ПРОТОКОЛ ИССЛЕДОВАНИЯ ДАННЫХ ДЛЯ ИЗБЕЖАНИЯ ОБЩИХ СТАТИСТИЧЕСКИХ ПРОБЛЕМ

**Резюме**

1. Во время преподавания статистики экологам ведущие авторы этой статьи заметили общие статистические проблемы. Если была выбрана случайная выборка их работ (в том числе научных статей), которая была сделана до проведения этих курсов, половина, вероятно, будет содержать нарушения основополагающих допущений применимых статистических методов.
2. Некоторые нарушения мало влияют на результаты или экологические выводы; но другие увеличивают ошибки типа I или типа II, что может привести к неправильным экологическим выводам. Большинство из этих нарушений можно избежать, применяя лучший анализ данных. Эти проблемы особенно сложны в прикладной экологии, где на карту поставлены управленческие и политические решения.
3. Здесь мы представляем протокол для исследования данных; Обсуждаем текущие инструменты для выявления выбросов, неоднородность дисперсии, коллинеарность, зависимость наблюдений, проблемы с взаимодействием, двойные нули при многомерном анализе, нулевую инфляцию в обобщенном линейном моделировании и правильный тип отношений между зависимыми и независимыми переменными; и даем рекомендации о том, как решать эти проблемы, когда они возникают. Мы также рассматриваем неправильные представления о нормальности и предоставляем рекомендации по преобразованию данных.
4. Исследование данных позволяет избежать ошибок типа I и типа II, среди других проблем, тем самым уменьшая вероятность ошибочных экологических выводов и плохих рекомендаций. Поэтому важно обеспечить качественное управление и политику на основе статистического анализа.

**Введение**

За последние три десятилетия произошло огромное расширение статистических инструментов, доступных для применения экологам. Краткий перечень доступных методов включает: линейную регрессию, обобщенное линейное (смешанное) моделирование, обобщенное аддитивное /добавленное (смешанное) моделирование, деревья регрессии и классификации, анализ выживаемости, нейронные сети, многомерный анализ со всеми его многочисленными методами, такими как анализ основных компонентов (PCA) , анализ канонического соответствия (CCA), (не) метрическое многомерное масштабирование (NMDS), различные временные ряды и пространственные методы и т. д. Создание более быстрых компьютеров и свободно доступного программного обеспечения, такого как R ([**R Development Core Team 2009)**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b33)), позволило регулярно применять сложные статистические методы для любого типа данных. В этой статье речь не идет об этих методах. Вместо этого речь идет о важном шаге, который необходим, но часто не предшествует применению этих методов.

У всех статистических методов есть общая проблема «мусора». В некоторых методах, например, один выброс может определять окончательные результаты и выводы. Неоднородность (различия в вариации) может вызвать серьезные проблемы в линейной регрессии и анализе моделей дисперсии (**[Fox 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b14" \o "Ссылка на библиографическую цитату)**) и с некоторыми многовариантными методами (Huberty, 1994).

Когда основной вопрос заключается в том, чтобы определить, какие ковариаты\переменные управляют системой, то самым сложным аспектом анализа является, вероятно, вопрос о коллинеарности (корреляции между ковариатами), что увеличивает ошибки типа II (т.е. невозможность отклонения нулевой гипотезы, когда она ложна). При многовариантном анализе, применяемом к данным об экологических сообществах, наличие двойных нулей (например, два вида, которые совместно отсутствуют на разных участках) способствует сходству в некоторых методах (например, СПС), но не в других. Однако другие многовариантные методы чувствительны к видам с ограниченным распределением и низким уровнем обилия (например, CCA). В одномерных методах анализа, таких как обобщенное линейное моделирование (GLM) для подсчета данных, нулевая инфляция переменной отклика может приводить к смещенным оценкам параметров (**[Cameron & Trivedi, 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b5" \o "Ссылка на библиографическую цитату)**). Когда многовариантные методы используют методы перестановок для получения значений *P* , например, в CCA и анализе избыточности (RDA, **[ter Braak & Verdonschot 1995](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b401" \o "Ссылка на библиографическую цитату)**) или в тесте Mantel (**[Legendre & Legendre 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b25" \o "Ссылка на библиографическую цитату)**), временная или пространственная корреляция между наблюдениями может увеличить тип I ошибки (отклонение нулевой гипотезы, когда она верна).

То же самое относится к методам регрессионного типа, применяемым во временных или пространственно-коррелированных наблюдениях. Одним из наиболее используемых и неправильно используемых методов является, без сомнения, линейная регрессия. Часто этот метод связан с линейными моделям? и нормальностью; обе концепции часто неправильно понимаются. Линейная регрессия более чем способна устанавливать нелинейные отношения, например, используя взаимодействия или квадратичные условия (**[Montgomery & Peck 1992](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b26" \o "Ссылка на библиографическую цитату)**). Термин «линейный» в линейной регрессии относится к способу использования параметров в модели, а не к типу отношения, которые моделируются. Знание того, имеем ли мы линейные или нелинейные закономерности между объясняющими переменными и переменными отклика, имеет решающее значение для того, как мы применяем линейную регрессию и связанные с ней методы. Нам также необходимо знать, сбалансированы ли данные, прежде чем включать взаимодействия. Например, **[Zuur, Ieno & Smith (2007)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b40" \o "Ссылка на библиографическую цитату)** использовали ковариации/переменные пол, место и месяц для моделирования гонадосоматического индекса (веса гонад по отношению к общей массе тела) кальмара. Однако оба пола не измерялись в каждом месте в каждом месяце из-за несбалансированности выборки. Фактически данные были настолько неуравновешенными, что имело смысл анализировать только подмножество данных и воздерживаться от включения определенных взаимодействий.

