Herd-Test-Date Clustering

T. Strabel
August Cieszkowski Agricultural University, Department of Genetics and Animal Breeding
Wolynska 33, 61-627 Poznañ, Poland; e-mail strabel@au.poznan.pl.

Introduction

Genetic evaluation of dairy cattle is based on the comparison of animal performances observed under similar environmental conditions. In traditional evaluations records are usually grouped in herd-year-season (HYS) classes. This grouping is replaced by the herd-test-date (HTD) classes in test-day (TD) models. Both these classifications are based on two sources of variation: herd - which represents mainly the level of production and time which specifies the environmental condition in a particular period of the year. In populations with small herd size, such as in Central Europe, records with no contemporaries appear when the typical fixed classes are used. The accuracy of genetic evaluations could be improved by increasing the number of records in contemporary groups (CG) (Schmitz et al 1991, Tosh and Wilton 1994).

In this paper, the CLUSTER program is presented that groups htd classes. The algorithm is based on a clustering procedure.

Objective

The aim of the program is to decrease the number of small CG. HTD classes are compared on the basis of mean herd production and date of data recording. Classes are joined if additional conditions are fulfilled, for instance CG with a big number of records are not joined with other big CG.

Computing methods

To find the most similar pairs of observations a typical cluster procedure (like The Cluster procedure in SAS) works on the distance matrix (Hartigan 1975) which consists of the distances between all observations. The size of this matrix is equal to the number of records. It is easy to meet time and computer storage limits when the number of observations is high. The CLUSTER program finds the smallest distances without calculating all the possible distances for instance between classes from herds which significantly differ in production level or if the observations in two classes were recorded in different seasons. The clustering process could be controlled by additional parameters.

The input file consists of HTD number, mean herd production, and date at which records were registered expressed in days.

Parameters

Levels of the HTD effect are combined according to two criteria: mean herd production and the date at which test day records were collected. The main aim of the program is to join small contemporary groups so that the size of the group (number of TD records it contains) is also taken into account.

The following parameters are used to control the clustering:

- maximum distance in days between groups - *maxd* - a group could be combined with another if the other one was recorded within the specified number of days,
- maximum distance in mean herd production - *maxs* - contemporary groups from two herds could be joined if the difference in mean herd production does not exceed the maximum specified,
- maximum size of one of the CG could not be bigger than the minimum specified - *maxw*; this is to avoid the clustering of two big CG.

To allow computing distances when the criteria used are measured in different units (days and kg) data standardization was carried out:

- mean herd production was divided by const_1
- date at which performance test was carried out (expressed in days) was also divided by a const_2

Because the values of those additional parameters could influence the preferences of clustering they could also be set by the user.

Ordering of the data

In order to calculate only those distances which could be accepted for clustering, all CG are sorted according to dates and two times divided into parts - strips. The width of the strips is two times bigger than the *maxd*. The difference of the borders of first and second pattern of strips is equal to *maxd*. In this way, computing distances within the strips of all the possible distances which are not bigger than *maxd* could be found. Within the strips all records are sorted by the standardized mean herd production, which increases the speed of calculations.

Scanning

Beginning with the first observation in the strip, quadratic Euclidean distances were computed. If the distance between two CG in the strip is bigger than *maxs*, the program starts to calculate distances between the second CG and all the remaining ones. This step is repeated as many times as many observations there are in the strip. If one of the groups is smaller than *maxw* the calculated distance is put into a separate distance file.

Clustering

The smallest distance is taken from the distances file and two CG are combined into a bigger one. Weighted mean of mean herd production and date are calculated. The size of the new CG is a sum of the sizes of CG which were combined. The distances between the joined CG and all the other ones are discarded from the file of distances. The new CG is put into the data file in the right strips. Then scanning is performed for the new CG.

The time of the scanning depends mainly on the width of the strips. This parameter decides how many potential calculations have to be done. This process could be speeded up by shortening the width of the strips. In such a case, it is not necessary to consider distances between two CG in standardized days which are bigger than half of the strip width. In this way scanning could be repeated several times and the number of calculations is reduced. During the last scanning the width of the strips has to be two times bigger than the value of the *maxd* parameter.

