# Heterogeneity in (Co)Variance Structures of Test-Day Yields

***N. Gengler [1,2], and G. R. Wiggans [3]***
*[1]National Fund for Scientific Research, B-1000 Brussels, Belgium*
*[2]Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium*
*[3]Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA*

## Abstract

First-lactation test-day milk, fat and protein yields from New York, Wisconsin, and California herds were adjusted additively for age and lactation stage. A random regression model with third-order Legendre polynomials for permanent environmental and genetic effects was used. This model also included a random effect with the same polynomial regressions for 2-yr intervals within herd (herd, time period of calving effect). Phenotypic variances were modeled using a mixed model. Heritabilities and permanent environment showed the expected pattern. Herd, time period effects explained some of the phenotypic variance differences especially at the beginning (12% to 20%) of the lactation. Variances increased with time, size of subclass and milking frequency. Month of test had only a very limited influence. Low and high milk production level showed increased variances, as did late and especially early lactation stages. Repeatabilities of variances observed for a given herd, test-day, frequency class across nested variance subclasses based on lactation stage were 14% to 17%.

## Introduction

Although a common assumption of genetic evaluation models is homogeneity of (co)variances, this assumption is often incorrect across time or herds. In test-day models an additional reason for unequal variances, is linked to lactation stage. For yield data, research done mostly on 305-d lactation data showed that the most important issue was total yield of the herds. Therefore, the objective of this study was to show existence of unequal phenotypic variances by modeling this heterogeneity of variance jointly with (co)variance estimation.

## Materials and Methods

### *Data*

First-lactation test-day yields from New York, Wisconsin, and California were adjusted additively for age and lactation stage. Adjustment factors were those obtained by Bormann et al. (2001). This data was used to create three data sets where herds were randomly selected. The data sets were very similar in size (72582 to 76641 records) and production levels (29.0 to 30.7 kg milk yield).

### *(Co)variance Component Estimation Model*

(Co)variance components were estimated using an accelerated EM-REML algorithm (Gengler et al., 1999) and the following random regression model:

$$\mathbf{y} = \mathbf{Xt} + \mathbf{Q(Hh} + \mathbf{Z^*a} + \mathbf{Zp)} + \mathbf{e}$$

which can be rewritten as

$$\mathbf{y} = \mathbf{Xt} + \mathbf{Qr} + \mathbf{e}$$

by setting $\mathbf{r} = \mathbf{Hh} + \mathbf{Z^*a} + \mathbf{Zp}$

where **y** = vector of test-day records for milk, fat, or protein yields; **t** = vector of fixed herd, test day, and milking frequency class effects; **h** = vector of random herd, time period (2-yr of calving) effects; **p** = vector of random permanent environment effects; **a** = vector of animal effects (breeding values); **e** = residual effect; **X** = incidence matrix linking **y** and **t**; **r** = vector of regressions; **Q** = matrix of constant, linear, and quadratic modified Legendre polynomials: $I_0 = 1$, $I_1 = 3^{0.5}x$, and $I_2 = (5/4)^{0.5}(3x^2 - 1)$, where $x = -1 + 2[(DIM - 1)/(365 - 1)]$, linking **y** and **r**; **H**, **Z** and **Z*** = incidence matrices linking y with

**h**, **p** and **a**. The herd, time period effect was introduced as an earlier study on the same data set showed that the portion of total variance explained by this effect was not negligible (Gengler and Wiggans, 2001).

