Three Methods to Validate the Estimation of Genetic Trend in Dairy Cattle

B. BONAITI, D. BOICHARD, A. BARBAT, and S. MATTALIA[1]

Station de Génétique Quantitative et Appliquée
Institut National de la Recherche Agronomique
[1] Institut de l'Elevage
78352 Jouy en Josas Cedex
France

ABSTRACT

Three methods are proposed to validate the estimation of genetic trend in dairy species. In the first one, the official proofs, generally derived from a repeatability animal model applied to several lactations, are compared to proofs derived from first lactation data only. Estimated genetic trends with both methods are expected to be similar. In the second method, the within-sire daughter yield deviations are analysed by production year and are expected to remain stable. The third method analyses variations of official proofs over time by regression. Any systematical trend associated with new daughters information is an estimate of the bias of the estimated genetic trend. The first two methods require a free access to raw data while the third one can be applied using published proofs. An application of these methods to the French Holstein evaluation for milk yield is presented.

(**Key words**: Animal Model, estimated genetic trend)

Interbull meeting, August 5-6 1994, Ottawa, Ontario, Canada

INTRODUCTION

In theory, if genetic parameters are known, the mixed model methodology provides the best estimate of genetic trend (Sorensen and Kennedy, 1984), particularly when it is applied to an animal model (Kennedy et al, 1988) which accounts for assortative matings and selection on every genes transmission pathway. However this property is true only if all underlying assumptions are fulfilled and particularly if the statistical model which describes the data is correct. Estimates of genetic trend have been published in dairy species (Wiggans and VanRaden, 1991; Bonaïti and Boichard, 1990; Banos et al, 1992; Barillet et al, 1992; Canon and Munoz, 1991) and are assumed to be unbiased because they have been obtained from the best methodology. However it may be easily demonstrated that omitting a factor of variation from the analysis may strongly bias the estimation of genetic trend. Bonaïti et al (1993) pointed out that preadjustments for age or parity effect may seriously affect the estimation of genetic trend. They reported that a 100 kg milk bias in the adjustment factor (i.e. about 5% of its range) leads to a 40 kg milk/year bias in the annual genetic trend estimate (i.e. 25-50 % of its true value).

Some studies already reported discrepancies between estimates of genetic trend. Banos et al (1992) reported quite different estimates obtained from the same data set of the Ayrshire breed analysed with the American and Canadian evaluation systems. Bonaïti et al (1993) showed that the genetic trends estimated in France and in the USA were not consistent. In an international comparison involving artificial insemination bulls from the USA, the Netherlands, Germany, Italy, and France with Schaeffer's method (1985), a large discrepancy among countries was observed between the expected and realized trends (Banos et al, 1993).

Although such a bias would probably have little effect on within-country selection efficiency, it provides a distorted picture of the real situation and strongly disturbs international germplasm exchanges. This paper describes three methods to validate the estimates of genetic trend in dairy cattle.

METHODS

Comparison of evaluations based on all lactations data or on first lactations data only

Most evaluation systems in dairy species use a standard repeatability model where successive performances of the same female are considered as repetitions of the same trait. In such a model, the estimation of genetic trend results from three different sources of information: the genetic superiority of the animals selected as parents over their contemporaries, i.e. the expected genetic trend; the difference between performances of contemporary daughters born from parents of different ages; and the difference between "true" contemporary performances of "pseudo"-contemporary animals raised in the same environment but born in different years. This third component depends on the definition of the contemporary group, which frequently gathers animals of different parities. Obviously, when only first lactation data are analysed, this component does not exist and cannot be suspected to affect the estimation of genetic trend. Moreover, the model is much simpler, since it is free of the effects of parity (and age to some extent), calving interval, days dry, and of the permanent environment effect. It would not be biased by incorrect age preadjustement factors. As the genotype*parity interaction is known to be small, both analyses are expected to provide similar estimates of genetic trend. In case of discrepancy between methods, the estimated genetic trend is more likely to be biased in the analysis of all lactations.

