ISSN 0868-8540 (Print)
ISSN 2413-5984 (*Online*)

БРЕН О.Г.[1] (https://orcid.org/0000-0001-7423-9258)
БРЕН О.А.[2] (https://orcid.org/0009-0003-7632-6285)
СОЛОНЕНКО А.М.[2] (https://orcid.org/0000-0002-3417-5146)
ПОДОРОЖНИЙ С.М.[2] (https://orcid.org/0000-0002-7702-7602)

[1]*Інститут ботаніки, Чеська академія наук,*
вул. Дукелска, 135, Тршебонь 37901, Чеська Республіка
oscillat8@gmail.com
[2]*Мелітопольський державний педагогічний університет імені Богдана Хмельницького,*
кафедра ботаніки і садово-паркового господарства,
вул. Наукового містечка, 59, Запоріжжя 69000, Україна

АВТОМАТИЗАЦІЯ ОБЧИСЛЕНЬ ІНДЕКСУ ДАЙСА (ЧЕКАНОВСЬКОГО-СЬОРЕНСЕНА) У ФІКОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Реферат. Розглядаються тенденції використання індексу Дайса (Чекановського-Сьоренсена) в дослідженнях водоростей та ціанопрокаріот. Зроблено стислий огляд особливостей його використання з урахуванням позитивних рис та недоліків цієї метрики. Актуальність роботи обумовлена потребою дослідників в автоматизації розрахунків індексу Дайса та побудови результуючої матриці. Пропонується автоматизація обчислень шляхом використання макросів у середовищі Excel. Здійснено огляд можливостей такого підходу та запропоновано власний макрос для швидкого й зручного обчислення індексу Дайса без побудови зведеної таблиці та сторонніх програм.

Ключові слова: водорості, ціанопрокаріоти, міра подібності, індекс Дайса (Чекановського-Сьоренсена)

Надійшла до редакції 30.07.2023. Після доопрацювання 27.11.2023. Підписана до друку 05.12.2023. Опублікована 20.03.2024

Цитування. Брен О.Г., Брен О.А., Солоненко А.М., Подорожний С.М. 2024. Автоматизація обчислень індексу Дайса (Чекановського-Сьоренсена) у фікологічних дослідженнях. *Альгологія*. 34(1): 80–90. https://doi.org/10.15407/alg34.01.080

Вступ

У теоретичних і практичних галузях знань широко використовуються різноманітні метрики подібності для розуміння, наскільки певні об'єкти або процеси об'єднані (подібні) або, навпаки, незалежні (відмінні) один від одного (Cheetham, Hazel, 1969; Hubalek, 1982; Keil 2019). Однією з таких метрик є індекс (або коефіцієнт) Дайса (інші назви: Сьоренсена, Дайса-Брея, Сьоренсена-Дайса, Чекановського-Сьоренсена) (Czekanowski, 1909; Dice, 1945; Sørensen, 1948; Bray, 1956), який широко застосовується в сучасних наукових пошуках і технологіях (Bertels et al., 2019; Li et al., 2020; Flores et al., 2022; Ataş, 2023; Eikelboom et al., 2023). При дослідженні видового різноманіття водоростей та ціанопрокаріот даний індекс використовується переважно для попарного порівняння різних пробних ділянок (водойм) або проб за подібністю якісного складу (наявністю спільних видів) (Hubalek, 1982; Berezovska, 2019; Graco-Roza et al., 2019; Peipoch et al., 2019; Mironyuk, Tkachenko, 2020; Zhang et al., 2021; Shcherbak et al., Semeniuk, Lutsenko, 2023).

Індекс Дайса обчислюється за формулою:

$$D = \frac{2c}{a+b},$$

де D – індекс Дайса, c – кількість спільних елементів (видів) між двома наборами даних (пробами або пробними ділянками), a, b – кількість елементів у першому та другому наборі даних. Значення коефіцієнта може варіювати від 0 до 1, де 0 вказує на відсутність спільних елементів, а 1 – на повну ідентичність двох наборів даних (Dice, 1945).

