НАЦІОНАЛЬНА АКАДЕМІЯ АГРАРНИХ НАУК УКРАЇНИ

ІНСТИТУТ РОСЛИННИЦТВА ім. В. Я. ЮР'ЄВА

МЕТОДИКА ОЦІНКИ СЕЛЕКЦІЙНОГО МАТЕРІАЛУ КУКУРУДЗИ ЗА ІНДЕКСАМИ ПОСУХОСТІЙКОСТІ ТА БАГАТОМІРНИМ ІНДЕКСОМ СТАБІЛЬНОСТІ

науково-методичні рекомендації

ХАРКІВ 2026

УДК 633.15:581.19: 631.527
М 54

Методика оцінки селекційного матеріалу кукурудзи за індексами посухостійкості та багатомірним індексом стабільності: науково-методичні рекомендації / С. Г. Понуренко, Н. М. Музафаров, І. П.Барсуков, О. В. Сікалова, Н. В. Кузьмишина, Л.Н. Кобизєва, Л. М. Чернобай / НААН, Інститут рослинництва ім. В. Я. Юр'єва. Харків, 2026. 24 с.

У методичних рекомендаціях представлено сучасний підхід до оцінки генотипів за показниками продуктивності, адаптивності та посухостійкості із використанням індексного та багатовимірного статистичного аналізу. Запропонована методика базується на інтеграції класичних індексів стійкості до стресу з методами аналізу головних компонент (PCA), кластеризації та багатокритеріальної оптимізації, що дозволяє здійснювати комплексний і об'єктивний добір селекційного матеріалу. Детально описано алгоритм обробки експериментальних даних, включаючи розрахунок індексів адаптивності (YI, YSI, STI, SSI, TOL, SSPI та ін.), формування інтегральних показників (Selection Index, MGIDI), оцінку інформаційної цінності ознак, а також використання Pareto-оптимальності для визначення найкращих генотипів. Особливу увагу приділено інтерпретації результатів, зокрема аналізу рангових показників, стабільності кластеризації та візуалізації даних за допомогою PCA-біплотів. Запропонований підхід дозволяє враховувати складну природу взаємозв'язків між ознаками, зменшувати вплив корельованих показників і підвищувати точність селекційного відбору. Використання інтегральних критеріїв забезпечує об'єктивну оцінку генотипів та сприяє ідентифікації форм із високим адаптивним потенціалом і стабільною продуктивністю в умовах стресу.

Методичні рекомендації можуть бути використані у наукових дослідженнях, селекційних програмах, а також у навчальному процесі для підготовки фахівців у галузі агрономії, селекції та рослинництва.

Рецензенти:

Коломацька В.П. – доктор с.-г. наук, старший науковий співробітник, заступник директора з наукової роботи Інституту рослинництва ім. В. Я. Юр'єва НААН
Рожков Р. В. – кандидат біологічних наук, доцент кафедри генетики, селекції та насінництва Державного біотехнологічного університету МОН

Рекомендовано до друку вченою радою
Інституту рослинництва ім. В. Я. Юр'єва НААН
від 24 жовтня 2026 р. протокол № 4

ВСТУП

Селекція на посухостійкість є одним із ключових напрямів сучасної селекції рослин, оскільки зміни клімату призводять до збільшення частоти та інтенсивності водного стресу в агроекосистемах. В умовах нестабільного зволоження традиційний відбір лише за врожайністю в оптимальних умовах є недостатнім, так як не відображає реальної адаптивної здатності генотипів. Тому особливого значення набуває комплексна оцінка, що враховує як потенційну продуктивність, так і стабільність її реалізації в стресових умовах.

Використання багатовимірних підходів дозволяє інтегрувати різні фізіологічні та агрономічні характеристики в єдину систему оцінювання. Це забезпечує більш точне розмежування генотипів за рівнем толерантності, пластичності та стабільності. У результаті селекційний процес переходить від одновимірного відбору до системного аналізу, де враховується взаємодія генотипу з середовищем.

Важливість цього напряму також полягає у підвищенні ефективності селекційних програм. Ідентифікація стабільних і високопродуктивних генотипів дозволяє скоротити цикл створення нових сортів та підвищити ймовірність їх успішної адаптації у різних ґрунтово-кліматичних умовах. Крім того, такі підходи сприяють раціональному використанню генетичних ресурсів та збереженню біорізноманіття.

Отже, селекція на посухостійкість є стратегічно важливим напрямом, що забезпечує продовольчу безпеку та стійкість агровиробництва в умовах глобальних кліматичних змін.

1. Індекси посухостійкості в селекції рослин

Посухостійкість є складною багатокомпонентною ознакою, яка включає фізіологічні, морфологічні та біохімічні механізми адаптації рослин до дефіциту вологи. Вона проявляється не лише у здатності виживати за стресових умов, але й у збереженні продуктивності. Саме тому у селекційній практиці важливо оцінювати не тільки абсолютні показники врожайності, але й реакцію генотипів на зміну умов вирощування. Генотипи, які демонструють високий потенціал у сприятливих умовах, не завжди зберігають свою продуктивність під час посухи, що обмежує їх практичну цінність. Натомість стабільні генотипи з помірним, але надійним рівнем урожайності часто є більш бажаними.

Традиційні підходи до оцінки посухостійкості, які базуються на порівнянні середніх значень урожайності, мають обмежену інформативність. Вони не дозволяють врахувати складні взаємозв’язки між показниками та не

відображають повною мірою адаптивний потенціал генотипів. Саме тому у сучасних дослідженнях широко застосовується індексний підхід, який передбачає використання спеціальних показників, розрахованих на основі урожайності в різних умовах.

Індекси посухостійкості дозволяють кількісно оцінити реакцію генотипів на стрес і порівнювати їх між собою за різними критеріями. Наприклад, індекси, що базуються на відношенні урожайності в стресових і оптимальних умовах, відображають стабільність продуктивності. Інші індекси оцінюють абсолютні або відносні втрати врожаю, що дозволяє визначити ступінь чутливості до стресу. Є також індекси, які поєднують інформацію про продуктивність у різних умовах і дозволяють виявити генотипи з оптимальним балансом характеристик.

Перевагою індексного підходу є його здатність інтегрувати різні аспекти адаптивності в узагальнені показники. Це значно підвищує точність оцінки та дозволяє уникнути однобічних висновків. Крім того, використання кількох індексів одночасно дає змогу врахувати різні типи реакції генотипів на стрес, що є важливим для формування різноманітного селекційного матеріалу.

Разом з тим, велика кількість індексів може ускладнювати інтерпретацію результатів, особливо якщо вони корелюють між собою. У цьому випадку виникає потреба у використанні інтегральних підходів, які дозволяють об’єднати інформацію з різних показників. Саме тому сучасні селекційні дослідження все частіше використовують методи багатовимірного аналізу, такі як аналіз головних компонент або багатокритеріальні індекси, що дозволяють зменшити розмірність даних і виділити найбільш інформативні ознаки.

