## Глава 2. Трудности приложения математической статистики к анализу данных мониторинга

Данные экологического мониторинга не всегда удовлетворяют требованиям метрологии, статистической воспроизводимости и другим условиям, которые позволили бы обоснованно использовать для их анализа традиционные методы математической статистики (Максимов и др., 1999). Важнейшие из этих требований — нормальность распределений переменных, однородность дисперсий, устойчивость частот. Остается только определить, к чему приводит нарушение этих требований.

Традиционный многомерный анализ матрицы экологических данных размерности $m \times n$ ($m$ — число строк, $n$ — число столбцов) часто начинают с расчета Пирсоновских коэффициентов парной корреляции, которые отражают сопряженность изменений во времени и пространстве любых двух переменных в исходной матрице данных (Максимов и др., 1999). Наличие таких сопряженностей, например, между гидрохимическими переменными указывает на некоторые характерные особенности гидрологии в исследуемом водоеме, а наличие статистической связи между отдельными гидрохимическими и биологическими переменными может характеризовать отклик биотической компоненты экосистемы на изменение абиотических условий. Кроме того, в результате расчета коэффициентов корреляции для всех пар переменных получается корреляционная матрица, которая служит основой для таких популярных методов многомерного анализа, как множественная регрессия, факторный анализ, некоторые модификации кластер-анализа.

Хорошо известно однако, что величина коэффициента корреляции Пирсона, как параметра 2-мерного нормального распределения, отражает тесноту связи между двумя выборками, только если они получены из нормально распределенных совокуп-

ностей, а сама связь является линейной. Поэтому прежде чем "запускать" соответствующую вычислительную программу, следовало бы по крайней мере убедиться в нормальности распределения всех переменных в исходной матрице данных. Однако даже в руководствах по многомерной статистике не удается найти примеры, где проводилась бы проверка нормальности исходных переменных. Между тем такая проверка не требует особых усилий, поскольку соответствующие вычислительные программы имеются в любом пакете прикладных статистических программ.

Можно, в частности, построить так называемые "ящики-с-усами" (*box-and-whiskers plots*) и сразу увидеть такие типичные отклонения от нормальности, как асимметрию распределения и наличие "выскакивающих" или "экстремальных" значений. Хорошим способом проверки является и построение полувероятностных графиков. В качестве наиболее надежного метода проверки нормальности часто рекомендуют расчет асимметрии и эксцесса эмпирических распределений, в особенности, когда объем выборки достаточно велик.

Чтобы показать, к чему может привести отклонение исследуемых переменных от нормальности, приведем для примера диаграмму рассеяния (рис. 2.1) для концентраций взвешенных веществ и хрома в р. Суре в районе Сурского водохранилища в 1994-1997 гг. (Максимов и др., 1999). Коэффициент корреляции для этой пары переменных равен 0.57 и его статистическая значимость формально очень высока, поскольку он рассчитан по 213 парам значений. Вид диаграммы на рис. 2.1 заставляет, однако, усомниться в том, что между этими переменными действительно имеется сколько-нибудь существенная связь.

Обратим внимание на “ящики-с-усами” сверху и сбоку диаграммы, характеризующие вид распределений переменных. Отклонения от нормальности этих распределений совершенно очевидны. На фоне довольно плотного облака точек в левом нижнем углу диаграммы резко выделяются экстремальные значения, явно "выскакивающие" и за пределы "усов". Если же исключить эти значения (как иногда рекомендуют в руководствах по статистике), т.е. всего 11 точек из 213, то для оставшихся 202 пар значений переменных “взвешенные вещества” и “хром” коэффициент корреляции будет равен всего 0.22. Формально этот коэффициент отличен от нуля с вероятностью 0.998, что обычно считается очень высоким уровнем значимости, однако следует иметь в виду, что этот уровень определяется для единственной корреляции. Иначе го-

