УДК 004.62

Голуб Б. Л.
кандидат технічних наук, завідувачка кафедри комп'ютерних наук факультету інформаційних технологій НУБІП України

Лещук Н.В.
аспірант, Український інститут експертизи сортів рослин

Циба С.В.
аспірант кафедри комп'ютерних наук факультету інформаційних технологій НУБІП України

ТЕХНОЛОГІЇПРОВЕДЕННЯ ЕКСПЕРТИЗИ НОВИХ СОРТІВ РОСЛИН

Анотація. *Розглянуто недоліки існуючої методики проведення експертизи нових сортів рослин, що використовується Українським інститутом експертизи сортів рослин, з організаційної та технологічної точки зору. Запропоновано використання сучасних технологій OLAPта DataMining для усунення вказаних проблем. Наведено структуру вітрини даних та визначені етапи подальшої роботи.*

Ключові слова*: експертиза сортів рослин, метод варіаційної статистики, аналіз даних, експертна система, сховище даних, вітрина даних, технологія OLAP, технологія DataMining.*

Вступ

Сортовим рослинним ресурсам належить особлива роль в економічному і соціальному розвитку України, насамперед у стабілізації та збільшенні обсягів виробництва продукції рослинництва як основи продовольчої безпеки держави. До Державного реєстру сортів рослин, придатних для поширення в Україні, занесено понад 8000 сортів, які належать до сільськогосподарських, лісових, декоративних та інших ботанічних таксонів. Подальше формування сортових рослинних ресурсів, передусім за рахунок сортів, які перебувають у комерційному обігу, потребує вдосконалення механізму його законодавчого, організаційного, науково-технічного,інформаційно-технологічного, фінансового, кадрового та іншого забезпечення. Результати комплексу польових і лабораторних досліджень за кваліфікаційної експертизи сортів рослин мають бути достовірними та об'єктивними для прийняття кінцевого рішення за заявкою на сорт рослин. Алгоритм та механізм обробки цих результатів з метою визначення придатності до розповсюдження на території

України сорту мають бути науково обґрунтованими, такими, що не викликають сумнівів у заявників сортів.

Огляд існуючих досліджень

Український інститут експертизи сортів рослин (УІЕСР) є державною установою, яка реалізує усі дії, пов'язані із прийняттям заявки на новий сорт, формуванням плану випробувань на дослідних станціях України, внесенням результатів польових випробувань у центральну базу даних та, на основі цих результатів, визначення можливості розповсюдження заявленого сорту на території України і, відповідно, можливості занесення його до Державного реєстру сортів рослин.

Збирання та облік урожаю – це завершальний етап польових досліджень у закладах експертизи. Формування середньої проби з кожної ділянки забезпечує єдиний підхід до всіх варіантів досліду.

Урожайність з приведенням її до стандартної вологості *(X)* визначають за формулою (1) :

$$X = \frac{Y \times (100 - B)}{100 - CB}, \quad (1)$$

де:

Y– урожайність за збирання, т/га;

B– вологість урожаю, %;

CB – стандартна вологість для виду, %.

Урожайність соломи *(X)* з гектара обліковують загалом по сорту та обчислюють за формулою (2):

$$X = \frac{а \times б}{в}, \quad (2)$$

де:

а – урожайність зерна, т/га;

б – соломи в загальній масі, %;

в – зерна в загальній масі, %.

Середню врожайність сорту визначають як середнє арифметичне з повторень. Такий спосіб обчислення застосовують незалежно від зменшення облікової площі ділянок окремих повторень у результаті виділення вилучок.

Порівняння сортів на одній станції за кілька років виводять як середнє (а не зважене) з урожаю сортів незалежно від зміни облікової площі ділянки в різні роки. Показники досліду багаторічних видів (трави, плодові, ягідні та ін.), який ведуть тривалий час в одному місці, встановлюють через суму врожайності за роки експертизи (сортовивчення). Статистично опрацьовують дані таких дослідів за кілька років за сумою врожаїв сортів у повтореннях.В агротехнічних дослідах окремі варіанти порівнюють із контрольним і між собою окремо кожного сорту, й загалом варіанту.

