



UPOV/DATA/BEI/04/5

ORIGINAL: English

DATE: May 25, 2004

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

WORKSHOP ON DATA HANDLING

organized by
the International Union for the Protection of
New Varieties of Plants (UPOV)

in cooperation with
the State Forestry Administration of China,
the Ministry of Agriculture of China and
the State Intellectual Property Office of China

with the financial assistance of
the Ministry of Agriculture, Forestry and Fisheries of Japan

Beijing, June 9 to 11, 2004

OUTLIER DETECTION AND DATA VALIDATION

Document prepared by an expert from Denmark

OUTLIER DETECTION AND DATA VALIDATION

Erik A. Lawaetz
Department of Variety Testing
Danish Institute of Agricultural Sciences

1. Introduction

Outlier detection is a constant search for observations that contribute great variation in the data recorded from the examination of Distinctness, Uniformity and Stability (“the DUS test”). Data validation is undertaken throughout the season. The speed of the process is very important to enable a revisit to the field at the same plant growth stage and, if possible, under the same environmental conditions. Consideration of outliers and data validation should be done in compliance with the principles for the assessment of uniformity laid down in the General Introduction (TG/1/3). The crop expert needs to be aware that disregarding observations will alter the results of the test for distinctness and uniformity. Computers and statistical programs are essential tools for the crop expert in the process of outlier detection and data validation of large data collections. The purpose of this paper is to provide guidance to the crop expert on how to carry out the actual outlier detection and data validation in DUS testing.

1.1 Definitions

In the General Introduction, TG/1/3, the term off-type is clarified as follows:

“CHAPTER 6 “EXAMINING UNIFORMITY”

6.4.1 Self-Pollinated and Vegetatively Propagated Varieties

6.4.1.1 *Determination of Off-Types by Visual Assessment*

A plant is to be considered an off-type if it can be clearly distinguished from the variety in the expression of any characteristic of the whole or part of the plant that is used in the testing of distinctness, taking into consideration the particular features of its propagation. This definition makes it clear that, in the assessment of uniformity, the standard for distinctness between off-types and a candidate variety is the same as for distinctness between a candidate variety and other varieties.”

It also explains that:

“6.5 Unrelated and Very Atypical Plants

The test material may contain plants that are very atypical or unrelated to those of the variety. These are not necessarily treated as off-types, or part of the variety, and may be disregarded, and the test may be continued, as long as the removal of these very atypical or unrelated plants does not result in an insufficient number of suitable plants for the examination, or make the examination impractical. In choosing the term “may be disregarded,” UPOV makes it clear that it will depend on the judgment of the crop expert.

In practice, in tests conducted with a small number of plants, just one single plant could interfere with the test, and therefore should not be disregarded.”

1.2 Data material recorded

Different methods can be used to observe the characteristics in the DUS test, according to the type of variety and characteristic being examined

- VG Visual assessment by a single observation of a group of plants or parts of plants
- MG Single measurement of a group of plants or parts of plants
- VS Visual assessment by observation of individual plants or parts of plants
- MS Measurement of a number of individual plants or parts of plants

The different methods of observing characteristics accumulate different quantities of data and demand different methods to detect recording/typing errors or systematical/methodical observation errors. This paper demonstrates a *spreadsheet* method and a *graphical* method to detect outliers and to perform data validation.

2. Spreadsheet Method

The spreadsheet method is a simple and efficient way to present data, especially for VG and MG observations. The recorded data can be sorted in order of variety code, plot number and 1st and 2nd replication. The difference between the maximum and minimum recordings (minmax) and average for the variety are used for finding recording errors between replications and possible recordings transposed between plots.

2.1 Large “minmax”

The following characteristics are taken from the UPOV Test Guidelines TG/19/10 for barley:

Char. 3: ‘Flag leaf: anthocyanin coloration of auricles’ and,

Char. 4: ‘Flag leaf: intensity of anthocyanin coloration of auricles’

These are two difficult characteristics which the crop expert must validate carefully and quickly after the assessment in the year of testing and over the years of testing.

The spreadsheets in Figures 1 and 2 present the same set of five varieties sorted in order of variety code. Plot number in 1st and 2nd replication, expression of characteristic in 1st and 2nd replication, average of expression and minmax are given in the Figures. The minmax column on the right side of the spreadsheets is quickly calculated and large differences can be detected.

