



Роль статистики в изучении разнообразия коллекции генетических ресурсов растений ВИР

Л. Ю. Новикова

Федеральный исследовательский центр Всероссийский институт генетических ресурсов растений имени Н.И. Вавилова, Санкт-Петербург, Россия

Автор, ответственный за переписку: Любовь Юрьевна Новикова, l.novikova@vir.nw.ru

Биологическая статистика начала развиваться синхронно с формированием первых коллекций генетических ресурсов растений (ГРР) в конце XIX – начале XX в. С момента создания коллекции ГРР ВИР в полевых и лабораторных условиях с использованием всего доступного диапазона экспериментальных методов ведется оценка разнообразия генофонда по фенотипическим и генотипическим показателям, выделение образцов с заданным набором характеристик для конкретных задач селекции. Ключевым вопросом параметризации образцов является выбор адекватного задаче метода статистического исследования. Целью данного обзора является сводка типовых статистических задач, возникающих при работе с фенотипическими оценками коллекций ГРР, и соответствующих им методов.

При анализе коллекций ГРР базовым элементом является определение качественных и количественных характеристик образцов. Методы описательной статистики позволяют оценить средние значения и стабильность хозяйственно ценных признаков на основе нескольких лет изучения в одном или нескольких пунктах. Эти показатели сводятся в базы, публикуются в каталогах, постоянно издаваемых ВИР, используются при выделении перспективных образцов для задач селекции.

Часто встает задача сравнения групп образцов разных видов, разного географического происхождения, результатов испытания в разных эколого-географических условиях, жизнеспособности при разных условиях хранения и т. д. Проверка гипотезы о различии нескольких признаков или вариантов опыта по выборочным данным осуществляется с использованием t-критерия Стьюдента, дисперсионного анализа. Активно развиваются современные методы оценки взаимодействия «генотип × среда».

Выявление погодно-климатических факторов, влияющих на хозяйственно ценные признаки образцов, такие как урожайность, крупнозерность, устойчивость к биотическим и абиотическим факторам, показатели качества, установление взаимосвязей этих признаков, которые необходимо учитывать при создании сорта заданной модели, осуществляются методами корреляционно-регрессионного анализа. Начавшееся в 1970-х гг. активное потепление климата вызвало интерес биологов к определению трендов урожайности сортов, использованию методов анализа временных рядов в регрессионном моделировании.

При описании объемной коллекции ГРР, исследовании филогенетических связей, выделении репрезентативной выборки, создании core-коллекций возникает задача структурирования коллекции, выделения групп схожих образцов. Задача классификации решается методами классификации без обучения, в частности кластерного анализа, или с обучением, например дискриминантного анализа.

При анализе образцов ГРР по большому количеству показателей стоит задача выявления наиболее информативных показателей, то есть снижения размерности пространства признаков, которая решается методами анализа главных компонент или факторного анализа с переходом от реальных измеренных переменных к небольшому количеству гипотетических величин, называемых факторами.

Изучение коллекций ГРР все более выходит за рамки классических полевых наблюдений урожайности, для которых в начале XX в. разработали методы биостатистики, основанные на предположении о нормальном характере распределения. В середине XX в. появились непараметрические аналитические методы и ресемплинг-методы имитационного компьютерного моделирования для решения тех же предметных задач.

Развитие компьютеров позволило создать интегрированные дата-платформы, собирающие огромные массивы разноплановых характеристик ГРР, масштабировать сложные многомерные методы, доступные только математикам, на широкий круг исследователей в различных предметных областях. Методы биостатистики, относящиеся к фенотипическим данным, продолжают развиваться, обеспечивая совершенствование оценки потенциала коллекций ГРР и являясь неотъемлемой информационной поддержкой современного геномного анализа.

Ключевые слова: коллекции ГРР, фенотипические характеристики, базы данных, статистические методы

Благодарности: Работа выполнена в рамках реализации Программы развития Национального центра генетических ресурсов растений по соглашению с Минобрнауки России от 15 февраля 2024 года № 075-02-2024-1090.

Для цитирования: Новикова Л.Ю. Роль статистики в изучении разнообразия коллекции генетических ресурсов растений ВИР. *Бюллетень ВИР*. 2024;(243):21-29.

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах. Автор благодарит рецензентов за их вклад в экспертную оценку этой работы. Мнение журнала нейтрально к изложенным материалам, авторам и их месту работы.

Review article

The role of statistics in studying the diversity of the collection of plant genetic resources at VIR

Liubov Yu. Novikova**N.I. Vavilov All-Russian Institute of Plant Genetic Resources, St. Petersburg, Russia****Corresponding author:** Liubov Yu. Novikova, l.novikova@vir.nw.ru

Biostatistics began to develop synchronously with the formation of the first collections of plant genetic resources (PGR) in the late 19th and early 20th centuries. Since the foundation of the PGR collection at VIR, the entire available range of experimental methods was employed in the field and in the laboratory to evaluate its genetic diversity according to phenotypic and genotypic indicators and to select accessions with a given set of characteristics for specific breeding tasks. The key point is to choose a statistical research method that is adequate to the task. The purpose of this review was to summarize typical statistical problems that arise during phenotypic assessments of PGR collections and methods appropriate for their solution.

When analyzing PGR collections, the basic element is to determine qualitative and quantitative characteristics of their accessions. Descriptive statistics methods make it possible to estimate the mean values and stability of economically valuable plant characters on the basis of several years of studying in one or more locations. These indicators are compiled into databases, included into catalogues constantly published by VIR, and used in the selection of accessions promising for breeding tasks.

It is often necessary to compare groups of accessions belonging to different species and having different geographic origin, test results in different ecogeographic environments, viability under various storage conditions, etc. The hypothesis of differences among several plant features or variants of the experiment based on sample data is tested using Student's t-test and the analysis of variance. Modern methods for assessing the genotype $\times$ environment interaction are being actively developed.

The correlation and regression analysis is employed to identify weather and climate factors affecting valuable agronomic characters of accessions, such as yield, grain weight, resistance to biotic and abiotic factors, and quality indicators, and to find interrelationships among such characters that should be taken into account while developing a cultivar according to a given model. The active warming of the climate, which started in the 1970s, aroused the interest of biologists in disclosing crop yield trends and using time series analysis methods in regression modeling.

