УДК 574.587+578.087(282.247.431.2)+595.7(476)2

ОБОБЩЕННЫЕ ЛИНЕЙНЫЕ СМЕШАННЫЕ МОДЕЛИ (GLMM) В ИССЛЕДОВАНИЯХ ЭКОЛОГИИ СООБЩЕСТВ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИЧЕСКОЙ СРЕДЫ R

Г. Г. СУШКО[1)], ***А. С. ТКАЧЁНОК***[1)]

[1)]*Витебский государственный университет им. П. М. Машерова,*
пр. Московский, 33, 210038, г. Витебск, Беларусь

Анализ данных в экологии сообществ часто имеет определенные трудности, так как стандартные параметрические методы неприменимы вследствие того, что экологические данные редко соответствуют закону нормального распределения, отсутствуют линейные соотношения между переменными, может быть коллинеарность между объясняющими переменными и избыточная дисперсия в наборах данных. В предложенном исследовании рассматривается подход, основанный на применении обобщенных регрессионных моделей со смешанными эффектами (GLMM), который позволяет анализировать данные синэкологических исследований с учетом указанных сложностей, а также включать в анализ не только количественные, но и качественные предикторы. С использованием собственных результатов исследований ассамблей жужелиц в нескольких типах лесов продемонстрированы этапы выполнения GLMM в статистической среде R. Предложен несложный программный код для GLMM, доступный для использования начинающим исследователям. Детально рассмотрен протокол разведочного анализа данных, обосновано использование соответствующих пакетов, включая lme4, performance, car и др.

Ключевые слова: экология сообществ; регрессионный анализ; GLMM; статистическая среда R.

GENERALIZED LINEAR MIXED MODELS (GLMM) IN COMMUNITY ECOLOGY STUDIES USING THE R STATISTICAL ENVIRONMENT

G. G. SUSHKO[a], ***A. S. TKACHENOK***[a]

[a]*Vitebsk State University named after P. M. Masherov,*
33 Moskovsky Avenue, Vitebsk 210038, Belarus
Corresponding author: G. G. Sushko (gennadis@rambler.ru)

Data analysis in community ecology often has certain difficulties, since standard parametric methods are inapplicable due to the fact that ecological data rarely normal distributed, there are no linear relationships between variables, there may be collinearity between explanatory variables and overdispersion in data sets. The proposed article considers an approach based on the use of regression generalized linear mixed models (GLMM), which allows analyzing data from synecological studies taking into account the above difficulties, as well as including not only quantitative but also qualitative predictors in the analysis.

Образец цитирования:
Сушко ГГ, Ткачёнок АС. Обобщенные линейные смешанные модели (GLMM) в исследованиях экологии сообществ с использованием статистической среды R. *Журнал Белорусского государственного университета. Экология.* 2025;1:37–47.
https://doi.org//10.46646/2521-683X/2025-1-37-47

For citation:
Sushko GG, Tkachenok AS. Generalized linear mixed models (GLMM) in community ecology studies using the R statistical environment. *Journal of the Belarusian State University. Ecology.* 2025;1:37–47. Russian.
https://doi.org//10.46646/2521-683X/2025-1-37-47

Авторы:
Геннадий Геннадьевич Сушко – доктор биологических наук, профессор; заведующий кафедрой экологии и географии.
Анастасия Сергеевна Ткачёнок – студентка, факультет химико-биологических и географических наук.

Authors:
Gennadi G. Sushko, doctor of science (biology), full professor; head of the department of ecology and geography.
gennadis@rambler.ru
Anastasia S. Tkachenok, student, faculty of chemical, biological and geographical sciences.
updown9700@gmail.com

Using our own results of studies of ground beetle assemblages in several types of forests, the stages of GLMM implementation in the R statistical environment are demonstrated. A simple program code for GLMM is proposed, available for use by novice researchers. The protocol for exploratory data analysis is considered in detail, the use of appropriate packages, including lme4, performance, car, etc. is justified.

Keywords: community ecology; regression analysis; GLMM; R statistical environment.

Введение

Одной из ключевых задач в исследованиях экологии сообществ является поиск факторов среды, оказывающих влияние на показатели структурной организации, видовой состав, число видов, обилие и др. В данном случае для формирования доказательной базы и обоснованных выводов требуются специфические методы анализа и статистической обработки данных. В наши дни, благодаря развитию компьютерной техники и программного обеспечения, инструментарий для аналитической обработки данных экологических исследований постоянно расширяется. В частности, в экологии все большее применение находят возможности статистической среды R, которая в последнее время является одной из наиболее динамично развивающихся платформ анализа данных, объединяет язык программирования высокого уровня и мощные библиотеки программных модулей для вычислительной и графической обработки данных [1].

