PN-AAQ-739
ISN 36064

6-6-1

Analysis of Sampling Strategies for Collecting Genetic Material[1]

THOMAS P. BOGYO,[2] ENRICO PORCEDDU[3] AND PIETRO PERRINO[3]

A study was made using the data obtained during a 1973 expedition to collect wheat in Sicily in order to determine possible sampling strategies for collecting genetic resource material. Based on these analyses tables have been developed for selecting optimum sample sizes depending on relative cost and variance component considerations.

It is almost impossible to make a general recommendation for optimum sampling strategy. Variation may be (and probably is) different in diverse areas, thus what is true in Sicily may not be true for Ethiopia, etc. Also sampling strategy should be different for characters which are inherited quantitatively from those which are controlled by one or two genes. Visual selection will reduce quantitative genetic variation and thus will defeat the primary purpose of collecting genetic resource material, which is to preserve as much genetic variation as possible. On the other hand, all collectors intuitively collect forms which seem new or unusual to them. Certainly when collecting for specific characters, such as disease resistance, dwarf growth type, stem strength, etc., collectors will not resort to random sampling. Many of the traits are rare and only large samples will ensure that genes responsible for these traits will be represented in the collection.

Bennett (1970) pointed out that there is an upper limit to the number of samples an exploring team may collect during any one season because of the time factor and stressed the need for careful exploration planning. Field sampling procedures are aimed at recovering the greatest amount of the genetic variation of species, irrespective of the relative frequency or rarity of any genes or linked genetic combinations. Allard (1970) states that the problem of how to sample is to include the maximum amount of genetically useful variability and still keep the number of items in the samples within practical limits. Sampling methods necessarily depend on the patterns in which genetic variability is distributed over the geographical range of the species, and on the extent of genetic variability within the local population. Marshall and Brown (1975) made certain recommendations regarding sample size and numbers of sites. These recommendations are based on certain theoretical distributions of alleles in natural populations. Specifically they suggested that the optimal strategy is to collect 50–100 individuals per "site" (i.e., "an area from which one bulked sample is to be drawn related to one set of recording kept in a single collecting instance"), to sample as many sites as possible within the time available and to ensure that sampling sites represent as broad a range of environments as possible. They argued that even in the cases of quantitative characters the average number of alleles per locus provides the simplest and least ambiguous measure of genetic diversity for the purpose of

[1] Submitted for publication October 25, 1979; accepted December 11, 1979

[2] Genetics Program and Statistical Services, Washington State University, Pullman, Washington.

[3] Consiglio Nazionale delle Ricerche, Laboratorio del Germoplasma, Via G. Amendola 165-A, Bari, Italy.

Scientific paper no. 5482, College of Agriculture Research Center, Washington State University.

Economic Botany, 34(2), 1980, pp. 160–174
© 1980, by the New York Botanical Garden, Bronx, NY 10458

1

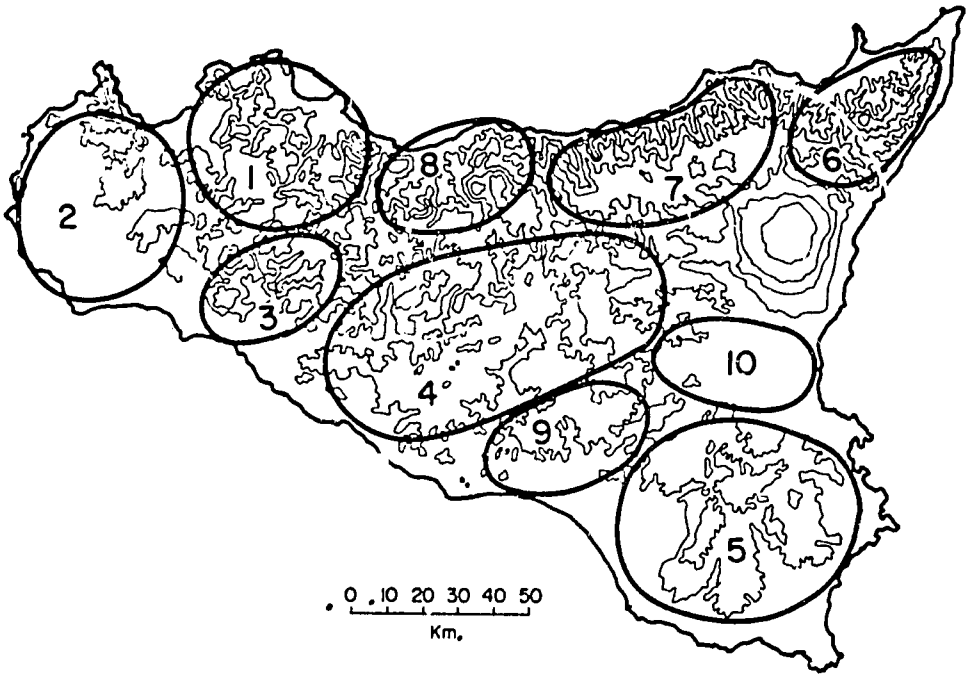

Fig. 1. Map of Sicily, Italy, showing the 10 areas of collection.

