его от других методов многомерного анализа.

В ихтиологических исследованиях LDA находит применение для решения задач визуализации и количественной оценки различий между группами, например при сравнении географических популяций, экологических форм или близких видов. Метод также позволяет идентифицировать признаки, вносящие наибольший вклад в разделение групп, что способствует выявлению биологически значимых характеристик. Благодаря этому LDA служит инструментом преобразования многомерных данных в интерпретируемые структуры, основанные на заданной классификации.

Методология

Математический аппарат LDA оперирует фундаментальными статистическими понятиями, выраженными в терминах ковариационных матриц. Внутригрупповая изменчивость характеризуется объединённой внутригрупповой ковариационной матрицей, которая описывает усреднённую форму и размер распределения точек внутри каждой группы с учётом ковариаций между признаками. Межгрупповая изменчивость представлена ковариационной матрицей между групповыми средними значениями, отражающей степень различий между центрами групп. Суть метода заключается в нахождении таких линейных комбинаций исходных признаков (дискриминантных функций), которые максимизируют отношение межгрупповой вариации к внутригрупповой вариации для проекций точек на эти направления.

Полученные дискриминантные оси строятся ортогональными друг другу в пространстве, преобразованном с учётом внутригрупповой ковариационной структуры, что обеспечивает статистическую независимость их вкладов в разделение групп.

LDA принципиально отличается от других методов многомерного анализа, таких как канонический анализ соответствий (ССА). Если LDA фокусируется исключительно на максимизации разделения между заранее заданными группами по числовым признакам, то ССА направлен на максимизацию корреляции между проекциями двух различных наборов переменных, исследуя связи между матрицами данных. Таким образом, хотя оба метода используют схожие принципы проецирования данных, их математические механизмы оптимизируют фундаментально различные целевые функции.

Метод основывается на двух ключевых статистических предположениях: о многомерной нормальности распределения признаков внутри каждой группы и о гомогенности ковариационных матриц между группами. Предположение о гомогенности ковариаций означает, что форма и размер облаков точек, представляющих разные группы, статистически одинаковы. Нарушение этого предположения может снижать эффективность анализа, однако метод демонстрирует определённую практическую устойчивость к умеренным отклонениям от идеальных условий. В случаях значительных нарушений статистических предпосылок могут применяться непараметрические или робастные аналоги метода.

В биологических исследованиях LDA служит мощным инструментом для преобразования многомерных данных в низкоразмерное пространство дискриминантных осей. Каждая такая ось представляет собой взвешенную комбинацию исходных признаков, где веса отражают вклад каждого признака в разделение групп с учётом ковариаций со всеми остальными переменными. Это позволяет выявлять комплексные паттерны различий между группами организмов, которые остаются неочевидными при анализе отдельных признаков, и визуализировать групповую структуру данных в интерпретируемой форме.

Характер данных

Линейный дискриминантный анализ (LDA) представляет собой статистический метод классификации и снижения размерности данных, эффективность применения которого строго обусловлена соответствием анализируемых данных ряду ключевых предпосылок,

связанных с их структурой. Метод предназначен для исследовательских ситуаций, в которых объекты анализа принадлежат заранее определённым категориальным группам (таким как таксономические единицы, географические популяции, экологические формы или половые категории), а описывающие их признаки являются непрерывными или близкими к непрерывным по своей природе (например, морфометрические показатели, весовые характеристики, физиологические параметры, биохимические концентрации). Фундаментальными математическими допущениями LDA выступают многомерная нормальность распределения признаков внутри каждой из групп и гомогенность ковариационных матриц между сравниваемыми группами. Нарушение этих условий — в частности, существенная гетерогенность ковариационных структур между группами или выраженные отклонения внутригрупповых распределений от многомерной нормальности — может привести к смещению результатов дискриминации, снижению точности классификации и содержательно искажённой интерпретации выявленных групповых различий.

В контексте биологических исследований LDA демонстрирует высокую эффективность при решении задач дискриминации и классификации в тех случаях, когда структура данных соответствует его базовым предпосылкам. Типичными примерами успешного применения метода являются: разграничение географических популяций мигрирующих видов по комплексу морфометрических и пластических признаков, где популяции априори определены по районам происхождения, а анализируемые признаки являются непрерывными и удовлетворяют условиям нормальности распределения и гомогенности ковариационных структур; дифференциация экологических форм по комплексу признаков, связанных с трофической специализацией и адаптацией к условиям среды, где группы чётко определены по экологическим критериям, а признаки являются непрерывными и соответствуют статистическим предпосылкам метода; а также таксономическая диагностика близкородственных форм на основе комплекса морфологических и остеологических характеристик, которые являются непрерывными или дискретными с большим числом градаций, при условии контроля онтогенетической изменчивости и соответствия внутригрупповых распределений нормальному закону. В перечисленных сценариях LDA позволяет выявить направления максимального межгруппового различия и построить эффективные классификационные функции.

Тем не менее, применение LDA становится проблематичным или методологически неоправданным при существенном нарушении его ключевых допущений или несоответствии типа данных и исследовательской цели. Метод неприменим для анализа данных по обилию таксонов в сообществах, где признаки являются категориальными, полуколичественными, характеризуются большим числом нулевых значений и резко асимметричными распределениями; для сравнения групп по частотным показателям, поскольку такие данные по своей природе категориальны и редко удовлетворяют условию многомерной нормальности; для идентификации категорий с возрастозависимым диморфизмом при анализе смешанных выборок, включающих онтогенетические стадии с различной морфологией; а также для выявления связи структуры сообществ с градиентами среды, поскольку LDA фокусируется исключительно на разделении предопределённых групп по признакам. В подобных ситуациях предпочтение следует отдавать альтернативным методам, соответствующим специфике данных и исследовательским задачам.

Таким образом, осознанный выбор метода многомерного анализа зависит от типа данных и конкретной исследовательской задачи. Обязательным этапом перед применением LDA является тщательная проверка его предпосылок — многомерной нормальности распределений признаков внутри групп и гомогенности ковариационных матриц. Серьезные нарушения этих допущений требуют осторожной интерпретации результатов или рассмотрения альтернативных методов анализа.

Пример анализа данных

Задача. На нагульных участках молоди трёх видов лососевых — семги, кумжи и арктического гольца — в водотоках северной Норвегии были оценены величины следующих факторов: глубина, степень затенённости русла, растворённый кислород, скорость течения, температура, плотность корма, число видов корма, конкуренция (через плотность рыб). Используя ординационный тест LDA, выявить руководящие факторы, определяющие предпочтение каждым из видов нагульных участков.

АНАЛИЗ ДАННЫХ

Линейный дискриминантный анализ LDA в PAST размещён в главном меню *Multivariate* → *Ordination* → *LDA*. Предварительный этап LDA включает стандартизацию величин абиотических/биотических факторов путем Z-преобразования. Анализ результатов линейного дискриминантного анализа осуществляется через систему специализированных вкладок с различными компонентами выходных данных.

Вкладка *Classifier*

Параметры вкладки *Classifier* демонстрируют эффективность построенной дискриминантной модели (рис. 14.1). В таблице сравнивается исходная видовая принадлежность молоди (*Given group*) с прогнозом, сделанным моделью на основе морфометрических параметров (*Classification*). Для 36 образцов, распределённых по трём видам (атлантический лосось — точки 1–12, кумжа — точки 13–23, голец арктический — точки 24–36), наблюдается полное совпадение значений во всех строках этих столбцов, что свидетельствует о 100% точности классификации обучающей выборки. Это означает, что выбранные признаки однозначно дифференцируют виды и позволяют безошибочно определять видовую принадлежность молоди в каждой точке отбора.

Для проверки надёжности модели и оценки её устойчивости к переобучению использован метод перекрёстной проверки *Jackknifed (leave-one-out)*, при котором каждый образец классифицируется моделью, построенной без его участия. Результат также показывает идеальное совпадение с исходными данными во всех 36 случаях, что подтверждает высокую прогностическую силу модели, её робастность и применимость для анализа новых, неизвестных проб. Таким образом, модель статистически доказывает, что исследуемые виды достоверно различаются по комплексу морфометрических признаков, что отражает наличие видоспецифичных адаптаций и различия в экологических нишах молоди.

Discriminant analysis			
Plot	Scores	Classifer	Confusion matrix
Centroids			
Point	Given group	Classification	Jackknifed
1	Атлантический лосось	Атлантический лосось	Атлантический лосось
2	Атлантический лосось	Атлантический лосось	Атлантический лосось
3	Атлантический лосось	Атлантический лосось	Атлантический лосось
4	Атлантический лосось	Атлантический лосось	Атлантический лосось
5	Атлантический лосось	Атлантический лосось	Атлантический лосось
6	Атлантический лосось	Атлантический лосось	Атлантический лосось
7	Атлантический лосось	Атлантический лосось	Атлантический лосось
8	Атлантический лосось	Атлантический лосось	Атлантический лосось
9	Атлантический лосось	Атлантический лосось	Атлантический лосось
10	Атлантический лосось	Атлантический лосось	Атлантический лосось
11	Атлантический лосось	Атлантический лосось	Атлантический лосось
12	Атлантический лосось	Атлантический лосось	Атлантический лосось
13	Кумжа	Кумжа	Кумжа
14	Кумжа	Кумжа	Кумжа
15	Кумжа	Кумжа	Кумжа
16	Кумжа	Кумжа	Кумжа
17	Кумжа	Кумжа	Кумжа
18	Кумжа	Кумжа	Кумжа
19	Кумжа	Кумжа	Кумжа
20	Кумжа	Кумжа	Кумжа
21	Кумжа	Кумжа	Кумжа
22	Кумжа	Кумжа	Кумжа
23	Кумжа	Кумжа	Кумжа
24	Кумжа	Кумжа	Кумжа
25	Гольц арктический	Гольц арктический	Гольц арктический
26	Гольц арктический	Гольц арктический	Гольц арктический
27	Гольц арктический	Гольц арктический	Гольц арктический
28	Гольц арктический	Гольц арктический	Гольц арктический
29	Гольц арктический	Гольц арктический	Гольц арктический
30	Гольц арктический	Гольц арктический	Гольц арктический
31	Гольц арктический	Гольц арктический	Гольц арктический
32	Гольц арктический	Гольц арктический	Гольц арктический
33	Гольц арктический	Гольц арктический	Гольц арктический
34	Гольц арктический	Гольц арктический	Гольц арктический
35	Гольц арктический	Гольц арктический	Гольц арктический
36	Гольц арктический	Гольц арктический	Гольц арктический

Рис. 14.1. Параметры вкладки *Classifer*

Матрица ошибок *Confusion matrix*

Матрица ошибок (рис. 14.2) позволяет количественно оценить точность построенной дискриминантной модели. Строки таблицы соответствуют реальной видовой принадлежности молодца (*Given group*), а столбцы — прогнозу, сделанному моделью (*Classification*).

Анализ матрицы демонстрирует идеальный результат: все 12 особей атлантического лосося, 12 особей кумжи и 12 особей арктического гольца были классифицированы верно. Нулевые значения вне главной диагонали указывают на полное отсутствие ошибок. Стопроцентная точность классификации (100% *correctly classified*) свидетельствует о том, что анализируемые признаки (вероятнее всего, морфометрические параметры) позволяют абсолютно точно дифференцировать виды. Это статистически подтверждает, что исследуемые группы достоверно различаются и не перекрываются в многомерном пространстве признаков, что может интерпретироваться как наличие чётких видовых различий (экологических или морфологических).

Discriminant analysis

Plot	Scores	Loadings	Classifier	Confusion matrix	Centroids
		Атлантичес	Кумжа	Голец аркти	Total
Атлантичес	12	0	0	12	
Кумжа	0	12	0	12	
Голец аркти	0	0	12	12	
Total	12	12	12	36	

Rows: Given groups
Columns: Predicted grps

☐ Jackknifed

% correctly classified:
100

Close Copy Print Help

Рис. 14.2. Параметры вкладки *Confusion matrix*

Вкладка *Loadings*

Вкладка числовые значения, которые показывают силу, направление и характер связи между исходными переменными (параметрами среды) и построенными дискриминантными осями (Рис. 14.3). *Радиометка «C»* данной вкладки выводит стандартизированные коэффициенты дискриминантной функции, показывающие, какой вклад вносит каждая переменная в построение оси и насколько сильно она влияет на разделение заранее заданных групп — чем больше абсолютное значение коэффициента, тем важнее роль этой переменной в дискриминации. *Радиометка «WcovC»* выводит коэффициенты корреляции между исходными переменными и дискриминантными осями. Они помогают интерпретировать смысл каждой оси, показывая, с изменением каких параметров среды наиболее тесно связано положение объектов вдоль этой оси.

Discriminant analysis

Plot	Scores	Loadings	Classifier	Confusion matrix
Matrix ☒ C ☐ WcovC				
	Axis 1	Axis 2		
Глу	5,173	-5,313		
Затененности	1,9743	-0,79333		
Субстрат	-1,1269	3,5117		
Кислород	-3,7133	9,7931		
Течение	-2,3977	5,4104		
Температура	13,536	10,773		
Плотность кс	0,72014	3,5168		
Число видов	1,9689	-1,1633		
Конкуренция	-2,6073	-5,7161		

Close Copy Print

Рис. 14.3. Параметры вкладки *loadings*

Вкладка *Centroids*

Вкладка содержит усреднённые координаты всех точек, принадлежащих к одной группе, на дискриминантных осях; они показывают «сердце» каждой группы или типичное её положение каждой в пространстве, построенном для максимального её отличия от других.

Интерпретация осей и общая структура графика

Вкладка *Plot*

Данная вкладка содержит ключевой результат анализа — триплот, представляющий собой комплексную визуализацию, совмещающую проекции объектов, положения групповых кластеров точек и векторы признаков в пространстве дискриминантных функций. Это позволяет одновременно интерпретировать разделение заранее определённых групп, взаимное расположение кластеров, направление и относительный вклад исходных признаков в формирование дискриминантных осей, а также корреляционные взаимоотношения между признаками, обеспечивая тем самым комплексное представление результатов анализа.

Ось LD1 является безусловно доминирующей, объясняя 80,8% общей дисперсии, что составляет подавляющую долю различий между группами (рис. 14.4). Её можно интерпретировать как интегральный градиент, связанный с гидродинамическими и температурными условиями. Отрицательный полюс оси LD1 чётко ассоциирован с арктическим гольцом — это условия, характерные для более глубоких участков со слабым течением. Положительный полюс оси LD1 представлен кумжей, где доминирующие абиотические параметры — сильное течение и каменистые грунты. Атлантический лосось также занимает положение в области положительных значений LD1, но при этом смещён по оси LD2, что указывает на его специфические требования к кислородно-температурному режиму.

Ось LD2, хотя и существенно менее весома, объясняет дополнительную, ортогональную (независимую) часть изменчивости, составляющую 19,2% от общей дисперсии. Эту ось можно интерпретировать как градиент кислородных, температурных и трофических условий. Положительный полюс оси LD2 сильно связан с атлантическим лососем — это зона с высоким содержанием кислорода, относительно высокой температурой и богатой кормовой базой. Отрицательный полюс оси LD2 в большей степени связан с кумжей и гольцом. Таким

образом, совместно эти две оси охватывают 100% объяснённой дисперсии, полностью описывая выделенные статистикой различия между видами в пространстве заданных переменных.

Квадрант I: Экологическая ниша атлантического лосося. Облако точек атлантического лосося формирует здесь плотный, хорошо обособленный кластер с самыми высокими значениями по оси LD2, что подтверждает занимаемую данным видом узкую и специфическую нишу. Этот квадрант не содержит ни одного вектора тестируемых параметров, что означает, что ниша лосося не определяется экстремальными значениями ни одного из факторов, но представляет собой комбинацию умеренных значений по всем параметрам (средняя глубина, течение, субстрат и т. д.), что и отражено в его положении на положительных значениях LD1 и максимальных положительных значениях LD2. Его позиция — это «золотая середина», смещённая в сторону благоприятных кислородно-трофических условий.

Квадрант II Переходная зона. В этот квадрант направлены векторы глубины, кислорода, субстрата, плотности и видового разнообразия корма. Это самый «нагруженный» квадрант с точки зрения векторов среды. Векторы средовых драйверов образуют с осью OX угол $\sim 45^\circ$, что означает одинаковую важность данных факторов для объяснения изменчивости по обеим осям, LD1 и LD2. Их совместное направление формирует общий градиент «благоприятности среды» по глубине, субстрату и кормовой базе, который противопоставлен векторам в Квадранте IV. Здесь практически отсутствуют точки исследуемых видов (исключение — единственная точка арктического гольца), что подчёркивает переходный характер данной зоны.

Квадрант III. Экологическая ниша арктического гольца. Облако точек арктического гольца образует плотный кластер в этом квадранте с крайне низкими значениями по LD1 и отрицательными по LD2. Это указывает на самую специфическую и отличную от других двух видов нишу. В этот квадрант не заходит ни один вектор. Это, как и в случае с лососем, означает, что ниша гольца определяется комплексом параметров среды, противоположных экстремальным значениям, представленным в других квадрантах: большие глубины, отсутствие сильного течения, умеренное содержание кислорода и более бедная кормовая база.

