doi: 10.15407/ukrbotj73.06.568

І.В. ГОНЧАРЕНКО
Інститут еволюційної екології НАН України
вул. акад. Лебедєва, 37, м. Київ, 03143, Україна
3604749@gmail.com

ЗАСТОСУВАННЯ МЕТОДУ DRSA – НЕПАРАМЕТРИЧНОГО КЛАСТЕРНОГО АНАЛІЗУ В КЛАСИФІКАЦІЇ РОСЛИННОСТІ

Goncharenko I.V. **Application of the DRSA technique, a non-parametric cluster analysis, in vegetation classification.** Ukr. Bot. J., 2016, **73**(6): 568–578.

Institute for Evolutionary Ecology, National Academy of Sciences of Ukraine
37, Acad. Lebedeva Str., Kyiv, 03143, Ukraine

Abstract. Advantages of the original clustering method of DRSA, or Distance-Ranked Sorting Assembling, for vegetation classification are discussed. Using ranks in determining distances between objects provides robust clustering in case of noisy and heterogeneous phytocoenotic data. Algorithm of objects agglomeration is based on ranking objects by the indices of freeness and connectedness as well as on assessing clusters within k-NN graph's framework. Clusters are assembled iteratively for some time to be finalized at the maximum of cluster's connectivity. We also consider in detail approaches to assess classification quality of phytocoenotic dataset including degree of cluster's (phytocoenon) compactess-distinctness and amount of differential species. We propose using nominal correlation coefficients to evaluate concordance of phytocoenotic classifications and contingency tables to compare frequencies of common releves between different classifications. Phytocoenon's compactness and distinctness are evaluated using well-known internal cluster validation indices, e.g. silhouette statistics. We introduced CDR-index (compactness / distinctness ratio) which is calculated from the score of average similarity of within-phytocoenon and between-phytocoenons releves. Total amount of faithful (differential) species and average amount of them per phytocoenon as floristic index of partitioning quality were used. We classified differential species on a statistical basis calculating specied-to-cluster fidelity index and selecting species with fidelity above defined fidelity's threshold. Using the sample phytocoenotic datasets we proved that both internal and floristic indices of classification quality improve after the exclusion of transient releves with ecotonic species composition. In the DRSA method, noise detection is carried out during cluster agglomeration; this objectifies rejecting ecotonic releves according to Braun-Blanquet approach as well as increases amount of differential species and thus improves phytocoenons interpretability.

Keywords: DRSA, cluster analysis, Braun-Blanquet approach, phytocoenon, quality of classification

Вступ

Класифікація таблиць фітоценотичних даних є початковим, аналітичним етапом класифікації рослинності. Використання методів автоматичної класифікації (кластерного аналізу) стикається з низкою труднощів і обмежень. Тому до 80-х рр. минулого сторіччя у європейській фітоценології панував підхід «ручного сортування» таблиць геоботанічних описів за методикою Браун-Бланке. З впровадженням комп'ютерних технологій для накопичення та обробки фітоценотичних даних інтерес до методів автоматичної класифікації у фітоценології почав зростати. В сучасних дослідженнях автоматична класифікація фітоценотичних даних, найчастіше з використанням ділячого політетичного алгоритму TWINSPAN (Hill, 1979; Hill, Šmilauer, 2005) передує ручному сортуванню. Автоматична класифікація покликана каналізувати процес подальшого ручного сортування, намітити «первинні» фітоценотичні кластери, які потім «доводяться» шляхом ручного сортування з використанням спеціальних геоботанічних комп'ютерних програм: Megatab (Hennekens, 1996), Ficen2 (Kosman et al., 1996) та ін. Як зовнішній модуль TWINSPAN використовується у програмі Juice (Tichý, 2002). Але слід пам'ятати, що TWINSPAN – це перш за все ординація, тому він чутливий до «шуму», а результат поділу на кожному кроці ділячого алгоритму істотно залежить від описів на протилежній частині градієнту: варто змінити співвідношення кількості різних описів і результат виявиться іншим.

Необхідність розробки нового методу викликана неможливістю чи неефективністю застосування до фітоценотичних даних строгих математичних методів, необхідністю обробки великих масивів даних широкого еколого-фітоценотичного діапазону з урахуванням їхньої неоднорідності, неповноти та зашумованості. Наявність випадкових видів, неповночленність фітоценозів, неоднорідність фітоценотичних даних та їхня неповнота – все це ро-

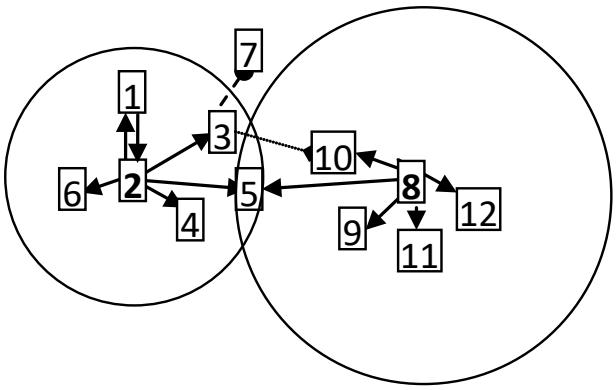

Рис. 1. Визначення *k*-найближчих сусідів
Fig. 1. Selection of *k*-nearest neighbors

бить фітоценотичні дані «проблемними» і вимагає застосування непараметричних методів. Ієрархічні агломеративні методи кластерного аналізу – неефективні для великих масивів даних, а також чутливі до вибору метрики чи алгоритму групування. Ітеративні методи кластерного аналізу, зокрема метод K-середніх, потребують апріорних знань про кількість кластерів у даних, а у фітоценолога така інформація найчастіше відсутня.

Для класифікації рослинності нами було розроблено алгоритм непараметричного кластерного аналізу «Distance-Ranked Sorting Assembling» (DRSA), метод «сортуючої зборки» з використанням рангів відстаней (Goncharenko, 2015a). Метод DRSA – агломеративний, неієрархічний метод кластерного аналізу. Математична основа його детально розглянута в окремих публікаціях (Goncharenko, 2015b, c). Особливості методу DRSA такі:

- відстань між об'єктами визначається рангами;
- результат групування мало залежить від обраної метрики чи коефіцієнта подібності;
- автоматичне визначення кількості кластерів у даних (немає необхідності задавати кількість кластерів, як у методі K-середніх, чи «розрізати» дендрограмму, як у агломеративних алгоритмах);
- щільні кластери (описи всередині фітоценонів значною мірою подібні за видовим складом);
- фільтрація шуму – визначення перехідних фітоценозів та їхнє виключення із кластерів;
- наявність параметру *k* (кількість найближчих сусідів, що враховуються у кожного об'єкта), яка дозволяє впливати на розміри та кількість кластерів.