Нами проведено розширений пошук по анотаціях наукових публікацій на платформах Web of Science (https://www.webofscience.com) та Scopus (https://www.scopus.com) за запитами щодо робіт, в яких є згадки про використання індексу Дайса під час дослідження водоростей та ціанопрокаріот (табл. 1).

Результати пошуку на обох платформах демонструють тенденцію до все більшого використання індексу Дайса в наукових дослідженнях водоростей та ціанопрокаріот, особливо з 2016 р. (рис. 1).

Не дивлячись на простоту розрахунку та поширеність у використанні, в індексу Дайса є певні недоліки. Він є досить чутливим до розміру порівнюваних множин – якщо вони досить відрізняються, то індекс може надавати неправильні оцінки схожості.

Таблиця 1. **Запити для пошуку публікацій фікологічної тематики зі згадуванням використання індексу Дайса**

Платформа	Web of Science	Scopus
Текст запиту	(AB=(dice index) OR AB=(dice coefficient) OR AB=(sorensen) OR AB=(tchekanovsky) OR AB=(chekanovsky) OR AB=(sørensen) OR AB=(czekanowski)) AND (AB=(algae) OR AB=(cyanobacteria) OR AB=(cyanoprokaryota) OR AB=(cyanoprocaryota))	(ABS (dice index) OR ABS (sorensen) OR ABS (tchekanovsky) OR ABS (chekanovsky) OR ABS (czekanowski)) AND (ABS (algae) OR ABS (cyanobacteria) OR ABS (cyanoprokaryota) OR ABS (cyanoprocaryota)))

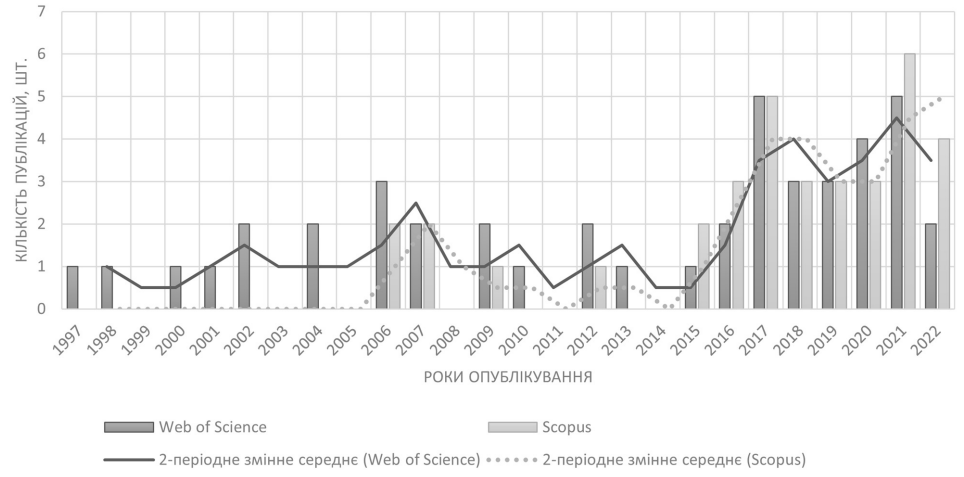


Рис. 1. Кількість публікацій з використанням індексу Дайса в дослідженнях, пов'язаних з водоростями та ціанопрокаріотами в базах, які індексуються у Web of Science (1997–2022 рр.) та Scopus (2004–2022 рр.)

Окрім того, індекс не є стійким до викидів у даних, якими, наприклад, можуть бути певні види, які значно відрізняються від загальної вибірки, що може призводити до значних похибок (McCune, Grace, 2002). Наявні певні недоліки в залежності від особливостей поширеності певних об'єктів у межах досліджуваного простору (Mainali et al., 2022). Серед іншого, ця метрика розглядає кожен елемент множини незалежно від контексту або розташування, не враховуючи просторовий розподіл або

структуру даних, тому інтерпретація результатів у певному контексті стає додатковим завданням самого дослідника. Ще однією вадою є відсутність можливості аналізувати дескриптивні множини.