Одним из методов оценки влияния факторов среды на различные параметры сообществ является регрессионный анализ. В общих чертах его можно охарактеризовать как сборный набор методов, который применяется для предсказания изменчивости переменной отклика (или зависимой переменной) в зависимости от вариации значений одной или более предсказывающих переменных (независимые, или объясняющие переменные) [2; 3]. Регрессионный анализ можно использовать для обнаружения независимых переменных, которые оказывают влияние на зависимую с целью описания типа взаимосвязи, а также для составления уравнения, позволяющего предсказать значения зависимой переменной по изменению значений независимых переменных [3]. В синэкологии под независимыми переменными понимают используемые биотические и абиотические факторы среды, которые могут быть как количественными (температура, влажность, концентрация химических веществ в воде, или почве), так и качественными (наличие или отсутствие признака, степень антропогенного воздействия) [4; 5]. Классификация методов регрессионного анализа достаточно обширна и ее детальное рассмотрение выходит за рамки данной публикации. Ознакомиться подробно с различными методами регрессионного анализа можно в специальной литературе [1; 3; 6; 7]. Чаще всего в русскоязычных научных публикациях по экологии можно увидеть анализ с использованием простой линейной регрессии, с помощью, которой оценивается влияние одной переменной (объясняющей) на другую (переменная отклика). Однако для ряда исследований возникает необходимость выявить влияние комплекса факторов среды на зависимую переменную, например на число видов определенного таксона в местообитании или их спектре. В данном случае уместно использование таких типов анализа, как обобщенные линейные (generalized linear models, GLM) и смешанные модели множественной регрессии (generalized linear mixed models, GLMM) [8–10]. Эти модели имеют важное значение. Они получили широкое применение, так как экологические данные редко соответствуют закону нормального распределения, а анализируемые переменные не всегда характеризуются линейными зависимостями [8; 9]. Применение GLM и GLMM с использованием биномиальной, отрицательной биномиальной и пуассоновской регрессии дает возможность решать эти проблемы [8; 10]. Отдельно следует выделить смешанные модели множественной регрессии (GLMM), которые позволяют детализировать оценку влияния независимых переменных, благодаря возможности включения в модель фиксированных и случайных эффектов, а также как количественные, так и качественные (категориальные) переменные [8]. Это является несомненным достоинством GLMM (по сравнению с другими методами, для исследований экологии сообществ).

Случайные эффекты – это переменные, которые могут варьировать случайным образом, уровни которых определяются случайной выборкой из совокупности всех возможных уровней. Например, измерения выполнены в разное время, в разных точках пространства, до и после воздействия фактора и др. [8; 11]. Таким образом, одни и те же детали, произведенные на разных станках или выполненные разными мастерами, как правило, имеют некоторые различия. Точно также организмы из одной популяции будут иметь те или иные различия. Фиксированные эффекты – это переменные, которые являются предметом интереса исследователя, то есть независимые переменные, уровни которых он контролирует [8; 11]. Например, мы хотим выяснить, какие факторы среды влияют на параметры биоразнообразия в нескольких биотопах. Для этого сами выбираем биотопы для исследований. Следовательно, тип биотопа можно рассматривать как фиксированный эффект. Далее мы выполняем измерения абиотических факторов среды для оценки их воздействия на зависимую переменную, которые также относятся к фиксированным эффектам.

В то же время на один и тот же фактор можно посмотреть и как фиксированный и как случайный в зависимости от целей исследования. В частности, как уже сказано, биотоп можно рассматривать как фиксированный фактор, так как мы заинтересованы в том, чтобы сделать выводы об этих конкретных местах обитания. Случайный эффект наблюдается, когда данные включают рандомную выборку из многих возможных уровней фактора. Важно то, что данные выборок (например, число видов, обилие и др.) из одного биотопа могут иметь некоторую корреляцию между собой (пространственная автокорреляция), поскольку выборки получены из локалитетов со сходными условиями окружающей среды. Даже если вас не интересует эффект каждого локалитета, где взята выборка, вы должны учитывать эту потенциальную корреляцию, чтобы сделать выводы о биотопах в целом. Следовательно, в таком случае биотоп следует рассматривать как случайный фактор. Переменные со случайным эффектом, как правило, являются качественными переменными (категориальными) [8; 9].

Нетрудно убедиться, что обобщенные линейные смешанные модели (GLMM) – это удобный инструмент для описания взаимосвязи между переменными в экологических исследованиях. В тоже время в доступной русскоязычной литературе описания методик (протоколов) выполнения данного анализа крайне ограничены. Как показала практика, исследователи, связанные с анализом данных, ограничиваются в своей работе, как правило, средствами *Microsoft Office Excel* или недостаточно гибкой программы *Statistica,* тогда как статистическая среда R является по-настоящему интеллектуальной, гибкой и богатой методами [1]. В тоже время для работы с R требуются навыки написания программного кода, что для многих биологов вызывает затруднение. В связи с этим цель данной работы – предоставить пример несложного программного кода в R для GLMM, а также ознакомить читателя с необходимыми этапами анализа и пакетами для их реализации.

Материалы и методы исследования

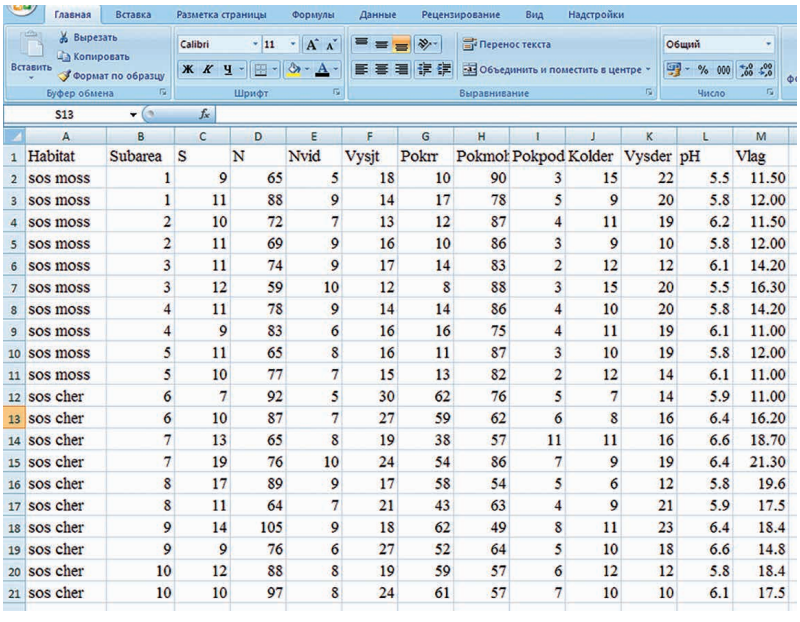
В качестве иллюстративного материала использованы результаты собственных исследований параметров биоразнообразия жужелиц в нескольких типах сосновых лесов в Белорусском Поозерье. В качестве переменных отклика были выбраны число видов и число особей в выборках (по 10 выборок в каждом биотопе). В качестве объясняющих переменных использованы: число видов высших сосудистых растений, высота травяно-кустарничкового яруса (см), проективное покрытие травяно-кустарничкового яруса (%), проективное покрытие мохового яруса (%), проективное покрытие подстилки (%), pH, влажность гумусового слоя (%) (измерены на 10 площадках 1х1 м); количество деревьев, высота деревьев (м) (измерены на 10 площадках 10×10 м).

