E

TWC/26/17
ORIGINAL: English
DATE: August 22, 2008

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Sixth Session
Jeju, Republic of Korea, September 2 to 5, 2008

SOME CONSEQUENCES OF REDUCING THE NUMBER OF PLANTS OBSERVED IN THE ASSESSMENT OF QUANTITATIVE CHARACTERISTICS OF REFERENCE VARIETIES[1]

Document prepared by experts from Denmark and Germany

Introduction

1. At its twenty-third session, held in Ottawa, Canada, from June 13 to 16, 2005, and at its twenty-fourth session, held in Nairobi, June 19 to 22, 2006, the Technical Working Party on Automation and Computer Programs (TWC) discussed the impact of the reduction in the number of plants to be observed in the assessment of distinctness and uniformity. The cost of assessing 60 plants of all reference varieties, especially in crops where the number of reference varieties is large, is one of the main reasons for considering the possibility for a reduction in the number of plants to be observed. Previous discussions have shown that some decisions on distinctness and uniformity after such a reduction may change (documents TWC/24/10 and TWC/24/12). Furthermore, such changes in the number of assessed plants may need some changes in the software used for analyzing the data using the COY-D and COY-U methods. This paper will focus on the changes that may be needed for COY-U.

[1] The term reference varieties here refers to established varieties which have been included in the growing trial and which have comparable expression of the characteristics under investigation

Considerations for the COY-U method

2. In order to consider this question we first looked at the model behind the present method in a more detailed formulation – in order to include the varying number of observations behind each observation.

3. We assumed that each original observation in a given year is as follows:

X_{vgp} is assumed to be $N(\mu_{vg}, \sigma_X^2)$

$Y_v = \log(SD_v + 1)$ has mean θ_v and variance σ_Y^2

where

X_{vgp} is the recorded value of plant p of variety v in replicate g

SD_v is the average standard deviation of variety v over the replicates, i.e. an estimate of σ_X

θ_v is the mean value of $\log(SD_v + 1)$ of variety v

$\sigma_Y^2 = \sigma_f^2 + \sigma_O^2$ is the variance of Y_v which is regarded as a sum of two components, one depending on f, the number of degrees of fredom in estimating SD_i and another depending on other sources of variation, such as variety and year and assumed independent of the number of degrees of freedom

$\sigma_f^2 = \dfrac{\sigma_X^2}{1+\sigma_X^2}\dfrac{0.5}{f}$ which for $\sigma_X^2 >> 1$ becomes $\sigma_f^2 \approx \dfrac{0.5}{f}$. The same value for σ_f^2 is obtained if $Y_v = \log(SD_v)$

4. In the following, we will assume that the variables are scaled so that $\sigma_X^2 >> 1$ - or that $Y_v = log(SD_v)$. If X_{vgp} is not normally distributed then the last term is changed to $\sigma_f^2 = \dfrac{0.5}{f} + \dfrac{\gamma}{4n}$, where γ and n are the kurtosis and number of observations, respectively.

5. After sorting the observations by the corresponding mean values, $X_{(v)}$, trend values, $T_{(v)}$, are calculated as mean of the values $Y_{(v-4)}$. $Y_{(v-3)}$, $Y_{(v-2)}$, $Y_{(v-1)}$, $Y_{(v)}$, $Y_{(v+1)}$, $Y_{(v+2)}$, $Y_{(v+3)}$, $Y_{(v+4)}$, with the exceptions that the 4 observations at each end are based on only the first/last 3, 3, 5 and 7 observations. This means that the variances of the trend values, $T_{(v)}$, varies between $\dfrac{1}{3}\sigma_Y^2$ and $\dfrac{1}{9}\sigma_Y^2$ and that the correlations between them varies between 0 and 1 with 8/9 being the most common value for "neighbours". From that the adjusted values are calculated for each reference variety as $A_v = Y_v - T_v + \overline{Y}_{.}$. The variance of A_v will then be:

$$Var(A_v) = \left(\frac{m-1}{m} + \frac{1}{r}\right)\sigma_Y^2,$$

where m is the number of observations used for calulating T_v and r is the number of reference varieties.