С учетом этого богатства потенциальных ловушек, гарантирующего, что ученый не обнаружит ложного ковариационного эффекта (ошибка типа I), неправильно отклонять модель с конкретным ковариатом (ошибка типа II) или производить результаты, определенные только несколькими влиятельными наблюдениями, требуется, чтобы подробное исследование данных было проведено до применения любого статистического анализа. Целью настоящего документа является предоставление протокола для исследования данных, который идентифицирует потенциальные проблемы ([рис.1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f1)). По нашему опыту, исследование данных может занимать до 50% времени, затраченного на анализ.

Хотя исследование данных является важной частью любого анализа, важно, чтобы оно было четко отделено от проверки гипотез. Решения о том, какие модели для тестирования должны быть сделаны априорно, основаны на биологическом понимании системы исследователем (**[Burnham & Anderson 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b4" \o "Ссылка на библиографическую цитату)**). Когда это понимание очень ограничено, исследование данных может использоваться как упражнение, генерирующее гипотезу, но это принципиально отличается от процесса, который мы защищаем в этой статье. Использование аспектов исследования данных для поиска моделей («углубление в данные») может служить руководством для будущей работы, но результаты следует рассматривать очень осторожно, а выводов о более широкой популяризации избежать. Вместо этого новые данные должны собираться на основе полученных гипотез и независимых тестов. Когда исследование данных используется таким образом, необходимо четко указать как используемый процесс, так и недостатки любых выводов.

На протяжении всей статьи мы фокусируемся на использовании графических инструментов (**[Chatfield 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b7" \o "Ссылка на библиографическую цитату)** , **[Gelman, Pasarica & Dodhia 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b15" \o "Ссылка на библиографическую цитату)**), но в некоторых случаях также можно применять тесты для нормальности или однородности. Однако статистическая литература предостерегает от определенных тестов и рекомендует графические инструменты (**[Montgomery & Peck 1992](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b26" \o "Ссылка на библиографическую цитату)** , **[Draper & Smith 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b9" \o "Ссылка на библиографическую цитату)** , **[Quinn & Keough, 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b32" \o "Ссылка на библиографическую цитату)**). **[Ляяра (2009)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b23" \o "Ссылка на библиографическую цитату)** представляет семь причин не применять предварительные испытания для нормальности, в том числе: большинство статистических методов, основанных на нормальности, устойчивы к нарушению; для больших наборов данных центральная предельная теория подразумевает приближенную нормальность; для небольших образцов мощность тестов низкая; и тесты для больших наборов данных чувствительны к малым отклонениям (что противоречит центральной предельной теории).

## Шаг 1: Есть ли выбросы в Y и X?

В результатах некоторых статистических методов преобладают выбросы; другие методы рассматривают их как любые другие значения. Например, выбросы могут приводить к избыточному расслоению/избыточной дисперсии в GLP или биномиальном GLM Пуассона, когда исход не является бинарным (**[Hilbe 2007](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b19" \o "Ссылка на библиографическую цитату)** ). Напротив, в NMDS с использованием индекса Jaccard (**[Legendre & Legendre 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b25" \o "Ссылка на библиографическую цитату)**) наблюдения, по существу, рассматриваются как наличие и отсутствие, поэтому выброс не влияет на результат анализа каким-либо особым образом. Следовательно, важно, чтобы исследователь понимал, как конкретный метод реагирует на наличие выбросов. На данный момент мы определяем выброс как наблюдение, которое имеет относительно большую или малую величину по сравнению с большинством наблюдений.

Графическим инструментом, который обычно используется для обнаружения выбросов, является коробчатая диаграмма. Она отображает медиану и рассеивание данных. В зависимости от используемого программного обеспечения медиана обычно представлена ​​в виде горизонтальной линии с 25% и 75% квартилями, образующими квадрат/коробку вокруг медианы, которая содержит половину наблюдений. Затем поводятся линии из квадратов/коробок, и любые точки за пределами этих линий обозначаются как выбросы. Некоторые исследователи обычно (но ошибочно) удаляют эти наблюдения. [На рисунке 2а](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f2) показан пример такого графика с использованием 1295 наблюдений морфометрической переменной (длина крыла солеарного воробья  Ammodramus caudacutus; **[Gjerdrum, Elphick & Rubega 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b17" \o "Ссылка на библиографическую цитату)**). График наводит на мысль (возможно, ошибочно, как мы увидим в мгновение), что существует семь выбросов.

Еще один очень полезный, но забытый графический инструмент для визуализации выбросов - это точечная диаграмма Кливленда (**[Cleveland 1993](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b8" \o "Ссылка на библиографическую цитату)**). Это график, в котором строчный номер наблюдения нанесен на график относительно значения наблюдения, тем самым предоставляя гораздо более подробную информацию, чем ящик. Точки, которые лежат далеко справа или далеко слева, это наблюдения, значения которых значительно больше или меньше, чем большинство наблюдений, и требуют дальнейшего изучения. Если такие наблюдения существуют, важно проверить необработанные данные на наличие ошибок и оценить, являются ли наблюдаемые значения разумными/приемлемыми. [Рисунок 2b](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f2) показывает точечную диаграмму Кливленда для данных длины крыла воробья; обратите внимание, что наблюдения, выявленные коробчатой диаграммой, не являются особенно экстремальными. Тенденция «вверх» на [рисунке 2b](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f2) просто возникает из-за того, что данные в электронной таблице сортируются по весу. Существует одно наблюдение длины крыла около 68 мм, которое выделяется слева примерно на полпути вверх по графику. Это значение не намного больше других значений, поэтому мы пока не можем сказать, что это выброс.

