

A METHODOLOGY OF GENETIC DIVERGENCE ANALYSIS BASED ON SAMPLE UNIT PROJECTION ON TWO-DIMENSIONAL SPACE

Cosme Damião Cruz and José Marcelo Soriano Viana

ABSTRACT

A methodology for the assessment of the dissimilarity among sample units (varieties, lines, etc.) in dispersion graph was presented. The coordinate of each unit was calculated from a dissimilarity matrix, adopting the criteria of minimizing the differences between the original distances calculated in the p -dimensional space ($p > 2$) and the two dimensional space distances. A comparison was made between the proposed methodology, canonic variables and principal components techniques. It was slightly superior in expressing genetic dissimilarity among cultivars.

INTRODUCTION

We present a new methodology for grouping sample units, suitable for genetic dissimilarity studies based on dispersion graph analysis. The coordinate of each unit is estimated from a dissimilarity (or similarity) matrix by statistical procedures based on the minimization of the differences between the original distances and those obtained in the two dimensional space.

METHODOLOGY

The technique consists in evaluating the dissimilarity between the sample units, based on their relative distances on dispersion graph in a two-dimensional space. The coordinate of each unit is obtained from the dissimilarity matrix calculated from the original data. The distances calculated from these coordinates will be identical to the original ones if there are only two variables or they will show minimal distortion in relation to the original estimates when the number of measured variables is greater than two.

Setting up a decreasing order of diversity among the sample units

Calculate the coordinates for the most divergent sample units and then for those that show, in decreasing order, greater dissimilarity with those already considered. As i and j are the two most dissimilar sample units among n studied units, the next most divergent will be that with the greatest value of $d_{(ij)k}$, given by:

$$d_{(ij)k} = d_{ik} + d_{jk}$$

where d_{ij} is the measure of dissimilarity between the units i and i' and $d_{(ii')i''}$ is the measure of dissimilarity between the group of units i and i' and the unit i'' .

This same criteria is used to set up the order of the other sample units. Therefore, the fourth unit in a decreasing order of divergence, will be that with the greatest value of $d_{(ijk)l}$ given by:

$$d_{(ijk)l} = d_{il} + d_{jl} + d_{kl}$$

Calculating the coordinates for projection in two dimensional space

The coordinates of the first two sample units are established arbitrarily. The coordinate of the third is obtained through mathematical equations and the coordinates of the others are calculated by a process of

distortion minimization between the real distances and those of the dispersion graph in two dimensional space.

Let i, j and k be the first three sample units of the ordered set according to the criteria presented. The coordinates of the first two (i and j) will be (0, 0) and (d_{ij}, 0), respectively, that is, X_i = 0, Y_i = 0, X_j = d_{ij} and Y_j = 0. If between the distances d_{ij}, d_{ik} and d_{jk} a geometrical figure of a triangle can be established the coordinate of the third sample unit can be obtained:

$$X_k = \frac{d_{jk}^2 - d_{ik}^2 - d_{ij}^2}{-2d_{ij}}$$

$$Y_k = (d_{ik}^2 - X_k^2)^{1/2}$$

The coordinates should be obtained from the previously transformed dissimilarity matrix when no triangular inequality is found, that is, when one distance value is not smaller than the sum of the other two. The square root transformation is recommended for Mahalanobis' generalized distance.

The coordinate of unit ℓ (fourth or fifth, etc.) is estimated from:

$$C = P^{-1} Q$$

where:

$$C' = [K \ X_{\ell} \ Y_{\ell}];$$

K is constant;

X_ℓ is the abscissa of sample unit ℓ;

Y_ℓ is the ordinate of sample unit ℓ;

$$P = \begin{bmatrix} M & -2 \sum_m X_m & -2 \sum_m Y_m \\ -2 \sum_m X_m & 4 \sum_m X_m^2 & 4 \sum_m X_m Y_m \\ -2 \sum_m Y_m & 4 \sum_m X_m Y_m & 4 \sum_m Y_m^2 \end{bmatrix}$$

M is the number of sample units with calculated coordinates; m = i, j, k, ..., that is, m assumes the values corresponding to the sample units with calculated coordinates;

$$Q = \begin{bmatrix} \sum_m d_{lm}^2 & - & \sum_m (X_m^2 + Y_m^2) \\ -2 \sum_m X_m d_{lm}^2 & + & 2 \sum_m X_m (X_m^2 + Y_m^2) \\ -2 \sum_m Y_m d_{lm}^2 & + & 2 \sum_m Y_m (X_m^2 + Y_m^2) \end{bmatrix}$$

