A General Approach for International Genetic Evaluations Robust to Inconsistencies of Genetic Trends in National Evaluations

Vincent Ducrocq[1], Isabelle Delaunay[1], Didier Boichard [1], Sophie Mattalia [2]

[1]*Station de Génétique Quantitative et Appliquée, Département de Génétique Animale, Institut National de la Recherche Agronomique,*

[2] *Département Génétique, Institut de l'Elevage*

78352 Jouy en Josas, France

email: vincent.ducrocq@dga.jouy.inra.fr

Abstract

This paper is a contribution from the PROTEJE project. It presents an approach which makes uses in a sire model of annual information (= average daughter yield deviations per year of performance and lactation number) for each sire. It is shown that with a suitable sire model (including a year x country effect, a lactation x country effect and a within country regression on age of the sire when his daughters are born), the genetic parameters and the sire solutions obtained are robust to over- or under-estimation of genetic trend in the national models. The approach has other benefits: validation of genetic trends is no longer critical (although obviously desirable), deregression becomes obsolete. An extension to categorical or survival traits is briefly described and a general strategy for the computation of equivalent daughter contributions and daughter yield deviations and their validation is proposed.

1. Introduction

The objectives of the PROTEJE (PROduction Traits European Joint Evaluation – Canavesi et al., 2002) project is to develop an alternative methodology for international evaluation, applicable to both bulls and cows, maintaining the modelling of environmental effects at the national level. One of the directions envisioned is the use of pre-corrected records, defined as performances adjusted for all fixed effects estimated in *national* evaluations. Admittedly, an international application of the project is not foreseeable in the short term. But pre-corrected records are not so different from another tool that has been proposed to summarise sire information: daughter yield deviations or DYD's (VanRaden and Wiggans, 1991). DYD's and their associated weighing factors, the EDC's (equivalent daughter contributions) are elements that could be used in multiple trait sire models for international MACE evaluations.

Over time, international evaluation methods have been improved but have been found highly sensitive to quality of national proofs and to changes in genetic correlations. The will to extend international evalua-tions to new traits described by complex, sometimes nonlinear models is also an incentive to look for a better, more robust approach. This paper presents a general strategy to compute characteristics similar to DYD's and EDC's, even for nonlinear models that could be used in more robust MACE evaluations.

2. Method

In this section, we will first look at a simple situation in which a linear trait with no repeated records is analysed nationally with an animal model. More complex situations will be considered afterwards.

The following procedure will be used: starting with a national animal model evaluation, the information needed to run an equivalent univariate sire model will be identified. Isolating the components that may be biased when national models are misspecified will lead to the splitting of DYD's and EDC's in items that can be

analysed with more robust international sire models that reveal and correct for inconsistencies.

2.1. From an animal model to a sire model in national evaluation

Consider the following simple animal model:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{a} + \mathbf{e} \qquad [1]$$

with the classical notations: **X** and **Z** are incidence matrices relating the observation vector **y** to the fixed and random effects vectors **b** and **a.** For any daughter f of a particular sire s with observation y_f and associated vector of (fixed) explanatory variables $\mathbf{x}_f$:

$$y_f = \mathbf{x}_f' \mathbf{b} + a_f + e_f \qquad [2]$$

The following notations are needed: let gs and gd represent the (possibly group of unknown) sire and the (group of unknown) dam of the sire s. Let m_f be the (group of unknown) dam of the cow f (= mate of sire s). Let d_i represent the fraction of the total genetic variance of animal i which is not explained by parental information (=1, ¾ or ½ depending on whether 0, 1 or 2 parents of i are known). Let $\alpha = \sigma_e^2/\sigma_a^2$ be the ratio of residual to genetic variance and **A** the numerator relationship matrix between all animals. Finally, let $u_i = 0.5\ a_i$ be the genetic transmitting ability of animal i.