It is easy to modify the program to influence the clustering process by:

- excluding the possibility of joining two CG from the same herd,
- accepting only the clustering of CG from different herds,
- requiring both CG to have size not bigger than maxw.

Using such modifications and setting different values of parameters, different results of clustering could be achieved.

Computing environment

The program is written in the Clipper language and runs under DOS/Windows systems. Free disk space required is about five-ten times bigger than the size of the input file. Additional RAM memory resources could speed up the program if the system works with a cache (Windows 95) or if files are stored in the RAM disk. Time of clustering procedure depends on parameters used and the size of the data set. Ten hours is enough for clustering 50 000 HTD groups when the program is run on the PC-Pentium machine.

AN EXAMPLE

Material

Data comprised 208 622 TD records of 23 088 lactations of Black and White heifers calving from August 1992 to July 1995 in 543 herds leading to 19905 herd-test-day levels.

Methods

For the analysis of TD the following mixed model was applied:

$$y_{ijk} = HTD_i + \sum_{l=1}^{6} b_l X_{lijk} + a_j + pe_j + e_{ijk}$$

where

y_{ijk} is a TD observation

HTD_i is a herd-test-day effect

b_{1-6} are regression coefficients

X_1 = DIM/305

X_2 = $(DIM/305)^2$

X_3 = ln(305/DIM)

X_4 = $(ln(305/DIM))^2$

X_5 = AGET

X_6 = $AGET^2$

where AGET are months of age at test day.

As a simple alternative to the proposed clustering methods herd-level-month of test contemporary groups were set using a 100 kg herd mean production step. For comparison, this model was indicated as HLMT.

The sets of parameters used during clustering are presented in Table 1.

Table 1. Description of models with clustered HTD.

Model	*const_1*	*const_2*	*maxd*	*maxs*	*maxw*	options
HTDR100	365	10000	65	100	1	*maxw*
HTDRR1	365	10000	65	100	3	*maxw / maxw*
HTDR1S	365	10000	65	-	3	*maxw ; clust. within herds*

Results

The ratio of residual to total variance, mean standard error of prediction and correlation between true and estimated breeding values for analyzed models as well as the number of levels of fixed effect and number of effective observations are presented in Table 2.

Table 2. The ratio of residual to total variance, mean standard error of prediction and correlation between true and estimated breeding values for analyzed models.

Model	σ_e^2 / σ_p^2	mean standard error of prediction	mean correlation between true and estimated breeding values	Number of levels of fixed effect	Number of effective observations
HTD	0,48	1,5276	0,6243	19905	467,56
HLMT	0,52	1,5637	0,5823	1995	577,34
HTDR100	0,48	1,5197	0,6269	17983	473,51
HTDRR1	0,48	1,5262	0,4396	17198	577,11
HTDR1S	0.48	1,8772	0,6240	16462	469,34

An analysis of the results shows that:

- the change in effective observations does not follow the pattern of changes in mean standard error of prediction;
- when clustering only classes from the same herd (HTDR1S) the mean standard error of prediction increases but there is no change in mean correlation between true and estimated breeding values when compared to nonclustered HTD classification.

Further evaluations of data sets with HTD clustered with different values of parameters: *maxd*, *maxs*, *maxw*, *const_1* and *const_2* will be carried out to investigate the influence of clustering on the accuracy of animal evaluations.

References

Hartigan, J.A. 1975. *Clustering algorithms.* John Wiley & Sons, New York.

Schmitz, F., Everett, R.W. and Quass, R.L. 1991 Herd-year-season clustering. *J. Dairy Sci. 74,* 629-636.

Tosh, J.J. and Wilton, J.W. 1994 Effect of data structure on variance of prediction error and accuracy of genetic evaluation. *J. Anim. Sci. 72,* 2568-2577.