6. Based on the trend values for the reference varieties, the trend values for the candidate varieties are calculated as weighted means of trend values for the 2 varieties for which it is located between, i.e.:

$$T_c = \frac{(x_c - x_{(v)})T_{(v+1)} + (x_{(v+1)} - x_c)T_{(v)}}{x_{(v+1)} - x_{(v)}}$$

7. The variance of T_c will depend on the variances of $T_{(v)}$ and $T_{(v+1)}$ and the covariance between them as well as the differences between the mean values, i.e. the differences between x_c, $x_{(v)}$ and $x_{(v+1)}$. The largest variance of T_c will occur when the mean of the candidate is located at the extremes, i.e. less than $x_{(2)}$ or larger than $x_{(r-1)}$ and the smallest variance will occur when the mean of the candidate is equal to the mean of two reference varieties located away from the extremes, i.e. $x_c = 0.5(x_{(v)}+x_{(v+1)})$ and $5 \leq (v) \leq r-4$. The largest and smallest variance of T_c will be between:

$((1/18)^2 + 8\times(1/9)^2 + (1/18)^2)\sigma_Y^2 \approx 0.105\sigma_Y^2$ and $(1/3)\sigma_Y^2 \approx 0.333\sigma_Y^2$. For candidates with means located between $x_{(5)}$ and $x_{(r-4)}$ the variance on T_c will be between $0.105\sigma_Y^2$ and $0.111\sigma_Y^2$

8. All trend values for the reference varieties in y years are submitted to a one-way ANOVA. Based on the variances of the trend values for the reference varieties the expected values of the residual sum of squares can be calculated as:

$$E(RSS) = y\left(\frac{2}{3}+\frac{2}{3}+\frac{4}{5}+\frac{6}{7}+(r-8)\frac{8}{9}+\frac{6}{7}+\frac{4}{5}+\frac{2}{3}+\frac{2}{3}+r\frac{1}{r}\right)\sigma_Y^2 \approx y\left(r\frac{8}{9}-0.13016\right)\sigma_Y^2$$

and the expected value of the residual value thus become

$E(RMS) = \frac{1}{r-1}\left(r\frac{8}{9}-0.13016\right)\sigma_Y^2$ which for large r will approach $0.8889\sigma_Y^2$

or

$$E(RMS) = \left(\frac{r}{r-1}\frac{8}{9}-\frac{0.13016}{r-1}\right)\sigma_Y^2 = \left(\frac{r}{r-1}\frac{8}{9}-\frac{0.13016}{r-1}\right)\left(\frac{0.5}{f_r}+\frac{\gamma}{4n_r}+\sigma_O^2\right)$$

which for large r will approach $\frac{0.4444}{f_r}+\frac{0.2222\gamma}{n_r}+0.8889\sigma_O^2$

and the term containing γ is for non-normally distributed data only

Note that this value is less than if the raw log SDs had been used and that the number of degrees of freedom for the reference varieties is less important if $\sigma_O^2 >> \sigma_f^2$.

9. The last step in the present method is to calculate the variance of the difference between the mean of the adjusted value for a candidate and all reference varieties, which is done as:

$Var(\overline{D}_c) = \left(\frac{1}{k}+\frac{1}{kr}\right)RMS$ which for large r gets close to $\frac{1}{k}RMS$.

10. The adjusted value for the candidate variety in each year can be written as: $A_c = Y_c - T_c + \overline{T}_.$, so the difference becomes $D_c = A_c - \overline{T}_. = Y_c - T_c + \overline{T}_. - \overline{T}_. = Y_c - T_c$, which is then averaged over k years. The variance of this mean difference depends on whether the mean of the response for the candidate is close to the extremes of the reference varieties or in

the middle part of the reference varieties. The extremes of this variance will – based on the above variances on Y_c and T_c – be between:

$$Var(\overline{D}_c) = \left(\frac{1}{k}+\frac{0.105}{k}\right)\sigma_Y^2 = \frac{1.105}{k}\sigma_Y^2 \text{ and } Var(\overline{D}_c) = \left(\frac{1}{k}+\frac{0.333}{k}\right)\sigma_Y^2 = \frac{1.333}{k}\sigma_Y^2$$

11. So in the present method the variance of the difference between the candidate and the mean of the reference varieties may vary by a factor of 1.21 and is approximately 1.24-1.5 times larger than the value calculated from residual mean square of the one-way ANOVA.