Those who describe a large PGR collection, examine phylogenetic relationships, select a representative sample, and build up core-collections would face a task of structuring the collection and identifying groups of similar accessions. The classification problem is solved by classification methods either without learning, such as the cluster analysis, or with learning, such as the discriminant analysis.

Analyzing PGR accessions for a large number of indicators requires identification of the most informative indicators, i.e., reducing the dimensions of the character space, accomplished by the principal component analysis or factor analysis techniques with a shift from real measured variables to a small number of hypothetical values, called factors.

The study of PGR collections is increasingly expanding beyond the limits of classical field observations of yield, for which biostatistics methods, based on the assumption of a normal distribution, were developed in the early 20th century. In the middle of the 20th century, nonparametric analytical methods and resampling techniques of computer simulation appeared.

The development of computers made it possible to make integrated data platforms that accumulated huge arrays of diverse PGR characteristics, and extend complex multidimensional methods available only to mathematicians to a wider range of researchers in various disciplines. Biostatistics methods associated with phenotypic data continue to evolve, ensuring improved assessment of the potential of PGR collections and providing integral information support to the modern genomic analysis.

Keywords: PGR collections, phenotypic characteristics, databases, statistical methods

Acknowledgments: The work was carried out within the framework of the implementation of the Development Program of the National Center for Plant Genetic Resources under the Agreement with the Ministry of Education and Science of Russia dated February 15, 2024 No. 075-02-2024-1090.

For citation: Novikova L.Yu. The role of statistics in studying the diversity of the collection of plant genetic resources at VIR. *VIR Bulletin*. 2024;(243):21-29. (In Russ.).

Financial transparency: The author has no financial interest in the presented materials or methods. The author thanks the reviewers for their contribution to the peer review of this work. The journal's opinion is neutral to the presented materials, the author, and his employer.

История биостатистики

Коллекции генетических ресурсов растений (ГРР) являются важнейшим ресурсом продовольственной безопасности страны. Эффективное оперативное управление таким сложным живым и динамичным объектом возможно только на основе мониторинга, численной оценки и экспертизы количественных и качественных характеристик образцов. На современном уровне цифровизации это означает интеграцию данных в дата-платформу как можно большей общности, с консолидированными данными различного характера, оборудованную аналитическими надстройками для эффективного анализа, прогнозирования и управления. С момента создания коллекции ГРР ВИР в полевых и лабораторных условиях с использованием всего доступного диапазона современных экспериментальных методов ведется оценка разнообразия генофонда по фенотипическим и генотипическим показателям, выделяются образцы с заданным набором характеристик для конкретных задач региональной селекции.

На современном уровне все численные массовые оценки должны быть выполнены с соблюдением статистических методов оценки значимости сделанных выводов. Ключевым моментом оценки численных данных является выбор адекватного задаче метода статистического исследования. Биостатистика (биологическая статистика) – наука об исследовании биологических объектов методами статистики. Каждый метод имеет свою область применения и свои возможности (Zaitsev, 1973, p. 5).

Многомерный анализ биологических данных начал активно развиваться в конце XIX – начале XX в. (Efimov, Kovaleva, 2008), синхронно с формированием первых коллекций ГРР. Представители математической статистики ставили своей задачей выявление закономерностей явлений на основе наблюдаемых вероятностных величин (Efimova et al., 2007). Началом биостатистики считается 1889 г., когда в книге Ф. Гальтона (1822–1911) появилось слово «*biometry*». Ф. Гальтон заложил основы корреляционного и регрессионного анализа. В 1901 г. при активном участии К. Пирсона (1857–1936) начал издаваться журнал «*Biometrika*». К. Пирсон предложил учение о форме распределения в биологии, коэффициент вариации, коэффициент парной линейной корреляции, критерий χ^2 , метод множественной регрессии, метод главных компонент. В 1908 г. В. Госсет (1876–1936) (под псевдонимом Стьюдент) создал теорию «малой выборки». Р. Фишер (1890–1962) в первой трети XX в. разработал методику планирования эксперимента и дисперсионный анализ. Эти методы легли в основу биостатистики и активно используются при анализе полевых экспериментов до сих пор, покрывая значительную часть потребностей анализа фенотипических данных генотипов. Основу большинства созданных на этом этапе методов составило представление о нормальном распределении исследуемых величин.

К середине XX в. стало ясно, что большое количество характеристик реальных объектов не подчиняется нормальному распределению. Получают развитие робастные методы, то есть нечувствительные к небольшим

отклонениям от нормального распределения (Dubrov et al., 2003), к концу 1940-х гг. приобретают популярность непараметрические методы, пригодные для анализа данных с любой формой распределения (Reznik, 2008).

50-60-е гг. XX века характеризовались развитием компьютерных методов, активно развиваются сложные расчетные процедуры многомерного разведочного анализа, служащие для выявления структуры совокупности объектов и показателей, снижения их размерности, такие как кластерный анализ, метод главных компонент, факторный анализ. Эти методы позволяют сокращать объемы обрабатываемой информации, не теряя при этом сути. Развивается байесовская статистика. Становятся возможными анализ, моделирование и прогнозирование огромных массивов любой информации, в том числе страниц интернета, методами Data Mining – имитационного моделирования, нейронных сетей, размножения выборок (Enslein et al., 1986; Dubrov et al., 2003).

XXI век характеризуется развитием методов биоинформатики (анализ генома, белковой структуры, биологических систем) с использованием инструментария прикладной математики, статистики и информатики.

