UPOV

International Union for the Protection of New Varieties of Plants

E

Technical Working Party on Automation and Computer Programs **TWC/35/6**

Thirty-Fifth Session
Buenos Aires, Argentina, November 14 to 17, 2017

Original: English
Date: October 20, 2017

METHOD OF CALCULATION OF COYU: PRACTICAL EXERCISE, PROBABILITY LEVELS, EXTRAPOLATION & SOFTWARE

Document prepared by experts from the United Kingdom

Disclaimer: this document does not represent UPOV policies or guidance

BACKGROUND

1. At its thirty-second session, the TWC received a presentation by an expert from the United Kingdom on the method for improving the calculation of COYU, including a demonstration version of a module for the DUST software.

2. The TWC agreed to invite users of the COYU method to test the new method and software.

3. The TWC agreed that participants should seek to define probability levels to match decisions using the previous COYU method for continuity in decisions and that the test should be run for rejection probabilities of 1, 2 and 5% levels. The TWC agreed that participants should assess whether the results are consistent in all crops.

4. At the thirty third session, the expert from the United Kingdom reported on the results of this practical exercise. Results and data were received from Finland, France, Kenya and the United Kingdom. It was concluded that further large data sets were required to define the probability levels required.

5. The Technical Committee at its fifty-second session agreed to invite UPOV members' experts to provide the further large data sets for this purpose. Following the invitation sent in April 2016, data sets were received from Denmark and Slovakia.

6. This document reports on the comparison of probability levels between previous and proposed COYU method, development of the software as well as consideration of the extrapolation issue previously raised.

PRACTICAL EXERCISE

7. In July 2014, an invitation to take part in the COYU Practical Exercise was sent to TWC members. Software and a document giving guidance on the software and instructions for the Exercise were sent to those expressing an interest in participation. The R software was made available in October 2015 and the DUST software in December 2015. Participants were asked to evaluate the software and to compare the results obtained by the current version of COYU with those produced by the proposed improved version.

8. A further invitation was sent out to UPOV members in April 2016 seeking more large data sets.

9. The following took part in the exercise or supplied data:

Country	Participant	Software/data	Crops
Denmark	Erik Lawaetz	Data	Oilseed rape
Finland	Sami Markannen	DUST	Timothy, meadow fescue, tall fescue, Canarian reed grass, red clover, white clover, turnip rape
France	Christophe Chevalier	R	Fescue
Kenya	Abraham Lagat	R	Wheat
Slovakia	Ľubomír Bašta	Data	Red fescue
United Kingdom	Sally Watson	DUST	Perennial ryegrass
United Kingdom	Haidee Philpott	DUST	Oilseed rape
United Kingdom	Tom Christie	DUST	Pea

10. In addition, an expert from Germany said that they currently use SAS for COYD and COYU, though they are likely to move to R in due course. They are developing software in SAS for COYU using splines and then would be interested to compare results.

11. Impressions of the R software were positive, with no problems noted. However problems with software installation delayed the start of the exercise for all participants using DUST. The main cause of these difficulties was related to installing software on secure government networks.

12. A review of the results obtained by the participants is presented in the Annex.

SOFTWARE DEVELOPMENT

13. Software has been developed using R. This is now available as an R library/package or as open-source R code. There is also an evaluation version of DUST that includes a module for the proposed method of COYU. This module calls the R version. It required an extension of the interface and the development of a new installation process to allow for the inclusion of R code.

14. As a result of the practical exercise, several areas for further improvement of the software have been identified, especially in the installation process for the DUST version, the clarity of error messages and dealing with problematic data sets.

15. The algorithm employed effectively carries out separate analyses in each year and then combines these. This avoids use of more complex statistical methods but gives equivalent results in the case of balanced data (see TWC/31/15). The practical exercise has highlighted that sometimes the comparable[1] varieties are not present in all two or three years. We need to verify whether this is of concern and if an alternative algorithm would be needed for such cases.

EXTRAPOLATION

16. Extrapolation is a key issue to be addressed. This is when the level of expression of a candidate variety lies outside that seen in the set of comparable varieties.

17. Extrapolation is a problem both in principle and technically. The General Introduction to the Examination of Distinctness, Uniformity and Stability and the Development of Harmonized Descriptions of New Varieties of Plants (TG/1/3) says:

> *"6.4.2.2.1 For measured characteristics, the acceptable level of variation for the variety should not significantly exceed the level of variation found in comparable varieties already known."*

If the level of expression of the candidate variety is very different from the set of "comparable" varieties, it might be questioned whether these varieties are actually comparable. Technically, the revised COYU

[1] Comparable varieties are known varieties believed to be of a similar type or nature to the new variety in question. This terminology is used in TG/1/3 and TGP/8/3 in preference to the term reference varieties in this context.

criterion is more likely to indicate candidates as being sufficiently uniform when they exhibit extrapolation. This is because the standard error associated with the COYU criterion (TWC/31/15 Corr. para 32) increases as the degree of extrapolation increases.

18. It is clearly not sensible to use the COYU criterion when the degree of extrapolation is high. In these cases, expert judgement should be used, perhaps in combination with the output from the COYU software (including the graphs). For candidates showing levels of expression just outside the comparable varieties, use of COYU may be considered reasonable because the "comparable" varieties are indeed comparable and the COYU method does not give excessive extra benefit from the statistical effects of extrapolation.

