E

UPOV

TWC/17/12
ORIGINAL: English
DATE: July 29, 1999

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Seventeenth Session
Helsinki, June 29 to July 2, 1999

SPECIAL APPLICATIONS OF DUS VARIETY DESCRIPTIONS

Document prepared by experts from Hungary

Special applications of DUS variety descriptions

1. Introduction

1. The DUS testing carried out in the National Institute for Agricultural Quality Control follows the guidelines laid down by UPOV.

2. One result of these tests is the preparation of variety descriptions.

3. The variety descriptions are n x m type matrices, where the rows are the varieties, the columns are the characteristics. The cellules of the matrices contain the state values of the characteristics, so here there are figures between 1 and 9. It is expressed in another way:

$\mathbf{F} = \mathbf{F}_{\mathbf{n\,x\,m}} = \| c_{ij} \|_{n\,x\,m}$, where

$i = 1,2, ..., n, \quad j = 1,2, ..., m$ and $c_{ij} \in [1,2,\ldots,9]$

4. The characteristics involved in DUS testing can be divided in two groups according to the types of data: measured or visually observed (coded, scored, state) data. Since only the latter type is used in the variety descriptions, measured data must be transformed to a 1 - 9 scale.

5. The characteristics occurring in the variety descriptions are of three types, depending on whether they represent simply a list of unrelated terms (nominal type) or can be considered as distances along the scale: they have but two state numbers (binary type) or more up to 9 state values (ordinal type).

2. The aim of this work

6. The number of varieties involved into the DUS testing is increasing year by year and this requires more work, more money and the modification of testing and evaluations**.** As an example, the reduction or the possibility of the reduction of the varieties involved into the DUS testing is a way**.** If we know the similarity groups where the member varieties in a group are more similar each to other than to the varieties belonging to other groups then there is a way to look after the more representative varieties of the similarity groups**.** Using this method the representative varieties will be the member varieties involved into the DUS testing, and so an important reduction of the number of varieties could be achieved.

7. So, there is a task to construct a similarity function of variety pairs, which is

1. very simple, well understandable;
2. the special points of view of the distinctness of varieties are under considerations in the calculation of similarity of the variety pairs;
3. the different kind of data types (ordinal, binary, nominal) participating in the variety descriptions are handled together in a unique way;
4. this function is properly elastic and so the characteristics involved into the DUS testing are taken under consideration according the weights they have in the DUS evaluations**.**

3. The investigated variety descriptions and the applied programs

8. A program package has been developed for the examination of similarity groups. It is part of the D3e program package of the National Institute for Agricultural Quality Control written in VBA (Visual Basic for Applications) macros and using Excel files. The winter barley example was tested using by this program package.

4. The calculation of the similarities between the varieties

9. The calculation of the similarities between the varieties is made by the following algorithm:

1. Each variety is compared to each other variety.
2. The state values are compared in each characteristics in each variety pair.
3. If the a state values are the same, that is to say their difference is 0, then the characteristic is rewarded by 10 points.
 If the absolute value of this difference > 0 and this value is less than the DUS distinctness threshold value of this characteristic then
 a/ if the difference = 1, then the reward is 5 points,
 b/ if the difference = 2, then the reward is 3 points,
 c/ if the difference = 3, then the reward is 1 points,
4. If difference >= the DUS distinctness threshold value of this characteristic, then the reward is 0.
5. If there are missing data in the comparison, then the reward point is 0.
6. The reward points are summed.
7. The possible maximal value of this sum is the 100 %. It equals to 10 * the number of characteristics involved into the comparison. (We meet the maximal similarity sum if each characteristics gives us maximal rewarded points (10 points)).
8. If the similarity sum of a given variety pair is related to the possible maximal sum (100 %), the similarity % of the variety pair is given.
9. Distance % = 100 % - similarity %.

10. An example: Let us have 29 characteristics examined. Now, the maximal sum of points we can collect, (if each characteristic has the same state vale, note in both varieties compared), 29 x 10 = 290 points. It is the 100 %. The similarity sums received after having compared the variety pairs have to be related to the maximal number (for example, if 261 points are collected, then the similarity % of the variety pair is 90 %., and so 100 % - 90 % = 10 % the distance % of the variety pair.

11. More generally speaking: For each (r, q) variety pair a similarity sum is calculated:

$u_{rq} = \sum_{k=1}^{m} g_k(|c_{rk} - c_{qk}|)$, where the g_k function is the same (in our case now) for each characteristic k, so instead of g_k can be written g function.

