E

UPOV

TWC/18/14
ORIGINAL: English
DATE: September 28, 2000

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Eighteenth Session
Kyiv, June 12 to 15, 2000

SUPPLEMENT TO THE SPECIAL APPLICATIONS OF DUS VARIETY DESCRIPTIONS

Document prepared by experts from Hungary

Supplement to the Special Applications of DUS Variety Descriptions

Zoltán Veress
National Institute for Agricultural Quality Control
Budapest

1. Introduction

A document named “Special Applications of DUS Variety Descriptions” (TWC/17/12) was prepared for the seventeenth session of the TWC (Helsinki – Turku, 1999). This document is a supplement to it.

The aims of this supplement are as follows:

- to study the types of data and the distance functions involved in the study of variety descriptions;
- comparisons of evaluations made on variety descriptions for a period of years;
- comparisons of different versions of method (each characteristics has the same importance vs. there are weighting of the characteristics).

2. The types of data

The types of data occurring in the variety descriptions have been collected in Figure 1.

Table 1.

(DUS data	Variety description)
Nominal	Nominal
Binary (dichotomous)	Binary
Ordinal	Ordinal
Transformation of data	
Interval	Ordinal
Ratio	Ordinal

In document TWC/17/6 entitled “Handling of Visually Assessed Characteristics”, by Uwe Meyer, Helsinki–Turku 1999, there is a good table for the different kind of data types with examples and definitions.

3. Similarity, distance functions, Gower measure

There is a good collection and system of the different kinds of distance functions in the document by J. Law, J. Hayter: Similarity, Clustering and Dendrograms, 1996.

According to this collection and referring to the data types occuring in the variety descriptions mentioned above we can conclude that there is a mixture of different kind of data types in the variety descriptions and therefore a good distance function in the case of variety descriptions is a complex problem. The most general distance function is the so-called Gower measure.

W. Pilarczyk applied this kind of distance function in his paper TWC/14/2 entitled “Application of Gower’s Similarity Coefficient to Detect Most Similar Varieties”.

The Gower measure:

$$1 - \left(\sum_{k=1}^{m} (w_i\, s_i) / \sum_{k=1}^{m} w_i \right)$$

where s_i is the score and w_i the weight. The score and weight for different types of variable are:

Table 2.

Data Type	s_i	w_i
Quantitative	1 - abs(xi - yi) / range 0 if $w_i = 0$	0 for missing values 1 for all other situations
Binary	1 for matches 0 for mis-matches 0 if $w_i = 0$	0 for negative values 1 for all other situations
Qualitative (only nominal ?!)	1 for matches 0 for mis-matches 0 if $w_i = 0$	0 for negative values 1 for all other situations

In my opinion the Gower measure is suitable for the binary and nominal data but is unsuitable in the case of ordinal data. (In this case there would be a loss of information contained by the ordinal data.) As a consequence we looked for another distance procedure.

4. The calculation of the distances between the varieties

4.1 There is no weighting among the characteristics

The distance values of variety pairs are calculated in the following way:

1. each variety is compared to each other variety;
2. the state values are compared in each characteristics in each variety pair;
3. the rewarded points according to the result of a comparison (ordinal type):

Table 3. “A” version

The absolute value of the difference of two scores (state values)	The rewarded points
0	10
1	5
2	3
3	1
> 3	0

In case of nominal and binary type: the reward is 10 for matches, 0 otherwise.

4. The reward points are summed:

For each (r, q) variety pairs (k characteristics):

$$u_{rq} = \underset{k=1}{\overset{m}{}} \quad g\,(|\,c_{rk} - c_{qk}\,|),$$

5. The possible maximal value of this sum is

$$u_{max} = 10 \ \text{x} \ m,$$

where m is the number of characteristics;

6. the similarity % :

$$u_{rq}\,\% = \ u_{rq} / u_{max} * 100$$

$$(u_{max} = 100\,\%)$$

the distance % :

$$d_{rq}\,\% = \ 100\,\% - u_{rq}\,\%$$

4.2 The calculation of the distances between the varieties (there is weighting among the characteristics)

In order to express the difference among the characteristics according to the stable or changeable character of the characteristics a weighting is proposed for the characteristics.

$$u_{rq} = \underset{k=1}{\overset{m}{}} \quad g_k(|\,c_{rk} - c_{qk}\,|),$$

(That is to say, now g_k is applied instead of g .)

