A Dissimilarity Coefficient and Its Use

CHARLES D. BONHAM

Abstract

A coefficient of dissimilarity was used to test hypotheses concerned with standing crop of individual plant species occurring on two soil types and subjected to two levels of cattle grazing. It was concluded that variations in relative biomass of individual species changed with grazing, but biomass by species did not vary over the growing season on deep sandy soils.

The need to study range ecological associations on a quantitative basis has long been recognized. As more quantitative range research is conducted, subsequent analysis of data will be facilitated if probabilistic statements can be used to make inferences or draw conclusions about the data. Non-normality of most ecological data limits their use in statistical analysis.

A specific need exists for testing values obtained from similarity coefficients, since these coefficients are used extensively in making range ecological comparisons. These coefficients are certainly useful in assessing differences among various groups of data which are thought to be similar in some respect, such as species composition of vegetation or dietary composition. However, the probability distributions of these coefficients have not been studied. Yet, these distributions could be used in making quantitative comparisons of various combinatorial occurrences of ecological similarities. Probability statements, coupled with ecological knowledge, could enhance the significance of studies of range ecological associations.

Quantitative approaches to plant community comparisons have been employed by plant ecologists for several decades (Gleason 1920, Archibald 1949, Cottam and Curtis 1949, Bray and Curtis 1957, Curtis 1959). Reviews have been made concerning the applications of Jaccard's (1902) community coefficient (Gleason 1920, Cottam and Curtis 1949, Bray and Curtis 1957, Curtis 1959). Jaccard compared community A with community B by expressing the number of species common to both communities as a percentage of the total number of species for A plus B. Cottam and Curtis (1949) concluded that the correlation coefficient previously proposed for species associations could not be used for stand comparisons.

Species presence, frequency, basal area, and density data from plant communities have been used singly and in combination to arrive at a community coefficient based on modifications of Jaccard's formula (Gleason 1920, Dice 1945, Cottam and Curtis 1949, Curtis and Greene 1949, Hanson 1955, Bray and Curtis 1957, Curtis 1959, Dix 1959). Studies of alpine communities have also used Czekanowski's index (Ward 1961, Holway 1962, Bonham 1966).

Another coefficient is relatively unknown but should be considered only because it allows some important ecological similarity hypotheses to be tested statistically. In particular, interest may be in determining whether the amount of similarity between two communities differs significantly from the similarity of two other communities. This approach in comparison is perhaps more easily tested statistically than a single value of similarity between two communities. The index is actually one which measures the degree of dissimilarity rather than similarity based on the fact that the amount of variance in a sample group is a common measure and indicates the degree of dissimilarity relative to the mean value. I developed the test procedure for R.M. Hansen to use for marmot diet comparisons (Hansen 1975). The value could be appropriately referred to as the Community Dissimilarity Coefficient (CDC) and defined as

$$CDC = \frac{\sum_{n}(f_{Ai} - f_{Bi})^2}{n} \quad (1)$$

where f_i is the relative measure of species i characteristics (cover, etc.) encountered in the community sampling process. The subscripts A and B represent communities A and B, respectively, which are being compared for their amount of dissimilarity and n is the total number of species encountered in the sample. It is desirable to know whether this coefficient differs significantly from another dissimilarity value for two other communities. In which case, a statistical test can be developed. The value of CDC is the average of squared differences as determined over all species in the two communities.

It can be shown that if

$$E(x) = \mu, \quad (2)$$

where μ is the measure of the population mean for a given x measurement, then the expected value, E(x), or mean of CDC, is

$$E(CDC) = V(f_A) + \mu^2_{f_A} + V(f_B) + \mu^2_{f_B} - 2\mu_{f_A}\mu_{f_B}, \quad (3)$$

where V is the variance and f_A and f_B are the characteristics of interest in communities A and B, respectively. If relative measurements are used for all species, and thus sum to 100%, and the same species are assumed to occur in both communities, then $\mu_{f_A} = \mu_{f_B}$ and equation (3) reduces to

$$E[CDC] = V(f_A) + V(f_B). \quad (4)$$

It follows then that the mean of the squared differences in equation (1) reduces to equation (3) for unequal means and equation (4) for equal means. The latter will always be the case when data are relative values such as the example in this paper.

