Strategies for Developing Core Collections of Safflower (*Carthamus tinctorius* L.) Germplasm – Part III. Obtaining Diversity Groups Based on an Information Measure

R Balakrishnan and KK Suresh
Department of Statistics, Bharathiar University, Coimbatore-641046 (Tamil Nadu)

Evaluation and passport data collected from a germplasm collection of 3250 safflower (*Carthamus tinctorius* L.) accessions were used to obtain diversity groups based on three different methods. In the first method, 30 diversity groups were obtained by using multivariate cluster analysis followed by further classification into geographical origin and plant types as proposed by Suresh and Balakrishnan (2001). In the second method 13 diversity groups were obtained based on the geographical origin of the accessions. In the third method, the evaluation data were used to compute an information measure (designated as the Length of Encoded Attribute Values or Length of Encoded Attribute Values (LEAV) of the accessions) and this measure was used to divide the whole collection into fourteen diversity groups. Estimates of phenotypic diversity of core samples of size ranging from 5% to 20% of the whole collection were obtained through stratified random sampling from each set of diversity groups obtained by these three methods. The sampling variance of the pooled Shannon Diversity Index (SDI) of 28 descriptors was compared among the three methods of grouping of the accessions. It was found that grouping of the accessions based on the proposed information measure (LEAV) resulted in the least sampling variance. The LEAV index was also used to obtain core samples of required size by ranking the accessions as per the magnitude of this index. It compared better than the core samples obtained through the Principal Component Score method proposed by Noirot et al (1996) in terms of diversity of several qualitative descriptors.

Key words: Core Sample, Information Measure, Safflower, Shannon Diversity Index, Stratified Sampling

One of the important aspects of obtaining a core sample from a large germplasm collection concerns with the sampling strategies that could help in selecting accessions that reproduce the variation for several characters in the whole collection to the maximum possible extent. The sampling strategies are mainly concerned with grouping of accessions into homogenous groups or clusters and selecting sub samples from each group to obtain a pooled core sample. The grouping approaches described could be hierarchical (Hintum, 1995; Peeters and Martenelli, 1989) or non-hierarchical cluster analysis methods using quantitative or a mixture of both quantitative and qualitative descriptors (Spagnoletti Zeuli and Qualset, 1993; Mahajan *et al.* 1996; Harch *et al.* 1996; Bisht *et al.* 1998). Grouping of the accessions based on their geographical origin had also been suggested by several of the authors. The most common method is stratified random sampling to obtain core sample of desired size. Several strategies had also been suggested for deciding appropriate sampling fraction from each group or strata. These methods included proportional allocation, log frequency allocation, square root frequency proportion allocation etc. (Brown 1989, Spagnoletti Zeuli & Qualset, 1993; Mahajan *et al.* 1999; Balakrishnan and Suresh, 2000). Noirot *et al.* (1996) suggested that the accessions could be ranked on the basis of their relative contribution to the overall variance and a desired proportion of top ranked accessions could be selected from each group to constitute the core sample.

In Part I of the present investigation (Suresh and Balakrishnan, 2001), the mean and variance of the pooled Shannon Diversity Index (SDI) were compared by drawing core samples from 30 diversity groups obtained from a large germplasm collection of 3250 safflower accessions. In this approach, multivariate cluster analysis was used to group the accessions into 6 major clusters based on 18 morphological and 10 agronomic characters. These major clusters were further divided into 30 diversity groups based on the geographical origin and plant type of the accessions. Simple random sampling and stratified random sampling with 5 methods of group allocation in the core samples were used to compare the diversity of the core sample with that of the whole collection. In the present investigation two more schemes of grouping of the accessions have been considered. They are (i) grouping according to geographical origin of the accessions and (ii) grouping based on an information measure that quantifies how far each accession deviated from the average density of the whole collection with respect to the set of descriptors. The usefulness of this measure designated as the Length of Encoded Attribute Values (LEAV index in short) computed for each accession in the germplasm collection was also studied in obtaining core samples by selecting desired percentages of the top ranked accessions on this index.