Table 4. “B1” version:

1: for the stable characteristics:

The absolute value of the difference of two scores (state values)	The rewarded points
0	20
1	10
2	5
3	3

2: for the average characteristics:

The absolute value of the difference of two scores	The rewarded points
0	10
1	5
2	3
3	1

3: for the non-stable characteristics:

The absolute value of the difference of two scores	The rewarded points
0	5
1	3
2	2
3	1

Table 5. “B2” version:

1: for the stable characteristics:

The absolute value of the difference of two scores	The rewarded points
0	50
1	20
2	5
3	3

2: for the average characteristics:

The absolute value of the difference of two scores	The rewarded points
0	10
1	5
2	3
3	1

3: for the non-stable characteristics:

The absolute value of the difference of two scores	The rewarded points
0	5
1	3
2	2
3	1

5. The search for similarity groups

The detailed demonstration of this method is written in the TWC/17/12 document.

A brief summary of this procedure is given as follows:

The distance (similarity) values of each variety pairs are calculated according to the way given above and their frequency histogram is constructed. In the frequency histogram the most similar variety pairs are near to 0 distance but we don’t know the other limit of the “similarity interval” (L_r%).

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 are approximately the same in both variety descriptions. The control variety description was made by using the random number generator of the Excel.

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 %. So, if there is a variety pair in the so defined “similarity interval” then these varieties are really similar to each other and not by chance. It can be declared that the probability of finding variety pairs in “similarity interval” which are similar to each other by chance is very low. It is the reason that this point L % value is chosen as the real similarity threshold % (L_r%).

6. An application

Calculations were made for data of Winter Barley (year 1996, 1997, 1998 and 1996-98, Excel, programs written in VBA).

As an example in the Table 6., the variety description of the Winter Barley data of 1996-98 is shown.

Table 6.

Variety description of Winter Barley (years: 1996-98)																													
Var	Characteristics (TG/19/10)																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
4	4	1	9	5	4	7	5	9	6	5	1	3	2	5		3	7	2	3		2	1	9	3	1	9	2	2	1
7	4	9	9	3	6	6	5	9	3	4	6	7	2	5		7	7	3	3		2	2	9	1	1	1	2	2	1
9	4	9	9	4	3	5	8	9	3	1	4	5	2	5		5	7	3	3		3	2	9	1	4	9	2	1	1
10	4	9	9	4	5	6	6	9	5	5	2	3	2	5		3	7	3	3		1	2	9	3	1	9	2	2	1
12	3	1	1		4	7	4	1		4	3	5	2	5		4	7	3	4		2	1	9	1	2	9	2	2	1
20	3	1	9	7	8	6	5	9	7	7	1	1	1	3		4	7	3	3	3	2	2	9	3	3	1	2	1	3
22	5	9	1		5	6	3	1		4	3	3	2	5		3	7	3	3		1	2	9	1	1	9	2	1	2
23	4	9	9	2	4	5	6	9	1		2	7	2	5		5	7	3	3		3	2	9	1	2	9	2	1	1
24	6	9	9	5	8	5	7	9	6	2	2	5	1	5		7	7	3	5	2	2	1	9	7	4	1	2	2	1
27	3	9	1		4	7	5	1		6	3	6	2	5		5	7	3	3		2	2	9	1	3	9	2	1	1
30	3	1	1		6	7	2	1		4	5	4	2	5		3	7	3	3		2	2	9	1	3	1	2	3	3
33	4	1	9	4	5	7	6	9	2	5	5	5	2	5		5	7	3	3		2	2	9	1	1	9	2	2	1
42	3	1	9	2	5	7	2	9	5	6	5	5	2	5		3	7	3	4		2	1	9	1	2	1	2	2	1
62	3	9	9	3	4	7	4	9	2	5	2	5	2	5		1	7	3	3		2	1	9	1	1	9	2	2	1
64	5	9	9	7	8	5	5	9	7	2	3	5	1	5		6	7	5	4	1	1	2	9	1	3	1	2	1	1
65	3	9	9	1	6	7	5	9	2	5	3	7	2	5		3	7	3	3		2	2	9	1	4	9	2	1	1
66	4	9	9	2	5	7	5	9	3	8	5	6	2	5		5	7	3	4		2	2	9	1	3	9	2	1	1
67	3	9	1		7	6	4	1		4	4	7	2	5		4	7	3	3		2	2	9	1	4	9	2	1	1
68	3	1	1		8	6	5	1		6	5	3	1	5		7	6	3	4	2	2	2	9	1		1	2	1	1
69	3	1	9	6	8	6	2	9	7	6	5	5	1	5		8	5	3	5	3	1	2	9	8	5	1	2	1	1
71	4	9	9	2	6	6	6	9	4	5	6	7	2	5		5	7	2	3		3	2	9	1	5	9	2	1	1
73	5	9	9	5	7	2	5	9	5	1	5	6	1	5		7	7	3	3	2	2	2	9	2	2	1	2	2	1

The number of varieties involved to this variety description is 22. The number of varieties belonging to one of the similarity groups is 14. Consequently, 8 varieties don’t belong to any similarity groups.