Індексний підхід також сприяє підвищенню об’єктивності відбору, оскільки зменшує вплив суб’єктивних оцінок і дозволяє базувати рішення на кількісних критеріях. Це особливо важливо в умовах великої кількості досліджуваних генотипів, коли ручний аналіз стає неефективним. Використання індексів у поєднанні з сучасними статистичними методами дозволяє автоматизувати процес відбору та підвищити його ефективність.

Крім того, індексний підхід є універсальним і може застосовуватися для різних культур і умов вирощування. Він дозволяє адаптувати систему оцінки до конкретних цілей селекції, наприклад, підвищення стійкості до посухи, жаростійкості або інших абіотичних факторів. Це робить його важливим інструментом у сучасній аграрній науці.

Таким чином, селекція на посухостійкість є одним із пріоритетних напрямів розвитку рослинництва в умовах кліматичних змін. Використання індексного підходу дозволяє більш точно оцінити адаптивний потенціал генотипів, врахувати складність взаємодії ознак і забезпечити обґрунтований

відбір. Поєднання індексів із методами багатовимірного аналізу відкриває нові можливості для підвищення ефективності селекційних програм і створення сортів, здатних забезпечувати стабільний урожай у змінних умовах середовища.

2. Особливості програмної реалізації селекції з використанням індексів посухостійкості.

Існує широкий спектр програмних засобів і веб-орієнтованих сервісів, які застосовуються для розрахунку індексів посухостійкості, аналізу врожайності в різних умовах середовища та інтеграції результатів у селекційні рішення. Вибір конкретного інструменту значною мірою залежить від складності дослідження, обсягу даних, необхідності автоматизації та рівня підготовки користувача. У сучасній практиці можна виділити три основні категорії: програмні середовища для статистичного аналізу, табличні процесори та спеціалізовані веб-платформи.

Одним із найпотужніших і найбільш гнучких інструментів є R — мова програмування та середовище для статистичних обчислень. Вона широко використовується в агрономії та селекції завдяки відкритому коду, великій кількості пакетів і можливості реалізації власних алгоритмів. У контексті розрахунку індексів посухостійкості R дозволяє не лише обчислювати класичні показники (YI, STI, SSI, TOL тощо), але й інтегрувати їх із багатовимірними методами аналізу. Наприклад, пакет FactoMineR забезпечує проведення аналізу головних компонент (PCA), що є ключовим етапом у зменшенні розмірності та виявленні структурних закономірностей у даних. Інший популярний пакет — agricolae — орієнтований на агрономічні дослідження і містить функції для аналізу експериментальних даних, включаючи порівняння середніх, дисперсійний аналіз та інші процедури.

Перевагою використання R є висока відтворюваність результатів: усі обчислення виконуються через скрипти, які можуть бути повторно використані або модифіковані. Це особливо важливо в селекційних програмах, де необхідно забезпечити прозорість і наукову обґрунтованість методів. Крім того, R дозволяє працювати з великими масивами даних, автоматизувати обробку та інтегрувати різні етапи аналізу — від попередньої обробки до побудови складних селекційних індексів. Водночас використання цього середовища вимагає певного рівня програмістських навичок, що може бути обмеженням для початківців.

Більш доступною альтернативою є табличні процесори, зокрема Microsoft Excel. Excel широко застосовується у селекційній практиці завдяки інтуїтивно зрозумілому інтерфейсу та простоті використання. Розрахунок індексів посухостійкості в Excel здійснюється через формули, що дозволяє швидко

отримати результати без необхідності програмування. Крім того, Excel підтримує базові інструменти аналізу даних, такі як кореляція, регресія та побудова графіків.

Однак Excel має низку обмежень. По-перше, він менш ефективний при роботі з великими наборами даних, де продуктивність може суттєво знижуватися. По-друге, реалізація складних багатовимірних методів (наприклад, PCA або кластеризації) потребує додаткових надбудов або зовнішніх інструментів. По-третє, відтворюваність аналізу в Excel є нижчою, оскільки зміни у формулах або структурі таблиці важко відслідковувати. Незважаючи на це, Excel залишається корисним інструментом для первинного аналізу, перевірки даних та навчальних цілей.

Серед професійних статистичних пакетів варто відзначити IBM SPSS Statistics, Statistica та SAS. Ці системи пропонують широкий набір стандартних методів аналізу даних, включаючи факторний аналіз, кластеризацію, регресійні моделі та інші статистичні процедури. Вони характеризуються зручним графічним інтерфейсом, що дозволяє виконувати аналіз без написання коду, а також високою надійністю та підтримкою з боку розробників.

У контексті розрахунку індексів посухостійкості SPSS, Statistica і SAS можуть використовуватися для обробки даних і подальшого аналізу, однак вони не мають вбудованих функцій для специфічних агрономічних індексів. Це означає, що користувач повинен самостійно реалізовувати формули або імпортувати результати з інших середовищ. Крім того, ці платформи є комерційними, що може обмежувати їх доступність.

Окрему категорію становлять веб-орієнтовані інструменти, які набувають дедалі більшої популярності завдяки простоті доступу та відсутності потреби у встановленні програмного забезпечення.

Перевагою таких сервісів є їх доступність і швидкість: користувач може виконати аналіз без глибоких знань статистики або програмування. Це особливо важливо для селекціонерів і агрономів. Водночас веб-платформи мають обмеження щодо гнучкості — зазвичай вони підтримують лише фіксований набір індексів і методів, що ускладнює реалізацію нестандартних або розширених підходів. Крім того, виникають питання щодо конфіденційності даних і залежності від доступу до інтернету.

У сучасних дослідженнях все частіше використовується комбінований підхід, коли різні інструменти застосовуються на різних етапах аналізу. Наприклад, первинна обробка даних може виконуватися в Excel, розрахунок індексів і багатовимірний аналіз — у R, а представлення результатів — у вигляді графіків або звітів за допомогою спеціалізованих пакетів або графічних

редакторів. Такий підхід дозволяє максимально використати переваги кожного інструменту.

Загалом, ключовим фактором вибору програмного забезпечення є баланс між гнучкістю, складністю та відтворюваністю. Для простих задач і невеликих наборів даних достатньо табличних процесорів або веб-сервісів. Проте для комплексного аналізу, що включає інтеграцію великої кількості індексів, багатовимірну статистику, кластеризацію та побудову селекційних індексів, найбільш ефективним є використання середовищ програмування, таких як R. Саме такі інструменти забезпечують необхідний рівень точності, гнучкості та наукової строгості, що є критично важливим для сучасної селекції.

3. Загальна характеристика індексів посухостійкості

Система оцінки посухостійкості, реалізована у даному підході, базується на комплексі класичних та модифікованих індексів, що дозволяють кількісно описати реакцію генотипів на водний стрес через співвідношення урожайності в оптимальних (Yp) та стресових (Ys) умовах. Ключовою особливістю такого підходу є перехід від одновимірної оцінки врожайності до багатовимірного опису адаптивності, де кожен індекс відображає окремий фізіологічний або агрономічний аспект реакції рослин.

