Применение графических возможностей программной среды R для анализа экспериментальных данных по селекции тритикале

А.Ф. ЧЕШКОВА[1,*], А.Ф. АЛЕЙНИКОВ[1,2], П.И. СТЕПОЧКИН[3]

[1] Сибирский федеральный научный центр агробиотехнологий РАН, Краснообск, Новосибирская область, Россия

[2] Новосибирский государственный технический университет, Россия

[3] Сибирский НИИ растениеводства и селекции – филиал ИЦиГ СО РАН, Краснообск, Новосибирская область, Россия

* Контактный e-mail: anna.cheshkova@sorashn.ru

Описан опыт применения программной среды R для визуализации и статистического анализа данных полевых опытов СибНИИРС по изучению образцов яровой и озимой тритикале.

Ключевые слова: программная среда R, статистический анализ данных, тритикале, селекция.

Введение

Статистический анализ результатов полевых опытов – неотъемлемая часть селекционных исследований. Широкий спектр биометрических методов используется на различных этапах селекционной работы: при изучении коллекций и выделении лучших исходных форм, подборе родительских пар для гибридизации, отборе наиболее перспективных гибридных образцов и на этапе сортоиспытания. Использование специализированного программного обеспечения облегчает расчеты и позволяет представлять результаты анализа в графической и понятной информативной форме. Прекрасной альтернативой многочисленным коммерческим программным продуктам в этой области является свободно распространяемая программная среда R, являющаяся динамично развивающейся статистической платформой общего назначения [1].

Исследования проводились с целью анализа результатов изучения коллекционных образцов тритикале, выделения групп сортов, различающихся по комплексу признаков, а также выявления закономерностей сопряженной изменчивости количественных признаков тритикале. Материалом для исследования послужили данные полевых опытов СибНИИРС по изучению образцов яровой и озимой тритикале из мировой коллекции ВИР 2009 и 2011 гг. В 2009 г. изучали 51 сорт яровой тритикале и 103 сорта озимой тритикале, в 2011 г. – 120 образцов озимой тритикале. При проведении эксперимента учитывали 12 морфологических признаков: X1 – продолжительность периода от всходов до колошения, дн.; X4 – высота растения, см; X5 – длина колоса, см; X6 – число колосков в колосе, шт.; X8 – число зерен в колосе, шт.; X9 – масса зерен колоса, г; X11 – масса 1000 зерен, г; X14 – натура

зерна (масса 1 л зерен), г; X15 – длина остей, см; X16 – диаметр шейки, см; X26 – число продуктивных побегов, шт. /м^2; X27 – общая масса зерен, г/м^2.

1. Разведочный анализ данных

При проведении статистического анализа следует начинать работу с тщательного ознакомления со свойствами полученных данных и проверки необходимых условий применимости соответствующих статистических методов. Этот начальный этап называют разведочным анализом. В литературе по статистике можно найти немало рекомендаций по выполнению разведочного анализа данных [1, 2]. Метод включает:

– выявление точек-выбросов;
– проверку однородности групповых дисперсий;
– проверку нормальности распределения данных;
– выявление коллинеарных переменных;
– выявление характера связи между анализируемыми переменными.

Разведочный анализ данных наиболее эффективен при использовании разнообразных графических средств, поскольку графики позволяют лучше понять структуру и свойства анализируемых данных в отличие от формальных статистических тестов [3].

Для *выявления выбросов* (или экстремальных значений) обычно используют диаграмму размахов – “ящик с усами”. В R для ее построения служит функция boxplot(). На рис. 1 приведен пример диаграммы размахов для признака X15 – “длина остей”, где отображены оценки центральной тенденции (медианы) и разброса (интерквартильного размаха – ИКР). “Усы” простираются от границ “ящика” до наибольшего (наименьшего) выборочного значения, находящегося в пределах расстояния 1.5 ИКР от этой границы. Наблюдения, находящиеся за пределами “усов”, потенциально могут быть выбросами.

Чувствительность разных статистических методов к наличию выбросов в данных неодинакова. В некоторых случаях требуется нормализовать исходные данные (например, логарифмированием). При большом объеме выборки допустимо отбрасывание аномальных значений, если они не отражают естественную вариабельность данных.

Проверка *однородности групповых дисперсий* необходима при использовании методов дисперсионного анализа данных, дискриминантного анализа и ряда регрессионных моделей. Гипотеза равенства дисперсий может быть проверена с помощью F-критерия Фишера (функция var.test() в R), теста Левена (функция leveneTest()) или теста Бартлетта (функция bartlett.test()). Кроме того, дисперсии можно сравнить визуально на графике, используя ту же функцию boxplot() для нескольких групп данных (рис. 1).

Проверять *нормальность распределения данных* требуется, в частности, в дискриминантном и дисперсионном анализе, в то время как для метода главных компонент нормальное распределение не является необходимым условием. Среди графических способов проверки на нормальность обычно используются гистограммы (функция hist() в R), графики ядерной оценки плотности (функция density()) и графики квантилей (функция qqplot() в R). Пример графика ядерной оценки плотности приведен на рис. 2.

