ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
IX, 1 (1966)

T. FREY (Tartu)

# ON THE SIGNIFICANCE OF CZEKANOWSKI'S INDEX OF SIMILARITY

A plant community may be more or less homogeneous. This means that the floristic composition of any given stand is more or less similar to any stand of that community. On the other hand, the assumption that a given plant association has a characteristic species composition raises the question of how much it differs from other kinds of communities. The comparison of stands in terms of the species present only is a crude way of classification. The question usually arises whether a more laborious quantitative examination is necessary to determine the degree of similarity of two or more stands. Hence, since the fundamental principles of any heretofore undertaken attempt at an objective evaluation of the degree of similarity between two species lists have never been consistently applied, this question is worth further consideration.

There are many methods whereby the actual degree of similarity (or diversity) of one stand to another as well as to the possible communities or associations may be assertained. Most of these (see Dagnelie [3], 1960) are based on the indices showing the number of species common to two species lists in relation to the total number of species in two areas (Jaccard's coefficient of community) or on suitable modifications (Greig-Smith [4], 1964).

Another widely used coefficient of similarity is generally attributed to Sørensen [13] (1948). According to Curtis [1] (1959) it was originally proposed by Czekanowski [2] (1913), and then developed with modifications by Polish ecologists (Kulczyński [7], 1927; Henzel [5], 1938; Matuszkiewicz [11], 1947; Motyka [12], 1947). Basically, it is a correlation coefficient intended for use when a number of different quantitatively expressed measures or attributes are available for the two compared entities. In connection with species present this becomes the number of species, common to the two areas, expressed as a percentage of the mean number of species per area; i. e. $2C \times 100/(A+B)$ where the two areas contain $A$ and $B$ species respectively and there are $C$ species in common.

The possibility of using an index of similarity in compiling a classification of a set of stands has been attracting attention for some time. Sørensen [13] (1948) calculated values of the index of similarity for all pairs of a set of stands and placed stands between which the value of the coefficient was at least 50 per cent in one group. Groups defined in this way were then associated into '2nd order groups' for which the limiting value of the coefficient was 40 per cent and so on. In this way an objective classification was produced.

Looman and Campbell (1960) [8] pointed out that the significance of a given value of the coefficient, as an indicator of the degree of similarity, varies with the total number of species involved and the number of species in each of the compared stands. They indicate that Sørensen's arbitrary levels did not lead to gross misgrouping. Some of their conclusions require modification. Namely, as Greig-Smith [4] (1964) points out, they used the $\chi^2$ value for two degrees of freedom instead of that for one degree of freedom which is proper for a $2 \times 2$ contingency table.

The way of calculation of the least number of species common to the two stands that will indicate association between the stands at, for example, 95 % point of probability, derives from Kendall's [6] formula

$$\chi^2 = r^2 S, \tag{1}$$

where $r$ is the correlation coefficient and $S$ is the total number of observations in the $2 \times 2$ contingency table. For that purpose, the following table may be used here:

| Number of species in | | Stand II | | Total |
|---|---|---|---|---|
| | | present + | absent − | |
| Stand I | present + | $C$ | $A-C$ | $A$ |
| | absent − | $B-C$ | $S-A-B+C$ | $S-A$ |
| Total | | $B$ | $S-B$ | $S$ |

$C$ — the number of species common to the two stands, $A$ — the number of species in the first, and $B$ — the number of species in the second of the stands being compared, $S$ — the total number of species in a given set of stands.

The correlation coefficient $r$ equals (Kendall, [6]):

$$r = \frac{C(S-A-B+C)-(A-C)(B-C)}{\sqrt{A(S-A)B(S-B)}} = \frac{CS-AB}{\sqrt{AB(S-A)(S-B)}}. \tag{2}$$

The critical value of $r$ at a given point of probability $P$, may be expressed according to (1) by the formula

$$r_P = \sqrt{\frac{\chi_P^2}{S}}, \tag{3}$$

where $\chi_P^2$ stands for the $P$-quantile in $\chi^2$ distribution with one degree of freedom.

Now, the critical number of species in common at the given level ($C_P$) must satisfy the equations (2) and (3). These yield the formula

$$C_P = \frac{AB + \sqrt{\frac{AB(S-A)(S-B)\chi_P^2}{S}}}{S}. \tag{4}$$

Such testing of the singificance of each comparison in isolation is definitely impractical owing to the amount of calculating time, but no simple test exists for the significance of the whole matrix of similarities (Greig-Smith, 1964), because $A$ and $B$ are in fact independent variables.