Within-bull variation of daughter yield deviation

Daughter yield deviations (DYD) of bulls are average performances adjusted for the dam breeding value and for all the effects included in the model, excepted daughter breeding value. Their expectation depends only on the bull, and they are theoretically independent of any environmental effect, and particularly of the year of calving. This property, simply derived from the fact that residuals are independently distributed, may be used to validate the estimation of genetic trend. The model to analyze the individual deviations may be defined as follows :

$$d_{ijk} = s_i + t_m + e_{ijk}$$

where d_{ijk} is the deviation of daughter k of sire i, obtained in year j, s_i is the fixed effect of sire i, t_m is the effect of the mth year of use of bull i, and e_{ijk} is the error. The year of use m should be defined within-bull as follows. Each daughter of sire i is characterized by the year of her first calving j. Let j_0 denote the year of first calving of the first daughter of sire i. Accordingly, let $m = j - j_0 + 1$. Usually, m equals 1 or 2 (rarely 3) for first crop daughters, whereas m is greater than or equal to 5 (rarely 4) for second crop daughters.

When the estimate of genetic trend is unbiased, the year effect has a zero expectation, and should not significantly differ from zero. Alternatively, the year effect shows a decreasing or increasing trend when the estimate of genetic trend is underestimated or overestimated, respectively. In practice, such an analysis does not require individual deviations, but only DYDs per sire and year of production, with appropriate weights. Moreover, this method is quite general and can be easily extended to validate the model regarding any environmental effects, such as regions, age classes, seasons, or management systems.

Analysis of bulls proof variations with time

The first methods presented above require a free access to the raw data, and their implementation is limited to scientists in charge of the domestic genetic evaluation. A third method requires only successive official male proofs released into the public domain, and can be applied by anybody. The underlying assumption is that successive proofs of the same bulls have the same expectation, equal to their true breeding value, and should present only random variations associated to new informations.

The basic idea of this method may be illustrated by the comparison of two groups of contemporary bulls, those eliminated after progeny test and those returned to service. Bulls from the first group always have almost the same amount of information over time, corresponding to the first crop of daughters and they should stable proofs. Bulls returned to service have two crops of daughters separated by a 3-year lag. If the evaluation system provides a biased estimate of genetic trend or, equivalently, a biased estimate of environmental trend, the second crop of daughters brings a biased information, and the contrast between both groups does not remain constant. This contrast (proofs of bulls returned to service minus proofs of bulls not returned to service) increases or decreases when the genetic trend is overestimated or underestimated, respectively.

This method may be formalized as follows. Let us assume that **u** and **v** are two vectors of proofs from two different releases and are estimates of the same breeding values. **u** is estimated with data of the first n_i daughters, whereas v depends both on the data used for **u** and on subsequent data. The absence of bias in estimated genetic trend may be tested with the following regression model :

$$\mathbf{v} = \mathbf{1}\,a + \mathbf{u}\,b + \mathbf{t}\,\delta + \mathbf{e}$$

where **e** is a vector of residuals. The a and b coefficients may be interpreted as conversion factors between releases. If the method is the same and if the reference bases remained unchanged, expected values for a and b are 0 and 1, respectively. Otherwise, a is an estimation of the difference between reference bases and b is an estimation of variability changes between methods. The δ coefficient is an estimate of the bias in estimated genetic trend, and should not significantly differ from 0. **t** is a known vector of general term

$$t_i = \sum_j \left(\frac{v_{ij}}{N_i}\,m\right)$$

with N_i being the number of daughters with information included in proof v_i, and v_{ij} being the number of additional daughters with first calving during year j ($N_i = n_i + \sum_j v_{ij}$). As above, m is defined within-bull such that $m = j - j_0$, where j_0 is the mean year of first calving of the daughters which contributed to **u**. In the particular case of the analysis of proofs variations between both crops of daughters, j_0 can be defined as the mean year of first calving of the first crop of daughters.
The error variance depends on the amount of information at both releases. In Appendix, it is shown that the error variance for bull i, assuming a sire model, is

$$\mathrm{Var}(e_i) = \mathrm{Var}(v_i \,/\, u_i) = \frac{\sigma_g^2}{(N_i + \lambda)^2}\left[n_i(n_i + \lambda)(1 - r^2) + v_{i0}\left(\lambda + \frac{v_{i0}\,\lambda}{n_i + \lambda}\right)\right]$$

where σ_g^2 is the additive genetic variance, $v_{i0} = N_i - n_i = \sum_j v_{ij}$, $\lambda = 4/h^2 - 1$, and r is the expected correlation between methods used to estimate **u** and **v**. Consequently, the weight of each observation in the regression analysis is

$$w_i = \frac{\sigma_g^2}{Var(e_i)} = (N_i + \lambda)^2 \Big/ \left[n_i(n_i + \lambda)(1 - r^2) + v_{i0}\left(\lambda + \frac{v_{i0}\lambda}{n_i + \lambda}\right)\right]$$

Experience from recent introduction of animal model shows that r may vary from 0.95 to 1. A value different from 1 has to be chosen to ensure the weight is always defined, even for bulls with no new information ($v_{i0} = 0$) but results are not sensitive to this value.