[На рисунке 3](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f3) показана многопанельная точечная диаграмма Кливленда для всех измеряемых морфометрических переменных; Обратите внимание, что некоторые переменные имеют несколько относительно больших значений. Такие экстремальные значения могут указывать на истинные погрешности измерения (например, некоторые соответствуют характеристикам «отвлечения наблюдателя» sensu **[Morgan 2004](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b27" \o "Ссылка на библиографическую цитату)**, в результате чего глаз наблюдателя обращается к неправильному числу на шкале измерения). Обратите внимание, что нельзя пытаться утверждать, что такие большие значения могли возникнуть случайно. Если бы они были, то промежуточные значения также должны быть созданы случайно, но ни одного не было. (Полезное упражнение состоит в том, чтобы неоднократно генерировать эквивалентное количество случайных наблюдений из подходящего распределения, например нормальное распределение, и определять, как количество экстремальных точек сравнивается с эмпирическими данными.) Когда наиболее вероятным объяснением является то, что крайние/экстремальные наблюдения являются ошибками измерения (наблюдателя), их следует отбросить, поскольку их присутствие, вероятно, будет доминировать в анализе. Например, мы применили дискриминантный анализ полного набора данных воробьев, чтобы увидеть, отличаются ли наблюдения наблюдателей, и обнаружили, что первые две оси главным образом определялись выбросами.

До сих пор мы свободно определяли «выброс», как наблюдение, которое отличается от остальных. Более строгий подход заключается в рассмотрении вопроса о том, оказывают ли необычные наблюдения чрезмерное влияние на анализ (например, по оценочным параметрам). Мы проводим различие между влиятельными наблюдениями в переменной отклика/зависимой переменной и в ковариатах. Примером последнего является то, когда численность видов моделируется как функция температуры, причем почти все значения температуры принадлежат промежутку между 15 и 20 ° C, но одно значение равно 25 ° C. В общем, это не идеальный образец выборки, потому что диапазон 20-25 ° C недостаточно отобран. Однако в полевом исследовании может быть только одна возможность опробовать более высокую температуру. При большом размере выборки такие наблюдения могут быть отброшены, но с относительными малыми наборами данных последующее уменьшение размера выборки может быть нежелательным, особенно если другие наблюдения имеют выбросы для других объясняющих переменных. Если исключить такие наблюдения нельзя, тогда подумайте о преобразовании объясняющих переменных.

В методах регрессионного типа выбросы в переменных отклика более сложны. Преобразование данных является опцией, но поскольку переменная отклика имеет первостепенный интерес, лучше выбрать статистический метод, который использует распределение вероятностей, которое допускает большее изменение для крупных средних значений (например, гамма для непрерывных данных, пуассоновское или отрицательное биномиальное значение для Count), поскольку это позволяет нам работать с исходными данными. Для многовариантных анализов этот подход не подходит, поскольку эти методы не основаны на распределении вероятностей. Вместо этого мы можем использовать другую меру связи. Например, евклидово расстояние довольно чувствительно к большим значениям, потому что оно основано на теореме Пифагора, тогда как расстояние Хорда понижает-взвешивает большие значения (**[Legendre & Legendre 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b25" \o "Ссылка на библиографическую цитату)**).

Некоторые статистические пакеты поставляются с целым рядом диагностических инструментов для определения влиятельных наблюдений. Например, статистика Cook в линейной регрессии (**[Fox 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b14" \o "Ссылка на библиографическую цитату)** ) дает информацию об изменении параметров регрессии, для каждого наблюдения последовательно и по-отдельности. Проблема с такими инструментами заключается в том, что при наличии нескольких «выбросов» с одинаковыми значениями они не будут обнаружены. Следовательно, перед применением статистического анализа следует исследовать наличие таких наблюдений с использованием графических инструментов, обсуждаемых в этой статье.

В конечном счете, эколог должен решить, что делать с выбросами. Из-за плохого планирования эксперимента могут возникать выбросы в ковариате, и в этом случае отбрасывание наблюдения или трансформация ковариата являются разумными решения. Ошибки наблюдателя и измерения являются обоснованным основанием для отбрасывания наблюдений. Но выбросы в переменной отклика могут потребовать более изящного подхода, особенно когда они представляют собой истинное изменение измеряемой переменной. Взятие детальных полевых или экспериментальных заметок может быть особенно полезно для документирования при возникновении необычных событий и, таким образом, предоставления объективной информации для повторного изучения выбросов. Независимо от того, как проблема устранена, важно знать, есть ли выбросы, и сообщить, как они были обработаны; разведка данных позволяет это сделать.

## Шаг 2: Есть ли у нас однородность дисперсии?

Однородность дисперсии является важным предположением в анализе дисперсии (ANOVA), других регрессионных моделях и в многомерных методах, таких как дискриминантный анализ. [На рис. 4](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f4) показаны условные диаграммы размаха с уровня потребления пищи гудзонским веретенником (Limosa haemastica), перелетной птицы, на берегу Аргентины(E. Ieno, неопубликованные данные). Чтобы применить ANOVA к этим данным для проверки того, отличаются ли уровни потребления в зависимости от пола, периода времени или комбинации этих двух переменных (т.е. взаимодействия), мы должны предположить, что (i) изменения наблюдений от полов схожи; (ii) изменения наблюдений трех временных периодов схожи; И (iii) различия между тремя временными периодами внутри полов схожи. В этом случае, по-видимому, в зимних данных для самцов наблюдается несколько меньшая вариативность и большая в летних. Однако такие небольшие различия в вариациях - это не причина для беспокойства. Более серьезные примеры нарушения можно найти в **[Zuur et al. (2009a)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)** . [**Fox (2008)**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b14) показывает, что для упрощенной модели линейной регрессии неоднородность серьезно ухудшает оценки наименьших квадратов, когда соотношение между наибольшей и наименьшей дисперсией составляет 4 (консервативный) или более.

