Vestnik zoologii, **40**(1): 63—69, 2006

Методика

УДК 57.087.1:599.323.4

ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ НЕКОТОРЫХ АЛГОРИТМОВ ДИСКРИМИНАНТНОГО АНАЛИЗА В ИДЕНТИФИКАЦИИ БЛИЗКИХ ВИДОВ НА ПРИМЕРЕ ЛЕСНЫХ МЫШЕЙ *SYLVAEMUS* (RODENTIA, MURIDAE)

И. И. Дзеверин, Е. И. Лашкова

Институт зоологии им. И. И. Шмальгаузена НАН Украины, ул. Б. Хмельницкого, 15, Киев, 01601 Украина

Принято 16 ноября 2004

Возможности и ограничения некоторых алгоритмов дискриминантного анализа в идентификации близких видов на примере лесных мышей *Sylvaemus* (Rodentia, Muridae). Дзеверин И. И., Лашкова Е. И. — На примере эмпирических данных об изменчивости лесных мышей и результатов численного моделирования обсуждаются проблемы использования дискриминантного анализа в построении алгоритмов идентификации близких видов по морфометрическим признакам. Как правило, идентифицируемые группы характеризуются разной степенью сходства. Однако поэтапное проведение дискриминантного анализа с целью идентификации сначала наиболее своеобразных видов, а потом — всех остальных, обычно не содействует улучшению точности идентификации. В большинстве случаев дискриминантный анализ целесообразно проводить по объединенным данным, не деля выборку на подгруппы.

Ключевые слова: дискриминантный анализ, морфометрия, идентификация, *Sylvaemus.*

Opportunities and Restrictions for Some Algorithms of the Discriminant Function Analysis in Identification of Closely Related Species: a Case of Wood Mice *Sylvaemus* (Rodentia, Muridae). Dzeverin I. I., Lashkova E. I. — Empirical data on wood mice variation and results of numerical modeling were used to discuss the problems of applying the discriminant function analysis in working out the algorithms of identification of closely related species from morphometric characters. As usual, the groups being identified differ from one another at different extent. However, step-by-step identification (when firstly the most peculiar groups should be determined and then — the others) as a rule does not increase the correctness of identification. In the most cases it is better to apply discriminant function analysis to pooled sample without dividing it into subgroups.

Key words: discriminant function analysis, morphometrics, identification, *Sylvaemus.*

Введение

В настоящей статье рассматриваются некоторые проблемы применения дискриминантного анализа к биологическим данным. В качестве модельной взята хорошо изученная ранее (Лашкова, 2003; Межжерин и др., 2005) выборка лесных мышей с территории Украины. В выборке представлены все четыре вида лесных мышей фауны Украины (*Sylvaemus uralensis, S. arianus, S. sylvaticus и S. tauricus*). Мы остановимся на двух проблемах: критериях выбора более экономной модели и возможности проведения дискриминантного анализа в два этапа.

Лесные мыши *Sylvaemus* (Rodentia, Muridae) интересны как модельный объект для апробации различных математических методов анализа морфологической изменчивости. Данная группа включает в себя (по разным оценкам) от 7 до 9 морфологически очень похожих видов. Надежную диагностику этих видов обеспечивают генетические методы. Определение видовой принадлежности по морфологическим признакам также возможно, но только при учете большого числа признаков одновременно. Поэтому для получения наглядной картины различий лесных мышей в размерах и форме черепа, а также оценки степени морфологической дивергенции видов целесообразно использовать методы многомерной статистики, прежде всего дискриминантный анализ (Загороднюк, Федорченко, 1993; Лавренченко, Лихнова, 1995; Лашкова, Дзеверин, 2002; Лашкова и др., 2005; Reutter et al., 1999; Van Der Straeten, Van Der Straeten-Harrie, 1977).

