Integrated Adjustment for Variance Heterogeneity of Test-Day Yield with a Structural Variance Model in the Walloon Genetic Evaluation

P. Mayeres [1], A. Gillon[1] and N. Gengler [1,2]
[1]*Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium*
[2]*National Fund for Scientific Research, B-1000 Brussels, Belgium*

Introduction

The heterogeneity of variance (**HV**) is a frequent issue in genetic evaluation systems (**GES**) and accounting for it is a great challenge. Several procedures have been proposed to correct for HV, as well for lactation as for test-day (**TD)** models. There are basically two methods: preadjustment of data [e.g. Wiggans and Vanraden, 1996] or integrated adjustments [e.g. Meuwissen *et al.,* 1996] which are nowadays often based on variance models. This last approach has many advantages and therefore there is a tendency to develop integrated HV adjustments inside GES. First, as location and dispersion parameters are evaluated simultaneously, it allows to take into account the heterogeneity due to age, month of calving, culling rate, breed composition, mating or selection practices [Robert *et al.,* 1999]. Second, the ability to use a structural approach for the variance allows to reduce the parameter space, by appropriate identification of meaningful sources of variation of variances [San Cristobal *et al.,* 1993].

Two related methodologies exist to adjust for variance heterogeneity simultaneously. The first, introduced by Foulley *et al.* [1992], directly model the variances in each environment with a structural model. The second, presented by Meuwissen *et al.* [1996], is a multiplicative model, the multiplicative factors being obtained with a structural variance model based on observed heterogeneity.

The first approach has the advantage of flexibility, while it does not force variance ratios to be constant. The multiplicative method is easier to compute because variance adjustment only affects the right-hand side of the mixed model equations. Moreover, if variance components are not re-estimated simultaneously, both methodologies become computationally affordable.

The aim of this paper was to show the method initially proposed by Meuwissen *et al.* [1996], but adapted and improved with the modifications presented by Gengler *et* Wiggans [2001], implemented for the Walloon production genetic evaluation TD model for milk production. The results of the structural variance model, and those of some validation tests are presented.

Materials and Methods

Data

Data were provided for the February routine run by the Walloon Breeding Association (AWE) which manages performance recording data in the Walloon Region of Belgium. Data edition was done to keep the TD records between 5 and 365 DIM during the first three lactation and to exclude unlikely ages for a given lactation or gestation lengths. Additionally, TD yields were limited to 0 to 85 kg for milk, 1.5 to 9.0 % for fat and 1.0 to 7.0 % for protein.

A total of 12,742,020 TD records from 731,442 lactating cows (essentially Holstein, but also other dairy and dual-purpose breeds) was finally used.

Mean Model

The multilactation, multitrait RRTDM used can be written as [Auvray and Gengler, 2002]:

$$y_c = Xb+Q(Wh+Za+Zp)+e$$

where y_c is a vector of adjusted milk, fat and protein TD records, **b** is a vector of fixed effects (herd*test date, stage of lactation, age at calving regressed inside stage of lactation*major breed composition*season of calving*period of calving, gestation stage), **h** is a vector of herd*period of calving environmental random regression coefficients, **a** is a vector of additive genetic random regression coefficients, **p** is a vector of permanent environmental random regression coefficient, **e** is a vector of random residuals, **X**, **W** and **Z** are incidence matrices, **Q** is the covariate matrix for the second order Legendre polynomials.

As presented by Mayeres *et al.* [2003], an integrated detection and correction of outliers, based on the check of residuals, was applied.

Variance Model

The following model based on Gengler *et Wiggans* [2001] was used to correct for HV in TD records:

$$y_c = Xb +(y- Xb) \exp(-1/2(S_i\beta- S_i\beta_{base}))$$

where S_i is the incidence matrix linking the homogeneity subclass, defined as 6 lactation stage inside herd test date, to the dispersion effects (β). The effects in β were:

- mean;
- subclass production level*period (5 levels*periods of 3 months);
- subclass mean DIM (10 DIM classes);
- region (5);
- subclass size (1 to 6 and more);
- random effect of herd test date.

Two major differences with the paper of Meuwissen *et al.* [1996] have to be underlined.

First, the model used did not scale fixed effects, which could be problematic when multiple classes of homogeneity are defined for the same fixed effects class [Gengler *et Wiggans*, 2001].

Second, the study of autocorrelograms showed no evidence of a first order autoregressive process for the random herd test date effect. For this reason no covariance structure between test date of the same herd was considered.

The variance base was defined in order to match with data used for variance component estimation, which were the TD records recorded between January 1990 and December 1999.

As in Gengler and Wiggans [2001] and Lidauer *et al.* [2001], the adjustment factors were estimated separately for each trait in each lactation.

A model without HV correction was launched by considering $(S_i\beta- S_i\beta_{base})$ to be equal to zero.