exploration and conservation. While this may be so, the explorers in the field need guidance regarding the sampling of phenotypes rather than that of genes or gene-complexes which cannot be determined without extensive studies. The statement of Bennett (1970) that the explorer is essentially interested in preserving at least one of each of the different alleles in the target species is true in the theoretical sense but since genes are invisible entities, any sampling strategy based on calculation of gene frequencies is tantamount to advising explorers to resort to random sampling instead of conscientious selection for visible phenotypical traits. The present study is based on phenotypic traits of durum wheat material collected by teams from the Germplasm Laboratory of the Italian Na-

TABLE 1. MAXIMUM, MINIMUM AND RANGE OF ALTITUDES OF THE SAMPLE COLLECTION SITES (m).

Area	Min	Max	Range
1	0	540	540
2	0	415	415
3	0	680	680
4	120	830	710
5	100	500	400
6	330	725	395
7	465	940	475
8	80	940	860
9	30	540	510
10	0	160	160

Best Available Document

TABLE 2. DETAILS OF SPIKE CHARACTERS STUDIED IN WHEAT (SICILY, 1973).

	Character	Type of specification		
		Cont. measur.	No. of classes	Count
Spike	Length with awns	cm		
	Length without awns	cm		
	Shape		4	
	Cross-section		3	
	Density		3	
	Awning		3	
	Number			
	Sterile spikelet			
	Apical			
	Adherence of flowering glume		4	
Awn	Color of base		5	
	Color of apex		5	
	Angle		3	
	Form		3	
	Persistence		4	
	Length		3	
	Barbing (apical)		2	
	Barbing (base)		2	
Glume	Color		6	
	Color		6	
	Color		6	
	Length	mm		
	Width	mm		
	Keel		2	
	Keel shape		3	
	Keel pubescence		3	
	First beak length	mm		
	Second beak length	mm		
	Beak shape			
	Shoulder length		3	
	Shoulder shape		6	
	Pubescence		4	
Rachis	Node distance		3	
Caryopsis	Color		6	
	Transparency		3	
	Form		3	
	Cross section		3	
	Length	mm		
	Dorsal profile		3	
	Furrow width		3	
	Furrow depth		3	
	Edge shape		3	
	Texture		3	
	Brush length		3	
	Brush density		3	
	Brush extension		2	
	Scutellum shape		2	
	Scutellum length			

TABLE 3. FREQUENCY DISTRIBUTION OF THE 6 MOST COMMON SPIKE TYPES IN THE 10 AREAS OF COLLECTION.

Ear type	Area 1	2	3	4	5	6	7	8	9	10	Total
4	387	117	155	522	78	75	187	193	9	10	1,733 (20.8%)
5	46	9	48	196	8	18	131	155	1	6	618 (7.4%)
18	122	117	86	8	87	0	2	2	0	9	433 (5.2%)
31	143	197	121	507	357	101	141	4	129	32	1,732 (30.6%)
40	135	283	159	1,016	356	21	76	140	140	227	2,553 (30.6%)
54	83	97	48	358	203	119	135	31	50	141	1,265 (15.2%)
Total	916	820	617	2,067	1,089	334	672	525	329	425	8,334
Percent	11	9.8	7.4	31.3	13.1	4.0	8.1	6.3	3.9	5.1	100

tional Research Council (C.N.R.), Bari, Italy, during the summer of 1973. The particular interest in this study was to formulate a sampling strategy based on phenotypic variation, to find the optimum allocation of samples per site and spikes per sample which would minimize the cost of collection. This would ensure that for a given expedition the maximum amount of genetically useful variability of wheat would be collected while at the same time keeping the number of samples within practical limits. The analysis is based on data furnished by an expedition to collect wheat genotypes, which was organized by the Germplasm Laboratory of C.N.R. (Bari) and conducted in 1973 by P. Perrino and E. Porceddu in cooperation with the Food and Agriculture Organization of the United Nations.

MATERIAL AND METHODS

For the purposes of collection, Sicily was arbitrarily subdivided into 10 collecting "areas" (Fig. 1). In each area a number of "samples" were taken (these samples correspond to Marshall and Brown's 1975 "sites") and each sample consisted of about 100 spikes. The average number of samples per area was about 12. For each sample the altitude of the site of collection was duly recorded. These altitudes ranged from sea level to slightly below 1,000 m. The minimum, maximum and range of altitudes in each area is listed in Table 1. The spikes were then arbitrarily sorted into 72 types by visual comparison. Exact measurements for each spike type were later made on a number of characters or classified into groups according to some others. The characters under consideration were those of the spikes, spikelets, awns, glumes, rachises and caryopses. They are listed in Table 2.