Оцінка відстаней між об'єктами

Спочатку розраховуються коефіцієнти подібності описів за видовим складом. Можливе використання будь-яких з відомих коефіцієнтів флористичної подібності (Sokal, Sneath, 1963; Vasilevich, 1969; Goodall, 1973; Legendre P., Legendre L., 1998). У подальшому в кожного об'єкта визначається *k* найближчих сусідів. Якщо впорядкувати об'єкти за подібністю щодо певного об'єкта X і присвоїти їм ранги, то об'єкт, що має *k*-й ранг сусідства, є *k-найближчим сусідом* об'єкта X.

Використання рангів замість відстаней дає, з точки зору фітоценолога, важливі переваги. По-перше, у разі заміни коефіцієнта подібності на інший значення відстаней між об'єктами зміняться, але часто це не позначається на порядку розташування об'єктів (рангах) *A, B, C*, що забезпечує відносну стійкість кластерів. По-друге, при використанні еквівалентних коефіцієнтів (Semkin, 1979) ми отримаємо ідентичні класифікації. По-третє, наявність викидів (аномальних об'єктів) майже не впливає на результат. Крім того, використання рангів дозволяє застосовувати метод DRSA у випадку значного варіювання бета-різноманіття (щільності кластерів), а також щодо даних широкого еколого-фітоценотичного діапазону, коли інші методи, що спираються на абсолютні значення відстаней, малоефективні. Усе це робить метод DRSA робастним (англ. robust – міцний). Непараметричні методи прийнято вважати менш потужними, ніж параметричні, але у випадку різнорідних, неповних, зашумованих фітоценотичних даних, втрата потужності за рахунок виграшу у робастності є цілком виправданою.

На рис. 1 показано відбір найближчих сусідів при $k = 5$ у об'єктів 2 та 8.

При $k = 5$ для об'єкта 2 найближчими сусідами є об'єкти 1, 3, 4, 5, 6, а для об'єкта 8 – 5, 9, 10, 11, 12. Об'єкт 7 є найближчим сусідом об'єкта 3 при $k = 5$, але не для об'єкта 2. При $k = 6$ об'єкт 7 стане *k*-найближчим сусідом також і для об'єкта 2. В об'єктів 1 та 2, об'єкт 2 є найближчим сусідом об'єкта 1 та навпаки, що показано подвійною стрілкою. Відстань від центрального об'єкта до найвіддаленішого сусіда в об'єктів 2 та 8 різна, хоча $k = 5$ в обох випадках.

Групування об'єктів

Алгоритм групування у методі DRSA розроблявся виходячи з уявлень про «природну кластеризацію», тобто таку, яку б інтуїтивно побудувала людина, якби могла бачити розподіл точок у просто-

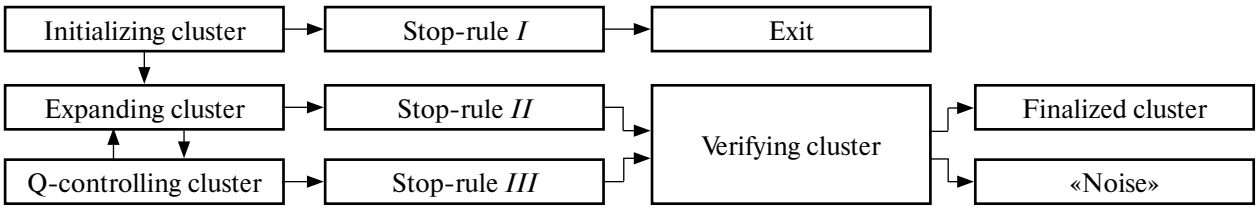

Рис. 2. Блок-схема алгоритму DRSA
Fig. 2. DRSA algorithm flowchart

рі. Комп'ютерний алгоритм відтворює цей процес шляхом сортування (ранжування) об'єктів (описів) і «збирання» з них кластерів (фітоценонів). Тому ми назвали метод DRSA «сортуючою зборкою» (англ. sorting assembling).

Сортування і відбір об'єктів спирається на індекси, які передають відстані «об'єкт–об'єкт» (індекс вільності) та «об'єкт–кластер» (індекс зв'язаності) у структурі *k-NN графа* (Goncharenko, 2015c). У структурі *k*-NN графа кластери DRSA нагадують кореляційні плеяди з однойменного методу П.В. Терентьєва, але плеяди виділяють при фіксованому значенні відстані, а у випадку DRSA це визначається порогом параметру *k*. Віднесення об'єкта до найближчого кластеру базується на тому ж принципі, що і у методі *k*-найближчих сусідів (Cover, Hart, 1967). Якщо певний об'єкт близький до кластеру, то серед його *k*-найближчих сусідів переважають об'єкти цього кластеру. Після ранжування відбирається черговий об'єкт, кластер нарощується і процес повторюється. Момент зупинки нарощування кластерів визначається максимізацією показника зв'язаності кластерів (плеяд) (Q-індекс). Поступово підвищуючи параметр *k* у методі DRSA, можна отримати серію кластерних рішень, що є «зрізами» кластерної структури на різних рівнях. Параметр *k* задається до початку групування: при більших значеннях *k* утворюється менше кластерів, але вони крупніші. Утворення кластерів відбувається почергово (послідовно): кластер проходить етапи ініціації, нарощування і фіналізації, після чого не змінюється. Наступний кластер ініціюється після фіналізації попереднього. У нарощуванні кластерів беруть участь лише вільні об'єкти, отже кластери не об'єднується, тому DRSA належить до неієрархічних методів кластерного аналізу.

На рис. 2 представлені основні етапи групування згідно до алгоритму DRSA.

Етап I. **Ініціація кластеру**. Утворення першого і чергового кластеру починається з одного об'єкта. Його вибір здійснюється за максимальним значенням *індексу вільності* (freeness index, FI) (Гончаренко, 2015b). Цей індекс – евристичний показник, що набуває максимального значення у об'єктів, розташованих далеко від утворених раніше кластерів, у центрі скупчень інших вільних об'єктів. Це дозволяє максимізувати відмежованість кластерів. Якщо в певний момент групування вільних об'єктів, що мають FI вищий за встановлений поріг, немає, ініціювати новий кластер неможливо (стоп-правило I), групування припиняється. Кількість утворених до цього моменту кластерів стає остаточною, а усі об'єкти поза кластерами визнаються шумом (перехідні описи).