12. This means that the present method underestimates the variance used for calculating the threshold above which the candidate should be rejected as non-uniform. The reason for this is partly that the expected value of RMS from the ANOVA is less than the expected value of Y_v and partly that only the number of varieties used in the local adjustment influences this variance and not the total number of reference varieties.

13. In the above calculations, it is assumed that all SDs are based on the same number of observations and thus have the same number of degrees of freedom. If the numbers are different, then σ_Y^2 in the above formulas for the variance of the difference between the candidate and the mean of the reference varieties becomes different from that of σ_Y^2 in the expression for *E(RMS)*. In order to take this into account we have to replace σ_Y^2 with $\sigma_{\mathrm{f}}^2 + \sigma_O^2$ where σ_O^2 is independent of the number of observations used for estimating the SD's. The formulas above then become

$$Var(\overline{D}_c) = \left(\frac{1}{k}\sigma_{f_c}^2 + \frac{0.105}{k}\sigma_{f_r}^2\right) + \left(\frac{1}{k}+\frac{0.105}{k}\right)\sigma_{\mathrm{O}}^2 = \frac{1}{k}\left(\frac{0.5}{f_c}+\frac{\gamma}{4n_c}+0.105\left(\frac{0.5}{f_r}+\frac{\gamma}{4n_r}\right)+1.105\sigma_{\mathrm{O}}^2\right)$$

and

$$Var(\overline{D}_c) = \left(\frac{1}{k}\sigma_{f_c}^2 + \frac{0.333}{k}\sigma_{f_r}^2\right) + \left(\frac{1}{k}+\frac{0.333}{k}\right)\sigma_{\mathrm{O}}^2 = \frac{1}{k}\left(\frac{0.5}{f_c}+\frac{\gamma}{4n_c}+0.333\left(\frac{0.5}{f_r}+\frac{\gamma}{4n_r}\right)+1.333\sigma_{\mathrm{O}}^2\right)$$

where f_c, f_r, n_c, and n_r is the number of degrees of freedom for estimating the the standard deviation and the number of observations for a candidate and a reference variety, respectively

14. If the data are normally distributed then this can be written as:

$$Var(\overline{D}_c) = \frac{1}{k}\left(\frac{0.5}{f_c}+\frac{0.052}{f_r}+1.105\sigma_{\mathrm{O}}^2\right) \text{ and } Var(\overline{D}_c) = \frac{1}{k}\left(\frac{0.5}{f_c}+\frac{0.167}{f_r}+1.333\sigma_{\mathrm{O}}^2\right)$$

15. The residual mean square of the one-way ANOVA depends on the number of degrees of freedom unless $\sigma_O^2 >> \sigma_f^2$. In order to examine that, γ, σ_O^2 and σ_f^2 are estimated for data on Oilseed Rape performed in Germany during the year 2002-2004. The results are shown in Table 1 together with the variance of the difference between a variety and the mean of the reference varieties. It is seen that the kurtosis was always estimated to be positive (for normal distributed variables the kurtosis is expected to be close to zero). The largest kurtosis was found for time of flowering. The variance component that depend on the number of degrees

of freedom (and number of plants per variety in each year) (σ_f^2) and the component that is assumed to be constant (σ_O^2) were of the same magnitude.

Table 1: Number of varieties, degrees of freedom for SDs together with estimated residual variance and estimated components of the variance of log(SD) for 12 characters recorded at station 6 in Germany in 2002-2004 using all recorded plants

Characteristic	Number of		Kurtosis	Variances			
	Vars	DF		RMS	σ_f^2	σ_O^2	$\overline{D}_c$
Cotyledon: length	301	57	0.66	0.01798	0.0115	0.0087	0.0060
Cotyledon: width	301	57	0.27	0.01517	0.0099	0.0071	0.0051
Leaf: length	300	57	0.25	0.01309	0.0098	0.0049	0.0044
Leaf: width	300	57	0.22	0.01278	0.0097	0.0047	0.0043
Leaf: number of lobes	300	57	0.21	0.01633	0.0097	0.0087	0.0055
Leaf: length of petiole	300	57	0.28	0.01354	0.0100	0.0052	0.0045
Plant: total length	297	57	0.35	0.02110	0.0102	0.0134	0.0071
Siliqua: length	297	57	0.42	0.02806	0.0105	0.0210	0.0094
Siliqua: length of beak	297	57	0.42	0.01823	0.0105	0.0099	0.0061
Siliqua: width	297	57	0.47	0.01903	0.0107	0.0106	0.0064
Siliqua: length of peduncle	297	57	1.15	0.03107	0.0136	0.0213	0.0104
Time of flowering	299	57	1.63	0.02797	0.0156	0.0158	0.0094

