ТЕХНИЧЕСКИЕ НАУКИ

TECHNICAL SCIENCES

Научная статья
УДК 581.45:519.23.001.33
ББК 28.56в63
А 45
DOI: 10.53598/2410-3225-2024-1-336-26-35

Алгоритм классификации популяций растений по морфологическим характеристикам листьев

(*Рецензирована*)

Марат Вячеславович Алиев[1], Ирина Владимировна Чернявская[2], Евгения Михайловна Еднич[3], Андрей Григорьевич Лобанов[4], Мирослав Игоревич Хизик[5], Александр Юрьевич Кузнецов[6]

[1–6] *Адыгейский государственный университет, Майкоп, Россия*
[1] *alievmarat@mail.ru*
[2] *chernyav.iv@mail.ru*
[3] *ednich@mail.ru*
[4] *andrey.gl.pro@gmail.com*
[5] *miroslavhizik@gmail.com*
[6] *sk.sasha64@gmail.com*

Аннотация. *Предлагается алгоритм для определения различий между популяциями с использованием заданного набора морфологических характеристик листьев и выявления закономерностей между обнаруженными различиями и условиями среды, в которых популяции произрастают.*

Ключевые слова: *морфологические характеристики листа, дисперсионный анализ, кластерный анализ, критерий Фишера, метод главных компонент*

Для цитирования: *Алгоритм классификации популяций растений по морфологическим характеристикам листьев / М. В. Алиев, И. В. Чернявская, Е. М. Еднич, А. Г. Лобанов, М. И. Хизик, А. Ю. Кузнецов // Вестник Адыгейского государственного университета. Сер. : Естественно-математические и технические науки. 2024. Вып. 1 (336). С. 26–35. DOI: 10.53598/2410-3225-2024-1-336-26-35*

Original Research Paper

Classification algorithm of plants populations under morphological characteristics of leaves

Marat V. Aliev[1], Irina V. Chernyavskaya[2], Evgeniya M. Ednich[3], Andrey G. Lobanov[4], Miroslav I. Khizik[5], Aleksandr Yu. Kuznetsov[6]

[1–6] *Adyghe State University, Maykop, Russia*
[1] *alievmarat@mail.ru*
[2] *chernyav.iv@mail.ru*
[3] *ednich@mail.ru*
[4] *andrey.gl.pro@gmail.com*
[5] *miroslavhizik@gmail.com*
[6] *sk.sasha64@gmail.com*

Abstract. *The algorithm for definition of distinctions between populations with use of the set of morphological characteristics of leaves and detecting of laws between the distinctions found and con-*

ditions of environment in which populations grow is proposed.

Keywords: *morphological characteristics of the leaf, analysis of variance, cluster analysis, Fisher's criterion, principal component method*

For citation: *Algorithm of classification of populations of plants under morphological characteristics of leaves / M. V. Aliev, I. V. Chernyavskaya, E. M. Ednich, A. G. Lobanov, M. I. Khizik, A. Yu. Kuznetsov // The Bulletin of the Adyghe State University. Ser. : Natural-Mathematical and Technical Sciences. 2024. Iss. 1* (*336*)*. P. 26–35. DOI: 10.53598/2410-3225-2024-1-336-26-35*

Введение

Морфологические характеристики листьев растений, такие как площадь листа, ширина и длина листа, флуктуирующая асимметрия и т. д., хорошо отражают различия как между видами растений [1, 2], так и изменчивость, обусловленную экологическими факторами (климатогеографическими, техногенными и т. д.) окружающей среды [3–5]. Определение различий и сходств между популяциями растений на основе морфологических характеристик листьев – сложная и трудоемкая задача. Она важна для понимания как адаптации растений, так и генетической основы морфологических характеристик в популяции. В данной статье предлагается алгоритм выявления морфологических параметров, по которым различаются популяции растений, и определения связи найденных различий с факторами среды произрастания, основанный на дисперсионном анализе, методе главных компонент, кластерном анализе и регрессионной модели, для выявления зависимостей между морфологическими характеристиками листа и факторами окружающей среды.

Алгоритм

Предлагаем следующий алгоритм для нахождения морфологических различий листьев растений различных популяций и выявления факторов среды произрастания, объясняющих данные различия, состоящий из следующих шагов:

Шаг 1. Сбор и обработка данных, фиксация измеримых морфологических характеристик;

Шаг 2. Проверка существования различий между популяциями;

Шаг 3. Выделение морфологических характеристик, по которым можно различать популяции;

Шаг 4. Интерпретация результатов: выявления закономерностей.