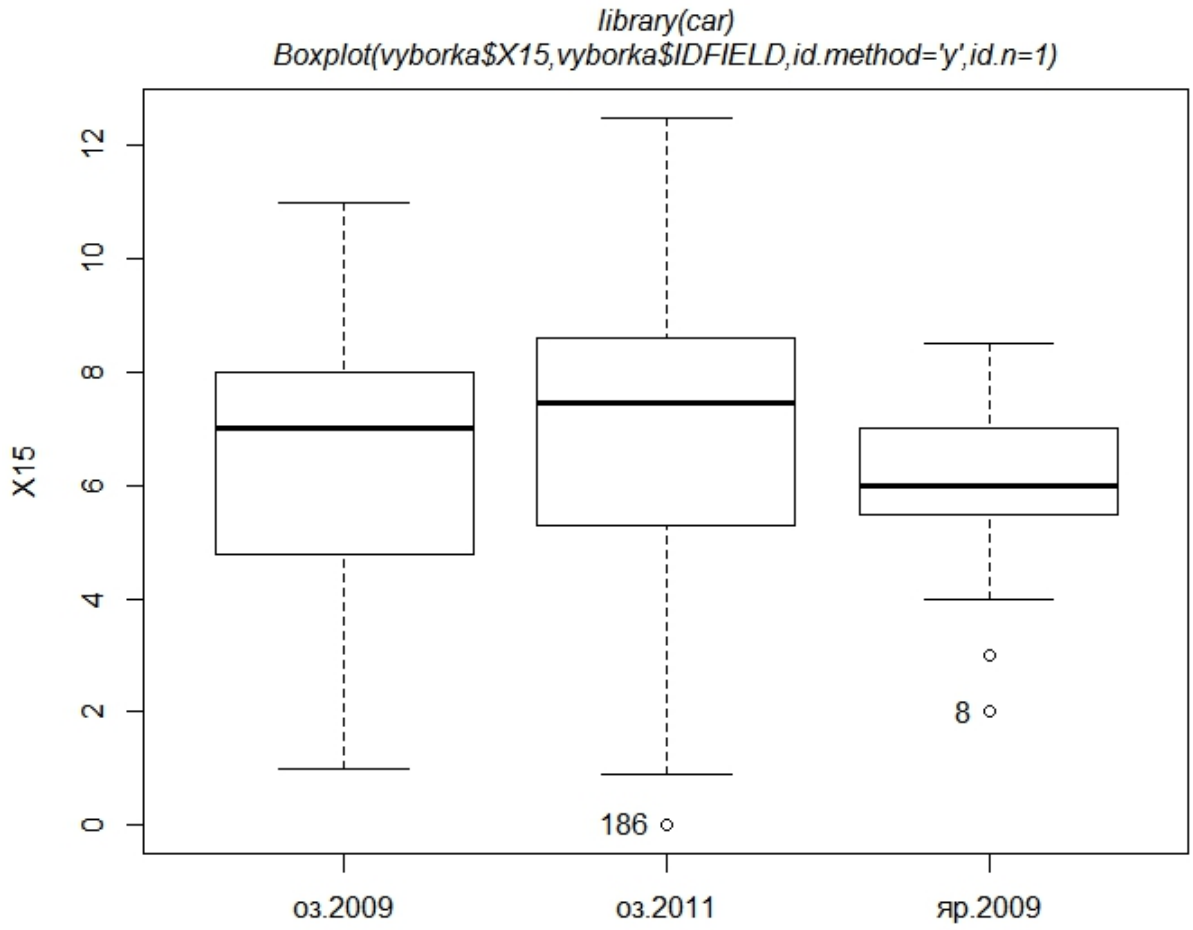

Рис. 1. Групповая диаграмма размахов признака X15

К формальным тестам на нормальность распределения относятся тест Пирсона (функция pearson.test()), тест Шапиро–Уилка (функция shapiro.test()), тест Колмогорова–Смирнова (функция lillie.test ()) и др.

Когда цель анализа заключается в нахождении переменных (предикторов), связанных со значениями зависимой переменной, важным этапом разведочного анализа данных является обнаружение *коллинеарности*. Под коллинеарностью понимают наличие линейной зависимости между двумя предикторами. В задачах с несколькими предикторами (например, при выполнении множественного регрессионного анализа) говорят также о мультиколлинеарности, т. е. наличии линейной зависимости сразу между несколькими переменными. Традиционным способом оценки коллинеарности является анализ корреляционной матрицы (функция corr.test() в R и ее графический вариант corrplot()). Пример графика корреляционной матрицы приведен на рис. 3.

На этапе разведочного анализа данных также важно исследовать *характер взаимодействия* между анализируемыми переменными, связь между которыми может быть нелинейная. Обнаруженные на этом этапе закономерности будут определять выбор статистической модели для описания данных. Удобным инструментом при анализе нескольких количественных переменных являются матричные диаграммы рассеяния (R-функция pairs()). Пример графика матричной диаграммы рассеяния приведен на рис. 4.