Materials and Methods

Details on the source data, the list of descriptors and the number of accessions from different geographical regions have been described in Part I of this investigation by Suresh and Balakrishnan (2001). The procedures for obtaining 30 diversity groups (of 3250 accessions) based on multivariate cluster analysis followed by further subdivision into geographical regions and plant types were also described. In the present investigation the grouping of the accessions based on the geographical origin of the accessions was also considered. There were 13 such groups. One more method, grouping the accessions based on an information measure was also considered and the details are explained in the following sections.

LEAV Index:

For each multi-state or qualitative descriptor '*d*' the probability of occurrence of a descriptor state '*m*' of the attribute in the whole collection was evaluated as:

$$p[m,d] = n[m,d] \ / \ n[d], \qquad ...(1)$$

n[m,d] denoting the number of accessions having attribute state *m* of the descriptor *d*; and n[d], the number of accessions having any known value of the attribute *d*. Based on information theory concepts, the length of the information code that can optimally indicate the possession of descriptor state *m* of attribute *d* is computed as

$$c[m,d] = -\log_e p[m,d] = -\log_e \{n[m,d]/\{n[d]\} \ ...(2)$$

For each continuous attribute d assumed to be normally distributed with mean *m* and standard deviation *s*, the length of the information code that can optimally indicate the possession of a value *x* by the attribute is given by (Wallace and Boulton, 1968):

$$c[d]= g+ (x -\mu)^2/(2\sigma^2) \qquad ...(3)$$

Where a distribution normalizing constant *g* is estimated by

$$g = \log_e (\sigma/(K^* \ \varepsilon)) \qquad ...(4)$$

and $K = 1/\sqrt{2\Pi}$. It is assumed that a measurement *x*[d] of the attribute *d* of an accession s is quoted to a least count of ε, i.e. to an accuracy of $\pm\varepsilon$ and that the probability of getting such a measurement form the distribution (μ, σ) is approximately

$$(K^* \ \varepsilon/\sigma)^* \ \exp \ (-(x-\mu)^2/ \ 2\sigma^2)$$

Each attribute value possessed by an entry in the collection can be regarded as a message about that entry. We consider the length of a message that can convey the description of all *N* accessions as is contained in the N x D attribute values. The optimum method of encoding this message is to use a Shannon-Fano code (Oliver, 1952), where the length of the message required is minus the logarithm of the probability of obtaining the given set of accessions. Assuming that the set of *N* accessions in the whole collection is a random sample from a very large genetic resource, this probability is the product of the probabilities of obtaining each individual accession. Therefore, for each accession *s* in the collection, a message length F[s] that is required to optimally encode all the d attributes of s using the joint density distribution of the whole collection, is then computed using the formula (Wallace and Boulton, 1968):

$$F[s] = \Sigma_{dis} \ c[x[d,s],d] + \Sigma_{Cont} \ \{g[d] + (x[d,s] - \mu[d])^2/2\sigma[d]^2\} \qquad ....(5)$$

where Σ_{dis} means summing over all discrete or qualitative attributes; Σ_{Cont} means summing over all continuous descriptors; x[d,s] indicates any descriptor state *m* of a qualitative attribute *d* possessed by the accession and it indicates the numerical value if *d* is a qualitative attribute. The message length that corresponds to each descriptor (discrete or continuous) can be obtained by substituting appropriate values from (2) and (3) and hence the total message length that corresponds to each of the accessions with a given set of attribute values can be computed from (5). This value is a numerical quantity measured in natural logarithm (or nits in short). When the accessions could be properly grouped based on this value, groups with larger average message length of the accessions would be far removed from the centroid of the whole collection than groups with smaller average message length. In other words, this quantity (LEAV) can be used as an index to quantify the dispersal of each accession from the centroid of the whole collection. For computing the values of LEAV, four descriptors *viz*., days to 50% elongation, days to primary branch initiation, days to first flowering and days to 50% flowering were omitted as they were directly correlated with 'days to physiological maturity'. One more descriptor *viz*., 'plant spread' was also omitted, which was closely associated with 'plant growth habit'. Thus 23 descriptors were used for computing the LEAV index. In the present investigation, as the quantitative descriptors were not normally distributed, they were categorized into class-intervals and these classes were treated as the descriptor states as in the case of a discrete descriptor. Hence the second part on the right-hand side of (5) was not used.