19. There is therefore a question as to how much extrapolation is too much. We could consider two scales:

a) Based on the level of expression: the percentage of extrapolation given by the ratio of the difference between the candidate mean level of expression and that of the nearest comparable variety divided by the range of mean levels of expression seen in the comparable varieties.
b) Based on the inflation of the COYU criterion: the ratio of square root of the prediction error (used in the COYU criterion – see TWC/31/15 Corr.) for the candidate over that for the nearest comparable variety (based on the level of expression).

Whilst scale (a) is more natural in many senses, scale (b) is more related to the effect of extrapolation. For that reason, we propose that scale (b) is used as a basis. Scale (b) for a particular candidate depends not only on level of expression (a), but also on the number of comparable varieties and to some extent their distribution. Fewer comparable varieties mean greater effects of extrapolation.

20. Next we need to identify the degree of inflation (scale b) that marks the level beyond which use of COYU is not recommended. This needs to be a balance between giving too much benefit of doubt to candidates and giving rise to too many cases where COYU can't be used. For example, an inflation factor of 120% might be used. Simulations show that this is the inflation level that would approximately correspond to 11% extrapolation with 41 comparable varieties but 7% with 11 comparable varieties.

21. In the Appendix, the number of cases of extrapolation of different degrees (using scale a) are shown for the data sets provided for the Practical Exercise. To date the inflation level (scale b) is not output by the new software.

CONCLUSIONS AND FUTURE WORK

22. The new method works well in practice. The fit of the spline adjustments seem to be fit for purpose.

23. Higher probability levels are likely to be required than for the current method; the example data sets indicate that a reasonable probability level required to match the often used 0.001 rejection level for the current method might be 0.0025 or 0.003. To match with 0.01 acceptance level (after the second year of a three year test), a probability level of 0.02 may be suitable.

24. The Practical Exercise highlighted the need to discuss what action should be taken when the candidate has a level of expression outside that seen in the set of comparable varieties. We suggest that cases of minor extrapolation can safely be ignored, but cases of major extrapolation should be considered by the crop expert. These could be indicated by the software based on the degree of effect on the COYD criterion. Proposals have been outlined as to how this might be achieved. The authors would value input from the TWC in order to take this forward.

25. The software worked well once installed. Key areas for improvement and development have been noted. These include:

- Improving the installation process, particularly for the DUST version;
- Improving error messages and access to these;
- Ensuring that problematic data sets can be dealt with appropriately;
- Improving the extrapolation flag based on the method agreed by the TWC;
- Ensuring that the algorithm works well for unbalanced data.

ACKNOWLEDGEMENTS

26. We are grateful for the essential input of Sally Watson and her AFBI colleagues. We are also very grateful to all those contributing to the Practical Exercise and for their useful suggestions for improvements.

Adrian Roberts and David Nutter
Biomathematics & Statistics Scotland

[Annex follows]

ANNEX

RESULTS OF PRACTICAL EXERCISE

Data sets

1. The table below summarises the data sets considered in the practical exercise:

Country	Crop	Number of sets of years	Number of years for each set	Probability level for COYU	Number of characters	Overall number of candidates
Denmark	Oilseed rape a)	3	2	0.001	15	570
Finland	Timothy	1	2	0.001	6	3
Finland	Timothy	2	3	0.001	1-7	6
Finland	Meadow fescue	1	2	0.001	6	2
Finland	Meadow fescue	1	3	0.001	6	2
Finland	Tall fescue	1	2	0.001	6	1
Finland	Canarian reed grass	1	3	0.001	8	1
Finland	Red clover	2	2	0.001	6	1
Finland	Red clover	2	3	0.001	7	1
Finland	White clover	1	2	0.001	9	1
Finland	White clover	1	3	0.001	9	1
Finland	Turnip rape	1	2	0.001	8	3
Finland	Turnip rape	1	3	0.001	8	1
France	Fescue	1	2	0.001	11	4
Kenya	Wheat	1	2	?	3	2
Slovakia	Red fescue	9	2	0.001	4-7	57
United Kingdom	Perennial ryegrass	2	3	0.001	30	46
United Kingdom	Winter oilseed rape b)	1	2	0.001	12	128
United Kingdom	Pea c)	5	2	0.001	13-19	47

a) Winter/spring by lines/hybrids
b) Lines/restored hybrids/conventional/hybrids/composite
c) Semi-leafless/conventional

Fit of splines to data

2. The principal change in the improved COYU method is the use of splines instead of the moving average method to map the relationship between uniformity and level of expression. This method also restricts the flexibility of the spline (see TWC/31/15 corr.). To assess the success of this, participants were asked to review the plots output by the new software. A number of these were also reviewed by the author. It was seen that the curves fitted the data adequately, without any tendency to over-fit. An example plot is shown in Figure A1.