Since the use of the above criterion (4) is however required in more detailed investigations, it was decided to employ the electronic computer to prepare the corresponding tables of $C_P$ at $P = 0.95$. This work was conducted at the Tartu State University Computing Centre for the following values of $S$, $A$ and $B$

$S = 20$; $A$ (and $B$) $= 2, 3, \ldots, 20$,

$S = 25, 30$; $A$ (and $B$) $= 5, 6, \ldots, 25$,

$S = 40, 50$; $A$ (and $B$) $= 6, 7, \ldots, 40$,

$S = 60, 70$; $A$ (and $B$) $= 11, 12, \ldots, 60$,

$S = 80, 90, 100, 125, 150, 200$; $A$ (and $B$) $= 11, 12, \ldots, 75$.

These intervals have been selected to correspond to the widest benefit. However they may turn out to be incomplete in some cases. The tables for $S = 25$; $A$ (and $B$) $= 5, 6, \ldots, 25$ and for $S = 50$; $A$ (and $B$) $= 6, 7, \ldots, 50$ are shown at the end of this paper.

The coefficient of Czekanowski (i.e. the $K = 2C/(A+B)$ of Sørensen) has been applied by many workers, because it offers a very speed procedure if compared with other indices of this kind. Unfortunately, this advantage disappears when the corresponding criterion ($K_P$) for the whole matrix is required.

One of the possible ways to evaluate the $K$ values in relation to the criterion $K_P = 2C_P/(A+B)$ is the ratio of them, resulting in a more simple expression $(K/K_P =)\, C/C_P$, which seems to be preferred.

Actually, in drawing up the matrix of similarities for classification purposes it is sufficient to employ the *Ratio of Similarity*, defined as $T_P = C/C_P$, instead of any coefficient of similarity. The author used the $T_P$ in examining the interrelations of 117 plots in Estonian spruce forests (unpublished). It was found that, if using the tables of $C_P$, this ratio can be calculated for one pair of quadrats in 10 sec. (i.e. as rapid as the determination of the value of $K$). Consequently, the procedure can be suggested for use in preliminary classification of vegetation as a speedy technique with statistical background.

Now, the question of using the other (different from $P = 0.95$) levels of probability arises. As the Chi-square in formula (4) is placed under the square-root with four variates, they all together will introduce a relatively small proportion to the value of $T_P$. Thus it is possible to say in advance that in many cases (for example, Estonian spruce forests, for which the Ratio of Similarity had frequent values ranging from 0.5 to 2.8) the value of $T_{95\%} = 1.5$ corresponds to the $T_{99\%} = 1.0$, and the $T_{95\%} = 2.0$ to $T_{99,9\%} = 1.0$.

In different tables and their different parts the relationship between $C_{0.95}$ and $C_P$ does not, however, remain constant. Therefore the more precise expression:

$$C_P = R_{P,0.95} C_{0.95} - \frac{AB}{S}(R_{P,0.95} - 1), \tag{5}$$

is required, where:

$$R_{P,0.95} = \sqrt{\frac{\chi_P^2}{\chi_{0.95}^2}},$$

so that the values $R_{0.99,0.95} = 1.727$ and $R_{0.999,0.95} = 2.813$ might be used.

Finally, the question not only of similarity, but also of diversity may arise.

If the species complement of stand I is determined by certain causal factors, the null hypothesis states that the assemblage of stand II is independent of the factors ($r_{I,II} = 0$). From the $2 \times 2$ contingency table the expected number of species common to I and II equals

$$C_0 = \frac{AB}{S},$$

and the expected number of species absent both in I and II is:

$$D_0 = \frac{(S-A)(S-B)}{S}.$$

The alternative hypothesis depends on the assumption that the species complements are caused either by more or less the same ($1 > r > 0$)

| 23 | 24 | 25 | $A$ / $B$ |
|---|---|---|---|
| 5 | 5 | 5 | 5 |
| 6 | 6 | 6 | 6 |
| 7 | 7 | 7 | 7 |
| 8 | 8 | 8 | 8 |
| 9 | 9 | 9 | 9 |
| 10 | 10 | 10 | 10 |
| 11 | 11 | 11 | 11 |
| 12 | 12 | 12 | 12 |
| 13 | 13 | 13 | 13 |
| 14 | 14 | 14 | 14 |
| 15 | 15 | 15 | 15 |
| 16 | 16 | 16 | 16 |
| | | | $B$ |

[2] J. Czekanowski, *Zarys metod statystycznych*, Warszawa 1913.

[3] P. Dagnelie, *Contribution à l'étude des communautés végétales par l'analyse factorielle*, Bull. Serv. Carte phytogéogr., Sér. B, 5 (1960), p. 7-71 and 93-195.

[4] P. Greig-Smith, *Quantitative plant ecology*, London 1964.

[5] T. Henzel, *Zagadnienia metodologiczne w określaniu rasowym*, Przegląd Antropologiczny 12 (1938 — quoted from Curtis, 1959).