Grouping of the Accessions Based on LEAV

By ordering the computed LEAV for the entries, an optimum strategy for stratification of the accessions was arrived at by dividing the accessions into L strata by finding the stratum boundaries x_1, x_2,.....$x(_{L-1})$ (subject to the condition $x_0 < x_1 < x_2 <$ ..... x_L where x_0 = min (x) and x_L=Max(x)) such that the pooled variance of LEAV evaluated through the stratification was minimized. Since LEAV could be treated as a continuous variable, the stratum boundaries were fixed by using the Dalenius formula (Jarque, 1981):

$$x_{(h)} = \frac{1}{2}\left\{\left[\int_{x(h-1)}^{x(h)} x.f(x)dx \Big/ \int_{x(h-1)}^{x(h)} f(x)dx\right] + \left[\int_{x(h)}^{x(h+1)} x.f(x)dx \Big/ \int_{x(h)}^{x(h+1)} f(x)\ \ dx\right]\right\} \qquad ...6$$

The values of x_h were computed in an iterative way. Since the optimum number of strata (L) was unknown initially, the process was initiated by dividing the whole distribution into 2 groups and the stratum boundary x, was first fixed. Subsequently, these two groups were further sub-divided in an orderly fashion, obtaining the stratum boundaries in each stage. For the computation purpose, the whole set of LEAV values was arranged in the form of a frequency table with a class-interval of 0.5 to get a continuous distribution. A computer program was developed to interactively divide each larger group into sub-groups. At each stage of partition (which is user defined), the ordered stratum boundaries, the variance of each stratum and the pooled variance computed after the most recent partitioning were made available by the program. Using the program, a larger group was sub-divided if the partition resulted in two sub-groups with much smaller variance than the parent group and/ or if there was an appreciable reduction in the overall pooled variance after the partition. Any partition that did not result in appreciable variance reduction was cancelled and an alternative partition was decided. Using this program the optimum number of strata and the stratum boundaries could be fixed very easily. There were two values that were discontinuous towards the tail of the distribution and they were excluded at the data-input stage. These two accessions were allocated to the last stratum after obtaining the groups.

Diversity Index

The diversity index was computed using the Shannon formula and for computing the diversity index for the numerical descriptors they were converted into appropriate class intervals and each class interval was treated as a descriptor state. The population Shannon Diversity Index (SDI) for each descriptor was computed using the formula : $SDI_i = -\Sigma_j P_{ij} * \log_e (P_{ij})$. A pooled diversity index SDI across all the 28 descriptors was then computed by adding the SDI values for the individual descriptors. Similarly in case of computing the pooled diversity index for a core sample of a given size, the same formula was used by replacing the population proportion $P_{i\,j}$ with the sample proportion p_{ij} for a given descriptor state.

Estimation of Mean and Variance of the Pooled SDI through Sampling

For obtaining core samples, 4 different sizes were considered. They were approximately fixed at 5, 10, 15, and 20% of the whole collection and were respectively 150, 330, 480, and 660. For drawing samples from each of the diversity groups, stratified random samples were considered. The group sizes in the core sample were fixed according to five different methods and they were explained in Part I of this investigation (Suresh and Balakrishnan, 2001). To estimate the expected value and sampling variance of the pooled SDI, 100 independent random samples of a given size were drawn without replacement from the given data set. It was also ensured that the number of accessions from each group was fixed as per the method of allocation. The sample pooled SDI was computed in each case and also the mean and variance of the pooled SDI was computed over the 100 samples. This procedure was repeated with regard to each of the three methods of grouping the accessions.