Table 2: Number of varieties, degrees of freedom for SDs together with estimated residual variance and estimated components of the variance of log(SD) for 12 characters recorded at station 6 in Germany in 2002-2004 using a random sample of 10 plants from each plot

Characteristic	Number of		Kurtosis	Variances			
	Vars	DF		RMS	σ_f^2	σ_O^2	$\overline{D}_c$
Cotyledon: length	301	27	0.71	0.02879	0.0244	0.0079	0.0096
Cotyledon: width	301	27	0.29	0.02534	0.0210	0.0074	0.0085
Leaf: length	300	27	0.28	0.02192	0.0209	0.0037	0.0073
Leaf: width	300	27	0.21	0.02317	0.0203	0.0057	0.0078
Leaf: number of lobes	300	27	0.23	0.02413	0.0204	0.0066	0.0081
Leaf: length of petiole	300	27	0.29	0.02206	0.0209	0.0038	0.0074
Plant: total length	297	27	0.37	0.03192	0.0216	0.0142	0.0107
Siliqua: length	297	27	0.38	0.03751	0.0217	0.0204	0.0125
Siliqua: length of beak	297	27	0.40	0.02997	0.0219	0.0118	0.0100
Siliqua: width	297	27	0.39	0.02865	0.0218	0.0104	0.0096
Siliqua: length of peduncle	297	27	1.19	0.04214	0.0284	0.0188	0.0141
Time of flowering	299	27	1.46	0.04105	0.0307	0.0154	0.0137

16. In Table 2 the same figures are shown for the case when only 10 randomly sampled observations from each plot were used for the calculations. As expected, the variance components that depend on the number of degrees of freedom were approximately doubled and the variance components that did not depend on the number of degrees of freedom were approximately unchanged. The variances of the difference between the log SD of the candidate and the reference varieties (last column of table 1 and 2) were in all cases larger in table 2 than in Table 1 because the residual mean square was increased.

Table 3: Calculated minimum and maximum variance of the difference between the candidate and the reference varieties for three selected intensities of observations for reference varieties. The calculations are based on the residual variance for all plants

	Minimum variance			Maximum variance		
No of observations / degrees of freedom in estimating SD	60/57	30/27	15/12	60/57	30/27	15/12
Cotyledon: length	0.0074	0.0079	0.0089	0.0076	0.0082	0.0094
Cotyledon: width	0.0063	0.0066	0.0075	0.0064	0.0069	0.0080
Leaf: length	0.0054	0.0058	0.0067	0.0055	0.0060	0.0071
Leaf: width	0.0053	0.0057	0.0065	0.0054	0.0059	0.0070
Leaf: number of lobes	0.0067	0.0071	0.0080	0.0069	0.0074	0.0085
Leaf: length of petiole	0.0056	0.0060	0.0069	0.0057	0.0062	0.0074
Plant: total length	0.0087	0.0091	0.0100	0.0089	0.0094	0.0106
Siliqua: length	0.0116	0.0120	0.0129	0.0119	0.0124	0.0136
Siliqua: length of beak	0.0075	0.0079	0.0089	0.0077	0.0082	0.0094
Siliqua: width	0.0079	0.0083	0.0092	0.0081	0.0086	0.0098
Siliqua: length of peduncle	0.0128	0.0133	0.0145	0.0132	0.0138	0.0153
Time of flowering	0.0116	0.0121	0.0134	0.0119	0.0126	0.0142

Table 4: Calculated minimum and maximum variance of the difference between the candidate and the reference varieties for three selected intensities of observations for reference varieties. The calculations are based on the residual variance for a random sample of 10 plants from each plot for the reference varieties