In the case where bulls from different ages were analysed simultaneously, a bull's birth year effect (g) could be added to the model :

$$\mathbf{v} = \mathbf{X}\,\mathbf{g} + \mathbf{u}\,b + \mathbf{t}\,\delta + \mathbf{e}$$

This effect could adjust for any bias in the estimation of **u**. As δ depends on the new daughters information only, it should not be affected by this additional factor.

This method can easily be extended to a situation where **u** and **v** are estimated breeding values for two different traits, with a genetic correlation ρ ($r < \rho < 1$). In such a case, the same model can be used but each weight w_i should be defined, as shown in Appendix :

$$w_i = (N_i + \lambda)^2 \Big/ \left[n_i(n_i + \lambda)(1 - r^2) + v_{i0}\left(\lambda + \frac{v_{i0}\lambda}{n_i + \lambda} + \frac{v_{i0}\, n_i (1 - \rho^2)}{n_i + \lambda}\right)\right]$$

The δ coefficient still estimates the bias in estimated genetic trend, and should not significantly differ from 0.

APPLICATION

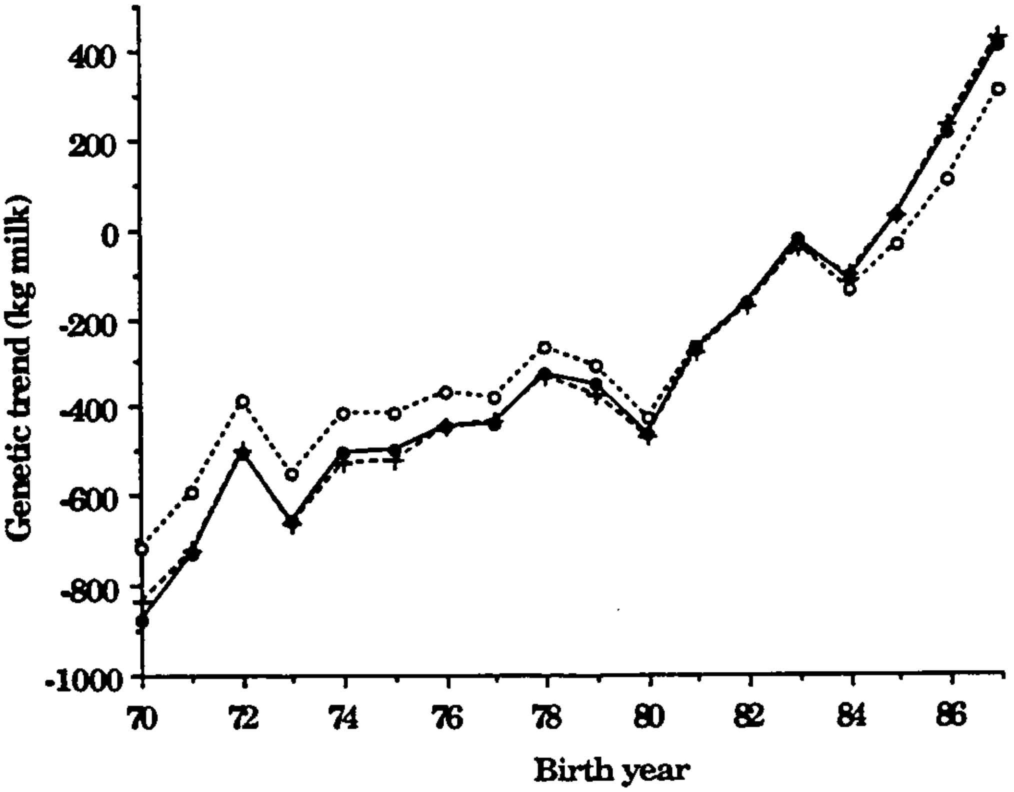
An application of these three methods to the French Holstein evaluation is presented. In France the animal model was implemented in 1990 (Bonaïti & Boichard,1990; Ducrocq et al, 1990). Data from parities 1 to 3 were analysed with a model including the additive genetic effect, a permanent environmental effect, genetic groups for unknown parents and the fixed environmental effects of herd-year, parity, age at first calving, calving interval within parity 2 and 3, and month of calving within parity. These last four fixed effects were nested within year and region. In order to correct for a putative underestimation of the genetic trend, this model was modified in 1993 by including an age effect within parity 2 and 3, and by replacing the calving interval effect by a days dry effect (Bonaïti et al, 1993). The effect of heterosis and recombination losses due to the Holstein x European Friesian crossbreeding was also added to the model (Boichard et al, 1993). These changes increased the estimated annual genetic trend for milk yield by 22 kg. The models used from 1990 to 1992 and since 1993 are called AM90 and AM93, respectively.