Purposive Selection of Core Samples

Apart from obtaining the core samples using the method of random sampling, purposive selection to obtain a higher diversity level in the core sample was also attempted. For this two procedures were followed. In the first method, the principal component scores of the individual accessions were evaluated based on 10 quantitative descriptors. The accessions' contribution to overall variance or generalized sum of squares (GSS) was evaluated and the top ranking accessions for a desired core sample size were selected (Noirot *et al.* 1996). In the second method, the accessions were ranked in the descending order of their LEAV and the top ranking accessions for a desired core sample size were selected. The SDI values for the individual descriptors in these core samples were evaluated and they were compared using a t-test (Hutcheson, 1970). In addition, a coefficient of similarity proposed by Harch *et al.* (1996) was computed for all possible pairs of accessions in the core samples obtained by these two methods. The relative frequency of pairs of accessions with different degrees of similarity was computed. This frequency pattern was used to assess which of the two methods resulted in core samples with lesser number of probable duplicates.

Results and Discussion

The values of LEAV for the individual accessions in the whole collection were computed. The frequency distribution of LEAV was divided into 78 class intervals with the width of a class interval of 0.5 nits. The stratum boundaries were decided based on Dalenius formula (equ. 6) using the computer program developed for the purpose. This resulted in 14 diversity groups. Also the pooled SDI and the mean LEAV were computed for each group. The relationship between the mean LEAV and pooled SDI is presented graphically in Fig. 1. It is seen from Fig. 1 that the relationship between mean LEAV and the pooled SDI for the groups is nearly linear in case of grouping of the accessions based on the LEAV index. Only in the last few groups, there was same reduction in the pooled SDI. This could be due to reduced group sizes and quite different distribution patterns in relation to the centroid of the whole collection, but resulting in SDI values that were nearer to those of the groups with lesser mean LEAV index. It was quite possible that two different distribution patterns of the descriptor states could still yield SDI values that were nearly equal. In the case of grouping of the accessions according to geographical origin, the relationship between mean LEAV and the pooled SDI for the groups was low and it was intermediate in the case of grouping as per the cluster analysis method.

The results pertaining to stratified random sampling from groups constituted by the three methods are presented in Table 1. It is clearly seen that the sampling variance of the pooled SDI of the core samples obtained from groups formed on the basis of geographical origin of the accessions had larger variance than those obtained from the other two methods for different sample sizes and frequency allocation methods. The stratification of the accessions using LEAV index resulted in only half the number of groups obtained by using the cluster analysis procedure. The results clearly indicated that the

