An Optimum Sampling Strategy for Plant Species Frequencies

CHARLES D. BONHAM

Highlight: *An optimum sampling strategy was developed for predicting frequency values for all plant species occurring in an area. The approach uses both multistage and double-sampling procedures to predict the frequency of occurrence of all species. The application of these procedures to one grassland area resulted in an average saving of 26% in the number of sample quadrats required to determine the frequency of all species.*

The author is with the Range Science Department, Colorado State University, Fort Collins, 80523.

This study was supported by the U.S. Atomic Energy Commission through contract No. AT (11-1)-2115 to the author.

Manuscript received April 30, 1975.

Natural vegetation is becoming increasingly more important in detecting, evaluating, and monitoring environmental impacts. Consequently, more quantitative information is needed for vegetation characteristics so that significant vegetation changes can be detected. A number of structural characteristics of vegetation including density, cover, and species composition have been used for many years by plant ecologists in describing plant communities. Rangeland managers have also found many of these characteristics to be valuable for determination and evaluation of trends in range condition.

Frequency of plant species has been

used as a measure to quantitatively describe vegetation. Frequency continues to be a popular measure since it is easily obtained in spite of disadvantages pointed out by some quantitative ecologists (Greig-Smith, 1964). Frequency is defined as the chance of finding a species in a quadrat (Greig-Smith, 1964). Plant frequencies have also been described as representing a blend of density and dispersion characteristics (Hyder et al., 1965). It has also been noted that frequency characteristics of perennial vegetation are seasonally stable and were useful for the classification of range sites and evaluating plant responses to grazing (Hansen, 1934). The advantages of simplicity, objectivity, and speed should be gained when frequency measurements are substituted for commonly used cover and basal area techniques (Hyder et al., 1965). Although collection of frequency data is rapid, the use of multistage and double-sampling techniques should further increase sampling efficiency.

This paper presents the results of a study of an optimum sampling procedure for estimating frequencies of plant species. The study was conducted on mixed grassland vegetation type at The Research Ranch, Elgin, Arizona. The basic approach was to combine methods for multistage sampling with those for double sampling to obtain an optimum sampling strategy.

Methods

The hypothesis in this study was that a measure of association between two species could be used to predict the frequency value for one of the species. That is, given the quantitative value for association between any two species of a sample area, the frequency of one species could be used to predict the frequency of the second species. Thus, basic assumptions made in the development and testing of the model from which species frequencies were estimated included:

1) The predictor plant species were widely distributed. That is, they had a wide range of environmental tolerances over the sampling area. This did not imply that predictor species had to have large frequency values, but rather that the species were ubiquitous.

2) The precision of the estimated frequencies for a species in the area must be independent of any individual species pattern occurring in the area. Since the species used in the prediction equation had to be widely distributed, any patchiness in the predicted species caused errors in predicting frequencies. The magnitude of these errors was assumed to vary according to the degree of patchiness, intensity of sampling, and size of the area under study.

3) The association index used in the model must have represented the true relationship that existed between the predictor species and the predicted species frequencies for the area sampled.

Like most assumptions underlying theoretical development of ecological models, these were somewhat relaxed in practical applications without invalidating the usefulness of the approach. For example, the importance of assuming that predictor plant species had to be widely distributed was relative in that the exact frequency value of a given species over the sample region is not known. Furthermore, the assumption concerning species patterns which occur throughout a sample area would be difficult to deal with in practice. Individual species have patterns which are peculiar to them under differing environmental conditions, and no single species can be described as to its pattern over an entire area under various environmental conditions. Therefore, any errors which arise as a result of variations in these patterns were assumed to be included in the error for predicting frequencies. The third assumption also includes one of linearity in the relationship that exists between two species. Such linearity seldom exists in nature but can be used for first order approximations to the frequencies of other plants. These assumptions appeared to be adequate for the development and applications of the approach.