В моделях регрессионного типа проверка однородности должна проводиться с использованием остатков модели; Т.е. путем построения остатков напротив установленных значений и создания аналогичного набора условных диаграмм размаха для остатков. Во всех этих графах остаточная вариация должна быть одинаковой. Решение проблемы разнородности дисперсии представляет собой либо трансформацию переменной отклика для стабилизации дисперсии, либо применение статистических методов, которые не требуют однородности (например, обобщенные наименьшие квадраты: **[Pinheiro & Bates 2000](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b30" \o "Ссылка на библиографическую цитату)** , **[Zuur et al., 2009a](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)**.

## Шаг 3: Являются ли данные нормально распределенными?

Различные статистические методы предполагают нормальность, и это привело к тому, что многие из наших аспирантов создали гистограмму после гистограммы их данных (например, [рис.5а](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f5)). Важно, однако, знать, предполагает ли используемый статистический метод нормальность, и что именно называется нормальным распределением? Например, PCA не требует нормальности (**[Jolliffe 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b21" \o "Ссылка на библиографическую цитату)**). Линейная регрессия предполагает нормальность, но достаточно надежна против нарушения предположения (**[Fitzmaurice, Laird & Ware 2004](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b13" \o "Ссылка на библиографическую цитату)**). Однако если вы хотите применить статистический тест, чтобы определить, существует ли значительное разделение групп в дискриминантном анализе, нормальность наблюдений конкретной переменной в пределах каждой группы важна (Huberty, 1994). Простые t- тесты также предполагают, что наблюдения в каждой группе имеют нормальное распределение; поэтому следует изучить гистограммы для необработанных данных каждой группы.

В линейной регрессии мы фактически принимаем нормальность всех повторяющихся наблюдений при определенном значении коварианта ([рис.6](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f6) , **[Montgomery & Peck, 1992](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b26" \o "Ссылка на библиографическую цитату)**), предположение, которое невозможно проверить, если у него много повторений при каждом выборочном значении коварианта. Однако нормальность исходных данных означает нормальность остатков. Поэтому мы можем построить гистограммы остатков, чтобы получить некоторое представление о нормальности (**[Quinn & Keough 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b32" \o "Ссылка на библиографическую цитату)**; **[Zuur et al., 2007](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b40" \o "Ссылка на библиографическую цитату)**), хотя мы не можем полностью проверить предположение.

Даже когда предположение о нормальности, по-видимому, нарушается, ситуация может быть более сложной, чем кажется. Например, форма гистограммы на [рис. 5а](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f5) указывает на асимметрию, которая может предполагать необходимость преобразования данных. [На рисунке 5b](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f5) показана многопанельная гистограмма для той же переменной, за исключением того, что данные отображаются по месяцам; это позволяет нам видеть, что асимметрия исходной гистограммы, вероятно, вызвана изменением веса воробья с течением времени. В этих условиях было бы нецелесообразно преобразовывать данные, поскольку различия между месяцами могут быть уменьшены и их будет сложнее обнаружить.

## Шаг 4: Много ли нулей содержат данные?

[**Elphick & Oring (1998, 2003)**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b11) исследовали влияние управления соломой на обилие водно-болотных птиц на затопленных рисовых полях. Одним из возможных статистических анализов является моделирование числа птиц в зависимости от времени, глубины воды, фермы, способа управления полем, температуры и т. д. Поскольку этот анализ включает в себя моделирование подсчета, обобщённая линейная модель является подходящим анализом. [На рисунке 7](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f7) показан частотный график, иллюстрирующий частоту каждого значения для общего количества водоплавающих птиц. Чрезвычайно большое количество нулей говорит нам о том, что мы не должны применять обычный пуассоновский или отрицательную биномиальную обобщенную линейную модель, поскольку они будут приводить к смещенным оценкам параметров и стандартным ошибкам. Вместо этого следует рассмотреть нуль - завышенные обобщённые линейные модели (**[Cameron & Trivedi 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b5" \o "Ссылка на библиографическую цитату)**; **[Zuur et al., 2009a](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)**).