[6] M. G. Kendall, *Rank correlation methods*, London 1948.

[7] S. Kulczyński, *Zespoły roślin w Pieninach*, Bull. Intern. Polon. Acad. Sci. Lett., Cl. Sci. Math. et Nat., Ser. B, 2 (1927), p. 57-203.

[8] J. Looman and J. E. Campbell, *Adaptation of Sørensen's K (1948) for estimating unit affinities in prairie vegetation*, Ecology 41 (1960), p. 409-416.

[9] E. Marczewski and H. Steinhaus, *On a certain distance of sets and the corresponding distance of functions*, Coll. Math. 6 (1958), p. 319-327.

[10] E. Marczewski and H. Steinhaus, *O odległości systematycznej biotopów*, Zastosow. Mat. 4 (1959), p. 195-203.

[11] W. Matuszkiewicz, *Zespoły leśne południowego Polesia*, Ann. Univ. M. Curie-Skłodowska, Sect. E., 2 (1947), p. 69-138.

[12] J. Motyka, *O celach i metodach badań geobotanicznych*, Ann. Univ. M. Curie-Skłodowska, Sect. C. Suppl. 1 (1947), p. 1-168.

[13] T. A. Sørensen, *Method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons*, Det Kongelige Danske Videnskabernes selskab, Biol. ski.-r 5 (1948), p. 1-34.

ACADEMY OF SCIENCES OF THE ESTONIAN S.S.R.,
INSTITUTE OF ZOOLOGY AND BOTANY, TARTU

*Received on 20. 4. 1965*

---

**T. FREY (Tartu)**

## O POZIOMIE ISTOTNOŚCI WSKAŹNIKA PODOBIEŃSTWA CZEKANOWSKIEGO

STRESZCZENIE

W pracy rozpatruje się zagadnienia spotykane przy porównywaniu stanowisk biologicznych na podstawie występowania różnych gatunków. Autor nie próbuje przedstawić kompletnej dyskusji wszystkich możliwych wzorów, jakie są proponowane w obszernym piśmiennictwie dotyczącym tego tematu, bo w zasadzie wszystkie dotychczasowe metody nie różnią się zasadniczo. Wszystkim można zarzucić, że nie uwzględniają liczby $S$ wszystkich gatunków występujących w zespole wszystkich badanych stanowisk. Ze względu na to nie można niestety obliczyć poziomów istotności obliczanych dotychczas wskaźników podobieństwa.

W pracy sformułowano hipotezę zerową o niezależności czynników warunkujących występowanie gatunków na dwu wybranych stanowiskach i zaproponowano metodę jej testowania. Dolna i górna wartość krytyczna liczby $C$ gatunków występujących jednocześnie na obu stanowiskach są odpowiednio równe

$$C_P^- = 2C_0 - C_P \quad \text{oraz} \quad C_P^+ = C_P,$$

gdzie $C_0$ oznacza wartość oczekiwaną liczby gatunków wspólnych dla dwu stanowisk w warunkach, gdy komplety czynników wpływających na zestawy gatunków $A$ i $B$ są stochastycznie niezależne.

Dla uproszczenia rachunków przy stosowaniu proponowanej metody sporządzono tablice górnych wartości krytycznych $C_P$ dla poziomu ufności $P = 0{,}95$. Przykłady takich tablic podane są w końcu pracy.

---

Т. ФРЕЙ (Тарту)

## О СТАТИСТИЧЕСКОМ УРОВНЕ ПРИ ИНДЕКСЕ СХОДСТВА ЧЕКАНОВСКОГО

РЕЗЮМЕ

В данной статье уделяется внимания вопросу о сравнении пробных площадей на основе присутствия разных видов. Здесь мы не пытаемся дать детального изложения и сравнения всевозможных формул, приведённых в обширной литературе, так как в принципе они существенно не различаются. Все эти попытки сравнения подвергаются критике в том, что не учитывается общее число видов $S$ в данной серии пробных площадей. При таком подходе, к сожалению, мы не имеем возможностей установить статистического уровня достоверности.

Данная статья содержит первоначальное определение нолевой гипотезы с соответствующим объяснением способа установления уровня достоверности.

Нижний и верхний критические уровни для статистической проверки нолевой гипотезы выражаются как

$$C_P^- = 2C_0 - C_P, \quad \text{и} \quad C_P^+ = C_P,$$

где $C_0$ обозначает математическое ожидание числа общих для пробных площадей I и II видов в обстановке, при которой сочетания факторов, определяющие видовые составы $A$ и $B$, являются статистически независимыми.

С целью сокращения вычислений подготовлены таблицы верхнего критического уровня $C_P$ для достоверности $P = 0{,}95$, примеры которых приведены (см. таблицы 1 и 2) в конце статьи.

---