Оцінка якості фітоцено тичної класифікації (теоретико-методичний аспект)

ІГОР ВІКТОРОВИЧ ГОНЧАРЕНКО

GONCHARENKO I.V. (2016). **Quantifying the quality of phytocoenotic classification (theoretical-methodological aspect)**. *Chornomors'k. bot. z.,* **12** (1): 41-50. doi:10.14255/2308-9628/16.121/4.

This article reviews quantitative approaches to assessing the quality of phytocoenons and phytocoenotic classification. Mathematical criterion is based on calculating phytocoenon's compactness / distinctness ratio computed from a distance matrix between releves by species composition. Floristic criterion assumes evaluating diagnostic power of species and takes into account total amount of differential species which are classified statistically with fidelity indexes. We also considered related methods which European phytocoenologists applied for the same purpose – among them indexes of sharpness and uniqueness of syntaxon and the Optimclass approach. We measured resemblance of phytocoenotic classifications of the dataset using contingency tables and nominal correlation coefficients. We determined stability of phytocoenons and the robustness of the cluster topology using bootstrapping methods.

Key words: classification of vegetation, cluster analysis, phytocoenon, classification quality indexes

ГОНЧАРЕНКО І.В. (2016). **Оцінка якості фітоценотичної класифікації (теоретико-методичний аспект).** *Чорноморськ. бот. ж.,* **12** (1): 41-50. doi:10.14255/2308-9628/16.121/4.

Розглянуто підходи до оцінки якості фітоценонів та фітоценотичної класифікації за кількісними показниками. Математичний критерій якості класифікації базується на оцінці співвідношення щільності і відмежованості фітоценонів з використанням матриці відстаней між описами за видовим складом. Флористичний критерій ґрунтується на класифікації видів за діагностичним значенням, кількості диференціюючих видів і індексах вірності видів. Розглянуто підходи, що використовуються у європейській фітоценології – індекси визначеності, унікальності синтаксонів, підхід Optimclass. Подібність фітоценотичних класифікацій набору даних оцінюється за таблицями спряженості і коефіцієнтами номінальної кореляції. Стійкість фітоценонів і робастність кластер-структури визначається з використанням методів бутстреппінгу.

Ключові слова: класифікація рослинності, кластерний аналіз, фітоценон, індекси якості класифікації

ГОНЧАРЕНКО И.В. (2016). **Оценка качества фитоценотической классификации (теоретико-методический аспект).** *Черноморск. бот. ж.*, **12** (1): 41-50. doi:10.14255/2308-9628/16.121/4.

Рассмотрены подходы к оценке качества фитоценонов и фитоценотической классификации по количественным показателям. Математический критерий качества классификации базируется на оценке соотношения плотности и обособленности фитоценонов с использованием матрицы расстояний между описаниями по видовому составу. Флористический критерий основывается на классификации видов по диагностическому значению, количеству дифференцирующих видов и индексах верности видов. Рассмотрены подходы, используемые в европейской фитоценологии – индексы определенности, уникальности синтаксонов, подход Optimclass. Сходство фитоценотических классификаций набора данных оценивается по таблицам сопряженности и коэффициентами номинальной корреляции. Устойчивость

Чорноморськ. бот. ж., **12** (1): 41-50.

фитоценонов и робастность кластер-структуры определяется с использованием методов бутстрэппинга.

Ключевые слова: классификация растительности, кластерный анализ, фитоценон, индексы качества классификации

Класифікація рослинності включає емпіричний (польовий), аналітичний та синтетичний етапи. При індуктивному підході на аналітичному етапі спочатку формують первинну об'єднану таблицю геоботанічних описів, а потім, застосовуючи різноманітні методи, виділяють блоки подібних за видовим складом описів – фітоценони. На жаль, у роботах вітчизняних геоботаніків якість одержаних фітоценонів (синтаксонів) і проведеної класифікації не оцінюється кількісно. Це ускладнює прийняття рішень на етапі інтерпретації, оскільки не зрозуміло чи можна було б для того ж набору даних отримати кращий результат, наприклад, з використанням інших методів класифікації чи з меншою або більшою кількістю фітоценотичних кластерів (фітоценонів). Запровадження кількісних індексів, які б оцінювали якість фітоценонів, кожного окремо та класифікації в цілому, є неодмінною умовою об'єктивізації фітоценотичної класифікації.