1. Індекси продуктивності (YI, YSI)

Yield Index (YI) та Yield Stability Index (YSI) є базовими відносними показниками, що характеризують ефективність реалізації генетичного потенціалу в різних умовах середовища. YI нормує урожайність у стресових умовах відносно середнього рівня Ys у популяції, що дозволяє ідентифікувати генотипи з відносно високою продуктивністю під стресом. YSI, у свою чергу, відображає стабільність продуктивності шляхом співвідношення Ys до Yp, що дозволяє оцінити ступінь збереження врожайності в умовах стресу.

2. Індекси толерантності та втрат (TOL, HM)

Tolerance index (TOL) є абсолютною мірою втрат урожайності під впливом стресу і визначається як різниця між Yp та Ys. Він відображає чутливість генотипу, однак має обмеження через залежність від абсолютних значень урожайності.

Harmonic Mean (HM) є інтегральним індексом, який поєднує Yp та Ys в одну метрику, надаючи перевагу генотипам із збалансованою продуктивністю в обох умовах. Використання гармонійного середнього дозволяє уникнути домінування одного з компонентів та забезпечує більш стабільну оцінку загальної продуктивності.

3. Індекси чутливості до стресу (SSI, SSPI)

Stress Susceptibility Index (SSI) є одним із найважливіших індикаторів чутливості генотипу до стресу. Він нормує відносне зниження врожайності генотипу на середньопопуляційне зниження, що дозволяє порівнювати генотипи незалежно від загального рівня стресу в експерименті. Значення SSI > 1 вказує на підвищену чутливість, тоді як SSI < 1 — на відносну толерантність.

Stress Susceptibility Percentage Index (SSPI) є спрощеною альтернативою, яка виражає втрати врожайності у відсотках. Він має високу інформативність і є зручним для практичного селекційного використання, хоча менш чутливий до нормалізації популяційних ефектів.

4. Індекси адаптивності та ефективності (STI, REI, RDI)

Stress Tolerance Index (STI) є одним із ключових індексів сучасної селекції, оскільки одночасно враховує як Yp, так і Ys, нормуючи їх відносно середнього рівня продуктивності. Це дозволяє ідентифікувати генотипи, які є високопродуктивними в обох середовищах, що є бажаною селекційною ознакою.

Relative Efficiency Index (REI) базується на добутку відносних значень Ys та Yp і відображає загальну ефективність використання генетичного потенціалу в різних умовах. Він підкреслює генотипи зі збалансованою реакцією на середовище.

Relative Drought Index (RDI) нормує співвідношення Ys/Yp відносно популяційного середнього, що дозволяє оцінити відносну стабільність продуктивності в умовах стресу.

5. Модифіковані та комбіновані індекси (MSTI-K1, MSTI-K2, DI)

Modified Stress Tolerance Indices (MSTI-K1 та MSTI-K2) є розширенням STI, які додатково враховують квадратичні ефекти Yp та Ys. Це дозволяє посилити диференціацію між високопродуктивними та стабільними генотипами, підвищуючи чутливість індексу до екстремальних значень.

Drought Index (DI) є комбінованим показником, що інтегрує Ys та відносну продуктивність Ys/Yp, нормовану на популяційний рівень. Він дозволяє оцінити як абсолютну, так і відносну адаптивність генотипу до дефіциту води.

6. Інтегральні індекси складної структури (ATI, SNPI)

Abiotic Tolerance Index (ATI) поєднує відносне зниження врожайності з геометричним середнім Yp та Ys. Це дозволяє одночасно враховувати втрати та загальний рівень продуктивності, що робить його корисним для комплексної оцінки адаптивності.

Stress Nonlinear Performance Index (SNPI) є найбільш складним індексом у системі. Він використовує кубічні корені та нелінійні трансформації для зменшення впливу екстремальних значень та підсилення стабільних патернів

реакції. SNPI забезпечує більш збалансовану оцінку генотипів у випадку високої варіабельності даних.

Уся система індексів має кілька ключових особливостей:

1. Багатовимірність — кожен індекс описує окремий аспект реакції на стрес (продуктивність, стабільність, чутливість, ефективність).

2. Нормалізація відносно популяції — більшість індексів використовують середні значення Yp та Ys, що дозволяє порівнювати генотипи між експериментами.

3. Комбінація лінійних та нелінійних підходів — від простих співвідношень до геометричних та кубічних трансформацій.

4. Зменшення масштабної залежності — використання відносних величин мінімізує вплив одиниць виміру та масштабів урожайності.

5. Підготовка до багатовимірного аналізу — індекси спеціально сконструйовані для подальшого використання в PCA, кластеризації та селекційних індексах.

4. Концептуальна модель селекційної системи оцінки посухостійкості

Нами пропонується система критеріїв добору, яка ґрунтується на принципі багатокритеріальної оптимізації, що дозволяє оцінити генотипи не за окремими показниками, а за їх сукупністю. Такий підхід є особливо актуальним у сучасній селекції, де необхідно одночасно враховувати продуктивність, адаптивність, стабільність та взаємозв'язки між ознаками. У межах цієї системи використовується кілька груп критеріїв, кожна з яких відображає певний аспект поведінки генотипів, а їх інтеграція забезпечує об'єктивність і надійність відбору.

Першою і базовою групою критеріїв є показники продуктивності, представлені урожайністю в оптимальних (Yp) та стресових умовах (Ys). Ці змінні формують фундамент оцінки, оскільки відображають як потенційну, так і реалізовану продуктивність генотипів. У контексті селекції особливу цінність має саме показник Ys, оскільки він характеризує здатність генотипу забезпечувати врожай у несприятливих умовах середовища. Генотипи з високими значеннями Ys розглядаються як більш адаптивні, тоді як високий рівень Yp без відповідної підтримки в Ys може свідчити про нестійкість до стресу.

Другою важливою групою є індекси стійкості та адаптивності, які формуються на основі співвідношення між Yp і Ys. Вони дозволяють оцінити не лише абсолютні значення урожайності, але й характер реакції генотипу на зміну умов. Наприклад, індекси YI та YSI відображають відносну продуктивність і стабільність, тоді як TOL та SSPI характеризують втрати

врожаю під впливом стресу. У цій групі також представлені складніші індекси, такі як STI, DI, RDI, ATI та інші, які інтегрують різні аспекти продуктивності та адаптивності. Сутність використання цих показників полягає в тому, що жоден окремий індекс не може повністю описати поведінку генотипу, тому їх сукупний аналіз дозволяє отримати більш точну картину.