Efficiency analysis

The efficiency of this procedure can be assessed by the following analyses:

a) the correlation between the estimated distances based on the obtained coordinates and the original distances;

b) the degree of distortion of the distances, due to projection in a two dimensional space, given by:

$$1 - (\sum \sum d_{ij}) / (\sum \sum do_{ij})$$

where,

$\sum \sum d_{ij}$ is the distances total on the dispersion graph;

$\sum \sum do_{ij}$ is the original distances total;

c) the stress coefficient proposed by Kruskal (1964), given by:

$$\{[\sum \sum (do_{ij} - d_{ij})^2] / (\sum \sum d^2 o_{ij})\}^{1/2}$$

Application

The coordinates for the dispersion graph of five individuals, whose dissimilarity measurements are shown in Table I, is calculated as an example.

Table I - Measures of dissimilarity among five individuals.

Individual	2	3	4	5
1	3.7	6.0	6.4	5.9
2		5.4	4.1	2.9
3			4.0	6.9
4				3.8

The most divergent individuals are 3 and 5. The divergences of the other individuals in relation to these two are given by:

$$d_{(35)1} = d_{13} + d_{15} = 11.9$$

$$d_{(35)2} = d_{23} + d_{25} = 8.3$$

$$d_{(35)4} = d_{34} + d_{45} = 7.8$$

Thus individual 1 is third in decreasing order of dissimilarity in relation to the first two (3 and 5). Individual 4 is in fourth place, because:

$$d_{(35)2} = 12.0$$

$$d_{(35)4} = 14.2$$

Using the same criteria, the order of the individuals for the coordinates estimation will be 3, 5, 1, 4 and 2 (or 5, 3, 1, 4 and 2). The coordinates for individuals 3 (i) and 5 (j) are (0, 0) and (6.9, 0), respectively. The coordinate for individual 1 (k) is:

$$X_1 = \frac{d_{31}^2 - d_{51}^2 - d_{35}^2}{-2d_{35}} = 3.5362$$

$$Y_1 = (d_{51}^2 - X_1^2)^{1/2} = 4.8471$$

The calculation of the coordinate for the fourth individual (ℓ = 4) is done with the following statistics:

Order	Individual	X _m	Y _m	X _m ²	Y _m ²	X _m Y _m
1	3	0.0000	0.0000	0.0000	0.0000	0.0000
2	5	6.9000	0.0000	47.6100	0.0000	0.0000
3	1	3.5362	4.8471	12.5047	23.4943	17.1403

Order	Individual	X _m ² + Y _m ²	X _m (X _m ² + Y _m ²)	Y _m (X _m ² + Y _m ²)	d _{4m} ²	X _m d _{4m} ²	Y _m d _{4m} ²
1	3	0.00	0.00	0.00	16.00	0.00	0.00
2	5	47.61	328.509	0.00	14.44	99.636	0.00
3	1	36.00	127.3032	174.50	40.96	144.8427	198.5372

Thus

$$P = \begin{bmatrix} 3.0000 & -20.8724 & -9.6942 \\ -20.8724 & 240.4588 & 68.5612 \\ -9.6942 & 68.5612 & 93.9772 \end{bmatrix} \text{ and } Q = \begin{bmatrix} -12.2100 \\ 422.6670 \\ -48.0832 \end{bmatrix}$$

Solving $C = P^{-1}Q$ the values $K = 16$, $X_4 = 3.5630$ and $Y_4 = -1.4606$ are obtained.

The coordinate of the fifth individual ($l = 2$) is similarly obtained and given by $X_2 = 4.9489$ and $Y_2 = 1.8660$.

Figure 1 shows the individual dispersion patterns obtained from the calculated coordinates. Similar individuals can be visually recognized or established by clustering technique(s), as is usually done when combined

procedures of cluster analysis and principal components or canonic variables are used. In this last case, Tocher's grouping analysis, quoted by Rao (1952), is routinely used.

Table II shows the estimates of the Euclidian distances (d_{eij}) between pairs of individuals obtained from the expression:

$$d_{eij} = [(X_i - X_j)^2 + (Y_i - Y_j)^2]^{1/2}$$

Table II - Measurements of dissimilarity among five individuals obtained from the dispersion graph coordinates

Individual	2	3	4	5
1	3.3079	6.0000	6.3078	5.9000
2		5.2854	3.5945	2.6929
3			3.8508	6.9000
4				3.6426