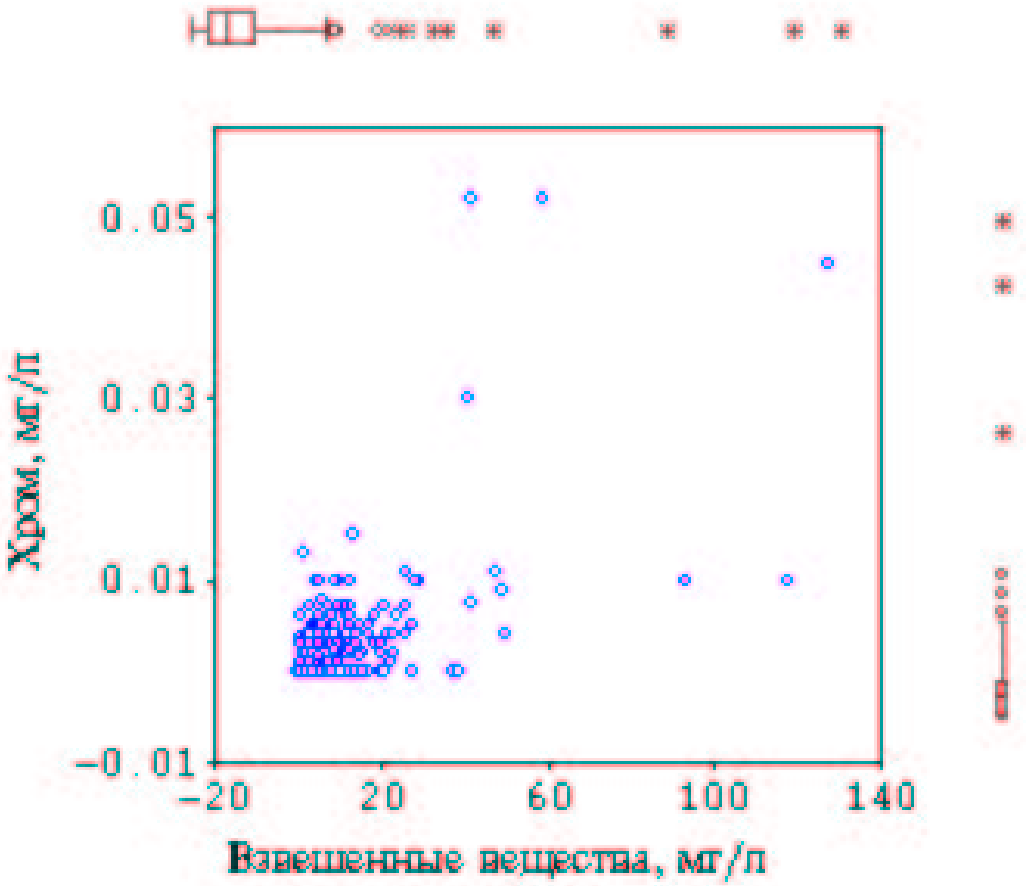


**Рисунок 2.1.** Диаграмма рассеяния для концентраций взвешенных веществ и хрома (справа "ящики с усами" для хрома, сверху — для взвешенных веществ)

воря, оценка, которую мы получаем по стандартной программе корреляционного анализа, рассчитывается только для пары “взвешенные вещества — хром” в отсутствие остальных переменных. А в нашей матрице данных, как уже было сказано, 44 переменных, так что корреляционная матрица должна содержать 946 коэффициентов парной корреляции. Поэтому при уровне значимости 1-0.998 = 0.002 можно ожидать, что по крайней мере 1-2 раза величина коэффициента корреляции $r$ = 0.22 может встретиться в такой матрице совершенно случайно. Разумеется, оценки коэффициентов в нашей матрице корреляций нельзя считать полностью независимыми, но тем не менее должно быть понятно, что для каждого из 946 коэффициентов статистическая оценка их значимости должна быть более строгой, чем для единственной пары переменных. По-видимому, для того, чтобы признать отличным от нуля коэффициент корреляции, равный 0.22, нужно принять уровень значимости во всяком случае не более 0.001.