Фітоценотична класифікація на флористичній основі має перш за все екологічний базис. Тому фітоценони, що виділяються при класифікації, повинні бути максимально відмінними екологічно і флористично. При цьому описи, що об'єднані в один фітоценон, повинні бути максимально подібними (гомогенними) за видовим складом. Масштаб фітоценотичної класифікації повинен бути настільки детальним, наскільки це можливо – так, щоб при подальшій спробі розділити фітоценони, виявити і пояснити екологічні фактори формування їх видового складу і екотопологічної приуроченості було б неможливо.

Підходи до оцінки якості класифікації ми умовно поділимо на суто математичні та біологічні (фітоценологічні). Математичні підходи широко застосовуються у методах кластерного аналізу. Так, у ітеративних методах кластерного аналізу, зокрема методі К-середніх, використовують різноманітні індекси якості поділу для того щоб з'ясувати, яка кількість кластерів для конкретного набору даних є оптимальною, адже метод К-середніх відрізняє те, що він потребує вказувати кількість кластерів до початку групування. Якщо вказати «неоптимальну» кількість кластерів, ми отримаємо неоптимальний поділ. У агломеративних методах кластерного аналізу будується повне дерево об'єднання, але для визначення рівня «розрізання» дендрограмми, також необхідно оцінити якість класифікації на декількох рівнях і вибрати оптимальний рівень.

У математичній статистиці індекси (критерії) якості поділу або, як прийнято їх називати, критерії валідації кластерів (англ. cluster validation) прийнято ділити на внутрішні та зовнішні [HALKIDI et al., 2001; RENDÓN et al., 2011]. Перші для оцінки якості кластерів використовують матрицю відстаней між об'єктами, тобто залучають лише ту інформацію, яка була використана у кластерному аналізі. Зовнішні критерії базуються на порівнянні одержаної класифікації з певними класами (групами) об'єктів того ж набору даних, що відомі апріорно, тому і називаються вони зовнішніми критеріями. Класичний приклад – «іріси Фішера», модельний набір даних для апробації методів кластерного, дискримінантного та інших багатовимірних методів.

Біологічний підхід до оцінки якості класифікації полягає в тому, що фітоценотичні кластери являють не лише групу описів, а і певний набір видів, що трапляються переважно у цих описах і відсутні у інших (ценофлора). Таким чином, крім суто математичного підходу, фітоценони можна аналізувати за флористичним критерієм: наявність численних диференціюючих видів є одночасно і критерієм якості

фітоценонів, і основою для порівняльно-флористичного аналізу фітоценонів (синтаксонів).

У даному повідомленні ми не ставили за мету зробити всебічний аналіз підходів до оцінки якості поділу і кластерної структури. Ми лише покажемо деякі можливості та необхідність такої оцінки при обробці фітоценотичних даних.

Оцінка фітоценонів за показниками «щільності-відмежованості»

Кожен кластер можна уявити як сукупність точок у просторі ознак [MANDEL, 1988]. Конфігурацію кластера можна описати положенням його центру та дисперсією відстаней. Щільність кластерів і їх відмежованість (непересічність) вважається ознакою якісного поділу.

Фітоценон, або кластер, теж можна охарактеризувати щільністю і відмежованістю. Задача оцінки певною мірою спрощується завдяки тому, що перед початком автоматичної класифікації та ординації фітоценозів зазвичай розраховують матрицю подібності (або відстаней) між описами за видовим складом. Для оцінки щільності кластерів достатньо розрахувати середню подібність між описами в межах кожного фітоценону, а для оцінки відмежованості – порівняти її із подібністю описів з різних фітоценонів.

Нами запропоновано індекс CDR, compactness / distinctness ratio, співвідношення «щільність-відмежованість», у якому порівнюється подібність описів усередині фітоценону, виділеного за результатами автоматичної чи експертної класифікації конкретного набору фітоценотичних даних (первинної таблиці з геоботанічними описами), з подібністю описів цього ж фітоценону щодо описів іншого найближчого фітоценону [GONCHARENKO, 2015].

Нехай, *wcs*, within-cluster similarity – середня подібність описів усередині кластеру (фітоценону); *bcs*, between-clusters similarity – середня подібність описів з різних кластерів (фітоценонів). Для кожного фітоценону визначається найбільш близький за видовим складом інший фітоценон, а потім розраховується індекс CDR наступним чином:

$$\mathbf{CDR = (wcs - max\ (bcs)) / (wcs + max\ (bcs))} \qquad (1)$$

Індекс CDR приймає значення від –1 до +1. Якщо CDR > 0, то кластер (фітоценон) є прийнятним, але чим більше значення CDR, тим якіснішим є кластер (фітоценон).