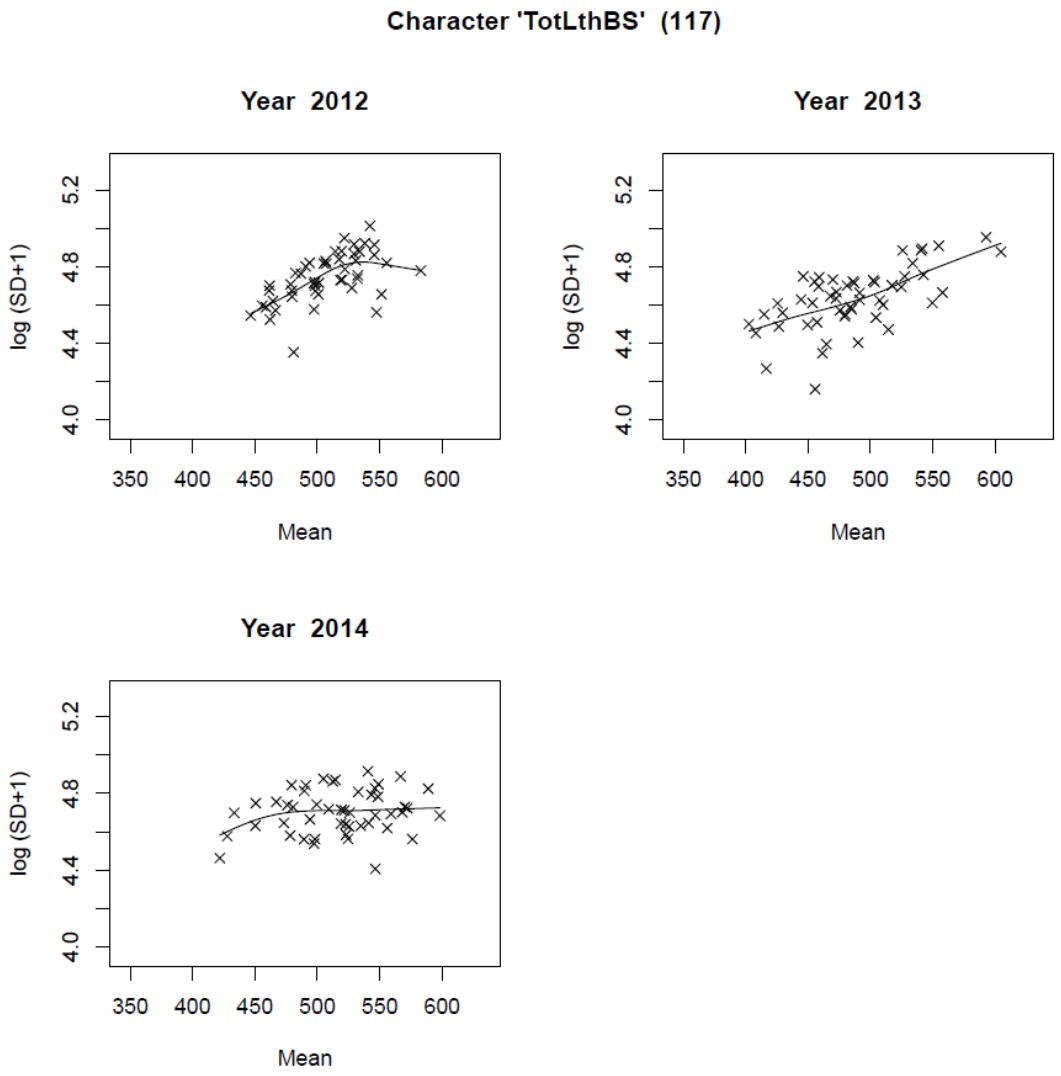


Figure A1: *Spline fits of uniformity to level of expression for characteristic 117 Perennial Ryegrass Tetraploid Intermediates – data from the United Kingdom*

Matching probability levels between the two COYU methods

3. All participants use the probability level of 0.001 with the current COYU method. The new method is expected to require the use of a higher probability level to achieve the same level of decision-making (TWC/31/15 corr.). In principle, it might be possible to assess what probability level might be required by looking at how many varieties would be found non-uniform with differing levels of probability. In practice, few varieties actually fail the COYU criterion; this means that a comparison of rejection rates would only give a very coarse idea of the probability level required for the new method. Instead, it is better to calculate and compare p-values for each candidate.

4. In the figures below, we compare the p-values for the current and new methods of COYU. The plots on the left hand side show all the candidates – the plots on the right hand side show only those results where the p-value for the current method is less than 0.01. Curves are fitted onto these graphs to give some idea of trend. Candidate varieties that show extrapolation were omitted (see below). Note that the plots typically cover the p-values for several characteristics and years.

Figure A2: *Comparison of p-values for Denmark oilseed rape data sets; solid red line is fitted curve, dashed blue line shows equality between the p-values*