For later use, consider the fixed effect solutions of the mixed model equations in the national animal model evaluation. We have:

$$\left[\mathbf{X}'\mathbf{X}\right]\hat{\mathbf{b}} = \mathbf{X}'(\mathbf{y} - \mathbf{Z}\,\hat{\mathbf{a}}) \qquad [3]$$

Hence, for a particular level p of fixed effect q that affects a group Ω of animals i, i=1, … n_{pq}:

$$n_{pq}\,\hat{b}_{pq} + \sum_{\Omega}\left(\sum_{j\neq q}\hat{b}_j\right) = \sum_{\Omega}(y_i - \hat{a}_i) \qquad [4]$$

Also, after absorption of the equations for other effects into the equations for the additive genetic value:

$$\left[\mathbf{Z}'\mathbf{M}\mathbf{Z} + \alpha\,\mathbf{A}^{-1}\right]\hat{\mathbf{a}} = \mathbf{Z}'\mathbf{M}\,(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \qquad [5]$$

where **M** is the absorption matrix. A close look at this equation leads to the following expressions:

- for the equation corresponding to sire s:

$$\left(\alpha d_s^{-1} + \sum_f \frac{\alpha}{4} d_f^{-1}\right)\hat{a}_s - \alpha d_s^{-1}\left(\frac{\hat{a}_{gs} + \hat{a}_{gd}}{2}\right) - \sum_f \frac{\alpha}{4} d_f^{-1}.2\left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) = 0 \qquad [6]$$

i.e., after some manipulations, assuming that all daughters have a known (d_f = ½) or an unknown (d_f = ¾) dam:

$$\sum_f\left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) = \frac{2}{d_f^{-1}}\left[\left(d_s^{-1} + \sum_f \frac{d_f^{-1}}{4}\right)\hat{a}_s - d_s^{-1}\left(\frac{\hat{a}_{gs} + \hat{a}_{gd}}{2}\right)\right] \qquad [7]$$

or, if N_s is the total number of daughters of sire s:

$$\sum_f\left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) = \left(\frac{4d_s^{-1}}{d_f^{-1}} + N_s\right)\hat{u}_s - \frac{4d_s^{-1}}{d_f^{-1}}\left(\frac{\hat{u}_{gs} + \hat{u}_{gd}}{2}\right) \qquad [8]$$

- for a particular daughter of sire s, with no progeny herself; ignoring the off-diagonal elements of **M**:

$$\left(w_f + \alpha d_f^{-1}\right)\hat{a}_f - \alpha d_f^{-1}\left(\frac{\hat{a}_s + \hat{a}_{mf}}{2}\right) = w_f\left(y_f - \mathbf{x}_f'\hat{\mathbf{b}}\right) \qquad [9]$$

where w_f is the diagonal element of **M** for the daughter f, that is, the weight of her performance accounting for the estimation of other effects. Often, the only effect that will be taken into account is the size n_{cg} of the contemporary group and $w_f = 1 - \frac{1}{n_{cg}}$.

The summation of the equations [9] over all the daughters of sire s leads to:

$$\sum_f \left(w_f + \alpha d_f^{-1}\right)\hat{a}_f - \alpha \sum_f d_f^{-1}\left(\frac{\hat{a}_s + \hat{a}_{mf}}{2}\right) = \sum_f w_f (y_f - \mathbf{x}_f'\hat{\mathbf{b}}) \quad [10]$$

Subtract half the dam estimated breeding value multiplied by w_f on both sides. Still considering that all daughters have either a known or an unknown dam:

$$\sum_f \left(w_f + \alpha d_f^{-1}\right)\left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) - N_s \alpha d_f^{-1}\hat{u}_s = \sum_f w_f (y_f - \mathbf{x}_f'\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2}) \quad [11]$$

Denote as $\overline{w}_s = \left(\sum_f w_f\right)/N_s$ the average weight of the records of the daughters of sire s. We will assume the following approximation:

$$\sum_f \left(w_f + \alpha d_f^{-1}\right)\left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) \approx \left(\overline{w}_s + \alpha d_f^{-1}\right)\sum_f \left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) \quad [12]$$