Квадрант IV: Экологическая ниша кумжи. Облако точек кумжи расположено в этом квадранте, имеющем высокие значения по LD1 и отрицательные по LD2. Кластер менее компактный, чем у лосося,

что может говорить о несколько большей пластичности вида. В этот квадрант направлены векторы течения, конкуренции, температуры и затенённости. Температура воды имеет самую большую длину проекции на LD1 (коэффициент 13,129), что указывает на её ключевую роль в определении ниши кумжи. Векторы течения и затенённости также сильно смещены в сторону положительных значений LD1, характеризуя биотопы кумжи как участки с быстрым течением. Положение вектора температуры с его положительной проекцией на LD1 указывает на то, что кумжа ассоциирована с зонами с более высокой температурой по сравнению с другими видами.

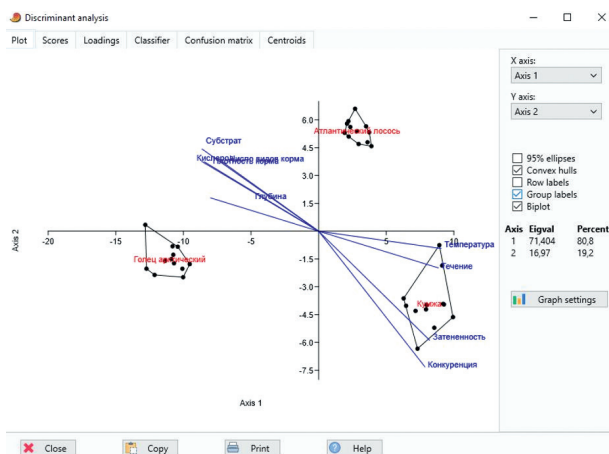


Рис. 14.4. Ординационная диаграмма — триплот — линейного дискриминантного анализа (LDA)

Сводка экологических портретов видов

Атлантический лосось. Молодь обитает на умеренных глубинах со средними скоростями течения; ключевыми для вида являются высокое содержание растворенного кислорода, сравнительно высокие температуры воды и богатая кормовая база: молодь населяет биотопы с оптимальным сочетанием условий без экстремумов.

Кумжа. Молодь тяготеет к водотокам с высокими скоростями течения; её ниша определяется затененным руслом и участками рек с высокой скоростью потока, низкой температурой воды и каменистым субстратом; характерно наличие высокой внутривидовой конкуренции.

Арктический голец. Занимает противоположную часть экологического пространства. Молодь ассоциирована с относительно глубокими участками водотоков, характеризующимися невысокими скоростями течения, пониженной концентрацией растворённого кислорода и относительно бедной кормовой базой.

Сила и направление влияния параметров

Длина вектора на биплоте прямо пропорциональна его вкладу в разделение групп. Температура является наиболее значимым фактором (самый длинный вектор), за ним следуют кислород и глубина. Направление влияния средовых векторов: 1) положительная корреляция с LD1: температура, глубина, число видов и плотность корма; 2) отрицательная корреляция с LD1: глубина, субстрат, кислород, конкуренция; 3) положительная корреляция с LD2: кислород, температура, плотность корма, субстрат; 4) отрицательная корреляция с LD2: глубина, затененность, число видов и конкуренция.

Анализ выявляет два ключевых ортогональных градиента условий среды: первый и наиболее мощный из них описывает переход от зон со слабой гидродинамикой и большими глубинами к участкам с интенсивным течением и каменистыми субстратами, в то время как второй, независимый градиент, отражает изменение от менее благоприятных кислородных и термических режимов с обедненной трофической базой к зонам с высокой концентрацией кислорода, повышенной температурой и богатыми кормовыми ресурсами.

Философское эхо

LDA — это штурман, прокладывающий маршрут по уже нанесённым на карту акваториям, чей компас неизменно держит направление вдоль оси наибольшего расхождения между обозначенными архипелагами. Его дар — не в открытии новых берегов, но в умении прокладывать оптимальный курс между известными островами, превращая зыбкие догадки наблюдателя в чёткие меридианы пространства признаков. Отдельный признак здесь становится парусом, улавливающим ветер изменчивости, но курс задаётся не случайными порывами ветра, а выверенным румбом дискриминантной функции, вдоль которого контрасты проступают с максимальной отчётливостью. Метод преобразует судовые журналы наблюдений в навигационные карты, где уже открытые архипелаги обретают строго очерченные берега, а отдельные

пробы — координаты, вычисленные методом морского счисления широта и долгота. Каждая точка здесь — результат линейного сложения ветров-признаков, застывший в чётких координатах. В этой геометрической ясности лежат и сила, и ограничение метода: когда острова расплываются в тумане неоднородной дисперсии, а внутренние течения групп поднимают волны неопределённости, линейные границы теряют контур, как береговая линия в шторм. Подобно лоцману, сверяющему курс с картой и учитывающему характер течений, исследователь должен чувствовать природу LDA: это инструмент прокладки маршрута по уже известным акваториям. Если реакции видов становятся нелинейными, а градиенты — как фарватер между рифами — извилистыми, метод уступает место иным навигационным системам, способным вести судно по петлям шхерных проливов. Но в руках того, кто соблюдает правила навигации и слышит ритм океана данных, LDA становится не набором навигационных пособий, а цельной логикой, превращающей гипотезу о различии в зримый морской ландшафт, где каждая проекция — проложенный курс, напоминающий, что даже в многомерном океане данных можно найти ось, вдоль которой мгла отступает, возвращая видимость берегов, а вектор проекции сохраняет чёткость навигационного створа.

15. CALGARY AND BROWN FACTOR ANALYSIS (CABFAC)

Введение

В условиях ограниченных вычислительных ресурсов (начало 1970-х гг.) анализ крупных биологических и геологических массивов данных был сопряжен со значительными трудностями. Особую сложность представляли задачи, требующие оценки сходства между множеством объектов (образцов, проб, особей) по десяткам признаков. Классический Q-режим факторного анализа сталкивался с вычислительными ограничениями из-за необходимости обработки матриц парных сходств большого размера. В данном технологическом контексте был разработан метод CABFAC (*Calgary and Brown Factor Analysis*) геологом Дж. Э. Клованом (Университет Калгари) и палеоклимато-

логом Дж. Имбри (Брауновский университет) (Klovan, Imbrie, 1971). Ключевым достижением метода стал алгоритмический подход, позволивший избежать трудоемкого вычисления матрицы парных сходств и эффективно анализировать до 1500 объектов, что существенно превосходило возможности стандартных методов того времени.

Изначально созданный для решения геологических и палеонтологических задач, таких как, например, анализ минерального состава пород, комплексов ископаемых фораминифер, метод CABFAC впоследствии нашёл широкое применение в биологии и экологии. Его основу составляет выявление латентной структуры сходства между объектами через анализ пропорций их признаков, а не абсолютных значений. Способность метода работать с относительным вкладом характеристик — морфометрическими соотношениями, видовыми обилиями, частотами аллелей — обусловила его применение в экологии, популяционной генетике и ихтиологии. CABFAC обеспечивает не только сокращение размерности данных, но и позволяет идентифицировать объективно существующие группировки и градиенты в многомерном пространстве признаков, основанные на внутренней структуре самих данных. Таким образом, для ихтиолога метод представляет собой инструмент перехода от первичных измерений к интерпретируемым биологическим закономерностям.

Методология

Анализ крупных экологических массивов данных, таких как характеристики множества станций наблюдения, оценивающих десятки параметров (видовое разнообразие, плотность популяций, биотические индексы, абиотические факторы), с использованием классического Q-режима факторного анализа часто сопряжён с вычислительными трудностями. Необходимость расчёта и обработки матрицы парных сходств между объектами (станциями) размером $N \times N$, где N — количество объектов, становится ресурсоёмкой задачей. Для 1500 станций размер такой матрицы превышает миллион элементов, что существенно ограничивает применимость стандартных методов.

Алгоритм CABFAC предлагает альтернативное решение, основанное на анализе взаимосвязей между параметрами, а не между объектами. Вместо построения матрицы сходств объектов алгоритм использует матрицу кросс-продуктов (R-матрицу) размером $M \times M$, где M — количество параметров. Это позволяет существенно сократить

вычислительную нагрузку, поскольку M , как правило, значительно меньше N . Математически ключевая информация о вариативности объектов может быть получена из анализа этой компактной матрицы.

Процедура включает несколько этапов: 1) нормализация данных: значения параметров для каждого объекта стандартизируются, что позволяет анализировать относительные пропорции признаков, а не их абсолютные величины; 2) построение матрицы кросс-продуктов: вычисляется матрица, отражающая ковариацию между всеми параметрами по всей выборке объектов; 3) выделение главных компонент: для R -матрицы находятся собственные векторы и значения, которые соответствуют основным скрытым факторам, объясняющим ковариацию параметров; 4) проецирование данных: исходные нормализованные данные проецируются на пространство выделенных факторов, что даёт факторные нагрузки (Q) для каждого объекта; эти нагрузки количественно выражают вклад каждого фактора в характеристику объекта; 5) вращение факторов: для упрощения интерпретации применяется вращение (например, варимакс), которое максимизирует дисперсию нагрузок и стремится к такой структуре, где каждый объект имеет высокие нагрузки на минимальное число факторов.

CABFAC позволяет эффективно обрабатывать данные, например, 1500 объектов по 50 параметрам, выделяя несколько главных факторов. Количество сохраняемых факторов определяется заданным порогом объяснённой дисперсии (параметр $QUIT$, например 95%). Алгоритм извлекает факторы до достижения заданного порога или до ограничения в 10 факторов. Для оптимизации интерпретируемости применяется стратегия последовательного вращения «снизу вверх»: после вращения всех факторов, удовлетворяющих $QUIT$, фактор с наименьшим вкладом исключается, а оставшиеся повторно вращаются. Процесс продолжается до тех пор, пока не останется два фактора или объяснённая дисперсия не упадёт ниже 75%. Это позволяет получить несколько альтернативных ротированных решений, из которых исследователь выбирает наиболее содержательное с экологической точки зрения. (*Пояснение.* Вращение — процедура поворота осей в многомерном пространстве.)

Важной особенностью CABFAC является возможность проекции новых данных на существующую факторную структуру. Если для эталонной выборки уже выявлены главные факторы изменчивости, новые объекты могут быть нормализованы аналогичным образом и

спроецированы на матрицу ротированных факторных счетов эталонной выборки. Это позволяет непосредственно сравнивать новые объекты с существующими группами и интерпретировать их в контексте ранее определённых факторов, таких как «антропогенная нагрузка» или «климатическая зональность». Таким образом, CABFAC представляет собой эффективный метод анализа больших многомерных экологических данных, обеспечивающий не только выявление структуры сходства между объектами, но и построение устойчивых моделей изменчивости, применимых к новым данным. Алгоритм сочетает вычислительную эффективность с содержательной интерпретацией результатов.

Характер данных

Методологическая инновация CABFAC — селективный скрининг (сортировка) значимых абиотических драйверов. В отличие от классических ординационных методов (RDA, CCA, LDA), интегрирующих все факторы среды в единую модель «среда–биота», CABFAC реализует принципиально иной подход: последовательное тестирование связи каждого отдельного абиотического параметра со структурой сообщества. Метод выявляет объективные оси биотической изменчивости (PC1, PC2) и строит регрессионную модель, преобразующую координаты сообщества в реконструированные значения среды. Это позволяет выявлять значимые драйверы (параметры с высоким R^2), отсеивать нерелевантные факторы (низкий R^2) и ранжировать параметры по силе влияния на биоту через метрики (R^2 /RMSE). Такой селективный скрининг трансформирует CABFAC в инструмент оптимизации мониторинга, смещая фокус на факторы, успешно реконструируемые через биотические паттерны, при игнорировании статистически незначимых переменных.

При этом важно подчеркнуть, что CABFAC относится к семейству методов прямого градиентного анализа, работающих с матрицами «факторы среды–биотические отклики», и оперирует теми же типами данных: количественными абиотическими параметрами (солёность, температура, кислород, pH и т. п.) и количественными характеристиками сообщества (плотность видов, биомасса, встречаемость). Однако его аналитическая философия радикально переосмысливает взаимодействие этих матриц. Если в RDA или CCA среда выступает как монолитный блок предикторов, совместно

объясняющих биотическую изменчивость через единую модель, то CABFAC деконструирует этот блок, рассматривая каждый фактор как самостоятельную сущность, чья связь с биотой тестируется изолированно.

Данный метод декомпозиции не требует принципиально иных данных, но раскрывает дополнительный аналитический потенциал стандартных гидробиологических таблиц. Например, данные по количественно-видовому составу ихтиоценозов, полученные вдоль градиентов факторов среды и традиционно анализируемые методами CCA/RDA, могут быть повторно проанализированы с помощью CABFAC. Этот метод оценивает информативность каждого параметра через степень его воспроизводимости на основе структуры биотических компонентов. Это особенно полезно для сложных систем с множеством потенциальных драйверов (температура, солёность, мутность, нитраты, фосфаты, pH и др.), где интегральные методы часто сталкиваются с проблемой мультиколлинеарности. CABFAC последовательно оценивает вклад каждого фактора, измеряя его объяснительную способность через коэффициент детерминации (R^2), что позволяет выделять статистически значимые воздействия среди множества переменных.

Таким образом, CABFAC не заменяет RDA или CCA, а дополняет арсенал методов, предлагая альтернативный путь интерпретации тех же исходных данных. Его преимущество заключается в способности преобразовывать стандартную матрицу «станция × параметр» в диагностическую карту значимости факторов, где ранжирование по R^2 позволяет объективно оценивать влияние средовых переменных. Это делает метод особенно полезным для: 1) скрининга гипотез о драйверах экосистемных сдвигов; 2) оптимизации программ мониторинга путём исключения малозначимых параметров; 3) ретроспективного анализа архивных данных, где доступны биотические сводки, но отсутствуют детальные измерения среды.

Пример анализа данных

Задача. В предустьевой части пяти рек бассейна Каспийского моря оценили величины ряда абиотических параметров и биомассу трёх видов кефалей — остроноса, лобана и сингиля. Используя метод CABFAC, оценить влияние отдельных факторов среды на динамику биомассы кефалевых рыб.

АНАЛИЗ ДАННЫХ

Тест CABFAC размещён в PAST в главном меню *Multivariate* → *Ordination* → *CABFAC*. Перед переносом данных из Excel в PAST биотические данные стандартизировать не требуется. Данные должны быть расположены в строгом порядке: 1) группирующая переменная, 2) абиотические факторы и 3) биотические параметры. Важно учитывать, что за одну аналитическую сессию CABFAC исследует зависимость биотического комплекса только от одного абиотического фактора. В связи с этим, если средовых параметров несколько, в листе данных PAST необходимо, удерживая клавишу «Ctrl», выделить столбцы с 1) группирующей переменной, 2) одним выбранным абиотическим фактором и 3) всеми биотическими показателями. После запуска CABFAC программа запросит: «*Does first column contain environmental data?*», на что следует ответить «Yes», после чего откроется основное меню анализа. По завершении работы с одним фактором процедура повторяется для следующего средового параметра.

Вкладка *Summary* содержит таблицу с собственными значениями (*Eigenvalues*) и процентом объяснённой дисперсии (% *variance*) для каждой оси (PC). Идентичность значений *Eigenvalues* и % *variance* для всех абиотических параметров обусловлена методологией CABFAC: факторная модель строится исключительно на основе ковариационной или корреляционной матрицы биотических данных и затем последовательно используется для реконструкции каждого абиотического параметра через регрессионные веса. Таким образом, таблица *Summary* отражает внутреннюю структуру вариации самого биологического сообщества, которая служит общей математической основой для всех последующих реконструкций, а не результаты анализа отдельных средовых факторов.

Вкладка *Scatter plot* содержит ординационную диаграмму рассеяния в пространстве первых двух главных компонент (*Factor 1 vs Factor 2*), где положение каждой точки определяется её факторными координатами (вкладка *Factors*). Данный график идентичен для всех анализируемых параметров среды, так как отображает структуру биотических данных, а не конкретного абиотического фактора. Вкладка *Scores* включает факторные нагрузки (вклад видов в формирование осей), которые также инвариантны относительно выбора абиотического фактора. На основании этих данных во вкладке *Score plot*

строятся дивергентные столбчатые диаграммы, наглядно демонстрирующие вклад каждого вида в формирование факторных осей; в правом верхнем углу интерфейса расположена опция для выбора отображаемой оси.

Вкладка *Reconstruction plot* содержит графики реконструкции — диаграммы рассеяния для сравнения измеренных (фактических) значений абиотического параметра с величинами, предсказанными моделью на основе координат сообщества в факторном пространстве. Близость точек к линии 1:1 ($y = x$) отражает степень, в которой данный средовый параметр «закодирован» в структуре биотического сообщества. Ключевыми метриками качества модели являются коэффициент детерминации R^2 и среднеквадратическая ошибка RMSE (*Root Mean Square Error*), отображаемые на вкладке *Reconstruction plot*. R^2 показывает долю дисперсии фактических значений параметра, объясняемую моделью CABFAC; чем ближе R^2 к 1, тем точнее модель описывает связь. RMSE отражает среднюю величину ошибки предсказания в физических единицах измерения параметра; чем меньше RMSE, тем ближе реконструированные значения к реально измеренным. Числовые данные для построения графика доступны на вкладке *Reconstruction*. Значения в столбце *Residual* (остатки) представляют собой разность между фактическими и реконструированными величинами для каждой конкретной выборки, что позволяет локализовать участки, где модель даёт наибольшую погрешность.