Также можно одновременно анализировать данные для нескольких видов, используя многовариантные методы. Для такого анализа нам необходимо рассмотреть, что означает, когда два вида совместно отсутствуют. Этот результат мог бы сказать что-то важное в отношении экологических характеристик участка, например, что он содержит условия, неблагоприятные для обоих видов. К тому же, когда два места имеют те же недостатки, это может означать, что места экологически схожи. С другой стороны, если виды имеют очень ограниченное/сгруженное распределение или просто редки, тогда совместные отсутствия могут возникать случайно и ничего не говорить о пригодности данной территории для видов, сходстве между необходимыми средами обитания для видов или экологическом сходстве мест. Таким образом, высокая частота нулей может значительно усложнить интерпретацию таких анализов. Независимо от нашего отношения к совместному отсутствию, нам нужно знать, есть ли в данных двойные нули. Это означает, что для каждой пары видов нам нужно рассчитать, как часто оба имеют множество нулей для одного и того же наблюдения (например, место). Мы можем либо представить эту информацию в таблице, либо использовать передовые графические инструменты, такие как кореллограмм ([Фиг.8](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f8); **[Саркар 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b36" \o "Ссылка на библиографическую цитату)**). В нашем примере с водоплавающей птицей частота двойных нулей очень высока. Все синие круги соответствуют видам, у которых более 80% их наблюдений совместно равны нулю. Этот результат соответствует биологией изучаемых видов, большинство из которых образуют крупные стаи и имеют очень сгруженные распределения. Метод главных компонент будет обозначать такие виды, как схожие, хотя их экологическое использование мест обитания часто отличается (например, **[Elphick & Oring 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b11" \o "Ссылка на библиографическую цитату)**). Альтернативные многомерные анализы, которые игнорируют двойные нули, обсуждаются в **[Legendre & Legendre (1998)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b25" \o "Ссылка на библиографическую цитату)** и **[Zuuret al. (2007)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b40" \o "Ссылка на библиографическую цитату)** .

## Шаг 5: Существует ли коллинеарность среди ковариатов?

Если основной вопрос в исследовании заключается в том, какие ковариаты управляют переменной (-ами) отклика, то самой большой проблемой для преодоления зачастую является коллинеарность. Коллинеарность - это наличие корреляции/взаимосвязи между ковариатами. Обычными примерами являются ковариаты, такие как вес и длина, или глубина воды и расстояние до береговой линии. Если коллинеарность проигнорирована, то, скорее всего, это закончится запутанным статистическим анализом, в котором ничего не значимо, но где опущение одного ковариата может сделать другие значительными или даже изменить знак оценочных параметров. Эффект коллинеарности проиллюстрирован в контексте множественной линейной регрессии, но схожие проблемы существуют при анализе дисперсий, моделей смешанных эффектов, RDA, CCA, GLM или GAM. [Таблица 1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#t1)дает результаты множественной линейной регрессии, в которой число солевых воробьев, охваченных на графике исследования, моделируется как функция ковариатов, которые описывают относительное обилие различных видов растений (подробнее см. **[Gjerdrum, Elphick & Rubega 2005](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b16" \o "Ссылка на библиографическую цитату)** , **[Gjerdrum et al., 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b17" \o "Ссылка на библиографическую цитату)**). Во втором столбце таблицы приведены примерные P- значения t- статистики для каждого параметра регрессии, когда все ковариаты включены в модель. Обратите внимание, что только один ковариат, который для процентного покрытия пика ситника Жерара, слабо значим на уровне 5%.

**Таблица 1.** P- значения t- статистики для трех моделей линейной регрессии и значений коэффициента разброса (VIF) для полной модели. В полной модели число полосатых воробьев, которое является мерой количества птиц, моделируется как функция ковариатов, перечисленных в первом столбце. Во втором и третьем столбцах представлены значения P-значений и VIF для полной модели (обратите внимание, что никакие переменные еще не удалены). В четвертом столбце P- значения представлены для модели после того, как коллинеарность была удалена путем последовательного удаления каждой переменной, для которой значение VIF было самым высоким, пока все оставшиеся VIF не были ниже 3. В последнем столбце остались только переменные со значительными  P-значениями, представляющие самое скудное объяснение количества воробьев на графике

В линейной регрессии выражение для параметров дисперсий β j вычисляется по формуле (**[Draper & Smith 1998](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b9" \o "Ссылка на библиографическую цитату)**; **[Fox 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b14" \o "Ссылка на библиографическую цитату)**):

Член S j зависит от ковариационных значений, n - размер выборки, а σ 2 - дисперсия остатков, но эти термины не имеют отношения к текущему обсуждению (и поэтому их математическая формулировка здесь не приводится). Это первое выражение, которое важно. Термин R j2представляет собой R 2 из линейной регрессионной модели, в которой в качестве переменной отклика используется ковариант X j , а все остальные ковариаты - как объясняющие переменные. Высокое значение R 2 в такой модели означает, что большая часть изменений в ковариате X j объясняется всеми другими ковариатами, что означает, что существует коллинеарность. Цена, которую приходится платить в этой ситуации, заключается в том, что стандартные ошибки параметров завышаются квадратным корнем из 1 / (1 -  R j2), также называемым коэффициентом вариации дисперсии (VIF), что означает, что P-значения становятся более крупными, что затрудняет обнаружение эффекта. Это явление проиллюстрировано в [таблице 1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#t1); третий столбец таблицы содержит значения VIF для всех ковариантов и показывает, что существует высокий уровень коллинеарности. Одна из стратегий решения этой проблемы состоит в том, чтобы последовательно отбрасывать ковариант с наивысшим VIF, пересчитывать VIF и повторять этот процесс до тех пор, пока все VIF не будут меньше предварительно выбранного порога. [**В Montgomery & Peck (1992)**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b26) использовалось значение 10, но более строгий подход заключается в использовании значения такого низкого как 3, как мы сделали здесь. Высокая или даже умеренная коллинеарность особенно проблематична, когда экологические сигналы слабы. В этом случае даже VIF от 2 может вызывать недостоверную оценку параметров по сравнению с ситуацией без коллинеарности. После этого процесса были отброшены три переменных из нашего анализа: высокая Spartina alterniflora , а также для роста растений и плотности стебля. При устранении проблемы коллинеарности показатель Juncus показан очень значимым ([таблица 1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#t1)). Последовательно отбрасывая дополнительные незначительные члены/условия по одному, получается модель только с Juncus и Shrub, но с небольшим дальнейшим изменением значений P , показывая, как отбрасывание коллинеарных переменных может оказать большее влияние на значения P, чем отбрасывание несущественных/недостоверных ковариатов.