Однако даже если признать статистически значимым коэффициент корреляции, равный 0.22, то остается весьма сомнительной его, так сказать, познавательная ценность. Квадрат коэффициента корреляции (коэффициент линейной детерминации) характеризует долю дисперсии переменной *y*, объясняемой линейной зависимостью *y* от *x*. Несколько огрубляя, это можно интерпретировать, как долю изменений переменной *y*, связанных именно с изменением *x*. Поскольку $0.22^2 = 0.0484$, возвращаясь к нашему примеру, можно заключить, что (без учета выпадающих значений) концентрация хрома в воде примерно на 5% обусловлена содержанием в воде взвешенных веществ или, наоборот, всего 5% изменений концентрации взвешенных веществ в воде непосредственно связано с содержанием в ней хрома.

Исходя из рис. 2.1 становится ясно, что в большинстве случаев концентрации взвешенных веществ в исследованных пробах не выходят за пределы 20-30 мг/л, а концентрации хрома — за пределы 0.01 мг/л. Кроме того, только в 4 случаях из 214 экстремально высокие значения для обеих переменных наблюдались в одной и той же пробе. При этом все эти пробы были отобраны в Сурском водохранилище на станциях B, D, E, F в мае 1997 г. И еще в 2-х пробах, на станции F в августе 1996 г. и у середины плотины в одной из проб, взятых в апреле 1994 г., экстремальные значения концентрации взвешенных веществ наблюдались одновременно с относительно высокой концентрацией хрома (0.01 мг/л). Вряд ли на этом основании можно заключить, что содержание хрома в воде р. Сура и Сурском водохранилище как-то связано с общим количеством взвеси. Итак, по сути дела все, что удалось узнать, сводится к обнаружению "выбросов" взвешенных веществ и хрома. Зато естественным образом возникает вопрос: в какой мере эти самые "выбросы" повлияли на сообщества организмов на упомянутых станциях.

При попытке применить анализ корреляций к данным о численности отдельных видов зоопланктона возникают еще более серьезные затруднения (Максимов и др., 1999). Для всех зоопланктеров р. Суры, выбранных для анализа из проб 1994-1997 гг., характерна не только сильная асимметрия, но и не менее значительный эксцесс. Поэтому *a priori* ясно, что прежде чем рассчитывать коэффициенты корреляции, следует каким-то способом преобразовать исходные данные, чтобы получить их распределение, близкое к нормальному. Однако помимо этого возникает еще проблема так называемых пропущенных значений.

Например, в нашей матрице данных кладоцера *Daphnia cucullata* была обнаружена в 67 пробах, а *D. longispina* — в 83, но только в 37 пробах оба эти вида присутствовали одновременно. Если рассчитать коэффициент корреляции между численностями этих видов, рассматривая отсутствие вида в пробе, как пропущенное значение (т.е. оставив пустыми соответствующие места в таблице исходных данных), то при вычислении все пробы, в которых отсутствовал хотя бы один из двух видов, будут исключены и получится коэффициент, характеризующий сопряженность их обилий в 37 пробах, тогда как общее количество проб в нашей матрице — 214. Этот коэффициент практически равен нулю ($r$ = 0.0006). Если заменить пропущенные значения нулями, как это чаще всего делается, то в расчете будут использованы все 214 проб. Коэффициент корреляции при этом оказывается на порядок больше ($r$ = 0.0065), но его отличие от нуля остается статистически незначимым.

Однако замена пропущенных значений нулями, когда дело касается численностей видов, представляется разумной только на первый взгляд. Прежде всего, при этом исключается возможность применить логарифмирование — наиболее популярный метод нормализации асимметричных распределений, поскольку нули не логарифмируются. Но гораздо важнее другое: ведь если ни одна особь данного вида не попала в пробу, это не значит, что его не было в водоеме в момент отбора пробы. В лучшем случае можно считать, что его обилие было меньше, чем обилие самого малочисленного вида, обнаруженного в этой пробе. Как правило, этот самый малочисленный вид бывает представлен в пробе единственной особью. Поэтому иногда пропущенные значения вместо нуля заменяют некоторыми условными величинами, скажем 1/5 или 1/6. Правда, при этом такую, пусть малую, численность можно приписать и тем видам, которых действительно не было в данный момент в водоеме.