Величина RMSE сама по себе не является абсолютным индикатором качества модели: одно и то же числовое значение может свидетельствовать как о высокой точности, так и о неприемлемой погрешности — всё зависит от масштаба измеряемого признака. RMSE = 0.02 мг/л для вещества с диапазоном 0–1 мг/л означает высокую точность — 2% от размаха вариации — и во многих прикладных задачах будет считаться приемлемой, тогда как RMSE = 20 мг/л для диапазона 0–100 мг/л соответствует уже 20% от размаха вариации, что указывает на сомнительную прогностическую пригодность модели. Таким образом, RMSE всегда следует оценивать в контексте диапазона или среднего значения целевой переменной. Окончательный вывод о качестве модели делается на основе совместного анализа RMSE и R^2 : первая метрика показывает абсолютную погрешность в единицах измерения, вторая — долю объяснённой вариации. Низкий RMSE при низком R^2 свидетельствует о том, что модель по сути предсказывает лишь

среднее значение; высокий RMSE при высоком R^2 указывает на наличие сильного тренда, но с большой ошибкой; оптимальная комбинация «низкий RMSE + высокий R^2 » характеризует точную и информативную модель.

Интерпретация значений смоделированных параметров

Солёность; $R^2 = 0.96$, $RMSE = 1.4$.

График реконструкции демонстрирует высокую степень соответствия модели данным (рис. 15.1, лев.). Реконструированные значения солёности (*Reconstructed*) очень близки к измеренным (*Salinity*) для большинства точек, особенно для рек с высокой солёностью (Кура, Самур, Артек), где точки плотно группируются вдоль линии 1:1. Остатки (*Residual*) малы — в основном в пределах ± 2 ед., большинство около 0 — за исключением первых точек по рекам Урал и Терек, где модель систематически занижает реальную солёность (отрицательные остатки до -2.7 , рис. 15.1, прав.). Это подтверждается высоким коэффициентом детерминации $R^2 = 0.96$, означающим, что 96% вариации солёности объясняется изменениями в сообществе кефалей, и относительно низким $RMSE = 1.4$ ед. солёности. (*Пояснение.* Реконструкция — процесс математического восстановления значений абиотического параметра среды на основе структуры биотического сообщества с использованием выявленных факторных закономерностей; это ключевой этап метода CABFAC, реализующий его основную концепцию «биота как индикатор среды».)

Распределение и плотность трёх видов кефалей — сингили, лобана и остроноса — чрезвычайно точно отражают градиент солёности в устьях исследованных рек. Модель на основе плотностных показателей трёх видов позволяет с высокой точностью предсказать уровень солёности, что свидетельствует о сильной зависимости их экологии от данного параметра. Солёность является доминирующим фактором, структурирующим сообщества кефалей в этом регионе. Видовой состав и обилие чётко меняются вдоль градиента от пресноводных условий (Урал) к морским (Артек), при этом каждый вид занимает специфическую нишу в рамках этого градиента.

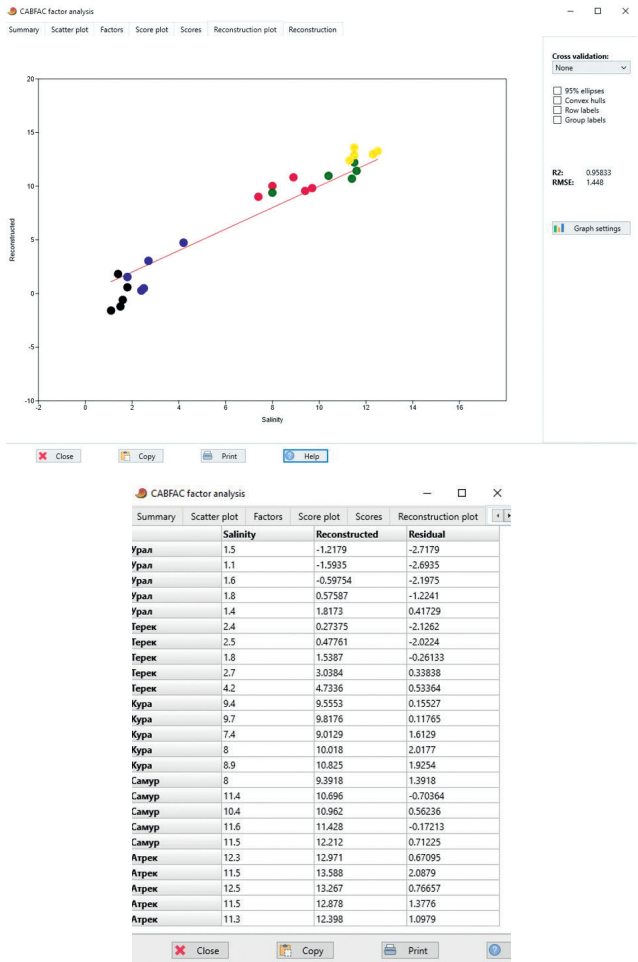


Рис. 15.1. Диаграмма рассеяния: реконструированные vs измеренные значения солёности; чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек, (верхний); *Reconstruction* результатов CABFAC-анализа (нижний)

Температура; $R^2 = 0.67$, $RMSE = 4.9$.

График реконструкции демонстрирует положительную, но слабую тенденцию с существенным разбросом точек вокруг линии тренда по сравнению с солёностью (рис. 15.2, верх.). Реконструированные значения (*Reconstructed*) часто заметно отличаются от измеренных.

Анализ остатков (*Residual*) выявляет систематические ошибки: для реки Урал (чёрные точки) модель значительно **занижает** температуру (большие отрицательные остатки, до -9.9 , рис. 15.2, ниж.), тогда как для рек Самур и Артек (зелёные и жёлтые точки) модель, наоборот, **завышает** показатели (положительные остатки до $+5.7$, рис. 15.2, ниж.). Высокое значение $RMSE = 4.9$ подтверждает низкую точность предсказания: в среднем ошибка модели составляет почти 5°C , что сопоставимо с естественным диапазоном колебаний температуры в исследуемых водоёмах. (*Пояснение:* Среднеквадратичная ошибка $RMSE = 4.9$ означает, что в среднем предсказания модели отклоняются от фактических измерений на 4.9 единицы; для температуры воды такая погрешность является критической и свидетельствует о том, что модель не может надёжно восстанавливать этот параметр).

Расположение рек в пространстве факторов, извлеченных из данных о рыбах, определяется в первую очередь градиентом солёности, а не температуры. Модель, обученная на данных о плотности кефалей, лишь умеренно ($R^2 = 0.67$) способна предсказать температуру воды, так как видовое распределение рыб в данном регионе лимитировано именно солёностью. Значительные остатки указывают на то, что вариация температуры не находит чёткого отражения в структуре сообщества кефалей. Вероятно, температура играет второстепенную роль или её влияние маскируется более сильным фактором солёности, из-за чего модель, «настроенная» на биотические реакции, не может выделить температурный сигнал. (*Пояснение:* Обучение модели — это процесс обобщения программой закономерностей из исходных данных; в данном случае программа «выучила», что рыбы реагируют на солёность, и пытается использовать эту связь для прогноза температуры, что приводит к ошибкам, так как прямой жёсткой связи «рыбы–температура» в данном наборе данных нет).

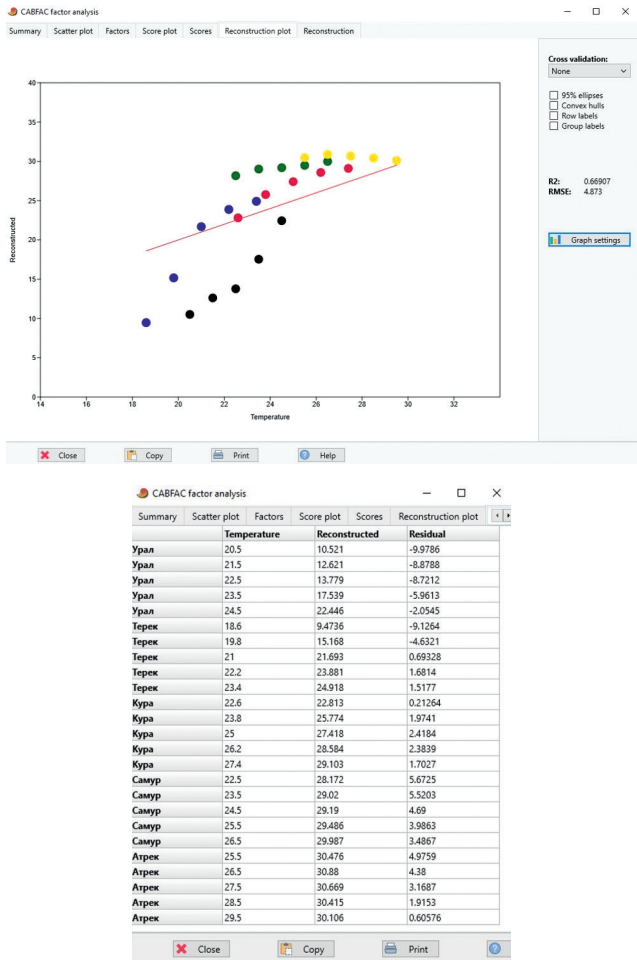


Рис. 15.2. Диаграмма рассеяния: реконструированные vs измеренные значения температуры; чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек (верхний); *Reconstruction* результатов CABFAC-анализа (нижний)

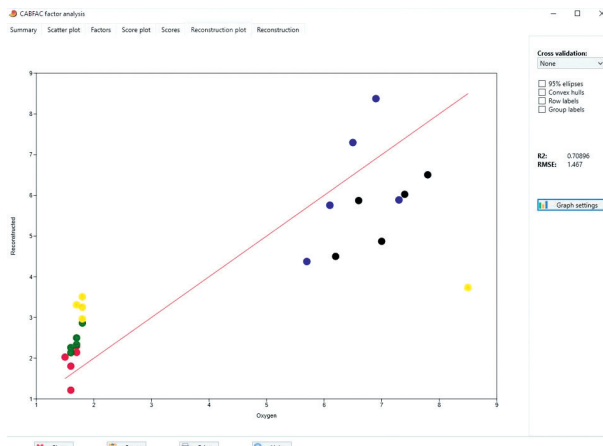
Кислород; $R^2 = 0.71$, $RMSE = 1.5$.

График реконструкции демонстрирует умеренную согласованность ($R^2 = 0.71$) (рис. 15.3, верх.). Анализ остатков (*Residual*) выявляет систематическую ошибку модели: она стремится «сжать» реальный градиент концентраций. Для опреснённых рек (Урал, Терек) модель

занижает уровень кислорода (реконструированные значения 4.5–7.3 мг/л против измеренных 6.2–7.8 мг/л, остатки отрицательные). Напротив, для высокосолёных рек (Самур, Артек) модель **завышает** показатели (реконструированные 2.1–3.7 мг/л против измеренных 1.6–1.8 мг/л, остатки положительные). Значение $RMSE = 1.5$ мг/л является приемлемым, но указывает на меньшую точность по сравнению с солёностью.

Особого внимания требует резкий выброс в данных по реке Артек: зафиксировано аномально высокое содержание кислорода (8.5 мг/л), тогда как модель, опираясь на общую структуру сообщества (характерную для солёных вод), предсказывает низкое значение (3.7 мг/л). Это даёт огромный отрицательный остаток (–4.76, рис. 15.3, ниж.), свидетельствующий о том, что в данной конкретной точке действуют локальные факторы (например, аэрация или сброс вод), не определяющий общий видовой состав рыб.

Резкое разделение рек на графике (Урал/Терек против Куры/Самура/Артека) обусловлено доминированием первой главной компоненты (PC1, 86.06%), которая отражает градиент солёности. Поскольку в данном регионе содержание кислорода обратно коррелирует с солёностью (в пресных водах его больше, в морских — меньше), модель успешно улавливает общий тренд. Однако недостаточная мощность второй компоненты (PC2, 13.48%) не позволяет скорректировать систематические ошибки «сжатия» диапазона. Таким образом, структура сообщества кефалей является хорошим, но не идеальным индикатором кислородного режима, так как локальные гидрохимические аномалии (как в случае с Артеком) не находят отражения в распределении рыб.



	Oxygen	Reconstructed	Residual
Урал	6.2	4.5008	-1.6992
Урал	6.6	5.8734	-0.7266
Урал	7	4.8714	-2.1286
Урал	7.4	6.0255	-1.3745
Урал	7.8	6.5049	-1.2951
Терек	5.7	4.3748	-1.3252
Терек	6.1	5.7575	-0.34246
Терек	6.5	7.2973	0.79729
Терек	6.9	8.3766	1.4766
Терек	7.3	5.8881	-1.4119
Кура	1.5	2.0265	0.52653
Кура	1.6	1.2149	-0.38511
Кура	1.6	1.8042	0.20425
Кура	1.7	2.142	0.44199
Кура	1.7	2.2951	0.59506
Самур	1.6	2.1354	0.5354
Самур	1.6	2.2609	0.66085
Самур	1.7	2.3351	0.63513
Самур	1.7	2.4989	0.79893
Самур	1.8	2.8633	1.0633
Артек	1.7	3.3104	1.6104
Артек	8.5	3.7358	-4.7642
Артек	1.8	3.5076	1.7076
Артек	1.8	3.2507	1.4507
Артек	1.8	2.9645	1.1645

Рис. 15.3. Диаграмма рассеяния: реконструированные vs измеренные значения кислорода; чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек, (верхний); *Reconstruction* результатов CABFAC-анализа (нижний)

Органика; $R^2 = 0.67$, $RMSE = 4.1$.

График реконструкции демонстрирует слабую положительную тенденцию с очень большим разбросом точек ($R^2 = 0.67$, $RMSE = 4.1$) (рис. 15.4, верх.). Остатки (*Residual*) достигают экстремальных значений, указывая на систематические ошибки модели для конкретных рек. Например, для реки Самур (зелёные точки) модель существенно **завышает** содержание органики (остатки положительные, до +8.4 (рис. 15.4, ниж.): при измеренных 3.7 ед. модель предсказывает 12.1 ед.). Напротив, для реки Кура (красные точки) модель грубо **занижает** показатели (остатки отрицательные, до -10.0 (рис. 15.4, ниж.): при измеренных 18.8 ед. модель предсказывает лишь 8.8 ед.). Высокое значение $RMSE = 4.1$ и визуальный хаос на графике свидетельствуют о том, что модель даёт неадекватные предсказания концентрации органического вещества на основе данных о рыбах; чёткой связи между измеренными и реконструированными значениями не прослеживается.

Пространство факторов, извлеченных из данных о сообществах рыб, практически не несёт полезной информации для предсказания содержания органики. Умеренное значение $R^2 = 0.67$, вероятно, возникает лишь как артефакт косвенной корреляции через солёность или речной сток

(например, в устьях крупных рек может наблюдаться специфический набор видов и одновременно высокий уровень органики), но сама по себе вариация органического вещества плохо отражается в изменении плотности и состава трёх видов кефалей. Органика, по-видимому, не является ключевым драйвером (лимитирующим фактором) для формирования данных конкретных сообществ в исследованных устьях, либо её влияние полностью маскируется доминирующим градиентом солёности.

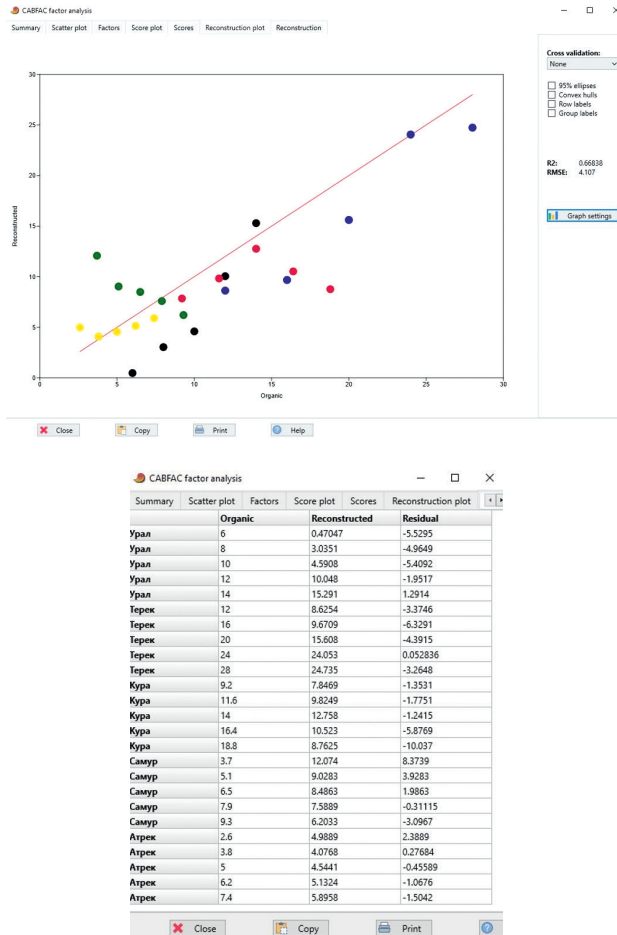


Рис. 15.4. Диаграмма рассеяния: реконструированные vs измеренные значения органики; чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек, (верхний); Reconstruction результатов CABFAC-анализа (нижний)

Кислотность; $R^2 = 0.22$, $RMSE = 2.1$.