In order to test the hypothesis, 25 stands of grassland vegetation with dimensions of 10 m x 10 m were randomly selected from a grid of 100 stands occurring at 200-m intervals in 10 rows and 10 columns. Each stand was relatively homogeneous within in respect to dominant species cover. A subsampling unit (referred to hereafter as a quadrat) 40 cm x 40 cm in size was used to obtain 50 random observations for all species frequencies within each stand. A complete description of the 324 ha (800 acres) has been described previously as to environmental and vegetational characteristics (Bonham, 1974).

The quadrat data were used to compute all possible species association values, taking two species at a time and arranging the data in a 2 x 2 table according to Dice (1945) (Fig. 1). Several other such measures are available, but this one was chosen because of the author's experience with it. Let A_1 be the association of species **B** with species **A**. Then

$$A_1 = \frac{a}{a + b} \qquad (1)$$

where **a** is the number of quadrats containing both species and **(a + b)** is the total number of quadrats containing species **A**. That is, A_1 equals the proportion of quadrats in which species **A** occurred in the presence of species **B**. A_2 is the association of species **A** with species **B**.

$$A_2 = \frac{a}{a + c} \qquad (2)$$

where **a** is previously defined and **(a + c)** is the total number of quadrats containing species **B**. Then A_2 equals the proportion of quadrats containing species **B** in the presence of species **A** (Fig. 1). These two indices were combined and used to predict all species frequencies in the area by the following relationships:

$$A_1 (a + b) = a \qquad (3)$$

and

$$A_2 (a + c) = a. \qquad (4)$$

Therefore,

$$A_1 (a + b) = A_2 (a + c) \qquad (5)$$

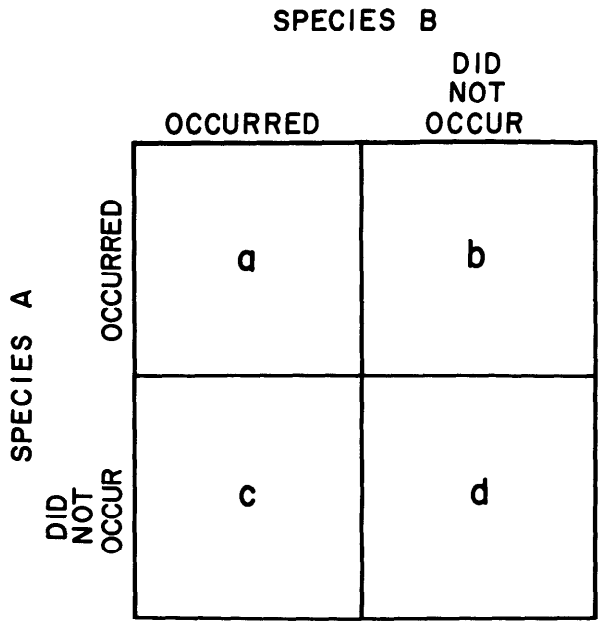
and

$$\frac{A_1}{A_2} (a + b) = (a + c) \qquad (6)$$

which is an estimate of the total number of quadrats containing species **B** had all quadrats been examined for its presence.

Thus, the frequency of a species can be estimated by dividing equation (6) through by the total number of quadrats sampled only for the presence of species **A**.

The selection of the species to be used in predicting the frequencies of other species was based on several criteria. A single species could have



a = number of joint occurrences of Species A and Species B
b = number of occurrences of Species A without Species B
c = number of occurrences of Species B without Species A
d = number of times neither species occurred

Fig. 1. *The quadrat data summary method to calculate species association values.*

been selected solely on the basis of its precision in estimating the occurrence of all other species. However, it was obvious that such an individual species did not exist. Furthermore, several species would estimate the occurrence of all remaining species more efficiently than any one species alone. The Central Limit Theorem of statistics was then applied in obtaining an estimate of the number of occurrences of all other species. That is, the precision of an estimate is related to the number of samples used in the estimation of the parameter and is some function of

$$\frac{1}{n_s} \qquad (7)$$

where n_s was the number of species used to obtain the estimate. Ideally, a set of species which produced a minimum standard error of the estimate should have been selected. While statistically this may be possible, it was not practical since each possible combination of species should have been tested for their combinatorial effect on the standard error of the estimate for all other species frequencies.