Етап II. **Нарощування кластеру**. Після ініціації кластеру, він нарощується. На кожному кроці відбирається і приєднується один об'єкт з максимальним значенням *індексу зв'язаності* (connectedness index, CI) (Гончаренко, 2015b). Цей показник, що залежить від відстані між об'єктом і кластером, набуває максимальних значень у найближчих об'єктів. Приєднання на кожному кроці групування об'єктів з максимальним значенням CI максимізує щільність кластерів. Якщо об'єктів зі значенням CI, вищим за поріг, немає (стоп-правило II), нарощування кластеру припиняється (фіналізація).

Етап III. **Контроль якості кластеру**. Під час нарощування кластеру розраховується показник, що оцінює «якість» кластеру, Q-індекс (Goncharenko, 2015c). Оцінка кластерів (англ. cluster validation) традиційно проводиться по завершенню кластерного аналізу (Halkidi et al., 2001), тобто оцінює результат пост-фактум. У методі DRSA оцінка кластерів здійснюється під час групування. Її метою є визначення «оптимального» моменту для фіксації (фіналізації) кластеру.

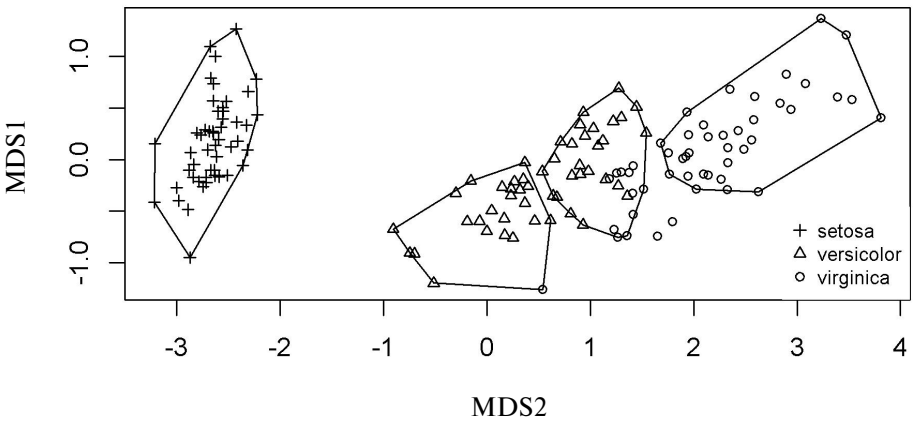

Рис. 3. Розподіл кластерів DRSA і трьох видів роду *Iris* набору даних «іриси Фішера»

Fig. 3. Allocation of clusters derived from DRSA and *Iris* species of Fisher's Iris dataset

Q-index залежить від *повноти*, *зв'язаності* та *відмежованості* кластеру, розрахунок яких базується на аналізі структури кластерів у *k*-NN графі. На початку росту кластер має низьку повноту (англ. integrity), оскільки більша частина об'єктів майбутнього кластеру вільна. Під час нарощування кластеру повнота зростає, але одночасно зменшується відмежованість (англ. separability) кластеру. Її можна оцінити кількістю зв'язків між вершинами *k*-NN графа з різних кластерів. Під час нарощування кластеру зростає його зв'язаність (англ. connectivity) – кількість зв'язків між вершинами *k*-NN графа одного кластеру. Максимізація Q-index (стоп-правило III) визначає момент фіналізації кластеру.

Оцінка якості класифікації фітоценотичних даних за кількісними критеріями

Після обробки фітоценотичного набору даних вкрай важливо оцінити якість класифікації (якість фітоценонів). Це дає можливість оцінити ефективність того чи іншого методу кластеризації, а також вибрати оптимальний поділ, якщо їх декілька. Оцінка якості проведеної класифікації фітоценотичних даних можлива:

- через візуальний аналіз меж кластерів (у площині ординації чи просторі ознак);
- за величиною кореляції з іншою, еталонною, класифікацією.
- за показниками щільності та відмежованості кластерів (фітоценонів);
- за кількістю диференціюючих видів.

Візуальний аналіз кластерів у ординаційній площині

Апробацію методів кластерного аналізу традиційно прийнято перевіряти класифікацією штучного набору даних «іриси Фішера». Ці дані (http://archive.ics.uci.edu/ml/datasets/Iris) містять інформацію про чотири ознаки будови квітки для 150 екземплярів трьох видів роду *Iris* L. Класифікуємо їх методом DRSA та співставимо розподіл об'єктів між кластерами та видами. Щоб оцінити відповідність класифікацій та непересічність кластерів, розглянемо положення кластерів у ординаційній площині 2-х перших осей багатовимірного шкалювання (нами використана функція metaMDS пакету vegan (Oksanen et al., 2010) середовища R), де кластери позначено полігонами по крайніх об'єктах (рис. 3).

Як бачимо, на рис. 3 кластери відмежовані. Отже, завдання кластерного аналізу – виділення відокремлених груп – вирішена. У класичному наборі даних було три види: *Iris setosa* Pall. ex Link, *I. versicolor* L., *I. virginica* L., які показано окремими позначеннями. Ми одержали чотири кластери, причому три з них чітко відповідають трьом видам, а четвертий становить збірну групу *I. versicolor* та *I. virginica*. Однак, з огляду на його відокремлене розташування, і він може вважатися самостійним. Таким чином, поєднуючи багатовимірне шкалювання (ординацію) та кластерний аналіз (класифікацію), що базуються на одній матриці відстаней, можна аналізувати відмежованість груп (класів, кластерів, фітоценонів), співставляти класифікації, використовуючи ординаційну площину у якості основи для візуального аналізу, прогнозувати наявність і формувати нові групи (класи) об'єктів, виявляти аномальні об'єкти та шум.

