УДК 595.7(476)

ИСПОЛЬЗОВАНИЕ МЕТОДОВ АНАЛИЗА ДАННЫХ В ЭНТОМОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

СУШКО Г. Г.
Витебский государственный университет им. П.М. Машерова, г. Витебск, Беларусь
e–mail: gennadis@rambler.ru

В работе приведен обзор методов анализа данных, наиболее часто используемых в исследованиях, связанных с насекомыми, за последние пять лет. Выявлены основные методы ординации и регрессионного анализа, проанализирована частота применения соответствующих пакетов анализа данных.

Введение. С компьютерной техникой информационные технологии все глубже проникают в различные сферы нашей жизни, особенно в научную деятельность. Не обошли они стороной и энтомологию. В настоящее время уже трудно провести грань между отраслями знаний. Так, насекомые являются подходящими объектами исследований в самых различных областях, например, при оценке состояния окружающей среды, биоиндикации, молекулярной генетики, теории эволюции, климатологии и т.д. С другой стороны, в большинстве энтомологических исследований присутствуют элементы экологии, зоогеографии, геоботаники и др. В этом не трудно убедиться, проанализировав наиболее популярную форму презентации энтомологических данных – аннотированный список видов, который является неотъемлемой частью диссертационных исследований и научных отчетов, а также основой самостоятельных публикаций. И с полной уверенностью можно утверждать, что ни одно современное энтомологическое исследование не обходиться без использования компьютерных технологий.

Появление мощных компьютеров позволило обрабатывать огромные массивы данных и использовать методы многофакторного анализа. Следует отметить и то, что перед исследователем возросла ответственность при разработке дизайна исследований, так как проведение различных типов анализа данных и достоверность результатов требуют определенного (зачастую большого) числа повторностей эксперимента и объема выборок.

Существенной проблемой является ограниченность русскоязычной литературы по современным методам анализа данных, которые появляются, уточняются и дополняются чуть ли не каждый год. Еще одна проблема – отсутствие многих методик анализа данных и методик работы с современным программным обеспечением в цикле биологических дисциплин, читаемых в настоящее время для студентов ВУЗов. Исключение составляет учебный план магистратуры, где уже предложены дисциплины, направленные восполнить данный пробел. Вследствие этого, а также из-за языкового барьера при чтении англоязычной литературы, может возникнуть непонимание методик анализа данных. В данной работе предпринята попытка дать краткий обзор методов анализа данных, используемых в современных исследованиях в области экологии насекомых.

Материалы и методы. Материалом для работы послужил анализ литературных источников международной базы данных PubMed Central (PMC). Эта база данных является бесплатным полнотекстовым архивом журнальных статей по биологии, медицине и наукам о живой природе в Национальной медицинской библиотеке Национального института здравоохранения США (NIH / NLM). Для обзора были выбраны такие показатели как использование регрессионного анализа, методов многомерного анализа (ординация, кластерный анализ, мультивариантные тесты). Также выполнен анализ статистических пакетов, применяемых для этих целей. Всего проанализировано 350 работ по синэкологии, в которых модельными объектами были группировки (ассамблеи) наземных насекомых различных таксонов, отобранных по соответствующим ключевым словам в названии и

аннотации. Работы по экологии отдельных видов насекомых не рассматривались, так как многие методы многофакторного анализа здесь не уместны.

Результаты. Среди применяемых в настоящее время в экологических исследованиях методов оценки влияния факторов среды на живые организмы, в том числе и на насекомых, наибольшее распространение получил регрессионный анализ (McCune & Grace 2002, Borcard et al. 2018). В 47,9% проанализированных литературных источников использованы модели множественной регрессии (среди них: 70,7% – GLM, 24,1% – GLMM, 5,2% – GAM). Следует обратить внимание, что в их числе только обобщенные модели, применяемые при несоответствии данных закону нормального распределения. Как видно, данные видового богатства, учетной плотности, числа зарегистрированных особей, разнообразия, как правило, соответствуют распределению Пуассона или отрицательному биноминальному. Кроме того, обобщенные модели множественной регрессии могут справиться и с другими ситуациями, такими как гетерогенность дисперсии, большое число нулевых наблюдений, наличие выбросов, отсутствие нормальности распределения остатков и др. Данные методы уместны как для количественных, так и для категориальных переменных (Borcard et al. 2018).

Широкое применение имеют различные методы ординации (содержат 65,6%, проанализированных публикаций).