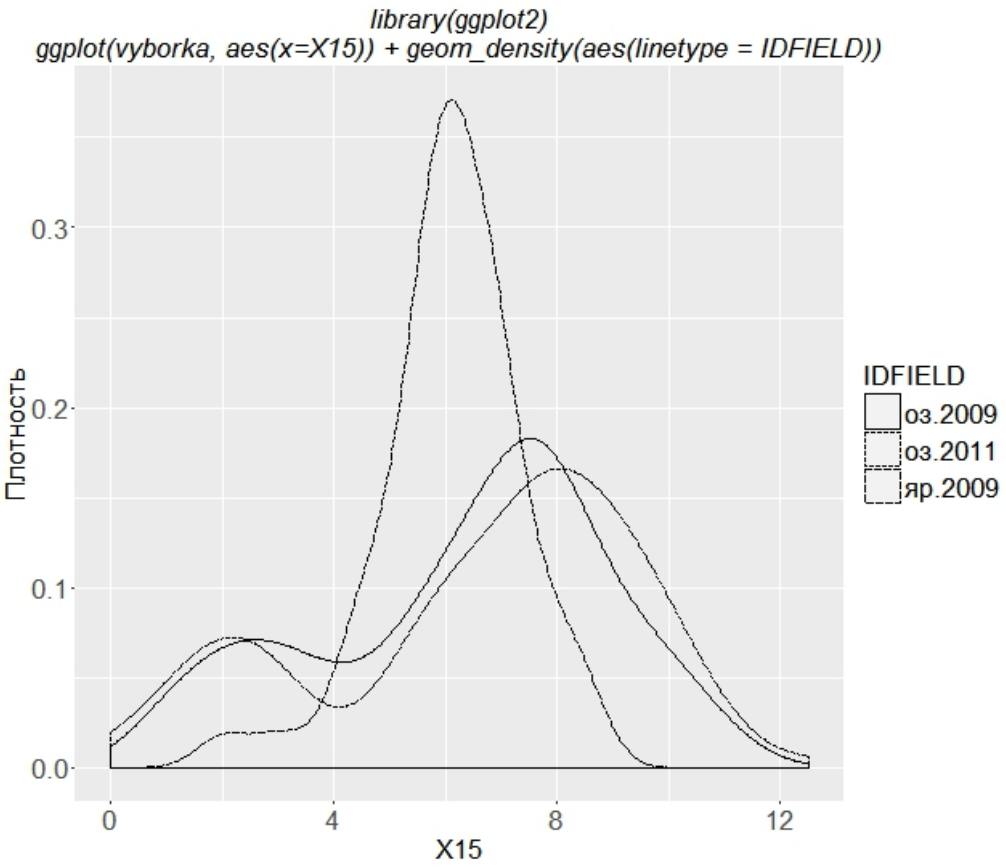

Рис. 2. График ядерной оценки плотности признака X15

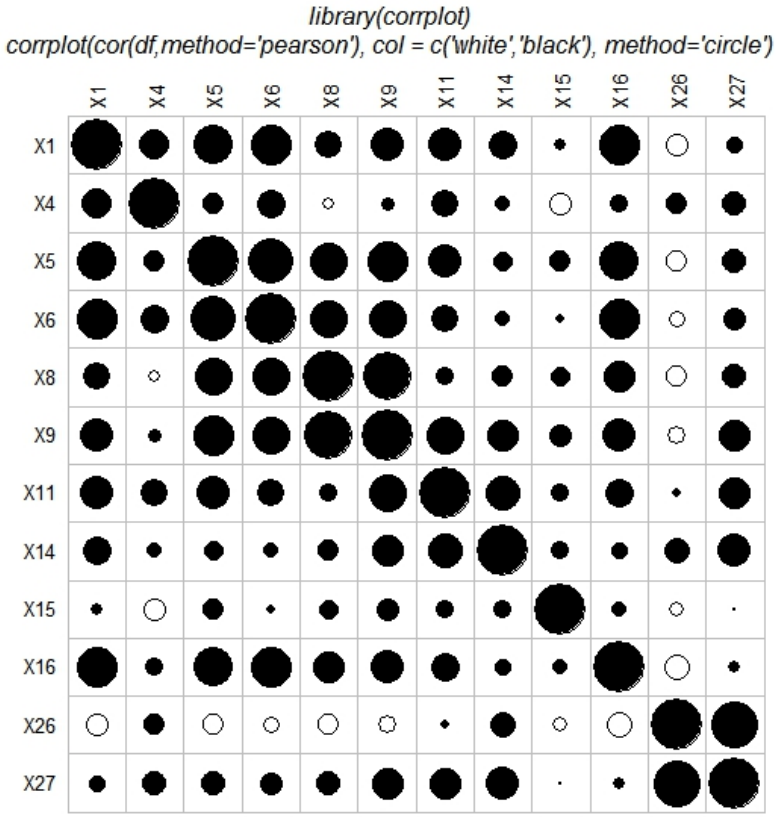

Рис. 3. График корреляционной матрицы для объединенных данных

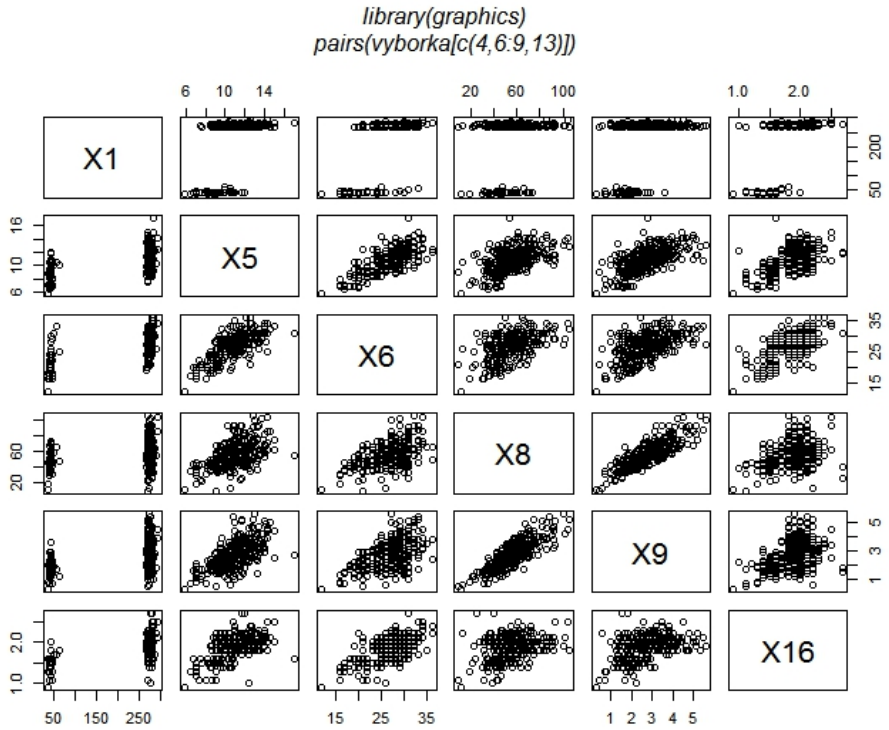

Рис. 4. Матричный график диаграмм рассеяния для объединенных данных

Разведочный анализ, проведенный по имеющимся данным полевых опытов, позволил выявить, что большинство исследуемых признаков не имеет точек выбросов либо единичные выбросы незначительно отклоняются от общей тенденции. Гипотеза нормальности распределения не подтверждается для некоторых из признаков, следовательно, к этим выборкам не следует напрямую применять статистические методы, необходимым условием которых является нормальность распределения данных. По признаку “длина остей” данные разбиваются на две группы, что следует учитывать при дальнейшем анализе.