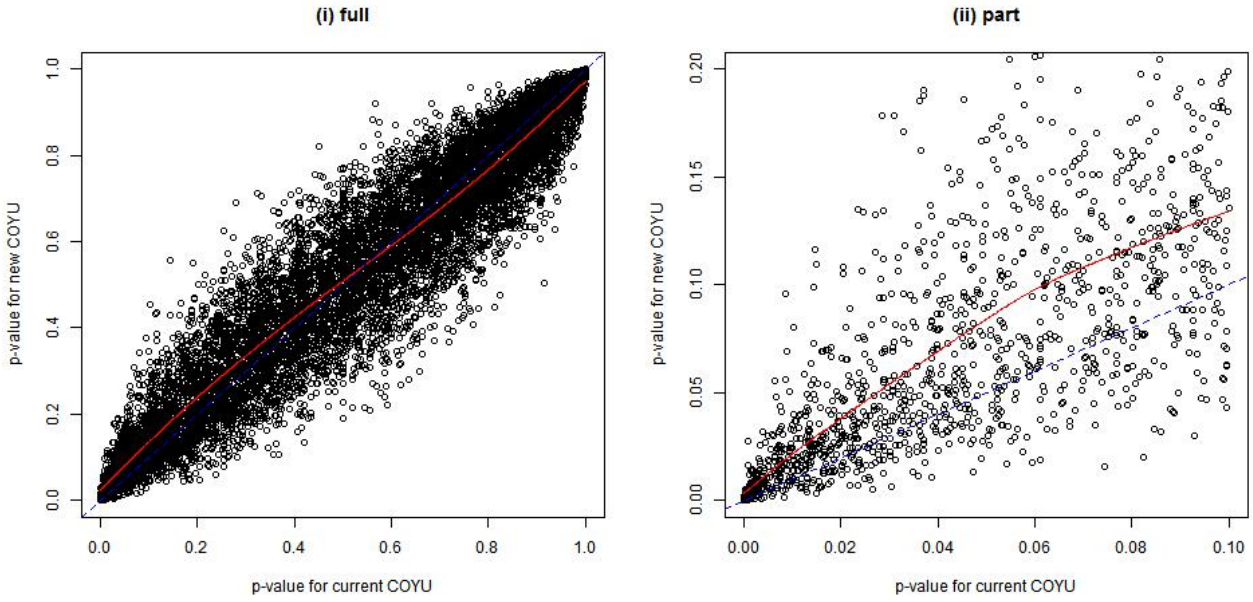


Figure A3: *Comparison of p-values for Finland data sets; solid red line is fitted curve, dashed line shows equality between the p-values*

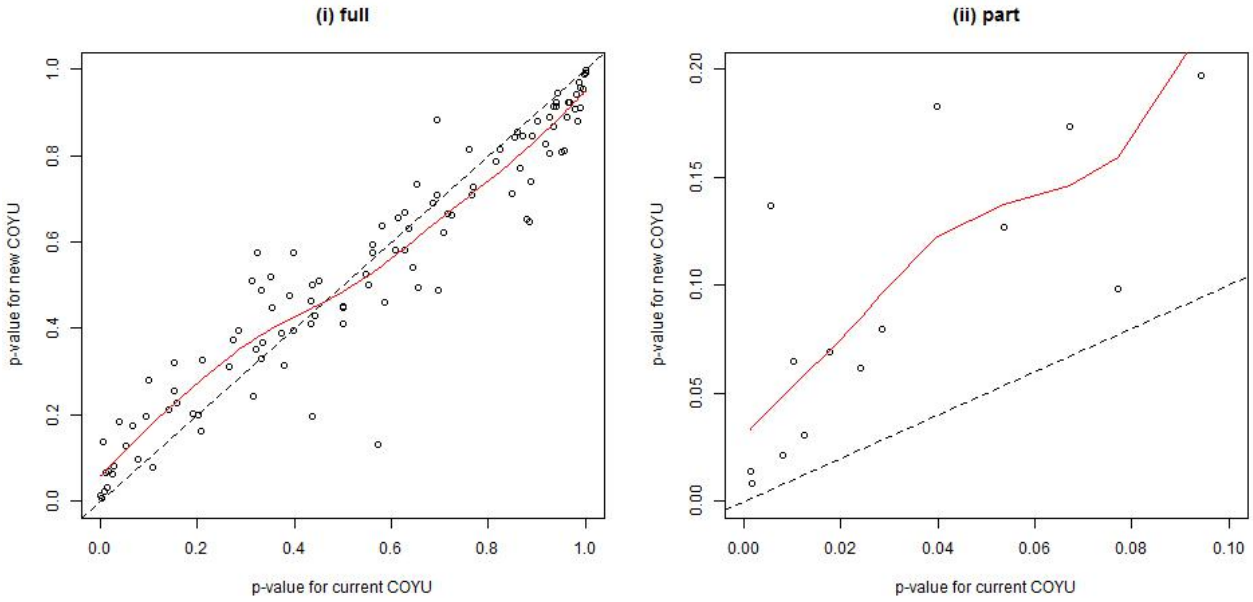


Figure A4: *Comparison of p-values for France data set; solid red line is fitted curve, dashed line shows equality between the p-values*