Задача дискриминантного анализа – отнесение произвольного объекта к одной из нескольких априорно заданных совокупностей, например, того или иного организма – к определенному виду. Для выполнения этой процедуры возможны самые разнообразные алгоритмы, но именно дискриминантный анализ позволяет найти оптимальные критерии идентификации (Дерябин, 1983). Именно поэтому данный метод является одной из наиболее употребительных в биологии статистических процедур. Как показал подсчет публикаций в ведущих биологических журналах за 80-е годы прошлого века, дискриминантный анализ оказался по частоте использования среди статистических методов на втором месте, уступив только анализу главных компонент (James, McCulloch, 1990). В последующие годы ситуация принципиально не изменилась. Дискриминантный анализ по-прежнему остается одним из основных методов в решении задач классификации (построение диагностических алгоритмов, ключей и т. п.) и ряда иных проблем биологии.

Обсуждаемые в статье алгоритмы основаны на использовании линейного дискриминантного анализа. Теория и методика применения этого раздела статистики изложены во многих сводках и пособиях, как классических (Андерсон, 1963: гл. 6, 12; Кендалл, Стьюарт, 1976: гл. 44), так и более современных (Айвазян и др., 1989: разд. 1; Афифи, Эйзен, 1982; Клекка, 1989; Справочник..., 1990: гл. 16). Основные принципы дискриминантного анализа детально рассмотрены в работах разного уровня сложности, предназначенных специально для биологов (Jolicoeur, 1959: 284–287, 298–299; Campbell, Atchley, 1981; Джефферс, 1981; Дерябин, 1983; Компьютерная..., 1990). Важные аспекты и проблемы применения дискриминантного анализа в биологии, особенно в экологических исследованиях, обсуждаются в работе Ф. Джеймса и Ч. Мак-Каллоха (James, McCulloch, 1990).

Все вычисления в нашей статье произведены стандартными методами с помощью компьютерной системы анализа данных «STATISTICA», версия 6 (StatSoft, Inc., 2001, США). Для каждой из моделей было определено значение статистики Уилкса (λ). Уровни значимости моделей (p) были оценены путем аппроксимации этой величины статистикой Фишера (F) со степенями свободы df1 и df2. На основе этих моделей были построены классификационные и канонические функции. Шаговые алгоритмы выбора признаков в настоящей работе не были использованы: возможности и ограничения их применения – это отдельная, весьма актуальная проблема, которой мы здесь не касаемся. Модельные выборки для иллюстрации возможностей поэтапного проведения дискриминантного анализа были сформированы с помощью системы технических вычислений MATLAB, версия 6 (MathWorks, Inc., 2001, США).

Наши рекомендации основаны на опыте применения дискриминантного анализа к конкретным наборам данных. Мы не претендуем на аналитическое решение поставленных проблем.

Канонические переменные и классификационные функции

Современный дискриминантный анализ представляет собой весьма детально разработанную систему алгоритмов выявления межгрупповых различий. Классические алгоритмы линейного дискриминантного анализа основаны на использовании одних и тех же исходных данных и содержат одну и ту же информацию. Однако эта информация представлена в разном виде, и на практике в разных ситуациях предпочтительными оказываются разные варианты анализа.

Наиболее экономное описание межгрупповых различий дают канонические переменные. Так, если исследуются различия между n группами, то вся информация об этих различиях, доступная линейным методам анализа, может быть представлена в виде n–1 линейных функций. Эти функции можно рассматривать как своего рода признаки, анализировать их, давать им функциональную интерпретацию и т. п. В этом отношении канонические переменные вполне аналогичны главным компонентам. При этом в отличие от главных компонент, канонические переменные представляют исходные данные таким образом, что минимизируют внутригрупповые различия и максимизируют межгрупповые, будучи таким образом весьма полезными для содержательной интерпретации тех и других различий.

Так, именно использование дискриминантного анализа в классическом исследовании П. Жоликёра и его соавторов позволило выявить и количественно охарактеризовать особенности строения и функционирования мозга в различных отрядах млекопитающих (Jolicoeur et al., 1984). Дискриминантный анализ может быть использован для описания географической изменчивости (например, в работе Jolicoeur, 1959). Яркий и имеющий отнюдь не только иллюстративное

значение пример исследования внутри- и межгрупповой изменчивости в человеческих популяциях путем применения канонического анализа к антропометрическим данным приведен В. Е. Дерябиным (1983: 146—149). В целом можно констатировать, что канонический анализ относится к числу весьма надежных инструментов в изучении закономерностей изменчивости.

