E

TWC/27/15
ORIGINAL: English
DATE: June 11, 2009

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Seventh Session
Alexandria, Virginia, United States of America
June 16 to 19, 2009

POTENTIAL APPROACHES TO IMPROVING COYU

Document prepared by experts from Denmark and the United Kingdom

POTENTIAL APPROACHES TO IMPROVING COYU

Introduction

Method of calculation of COYU

1. At its twenty-sixth session held in Jeju, Republic of Korea, from September 2 to 5, 2008, the TWC considered document TWC/26/17 "Some consequences of reducing the number of plants observed in the assessment of quantitative characteristics of reference varieties[1]" and a presentation by Mr. Kristian Kristensen (Denmark), a copy of which was reproduced as document TWC/26/17 Add..

2. Document TWC/26/17 stated the following with regard to the current method of calculation of COYU:

> "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 [residual mean square] 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[s] 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."

3. The TWC noted the following possible actions to address the bias in the present method of calculation of COYU, as identified and commented on by Mr. Kristensen:

- (i) Ignore the biases
  (comment: the test will most probably be too liberal);
- (ii) Correct only for the bias introduced by the smaller sample sizes
  (comment: the test will be too liberal, but will be comparable to those in the past);
- (iii) Correct only for the present bias
  (comment: the test will be conservative, but not comparable to the past);
- (iv) Correct for all biases
  (comment: there will be no biases, but the tests will not be comparable to the past)

4. The TWC agreed that Denmark and the United Kingdom should prepare a new document, including a simulation using the smoothing spline method. It was noted that that would also allow experts further time to reflect on the situation and possible ways forward.

Method of calculation of COYU

---

[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.

5. At its forty-fifth session held in Geneva, on March 30 and April 1, 2009, the TC noted the discussions concerning the current method of calculation of COYU, as set out in document TC/45/3, and agreed that the Technical Working Parties should be informed about those discussions at their sessions in 2009. The TC requested the TWC to make its recommendations to the TC concerning the proposals set out in paragraph 3 of this document.

Description of methods

6. The biases discovered in TWC/26/17 arise in part to the use of the moving average approach to adjust for the mean-variance relationship. In document TWC/26/17, it was noted that the bias can be corrected but it may also be possible to use other methods of adjustment. Here we recap the moving average method and describe two alternatives.

Moving average (current method)

7. As shown in document TWC/26/17, the present method underestimates the variance of the difference between the estimate of uniformity of the candidate and the average of the reference varieties. We refer back to TWC/26/17 for notation. The present method is based on analysing the logarithm of the standard deviations, i.e. $Y_i = \log(SD_i + 1)$ where SD_i is the pooled stadard diviation for variety i. These are adjusted using the moving average method to give A_i. The comparison of a candidate with the reference varieties is given by: $D_c = A_c - \overline{T} = Y_c - T_c + \overline{T} - \overline{T} = Y_c - T_c$, which is then averaged over k years. Here A_c is the adjusted value for the candidate variety, $\overline{Y}$ the average of the reference varieties, Y_c is the recorded value for the candidate, T_c is the calculated trend value for the candidate and $\overline{T}$ is the mean of the reference trend values. The correct variance on this difference will lie between $1.105k^{-1}\sigma_Y^2$ and $1.333k^{-1}\sigma_Y^2$ whereas the variance applied in the methods will lie between $0.8889k^{-1}\sigma_Y^2$ and $1.0705k^{-1}\sigma_Y^2$. Here the smallest value is for infinite number of reference varieties and the largest value is for 10 reference varieties, which is about the smallest number of reference varieties that should be used in the method.