12. The general formula can be modified if we take under consideration the distinctness thresholds ($\mathbf{k}_k$) and weights ($\mathbf{b}_k$) of the characteristics:

$$u_{rq} = \sum_{k=1}^{m} g\left(|c_{rk} - c_{qk}|, \mathbf{k}_k, \mathbf{b}_k\right).$$

The maximal similarity sum is:

$u_{max} = u_{r r}$ for each row .

The similarity % is for each (r, q) variety pair:

$h_{r q} = u_{r q} / u_{max}$

$h_{r q}\% = u_{r q} / u_{max} * 100$

The value of $h_{r q}$ is between 0 and 1. More two varieties are similar, more is the value of the similarity %.

On the base of similarity % the distance % can be defined: for each (r, q) variety pair:

$t_{r q} = 1 - h_{r q}$

$t_{r q}\% = t_{r q} * 100$

The value of $t_{r q}$ is between 0 and 1, too. More two varieties are similar, less is the value of the distance %.

5. The number of variety pairs, the histogram of distance values of variety pairs

13. Let us have n pieces of varieties in a comparison procedure and each variety is compared to each other, excluding itself. The number of variety pairs is so $A = n \times (n-1) / 2$. The number „A” is the maximal number of variety pairs.

14. The distance (similarity) values of variety pairs have to be compared and after it the distance values are sorted. The total 0% and 100 % interval is divided equally into subintervals. Let us see the number of variety pairs belonging to the different subintervals (see Figure 2).

6. The distance threshold of variety pairs and the real distance threshold %: $L_r\%$, similarity groups, the representative variety of a variety group

15. If a given value % is taken, the distance threshold (L%), then the variety pairs which have less distance value than the given distance threshold, can be selected and so the varieties belonging to these groups, as well. The similarity groups of varieties for the given distance threshold are built up with these varieties.

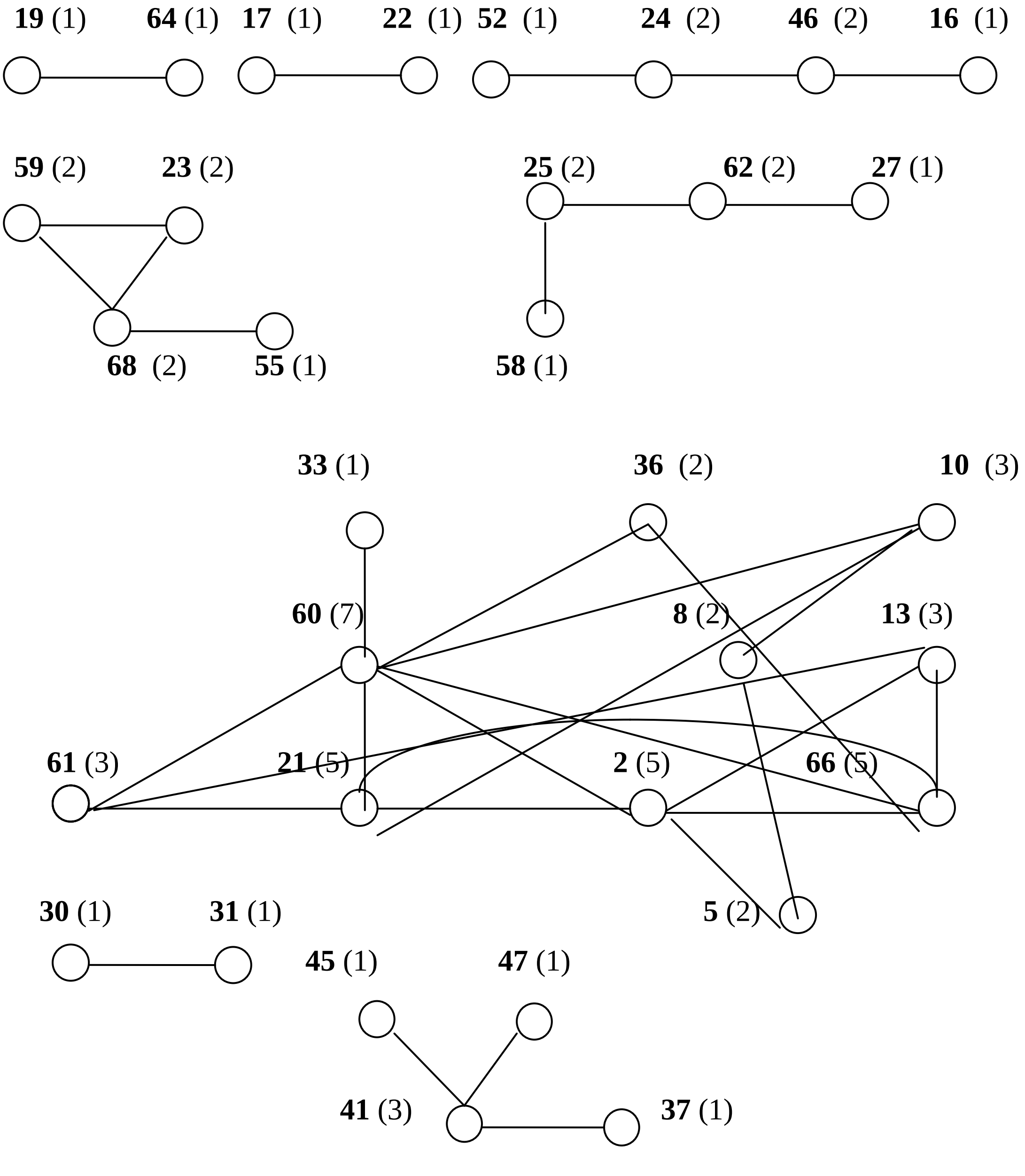
16. The number of groups and also the number of varieties belonging to a similarity group depends on the distance threshold. Less is the distance threshold, less are the numbers of the varieties and groups with similarity connections.