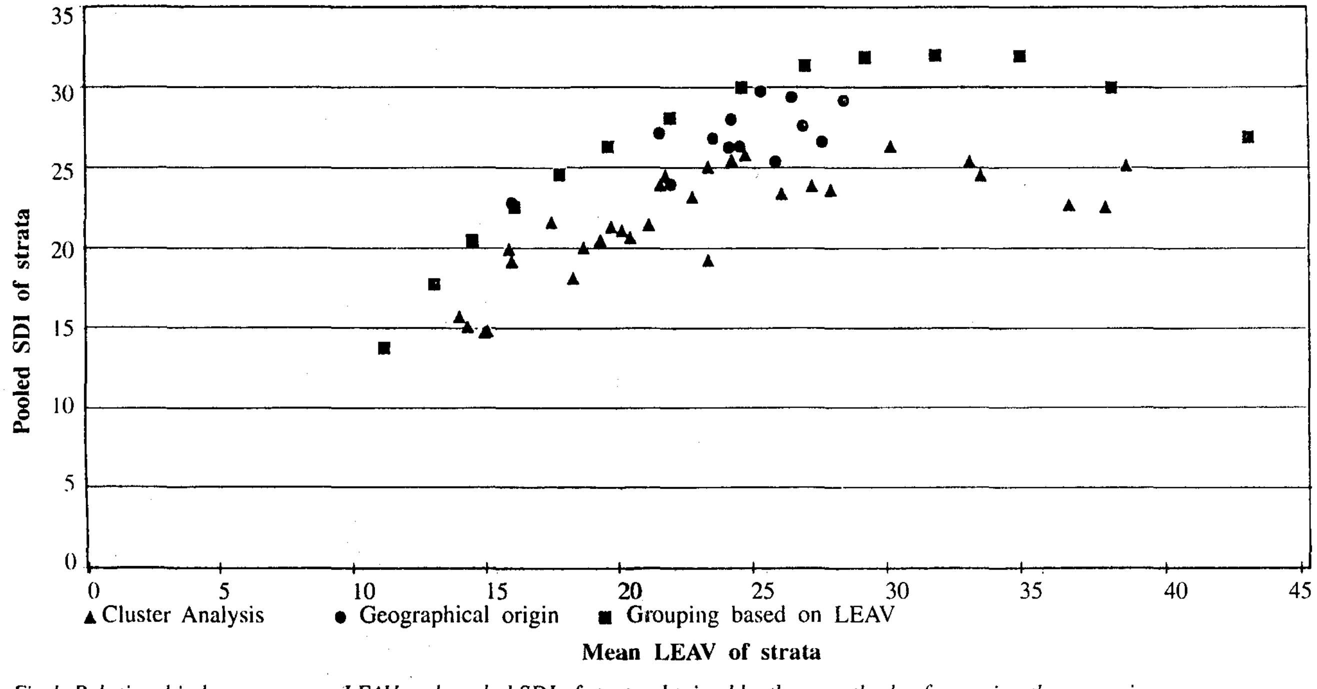

Fig 1. Relationship between mean LEAV and pooled SDI of strata obtained by three methods of grouping the accessions

Table 1. Mean Diversity and its sampling variance for the core samples drawn from the whole collection through stratified random sampling using different stratification procedures

Sl. No.	Sample Size	30 groups based on cluster analysis + geographical origin & Plant type		13 groups based on geographical origin		14 groups based on information measure (LEAV)	
		Mean SDI*	Variance (SDI)	Mean SDI	Variance (SDI)	Mean SDI	Variance (SDI)
1.	Simple random sampling						
	150	25.72	0.3876	Common to all the 3 methods of grouping			
	330	26.00	0.2163	—	—	—	—
	480	26.00	0.1230	—	—	—	—
	660	26.04	0.0900	—	—	—	—
2.	Frequency proportion method						
	150	25.97	0.1336	25.74	0.2899	25.78	0.0610
	330	26.07	0.0509	25.98	0.1152	26.00	0.0123
	480	26.13	0.0376	26.05	0.0865	26.03	0.0098
	660	26.00	0.0289	26.08	0.0499	26.10	0.0063
3.	Square root proportion method						
	150	28.28	0.1068	28.83	0.1694	28.38	0.0283
	330	28.50	0.0581	29.08	0.0676	28.62	0.0155
	480	28.58	0.0228	29.16	0.0313	28.67	0.0053
	660	28.62	0.0211	29.20	0.0258	28.73	0.0060
4.	Log frequency method						
	150	29.22	0.1147	29.40	0.1754	29.91	0.0402
	330	29.57	0.0335	29.68	0.0443	30.22	0.0108
	480	29.60	0.0209	29.75	0.0311	30.19	0.0112
	660	29.60	0.0130	29.62	0.0175	30.28	0.0044
5.	Diversity proportional method						
	150	26.98	0.1218	26.14	0.2777	27.50	0.0354
	330	27.08	0.0420	26.34	0.1563	27.78	0.0160
	480	27.08	0.0393	26.53	0.0807	27.75	0.0100
	660	27.11	0.0237	26.53	0.0709	27.85	0.0083
6.	Equal frequency method						
	150	29.85	0.0885	29.56	0.1350	30.65	0.0367
	330	30.03	0.0260	—	—	30.96	0.0145
	480	30.01	0.0223	—	—	31.07	0.0052
	660	—	—	—	—	31.06	0.0040

* – Pooled Shannon Diversity Index based on 28 descriptors

mean SDI of the core samples drawn from the groups constituted on the basis of LEAV index was comparable from those obtained from the other two methods of grouping. The most striking aspect was that this method yielded core samples whose sampling variance of the pooled SDI was very much smaller than that obtained by the other two methods of grouping.