График реконструкции демонстрирует крайне слабую связь между измеренными и предсказанными значениями ($R^2 = 0.22$), что указывает на неэффективность модели для данного параметра (рис. 15.5, верх.). Данные имеют выраженный бимодальный характер: значения pH группируются либо в зоне крайне низких показателей (~ 1.7 – 1.8), либо в зоне высоких (~ 8.1 – 8.5). Модель CABFAC не способна воспроизвести эту структуру, выдавая прогнозы преимущественно в промежуточном диапазоне (от 1.7 до 5.0). В результате для точек с высоким pH (8.1–8.5) модель даёт сильное занижение (например, для Самура при измеренных 8.1 модель предсказывает 2.8, остаток -5.3 , рис. 15.5, ниж.), а для части точек с низким pH (1.7–1.8) — завышение (например, для Артека при измеренных 1.8 модель предсказывает 4.7, остаток $+2.9$, рис. 15.5, ниж.). Высокое значение $RMSE = 2.1$ подтверждает, что средняя ошибка предсказания сопоставима с половиной всего диапазона изменчивости параметра.

Пространство факторов, сформированное сообществом кефалей, не содержит информации, необходимой для реконструкции кислотности среды. Столь низкий коэффициент детерминации ($R^2 = 0.22$) свидетельствует о том, что вариация pH не отражается в изменении плотности и видового состава рыб. Вероятно, кислотность не является лимитирующим фактором для данных видов в исследуемом диапазоне, либо ее влияние полностью маскируется более мощным градиентом солёности. Наличие аномальных значений pH (как крайне низких, так и высоких) при сохранении стабильного рыбного сообщества указывает на высокую толерантность исследуемых видов к колебаниям кислотности или на то, что данный параметр не играет роли в структурообразовании ихтиоценозов в устьях данных рек.

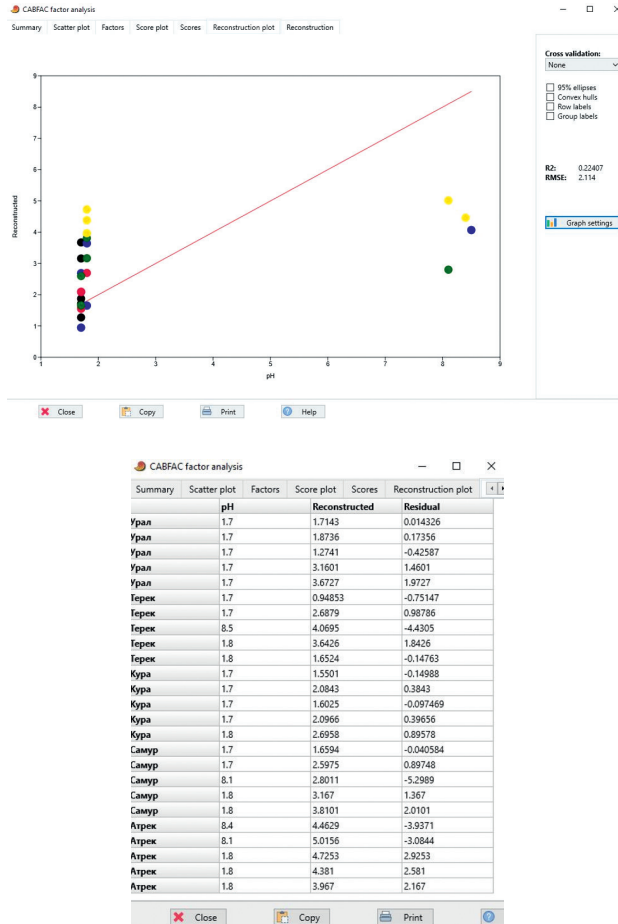


Рис. 15.5. Диаграмма рассеяния: реконструированные vs измеренные значения кислотности; чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек, (верхний); *Reconstruction* результатов CABFAC-анализа (нижний)

Диаграмма рассеяния в пространстве главных компонент

Идентичность графиков рассеяния («*Factor 1* vs. *Factor 2*») для всех анализируемых параметров среды обусловлена методологией CABFAC. На первом этапе анализа строится единая факторная модель, основанная исключительно на матрице биологических данных (обилии видов). Эта модель фиксирует внутреннюю структуру

ихтиоценоза, выделяя основные градиенты изменчивости: доминирующий фактор (PC1), объясняющий 89,65% дисперсии (который, как установлено ранее, соответствует градиенту солёности), и второй фактор (PC2), описывающий 10,31% изменчивости (рис. 15.6). Поскольку координаты образцов в этом факторном пространстве определяются только биотическими данными, ординационная диаграмма остается неизменной при переключении между различными абиотическими переменными.

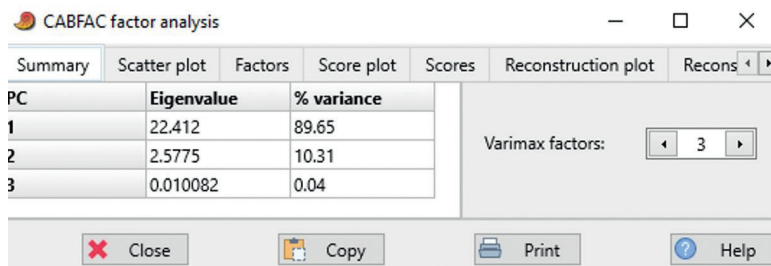


Рис. 15.6. Вкладка *Summary*, % объяснённой дисперсии для первых трёх главных компонент

Контраст графиков реконструкции возникает на втором этапе анализа CABFAC: после выделения единого факторного пространства на основе матрицы биотических данных, для каждого абиотического параметра строится индивидуальная регрессионная модель. Качество реконструкции напрямую отражает силу связи параметра с биотической структурой: от практически идеального соответствия для солёности ($R^2 = 0.96$) до хаотичного разброса для pH ($R^2 = 0.22$) и выраженных систематических ошибок для температуры и кислорода. Таким образом, CABFAC визуализирует общую структуру сообщества через единый ординационный график, но отдельно и количественно оценивает пригодность этой структуры для восстановления конкретных средовых условий.

График на рис. 15.7 демонстрирует чёткую линейную упорядоченность станций вдоль оси Factor 1. Значения первого фактора монотонно возрастают от р. Урал (диапазон ~ 0.02 – 0.73) и р. Терек (~ 0.28 – 0.88) к р. Кура (~ 0.82 – 0.95), р. Самур и р. Артек (~ 0.95 – 0.96). Эта последовательность полностью согласуется с установленным градиентом солёности: от пресноводных/опреснённых участков (Урал,

Терек) через переходные зоны (Кура) к высокоминерализованным водам (Самур, Артек). Следовательно, Factor 1 выступает интегральным индикатором солёности и основным драйвером пространственного структурирования ихтиоценоза. Ось Factor 2 характеризуется стабильно отрицательными значениями для всех объектов (от -0.3 до -1.0) и объясняет лишь незначительную долю дисперсии (10.31%); она слабо дифференцирует реки и, вероятно, отражает второстепенные внутригрупповые вариации или артефакты процедуры вращения, не имеющие выраженного экологического смысла в данном контексте.

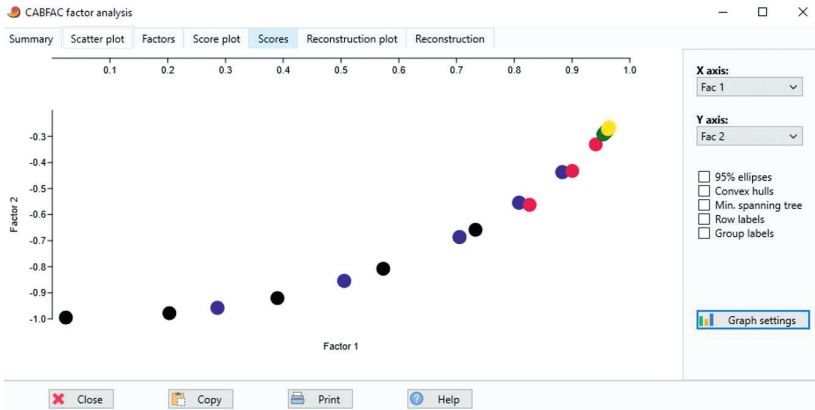


Рис. 15.7. Диаграмма рассеяния в пространстве главных компонент (*Factor 1 vs Factor 2*); чёрный — р. Урал, синий — р. Терек, красный — р. Кура, зелёный — р. Самур, жёлтый — р. Артек

Факторные нагрузки

Вкладка *Scores* (рис. 15.7) содержит матрицу факторных нагрузок (*factor loadings*) для трёх видов кефалей. Значения нагрузок идентичны для всех пяти анализируемых абиотических параметров, поскольку метод CABFAC на первом этапе строит единую факторную модель исключительно на основе биотических данных — структуры ихтиоценоза, не зависящей от конкретных параметров среды. Эта модель фиксирует объективные оси изменчивости сообщества: Factor 1, объясняющий 86,06% дисперсии (доминирующий градиент солёности), и Factor 2, объясняющий 13,48% дисперсии. Каждому виду присваиваются неизменные коэффициенты корреляции с этими осями. Последующий анализ параметров среды не пересчитывает биотические

нагрузки, а применяет к факторным координатам станций индивидуальные регрессионные модели для реконструкции значений. Таким образом, постоянство нагрузок является прямым следствием методологии CABFAC: биотическая структура первична, а её связь со средой диагностируется через точность реконструкции (R^2/RMSE), а не через модификацию вкладов видов. Распределение рыб формирует «статичную карту» сообщества, а параметры среды проверяются на соответствие этой карте через отдельные регрессионные «ключи».

Сингиль (*Liza aurata*). Характеризуется высокой положительной нагрузкой на Factor 1 (1,4408), что указывает на его роль как доминантного вида, строго ассоциированного с градиентом солёности (положительный полюс оси). Практически нулевая нагрузка на Factor 2 (0,013) свидетельствует о слабой зависимости от второстепенных факторов, что подтверждает его узкую экологическую нишу как индикатора морских (высокосолёных) условий.

Лобан (*Mugil cephalus*). Демонстрирует умеренную положительную нагрузку на Factor 1 (0,861), что отражает его эвригалинность — способность осваивать широкий диапазон солёности, хотя и с меньшим сродством к морским условиям, чем у сингиля. Отрицательная нагрузка на Factor 2 (–0,788) предполагает чувствительность к дополнительным факторам (например, кислородному или температурному режиму), однако их вклад в распределение вида субдоминантен относительно основного градиента солёности.

Остронос (*Liza saliens*). Имеет отрицательную нагрузку на Factor 1 (–0,427), что идентифицирует его как вид, тяготеющий к опреснённым участкам (отрицательный полюс оси) и избегающий высокой солёности. Экстремально низкая (отрицательная) нагрузка на Factor 2 (–1,543) указывает на то, что данный вид является ключевым индикатором специфических условий, описываемых второй осью (возможно, связанных с характеристиками субстрата или кормовой базы в пресных водах), которые не совпадают с доминирующим градиентом солёности.

CABFAC factor analysis

Summary	Scatter plot	Factors	Score plot	Scores	Reconstruction plot	Reconstruction
		Factor 1	Factor 2	Factor 3		
сингиль		1.4408	0.013118	-0.96117		
лобан		0.86098	-0.78777	1.2799		
остронос		-0.42746	-1.5425	-0.66183		

Close Copy Print Help

Рис. 15.8. Таблица факторных нагрузок для трёх видов кефалей

Общий вывод

Ключевой результат анализа — исключительно сильная связь структуры сообществ кефалей с солёностью ($R^2 = 0.96$, $RMSE = 1.4$) — имеет чёткую гидрологическую основу. Исследуемые реки (Урал, Терек, Кура, Самур, Артек) существенно различаются по размерам водосборов и величине речного стока. Крупные реки, такие как Урал и Терек, приносят в Каспийское море огромные объёмы пресной воды, формируя мощные зоны опреснения в эстуариях. Напротив, относительно небольшие реки, особенно Артек, имеют значительно меньший сток, и их влияние на солёность прилегающих морских вод слабее. Таким образом, существует обратная зависимость между размером реки/объёмом стока и солёностью в устьевой зоне: чем крупнее река, тем сильнее опреснение. Выявленный градиент расположения рек вдоль Factor 1 (Урал → Терек → Кура → Самур → Артек) в пространстве, определяемом плотностью кефалей, в первую очередь отражает именно этот градиент солёности. Виды кефалей демонстрируют чёткую видовую специфичность и различия в обилии вдоль данного градиента, выступая надёжными биоиндикаторами ключевого абиотического параметра эстуарных экосистем Каспийского моря, что и обуславливает высокую точность реконструкции солёности моделью CABFAC.

Температура и кислород демонстрируют умеренную связь с сообществами рыб ($R^2 = 0.67$ и 0.71 соответственно), но эта связь, вероятно, опосредована солёностью, поскольку градиенты этих параметров часто сопряжены. Их собственное независимое влияние либо слабее, либо имеет более сложный характер, о чём свидетельствуют более высокие ошибки предсказания ($RMSE = 4.9$ и 1.5) и наличие систематических выбросов, особенно для кислорода. Содержание органики

и уровень кислотности (pH) практически не отражаются в структуре исследованных сообществ кефалей. Модель показывает неудовлетворительные результаты реконструкции этих параметров: для органики при формально умеренном $R^2 = 0.67$ наблюдается неприемлемо высокая ошибка $RMSE = 4.1$, а для pH модель практически неработоспособна ($R^2 = 0.22$). Следовательно, органика и кислотность не являются значимыми драйверами распределения данных видов в исследованных устьевых зонах.

Философское эхо

CABFAC — это повар сходства, чья кухня — матрица признаков, а ингредиенты — исходные меры объектов. Он готовит блюда, где приправой служит нормировка, а вкус рождается из выверенных пропорций, уравнивающих силу каждого компонента. Исходные данные подобны рецептуре: букет признаков в пробе — как сочетание специй, а частоты — точные дозировки, формирующие вкусовой профиль блюда. Как опытный шеф доверяет не отдельным компонентам, а их балансу и взаимодействию, так и метод раскрывает смысл через согласованность сочетаний, а не через изолированные величины. Сила CABFAC — в кулинарной редукции: он не перебирает все пары сходств, а уваривает перекрёстные связи между объектами, концентрируя суть в плотном экстракте, из которого выделяются факторы — дегустационные шкалы, выявляющие скрытые профили и формирующие базовый набор вкусов для групп объектов. Однако избыток ингредиентов нарушает чистоту бульона: россыпь бинарных данных делает текстуру неоднородной, а сложные пространственно-временные зависимости, подобно взвеси, мутят отвар и сбивают вкусовой баланс. Тем не менее, повар, мастерски варьируя режимы приготовления — перестановочные тесты, — проводит слепую дегустацию, случайным образом меняя составы, чтобы оценить устойчивость результата. Важно помнить: CABFAC предлагает меню родства, где группы объектов — узнаваемые блюда по их гармоничному набору компонентов, а факторы — фундаментальные вкусы (умами, сладость, кислота), задающие палитру различий. В руках исследователя-шефа, умеющего переводить язык пропорций в гастрономическое описание, метод превращает таблицу мер в выверенное меню, где каждая строка — рецепт состояния, а факторные оси — линии дегустации, раскрывающие сущность объекта через призму его собственных, согласованных мер.

16. МЕТОД ANOSIM

Введение

Водные экосистемы характеризуются сложной организацией сообществ, видовой состав и структура которых определяются комплексом абиотических и биотических факторов: температурным режимом, кормовой базой, антропогенным воздействием и естественными циклами. При анализе многомерных данных по встречаемости и обилию видов на множестве станций возникает ключевая задача — выявление и статистическое подтверждение различий между априорно определёнными группами проб. В конце 1980-х – начале 1990-х годов эта проблема приобрела особую актуальность в связи с необходимостью анализа сложных экологических массивов. Параметрические методы дисперсионного анализа часто оказывались неприменимы вследствие специфики биологических данных: нарушения предположения о нормальности распределения, высокой доли нулевых значений и сложных нелинейных взаимосвязей. Разработка надёжного статистического инструмента для анализа различий в многомерном пространстве стала насущной необходимостью, что привело к созданию метода ANOSIM (*Analysis of Similarities*).

Метод ANOSIM был разработан в Плимутской морской лаборатории (Великобритания) К. Р. Кларком (K. R. Clarke) и представлен в работе 1993 года. Кларк систематизировал подход, основанный на ранговом анализе матриц несходства, который активно развивался в контексте интерпретации результатов неметрического многомерного шкалирования nMDS. ANOSIM был предложен как универсальный непараметрический инструмент для проверки гипотезы о наличии статистически значимых различий в видовом составе между заранее заданными группами образцов. Разработка метода была обусловлена практической потребностью экологов в анализе зашумлённых данных реальных сообществ, где традиционные параметрические подходы не обеспечивали достоверных результатов.

Методологическая основа ANOSIM предполагает анализ матрицы несходства, рассчитанной между всеми парами проб. Метод использует экологически релевантные индексы, такие как Брея–Кёртиса или Жаккара, учитывающие качественные и/или количественные аспекты данных. Полученные значения несходства преобразуются в ранги, после чего

проводится сравнение средних рангов для пар внутри групп и между группами. Статистическая значимость различий оценивается с помощью R-статистики, количественно характеризующей степень разделения групп.