Входные данные представляют собой набор таблиц, каждая из которых соответствует определенной популяции. Строки таблицы соответствуют отдельным листьям, а столбцы – морфологическим характеристикам листьев.

Шаг 1. Сбор и обработка данных – отбор образцов листьев растений из различных популяций и измерение их морфологических характеристик, таких как форма листа, размер, цвет, жилкование, опушение и т. д., с последующей стандартизацией и нормализацией полученных данных для обеспечения их сопоставимости.

Шаг 2. Применение дисперсионного анализа для проверки существования различий между популяциями по морфологическим характеристикам листьев [6]. Для проверки гипотезы будем использовать критерий Фишера, заданный следующим образом:

$$F = \frac{S_1^2}{S_2^2},$$

где S_1^2 – выборочная межгрупповая дисперсия, а S_2^2 – выборочная внутригрупповая дисперсия.

В качестве нулевой гипотезы принимается утверждение о том, что различия между группами по средним значениям наблюдений отсутствуют. Если гипотеза верна, то F имеет распределение Фишера со степенями свободы $k_1 = m - 1$ и $k_2 = mn - m$, где

m − число групп, n − число наблюдений в каждой группе. Соответственно, если вероятность (p-значение) получить фактическое значение F ниже уровня значимости, то гипотеза отвергается, т. е. между группами имеются различия по средним значениям.

Шаг 3. Если по каким-либо морфологическим характеристикам между популяциями в результате дисперсионного анализа обнаружено различие, то далее возможно выявление морфологических различий уже между отдельными популяциями и выявление морфологических характеристик, которые вносят наибольший вклад в эти различия. Для этих целей предлагается использовать метод главных компонент и кластерный анализ [7, 8].

Метод главных компонент основан на создании линейных комбинаций изначальных переменных. Каждая такая линейная комбинация (компонента) выбирается так, чтобы максимизировать дисперсию при проекции данных на нее. Соответственно, наибольшая дисперсия будет у первой компоненты, вторая по величине – у второй и т. д. В [3] для всех популяций строилась диаграмма рассеяния по значениям первых двух компонент, на которой возможно обнаружить визуальное отличие одних групп популяций от других.

Далее строится матрица корреляций найденных компонент с изначальными переменными. Если на диаграмме наблюдается различие между популяциями по какой-либо из компонент, то в таблице находятся морфологические характеристики листа, имеющие наибольшую корреляцию с данной компонентой. Таким образом делается вывод, что именно эти морфологические характеристики вносят наибольший вклад в морфологические различия популяций.