Of the many characters recorded we utilized the eight with continuous measurements for studying variability and the information on spike-type distribution to study diversity.

TABLE 4. FREQUENCY DISTRIBUTION OF THE 6 MOST COMMON SPIKE TYPES AT 10 DIFFERENT ALTITUDES.

	0	100	200	300	400	500	600	700	800	900	Total	%
	41	31	150	482	272	133	442	42	77	63	1,733	21.13
	157	13	0	39	104	66	60	12	159	8	618	7.53
	13	168	110	30	107	0	3	0	0	2	433	5.28
	176	235	103	277	490	175	161	19	83	13	1,732	21.12
	214	561	74	326	535	405	227	5	62	12	2,421	29.52
	177	178	22	221	251	102	165	31	66	52	1,265	15.42
Total	778	1,186	459	1,375	1,759	881	1,058	109	447	150	8,202	
Percent	9.49	14.46	5.60	16.76	21.45	10.74	12.90	1.33	5.45	1.83		100.00

$\chi^2(45df) = 2{,}739.2 P(\chi^2 > 2{,}739.2) < 0.0001$.

TABLE 5. AVERAGE VALUES BY AREA (SICILY, 1973).

Character	Area 1	2	3	4	5	6	7	8	9	10
Spike length with awns (cm)	19.54	18.83	19.46	18.91	18.56	18.17	19.21	20.68	18.26	18.76
Spike length without awns (cm)	8.46	8.16	8.49	8.37	8.10	9.43	8.59	8.92	7.93	8.81
Spikelets number per spike	21.59	21.42	21.82	21.91	22.00	22.90	22.73	22.34	22.52	21.53
Number of sterile spikelets	1.56	1.57	1.66	1.62	1.86	1.78	1.47	1.28	1.57	1.91
Glume length (mm)	10.10	9.78	9.93	9.75	9.50	8.82	9.49	10.47	9.32	9.19
Glume width (mm)	3.91	4.03	3.99	4.13	3.99	4.06	4.15	4.17	4.01	4.03
Length of first beak	2.10	2.02	1.79	1.98	1.80	1.77	2.09	2.24	1.47	2.59
Length of second beak	3.98	4.12	3.10	4.06	3.49	5.23	4.75	4.08	3.05	4.55

TABLE 6. AVERAGE VALUES BY ALTITUDE.

Character	Altitude (m) 0–99	100–199	200–299	300–399	400–499	500–599	600–699	700–799	800–899	900–1,000
Spike length with awns (cm)	19.41	18.74	18.40	18.64	18.96	19.28	19.26	17.92	20.21	18.78
Spike length without awns (cm)	8.84	8.25	7.78	8.43	8.31	8.68	8.59	8.68	9.04	8.12
Number of spikelets per spike	22.00	21.62	21.61	21.94	21.85	22.07	22.15	24.12	23.00	22.34
Number of sterile spikelets	1.56	1.90	1.82	1.71	1.61	1.53	1.60	1.22	1.03	1.93
Glume length (mm)	9.66	9.45	9.63	9.67	9.69	9.86	9.97	8.85	9.69	9.39
Glume width (mm)	4.10	3.94	3.80	4.01	4.07	4.19	4.04	4.22	4.33	4.00
Length of first beak (mm)	2.01	1.95	1.76	1.94	1.90	2.00	2.08	1.73	2.12	2.07
Length of second beak (mm)	4.07	3.56	3.20	4.41	3.88	3.85	4.21	2.75	4.60	5.24

TABLE 7. ANALYSIS OF VARIANCE OF SPIKE LENGTHS WITH AWNS (SICILY, 1973) (cm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	30.16	3.35	0.176	11.39
Samples within areas	110	150.18	1.37	1.365	88.61

TABLE 8. ANALYSIS OF VARIANCE OF SPIKE LENGTHS WITHOUT AWNS (SICILY, 1973) (cm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	13.85	1.54	0.073	9.47
Samples within areas	110	77.53	0.70	0.705	90.53

TABLE 9. ANALYSIS OF VARIANCE OF GLUME LENGTHS (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	14.27	1.59	0.11	23.41
Samples within areas	110	39.11	0.36	0.36	76.59

TABLE 10. ANALYSIS OF VARIANCE OF GLUME WIDTHS (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	0.75	0.08	0.0004	0.48
Samples within areas	110	8.69	0.08	0.08	99.52

TABLE 11. ANALYSIS OF VARIANCE OF FIRST BEAK LENGTHS (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	4.31	0.48	0.026	12.24
Samples within areas	108	20.25	0.19	0.190	87.76

TABLE 12. ANALYSIS OF VARIANCE OF SECOND BEAK LENGTHS (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	29.76	3.31	0	0
Samples within areas	108	375.66	3.48	3.48	100

TABLE 13. ANALYSIS OF VARIANCE OF NUMBERS OF SPIKELETS (SICILY, 1973).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	22.65	2.52	0.111	8.11
Samples within areas	110	138.53	1.26	1.259	91.89

6

TABLE 14. ANALYSIS OF VARIANCE OF NUMBERS OF STERILE SPIKELETS (SICILY, 1973).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	2.30	0.26	0	0
Samples within areas	110	31.20	0.28	0.28	100

TABLE 15. TWO-WAY NESTED ANALYSIS OF VARIANCE OF SPIKE LENGTHS WITH AWNS (cm) BY AREA.