A comparison of both models provides a good illustration of the validation methods proposed in this paper, with one example of biased method (AM90) and one example of a supposedly unbiased method (AM93).

For the population of pure Holstein AI bulls born from 1978 to 1986, the figure 1 shows a good agreement between the estimated genetic trend with AM93 (+73 kg/year) and the first lactation analysis (+76 kg/year), while the previous model (AM90) presents a lower estimated genetic trend (+54 kg/year).

In the within-bull analysis of DYD (figure 2), a negative trend was observed with the AM90 model. On average, the DYD decreased by 50 kg milk between crops of daughters and by 10 kg per year within the service period. Moreover, this trend was not constant for all bulls but was most pronounced for recent bulls (born after 1978), while it was of smaller magnitude for bulls born before 1975. Although these trends were quite small compared to the large variation in the environmental conditions and the level of the mates, they reveal a 25% bias in the estimated genetic

FIGURE 1. Estimated genetic trend for milk yield in the Holstein AI bulls population according to the evaluation model (——●—— method used since 1993 (AM93), ...o.... method used from 1990 to 1992 (AM90), ... +.... first lactation data analysis)



trend. Results with AM93 were consistently better, in spite of a small decreasing trend (-20 kg milk) of DYD between the first and second crops of daughters.

In the analysis of proofs variations over time, the population considered included 2607 French AI bulls born from 1980 to 1984 with at least 20 and at most 200 daughters in March 1990. These bulls were already evaluated in 1990 but not yet returned to service. **u** was the vector of proofs released in March 1990 and v was the vector of proofs released either in December 1992 (AM90) or in March 1994 (AM93). The assumed heritability was .3 ($\lambda = 12.33$) and the correlation between methods was assumed to be .99 in both cases. Results are shown in Table 1. The b coefficients were very close to 1 as expected. The a coefficients were strongly negative and illustrated the effect of the reference rolling basis updated each year. When comparing proofs released in 1992 and 1990, the δ coefficient was found to be strongly negative, showing that the AM90 model underestimated the genetic trend. But when comparing proofs released in 1994 and 1990, the δ coefficient was found to be not significantly different from zero, showing that the AM93 provided an unbiased estimate of genetic trend. Table 2 provides similar results corresponding to the model including a birth year effect. Although AM90 was proven to underestimate the genetic trend, estimates for b and δ were virtually unchanged.

FIGURE 2. Within-bull evolution of daughter yield deviation for milk yield, according to production year (—●— method used since 1993 (AM93), ...o... method used from 1990 to 1992 (AM90))