Другие способы обнаружения коллинеарности включают в себя парные диаграммы рассеяния, сравнивающие ковариаты, коэффициенты корреляции или бипот PCA (**[Jolliffe 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b21" \o "Ссылка на библиографическую цитату)**), применимый ко всем ковариатам. Коллинеарность также можно ожидать, если временные (например, месяц, год) или пространственные переменные (например, широта, долгота) используются вместе с ковариатами, такими как температура, количество осадков и т. д. Поэтому всегда следует сопоставлять на графике все ковариаты напротив временных и пространственных ковариантов. Самый простой способ разрешения коллинеарности - это отбросить коллинеарных ковариаций. Выбор того, какой ковариант будет отброшен, может основываться на VIF, а может быть, лучше, на здравом смысле или биологическом знании. Альтернативное рассмотрение, особенно когда будущая работа над этой темой будет сделана, заключается в том, насколько легко альтернативные ковариаты должны измеряться с точки зрения усилий и затрат. Всякий раз, когда два ковариата X и Z коллинеарны, и Z используется в статистическом анализе, биологическое обсуждение, в котором объясняется эффект Z, должно включать упоминание коллинеарности и признание того, что, может быть, это X управляет системой (см. **[Gjerdrum et al., 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b17" \o "Ссылка на библиографическую цитату)**). Для обсуждения коллинеарности в сочетании с ошибками измерения на ковариатах см. **[Carroll et al. (2006)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b6" \o "Ссылка на библиографическую цитату)** .

## Шаг 6: Каковы отношения между переменными Y и X?

Еще одна важная часть исследования данных, особенно при одномерном анализе, представляет собой график зависимости переменной отклика от каждого ковариата ([рис.9](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f9)). Обратите внимание, что переменная для процента высокой осоки на графике (% Tall sedge) должна быть исключена из любого анализа, так как она имеет только одно ненулевое значение. Этот результат показывает, что диаграммы размахов и точечные диаграммы Кливленда должны применяться не только к переменной отклика, но также и к ковариатам (т.е. мы не должны были вычислять VIF с % Tall sedck, включенными в предыдущий раздел). На [рисунке 9](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f9) нет четких закономерностей между ответными и объясняющими переменными, за исключением, возможно, количества Juncus (см. Также [таблицу 1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#t1)). Обратите внимание, что отсутствие четких шаблонов не означает, что нет никаких отношений; это просто означает, что нет четких двусторонних отношений. Модель с несколькими объясняющими переменными может по-прежнему обеспечивать хорошую подгонку.

Помимо визуализации связей между переменными, диаграммы рассеяния также полезны для обнаружения наблюдений, которые не соответствуют общей схеме между двумя переменными. [На рисунке 10](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f10) показана многопанельная диаграмма рассеяния (также называемая парным сюжетом) для 1295 воробьев, для которых мы имеем морфологические данные. Любое наблюдение, которое лежит вне черного облака, требует дальнейшего изучения; это могут быть разные виды, ошибки измерений, опечатками или они могут быть правильными значениями. Обратите внимание, что наблюдение с длинными крыльями, которое мы взяли с помощью точечной диаграммы Кливленда на [рис. 2b,](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f2) имеет средние значения для всех других переменных, предполагая, что это действительно то, что нужно проверить. Нижние панели на [фиг.10](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f10) содержат коэффициенты корреляции Пирсона, на которые могут влиять выбросы, что означает, что выбросы могут даже способствовать коллинеарности.

## Шаг 7: Должны ли мы рассматривать взаимодействия?

Останавливаясь на морфометрических данных воробья, предположим, что существует вопрос о том, изменяется ли соотношение между длиной крыла и весом в течение месяцев и отличается между полами. Общим подходом к этому анализу является применение модели линейной регрессии, в которой вес представляет собой переменную отклика, а длина (непрерывная) крыла, пол (категорический) и месяц (категориальный) являются ковариатами. Результаты показали, что трехстороннее взаимодействие является значительным, что указывает на то, что соотношение между весом и длиной крыла действительно меняется в течение месяцев и между полами. Однако с этим анализом существует проблема. [Рисунок 11](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f11) показывает данные в coplot, который является отличным графическим инструментом для визуализации потенциального наличия взаимодействий. График содержит несколько диаграмм рассеивания длины и веса крыла; по одному на каждый месяц и пол. В каждую диаграмму рассеяния добавляется двумерная линейная регрессионная линия; если все линии параллельны, то, вероятно, не существует значительного взаимодействия (хотя только регрессионный анализ может рассказать нам, действительно ли это так). В нашем примере линии имеют разные наклоны, что указывает на потенциальное наличие взаимодействий. Однако в некоторые месяцы число наблюдений очень мало, и в сентябре нет данных по самцам. Разумным подходом было бы повторить анализ только в период с июня по август.

## Шаг 8: Независимы ли наблюдения от переменной отклика?

Важнейшим предположением большинства статистических методов является то, что наблюдения независимы друг от друга (**[Hurlbert, 1984](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b20" \o "Ссылка на библиографическую цитату)**), что означает, что информация из любого одного наблюдения не должна предоставлять информацию о другом после того, как были учтены другие переменные. Эта концепция лучше всего объясняется примерами.