In Table 2 the diversity measures of the core samples obtained by selecting the top ranked accessions based on the principal component scores evaluated on 10 quantitative descriptors are presented along with those of core samples obtained on the basis of ranked values of LEAV. The principal component scores method was mainly aimed at increasing the diversity of the core samples with respect to quantitative descriptors and the accessions in the core samples accounted for higher percentage of the GSS than those obtained by selecting the top ranked accessions on the basis of LEAV index. However, the core samples obtained by the latter procedure had higher pooled SDI in respect of qualitative and quantitative descriptors combined together. The core samples obtained on the basis of LEAV values had higher pooled SDI with respect to 18 qualitative descriptors and marginally lesser pooled SDI with respect to 10

Table 2. Diversity measures of the core samples obtained by selecting the top ranked accessions based on principal component scores evaluated on 10 quantitative descriptors and the LEAV index

Criterion	% Accessions Selected	%GSS accounted for by the Core Sample	Pooled SDI for 18 qualitative descriptor	Pooled SDI for 10 quantitative descriptors	Total pooled SDI$
Principal	Top 10%	37.70	12.82	18.39	31.21
Component	Top 15%	46.30	12.88	18.23	31.11
Score	Top 20%	53.80	13.04	18.13	31.17
Information	Top 10%	18.15	16.13	16.56	32.69
Measure	Top 15%	26.28	16.44	16.81	33.25
(LEAV)	Top 20%	34.03	16.24	16.89	33.13

$ Pooled SDI for the whole collection of 3250 accessions based on 28 descriptors = 26.14

quantitative descriptors. But both the methods yielded core samples that had much higher pooled SDI than that of the whole collection (= 26.14 nits).

In Table 3 the standardized SDI values for the individual descriptors in the core samples obtained by selecting the top ranked accessions based on principal component scores and the LEAV index are presented for a core sample size of 15% of the whole collection. The results indicated that the core samples by either of the two methods had significantly higher SDI than the whole collection with respect to almost all the descriptors. The SDI values in respect of qualitative descriptors were significantly higher in core samples obtained on the basis of LEAV index, except for 'growth habit' for which the SDI value was significantly lower. The SDI values for Shape of lower stem leaf and attitude of OIB to head were statistically at par in both the core samples. Also, in respect of descriptors that had low diversity in the whole collection, the core samples obtained on the basis of LEAV had substantially higher diversity than the whole collection. However, in the case of quantitative descriptors, except in the case of 'internode length' and 'main capitula diameter', the core samples obtained on the basis of principal component scores had significantly higher SDI. This was also evident from the results presented in Table 2, where these core samples had accounted for much higher GSS%. Same trends were observed for the core samples of 10% and 20% size.

Table 4 presents the relative frequencies of pairs of accessions with different degrees of phenotypic similarity for core samples drawn on the basis of ranked values of principal component scores and LEAV. The frequency patterns indicated that nearly 40% of accession pairs had a high degree of similarity (0.7 and above) in the case of core samples obtained on the basis of

Table 3. Diversity measure for individual descriptors in the core samples obtained by selecting the top ranked accessions based on principal component scores (PCS) and the LEAV index (15% sample size)

Descriptor Name	Whole Collection	Core sample based on PCS	Core sample based on LEAV
Shape of lower stem leaf	0.543	0.497	0.570
Margin of lower stem leaf	0.322	0.399	**0.571
Primary head shape	0.376	0.501	** 0.862
Texture of upper LEAVs	0.624	0.738	** 0.862
Shape of upper stem leaf	0.396	0.621	** 0.724
Margin of upper stem leaf	0.535	0.656	** 0.836
No. of Spines on upper stem leaf	0.520	0.638	**0.830
Attitude of OIB to head	0.731	0.918	0.900
OIB cross section shape	0.599	0.745	**1.000
Location of spines on OIB	0.249	0.325	**0.766
No. of spines on OIB	0.535	0.644	**0.964
Length of spines on OIB	0.607	0.727	**0.905
Bracts enclosing head	0.319	0.452	**0.928
Growth habit	0.893	0.921	*0.868
Branch Location on main stem	0.809	0.888	*0.939
Pollen production	0.822	0.903	**0.968
Pappus on the acheme	0.246	0.263	*0.355
Hull thickness	0.638	0.696	*0.796
Days to 50% elongation	0.682	0.839	**0.657
Days to primary branch initiation	0.704	0.809	**0.731
Days to 1st flowering	0.809	0.894	*0.860
Days to 50% flowering	0.808	0.906	*0.874
Days to physiol. Maturity	0.650	0.709	**0.647
Plant spread	0.705	0.870	**0.817
No. of primary branches	0.723	0.857	**0.817
No. of capitula/plant	0.684	0.882	**0.768
Internode length	0.559	0.730	0.691
Main capitula diameter	0.556	0.709	0.702
No. of accessions	3250	490	490

$: Diversity measure expressed as Shannon Diversity Index in standardized form

*,** SDI significantly different at 5% and 1% levels respectively (t-test)

Table 4. Relative frequencies of pairs of accessions with different degrees of similarity in core samples obtained on the basis of higher values of principal component scores (PCS) and the LEAV.