Linear regression

8. This method is very simple as it is based on regressing the *Y's* i.e. the log(SD) against the means of the characteristic in each year and then adjusting the log(SD) of the candidate according to that relation. The method can be formulated as: For an individual year the variance on the sought difference, D_c will be:

$$\mathrm{var}(D_c) = \mathrm{var}(Y_c - \hat{Y}_c) = \sigma_{reg}^2 + \frac{\sigma_{reg}^2}{r} + \frac{\sigma_{reg}^2 (x_c - \overline{x})^2}{\sum_{i \in reference} (x_i - \overline{x})^2}$$

assuming the model is correct. Here $\hat{Y}_C$ is estimated by linear regression and x_c is the mean value of candidate c, x_i is the mean value for reference variety i and $\overline{x}$ is the mean of the reference varieties. σ_{reg}^2 is the residual variance for the regression of log(SD+1) (or Y) on x.

Here we have $\mathrm{var}(D_c) \geq \sigma_{reg}^2$, with approximate equality when r is large and/or when $x_c = \overline{x}$.

$E(RMS) = \sigma^2_{reg}$ by consideration of the ANOVA for the regression, but care must be taken to divide estimates by the appropriate degrees of freedom allowing for the linear regression fit. So in the style of current method var(D_c) is estimated by
$s^2_D = \left(1+\frac{1}{r}\right)\sigma^2_{reg}$

So $\frac{\operatorname{var}(D_C)}{s^2_D} = 1 + \frac{(x_c-\bar{x})^2}{\left(1+\frac{1}{r}\right)\sum_{i\in reference}(x_i-\bar{x})^2} = 1 + \frac{(x_c-\bar{x})^2}{\left(1+\frac{1}{r}\right)(r-1)\operatorname{var}(x_{ref})} \approx 1$ when r is large and/or when the candidate has a mean expression similar to the mean of the reference varieties. However, if $(x_c-\bar{x})^2 >> Var(x_{ref})$, i.e. the candidate has an extreme mean value, this quotient can be considerable larger than 1.

Now taking average over *k* years we get: $\operatorname{var}(\bar{D}_c) = k^{-1}\operatorname{var}(D_c)$ and that the residual variance from the analysis over years becomes:
$E(RMS) = \frac{1}{k(r-1)}\sum_k\sum_{i\in ref}\left(\hat{Y}_{ki}-\bar{Y}_k\right)^2 = E\left(\hat{\beta}(x_i-\bar{x})\right)^2 = E\left(\hat{\beta}x_i-\hat{\beta}\bar{x}\right)^2$ which we think should equal σ^2_{reg}

For more general linear regression, (1) extends to the form $\operatorname{var}(D_c) = \sigma^2_{reg} + X_c^T\left(X^TX\right)^{-1}X_c\sigma^2_{reg}$.

Smoothing spline

9. This method is like the moving average approach in that it is flexible enough to fit any mean-variance relationship. However, unlike the moving average approach, adjustments for any one point are based on all observations and the method can be set in an additive model framework (like linear regression). It can be seen as a generalization of linear regression. Further explanations can be found in the book by Hastie and Tibshirani (1990). Here we restrict our attention to cubic splines. Büchse et al. (2007) refer to another method, called LOESS, as well.

10. The flexibility of the adjustment can be controlled by the smoothing parameter. Since we do not expect a very complex relationship between the log(SD) values and the means, we have set the smoothing parameter so that there are two degrees of freedom associated with the spline. Note that the linear regression approach uses one degree of freedom, and a quadratic regression would also use two degrees of freedom. As part of future work, more flexible splines might be considered.

11. It is not easy/possible to write down formulae in the way that we can for linear regression. However, it does seem likely that $\frac{\operatorname{var}(D_C)}{s^2_D}$ has a similar form so that when *r* is large this ratio is approximately 1 (so no bias). This is confirmed below.