	Minimum variance			Maximum variance		
No of observations / degrees of freedom in estimating SD	60/57	30/27	15/12	60/57	30/27	15/12
Cotyledon: length	0.0072	0.0077	0.0087	0.0074	0.0080	0.0093
Cotyledon: width	0.0064	0.0068	0.0077	0.0066	0.0071	0.0082
Leaf: length	0.0050	0.0054	0.0063	0.0052	0.0056	0.0068
Leaf: width	0.0057	0.0060	0.0069	0.0058	0.0063	0.0074
Leaf: number of lobes	0.0060	0.0064	0.0073	0.0062	0.0067	0.0078
Leaf: length of petiole	0.0051	0.0055	0.0064	0.0052	0.0057	0.0068
Plant: total length	0.0090	0.0094	0.0103	0.0093	0.0098	0.0109
Siliqua: length	0.0113	0.0117	0.0126	0.0116	0.0121	0.0133
Siliqua: length of beak	0.0082	0.0086	0.0095	0.0084	0.0089	0.0101
Siliqua: width	0.0076	0.0080	0.0090	0.0078	0.0083	0.0095
Siliqua: length of peduncle	0.0120	0.0125	0.0137	0.0123	0.0129	0.0144
Time of flowering	0.0111	0.0117	0.0129	0.0114	0.0121	0.0137

17. Tables 3 and 4 show the estimated variance of the difference between adjusted Log SD and the mean of the reference varieties when all 60 plants were recorded for the candidate variety and 60, 30 or 15 plants were recorded for the reference varieties. In Table 3 the calculations are based on an ANOVA using all plants whereas in table 4 only a random sample of 10 plants from each plot was used. The two cases gave approximately the same results. The change caused by going from 60 plants to 30 plants changes the variance by approximately the same amount as that of going from a candidate with minimum variance (i.e. with mean in the middle of the reference collection) to a candidate with maximum variance (i.e. with mean at the extremes of the reference collection). If the number of recorded plants were further decreased to 15, then the variance was increased by more than the difference between candidates with the minimum and maximum variance in the present system.

Conclusions

18. From the above it can be concluded that the variances calculated in the present system do not reflect the expected value of the true variance as they are too small, partly because the expected value of RMS from the ANOVA is less than the expected value of $Var(Y_v)$ and partly because only the number of varieties used in the local adjustment influence this variance (and not the total number of reference varieties). However, the present method probably adjusts for this bias by using a large t-value (by using a small α-value). Also it can be concluded that the residual mean square (RMS) may depend significantly on the number of observations recorded as the component of RMS that depends on the number of observations (degrees of freedom) was not a negligible part. Therefore, it is suggested that an adjusted RMS and the upper threshold are calculated as follows:

$$RMS_{Adj} = RMS - \frac{0.5}{f_r} - \frac{\gamma}{4n_r} + \frac{0.5}{f_c} + \frac{\gamma}{4n_c}$$

$$UC_p = \overline{SD_r} + t_p\sqrt{\frac{RMS_{Adj}}{k} + \frac{RMS}{rk}}$$

where the RMS_{Adj} is the value that would be expected if all (n_c) plants were recorded. This value is used when calculating the contribution from the candidate variety, while the *RMS* (unadjusted) is used when calculating the contribution from the reference varieties. Note that these modifications do not correct for the bias in the present method. A correction also for the bias (caused by a too low *RMS* and a too large divisor when calculating the contribution from the reference varieties) in the present method could be done using the following formulas:

$$\sigma_{f_r}^2 = \frac{0.5}{f_r} + \frac{\gamma}{4n_r} \qquad \sigma_{f_c}^2 = \frac{0.5}{f_c} + \frac{\gamma}{4n_c} \qquad \sigma_O^2 = RMS(r-1)\left(\frac{8r}{9} - 0.13016\right)^{-1} - \sigma_{f_r}^2$$

$$UC_p = \overline{SD_r} + t_p\sqrt{\frac{\sigma_{f_c}^2 + \sigma_O^2}{k} + \frac{c\left(\sigma_{f_r}^2 + \sigma_O^2\right)}{k}}$$

with c depending on the location of the candidate (ranging between 0.105 and 0.333).

19. The constant c is approximately equal to $1/n$, where n is the average number of reference varieties used for calculating the two T_v's on which T_c is based (n varies between 3 and 9).

[End of document]