Наблюдения из набора данных о численности воробьев были взяты в нескольких местах. Если птицы в местах, близких друг к другу, имеют характеристики, которые более похожи друг на друга, чем птицы из мест, разделенных большими расстояниями, тогда мы будем нарушать предположение о независимости. Другим примером является выборка отдельных особей одного семейства (например, всех молодых из одного гнезда); Эти особи могут быть больше похожи друг на друга, чем случайные особи в популяции, потому что они имеют сходный генетический состав и схожую историю родительской подготовки.

Когда возникает такая зависимость, статистическая модель, используемая для анализа данных, должна учитывать ее. Например, путем моделирования любых пространственных или временных отношений или путем размещения данных в иерархической структуре (например, птенцы могут быть вложены в гнезда). Однако проверка на независимость не всегда легка. В **[Zuur et al. (2009a)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)** было проанализировано большое количество наборов данных, в которых зависимость между наблюдениями сыграла свою роль. Примеры включают количество биолюминесценции на участках вдоль океанического градиента глубины, отношение изотопов азота в зубах китов в зависимости от возраста, значений рН в ирландских реках, количество амфибий, убитых автомобилями в разных местах вдоль дороги, поведение кормления различных веретенников на пляже, количество болезнетворных споров, поражающих личиночных медоносных пчел из нескольких ульев и количество криков птенцов совы по прибытии родителя. Другая часто встречающаяся ситуация, в которой необходимо обеспечить независимость, - когда в наборе данных существует филогенетическая структура (т. е. зависимость, обусловленная совместной родословной).

Существует множество способов включения временной или пространственной структуры зависимостей в модель для анализа. К ним относятся использование отстающих переменных отклика в виде ковариатов (**[Brockwell & Davis 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b2" \o "Ссылка на библиографическую цитату)**), моделирование смешанных эффектов (**[Pinheiro & Bates 2000](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b30" \o "Ссылка на библиографическую цитату)**), наложение остаточной корреляционной структуры с использованием обобщенных наименьших квадратов (**[Zuur et al., 2009a](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)**) или изменение параметров регрессии со временем (**[Harvey, 1989](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b18" \o "Ссылка на библиографическую цитату)**). Также возможно подобрать модель с корреляционной структурой и без нее и сравнить модели с использованием критерия выбора или проверки гипотезы (**[Pinheiro & Bates 2000](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b30" \o "Ссылка на библиографическую цитату)**). Наличие структуры зависимости в исходных данных может быть смоделировано с помощью ковариата, такого как месяц или температура, или включение функции сглаживания времени или двумерных более сглаженных пространственных координат (**[Wood 2006](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b1000" \o "Ссылка на библиографическую цитату)**). Независимо от используемого метода, остатки модели не должны содержать никакой структуры зависимости. Довольно часто остаточная корреляционная структура бывает вызвана важным ковариатом, который не был измерен. Если это так, может быть невозможно решить проблему.

При использовании методов регрессии предположение о независимости довольно важно, и нарушение может увеличить ошибку типа I. Например, **[Ostrom (1990)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b29" \o "Ссылка на библиографическую цитату)** показал, что игнорирование автокорреляции может дать P-значения, которые завышены на 400%. Следовательно, важно проверить, есть ли зависимость в исходных данных перед проведением анализа, а также остатки после этого. Эти проверки можно выполнить, построив переменную отклика, напротив времени или пространственных координат. Любая четкая картина является признаком зависимости. Этот подход более сложен, если нет четкой последовательности наблюдений (например, несколько наблюдений на одном объекте), но в этом случае можно включить структуру зависимости с использованием случайных эффектов (**[Pinheiro & Bates 2000](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b30" \o "Ссылка на библиографическую цитату)**; **[Fitzmauriceи другие. 2004](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b13" \o "Ссылка на библиографическую цитату)**; **[Brown & Prescott 2006](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b3" \o "Ссылка на библиографическую цитату)**; **[Zuur et al. 2009a](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)**). [На рис. 12а, с](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f12) показаны короткие временные ряды, иллюстрирующие наблюдаемое множество двух видов птиц на берегу Аргентины в течение 52-недельного периода (E. Ieno, неопубликованные данные). В первом ряду показано большое количество куликов белого калибра Calidris fuscicollis в течение первых 20 недель, за которыми следуют нули (поскольку вид мигрирует), а затем через 38 недель их количество основа увеличивается. Второй временной ряд не показывает четкой картины изобилия водорослей (Larus dominicanus).

Более формальным способом оценки временной зависимости является построение автокорреляционных функций (ACF) для регулярных интервалов времени или вариограмм для нерегулярно разнесенных временных рядов и пространственных данных (**[Schabenberger & Pierce 2002](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b37" \o "Ссылка на библиографическую цитату)**). ACF вычисляет корреляцию Пирсона между временными рядами и теми же временными рядами, сдвинутыми на k единиц времени. [На рисунках 12b, d](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f12) показана автокорреляция временных рядов на панелях (a) и (c). На панели (б) показана существенная корреляция с временной задержкой k  =  1 и k  =  2. Это означает, что множество в момент времени t зависит от численности в момент времени t  - 1 и t - 2, и любой из упомянутых выше методов может быть применен. Для временного ряда L. dominicanus нет значительной автокорреляции.