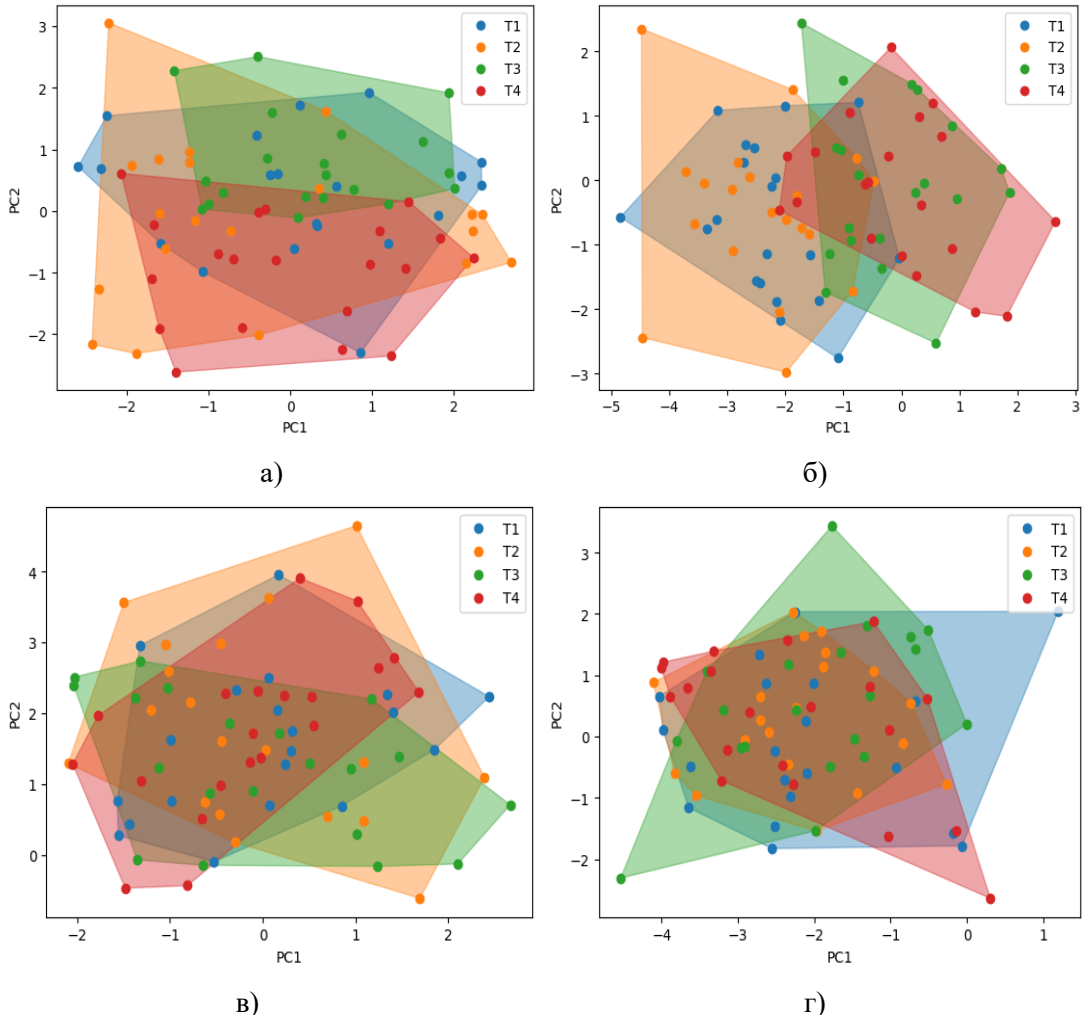
К примеру, на диаграмме, изображенной на рисунке 1а), наблюдается различие между популяциями T3 и T4 – они различаются по значениям второй компоненты. На рисунке 1б) – различие между T1, T2 и T3, T4 по первой компоненте. Однако на рисунке 1в) и рисунке 1г) все популяции сосредоточены в одной области диаграммы и визуальных различий между ними не наблюдается. В таком случае после построения матрицы корреляций невозможно сказать, какие именно морфологические характеристики оказывают влияние на различия, обнаруженные на шаге 2.

Так как на рисунке 1а) по второй компоненте обнаружены различия, то для выявления морфологических характеристик, их обуславливающих, строится таблица корреляций найденных компонент с изначальными переменными. На рисунке 2а) изображена данная таблица, и на основе нее делается вывод, что морфологические характеристики L3, L4 и L7 вносят наибольший вклад в различия популяций T3 и T4, так как они имеют наибольшую корреляцию со второй компонентой.

Также для определения морфологических свойств и различий используется кластерный анализ, а именно иерархическая агломеративная кластеризация [3]. Так как после использования метода главных компонент выводы делаются на основе визуального анализа полученной диаграммы, не подкрепленного строгими математическими критериями, кластерный анализ служит более надежным методом выявления различий между популяциями.

Так как каждая популяция представлена матрицей, то для проведения кластеризации можно усреднить для каждой популяции значения каждой из морфологических характеристик, т. е. представить ее в виде вектора, размерность которого соответствует количеству морфологических характеристик. Связано это с тем, что в существующей литературе не удалось найти примеров применения кластерного анализа к набору экземпляров, которые были бы заранее объединены в группы.

Алгоритм кластерного анализа имеет два основных параметра – метрика расстояния и вид связи. В качестве метрики расстояния обычно используется евклидово расстояние. Для вида связи чаще всего используются следующие метрики [9].

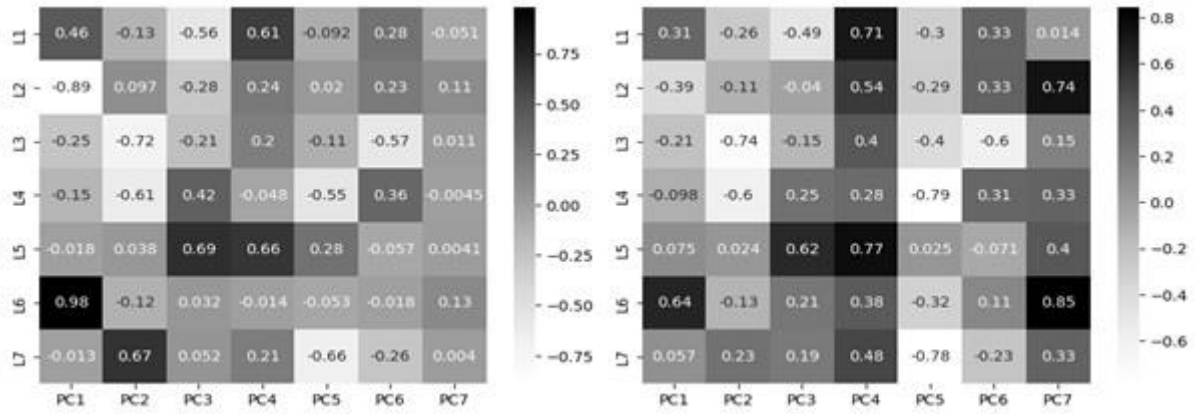


а) б)