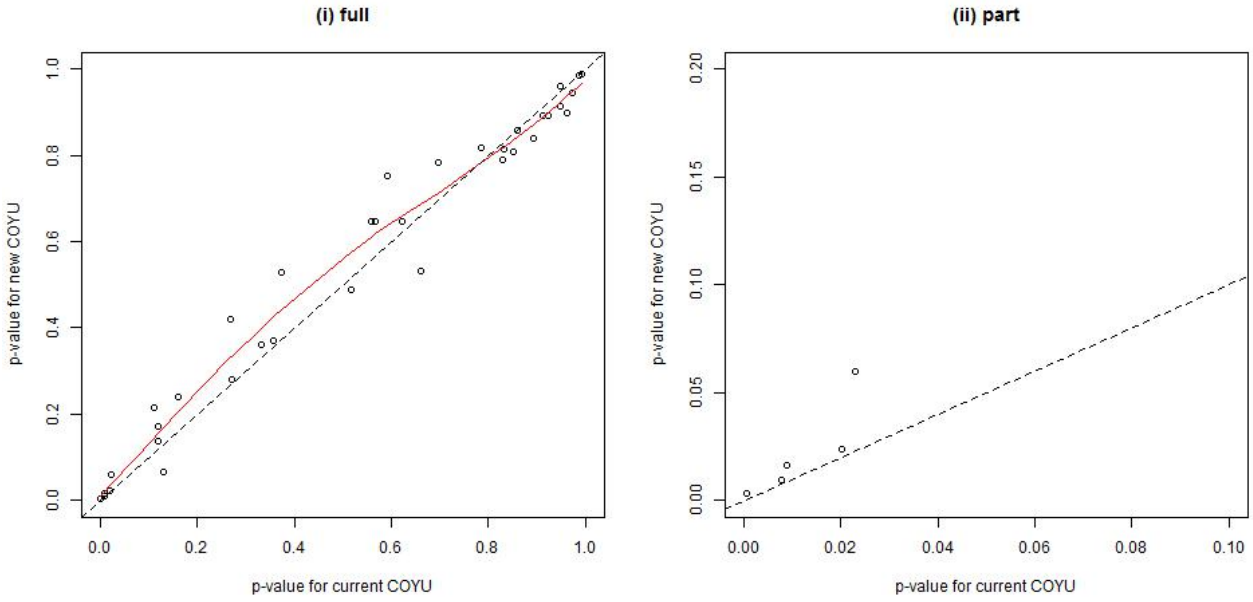


Figure A5: *Comparison of p-values for Kenya data set; (fitted curve omitted) dashed line shows equality between the p-values*

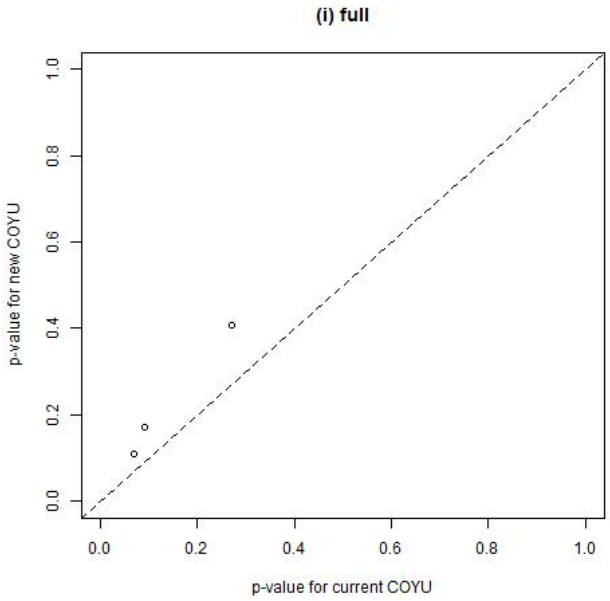


Figure A6: *Comparison of p-values for Slovakia red fescue data sets; solid red line is fitted curve, dashed line shows equality between the p-values*

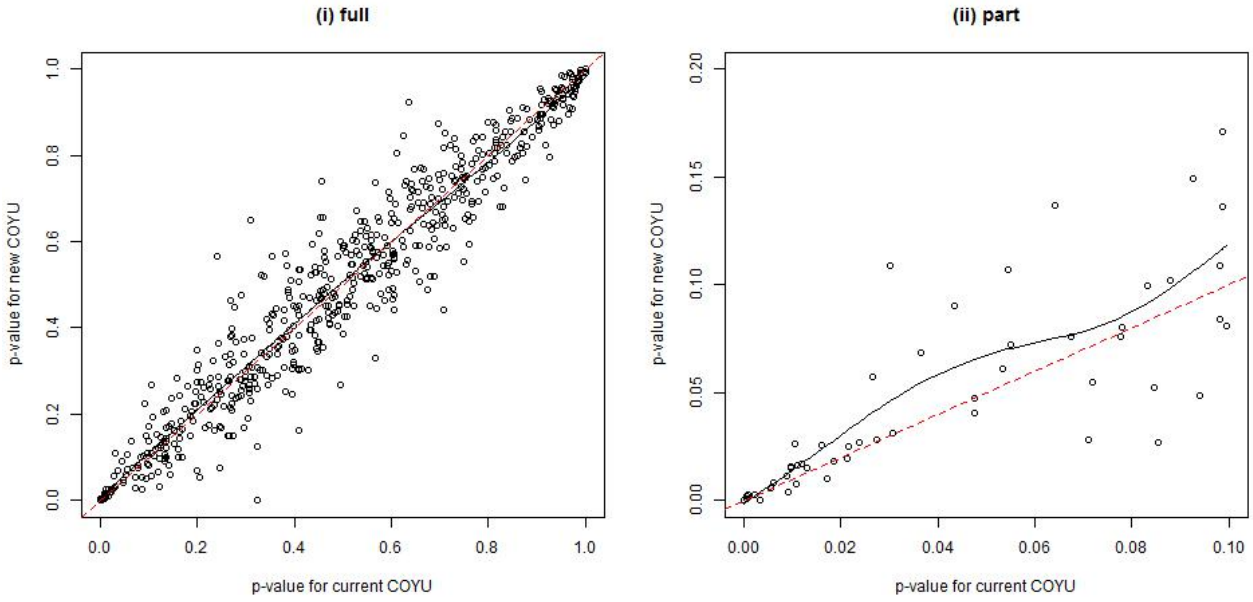


Figure A7: *Comparison of p-values for the United Kingdom perennial ryegrass data set; solid red line is fitted curve, dashed line shows equality between the p-values*