Наступною групою критеріїв є рангові показники. Для кожного індексу здійснюється ранжування генотипів, що дозволяє привести всі показники до єдиної шкали та уникнути впливу різних одиниць вимірювання. Середній ранг (Mean_Rank) відображає загальну позицію генотипу серед інших, тоді як стандартне відхилення рангів (SD_Rank) характеризує узгодженість оцінки. Генотипи з низьким середнім рангом і низькою варіабельністю рангів вважаються більш стабільними і збалансованими. Таким чином, ранговий підхід виступає як допоміжний інструмент, що підсилює інтерпретацію основних індексів.

Центральне місце у системі критеріїв займає інтегральний показник MGIDI (Multi-trait Genotype–Ideotype Distance Index). Його сутність полягає у визначенні відстані кожного генотипу до умовного ідеалу, який має найкращі значення за всіма досліджуваними ознаками. Для цього використовується метод головних компонент, що дозволяє зменшити розмірність даних і врахувати кореляційну структуру показників. MGIDI інтегрує інформацію з усіх індексів і переводить її у єдиний критерій, який легко інтерпретується: чим менше значення, тим ближче генотип до ідеального. Саме цей показник виступає основним критерієм ранжування і відбору генотипів.

Паралельно з MGIDI використовується селекційний індекс SI (Selection Index), який також має інтегральний характер, але базується на іншому підході. Він є зваженою сумою стандартизованих значень індексів, де ваги визначаються на основі інформаційної цінності показників. Зокрема, враховується їх варіабельність та середній рівень кореляції з іншими індексами. Такий підхід дозволяє зменшити вплив надлишкової інформації і підсилити роль найбільш інформативних ознак. SI виступає як незалежна перевірка результатів MGIDI і дозволяє підвищити надійність відбору.

Ще одним важливим критерієм є належність до множини Парето. Цей підхід базується на концепції багатокритеріальної оптимальності і дозволяє визначити генотипи, які не можуть бути покращені за одним показником без погіршення іншого. У даному випадку оцінюється одночасно Yp і Ys. Генотипи, що входять до множини Парето, представляють оптимальний компроміс між продуктивністю в різних умовах і тому мають високу селекційну цінність. Включення цього критерію дозволяє уникнути ситуацій, коли відбір здійснюється лише за одним показником.

Суттєву роль у системі відіграє також кластеризація, яка дозволяє групувати генотипи за подібністю їх характеристик. Використання методу k-середніх у просторі головних компонент забезпечує ефективне розділення генотипів на однорідні групи. Це дозволяє виявити структуру вибірки, визначити типи реакції на стрес і спростити подальший аналіз. Оцінка якості кластеризації здійснюється за допомогою коефіцієнта силуету, який показує, наскільки добре генотип відповідає своєму кластеру. Додатково використовується показник стабільності, розрахований методом бутстрепу, який відображає надійність кластеризації. Висока стабільність кластеризації означає, що результати групування не є випадковими.

Додатковим геометричним критерієм є відстань до ідеотипу в просторі головних компонент. Вона доповнює MGIDI і дозволяє оцінити близькість генотипів до оптимальної точки у зменшеному просторі. Цей показник є зручним для візуальної інтерпретації, особливо у поєднанні з PCA-біплотами.

Завершальним етапом є формування інтегрального рішення щодо кожного генотипу. Воно базується на поєднанні значень MGIDI та стабільності. Генотипи з найкращими значеннями MGIDI та достатньою стабільністю відносяться до категорії SELECT, що означає їх високу селекційну цінність і рекомендацію для подальшого використання. Генотипи з проміжними значеннями потрапляють до категорії TEST і потребують додаткового дослідження. Ті, що мають найгірші показники, класифікуються як REJECT.

Таким чином, пропонована система критеріїв добору є комплексною та багаторівневою. Вона поєднує класичні показники продуктивності з сучасними методами статистичного аналізу, такими як PCA, кластеризація та багатокритеріальна оптимізація. Її основною перевагою є здатність інтегрувати велику кількість інформації в узагальнені показники, що значно спрощує процес прийняття рішень і підвищує їх обґрунтованість.

5. Програмна реалізація селекційної системи оцінки посухостійкості

Для автоматизації комплексної оцінки посухостійкості нами був створений скрипт для середовища R, текст якого наведено в додатку. Нижче ми подаємо опис всіх етапів аналізу з переліком використаних програмних пакетів та функцій.

Крок 0. Очищення середовища. На початку скрипт повністю очищає робоче середовище за допомогою `rm(list = ls())`, що видаляє всі об'єкти з пам'яті, та викликає `gc()` для примусового звільнення пам'яті. Далі підключаються ключові пакети: `readxl` для імпорту Excel-файлів, `openxlsx` для експорту результатів, `dplyr` для обробки даних, `ggplot2` і `ggrepel` для

візуалізації, `cluster` для кластерного аналізу та `FactoMineR` для PCA. Функція `set.seed(123)` фіксує генератор випадкових чисел, забезпечуючи відтворюваність результатів, зокрема для кластеризації та бутстрепу.

Крок 1. Вибір файлу. Скрипт використовує `file.choose()` для інтерактивного вибору Excel-файлу користувачем. Якщо файл не обрано, виконується `stop()`, що припиняє виконання з повідомленням про помилку. Це гарантує, що подальші обчислення не почнуться без вхідних даних.

Крок 2. Завантаження та підготовка даних. Дані зчитуються через `read_excel()`, після чого залишаються лише перші три стовпці, які перейменовуються у `Genotype`, `Yp` і `Ys`. Далі типи приводяться до відповідних форматів: текстовий для генотипу та числовий для показників. За допомогою `dplyr::filter()` видаляються рядки з пропущеними значеннями, а також ті, де `Yp` $\leq$ `0` або `Ys` $<$ `0`, що забезпечує коректність подальших розрахунків.

Крок 3. Обчислення середніх значень Скрипт обчислює середні значення `Yp_mean` і `Ys_mean` за допомогою функції `mean()`. Ці величини використовуються як базові параметри для нормування індексів та розрахунку відносних показників стійкості і продуктивності.

Крок 4. Безпечні математичні функції Створюються допоміжні функції `safe_div`, `safe_sqrt` і `safe_cbrt`, які запобігають помилкам обчислення. Вони обробляють випадки ділення на нуль, від'ємних значень під коренем і кубічного кореня для від'ємних чисел. Це критично для стабільності розрахунків індексів.

Крок 5. Розрахунок індексів Формується датафрейм `traits`, що містить набір селекційних індексів, таких як YI, YSI, TOL, STI, SSI та інші. Вони оцінюють продуктивність, стійкість і адаптивність генотипів. Використовуються як прості формули, так і комбіновані з використанням безпечних функцій для уникнення числових помилок.

Крок 6. Кореляційна матриця Функція `cor()` обчислює матрицю кореляцій між індексами та показниками `Yp` і `Ys`. Параметр `use = "pairwise.complete.obs"` дозволяє використовувати всі доступні дані без повного видалення рядків із пропущеними значеннями.