Therefore:

$$\left(\overline{w}_s + \alpha d_f^{-1}\right)\sum_f \left(\hat{a}_f - \frac{\hat{a}_{mf}}{2}\right) - N_s \alpha d_f^{-1}\hat{u}_s \approx \overline{w}_s \sum_f (y_f - \mathbf{x}_f'\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2}) \quad [13]$$

Plugging [8] into [13], we get, after some manipulations:

$$\left(N_s \overline{w}_s + \gamma d_s^{-1}\right)\hat{u}_s - \gamma d_s^{-1}\left(\frac{\hat{u}_{gs} + \hat{u}_{gd}}{2}\right) = \overline{w}_s \sum_f (y_f - \mathbf{x}_f'\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2}) \quad [14]$$

with $\gamma = 4\left(\frac{\overline{w}_s}{d_f^{-1}} + \alpha\right)$. Here, $N_s \overline{w}_s = \sum_f w_f$ is the equivalent daughter contribution EDC_s for sire s. Define the average daughter yield deviation as $DYD_s = \sum_f (y_f - \mathbf{x}_f'\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2})/N_s$. Then equation [14] becomes:

$$\boxed{\left(EDC_s + \gamma d_s^{-1}\right)\hat{u}_s - \gamma d_s^{-1}\left(\frac{\hat{u}_{gs} + \hat{u}_{gd}}{2}\right) = EDC_s \times DYD_s} \quad [15]$$

Expression [15] can be regarded as a typical equation of the mixed model equations for a sire model:

$$\left[\mathbf{Z}_s'\mathbf{D}\mathbf{Z}_s + \gamma \mathbf{A}_s^{-1}\right]\hat{\mathbf{u}} = \mathbf{D}\mathbf{r} \quad [16]$$

where **r** is a vector of daughter yield deviations DYD_f, **D** is a diagonal matrix of equivalent daughter contributions, $\mathbf{Z}_s$ is the incidence matrix relating observations to sires, $\mathbf{A}_s$ is the numerator relationship matrix between sires.

Although not directly related to the topic of this paper, it is interesting to note that

$$\gamma = 4\left(\frac{\overline{w}_s}{d_f^{-1}} + \frac{\sigma_e^2}{\sigma_a^2}\right) = \frac{\sigma_e^2 + \overline{w}_s d_f \sigma_a^2}{1/4\,\sigma_a^2} \quad [17]$$

When all daughters have an unknown dam (d_f = ¾) and contemporary groups are large ($\overline{w}_s \approx 1$), γ is the usual variance ratio for sire models, with $\overline{w}_s d_f = ¾$, *but this is not the case* when contemporary groups are small or when all daughters have a known dam (d_f = ½), because part of the genetic variance in the daughters' observations not accounted for through the sire genetic effect "has been used" to compute the daughter yield deviation.

2.2 Incorrect national genetic trend

Suppose that there is a problem in the national evaluation such that the "national" estimation of the genetic gain in the population is incorrect. If true and estimated average genetic merits coincide for a particular reference year Y_0, this is no longer the case for animals i born in year $Y_{(i)} = Y_0 + t_{(i)}$:

$$\hat{a}_i = \hat{a}_{Ti} + t_{(i)}\,\Delta \qquad [18]$$

In [18], $\hat{a}_{Ti}$ refers to the ("true") estimated breeding value under the correct model and Δ represents the over- or under- estimation of the annual genetic trend. In equation [4], this implies:

$$n_{pq}\,\hat{b}_{pq} + \sum_{\Omega}\left(\sum_{j\neq q}\hat{b}_j\right) = \sum_{\Omega}\left(y_i - \hat{a}_{Ti} - t_{(i)}\Delta\right)$$