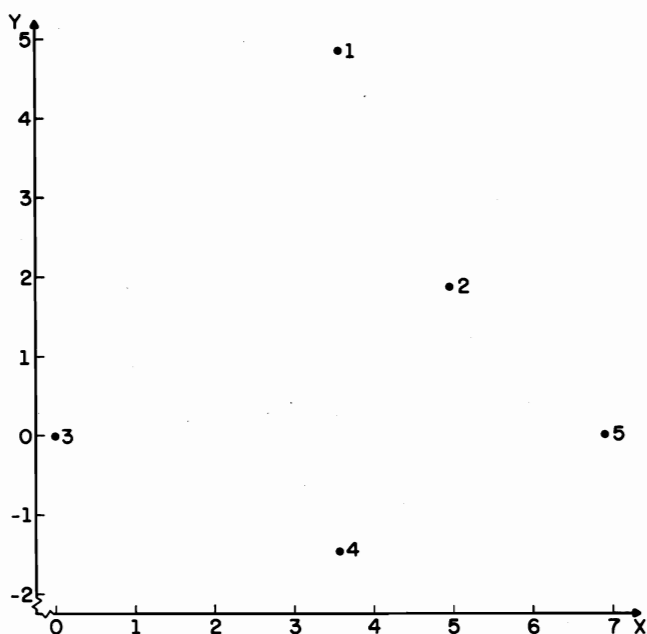


Figure 1 - Dispersion of five individuals based on coordinates estimated from a dissimilarity matrix.

From data in Tables I and II the stress value (4.49%), the correlation between the original and estimated distances (0.9965) and the relation between the original and estimated total distances (0.9670) (3.3% distortion) were obtained.

Two studies of genetic divergence based on canonic variables, principal components techniques and the proposed methodology were made. These analyses were carried out using the GENES I computer program, developed by the Genetics Department of the Federal University of Viçosa.

The first study used data from four characteristics assessed in eight barley cultivars (Singh and Chaudhary, 1979).

Table III presents the Mahalanobis' generalized distances (D^2) and the estimates of the average Euclidian distances obtained from the scores of the first two canonic variables (d_{vc}) and the coordinates obtained by the proposed methodology (d).

Table III - Estimates of Mahalanobis' distance (D^2) and the average Euclidian distances obtained from the scores of the two first canonic variables (d_{vc}) and the coordinates calculated by the proposed methodology (d) among eight barley cultivars.

Cultivars	D^2	d_{vc}	d
1, 2	17.85	2.81	36.96
1, 3	67.38	5.65	75.13
1, 4	24.73	3.40	40.73
1, 5	24.51	3.08	43.80
1, 6	30.91	3.15	38.69
1, 7	15.36	2.67	28.63
1, 8	20.38	2.96	30.35
2, 3	109.08	7.34	109.08
2, 4	51.89	4.73	65.20
2, 5	78.77	5.89	78.77
2, 6	72.85	5.68	72.74
2, 7	49.07	4.73	61.43
2, 8	67.93	5.76	66.23
3, 4	23.56	2.62	49.56
3, 5	60.74	5.23	60.74
3, 6	36.80	3.23	36.59
3, 7	25.84	2.98	47.68
3, 8	63.67	5.39	59.76
4, 5	44.95	4.33	60.63
4, 6	27.58	2.55	23.38
4, 7	4.17	1.28	18.92
4, 8	45.26	4.40	49.33
5, 6	40.09	2.04	39.90
5, 7	28.87	3.05	41.93
5, 8	15.51	0.22	13.82
6, 7	12.02	1.37	11.48
6, 8	12.54	2.18	31.42
7, 8	22.30	3.12	30.41

The use of the first two canonic variables resulted in a distortion of 18.83% (100% minus the percentage of the total variance they express). The distances estimated from the coordinates had a 17.28% distortion. The correlations between D^2 and the statistics d_{vc} and d were, respectively, 0.9289 and 0.9357. These results indicate that the canonic variables analysis and sample unit projection in the two dimensional space technique are equally

efficient in demonstrating the genetic diversity in this example.

A study of the genetic dissimilarity among 33 cotton cultivars was also carried out, using measurements on eight agricultural traits presented by Singh and Gupta (1968). The genetic dissimilarity was assessed by the dispersion graph in relation to the first two principal components (Figure 2) and to the cartesian axis coordinates obtained by the proposed method (Figure 3).