в) г)

Рис. 1. а), б) – диаграммы, на которых обнаружены визуальное различие между популяциями; в), г) – примеры диаграмм, на которых различий не обнаружены

Fig. 1. a), b) – diagrams on which visual differences between populations are found; c), d) – examples of diagrams on which differences are not found



а) б)

Рис. 2. Матрица корреляций компонент с морфологическими характеристиками листьев: а) для диаграммы на рисунке 1а); б) для диаграммы на рисунке 1б), где ни одна из компонент не показывает существенного различия между популяциями

Fig. 2. A matrix of correlations of components with morphological characteristics of leaves: a) for the diagram in figure 1a); b) for the diagram in figure 1b), where none of the components shows a significant difference between populations

Метод одиночной связи [9] – за расстояние между кластерами принимается минимальное расстояние между принадлежащими им объектами. В таком случае даже две непохожие популяции, имея всего по одному листу, близких друг к другу, будут объединены в один кластер, поэтому данный метод плохо подходит для выявления различий между популяциями;

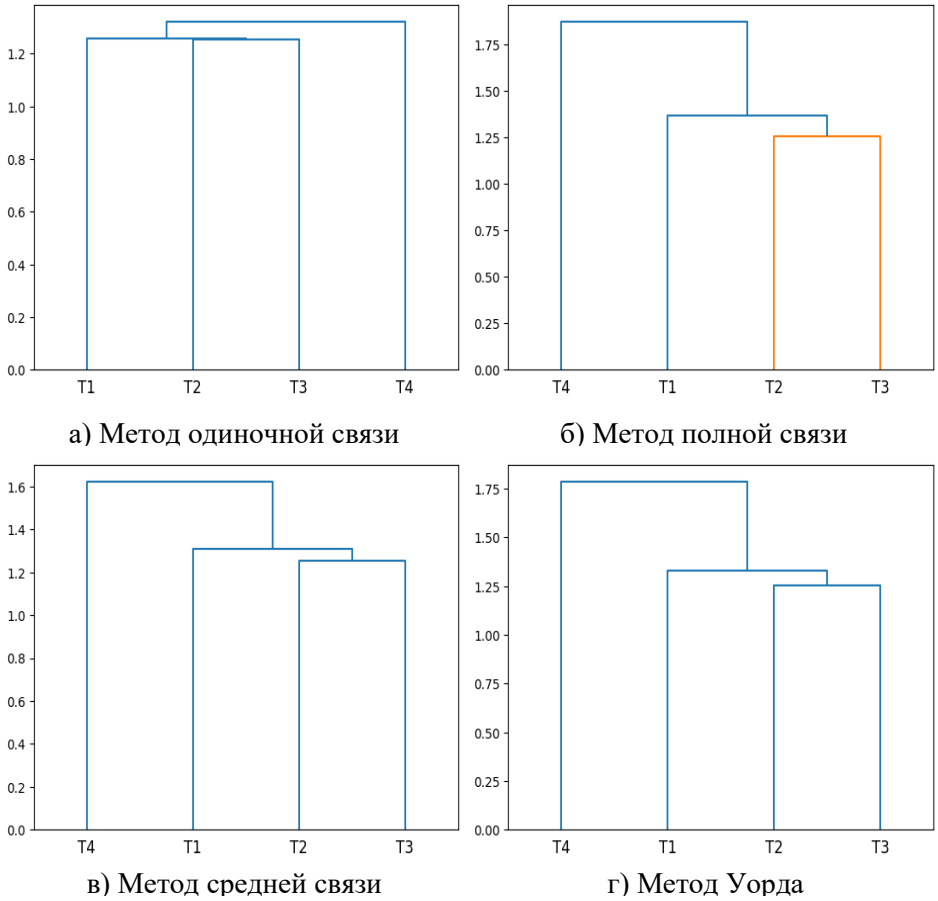
Метод полной связи [9] – за расстояние между кластерами принимается максимальное расстояние между принадлежащими им объектами;

Метод средней связи [9] – за расстояние между кластерами принимается среднее расстояние между принадлежащими им объектами;

Центроидный метод [9] – за расстояние между экземплярами принимается квадрат расстояния между их центроидами – «усредненными» по всему кластеру объектами. При использовании этого метода отсекаются листы, чьи морфологические характеристики сильно отличаются от средних у данной популяции, т. е. таким образом удаляется «шум»;

Метод Уорда [9, 10] – за расстояние между кластерами принимается квадрат расстояния между их центроидами, умноженный на отношение произведения числа объектов в кластерах к их сумме. Такой метод позволяет минимизировать дисперсию в получаемых кластерах.