Для выполнения анализа использованы пакеты factoextra, corrplot, Hmisc, lme4, MuMIn, performance и car статистической среды R [12; 13].

Результаты исследования и их обсуждение

При подготовке таблицы данных следует учесть, что переменные должны быть расположены в столбцах, которые нужно обозначить латинскими буквами, сократив названия, а в строках приведены данные по выборкам. В нашем образце использованы следующие сокращения: S – число видов, N – число особей в выборке, Nvid – число видов высших сосудистых растений, Vysjt – высота травяно-кустарничкового яруса, Pokrr – проективное покрытие травяно-кустарничкового яруса, Pokmoh – проективное покрытие мохового яруса, Pokpod – проективное покрытие подстилки, Kolder – количество деревьев, Vysder – высота деревьев, pH – водородный показатель, Vlag – влажность гумусового слоя. Отдельно поясним наполнение первых двух столбцов. Habitat – категориальная переменная, где каждая строка – это название типа леса, в котором получены выборки (sos moss – сосняк мшистый, sos cher – сосняк черничный). Subarea – категориальная переменная, где цифра в строке обозначает более точное место проведения исследований (это может быть подрегион в пределах региона, страта, синузия в фитоценозе и др.). В частности, учеты, выполненные в лесу в одном регионе, обозначены номером 1, в другом – номером 2, в третьем – номером 3 (рис. 1). Назовем их условно подрегионами. Такая детализация нужна для снятия проблемы автокорреляции, которая может возникнуть, если точки учетов расположены близко. Если вы уверены, что пространственная автокорреляция не является проблемой, так как выборки получены в локалитетах на больших расстояниях, эту переменную можно не вводить. Сохраняем таблицу в формате .csv загружаем ее в RStudio, указав ее расположение: demo<- read.csv («C:\\Users\\datapinefor.csv», header = T, sep = «;»). С помощью view(demo) убеждаемся, что она загружена правильно.

Анализ состоит из трех блоков: разведочный анализ данных, собственно построение GLMM и тестирование построенной модели. Каждый из этих блоков включает ряд этапов.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Habitat	Subarea	S	N	Nvid	Vysjt	Pokrr	Pokmol	Pokpod	Kolder	Vysder	pH	Vlag
2	sos moss	1	9	65	5	18	10	90	3	15	22	5.5	11.50
3	sos moss	1	11	88	9	14	17	78	5	9	20	5.8	12.00
4	sos moss	2	10	72	7	13	12	87	4	11	19	6.2	11.50
5	sos moss	2	11	69	9	16	10	86	3	9	10	5.8	12.00
6	sos moss	3	11	74	9	17	14	83	2	12	12	6.1	14.20
7	sos moss	3	12	59	10	12	8	88	3	15	20	5.5	16.30
8	sos moss	4	11	78	9	14	14	86	4	10	20	5.8	14.20
9	sos moss	4	9	83	6	16	16	75	4	11	19	6.1	11.00
10	sos moss	5	11	65	8	16	11	87	3	10	19	5.8	12.00
11	sos moss	5	10	77	7	15	13	82	2	12	14	6.1	11.00
12	sos cher	6	7	92	5	30	62	76	5	7	14	5.9	11.00
13	sos cher	6	10	87	7	27	59	62	6	8	16	6.4	16.20
14	sos cher	7	13	65	8	19	38	57	11	11	16	6.6	18.70
15	sos cher	7	19	76	10	24	54	86	7	9	19	6.4	21.30
16	sos cher	8	17	89	9	17	58	54	5	6	12	5.8	19.6
17	sos cher	8	11	64	7	21	43	63	4	9	21	5.9	17.5
18	sos cher	9	14	105	9	18	62	49	8	11	23	6.4	18.4
19	sos cher	9	9	76	6	27	52	64	5	10	18	6.6	14.8
20	sos cher	10	12	88	8	19	59	57	6	12	12	5.8	18.4
21	sos cher	10	10	97	8	24	61	57	7	10	10	6.1	17.5

Рис. 1. Фрагмент анализируемой таблицы данных

Fig. 1. Fragment of the analyzed data table

Преимущества GLMM для использования в экологии обусловлено тем, что не требуется предположений о нормальности распределения данных, линейных взаимоотношений зависимой и независимой переменной и однородности дисперсии. Основными этапами разведочного анализа являются проверка на выбросы, коллинеарность и чрезмерную дисперсию [14].

Проверка на выбросы. Причиной наличия выброса может быть ошибка ввода или наблюдение, которое имеет относительно большую или малую величину по сравнению с большинством других наблюдений. Графические инструменты, которые обычно используются для обнаружения выбросов – диаграммы рассеяния и диаграммы размаха (боксплоты) [14]. Ниже приведен код построения боксплотов для зависимых переменных – число видов и число особей в выборках.

```
demo<- read.csv("C:\\Users\\datapinefor.csv", header = T, sep = ";")
par(mfrow = c(1,2), mar = c(5,5,4,1))
boxplot(demo[,3:3], outpch = 16, outcol = "darkorange1", outcex = 1, boxwex = 0.45, boxlwd = 0.1,
whisklwd = 2, whiskcol = c("orange4", "#0e598f"), xlab = "S")
boxplot(demo$N, outpch = 16, outcol = "darkorange1", outcex = 1, boxwex = 0.45, boxlwd = 0.1, whisklwd = 2,
whiskcol = c("orange4", "#0e598f"), xlab = "N")
```

Для улучшения визуализации боксплота использована базовая графика с установкой параметров с помощью функции par() для выведения двух графиков на одну панель.