Найбільшою проблемою використання метрик подібності (у т. ч. індексу Дайса) окрім їхніх методологічних особливостей є те, що розрахунки та робота з вихідною інформацією є дуже витратними з огляду на час і ресурси, особливо при роботі з великими обсягами даних. Насамперед це стосується випадків, коли порівнюються не лише дві пробні ділянки або проби, а значно більша їхня кількість, наприклад декілька десятків або сотень. Ручне обчислення або використання сторонніх програм дозволяють вирішити це завдання, але значно уповільнюють процес дослідження та аналізу даних. Обмежує використання й потреба в знанні певних мов програмування або розуміння особливостей користування спеціальними програмами. Окрім того, неминучими є помилки та неточності в розрахунках та при перенесенні даних з одного додатку або файлу в інший, які обумовлені людським фактором.

Постає потреба в автоматизації розрахунків метрик подібності в цілому, зокрема індексу Дайса. В цьому напрямку вже виконувалися роботи (Sneath, 1957; Rogers, Tanimoto, 1960; Austin, Colwell, 1977; Sinnott, 1981; Hammer et al., 2001), але й досі відсутні достатньо прості з методичної точки зору рішення, що й обумовлює актуальність нашої роботи.

Матеріали та методи

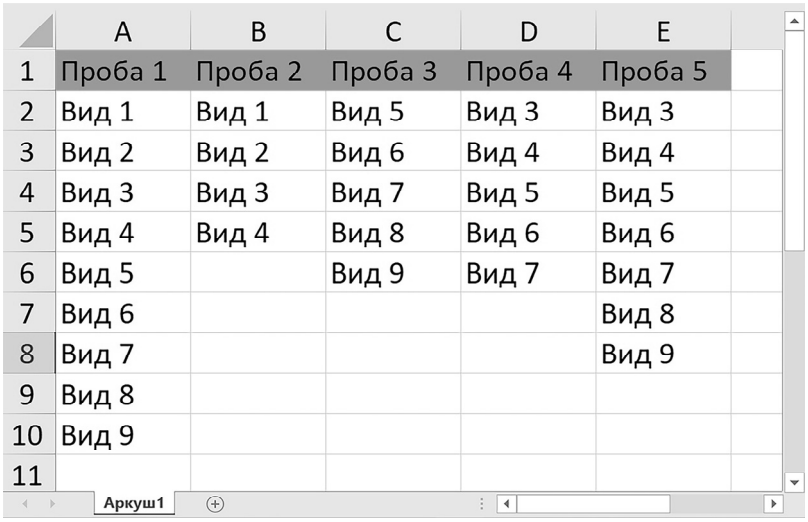
Для розробки макросу для обчислення індексу Дайса нами було використано програмне забезпечення Microsoft Excel 365. Налаштування розділу «Розробник» («Developer») здійснювалось у відповідності з офіційною підтримкою продукту за електронною адресою: https://support.microsoft.com/en-us/excel. У процесі розробки макросу використовувалася мова програмування VBA (Visual Basic for Applications).

Результати та обговорення

Розроблений нами макрос здійснює попарне порівняння стовпців вихідного аркушу Microsoft Excel на предмет подібності. Результати порівняння використовуються для розрахунку індексу Дайса з подальшим перенесенням усіх розрахованих індексів в матрицю.

Вихідний аркуш повинен містити стовпці, які представляють собою певні пробні ділянки або проби. Кожна комірка в стовпці – назва виду, виявленого на даній пробній ділянці або в даній пробі. В межах одного стовпця всі комірки мають бути унікальними.

Перший рядок усіх стовпців на вихідному аркуші – заголовки (назви) цих пробних ділянок або проб (при порівнянні стовпців комірки цього рядка не використовуються) (рис. 2).