A set of species was selected which satisfied the stated assumptions and which gave a minimum variance for occurrences in stands as well as quadrats. Ecologically, these were species which were either dominant or major associates that occurred throughout the area under consideration. In addition to these possible selections, there were also ubiquitous species which were not necessarily abundant but did occur throughout the area. That is, if a species occurs consistently then its variance of occurrence is smaller. It follows that the variance of occurrences for these species in stands would not be as large over the area as for species which were more localized in their occurrence.

Hairy grama (*Bouteloua hirsuta*), Havard threeawn (*Aristida barbata*), spidergrass (*Aristida ternipes*), sprucetop grama (*Bouteloua chondrosioides*), and curly mesquite (*Hilaria belangeri*) were five grasses that occurred throughout the area. These species were not necessarily equally abundant in the area, but all fit the basic assumptions outlined. These five species were used to develop an optimum sampling procedure for estimating the frequency of all other species.

Multistage Sampling Procedures

A cost model for the relationship of sampling stands and quadrats was assumed to be of the form (Cochran, 1953)

$$C = n c_s + n m c_q \qquad (8)$$

where **n** was the number of stands and **m** was the number of quadrats. **C** was the total cost, while c_s and c_q were the costs for sampling each stand and quadrat, respectively. Obviously, it was more costly to sample an entire stand than to sample a single quadrat. Therefore, c_s was much larger than c_q and a combination of **n** and **m** was used to minimize the variance of the estimate, a+c from equation (6), for a given cost. However, my interest was in obtaining an estimate of the proportion of quadrats occupied, so the estimate was a+c/nm. This proportion was converted to a percentage at the end of the calculations since the development was based on proportions.

The following relation exists between number of quadrats needed and variances and costs for stands and quadrats:

$$m = \sqrt{S_q^2 C_s / S_s^2 C_q} \qquad (9)$$

where S_s^2 and S_q^2 were the variance of a + c/nm from stands and quadrats, respectively. Again, the variable to be estimated by sampling stands and quadrats was the proportion of the total number of quadrats occupied by a species. The value for this variable ranged between 0 and 1, and its variances, S_q^2 was estimated by (Hyder et al., 1965) as

$$S_q^2 = m (\Sigma p_i q_i) / n (m-1) \qquad (10)$$

where p_i was the frequency of a given species in stand **i**, $q_i = (1-p_i)$, **m** equaled number of quadrats per stand and **n** was the number of stands sampled.

The stand variance, S_s^2 was estimated by (Hyder et al., 1965)

$$S_s^2 = \frac{\Sigma p_i^2 - (\Sigma p_i)^2 / n}{n-1} - \frac{S_q^2}{m} \qquad (11)$$

where i = 1, 2, . . . **n** (the number of stands sampled). The relative cost of observing stands compared to quadrats was estimated to be 625 to 1, since a complete stand contained 625 quadrat-areas and other factors such as location time were about equal for this study. Equation (8) becomes

$$n = 100 / 1 + .16(m) \quad (12)$$

on a relative basis (Cochran, 1953).

Table 1 provides the variances for the five species selected as predictor species as well as the optimum number of quadrats to be taken within stands. The optimum number of stands needed to sample the occurrences of these species was also included (Table 1). The latter numbers were obtained by using the respective variances to solve equations (9) and (12). An overall average number of quadrats and stands was used to sample for occurrences of the five predictor species. These averages were also used to determine the double-sample size necessary to observe the occurrences of all species in relation to the five predictor species.

Double-Sampling Procedures

Eleven stands were randomly selected from the remaining 75 stands and 50 quadrats were randomly sampled within each of these stands. These 11 stands and 50 quadrats were the average values obtained by solving equations (9) and (12) for the five predictor species (Table 1). These numbers were the optimum multistage sampling numbers to use in further sampling.