Обязательным элементом современного биологического исследования является статистическое подтверждение выводов из массовых экспериментальных данных. Однако математическая теория статистических методов сложна для не математиков. С появлением и совершенствованием современных программ обработки данных статистическая обработка поднялась на новый уровень. Теперь исследователь может и не иметь математической подготовки. Практическим решением проблемы статистической обоснованности выводов стало понимание статистической парадигмы и использование апробированных статистических пакетов. Выбор конкретного пакета и метода компьютерного статистического анализа определяется задачей, типом переменных (качественные, количественные), видом распределения переменной (нормальное, отличное от нормального). Статистические программные продукты, которых насчитываются тысячи, делятся на интегрированные пакеты общего назначения (из которых при исследовании коллекций ГРР наиболее часто встречаются пакеты R, Statistica, SPSS, SAS, Statgraphics), методо-ориентированные пакеты (например, только для разведочного анализа, анализа нечисловых данных, как пакет DARwin), предметно-ориентированные пакеты (биомедицинские, сельскохозяйственные, генетические, биоинформационные, агрометеорологические) и обучающие программы. Пакет Statistica хорош огромным количеством реализованных методов, видов графиков, обилием русскоязычной литературы, дружелюбным интерфейсом (Borovikov, 2003, 2013; Vukolov, 2004; Khalafyan, 2010). Статистическая среда R принадлежит к наиболее динамично развивающимся в своем классе, позволяет проводить любой стандартный анализ данных, имеет бесплатную версию.

Всероссийский институт генетических ресурсов растений имени Н.И. Вавилова (ВИР) ведет обширный спектр исследований в области генетических ресурсов растений,

что позволяет определить наиболее актуальные задачи и соответствующие им статистические методы (Novikova, 2024). Статистический анализ данных коллекций ГРП имеет ряд особенностей: это возможность поставить активный эксперимент, близость распределения фенотипических характеристик к нормальному. При анализе разнообразия коллекций ГРП можно выделить типовые задачи и определенную последовательность их выполнения.

Целью данного обзора является сводка типовых статистических задач, возникающих при работе с фенотипическими оценками коллекций ГРП, и соответствующих им методов.

Описательная статистика и анализ распределения

При анализе коллекций ГРП базовым элементом является описание качественных и количественных характеристик образцов. Обязательно оцениваются средние значения и стабильность хозяйственно ценных признаков на основе трех или более лет изучения в одном или нескольких пунктах по методическим указаниям ВИР для каждой группы культур. Эти базовые статистики сводятся в базы описательных и оценочных характеристик и затем используются при выделении перспективных образцов для задач селекции, принятии решений о более глубоком изучении генотипов. Обработка первичных данных об исследованном фрагменте коллекции производится методами описательной статистики. Существенным элементом описательной статистики является представление результатов в виде таблиц и графиков.

Первый этап – это визуализация данных. Признаком хорошего статистического качества исследования являются грамотно подобранные компактные графики, позволяющие охватить и осмыслить большие массивы информации (Efimova et al., 2007). Современные статистические пакеты – Statistica, Statgraphics и др., табличный процессор MS Excel – предлагают широкий ассортимент графических подходов. Каждый тип графика предназначен для визуализации определенного типа данных (Efimova et al., 2007; Khalafyan, 2010): линейный график – для представления динамики количественных переменных, связи между количественными переменными; гистограмма – для изображения вариационного ряда как по количественным, так и качественным переменным; секторные диаграммы – для изображения структуры, то есть переменных, представляющих долю или процент целого, дающих в сумме 1 или 100% и т. д. В результате визуализации распределения данных делается вывод об отсутствии/наличии ошибок заполнения базы данных, однородности/неоднородности выборки, корректируется выборка (удаляются явно не принадлежащие ей элементы) или разбивается на несколько однородных выборок в случае выявления неоднородности; делается вывод о визуальной нормальности/отсутствии нормальности распределения количественных переменных.

Следующим этапом (Zaitsev, 1973; Efimova et al., 2007; Khizhnyak, Puchkova, 2019) становится расчет описатель-

ных статистик – точечных и интервальных оценок средних, показателей вариации, вида распределения признаков по образцам, годам, пунктам исследования. Выделяются два типа характеристик вариационного ряда – степенные и структурные (порядковые). Степенные характеристики хорошо описывают вариабельность ряда, близкого к нормальному распределению. Структурные характеристики характеризуют любой вид распределения переменной, измеренной в шкале не ниже порядковой, при любых параметрах распределения, то есть являются непараметрическими. К ним относят такие универсальные характеристики, как минимум, максимум, мода, медиана, квантили и т. д. Если распределение нормально, к нему может быть применен мощный, хорошо разработанный аппарат соответствующих статистических методов; если отличается от нормального, используют, как правило, методы непараметрической статистики. Основой применения таких методов, как дисперсионный, регрессионный, факторный, дискриминантный анализ, является нормальность распределения исследуемой выборки. Хотя все эти методы робастны, то есть хорошо работают и при небольшом отклонении от нормальности, после визуальной проверки однородности выборки приводят численные оценки вероятности гипотезы о нормальности распределения. Критерии согласия предназначены для проверки гипотезы о согласии между анализируемым эмпирическим и теоретическим распределением (Glotov et al., 1982; Eliseeva et al., 2006; Khizhnyak, Puchkova, 2019). Если распределение отлично от нормального, пытаются определить его вид, параметры, преобразовать распределение к нормальному. Нормальность проверяется такими критериями, как Шапиро – Уилка, χ^2 , Колмогорова – Смирнова. Критерии χ^2 и Колмогорова – Смирнова подходят для проверки гипотезы об соответствии формы экспериментального распределения любому теоретическому распределению – нормальному, равномерному, Пуассона и т. д. (Glotov et al., 1982; Dospekhov, 1985; Reznik, 2008).

Описательные и оценочные характеристики образцов фрагментов коллекций, сводки выделяющихся по каким-либо признакам образцов публикуются в каталогах, постоянно издаваемых ВИР.