Variety-code	type	1	2	1	2	Average recording	MINMAX
		replication plot	replication plot	replication recording	replication recording		
18148	31	85	300	9	9	9	0
18149	31	87	304	99	9	54	90
18150	36	86	301	9	9	9	0
18151	31	88	303	9	1	5	8
18152	31	84	302	1	1	1	0

Figure 1: Spring Barley. Flag leaf: anthocyanin coloration of auricles

In Figure 1, the variety entry '18149' has a large minmax because of an obvious typing error (99 instead of 9) in plot 87. This type of recording error does not require any revisit to the field. When minmax equals 2 or more the crop expert should consider a revisit to the field for re-examination and to determine if the high minmax is caused by a recording error, environmental conditions or some other factor. Variety entry '18151' in Figure 1 and variety entry '18150' in Figure 2 have large differences between replications, which require a revisit to the field in order to validate the assessments. However, the minmax for variety entry '18151' in Figure 2, on the other hand, does not make it obvious (minmax = 0) to revisit the field. This example shows how important it is that the validation is performed for each characteristic as soon as possible after it is recorded in the field. For VG characteristics like 'Flag leaf: anthocyanin coloration of auricles', where the expression is either 'absent' or 'present', validation over years, in addition to that within the year, is necessary to determine the expression of the variety for the process of examining distinctness and for the final description of the variety.

Variety-code	type	1	2	1	2	Average recording	MINMAX
		replication plot	replication plot	replication recording	replication recording		
18148	31	85	300	7	5	6	2
18149	31	87	304	2	3	3	1
18150	36	86	301	4	8	6	4
18151	31	88	303	2	-	2	0
18152	31	84	302	-	-	-	-

Figure 2: Spring Barley. Flag leaf: intensity of anthocyanin coloration of auricles

2.2 Transposition of recordings

By sorting the spreadsheet by replication and plot number, possible displacement of observations can be detected. For example, if two or more adjacent plots had high minmax observations this might be due to transposition and revisiting the field would be necessary.

Variety code	type	1	2	1	2	Average recording	MINMAX
		replication plot	replication plot	replication recording	replication recording		
18148	31	85	300	7	5	6	2
18150	36	86	301	4	8	6	4
18152	31	84	302	-	-	-	-
18151	31	88	303	2	-	2	0
18149	31	87	304	2	3	3	1

Figure 3: Spring Barley. Flag leaf: intensity of anthocyanin coloration of auricles

When sorting the spreadsheet in Figure 2 by plot number in 2 replication instead of variety entry code order, the result will be the spreadsheet in Figure 3. The high minmax values for the varieties entries '18148' and '18150' are now next to each other and a possible transposition of recordings could be the cause. If the recordings in plots 300 and 301 are reversed the minmax will change to 1 for both varieties. Transposition of observations is usually due to human error at the time of recording, or when the data is transferred from written to electronic records.

2.3 Missing values

Missing values will be recorded where plots are discarded, expressions of characteristics cannot be recorded and where plant or parts of plants are missing. It is important that the missing values are recorded as missing values and not as a zero. If a missing value is recorded as a zero by mistake, it will corrupt the results of the test for distinctness and uniformity. The extent of the corruption will depend on the methods by which the characteristic is observed.

Data validation is a useful system for the crop expert to check the quality of the work throughout the growing season.

3. Graphical method

The graphical method is a good method for presenting large quantities of data. For example, the DUS trial for rape seed normally includes many varieties with 2 or 3 replications. As the basis for the DUS testing of rape seed, Test Guidelines TG/36/6 are used. In rape seed, many characteristics are observed on single plants or parts of plants (VS or MS); characteristic no. 8 'Leaf: length' recorded at stage 23-27 is given as an example. Figures 4-6 show the same data for the characteristic 'leaf length' in three different ways.

The data in Figure 4 shows a trend for the 'leaf length' to increase from plot 1 to 426. In this case, a systematical/methodical-harvesting trend occurs because it is not physically possible to harvest and record the leaf characteristics of the 426 plots in one day. To treat all varieties equally and minimize the effect of plant growth, each replication is harvested and recorded separately and if possible within one day. The average 'leaf length' in replications 1, 2 and 3 were 243, 254 and 284 mm. The minmax recorded observations within the plots are shown graphically in Figure 5. In the system used here you can click on the observation and obtain the variety code and at the same time the particular observation will be indicated in the other figures as well.