## Обсуждение

Все проблемы, описанные в этом документе, и стратегии их решения, применяются во всех экологических исследованиях, но они особенно важны, когда результаты должны использоваться для руководства управленческими решениями или общественной политикой из-за последствий совершения ошибки. В последние годы все большее внимание уделяется совокупности данных, поддерживающих конкретные методы управления (**[Roberts, Stewart & Pullin 2006](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b34" \o "Ссылка на библиографическую цитату)** , **[Pullin & Knight 2009](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b31" \o "Ссылка на библиографическую цитату)**), и экологи становятся все более изощренными в статистических методах, которые они используют (например, **[Ellison 2004](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b10" \o "Ссылка на библиографическую цитату)**; **[Stephens и др. 2005](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b39" \o "Ссылка на библиографическую цитату)** ; **[Robinson & Hamann 2008](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b35" \o "Ссылка на библиографическую цитату)** , [**Копер и Manseau 2009**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b22) ; [**Закон и другие. 2009**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b24) ; **[Sonderegger et al.2009](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b38" \o "Ссылка на библиографическую цитату)**). Но более фундаментальные вопросы о целесообразности базовых данных для данного анализа могут быть столь же важными для обеспечения наилучшие политики, и были получены из экологических исследований.

В этой статье мы обсудили ряд подводных камней, которые могут серьезно повлиять на результаты анализа. Некоторые из этих проблем хорошо известны, в некоторых случаях, но даже общеизвестные предположения по-прежнему часто нарушаются в экологической литературе. Во всех случаях проблемы могут привести к неправильным статистическим моделям. Таких проблем можно избежать только путем применения систематической разведки данных, до их анализа ([рис.1](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#f1)).

Хотя мы представили наш протокол как линейную последовательность, его следует использовать гибко. Не каждый набор данных требует каждого шага. Например, некоторые статистические методы не требуют нормальности (например, PCA), и поэтому нет смысла делать гистограммы. Порядок применения этапов может также зависеть от конкретного набора данных. И для некоторых методов допущения могут быть проверены только путем применения шагов исследования данных после проведения анализа. Например, в линейной регрессии нормальность и однородность должны быть проверены с использованием остатков, созданных моделью. Вместо того чтобы просто следовать протоколу, просто отмечая каждую точку, мы призываем пользователей подвергнуть его некоторому сомнению и с большим вниманием отнестись к данным. Убедившись, что каждый вопрос был адекватно рассмотрен таким образом, который имеет биологический смысл, набор данных должен быть готов для основного анализа.

Данные по экологическим полям, как правило, являются шумными, полевые условия непредсказуемыми и предварительные знания часто ограничены. В прикладной сфере изменения в финансировании, правилах проведения и приоритетах исследований еще больше усложняют ситуацию. Такая ситуация особенно характерна для долгосрочных исследований, где первоначальные цели часто меняются с учетом обстоятельств (например, использование множества наборов данных для изучения реагирования видов на изменение климата). По всем этим причинам идеализированная ситуация, при которой эколог тщательно разрабатывает свой анализ априорно, а затем собирает данные, может быть скомпрометирована или неактуальна.  Для приспособления анализов к таким обстоятельствам наличие аналитической гибкости является важным навыком для прикладного эколога, но для этого требуется глубокое понимание сдерживающих/принудительных предположений, налагаемых данным набором данных.

Лучшие решения возникающих проблем могут различаться. Однако, зачастую, экологи просто преобразуют данные, чтобы избежать нарушения предположений. Существует три основные причины трансформации:  уменьшение эффекта выбросов (особенно в ковариатах), стабилизация дисперсии и структуризация/линеаризация отношений. Однако, используя более совершенные методы, такие как GLS и GAM, проблемы гетерогенности и нелинейности могут быть решены, что делает трансформацию менее важной. **[Zuur et al. (2009a)](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b41" \o "Ссылка на библиографическую цитату)** показал, как использование преобразования данных привело к различным выводам о долгосрочных тенденциях по сравнению с соответствующим анализом с использованием нетрансформированных данных; поэтому лучше избегать преобразования переменных отклика. Если используется трансформация, инструменты автоматического выбора, такие как Мостеллер и правило [**выталкивания Тьюки**](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full#b28) (**[Mosteller & Tukey 1977](http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/full" \l "b28" \o "Ссылка на библиографическую цитату)**) следует использовать с большой осторожностью, потому что эти методы игнорируют эффекты ковариатов. Другим аргументом против преобразований является необходимость последующего обратного преобразования значений для прогнозирования; может быть не всегда понятно, как это сделать и все еще иметь возможность интерпретировать результаты по исходной шкале переменной отклика. Также важно убедиться, что преобразование действительно решает проблему; даже обычно рекомендуемые преобразования не всегда работают. Суть в том, что выбор конкретного преобразования - это вопрос проб и ошибок.

Факт в том, что разведка данных не должна использоваться для определения вопросов, которые исследование намеревается проверить. Следует сообщать о каждом этапе разведки, и любое удаление выброса должно быть обосновано и упомянуто. Основания для преобразования данных должны быть обоснованы на основе исследовательского анализа (например, доказательства того, что предположения модели были нарушены и, что трансформация исправила ситуацию).

Следует избегать применения исследования данных (например, диаграммы рассеяния для визуализации связей между объясняющими и переменными отклика) для создания гипотез, чтобы затем использовать те же данные для проверки этих гипотез. Если кто-то имеет ограниченное априорное знание, то обоснованный подход заключается в создании двух наборов данных; применить исследование данных в первом наборе данных для создания гипотез и использовать второй набор данных для проверки гипотез. Однако такой процесс применим только для больших наборов данных. Независимо от конкретной ситуации, регулярное использование и понятная отчетность по систематическому исследованию данных улучшат качество экологического исследования и любые прикладные рекомендации, которые он дает.