Comparison of models using simulations

Method of simulation

12. In order to compare the different methods with the present method and a standard using no adjustment the four methods were each applied to simulated data. Eight sets of simulated data were used for four methods (no adjustment, moving average (present method), linear regression and smoothing spline with two degrees of freedom). The eight sets were obtained using the combinations of the following 3 parameters

1. Number of reference varieties: r=10 or r=50
2. Interaction between year and variety: $\sigma_{VY}{}^{2}$=0 or $\sigma_{VY}{}^{2}$=100
3. SD's dependence on variety mean: β=0, $\sigma_{V}{}^{2}$=0 or β=0.1, $\sigma_{V}{}^{2}$=100

13. In all cases we simulated the data for 3 years, using complete block design with 3 blocks each with 20 recorded plants, r reference varieties, c=10 candidate varieties. The simulation was performed according to the following model:

$$X_{yvbp} = \mu + A_y + B_v + C_{yv} + D_{yb} + E_{yvb} + F_{yvbp}$$

where

X_{yvbp} is the recorded value for plant p of variety v in blok b in year y

μ is the average value of the characteristic, here $\mu = 200$

A_y is the effect of year y, here $A_y = 0$

B_v is the effect of variety v, here $B_v \sim N(0, \sigma_V^2)$, $\sigma_V^2 = 0$ or 125

C_{yv} is the interaction effect of year and variety, here $C_{yv} \sim N(0, \sigma_{YV}^2)$, $\sigma_{YV}^2 = 0$ or 100

D_{yb} is the effect of block b in year y, here $D_{yb} = 0$

E_{yvb} is the effect of the plot with variety v in block b in year y

F_{yvbp} is the effct of plant p of variety v in blok b in year y, here $F_{yvbp} \sim N(0, \sigma^2)$, here $\sigma = \sqrt{200} + \beta B_v$

with $\beta = 0$ when σ_V^2=0 or β=0.1 when σ_V^2 =125

$y = 1,2,3;\ v = 1,2,3,\ldots,r+c;\ b = 1,2,3;\ p = 1,2,3,\ldots,20;\ r = 10$ or 50; $c = 10$

14. The simulations were repeated 500 times for each of the 8 combinations mentioned above. For each simulation the variety mean and the pooled standard deviation were calculated for each combination of year and variety. The pooled standard deviations were log-transformed, Y=log(sd). The *logsd*'s=Y's were adjusted according to each of the described methods and the *Y's* for the reference varieties were analyzed using a linear model including the fixed effect of year:

$$Y_{yv} = \mu + \alpha_y + E_{yv}$$

Based on the residual variance from this model, the relative number of candidates that were significantly non-uniform at the 5% level of significance was recorded.

15. Note that the moving average method has not been adjusted for bias.

Results

16. The method of moving average clearly gives too many significant results for all sets of simulations (table 1). The relative number of significant non-uniform candidates that was almost independent of the simulation set seemed to vary around 11-12%, which was about twice as many as required given the level of significance. The best method – among those tested – seemed to be the adjustment using linear regression. For this method, the relative number of significant non-uniform candidates was about 5% when 50 reference varieties were used and about 7% when only 10 varieties were used, indicating that this method may give satisfactory results when a sufficiently large number of reference varieties is used. The smoothing spline method gave a somewhat higher number of significant non-uniform candidates and also here the number is higher when few reference varieties are used, indicating that this may give correct results for very large numbers of reference varieties. If no adjustment is applied, the relative number of significant non-uniform candidates was very close to the expected 5% when the *SD*'s were independent of the mean (simulation set number 1, 2, 5 and 6), but too high if *SD*'s were dependent on the mean.

Table 1: Relative number of significant comparisons using for four different methods under varying assumptions when using alpha=0.05

Set No	Assumptions in simulations			Method			
	No reference varieties, r	Variety,σ_V^2/ Slope, β	Interaction,σ_{YV}^2	No adjustment	Moving average	Linear regression	Smoothing spline
1	50	0/0	0	0.045	0.111	0.048	0.056
2	10	0/0	0	0.050	0.121	0.074	0.125
3	50	125/0.1	0	0.111	0.111	0.049	0.054
4	10	125/0.1	0	0.121	0.119	0.071	0.093
5	50	0/0	100	0.045	0.117	0.047	0.057
6	10	0/0	100	0.050	0.123	0.075	0.119
7	50	125/0.1	100	0.093	0.108	0.047	0.056
8	10	125/0.1	100	0.099	0.116	0.069	0.116