Как следует из рис. 2, переменная S характеризуется наличием двух выбросов. При детальном рассмотрении значений, соответствующих выбросам, убеждаемся, что это не ошибка ввода, а наблюдения с высокими значениями (в двух выборках число видов было выше, по сравнению с другими).

Проверка коллинеарности. Коллинеарность – это наличие корреляции между независимыми переменными. Если коллинеарность проигнорирована, то результаты анализа могут привести к неправильным и запутанным выводам. Например, все использованные независимые переменные будут иметь значимое влияние, хотя некоторые из них коррелируют между собой [14]. Поэтому логично, что ключевое влияние будет оказывать одна из двух или более переменных, которые скоррелированы. Выбрать наиболее важную из двух или более переменных задача не всегда простая. Для этого используют специальный коэффициент – фактор инфляции дисперсии (Variance Inflation Factor, VIF). Чем он выше для j-го предиктора, тем сильнее линейная связь между этим и остальными предикторами. При определенном значении VIF переменная исключается из модели. Есть разные мнения по поводу пороговых значений VIF. Чаще всего критическим считают значение VIF = 5, несколько реже VIF = 10 [14]. Функции для автоматического расчета VIF реализованы в нескольких пакетах для R. Одним из примеров таких функций является vif() из пакета car. Для проверки наличия коллинеарности на начальном этапе используют разные подходы. Чаще всего в их числе анализ главных компонент (PCA) и построение корреляционных матриц. Далее, в случае выявления коллинеарности, одна или несколько переменных удаляется из модели по значению VIF.

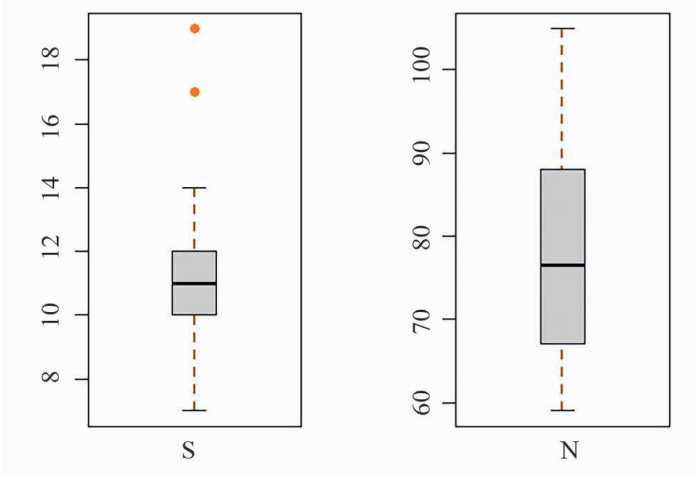


Рис. 2. Проверка на наличие выбросов переменных отклика

Fig. 2. Testing for outliers in response variables

Для выполнения PCA с помощью функции prcomp() используем следующий код:

```
demo<- read.csv("C:\\Users\\datapinefor.csv", header = T, sep = ";")
demodata<- demo[,5:ncol(demo)]
res.pca<- prcomp(demodata, scale = T)
print(res.pca)
summary(res.pca)
library(factoextra)
fviz_pca_var(res.pca, col.var = "green4")
```

Для улучшения визуализации биплота использован пакет factoextra и функция этого пакета fviz_pca_var().

Диаграмма ординации (рис. 3) демонстрирует коллинеарность некоторых переменных, так как соответствующие им вектора либо накладываются друг на друга, либо расположены рядом. Например, pH и проективное покрытие травяно-кустарничкового яруса, количество деревьев и высота деревьев.

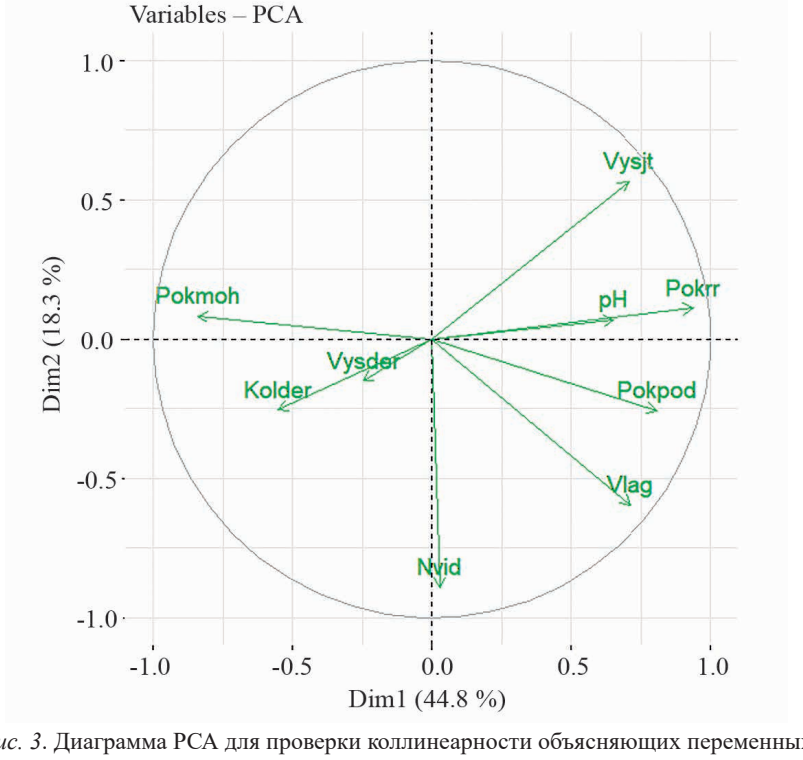


Рис. 3. Диаграмма PCA для проверки коллинеарности объясняющих переменных

Fig. 3. PCA diagram for testing collinearity of explanatory variables

Выполним корреляционный анализ, рассчитаем коэффициент корреляции Спирмена. Пары переменных, которые имеют корреляцию между собой, выберем при коэффициенте корреляции Rs > 6 [14].