Крок 7. Інформаційні ваги Скрипт оцінює інформативність кожного індексу через стандартне відхилення (`apply(..., sd)`) та середню абсолютну кореляцію як штраф. Формується показник `info_score`, який нормується до ваг `weights`. Це дозволяє зменшити вплив сильно корельованих ознак.

Крок 8. Значення та ранги Для кожного індексу обчислюються ранги за допомогою `rank()` зі спадним порядком. Дані об'єднуються через `bind_cols()`. Додатково розраховуються середній ранг і стандартне відхилення рангів для оцінки стабільності генотипу.

Крок 9. Ефективність індексів Створюється таблиця ефективності, де кожен індекс оцінюється як середнє значення кореляції з `Ys` і рангової кореляції. Це дозволяє визначити, які індекси найкраще відображають продуктивність.

Крок 10. Селекційний індекс (SI) Ознаки масштабуються через `scale()`, після чого обчислюється інтегральний показник SI як зважена сума (`matrix multiplication`). Це агрегований критерій для оцінки генотипів.

Крок 11. MGIDI Виконується PCA через `prcomp()`, визначається кількість значущих компонент, і обчислюється відстань кожного генотипу до ідеальної точки. Це показник MGIDI, який оцінює загальну бажаність генотипу.

Крок 12. PCA візуалізація Використовується `FactoMineR::PCA()` для отримання координат головних компонент і внеску змінних. Дані масштабуються для побудови біплоту, де одночасно відображаються генотипи і змінні.

Крок 13. Кластеризація Алгоритм `kmeans()` групує генотипи у 4 кластери за координатами PCA. Параметр `nstart = 50` підвищує стабільність результату.

Крок 14. Силуетний аналіз Функція `silhouette()` оцінює якість кластеризації. Значення силуету додаються до таблиці, що дозволяє визначити, наскільки добре кожен об'єкт належить до свого кластеру.

Крок 15. Стабільність кластерів Через бутстреп (50 ітерацій) перевіряється стабільність кластеризації. Для кожного генотипу обчислюється частота належності до одного кластеру.

Крок 16. Парето-оптимальність Перевіряються переваги між генотипами за `Yp` і `Ys`. Якщо інший генотип кращий за обома показниками, поточний не входить до Парето-оптимального набору.

Крок 17. Ідеотип Обчислюється відстань до ідеальної точки в просторі перших двох компонент PCA. Це ще один критерій близькості до оптимального генотипу.

Крок 18. Правило відбору На основі квантилів MGIDI та стабільності формується рішення: SELECT, TEST або REJECT. Це фінальний селекційний висновок.

Крок 19. Біплот (аналітичний) За допомогою `ggplot2` будується графік PCA з кластерами, підписами (`ggrepel`) і векторами змінних. Він показує структуру даних.

Крок 20. Біплот (селекційний) Другий графік додає позначення Парето-оптимальних генотипів та ідеотипу, що полегшує інтерпретацію відбору.

Крок 21. Експорт результатів Використовуючи `openxlsx`, створюється Excel-файл з кількома листами: дані, фінальна таблиця, індекси, ефективність, кореляції та силует. Це забезпечує зручне збереження результатів.

Крок 22. Підсумок Скрипт виводить у консоль коротке резюме з кількістю відібраних генотипів, Парето-оптимальних рішень та шляхом до збереженого файлу, завершуючи роботу системи.

6. Форма представлення результатів аналізу та основні характеристики показників системи оцінки посухостійкості.

Файл результатів, сформований у процесі виконання аналізу, є комплексною структурованою системою, яка відображає повний цикл оцінки генотипів за показниками продуктивності, адаптивності та стабільності. Його побудова забезпечує можливість не лише отримання числових значень індексів, але й їх глибокої інтерпретації в контексті селекційного відбору. Усі дані організовані у вигляді окремих аркушів, кожен з яких виконує специфічну функцію та доповнює загальну картину аналізу.

Початковою основою для всіх подальших розрахунків є аркуш Data, який містить як вихідні, так і похідні показники для кожного генотипу. У цьому аркуші наведено ідентифікатор генотипу, а також два ключові параметри — урожайність в оптимальних умовах (Yp) і урожайність у стресових умовах (Ys). Саме ці два показники формують базу для оцінки адаптивності, оскільки Yp відображає потенційну продуктивність генотипу, тоді як Ys характеризує його здатність реалізовувати цей потенціал за несприятливих умов. У сучасній селекції особливе значення має саме показник Ys, адже він визначає практичну цінність генотипу в умовах кліматичної нестабільності.

На основі цих двох показників розраховується система індексів, які дозволяють оцінити реакцію генотипів на стрес з різних сторін. До базових індексів належать YI, YSI, TOL та HM. Індекс YI (Yield Index) відображає відносну продуктивність генотипу в умовах стресу порівняно із середнім значенням по вибірці. Його високі значення свідчать про те, що генотип перевищує середній рівень адаптивності. Індекс YSI (Yield Stability Index) є відношенням Ys до Yp і показує, наскільки стабільно генотип зберігає свою продуктивність при переході від оптимальних до стресових умов. Значення,

близькі до одиниці, вказують на високу стабільність. Індекс TOL (Tolerance Index), навпаки, відображає абсолютні втрати урожайності і є різницею між Yp і Ys. Чим менше значення TOL, тим стабільнішим є генотип. Гармонійне середнє HM поєднує Yp і Ys в один показник, підкреслюючи баланс між потенційною та реальною продуктивністю.

Окрім базових індексів, у файлі представлені розширені показники, які дозволяють більш глибоко проаналізувати поведінку генотипів. Індекс SSI (Stress Susceptibility Index) характеризує чутливість до стресу і враховує відносне зниження урожайності. Низькі значення SSI є бажаними, оскільки вони свідчать про стійкість генотипу. Індекс STI (Stress Tolerance Index) оцінює здатність генотипу поєднувати високу продуктивність в обох умовах і є одним із ключових показників для відбору. Додаткові індекси, такі як DI, RDI, ATI, SNPI та REI, представляють різні математичні комбінації вихідних показників і спрямовані на підвищення точності оцінки адаптивності. Наприклад, DI та RDI дозволяють врахувати співвідношення між умовами, тоді як ATI і SNPI інтегрують нелінійні залежності між показниками. Індекс SSPI (Stress Susceptibility Percentage Index) виражає втрати у відсотках і є зручним для практичної інтерпретації.

Важливою складовою структури є система ранжування. Для кожного індексу обчислюється ранг, який дозволяє порівнювати генотипи між собою незалежно від масштабу значень. Додатково розраховуються середній ранг (Mean_Rank) і стандартне відхилення рангів (SD_Rank). Середній ранг відображає загальну позицію генотипу серед інших, тоді як стандартне відхилення показує узгодженість оцінки: низьке значення свідчить про стабільно високі або низькі позиції за всіма індексами, а високе — про неоднорідність результатів.