Впрочем, на величине коэффициента корреляции такая замена нуля на любую достаточно малую постоянную величину не сказывается, а отклонение наших данных от нормального закона при этом не уменьшается, так что неплохо было бы убедиться, что найденное выше отсутствие корреляции между численностями *D. cucullata* и *D. longispina* действительно указывает на отсутствие сопряженности в их изменениях.

При работе с заведомо "ненормально" распределенными переменными, в частности при описании структуры экологических сообществ, рекомендуется использовать непараметрические методы, среди которых наиболее популярен коэффициент

корреляции рангов Спирмена (см. например, Михайловский, 1988). Если рассчитать этот коэффициент для той же пары видов, обозначив нулями отсутствие вида в пробе, получим $r_S$ = 0.293, что для 214 сравниваемых рангов отличается от нуля при очень высоком уровне значимости ($p < 10^{-4}$). Если же рассматривать отсутствие вида в пробе, как пропущенное значение, то для 37 проб, в которых встречались оба вида дафний, то корреляция оказывается не только статистически значимой, но и весьма существенной с экологической точки зрения: $r_S$ = 0.783 ($p < 10^{-6}$).

Итак, в зависимости от того, как оценивается отсутствие вида в пробе и какой метод анализ используется при расчете корреляции, получаются принципиально различные результаты.

Причина этого становится понятной, если построить диаграмму рассеяния для рангов, рассчитанных по возрастанию численности *D. cucullata* и *D. longispina* так, что отсутствию вида в пробе приписывали средний минимальный ранг. Вид этой диаграммы (рис. 2.2) показывает, во-первых, что коэффициент линейной корреляции

рангов (коэффициент Спирмена) ни в какой мере не характеризует связь между рангами обилия для данной пары видов. Во-вторых, хорошо видна положительная корреляция между рангами в пробах, где оба вида встречались одновременно, что собственно и отражает коэффициент $r_S$ = 0.783. Но, кроме того, нельзя не обратить внимания на совокупности точек по краям диаграммы. Они соответствуют пробам, в которых встречался лишь один из двух видов. Расположение этих точек заставляет усомниться в том, что между *D. cucullata* и *D. longispina* вообще имеется какая-либо связь, поскольку их число сравнимо с числом проб, содержащих оба вида.

Приходится признать, что "классический" метод корреляционного анализа оказывается малоэффективным в применении к данным экологического мониторинга. Подобный вывод справедлив и для упомянутых выше факторного и регрессионного анализов, так как их алгоритмы также строятся исходя из обязательности нормального распределения переменных.

Вернемся, однако, к диаграмме рассеяния для концентраций взвешенных веществ и хрома (рис. 2.1). Если более внимательно рассмотреть ее левый нижний угол в увеличенном масштабе (рис. 2.3), то обнаружится, что почти все точки на этой диаграмме расположены на горизонталях, соответствующих некоторым округленным значениям концентрации хрома. Из 214 значений величина 0.001 встретилась 14 раз,