The results of the calculations searching the similarity groups are shown in the following five tables. The similarity groups (1, 2, 3, and their versions) are shown in different order according to years, procedure versions (A, B1 and B2). According to these results the difference among the years and procedure versions is relatively small. It is a quantitative conclusion and only for Winter Barley. Now, a work is under way: this procedure will be made for the main crops and after having finished this work a more general statement can be done.

1. Similarity groups (III) according to the different years (I) and *procedure versions* (II)													
I	II	III	Varieties										
96	A	1	4	7	9	10	23		33	62	65	66	71
97	A	1							33	62	65	66	71
98	A	1	4		9	10	23	27	33	62	65	66	71
6-8	A	1	4	7	9	10	23		33	62	65	66	71
97	A	1b	4			10							
97	A	1c			9		23						
96	A	2	22	27		67							
97	A	2	22	27		67							
98	A	2		27		67							
6-8	A	2		27		67							
96	A	3			24	64							
98	A	3				64	73						
97	A	3a		7			73						
96	B1	1				10	23			62	65	66	71
97	B1	1			9		23			62	65	66	71
98	B1	1			9		23				65	66	71
6-8	B1	1			9		23				65	66	71
96	B1	2		27		67							
97	B1	2		27		67							
98	B1	2		27		67							
6-8	B1	2		27		67							
96	B1	3			24	64							
97	B1	3		7	24		73						
98	B1	3				64	73						
96	B2	1				10	23			62	65	66	71
97	B2	1			9	10	23			62	65	66	71
98	B2	1			9	10	23			62	65	66	71
6-8	B2	1			9	10	23			62	65	66	71
96	B2	1a	4						33				
98	B2	1a	4						33				
96	B2	2	22	27		67							
97	B2	2	22	27		67							
98	B2	2		27		67							
6-8	B2	2		27		67							
96	B2	3		7	24	64	73						
97	B2	3		7	24	64	73						
98	B2	3			24	64	73						
6-8	B2	3		7	24	64	73						

2. Similarity groups (III) according to the different *years* (I) and *procedure versions* (II)													
I	II	III	Varieties										
96	A	1	4	7	9	10	23		33	62	65	66	71
96	B1	1				10	23			62	65	66	71
96	B2	1				10	23			62	65	66	71
96	B2	1a	4						33				
97	A	1							33	62	65	66	71
97	A	1b	4			10							
97	A	1c			9		23						
97	B1	1			9		23			62	65	66	71
97	B2	1			9	10	23			62	65	66	71
98	A	1	4		9	10	23	27	33	62	65	66	71
98	B1	1			9		23				65	66	71
98	B2	1			9	10	23			62	65	66	71
98	B2	1a	4						33				
96	A	2	22	27		67							
96	B1	2		27		67							
96	B2	2	22	27		67							
97	A	2	22	27		67							
97	B1	2		27		67							
97	B2	2	22	27		67							
98	A	2		27		67							
98	B1	2		27		67							
98	B2	2		27		67							
96	A	3			24	64							
96	B1	3			24	64							
96	B2	3		7	24	64	73						
97	A	3a		7			73						
97	B1	3		7	24		73						
97	B2	3		7	24	64	73						
98	A	3				64	73						
98	B1	3				64	73						
98	B2	3			24	64	73						
6-8	A	1	4	7	9	10	23		33	62	65	66	71
6-8	B1	1			9		23				65	66	71
6-8	B2	1			9	10	23			62	65	66	71
6-8	A	2		27		67							
6-8	B1	2		27		67							
6-8	B2	2		27		67							
6-8	B2	3		7	24	64	73						