Сравнение нескольких признаков или вариантов опыта

Если гипотеза о соответствии вида распределения нормальному на заданном уровне значимости может быть принята, переходят к более сложным исследованиям структуры коллекции методами параметрической статистики. Часто встает задача сравнения групп образцов разных видов, разного географического происхождения, результатов испытания в разных эколого-географических условиях, жизнеспособности при разных условиях хранения и т. д. Проверка гипотезы о различии/схождении нескольких признаков или вариантов опыта по выборочным данным осуществляется сравнением средних значений исследуемых выборок (Zaitsev, 1973; Dospekhov, 1985;

Eliseeva et al., 2006; Efimova et al., 2007). Если необходимо доказать различие двух нормально распределенных выборок, наиболее часто используется t-критерий Стьюдента. **t-критерий Стьюдента для независимых выборок** показывает, достоверно ли отличие среднего значения признака одной выборки от среднего другой выборки на принятом (как правило, 5-процентном) уровне значимости. **t-критерий Стьюдента для связанных выборок** (Dospekhov, 1985, p. 194–195) используется для анализа различий характеристик выборок до и после некоторого воздействия. Для расчета критерия рассчитывается вспомогательная величина – разница значения признака до и после воздействия для каждого объекта. В случае опыта «до – после» применение критерия для связанных выборок позволяет определить больше достоверных различий, чем использование критерия для независимых выборок, так как значительно уменьшает влияние разнообразия выборки на результат опыта. t-критерий Стьюдента не подходит для множественных сравнений, поскольку при каждом последующем сравнении вероятность ошибки множественного сравнения увеличивается, поэтому задачу сравнения более двух выборок решает дисперсионный анализ.

Дисперсионный анализ (Scheffé, 1963; Glotov et al., 1982; Dospekhov, 1985; Eliseeva et al., 2006) является самым востребованным видом статистического анализа при анализе полевого изучения коллекций ГРП. Дисперсионный анализ может быть однофакторным (исследуется влияние одного качественного признака на количественную характеристику, например влияние года на среднюю урожайность коллекции) и многофакторным (исследуется влияние нескольких факторов – года, пункта, сорта). Опыт с числом факторов более трех сложен для осмысления, поэтому, как правило, ограничиваются не более чем трехфакторными планами. Для полноценного многофакторного анализа нужен полный факторный эксперимент: каждый уровень фактора должен быть исследован с каждым, что ставит задачу правильного планирования эксперимента. Для снижения ошибки эксперимента необходимо наличие повторностей в каждом опыте, рандомизированное (случайное) размещение опытных делянок (Dospekhov, 1985). Дисперсионный анализ дает ответ на вопрос, есть ли хотя бы одна выделяющаяся из общей совокупности группа. Чтобы определить, средние каких групп друг от друга отличаются, нужно проводить дальнейший, апостериорный анализ. Множественные сравнения позволяют сделать вывод о достоверном отличии среднего значения одной группы от всех, оставаясь на принятом уровне значимости. Могут быть использованы разные критерии: наименьшей существенной разницы (НСР), Данкана, Шеффе, Тьюки (Scheffé, 1963, p. 115–118). Наиболее распространен критерий Тьюки. Востребованной опцией в российской традиции анализа коллекций ГРП является определение НСР и публикация их значений для каждого признака в таблицах с описательными характеристиками образцов, что дает возможность самостоятельно по таблице подтвердить, например, достоверность отличия средней урожайности образцов друг

от друга и от стандарта на фоне межгодовой вариабельности (Dospekhov, 1985, p. 214–225). Дисперсионный анализ является основой выделения перспективных доноров и источников для задач селекции на разные признаки.

Специальные методы анализа взаимодействия «генотип × среда» (GE) отражают существенное развитие дисперсионного анализа, позволяют сравнить стабильность урожайности образцов данной выборки в контрастных средах – пунктах, годах исследования (Eberhart, Russell, 1966; Pakudin, 1976; Kilchevsky, Khotyleva, 1989). Современным развитием метода является подробная оценка особенности реакции генотипа на среду моделями типа AMMI (Zobel et al., 1988), GGE (Yan et al., 2000).

Определение связи нескольких признаков

Создание модели сорта, понимание адаптационных механизмов культур требует выявления связей хозяйственно ценных признаков образцов, таких как урожайность, показатели качества, устойчивость к биотическим и абиотическим факторам, друг с другом и с погодно-климатическими факторами. Оценка этих связей осуществляется методами корреляционно-регрессионного анализа, по сути представляющими собой частный случай дисперсионного анализа (Enslein et al., 1986).

Корреляционный анализ является одним из методов статистической оценки степени связи нескольких признаков (Glotov et al., 1982; Ayvazyan et al., 1985; Dospekhov, 1985; Enslein et al., 1986; Ayvazyan, Mkhitarian, 2001; Eliseeva et al., 2006; Efimova et al., 2007). Парный коэффициент корреляции Пирсона служит мерой линейной связи между двумя признаками, коэффициент множественной линейной корреляции служит для измерения тесноты большего, чем два, числа признаков между собой. Для коэффициента корреляции, как и любой выборочной характеристики, необходимо показать достоверность, то есть значимость его отличия от нуля; кроме того, характеризуют направление (положительная/отрицательная) и силу корреляционной связи. Сильной считается связь при значении модуля коэффициента корреляции $|r| > 0,7$, средней – при $0,3 \leq |r| \leq 0,7$, слабой – при $|r| < 0,3$ (Dospekhov, 1985, p. 271).

Регрессионный анализ – построение уравнения, описывающего зависимость признака-результата от одного или нескольких признаков-факторов (Glotov et al., 1982; Dospekhov, 1985; Enslein et al., 1986; Vuchkov et al., 1987; Ayvazyan, Mkhitarian, 2001; Eliseeva et al., 2006; Efimova et al., 2007). Показателем качества уравнения регрессии является коэффициент детерминации, показывающий, какую долю общей вариабельности результирующего признака объясняет функция регрессии. Поскольку, как правило, уравнение получено по выборочным данным, необходимо определить значимость каждого коэффициента и уравнения в целом. При большом числе возможных влияющих признаков – предикторов – встает вопрос, какие признаки включать в регрессионную модель. Решение может быть осуществлено разными вычислительными процедурами – с последовательным включением

переменных, когда на каждом следующем шаге выбирается наиболее значимая переменная, или с исключением переменных (Enslein et al., 1986). Преимущество метода с последовательным включением переменных заключается в том, что исходное число потенциальных предикторов может быть больше числа измерений.

Начавшееся в 1970-х гг. активное потепление климата вызвало интерес биологов к **анализу временных рядов**, в значительной степени развивавшихся ранее в экономике (Box, Jenkins, 1974; Afanasyev et al., 2006; Eliseeva et al., 2007; Efimov, Kovaleva, 2008; Sirotenko, 2012). Анализ зависимостей хозяйственно ценных признаков растений за последние 30–50 лет осложнен наличием климатических и агротехнических трендов. Устранение трендов из временных рядов позволяет повышать качество регрессионных моделей и удалять ложные корреляции (Box, Jenkins, 1974; Ayvazyan et al., 1985; Eliseeva et al., 2007; Vukolov, 2004; Sirotenko, 2012; Iler et al., 2017).