У разі випадіння та наступного відновлення статистичним методом урожайних даних середню врожайність по сорту визначають з урахуванням відновлених даних. Усі відновлені дані беруть у дужки. Інші показники, що

випали, статистичним методом не відновлюють, а середнє виводять як середнє з прийнятих до обліку спостережень.

За умов опрацювання даних експертизи сортів, передбачених Програмою по зонах, групах видів і роках, дисперсійний аналіз статистичних даних не в змозі забезпечити вирішення поставленого завдання з виявлення найкращих сортів у порівнянні з умовними (національними) стандартами.

Для цієї мети підходить метод варіаційної статистики, суть якого викладено нижче, ґрунтуючись на конкретних прикладах.

Спочатку визначають середню врожайність з приведенням її до стандартноївологості для наступних видів: зернові, зернобобові (горох, соя), ріпак та ін. за формулою (1).

Для результатів урожайності в повтореннях потрібно встановити додатковий контроль їхньої вірогідності за величиною відносної помилки середнього значення:

$$P = \frac{m_{\bar{X}}}{\bar{X}} \times 100, \qquad (3)$$

де:

P – помилка спостереження, %;

$m_{\bar{X}}$ – помилка середнього значення, т/га;

$\bar{X}$ – середнє значення, т/га.

Якщо P>5%, дані врожаю слід бракувати, сорт позначити і сповістити про недостовірність даних урожаю цього сорту.

Аналіз результатів експертизи починається з добору з таблиці результатів даних для статистичного опрацювання.

Статистичнеопрацювання даних здійснюють за алгоритмом варіаційного аналізу. Порівняння сортів ведеться з умовним (національним) стандартом, який розраховується на кожний рік, на ґрунтово-кліматичну зону, для кожного виду рослин, за блоками.

Першим кроком статистичного аналізу є оцінка однорідності варіант варіаційного ряду. Оцінку однорідності слід виконувати для показника як за обчислення результатів експертизи, так і за обчислення показників умовного (національного) стандарту.

Для показників стійкості до стресових явищ, шкідливих організмів статистичне опрацювання не застосовують. Для них у звіт вводять *максимальні значення* з масиву даних.

Алгоритм оцінки.Варіаційний ряд упорядковують по висхідній.

1. Обчислюють середнє квадратичне відхилення (σ) варіаційного ряду за формулою:

$$\sigma = \sqrt{\frac{\sum (x_i - M)^2}{N-1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N-1}}, \qquad (4)$$

де:

M – середнє значення варіаційного ряду;

x_i – окрема дата варіаційного ряду;

N – кількість дат варіаційного ряду.

2. Для оцінки вірогідності належності максимальної дати (υ_N) – до розподілу визначають критерії цієї вірогідності за формулою:

$$\upsilon_N = \frac{x_N - x_{N-1}}{\sigma}, \quad (5)$$

де:

x_N – максимальна дата ряду;

x_{N-1} – дата ряду, попередня максимальної;

σ – середнєквадратичневідхилення ряду.

3. Для оцінки вірогідності належності мінімальної дати (υ_1) до розподілу вираховують її критерії за формулою:

$$\upsilon_1 = \frac{x_2 - x_1}{\sigma}, \quad (6)$$

де:

x_1 – мінімальна дата ряду;

x_2 – наступна за мінімальною дата ряду;

σ – середнєквадратичневідхилення ряду.

4. За таблицею 1 оцінюють критерії вірогідності для N(приймаючи найближче, менше за N, число). Якщо υ_Nі $\upsilon_1 < \upsilon_{табл.}$, то максимальну та мінімальну дати не вилучаємо з ряду (і навпаки).

5. Якщо будь-яку дату вилучають з ряду, оцінюють новий ряд за формулами (4)–(6).