Таблиця 1. **Матриця коефіцієнтів подібності кластерів автоматичної класифікації DRSA і синтаксонів експертної класифікації на прикладі модельного набору даних 203 × 596**

Table 1. **Matrix of similarity coefficients between clusters of automatic classification derived from the DRSA technique and syntaxa of expert classification of the sample 203 × 596 dataset**

	01	02	03	04	05	00
32BA10	100	0	0	0	0	0
32BA03a	0	50	0	0	0	35
32BA05	0	0	60	15	0	34
32BA08	0	0	0	89	0	6
32BA02	0	0	0	0	88	7
32BA03b	0	32	0	0	0	8
32BA03c	0	7	10	0	0	18
32BA06	0	4	0	0	0	42
32BA07	0	41	0	0	0	14
32BA09	0	0	0	0	0	36

Примітка: Коди синтаксонів (Chytrý, Horák, 1997): код 32 – клас *Querco-Fagetea*, 32B – порядок *Quercetalia pubescenti-petraeae*, 32BA – союз *Quercion pubescenti-petraeae*, 32BA02 – асоціація *Pruno mahaleb-Quercetum pubescentis*, 32BA03 – *Sorbo torminalis-Quercetum*, 32BA03a – *Sorbo torminalis-Quercetum typicum*, 32BA03b – *Sorbo torminalis-Quercetum caricetosum humilis*, 32BA03c – *Sorbo torminalis-Quercetum poetosum*, 32BA05 – *Corno-Quercetum*, 32BA06 – *Potentillo albae-Quercetum*, 32BA07 – *Genisto pilosae-Quercetum petraeae*, 32BA08 – *Quercetum pubescenti-roboris*, 32BA09 – *Carici fritschii-Quercetum roboris*, 32BA10 – *Asplenio cuneifolii-Quercetum petraeae*.

Оцінка кореляції фітоценотичних класифікацій

Для вимірювання кореляції класифікацій існують кількісні індекси – коефіцієнти кореляції номінальних ознак. Відомі статистика Крамера (Cramer's V), індекс Фолкса-Меллоуса (FM-index) та ін. Індекси приймають значення або від –1 до +1 (ті, що враховують d-клітинку таблиці спряженості і вимірюють також негативну кореляцію), або від 0 до 1 (ті, що d-клітинку не враховують). Значення +1, або 100%, вказує на повну ідентичність двох класифікацій.

Класифікаційна належність фітоценозів (описів, об'єктів) до певних кластерів (фітоценонів, синтаксонів) – номінальна ознака, а зазначені індекси дозволяють оцінити «узгодженість» класифікацій. Якщо одна з класифікацій приймається за еталон, то розрахунок кореляції стає методом верифікації іншої класифікації. Значення індексів більше 0,8 можна прийняти як свідчення високої кореляції класифікацій. Якщо обидві класифікації рівнозначні і жодна з них не може вважатися еталоном, то висока кореляція – можливе свідчення природності кластерів, їх відповідності дійсній структурі даних: якщо різні методи дають схожі класифікації, ймовірно, кластери природні.

Щоб з'ясувати відповідність конкретних кластеру та класу, необхідно дослідити розподіл об'єктів альтернативних класифікацій, використовуючи M × N таблиці спряженості, де M та N – кількість груп (кластерів) порівнюваних класифікацій. У табл. 1 представлені коефіцієнти подібності для модельного набору даних (203 описів × 596 видів) між кластерами автоматичної класифікації DRSA (по горизонталі) та синтаксонами експертної класифікації Браун-Бланке (по вертикалі), що наведена у першоджерелі (Chytrý, Horák, 1997). Схожість пари «кластер-синтаксон» розрахована виходячи з кількості спільних описів, що увійшли до одного кластеру та синтаксону. Застосовано коефіцієнт Охаї (Ochiai, 1957). За результатами автоматичної класифікації усього було виділено п'ять фітоценонів та «шум», кластер «00». Синтаксони розташовували таким чином, щоб найбільший коефіцієнт подібності знаходився на умовній діагоналі. Для кращого візуального сприйняття у комірках табл. 1 вміщено гістограми.

Перші п'ять синтаксонів (32BA10, 32BA03a, 32BA05, 32BA08, 32BA02) з високою подібністю відповідають п'яти кластерам автоматичної класифікації (01-05), інші – більшою (32BA06, 32BA09) або меншою (32BA03b, 32BA03c, 32BA07) мірою скла-

Таблиця 2. **Оцінка щільності та відмежованості фітоценонів автоматичної класифікації DRSA на прикладі модельного фітоценотичного набору даних 210 × 574**

Table 2. **Assessment of phytocoenons compactness and distinctness of automatic DRSA classification of the sample 210 × 574 dataset**

No. cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
No. of releves	7	8	10	33	6	12	12	16	20
1	0,47	0,27	0,10	0,05	0,06	0,03	0,02	0,04	0,03
2	0,27	0,47	0,23	0,07	0,08	0,04	0,01	0,02	0,00
3	0,10	0,23	0,51	0,15	0,14	0,11	0,03	0,06	0,03
4	0,05	0,07	0,15	0,43	0,25	0,28	0,06	0,14	0,10
5	0,06	0,08	0,14	0,25	0,62	0,14	0,07	0,13	0,09
6	0,03	0,04	0,11	0,28	0,14	0,52	0,20	0,20	0,10
7	0,02	0,01	0,03	0,06	0,07	0,20	0,48	0,26	0,19
8	0,04	0,02	0,06	0,14	0,13	0,20	0,26	0,45	0,33
9	0,03	0,00	0,03	0,10	0,09	0,10	0,19	0,33	0,45
10	0,05	0,00	0,01	0,05	0,06	0,05	0,07	0,14	0,25
11	0,05	0,02	0,03	0,17	0,15	0,09	0,07	0,13	0,21
wcs *	0,47	0,47	0,51	0,43	0,62	0,52	0,48	0,45	0,45
bcs	0,27	0,27	0,23	0,28	0,25	0,28	0,26	0,33	0,33
CDR	0,28	0,28	0,39	0,21	0,42	0,30	0,30	0,15	0,15

* Розшифрування див. у тексті статті

даються переважно з шумових об'єктів (табл. 1). Синтаксон 32BA10 і кластер 01 мають повну відповідність. Переважне потрапляння описів декількох синтаксонів (32BA03a, 32BA03b, 32BA07) в один кластер 02 свідчить про значну їхню подібність. Таким чином, таблиці спряженості дозволяють оцінити відповідність експертних синтаксонів окремим кластерам автоматичної класифікації.

Оцінка щільності та відмежованості фітоценонів

Головним завданням кластерного аналізу є виділення щільних та відмежованих груп об'єктів. Для оцінки якості кластерів у математичній статистиці запропонована значна кількість індексів, які прийнято називати внутрішніми, оскільки вони базуються виключно на матриці відстаней (Rendon et al., 2011). Серед найбільш відомих статистика силуетів, індекс Калінського-Харабаша (Calinski, Harabasz, 1974) та ін. При розрахунку внутрішніх індексів враховують відстані від певного об'єкту до об'єктів «свого» кластеру та до об'єктів у інших кластерах. Отже, середня подібність за видовим складом описів усередині фітоценонів у порівнянні з подібністю цих описів з описами з інших фітоценонів є аналогом згаданих внутрішніх критеріїв у фітоценології.