Классификация объектов

При описании объемной коллекции ГРП, исследовании филогенетических связей, выделении репрезентативной выборки, создании соге-коллекций возникает задача классификации, выделения групп схожих образцов. Возникает вопрос: нельзя ли снизить количество описываемых объектов, выделив однородные группы? Задача классификации решается методами классификации без обучения (кластерный анализ) и с обучением (дискриминантный анализ) (Enslein et al., 1986; Ayvazyan et al., 1989; Dubrov et al., 2003; Vukolov, 2004; Efimov, Kovaleva, 2008; Simchera, 2008; Khizhnyak, Puchkova, 2019).

Кластерный анализ – группа методов, используемых для классификации объектов в гомогенные (однородные) группы – кластеры. Началом кластерного анализа была книга Р. Сокзла и А. Спита «Численная таксономия» 1957 г. (Enslein et al., 1986, p. 19; Dubrov et al., 2003). Кластерный анализ позволяет выделить однородные группы по совокупности всех известных характеристик образцов. Результатом кластерного анализа является дендрограмма (дерево), объединяющая все образцы по мере их сходства. Разработано множество метрик сходства и различия образов и методов выделения однородных групп. Существуют несколько способов рассчитать расстояние между объектами по совокупности изученных количественных показателей: евклидово, Манхэттен, Чебышева, Минковского, для бинарных переменных – процент несогласия и др. При одной и той же матрице расстояний между объектами можно получить разные дендрограммы с использованием разных методов определения близости объектов и групп объектов друг другу (Ayvazyan, Mkhitarian, 2001; Dubrov et al., 2003; Borovikov, 2003, 2013; Khalafyan, 2010): ближайшего соседа, дальнего соседа, попарного среднего, невзвешенный центроидный, Варда и др. Различные численные методы приводят к разным деревьям классификации, и выбор окончательного решения осуществляется в соответствии с априорными представлениями. Если появляются группы из единичных образ-

цов, можно использовать взвешенные методы, в которых более крупные группы «притягивают» мелкие, так как в качестве весового коэффициента в них используется размер кластера. Метод *k*-средних позволяет разбить исследуемую выборку на заданное число классов *k*.

Дискриминантный анализ решает задачу классификации объектов на заданное количество классов на основании обучающей выборки, у элементов которой известна принадлежность их к классам (популяциям) (Malinovsky, 1979; Enslein et al., 1986, p. 113; Ayvazyan, Mkhitarian, 2001; Dubrov et al., 2003; Khizhnyak, Puchkova, 2019). Анализируя эти группы объектов, определяют набор наиболее информативных признаков, отвечающих за разделение на классы, дискриминантную функцию и решающее правило (дискриминантную функцию), в соответствии с которым можно однозначно отнести объект к определенному классу по его значениям признаков. О качестве модели можно судить по числу неправильно классифицированных объектов обучающих выборок. Создание дискриминантных функций дает возможность решить вопрос о вероятности принадлежности нового объекта к исследованным популяциям (Enslein et al., 1986, p. 113).

Снижение числа анализируемых признаков и отбор наиболее информативных показателей

При анализе образцов ГРП по большому количеству показателей стоит задача выявления наиболее информативных показателей, то есть снижения размерности пространства признаков. При этом происходит переход от реальных измеренных переменных к небольшому количеству гипотетических величин, называемых факторами (Enslein et al., 1986; Ayvazyan et al., 1989; Ayvazyan, Mkhitarian, 2001; Borovikov, 2003, 2013; Vukolov, 2004; Simchera, 2008; Efimov, Kovaleva, 2008; Khalafyan, 2010; Khizhnyak, Puchkova, 2019). Основные предпосылки, дающие возможность снижения числа исходных признаков, – дублирование информации сильно связанными признаками; неинформативность признаков, мало меняющихся при переходе от одного объекта к другому; возможность агрегирования (суммирования) некоторых признаков (Ayvazyan, Mkhitarian, 2001, p. 142).

Метод главных компонент (МГК) был предложен К. Пирсоном в 1901 г., а затем детально разработан американским экономистом и статистиком Г. Хотеллингом. Метод заключается в получении нескольких новых показателей – главных компонент, являющихся линейными комбинациями (взвешенными суммами) исходных признаков, полученных в результате эксперимента (Dubrov et al., 2003, p. 134–171). Главные компоненты упорядочиваются в порядке убывания той дисперсии, которую они объясняют. Первый фактор (главная компонента) объясняет максимальную часть дисперсии исходного многомерного пространства признаков, вторая – меньшую и т. д. Общее число компонент равно исходному количеству переменных, однако некоторые факторы вносят очень маленький вклад в общую дисперсию. От исходного набора признаков переходят к набору факторов (глав-

ных компонент), которые могут объяснить значительную часть дисперсии исходной выборки. Существуют различные способы определения числа факторов (главных компонент), которое надо оставить для анализа. Критерий Кайзера предлагает оставить факторы со значением собственных чисел (числа включенных в него реальных переменных) больше 1; критерий «каменистой осыпи» – то число факторов, после которого добавление новых факторов слабо увеличивает объясненную дисперсию (определяется графически), или оставляются факторы, которые объясняют не менее 60–70% общей дисперсии (Simchera, 2008, р. 112). Коэффициенты, с помощью которых можно узнать, какие исходные переменные наибольшим образом связаны с факторами (главными компонентами), называются факторными нагрузками, они позволяют содержательно интерпретировать факторы. Координаты объектов в пространстве новых признаков – факторов – называются факторными координатами. Графики расположения исследуемых объектов в пространстве первых факторов (главных компонент) позволяют выявить группы схожих объектов. Для достижения лучшего соответствия априорным данным далее может быть использовано вращение системы новых координат.