	A	B	C	D	E
1	Проба 1	Проба 2	Проба 3	Проба 4	Проба 5
2	Вид 1	Вид 1	Вид 5	Вид 3	Вид 3
3	Вид 2	Вид 2	Вид 6	Вид 4	Вид 4
4	Вид 3	Вид 3	Вид 7	Вид 5	Вид 5
5	Вид 4	Вид 4	Вид 8	Вид 6	Вид 6
6	Вид 5		Вид 9	Вид 7	Вид 7
7	Вид 6				Вид 8
8	Вид 7				Вид 9
9	Вид 8				
10	Вид 9				
11					

Аркуш1

Рис. 2. Приклад вихідної таблиці для обробки даних запропонованим нами макросом

Нижче наводимо розроблений нами код макросу для розрахунку індексу Дайса. Кожна операція коду має порядковий номер та пояснювальний коментар після символу «'».

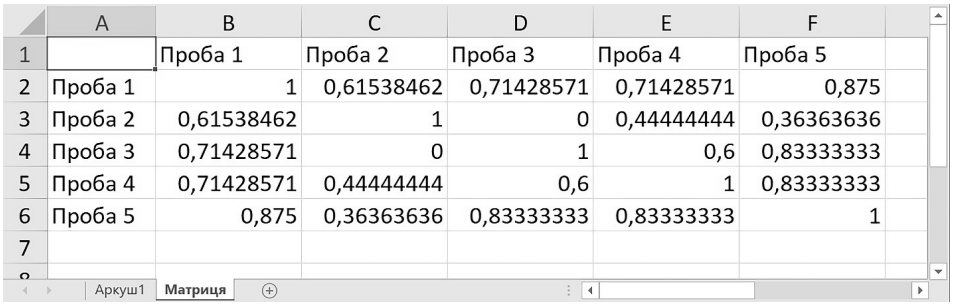
```
Sub CalculateDiceMatrix()
    Dim sourceSheet As Worksheet      ' 1. Вихідний аркуш
    Dim targetSheet As Worksheet      ' 2. Цільовий аркуш
    Dim sourceRange As Range                ' 3. Вихідний діапазон даних
    Dim targetRange As Range                ' 4. Цільовий діапазон даних
    Dim sourceColumnCount As Integer        ' 5. Кількість стовпців у вихідному діапазоні
    Dim sourceRowCount As Integer     ' 6. Кількість рядків у вихідному діапазоні
    Dim i As Integer                  ' 7. Змінна для ітерації по стовпцях
    Dim j As Integer                  ' 8. Змінна для ітерації по стовпцях
    Dim numerator As Integer          ' 9. Лічильник чисельника
```

```
    Dim denominator As Integer                    ' 10. Знаменник
    Dim index As Double                    ' 11. Індекс
' 12. Встановлюємо вихідний аркуш як активний
    Set sourceSheet = ActiveSheet
' 13. Створюємо новий аркуш для результатів
    Set targetSheet = Worksheets.Add(After:=Worksheets(Worksheets.Count))
    targetSheet.Name = "Матриця індексів Дайса"
' 14. Встановлюємо вихідний діапазон на вихідному аркуші
    Set sourceRange = sourceSheet.UsedRange
    sourceColumnCount = sourceRange.Columns.Count
    sourceRowCount = sourceRange.Rows.Count
' 15. Встановлюємо цільовий діапазон даних, починаючи з верхньої лівої комірки
    Set targetRange = targetSheet.Cells(2, 2).Resize(sourceColumnCount, sourceColumnCount)
' 16. Записуємо заголовки стовпців у перший рядок і перший стовпчик цільового діапазону
    For i = 1 To sourceColumnCount
        targetSheet.Cells(i + 1, 1) = sourceRange.Cells(1, i).Value
        targetSheet.Cells(1, i + 1) = sourceRange.Cells(1, i).Value
    Next i
' 17. Ітеруємося по кожній парі стовпців і обчислюємо індекс
    For i = 1 To sourceColumnCount
        For j = 1 To sourceColumnCount
            If i = j Then
                ' 18. Якщо стовпці однакові, встановлюємо індекс на 1
                targetRange.Cells(i, j) = 1
            ElseIf j < i Then
' 19. Якщо комірка знаходиться в нижній трикутній частині, копіюємо значення з відповідної комірки
                targetRange.Cells(i, j) = targetRange.Cells(j, i).Value
            Else
' 20. Обчислюємо індекс для поточної пари стовпців
                numerator = 0
                denominator = 0
                For Each sourceCell1 In sourceRange.Columns(i).Cells
                    If Not IsEmpty(sourceCell1) Then
                        For Each sourceCell2 In sourceRange.Columns(j).Cells
```

```
                    If Not IsEmpty(sourceCell2) Then
                        If sourceCell1.Value = sourceCell2.Value Then
                            numerator = numerator + 1 ' 21. Збільшуємо чисельник на 1
                            Exit For ' 22. Переходимо до наступної комірки в першому стовпчику
                        End If
                    End If
                Next sourceCell2
            End If
        Next sourceCell1
' 23. Обчислюємо індекс
        denominator = WorksheetFunction.CountA(sourceRange.Columns(i)) - 1 +
WorksheetFunction.CountA(sourceRange.Columns(j)) - 1
        If denominator > 0 Then
            index = 2 * numerator / denominator
        Else
            index = 0
        End If
' 24. Записуємо індекс у матрицю
        targetRange.Cells(i, j) = index
      End If
    Next j
  Next i
' 25. Автоматично підбираємо ширину стовпців у цільовому діапазоні
  targetRange.EntireColumn.AutoFit
End Sub
```