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	3,354.4	372.7	0.178	5.18
Samples within areas	111	15,526.0	139.9	1.457	42.36
Spikes within samples	11,594	20,930.6	1.8	1.805	52.47

TABLE 16. TWOFOLD NESTED ANALYSIS OF SPIKE LENGTH WITH AWNS BY ALTITUDE (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among altitudes	9	2,527.8	280.9	0.084	2.45
Among samples within altitudes	111	16,352.6	147.3	1.536	44.84
Spikes within samples	11,594	20,930.6	1.8	1.805	52.71

TABLE 17. THREEFOLD NESTED ANALYSIS OF SPIKE LENGTH (WITHOUT AWNS) (mm).

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	1,405.12	156.12	0.033	1.23
Among altitudes within areas	31	3,107.33	100.24	0.121	4.49
Among samples within altitude	80	4,619.30	54.74	0.593	21.98
Spikes within samples	12,039	23,500.00	1.95	1.952	72.30

TABLE 18. THE CHOICE OF N_2 (NO. OF SAMPLES PER AREA).

RATIO (σ_2^2/σ_1^2)	Ratio c_1/c_2							
	25	16	9	4	1	.5	.25	.1
25	25	20	15	10	5	4	3	2
16	20	16	12	8	4	3	2	2
9	15	12	9	6	3	3	2	1
4	10	8	6	4	2	2	1	1
1	5	4	3	2	1	1	1	1
.5	4	3	3	2	1	1	1	1
.25	3	2	2	1	1	1	1	1
.1	2	2	1	1	1	1	1	1
.04	1	1	1	1	1	1	1	1

RESULTS

Frequency tables of spike types by area and spike types by altitude were constructed; however, because of their sizes, the tables are not reproduced here. χ^2 tests of independence were calculated for both tables and they show that spike type is neither independent of area nor of altitude. Spike types 1, 18, 19 and 50 essentially only occur at lower elevations; types 8, 29, 35 and 56 at intermediate; 21, 34, 37, 38 and 39 at higher elevations. The tables also show that out of 72 spike types only 6 occur with a frequency of greater than 0.03 and only 4 with greater than 0.05. The majority (94.4%) of spike types collected would be classified as "rare" and only 5.6% as "common" using the arbitrary, but quite acceptable criterion of rareness established by Marshall and Brown (1975).

Tables 3 and 4 record the frequencies of the "common" spike types, i.e., those which occur 3 or more percent of the total collection, representing 91.7% of all spike types and about ⅔ of all the spikes, collected. These two tables show that even the common spike types are not distributed independently of areas or altitudes. With respect to altitudes there seem to be no trend of any relationship between types and altitude, except perhaps, that the most common spike type (No. 31) seems to be more abundant in the 400–500 m elevation than elsewhere, whilst spike type No. 5 seems to prefer extreme elevations; more than 50% of this spike type coming from either sea level or from 800–900 m elevations. Tables 5 and 6 list average values of the eight continuous variables used to estimate variability in this study. Both tables show that there is very little variation of mean values for areas or for elevations.

To study the components of variation in those quantitave characters which were characterized by continuous metric measurements, eight analyses of variance were performed on the characters: spike length with awns, spike length without awns, glume length, glume width, first and second beak lengths, (Tables 7–12) subdividing the variation into two parts: among areas, and among samples within areas. It is quite clear from these tables that most of the variation is due to the "within areas" component and only a small proportion (0-23%) to the "among areas" one. Similar results were obtained by analysing the characters characterized by counts: numbers of spikelets and numbers of sterile spikelets (Table 13, 14).

In Table 15 the spike length (with awns) data were further reanalyzed subdividing the "within areas" component into two parts: "among samples within areas" and "among ears within samples." This analysis reveals that the variance components of the lowest two hierachies are essentially of the same size, i.e., there is about the same amount of variation among spikes within a sample as there is among samples. The same kind of analysis was then repeated grouping the samples by altitude rather than by area. Table 16 shows identical results when classified by altitude: the amount of variation among samples within altitudes is essentially identical to the amount of variation among spikes within samples.

A threefold nested analysis of spike lengths without awns was then performed subdividing the variation into four components: "among areas," "among altitudes within areas," "among samples within altitudes" and "among spikes within samples" (Table 17). The relative sizes of the variance components in this model seem to increase with the decreasing level of hierarchy: the ratios being 3.6:1 of

9

the highest two levels (altitude to areas), 4.9:1 of the next two (samples to altitudes) and 3.3:1 of the lowest two (spikes to sample).