Факторный анализ выполняет те же задачи снижения размерности пространства переменных и объектов, но в основе лежит другая расчетная процедура – выделяется общая для нескольких переменных дисперсия. Факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. В методе главных компонент число исходных переменных равно числу факторов, а в факторном анализе, вообще говоря, существенно меньше. Некоторые исследователи считают МГК разновидностью факторного анализа (Enslein et al., 1986, р. 184).

Непараметрические методы

Изучение коллекций ГРП все более выходит за рамки классических полевых наблюдений урожайности, для которых в начале XX в. были разработаны виды анализа, основанные на предположении о нормальном характере распределения. Все более популярными делаются методы непараметрической статистики, которые не относятся к какому-либо конкретному распределению (Vukolov, 2004; Efimov, Kovaleva, 2008; Reznik, 2008).

Для данных, имеющих существенное отклонение от нормального распределения, возможны 4 подхода к обработке: 1) нормализующее преобразование (логарифмирование, арксин-преобразование и др.) и применение далее параметрических методов; 2) использование робастных методов (усеченные средние, абсолютные отклонения вместо среднеквадратичных); 3) использование непараметрических критериев; 4) использование ресэмплинг-методов имитационного компьютерного моделирования (джекнайф, бутстреп, Монте-Карло и др.).

Методы робастного оценивания параметров распре-

делений устойчивы к выбросам и малым отклонениям от известных распределений. Возникли методы выявления грубых ошибок (метод Смирнова – Граббса), отсекающие грубые ошибки (методы Пуанкаре, Винзора), группировки наблюдений для получения робастных оценок (методы анализа смеси распределений Тьюки – Хубера) (Dubrov et al., 2003, р. 301-303).

В середине XX в. появились непараметрические методы, не обращающиеся к расчету параметров распределения, но решающие те же задачи, что и параметрические (Glotov et al., 1982; Reznik, 2008). Аналогом t-критерия Стьюдента для независимых выборок является U-критерий Манна – Уитни; для зависимых выборок – критерий Вилкоксона, МакНемара; для нескольких зависимых выборок – Q-критерий Кохрана, знаков Фридмана; дисперсионного анализа – Краскела – Уоллиса, медианный; коэффициента корреляции Пирсона – коэффициенты ранговой корреляции Спирмена и Кендалла; для сравнения двух эмпирических распределений качественных признаков могут быть использованы критерии χ^2 или точный тест Фишера (для таблицы 2×2 при числе значений меньше 30). Непараметрические методы целесообразнее применять при сочетании трех факторов (Reznik, 2008, р. 55): неизвестен закон распределения; выборка мала; исходные данные представляют собой качественные переменные – порядковые (ранговые) или номинальные (неупорядоченные) данные: дегустационные оценки, балльные оценки устойчивости к болезни, название сорта, вида и т. п.

Анализ связи качественных (категоризированных, текстовых) переменных осуществляется методом анализа таблиц сопряженности (Ayvazyan, Mkhitarian, 2001; Efimova et al., 2007). В случае нелинейной связи между переменными при неизвестном виде зависимости в качестве меры связи между переменными может быть использовано корреляционное отношение η , предложенное К. Пирсоном в 1896 г. (Efimova et al., 2007, р. 240). Расчет корреляционного отношения основан на дисперсионном анализе, квадрат корреляционного отношения определяется как отношение межгрупповой дисперсии выборки к общей, группы определяются по одной из переменных. В качестве критерия наличия связи между качественными показателями с большим количеством градаций может быть использован критерий χ^2 (Efimova et al., 2007, р. 254). На основе χ^2 определяются такие показатели тесноты связи, как коэффициент сопряженности К. Пирсона, коэффициент взаимной сопряженности А. А. Чупрова (Efimova et al., 2007, р. 254-255).

Развитие компьютеров, повышение скорости вычислений привели к возможности использования методов, основанных на множественных расчетах – размножении выборок, моделировании случайных воздействий методом Монте-Карло (Enslein et al., 1986; Dubrov et al., 2003). Например, непараметрический аналог дисперсионного анализа PERMANOVA позволяет применить дисперсионный анализ к данным, обобщенным с помощью любого показателя расстояния (Anderson et al., 2017), и получить статистические оценки с использованием размножения

выборок.

Многомерное шкалирование является непараметрическим аналогом факторного анализа (Enslein et al., 1986, p. 301; Ayvazyan et al., 1989; Efimov, Kovaleva, 2008). Входная информация для анализа представляет собой матрицу чисел, содержащих информацию о порядковых отношениях между объектами. Выходная информация – точки в пространстве небольшой размерности, в графической форме или в виде координат. Фактически метод получил распространение после компьютерной реализации в 1962 г. Б. Шепардом (Enslein et al., 1986, p. 301–304).

Заключение

Каждый из методов статистического исследования решает определенную задачу для определенного типа данных, и можно говорить о типовых подходах в растениеводстве и управлении коллекциями ГРП. Развитие компьютеров позволило создать интегрированные дата-платформы, собирающие огромные массивы разноплановых характеристик ГРП, масштабировать сложные многомерные методы, ранее доступные только математикам, на широкий круг исследователей в различных предметных областях. Методы биостатистики, относящиеся к фенотипическим данным, продолжают развиваться, обеспечивая совершенствование оценок потенциала коллекций ГРП и являясь неотъемлемой информационной составляющей современного геномного анализа.