Therefore, to test CDC_1 (communities A and B) against CDC_2 (communities C and D) for the two community groups, respectively, for statistical significance involves a variance ratio test because the mean squared differences are a sum of two variances when two groups of communities are independent and the means are equal.

A test of CDC differences is easily made by forming a ratio of the larger variance to a smaller variance which follows the *F*-distribution with (n_1-1) and (n_2-1) degrees of freedom and where n_1 and n_2 are sample sizes from community groups 1(A and B) and 2 (C and D), respectively. The condition $n_1 = n_2$ must be met to obtain equation (4). Therefore,

$$F_{(n_1-1,\, n_2-1)} = \frac{\text{Larger CDC}}{\text{Smaller CDC}}. \quad (5)$$

Author is professor, Department of Range Science, Colorado State University, Fort Collins, Colorado 80523.

Research was supported by the Colorado State University Experiment Station and published as Scientific Series Paper No. 2552.

Manuscript received May 23, 1980.

Table 1. Values of the CDC calculated for herbage yields of species on two soil types, on grazed and ungrazed areas, and for two seasons of growth.

1	2	3	4	5	6	7	8	Community	Growth period	Grazing	Soil type
—	.010	.001	.015	.009	.006	.010	.013	1	July	Grazed	Deep sand
		.008	.005	.032	.013	.034	.019	2	July	Ungrazed	Deep sand
			.014	.014	.007	.015	.016	3	Sept.	Grazed	Deep sand
				.035	.009	.037	.017	4	Sept.	Ungrazed	Deep sand
					.013	.000	.025	5	July	Grazed	Sandy plains
						.015	.001	6	July	Ungrazed	Sandy plains
							.027	7	Sept.	Grazed	Sandy plains
							—	8	Sept.	Ungrazed	Sandy plains

Confidence limits for an individual value of CDC are calculated using Chi-square values and

$$P\left[\frac{CDCn}{\chi^2_{(1-\alpha),\,n}} < CDC < \frac{CDCn}{\chi^2_{(\alpha),n}}\right] = 1 - \alpha. \qquad (6)$$

An estimation of the individual variance terms from equation (4) will facilitate an understanding of the testing procedure. By definition,

$$V(f_{Ai}) = E\,[f_{Ai} - \mu_{Ai}]^2. \qquad (7)$$

Since each f_{Ai} is an average or mean, the Central Limit Theory of statistics is useful for obtaining an estimate of the $V(f_{Ai})$ and μ_{Ai}. That is, means tend to be normally distributed with a mean μ and a variance of σ^2/n, where σ^2 is the variance of the observations. Since E(CDC) is a variance, E(CDC) in equation (3) can be obtained from

$$V(f_{Ai}) = \frac{\sum^{n}(f_{Ai} - \bar{f}_{Ai})^2}{n-1}\left(\frac{1}{n}\right), \qquad (8)$$

where

$$\bar{f}_{Ai} = \frac{\sum^{n} f_{Ai}}{n} = \frac{100}{n} \qquad (9)$$

since f_{Ai} is a relative measure and sums to 100%. Furthermore, a rearrangement of equation (8) results in

$$V(f_{Ai}) = \frac{\sum f_{Ai}^2 - \frac{(\sum f_{Ai})^2}{n}}{n-1}\left(\frac{1}{n}\right) \qquad (10)$$

$$= \frac{(\sum f_{Ai}^2) - \frac{(10{,}000)}{n}}{n(n-1)} \qquad (11)$$

An Example

I chose a data set consisting of individual species standing crop measured on two soil types (deep sand and sandy plains) with and without a history of cattle grazing and used two seasons for observations, July and September. The hypotheses of interest included (a) there are no differences in individual species standing crop between soil type, (b) there are no species biomass composition changes due to cattle grazing on a given soil type, and (c) there are no differences in seasonal responses of plants relative to biomass with respect to soil types and grazing.