Щоб оцінити якість поділу в цілому, необхідно узяти до уваги щільність-відмежованість кожного кластеру (фітоценону). Індекс PQI, partitioning quality index – індекс якості поділу, розраховується як середнє арифметичне CDR кожного кластеру, де N – загальна кількість кластерів.

$$\mathbf{PQI = avg\ (CDR) = \sum CDR / N} \qquad (2)$$

Як і CDR, PQI приймає значення від –1 до +1, причому чим вище PQI, тим якіснішим є результат фітоценотичної класифікації. Якщо для одного набору даних одержано декілька класифікацій, то слід вибрати той результат, для якого індекс PQI виявиться більшим.

Для оцінки кластерів по матриці відстаней можна використовувати і класичні індекси, що широко застосовуються у математичній статистиці – зокрема, статистику силуетів (Silhouette statistic) [ROUSSEEUW, 1987]. Спочатку для кожного опису розраховується його середня відстань до усіх інших описів «свого» фітоценона (кластера) і середня відстань до описів в іншому, найближчому кластері. Оцінка кластера за статистикою силуетів зводиться до розрахунку середнього значення статистики силуетів для кожного об'єкта кластеру (опису, фітоценозу).

$$\mathbf{clus.avg.silwidths = avg\ ((b_i - a_i) / max\ (b_i; a_i))} \qquad (3);$$

$$\mathbf{avg.silwidth = avg\ (clus.avg.silwidths)} \qquad (4),$$

де *clus.avg.silwidths* – статистика силуетів окремих кластерів, a_i – середня відстань між і-м об'єктом та об'єктами всередині кластеру, b_i – найменша середня відстань між і-м об'єктом та об'єктами з іншого кластеру, avg – математична операція розрахунку середнього арифметичного, *avg.silwidth* – статистика силуетів для класифікації в цілому. Показник *clus.avg.silwidths* подібний до індексу CDR, оскільки характеризує кластер окремо, а *avg.silwidth* – до індексу PQI, що характеризує якість класифікації (групування, поділу) в цілому.

Слід зазначити, що на цьому перелік індексів, які називають внутрішніми критеріями валідації кластерів і які базуються на матриці відстаней, не обмежується. Відомі також інші, наприклад індекс Данна (Dunn Index), який оцінюється як співвідношення мінімальної відстані між об'єктами в різних кластерах до максимальної відстані між об'єктами одного кластеру [DUNN, 1974].

Бібліотека (пакет) fpc для середовища R (http://cran.r-project.org) містить функцію cluster.stats, яка дозволяє розраховувати різноманітні індекси – крім статистики силуетів, індекса Данна, також Gap-статистику, ентропійний індекс та ін. [HENNIG, 2013]. Увесь цей арсенал кількісної оцінки кластерів може бути застосований до фітоценотичних кластерів (фітоценонів). Для цього необхідно мати розраховану матрицю відстаней між описами (або подібності за видовим складом) та вектор, де кожному опису відповідає умовний номер фітоценотичного кластеру (фітоценону). Врешті-решт, така оцінка дозволить виміряти щільність та відмежованість фітоценонів, виділених з використанням певних підходів чи методів, або різним авторами, і зробити висновок про якість поділу (класифікації) за математичним критерієм, а саме матрицею відстаней між об'єктами (описами).

Оцінка кореляції (подібності) фітоценотичних класифікацій

Якщо для одного набору даних одержано декілька альтернативних класифікацій, наприклад, з використанням різних методів або різними авторами, можна розрахувати коефіцієнти кореляції, що оцінюють «узгодженість» пари класифікацій. Кожна з класифікацій – це номінальна ознака. Для вимірювання кореляції номінальних ознак можна використати статистику Крамера (Cramer's V) [CRAMER, 1964], індекс Фолкса-Меллоуса (FM-index) [FOWLKES, MALLOWS, 1983], індекс Жаккара [JACCARD, 1901] та ін. Індекси приймають значення 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.

Якщо одна з класифікацій може бути прийнята за еталон (еталонна класифікація відображає дійсний розподіл об'єктів за деякими класами (природними групами)), можна порівняти розподіл об'єктів двох класифікацій – випробуваної і еталонної, з використанням таблиці спряженості (крос-табуляції), або як ще називають її – матриці помилок (confusion matrix). Еталонна класифікація може бути означена експертом або отримана іншим верифікованим методом класифікації.

Можливо чотири випадки розподілу об'єктів (для еталонної класифікації будемо вживати термін «клас», для випробуваної – «кластер») (табл. 1):

- об'єкти належать одному кластеру і одному класу (true positive, справжньо-позитивна класифікація) – p_{11};
- об'єкти належать одному кластеру, але різним класам (false positive, помилково-позитивна класифікація, помилка 1-го роду) – p_{12};
- об'єкти належать різним кластерам, але одному класу (false negative, помилково-негативна класифікація, помилка 2-го роду) – p_{21};
- об'єкти належать різним кластерам і різним класам (true negative, справжньо-негативна класифікація) – p_{22}.