В исследованиях водных сообществ ANOSIM получил широкое применение вследствие своей устойчивости к особенностям биологических данных. Метод эффективен для анализа структуры ихтиоценозов и позволяет статистически подтверждать различия, визуально наблюдаемые на ординационных графиках. С помощью ANOSIM стало возможным количественно оценивать воздействие антропогенных факторов, выявлять различия между сообществами на охраняемых и эксплуатируемых акваториях, а также анализировать временную динамику. Основное назначение метода заключается в проверке глобальной гипотезы о наличии статистически значимых различий между группами проб.

Таким образом, ANOSIM представляет собой непараметрический подход к анализу многомерных экологических данных, основанный на сравнении ранговых мер несходства. К преимуществам метода относятся его непараметрическая природа, возможность работы с типичными «разреженными» биологическими данными и интерпретируемость R-статистики как меры выраженности групповых различий. Многолетнее успешное применение ANOSIM в разнообразных водных средах подтверждает его надёжность как инструмента для объективного выявления различий в структуре сообществ.

Методология

Методологическую основу ANOSIM составляет математический аппарат, позволяющий трансформировать многомерную информацию о составе биологических сообществ в формализованную статистическую гипотезу о значимости межгрупповых различий. Каждая проба (в контексте ихтиологических исследований — проба ихтиофауны) концептуально представляет собой точку в многомерном пространстве, где каждая ось соответствует обилию конкретного вида. Для преодоления аналитических сложностей, связанных с высокоразмерным пространством, ANOSIM оперирует не исходными координатами, а матрицей попарных расстояний между всеми пробами. Эти расстояния рассчитываются с использованием экологически релевантных мер несходства (например, индекс Брея–Кёртиса), количественно отражающих степень различия в видовом составе. Меньшие значения расстояния указывают на большее сходство проб, тогда как большие значения свидетельствуют о существен-

ных различиях. Исходные многомерные данные таким образом преобразуются в симметричную матрицу несходства, содержащую полный набор дистанционных мер между всеми возможными парами проб.

Критически важный этап, определяющий непараметрическую природу ANOSIM, — процедура ранжирования всех парных расстояний, извлечённых из матрицы несходства. Все возможные пары проб упорядочиваются в единый ранжированный ряд: от минимального расстояния (ранг 1) до максимального (максимальный ранг). Процедура ранжирования учитывает исключительно относительный порядок величин, абстрагируясь от их абсолютных значений. Такой переход к порядковой шкале эффективно снижает чувствительность метода к выбросам, отклонениям от параметрических предположений о распределении данных и другим источникам вариабельности, характерным для экологических массивов.

Фундаментальная логика ANOSIM заключается в сравнительном анализе рангов расстояний для двух категорий пар: внутригрупповых (пробы, принадлежащие к одной заранее определённой группе) и межгрупповых (пробы из разных групп). Если группы действительно экологически различимы, то ранги расстояний для внутригрупповых пар должны статистически значимо уступать таковым для межгрупповых пар. Метод вычисляет средние ранги для обеих категорий, после чего рассчитывается ключевая R-статистика как нормированная разность между средним рангом межгрупповых и внутригрупповых расстояний: $R = (\bar{r}_B - \bar{r}_W) / \frac{N(N-1)}{4}$, где $\bar{r}_B$ — средний ранг всех межгрупповых попарных расстояний. $\bar{r}_W$ — средний ранг всех внутригрупповых попарных расстояний. N — общее количество проб (образцов) во всех группах вместе, а $\frac{N(N-1)}{4}$ — теоретический максимум разности рангов при идеальном разделении. Значение R, приближающееся к +1, указывает на выраженное разделение групп (межгрупповые различия существенно превышают внутригрупповые). Значение около 0 свидетельствует об отсутствии различий, а отрицательные значения (встречающиеся редко) указывают на большую схожесть между группами, чем внутри них, что обычно сигнализирует о проблемах в определении группирующих факторов.

Вычисленное значение R-статистики само по себе не является доказательством статистической значимости. Для оценки значимости ANOSIM использует процедуру перестановочного тестирования (*permutation test*),

моделирующую распределение статистики при справедливости нулевой гипотезы об отсутствии различий. Метки групп случайным образом перераспределяются между пробами многократно, при этом исходная матрица несходства остаётся неизменной. Для каждой перестановки вычисляется новое значение R , совокупность которых формирует эмпирическое нулевое распределение. Статистическая значимость наблюдаемого значения оценивается как доля случайных перестановок, в которых полученное значение R было равно или превышало наблюдаемое. Малая доля (обычно $p < 0.05$) обеспечивает основание для отклонения нулевой гипотезы. Эта процедура делает выводы ANOSIM устойчивыми к нарушениям параметрических предположений. Таким образом, математическая сущность метода включает последовательное преобразование многомерных данных в матрицу несходства, ранжирование расстояний, сравнение групп через R -статистику и оценку значимости через перестановочный тест.

Характер данных

Метод ANOSIM представляет собой ценный инструмент для проверки гипотез о значимых различиях в структуре сообществ между заранее определёнными группами. В отличие от методов, требующих нормальности распределения и гомогенности дисперсий, ANOSIM предлагает устойчивый подход, основанный на анализе рангов мер несходства, что делает его особенно полезным при работе с типичными экологическими данными. Метод находит применение в сценариях, где исследователь имеет чёткие априорные гипотезы о группировке проб: оценка воздействия антропогенных факторов, анализ сезонной динамики или сравнение разных типов местообитаний.

Метод демонстрирует наибольшую эффективность при анализе данных с выраженной групповой структурой, где внутригрупповое сходство существенно превышает межгрупповое. Он успешно применяется для работы с количественными данными по обилию видов, данными присутствия–отсутствия, а также с производными показателями (биомасса, индексы разнообразия). Типичные примеры использования: сравнение ихтиоценозов выше и ниже плотины, анализ сезонной динамики, оценка эффективности морских охраняемых акваторий, изучение вертикальной зональности. Важным условием является наличие достаточного количества реплик в каждой группе и относительная гомогенность внутригрупповой вариабельности. Метод особенно результативен при сравнении сообществ, подвер-

женных воздействию дискретных факторов среды (типы донных субстратов, уровни антропогенной нагрузки). Высокая чувствительность достигается при контрастных условиях, формирующих чётко дифференцированные группировки гидробионтов. Важную роль играет корректный выбор меры сходства, адаптированной к специфике данных. При соблюдении методологических предпосылок анализ обеспечивает надёжное выявление статистически значимых различий даже в условиях высокой пространственно-временной изменчивости.

Вместе с тем применение ANOSIM связано с определёнными ограничениями. Метод малоэффективен для анализа плавных градиентов, где искусственное разделение на дискретные группы может искажать реальные закономерности. Он также чувствителен к гетерогенности дисперсий: различия в вариабельности между группами могут интерпретироваться как различия в положении центроидов. Мощность метода снижается при большом количестве групп с малым размером выборки, а также при анализе слабых эффектов на фоне высокой естественной вариабельности. Важным ограничением является неспособность метода анализировать взаимодействия между факторами в сложных экспериментальных планах, в отличие от более современных подходов, таких как PERMANOVA. Метод может давать ложноположительные результаты при нарушении предположения о внутригрупповой гомогенности, когда высокая вариабельность внутри групп маскирует истинные межгрупповые различия. Кроме того, ANOSIM не предоставляет информации о величине эффекта отдельных факторов или о вкладе конкретных видов в наблюдаемые различия. Эти ограничения следует учитывать при планировании исследования и выборе статистического инструмента.

К ключевым преимуществам ANOSIM относятся его непараметрическая природа, устойчивость к нарушениям предположений о распределении данных и интуитивно интерпретируемая R-статистика. Метод обеспечивает надёжную оценку значимости через перестановочные тесты и хорошо подходит для первоначальной оценки различий между группами. Однако к его недостаткам следует отнести неспособность выявлять конкретные пары различающихся групп без дополнительных попарных сравнений, а также невозможность определения вклада отдельных видов в наблюдаемые различия. Выбор ANOSIM должен основываться на соответствии структуры данных его методологическим предпосылкам и конкретным задачам исследования.

Пример анализа данных. One-way ANOSIM

При данном типе анализе используются один категориальный фактор.

Задача. Оценили скорость плавания молоди плотвы в длинах тела в секунду (BL/s) при 4-х различных температурах на фоне 4-х различных уровней кислотности среды. Имеются ли различия в скорости плавания при различных температурах на фоне разной кислотности? Если имеются, то какой. Построить диаграммы размаха.

АНАЛИЗ ДАННЫХ

Тест ANOSIM размещён в PAST в главном меню *Multivariate* → *Tests* → ANOSIM.

Вкладка *Summary*

На основе проведённого однофакторного анализа сходства (1-way ANOSIM) получены статистически значимые результаты, позволяющие утверждать, что температура воды оказывает существенное влияние на многомерный профиль скорости плавания молоди плотвы, измеренный на фоне разной кислотности среды (рис. 16.1). Значение R-статистики, равное 0,1582, указывает на наличие статистически значимого, хотя и умеренного по величине эффекта группировки: сходство особей внутри температурных групп превышает их сходство с особями из других групп, однако разделение не является абсолютным. Величина *p-value* (0,0021) значительно ниже общепринятого порога значимости 0,05, что позволяет уверенно отвергнуть нулевую гипотезу об отсутствии различий и сделать вывод о наличии статистически значимых различий в многомерном отклике скорости плавания как минимум между одной парой температурных условий.

Высокая достоверность этого вывода подкреплена строгой методологией проведения анализа: в качестве меры сходства использовался индекс Брея–Кертиса, адекватный для анализа экологических данных такого типа, а для расчета значимости было применено 9999 перестановок (*permutations*), что обеспечивает высокую надёжность оценки *p-value*. Это означает, что выявленные различия не являются случайными и действительно обусловлены влиянием температурного фактора, который модулирует поведение рыб в условиях различного pH.

Таким образом, можно заключить, что температура воды является значимым фактором, определяющим скорость плавания молоди плотвы. Для выявления конкретных пар температурных групп, между которыми существуют различия, необходимо провести попарные сравнения (вкладка *Pairwise*).

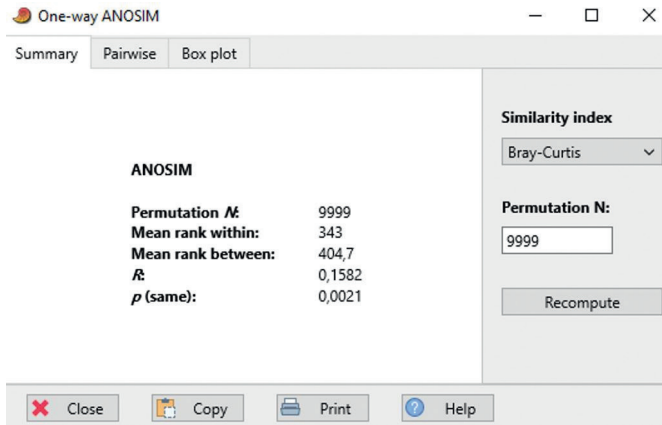


Рис. 16.1. Результаты однофакторного анализа сходства (ANOSIM) на основе индекса Брея-Кёртиса

Вкладка *Pairwise*

При проведении попарных сравнений между группами возникает проблема множественного тестирования, увеличивающая вероятность ложноположительных результатов (рис. 16.2). Для коррекции *p-values* в PAST предлагается три метода. Нескорректированные *p-values* (*uncorrected significance*) подходят для предварительного разведочного анализа, но существенно повышают риск ошибки первого рода. Классическая поправка Бонферрони (*Bonferroni-corrected p-values*) является наиболее консервативной и строгой, однако значительно увеличивает риск ложноотрицательных результатов. Последовательная поправка Бонферрони (*sequential Bonferroni significance*) представляет собой сбалансированный подход, контролирующий общую ошибку при сохранении более высокой чувствительности к реальным различиям по сравнению с классическим методом. Учитывая исследовательский характер анализа, для выявления потенциально значимых закономерностей рекомендуется использование именно последовательной поправки Бонферрони.

Результаты попарного сравнения с применением последовательной поправки Бонферрони демонстрируют наличие статистически значимых различий в скорости плавания молоди плотвы между определёнными температурными группами (матрица симметрична относительно диагонали). Установлено, что основные различия обусловлены специфическим положением группы Temp_2. Достоверные различия выявлены между группой Temp_2 и группами Temp_3 ($p = 0,0018$) и Temp_4 ($p = 0,007$). В то же время отсутствие значимых различий между другими парами групп (Temp_1 vs Temp_2: $p = 0,1492$; Temp_1 vs Temp_3: $p = 0,1503$; Temp_1 vs Temp_4: $p = 0,0956$; Temp_3 vs Temp_4: $p = 0,1089$) позволяет выделить два кластера условий: обобщённую группу (Temp_1, Temp_3, Temp_4), в рамках которой изменение температуры не приводит к существенным изменениям плавательной активности, и изолированную группу Temp_2, оказывающую выраженное специфическое воздействие на изучаемый признак.

Полученные результаты свидетельствуют о том, что выявленные различия носят не линейный характер, а обусловлены специфическим эффектом температуры 2. Это указывает на наличие определённого температурного порога или оптимума, при котором реакция организма (скорость плавания) на комплексное воздействие температуры и кислотности среды проявляется наиболее выражено. Данный вывод имеет значение для понимания физиологических механизмов адаптации рыб к многофакторному воздействию абиотических факторов и может быть использован для прогнозирования реакций гидробионтов в изменяющихся условиях среды.

One-way ANOSIM

Summary Pairwise Box plot

p-values, sequential Bonferroni significance

	Temp_1	Temp_2	Temp_3	Temp_4
Temp_1		0,1492	0,1503	0,0956
Temp_2	0,1492		0,0018	0,007
Temp_3	0,1503	0,0018		0,1089
Temp_4	0,0956	0,007	0,1089	

Close Copy Print Help

Рис. 16.2. Матрица попарных сравнений групп методом ANOSIM с последовательной поправкой Бонферрони (*sequential Bonferroni correction*)

Вкладка *Box plot*

Анализ диаграмм размаха, построенных на основе ранжированных расстояний между объектами, позволяет визуализировать характер распределения сходства внутри температурных групп и между ними (рис. 16.3). По оси ординат отложены ранги расстояний — преобразованные в порядковые величины значения меры различия между объектами, где более высокий ранг соответствует большему различию. Особый интерес представляет группа Temp_3, которая демонстрирует наименьший межквартильный размах и самое низкое медианное значение ранжированного расстояния (около 150). Это указывает на высокую однородность и стабильность поведения молоди плотвы при данной температуре: особи демонстрируют практически идентичные профили скорости плавания, то есть внутригрупповые различия минимальны.

Медианные значения и межквартильный размах демонстрируют, что различия между температурными группами носят не случайный, а систематический характер. Высокая однородность группы Temp_3 (низкий медианный ранг внутригрупповых расстояний) контрастирует с группой Temp_2, которая характеризуется более высоким медианным значением (около 450) и большим разбросом показателей. Именно компактность кластера Temp_3 и его низкая внутригрупповая вариабельность являются ключевыми факторами, объясняющими статистическую значимость различий с другими группами, выявленную в попарных сравнениях. Визуальное разделение ящиков, особенно для группы Temp_3, подчеркивает надёжность обнаруженных различий между температурными условиями.

Полученные результаты свидетельствуют о том, что температура 3 представляет собой условие, стабилизирующее плавательную активность молоди плотвы. Данный температурный режим, по-видимому, минимизирует влияние вариаций кислотности на поведение рыб, что указывает на наличие специфического температурного оптимума для данного вида. Этот вывод имеет важное значение для понимания механизмов адаптации гидробионтов к комплексному воздействию абиотических факторов в изменяющейся среде.

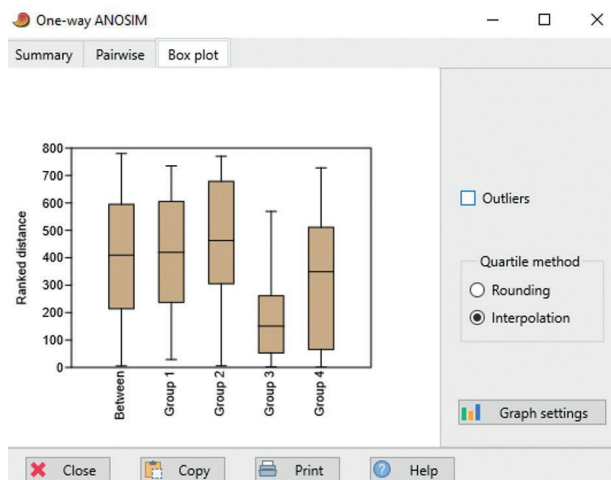


Рис. 16.3. Диаграмма размаха для значений четырёх температур

Пример анализа данных. Two-way ANOSIM

Частный случай n -way ANOSIM; при данном типе анализа используются два категориальных фактора.

Задача. У двухгодовиков четырёх пород карпа — парский (П), ангелинский (А), московский (М) и чувашский (Ч) — измерили приросты массы тела в граммах в единицу времени, $\Delta W/\Delta t$, при выращиваниях на кормах четырёх типов, А–D, и двух величинах температуры, 1–2. Различаются ли величины приростов у данных видов в зависимости от типа корма и температуры? Оценить эффект взаимодействия.