DISCUSSION

In discussing strategies of sampling procedures one may utilize the well known results of multistage sampling theory (Cochran, 1963) considering both variance components and the costs of the various stages of sampling.

For two stage sampling let n_1 be the number of areas and n_2 the average number of samples per area. Also let c_1 be the cost of adding a new area to the expedition, this cost not being dependent on the number of samples taken in an area. For example, irrespective of how many samples will be collected, c_1 would include the cost of the time spent on traveling to a new area, or, if the new area has to be sampled simultaneously with the old ones, the cost of adding a new collecting team, or another vehicle, etc., but c_1 would be independent of the cost of taking the samples themselves. On the other hand, let c_2 be the average cost of taking one sample of a given number of ears. Then the total cost is

$$T = c_1 n_1 + c_2 n_1 n_2.$$

Also the variance of an area mean is

$$V_{\bar{y}} = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_1 n_2},$$ where $\hat{\sigma}_1^2$ is the estimate of the variance component for "areas" and $\hat{\sigma}_2^2$ the variance component for "samples." One wishes to minimize the product $TV_{\bar{y}}$

$$TV_{\bar{y}} = (c_1 n_1 + c_2 n_1 n_2)\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_1 n_2}\right).$$

Cochran (1963) shows that this quantity will have the smallest value when

$$n_2 = \sqrt{\frac{c_1 \hat{\sigma}_2^2}{c_2 \hat{\sigma}_1^2}}$$ irrespective of n_1

In our analyses the ratio $\sqrt{\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}}$ varies from ∞ to 3.

Taking 9:1 as an average ratio of $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$ for the 8 characters studied

$$n_2 = 3\sqrt{\frac{c_1}{c_2}}$$

Thus, if c_1 (the cost of adding a new area) would be about 16 times as large as c_2 (the cost of taking an extra sample), then the suggested optimum number of samples per area would be 12 (which was the actual number of samples collected per field). Looking at it from another point of view, n_2 would be equal to 1 if c_2 would be 9 times as much as going to a new area, then it would be the optimum strategy to take a single sample in each area to be sampled. Table 18 gives the best choices of n_2 (number of samples per area) given different ratios of c_1/c_2 and $\hat{\sigma}_2^2/\hat{\sigma}_1^2$.

One may consider the question of how many spikes to collect for each sample. For this we have to consider a three stage sampling scheme. For such a sampling scheme the cost equation

$$T = c_1n_1 + c_2n_1n_2 + c_3n_1n_2n_3, \quad \text{where } c_1, c_2, n_1 \text{ and } n_2$$

remains as in the two stage equation, c_3 is the cost of collecting a single spike and n_3 is the number of spikes per sample. The variance of an area mean becomes

$$V_y = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_1n_2} + \frac{\hat{\sigma}_3^2}{n_1n_2n_3} \quad \text{where } \hat{\sigma}_3^2 \text{ is the}$$

"spike" variance component. The product of the two equations can now be solved for optimum n_2 and n_3 by taking the first derivatives of the product of the two equations with respect to n_2 or n_3 and equating these to zero. The solutions are

$$n_2 = \sqrt{\frac{c_1}{c_2}\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}} \quad \text{(as before) and}$$

$$n_3 = \sqrt{\frac{c_2}{c_3}\frac{\hat{\sigma}_3^2}{\hat{\sigma}_2^2}}.$$

It is noteworthy that the form of the equation for n_3 is similar to that of n_2 and Table 18 may be used for calculating n_3 by substituting the digits 2 and 3 for 1 and 2 respectively.

In this study of the variation of spike length (Table 15, 16) $\hat{\sigma}_2^2$ is approximately equal to $\hat{\sigma}_3^2$ and assuming that this is true for all those characters for which analyses have not been undertaken, the number of spikes per sample should be approximately equal to the square root of the ratio of the cost of taking a sample to the cost of collecting a spike. Since this ratio will always be less than 2500, Marshall and Brown's (1975) recommendation of taking 50–100 units per site (=spikes per sample) in our situation is definitely an overestimate.

It is very difficult to estimate the c_1, c_2 and c_3 values for various reasons. If, for instance, two distant areas have to be sampled where the maturities are similar, c_1 would essentially mean the cost of duplicating the exploring teams, equipment, vehicles, etc. On the other hand, one may look at c_1, c_2 and c_3 as simple functions of time or distance, i.e., c_1/c_2 would be the ratio of distances of travel from one area to another to the distance of travel from one field to another in the same area, and c_2/c_3 could be the ratio of time consumed by traveling from one field to another to the time taken to cross the field to gather the spikes. Should c_1, c_2 and c_3 be all the same, then according to Table 15 n_2 should be about 3 and n_3 about 2.