У результаті обробки вихідної таблиці макросом будується квадратна матриця з занесеними до неї індексами Дайса для кожної пари порівнюваних стовпців (рис. 3).

Перший рядок та перший стовпчик на аркуші, де будуватиметься матриця, мають координати-заголовки, які автоматично переносяться з вихідного аркуша і є заголовками стовпців вихідного аркуша. Головна діагональ матриці в усіх випадках має значення «1», що означає повне співпадіння видів певного стовпця з самим собою. За замовчуванням значення розрахованого індексу подаються до восьмого символу після коми.



	A	B	C	D	E	F
1		Проба 1	Проба 2	Проба 3	Проба 4	Проба 5
2	Проба 1	1	0,61538462	0,71428571	0,71428571	0,875
3	Проба 2	0,61538462	1	0	0,44444444	0,36363636
4	Проба 3	0,71428571	0	1	0,6	0,83333333
5	Проба 4	0,71428571	0,44444444	0,6	1	0,83333333
6	Проба 5	0,875	0,36363636	0,83333333	0,83333333	1
7						

Рис. 3. Приклад збудованої матриці з розрахованими індексами Дайса

Макрос може обробити до 16 384 стовпці у вихідній таблиці. Розміри побудованої квадратної матриці також обмежуються розміром до 16 384 × 16 384 комірок, що обумовлено максимальною кількістю стовпців, яка підтримується в Excel версії 2007 і вище.

Розроблений нами макрос є досить зручним та простим рішенням для розрахунку індексу Дайса і може бути використаний не лише у флористичних або екологічних дослідженнях, а й в інших галузях. Використовуючи цей макрос, дослідники можуть з легкістю проводити обчислення індексу Дайса без необхідності вручну прописувати формули або використовувати сторонні програми, оскільки всі маніпуляції з даними здійснюються автоматично в межах лише одного середовища – Microsoft Excel. До того ж, відсутня потреба в побудові зведеної таблиці, що є обов'язковою вимогою для більшості програмних продуктів, які здійснюють подібні розрахунки. Все це дозволяє значно скоротити час, необхідний для обробки та аналізу даних, і спростити процес обчислення, а автоматизація зменшує не тільки часо- та трудозатратність, а й значно зменшує можливі помилки при розрахунках.