Cochran (1953) provided a method for determining the number of quadrats necessary for sampling the occurrences of all other species in a double-sampling fashion by the relationship

$$\frac{k}{\sqrt{V_k c_m}} = \frac{m}{\sqrt{V_m c_k}} \quad (13)$$

where **k** was the number of quadrats to be sampled for the occurrences of all species, **m** was the number of quadrats to be sampled for the occurrences of the five species, V_k was the variance due to regression between the occurrences of one of the five species and the predicted occurrences of another species, and V_m was the variance of the deviation from the regression just described. The values of c_m and c_k were the estimated costs for obtaining **m** and **k** samples, respectively, and were estimated to be in a ratio of 1:4, respectively. That is, it took four times longer to observe and record the occurrence of all species in a quadrat than to observe and record only the occurrence of the five selected species used as predictors.

Equation (13) becomes

$$\frac{k}{m} = \sqrt{(\Sigma y^2)(r^2)} / \sqrt{(\Sigma y^2)(1-r^2)(4)} \quad (14)$$

where **r** was the measure of linear correlation between the predicted species occurrences and the predictor species occurrences and Σy^2 was the sum of squares of the number of occurrences of the species to be predicted. The ratio of **k/m** was used in the cost equation for **k** and **m** as follows:

$$C = m + k \quad (15)$$
$$= 50.$$

This restriction of relative cost of total sampling within a stand was the result of the optimum number of quadrats needed for obtaining frequencies of the five predictor species. Therefore, this imposed maximum of 50 quadrats was used to determine optimum double-sample numbers. These optimum numbers of quadrats differed for each species frequency to be predicted, as well as for predictor species. Equations (14) and (15) were solved for each species frequency to be predicted using the five predictor species at one time. Average number of stands and quadrats/stands needed for double sampling all species were then obtained for each predictor species (Table 1).

Eleven stands of vegetation were selected as previously described with 50 quadrats being located at random within each of these stands. Observations were made for the occurrences of the five species in the 50 quadrats, while 37 quadrats were observed for the presence of all other species. The 37 quadrats were used to determine the association values of equations (1) and (2). The association ratios were formed and multiplied by the total number of quadrats occupied by the occurrence of one of the predictor species in order to obtain an estimate of the total number occupied by the predicted species [Equation (6)]. That is, the estimate was obtained for the number of quadrats occupied by a predicted species had all 50 quadrats been sampled for its presence.

Determination of Precision

Finally, the precision in estimating the total number of quadrats having an individual species using the model was determined. The variance of this estimate for an individual species occurrence using an individual predictor species was obtained by taking the variance of equation (6) which was

$$V(a+c) = (A_1/A_2)^2 \; V(a+b) \quad (16)$$

where **V(a+b)** was the estimated variance for number of occurrences of an individual predictor species obtained from equation (17). This latter variance was obtained from

$$V(a+b) = \sum_{i=1}^{11} [(a+b)_i - \overline{(a+b)}]^2 / 10 \quad (17)$$

where 50 quadrats in each of 11 stands were used to obtain estimates of $(a+b)_i$, the number of occurrences of an individual predictor species in the i^{th} stand. The mean occurrences of the predictor species over all 11 stands was $\overline{(a+b)}$. The ratio, A_1/A_2, was considered as a constant over the sample area as stated in the assumptions and was calculated from equations (1) and (2) using the 407 quadrats (11 times 37). This was the total number of quadrats that were double sampled.

The variances for the average predicted number of occurrences of individual species using five predictor species included two sources. One was caused by the variation in the number of occurrences of the five predictor species in the sampling process, while the other involved the association ratios which varied from one predictor species to another. Therefore, the standard error of the estimated total number of occurrences of a species was calculated using five observations since only five species were used as predictors.

Table 1. Variances of species proportions for quadrats and stands and number of sampling units needed for five predictor species. (The numbers in parentheses are double samples needed based on a sample size of 50 for the predictor species.)