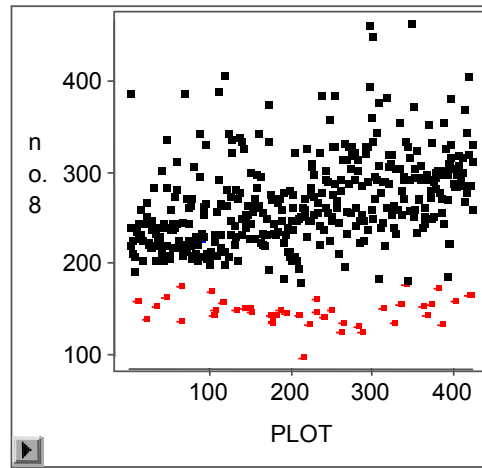


Figure 4: Plot means vs. Plot no.

Figure 5 shows quite a few plots where the minmax is around 150 mm. or above - all these plots should be re-examined for possible recording errors. Re-examination of observations having large minmax for the characteristic is not just a search for outliers but also for 'off-types'. For example, re-examination of the observations for the variety entry '14907' did not show any obvious outliers in the plot and the cross check with other characteristics observed on the same part of plants did not indicate any errors. The conclusion of variety entry 14907 is that there is no correction of recordings to be made in this plot (89).

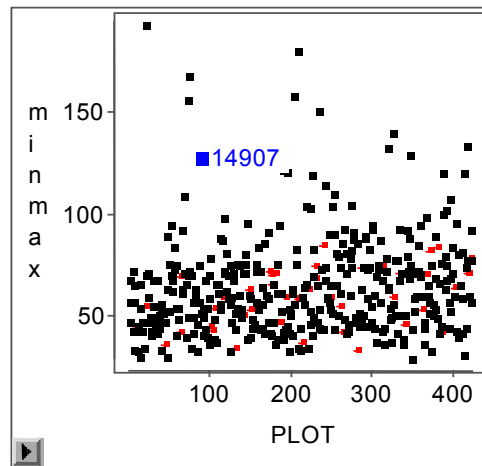


Figure 5: Minmax vs. Plot no.

Figure 6 shows the plot of the residual versus the predicted values of 'leaf length'. This is a way to check the heterogeneity of the variance and deviating plots (caused by e.g. sowing error, transposition of recordings or damage to the plot). The data should be evenly distributed around zero. In cases where the variance increases with the mean, the observations will fall approximately in a funnel with the narrow end pointing to the left. Outlying observations, which may be errors, will be shown in such a figure as observations that clearly have escaped from the horizontal band formed by most of the others. In the actual data, the following is apparent: besides the systematical/methodical-harvesting trend observed in Figure 4, the Figure shows 40 plots (red dots) where the 'leaf length' is considerably shorter than in the rest of the plots. In Figure 6, the variation of these 40 plots are clearly different from the rest of the plots.

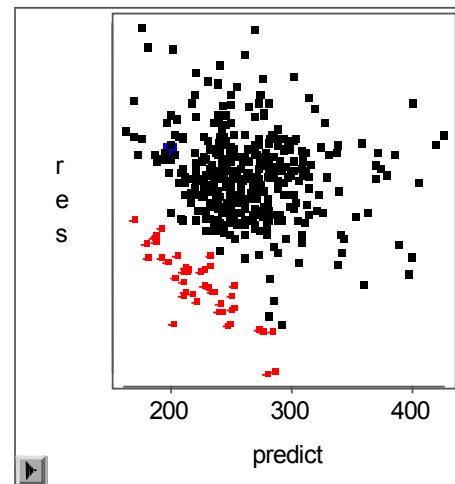


Figure 6: Residuals vs. predicted

The reason for this is that the recorder misunderstood how to record the characteristic ‘Leaf: length’. In the Test Guidelines ‘Leaf: length’ is described as length of blade and petiole, but the recorder only measured the ‘blade’ without the ‘petiole’. This example clearly shows the importance for the crop expert to instruct and explain to the staff in detail how to record each characteristic according to the Test Guidelines.

For rape seed, many characteristics are observed on single plants or parts of plants (VS or MS) as for example UPOV characteristic no. 16 ‘Plant: height at full flowering’ as shown in Figures 7-9. The data are from DUS trial of winter oilseed rape with 426 varieties in 3 replications. Figure 7 shows the ‘Plant: height at full flowering’ mean per plot and the plot numbers. The mean height in the collection is 137 cm, with two distinct short varieties (blue dots) under 80 cm height.