6. Обчислення параметрів показника й умовного стандарту. Після оцінки належності варіант до варіаційного ряду статистики окремого показника за кожний рік обчислюють методом варіаційного аналізу за формулами:

$$\overline{X}_j = \frac{\sum x_i}{N_1}, \quad (7)$$

де:

$\overline{X}_j$ – середнє значення показника за рік;

x_i – часткове значення показника;

N_i – об’єм вибірки;

$$\sigma_j^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N_i}}{N_i - 1}, \quad (8)$$

де:

σ_j^2 – дисперсія показника;

x_i – часткове значення показника;

N_i – об’єм вибірки;

$$\sigma_j = \sqrt{\sigma_j^2}, \quad (9)$$

де:

σ_j – середнє квадратичне відхилення показника;

σ_j^2 – дисперсія показника.

За отриманими статистиками обчислюють основні показники умовного стандарту на рік аналізу за формулами:

$$M = \frac{\sum N_i \times \overline{X_j}}{\sum N_i}, \tag{10}$$

де:

M – середнє значення показника умовного стандарту;

$\overline{X}_j$ – середнє значення показника за окремий рік;

N_i– об'єм вибірки;

$$\sigma_M = \sqrt{\frac{\sum (N_i - 1) \times \sigma_j^2}{\sum N_i - k}}, \tag{11}$$

де:

σ_M – середнє квадратичне відхилення показника умовного стандарту;

σ_j^2 – дисперсія окремого показника;

k – кількість років, узятих для розрахунку;

N_i – об'єм вибірки;

$$m_M = \frac{\sigma_M}{\sqrt{\sum N_i}}, \tag{12}$$

де:

m_M – помилка середнього значення показника умовного стандарту;

σ_M – середнє квадратичне відхилення показника умовного стандарту;

N_i – об'єм вибірки;

$M - t_{05} m_M < ДІ < M + t_{05} m_M,$ (13)

де: $ДІ$ – довірчий інтервал для показника умовного стандарту;

M– середнє значення показника умовного стандарту;

m_M – помилкасереднього значення показника умовного стандарту;

t_{05} – критерій Стьюдента на рівні значущості 95% за кількості ступенів свободи $\nu = \sum N_i - 1$.

На рік аналізу значення показника сорту обчислюють за формулами (7)–(9).

Вірогідність різниці значення показника сорту від значення показника умовного стандарту обчислюють за формулою:

$$|t| = \frac{M_c - M_{yc}}{\sqrt{\frac{\sigma_c^2}{N_c} + \frac{\sigma_{yc}^2}{N_{yc}}}}, \tag{14}$$

з кількістю ступенів свободи:

$$v = \left(N_c + N_{yc} - 2\right) \times \left[\frac{1}{2} + \left(\frac{\sigma_c^2}{\sigma_{yc}^2} + \frac{\sigma_{yc}^2}{\sigma_c^2}\right)^{-1}\right], \tag{15}$$

де:

$|t|$ – критерій вірогідності показника;

M_c – середнє значення показника сорту;

M_{yc} – середнєзначення показника умовного стандарту;

σ_c^2 – дисперсія показника сорту;

σ_{yc}^2 – дисперсія показника умовного стандарту;

N_c – кількість значень показника сорту;

N_{yc} – кількість значень показника умовного стандарту.

Кількість ступенів свободи (ν) заокруглюють до цілого числа. За таблицею значень критерію Стьюдента на довірчому рівні 95% знаходять табличне значення t_{05} і порівнюють його з обчисленим $|t|$. Якщо $t_{05}<|t|$, різниця між показниками сорту і показником умовного стандарту значуща. Коли різниця значуща, у звіті поряд з цим значенням виводять відповідну примітку (*).

Порівняння з показниками сорту ведуть обчисленням критерію вірогідності різниці середніх значень (14) – (15).

Постановка проблеми

Наведена методика обробки результатів польових випробувань сортів рослин має цілий ряд недоліків. Наведемо найважливіші з них.

1. Визначені параметри польових випробувань на дослідних станціях передаються до УІЕСР у паперовому вигляді (форма 1). Ці дані надалі вносяться до бази даних спеціалістами УІЕСР. На такому довгому шляху достовірність такої інформації сумнівна.