На рисунке 3а) видно, что при использовании метода одиночной связи расстояние между полученными кластерами крайне мало по сравнению с другими типами связи, соответственно и выявление различий между популяциями достаточно слабое. На рисунке 3б) также заметно, что метод полной связи обеспечивает более существенное разделение популяций, нежели метод средней связи или метод Уорда.



а) Метод одиночной связи

б) Метод полной связи

в) Метод средней связи

г) Метод Уорда

Рис. 3. Возможные результаты кластерного анализа при разных типах связи

Fig. 3. Possible results of cluster analysis for different types of communication

Если популяции, между которыми посредством анализа главных компонент выявлены морфологические различия, находятся в разных кластерах, то можно сделать заключение о том, что кластерный анализ подтвердил вывод, полученный при анализе диаграммы для метода главных компонент.

Шаг 4. Интерпретация результатов: анализ и интерпретация полученных результатов для выявления закономерностей. Для этой цели будет использоваться линейная регрессия [11]. Уравнение линейной регрессии выглядит следующим образом:

$$y = w^T x,$$

где $w \in R^m$ – оцениваемые параметры модели;

m – количество признаков;

R – вектор признаков, т. е. в нашем случае – морфологических характеристик листов;

$\epsilon \in R$ – случайный шум с нулевым математическим ожиданием и конечной дисперсией;

$y \in R$ – зависимая переменная.

Для каждой независимой переменной x_i в качестве нулевой гипотезы принимается предположение о том, что ее коэффициент w_i равен 0, т. е. данная переменная не оказывает никакого влияния на зависимую. Соответственно, если *p*-значение мало, то независимая переменная оказывает существенное влияние на зависимую. В качестве зависимой переменной будет выступать одна из морфологических характеристик листа (т. е. для каждой морфологической характеристики листа регрессионный анализ проводится отдельно), а в качестве независимых − значения климатогеографических факторов. После проведения регрессионного анализа для каждой пары климатогеографического фактора и морфологической характеристики листа, *p*-значение которой меньше уровня значимости, делается вывод, что данный фактор среды произрастания оказывает существенное влияние на морфологическую характеристику листа.

Пример использования предложенной методики

Имеется 4 популяции листьев клена ясенелистного в 4-х местах произрастания – Майкоп, Ботанический сад АГУ, станица Даховская и станица Гиагинская. У листов были измерены следующие семь морфологических характеристик: максимальная длина листовой пластинки верхнего листочка (L1), максимальная ширина листовой пластинки верхнего листочка (L2), площадь листовой пластинки верхнего листочка (L3), толщина листовой пластинки верхнего листочка (L4), угол между центральной и боковой жилкой верхнего листочка (L5), индекс ксероморфизма (L6=L1/L2), флуктуирующая асимметрия (L7). Выборка составлена из 20 листьев каждой популяции. Также имеются следующие характеристики мест произрастания (табл. 1): географические координаты, высота над уровнем моря, средние значения температуры, количества осадков и влажность (%).

Таблица 1

Значения климатогеографических параметров для каждой популяции

Table 1. Values of climatic and geographical parameters for each population

	Широта	Долгота	Высота, н.у.м.	Температура	Осадки	Влажность, %
Майкоп	44° 36' 28"	40° 06' 20"	212	27	59,5	60
Бот. сад	44° 32' 04"	40° 06' 09"	238	24	83	62
Гиагинская	44° 52' 10"	40° 03' 58"	129	27	79,3	57
Даховская	44° 13' 58"	40° 11' 51"	570	23,5	94,4	62

В таблице 2 приведены *p*-значения для каждой морфологической характеристики листов, полученные в результате дисперсионного анализа.

Жирным шрифтом в таблице 2 выделены морфологические характеристики, имеющие *p*-значения ниже уровня значимости (p=0,05). Как видно из таблицы 2,

наиболее существенные различия между популяциями наблюдаются в морфологических характеристиках L3, L4, L5.