References / Литература

- Afanasyev V.N., Yuzbashev M.M., Gulyaeva T.I. Econometrics (Ekonometrika). Moscow: Finansy i Statistika; 2006. [in Russian] (Афанасьев В.Н., Юзбашев М.М., Гуляева Т.И. Эконометрика. Москва: Финансы и статистика; 2006).
- Anderson M.J. Permutational multivariate analysis of variance (PERMANOVA). In: N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, J.L. Teugels (eds). *Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ: John Wiley & Sons, Ltd.; 2017. p.1-15. DOI: [10.1002/9781118445112.stat07841](https://doi.org/10.1002/9781118445112.stat07841)
- Ayvazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. Applied statistics: classification and dimensionality reduction (Prikladnaya statistika: klassifikatsiya i snizheniye razmernosti). Moscow: Finansy i Statistika; 1989. [in Russian] (Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. Москва: Финансы и статистика; 1989).
- Ayvazyan S.A., Enyukov I.S., Meshalkin L.D. Applied statistics: a study of dependencies (Prikladnaya statistika: issledovaniye zavisimostey). Moscow: Finansy i Statistika; 1985. [in Russian] (Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: исследование зависимостей. Москва: Финансы и статистика; 1985).
- Ayvazyan S.A., Mkhitarian V.S. Applied statistics in tasks and exercises (Prikladnaya statistika v zadachakh i uprazhneniyakh). Moscow: Unity-Dana; 2001. [in Russian] (Айвазян С.А., Мхитарян В.С. Прикладная статистика в задачах и упражнениях. Москва: Юнити-Дана; 2001).
- Borovikov V.P. A popular introduction to modern data analysis in the STATISTICA system (Populyarnoye vvedeniye v sovremennyy analiz dannykh v sisteme STATISTICA). Moscow: Goryachaya Liniya – Telekom; 2013. [in Russian] (Боровиков В.П. Популярное введение в современный анализ данных в системе STATISTICA. Москва: Горячая линия – Телеком; 2013).
- Borovikov V.P. STATISTICA. The art of data analysis on a computer: for professionals (STATISTICA. Iskustvo analiza dannykh na kompyutere: dlya professionalov). 2nd ed. St. Petersburg: Piter; 2003. [in Russian] (Боровиков В.П. STATISTICA. Искусство анализа данных на компьютере: для профессионалов. 2-е изд. Санкт-Петербург: Питер; 2003).
- Box G., Jenkins G. Time series analysis. Forecasting and control. Issue 1. Moscow: Mir; 1974. [in Russian] (Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. Вып. 1. Москва: Мир; 1974).
- Dospikhov B.A. Methodology of field trial (with fundamentals of statistical processing of research results) (Metodika polevogo opyta [s osnovami statisticheskoy obrabotki rezultatov issledovaniy]). 5th ed. Moscow: Agropromizdat; 1985. [in Russian] (Доспихов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). 5-е изд. Москва: Агропромиздат; 1985).
- Dubrov A.M., Mkhitarian V.S., Troshin L.I. Multidimensional statistical methods (Mnogomernyye statisticheskiye metody). Moscow: Finansy i Statistika; 2003. [in Russian] (Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. Москва: Финансы и статистика; 2003).
- Eberhart S.A., Russell W.A. Stability parameters for comparing varieties. *Crop Science*. 1966;6(1):36-40. DOI: [10.2135/cropsci1966.0011183X00060001001x](https://doi.org/10.2135/cropsci1966.0011183X00060001001x)
- Efimov V.M., Kovaleva V.Yu. Multidimensional analysis of biological data: a training manual (Mnogomernyy analiz biologicheskikh dannykh: uchebnoye posobiye). 2nd ed. St. Petersburg: VIZR; 2008. [in Russian] (Ефимов В.М., Ковалева В.Ю. Многомерный анализ биологических данных: учебное пособие. 2-е изд. Санкт-Петербург: ВИЗР; 2008).
- Efimova M.R., Petrova E.V., Rumyantsev V.N. General theory of statistics: a manual (Obshchaya teoriya statistiki: uchebnik). 2nd ed. Moscow: INFRA-M; 2007. [in Russian] (Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики: учебник. 2-е изд. Москва: ИНФРА-М; 2007).
- Eliseeva I.I., Isotov A.V., Kapralova E.B., Flud N.A., Shchirina A.N. Statistics (Statistika). I.I. Eliseeva (ed.). Moscow: KNORUS; 2006. [in Russian] (Елисеева И.И., Изотов А.В., Капралова Е.Б., Флуд Н.А., Щирина А.Н. Статистика / под ред. И.И. Елисеевой. Москва: КНОРУС; 2006).
- Eliseeva I.I., Kuryshcheva S.V., Kosteeva T.V., Pantina I.V., Mikhailov B.A., Neradovskaya Yu.V., Strohe H.G., Bartels K., Rybkina L.R. Econometrics (Ekonometrika). Moscow: Finansy i Statistika; 2007. [in Russian] (Елисеева И.И., Курышчева С.В., Костеева Т.В., Пантина И.В., Михайлов Б.А., Нерадовская Ю.В., Штрое Г.Г., Бартелс К., Рыбкина Л.Р. Эконометрика / под ред. И.И. Елисеевой. Москва: Финансы и статистика; 2007).
- Enslein A., Ralston A., Wilf H.S. (eds). Statistical methods for digital computers. Moscow: Nauka; 1986. [in Russian] (Статистические методы для ЭВМ / под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. Москва: Наука; 1986).
- Glotov N.V., Zhivotovsky L.A., Khovanov N.V., Khromov-Borisov N.N. Biometrics (Biometriya). Leningrad: Leningrad State University; 1982. [in Russian] (Глотов Н.В., Животовский Л.А., Хованов Н.В., Хромов-Борисов Н.Н. Биометрия. Ленинград: Ленинградский государственный университет; 1982).
- Iler A.M., Inouye D.W., Schmidt N.M., Høye T.T. Detrending phenological time series improves climate-phenology analyses and reveals evidence of plasticity. *Ecology*. 2017;98(3):647-655. DOI: [10.1002/ecy.1690](https://doi.org/10.1002/ecy.1690)
- Khalafyan A.A. Statistica 6. Statistical data analysis (Statistica 6. Statisticheskiy analiz dannykh). Moscow: Binom; 2010. [in Russian] (Халафян А.А. Statistica 6. Статистический анализ данных. Москва: Бином; 2010).
- Khizhnyak S.V., Puchkova E.P. Mathematical methods in agroecology and biology: a training manual (Matematicheskiye metody v agroekologii i biologii: uchebnoye posobiye). Krasnoyarsk: Krasnoyarsk State Agrarian University; 2019. [in Russian] (Хижняк С.В., Пучкова Е.П. Математические методы в агроэкологии и биологии: учебное пособие. Красноярск: Красноярский государственный аграрный университет; 2019).
- Kilchevsky A.V., Khotyleva L.V. Genotype and environment in plant breeding (Genotip i sreda v selektsii rasteniy). Minsk: Nauka i Tekhnika; 1989. [in Russian] (Кильчевский А.В., Хотылева Л.В. Генотип и среда в селекции растений. Минск: Наука и техника; 1989).
- Malinovskiy L.G. Classification of objects by means of discriminant analysis (Klassifikatsiya obyektov sredstvami diskriminantnogo analiza). Moscow: Nauka; 1979. [in Russian] (Малиновский Л.Г. Классификация объектов средствами дискриминантного