Obviously, the incorrect genetic trend will be associated with incorrect estimates of the fixed effects. The ones that are defined "globally", that is, irrespective of the time when the animals have their performance are certainly the least affected, because $\sum_{\Omega} t_{(i)}\Delta$ combines deviations from the true situation over a long period of time: these deviations are "averaged" out over the time axis. Conversely, all fixed effects defined on a time basis (contemporary groups or, say, age x year effects) will be biased, as they will be associated, for a given year, to one (or very few) values of $t_{(i)}$. Let Δ^*_τ represents the average over- or under-estimation of $\mathbf{x}'_f\hat{\mathbf{b}}$ for a typical record observed in year $Y_0+\tau$, as a consequence of the under- or over-estimation of the genetic trend or of missing important environmental factors that change annually. Note that nothing constrains Δ^*_τ to be a simple function of Δ.

For consistency, equation [14] will now be rewritten on the additive genetic value scale, i.e. working with 0.5 a_s instead of u_s. Also, daughter yield deviations and equivalent daughter contributions will be grouped by year $Y_0+\tau$ of performance. Let $n_{s,\tau}$ be the number of daughters of sire s with performance on year $Y_0 + \tau$. This leads to:

$$\left(\sum_\tau n_{s,\tau}\overline{w}_{s,\tau} + \gamma d_s^{-1}\right)\frac{\hat{a}_s}{2} - \gamma d_s^{-1}\left(\frac{\hat{a}_{gs}+\hat{a}_{gd}}{4}\right) = \sum_\tau \overline{w}_{s,\tau}\sum_{f\in Y_0+\tau}\left(y_f - \mathbf{x}'_f\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2}\right) \qquad [19]$$

Explicitly revealing the biases, we have:

$$\left(\sum_\tau n_{s,\tau}\overline{w}_{s,\tau} + \gamma d_s^{-1}\right)\left(\frac{\hat{a}_{Ts}+t_{(s)}\Delta}{2}\right) - \gamma d_s^{-1}\left(\frac{\hat{a}_{Tgs}+t_{(gs)}\Delta+\hat{a}_{Tgd}+t_{(gd)}\Delta}{4}\right) = \sum_\tau \overline{w}_{s,\tau}\sum_{f\in Y_0+\tau}\left(y_f - \mathbf{x}'_f\hat{\mathbf{b}}_T - \Delta^*_\tau - \frac{\hat{a}_{Tmf}+t_{(mf)}\Delta}{2}\right) \qquad [20]$$

Indeed, we are looking for a sire model making use of the "national" DYD's in [15] but with sire solutions as close as possible to the true ones, that is, such that:

$$\left(\sum_\tau n_{s,\tau}\overline{w}_{s,\tau} + \gamma d_s^{-1}\right)\left(\frac{\hat{a}_{Ts}}{2}\right) - \gamma d_s^{-1}\left(\frac{\hat{a}_{Tgs}+\hat{a}_{Tgd}}{4}\right) = \sum_\tau \overline{w}_{s,\tau}\sum_{f\in Y_0+\tau}\left(y_f - \mathbf{x}'_f\hat{\mathbf{b}}_T - \frac{\hat{a}_{Tmf}}{2}\right) \qquad [21]$$

We will call Q_s the difference between the right hand sides in [19] and [21], i.e., between the available and "correct" right hand sides for sire s:

$$Q_s = \sum_\tau \overline{w}_{s,\tau}\sum_{f\in Y_0+\tau}\left(y_f - \mathbf{x}'_f\hat{\mathbf{b}} - \frac{\hat{a}_{mf}}{2}\right) - \sum_\tau \overline{w}_{s,\tau}\sum_{f\in Y_0+\tau}\left(y_f - \mathbf{x}'_f\hat{\mathbf{b}}_T - \frac{\hat{a}_{Tmf}}{2}\right) \qquad [22]$$