Таблица 2

P-значения, полученные по результатам дисперсионного анализа

Table 2. *P*-values obtained from the results of the analysis of variance

Морфологическая характеристика листа	L1	L2	L3	L4	L5	L6	L7
p-значение	0,30672	0,06351	**0,00061**	**0,00047**	**0,00240**	0,79210	0,08107

По результатам метода главных компонент построена диаграмма (рис. 4), на которой наблюдается различие между популяциями ст. Даховской и ст. Гиагинской по второй компоненте (ось PC2).

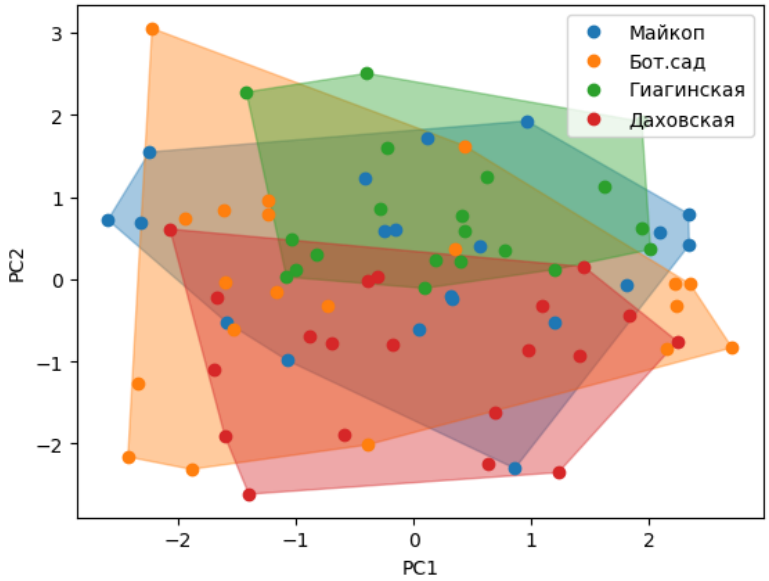


Рис. 4. Диаграмма рассеяния для первых двух компонент

Fig. 4. The scattering diagram for the first two components

Наибольшую корреляцию вторая компонента имеет с L7, L4 и L3 (рис. 5).

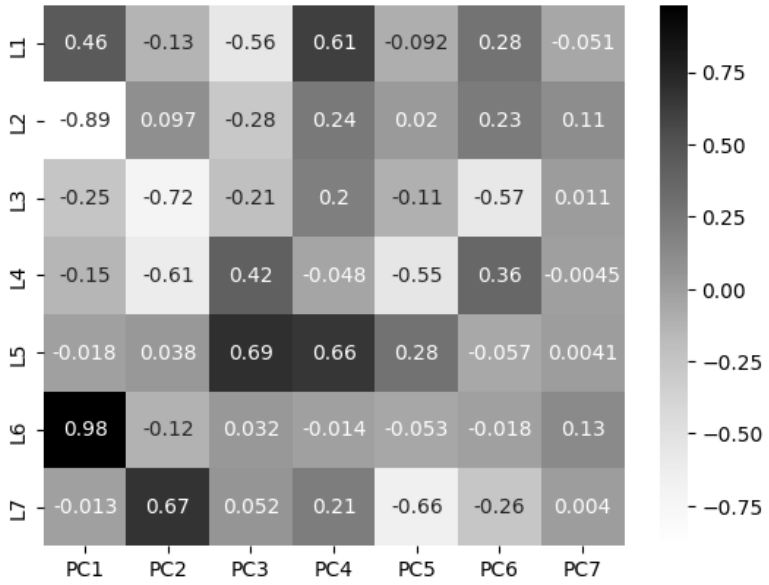


Рис. 5. Матрица корреляций главных компонент с изначальными переменными

Fig. 5. The matrix of correlations of the main components with the initial variables

Кластерный анализ с использованием полной связи подтверждает результаты метода главных компонент, т. к. популяции ст. Даховской и ст. Гиагинской находятся в разных кластерах (рис. 6).