Таблиця 1

Таблиця спряженості при порівнянні випробовуваної та еталонної класифікації

Table 1

Contingency table for comparing tested and benchmark classification

	Same cluster	Different clusters	∑ Сума
Same class	p_{11}	p_{21}	$p_{1\_} = p_{11} + p_{21}$
Different classes	p_{12}	p_{22}	$p_{2\_} = p_{12} + p_{22}$
∑ Сума	$p_{1} = p_{11} + p_{12}$	$p_{2} = p_{21} + p_{22}$	$N = p_{11} + p_{12} + p_{21} + p_{22}$

У табл. 1 показник p_{11} позначає правильно (обопільно) класифіковані об'єкти, тобто об'єкти, розподіл яких у випробуваної і еталонної класифікації однаковий. Показники $p_{1\_}$ і $p_{\_1}$ відповідають загальній кількості об'єктів кластера і класу; p_{12} і p_{21} – це кількість помилково класифікованих об'єктів (у класифікації рослинності – описів, фітоценозів), що іменується помилками 1-го і 2-го роду.

Точність (англ. precision) – це частка обопільних об'єктів класу і кластеру щодо загальної кількості об'єктів кластеру. Точність показує наскільки ймовірно, знаючи розподіл об'єктів по кластерам, передбачити фактичний склад класів (еталонну класифікацію).

$$\mathbf{precision = p_{11} / (p_{11} + p_{12}) = p_{11} / p_{1\_}} \quad (5),$$

де precision – точність випробуваної класифікації по відношенню до еталонної, а позначення p_{11}, p_{12}, $p_{1\_}$ відповідають аналогічним в табл. 1.

Повнота (англ. recall) – це частка обопільних об'єктів класу і кластеру щодо загальної кількості об'єктів класу.

$$\mathbf{recall = p_{11} / (p_{11} + p_{21}) = p_{11} / p_{\_1}} \quad (6),$$

де recall – повнота, позначення p_{11}, p_{21}, $p_{\_1}$ відповідають аналогічним в табл. 1.

Точність і повнота окремо не свідчать про кореляцію двох класифікацій, оскільки є асиметричними метриками. У літературі вони ще відомі як індекси Уоллеса, asymmetric Wallace's indices I і II [WALLACE, 1983]. Зрозуміло, що чим вище точність і повнота, тим краще. Але отримати максимальні точність і повноту одночасно в реальних дослідженнях практично неможливо. Тому доводиться шукати певний баланс. Хотілося б мати метрику, яка об'єднувала б у собі обидві складові. Часто використовують F-міру (F-score), яка максимальна і дорівнює одиниці у разі повної ідентичності двох класифікацій, і визначається вона як середнє гармонічне точності і повноти. Математично F-міра еквівалентна коефіцієнту Соренсена [SØRENSEN, 1948]:

$$\mathbf{F\text{-}score = 2 \times precision \times recall / (precision + recall)} \quad (7)$$

Таким чином, для оцінки кореляції двох класифікацій можна використати таблиці спряженості, а також різноманітні коефіцієнти кореляції для номінальних ознак. Якщо одна з класифікацій приймається за еталон, то розрахунок кореляції стає методом верифікації іншої класифікації. Значення індексів більше 0,7–0,8 свідчить про значну близькість класифікацій. Склад фітоценотичних кластерів при цьому суттєво не відрізняється. Якщо обидві класифікації рівнозначні і жодна з них не може вважатися еталоном, то висока кореляція може свідчити про природність кластерів, їх відповідність дійсній структурі даних: якщо у даних дійсно є кластер-структура, то різні методи дадуть схожі класифікації, отже – кластери природні.

Оцінка фітоценонів за індексами визначеності та унікальності (Sharpness and Uniqueness Index)

У 2003 році М. Хитри (Чехія) та Л. Тихий (Чехія) запропонували для оцінки якості та унікальності синтаксонів індекси визначеності та унікальності [CHYTRÝ, TICHÝ, 2003]. Ними було застосовано цей статистичний підхід для оцінки класів та союзів рослинності Чехії. У 2010 р. аналогічний підхід І. Яролімек з колегами

застосували для оцінки класів рослинності Словакії [JAROLIMEK et al., 2010]. Ці індекси базуються на розрахунках вірності видів [CHYTRÝ et al., 2002] і доступні у програмі Juice 7.0 [TICHÝ, 2002].