2. База даних структурована під операції для транзакцій, а не для аналізу.

3. Алгоритм обробки параметрів польових випробувань реалізований у коді програми, для його зміни необхідно долучати розробників програми та компілювати увесь код.

4. Заявники сортів рослин, отримуючи негативний результат, висловлюють сумніви у коректності обчислень, що викликає цілий ряд спірних питань.

Мета і задачі дослідження

Виходячи з вищезазначеного, метою дослідження є розробка інформаційної технології, яка б дозволили проводити експертизу нових сортів рослин, спираючись лише на дані, накопичені шляхом проведення польових випробувань.

Задача дослідження – розробка експертної системи нових сортів рослин на основі сучасних технологій OLAPi DataMining.

Технологія OLAP(аналіз даних у режимі реального часу) дозволяє перевірити гіпотези та передбачає побудову звітів, які ілюструють у належному вигляді як отримані параметри польових випробувань, так і статистично оброблені показники.

Технологія DataMining(добування знань) дозволяє отримати нові, до цього моменту невідомі знання з накопичених даних.

Обидві технології вимагають представлення даних у спеціально структурованому вигляді – сховищі даних. Структура сховища даних має відповідати задачам аналізу.

Отримані результати

У ході проведеного дослідження у середовищі MSSQLManagementStudioбуло побудовано частину сховища даних – вітрину даних. На рис. 1 представлена структура вітрини даних, як частини сховища.

Вітрина складається з однієї таблиці фактів та трьох таблиць вимірів. Вітрина має структуру, що відповідає схемі «зірка». Це означає, що у таблиці фактів разом з детальними даними зберігаються узагальнені дані, а таблиці вимірів не нормалізовані.

Детальні дані для таблиць фактів та вимірів завантажуються у вітрину з бази даних УІЕСР. Узагальнені дані розраховуються за формулами, представленими вище.

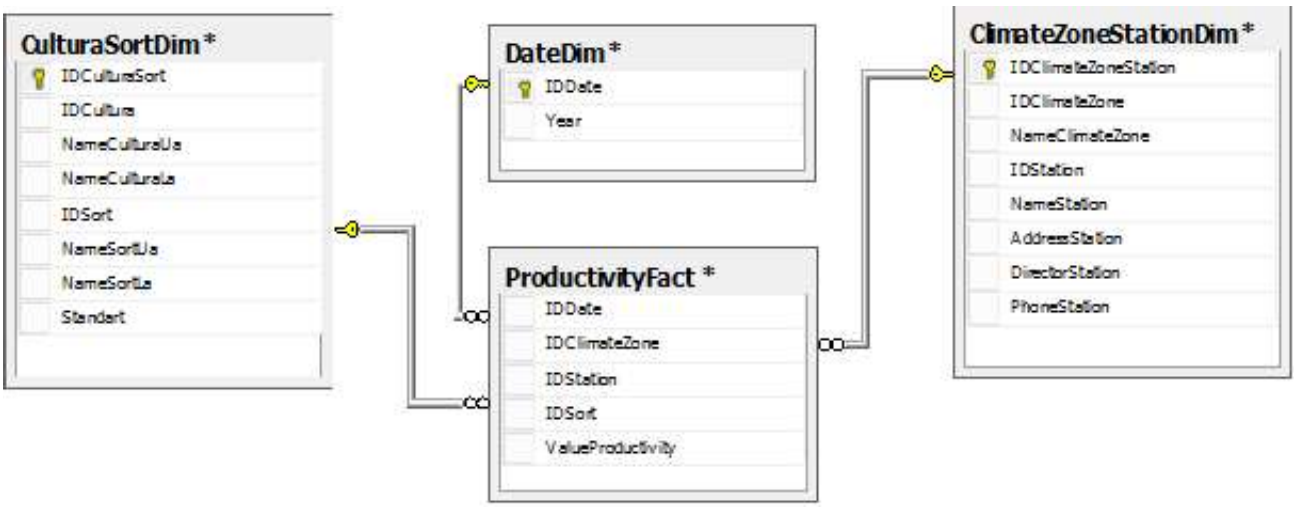

Рис 1. Вітрина даних

Саме така структура дозволить отримати звіти у розрізі будь якого сорту рослини, станції, кліматичної зони, певного часового періоду. У разі необхідності у вітрину можуть бути додані додаткові виміри, збільшено кількість фактів відповідно досліджуваних параметрів сортів рослин. Для побудови звітів використовується середовище BIMSSQLServer.