Анализ корреляционных матриц и диаграмм рассеяния показал, что сильная линейная зависимость наблюдается между признаками “число продуктивных побегов” и “общий вес зерен”, зависимость средней силы наблюдается между признаками “вес зерен с колоса” и “масса 1000 зерен”, а признаки “длина колоса”, “число колосков в колосе”, “число зерен в колосе” и “вес зерен с колоса” образуют корреляционную группу.

2. Кластерный анализ данных

Селекционная работа начинается, как правило, с изучения многочисленных коллекций и выделения лучшего исходного материала. Тщательный подбор родительских форм для гибридизации во многом определяет результативность рекомбинационной селекции растений.

Работу по гибридизации можно сделать более целенаправленной и рациональной, если предварительно классифицировать образцы селекционных коллекций, разделив их на несколько отличающихся друг от друга групп. Отдельные представители групп и могут быть использованы в скрещиваниях.

Для выделения групп, сходных по комплексу признаков, применяются методы кластерного анализа данных [4].

Для кластеризации использована одна из агломеративных иерархических процедур – метод Уорда с евклидовым расстоянием в качестве меры сходства [5]. В этом методе кластерного анализа объединение двух классов минимизирует приращение общей дисперсии и приводит к формированию дендрита, разрезание которого на выбранном уровне и позволяет выделить группы. Результат кластеризации в этом случае представляется в виде дендрограммы. В R для ее построения используют функцию hclust().

На рис. 5 представлен результат кластеризации 51 образца яровой тритикале по комплексу из 12 количественных признаков. Анализ полученной дендрограммы показал, что предположительно в коллекции можно выделить три кластера.

После построения дендрограммы требуется определить количество кластеров, которые следует оставить, т. е. уровень обрезания “дерева”. Как правило, для этого используется эвристический подход. Можно также графически изобразить число получаемых из иерархического дерева кластеров как функцию коэффициента слияния и найти значимые “скачки” значения коэффициента. Скачок означает, что объединяются два довольно несхожих кластера. Таким образом, число кластеров, предшествующее этому объединению, является наиболее вероятным решением. Для построения графика коэффициентов слияния в R была использована функция plot().

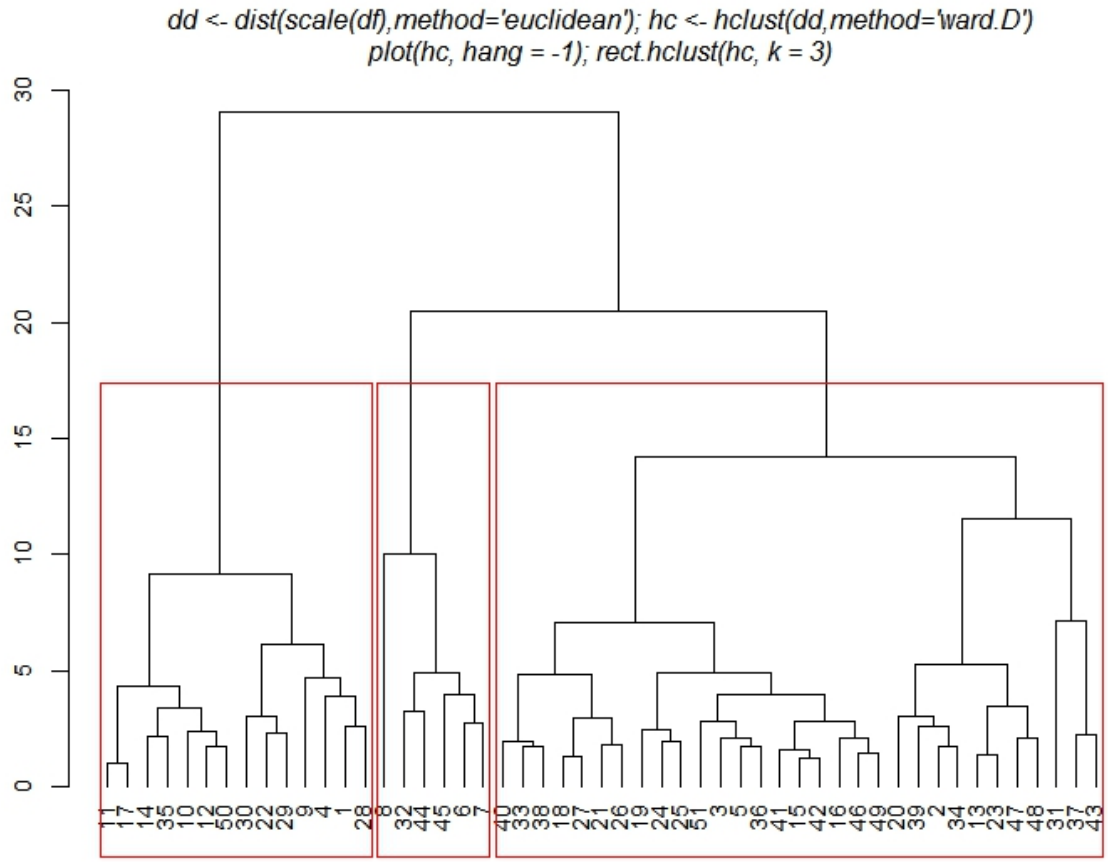

Рис. 5. Дендрограмма кластерного разбиения коллекции образцов яровых тритикале 2009 г. по комплексу признаков

Следует отметить, что кластерный анализ может “навязать” данным структуру, которой у них на самом деле нет. Поэтому необходимо проверять, насколько “хороши” получившиеся кластеры. Существуют бутстреп-методы, позволяющие оценить степень повторяемости кластерного решения в серии наборов данных [6]. Если для различных выборок из одной и той же генеральной совокупности получается одинаковое кластерное решение, то делается вывод, что это решение присуще всей совокупности. В программе R такую проверку позволяет сделать функция pvclust() из одноименного пакета.