17. So, the determination of the real similarity groups is very important and because it depends on the real distance threshold ($L_r\%$), the correct determination of the $L_r\%$ value is a crucial point. Let us mark with B the variety pairs whose distance % values are <= a given L%. Let us C where > than L%. And so let us mark with B_r the B, if the given L% equals the $L_r\%$ itself. It means that the members of B_r are the varieties constructing the real similarity groups. The explanation of C_r follows the before writing.
So $A = B + C = B_r + C_r$, and $B = G_1 + G_2 + ...$ are the similarity groups at a given distance threshold and a $B_r = G_{1r} + G_{2r} + ...$ are the real similarity groups using the $L_r\%$ value.

18. A similarity group is represented the best by the variety which has the most similarity connections in the given similarity group. It is why the number of connections in a similarity groups are shown in brackets.

19. The similarity groups of winter barley variety description are demonstrated:

Figure 1: The connection net of varieties (winter barley, 1996-1998, DUS data, 29 characteristics, the distance threshold % : 25 %. Numbers printed in bold: identifiers of varieties, numbers in brackets: the number of connections of the variety within a similarity group)



20. As an example, let us see the last group Figure 1 (the members are: varieties 45, 47, 41 and 37). Here, the representative variety is variety no. 41, because it has similarity connections with 3 varieties and others have but one each.

7. The determination of the real threshold distance (L_r%) and of the control variety description generated by uniformly distributed random numbers

21. So, the determination of the real threshold distance (L_r%) and the real similarity groups is a very important question.

22. The determination of the desired L_r% is made in the following way. A control variety description is built up with some conditions. It is a matrix with the same number of rows (varieties) and columns (characteristics) as the original variety description has. The numbers of the different state values, notes in a given column (characteristic) is about the same in both variety descriptions that is to say the frequencies of the different state values of the different columns approximately the same in both variety descriptions. The control variety description was made by using the random number generator of the Excel.

23. In both cases frequency histograms are made where distance intervals are demonstrated in axis x and numbers of variety pairs belonging to a given distance interval in axis y. The point where the histogram of the control variety description approaches the axis x is the point of L_r %. According to the explanation given in section 6 B_r signifies the interval between the points 0 % and L_r%. By definition, here there is no variety pair coming from the random number generated variety pairs. So, if there is a variety pair in B_r then these varieties are really similar each to other and not by chance. It can be declared that the probability of finding variety pairs in B_r -ben which are similar each to other by chance is very low. It is the reason that this point L % value is chosen as the real similarity threshold % (L_r%).