(Relative frequency as % of total number of all possible pairs of accessions)

Range of similarity coefficient	Top 10% (=325) accessions selected based on		Top 15% (=490) accessions selected based on		Top 20% (=650) accessions selected based on	
	PCS	LEAV	PCS	LEAV	PCS	LEAV
0.0-0.1	0.00	0.00	0.00	0.00	0.00	0.00
0.1-0.2	0.00	0.00	0.00	0.00	0.00	0.00
0.2-0.3	0.00	0.03	0.00	0.03	0.01	0.01
0.3-0.4	0.64	2.17	0.51	1.94	0.73	1.28
0.4-0.5	7.53	17.53	7.31	17.15	7.26	13.97
0.5.-0.6	19.85	35.12	19.52	37.38	19.23	36.27
0.6-0.7	31.78	28.58	30.97	29.97	31.94	33.55
0.7-0.8	25.29	12.87	26.78	11.12	27.17	12.67
0.8-0.9	12.41	3.42	12.73	2.24	11.81	2.10
0.9-1.0	2.50	0.28	2.18	0.17	1.85	0.15

ranked principal component scores. In contrast, about 15% of the pairs of accessions had a high degree of similarity in the case of core samples obtained on the basis of ranked LEAV. This indicated that core samples obtained by the latter method are likely to contain lesser number of probable duplicates than the former method.

Our main objective was to classify the accessions based on the information contained in N x D attribute measurements. If there were not much diversity in the whole collection, then the accessions would be concentrated in a small region around the centroid of the collection. In such a case any random sample of desired size should constitute a good core sample. There would be scope for classification, if the accessions were distributed non-uniformly in groups and sets of accessions concentrated in a small area around the group centroids in the measurement space. In this context, the attributes' values for each accession in the collection may be regarded as a message about that accession. The messages here nominate the positions in the measurement space of the *N* points representing the attribute values of the entries. Shannon (1948) showed that information needed to record a series of such messages would be minimized if the messages were encoded such that the length of each message was proportional to minus the logarithm of the relative frequency of occurrence of the event what it represented. The message length is greatest when all the frequencies are equal. If the expected density of the points in the measurement space is everywhere uniform, the N x D points can not be encoded more briefly than by a simple listing of the attribute values. However, if the observed distribution of the *N* given points is markedly non-uniform, the average density distribution of the *N* points can be used as the basis for encoding the attribute messages. Given that the measurements for each accession is encoded on the basis of the average density distribution, equation (5) can be regarded as minus the logarithm of the probability that any arbitrary member of the collection would be found to have measurement x[d,s], be it a discrete or a continuous attribute (Wallace and Boulton, 1968). Accessions that have smaller LEAV would be concentrated near the centroid of the collection, whereas those with larger values would be concentrated farther away from the centroid. That is, the accessions that have relatively same values of LEAV are expected to be concentrated in a group or cluster. Hence, instead of the optimally encoded values for each accession about which we are not directly concerned, we could as well make use of the empirical value given by (5) to quantify each accession's dispersal from the centroid of the whole collection as an index for classification. In the present investigation the LEAV index has been viewed as one of the possible criteria for classification of the accessions apart from other criteria cited earlier.