- анализа. Москва: Наука; 1979).
- Novikova L.Yu. Methods for statistical processing of phenotypic data in plant genetic resources collections. St. Petersburg: VIR; 2024. [in Russian] (Новикова Л.Ю. Методы статистической обработки фенотипических данных коллекций генетических ресурсов растений. Санкт-Петербург: ВИР; 2024).
- Pakudin V.Z. Parameters of environmental plasticity assessment in cultivars and hybrids (Parametry otsenki ekologicheskoy plastichnosti sortov i gibridov). In: L.V. Khotyleva (ed.). *Theory of Selection in Plant Populations (Teoriya otbora v populyatsiyakh rasteniy)*. Novosibirsk: Nauka; 1976. p.178-189. [in Russian] (Пакудин В.З. Параметры оценки экологической пластичности сортов и гибридов. В кн.: *Теория отбора в популяциях растений* / под ред. Л.В. Хотылевой. Новосибирск: Наука; 1976. С.178-189).
- Reznik A.D. A book for those who do not like statistics, but are forced to use it. Nonparametric statistics in examples, exercises, and drawings (Kniga dlya tekhn, kto ne lyubit statistiku, no vyzhuden yeyu polzovatsya. Neparametricheskaya statistika v primerakh, uprazhneniyakh i risunkakh). St. Petersburg: Rech; 2008. [in Russian] (Резник А.Д. Книга для тех, кто не любит статистику, но вынужден ею пользоваться. Непараметрическая статистика в примерах, упражнениях и рисунках. Санкт-Петербург: Речь; 2008).
- Scheffé H. The analysis of variance. Moscow: State Publishing House of Physico-Mathematical Literature; 1963. [in Russian] (Шеффе Г. Дисперсионный анализ. Москва: Государственное издательство физико-математической литературы; 1963).
- Simchera V.M. Methods of the multidimensional analysis of statistical data (Metody mnogomernogo analiza statisticheskikh dannykh). Moscow: Finansy i Statistika; 2008. [in Russian] (Симчера В.М. Методы многомерного анализа статистических данных. Москва: Финансы и статистика; 2008).
- Sirotenko O.D. Fundamentals of agricultural meteorology. Vol. II. Methods of calculations and forecasts in agrometeorology. Book 1. Mathematical models in agrometeorology (Osnovy selskokhozyaystvennoy meteorologii. T. II. Metody raschetov i prognozov v agrometeorologii. Kniga 1. Matematicheskie modeli v agrometeorologii: uchebnoye posobiye). Obninsk: VNIIGMI-MTsD; 2012. [in Russian] (Сиротенко О.Д. Основы сельскохозяйственной метеорологии. Т. II. Методы расчетов и прогнозов в агрометеорологии. Книга 1. Математические модели в агрометеорологии: учебное пособие. Обнинск: ВНИИГМИ-МЦД; 2012).
- Vuchkov I., Boyadjieva L., Solakov E. Applied linear regression analysis (Prikladnoy lineyny regressiionny analiz). Moscow: Finansy i Statistika; 1987. [in Russian] (Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. Москва: Финансы и статистика; 1987).
- Vukolov E.A. Fundamentals of statistical analysis: a workshop on statistical methods and operations research using the STATISTICA and EXCEL packages (Osnovy statisticheskogo analiza: praktikum po statisticheskim metodam i issledovaniyu operatsiy s ispolzovaniyem paketov STATISTICA i EXCEL). Moscow: FORUM; 2004. [in Russian] (Вуколов Э.А. Основы статистического анализа: практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. Москва: ФОРУМ; 2004).
- Yan W., Hunt L.A., Sheng Q., Szlavnic Z. Cultivar evaluation and mega-environment investigation based on GGE biplot. *Crop Science*. 2000;40(3):596-605. DOI: [10.2135/cropsci2000.403597x](https://doi.org/10.2135/cropsci2000.403597x)
- Zaitsev G.N. Methodology of biometric calculations. Mathematical statistics in experimental botany (Metodika biometricheskikh raschetov. Matematicheskaya statistika v eksperimentalnoy botanike). Moscow: Nauka; 1973. [in Russian] (Зайцев Г.Н. Методика биометрических расчетов. Математическая статистика в экспериментальной ботанике. Москва: Наука; 1973).
- Zobel R.W., Wright A.J., Gauch Jr. H.G. Statistical analysis of a yield trial. *Agronomy Journal*. 1988;80(3):388-393. DOI: [10.2134/agronj1988.00021962008000030002x](https://doi.org/10.2134/agronj1988.00021962008000030002x)

Информация об авторе

Любовь Юрьевна Новикова, доктор сельскохозяйственных наук, ведущий научный сотрудник, и. о. заведующего отделом, Федеральный исследовательский центр Всероссийский институт генетических ресурсов растений имени Н.И. Вавилова, 190000 Россия, Санкт-Петербург, ул. Б. Морская, 42, l.novikova@vir.nw.ru, <https://orcid.org/00000003-4051-3671>

Information about the author

Liubov Yu. Novikova, Dr. Sci. (Agriculture), Leading Researcher, Acting Head of a Department, N.I. Vavilov All-Russian Institute of Plant Genetic Resources, 42 Bolshaya Morskaya Street, St. Petersburg 190000, Russia, l.novikova@vir.nw.ru, <https://orcid.org/0000-0003-4051-3671>

Вклад автора: автор сделал самостоятельный вклад в подготовку публикации.

Contribution of the author: the author contributed solely to this article.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Conflict of interests: the author declares no conflict of interests.

Статья поступила в редакцию 02.12.2024; одобрена после рецензирования 11.12.2024; принята к публикации 19.12.2024.

The article was submitted on 02.12.2024; approved after reviewing on 11.12.2024; accepted for publication on 19.12.2024.