В то же время канонические переменные часто неудобны, если нужно определить групповую принадлежность произвольного объекта. В общем случае для этого нужно определить характеризующие данный объект значения канонических переменных (n–1 линейных функций по признакам объекта), а потом вычислить расстояния до n групповых центроидов в n–1-мерном пространстве, да еще и с поправкой на неравную численность групп (вычисление еще n функций; объект считается относящимся к той группе, расстояние до центроида которой оказалось наименьшим). Таким образом, для определения групповой принадлежности одного объекта нужно проделать 2n–1 весьма сложных расчетов. Определение групповой принадлежности по каноническим переменным оказывается поэтому весьма громоздкой процедурой. Не случайно данный алгоритм не реализован программно во многих употребительных компьютерных статистических пакетах, в том числе в STATISTICA.

На практике, конечно, можно пренебречь частью канонических переменных и использовать только первые 2—3 из них. Это позволяет несколько упростить вычисления. И все же наиболее удобной для диагностики объекта представляется модификация метода, позволяющая построить на основании исходных данных набор из n классификационных функций, каждая из которых представляет собой оценку принадлежности объекта к определенной группе. Предполагается, что объект относится к той группе, значение классификационной функции которой оказалось наибольшим. Таким образом, определение групповой принадлежности по классификационным функциям сводится к вычислению только n значений, что несомненно более экономно, если принять во внимание следующее: а) итоговый результат применения обеих методик одинаков; б) вычисление одной классификационной функции требует тех же затрат, что и вычисление канонической переменной; в) выбор максимального значения из n вариантов не представляет никакой сложности; г) (2n–1) превосходит n, если $n > 1$.

Итак, при определении групповой принадлежности изучаемых объектов предпочтительнее работать с классификационными функциями, а не с каноническими переменными. Важным исключением является, однако, ситуация, когда $n = 2$ (например, при разграничении двух видов или при изучении различия между самцами и самками). В этом случае канонический анализ сведется к элементарному алгоритму дискриминантного анализа. Для определения групповой принадлежности произвольного объекта достаточно знать значение межгрупповой границы и вычислить значение дискриминантной функции. Если оно окажется больше значения межгрупповой границы, то объект следует отнести к одной группе, если меньше — то к другой. В этом случае использование классификационных функций себя не оправдывает.

Поэтапное проведение дискриминантного анализа

Весьма часто некоторые из групп, которые являются объектами дискриминантного анализа, более сходны между собой, чем с другими группами. Так, по большинству признаков желтогорлая мышь отличается от лесной, малой и степной заметно больше, чем эти три вида друг от друга. Эта давно известная особенность отмечена и у лесных мышей с территории Украины (Лашкова, 2003; Лашкова и др., 2005; Межжерин, 1993). Разная степень межгруппового сходства позволяет строить иерархические классификации (например, методами клас-

терного анализа). Отдельная проблема состоит, однако, в том, нужен ли учет этой иерархии в алгоритмах дискриминантного анализа. Применительно к лесным мышам можно предположить, что целесообразно проводить диагностику особей рода, основанную на применении дискриминантного анализа, в два этапа. На первом этапе определяется принадлежность особи к одной из двух групп: к крупным мышам (*S. tauricus*) или к мелким (*S. sylvaticus*, *S. arianus* или *S. uralensis*). Если особь отнесена к группе мелких лесных мышей, то на втором этапе определяется, к какому конкретно из трех видов она относится.

Первый этап может быть осуществлен с помощью одной дискриминантной функции, а второй – с помощью набора из трех классификационных функций (см. выше). Таким образом, для определения видовой принадлежности произвольных особей понадобится вычислить значения в среднем примерно 3,25 функций в пересчете на особь, что даже несколько меньше, чем нужно при вычислении принадлежности особи к одному из четырех видов в один этап.