To test any one of the several hypotheses, I used standing crop data for 40 species which occurred in at least one of the areas sampled. Some species occurred only under one condition, i.e., ungrazed deep sand area. The data were then converted to a relative percentage basis and these were used to calculate a CDC value (equation (1)) for each combination of soil type, grazed condition, and season (Table 1).

A test of differences in variation of relative species biomass for grazing between July and September and ungrazed plants on deep sand involved the ratio of community groups 1 and 3 compared to 2 and 4 (Table 1). This ratio is $0.005/0.001 = 5.0$ which is an F value with 33 d.f. for each group. Degrees of freedom depend only on the number of species that occurred in at least one community. The tabular F is 1.80, which is significant at $p = 0.05$. The sandy plains soil type is also significantly different at $p = 0.05$. The latter ratio resulted from 5 and 7 compared to 6 and 8 (Table 1). Therefore, it was concluded that the variation in relative biomass of individual species did change with grazing. Ecological implications were not assessed for this paper.

A test of differences in variation caused by grazing vegetation on the soil types involved the ratio of community groups 2 and 4 compared to 6 and 8 (Table 1). This ratio was significant at $p = 0.05$. Thus, variability was different for the two soil types when ungrazed because the deep sand had a larger variation in species relative biomass. The grazed areas likewise differed by soil type and the difference occurred since species biomass varied between July and September in sandy plains soil.

Does variability in plant species biomass differ from grazing vegetation of deep sandy soils from July to September? This question is answered by the ratio of 3,4 over 1,2 from Table 1. $F = 0.14/.010 = 1.4$, which is not significant at $p = 0.05$. Therefore, species biomass does not vary over the season as a result of grazing vegetation of deep sandy soils.

Any two communities can be compared by splitting each community into halves with regard to sampling, calculating the variance (CDC) for each community and comparing this variance (CDC) to another community variance (CDC) of interest. Recall that the CDC is calculated by differences on a specie-by-specie basis (equation (1)). Temporal comparisons can be made for a single community by partitioning the community into two parts and observing the measured characteristics over two or more time periods.

The proposed coefficient which measures the amount of dissimilarity existing between two groups of communities has wide applications. The approach permits testing for statistical differences as long as communities (or parts) can be paired on a basis that is ecologically meaningful, calculating a variance and testing the latter value against another variance of a pair of communities. A confidence interval which gives the range of variations expected to occur between the two communities can also be placed on the CDC of any pair-wise communities.

Literature Cited

Archibald, E.E.A. 1949. The specific character of plant communities. II. A quantitative approach. J. Ecol. 37:274-288.

Bonham, C.D. 1966. An ordination of alpine (*Deschampsia caespitosa* (L.) Blauv.) meadows. Ph.D. Diss., Colorado State Univ., Fort Collins.

Bray, J.R., and J.T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27:325-349.

Cottam, G., and J.T. Curtis. 1949. A method for making rapid surveys of woodlands by means of pairs of randomly selected trees. Ecology 30:101-104.

Curtis, J.T. 1959. The Vegetation of Wisconsin: An Ordination of Plant Communities. Univ. Wisconsin Press, Madison. 657 p.

Curtis, J.T., and H.C. Greene. 1949. A study of relic Wisconsin prairies by the species-presence method. Ecology 30:83-92.

Dice, L.R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297-302.

Dix, R.L. 1959. The influence of grazing on the thin-soil prairies of Wisconsin. Ecology 40:36-49.

Gleason, H.A. 1920. Some applications of the quadrat methods. Bull. Torrey Bot. Club 47:21-33.

Hansen, R.M. 1975. Foods of the hoary marmot on the Kenai Peninsula, Alaska. Amer. Midl. Natur. 94:348-353.

Hanson, H.C. 1955. Characteristics of the *Stipa comata-Bouteloua gracilis-Bouteloua curtipendula* association in northern Colorado. Ecology 36:269-280.

Holway, J.G. 1962. Phenology of Colorado alpine plants. Ph.D. Diss. Colorado State Univ., Fort Collins.

Ward, R.T. 1961. An ordination of alpine vegetation. Bull. Ecol. Soc. Amer. 43:147-148.