Species	Quadrat variance	Stand variance	Number of Quadrats	Number of Stands
Aristida barbata	0.30	0.09	46 (36)	12
Aristida ternipes	0.29	0.05	60 (39)	9
Bouteloua chondrosioides	0.25	0.12	39 (39)	14
Bouteloua hirsuta	0.27	0.12	37 (34)	14
Hilaria belangeri	0.31	0.04	70 (37)	8
Average samples needed			50 (37)	11

Results and Discussion

Estimates for the occurrences of 31 additional species were obtained using the five predictor species and the optimum sampling strategy. Table 2 gives the number of times a species was predicted to occur based on an average value obtained from the five

predictor species. The standard error of these estimates is small relative to their respective means. Furthermore, most of the standard errors of the predicted frequencies are zero for practical purposes. This can be verified by multiplying the standard errors of Table 2 by (100/550) which is necessary for conversion to a percentage value.

Species which have a contagious pattern of occurrences were not predicted as precisely as the more ubiquitous species. These patterns for some species restricted their joint occurrence to only one of the predictor species. The average predicted value is very low if all five predictor species are assumed of equal weighted importance. For example, some species may be widespread but locally abundant in at least one of the sample stands. This would present some problems in predicting their average frequency values with a desirable level of precision. Therefore, it becomes a matter of determining the average frequency value of a species in the general area or to be specific to individual stands. If the average species frequency is desirable over the area, then all predictor species should be used to predict the remaining species frequencies. However, if interest centers on the localized frequency of an individual species, then only the predictor species having some joint association with that particular species should be used for predictive purposes.

A wide range in number of quadrats to be double sampled was obtained for the various species to be predicted. Blue grama (*Bouteloua gracilis*), for example, required only 20 plots to be sampled out of 50 when using spidergrass as the predictor species. This was in contrast to black grama (*Bouteloua eriopoda*) where all 50 plots needed to be sampled using spidergrass as the predictor species. However, 38 plots were required to double sample black grama compared to 23 double-sampled quadrats for blue grama when hairy grama was used as the predictor species.

Hairy grama, compared to the other predictor species, required the least number of double samples on the average to estimate all other species occurrences in an optimum way for the area (Table 1). An average of 34 double samples were required using this predictor species, whereas 39 double samples were required when spidergrass and sprucetop grama were used as individual predictor species.

Havard threeawn and curly mesquite were intermediate in their double sampling efficiency. That is, these two species needed between 30 and 37 double samples to predict all remaining species of the area. This was in contrast to spidergrass, which had a large range in double samples required to estimate the occurrences of other species. This latter species required from 20 to 50 samples for an optimum double sample for the occurrence of another species, depending on the individual species. For example, the occurrences of blue grama, common evolvulus (*Evolvulus sericeus*), and bundleflower (*Desmanthus cooleyi*) could be adequately estimated using 20 double samples with spidergrass as a predictor species. On the other hand, at least 49 samples were needed to estimate the occurrences of sida (*Sida procumbens*), leatherweed croton (*Croton corymbulosus*), and wolftail (*Lycurus phleoides*) based on spidergrass occurrence.

Hairy grama, when used as a predictor species, followed about the same pattern as spidergrass in predicting occurrences of certain species. That is, only 19 to 20 double samples were required to predict the occurrences of blue grama, common evolvulus, and bundleflower on the area. This is in contrast to the 35 to 40 quadrats required for sampling the occurrences for most remaining species when these two species were used as predictor species.

The efficiency of double sampling for species frequencies ranged from 22 to 32% savings in number of sample quadrats required. The average saving was 26% when estimating occurrences of all species of an area using the five predictor species.

When emphasis is centered on predicting the occurrences of certain species, then it is not advisable to use the average value of several species for predictive purposes. Instead, an individual species should be selected which has the highest overall association values with the species to be predicted. This would result in an increased precision for predicting occurrences of all other species than would an average of several predictor species. However, in this study it was found that several predictor species served to predict the occurrences of all other species in the area more precisely than did the use of only one or two predictor species. This undoubtedly was the result of at least some heterogenous patterns that existed for a few of the species.

In any case, the use of more than one species to predict the occurrences of other species involves the averaging

Table 2. Average number of predicted occurrences and standard errors of species using five predictor species based on 550 quadrats.