Нами запропоновано індекс *CDR* (compactness/distinctness ratio) (формула 1). Він дозволяє оцінювати щільність окремих фітоценонів, оскільки вираховується як середнє значення подібності описів за видовим складом, тому його можна вважати індексом флористичної гомогенності ценофлор виділених фітоценонів. Для оцінки якості класифікації фітоценотичного набору даних в цілому запропоновано індекс *PQI* (partitioning quality index), який розраховується як середнє *CDR* усіх кластерів (формула 2):

$$CDR = (wcs - max\,(bcs)) / (wcs + max\,(bcs)) \quad (1),$$

$$PQI = avg\,(CDR) = \sum CDR / N \quad (2),$$

де *wcs* (within-cluster similarity) – подібність описів усередині кластеру (фітоценону); *bcs* (between-clusters similarity) – подібність описів різних кластерів (фітоценонів); *CDR* (compactness/distinctness ratio) – співвідношення щільність–відмежованість; *PQI* (partitioning quality index) – індекс якості поділу, *N* – загальна кількість кластерів.

У табл. 2 наведено середні значення коефіцієнтів подібності між описами усередині фітоценонів (на діагоналі), виділених за результатами DRSA, та між описами різних фітоценонів (поза діагоналлю). Для розрахунків середнього значення подібності між описами усередині та між кластерами було узято вхідну матрицю подібності за видовим складом між описами (210 описів), розраховану за коефіцієнтом Охаї, після чого здійснили розрахун-

Таблиця 3. **Кількість вірних видів фітоценонів автоматичної класифікації DRSA на прикладі набору даних 780 × 728 за різних значень *k***

Table 3. **Number of faithful species of phytocoenons derived from automatic DRSA classification of the sample 780 × 728 dataset at different values of the parameter *k***

Параметр *k*	Описи в кластерах, %	N_total	N_good	A_total	A_*avg*
3	49	49	11	125	2.6
4	52	44	17	133	3
5	53	43	18	133	3.1
6	**53**	**35**	**24**	**145**	**4.1**
7	54	30	23	133	4.4
8	54	26	21	141	5.4
9	51	25	23	142	5.7
10	50	22	19	111	5

Примітка: N_total – загальна кількість кластерів (фітоценонів), N_good – кількість «добрих» кластерів (фітоценонів), що мають мінімум два вірних види, для яких fidelity > 50%, A_total – загальна кількість вірних видів усіх фітоценонів, A_*avg* – кількість вірних видів у середньому на фітоценон.

ки *wcs*, *bcs*, *CDR* та *PQI*. У якості модельного набору даних обрано дані з лісової рослинності національного парку Тайяталь, Австрія (Chytrý, Vicherek, 1995).

Як видно з табл. 2, найбільші значення подібності розташовуються на діагоналі, отже у всіх фітоценонів видовий склад більш подібний у описів всередині одного фітоценону, ніж поміж фітоценонами. Фітоценони значною мірою гомогенні за видовим складом, подібність між описами усередині кластерів коливається від 0,43 до 0,62 і є значною. Виділені фітоценони мають приблизно однаковий «об'єм» або рівень подібності між описами усередині фітоценонів, тобто відповідають одному рангу.

Приклад розрахунку *CDR*: для кластеру 1 *wcs* = 0,47, найближчим до нього є кластер 2 (середня подібність між описами цих кластерів *bcs* = 0,27). Таким чином, *CDR* = (0,47 – 0,27) / (0,47 + 0,27) = 0,28. Індекс *CDR* приймає значення від –1 до +1. Позитивний індекс *CDR* свідчить про відмежованість фітоценотичного кластеру. Найбільш щільним серед 11 виділених кластерів є кластер 5: *wcs* = 0,62, найменш щільним – кластер 4, *wcs* = 0,43. Індекс *CDR* через значну подібність видового складу фітоценонів 8 та 9 найменший в кластеру 8 і дорівнює 0,15. Загалом, індекс *CDR* в одержаних фітоценонів коливається від 0,15 (кластер 8) до 0,42 (кластер 5), а з урахуванням усіх 11 фітоценонів *PQI* = *avg* (*CDR*) = 0,29. Це свідчить про задовільну якість класифікації.

Оцінка якості фітоценотичної класифікації кількістю вірних (диференціюючих) видів

Метод DRSA здійснює автоматичну класифікацію фітоценозів (описів), тобто виділення фітоценонів. Інтерпретація останніх проводиться за видовим складом. Говорити про природність фітоценотичних кластерів і їх екологічну своєрідність можна лише у тому випадку, якщо за результатами класифікації видів у фітоценонів численні диференціюючі види. Їх кількість є флористичним критерієм якості класифікації. Оцінка діагностичної сили видів здійснюється на статистичній основі розрахунком індексів вірності (англ. – fidelity index) (Bruelheide, 2000; Chytrý et al., 2002; De Cáceres et al., 2008). Згідно до підходу, що одержав назву Optimclass (Tichy, 2010), кількість кластерів, а також якість класифікації пропонується визначати за максимальною кількістю вірних (зі значеннями *fidelity* вище порогу) видів або за кількістю «добрих» фітоценонів, у яких кількість вірних видів більша за обраний поріг.

У табл. 3 показано індикативні показники класифікації фітоценотичного набору даних 780 описів × 728 видів (Goncharenko, 2003) за методом DRSA із різним значенням *k*. Ми проводили кластерний аналіз із певним значенням *k,* фіксували кількість кластерів (N_total) та описів (об'єктів), включених до складу кластерів. Потім виконували класифікацію видів. Види із значенням fidelity > 50% включали до списку вірних видів та підраховували їхню загальну кількість (A_total) та показник середньої кількості вірних видів на один фітоценон (A_*avg*).