There are, of course, several complicating factors that have to be considered.

1) Both Marshall and Brown's (1975) recommendations and those of this report are based on random sampling. Explorers in the field, even when trying to collect spikes at random, often find themselves collecting "unusual" types rather than collecting spikes by some predetermined system of randomization. Gevers (1976) has demonstrated that selection for agronomic characters in a reciprocal recurrent selection program in maize (*Zea mays*) reduces the variability of the genetic background to such an extent that progress in

improving yield will be inhibited. In his control material where plants were chosen by random numbers, sufficient progress in improving yield was made. It is quite likely that the same principle will apply for collectors of quantitative genetic resources in wheat.

The samples collected in this study are probably not random. It is impossible to assess what effect this would have on the recommendation of optimal sampling strategy. Based on the results of this study and assuming all variance components similar to the ones in Sicily, one would have to recommend (also using common sense) that one should sample as many areas and as many fields as is feasible, possibly reducing the number of spikes per sample from the 100 used in Sicily to a much lower number.

2) Sicily cannot be regarded as a typical area for wheat exploration. Wheat on this island has been grown under traditional agricultural practices for many centuries. Here the germ plasm is mainly constituted from selected varieties. One would expect an entirely different situation, e.g., in Ethiopia where the germ plasm is formed from local varieties, which possess a very high variability and a certain amount of heterozygosity. Porceddu and others (1975) demonstrated that in Ethiopia much more diversity was found along the sides of the main routes than in areas distant from such lines of communication. It is therefore very probable that sampling procedures in the less developed areas will have to be different from those in Sicily.

3) This study is based on eight morphological characters. It is quite possible that an analysis of various physiological characters would yield different results and would result in different recommendations.

The study of diversity

The Shannon-Weaver (1963) index of diversity was calculated for each sample collected. Although this index is widely used, for the understanding of the probabilistic background of this index it may be useful to recall its derivation. Suppose that at a given site N individual ears have been collected belonging to s spike types and that of these n_j belong the the jth type. Furthermore, assume this site to be a random sample from an infinitely large population n_j/N. The probability of selecting a spike type at random and the probability of selecting the n_j such individual spikes is $(n_j)^{n_j}/N$. The probability of selecting the observed numbers of spikes of each of the other types may be similarly estimated. Then the probability of collecting a sample such as the one found in the field (P_a) is the joint probability of finding each type and thus

$$P_a = \frac{(n_1)^{n_1}}{N} \frac{(n_2)^{n_2}}{N} \cdots \frac{(n_s)^{n_s}}{N}. \qquad (1)$$

Taking logarithms of both sides of equation (1) this becomes:

$$\begin{aligned} \text{Log } P_a &= n_1 \log n_1 + n_2 \log n_2 + \ldots n_s \log n_s \\ &\quad - n_1 \log N - n_2 \log N - \ldots n_s \log N \\ &= \sum_{j=1}^{s} n_j \log n_j - \sum_{j=1}^{s} n_j \log N \\ &= \sum_{j=1}^{s} n_j \log n_j - N \log N \qquad (2) \end{aligned}$$

since $$\sum_{j=1}^{s} n_j = N$$

Thus the basic equation of diversity, H, can be expressed as

$$H = -\log P_a = N \log N - \sum_{j=1}^{s} n_j \log n_j \tag{3}$$

which is the original measure of diversity proposed by Shannon in 1948. It has a minimum value of 0 (where $N = n_j$) or when all individuals in the sample belong to a single type. The measure H will be a maximum if each type in the sample will be represented by a single individual, in which case $H = N \log N$. H, however, is dependent on N, the sample size, and thus not suitable for purposes of comparison, unless all samples are of the same size. Thus an "average H" or H_1 can be calculated as

$$H_1 = H/N \quad \text{or}$$

$$H_1 = \log N - \frac{1}{N} - \sum_{j=1}^{s} n_j = - \sum_{j=1}^{s} p_j \log P_j \tag{4}$$

where $p_j = \frac{n_j}{N}$.

Equation (4) is the well-known Shannon-Weaver formula of 1963. It should be noted that

1) H_1 is at maximum when $p_j = \frac{1}{s}$ for all j;
2) Given two completely even sites the one with the larger number of types has the greater value of H_1;
3) H_1 can be split into components that are additive. This may become very convenient when we wish to subdivide the samples;
4) It is immaterial which base is used to calculate the logarithms. Some investigators recommend the use of $\log_2$, in which case, they say, the unit of measurement becomes a "nat." However, since most electronic hand calculators do not have $\log_2$ keys, this may have some practical disadvantages.

A one-way analysis of variance of dispersion indices (H_1) (Table 19) was performed to estimate relative variance components. The analysis reveals that approximately 94% of the variation in dispersion indices occurs within areas and only 6% among areas. However, this information alone is not sufficient to assist in determining a sampling strategy since one may obtain very similar diversity indices on completely different sets of types. The significant χ^2 values of the original frequency tables (not reproduced here) indicate that the information would be inappropriate in this case for recommending sampling procedures. A

TABLE 19. ANALYSIS OF VARIANCE OF DIVERSITY INDICES (SICILY, 1973) BY AREA.

