

**A Study on Heterogeneity of Variances Adjustment in Genetic Evaluations in Spain**

**Ibáñez, M., Carabaño, M.J. and Alenda, R.**  
Departamento de Producción Animal  
CIT-INIA Apdo. 8111  
28080 Madrid - Spain

**Introduction**

Homogeneity of genetic and environmental variances is assumed when predicting BV in the H-Friesian national evaluations in Spain. However, several studies (Hill et al. 1983; De Veer and Van Vleck, 1987; Van der Werf and de Boer, 1989) have shown heterogeneity of variances at different production levels. This has also been observed in the Spanish population (Ibáñez et al. 1991). In addition, estimates of heritability obtained with data from the N-NE region of the country yielded larger heritabilities than the ones obtained with the national data set (Ben Gara et al. 1991). Results of both studies are shown in Table 1.

Theoretical studies (Hill, 1984; Vinson, 1987) show that more animals are selected from the larger variance environments. Bias appears in bulls EBV with a larger proportion of daughters in one environment and in cows EBV (Boldman and Freeman, 1990).

BLUP methodology takes account of heterogeneity of variances if the (co)variance structure is known (Gianola, 1986). Unfortunately, that structure is generally unknown. Assuming that heterogeneity of variances derives from different environmental conditions and genetic practices in different contemporary groups (CG), defined as herds, herd-year or herd-year-season classes, seems appealing. However, estimation of genetic parameters would be inaccurate due to the small amount of information to estimate within CG variances and computationally unfeasible in most cases. A computationally easy procedure used in several countries (Wiggans and Van Raden, 1991; Bagnato and Jensen, 1992, personal communication) is to obtain estimates of the phenotypic standard deviation within CG and to standardize the data to be used afterwards in the national evaluation following current procedures. In order to improve the accuracy of estimation of phenotypic variance within CG, the individual estimate of CG variance is combined with an estimate of the variance of observations obtained in CG assumed to be in a similar variance class (Wiggans and Van Raden, 1991; Bagnato and Jensen, 1992; Weigel and Gianola, 1993).

A different approach that could be used to accommodate heterogeneity of variances is to investigate sources of heterogeneity in terms of factors such as production level, year of calving, management practices, parity, region, etc. This would allow clustering of CG within similar variance classes, which would largely decrease the number of genetic parameters to be estimated. A procedure to test goodness of fit for a log-linear model describing heterogeneity of variances among CG in terms of main effects and interactions has been proposed by Foulley et al. (1990). This procedure also requires large amounts of computation, particularly when interactions among factors are considered. An application of the procedure by Weigel et al. (1993) to a subset of the US Holstein population could not detect significant differences in within herd sire variances when using a model considering several management factors.

The objective of this study was, first, to investigate sources of heterogeneity of variance in the Spanish

Friesian population at a phenotypic level by using computationally simple multi-variate statistical analyses. This was done to establish homogeneous variance CG classes to help in improving estimates of CG phenotypic standard deviation and in defining a log-linear model to be further used in studying heterogeneity of variances at genetic and residual level. Second, the impact of standardization of data by an estimate of the CG standard deviation on genetic evaluations was also evaluated.

## Material and Methods

Data from the 1992 H-Friesian evaluation were provided by the Spanish Friesian Association (CONAFE). The data set consisted of 400989 milk production records and genetic evaluations for milk yield obtained by CONAFE from non-standardized data. CG used in the analyses were composed of animals producing in the same herd and period of time. Time periods were different for each CG to allow a minimum of 4 observations. Summary statistics about average yield, standard deviation, correlation between mean and standard deviation and size of CG are presented in Table 2.

Analyses aimed at, first, establishing environmental sources causing heterogeneity of variances in order to define homogeneous CG classes, and, second, investigating changes in genetic proofs after standardization by estimates of CG standard deviation.

### 1. Establishing sources of heterogeneity and CG group classes.

Three factors, region, time period and CG management level were investigated. Parity and CG production level, which could be thought to have a effect on determining heterogeneity, were not consider because of small CG size and because of the high correlation between management and production levels (0.94), respectively.