Як бачимо з табл. 3, найбільша кількість вірних видів – 145 (або 19,9% усіх видів) спостерігається у випадку *k* = 6. При цьому утворюється 35 кластерів, 24 з них (або 69% загальної кількості) мають щонайменше два вірних види. Таким чином, для даного фітоценотичного набору оптимальним є значення *k* = 6 для кластерного аналізу за методом DRSA. Співвідношення класифікованих описів/шуму із зростанням параметру *k* лишається майже незмінним (від 49% до 54%), оскільки цей показник залежить від особливостей даних (головним чином

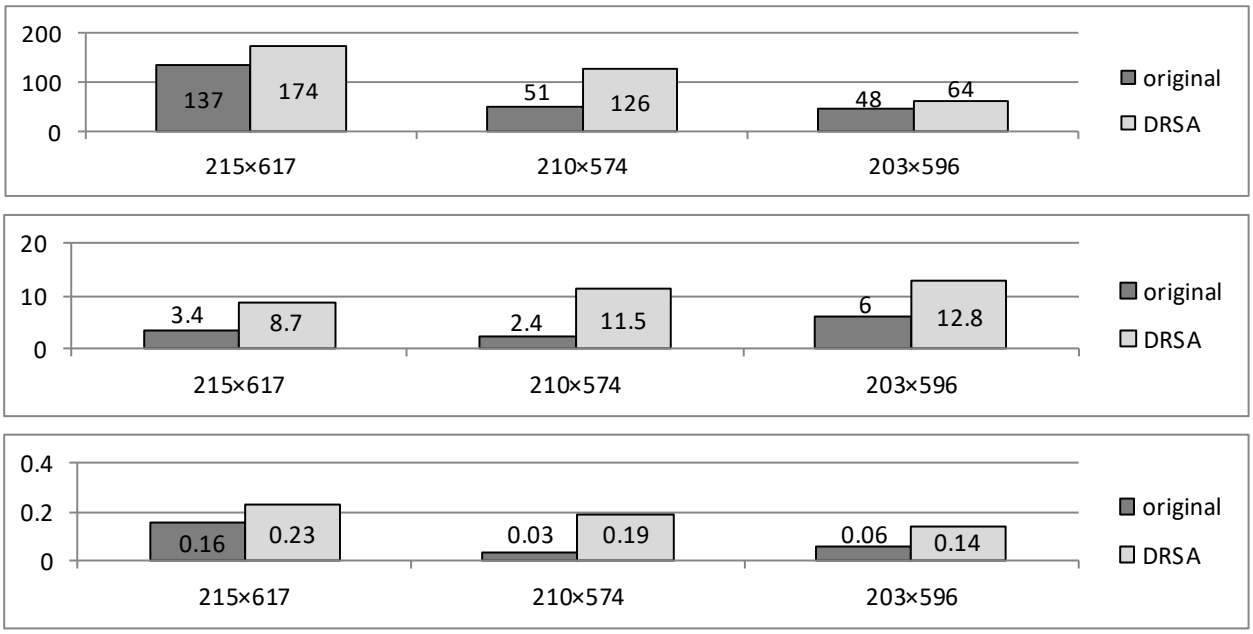

Рис. 4. Загальна кількість вірних видів (*a*), кількість вірних видів у середньому на фітоценон (*b*), статистика силуетів для автоматичної (*DRSA*) та еталонної (оригінальної) класифікації (*c*)

Fig. 4. Total amount of faithful species (*a*), average amount of faithful species per phytocoenon (*b*), silhouette statistics for automatic (DRSA) and reference (original) classification (*c*)

бета-різноманітності даних) і не залежить від кількості виділених кластерів. Із зростанням параметру *k* кількість кластерів монотонно зменшується і зростає також показник A_*avg*. Це пов'язано з тим, що одночасно із укрупненням фітоценотичних кластерів, вони стають більш відмінними за видовим складом. Як наслідок, кількість диференціюючих видів на фітоценон A_*avg* зростає. Аналогічну тенденцію ми спостерігаємо при переході від рівня асоціацій до рівня союзів, порядків і т. ін.

Вплив бракування перехідних фітоценозів на якість класифікації фітоценотичних даних

Згідно до методики Браун-Бланке бракування перехідних описів (фітоценозів із екотонним видовим складом) становить невід'ємну частину аналітичного етапу класифікації. Воно може складати до 60% загальної кількості описів залежно від даних: збільшення середньої подібності описів (зменшення еколого-фітоценотичного діапазону), як правило, призводить до збільшення бракування.

Метод DRSA здійснює визначення шумових об'єктів (перехідних описів) під час та по завершенню групування. Цей процес відбувається на основі інформації з матриці відстаней між об'єктами, таким чином здійснюється на кількісній основі. Це значно об'єктивізує визначення перехідних фітоценозів, адже у методі Браун-Бланке воно відбувається на розсуд фітоценолога і є суб'єктивним. Відсоток описів, включених до кластерів DRSA, склав 49–54%, відповідно друга частина описів – шум (див. табл. 3).

Бракування перехідних описів має важливе значення. Здебільшого фітоценотичні набори даних континуальні. Континуум є фундаментальною основою організації рослинного покриву і трапляється значно частіше, ніж дискретні дані з чітко оформленими синтаксонами. Оскільки вірними (диференціюючими) видами є види, що тяжіють до одного синтаксону (фітоценону) та відсутні в інших, кількість диференціюючих видів, як правило, незначна, але збільшується внаслідок бракування перехідних описів. При цьому відмінності видового складу між фітоценонами зростають, збільшується кількість статистично вірних видів, зростають показники якості класифікації. Отже бракування дозволяє суттєво покращити результат класифікації.

На рис. 4 представлені показники кількості вірних видів та статистики силуетів для трьох мо-

дельних фітоценотичних наборів фітоценотичних даних 215 × 617, 210 × 574 та 203 × 596 після автоматичної їхньої класифікації за методом DRSA. Для порівняння наведено аналогічні показники для цих самих даних, розраховані для оригінальних авторських (еталонних) класифікацій. Набір даних № 1 – 215 описів × 617 видів, рослинність у долинах річок Ослави, Їглави та Рокитної (Чехія) (Chytrý, Vicherek, 1996), набір даних № 2 – 210 описів × 574 види, лісова рослинність Національного парку Тайаталь (Австрія) (Chytrý, Vicherek, 1995), набір даних № 3 – 203 описи × 596 видів, термофільні ліси Моравії (Чехія) (Chytrý, Horák, 1997).