Тем не менее опыт применения этой схемы диагностики к конкретным данным по изменчивости лесных мышей показывает, что она почти никогда не приводит к лучшим результатам, чем традиционная схема. Как правило, результаты определения на первом этапе очень надежные (для некоторых наборов признаков – даже 100%-ные), зато качество разграничения трех видов на втором этапе остается прежним или становится даже худшим, чем при работе с четырьмя видами. Лишь для отдельных наборов признаков качество диагностики незначительно улучшается. Общий выигрыш от применения двухэтапной методики, если он вообще наблюдается, не превышает нескольких процентов. Возможные причины такой ситуации обсуждаются далее.

Сказанное можно проиллюстрировать конкретным примером: исходная модель дана в работе Е. И. Лашковой с соавт. (2005). В этой работе описано определение видовой принадлежности взрослых лесных мышей по трем экстерьерным признакам (длины хвоста, ступни и уха). Полученная модель статистически значима ($\lambda = 0{,}079$, $F = 155{,}20$, $df1 = 9$, $df2 = 757$, $p < 10^{-4}$) и позволяет правильно определять видовую принадлежность 93,7% особей. Лучше всего диагностируются желтогорлые мыши (97,6%), хуже всего – степные мыши (71,4%).

Если разработать двухэтапный алгоритм определения видовой принадлежности лесных мышей по тем же признакам, то дискриминантная функция для первого этапа ($\lambda = 0{,}18$, $F = 466{,}66$, $df1 = 3$, $df2 = 315$, $p < 10^{-4}$) позволит правильно определять 98,1% особей, в том числе 97,4% мелких мышей и 98,8% – желтогорлых. Набор классификационных функций для второго этапа ($\lambda = 0{,}16$; $F = 72{,}20$; $df1 = 6$; $df2 = 290$; $p < 10^{-4}$) позволяет правильно определять 89,3% мышей, в том числе 94,5% *S. uralensis*, 64,3% *S. arianus*, 95,9% *S. sylvaticus*. Итоговая (по двум моделям) доля правильно определенных особей – 92,9%. Таким образом, результаты применения двухэтапной методики оказались незначительно худшими, чем при работе с четырьмя группами. Это касается как всей совокупности, так и в первую очередь особей *S. arianus*, идентифицировать которые (вследствие сходства в размерах как с *S. uralensis*, так и с *S. sylvaticus*) особенно сложно. Двухэтапная диагностика именно этого вида оказалась наименее удачной сравнительно с одноэтапной методикой.

Качественно сходные результаты дает применение двухэтапной методики и в других случаях. В большинстве случаев разделение выборки сразу на четыре видовые группы дает лучший результат. Объяснить эту особенность изучаемой группы видов можно с помощью схемы (рис. 1). На этой схеме представлены результаты дискриминантного анализа отличий между четырьмя условными видами, похожими на виды лесных мышей. Обозначены как отдельные особи, так и центроиды четырех видовых групп (I–IV), а также центроид условной

группы, объединяющей первые три вида (V). Как и в реальных выборках лесных мышей, вид 4 больше отличается от видов 1, 2 и 3, чем эти последние друг от друга. Предполагается также, что вид 3 занимает промежуточное положение между 4, с одной стороны, и 1 и 2 — с другой. Для простоты априорные вероятности принадлежности особей к группам предполагаются равными. Решающее правило дискриминантного анализа состоит в том, что произвольная особь может быть отнесена к той группе, к центроиду которой она расположена ближе всего в пространстве итоговых переменных. Мы видим, что для большинства особей это правило выполняется. Кроме того, в большинстве случаев нет разницы, учитывается расстояние особи до центроида вида I, II или III, или же до объединенного центроида V. Если особь первого вида ближе к центроиду I, чем к центроиду IV, то она ближе и к центроиду V, чем к центроиду IV.