АНАЛИЗ ДАННЫХ

Программа PAST применительно к ANOSIM позволяет анализировать влияние не более двух факторов одновременно. Для исследований с более сложным дизайном, включающим три и более факторов, необходимо использовать специализированные программные пакеты, такие как Primer-e или Canoco. Однако принципы интерпретации результатов многомерного анализа универсальны независимо от программного обеспечения и количества анализируемых факторов. Следует также отметить, что для 2-way ANOSIM в PAST не предусмо-

трены попарные сравнения и построение диаграмм размаха, а также оценка взаимодействия факторов (*interaction effect*), что является существенным ограничением метода в данной программе.

На основании проведённого двухфакторного анализа сходства (2-way ANOSIM) установлено (рис. 16.4), что как фактор 1 (ФАКТОР_1), так и фактор 2 (ФАКТОР_2) оказывают статистически значимое влияние на величину прироста массы тела у двухгодовиков исследуемых пород карпа. Анализ, основанный на метрике Брея–Кёртиса и подтверждённый перестановочным тестом с 9999 повторениями, показал статистически значимые различия для обоих факторов: для ФАКТОР_1 ($R = 0,12283$; $p = 0,0238$) и для ФАКТОР_2 ($R = 0,18229$; $p = 0,03$). Статистика R в ANOSIM оценивает степень разделения групп: значение 0 указывает на полное отсутствие различий (случайное распределение), а 1 — на максимально возможное разделение (полное непересечение групп). Значения R в диапазоне 0,12–0,18 свидетельствуют о слабом, но статистически детектируемом эффекте, что означает наличие устойчивых различий между группами, которые, однако, не являются выраженными на фоне внутригрупповой вариабельности. Выбор метрики Брея–Кёртиса особенно актуален для биологических данных, так как она учитывает относительные пропорции показателей и менее чувствительна к выбросам и нулевым значениям по сравнению с евклидовым расстоянием.

Таким образом, можно заключить, что оба изучаемых фактора вносят самостоятельный статистически значимый вклад в изменчивость показателей прироста массы, причём фактор 2 ($R = 0,18$) оказывает несколько более сильное влияние, чем фактор 1 ($R = 0,12$). Полученные результаты подтверждают необходимость учёта обоих факторов в рыбоводстве, но также указывают на важность других неучтённых факторов (возможно, индивидуальных особенностей рыб, взаимодействия факторов или условий среды), влияющих на рост рыб. Отсутствие в PAST возможности оценки взаимодействия факторов в рамках 2-way ANOSIM не позволяет определить, зависит ли эффект одного фактора от уровня другого, что требует применения альтернативных методов (например, PERMANOVA) для более полного анализа.

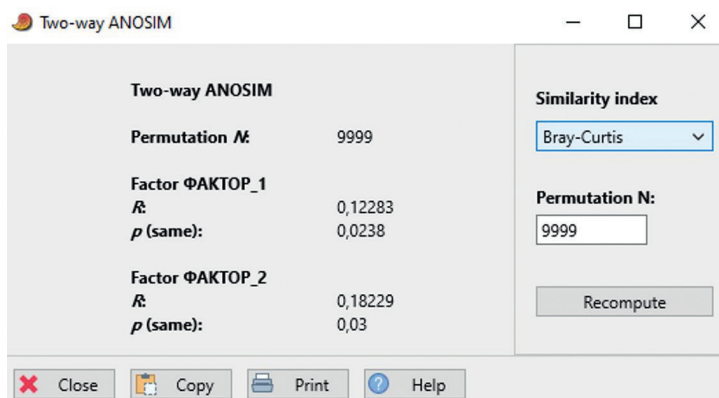


Рис. 16.4. Результаты двухфакторного анализа сходства (ANOSIM) на основе индекса Брея-Кёртиса

Философское эхо

ANOSIM выступает главным арбитром на поле многомерных данных, где группы проб выстраиваются в тактические схемы и вступают в состязание различий. Он оценивает не отдельные стыки, а общую структуру противостояния: как плотно игроки держатся внутри своей зоны и как их позиции расходятся с линией соперника. Вместо прямого измерения дистанций метод переводит все парные контакты в ранговый протокол, фиксируя, где борьба идёт плотнее — на своей или на чужой половине поля. Он заменяет абсолютные расстояния их порядковым строем, подобно судье, использующему правило преимущества: не сверяясь с разметкой, он считывает динамику — где мяч удерживается внутри командной связки, а где неизбежно уходит в единоборство, — и выносит решение по балансу давлений. Итоговая R -статистика становится его главным маркером: значение, стремящееся к единице, означает уверенное преимущество, нулевая отметка — тактический баланс в центральной части поля, где мяч не закрепляется ни за одной из сторон; а отрицательный результат — потерю компактности, когда дистанция между своими игроками превышает расстояние до соперника, превращая расстановку в позиционный разброд. Для проверки устойчивости исследователь запускает серию контрольных прогонов, подобно тренерскому штабу, проигрывающему на тактической доске расстановку против разных схем соперника,

проверяя, сохранится ли преимущество при изменении условий, тогда как арбитр оценивает лишь итог единственного сыгранного матча. Если внутригрупповой разброс велик, а границы зон сливаются в плотной борьбе, ранговая шкала не находит оснований для свистка — арбитр не останавливает матч, поскольку в данных нет статистического нарушения, требующего вмешательства. Но когда ранги между группами начинают устойчиво опережать ранги внутри них, пересечение порога значимости срабатывает как назначенный пенальти: арбитр выносит вердикт о доказанном отклонении от случайности, останавливая игру на пороге значимости, а контрольные перестановки становятся инструментальной проверкой, подтверждающей, что фиксация различий — закономерный итог, а не арбитражная ошибка. Таким образом, для исследователя, чувствующего игру данных, метод превращает матрицу дистанций в прозрачную турнирную таблицу, где ранги выстроены по результатам, а значимая R-статистика становится вердиктом арбитра: финальный свисток фиксирует итог, после которого статус каждой группы подтверждён и больше не требует пересмотра.

17. МЕТОД PERMANOVA

Введение

Анализ сложных водных сообществ требует решения фундаментальной методологической задачи: статистического сравнения структуры многомерных систем. Исследователям необходимо выявлять закономерности в распределении гидробионтов — различия в составе ихтиоценозов между участками водотоков, влияние антропогенных факторов или эффективность искусственных нерестилищ. Традиционные параметрические методы, предполагающие нормальность распределения и гомогенность дисперсий, часто неприменимы к экологическим данным, характеризующимся высокой долей нулей, асимметричным распределением и наличием выбросов. Ответом на этот методологический вызов стало появление метода PERMANOVA (*Permutational Multivariate Analysis of Variance*) в начале 2000-х годов.

Метод PERMANOVA был разработан для анализа многомерных данных в условиях, когда классические параметрические подходы неприменимы. Его создание связано с работами Марти Дж. Андерсон

(Marti J. Anderson, Оклендский университет, Новая Зеландия) (Anderson, 2001). Цель разработки состояла в создании универсального статистического инструмента для проверки гипотез о различиях многомерных сообществ под влиянием факторов среды (тип биотопа, антропогенная нагрузка, сезонные изменения) или экспериментальных воздействий. PERMANOVA представляет собой развитие ранних непараметрических подходов, в частности ANOSIM. Хотя ANOSIM также использует перестановочную процедуру для оценки значимости групповых различий на основе матрицы сходств, он обладает существенными ограничениями: низкой мощностью при сложных экспериментальных планах и невозможностью явного анализа взаимодействий факторов.

Методологические преимущества PERMANOVA определяются тремя ключевыми особенностями. Во-первых, метод позволяет использовать различные меры несходства, адекватные исследовательской задаче — индекс Брея–Кёртиса для количественных данных, коэффициент Жаккара для данных присутствия–отсутствия или специализированные меры, учитывающие функциональные либо филогенетические характеристики видов. Во-вторых, PERMANOVA использует перестановочную процедуру для построения нулевого распределения тестовой статистики (псевдо-F-отношения), что позволяет обойти требования к нормальности распределения и многомерной нормальности. В-третьих, метод обобщает логику дисперсионного анализа на многомерное пространство, позволяя разлагать общую многомерную изменчивость на компоненты, объяснённые изучаемыми факторами, их взаимодействиями и остаточной вариацией. Это делает метод особенно ценным для работы со сложными экспериментальными планами, включающими многофакторные схемы, ковариаты и повторные измерения.

Таким образом, PERMANOVA представляет собой методологически строгий инструмент для анализа многомерных данных в экологических исследованиях. Унаследовав перестановочный подход ANOSIM, метод преодолел его ограничения, предложив полноценное многомерное обобщение дисперсионного анализа. Возможность работы с различными мерами несходства, адаптация к сложным экспериментальным схемам и способность тестировать взаимодействия факторов делают PERMANOVA незаменимым инструментом для изучения реакции водных сообществ на комплекс факторов среды.

Методология

Математическая основа метода PERMANOVA представляет собой многомерное обобщение принципов классического дисперсионного анализа (ANOVA). Аналогично тому, как одномерный ANOVA разделяет общую вариацию единственного отклика на компоненты, объяснённые факторами модели и остаточной вариацией, PERMANOVA осуществляет разложение общей многомерной изменчивости, выраженной через матрицу мер несходства. Первоначально вычисляется общая изменчивость всего набора проб как сумма квадратов расстояний от всех проб до общего центроида — условной многомерной точки, представляющей среднее положение всех проб в пространстве признаков. Затем эта общая изменчивость разделяется на составляющие компоненты. Первая компонента отражает изменчивость, объяснённую исследуемым фактором, и рассчитывается как взвешенная сумма квадратов расстояний от групповых центроидов до общего центроида. Вторая компонента представляет собой остаточную изменчивость, вычисляемую как сумму квадратов расстояний от отдельных проб до центроида своей группы.

Поскольку данные о видовом составе сообществ редко удовлетворяют предположениям параметрической статистики, PERMANOVA использует процедуру перестановочного тестирования для оценки статистической значимости влияния факторов. Вместо обращения к теоретическим распределениям метод рассчитывает псевдо-F-статистику, представляющую отношение объяснённой изменчивости к необъяснённой с учётом степеней свободы. Для определения статистической значимости наблюдаемого значения псевдо-F осуществляется многократное случайное перераспределение меток групп между пробами при сохранении исходной матрицы несходства. Для каждой перестановки вычисляется новое значение псевдо-F, формирующее эмпирическое нулевое распределение. Значимость определяется как доля перестановок, в которых полученное значение статистики превышало или равнялось наблюдаемому.

Важно отметить методологическое отличие PERMANOVA от ANOSIM: если ANOSIM оперирует рангами мер несходства, то PERMANOVA работает непосредственно с их величинами и осуществляет явное разложение вариации. Это обеспечивает PERMANOVA ряд преимуществ, включая чувствительность к различным типам структур данных, возможность тестирования сложных моделей с

несколькими факторами и их взаимодействиями, а также количественную оценку вклада факторов в объяснённую изменчивость. Указанные характеристики делают данный метод ценным инструментом для анализа сложных экологических данных, в частности при исследовании структуры ихтиоценозов в разнообразных условиях среды.

Характер данных

Методологическая основа PERMANOVA связана с важной предпосылкой о гомогенности многомерных дисперсий, то есть близости степени разброса проб внутри сравниваемых групп в пространстве выбранной меры несходства. Нарушение этого условия может привести к тому, что различия в вариабельности будут ошибочно интерпретированы как различия в положении групповых центроидов. Поэтому проверка гомогенности дисперсий с помощью специализированных процедур, таких как PERMDISP, является обязательным этапом перед интерпретацией результатов. При этом, в отличие от классических параметрических методов, PERMANOVA не требует нормальности распределения признаков и допускает использование различных мер несходства, адаптированных к специфике биологических данных.

Метод демонстрирует наибольшую эффективность при анализе многомерных данных о видовом составе сообществ при сравнении чётко определённых категориальных групп. PERMANOVA успешно применяется для оценки влияния различных типов местообитаний, анализа комплексных воздействий факторов среды, а также для обработки планов с повторными измерениями. Его ключевое преимущество заключается в способности анализировать многофакторные схемы и оценивать взаимодействия между факторами, предоставляя количественную оценку вклада каждого фактора в общую изменчивость сообщества. Особенно ценен метод при работе с данными, не соответствующими требованиям параметрической статистики, поскольку перестановочная природа теста позволяет обойти ограничения, связанные с нормальностью распределения. Гибкость в выборе меры несходства позволяет адаптировать анализ к различным типам экологических данных — от количественных показателей обилия до данных присутствия–отсутствия. Возможность интеграции в сложные экспериментальные схемы делает PERMANOVA важным инструментом при оценке совместного влияния природных и антропогенных факторов на структуру сообществ.

К существенным ограничениям метода относится его чувствительность к различиям в дисперсиях между группами: при выраженной гетерогенности дисперсий PERMANOVA может давать ложноположительные результаты, приписывая эффекту «сдвига центроидов» то, что фактически обусловлено различием в разбросе. Метод также требует достаточного числа проб в каждой группе; при крайне малых выборках и сильной гетерогенности внутри групп мощность теста снижается, а результаты становятся трудно интерпретируемыми. PERMANOVA менее информативен при анализе плавных непрерывных градиентов, где жёсткая категоризация проб приводит к потере части информации, и лучше дополняется методами ординации или градиентного анализа. Кроме того, базовая реализация метода не учитывает пространственную или временную автокорреляцию без соответствующего включения этих факторов в дизайн, что важно для корректной интерпретации результатов в исследованиях с выраженной пространственно-временной структурой.

Таким образом, выбор PERMANOVA в качестве аналитического инструмента требует тщательной оценки соответствия структуры данных его методологическим предпосылкам и целям исследования. Метод обладает значительными преимуществами при анализе категориальных различий в сложных экспериментальных планах, но имеет чёткие ограничения, связанные с типом данных и характером исследуемых эффектов. Корректное применение предполагает обязательную проверку гомогенности дисперсий, адекватный выбор меры несходства и последующую интерпретацию результатов с учётом выявленных особенностей данных. PERMANOVA следует рассматривать как часть комплексного аналитического подхода, дополняемую методами визуализации, апостериорными тестами и, при необходимости, альтернативными статистическими методами в случаях выраженного нарушения ключевых предпосылок.

Пример анализа данных. One-way PERMANOVA

При данном типе анализа используется один категориальный фактор.

Задача. Оценили скорость плавания молодежи плотвы в длинах тела в секунду (BL/s) при 4-х различных температурах на фоне 4-х различных уровней кислотности среды. Имеются ли различия в скорости плавания при различных температурах на фоне разной кислотности? Если имеются, то какой. Построить диаграммы размаха.

АНАЛИЗ ДАННЫХ

Тест PERMANOVA размещён в PAST в главном меню *Multivariate* → *Tests* → *PERMANOVA*. Анализ данных скорости плавания молоди плотвы, полученных при разных температурах и вариативной кислотности, был дополнен методом PERMANOVA на том же массиве данных, который использовался для демонстрации работы ANOSIM. Такой подход методологически корректен как в дидактическом, так и в аналитическом аспектах. Сравнительный анализ единого набора данных разными многомерными методами позволяет наглядно продемонстрировать их специфику и взаимодополняемость, а также обеспечивает верификацию результатов за счёт совпадения выводов, полученных разными методами. Важно подчеркнуть, что оба метода — ANOSIM и PERMANOVA — предназначены для работы с сопоставимыми типами многомерных данных, что делает их последовательное применение к единому массиву наблюдений методологически оправданным для получения комплексной статистической картины.

Вкладка *Summary*

Результаты однофакторного PERMANOVA-анализа с использованием метрики Брея–Кертиса подтверждают наличие статистически значимых различий в скорости плавания молоди плотвы при разных температурах на фоне различной кислотности среды (рис. 17.1). Выбор меры Брея–Кертиса является методологически обоснованным, поскольку эта метрика оптимально подходит для анализа экологических данных, учитывая пропорциональные различия между выборками и оставаясь устойчивой к выбросам и нулевым значениям. Анализ, основанный на 9999 перестановках, демонстрирует достоверное разделение групп ($p = 0,0026$). Значение F-критерия, равное 3,08, указывает на то, что межгрупповая вариативность существенно превышает внутригрупповую, что свидетельствует о систематическом влиянии температурного фактора на плавательную активность рыб.

Сравнение сумм квадратов отклонений показывает, что общая вариативность данных (*Total sum of squares* = 0,08821) в значительной степени обусловлена различиями между температурными группами, в то время как внутригрупповая вариативность (*Within-group sum of squares* = 0,07019) составляет меньшую долю. Это подтверждает, что различия в скорости плавания носят не случайный характер, а связаны

с конкретными температурными условиями. Полученные результаты согласуются с данными ранее проведенного ANOSIM-анализа, где также было выявлено статистически значимое разделение групп ($p = 0,0021$, $R = 0,1582$).

Важно отметить, что использование перестановочного метода (*permutation test*) с большим числом повторений (9999) обеспечивает высокую надёжность и достоверность полученных результатов, делая анализ устойчивым к нарушениям параметрических предположений. Совокупность результатов PERMANOVA и ANOSIM позволяет сделать уверенный вывод о том, что температура воды является значимым фактором, влияющим на многомерный профиль скорости плавания молоди плотвы в условиях различной кислотности среды.