Sixteen regions, four time periods (1980-84; 1985-87; 1988-89 and 1990-92) and three management levels (high, medium and low, defined from BLUP estimates of CG effect) were considered. Clustering techniques (Hierarchical Cluster Analysis and Multiple Correspondence Analysis) were used to define homogeneous CG classes in terms of combinations of region, time period and management level.

### 2. Standardization and impact on genetic evaluations.

Once CG classes were defined, estimates of CG standard deviation were obtained following the procedure suggested by Weigel and Gianola (1993) that uses empirical Bayesian methods. Standardized records were then obtained in the following manner:

$$y_{ij}^* = \frac{y_{ij} - \bar{y}_i}{s_i} s_b + \bar{y}_i$$

where  $y_{ij}^*$  = standardized  $j^{\text{th}}$  record belonging to  $i^{\text{th}}$  CG

$y_{ij}$  = original, non-standardized  $j^{\text{th}}$  record in  $i^{\text{th}}$  CG

$\bar{y}_i$  = estimate of CG mean

$\hat{\sigma}_i$  = estimate of CG standard deviation

$s_n$  = new CG standard deviation

New CG standard deviation was fixed to be 1274.34, which corresponds to the average standard deviation of CG in high management level and time period 1990-92.

BLUP animal model genetic evaluation was run using standardized data and several analyses were performed to investigate differences with 1992 national evaluation, performed on non-standardized data. Correlations among EBV obtained in both runs and individual and average changes in ranking were obtained for all and top bulls and cows. Correlations between daughter yield deviation (DYD), obtained as defined in Van Raden and Wiggans (1991), and parent average (PA) were also calculated. The accuracy of both procedures to predict future records was also investigated. This accuracy was measured by calculating correlations between PA of heifers calving in 1992 obtained from BLUP evaluations that only used data until 1991 and actual records of those heifers adjusted by age and month of calving. Correlations were obtained within CG to avoid the effect of CG management level and averaged for all CG afterwards.

## Results

### 1. Establishing sources of heterogeneity and CG group classes.

A Hierarchical cluster analysis grouped initial 36781 CG in 5 homogeneous classes according to CG standard deviation. The clustering algorithm defines homogeneous classes by maximizing variability among classes and minimizing variability within classes, maintaining total initial variability. After clustering, reduction in variability among classes was only 2.26%.

The Multiple Correspondence Analyses showed a positive association between the homogeneous clusters and combinations of levels of region, time period and management level. This result indicates that those three factors should have an effect in defining heterogeneity among CG

A final cluster analysis established groups of region-time-management classes according to the distribution of those classes in the 5 homogeneous clusters. 37 groups of region-time-management classes were finally obtained.

### 2. Standardization and impact on genetic evaluations.

After standardization, average CG standard deviation was 1200 Kg, correlation between standardized CG mean and standard deviation decreased to 0.17 and expected values of individual CG standard deviations were around 1274. A generalized likelihood ratio test performed to test homogeneity of variances after standardization indicated homogeneity.

Correlations between EBV, changes in ranking, correlations between DYD and PA for bulls and between PA and future records of heifers for both standardized and non-standardized data are shown in Table 3. Correlations between estimated breeding values under both procedures was high (0.9887). Changes in ranking were larger for cows (20% cows were excluded from the 1000 best cows list after standardization) than for bulls (9% bulls were not in the 100 best bulls list after standardization). Correlation between DYD and PA for bulls did not greatly improve after standardization (0.7698 vs. 0.7553 for standardized and non-standardized data, respectively). A similar result was observed for correlation between PA and future records for heifers (0.2079 vs. 0.2033 for standardized and non-standardized data, respectively).

## Conclusions

Standardization by CG phenotypic standard deviation did not greatly change genetic evaluations. However, individual changes in ranking were detected, specially, for cows. This would be important regarding MOET techniques to be applied within this population and to produce more fair evaluations. Further studies need to be done to investigate goodness of clustering techniques in terms of explaining total variability among CG and to determine whether heterogeneity of variances exists at genetic and environmental level.