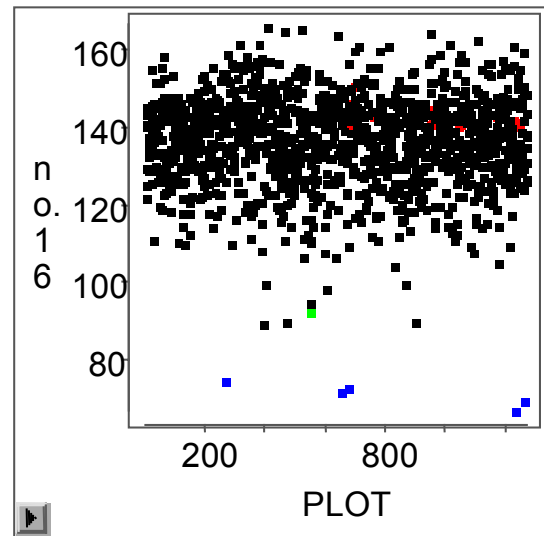


Figure 7: “Plant: height at full flowering

Figure 8 shows the minmax (within the plot) versus plot number. The minmax is the difference between min and max of the observations within the plot and not between replications. For most of the plots minmax is between 10-20 cm, but for 9 plots (red dots) outlier observations are detected. The observations of these plots demand a re-examination. The re-examination showed that one plant in each plot was measured to be 100 cm shorter than the rest of the plants in the plot, which strongly indicates that the staff person has missed the ‘1’ key on the data logger during the recording. These 9 outlier observations will be disregarded and not influence the test final tests for Distinctness and Uniformity. Some data loggers can be programmed with a limited range of the recordings for each characteristic to be accepted.

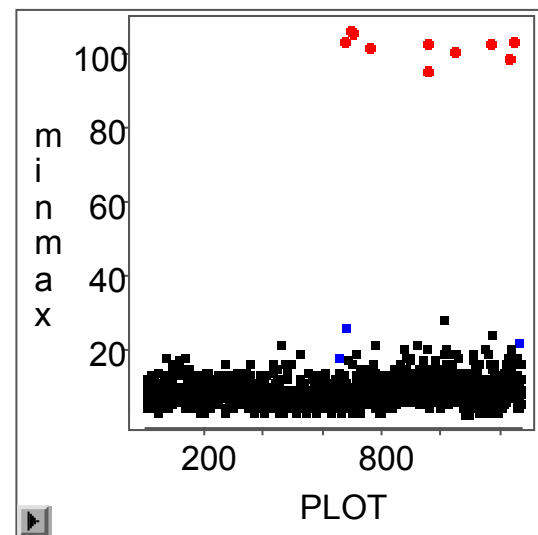


Figure 8: Minmax vs. Plot no.

For example, a limited range for 'Plant: length' in Figure 7, between 60 and 200 cm. would have notified and stopped the recorder in the typing mistakes (red dots in Figure 8) in the field. Figure 9 shows the residual versus predicted values of residual 'Plant: height at full flowering'. The heterogeneity of the variation of the varieties is satisfactory with an even distribution around zero. The two short varieties clearly differ in height from the rest of the varieties, but still within the acceptable range of variation around zero.

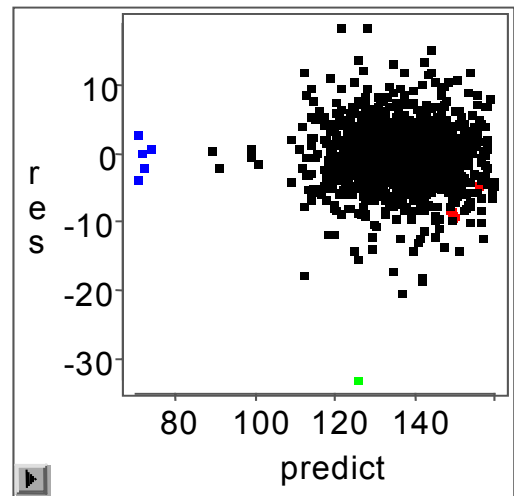


Figure 9: Residuals vs. predicted

4. Conclusion

1. All details concerning the trial need to be organized before the actual data validation
2. Outlier detection and data validation are important procedures in DUS testing and should be performed during the season in cooperation with the crop expert. It is essential to be able to revisit the field at the same plant stage to allow correction of data.
3. It is recommended to re-examine the plots to establish if other characteristics within the same plot differ. When characteristics of plants or parts of plants are observed and an off-type has been identified, it should be checked if that plant or part of plant is also off-type for other characteristics too.
4. Be aware that correction of data which alters the variation between replications or within the plots will affect the test for distinctness and uniformity.

[End of document]