На рис. 6 представлен результат проверки устойчивости кластерного разбиения бутстреп-методом. Над каждым узлом печатаются p-значения (au), связанные с устойчивостью кластеров в процессе репликации исходных данных. Значения 70–80 % считаются “хорошими”. В нашем случае мы видим “хорошую” устойчивость трех основных кластеров.

Графическое представление полученной кластеризации в пространстве главных компонент (рис. 7) позволило визуально оценить качество кластерного решения. Из рисунка видно, что полученные три кластера не пересекаются, но у них нет четких границ и они расположены близко друг к другу.

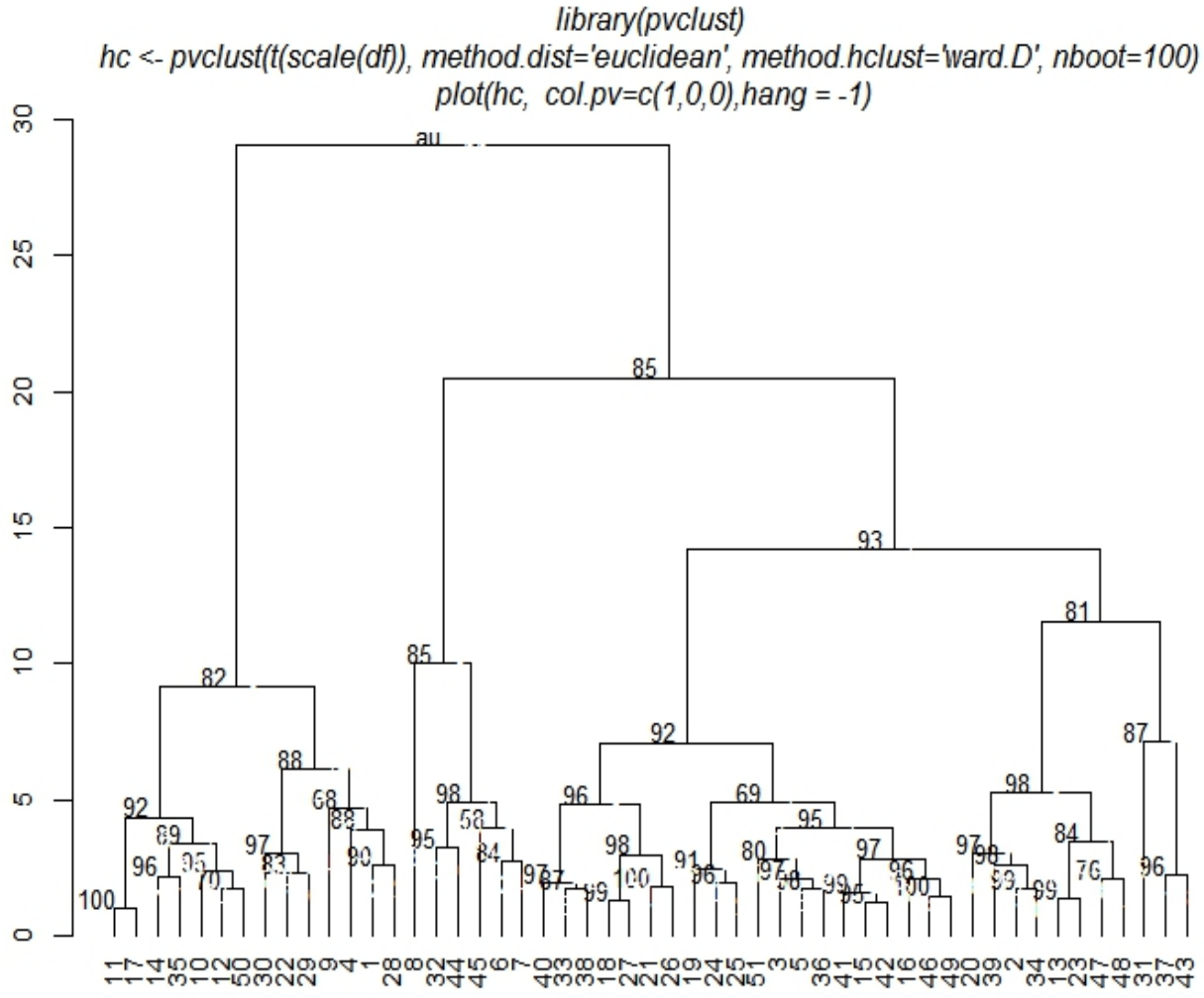

Рис. 6. Результаты проверки устойчивости кластерного решения для яровых тритикале 2009 г.