Estimation of genetic variance within year

A validation test, based on Mendelian sampling deviation, as proposed by Fikse *et al.* [2003], was implemented for the model with and without HV adjustment. The genetic variance within year was estimated, for animals with known parents, as:

$$\hat{\sigma}_{ui}^2 = \frac{1}{q_i}\Sigma\left[\hat{m}_k^2 + PEV(\hat{m}_k)\right],$$

with q_i the number of animals in year i, $\hat{m}_k$ is the Mendelian sampling deviation and $PEV(\hat{m}_k)$ is the prediction error variance of $\hat{m}_k$, estimated directly from animals expected daughter contributions. Birth years between 1975 and 2000 were considered for this analysis.

Results and Discussion

Structural Variance Model Solutions

We will present the β solutions for the different effects of the structural model. The exponential of the opposite of these values is used for the adjustment of **y**: a β solution superior to zero correspond to a more variable environment, and

the constructed adjustment factor of y will be reduced.

The following figures present the solutions for the fixed effects of subclass production level*period (Figure 1), subclass mean DIM (Figure 2), subclass size (Figure 3), for milk in first lactation. The results for the other traits and lactations were similar.
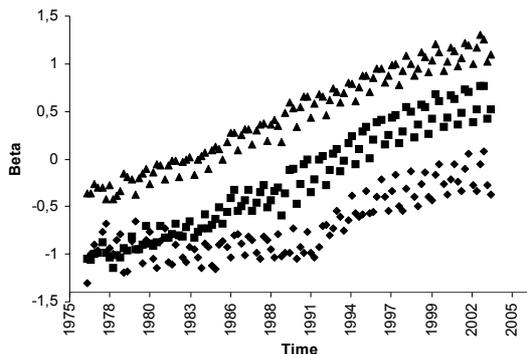

Figure 1. Evolution of variance model solutions for the subclass production level*period effect at three levels of production (◆=low, ■=intermediary, ▲=high).

The variability tended to increase with production level and time. The trends are different for the different production levels: it is more flat for low production level, what means that the variability is more stable with time than for high production levels. The three other levels (only the third is presented here) are intermediary. The variation inside year of approximately 0.5 that appear in Figure 1 for each production level is due to the separation of time in periods of 3 months in place of years. So two time trends are superposed: one through years and one inside years.
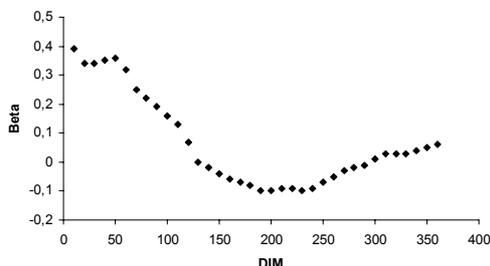
Figure 2. Evolution of variance model solutions for the subclass mean DIM effect.

Variance is larger at the beginning and to a less extent at the end of the lactation.
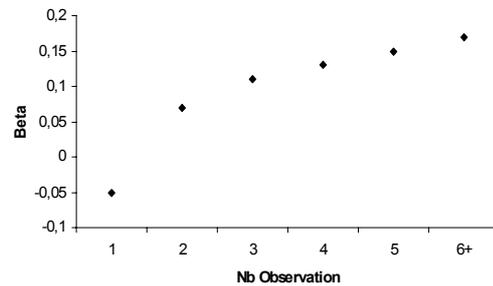
Figure 3. Evolution of variance model solutions for the subclass size effect.

Figure 3 showed that larger subclass contemporary groups are more variable.

The solutions for the region effect went up from 0.09 for the less variable to 0.20 for the more variable region.

The variability is mostly explained by the subclass production level*period effect, then in decreasing order by the subclass mean DIM effect, the subclass size effect and the region effect.

Table 1. Mean, standard deviation (std), minimum (min) and maximum (max) solutions of the random herd test date effect for the different traits and lactations.

Trait	Lact	Mean	Std	Min	Max
Milk	1	0.00	0.155	-0.81	1.81
Fat	1	0.00	0.205	-1.06	3.36
Protein	1	0.00	0.179	-0.87	2.04
Milk	2	0.00	0.159	-0.95	2.07
Fat	2	0.00	0.179	-0.99	2.81
Protein	2	0.00	0.177	-0.91	2.55
Milk	3	0.00	0.142	-1.10	2.23
Fat	3	0.00	0.178	-1.17	2.56
Protein	3	0.00	0.149	-1.14	2.31

Table 1 gives the distribution parameters of the random herd test date effect. Most random herd test date solutions were close to zero. Some were however more extreme. To avoid too important corrections that could be artefacts and create distortions, it was decided to limit these solutions to a [−1 , +1] interval.

The matrix ($\mathbf{S_i\beta}$- $\mathbf{S_i\beta_{base}}$), containing the adjustment factors, was constructed (Table 2).

Table 2. Mean, standard deviation (std), minimum (min) and maximum (max) values of adjustment factors for the different traits and lactations.