Source	D.F.	S.S.	M.S.	V.C.	%
Among areas	9	3.17	0.352	0.016	5.83
Within areas	111	22.87	0.206	0.206	94.17

18

TABLE 20. ANALYSIS OF VARIANCE OF DIVERSITY INDICES BY ALTITUDE.

Source	D.F.	S.S.	M.S.	V.C.	%
Among altitudes	9	4.32	0.480	0.024	11.08
Among samples within altitudes	111	21.72	0.196	0.196	88.92

similar analysis of the diversity indices (Table 20) by altitude reveals that the variance component for samples is about 8 times as large as the variance component for altitudes.

Rare spike types

More than 94% of the spike types collected are rare by the definition of Marshall and Brown (1975), since 68 out of 72 ear types occurred with a frequency (p) of less than .05. One would perhaps be able to state that for an explorer a "rare" type is more interesting than a "common" one, and that collectors are, correctly or incorrectly, deliberately selecting types which look rare to them. The sample size to be taken will depend on the degree of rarity and, obviously the smaller the true frequency, p, of a spike type, the larger the sample has to be for it to contain at least one of them.

Let us assume that the probability of finding a very rare form of spike is p in a large sample of size n. Then the probability of finding x = 1, 2, 3, . . . such spikes are

$$P(X = x) = \frac{e^{-np}(np)^x}{x!} = p(x) \qquad (5)$$

and the probability of finding one or more such spike is

$$\sum_{x=1}^{\infty} p(x) = 1 - e^{-np}. \qquad (6)$$

We may use this formula to find the number of samples necessary to find at least one spike with a given probability. Since

$$P(x \geq 1) = 1 - e^{-np} \qquad (7)$$

$$1 - P(x \geq 1) = e^{-np} \qquad \text{then} \qquad (8)$$

TABLE 21. SAMPLES SIZES REQUIRED FOR RARE FORMS.

p	P(X ≥ 1) = .99	P(X ≥ 1) = .95
.05	93	60
.04	116	75
.03	154	100
.02	231	150
.01	461	300
.005	922	600
.001	4,606	3,000
.0005	9,211	5,992
.0001	46,052	29,958

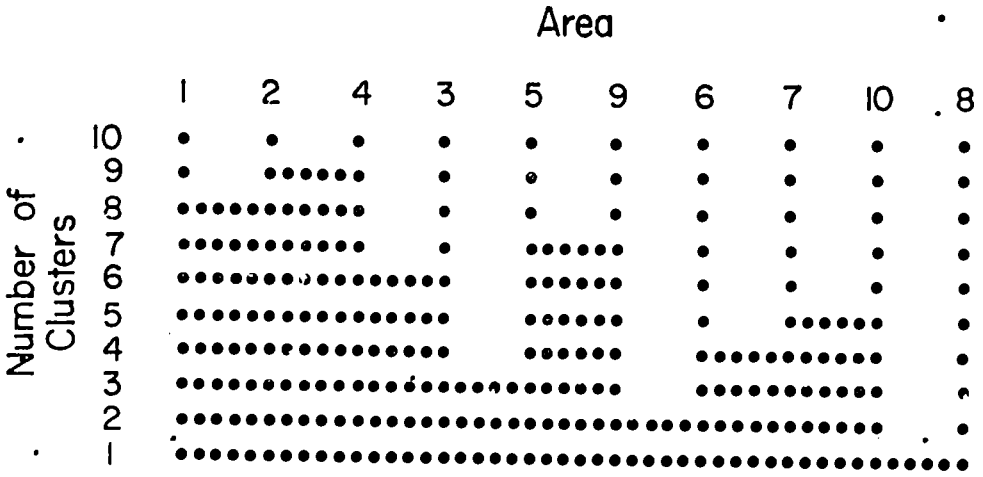

Fig. 2. Results of the cluster analysis of areas based on the 8 continuous spike variables measured in this study.

$$\log_e[1 - P(x \geq 1)] = -np \quad \text{and}$$

$$n = \frac{-\log_e[1 - P(x \geq 1)]}{p} \tag{9}$$

For example, to be 99% confident of finding a rare spike which has a frequency of $P = .005$ (.5%) n (the sample size) has to be at least

$$n = \frac{-\log_e(.01)}{.005} = 922.$$

Table 21 lists the sample sizes required for various values of p and $P(x \geq 1) = .99$ and .95. In the present study the rarest spike types occured with $p = .0002$ which means that to be 99% confident of finding such a spike type and assuming random selection one would need to have a sample size of 23,026. By reducing the confidence level from 99% to 95% the needed sample size would reduce to 14,979.