Визначеність синтаксону (фітоценону) вимірюється як частка диференціюючих видів щодо середньої кількості видів в описах:

$$\mathbf{S_j = (1 + \sum \Phi_{ij} \times 100) / R_j} \quad (8),$$

де S_j – індекс визначеності фітоценону j, Φ_{ij} – вірність і-го виду фітоценону j, R_j – середня кількість видів в описах фітоценону j.

Таким чином, індекс визначеності можна трактувати як діагностованість певного фітоценону (синтаксону) за видовим складом: чим більше у фітоценону (синтаксону) вірних (диференціюючих) видів, тим краще він діагностується (інтерпретується). Для розрахунків можна застосовувати будь-які індекси вірності, головне, щоб вони були нормовані, приймали значення від 0 до 1. Зокрема, такі індексі вірності, як phi-коефіцієнт [CHYTRÝ et al., 2002], IndVal [DUFRÊNE, LEGENDRE, 1997], коефіцієнт Охаї [DE CÁCERES et al., 2008] та ін. Одним із недоліків індексу визначеності можна вважати те, що він є ненормованим і може приймати значення від 0 до +∞. Отже, для певного синтаксону важко сказати, чи є досягнуте значення індексу визначеності максимально можливим. Інший недолік полягає в тому, що заміна індексу вірності призведе до зміни розрахованих за формулою 8 значень S_j, що теж утруднює порівняння.

Унікальність синтаксону (фітоценону) визначається часткою діагностичних (вірних, диференціюючих) видів, які є діагностичними лише для одного синтаксону [CHYTRÝ, TICHÝ, 2003]. Низькі значення унікальності свідчать про можливість об'єднання синтаксону з іншими синтаксонами того ж рангу. Синтаксон є унікальним, якщо жоден з його діагностичних видів не має такого ж статусу у інших синтаксонах. Розрахунок індексу унікальності синтаксону здійснюється у два етапи. Спочатку розраховується т.з. асиметричний індекс подібності для кожної пари синтаксонів j і k (формула 9). Після цього індекс унікальності синтаксону визначається за формулою 10.

$$\mathbf{T_{jk} = \sum \Phi_{ij} \times \Phi_{ik} / \sum \Phi_{ij}^2} \quad (9);$$

$$\mathbf{U_j = 1 / \sum T_{jk}} \quad (10),$$

де T_{jk} – асиметричний індекс подібності пари синтаксонів j і k, Φ_{ij} та Φ_{ik} – вірність і-го виду щодо синтаксонів j і k відповідно, U_j – індекс унікальності синтаксону j. Індекс унікальності, на відміну від індексу визначеності, є нормованим. Вищі значення U_j свідчать про значні відмінності синтаксону j від усіх інших синтаксонів за видовим складом.

Чіткість класифікації (Crispness of classification)

У 2005 році З. Ботта-Дукат (Венгрія) з колегами запропонували оцінювати ламкість (англ. crispness – ламкість) (*прим. авт.* – доречніше перекладати як «чіткість») класифікації з використанням 2*c таблиць спряженості «вид-кластер» та G-статистики [BOTTA-DUKÁT et al., 2005].

G-статистика (G-statistic), або G-тест, як і Хі-квадрат-статистика Пірсона (Pearson's chi-square statistic), використовується для перевірки, чи є зв'язок між двома категоріальними ознаками (*прим. авт.* – не варто плутати поняття перевірки наявності взаємозв'язку між двома ознаками та вимірювання сили зв'язку (див. розділ «Оцінка кореляції (подібності) фітоценотичних класифікацій»), де також використовуються таблиці спряженості). Згідно нульової гіпотези H_0 ознаки незалежні. Якщо фактичні і очікувані (теоретичні) частоти близькі, значення G-статистики буде невеликим, гіпотеза H_0 не може бути відкинута, ознаки незалежні. G-статистика заснована не на різниці фактичної і теоретичної частоти, як Хі-квадрат статистика, а на їх відношенні, що робить її адитивною і більш зручною у багатовимірному аналізі:

$$\mathbf{G = 2*\sum[O \times \ln(O/E)]} \quad (11),$$

де G – G-статистика, O (Observed) – фактичні частоти, E (Expected) – теоретичні частоти, виходячи з гіпотези про незалежність ознак.

Розглянемо таблицю спряженості «вид-кластер», де кластерами виступають окремі фітоценони, виділені у результаті класифікації таблиці даних (описів) (табл. 2).