***Integrated Heterogeneous Variance Adjustment***

Meuwissen et al. (1996) developed a method to allow joint estimation of breeding values and heterogeneous variances which is basically a multiplicative mixed model that scales milk production records toward a common phenotypic variance through computation of a heterogeneity parameter at each iteration. Then adjustment factors are obtained by modeling those heterogeneity parameters and extracting an expected variance estimate. This method is appealing because it accounts for (co)variances among observations and heterogeneity can be modeled in a flexible manner. Pool and Meuwissen (2000) applied it in a slightly modified manner to correct for unequal variances due to lactation stage in the estimation of (co)variance components. Their method can be improved in two ways. First, it should be possible to use the original method by Meuwissen et al. (1996) because of its greater flexibility and the possibility of writing a variance model. Second, the scaling of fixed effects can be quite problematic especially if several variance subclasses exist for the same level. Following a suggestion from Pool and Meuwissen (2000) we adapted the methods by precorrecting for fixed effects at every EM round. The general model solved in EM round n+1 can be written as:

$$\mathbf{y}_C^{n+1} = \mathbf{X}\mathbf{t}^{n+1} + \mathbf{Q}\mathbf{r}_C^{n+1} + \mathbf{e}_C^{n+1}$$

where the pre-corrected data vector was obtained from:

$$\mathbf{y}_C^{n+1} = \mathbf{X}\hat{\mathbf{t}}^{n} + \left(\boldsymbol{\Gamma}^{n}\right)^{-1}\left[\mathbf{y} - \mathbf{X}\hat{\mathbf{t}}^{n}\right]$$

One should note that as soon as $\mathbf{X}\hat{\mathbf{t}}^{n+1} \approx \mathbf{X}\hat{\mathbf{t}}^{n}$ and $\boldsymbol{\Gamma}^{n+1} \approx \boldsymbol{\Gamma}^{n}$ the model can be written as:

$$\mathbf{y}^{n+1} = \mathbf{X}\mathbf{t}^{n+1} + \boldsymbol{\Gamma}^{n+1}\left[\mathbf{Q}\mathbf{r}_C^{n+1} + \mathbf{e}_C^{n+1}\right]$$

which is a random regression model with scaled random effects and a modified version of the model of Meuwissen et al. (1996). The diagonal matrix $\boldsymbol{\Gamma}^{n+1}$ contains the scaling coefficients $\exp(\gamma_k / 2)$ for every variance subclass k obtained from the solutions of the model used to describe heterogeneous variances.

***Heterogeneity Parameter and Model***

A feature of the method of Meuwissen et al. (1996) is that the modeling of the heterogeneity parameter uses a weighted mixed model on pseudovariates obtained by summing current $\gamma_k$ with the remaining heterogeneity within variance subclass. Based on Meuwissen et al. (1996). the heterogeneity parameter called z for variance subclass k could be developed:

$$z_k = \left[\frac{1}{\sigma_e^2}\left(\mathbf{y}_{Ck} - \mathbf{X}_k\hat{\mathbf{t}}_k\right)'\hat{\mathbf{e}}_{Ck} - n_k\right]/2$$

where subscript k denotes blocks of matrices or vectors associated with records in variance subclass k, $n_k$ is the number of records in subclass k and $\sigma_e^2$ is the residual variance. This formula is conceptually similar to a quadratic form but for a log-normal distribution,. The variance associated with this heterogeneity parameter was estimated as:

$$\mathrm{Var}(z_k) = \left[\frac{1}{\sigma_e^2}\left(\mathbf{Q}_k\hat{\mathbf{r}}_{Ck}\right)'\left(\mathbf{Q}_k\hat{\mathbf{r}}_{Ck}\right) + 2n_k\right]/4$$

We assume in this development equal weights for every test-day record but the formulas can be easily modified. The weighted mixed model on pseudovariates $\mathrm{diag}(\gamma_k) + \mathbf{W}^{-1}\mathbf{z}$ was written as:

$$\left(\mathbf{S}'\mathbf{W}\mathbf{S} + \boldsymbol{\Lambda}^{-1}\right)\boldsymbol{\beta} = \mathbf{S}'\mathbf{W}\left[\mathrm{diag}\left(\gamma_k\right) + \mathbf{W}^{-1}\mathbf{z}\right]$$

where $\boldsymbol{\beta}$ = solutions, $\mathbf{S}$ = design matrix linking pseudovariates and $\boldsymbol{\beta}$; $\mathbf{W}$ = diagonal matrix of iterative weights with $\mathbf{W} = \mathrm{diag}[\mathrm{Var}(z_k)]$ and $\mathrm{Var}(\boldsymbol{\beta}) = \boldsymbol{\Lambda}$.