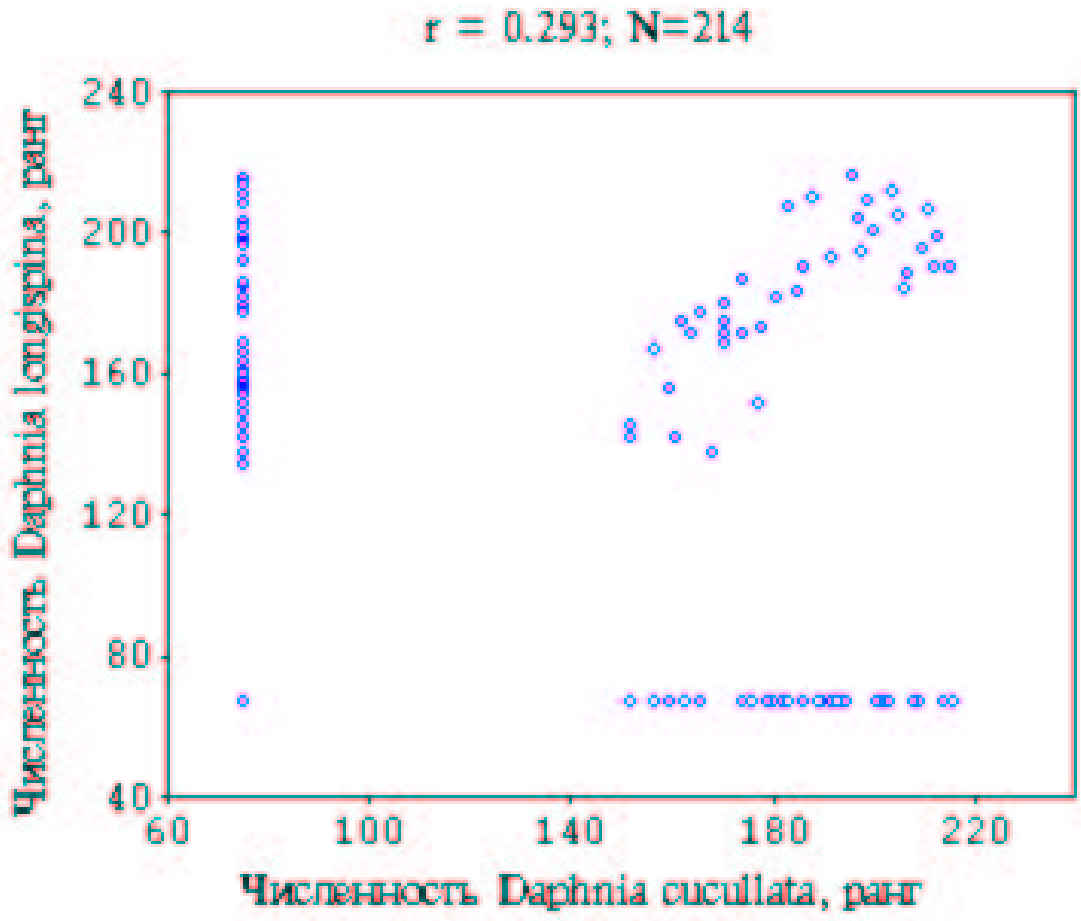


**Рисунок 2.2.** Диаграмма рассеяния для рангов численности *D. cucullata* и *D. longispina*

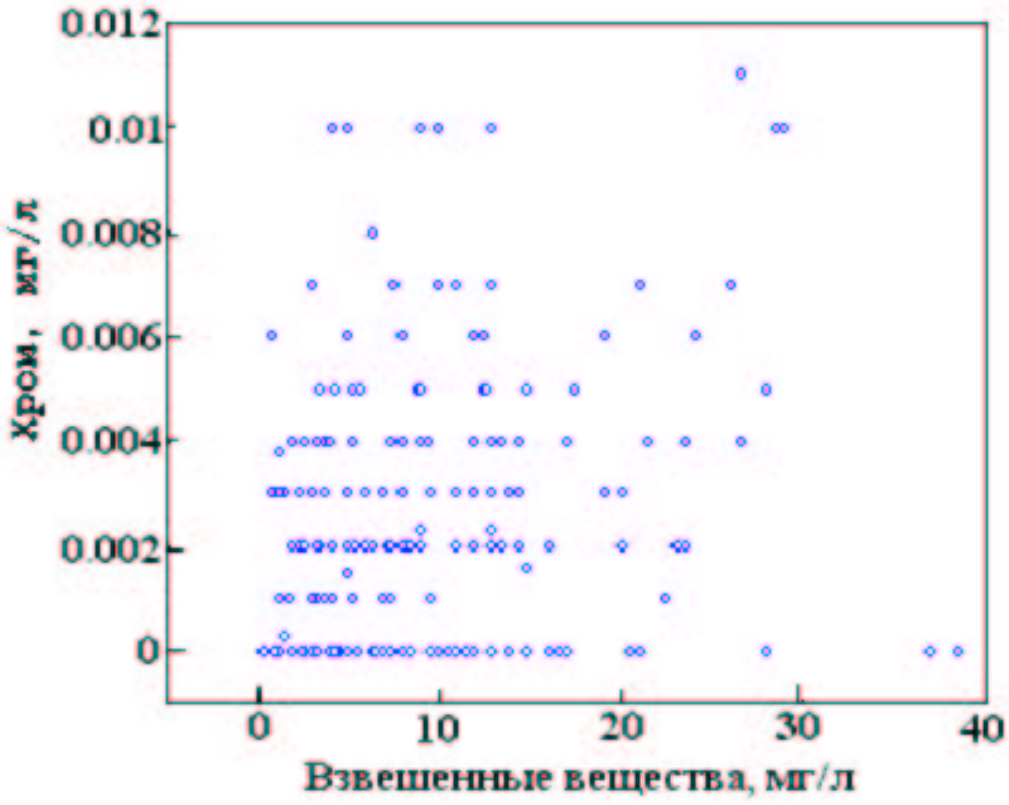


**Рисунок 2.3.** Диаграмма рассеяния для малых концентраций взвешенных веществ и хрома

0.002 — 37 раз и т.д. Это конечно связано с методическими особенностями способа определения хрома. То же характерно и для ряда других переменных, прежде всего ионов металлов, определявшихся в пробах (Cu, Zn, Pb, Ni, Mn, Co), а также для нитритов и фенолов. Но это значит, что мы имеем дело не с непрерывными, а с дискретными величинами, и не выполняется еще одно требование большинства традиционных статистических методов и в частности корреляционного и факторного анализа — требование непрерывности измеряемых переменных.

Для этого случая существует специально разработанный метод обнаружения связи между переменными — метод таблиц сопряженности или перекрестных таблиц (*contingency tables, cross-tabulation*). Чаще всего он применяется при анализе согласованности изменений качественных признаков, для которых существуют некоторые естественные градации. В простейшем варианте так называемых таблиц 2×2 таких градаций всего 2: 1) признак есть и 2) признака нет. Если же дискретность появляется (как в описанном примере) из-за недостатка точности измерений, предлагается более или менее произвольно разбивать их шкалу на классы и подсчитывать частоты попадания результатов измерений в каждый из классов (Максимов и др., 1999).