17. The pooled *SD*'s used here each have 57 degrees of freedom. As the simulated data were normally distributed, we should expect a residual variance of $0.5/57\approx0.0088$ when no interaction between year and variety are present and the method properly takes into account the dependence between the *SD*'s and the variety means. For the method where no adjustment is made, the residual variance is very close to the expected value when the *SD*'s were independent of the mean (simulation set number 1, 2, 5 and 6). The residual variance for the linear regression method seems to be fairly close to the expected value when 50 reference varieties are used, but too low if only 10 reference varieties are used. For the methods using smoothing spline and moving average, the residual variance was always lower than expected, which may partly explain the high frequency of significant non-uniform candidates. However, the average residual variance for set number 8 is rather high for the method using moving average. This effect was also seen in using further sets of simulated data with these parameter settings, but we have been unable to explain why this combination had a larger residual variance than expected, but still too many significant differences. However, it might be the interaction between variety and year that produces higher residual variance – and in combination with the known bias it still gives too many significant comparisons, but this has to be investigated further.

Table 2: Average residual variance in the over years analyses for each of 4 different methods under varying assumptions

Set No	Assumptions in simulations			Method			
	No reference varieties, r	Variety, σ_V^2/ Slope, β	Interaction, σ_{YV}^2	No adjustment	Moving average	Linear regression	Smoothing spline
1	50	0/0	0	0.0089	0.0079	0.0087	0.0084
2	10	0/0	0	0.0088	0.0075	0.0078	0.0064
3	50	125/0.1	0	0.0154	0.0081	0.0089	0.0086
4	10	125/0.1	0	0.0151	0.0083	0.0080	0.0066
5	50	0/0	100	0.0089	0.0079	0.0087	0.0084
6	10	0/0	100	0.0088	0.0075	0.0078	0.0064
7	50	125/0.1	100	0.0208	0.0082	0.0090	0.0086
8	10	125/0.1	100	0.0203	0.0091	0.0080	0.0065

Discussion and conclusions

18. All tested methods seem to give incorrect results in some cases at least when the number of reference varieties is low. For these data sets, linear regression performed best. The smoothing spline method performed satisfactorily when there were sufficient reference varieties. The moving average method performed poorly compared to the others.

19. If the form of the relationship between log(SD) and the means were known (through studies on historical data) it would be possible to select a suitable method. For example, this could be no adjustment, linear regression, an appropriate transformation or multiple linear regression. However, this would require a complete change of approach from the current one, where the form of the relationship is derived from the current data set.

20. However, it seems that in practice the current approach is likely to be preferred by most testing authorities. Although the linear regression approach worked well here, it may not be sufficiently flexible to cover all the different types of dependence between the log *SD*'s and the variety means. The smoothing spline method has this flexibility but suffers more when the number of reference varieties is low. Further investigation is required to:

(a) Evaluate the capability of the regression method to cope with the different types of dependence between the *SD*'s and the variety means
(b) Investigate other methods for adjustment based on the current data set, and compare with linear regression, e.g.
    (i) Smoothing splines (including those with more than 2 degrees of freedom)
    (ii) Polynomial regression
    (iii) Transformation of the characteristic in order to obtain approximately constant variance, through Box-Cox transformations, as suggested by Büchse et al. (2007).
    (iv) Incorporate the relation between log *SD's* and variety mean directly in the over years analysis (e.g. include the mean or a function of the mean as a covariate) as suggested by Büchse et al. (2007).
(c) Consider other aspects of the paper by Büchse et al. (2007).

21. When an acceptable method has been chosen the method should be adapted to handle the use of different number of recorded plants for reference varieties and for candidate varieties.

References

Büchse, Andreas.; Piepho, Hans-Peter; Meyer, Uwe (2007). Examination of statistical procedures for checking uniformity in variety trials. Biuletyn Oceny Odmian (Cultivar testing Bulletin). 32, 7-27.

Hastie and Tibshirani (1990). Generalized Additive Models, Chapman & Hall, 1990. 335 pp

[End of document]