Using [20] and [21] to compute Q, it follows that :

$$Q_s = \gamma d_s^{-1}\left(t_{(s)} - \frac{t_{(gs)}+t_{(gd)}}{2}\right)\frac{\Delta}{2} + \sum_\tau n_{s,\tau}\overline{w}_{s,\tau}\Delta^*_\tau + \sum_\tau n_{s,\tau}\,\overline{w}_{s,\tau}\left(\frac{t_{(s)}+\bar{t}_{(mf)}}{2}\right)\Delta \qquad [23]$$

Q_s is composed of three terms:

- The first term is a function of the generation interval on the male side. This information is usually available and could be directly included in what follows. But its contribution is probably small as it is not a function of the number of daughters of sire s. As the average generation interval is not very variable from year to year, this term will be regarded as a constant;

- The second term is a weighted sum of yearly biases, i.e. errors specific to all cows having a performance on a given year. The weights are the equivalent daughter contributions $EDC_{s,\tau}$ related to all daughters of sire s with a performance in year τ.
- The third term is a weighted sum of generation intervals on the female side, with weights $EDC_{s,\tau}$ again. Let $\lambda=0=\tau_0, \ldots, \tau_{last}$ represent the successive years of use of a bull in AI, starting at year $Y_0+\tau_0$ when his initial first crop daughters are born. The generation interval increases by 0.5 year when λ increases by 1.

As a consequence, if EDC's and DYD's are available for each year of production of the daughters of each sire, the sire solutions obtained considering the system with the following typical equation should be robust to improper genetic trends:

$$\left(\sum_{\tau} EDC_{s,\tau} + \gamma d_s^{-1}\right) u_s - \gamma d_s^{-1}\left(\frac{u_{gs}+u_{gd}}{2}\right) + \sum_{\tau} EDC_{s,\tau}\,\varphi_\tau + \sum_{\tau=\tau_0}^{\tau=\tau_{last}} \left((\tau-\tau_0)\ EDC_{s,\tau}\right)\delta = \sum_{\tau} EDC_{s,\tau} \times DYD_{s,\tau} \qquad [24]$$

In [24], the usual sire model equations are expended to include an effect of the (country by) year of production of the daughters and a regression term on the number of years of use of sire s, both weighted by the equivalent daughter contributions per year of production. From the knowledge of δ and the φ_r's, it is theoretically possible to compute Δ and Δ_τ^*.

2.3 Extension to a repeatability model

Consider now a repeatability model, for which the k^{th} performance of cow f is analysed in the national evaluation as:

$$y_{fk} = \mathbf{x}'_{fk}\,\mathbf{b} + a_f + p_f + e^*_{fk} \qquad [25]$$

The existence of repeated records introduces two important differences with the previous model. First, the computation of EDC's is obviously modified. In equation [5], the matrix **M** is now obtained after absorption of the contemporary group effects and the permanent environment effect. A typical diagonal element of **M** for an animal with L records indexed by k (k=1,..L) is:

$$w_f = \sum_k \left(1-\frac{1}{n_{cgk}}\right) - \frac{\left[\sum_k \left(1-\frac{1}{n_{cgk}}\right)\right]^2}{\sum_k \left(1-\frac{1}{n_{cgk}}\right) + \frac{\sigma^2_{e*}}{\sigma^2_{pe}}} \qquad [26]$$

where σ^2_{pe} is the permanent environment variance and n_{cgk} is the contemporary group size for lactation k. The variance σ^2_{e*} of the residual term e^*_{fk} in [25] does not include the variance of the permanent environment and therefore, an adjustment of w_f should be made before its use in [15] or [16] for which γ is a function of σ^2_e, not σ^2_{e*}. The adjustment factor is a scale factor equal to $\sigma^2_e / \sigma^2_{e*}$ The adjusted weight $w_{f(adj.)}$ can be decomposed as a sum over all lactations of a contribution $w_{f,m}$ per lactation:

$w_{f(adj.)} = \sum_k w_{f,k}$ with:

$$w_{f,k} = \frac{\sigma^2_e}{\sigma^2_{e*}}\,\upsilon_f\left(1-\frac{1}{n_{cgk}}\right) \qquad [27]$$

and:

$$\upsilon_f = 1 - \frac{\left[\sum_k \left(1-\frac{1}{n_{cgk}}\right)\right]}{\sum_k \left(1-\frac{1}{n_{cgk}}\right) + \frac{\sigma^2_{e*}}{\sigma^2_{pe}}}. \qquad [28]$$

For example, assume $\sigma^2_a=0.3, \sigma^2_{pe}=0.2, \sigma^2_{e*}=0.5$ and L=3 and the size of the contemporary groups is large enough so its effect on the weight can be ignored. Then, $\upsilon_f=0.455$, $w_{f,k}=0.636$ and $w_{f(adj.)}=1.909$.

The second difference with respect to the initial simple model is that the right hand side in [11] is now:

$$\sum_f \sum_k w_{f,k}\,(y_{fk} - \mathbf{x}'_{fk}\hat{\mathbf{b}} - \hat{p}_k - \frac{\hat{a}_{mf}}{2}) \qquad [29]$$

In principle, the development presented in section 2.1 can be repeated here to obtain formula [15] with a definition of EDC_s and DYD_s adapted to the repeated records situation. However, looking at DYD's over all lactations impairs the analysis made in section 2.2 on two aspects:

- the daughter yield deviations are no longer available on a yearly basis: the contribution to the computation of DYD_s of one daughter with three lactations will be scattered over year $Y_0+\tau$, $Y_0+\tau+1$ and $Y_0+\tau+2$. The estimation of φ_τ in [24] is no longer directly possible without extra approximations.
- often, some fixed effects in the model are defined on a lactation basis. Examples are corrections for age or for days open. A wrong specification of these will affect their estimate which in turn will lead to biased estimates of the genetic trend (see Bonaiti et al., 1993, for a concrete example). Working with DYD's averaged over several lactations prevents us from explicitly representing the fixed effect biases in the right hand side of [20].

Instead, define DYD's and EDC's within year τ and lactation K as:

$$EDC_{s,\tau,K} = \sum_{\substack{f \in Y_0+\tau \\ k=K}} w_{f,\tau,k} \qquad [30]$$

and

$$DYD_{s,\tau,K} = \frac{\sum_{\substack{f \in Y_0+\tau \\ k=K}} \left(y_{fk} - \mathbf{x}'_{fk}\hat{\mathbf{b}} - \hat{p}_k - \frac{\hat{a}_{mf}}{2} \right)}{EDC_{s,\tau,K}} \qquad [31]$$

Then [24] can be extended to:

$$\left(\sum_\tau \sum_k EDC_{s,\tau,k} + \gamma d_s^{-1} \right) u_s - \gamma d_s^{-1} \left(\frac{u_{gs} + u_{gd}}{2} \right) + \sum_\tau \left(\sum_k EDC_{s,\tau-k+1,k} \right) \varphi_\tau + \sum_k \left(\sum_\tau EDC_{s,\tau,k} \right) \theta_k + \sum_{\tau=\tau_0}^{\tau=\tau_{last}} \left((\tau-\tau_0) \sum_k EDC_{s,\tau,k} \right) \delta = \sum_\tau \sum_k EDC_{s,\tau,k} \times DYD_{s,\tau,k} \qquad [32]$$

Now, the variables associated with the (country x) year of production of the daughters are now the desired ones and the estimation of the bias θ_k between lactations and for each country becomes feasible.

Expression [32] corresponds to a typical sire equation for the following sire model:

$$DYD_{s,\tau,k} = \varphi_\tau + \theta_k + (\tau-\tau_0)\,\delta + u_s + e_{s,\tau,k} \qquad [33]$$

with $\mathrm{var}(e_{s,\tau,k}) = \sigma_e^2 / EDC_{s,\tau,k}$

2.4 Relationship with methods to validate genetic trend

It is possible to relate the fixed effect in [33] to the first two methods recommended for trend validation (see http://www-interbull.slu.se/service_documentation/ or Boichard et al (1995)).