Ключову роль у системі відіграють інтегральні показники, зокрема Selection Index (SI) та MGIDI. Індекс SI є зваженою сумою стандартизованих значень усіх індексів і враховує їх інформаційну цінність. Ваги визначаються на основі варіабельності показників та їх кореляційної залежності, що дозволяє зменшити вплив дублюючої інформації. Високе значення SI свідчить про те, що генотип демонструє високі результати за найбільш інформативними показниками. У свою чергу, MGIDI (Multi-trait Genotype–Ideotype Distance Index) є центральним інструментом оцінки. Він базується на методі головних компонент і визначає відстань кожного генотипу до умовного ідеалу, який має максимальні значення за всіма бажаними ознаками. Чим менше значення MGIDI, тим ближче генотип до ідеалу, і тим вищою є його селекційна цінність.

Для зменшення розмірності даних і виявлення прихованих структур використовується аналіз головних компонент (PCA). У файлі представлені

координати генотипів за першими двома головними компонентами (PC1 і PC2), які пояснюють найбільшу частку варіації. Це дозволяє візуалізувати розташування генотипів у двовимірному просторі та оцінити їх подібність. Змінні також проєктуються у цей простір, що дає змогу визначити, які саме індекси найбільше впливають на розподіл.

На основі координат PCA виконується кластеризація методом k-середніх. Кожному генотипу присвоюється номер кластера, що дозволяє групувати їх за подібністю. Для оцінки якості кластеризації використовується коефіцієнт силуету (Silhouette), який показує, наскільки добре об’єкт відповідає своєму кластеру порівняно з іншими. Значення, близькі до одиниці, свідчать про чітку належність, тоді як низькі або від’ємні значення можуть вказувати на невизначеність або помилки класифікації. Додатково розраховується стабільність кластеризації (Stability) за допомогою бутстреп-підходу. Вона відображає, наскільки часто генотип потрапляє до одного й того ж кластера при повторних вибірках. Висока стабільність свідчить про надійність групування.

Окремим елементом аналізу є визначення множини Парето. Генотип вважається Парето-оптимальним, якщо не існує іншого генотипу, який перевищує його одночасно за Yp і Ys. Такий підхід дозволяє виявити генотипи, що представляють найкращі компроміси між продуктивністю в різних умовах. Належність до множини Парето є важливим аргументом на користь селекційної цінності генотипу.

Додатково обчислюється відстань до ідеотипу у просторі головних компонент (IdeotypeDist). Вона є геометричним показником і доповнює MGIDI, дозволяючи оцінити близькість генотипу до оптимального поєднання ознак у зменшеному просторі. Менші значення свідчать про кращі характеристики.

Підсумкове рішення щодо кожного генотипу представлено у змінній Decision, яка приймає значення SELECT, TEST або REJECT. Це рішення базується на поєднанні MGIDI та стабільності. Генотипи з найнижчими значеннями MGIDI та достатнім рівнем стабільності відносяться до категорії SELECT і рекомендуються для подальшого використання. Генотипи з проміжними значеннями відносяться до категорії TEST і потребують додаткового дослідження. Ті, що мають високі значення MGIDI, класифікуються як REJECT.

Аркуш Final є узагальненням усіх результатів і містить відсортований список генотипів за значенням MGIDI. Це дозволяє швидко визначити найперспективніші варіанти. Аркуш Index_Info надає інформацію про статистичні характеристики індексів, зокрема їх стандартне відхилення, середню кореляцію та ваги. Це дозволяє зрозуміти, які показники мають найбільший вплив на інтегральні оцінки. Аркуш Efficiency містить оцінку

ефективності індексів на основі їх кореляції з Ys, що дає змогу визначити найбільш інформативні показники адаптивності. Кореляційна матриця дозволяє виявити взаємозв'язки між індексами і уникнути дублювання інформації. Аркуш Silhouette містить деталізовану інформацію про якість кластеризації для кожного генотипу.

Окрім табличних результатів, важливою складовою аналізу є графічна візуалізація даних, яка представлена у вигляді двох основних біплотів, побудованих на основі методу головних компонент (PCA). Ці графіки виконують не лише ілюстративну, але й аналітичну функцію, дозволяючи глибше зрозуміти структуру взаємозв'язків між генотипами та показниками, а також обґрунтувати селекційні рішення.

Перший графік, умовно позначений як «аналітичний PCA-біплот (PCA Analytical)», відображає розподіл генотипів у просторі перших двох головних компонент (PC1 та PC2), які пояснюють найбільшу частку загальної варіації даних. Кожна точка на графіку відповідає окремому генотипу, а її координати визначаються значеннями PC1 і PC2. Це дозволяє зменшити багатовимірний простір індексів до двох вимірів без суттєвої втрати інформації.

Генотипи на графіку додатково диференційовані за допомогою кольору та форми відповідно до результатів кластеризації методом k-середніх. Це дає змогу візуально ідентифікувати групи генотипів зі схожими характеристиками. Генотипи, що належать до одного кластера, зазвичай розташовані близько один до одного, що свідчить про подібність їх реакції на стрес та структури індексів. Натомість значна відстань між кластерами вказує на суттєві відмінності між групами.

Важливим елементом аналітичного біплоту є вектори змінних (індексів), які виходять із початку координат. Кожен вектор відповідає певному показнику і має напрямок та довжину. Напрямок вектора відображає внесок змінної у формування головних компонент, а довжина — силу цього внеску. Чим довший вектор, тим більший вплив відповідного індексу на структуру даних.

Інтерпретація взаємного розташування векторів дозволяє оцінити кореляційні зв'язки між показниками. Вектори, спрямовані в один бік, свідчать про позитивну кореляцію, тоді як протилежно спрямовані — про негативну. Перпендикулярне розташування вказує на відсутність суттєвого зв'язку. Таким чином, графік дозволяє візуально підтвердити результати кореляційного аналізу.

Розташування генотипів відносно векторів має важливе інтерпретаційне значення. Генотипи, які знаходяться у напрямку певного вектора, характеризуються високими значеннями відповідного показника. Наприклад, якщо генотип розташований у напрямку векторів STI або YI, це свідчить про

його високу толерантність до стресу. Якщо ж генотип знаходиться у протилежному напрямку, це означає низькі значення відповідного індексу.

Лінії, що проходять через нульові значення PC1 та PC2, поділяють графік на чотири квадранти, кожен з яких може бути інтерпретований як зона з певною комбінацією характеристик. Наприклад, один квадрант може відповідати генотипам із високою продуктивністю в обох умовах, тоді як інший — генотипам із низькою адаптивністю.

Другий графік, «PCA Selection (Pareto + Ideotype)», є розширеною версією першого і орієнтований безпосередньо на прийняття селекційних рішень. Він зберігає всі елементи аналітичного біплоту, але доповнюється додатковими маркерами, які підкреслюють ключові об’єкти.