Таблиця 2

Таблиця спряженості 2*c для розрахунку вірності видів

Table 2

2*c contingency table for fidelity measures calculation

	1st cluster	2nd cluster	c-th cluster
containing the species	n_{11}	n_{12}	n_{1C}
not containing the species	n_{01}	n_{02}	n_{0C}
∑ Сума	N_1	N_2	N_C

Кількість колонок таблиці спряженості дорівнює *c*, кількості виділених кластерів (фітоценонів), частоти – абсолютному траплянню видів у фітоценонах (n_{11}, n_{12}, …, n_{1C}), а $N_1, N_2, ..., N_C$ – це кількість описів фітоценонів. Загальне трапляння *i*-виду визначається як $freq = n_{11} + n_{12} + \ldots + n_{1C,}$ а загальна кількість описів в таблиці даних – $N = N_1 + N_2, + \ldots + N_C$.

Автори методу визначення «чіткості класифікації» зазначають, що G-статистика не оцінює вірність видів окремим кластерам (фітоценонам), а їх спроможність розрізняти кластери в цілому, тому її розглядають як «діагностичну силу» (separation power of species), а не вірність [BOTTA-DUKÁT et al., 2005]. Усі розрахунки виконуються в програмі Juice 7.0 [TICHÝ, 2002]. Спочатку розраховують середнє значення G-статистики одержане для усіх видів – це т.з. нескореговане значення COC, crispness of classification, а потім одержують скореговане значення COC, яке дає можливість порівнювати класифікації з різною кількістю кластерів. Це зрозуміло, адже метод визначення чіткості класифікації, подібно до Optimclass підходу, про який піде мова далі, розроблявся для вибору оптимальної класифікації з різною кількістю кластерів, а також для порівняльної оцінки якості класифікацій одного набору даних.

Це потужний кількісний метод, який дозволяє оцінювати і верифікувати класифікацію геоботанічних таблиць, але, на жаль, у фітоценології пострадянського простору цей та інші методи оцінки якості класифікації, запропоновані європейськими фітоценологами, застосовуються дуже рідко [GOLUB et al., 2012; GOLUB et al., 2013].

Optimclass підхід

Оцінка результатів класифікації з використанням матриці відстаней дозволяє виміряти щільність і відмежованість кластеру, але не говорить про те, чим саме даний кластер відрізняється від інших і яка його екологічна специфіка. Привнести якісний аспект в оцінку фітоценотичної класифікації може лише класифікація видів. Цей напрямок отримав назву «аналіз індикаторних видів» (indicator species analysis). Кожен вид несе певну екологічну інформацію, а по відношенню до синтаксонів – види мають різну диференціюючу силу (вірність, аффінність).

В оцінці вірності видів певного синтаксону важливі дві складові – константність (постійність, трапляння) і характерність (вибірковість, специфічність) виду. Якщо вид константний (постійний), але при цьому його амплітуда перекриває кілька синтаксонів, то його диференціююча сила незначна. Характерні види є добрими індикаторами внаслідок вибірковості, однак не всі характерні види можуть бути диференціюючими: якщо вид зустрічається менш ніж в 40 % описів певного синтаксону (I-II клас константності), його не можна використовувати для діагностики.

У 2010 році Л. Тихим і колегами [TICHÝ et al., 2010] було запропоновано критерій для оцінки якості класифікації за кількістю діагностичних (диференціюючих)

видів. Цей підхід отримав назву Optimclass і використовується в програмі Juice [TICHÝ, 2002]. У модифікації Optimclass1 пропонується підраховувати загальну кількість вірних видів і вибирати ту класифікацію (поділ) або таку кількість кластерів (наприклад, в методі k-середніх), при якому цей показник максимізується. Модифікація Optimclass2 має таке ж призначення, але враховує кількість «добрих» фітоценонів, що містять кількість вірних видів не менше встановленого порогу (у оригінальній роботі [TICHÝ et al., 2010] використаний поріг 4 види).

Слід зазначити, що застосування підходу Optimclass передбачає виконання певних умов. По-перше, екстрагування вірних видів має бути проведено статистично, без експерта, оскільки об'єктивний розрахунок кількості вірних видів можливий тільки у разі «статистичної» класифікації видів. По-друге, Optimclass можна застосовувати лише для порівняння класифікацій одного набору даних, наприклад, різними методами автоматичної класифікації або різними експертами, але не для різних наборів даних. По-третє, спосіб розрахунку індексів вірності та встановлений поріг значення fidelity, при яких вид включається в список «статистично» вірних, може бути будь-яким, але однаковим при порівнянні якості класифікацій, оскільки спосіб розрахунку та поріг вірності прямо впливають на кількість диференціюючих видів, а отже і на оцінку якості класифікації згідно підходу Optimclass.