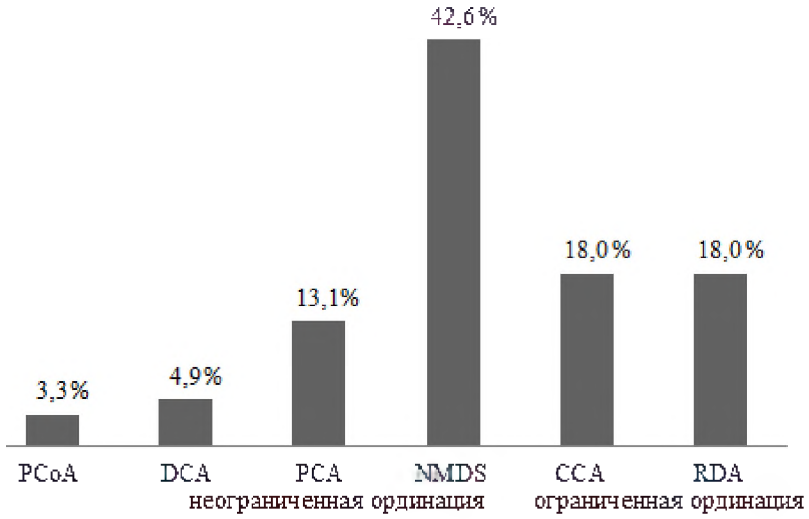

Рисунок 1 – Использование методов ординации в энтомологических исследованиях

Методы ординации делят на неограниченные и ограниченные. Первая группа методов не ограничена конкретными переменными и показывает изменение видового состава вдоль некоторой абстрактной оси (градиента), которая отражает максимальную изменчивость в структуре данных, тогда как прямая ординация отображает изменение видового состава вдоль некоторого выбранного исследователем физического градиента среды (уровень влажности, изменения рельефа, временная динамика, уровень загрязнения и т.д.) (Ramette 2007). Наиболее распространенными оказались методы непрямой ординации и, в частности, неметрическое многомерное шкалирование – NMDS (42,6%). Главным преимуществом NMDS является то, что от исходных данных не требуется никаких априорных предположений о характере распределения, и, в частности, соответствия закону нормального распределения (McCune & Grace 2002).

Следует отметить, что такие популярные в русскоязычных публикациях методы, как кластерный анализ (5,02%) и дискриминантный анализ (1%) не имеют столь широкого распространения в исследованных публикациях.

Проверку того, значимо ли отличаются группы объектов (например, данные обилия видов в нескольких местообитаниях) можно осуществить и с помощью тестов, основанных на статистических гипотезах, сопоставляющих различия внутри групп с различиями между группами. Такие методики выявлены в 29,2% публикаций. Чаще всего применяют

непараметрические тесты, такие как многофакторный дисперсионный анализ PERMANOVA (53,5%) и анализ сходства ANOSIM (28,5%). Отмечены также тест Мантеля (Mantel test) – 10,7% и SIMPER (Similarity Percentage) – 7,1%.

Для выполнения различных методов анализа данных были использованы как стандартные пакеты статистического анализа (STATISTICA, SPSS, PAST), так и специализированные, например, разработанные для выполнения ординации данных (PC-Ord, CANOCO). Однако, чаще всего, анализ выполнялся в R (62,5%). Статистическая среда R в настоящее время предоставляет наиболее широкие возможности для обработки данных. В частности, пакеты vegan, labdsv, BiodiversityR и многие другие позволяют анализировать биологические данные. Другие, в том числе и наиболее популярные до недавнего времени пакеты анализа, применяются на много реже: PRIMER (12,5%), CANOCO (9,3%), SPSS (7,3%), STATISTICA и PAST (по 6,2%), SAS (3,1%), PC-ORD (2%).

Список использованных источников:

Borcard D., Gillet F., Legendre P. 2018. Numerical Ecology with R. 435 p.

McCune B., Grace J.B. 2002. Analysis of ecological communities. 304 p.

PubMed Central (PMC) [Electronic resource]. – Mode of access : https://www.ncbi.nlm.nih.gov/pmc/ – Date of access : 10.08.2019.

Ramette A. 2007. Multivariate analyses in microbial ecology // Microbiol. Ecol. Vol. 62. № 2. P. 142–160.

Use of data analysis methods in entomological research

G.G. Sushko

The paper provides an overview of the data analysis methods most commonly used in insect-related studies over the past five years. The main methods of ordination and regression analysis are identified; the frequency of application of the corresponding data analysis packages is analyzed.