## References

- Boldman, K.G. and A.E. Freeman. 1990. Adjustment for heterogeneity of variances by herd production level in dairy cow and sire evaluation. *J. Dairy Sci.* 73:503.
- Ben Gara, A., R. Alenda, A. Alonso. 1991. Estima de los componentes de varianza con dos modelos diferentes en dos poblaciones de ganado Frisón en España. ITEA. Vol extra 11 (II):565.
- De Veer, J.C. and L.D. Van Vleck. 1987. Genetic parameters for first lactation milk yields at three levels of herd production. *J. Dairy Sci.* 70:1434.
- Foulley, J.L., M. San Cristobal, D. Gianola, and S. Im. 1990. A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *J. Dairy Sci.* 73:1612.
- Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogeneous. *Theor. Appl. Genet.* 72:671.
- Hill, W.G. 1984. On selection among groups with heterogeneous variance. *Anim. Prod.* 39:473.
- Hill, W.G., M.R. Edwards, M.K.A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition at different levels and variability of production. *Anim. Prod.* 36:59.
- Ibáñez, M.A., R. Alenda, M.J. Carabaño. 1991. Interacción genotipo medio ambiente, nivel de producción de leche y tamaño de los animales en la población Frisona española. ITEA. Vol extra 11(II):559.
- Van der Werf, J.H.J. and W. De Boer. 1989. Estimation of genetic parameters in a crossbred population of Black and White dairy cattle. *J. Dairy Sci.* 72:2615.
- Vinson, W.E. 1987. Potential bias in genetic evaluations from differences in variation within herds. *J. Dairy Sci.* 70:2450.
- Weigel, K.A. and D. Gianola. 1993. A computationally simple Bayesian method for estimation of heterogeneous within-herd phenotypic variances. *J. Dairy Sci.* 76:1455.
- Weigel, K.A., D. Gianola, B.S. Yandell and J.F. Keown. 1993. Identification of factors causing heterogeneous within-herd variance components using a structural model for variances. *J. Dairy Sci.* 76:1466.
- Wiggans, G.R. and P.M. VanRaden. 1991. Method and effect of adjustment for heterogeneous variance. *J. Dairy Sci.* 74:4350.

**Table 1. Evidence of heterogeneity of variances in the Spanish- Friesian population.**

Estimates of genetic and phenotypic variances (kg<sup>2</sup>) and heritability for high and low production level (Ibáñez et al., 1991).

	High (> 6000 kg)	Low (< 5000 kg)
N	24768	14194
Genetic Variance	352854	124428
Phenotypic Variance	1244638	478560
Heritability	0.28	0.26

Estimates of genetic and phenotypic variances (kg<sup>2</sup>) and heritability for all regions and N-NE data (Ben Gara et al., 1991)

	All regions	N - NE
N	63250	15294
Genetic Variance	232388	241780
Phenotypic Variance	941121	687232
Heritability	0.25	0.35

**Table 2. Number of contemporary groups (CG), average milk yield, average standard deviation and average size of CG, and correlation between average mean and standard deviation of CG used in the analyses (standard deviations in parenthesis).**

No. of CG	36781
Average milk yield (kg) <sup>1</sup>	5791 (1155)
Average standard deviation (kg)	1019 (407)
Average size	11 (10)
Correlation CG mean-standard deviation	0.54

<sup>1</sup> Milk yield was not adjusted to mature equivalent.

**Table 3. Correlations between estimated breeding values ( $R_{S,NS}$ ), changes in ranking, correlation between daughter yield deviation and parent average for bulls ( $R_{D\Delta PA}$ ) and correlation between parent average and future records of heifers ( $R_{PA-FR}$ ) for both standardized (S) and non-standardized (NS) data.**

	All data N=302012	Cows N=283519	Bulls N=19493	1000 top cows	100 top bulls
$R_{S,NS}$	0.9887	0.9871	0.9877	0.8184	0.9880
Changes rank.					
Average		8247.7	552.6	242.4	11.4
Maximum				+668 -2376	+36 -72

	S-data	NS-data
$R_{D\Delta PA}$ (N=1254) <sup>1</sup>	0.7698	0.7553
$R_{PA-FR}$ (N=970) <sup>2</sup>	0.2079	0.2033

<sup>1</sup> Data on bulls with more than 20 daughters in at least 5 herds with identified dam and sire.

<sup>2</sup> Data on heifers born in 1992-93 producing in CG with 5 or more observations with identified dam and sire.