Флористичний критерій (рис. 4, *a*, *b*), а саме – кількість вірних видів, та математичний критерій (рис. 4, *c*) виявилися кращими, ніж для оригінальних класифікацій, наведених чеськими фітоценологами. Цей факт пояснюється тим, що у випадку класифікації DRSA до кластерів увійшли не всі описи, частина їх була виключена зі складу фітоценонів (шум). Так, для фітоценотичного набору даних № 1, що нараховував 215 описів, до результуючих 20 фітоценотичних кластерів увійшло 169 описів (або 79% їхньої загальної кількості). Отже, бракування становило 21% описів. Для порівняння у оригінальній роботі (Chytrý, Vicherek, 1996) було виділено 40 синтаксонів рангу асоціації та субасоціації. Таким чином, кількість фітоценонів у випадку DRSA менша вдвічі. Але при цьому загальна кількість вірних видів (при порозі fidelity > 50%) зросла з 137 до 174 (збільшилася у 1,27 рази) (рис. 4, *a*), у середньому на фітоценон – з 3,4 видів/фітоценон до 8,7 видів/фітоценон (збільшилася в 2,56 рази) (рис. 4, *b*). Аналогічно і для статистики силуетів: вона збільшилася з 0,16 до 0,23 (у 1,44 рази) (рис. 4, *c*). Таким чином, унаслідок бракування описів та укрупнення фітоценонів вдалося покращити індекси якості класифікації в порівнянні з оригінальними класифікаціями, наведеними у першоджерелах. Чи правильно визначаються перехідні описи під час класифікації DRSA, адже цей процес відбувається без участі експерта? Якщо би з того ж масиву даних ми видалили 21% описів, відібраних випадковим чином, то показники би якості класифікації не збільшилися. Отже, у DRSA перехідні описи (шум) визначаються вірно, оскільки зростає кількість вірних видів.

Висновки

Нами розглянуто метод кластерного аналізу, що має переваги для класифікації рослинності. Як відомо, до непараметричних методів вдаються у випадку зашумованих, неоднорідних, неповних даних, таких, що відхиляються від нормального розподілу.

Під час групування у методі DRSA частина описів (об'єктів) виключається зі складу кластерів (фітоценонів), т. з. шум. Визначення перехідних описів (шуму) здійснюється на кількісній основі і це об'єктивізує бракування перехідних описів, покращує флористичні та математичні критерії якості класифікації, дозволяє отримувати більш дискретні фітоценотичні кластери, які мають численні диференціюючі види і краще інтерпретуються.

У цій статті ми також розглянули різні підходи до оцінки якості фітоценотичної класифікації. Кожний з аспектів оцінки якості класифікації доповнює інший. Так, використовуючи коефіцієнти номінальної кореляції або таблиці спряженості, можна порівнювати декілька фітоценотичних класифікацій. За наявності матриці відстаней між описами за видовим складом можна оцінити щільність та відмежованість одержаних фітоценонів. Збільшення показника середньої подібності між описами одного фітоценону (синтаксону) у порівнянні з подібністю інших фітоценонів (синтаксонів) свідчить про якісний поділ. Використання кількості диференціюючих видів дозволяє не лише оцінювати якість класифікації за флористичним критерієм, а й проводити порівняльний аналіз ценофлор синтаксонів.

Таким чином, розглянутий метод кластерного аналізу DRSA, є перспективним при аналізі фітоценотичних зашумованих, неповних, багатоознакових, різнорідних даних.

СПИСОК ЛІТЕРАТУРИ

Bruelheide H. A new measure of fidelity and its application to defining species groups, *J. Veget. Sci.*, 2000, **11**: 167–178.

Calinski R.B., Harabasz J. A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 1974, **3**: 1–27.

Chytrý M., Horák J. Plant communities of the thermophilous oak forests in Moravia, *Preslia*, 1997, **68**: 193–240.

Chytrý M., Tichý L., Holt J., Botta-Dukát Z. Determination of diagnostic species with statistical fidelity measures, *J. Veget. Sci.*, 2002, **13**: 79–90.

Chytrý M., Vicherek J. *Lesní vegetace Národního parku Podyjí/Thayatal. Die Waldvegetation des Nationalparks Podyjí/Thayatal*, Praha, 1995, 166 pp.

Chytrý M., Vicherek J. Přirozená a polopřirozená vegetace údolí řek Oslavy, Jihlavy a Rokytné, *Přírod. Sborn. Zápa domorav. Muz. Třebíč*, 1996, **22**: 1–125.

Cover T.M., Hart P.E. Nearest neighbor pattern classification, *Inform. Theory*, 1967, **13**: 21–27.

De Cáceres M., Font X., Oliva F. Assessing diagnostic species value in large data sets: A comparison between phi-coefficient and Ochiai index, *J. Veget. Sci.*, 2008, **19**: 779–788.

Goncharenko I.V. Analiz roslynnogo pokryvu pivnichnoskhidnogo Lisostepu Ukrainy. Monografiya. In: *Ukr. Phytosoc. Col.* (spec. issue), 2003, **1**(19): 203 pp. [Гончаренко І.В. Аналіз рослинного покриву північно-східного Лісостепу України. Монографія // *Укр. фітоценот. зб.* (спец. вип.). – 2003. – **19**(1). – 203 с.].

Goncharenko I.V. *DRSA (distance-ranked sorting assembling) – metod sortuyuchogo klasternogo analizu*. Svidotstvo pro reyestratsiyu avtorskogo prava, № 58837, publ. 26.02.2015, 2015a, Byull. no 36. [Гончаренко І.В. *DRSA (distance-ranked sorting assembling) – метод сортуючого кластерного аналізу* // Свід-во про реєстрацію авторського права на збірку наукових творів № 58837 від 26.02.2015 р. – 2015a. – Бюл. № 36.].

Goncharenko I.V. *Reports of the National Academy of Sciences of Ukraine*, 2015b, **9**: 129–136. [Гончаренко І.В. Метод «сортуючої» кластеризації (DRSA) для класифікації рослинності // *Доп. НАН України*. – 2015b. – **9**. – С. 129–136].

Goncharenko I.V. *Vegetation of Russia*, 2015c, **27**: 125–138. [Гончаренко И.В. DRSA: алгоритм неиерархической кластеризации с использованием k-NN графа и его применение в классификации растительности // *Растительность России*. – 2015c. – **27**. – С. 125–138].

Goodall D.W. Numerical classification. In: *Handbook of vegetation Science. Part V: Ordination and Classification of Vegetation*. Ed. R.H. Whittaker, The Hague: Junk, 1973, pp. 105–156.

Halkidi M., Batistakis Y., Vazirgiannis M. On Clustering Validation Techniques, *J. Intell. Inform. Systems*, 2001, **17**: 107–145.

Hennekens S.M. MEGATAB – a visual editor for phytosociological tables. Version 1.0. October 1996. Ulft., 1996, 11 pp.

Hill M.O. *TWINSPAN – A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Program manual,* Ithaca; New York: Cornell Univ., 1979, 90 pp.