Зокрема, на цьому графіку виділяються генотипи, що входять до множини Парето. Вони позначаються окремим символом (зазвичай іншим кольором або формою) і представляють генотипи, які не перекриваються іншими за одночасною оцінкою Yp та Ys. Їх розташування на графіку дозволяє швидко ідентифікувати найбільш збалансовані варіанти.

Окремо на графіку позначається ідеотип — умовна точка, яка відповідає максимальним значенням головних компонент. Вона відображає теоретично оптимальний генотип із найкращими характеристиками за всіма показниками. Відстань реальних генотипів до цієї точки є візуальним відображенням їх селекційної цінності: чим ближче генотип до ідеотипу, тим він перспективніший.

Поєднання інформації про кластери, вектори змінних, Pareto-оптимальність та ідеотип дозволяє здійснювати комплексну інтерпретацію результатів безпосередньо на графіку. Це значно спрощує процес аналізу та робить його більш наочним.

Важливим аспектом є також відсоток поясненої варіації, який зазначається на осях PC1 та PC2. Він показує, яку частку загальної варіації даних пояснює кожна компонента. Високі значення свідчать про те, що двовимірне представлення є достатнім для адекватного відображення структури даних.

Таким чином, графіки виконують роль інтегруючого інструменту, який поєднує результати різних методів аналізу — від індексів і кореляцій до кластеризації та оптимізаційних підходів. Вони дозволяють не лише підтвердити числові результати, але й виявити закономірності, які складно побачити в табличному вигляді. Використання PCA-біплотів є особливо цінним у селекційних дослідженнях, оскільки забезпечує наочне представлення складних багатовимірних залежностей і сприяє прийняттю обґрунтованих рішень щодо відбору генотипів.