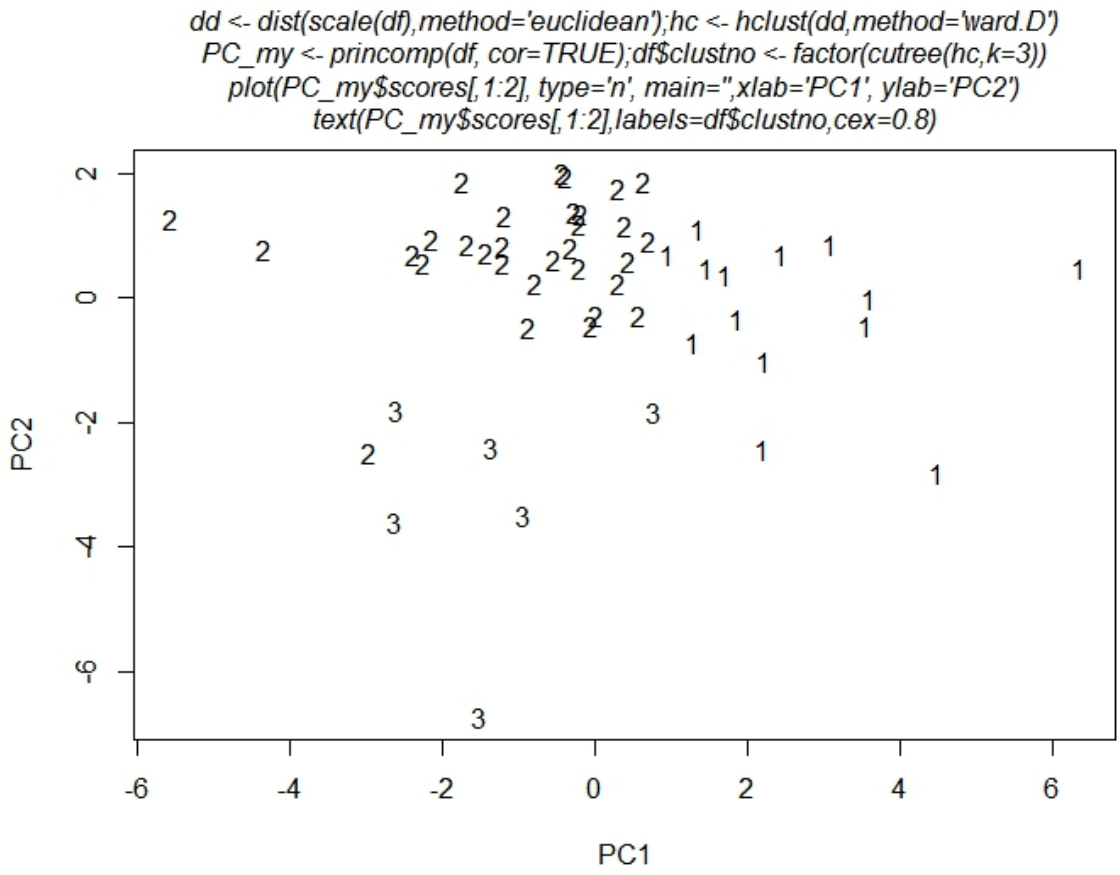

Рис. 7. Кластерное разбиение в пространстве главных компонент для яровых тритикале 2009 г.

Аналогичные данные получены и для двух других коллекций. По результатам кластерного анализа коллекционные образцы разделены на три группы, различающиеся по комплексу признаков таким образом, чтобы различие между группами было больше, чем внутри групп. Выбирая для скрещивания сорта из разных кластеров, можно добиться большего генетического разнообразия.

3. Анализ сопряженной изменчивости признаков методом главных компонент

В селекционных исследованиях необходимо учитывать корреляционные зависимости и закономерности сопряженной изменчивости признаков для определения среди них ключевых, по которым можно вести отбор с прогнозом эффекта по другим показателям. Метод главных компонент позволяет выделять корреляционные плеяды тесно связанных признаков, что весьма важно для правильного выбора исходного материала и проведения отбора по комплексу признаков [7].

В результате проведенного компонентного анализа данных полевых опытов установлено, что изменчивость количественных признаков тритикале определяется тремя–четырьмя главными компонентами, на долю которых приходится от 70 до 80 % суммарной дисперсии [8]. Во всех трех опытах в первую компоненту с относительно большими коэффициентами нагрузки вошли коррелирующие

между собой признаки X5 – длина колоса, X6 – число колосков в колосе, X8 – число зерен в колосе, X9 – масса зерен колоса, что дало основание интерпретировать первую компоненту как “продуктивность колоса”.

Для облегчения выявления корреляционных плеяд признаков построены графики распределения признаков в пространстве первых двух главных компонент для каждой из коллекций и для объединенных данных (рис. 8). Из рисунка видно, что признаки X5, X6, X8, X9 расположены близко друг к другу.

Проведенный кластерный анализ подтвердил результаты компонентного анализа (рис. 9). Исследуемые признаки сформировали три кластера. Для всех трех опытов в первый кластер вошли все те же признаки продуктивности колоса – X5, X6, X8, X9.

В результате анализа сопряженной изменчивости 12 количественных признаков тритикале выявлено 8 значимых корреляционных связей. Установлено, что для тритикале наблюдается стабильная взаимосвязь признаков: длина колоса, число колосков в колосе, число зерен в колосе и масса зерен колоса, образующих корреляционную группу. Они определяют основную долю совокупной изменчивости комплекса признаков, любой из них можно использовать в качестве ключевого и вести с его помощью отбор при селекции тритикале с прогнозом эффекта по остальным признакам из этой корреляционной плеяды.