Species	Average	S.E.	Frequency (%)	S.E.*
Asclepias asperula	11	0.1	2.0	–
Aster tanacetifolius	18	0.1	3.3	–
Bailea multiradiata	4	0.0	0.7	–
Bouteloua curtipendula	22	0.2	4.0	–
Bouteloua eriopoda	1	0.0	0.2	–
Bouteloua filiformis	1	0.0	0.2	–
Bouteloua gracilis	151	1.2	27.5	0.2
Brayulinea densa	109	0.8	19.8	0.1
Chloris virgata	6	0.0	1.1	–
Croton corymbulosus	24	0.5	4.4	0.1
Desmanthus cooleyi	60	0.3	10.9	0.1
Eragrostis intermedia	36	0.3	6.5	0.1
Eriogonum wrightii	1	0.0	0.2	–
Evolvulus arizonicus	65	0.5	11.8	0.1
Evolvulus sericeus	158	1.2	28.7	0.2
Haplopappus gracilis	77	0.6	14.0	0.1
Lycurus pheoides	171	1.3	31.1	0.2
Mimosa dysocarpa	23	0.2	4.2	–
Panicum hallii	9	0.1	1.6	–
Panicum obtusum	3	0.0	0.5	–
Penstemon dasyphyllum	15	0.1	2.7	–
Psoralea tenuiflora	3	0.0	0.5	–
Senecio douglasii	3	0.0	0.5	–
Sida procumbens	95	0.7	17.3	0.1
Trichachne californica	69	0.5	12.5	0.1
Zinnia grandiflora	2	0.0	0.4	–

*Only one significant decimal recorded and the dash = < 0.1%.

of phytosociological relationships among species over their differing responses to habitats. If species responded in the same way to environmental conditions, then the accuracy of the predicted occurrences obviously would be greater. This is not usually the case and any differences in environmental responses must be average among several species.

Once the optimum number of stands has been determined, selection of stands to be sampled should probably not be made on a random basis. Randomization was used in this study only to determine the statistical precision of the model. However, under normal field sampling conditions, one is usually interested in estimating the occurrence for all species found in the area. Therefore, the optimum number of stands should be allocated in such a way as to include the occurrence of as many species as possible. This procedure does not violate any assumptions of the model, since the purpose is to have all species represented. Moreover, the same approach could be used to study species occurrences or frequencies according to vegetation types. In this case, the procedures developed in this paper should be applied to each of the vegetation types and sampling should be conducted using the optimum strategy for each of the types.

Computer programs were developed for data processing in this study. The ideal application of applying optimum sampling strategies for plant frequencies as developed in this study would include the use of a stepwise computer procedure. That is, a small sample should be obtained initially for processing through the multistage double-sampling model to obtain optimum stand and quadrat numbers. These additional sample numbers could then be collected and processed until sufficient sampling has occurred in a multistage double-sampling fashion.

The procedure developed in this paper might become an important application in field sampling, as more emphasis is placed on determination of quantitative relations among plant species. The optimum sampling strategies developed for estimating species occurrences should greatly reduce the amount of field work necessary to determine frequency values for species in a given area. In turn, such a procedure will permit a collection of objective data for description of vegetation structure in a more efficient manner.

It is believed that much more efficiency can be derived by using this approach than has been shown in this developmental work. The vegetation of the study area, while homogeneous in some respects, also had some habitat diversity. Obviously, more efficiency in sampling for species frequencies is gained by using the model developed in this study as the homogeneity of the vegetation composition increases.

Literature Cited

Bonham, Charles D. 1974. Classifying grassland vegetation with a diversity index. J. Range Manage. 27:240-243.

Cochran, W. G. 1963. Sampling Techniques. 2nd Ed. John Wiley and Sons, Inc. New York. 330 p.

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

Greig-Smith, P. 1964. Quantitative Plant Ecology. 2nd Ed. Butterworth, London. 256 p.

Hanson, Herbert C. 1934. A comparison of methods of botanical analysis of the native prairie in western North Dakota. J. Agr. Res. 49:815-842.

Hyder, D. N., R. E. Bement, E. E. Remmenga, and C. Terwilliger, Jr. 1965. Frequency sampling of blue grama range. J. Range Manage. 18:90-93.