Перспективи подальших досліджень

З огляду вищезазначеного, проведені дослідження є лише початковим етапом. Надалі дослідження будуть виконуватися за такими етапами:

- ✓ побудова рішення у середовищі BIMSSQLServer для реалізації задач аналізу даних відповідно технології OLAP;
- ✓ реалізація механізму технології DataMining для проведення достовірної експертизи нових сортів рослин;

- ✓ побудова повної архітектури системи.

Висновки

Методика, що використовуються натепер УІЕСР для проведення експертизи нових сортів рослин, має суттєві недоліки з точки зору організації та технологій обробки даних. Тому пропонується дослідити доцільність і ефективність використання сучасних технологій аналізу даних, а саме – технології OLAPiDataMining. Таке дослідження ґрунтується на методах математичної статистики, що використовується натепер УІЕСР для проведення експертизи, але дані збираються, очищуються та оброблюються на шляху з дослідних станцій та бази даних УІЕСР до сховища даних. Сховище даних структуровано спеціально для зручності проведення аналізу даних. У подальшому буде проведено дослідження щодо використання технології DataMining.

Подяка

Висловлюємо подяку магістру першого року навчання за спеціальністю «Інформаційні управляючі системи і технології» НУБіП України Трохименку В.Ю. за активну участь у дослідженнях.

Список використаних джерел

1. Доспехов Б.А. Методика полевогоопыта / Б.А. Доспехов. – М.: Агропромиздат, 1985. – 351 с.

2. Голуб Б.Л. Автоматизація обліку показників придатності сортів до поширення в Україні / Голуб Б.Л., Трохименко В.Ю.// Енергетика і автоматика: електрон. наук. фах. вид. / Національний університет біоресурсів і природокористування України. – Київ: ВЦ НУБіП України, 2016. – Вип. 1. – С. 129-134.

Голуб Б. Л.
кандидат технических наук, заведующая кафедрой компьютерных наук факультета информационных технологий НУБИП Украины

Лещук Н.В.
аспирант, Украинский институт экспертизы сортов растений

Цыба С.В.
аспирант кафедры компьютерных наук факультета информационных технологий НУБИП Украины

ТЕХНОЛОГИИ ПРОВЕДЕННЯ ЭКСПЕРТИЗЫ НОВЫХ СОРТОВ РАСТЕНИЙ

Аннотация. *Рассмотрены недостатки существующей методики проведения экспертизы новых сортов растений, используется Украинским институтом экспертизы сортов растений, с организационной и технологической точки зрения. Предложено использование современных технологий OLAPта DataMining для устранения указанных проблем. Приведена структура витрины данных и определены этапы дальнейшей работы.*

Ключевые слова: *экспертиза сортов растений, метод вариационной статистики, анализ данных, экспертная система, хранилище данных, витрина данных, технология OLAP, технология DataMining.*

Golub B.L.
candidate of technical sciences, Head of the Department of Computer Science Faculty of Information Technology NULES Ukraine

Leshchuk N.V.
graduate student, Institute of Ukrainian examination of plant varieties

Tsyba S.V.
graduate student of Computer Science Faculty of Information Technology NULES Ukraine

Annotation.*Considered shortcomings of the existing methods for examination of plant new varieties used by the Ukrainian Institute examination of plant varieties organizational and technological perspective. It is proposed the use of modern technology OLAP and Data Mining to address these problems. Presented the structure data marts and determined steps further work.*

Keywords: *examination of plant varieties, method of variation statistics, data analysis, expert system, data warehouse, data marts, technology OLAP, Data Mining technology.*