In contrast to Meuwissen et al. (1996), $\gamma_k$ were scaled towards a common base:

$$\gamma_k = \mathbf{S}_k\boldsymbol{\beta} - \gamma^{base}$$

The variance subclass was defined as the portion of the records in a given herd, test day, and milking frequency class that were as homogeneous as possible for lactation stage. Therefore, test-day yields were subdivided according to days in milk (6-65, 66-125, 126-185, 186-245, 246-305, 306-365).

The effects used in the variance model were:

- mean;
- year-season of test (6 month period);
- month of test (12 months);
- milking frequency (2x or 3x and more);
- subclass size (11 classes);
- subclass milk yield (26 classes);
- subclass mean days in milk (36 classes of 10 days).

In addition to all these fixed effects, a random herd, test-day, and milking frequency effect was fitted. Repeatabilities of the random effects were estimated using Method R and the accelerator described by Druet et al. (2001).

## Results and Discussion

Only results for variance components are given. Tables 1 to 3 show the evolution of relative variances over the lactation compared to total variance. Heritabilities and permanent environment followed the expected pattern. Heritabilities were intermediate compared to results for high, medium and low producing herds found in previous studies on the same data (Gengler and Wiggans, 2001) as in the present study no distinction was made between herd production levels. Herd, time period effects explained some of the phenotypic variances especially at the beginning (12% to 20%) of the lactation.

Residual variances were kept constant in absolute values over the whole lactation. The changes in the residual relative variance reflect changes in the total variance.

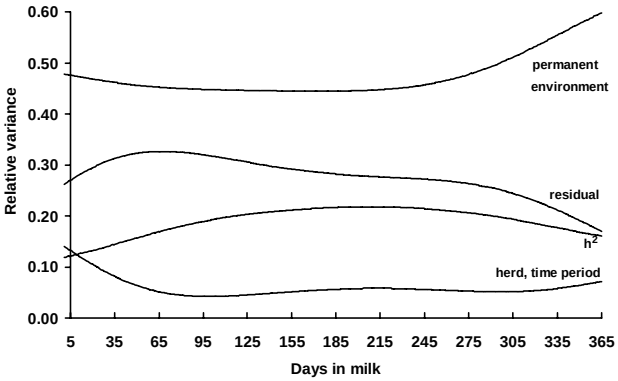


Figure 1. Relative variances for milk yield.

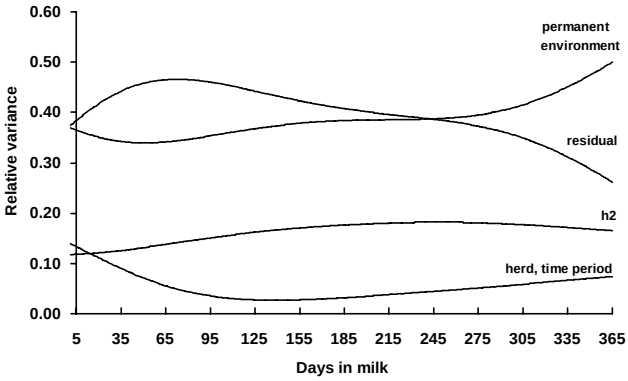


Figure 2. Relative variances for fat yield.