Первое затруднение, которое возникает при попытке воспользоваться методом таблиц сопряженности, также связано с многомерностью задачи. В сущности безразлично (при компьютерной обработке данных), будут ли мы рассматриваться диаграммы рассеяния для каждой пары переменных или строиться перекрестные таблицы для той же пары. И в том и в другом случае для 44 переменных должно получиться или 946 диаграмм, или 946 таблиц. При этом в таблицах с числом классов больше 2-х для каждой переменной в некоторые клетки таблицы могут попасть слишком малые частоты даже при сравнительно большом числе проб и тогда становятся мало эффективными соответствующие статистические процедуры. Это тем более справедливо для таблиц сопряженности более чем двух переменных, и, кроме того, такие таблицы становятся трудно обозримыми. Построение таких таблиц оказывается слишком трудоемким даже для современной вычислительной техники.

Другое затруднение связано с тем, что разбиение непрерывной количественной шкалы на классы само по себе может повлиять на результат анализа совершенно так же, как оно влияет, например, на вид гистограммы для эмпирической функции распределений (см. например: Плохинский, 1970). В известной мере оба эти затруднения

преодолеваются в той модификации метода таблиц сопряженности, которая реализована в алгоритме детерминационного анализа (ДА) (глава 3). Метод обработки многомерных экологических данных с помощью ДА может стать успешной альтернативой традиционным статистическим подходам и программной основой для реализации технологии экологического контроля.