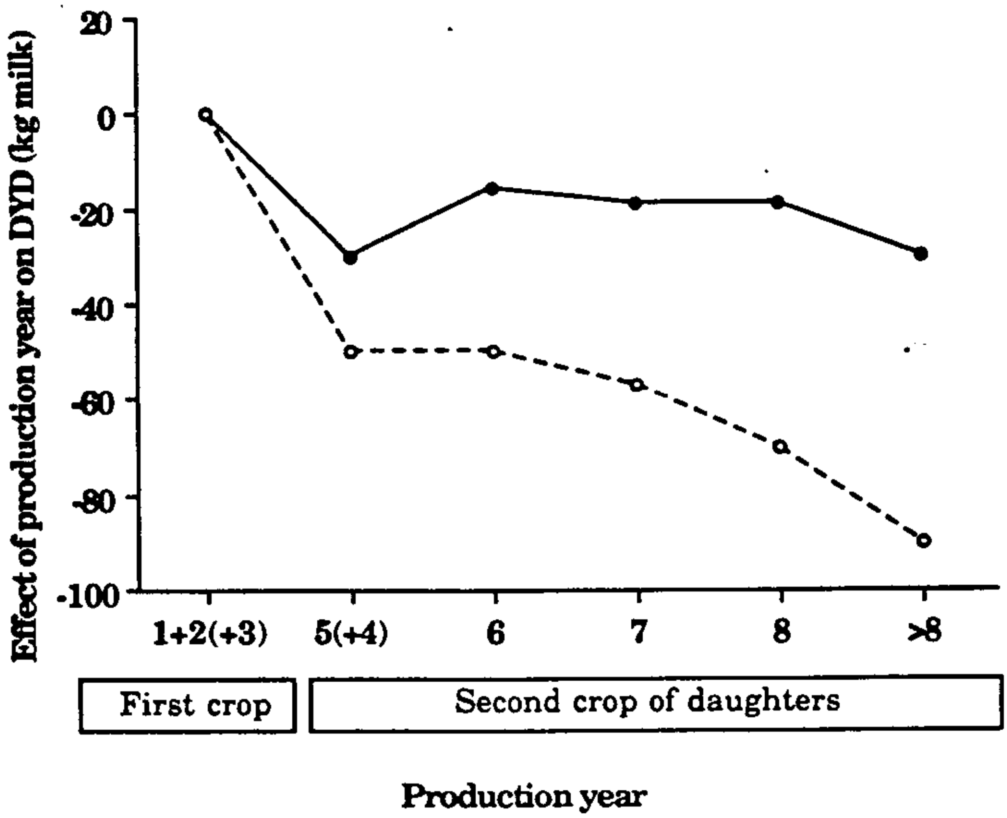


TABLE 1. Estimates of the regression analysis of milk yield proofs variation over time

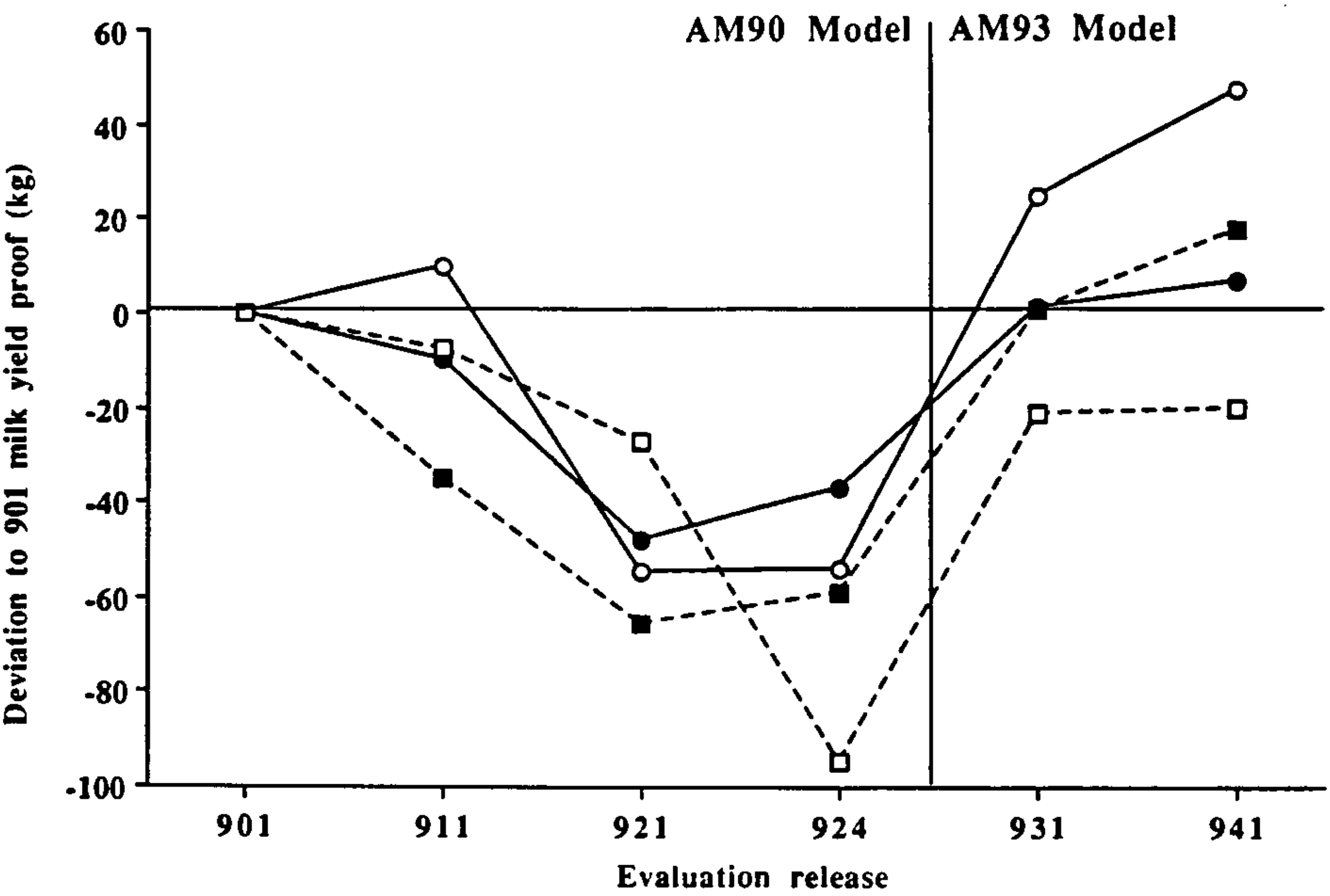
Evaluation method	a (kg)	b (kg)	δ (kg)
AM90	-303 ± 3	1.017± .004	-18.6 ± 4.9
AM93	-484± 3	1.005± .005	4.2 ± 4.3