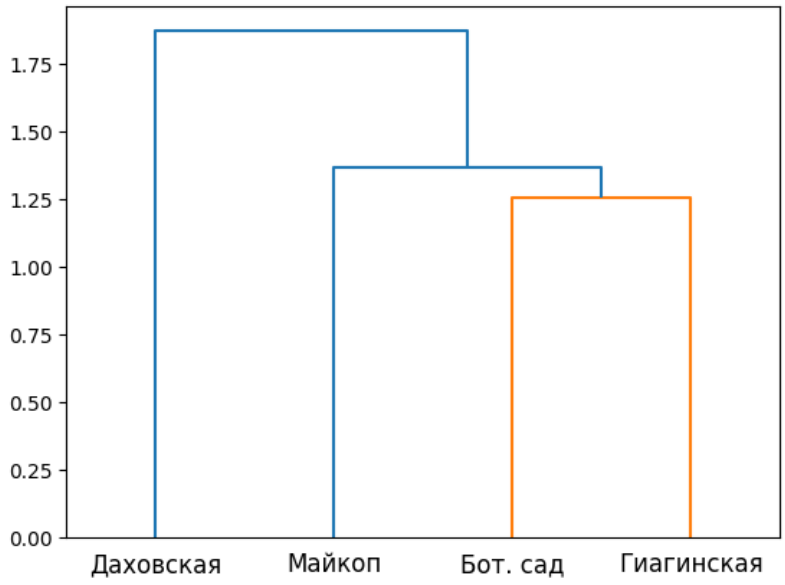


Рис. 6. Дендрограмма, построенная по результатам кластерного анализа

Fig. 6. A dendrogram based on the results of cluster analysis

В таблице 3 приведены *p*-значения, полученные в результате линейного регрессионного анализа (уровень значимости принят *p*=0,05). Как видно, высота оказывает влияние на L2, L4, осадки – на L5, влажность – на L1, L2 и L4. Таким образом, влиянием высоты и влажности на L4 можно объяснить найденные различия между популяциями ст. Гиагинской и ст. Даховской, которые имеют наибольшее различие по этим двум характеристикам.

Таблица 3

P-значения, полученные в результате линейной регрессии

Table 3. *P*-values obtained as a result of linear regression

	L1	L2	L3	L4	L5	L6	L7
Высота	0,11350	**0,02447**	0,62252	**0,01594**	0,08920	0,37181	0,28911
Температура	0,34268	0,11290	0,16239	0,69109	0,37678	0,12327	0,22163
Осадки	0,73129	0,71822	0,25999	0,19344	**0,00003**	0,81820	0,90930
Влажность	**0,00195**	**0,00170**	0,06188	**0,00450**	0,97685	0,94427	0,90908

Примечание. Жирным шрифтом выделены *p*-значения ниже уровня значимости (*p*=0,05).

Выводы

Получен универсальный алгоритм для нахождения различий морфологических характеристик листьев растений различных популяций и условий среды. Он позволяет выявить морфологические характеристики листьев, вносящих наибольший вклад в различия популяций, и выявить влияния условий среды на данные характеристики. Предложенный алгоритм был опробован на 4-х популяциях клена ясенелистного и позволил выделить различие между популяциями ст. Даховской и ст. Гиагинской, а также объяснить найденное различие разницы произрастания популяций по высотному градиенту и влажности в местах произрастания данных популяций.

Примечания

1. Оценка значимости морфологических признаков у культиваров Pinus Mugo Turra для их определения методом дисперсионного анализа / М. В. Симахин, В. А. Крючкова, А. В. Исачкин, А. М. Покиньчереда, В. Г. Донских, А. В. Евтюхова, Е. А. Козлова // Вестник Красноярского государственного аграрного университета. 2020. № 11 (164). С. 61–66. URL: https://cyberleninka.ru/article/n/otsenka-znachimosti-morfologicheskih-priznakov-u-kultivarov-pinus-mugoturra-dlya-ih-opredeleniya-metodom-dispersionnogo-analiza (дата обращения: 28.11.2023).

2. Кузьмичева Н. А. Таксономическая значимость морфологических признаков листа и побега восточноевропейских видов ив // Вестник фармации. 2008. № 4 (42). С. 12–22. URL: https://cyberleninka.ru/article/n/taksonomicheskaya-znachimost-morfologicheskih-priznakov-lista-i-pobega-vostochnoevropeyskih-vidov-iv (дата обращения: 17.12.2023).

3. Morphologic variability of the Acer campestre L. populations in Bosnia and Herzegovina / S. Kvesić, M. M. Hodžić, M. Čater, D. Ballian // Acta Biologica Sibirica. 2021. No. 7. P. 327–343.

4. Кузьмичева Н. А., Капустина Д. А. Климатически обусловленная изменчивость листьев и побегов ивы пурпурной (Salix purpurea L. S. L.) // Вестник фармации. 2019. № 2 (84). С. 16–28. URL: https://cyberleninka.ru/article/n/klimaticheski-obuslovlennaya-izmenchivost-listiev-i-pobegov-ivy-purpurnoy-salix-purpurea-l-s-l (дата обращения: 17.12.2023).