Перспективним продовженням роботи може бути оптимізація коду та розширення функціональних можливостей макросу.

Висновки

Індекс Дайса (Чекановського-Сьоренсена) – бінарна міра подібності, яка використовується для порівняння двох статистичних вибірок. Застосовується в різних наукових та прикладних галузях, в т. ч. в біологічних дослідженнях, зокрема у фікології, що підтверджується наведеним нами аналізом публікацій з Web of Science та Scopus.

Обмеження використання індексу Дайса пов'язані з чутливістю до рівновеликості порівнюваних вибірок, значним впливом викидів у даних

(низька робастність) і неможливістю аналізувати дескриптивні множини. Ускладнення щодо використання індексу може бути обумовлене великим обсягом вихідних даних, що призводить до часо- та трудомісткості розрахунків, а також помилок у результатах розрахунків.

Автоматизація при розрахунку індексів Дайса та побудова результуючої матриці може бути здійснена за допомогою різних сучасних статистичних додатків та деяких мов програмування, але не всі вони зручні у використанні та потребують спеціальних знань, що також ускладнює процес обробки даних.

Нами запропоновано власно розроблений макрос для Microsoft Excel з використанням мови програмування VBA для автоматизації розрахунків індексу Дайса. Цей макрос порівнює стовпці вихідного аркуша на предмет подібності та обчислює індекс Дайса для кожної пари порівнюваних стовпців без потреби створення зведеної таблиці. Результати порівнянь відображаються в квадратній матриці.

Список літератури

Ataş İ. 2023. Performance Evaluation of Jaccard-Dice Coefficient on Building Segmentation from High Resolution Satellite Images. *Balkan J. Electrical Comp. Engineer*. 11(1): 100–106.

Austin B., Colwell R.R. 1977. Evaluation of Some Coefficients for Use in Numerical Taxonomy of Microorganisms. *Int. J. Syst. Bacteriol*. 27(3): 204–210.

Berezovska V. 2019. Species Diversity of Algae of the Kiev Upland Rivers (Ukraine). *Int. J. Algae*. 21(1): 43–66. https://doi.org/10.1615/InterJAlgae.v21.i1.30

Bertels J., Eelbode T., Berman M., Vandermeulen D., Maes F., Bisschops R., Blaschko M.B. 2019. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Proc. 22nd Int. Conf. Shenzhen (China). Pp. 92–100.

Bray J.R. 1956. A study of mutual occurrence of plant species. Ecology. 37(1): 21–28.

Cheetham A.H., Hazel J.E. 1969. Binary (Presence-Absence) Similarity Coefficients. *J. Paleontol.* 43(5): 1130–1136.

Czekanowski J. 1909. Zur differential Diagnose der Neandertalgruppe. *Korrespbl. Dtsch. Ges. Anthropol*. 40: 44–47.

Dice L.R. 1945. Measures of the amount of ecological association between species. Ecology. 26(3): 297–302.

Eikelboom W., Van den Berg E., Coesmans M. Goudzwaard J. et al. 2023. Effects of the DICE Method to Improve Timely Recognition and Treatment of Neuropsychiatric Symptoms in Early Alzheimer's Disease at the Memory Clinic: The BEAT-IT Study. *J. Alzheimer's Dis*. 93(4): 1407–1423.

Flores P., Salicrú M., Sánchez-Pla A., Ocaña J. 2022. An equivalence test between features lists, based on the Sorensen–Dice index and the joint frequencies of GO term enrichment. *BMC Bioinform*. 23(1): 1–21.

Graco-Roza C., Santos J., Huszar V., Domingos P., Soininen J., Marinho M. 2019. Downstream transport processes modulate the effects of environmental heterogeneity on riverine phytoplankton. *Sci. Total Environ*. 703(3): 1–10.