Оцінка стійкості фітоценотичних кластерів

Одна з проблем класифікації рослинності – це її низька стійкість. Нерідко діагностичні блоки видів «розсипаються», якщо до вихідної таблиці додати нові описи. Після «вливань» синтаксони (особливо дрібні) можуть зникнути, а їхні оптимуми постійно дрейфують. Отож, важливий аспект – стійкість фітоценотичних кластерів (фітоценонів) і синтаксонів. Власне, перевірка стійкості кластер-структури є перевіркою її достовірності. Існує емпіричне правило – стійка типологія повинна зберігатися при заміні методів групування, тобто класифікація того ж набору описів з використанням іншого методу чи підходу повинна давати схожий результат.

Можна виділити два методологічно різні підходи щодо перевірки стійкості кластер-структури. При першому підході дані залишаються незмінними, але змінюють метод групування, після чого розраховують кореляцію класифікацій. При другому підході змінюють дані – щодо них застосовують різні схеми бутстреппінгу [SHYTYKOV, 2012] – технології множення вибірок на основі вихідної сукупності, яка останнього часу набула значної популярності у математичній статистиці як для перевірки статистичних гіпотез, так і для розрахунку різноманітних статистик і багатовимірного аналізу. Отож, одержані з вихідної сукупності тест-підмножини кластеризують і порівнюють результат з класифікацією вихідного набору даних.

У 2011 р. Л. Тихий з колегами [TICHÝ et al., 2011] запропонували оцінювати стійкість фітоценотичних кластерів застосовуючи бутстреп без повернення (without-replacement bootstrap resampling) до таблиці даних (описів), після чого оцінювати кореляцію класифікацій повного набору даних та бустреп-підмножин, використовуючи лямбду Гудмена-Краскела (Goodman-Kruskal's lambda index) [GOODMAN, KRUSKAL, 1954]. Таку стійкість кластер-структури слід трактувати як «відтворюваність», стабільність (repeatability).

Але схеми ресемплінгу (resampling), і бутстреппінгу як одного з його різновидів надзвичайно різноманітні – існують ще метод складного ножа (jackknifing), метод перехресної перевірки або крос-валідації (cross-validation) та ін. Ми пропонуємо ще одну модифікацію, яку слід трактувати як метод перевірки «міцності», резистентності кластер-структури фітоценонів. Разом зі стабільністю, про яку йшла мова вище, це один із аспектів робастності (robustness) класифікації.

Згідно нашої схеми проводиться «відсікання» даних:

1. Повний масив описів піддається обробці (класифікується), виділяються фітоценони, належність описів до фітоценонів (кластерів) запам'ятовуємо як еталонну класифікацію.

2. З повного масиву видаляють N % описів, вибраних випадковим чином, причому N (крок) визначається довільно – 5, 10 % і т.п.

3. Проводять класифікацію «усіченого» набору описів тим же методом кластерного аналізу, який був застосований до вихідного набору даних.

4. На кожному кроці «усічення» даних розраховують кореляцію класифікацій «повного» і «усіченого» набору описів за будь-яким коефіцієнтом номінальної кореляції – Crasmer's V, FM-index і т.п. (див. розділ «Оцінка кореляції (подібності) фітоценотичних класифікацій»).

5. Відмічають на графіку ступінь «усічення», крок (вісь X) та значення кореляції класифікацій у % (вісь Y). Повторюють п. 2, 3, 4 і т.д.

6. Знаходять ступінь «усічення», при якому кореляція класифікацій досягає критично малих значень, наприклад – 0,7. Також по ходу усікання даних відмічають, які фітоценони (кластери) зникають першими, а які є більш опірними (стабільними).

Висновки

Загалом, ми розглянули різні аспекти оцінки якості фітоценотичної класифікації – математичний критерій (за відстанями між об'єктами), флористичний (за кількістю диференціюючих видів), оцінку стійкості фітоценотичних кластерів. На нашу думку, наведення у вітчизняних публікаціях, присвячених класифікації рослинності, зазначених індексів дозволило б аргументувати результат класифікації, оцінити її міцність, переконати у дійсній наявності фітоценонів, розрахувати їх щільність (гомогенність), забракувати невизначені чи неунікальні фітоценотичні кластери, глибше пізнати структуру аналізованих даних.

References

BOTTA-DUKÁT Z., CHYTRÝ M., HÁJKOVÁ P., HAVLOVÁ M. (2005). Vegetation of lowland wet meadows along a climatic continentality gradient in Central Europe. *Preslia*, **77**: 89-111.