Для построения корреляционной матрицы и ее визуализации (оставляем только статистически значимые коэффициенты корреляции, $p < 0.05$) с помощью пакетов Hmisc (расчет коэффициентов корреляции, применена функция rcorr() для вычисления коэффициента корреляции и его p-уровня) и corrplot (визуализация корреляционной матрицы) используем следующий код:

```
library(Hmisc)
library(corrplot)
demo <- read.csv("C:\\Users \\datapinefor.csv", header = T, sep = ";")
demodata<- demo[,5:ncol(demo)]
res2 <- rcorr(as.matrix(demodata), type = "spearman")
corrplot(res2$r, type="upper", order="hclust",
p.mat = res2$P, sig.level = 0.05, insig = "blank", addCoef.col = "black")
```

Анализируя корреляционую матрицу (рис. 4), нетрудно заметить, что среди переменных есть такие, которые имеют высокую значимую корреляцию. В частности, между Pokmoh и Pokrr (rs = - 0.82, $p < 0.05$), Pokmoh и Pokpod (rs = - 0.70, p < 0.05), Vlag и Pokpod (rs = 0.64, p < 0.05) и др. Далее, в ходе подгонки модели будем принимать решение об исключении конкретных переменных с VIF > 5.

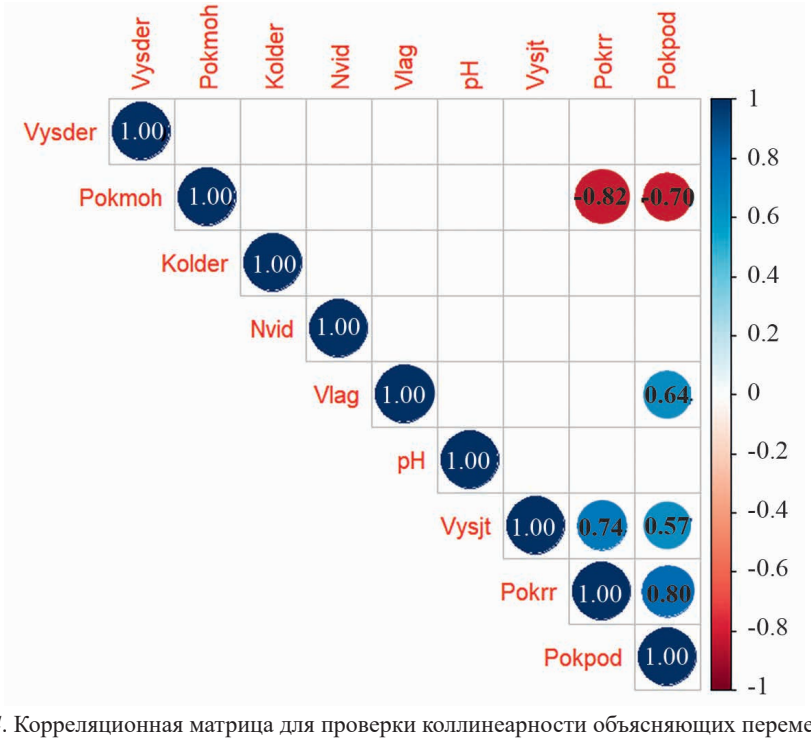


Рис. 4. Корреляционная матрица для проверки коллинеарности объясняющих переменных

Fig. 4. Correlation matrix for testing collinearity of explanatory variables

Следующим важным этапом разведочного анализа является проверка на наличие чрезмерной дисперсии (overdispersion). Такая дисперсия возникает в данных, если она превышает среднее значение. Чрезмерная дисперсия возникает еще и тогда, когда наблюдаемая дисперсия данных больше, чем предсказывает распределение Пуассона. При распределении Пуассона среднее значение и дисперсия равны, что может привести к занижению стандартных ошибок и неправильным выводам [8; 9]. Наличие чрезмерной дисперсии имеет значение для выбора типа распределения при построении модели. Во многих случаях, если данные не соответствуют закону нормального распределения, модели строят с распределением Пуассона, но только в случае отсутствия чрезмерной дисперсии. Если она выявлена, следует выбрать другой тип распределения, например, отрицательное биноминальное.

Данный этап совпадает с предварительной подгонкой модели с использованием всех переменных. Для построения GLMM используем пакет lme4, а для выявления чрезмерной дисперсии – пакет performance. Мы использовали также встроенную функцию пакета lme4, которая указывает количество итераций (прогонов или повторов выполнения модели) glmer Control [13]. Обратите внимание также и на то, что в модели добавлены два случайных фактора (1|Habitat) и (1|Subarea).

```
demo<- read.csv("C:\\Users\\ datapinefor.csv", header = T, sep = ";")
library(lme4)
model_global<- glmer(S ~ Nvid + Vysjt + Pokrr + Pokmoh + Pokpod + Kolder + Vysder + pH  + Vlag +
(1|Habitat) + (1|Subarea), data = demo, family = poisson(),
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000)))
library(performance)
check_overdispersion(model_global)
```

Результаты проверки, которые приведены ниже, показали отсутствие чрезмерной дисперсии. Следовательно, использование распределения Пуассона здесь уместно.

```
>check_overdispersion(model_global)
Dispersion ratio = 0.116
Pearson's Chi-Squared = 0.928
    p-value = 0.999
No overdispersion detected
```