The use of the first two components caused a distortion of 47.66% (100% minus the total variance percentage they express). When the proposed methodology was used the distortion was 40.68%. The results show that the present method was slightly superior

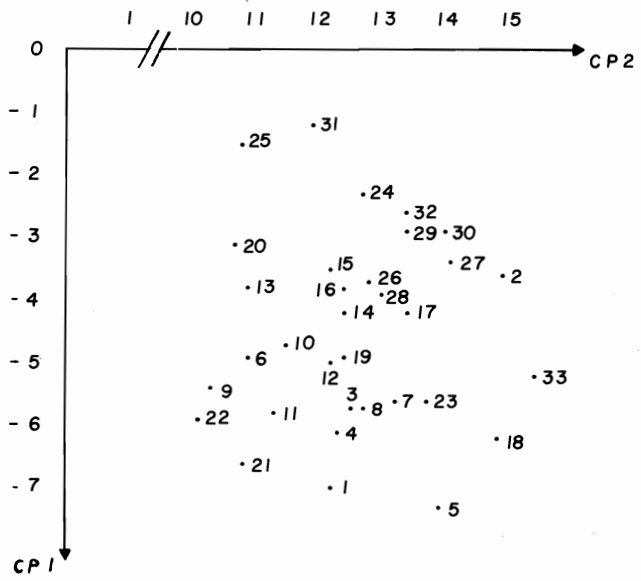


Figure 2 - Dispersion of scores of 33 cotton cultivars in relation to the first two principal components.

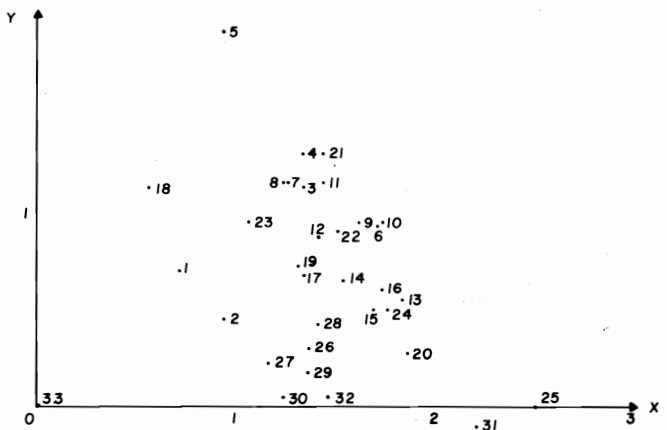


Figure 3 - Dispersion of 33 cotton cultivars in the two dimensional space, using coordinates obtained from a dissimilarity matrix by statistical procedure based on the minimization of the differences between the original distances and those in the graph.

to the principal components method in expressing genetic dissimilarity among cultivars. The discrepancies, and or, agreements between these techniques can be seen in Figures 2 and 3.

DISCUSSION

The proposed technique is suitable for the study of the level of similarity between sample units, independent of their nature and of the distance measurement under consideration. The studied units can be genetic material (varieties, lines, etc.), environments or even a set of traits. The most diverse types of indices, distances, coincidence coefficients, correlations, etc. can be used as measures of dissimilarity.

The graphic presentation for the assessment of the similarity has the following advantages in relation to the conventional clustering techniques: a) simplicity, as graphic interpretations do not offer any difficulty, since they are based on information that can be examined visually; b) consistency, as the cluster pattern is based on the relative position of the units presented together on the graph (on the contrary, conventional conglomerate methodologies use procedures based on various optimization principles which when applied to the same set of data can generate different clusters, giving distinct conclusions); c) the available information is concentrated, since visual assessment of the dissimilarity of, say, fifty genotypes is possible from a dispersion graph, but not from the analysis of a dissimilarity matrix with 1225 distances between pairs of genotypes; d) comprehensive, because although some genetic diversity studies use dispersion graphs based on principal component or canonic variables, others do not allow these types of analysis and

consequently present certain difficulties of analysis and interpretation of the results. The proposed method does not have such limitations, and can be more efficiently used in cases where the principal components and canonic variables techniques are usually applied.

RESUMO

Foi apresentada uma metodologia para avaliação da dissimilaridade entre unidades amostrais (variedades, linhagens etc.) em gráfico de dispersão. A coordenada de cada unidade é estimada a partir de uma matriz de dissimilaridade, adotando-se o critério de minimizar as diferenças entre as distâncias originais e as distâncias no espaço bidimensional.

Foram apresentados resultados comparativos do emprego da metodologia proposta e das técnicas das variáveis canônicas e dos componentes principais. Utilizou-se, nesses estudos, dois conjuntos de dados disponíveis na literatura.

REFERENCES

- Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* 29: 1-127.
- Rao, R.C. (1952). *Advanced Statistical Methods in Biometric Research*. John Wiley and Sons, New York, pp. 390.
- Singh, R.K. and Chaudhary, B.D. (1979). *Biometrical Methods in Quantitative Genetic Analysis*. Kalyani Publishers, Ludhiana, pp. 304.
- Singh, R.B. and Gupta, M.P. (1968). Multivariate analysis of divergence in upland cotton. *Indian J. Genet. Plant Breed.* 28: 151-157.

(Received October 20, 1992)