Тем не менее мы видим ряд небезынтересных исключений. В отдельных случаях расстояние особи, принадлежащей к виду 3, до центроида IV больше, чем расстояние до центроида III, но меньше, чем до центроида V. Это, как правило, аберрантные особи, резко отличающиеся от типичных представителей изучаемых групп (например, особь A на схеме), или, наоборот, промежуточные по своим признакам особи (например, особь B). Двухэтапная методика определит видовую принадлежность таких особей ошибочно. Противоположная ситуация (расстояние особи одного из трех близких видов до центроида IV меньше, чем до центроида собственного вида, но больше, чем до центроида V; в

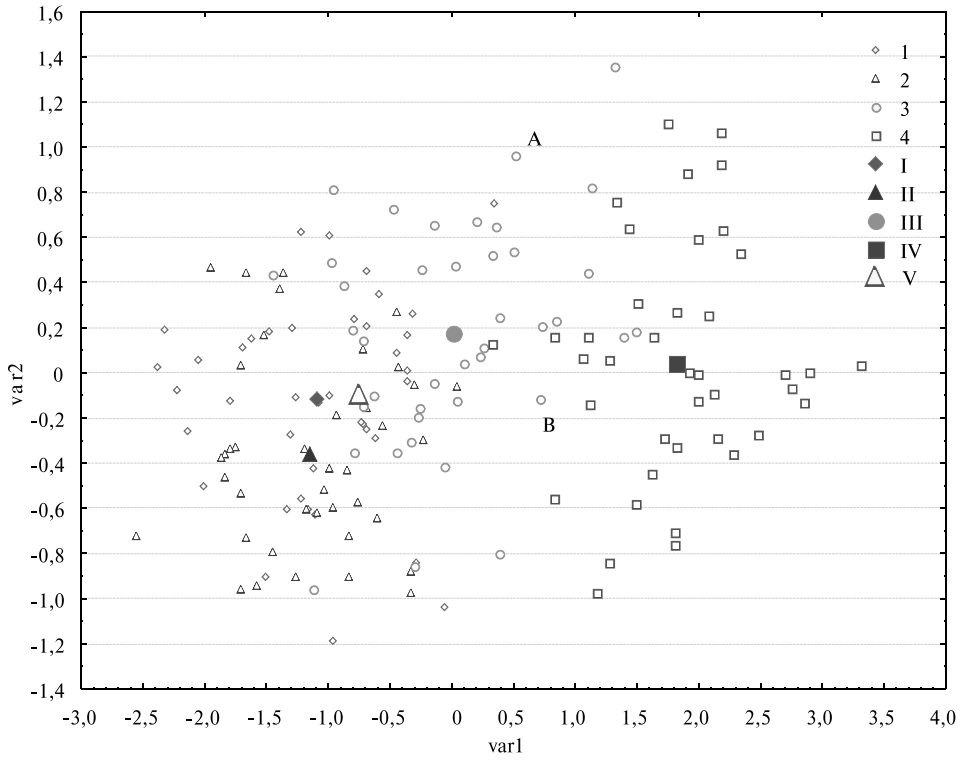

Рис. 1. Применимость двухэтапной методики дискриминантного анализа к совокупности из четырех слабо различающихся групп; 1—4 — особи условных видов: I—IV — центроиды выборок каждого из этих видов; V — центроид группы наиболее сходных видов (1—3); A, B — неправильно классифицируемые особи.

Fig. 1. The applicability of two-step technique in discriminant function analysis to the sample containing the specimens from four poorly differing species: 1—4 — specimens of modelled species; I—IV — sample centroids of each of these species; V — centroid of group of the most similar species (1—3); A, B — incorrectly classified specimens.

идентификации такой особи ошибку допустит уже одноэтапная методика) представляется неправдоподобной. Поэтому за счет части немногочисленных аберрантных и промежуточных особей двухэтапная методика оказывается в общем случае менее точной, чем одноэтапная.

Какую долю общей выборки составят ошибочные определения и как они распределятся между видами, — это, конечно, зависит от конкретной выборки и конкретного набора признаков. В большинстве случаев точность определения двухэтапным методом ниже на несколько процентов. В отдельных выборках, о чем уже шла речь выше, применение двухэтапной методики дало тот же или непринципиально лучший результат, чем одноэтапной.

Определенно рекомендовать двухэтапный метод можно, по-видимому, в тех случаях, когда вид 4 отличается от 1, 2 и 3 настолько значительно, что ситуация, описанная выше, в принципе невозможна (рис. 2). В этом случае определение даже аберрантных особей не вызывает никаких проблем, однако в подобных случаях, скорее всего, определение легко осуществимо даже без дискриминантного анализа. Тем не менее, если по тем или иным причинам для тех признаков, для которых выполняется ситуация, проиллюстрированная на рисунке 2, нужно прибегнуть к дискриминантному анализу, то имеет смысл протестировать двухэтапные модели.