5. Жуйкова Татьяна Валерьевна, Попова Анастасия Сергеевна, Мелинг Элеонора Васильевна. Морфологическая изменчивость листьев Betula pendula Roth в условиях техногенной трансформации окружающей среды // Самарский научный вестник. 2021. Т. 10, № 1. С. 65–73. URL: https://cyberleninka.ru/article/n/morfologicheskaya-izmenchivost-listiev-betula-pendula-roth-v-usloviyah-tehnogennoy-transformatsii-okruzhayuschey-sredy (дата обращения: 17.12.2023).

6. Source code for statsmodels.stats.anova // Statsmodels. URL: https://www.statsmodels.org/stable/_modules/statsmodels/stats/anova.html#anova_lm

7. Github. URL: https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/decomposition/_pca.py#L118

8. Github. URL: https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/cluster/_agglomerative.py#L763

9. Райзин Дж. Вэн. Классификация и кластер. Москва : Мир, 1980. 390 с.

10. Ward J. H. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. 1963. No. 58. P. 236–244.

11. Source code for statsmodels.regression.linear_model // Statsmodels. URL: https://www.statsmodels.org/dev/_modules/statsmodels/regression/linear_model.html#OLS

References

1. The assessment of the significance of morphological characteristics in Pinus Mugo Turra crops for their determination by the method of dispersion analysis / M. V. Simakhin, V. A. Kryuchkova, A. V. Isachkin, A. M. Pokinchereda, V. G. Donskikh, A. V. Evtyukhova, E. A. Kozlova // Bulletin of Krasnoyarsk State Agrarian University. 2020. No. 11 (164). P. 61–66. URL: https://cyberleninka.ru/article/n/otsenka-znachimosti-morfologicheskih-priznakov-u-kultivarov-pinus-mugoturra-dlya-ih-opredeleniya-metodom-dispersionnogo-analiza (access date: 28.11.2023).

2. Kuzmicheva N. A. Taxonomical value of morphological characters of leaf and twig of East European willow species // Pharmacy Bulletin. 2008. No. 4 (42). P. 12–22. URL: https://cyberleninka.ru/article/n/taksonomicheskaya-znachimost-morfologicheskih-priznakov-lista-i-pobega-vostochnoevropeyskih-vidov-iv (access date: 17.12.2023).

3. Morphologic variability of the Acer campestre L. populations in Bosnia and Herzegovina / S. Kvesić, M. M. Hodžić, M. Čater, D. Ballian // Acta Biologica Sibirica. 2021. No. 7. P. 327–343.

4. Kuzmicheva N. A., Kapustina D. A. Climatic variability of purple willow leaves and shoots (Salix Purpurea L.S.L.) // Pharmacy Bulletin. 2019. No. 2 (84). P. 16–28. URL: https://cyberleninka.ru/article/n/klimaticheski-obuslovlennaya-izmenchivost-listiev-i-pobegov-ivy-purpurnoy-salix-purpurea-l-s-l (access date: 17.12.2023).

5. Zhuykova Tatyana Valeryevna, Popova Anastasiya Sergeevna, Meling Eleonora Vasilyevna. Morphological variability of Betula pendula Roth leaves under conditions of technogenic transformation of the environment // Samara Scientific Bulletin. 2021. Vol. 10, No. 1. P. 65–73. URL: https://cyberleninka.ru/article/n/morfologicheskaya-izmenchivost-listiev-betula-pendula-roth-v-usloviyah-tehnogennoy-transformatsii-okruzhayuschey-sredy (access date: 17.12.2023).

6. Source code for statsmodels.stats.anova // Statsmodels. URL:

https://www.statsmodels.org/stable/_modules/statsmodels/stats/anova.html#anova_lm

7. Github. URL: https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/decomposition/_pca.py#L118

8. Github. URL: https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/cluster/_agglomerative.py#L763

9. Rayzin Dzh. Ven. Classification and clustering. Moscow : Mir, 1980. 390 p.

10. Ward J. H. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. 1963. No. 58. P. 236–244.

11. Source code for statsmodels.regression.linear_model // Statsmodels. URL: https://www.statsmodels.org/dev/_modules/statsmodels/regression/linear_model.html#OLS

Авторы заявляют об отсутствии конфликта интересов.

Статья поступила в редакцию 19.01.2024; *одобрена после рецензирования 11.02.2024*; *принята к публикации 12.02.2024.*

The authors declare no conflicts of interests.

The article was submitted 19.01.2024; *approved after reviewing 11.02.2024*; *accepted for publication 12.02.2024.*