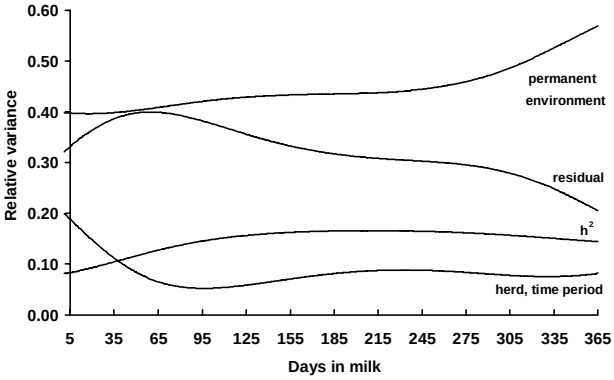


Figure 3. Relative variances for protein yield.

### *Variance Model Solutions*

Solutions for fixed effects of the variance model reflect on a linear scale the logarithm of variance that can be interpreted transforming them into multiplicative scaling effects computed as $\exp(-\text{solution}/2)$.

Mean effects were 0.25, 0.26 and 0.19 for milk, fat and protein yields. Figure 4 shows the changes in variances over time. Except for a certain stagnation in 1992 to 1994 there is a clear trend for milk, fat and protein. As milk production is in the model, this trend should not be an artifact of increased production.

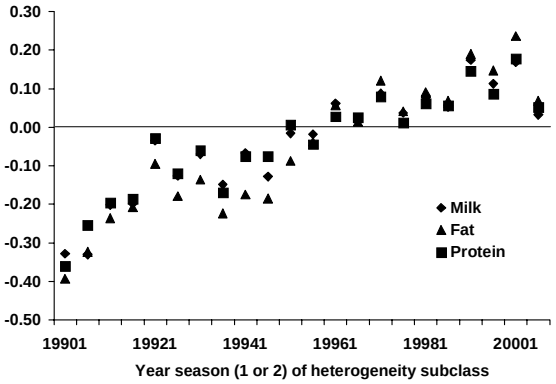


Figure 4. Variance model solutions for year season of variance subclass effects.

Figure 5 shows, however, that the month of test had only a very slight influence on the variances.

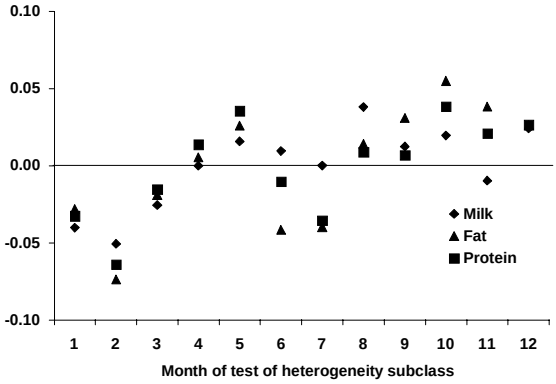


Figure 5. Variance model solutions for month of test of variance subclass effects.

Variances increased with subclass size, as expected (Figure 6). For very large subclasses variances stabilized or tended to decrease slightly.

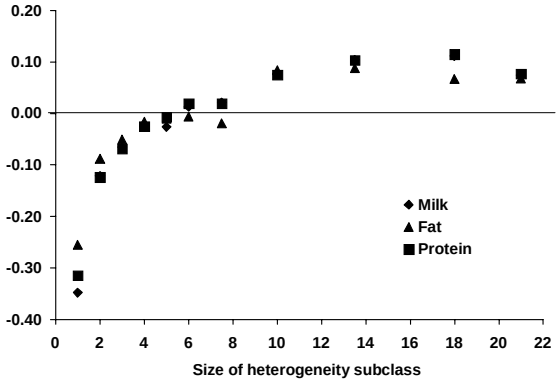


Figure 6. Variance model solutions for subclass size effects.

Low and high milk production level showed increased variances (Figure 7) as did late and especially early lactation stages (Figure 8).