ДОДАТОК

```
# ================================
# 0. CLEAN ENVIRONMENT
# ================================
rm(list = ls())
gc()
library(readxl)
library(openxlsx)
library(dplyr)
library(ggplot2)
library(ggrepel)
library(cluster)
library(FactoMineR)
set.seed(123)
# ================================
# 1. FILE INPUT
# ================================
file_path <- file.choose()
if (file_path == "") stop("Файл не обрано!")
# ================================
# 2. LOAD DATA
# ================================
df <- read_excel(file_path)
df <- df[, 1:3]
colnames(df) <- c("Genotype", "Yp", "Ys")
df$Genotype <- as.character(df$Genotype)
df$Yp <- as.numeric(df$Yp)
df$Ys <- as.numeric(df$Ys)
df <- df %>%
  filter(!is.na(Yp), !is.na(Ys), Yp > 0, Ys >= 0)
# ================================
# 3. MEANS
# ================================
Yp_mean <- mean(df$Yp)
Ys_mean <- mean(df$Ys)
# ================================
# 4. SAFE FUNCTIONS
# ================================
safe_div <- function(a,b) ifelse(is.na(b) | b == 0, NA, a/b)
safe_sqrt <- function(x) ifelse(x < 0, NA, sqrt(x))
safe_cbrt <- function(x) sign(x) * abs(x)^(1/3)
# ================================
# 5. INDICES (VALUES)
# ================================
traits <- data.frame(
  YI  = df$Ys / Ys_mean,
  YSI = df$Ys / df$Yp,
  TOL = df$Yp - df$Ys,
  HM  = safe_div(2*df$Yp*df$Ys, (df$Yp + df$Ys)),
  SSI = (1 - (df$Ys / df$Yp)) * (df$Yp / Yp_mean),
  STI = (df$Yp * df$Ys) / (Yp_mean^2),
  MSTI_K1 = ((df$Yp^2)/(Yp_mean^2)) * ((df$Yp * df$Ys) / (Yp_mean^2)),
  MSTI_K2 = ((df$Ys^2)/(Ys_mean^2)) * ((df$Yp * df$Ys) / (Yp_mean^2)),
```

```
  DI  = df$Ys * (df$Ys / df$Yp) / Ys_mean,
  RDI = (df$Ys / df$Yp) / (Ys_mean / Yp_mean),
  ATI = ((df$Yp - df$Ys)/(Yp_mean - Ys_mean + 1e-9)) * safe_sqrt(df$Yp*df$Ys),
  SNPI = safe_cbrt((df$Yp + df$Ys)/(df$Yp - df$Ys + 1e-9)) *
    safe_cbrt(df$Yp * df$Ys^2),
  REI = (df$Ys / Ys_mean) * (df$Yp / Yp_mean),
  SSPI = ((df$Yp - df$Ys)/(2*df$Yp))*100
)
# ================================
# 6. CORRELATION
# ================================
cor_matrix <- cor(cbind(traits, Yp = df$Yp, Ys = df$Ys),
                  use = "pairwise.complete.obs")
# ================================
# 7. INFORMATION WEIGHTS
# ================================
sd_vals <- apply(traits, 2, sd)
cor_penalty <- apply(abs(cor(traits)), 2, mean)
info_score <- sd_vals / cor_penalty
weights <- info_score / sum(info_score)
info_table <- data.frame(
  Index = names(traits),
  SD = sd_vals,
  CorPenalty = cor_penalty,
  InfoScore = info_score,
  Weight = weights
)
# ================================
# 8. VALUE + RANK
# ================================
rank_matrix <- as.data.frame(lapply(traits, function(x)
  rank(-x, ties.method = "average")
))
value_df <- traits
colnames(value_df) <- names(traits)
rank_df <- rank_matrix
colnames(rank_df) <- paste0(names(traits), "_rank")
df <- bind_cols(df, value_df, rank_df)
df$Mean_Rank <- rowMeans(rank_matrix)
df$SD_Rank <- apply(rank_matrix, 1, sd)
# ================================
# 9. SELECTION EFFICIENCY
# ================================
ys_rank <- rank(-df$Ys)
eff_table <- data.frame(
  Index = names(traits),
  Efficiency = sapply(traits, function(x) {
    mean(c(
      cor(x, df$Ys),
      cor(rank(-x), ys_rank)
    ), na.rm = TRUE)
  })
) %>%
  arrange(desc(Efficiency))
```

```
# ================================
# 10. SI
# ================================
traits_scaled <- scale(traits)
df$SI <- as.matrix(traits_scaled) %*% weights
# ================================
# 11. MGIDI
# ================================
pca <- prcomp(traits_scaled)
expl <- summary(pca)$importance[2,]
n_comp <- sum(expl > 0.10)
scores <- pca$x[,1:n_comp]
ideal <- apply(scores, 2, max)
df$MGIDI <- apply(scores, 1, function(x)
  sqrt(sum((x - ideal)^2))
)
# ================================
# 12. PCA VISUALIZATION
# ================================
pca_vis <- PCA(cbind(traits_scaled,
                     scale(df[,c("Yp","Ys")])),
               graph = FALSE)
df$PC1 <- pca_vis$ind$coord[,1]
df$PC2 <- pca_vis$ind$coord[,2]
var_coords <- as.data.frame(pca_vis$var$coord[,1:2])
colnames(var_coords) <- c("Dim1","Dim2")
var_coords$Variable <- rownames(var_coords)
var_exp <- pca_vis$eig[,2]
pc1_var <- sprintf("%.1f", var_exp[1])
pc2_var <- sprintf("%.1f", var_exp[2])
mult <- min(
  diff(range(df$PC1)) / diff(range(var_coords$Dim1)),
  diff(range(df$PC2)) / diff(range(var_coords$Dim2))
) * 0.7
var_coords$Dim1 <- var_coords$Dim1 * mult
var_coords$Dim2 <- var_coords$Dim2 * mult
# ================================
# 13. CLUSTERING
# ================================
set.seed(123)
km <- kmeans(df[, c("PC1","PC2")], centers = 4, nstart = 50)
df$Cluster <- as.factor(km$cluster)
# ================================
# 14. SILHOUETTE
# ================================
sil <- silhouette(km$cluster, dist(df[, c("PC1","PC2")]))
df$Silhouette <- sil[,3]
sil_df <- data.frame(
  Genotype = df$Genotype,
  Cluster = km$cluster,
  Silhouette = sil[,3]
)
# ================================
# 15. STABILITY
# ================================
B <- 50
```

```
mat <- matrix(NA, nrow = nrow(df), ncol = B)
for (b in 1:B) {
  idx <- sample(1:nrow(df), replace = TRUE)
  boot <- df[idx, c("PC1","PC2")]
  km_b <- kmeans(boot, centers = 4, nstart = 25)
  mat[idx, b] <- km_b$cluster
}
df$Stability <- apply(mat, 1, function(x){
  x <- na.omit(x)
  if(length(x)==0) return(NA)
  max(table(x))/length(x)
})
# ================================
# 16. PARETO
# ================================
df$Pareto <- TRUE
for (i in 1:nrow(df)) {
  for (j in 1:nrow(df)) {
    if (all(df$Yp[j] >= df$Yp[i]) &&
        all(df$Ys[j] >= df$Ys[i]) &&
        (df$Yp[j] > df$Yp[i] || df$Ys[j] > df$Ys[i])) {

      df$Pareto[i] <- FALSE
      break
    }
  }
}
# ================================
# 17. IDEOTYPE
# ================================
scores_all <- pca$x[,1:2]
ideotype <- apply(scores_all, 2, max)
df$IdeotypeDist <- sqrt(
  (scores_all[,1] - ideotype[1])^2 +
    (scores_all[,2] - ideotype[2])^2
)
# ================================
# 18. SELECTION RULE
# ================================
q <- quantile(df$MGIDI, c(0.2, 0.6))
df$Decision <- ifelse(
  df$MGIDI <= q[1] & df$Stability >= 0.6, "SELECT",
  ifelse(df$MGIDI <= q[2], "TEST", "REJECT")
)
final_table <- df %>%
  select(Genotype, Yp, Ys,
         everything(),
         MGIDI, SI,
         Mean_Rank, SD_Rank,
         Silhouette,
         Pareto,
         IdeotypeDist,
         Cluster, Stability, Decision) %>%
  arrange(MGIDI)
```

```
# ==============================
# 19. BIPLOT 1 (ANALYTICAL)
# ==============================
p1 <- ggplot(df, aes(PC1, PC2)) +
  geom_point(aes(color = Cluster, shape = Cluster), size = 3) +
  geom_text_repel(aes(label = Genotype), size = 3) +
  geom_segment(
    data = var_coords,
    aes(x = 0, y = 0, xend = Dim1, yend = Dim2),
    inherit.aes = FALSE,
    arrow = arrow(length = unit(0.2,"cm")),
    color = "black"
  ) +
  geom_text(
    data = var_coords,
    aes(x = Dim1, y = Dim2, label = Variable),
    inherit.aes = FALSE,
    size = 3
  ) +
  geom_hline(yintercept = 0, linewidth = 0.6, color = "grey30") +
  geom_vline(xintercept = 0, linewidth = 0.6, color = "grey30") +
  labs(
    title = "PCA Analytical",
    x = paste0("PC1 (", pc1_var, "%)"),
    y = paste0("PC2 (", pc2_var, "%)")
  ) +
  theme_minimal()

print(p1)
# ==============================
# 20. BIPLOT 2 (SELECTION)
# ==============================
p2 <- ggplot(df, aes(PC1, PC2)) +
  geom_point(aes(color = Cluster, shape = Cluster), size = 3) +
  geom_point(aes(shape = "Pareto"),
             data = df[df$Pareto,],
             color = "gold", size = 4) +
  geom_point(aes(shape = "Ideotype"),
             x = ideotype[1], y = ideotype[2],
             color = "red", size = 5) +
  geom_text_repel(aes(label = Genotype), size = 3) +
  geom_segment(
    data = var_coords,
    aes(x = 0, y = 0, xend = Dim1, yend = Dim2),
    inherit.aes = FALSE,
    arrow = arrow(length = unit(0.2,"cm")),
    color = "black"
  ) +
  geom_text(
    data = var_coords,
    aes(x = Dim1, y = Dim2, label = Variable),
    inherit.aes = FALSE,
    size = 3
  ) +
  geom_hline(yintercept = 0, linewidth = 0.6, color = "grey30") +
  geom_vline(xintercept = 0, linewidth = 0.6, color = "grey30") +
```

```
  scale_shape_manual(values = c(
    "Cluster 1" = 16,
    "Cluster 2" = 17,
    "Cluster 3" = 15,
    "Cluster 4" = 18,
    "Pareto" = 17,
    "Ideotype" = 8
  )) +
  labs(
    title = "PCA Selection (Pareto + Ideotype)",
    shape = "Legend",
    color = "Cluster",
    x = paste0("PC1 (", pc1_var, "%)"),
    y = paste0("PC2 (", pc2_var, "%)")
  ) +
  theme_minimal()

print(p2)
# ==============================
# 21. EXPORT
# ==============================
out <- sub("\\.xlsx$", "_v13.xlsx", file_path)
wb <- createWorkbook()
addWorksheet(wb, "Data")
writeData(wb, "Data", df)
addWorksheet(wb, "Final")
writeData(wb, "Final", final_table)
addWorksheet(wb, "Index_Info")
writeData(wb, "Index_Info", info_table)
addWorksheet(wb, "Efficiency")
writeData(wb, "Efficiency", eff_table)
addWorksheet(wb, "Correlation")
writeData(wb, "Correlation", cor_matrix)
addWorksheet(wb, "Silhouette")
writeData(wb, "Silhouette", sil_df)
saveWorkbook(wb, out, overwrite = TRUE)
# ==============================
# 22. SUMMARY
# ==============================
cat("\n====================\n")
cat("BREEDING SYSTEM v13 COMPLETE\n")
cat("SELECT:", sum(df$Decision=="SELECT"), "\n")
cat("Pareto:", sum(df$Pareto), "\n")
cat("Saved:", out, "\n")
cat("====================\n")
```