Hammer O., Harper D., Ryan P. 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontol. Electron.* 4(1): 1–9.

Hubalek Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.* 57(4): 669–689.

Keil P. 2019. Z-scores unite pairwise indices of ecological similarity and association for binary data. *Ecosphere*. 10(11): e02933.

Li X., Wang C., Zhang X., Sun W. 2020. Generic SAO Similarity Measure via Extended Sørensen-Dice Index. *IEEE Access*. 8: 66538–66552.

Mainali K.P., Slud E., Singer M.C., Fagan W.F. 2022. A better index for analysis of co-occurrence and similarity. *Sci. Adv.* 8(4): eabj9204.

McCune B., Grace J. 2002. *Analysis of Ecological Communities*. Glenden Beach: MjM Software Design. 307 p.

Mironyuk A., Tkachenko F. 2020. Species Composition of Algae in Small Rivers of the Northwestern Black Sea Region. *Int. J. Algae.* 22(4): 359–372. https://doi.org/10.1615/InterJAlgae.v22.i4.50

Peipoch M., Miller S., Antao T., Vallett H. 2019. Niche partitioning of microbial communities in riverine floodplains. *Sci. Rep.* 9(1): 16384.

Rogers D.J., Tanimoto T.T. 1960. A computer program for classifying plants. *Science*. 132(3434): 1115–1118. https://doi.org/10.1126/science.132.3434.1115

Shcherbak V., Semeniuk N., Lutsenko D. 2023. Diversity and Ecological Characteristics of Algae in the Water Column in the Subbasin of the Large Danube Lakes During the Autumn-Winter Period (Ukraine). *Int. J. Algae.* 25(1): 71–94. https://doi.org/10.1615/InterJAlgae.v25.i1.50

Sinnott Q.P. 1981. A Simple Similarity Coefficient for Taxonomic Comparisons. *Taxon.* 30(1): 18–26.

Sneath P.H.A. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* 17(1): 201–226.

Sørensen T. 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*. *Biol. Skrifter/Kongel. Danske Videnskab. Selskab.* 5(4): 1–34.

Zhang M., Shi X., Chen F. Yang Z. 2021. The underlying causes and effects of phytoplankton seasonal turnover on resource use efficiency in freshwater lakes. *Ecol. Evol.* 11(41): 1–13.

Bren O.G. [1] (https://orcid.org/0000-0001-7423-9258)
Bren O.A. [2] (https://orcid.org/0009-0003-7632-6285)
Solonenko A.M. [2] (https://orcid.org/0000-0002-3417-5146)
Podorozhnyi S.M. [2](https://orcid.org/0000-0002-7702-7602)

[1] Institute of Botany, Czech Academy of Sciences,
135 Dukelská Str., Třeboň 37901, Czech Republic
[2] Bohdan Khmelnytskyi Melitopol State Pedagogical University,
Department of botany and horticulture,
59 Naukovoho mistechka Str., Zaporizhzhya 69000, Ukraine

Automation of Dice (Czekanowski-Sørensen) similarity index calculations in phycological research

This paper examines the trends in the use of the Dice (Czekanowski-Sørensen) similarity index in studies of algae and cyanoprokaryotes. A concise overview of the characteristics of this metric is provided, considering its positive aspects and limitations. The relevance of the work is justified by the researchers' need for automation of Dice index calculations and the construction of resulting matrices. The article proposes a method for automating calculations using macros in the Excel environment. The authors provide an overview of the possibilities of this approach and offer their own macro for fast and convenient calculation of the Dice index without the need for third-party programs or formulas.

K e y w o r d s : algae, cyanoprokaryotes, similarity measure, Dice (Czekanowski-Sørensen) index

C i t a t i o n . Bren O.G., Bren O.A., Solonenko A.M., Podorozhnyi S.M. 2024. Automation of Dice (Czekanowski-Sørensen) similarity index calculations in phycological research. *Algologia*. 34(1): 80–90. https://doi.org/10.15407/alg34.01.080