Trait	Lact	mean	Std	Min	Max
Milk	1	-0.21	0.72	-2.01	2.54
Fat	1	-0.27	0.88	-2.71	2.90
Protein	1	-0.20	0.75	-2.18	2.90
Milk	2	-0.22	0.78	-2.32	2.66
Fat	2	-0.28	0.94	-3.07	3.11
Protein	2	-0.20	0.82	-2.27	3.07
Milk	3	-0.22	0.86	-2.57	3.04
Fat	3	-0.26	1.02	-3.05	3.48
Protein	3	-0.19	0.92	-2.40	3.37

Mean deviations stayed still negative, what is due to the time trend correction. As for the random herd test date effect, it was decided to limit these solutions to the [-2.5 , +2.5] interval.

Validation tests

INTERBULL validation test III was performed for the model with HV adjustment: for the 3 traits, the time trend stays below the 2% genetic standard deviation limitation.
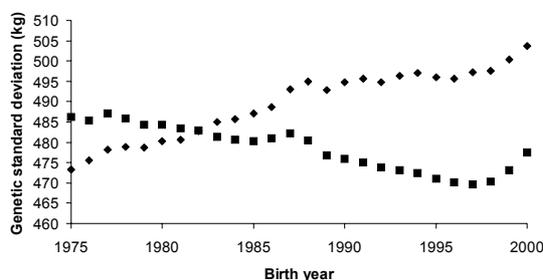
Figure 4. Evolution the estimated genetic standard deviation of milk for the model with (■) and without (◆) correction for variance heterogeneity.

Figure 4 shows the evolution of the estimated genetic standard deviation within year ($\hat{\sigma}_{ui}$). For the model without HV correction, a positive trend is observed. The modeling of HV reduces this trend, to the point that it becomes negative. The linear coefficient of the regression of $\hat{\sigma}_{ui}$ of birth year changes from 103 to −34.1 from the model without to the model with HV correction. Similarly, the quadratic coefficient goes from −0.0256 to 0.00848. Even if standards of this test are not yet defined, and if the birth years or

animals taken in account should have a great influence on the results, it seems that these results indicate that we have been able to reduce the effect of HV on Mendelian sampling and that the HV model is more stable than the model without correction.

March test run participation

Breeding values of 1453 sires were send to INTERBULL for the March test run. New international correlations were estimated, and those between BEL (Walloon Region) and other countries generally felt from 1 to 5 %, even with countries having a similar HV correction, as Finland. This large decrease was surprising and needs to be investigated.

Conclusions and Implications

This study presents the developments made to incorporate an adjustment for variance heterogeneity in the Walloon TD model. The choice was to use a multiplicative mixed model, with a variance structuration, based on the work of Meuwissen *et al.* [1996].

The model seems to fit adequately the data, but a small decrease in genetic standard deviation through years and a fall in the INTERBULL correlations with other countries were observed.

It was decided not to implement the system in routine yet, and to study the reasons of these findings.

Our main suspicion is that we overcorrect HV with our current procedure. Some work is now in progress around the correction for HV linked to production level. Another hypothesis is that while the correction for production level seems correct globally, some herds may not follow this general tendency. The use of a first order autoregression could help to link random effects of adjacent test date, leading to better estimates of this effect in the structural model.

These studies are ongoing, and we intend to submit a modified and improved GES to the September test run.

Acknowledgments

References

Auvray, B. & Gengler, N. 2002. Feasibility of a Walloon test-day model and study of its potential as tool for selection and management. *Interbull Bulletin 29,* 123-127.

Fikse, W.K., Klei, L., Liu, Z. & Sullivan, P.G. 2003. Procedure of validation of trends in genetic variance. *Interbull Bulletin 31,* 30-36.

Foulley, J.L., San Cristobal, M., Gianola, D. & Im, S. 1992. Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Comput. Stat. Data Anal. 13,* 291-305.

Gengler, N. & Wiggans, G. 2001. Heterogeneity in (co)variance structure of test-day yield. *Interbull Bulletin 27,* 179-183.

Lidauer, M. & Mäntysaari, E.A. 2001. A multiplicative random regression model for test-day data with heterogeneous variance. *Interbull Bulletin 27,* 167-171.

Meuwissen, T.H.E., De Jong, G. & Engel, B. 1996. Joint estimation of breeding values and heterogeneous variances of large data files. *J. Dairy Sci. 79,* 310-316.

Robert-Granié, C., Bonaïti, B., Boichard, D. & Barbat, A. 1999. Accounting for variance heterogeneity in French dairy cattle genetic evaluation. *Livest. Prod. Sci. 60,* 343-357.

San Cristobal, M., Foulley, J.L. & Manfredi, E. 1993. Inference about multiplicative heteroskedastic components of variance in mixed linear Gaussian model with an application to beef cattle breeding. *Genet. Sel. Evol. 25,* 3-30.

Wiggans, G. & VanRaden, P.M. 1996. Method and effect of adjustment for heterogeneous variance. *J. Dairy Sci. 74,* 4340-4257.