Clustering of areas

A hierarchal cluster analysis of the data by areas resulted in a clustering as in the diagram of Fig 2. Arbitrarily choosing the number of clusters to be four (based

TABLE 22. MEAN VALUES OF CLUSTERS.

Cluster	Spike length with awns (cm)	Spike length without awns (cm)	Glume length (mm)	Glume width (mm)	First beak length (mm)	Second beak length (mm)	No. of spikelets	No. of sterile spikelets
1	19.19	8.37	9.89	4.01	3.81	1.97	21.68	1.60
2	18.41	8.01	9.41	4.00	3.27	1.63	22.26	1.72
3	18.71	8.94	9.17	4.08	4.84	2.15	22.38	1.72
4	20.68	8.92	10.47	4.17	4.08	2.24	22.34	1.28

14

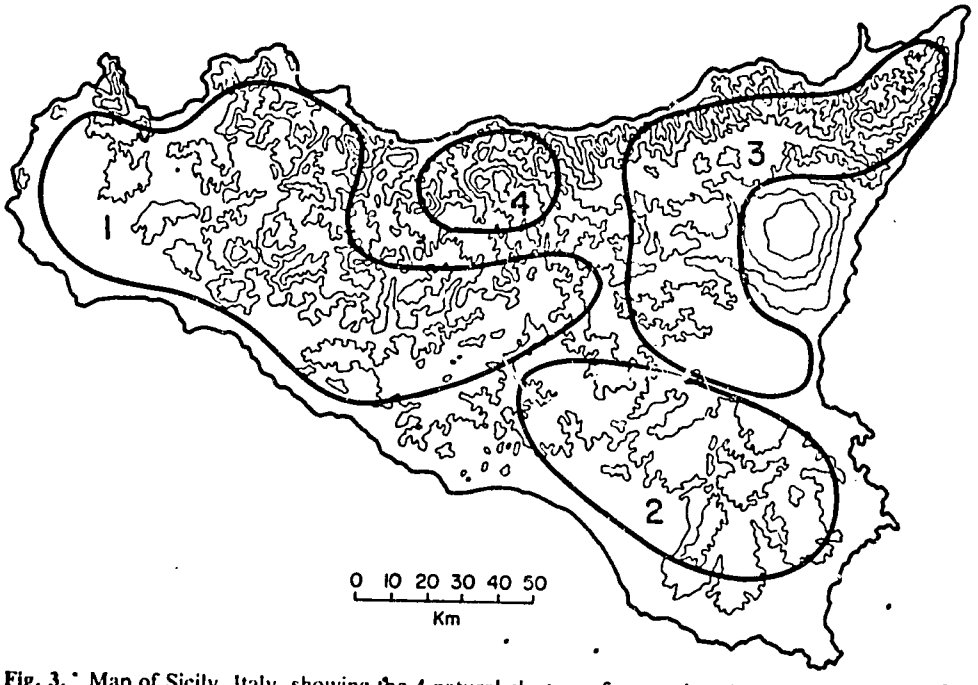

Fig. 3. Map of Sicily, Italy, showing the 4 natural clusters of areas given by the cluster analysis.

on all eight continuous variables studied), four natural clusters of areas are formed: No. 1 being the western lowlands, No. 2 the southeastern lowlands, No. 3 the highland area around Mount Etna, and No. 4 the northcentral highlands (Fig. 3). The mean values for these four areas for all eight variables are given in Table 22.

LITERATURE CITED

Allard, R. W. 1970. Population structure and sampling methods. *In* Genetic Resources in Plants. Their exploration and conservation. O. H. Frankel and E. Bennett, eds. IBP Handbook No. 11, pp. 97–107. Blackwell, Oxford and Edinburgh.

Bennett, E. 1970. Tactics of plant exploration. *In* Genetic Resources in Plants. Their exploration and conservation. O. H. Frankel and E. Bennett, eds. IBP Handbook No. 11, pp. 157–179. Blackwell, Oxford and Edinburgh.

Cochran, W. G. 1963. Sampling Techniques. 2nd edit. John Wiley, New York.

Gevers, H. O. 1976. Unpublished Ph.D. dissertation. Univ. Natal, Pietermaritzburg, Natal, Republic of South Africa.

Marshall, D. R., and H. D. Brown. 1975. Optimum sampling strategies in genetic conservation. *In* Plant Genetic Resources for Today and Tomorrow. O. H. Frankel and J. G. Hawkes, eds. IBP Vol. 2, pp. 53–80. Cambridge Univ. Press.

Porceddu, E., P. Perrino, and G. Olita. 1975. Preliminary information on an Ethiopian wheat germplasm collecting mission. Proc. Symp. Genetics and Breeding of Durum Wheat, Bari, Italy, 14–18 May, 1973. pp. 181–200.

Shannon, C. E., and W. Weaver. 1963. The Mathematical Theory of Communication. Univ. Illinois Press, Urbana.

15