TABLE 2. Estimates of the regression analysis of milk yield proofs variation over time, with the model including a bull's birth year effect

Evaluation method	b (kg)	δ (kg)
AM90	1.004± .004	-15.9 ± 4.8
AM93	0.989± .005	5.4 ± 4.2

Figure 3 illustrates these results. In contrast to bulls not returned to service which proofs remained essentially unchanged when expressed in the same reference basis, the proofs of bulls returned to service decreased by 40-100 kg with the new daughters from 1990 to 1992, and recovered their initial level thereafter with the implementation of the new model (AM93).

FIGURE 3. Evolution over time of the milk yield proofs of bulls returned to service, in deviation to 901 milk yield proofs
(——●—— bulls born in 1980 , ——○—— bulls born in 1981,
- - - -■- - - - bulls born in 1982, - - - -□- - - -bulls born in 1983)



CONCLUSION

Three simple methods have been proposed to validate the models used in genetic evaluation of dairy species. Although emphasis was put on the estimation of genetic trend, these methods, and particularly the within-sire analysis of daughter yield deviations, can be used for more general purpose. In the French situation, these methods have proven to be very useful to investigate results from the AM90 model and to validate the present system (AM93). The first two methods are very accurate but they require an analysis of the raw data. Their application is restricted to the scientists in charge of the official evaluations, who should be encouraged to use them routinely. Although the third one is probably less accurate and requires cumulated evaluations over several years, it is based on public domain information only. Therefore it could be used by anybody, and particularly by Interbull to validate national results, before using them to compute an international evaluation, for example based upon Schaeffer's method (1985).

REFERENCES

Banos, G., B. Bonaïti, M. Carabano, J. Claus, P. Leroy, P. Rozzi, J. Philipsson, G. Swanson, and J. Wilmink. 1993. Report of a joint research project between INTERBULL and COPA/COGECA on the feasibility of a simultaneous genetic evaluation of Black-and-White dairy bulls across the

European Community countries. Interbull Bull. 9

Banos, G. , G.R. Wiggans, and J.A.B. Robinson. 1992. Comparison of methods to rank bulls across countries. J. Dairy Sci. 75: 2560-2568

Barillet, F., D. Boichard, A. Barbat, J.M. Astruc, B. Bonaïti. 1992. Use of an animal model for genetic evaluation of the Lacaune dairy sheep. Livest. Prod. Sci. 31: 287-299

Boichard, D., B. Bonaïti, and A. Barbat. 1993. Effect of Holstein crossbreeding in the French Black-and-White cattle population. J. Dairy. Sci. 76: 1157-1162

Bonaïti, B., and D. Boichard 1990. Benefits from animal model evaluation of dairy cattle in France. Proc. 4th World Congr. Genet. Appl. Livest. Prod., Edinburgh, Scotland, 13: 364-373.