Далее переходим к следующему блоку анализа – непосредственному построению модели. Первую модель построим для зависимой переменной – число видов (S). Вначале нужно определиться, какие независимые переменные следует оставить. Для расчета VIF используем пакет car. При этом в модели используем все имеющиеся независимые переменные.

```
demo<- read.csv("C:\\Users\\ datapinefor.csv", header = T, sep = ";")
library(lme4)
library(car)
model_S<-glmer(S ~ Nvid+Vysjt+Pokrr+Pokmoh+Pokpod+Kolder+Vysder+pH+Vlag+
(1|Habitat)+(1|Subarea), data = demo, family = poisson(), glmerControl(optimizer = "bobyqa", optCtrl =
list(maxfun = 100000)))
summary(model_S)
vif(model_S)
```

Получены следующие значения VIF:

```
Nvid Vysjt Pokrr Pokmoh Pokpod Kolder Vysder pH  Vlag
4.071 7.23 14.52 5.80   3.23  1.67  1.32  2.20 6.13
```

Как видно, несколько переменных имеют значение VIF > 5. Уберем переменную с самым высоким значением VIF = 14.52 (Pokrr) и повторим анализ еще раз. Полученные результаты демонстрируют пороговое значение VIF = 5.08 для переменной Vlag, следовательно, ее также исключаем из модели. Теперь все независимые переменные имеют VIF > 5. Таким образом, в модель мы включаем следующие переменные: Nvid, Vysjt, Pokmoh, Pokpod, Kolder, Vysder и pH. Далее необходимо выполнить подгонку моделей и выбрать лучшую модель, используя различные сочетания независимых переменных. Такие сочетания строятся путем пошагового удаления или добавлений той или иной переменной для получения в итоге такого сочетания переменных, которое наилучшим образом сможет описать зависимости, если такие имеются.

Пакет MuMIn содержит функцию dredge(), которая осуществляет построение всех возможных вариантов моделей из различных комбинаций независимых переменных, включенных в изначальную полную модель [13]. Для оценивания моделей в данном случае используется информационный критерий Акаике (AICc) и показатель ΔAICc. Лучшей считается та модель (gi), которая среди всех построенных моделей (Gr), имеет наименьшее значение AICc. Значение второго показателя, как правило, устанавливается в определенных пределах. В частности, для лучших моделей ΔAICc < 3 [8; 9]. Еще один показатель, определяемый как нормированное значение силы обоснованности модели – это относительная вероятностная мера каждой модели gi в ансамбле из Gr моделей, или ее вес (weight). Найденные веса позволяют оценить относительную важность каждой независимой переменной как сумму весов тех моделей, в которых присутствует соответствующая независимая переменная. Все эти показатели включены в итоговую таблицу анализа. Теперь построим серию моделей с разными комбинациями независимых переменных используя функцию dredge() пакета MuMIn.

```
demo<- read.csv("C:\\Users\\ datapinefor.csv", header = T, sep = ";")
library(lme4)
library(MuMIn)
model_S_2 <- glmer(S ~ Nvid + Vysjt + Pokmoh + Pokpod + Kolder + Vysder + pH      + (1|Habitat) +
(1|Subarea), data = demo, family = poisson(),
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000)))
```

options(na.action=na.fail)

dredge_ S <- dredge(model_S_2)

Как показали результаты построения моделей, наилучшим является использование в итоговой модели одной переменной Nvid, так как она характеризуется наибольшим весом (0.287) и наименьшими значениями AICc (100.8) и ΔAICc (0.00) (рис. 5).

Model selection table

(Intrc)	Koldr	Nvid	pH	Pokmh	Pokpd	Vysdr	Vysjt	df	logLik	AICc	delta	weight
1.4620		0.1220						4	-45.043	100.8	0.00	0.287
1.3500		0.1170			0.03057			5	-44.517	103.3	2.57	0.080
1.7310		0.1256		-0.00408				5	-44.681	103.6	2.90	0.067
0.4156		0.1263	0.16					5	-44.722	103.7	2.98	0.065
1.6770		0.1224						5	-44.801	103.9	3.14	0.060
1.1550		0.1363					0.010345		-44.810	103.9	3.15	0.059
1.3880		0.1223				0.00422		5	-45.010	104.3	3.55	0.049
2.4290								3	-48.492	104.5	3.73	0.044
2.2310					0.04004			4	-47.597	105.9	5.11	0.022
2.6850				-0.0034				4	-48.238	107.1	6.39	0.012

Рис. 5. Фрагмент таблицы выбора лучшей модели (скриншот RStudio)

Fig. 5. Fragment of the table for selecting the best model (screenshot RStudio)

Выполним анализ с использованием выявленного лучшего набора переменных (в нашем случае одной переменной) используя код:

best_model_S<- glmer(S ~ Nvid + (1|Habitat) + (1|Subarea), data = demo, family = poisson())

summary(best_model_S).

Получаем результаты, приведенные ниже.

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmer Mod']
Family: poisson
Formula: S ~ Nvid + (1 | Habitat) + (1 | Subarea)
Data: demo
AIC BIC log Lik deviance df.resid
98.1 102.1 -45.0 90.1 16
Scaled residuals:
Min 1Q Median 3Q Max
-0.68369 -0.45586 -0.04234 0.27784 1.14744
Random effects:
Groups Name Variance Std.Dev.
Subarea (Intercept) 0.000e+00 0.000e+00
Habitat (Intercept) 1.776e-17 4.215e-09
Number of obs: 20, groups: Subarea, 10; Habitat, 2
Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.46202 0.38574 3.790 0.000151 ***
Nvid 0.12199 0.04719 2.585 0.009729 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Результаты демонстрируют значимую зависимость числа видов жужелиц (p = 0,009) от числа видов растений в местообитаниях. Экологическая интерпретация здесь вполне очевидна, так как с увеличением видового богатства растений, возрастает вероятность увеличения разнообразия фитофагов, которые могут быть жертвами большего числа видов жужелиц, которые являются преимущественно зоофагами.