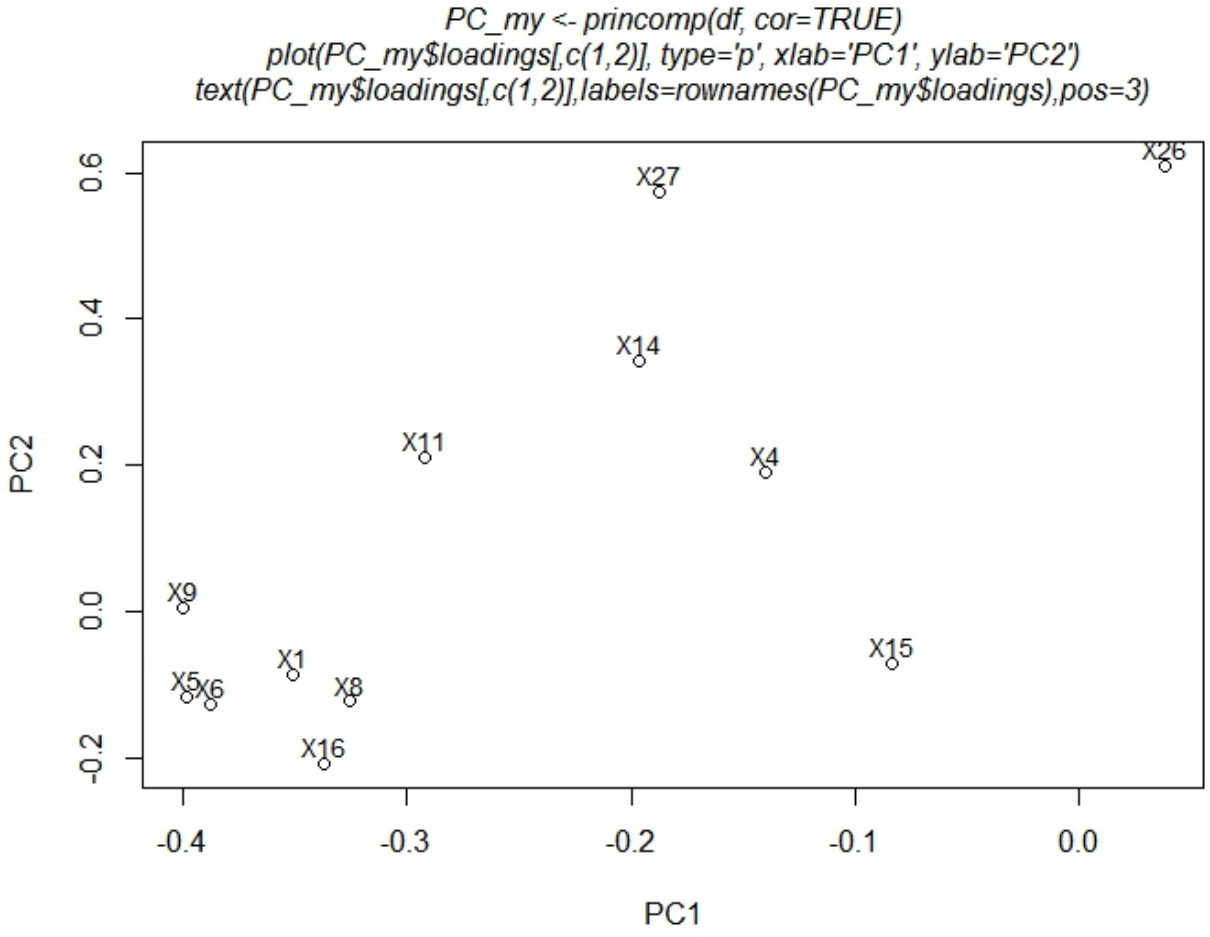

Рис. 8. Распределение комплекса признаков на плоскости главных компонент для объединенных данных

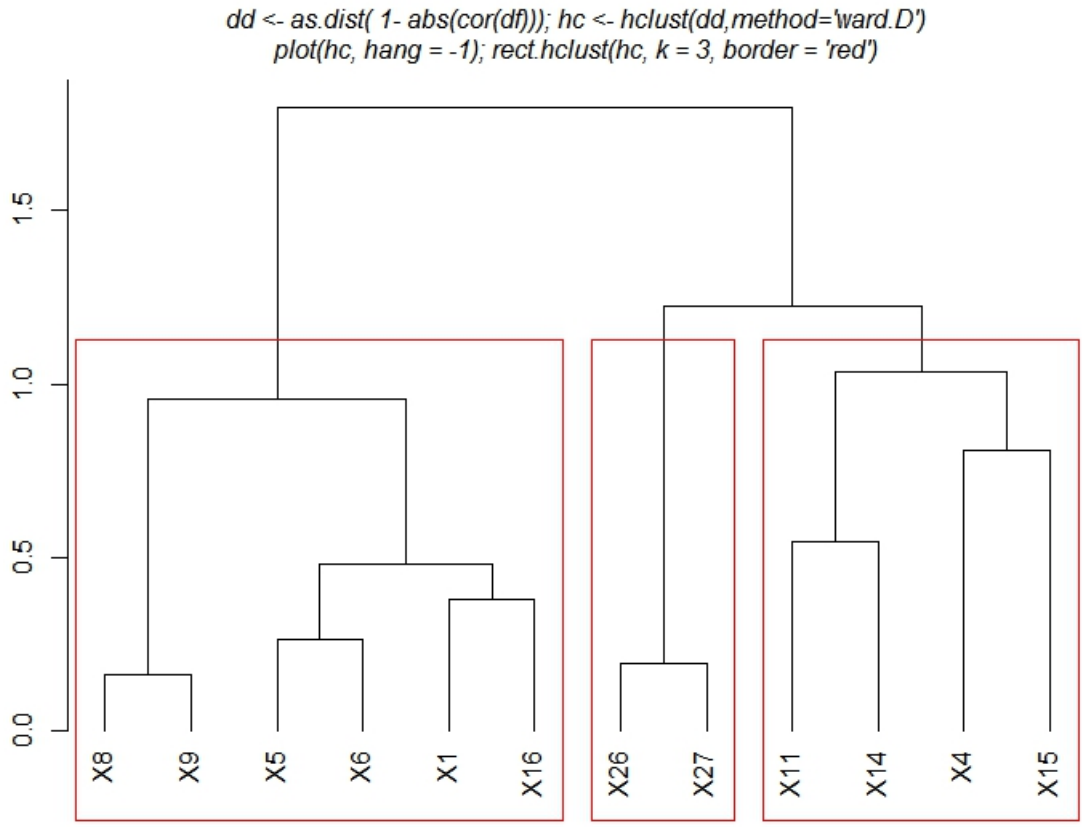

Рис. 9. Дендрограмма кластерного анализа признаков для объединенных данных

Таким образом, использование возможностей программной среды R позволило провести разведочный, кластерный и компонентный анализ данных и наглядно продемонстрировать их результаты.