Применительно к лесным мышам логически возможна еще одна двухэтапная схема проведения дискриминантного анализа. На первом этапе можно разделить выборку на 3 группы, первая из которых включает в себя особей вида

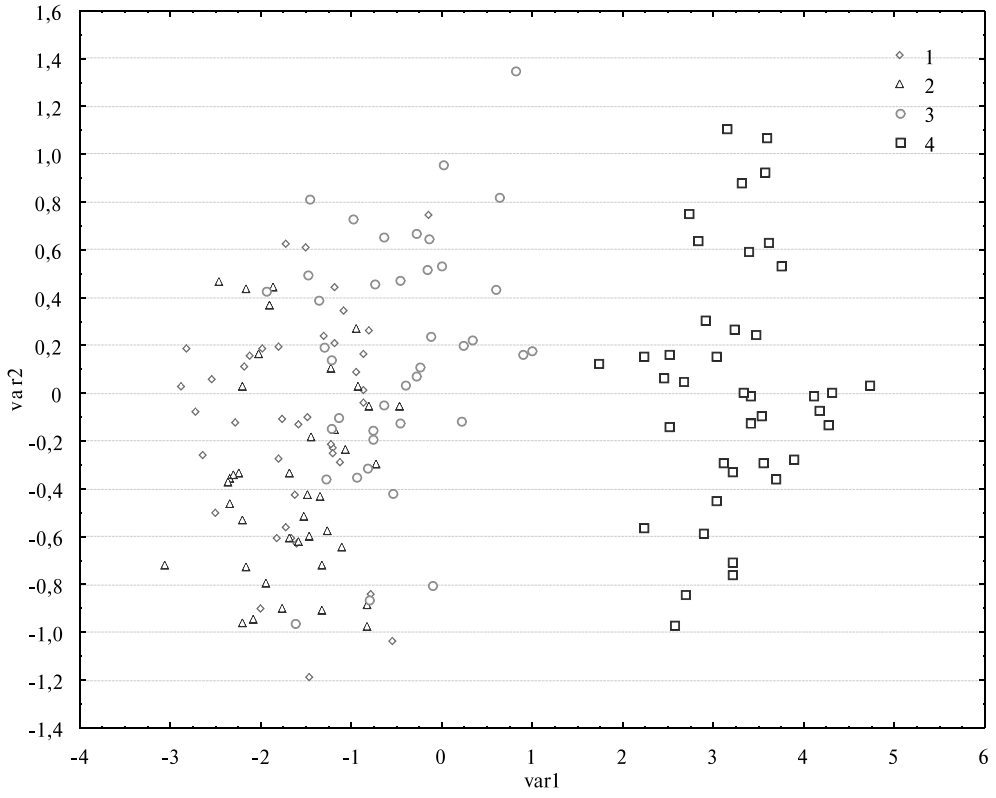

Рис. 2. Применимость двухэтапной методики дискриминантного анализа к совокупности из четырех групп, одна из которых весьма значительно отличается от остальных: 1—4 — особи условных видов.

Fig. 2. The applicability of two-step technique in discriminant function analysis to the sample containing the specimens from four species, one of them extremely differs from the others: 1—4 — specimens of modelled species.

S. uralensis, вторая – *S. tauricus*, а третья – особей двух видов, *S. arianus* и *S. sylvaticus*. Это объединение можно обосновать, во-первых, большим морфологическим сходством двух последних видов, во-вторых, тем, что они парапатричны. При необходимости (например, в спорных случаях) на втором этапе можно идентифицировать и особей этих видов. Однако такой вариант двухэтапной методики неудачен по тем же соображениям, что и вариант, описанный выше. Использование такой методики при анализе конкретных данных приводило к заметному ухудшению качества моделей практически во всех случаях.

Таким образом, при построении математических моделей диагностики морфологически сходных видов по количественным признакам в большинстве случаев целесообразно сразу делить изучаемую совокупность на группы, соответствующие априорно известным видам.