Далее построим GLMM для переменной N (число особей в выборках), следуя этапам, охарактеризованным выше с целью демонстрации написания кода.

#загружаем в RStudio таблицу с нашими данными, загружаем пакет lme4 и указываем зависимую переменную (N) и через тильду все независимые переменные; указываем тип распределения (с начала распределение Пуассона)

```
demo<- read.csv(«C:\\Users\\datapinefor.csv», header = T, sep = «;»)
library(lme4)
model_global<- glmer(N ~ Nvid +Vysjt + Pokrr + Pokmoh + Pokpod + Kolder +Vysder + pH + Vlag + (1|Habitat) + (1|Subarea), data = demo, family = poisson(), glmerControl(optimizer = “bobyqa”, optCtrl = list(maxfun = 100000)))
#проверяем наличие избыточной дисперсии с помощью пакета performance
library(performance)
check_overdispersion(model_global)
#получаем результаты теста; как видно избыточная дисперсия не выявлена, следовательно, распределение Пуассона остается
Dispersion ratio = 0.774
Pearson’s Chi-Squared = 6.191 p-value = 0.626
No overdispersion detected.
#так как ранее PCA и корреляционный анализ показали наличие коллинеарности независимых переменных, удаляем из анализа те, которые имеют высокую корреляцию, с помощью пакета car и функции VIF().
library(car)
model_N<- glmer(N ~ Nvid +Vysjt + Pokrr +Pokmoh + Pokpod + Kolder +Vysder + pH + Vlag + (1|Habitat) + (1|Subarea), data = demo, family = poisson(), glmerControl(optimizer = “bobyqa”, optCtrl = list(maxfun = 100000)))
vif(model_N)
#результаты показали VIF > 5 для 4 переменных; удаляем переменные с наибольшим VIF
Nvid Vysjt Pokrr Pokmoh Pokpod Kolder Vysder pH Vlag
4.31 8.41 15.29 6.33  3.03  1.74  1.32 1.94 5.73
#повторяем тоже самое без переменной Pokrr
model_N_1 <- <- glmer(N ~ Nvid +Vysjt + Pokrr +Pokmoh + Pokpod + Kolder +Vysder + pH + Vlag + (1|Habitat) + (1|Subarea), data = demo, family = poisson(), glmerControl(optimizer = “bobyqa”, optCtrl = list(maxfun = 100000)))
vif(model_N_1)
#результаты показали VIF < 5 для всех переменных
Nvid Vysjt Pokmoh Pokpod Kolder Vysder pH Vlag
4.22 3.63  3.36  2.92  1.63  1.29 1.82 4.95
#строим серию моделей с разными комбинациями независимых переменных, используя функцию dredge() пакета MuMIn
library(MuMIn)
options(na.action=na.fail)
dredge_N<- dredge(model_N_1)
dredge_N
#получаем таблицу результатов (рис. 6); наилучшим является использование в итоговой модели переменных Nvid, Pokmoh, Vlag, так как эта модель характеризуется наибольшим весом (0.258) и наименьшими значениями AICc (158.5) и ΔAICc (0.00).
```

Model selection table

(Intrc)	Koldr	Nvid	pH	Pokmh	Pokpd	Vysdr	Vysjt	Vlag	df	logLik	AICc	delta	weight
4.845		0.0002		-0.006634				0.0293	4	-73.919	158.5	0.00	0.258
4.974	-0.02006			-0.005577					5	-72.734	159.8	1.25	0.138
4.721		0.0002		-0.006042			0.0042		5	-73.623	161.5	3.03	0.057
4.897				-0.006407		-0.004			5	-73.718	161.7	3.22	0.052
4.708	-0.03390								4	-75.678	162.0	3.52	0.044
4.889				-0.007007	-0.0034				5	-73.895	162.1	3.57	0.043
4.876			-0.004	-0.006681					5	-73.917	162.1	3.62	0.042
4.847				-0.006634					5	-73.919	162.1	3.62	0.042
4.556	-0.02936				0.0215				5	-74.608	163.5	5.00	0.021
4.989	-0.01916			-0.005528		-0.001			6	-72.706	163.9	5.37	0.018
5.006	-0.01999			-0.005859	-0.0025				6	-72.722	163.9	5.40	0.017
5.068	-0.02018		-0.013	-0.005708					6	-72.723	163.9	5.40	0.017
4.945	-0.01912			-0.005516			0.0008		6	-72.727	163.9	5.41	0.017
4.962	-0.02022	0.0017		-0.005569					6	-72.731	163.9	5.42	0.017

Рис. 6. Фрагмент таблицы выбора лучшей модели (скриншот RStudio)