One additional advantage with the proposed information measure as a criterion for classification is that it is computationally simpler than principal component analysis and it reduces to a single value combining several attributes that are both discrete and continuous. The group average for LEAV has a very high correlation with the group diversity when the accessions are stratified on this criterion. Fixing the stratum boundaries on the basis of LEAV values using Dalenius method is much simpler compared to other multivariate techniques where mainly attributes that are continuous in nature are appropriate. Stratified random sampling of the accessions from strata constituted on the basis of LEAV resulted in pooled SDI with much smaller sampling variance. Purposive sampling of the accessions on the basis of ranked values of LEAV resulted in core samples having significantly higher levels of diversity for many qualitative attributes as compared to core samples selected on the basis of ranked principal component scores. This was due to the fact that LEAV index combined attribute values which were either discrete or continuous, whereas the principal component score was mainly concerned with quantitative descriptors. Again the core samples obtained on the basis of ranked values of LEAV had far less chances of containing duplicate accessions as seen from Table 4.

Conclusion

In the present investigation a new method has been proposed to stratify the accessions in a large germplasm collection based on an information measure that can be expressed as a single valued parameter. An optimum stratification strategy has been suggested to group the accessions based on this information measure. The grouping of the accessions based on this information measure resulted in core samples that had far less sampling variance for the pooled SDI as compared to groupings based on the other two methods. The proposed information measure could also be used as a criterion for purposive selection of core samples by ranking the accessions on the empirical value of LEAV and selecting a core sample of desired size.

Acknowledgements

The authors are highly grateful to the Project Director, Directorate of Oilseeds Research, Hyderabad for permission to use the data on safflower accessions published by the Institute.

References

Balakrishnan R and KK Suresh (2000) Some strategies for obtaining core samples from germplasm collections using strata of geographical origins - a case study in safflower (*Carthamus tinctorius* L.). *Statistics and Applications* **2**: 49-64.

Bisht IS, RK Mahajan, TR Loknathan and RC Agrawal (1998) Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. *Genet. Res. Crop Ev.* **45**: 325-335.

Brown AHD (1989) The case of core collections. In: AHD Brown, OH Frankel, DR Marshall and JT Williams (eds) *The Use of Plant Genetic Resources*. Cambridge University Press, Cambridge. pp 136-156.

Harch BD, KE Basford, IH DeLacy, PI Lawrence and A Cruickshank (1996) Mixed data types and the use of pattern analysis on the Australian groundnut germplasm data. *Genet. Res. Crop Ev.* **43**: 363-376.

Hintum TJL van (1995) Hierarchical approaches to the analysis of genetic diversity in crop plants. *In:* T Hodgkin, AHD Brown, TJL van Hintum and EAV Morales (eds) *Core Collections of Plant Genetic Resources*, John Wiley & Sons. pp 23-34.

Hutcheson (1970) A test for comparing diversities based on the Shannon formula. *J. Theor. Biology*. **29**: 151-154.

Jarque CM (1981) A solution to the problem of optimum stratification in multivariate sampling. *Appl. Statistics*. **30**: 163-169.

Mahajan RK, IS Bisht, RC Agrawal and RS Rana (1996) Studies on South Asian okra collection: Methodology for establishing a representative core set using characterization data. *Genetic Resources and Crop Evolution*. **43**: 249-255.

Mahajan RK, IS Bisht and PL Gautam (1999) Sampling strategies for developing Indian sesame core collection. *Indian J. Pl. Genet. Resources*. **12**: 1-9.

Noirot M, S Hamon and F Anthony (1996) The principal component scoring: a new method of constituting a core collection using quantitative data. *Genet. Res. Crop Ev.* **43**: 1-6.

Oliver BM (1952) Efficient coding. *Bell System Tech. J.* **31**: 724-750.

Peeters JP and JA Martinelli (1989) Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theor. Appl. Genet.* **78**: 42-48.

Shannon CE (1948) A mathematical théory of communication. *Bell System Tech. J.* **27**: 379 p. and 623 p.

Spagnoletti Zeuli PL and CO Qualset (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theor. Appl. Genet.* **87**: 295-304.

Suresh KK and R Balakrishnan (2001) Strategies for developing core collections of safflower (*Carthamus tinctorius* L.) germplasm - Part I. Sampling from diversity groups of quantitative-morphological descriptors. *Ind. J. Pl. Genet. Resources*. **14:** 22-31.

Wallace CS and DM Boulton (1968) An information measure for classification. *Computer J.* **11**: 185-194.