Благодарности. Статья опубликована при финансовой поддержке РФФИ (грант № 16-07-20001).

Список литературы / References

[1] **Мастицкий С.Э., Шитиков В.К.** Статистический анализ и визуализация данных с помощью R. Адрес доступа: http://r-analytics.blogspot.com (дата обращения 26.09.2016).

Mastitskiy, S.E., Shitikov, V.K. Statistical analysis and data visualization with R. Available at: http://r-analytics.blogspot.com (accessed 26.09.2016). (In Russ.)

[2] **Zuur, A.F., Ieno, E.N., Elphick, C.S.** A protocol for data exploration to avoid common statistical problems // Methods in Ecology and Evolution. 2009. No. 1. P. 3–14.

[3] **Chang, W.** R Graphics Cookbook. O'Reilly Media, 2012. 413 p.

[4] **Ефимов В.М., Ковалева В.Ю.** Многомерный анализ биологических данных: Учеб. пособие. СПб.: ВИЗР, 2008. 98 с.

Ephimov, V.M., Kovaleva, V.Yu. Textbook on multivariate analysis of biological data: schoolbook. St-Peterburg: VIZR, 2008. 98 p. (In Russ.)

[5] **Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р., Олдендерфер М.С., Блэшфилд Р.К.** Факторный, дискриминантный и кластерный анализ: пер. с англ. М.: Финансы и статистика, 1989. 215 с.

Kim, J.-O., Mueller, C.W., Klecka, W.R., Aldenderfer, M.S., Blashfield, R.K. Factor, discriminant and cluster analysis. M.: Finance and Statistics, 1989. 215 p. (In Russ.)

[6] **Шитиков В.К., Розенберг Г.С.** Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. Тольятти: Кассандра, 2013. 314 с.

Shitikov, V.K., Rozenberg, G.S. Randomization and bootstrap: statistical analysis in biology and ecology using R. Tol'yatti: Kassandra, 2013. 314 p. (In Russ.)

[7] **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. М.: Изд-во МСХА, 1992. 269 с.

Smiryaev, A.V., Martynov, S.P., Kil'chevskiy, A.V. Biometrics in genetics and plant breeding. Moscow: MSKhA, 1992. 269 p. (In Russ.)

[8] **Чешкова А.Ф., Алейников А.Ф., Степочкин П.И.** Анализ сопряженной изменчивости количественных признаков тритикале // Достижения науки и техники АПК. 2016. Т. 30, № 5. С. 50–52.

Cheshkova, A.F., Aleynikov, A.F., Stepochkin, P.I. Analysis of covariation of quantitative characters of triticale // Achievements of Science and Technology of AIC. 2016. Vol. 30, No. 5. P. 50–52. (In Russ.)

Поступила в редакцию 20 октября 2016 г.

Application of graphical features of the R programming environment for analysis of experimental data on the breeding of triticale

CHESHKOVA, ANNA F.[1,*], ALEYNIKOV, ALEXANDR F.[1,2], STEPOCHKIN, PETR I.[3]

[1] Siberian Federal Scientifice Center of Agro-BioTechnologies, Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, 630501, Russia

[2] Novosibirsk State Technical University, 630092, Russia

[3] Siberian Institute of Plant Growing and Breeding – Branch of the Institute of Cytology and Genetics, Krasnoobsk, Novosibirsk region, 630501, Russia

* Corresponding author: Cheshkova, Anna F., e-mail: anna.cheshkova@sorashn.ru

Purpose. The paper describes an application of the R environment for visualization and statistical analysis in breeding studies. The material for research includes the data of field experiments conducted by the GNU SibNIIRS during spring and winter triticale world VIR collection samples in 2009 (51 samples of spring triticale and 103 samples of winter one) and 2011 (120 samples of winter triticale). 12 morphological characters were taken for evaluation of the samples. The studies were conducted in order to define groups of triticale samples that differ from each other on a range of traits, as well as to detect regularities in correlated variability of triticale quantitative traits.

Methodology. We used standard statistical methods of multivariate data analysis (Pearson's method, principal component analysis, cluster analysis by Ward method), implemented in the R software environment.

Findings. First, the data exploration was fulfilled to check the necessary conditions for applying the presented adequate statistical techniques. The protocol of data exploration includes: detecting outliers and collinearity, testing on homogeneity of variance and normality, revelation of the relationships character between variables.

Then we applied cluster analysis technique to classify triticale samples and to divide them into several groups. By choosing samples from different clusters for hybridization it is possible to achieve greater genetic diversity. The effectiveness of the division to clusters was checked by bootstrap methods.

Finally. the primary components analysis (PCA) was carried on to identify correlated variability of triticale quantitative traits. It was revealed that the variability of quantitative traits of triticale was determined by 3–4 main components, which accounted for 70 to 80 % of the total variance. The first component includes with large loading coefficients such traits as “length of spike”, “number of spikelet per ear”, “number of grains per spike” and “spike grains weight” that form a correlation group, which gives the opportunity to interpret this component as “spike productivity”.

Conclusion. Use of the R software environment allowed to carry out data exploration, cluster and component analysis and to demonstrate their results.

Keywords: R environment, statistical analysis, triticale, breeding.

Acknowledgements. The article publication was supported by RFBR (grant No. 16-07-20001).

Received 20 October 2016