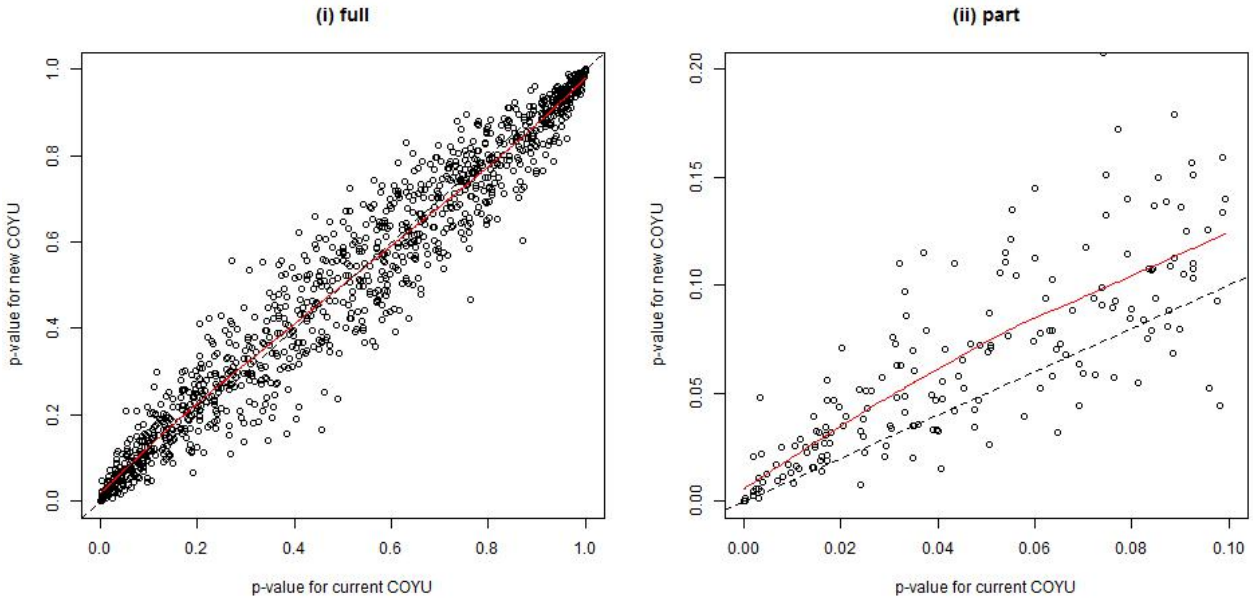


Figure A8: *Comparison of p-values for the United Kingdom winter oilseed rape data set; solid red line is fitted curve, dashed line shows equality between the p-values*