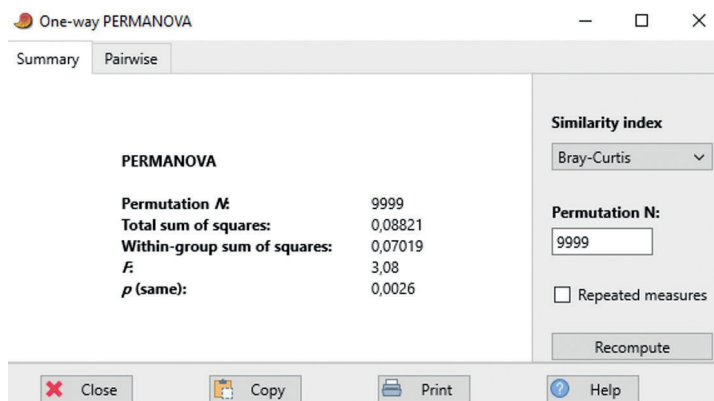


Рис. 17.1. Результаты однофакторного анализа PERMANOVA с использованием индекса Брея-Кёртиса

Вкладка *Pairwise*

Результаты попарного сравнения методом PERMANOVA с последовательной поправкой Бонферрони позволяют детализировать выявленные ранее различия между температурными группами (рис. 17.2). Анализ демонстрирует чёткую и статистически значимую специфику влияния температуры 2 на скорость плавания молоди плотвы. Достоверные различия зафиксированы между температурой 2 и температурой 3 ($p = 0,0071$), а также между температурой 2 и температурой 4 ($p = 0,0075$). Эти значения, остающиеся значимыми после строгой поправки на множественные сравнения, подчеркивают особый характер воздействия именно данного температурного режима.

В отличие от этого, сравнения между другими группами не показывают статистической значимости. Особенно показательное отсутствие различий между температурами 1, 3 и 4 (все $p > 0,0794$), формирующее чёткий кластер сходных по своему эффекту условий. Температура 1 не демонстрирует значимых отличий ни в одном из сравнений, включая температуру 2 ($p = 0,1284$), что указывает на ее переходный характер или промежуточное положение в спектре температурных воздействий.

Полученная картина попарных различий полностью согласуется с результатами ранее проведенного ANOSIM-анализа, где также были выявлены значимые различия между температурой 2 и температурами 3 ($p = 0,0018$) и 4 ($p = 0,007$). Такое совпадение ключевых выводов, полученных разными многомерными методами, демонстрирует методологическую устойчивость выявленных закономерностей и усиливает надёжность заключения о том, что температура 2 представляет собой критическую точку в температурном диапазоне, вызывающую существенные изменения в плавательной активности молоди плотвы на фоне варьирующей кислотности среды.

One-way PERMANOVA

Summary Pairwise

p-values, sequential Bonferroni significant

	Темп_1	Темп_2	Темп_3	Темп_4
Темп_1		0,1284	0,3661	0,0794
Темп_2	0,1284		0,0071	0,0075
Темп_3	0,3661	0,0071		0,3637
Темп_4	0,0794	0,0075	0,3637	

Close Copy Print Help

Рис. 17.2. Матрица попарных сравнений температурных групп методом PERMANOVA с последовательной поправкой Бонферрони (*p-values, sequential Bonferroni significant*)

Опция *Repeated measures*

В дидактических целях допустим, что выборки, для которых проведен вышеизложенный анализ, являются связанными, т. е. один и тот же организм последовательно участвует во всех опытах. Активация

опции *Repeated measures* (повторные измерения) на вкладке *Summary* при проведении PERMANOVA отражает специфический дизайн эксперимента со связанными выборками (рис. 17.3). Данная опция применяется в случаях, когда одна и та же особь измеряется многократно при различных условиях эксперимента, что требует особого подхода к расчету статистической значимости.

При использовании этой опции программа учитывает корреляционную структуру данных: четыре измерения скорости плавания в каждой строке рассматриваются как повторные замеры одной и той же особи при изменении уровня кислотности. Это находит отражение в расчете показателя *Error sum of squares* (сумма квадратов ошибок = 0,6053), который количественно выражает вариативность внутри индивидуальных траекторий изменения скорости плавания, отделяя её от межгрупповой вариабельности. Полученные значения F-статистики (3,227) и *p-value* (0,0014) следует интерпретировать исключительно в контексте модели со связанными измерениями. Более низкое *p*-значение по сравнению с анализом без учёта повторных измерений ($p = 0,0014$ против $p = 0,0026$) свидетельствует о повышении чувствительности теста при правильном учёте структуры данных.

Анализ попарных сравнений PERMANOVA с учётом повторных измерений выявил избирательное влияние температуры на скорость плавания молоди плотвы. Статистически значимые различия обнаружены между температурами 2 и 3 ($p = 0,0065$), 2 и 4 ($p = 0,0121$), а также 1 и 4 ($p = 0,0267$), в то время как остальные сравнения не показали значимых различий. Это указывает на нелинейный характер воздействия температурного фактора. Наиболее выраженные эффекты демонстрирует температура 2, которая достоверно отличается от температур 3 и 4, что подтверждает наличие специфического температурного порога, критического для плавательной активности рыб в условиях варьирующей кислотности.

Таким образом, применение опции *Repeated measures* является необходимым условием корректного анализа для данных с повторными измерениями. Такой подход обеспечивает валидность статистических выводов, поскольку адекватно учитывает фундаментальное свойство данных — наличие зависимых замеров у каждой особи, что позволяет отделить индивидуальный эффект от эффекта экспериментального воздействия и повысить точность обнаружения истинных различий между группами.

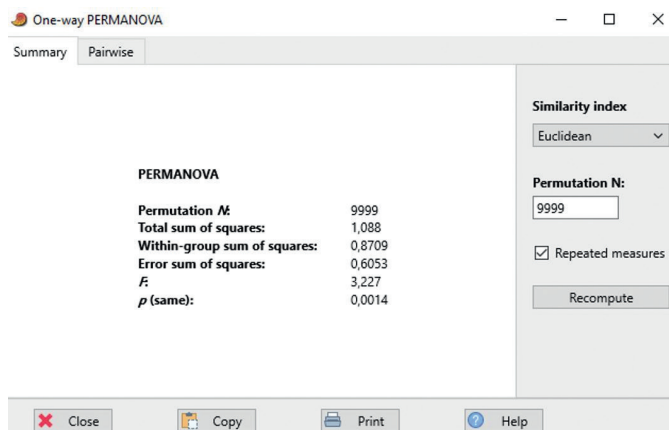


Рис. 17.3. Результаты однофакторного анализа PERMANOVA с использованием Евклидова расстояния при включенной опции *Repeated measures*

Пример анализа данных. Two-way PERMANOVA

Частный случай n -way PERMANOVA, где учитываются два категориальных фактора.

Задача. У двухгодовиков четырёх пород карпа — парский (П), ангелинский (А), московский (М) и чувашский (Ч) — измерили приросты массы тела в граммах в единицу времени, $\Delta W/\Delta t$, при выращиваниях на кормах четырёх типов, А–D, и двух величинах температуры, 1–2. Различаются ли величины приростов у данных видов в зависимости от типа корма и температуры? Оценить эффект взаимодействия.

АНАЛИЗ ДАННЫХ

Для анализа влияния типа корма и температуры на прирост массы тела у карпа был применен двухфакторный PERMANOVA с использованием метрики Брея–Кёртиса и 9999 перестановок для оценки значимости. Выбор именно этой метрики является методически обоснованным, поскольку индекс Брея–Кёртиса предназначен для работы с многомерными биологическими данными, устойчив к нулевым значениям и акцентирует внимание на различиях в составе (пропорциях), а не только на абсолютных величинах.

Результаты анализа показывают статистически значимое влияние типа корма (ФАКТОР_1: $F = 3,0892$; $p = 0,007$). Степени свободы ($df = 3$) подтверждают, что данный фактор соответствует четырем типам корма. Доля объясненной дисперсии для этого фактора составляет около 22,5% (сумма квадратов фактора / общая сумма квадратов), что указывает на его существенную роль в изменении прироста массы. В то же время влияние температуры (ФАКТОР_2: $F = 2,7058$; $p = 0,0611$) и эффект взаимодействия между факторами (Interaction: $F = 1,7033$; $p = 0,1161$) не достигли уровня статистической значимости при пороге $\alpha = 0,05$. Значение $p = 0,0611$ для температуры свидетельствует о наличии выраженной тенденции (тренда), но не позволяет уверенно отвергнуть нулевую гипотезу. Степень свободы $df = 1$ для этого фактора соответствует двум уровням температуры. Отсутствие значимого эффекта взаимодействия (*Interaction*) указывает на аддитивность факторов: влияние типа корма на прирост массы остается стабильным независимо от температурного режима.

Высокая доля остаточной дисперсии (*Residual SS*) в общей дисперсии составляет 58,4%, что свидетельствует о наличии существенных неучтённых факторов (индивидуальная изменчивость рыб, микроусловия и др.), влияющих на изменчивость показателей прироста массы. Тем не менее, модель в целом объясняет около 41,6% вариации данных, что является приемлемым результатом для биологических экспериментов.

Сравнение с ранее проведенным анализом ANOSIM этих же данных показывает определённую согласованность, но с важным уточнением: в ANOSIM оба фактора демонстрировали значимость ($p < 0,05$), тогда как более мощный метод PERMANOVA позволил дифференцировать их влияние, показав, что температура находится на границе значимости. PERMANOVA предоставляет более детализированную информацию о вкладе каждого фактора через суммы квадратов (*Sum of sqrs*), позволяя количественно оценить силу эффекта.

Таким образом, применение PERMANOVA подтвердило доминирующее влияние типа корма на прирост массы тела у карпа при отсутствии статистически значимого влияния температуры и эффекта взаимодействия. Полученные результаты подчеркивают критическую важность правильного выбора кормовых смесей в рыбоводстве, в то время как температурный фактор в данном диа-

пазоне оказывает лишь умеренное воздействие. Отсутствие взаимодействия факторов упрощает рекомендации по кормлению: оптимальный корм будет эффективен при обеих исследуемых температурах.

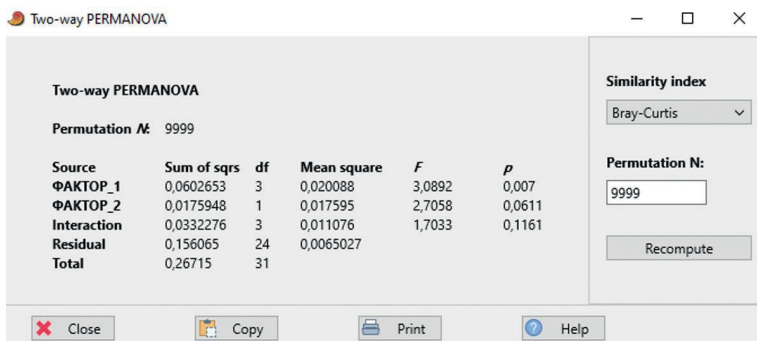


Рис. 17.4. Результаты двухфакторного анализа PERMANOVA с использованием индекса Брея-Кёртиса

Философское эхо

PERMANOVA — дирижёр многомерного оркестра, читающий партитуру изменчивости не по усреднённой громкости секций, а по точному взаимному расположению и взаимодействию каждого инструмента-пробы в пространстве звучания, заданном матрицей акустических расстояний. Его гений не просто в констатации диссонанса между группами; он стремится понять, как различные партии-факторы с их уникальными тембрами, а также их взаимодействия и переплетения голосов формируют общую музыкальную ткань произведения. Его суть — в разложении сложной симфонии изменчивости на партии, обусловленные каждым фактором, где вклад каждой составляющей можно услышать отдельно и в ансамбле. Он воспринимает сообщество как живое исполнение, где итоговое звучание рождается из слияния тембров — отдельных влияний, их наложений и индивидуальных нюансов исполнения, формирующих остаточную вариацию. Он не теряется в плавных крещендо градиентов, воспринимая их как естественную динамику исполнения, и уверенно работает с многослойной партитурой факторов, улавливая нюансы их взаимодействий, словно согласованное вступление инструментальных групп. Однако мастерство дирижёра требует дисциплины ансамбля — относительной одно-

родности звучания внутри каждой секции: если одна группа инструментов держит строй, а другая «разбегается» по широте диапазона, дирижёр рискует принять внутреннюю несогласованность звучания за новый выразительный мотив и приписать ей значимость фактора. Его интерпретация остаётся точной лишь там, где ансамбль сохраняет внутреннюю сбалансированность. В руках исследователя, слышащего структуру данных, PERMANOVA становится инструментом понимания многомерной партитуры: он не просто фиксирует различия, а показывает, какие партии действительно задают тон всему произведению, а какие остаются фоновым шумом. Каждая значимая партия в его разборе — не сухая цифра, а подтверждённый перестановками голос, звучащий в общей симфонии. PERMANOVA — инструмент для тех, кто слышит в экологии гармонию взаимодействий и стремится понять не только где и почему возникает различие, но и насколько силен вклад каждого фактора в общий гармонический строй сообщества.

18. МЕТОД SIMPER

Введение

Исследователям, анализирующим многомерные экологические данные, такие как структура сообществ рыб или других организмов, часто недостаточно констатации статистически значимых различий между группами проб (например, из разных местообитаний, временных периодов или условий воздействия). Ключевой задачей становится идентификация конкретных видов, вносящих основной вклад в наблюдаемое несходство. Для решения этой задачи предназначен метод SIMPER (*Similarity Percentage*), обеспечивающий детальную интерпретацию многомерных результатов.

Метод SIMPER, разработанный К. Р. Кларком (K. R. Clarke) в начале 1990-х годов в рамках методологического арсенала программного пакета PRIMER, первоначально применялся в морской экологии для анализа бентосных сообществ и впоследствии получил широкое распространение в различных разделах экологии. Его суть заключается в количественной оценке вклада каждого вида в среднее различие между сравниваемыми группами. Используя выбранную меру несходства (чаще всего дистанцию Брея–Кёртиса, хорошо подходящую

для данных по обилию), метод проводит попарное сравнение проб из разных групп. В результате виды ранжируются по их относительному вкладу в общее несходство, что позволяет оценить, насколько стабильно и существенно разница в обилии каждого конкретного вида влияет на итоговый показатель различий между группами.

Важным критерием надёжности интерпретации является показатель *Contrib/SD* — соотношение среднего вклада вида к стандартному отклонению этого вклада. Высокое значение *Contrib/SD* указывает на стабильно большой и статистически устойчивый вклад вида в межгрупповые различия, тогда как низкое значение свидетельствует о значительной вариабельности его обилия внутри групп, что снижает его диагностическую значимость. Дополнительно метод рассчитывает кумулятивный вклад ведущих видов, позволяя выделить ограниченный набор таксонов, который объясняет основную часть наблюдаемого несходства.

Для экологов, изучающих структуру, динамику и реакции сообществ на различные факторы (антропогенные воздействия, природные изменения, эффективность природоохранных мер), метод SIMPER служит важным инструментом интерпретации многомерных различий. Он обеспечивает переход от общего статистического вывода о наличии значимого различия, полученного методами вроде PERMANOVA или ANOSIM, к конкретному пониманию его биологической основы. Метод позволяет идентифицировать виды, демонстрирующие наиболее выраженные и стабильные изменения в обилии, формирующие основу различий между сравниваемыми условиями. Таким образом, SIMPER трансформирует абстрактные статистические паттерны в конкретное знание о том, какие именно элементы сообщества вносят основной вклад в наблюдаемые изменения, что создаёт основу для обоснованных управленческих решений и выдвижения проверяемых гипотез о механизмах этих изменений.

Методология

Метод SIMPER предназначен для идентификации и количественной оценки вклада отдельных видов в наблюдаемые различия между предварительно определёнными группами проб, статистическая значимость которых установлена методами многомерного анализа, такими как ANOSIM или PERMANOVA. В отличие от подходов, направленных на снижение размерности данных или проверку глобальных

гипотез о различиях, SIMPER служит инструментом интерпретации уже выявленной групповой структуры. Его основная задача заключается в определении видов, вносящих наибольший вклад в меру различий между парами групп.

Математический аппарат метода основан на разложении общей меры несходства между группами на составляющие, обусловленные отдельными видами. Алгоритм предполагает вычисление выбранной меры несходства (чаще всего дистанции Брея–Кёртиса, учитывающей различия в обилии организмов) для всех возможных пар проб, принадлежащих разным группам. Общее несходство между двумя группами рассчитывается как среднее значение этих парных мер. Ключевой этап анализа включает определение вклада каждого вида в несходство для каждой межгрупповой пары проб. Для этого вычисляется, насколько различие в обилии конкретного вида между двумя пробами вносит вклад в общую меру несходства для данной пары. Затем для каждого вида производится усреднение этих вкладов по всем возможным парам проб из сравниваемых групп, что даёт показатель среднего вклада (*Contrib*). Этот показатель отражает, в какой степени различия в обилии данного вида устойчиво влияют на наблюдаемое межгрупповое несходство.