CHYTRÝ M., TICHÝ L. (2003). Diagnostic, constant and dominant species of vegetation classes and alliances of the Czech Republic: a statistical revision. *Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis*, Biologia, **108**: 1-231.

CHYTRÝ M., TICHÝ L., HOLT J., BOTTA-DUKÁT Z. (2002). Determination of diagnostic species with statistical fidelity measures. *Journal of Vegetation science*, **13** (1): 79-90.

CRAMÉR H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press. 282 p.

DE CÁCERES M., FONT X., OLIVA F. (2008). Assessing diagnostic species value in large data sets: A comparison between phi-coefficient and Ochiai index. *Journal of Vegetation science,* **19**: 779-788.

DUFRÊNE M., LEGENDRE P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.*, **67**: 345-366.

DUNN J.C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**: 95-104.

FOWLKES E.B., MALLOWS C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, **78** (383): 553-569.

GOLUB V.B., SOROKIN A.N., MALTSEV M.V., CHUVASHOV A.V. (2012). *Vestnik Volzhskogo universiteta im. V.N. Tatishcheva*, **3**: 308–317. [ГОЛУБ В.Б., СОРОКИН А.Н., МАЛЬЦЕВ М.В., ЧУВАШОВ А.В. (2012). Почвы и растительность многолетней залежи в дельте Р. Волги. *Вестник Волжского университета им. В.Н. Татищева*, **3**: 308-317]

GOLUB V.B., STARICHKOVA K.A., BARMIN A.N., IOLIN M.M., SOROKIN A.N., NIKOLAICHUK L.F. (2013). Estimate of vegetation dynamics in the Volga delta. *Arid ecosystems*, **3**(3): 156-164.

GONCHARENKO I.V. (2015). *Vegetation of Russia*. **27**: 125-138. [ГОНЧАРЕНКО И.В. (2015). DRSA: алгоритм неиерархической кластеризации с использованием k-NN графа и его применение в классификации растительности. *Растительность России*. **27**: 125-138]

GOODMAN L.A., KRUSKAL W.H. (1954). Measures of association for cross-classification. *J. Am. Stat. Assoc*., **49**: 732-764.

HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, **17**: 107-145.

HENNIG C. (2013). fpc: Flexible Procedures For Clustering. R Package Version 2.1-6. available at: http://www.homepages.ucl.ac.uk/~ucakche/ (accessed 01 April 2016)

JACCARD P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. Vaudoise sci. Natur*, **37** (140): 241-272.

JAROLÍMEK I., ŠIBÍK J., TICHÝ L., KLIMENT J. (2010). Sharpness and uniqueness of the phytosociological classes of Slovakia. *Annali di Botanica*, Nuova Serie, Roma, **10**: 11-18.

MANDEL I.D. (1988). *Klasternyi analiz*. M.: Finansy i statistika. 176 p. [МАНДЕЛЬ И.Д. (1988). *Кластерный анализ*. М.: Финансы и статистика. 176 с.]

RENDÓN E., ABUNDEZ I., ARIZMENDI A., QUIROZ E. (2011). Internal versus external cluster validation indices. *Int. J. Computers and Communications*, **5**: 27-34.

ROUSSEEUW P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**: 53-65.

SHITIKOV V.K. (2012). *Printsipy ekologii*, **1**: 4-24. [ШИТИКОВ В.К. (2012). Использование рандомизации и бутстрепа при обработке результатов экологических наблюдений. *Принципы экологии*, **1**: 4-24]

SØRENSEN T.A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab. Biol. krifter*, **5**: 1-34.

TICHÝ L. (2002). JUICE, software for vegetation classification. *Journal of vegetation science*, **13** (3): 451-453.

TICHÝ L., CHYTRÝ M., HÁJEK M., TALBOT S. S., BOTTA-DUKÁT Z. (2010). OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation science,* **21**: 287-299.

TICHÝ L., CHYTRÝ M., SMARDA P. (2011). Evaluating the stability of the classification of community data. *Ecography,* **34** (5): 807-813.

WALLACE D.L. (1983). Comment. *Journal of the American Statistical Association*, **78** (383): 569-576.

Рекомендує до друку
Д.В. Дубина

Отримано 05.04.2016

Адреса автора:
І.В. Гончаренко
Інститут еволюційної екології НАН України
вул. Академіка Лебєдєва, 37
03143, м. Київ, Україна
e-mail: 3604749@gmail.com

Author's address:
I.V. Goncharenko
Institute for evolutionary ecology, National Academy of Sciences of Ukraine
37, Lebedeva str.
03143, Kyiv, Ukraine
e-mail: 3604749@gmail.com