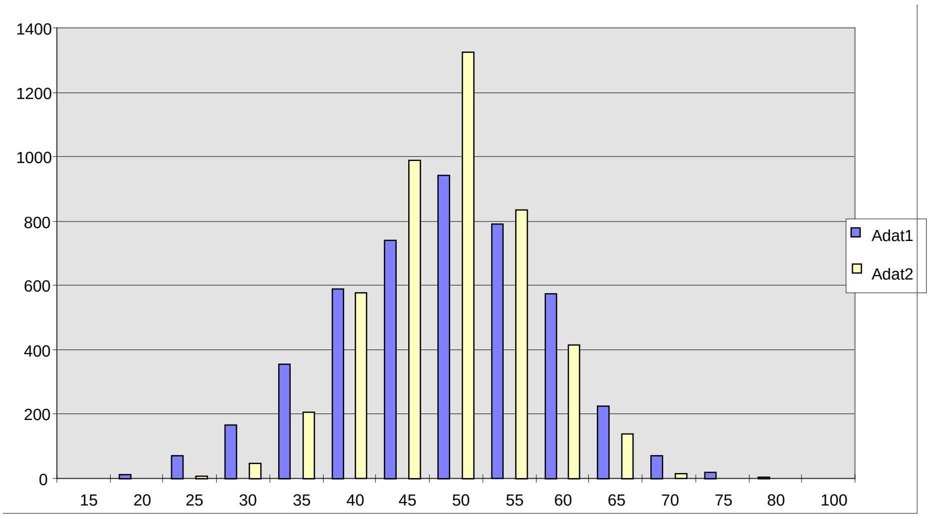
24. As an example of the above written, let us see the variety description of winter barley of years 1996-98 and its control variety description.

Table 1: The frequency table of the variety pairs according the distance value % (on the base of the variety description of winter barley, data from 1996-1998)

Distance value % intervals **(= 100 % - Similarity value % intervals)**	**Number of variety pairs in the given interval**	
	Original variety description list for winter barley	*Variety description list generated using random numbers*
0 - 15	0	0
-20	12	0
-25	72	6
-30	166	48
-35	354	204
-40	590	578
-45	738	988
-50	940	1324
-55	792	836
-60	574	416
-65	224	140
-70	72	16
-75	18	0
-80	2	0
-100	0	0

25. It can be seen from the table that the variety population generated using random numbers had a normal distribution. The distribution of the other population was flatter and contained a larger number of varieties (84) in the lowest (15-25 %) similarity value intervals than in the random number list (6). This range thus indicates the part of the distribution curve which is different due to the variety similarities present in the original variety population compared to the variety independence characteristic of the random number list. The threshold value, 25 %, indicates the level at which it is worth looking for similarity groups (this value was also found to be the most favorable when studying similarity groups in practice).

Figure 2. The histograms of variety pairs according the distance value %(on the base of the variety description of winter barley, data from 1996-1998 and its control variety description)



26. The „adat1" of figure 2 belong to the variety description of winter barley (data from 1996-1998) and the „adat2" belong to the control variety description.

27. Studies were made on the varieties included in the 84 variety pairs found in the 15-25 % interval in the original description list for winter barley. These include only those varieties which make up similarity groups below the 25 % similarity threshold in the earlier original winter barley variety description list.

28. These number 33 out of a total 68 varieties in the original winter barley variety description list. These 33 varieties, obtained in two ways in the similarity groups, represent those which are related to each other. The proportion of these to the total number of varieties (48.5 %) demonstrates the distance of the variety population from a population consisting of completely independent varieties, generated using random numbers, i.e. ratio of relatedness as regards origin.

8. Study on variety population involved into the DUS testing of some filed crops based on similarity groups determined by the real distance threshold %

29. If the real similarity groups can be determined correctly, then the varieties and the number of varieties belonging to the real groups (V_B) can be determined, as well. Let V_C the number of varieties which do not belong to any groups („solitary” varieties). The sum of this two equals to the total number of varieties ($V_B + V_C$). The varieties belonging to V_B have similarity connections with a variety or varieties, but the varieties of V_C have no similarity connections, so that way the ratio V_B / ($V_B + V_C$) gives us information about the ratio of „dependent” varieties of the given crop, maybe with a genetical background.

Table 2: The ratio of varieties belonging a similarity related to the total number of varieties

Crops	*The ratio of the varieties belonging a similarity group*	*The total number of varieties involved into the investigation*
winter barley, 1996-98.	48,5	68
spring barley, 1998.	29,1	24
winter wheat, 1997.	37,0	116

9. Results

30. 1. Similarity ratio of varieties of a given crop: V_B / ($V_B + V_C$);

2. The determination of L_r% by using control variety description;

3. The determination of real similarity groups and their representatives (reduction of the number of varieties involved to DUS comparisons).

10. Simplifications made

31. 1. The way of the determination of L_r%. According the L_r% used now, the variety pairs similar each to other by chance are excluded very strictly from interval B_r (the real similarity groups) and the variety pairs which are really similar each to other can be excluded from interval B_r, as well but with a higher probability.

2. The random number generated variety descriptions were made by neglecting the correlations between characteristics;

3. Calculating the sum of the similarity of the variety pairs the rewarded points are the same for each characteristic, that is to say the differences between the degree of importance of the characteristics are also neglected.

[End of document]