3. Similarity groups (III) according to the different *years* (I) and procedure versions (II)													
I	II	III	Varieties										
6-8	A	1	4	7	9	10	23		33	62	65	66	71
6-8	A	2		27		67							
6-8	B1	1			9		23				65	66	71
6-8	B1	2		27		67							
6-8	B2	1			9	10	23			62	65	66	71
6-8	B2	2		27		67							
6-8	B2	3		7	24	64	73						
96	A	1	4	7	9	10	23		33	62	65	66	71
96	A	2	22	27		67							
96	A	3			24	64							
96	B1	1				10	23			62	65	66	71
96	B1	2		27		67							
96	B1	3			24	64							
96	B2	1				10	23			62	65	66	71
96	B2	1a	4						33				
96	B2	2	22	27		67							
96	B2	3		7	24	64	73						
97	A	1							33	62	65	66	71
97	A	1b	4			10							
97	A	1c			9		23						
97	A	2	22	27		67							
97	A	3a		7			73						
97	B1	1			9		23			62	65	66	71
97	B1	2		27		67							
97	B1	3		7	24		73						
97	B2	1			9	10	23			62	65	66	71
97	B2	2	22	27		67							
97	B2	3		7	24	64	73						
98	A	1	4		9	10	23	27	33	62	65	66	71
98	A	2		27		67							
98	A	3				64	73						
98	B1	1			9		23				65	66	71
98	B1	2		27		67							
98	B1	3				64	73						
98	B2	1			9	10	23			62	65	66	71
98	B2	1a	4						33				
98	B2	2		27		67							
98	B2	3			24	64	73						

4. *Similarity groups* (III) according to the different years (I) and procedure versions (II)													
I	II	III	Varieties										
6-8	A	1	4	7	9	10	23		33	62	65	66	71
6-8	B1	1			9		23				65	66	71
6-8	B2	1			9	10	23			62	65	66	71
96	A	1	4	7	9	10	23		33	62	65	66	71
96	B1	1				10	23			62	65	66	71
96	B2	1				10	23			62	65	66	71
96	B2	1a	4						33				
97	A	1							33	62	65	66	71
97	A	1b	4			10							
97	A	1c			9		23						
97	B1	1			9		23			62	65	66	71
97	B2	1			9	10	23			62	65	66	71
98	A	1	4		9	10	23	27	33	62	65	66	71
98	B1	1			9		23				65	66	71
98	B2	1			9	10	23			62	65	66	71
98	B2	1a	4						33				
6-8	A	2		27		67							
6-8	B1	2		27		67							
6-8	B2	2		27		67							
96	A	2	22	27		67							
96	B1	2		27		67							
96	B2	2	22	27		67							
97	A	2	22	27		67							
97	B1	2		27		67							
97	B2	2	22	27		67							
98	A	2		27		67							
98	B1	2		27		67							
98	B2	2		27		67							
6-8	B2	3		7	24	64	73						
96	A	3			24	64							
96	B1	3			24	64							
96	B2	3		7	24	64	73						
97	A	3a		7			73						
97	B1	3		7	24		73						
97	B2	3		7	24	64	73						
98	A	3				64	73						
98	B1	3				64	73						
98	B2	3			24	64	73						

5. *Similarity groups* (III) according to the different *years* (I) and procedure versions (II)													
I	II	III	Varieties										
6-8	A	1	4	7	9	10	23		33	62	65	66	71
96	A	1	4	7	9	10	23		33	62	65	66	71
97	A	1							33	62	65	66	71
97	A	1b	4			10							
97	A	1c			9		23						
98	A	1	4		9	10	23	27	33	62	65	66	71
6-8	B1	1			9		23				65	66	71
96	B1	1				10	23			62	65	66	71
97	B1	1			9		23			62	65	66	71
98	B1	1			9		23				65	66	71
6-8	B2	1			9	10	23			62	65	66	71
96	B2	1				10	23			62	65	66	71
96	B2	1a	4						33				
97	B2	1			9	10	23			62	65	66	71
98	B2	1			9	10	23			62	65	66	71
98	B2	1a	4						33				
6-8	A	2		27		67							
96	A	2	22	27		67							
97	A	2	22	27		67							
98	A	2		27		67							
6-8	B1	2		27		67							
96	B1	2		27		67							
97	B1	2		27		67							
98	B1	2		27		67							
6-8	B2	2		27		67							
96	B2	2	22	27		67							
97	B2	2	22	27		67							
98	B2	2		27		67							
96	A	3			24	64							
97	A	3a		7			73						
98	A	3				64	73						
96	B1	3			24	64							
97	B1	3		7	24		73						
98	B1	3				64	73						
6-8	B2	3		7	24	64	73						
96	B2	3		7	24	64	73						
97	B2	3		7	24	64	73						
98	B2	3			24	64	73						

[End of document]