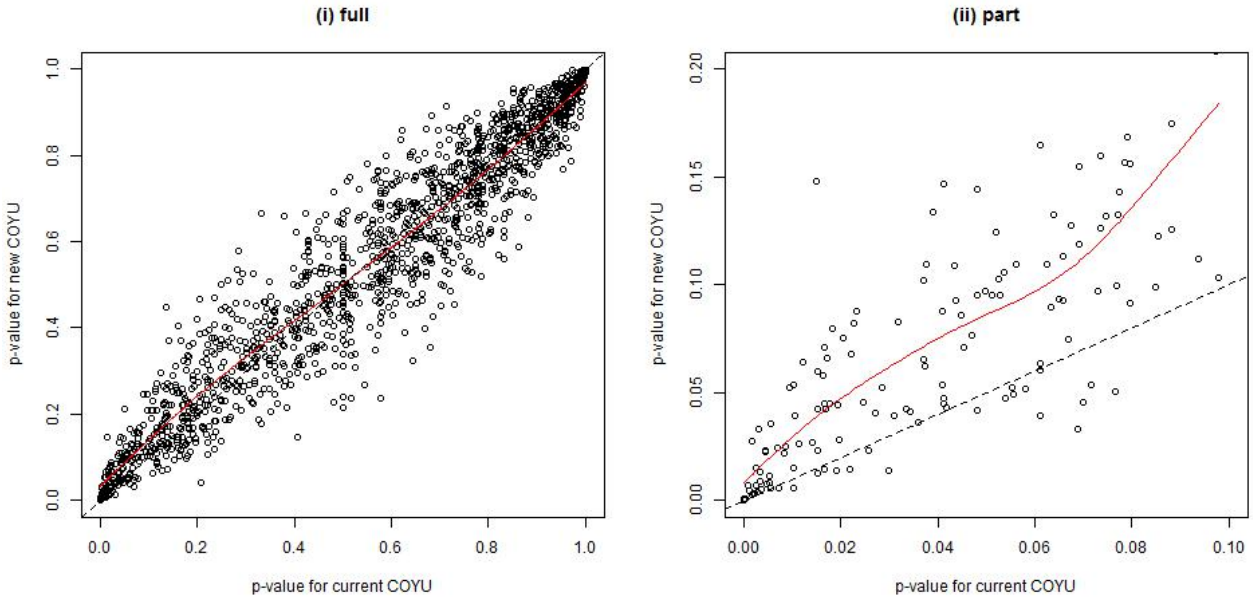
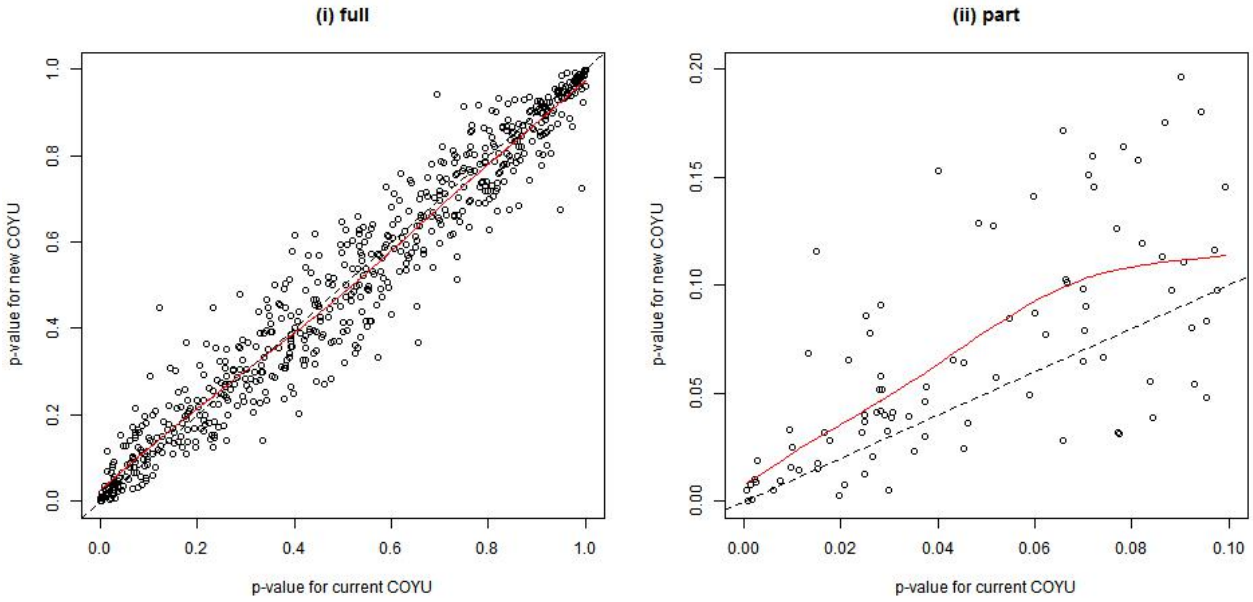


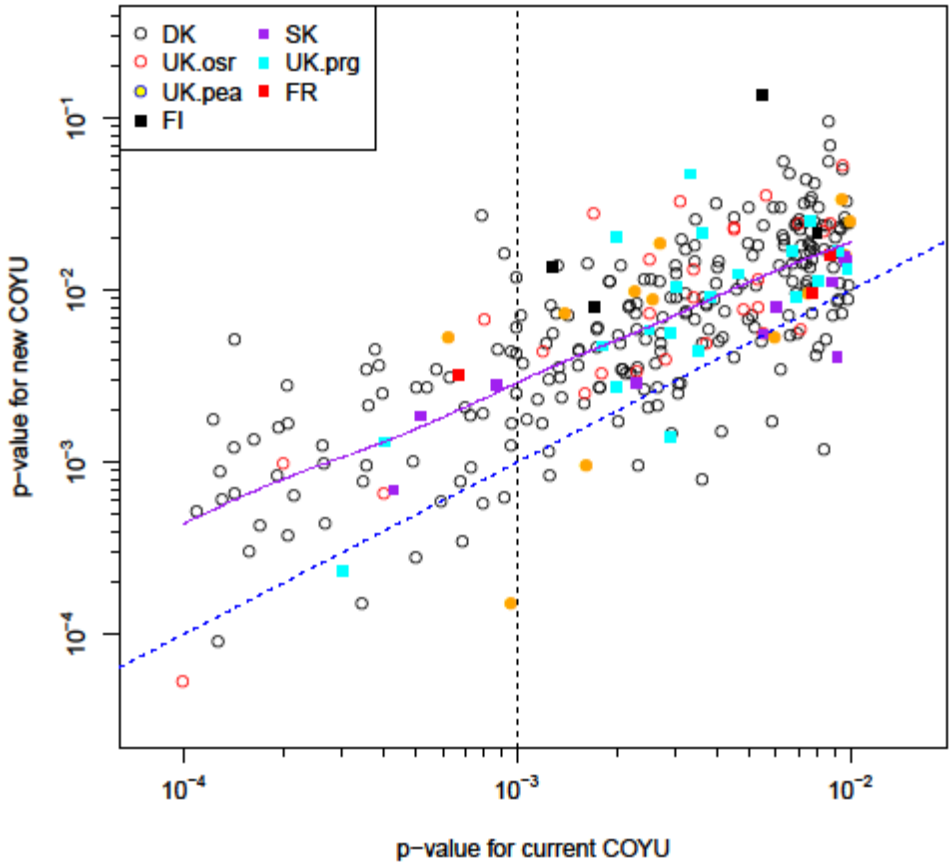
Figure A9: *Comparison of p-values for the United Kingdom pea data sets; solid red line is fitted curve, dashed line shows equality between the p-values*



5. The general pattern was similar across the different data sets, with a high degree of correlation between the two methods but the new method having slightly higher p-values for less uniform varieties.