Bonaïti, B., D. Boichard, A. Barbat, and S. Mattalia. 1993. Problems arising with genetic trend estimation in dairy cattle. Proc. Interbull Meet., Aarhus, Denmark, August 19-20 1993. Interbull Bull. 8

Canon, J., and A. Munoz. 1991. Genetic trend for milk production in Spanish Holstein population. J. Anim. Breed. Genet. 108: 41-47

Ducrocq, V., D. Boichard, B. Bonaïti, A. Barbat, and M. Briend. 1990. A pseudo-absorption strategy for solving animal model equations for large data files. J. Dairy Sci. 73: 1945-1955

Kennedy, B.W., L.R. Schaeffer, and D.A. Sorensen. 1988. Genetic properties of animal models. J. Dairy Sci. 71(suppl. 2): 17-26.

Schaeffer, R.L. 1985. Model for international evaluation of dairy sires. Livest. Prod. Sci. 12: 105-115

Sorensen, D.A. and B.W. Kennedy. 1984. Estimation of response to selection using least-squares and mixed model methodology. J. Anim. Sci. 58:1097-1106

Wiggans, G.R., and P.M. VanRaden. 1991. Methods and effect of adjustment for heterogeneous variance. J. Dairy Sci. 74: 4350-4357

APPENDIX

Derivation of the weight of each observation in the regression analysis

The weight of each record is proportional to $1/\mathrm{Var}(v_i/u_i)$.

Each bull gets a first proof u_i based on n_i daughters and a second proof v_i based on N_i daughters ($N_i = n_i + v_{io}$). In a first step, u_i and v_i are assumed to be predictions of the same trait. The expected correlation between methods used to predict u_i and v_i is r. For convenience, both releases are assumed to be obtained with a sire model.

The proof u_i is derived from the daughter mean x_i of the first n_i daughters :

$$u_i = 2\left(\frac{n_i}{n_i + \lambda}\right)x_i$$

with

$$\lambda = \frac{4}{h^2} - 1$$

The second proof v_i is derived from two different means : x_i^* from the first n_i daughters and y_i from the v_{io} additional daughters. The means x_i and x_i^* include the same raw performances but may differ if the precorrections or the model have varied between both releases. The correlation between x_i and x_i^* is, by definition, r. We have :

$$v_i = 2\left(\frac{n_i x_i^* + v_{io} y_i}{N_i + \lambda}\right)$$

Let us define u_i^*

$$u_i^* = 2\left(\frac{n_i}{n_i+\lambda}\right)x_i^* \qquad [1]$$

Then

$$\mathrm{Var}\left(u_i^*\right) = \left(\frac{n_i}{n_i+\lambda}\right)\sigma_g^2$$

and

$$\mathrm{Var}\left(u_i^*/u_i\right) = \left(\frac{n_i}{n_i+\lambda}\right)\left(1-r^2\right)\sigma_g^2 \qquad [2]$$

Then, from [1] and [2]

$$\mathrm{Var}(x_i^*/u_i) = \frac{1}{4}\left(\frac{n_i+\lambda}{n_i}\right)\left(1-r^2\right)\sigma_g^2 \qquad [3]$$

We also have :

$$\mathrm{Var}\left(y_i/u_i\right) = \frac{1}{4}\sigma_g^2 \quad (1+\frac{\lambda}{\nu_{io}}) \qquad [4a]$$

We will assume that:

$$\mathrm{Cov}\left(x_i^*, y_i / u_i\right) = 0$$

Finally, by combining [3] and [4a]:

$$\mathrm{Var}\left(v_i / u_i\right) = \frac{\sigma_g^2}{\left(N_i+\lambda\right)^2}\left[n_i\left(n_i+\lambda\right)\left(1-r^2\right)+\nu_{io}\left(\lambda+\frac{\nu_{io}\lambda}{n_i+\lambda}\right)\right]$$

If two different traits are evaluated, with genetic correlation ρ, equation [4a] becomes

$$\mathrm{Var}\left(y_i/u_i\right) = \frac{1}{4}\sigma_g^2\left(\frac{\lambda}{\nu_{io}}+1-\frac{n_i\rho^2}{n_i+\lambda}\right) \qquad [4b]$$

and

$$\mathrm{Var}\left(v_i / u_i\right) = \frac{\sigma_g^2}{\left(N_i+\lambda\right)^2}\left[n_i\left(n_i+\lambda\right)\left(1-r^2\right)+\nu_{io}\left(\lambda+\frac{\nu_{io}\lambda}{n_i+\lambda}+\frac{\nu_{io}n_i(1-\rho^2)}{n_i+\lambda}\right)\right]$$