Полученные средние вклады выражаются в процентах от общего несходства, что позволяет ранжировать виды по их относительному вкладу в межгрупповые различия. Для оценки надёжности выявленных вкладов рассчитывается соотношение среднего вклада к его стандартному отклонению (*Contrib/SD*). Высокие значения этого показателя (как правило, >1.0 – 1.5) свидетельствуют о стабильном и воспроизводимом вкладе вида, тогда как низкие значения указывают на существенную вариабельность его влияния на несходство между группами. Таким образом, метод SIMPER обеспечивает систематический подход к анализу вклада отдельных видов в межгрупповые различия, позволяя перейти от статистически значимых различий к идентификации конкретных биологических компонентов, их обуславливающих.

Характер данных

Метод SIMPER раскрывает свою аналитическую ценность в ситуациях, когда исследователь располагает заранее чётко определёнными и статистически значимо различающимися группами проб, что должно быть подтверждено предварительным применением соответствующих многомерных методов, таких как PERMANOVA или ANOSIM.

Его применение методологически оправдано при работе с количественными данными по обилию организмов, например с показателями уловистости, данными гидроакустических съёмок или результатами визуальных учётов, собранными в рамках строго определённой схемы исследования, основанной на категориальных факторах. К типичным ихтиологическим примерам корректного применения SIMPER относятся: оценка последствий выраженного антропогенного воздействия с резкой границей (сравнение структуры ихтиоценоза до и после строительства гидротехнического сооружения, создающего барьер для миграции видов); сопоставление принципиально различных типов водных экосистем в пределах одного региона (глубоководные озёра тектонического происхождения и мелководные пойменные водоёмы); анализ сезонной динамики сообществ рыб на критически важных биотопах (например, нерестилища в разные периоды года); оценка результативности природоохранных мероприятий путём сравнения ихтиоценозов строго охраняемых акваторий с соседними эксплуатируемыми участками. Именно в таких чётко определённых сценариях SIMPER наиболее эффективно выполняет свою функцию по идентификации и количественной оценке вклада конкретных таксонов в наблюдаемые межгрупповые различия, последовательно выделяя виды, чьё обилие существенно и закономерно отличается между сравниваемыми условиями.

Успешное применение и корректная интерпретация результатов SIMPER основаны на соблюдении нескольких ключевых предпосылок и учёте специфики данных. Фундаментальное условие — априорное определение и статистическое обоснование значимых различий между группами проб: метод не выполняет функцию кластеризации и полностью доверяет заданной пользователем группировке, поэтому его использование становится методологически необоснованным, если группы выделены произвольно или предварительные тесты не выявили достоверных различий. Критически важным является характер внутригрупповой вариабельности: SIMPER надёжно идентифицирует только те виды, которые демонстрируют стабильную и воспроизводимую разницу в обилии между группами; виды с высокой внутригрупповой вариабельностью, даже при большом среднем вкладе (*Contrib*), характеризуются высоким стандартным отклонением (*SD*), что выражается в низком отношении *Contrib/SD* и указывает на их ненадёжность в качестве диагностических индикаторов межгрупповых различий.

На результаты существенно влияет выбор меры несходства (чаще всего Брея–Кёртиса), преимущественно чувствительной к изменениям обилия доминирующих видов, что может приводить к недооценке вклада редких таксонов, различия которых проявляются главным образом в паттернах присутствия/отсутствия. Кроме того, метод изначально рассчитан на попарное сравнение групп; при наличии трёх и более групп анализ выполняется для каждой пары отдельно, что требует последующей осторожной интеграции результатов, поскольку вид, значимый для одной пары групп, может быть статистически незначимым для другой.

Тем не менее, существуют чётко определённые ограничения, при которых применение SIMPER становится проблематичным и может давать малоинформативные или даже вводящие в заблуждение результаты. Метод плохо адаптирован для анализа плавных градиентных изменений без естественных дискретных границ между группами, например, при изучении ихтиоценоза вдоль непрерывного градиента загрязнения: искусственное разбиение проб на группы в таких случаях неминуемо создаёт статистические артефакты в определении «основных вкладчиков». SIMPER также демонстрирует низкую надёжность при работе с данными, характеризующимися преобладанием редких видов и крайне высокой долей нулевых значений в матрице обилий; высокая стохастичность встреч делает оценки вкладов отдельных видов неустойчивыми и трудно интерпретируемыми. Аналогичные трудности возникают при анализе крайне гетерогенных выборок с малым числом проб на группу, когда высокая внутригрупповая вариабельность нарушает допущение о стабильности вкладов видов. Таким образом, SIMPER является мощным инструментом для декомпозиции установленных многомерных различий в стабильных и хорошо структурированных группах, где изменения преимущественно затрагивают обычные и многочисленные виды, но его корректное использование требует строгого учёта границ применимости и ограничений, связанных со структурой данных и дизайном исследования.

Пример анализа данных

Задача. В различных тканях окуня — кости, жабры, кожа, мышцы и печень — возрастных групп 2+ – 6+ из водоёма Красноярского края измерили концентрацию цинка. Оценить вклады концентраций

металла из отдельных тканей в общую изменчивость по данному показателю между отдельными возрастными группами, а также степень несходства общей концентрации цинка между отдельными тканями.

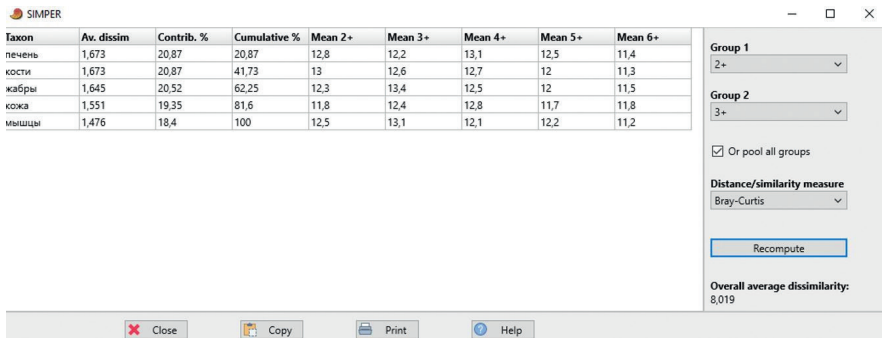
АНАЛИЗ ДАННЫХ

Тест SIMPER размещён в PAST в главном меню *Multivariate* → *Tests* → *SIMPER*. Применение данного теста позволяет оценить, какие именно ткани вносят наибольший вклад в различия по концентрации цинка между возрастными группами окуня. Полная интерпретация результатов становится возможной только при активации опции объединения всех групп — *Or pool all groups*, — что обеспечивает комплексное сравнение по всем возрастным категориям одновременно. Общее среднее несходство (*Overall average dissimilarity*) — усреднённая величина различий между сравниваемыми группами — составляет 8,019, что указывает на умеренный уровень различий в накоплении цинка между возрастными группами (рис. 18.1).

В анализе используются три ключевые метрики: *Av. dissim* (среднее несходство) показывает абсолютный вклад каждой ткани в общее различие между группами; *Contrib. %* (процент вклада) отражает относительный вклад каждой ткани в общее несходство; *Cumulative %* (кумулятивный процент) демонстрирует суммарный вклад тканей в порядке убывания их важности. Наибольший вклад в общее несходство вносят печень (20,87%) и кости (20,87%), что свидетельствует о ключевой роли этих тканей в аккумуляции цинка и в зависимой от возраста вариабельности его содержания. Жабры демонстрируют сопоставимый уровень вклада (20,52%), подчеркивая их важную роль в регуляции металлического гомеостаза. Кожа (19,35%) и мышцы (18,4%) вносят несколько меньший, но тем не менее существенный вклад в общую вариабельность. Совокупный вклад первых трёх тканей (печень, кости, жабры) превышает 62%, что позволяет считать их основными индикаторами возрастных изменений в накоплении цинка.

При анализе возрастной динамики прослеживаются определённые закономерности. В печени наблюдается постепенное снижение концентрации цинка от группы 2+ (12,8) к группе 6+ (11,4), что может свидетельствовать об изменении метаболических процессов или эффективности выведения металла с возрастом. Кости демонстри-

руют сходную тенденцию уменьшения содержания цинка в старших возрастных группах (с 13,0 до 11,3). Жабры отличаются наибольшей стабильностью показателей по всем возрастным группам (диапазон 12,0–13,4), в то время как кожа и мышцы не показывают выраженной возрастной динамики. Полученные результаты подчеркивают специфический характер аккумуляции цинка в различных тканях и важность учёта возрастных особенностей при оценке нагрузки загрязнения у гидробионтов. Наибольший вклад в возрастную изменчивость вносят печень и кости, что следует учитывать при разработке программ мониторинга и оценке экологического состояния водоёмов.



Таксон	Av. dissim	Contrib. %	Cumulative %	Mean 2+	Mean 3+	Mean 4+	Mean 5+	Mean 6+
печень	1,673	20,87	20,87	12,8	12,2	13,1	12,5	11,4
кости	1,673	20,87	41,73	13	12,6	12,7	12	11,3
жабры	1,645	20,52	62,25	12,3	13,4	12,5	12	11,5
кожа	1,551	19,35	81,6	11,8	12,4	12,8	11,7	11,8
мышцы	1,476	18,4	100	12,5	13,1	12,1	12,2	11,2

Group 1
2+

Group 2
3+

☒ Or pool all groups

Distance/similarity measure
Bray-Curtis

Recompute

Overall average dissimilarity:
8,019

Close Copy Print Help

Рис. 18.1. Результаты анализа SIMPER с использованием меры Брея-Кёртиса

Философское эхо

SIMPER выступает в роли финансового аудитора, который раскладывает совокупную разницу между бизнес-сегментами на индивидуальные вклады статей отчёта, выявляя ключевые драйверы расхода — те позиции отчётности, чьи отклонения формируют основной разрыв между подразделениями. Метод не пересчитывает баланс с нуля, а последовательно декомпозирует уже зафиксированное отклонение, превращая общую диспропорцию в прозрачную аналитическую ведомость, где каждая строка фиксирует свою долю в суммарном расхождении. Его сила — не только в оценке величины вклада, но и в проверке его устойчивости: какие статьи стабильно обеспечивают разницу, а какие лишь случайно всплывают на фоне операционного шума. Когда сегменты разделены чётко, а внутренняя вариативность показателей невелика, SIMPER уверенно выделяет позиции, формирующие костяк расхождения; но если внутригрупповой разброс велик,

а отчётность пестрит случайными выбросами, аудит теряет фокус — первоначальные лидеры оказываются лишь статистическим шумом, и итоговая картина рискует остаться формально точной, но аналитически пустой. В такие моменты специалист сталкивается с поверхностными отчётами, где цифры на месте, но структура расхождения размывается, и отделить устойчивый драйвер от фоновых колебаний становится делом профессионального чутья. Поэтому SIMPER требует не механического считывания, а понимания природы данных: в руках внимательного аналитика метод превращает разрозненные показатели в структурированное аудиторское заключение, где каждый значимый вклад — не абстрактная сумма, а стабильно воспроизводимый параметр, чётко показывающий, какие именно статьи формируют разрыв и насколько устойчиво их влияние.

19. СРАВНИТЕЛЬНЫЙ АНАЛИЗ ANOSIM, PERMANOVA И SIMPER

Вычислительные возможности PAST позволяют быстро реализовать все три теста и в полном объёме получить необходимые результаты. ANOSIM и PERMANOVA отвечают на вопрос о статистической значимости различий между априори заданными группами (например, возрастными, пространственными или экспериментальными), тогда как SIMPER позволяет определить конкретные переменные (виды, ткани, показатели), вносящие наибольший вклад в эти различия. ANOSIM и PERMANOVA являются методами проверки статистических гипотез: первый оценивает степень разделения групп через ранжирование попарных расстояний и рассчитывает R-статистику как меру выраженности эффекта; второй проводит дисперсионный анализ на основе матриц несходства с вычислением псевдо-F-статистики и доли объяснённой вариации (pseudo-R^2).

Метод PERMANOVA, как правило, обладает большей статистической мощностью и гибкостью: он лучше приспособлен к сложным многофакторным планам, позволяет оценивать вклад отдельных факторов и их взаимодействий, вводить ковариаты и явным образом разлагать общую вариацию на объяснённые и остаточные компонен-

ты. Поэтому в большинстве практических случаев, включая однофакторные схемы, предпочтение отдаётся PERMANOVA, тогда как ANOSIM остаётся уместным, когда принципиально требуется интуитивно интерпретируемая ранговая R-статистика как мера эффекта или необходима сопоставимость с данными, ранее проанализированными этим методом. Оба метода отвечают на вопрос: «есть ли различия и какова их величина в многомерном пространстве?», но сами по себе не идентифицируют конкретные переменные, ответственные за эти различия.

Тест SIMPER, в свою очередь, служит инструментом последующей детализации уже выявленных межгрупповых различий. Этот метод рассчитывает средний вклад каждой переменной в общее несходство (*average dissimilarity*) между группами, ранжируя переменные по убыванию вклада и предоставляя кумулятивный процент, что позволяет выделить ограниченный набор наиболее «влиятельных» видов или показателей. Интерпретация результатов SIMPER методологически оправдана только после установления статистической значимости различий между группами методом PERMANOVA (или, реже, ANOSIM): применение SIMPER к данным без предварительного подтверждения межгрупповых различий может приводить к ложным выводам о «важности» переменных при статистически неразличимых группах. Кроме того, SIMPER чувствителен к высокой внутригрупповой вариабельности и склонен переоценивать вклад редких, но сильно варьирующих переменных; поэтому для интерпретации существенны показатели стабильности вклада (например, *Contrib/SD*) и учёт структуры данных.

В совокупности эти методы образуют последовательную аналитическую цепочку: сначала глобальными тестами (PERMANOVA, при необходимости диагностикой дисперсий через PERMDISP) проверяется наличие и однородность значимых различий между группами, затем с помощью SIMPER выполняется апостериорная декомпозиция несходства по переменным. Такой комплексный подход позволяет не только констатировать наличие групповых эффектов, но и раскрыть их биологическую природу — какие таксоны или показатели и в каком направлении меняются, что особенно важно в экологических, токсикологических и рыбоводных исследованиях при формировании адресных рекомендаций по мониторингу и управлению.

20. СПИСОК ЛИТЕРАТУРЫ

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46.

Benzécri, J.-P. (1969). Statistical analysis as a tool to make patterns emerge from data. In *Methodologies of Pattern Recognition* (S. Watanabe, Ed.), pp. 35-74. New York: Academic Press.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4), 325–338. DOI: <https://doi.org/10.1093/biomet/53.3-4.325>.

Hill, M. D., & Smith, A. J. (1989). Evaluating Associativity in CPU Caches. *IEEE Transactions on Computers*, 38(12), 1612–1630.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441.

Imbrie, J. & N.G. Kipp. 1971. A new micropaleontological method for quantitative paleoclimatology: Application to a late Pleistocene Caribbean core. In: *The Late Cenozoic Glacial Ages*, edited by K.K. Turekian, pp. 71-181, Yale Univ. Press, New Haven, CT.

Clarke K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1), 117–143. <https://doi.org/10.1111/j.1442-9993.1993.tb00438.x>

Klovan, J.E. & J. Imbrie. 1971. An algorithm and FORTRAN-IV program for large scale Q-mode factor analysis and calculation of factor scores. *Mathematical Geology* 3:61-77.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27.

Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (3rd ed.). Elsevier.

Pearson, K. (1901). «On Lines and Planes of Closest Fit to Systems of Points in Space».

Philosophical Magazine, 2(11), 559–572.

21. СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Нохрин, Д. Ю. Лабораторный практикум по биостатистике / Д. Ю. Нохрин. – Челябинск : Изд-во Челяб. гос. ун-та, 2018. – 289 с. – (Классическое университетское образование). – ISBN 978-5-7271-1487-2.
2. Gauch, H. G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press.
3. Greenacre, M. (2010). *Biplots in Practice*. BBVA Foundation.
4. Kenkel, N. C., & Orlóci, L. (1986). Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results. *Ecology*, 67(4), 919-928.
5. Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (3rd ed.). Elsevier.
6. Leps, J., & Šmilauer, P. (2003). *Multivariate Analysis of Ecological Data using CANOCO*. Cambridge University Press.
7. McCune, B., & Grace, J. B. (2002). *Analysis of Ecological Communities*. MjM Software Design.
8. McGarigal, K., Cushman, S. A., & Stafford, S. G. (2000). *Multivariate Statistics for Wildlife and Ecology Research*. Springer.
9. Quinn, G. P., & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.
10. ter Braak, C. J. F., & Šmilauer, P. (2012). *Canoco Reference Manual and User's Guide: Software for Ordination, version 5.0*. Microcomputer Power.
11. Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). *Analysing Ecological Data*. Springer.

Лобырев Ф.С.

ОРДИНАЦИОННЫЕ МЕТОДЫ

*Учебное пособие по курсу лекций
«Практическая биостатистика»*

Подписано в печать 17.06.2026. Формат 70х90/16.
Усл. печ. л. 13,6. Печать офсетная. Бумага офсетная.
Тираж 120 экз. Заказ 26-01758

Издатель ООО «КОНСТАНТА»
308519, Белгородская обл., Белгородский р-н, пос. Северный, ул. Березовая, 1/12
Тел./факс (4722) 300-720, www.konstanta-print.ru

Отпечатано в ООО «КОНСТАНТА»
308519, Белгородская обл., Белгородский р-н, пос. Северный, ул. Березовая, 1/12.
Тел./факс (4722) 300-720, www.konstanta-print.ru