Авторы благодарны С. В. Межжерину за идею настоящей работы и обсуждение ее результатов, В. Н. Пескову и Е. И. Кожуриной – за ценные рекомендации.

Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности : Справочное издание. – М. : Финансы и статистика, 1989. – 608 с.

Андерсон Т. Введение в многомерный статистический анализ : Пер. с англ. – М. : Физматгиз, 1963. – 500 с.

Афифи А., Эйзен С. Статистический анализ: подход с использованием ЭВМ : Пер. с англ. – М. : Мир, 1982. – 488 с.

Дерябин В. Е. Многомерная биометрия для антропологов. – М. : Изд-во Моск. ун-та, 1983. – 227 с.

Джефферс Дж. Введение в системный анализ: применение в экологии : Пер. с англ. – М. : Мир, 1981. – 253 с.

Загороднюк И. В., Федорченко А. А. Мыши рода Sylvaemus Нижнего Дуная. Сообщ. 1. Таксономия и диагностика // Вестн. зоологии. – 1993. – № 3. – С. 41–49.

Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды : Пер. с англ. – М. : Наука, 1976. – 736 с.

Клекка У. Р. Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ : Пер. с англ. – М. : Финансы и статистика, 1989. – С. 78–138.

Компьютерная биометрика. – М. : Изд-во Моск. ун-та, 1990. – 232 с.

Лавренченко Л. А., Лихнова О. П. Аллозимная и морфологическая изменчивость трех видов лесных мышей (Rodentia, Muridae, Apodemus) Дагестана в условиях симбиотопии // Зоол. журн. – 1995. – **74**, № 5. – С. 107–119.

Лашкова Е. И. Морфометрическая изменчивость лесных мышей, Sylvaemus (Muridae), фауны Украины // Вестн. зоологии. – 2003. – **37**, № 3. – С. 31–41.

Лашкова Е. И., Дзеверин И. И. Одонтометрическая изменчивость и идентификация видов лесных мышей, Sylvaemus (Muridae, Rodentia), фауны Украины // Вестн. зоологии. – 2002. – **36**, № 3. – С. 25–33.

Лашкова Е. И., Межжерин С. В. Идентификация видов лесных мышей фауны Украины по экстерьерным и черепным признакам методами многомерного анализа // Вестн. зоологии. – 2005. – **39**, № 3. – С. 23–28.

Межжерин С. В. Лесные мыши рода Sylvaemus Ognev et Vorobiev, 1924 фауны Украины // Млекопитающие Украины. – Киев : Наук. думка, 1993. – С. 55–63.

Справочник по прикладной статистике : Пер. с англ. – М. : Финансы и статистика, 1990. – Т. 2. – 528 с.

Campbell N. A., Atchley W. R. The geometry of canonical variate analysis // Syst. Zool. – 1981. – **30**, N 3. – P 268–280.

James F. C., McCulloch C. E. Multivariate analysis in ecology and systematics: panacea or Pandora's box? // Ann. Rev. Ecol. Syst. – 1990. – **21**. – P 129–166.

Jolicoeur P. Multivariate geographical variation in the wolf Canis lupus L. // Evolution. – 1959. – **13**, N 3. – P. 283–299.

Jolicoeur P., Pirlot P., Baron G., Stephan H. Brain structure and correlation patterns in Insectivora, Chiroptera, and Primates // Syst. Zool. – 1984. – **33**, N 1. – P. 14–29.

Reutter B. A., Hausser J., Vogel P. Discriminant analysis of skull morphometric characters in Apodemus sylvaticus, A. flavicollis, and A. alpicola (Mammalia; Rodentia) from the Alps // Acta Theriol. – 1999. – **44**, N 3. – P. 299–308.

Straeten E. Van Der, Straeten-Harrie B. Van Der Étude de la biometrie crânienne et de la repartition d'Apodemus sylvaticus (Linnaeus, 1758) et d'Apodemus flavicollis (Melchior, 1834) en Belgique // Acta Zoologica et Pathologica Antverpiensia. – 1977. – **69**. – P. 169–182.