6. Figure A10 compares the p-values when the COYU p-value for the current method is between 0.0001 and 0.01. The axes are log-transformed.

Figure A10: *Comparison of p-values for all data sets restricted to current COYU method p-values lying between 0.0001 and 0.01; solid purple line is fitted curve, dashed blue line shows equality between the p-values*



7. The relationship between the p-values for the two methods was examined using additive models (linear models with spline curves for covariates). The p-values for the proposed COY method were regressed on those for the current method, after log-transforming both. These models were used to indicate "equivalent" p-values for the proposed method when the p-value for the current method was set at 0.01 (commonly used for second year acceptance in a three-year test) and 0.001 (commonly used for rejection in both two and three year tests).

8. The table below gives spline-based predictions of the equivalent p-value for the new COYU compared to a p-value of **0.001** for the current method. Predictions were based on the log-log additive models, for all data sets and based on separate curves for each data set (in this case, parallel curves fitted best). The models were fitted on only those cases where the COYU p-value for the current method is between 0.00001 and 0.1 and excludes cases of extrapolation. There was some apparent variation between data sets, but from this (and bearing in mind sample sizes); it seems that a p-value of 0.0025 or 0.003 may work for the new method to give broadly equivalent results to the current method.

Data set	Number of cases	Equivalent p-value	95% confidence interval	
			lower	upper
Overall	**1756**	**0.00288**	**0.00258**	**0.00322**
Denmark	1269	0.00284	0.00255	0.00317
Finland	15	0.00629	0.00450	0.00879
France	5	0.00293	0.00167	0.00514
Kenya	2	0.00372	0.00154	0.00898
Slovakia	51	0.00223	0.00182	0.00274
GB oilseed	137	0.00339	0.00293	0.00394
GB pea	92	0.00256	0.00217	0.00303
GB ryegrass	185	0.00273	0.00237	0.00315

9. The table below gives spline-based predictions of the equivalent p-value for the new COYU compared to a p-value of **0.01** for the current method. Predictions were based on the log-log additive models, for all data sets and based on separate curves for each data set (again, parallel curves fitted best). The models were fitted on only those cases where the COYU p-value for the current method is between 0.001 and 0.2 and excludes cases of extrapolation. There was some variation between data sets, but from this (and bearing in mind sample sizes); it seems that a p-value of 0.02 may work for the new method to give broadly equivalent results to the current method.

Data set	Number of cases	Equivalent p-value	95% confidence interval	
			lower	upper
Overall	**1674**	**0.0188**	**0.0179**	**0.0198**
Denmark	1200	0.0186	0.0177	0.0196
Finland	15	0.0410	0.0302	0.0557
France	4	0.0171	0.0095	0.0309
Kenya	2	0.0245	0.0106	0.0566
Slovakia	47	0.0141	0.0119	0.0169
GB oilseed	133	0.0226	0.0203	0.0251
GB pea	90	0.0171	0.0150	0.0195
GB ryegrass	183	0.0180	0.0164	0.0198

Extrapolation

10. Both the current and proposed new methods for COYU use adjustments based on fitting a curve to the relationship between uniformity (represented by the log of the standard deviation plus 1) and the level of expression (represented by the mean) over the comparable varieties. This curve is used to adjust uniformity data for both the comparable varieties and candidates. As noted in TWC/31/15 corr., there is an issue if a candidate exhibits a level of expression outside the range seen in the comparable varieties; this is extrapolation. This issue needs careful consideration and it was an aim of this Practical Exercise to evaluate the frequency of extrapolation cases in practice.

11. The effect of extrapolation is different for the two versions of COYU, current and new. Overall the proposed method is more likely to indicate such a candidate variety as uniform – it gives it the benefit of doubt. However with both versions, it is better that such cases are evaluated apart and with care by the crop experts. To this end, the new software indicates cases of extrapolation that would require expert review. A future version may indicate the extent of extrapolation.

12. The table below indicates the frequency of extrapolation cases and, in some cases, the extent of extrapolation (scale a)) in relation to the range of expression in the comparable varieties (in other cases, these figures were not available). Note that some example data sets had few varieties so these provide only a rough indication of more general levels of extrapolation.

Country	Data set	Number of cases	Frequency of extrapolation	Cases > 10% extrapolation	Cases > 20% extrapolation
Denmark	Oilseed rape	11,910	2%	0.8%	0.3%
France	Fescue	36	0%	n/a	n/a
Kenya	Wheat	6	50%	n/a	n/a
Finland	Various	137	19%	n/a	n/a
GB	Perennial ryegrass	1,381	13%	7%	4%
GB	Winter oilseed rape	1,536	1%	0.2%	0%
GB	Pea	698	8%	4%	2%
Slovakia	Red fescue	738	20%	15%	7%

13. The large number of extrapolation cases for the United Kingdom perennial ryegrass (tetraploid) data set was investigated further. Much of this was due to a single candidate, which was very different to the comparable varieties. Most of the remaining large extrapolations in this data set were due to two more candidates in one character.

[End of Annex and of document]