Fig. 6. Fragment of the table for selecting the best model (screenshot RStudio)

```
# выполняем анализ с использованием выявленного лучшего набора переменных
best_model_N<- glmer(N ~ Nvid + Pokmoh + Vlag + (1|Habitat) + (1|Subarea), data = demo, family = poisson())
summary(best_model_N)
#получаем результаты
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: poisson
Formula: N ~ Nvid + Pokmoh + Vlag + (1 | Habitat) + (1 | Subarea)
 Data: demo
AIC  BIC  logLik deviance df.resid
 155.8 159.8 -73.9 147.8  16
Scaledresiduals:
Min   1Q   Median  3Q  Max
-2.36661 -0.63351 0.09057 0.66928 1.73795
Random effects:
Groups Name  Variance Std.Dev.
Subarea (Intercept) 0.000e  0.000e
Habitat (Intercept) 1.246e-12 2.417e-09
Number of obs: 20, groups: Subarea, 10; Habitat, 2
Fixed effects:
Estimate  Std. Error z value Pr(>|z|)
(Intercept) 4.845016  0.135709 35.702 < 2e-16 ***
Nvid   0.13142  0.03737  2.483 0.003729 **
Pokmoh  -0.006634 0.001849 -3.588 0.000333 ***
Vlag   -0.034544 0.037667 -2.456 0.045479 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Результаты демонстрируют значимую зависимость числа особей жужелиц ($p < 0,05$) от числа видов растений в местообитаниях, проективного покрытия мохового яруса и влажности. Экологическая интерпретация зависимостей заключается в том, что с увеличением числа видов растений расширяется спектр фитофагов с ними связанных, что способствует расширению кормовой базы хищных жужелиц. С покрытием мохового яруса выявлена отрицательная зависимость. Можно предположить, что плотный моховый покров затрудняет перемещение жужелиц, которые передвигаются по поверхности почвы, так как большинство видов не способны летать. Отрицательная зависимость от режима увлажнения также вполне очевидна.

Заключение

Методология оценки влияния факторов среды на представителей различных таксонов, входящих в состав биотических сообществ, достаточно специфична, так как многие биоценотические показатели не соответствуют закону нормального распределения. Трансформация данных для их подгонки к нормальному распределению не всегда дает хорошие результаты. С другой стороны, во многих случаях исследователи ограничиваются в свой работе возможностями *Microsoft Office Excel* и *Statistica,* что крайне сужает спектр аналитических возможностей. В наши дни статистическая среда R является общепризнанным стандартом в научных публикациях в мире и мощным инструментом анализа данных. Однако у биологов возникает проблема с написанием программного кода. Несмотря на то что в последнее время появилось много литературы на русском языке по R, отдельные вопросы требуют более детального рассмотрения. В частности, методика построения обобщенных линейных смешанных моделей (GLMM), которые нашли широкое применение в экологических исследованиях. Предложенный программный код может стать полезным широкому кругу исследователей для анализа данных в области экологии сообществ, включая выявление зависимостей биоценотических показателей от биотических и биотических факторов среды, и формирования обоснованных выводов.

Библиографические ссылки

1. Шитиков ВК, Мастицкий СЭ. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. [Place unknown]: Электронная книга; 2017. 351 с.
2. Legendre P, Legendre L. Numerical Ecology. 3rd edition. Amsterdam: Elsevier Science. BV; 2012. 990 p.
3. Кабаков РИ. *R в действии. Анализ и визуализация данных в программе R*. Москва: ДМК Пресс; 2014. 580 с.
4. Borcard D, Gillet F, Legendre P. Numerical Ecology with R. Wien: Springer Nature; 2011. 319 p.
5. McCune B, Grace JB. Analysis of ecological communities. Gleneden Beach: MjMSoftware Design; 2002. 304 p.
6. Мастицкий СЭ, Шитиков ВК. *Статистический анализ и визуализация данных с помощью R*. Москва: ДМК Пресс; 2015. 496 с.
7. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. New York: Springer; 2009. 745 p.
8. Zuur AF, Ieno EN, Walker N, Saveliev, AA, Smith GM. Mixed Effects Models and Extensions in Ecology With R. New York: Springer; 2009. 574 p.
9. Zuur AF, Ieno EN, Smith GM. Analyzing Ecological Data. New York: Springer; 2007. 672 p.
10. Faraway JJ. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Volume 124. Chapman & Hall/CRC Taylor & Francis Group; 2017. 331 p.
11. Четвериков АА, Линейные модели со смешанными эффектами в когнитивных исследованиях. *Российский журнал когнитивной науки*. 2015;2(1):41–51.
12. R Development Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing. [Internet, cited 2024 December 12]. Vienna: Austria. Available from: https://www.R-project.org.
13. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. 2015;67:1–48.
14. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. *Methods of Ecology and Evolution*. 2010;1:3–14.

References

1. Shitikov VK, Mastitskiy SE. *Klassifikatsiya, regressiya i drugiye algoritmy Data Mining s ispol'zovaniyem R* [Classification, Regression and Other Data Mining Algorithms Using R]. [Place unknown]: Elektronnaya kniga; 2017. 351 p. Russian.
2. Legendre P, Legendre L. Numerical Ecology. 3rd edition. Amsterdam: Elsevier Science. BV; 2012. 990 p.
3. Kabacov RI. R in Action. Data analysis and graphics with R. Shelter Island, New York: Manning Publications Co.; 2012. 580 p.
4. Borcard D, Gillet F, Legendre P. Numerical Ecology with R. Wien: Springer Nature; 2011. 319 p.
5. McCune B, Grace JB. Analysis of ecological communities. Gleneden Beach: MjMSoftware Design; 2002. 304 p.
6. Mastitskiy SE, Shitikov VK. *Statisticheskiy analiz i vizualizatsiya dannykh s pomoshch'yu R* [Statistical analysis and data visualization using R]. Moscow: DMK Press; 2015. 496 p. Russian.
7. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. New York: Springer; 2009. 745 p.
8. Zuur AF, Ieno EN, Walker N, Saveliev, AA, Smith GM. Mixed Effects Models and Extensions in Ecology With R. New York: Springer; 2009. 574 p.
9. Zuur AF, Ieno EN, Smith GM. Analyzing Ecological Data. New York: Springer; 2007. 672 p.
10. Faraway JJ. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Volume 124. Chapman & Hall/CRC Taylor & Francis Group; 2017. 331 p.
11. Chetverikov AA, *Lineynyye modeli so smeshannymi effektami v kognitivnykh issledovaniyakh* [Linear models with mixed effects in cognitive research]. *Rossiyskiy zhurnal kognitivnoy nauki*. 2015;2(1):41–51. Russian.
12. R Development Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing. [Internet, cited 2024 December 12]. Vienna: Austria. Available from: https://www.R-project.org.
13. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. 2015;67:1–48.
14. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. *Methods of Ecology and Evolution*. 2010;1:3–14.

Статья поступила в редколлегию 29.01.2025.
Received by editorial board 29.01.2025.