Hill M.O., Šmilauer P. *TWINSPAN for Windows version 2.3*, Huntingdon & České Budějovice: Centre for Ecology and Hydrology & Univ. of South Bohemia, 2005, 29 pp.

Kosman Ye.H., Sirenko I.P., Solomakha V.A., Shelyah-Sosonko Yu.R. *Ukr. Bot. J.*, 1991, **48**(2): 98–104. [Косман Є.Г., Сіренко І.П., Соломаха В.А., Шеляг-Сосонко Ю.Р. Новий комп'ютерний метод обробки описів рослинних угруповань // *Укр. ботан. журн.* – 1991. – **48**(2). – С. 98–104].

Legendre P., Legendre L. *Numerical ecology*, 2nd English ed., Amsterdam: Elsevier, 1998, 853 pp.

Ochiai A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Japan. Soc. Fish Sci.*, 1957, **22**(9): 526–530.

Oksanen J., Blanchet F.G., Kindt R., Legendre P., O'Hara R.G., Simpson G.L., Solymos P., Stevens M.H.H., Wagner H. Vegan: Community Ecology Package, 2010, available at: http://cran.r-project.org/web/packages/vegan/ (accessed 22 March 2016).

Rendon E. Abundez I., Arizmendi A., Quiroz E.M. Internal versus external cluster validation indices, *Intern. J. Computers and Communications*, 2011, **5**(1): 27–34.

Semkin B.I. Ekvivalentnost mer blizosti i ierarkhicheskaya klassifikatsiya mnogomernykh dannykh. In: *Ierarkhicheskie klassifikatsionnye postroeniya v geograficheskoy ekologii i sistematike*. Ed. B.I. Semkin, Vladivostok, DVNTs AN USSR, 1979, pp. 97–112. [Сёмкин Б.И. Эквивалентность мер близости и иерархическая классификация многомерных данных // *Иерархические классификационные построения в географической экологии и систематике* / Отв. ред. Б.И. Семкин. – Владивосток: ДВНЦ АН СССР. – С. 97–112].

Sokal R., Sneath P. *Principles of Numerical Taxonomy*, San Francisco, CA: Wit. Freeman, 1963, 573 pp.

Tichý L. JUICE, software for vegetation classification, *J. Veget. Sci.*, 2002, **13**: 451–453.

Tichý L., Chytrý M., Hájek M., Talbot S.S., Botta-Dukát Z. OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities, *J. Veget. Sci.*, 2010, **21**: 287–299.

Vasilevich V.I. *Statisticheskie metody v geobotanike*, Leningrad: Nauka, 1969, 232 pp. [Василевич В.И. *Статистические методы в геоботанике*. – Л.: Наука, 1969. – 232 с.].

Рекомендує до друку
Я.П. Дідух

Надійшла 04.04.2016

Гончаренко І.В. **Застосування методу DRSA – непараметричного кластерного аналізу в класифікації рослинності.** – Укр. ботан. журн. – 2016, **73**(6): 568–578.

Інститут еволюційної екології НАН України
вул. акад. Лебедєва, 37, м. Київ, 03143, Україна

Розглядаються переваги застосування методу кластерного аналізу «Distance-Ranked Sorting Assembling» (DRSA) для класифікації рослинності. Використання рангів при визначенні відстаней між об'єктами забезпечує робастність і ефективність при обробці зашумованих, різнорідних фітоценотичних даних. Алгоритм групування об'єктів базується на ранжуванні об'єктів за індексами вільності та зв'язаності і виділенні кластерів у структурі *k*-NN графа. Нарощування кластерів припиняється по досягненню максимуму зв'язаності кластерів. Детально розглядаються підходи до оцінки якості класифікації фітоценотичних даних – за показниками щільності та відмежованості кластерів (фітоценонів), за кількістю диференціюючих видів. Для оцінки кореляції фітоценотичних класифікацій пропонується використовувати коефіцієнти кореляції номінальних ознак та таблиці спряженості альтернативних класифікацій. Оцінювати щільність та відмежованість фітоценонів пропонується з використанням внутрішніх індексів валідації кластерів, зокрема статистики силуетів. Запропоновано індекс *CDR* (compactness / distinctness ratio), який враховує співвідношення подібності описів за видовим складом всередині фітоценонів та між фітоценонами. Загальна кількість диференціюючих видів та їхня середня кількість на фітоценон використані як флористичний критерій для оцінки якості класифікації. Виділення диференціюючих видів проведено на статистичній основі з використанням індексів вірності видів. На модельних фітоценотичних наборах даних показано, що бракування перехідних описів покращує і внутрішні, і флористичні критерії якості класифікації.

Ключові слова: DRSA, кластерний аналіз, метод Браун-Бланке, фітоценон, якість класифікації

Гончаренко И.В. **Применение метода DRSA – непараметрического кластерного анализа в классификации растительности.** – Укр. ботан. журн. – 2016, **73**(6): 568–578.

Институт эволюционной экологии НАН Украины
ул. акад. Лебедева, 37, г. Киев, 03143, Украина

Рассматриваются преимущества использования метода кластерного анализа «Distance-Ranked Sorting Assembling» (DRSA) в классификации растительности. Использование рангов при определении расстояний между объектами обеспечивает робастность и эффективность при обработке зашумленых, разнородных фитоценотических данных. Алгоритм группировки объектов базируется на ранжировании объектов по индексам свободности-связанности и выделении кластеров в структуре *k*-NN графа. Наращивание кластеров прекращается при достижении максимума связности кластеров. Подробно рассматриваются подходы к оценке качества классификации фитоценотических данных – с использованием индексов плотности-обособленности кластеров (фитоценонов) и по количеству дифференцирующих видов. Для оценки корреляции фитоценотических классификаций предлагается использовать коэффициенты корреляции номинальных признаков и таблицы сопряженности альтернативных классификаций. Оценить плотность и обособленность фитоценонов предлагается с использованием внутренних индексов валидации кластеров, в частности статистики силуэтов. Предложен индекс *CDR* (compactness / distinctness ratio), учитывающий соотношение сходства описаний по видовому составу внутри фитоценонов и между фитоценонами. Общее количество дифференцирующих видов и их среднее количество на фитоценон используются как флористический критерий оценки качества классификации. Выделение дифференцирующих видов проведено на статистической основе с использованием индексов верности видов. На модельных фитоценотических наборах данных показано, что браковка переходных описаний улучшает и внутренние, и флористические критерии качества классификации.

Ключевые слова: DRSA, кластерный анализ, метод Браун-Бланке, фитоценон, качество классификации