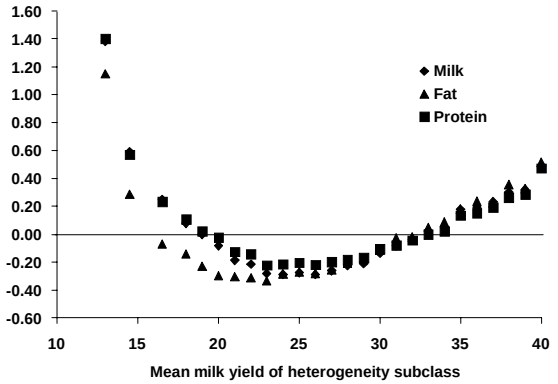


Figure 7. Variance model solutions for mean milk yield of variance subclass effects.

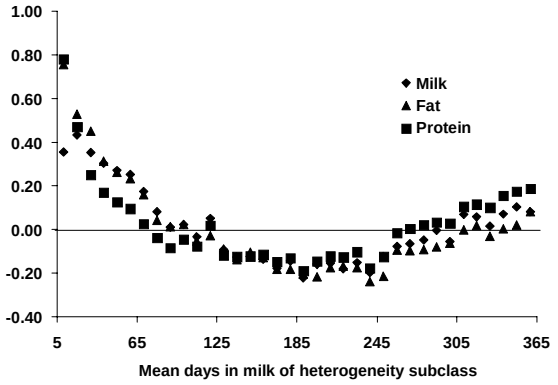


Figure 8. Variance model solutions for mean days in milk of variance subclass effects.

Variances increased with the frequency of milking (Table 1). As this is estimated jointly with the production level there should be no confounding between these effects.

Table 1. Variance model solutions for milking frequency.

| Milking | Trait | | |
|---|---|---|---|
| | Milk | Fat | Protein |
| 2x | -0.10 | -0.13 | -0.09 |
| 3x and more | 0.20 | 0.27 | 0.19 |

Repeatabilities of variances observed for a given herd, test-day, frequency class across nested variance subclasses based on lactation stage were 14% to 17%.

## Conclusions

First, this study presents and uses successfully a method to model and estimate jointly (co)variances and variance heterogeneity in test-day models. Relative variances followed the expected pattern. Heritabilities were intermediate compared to results for high, medium and low producing herds found in previous studies based on the same data. Therefore, one should account for these heritability differences in future studies, possibly by adjusting weights of test days as it is done currently for lactation records. Herd, time period effects explained some of the phenotypic variances especially at the beginning of the lactation. Variances increased with time, size of subclass and milking frequency. Month of test had only a very limited influence. Low and high milk production level showed increased variance as did late and especially early lactation stages. Repeatabilities of variances for a given herd, test-day, frequency class were 14% to 17%.

## Acknowledgements

Nicolas Gengler, who is Research Associate of the National Fund for Scientific Research, Brussels, Belgium, acknowledges its financial support.

## References

Bormann, J., Wiggans, G.R., Philpot, J.C., Druet, T. & Gengler, N. 2001. Estimation of test-day variance components for permanent environmental effects within and across parities and lactation stage, age, and pregnancy effects. *J. Dairy Sci.* (Submitted).

Druet,T., Misztal, I., Duangjinda, M., Reverter, A. & Gengler, N. 2001. Estimation of genetic covariances with Method R. *J. Anim. Sci. 79,* 605-615.

Gengler, N., Tijani, A., Wiggans, G.R. & Misztal, I. 1999. Estimation of (co)variance function coefficients for test day yield with a expectation-maximization restricted maximum likelihood algorithm. *J. Dairy Sci. 82,* (Aug) online.

Gengler, N. & Wiggans, G.R. 2001. Variance of effects of lactation stage within herd by herd yield. *J. Dairy Sci. 84(Suppl. 1),* 216 (abstr. 896).

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.

Pool, M.H. & Meuwissen, T.H.E. 2000. Reduction of the number of parameters needed for a polynomial random regression test-day model. *Livest. Prod. Sci. 64,* 133-145.