- θ_k is a measure of what Boichard et al (1995) called the "*difference between contemporary performance of pseudo-contemporary animals performing in the same environment but born on different years*". With the "testing method 1", it is hoped to find that $\theta_k = 0$ for all k.

- the regression parameter δ is related to "*the difference between performances of "true" contemporary daughters born from parents of different ages*". The year effects φ_r should be all equal and δ should be 0 if DYD's are independent from the year of calving of the bulls' daughters, as expected with the Interbull 'testing method 2'.

- However, nonzero estimates of φ_r's, θ_k 's or δ in the sire model [33] should still lead to unbiased international sire proofs.

3. A simple illustration

Consider a simulated international population as in Delaunay et al. (2001). Data over 5 generations with 10000 cows per generation in each of 4 countries (= a total of 200000 records) were simulated with a simple model combining a herd effect (100 herds x 5 generations x 4 countries, normally distributed with variance 1 but analysed later as the only fixed effect), an additive genetic effect and a residual. The chosen genetic and residual variances were 0.25 and 0.75.

At each generation, proven sires and dams of the next generation of cows were selected based on the results of a simulated *national* evaluation. The male population was structured in the following way: for each generation and each country, 80 young bulls sired 50 daughters each, 20 proven sires had 250 second-crop daughters each in their country. Connection between countries was simulated through the use of 8 sires of sons at each generation, selected after a simulated (MACE) international evaluation. Also, exchanges of semen were modelled: in each country, 10 foreign sires from "neighbouring" countries were picked – again based on MACE results - to generate 100 daughters each. It was assumed that the genetic correlation between any pair of countries was 0.9.

The data set was analysed first with an animal model (AM) with the full pedigree, without groups of unknown parents and assuming either a genetic correlation of 0 between countries (=national evaluations) or the "true" correlation of 0.9 between countries (correct international evaluation). Figure 1 represents the average genetic trends for both evaluations and the average herd solutions per generation in the national evaluations. Obviously, the herd effects are biased with an increasing bias for later generations. A likely reason is that selection taking place in other countries was not properly accounted for. As a result, the genetic trend in the national evaluations was systematically under-estimated. This was so *despite the fact that these national evaluations would "pass" the validation tests,* as the true model and the true genetic parameters were used.

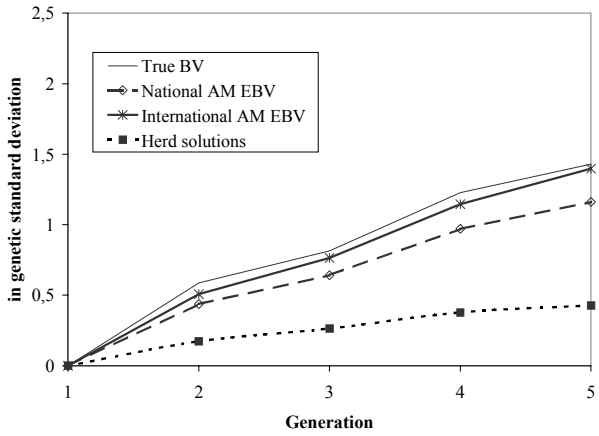

Figure 1: Average genetic trends and herd solutions in a simulated population of four countries and five generations when an animal model is used

(True BV: simulated breeding values, National and International AM EBV: Average estimated breeding values using an animal model and assuming a correlation of 0 (National) or 0.9 (International) between countries).

From the national evaluations, pre-corrected data were computed by simply subtracting the herd solution. These pre-corrected records were prepared for an animal model international evaluation in a study part of the PROTEJE project (Canavesi et al., 2002). But here, these pre-corrected record were considered as incorrectly computed DYD's, since there was no correction for the dam genetic effect, herd effects were biased and selection on the other countries was not properly accounted for. These DYD's were used in sire models comparable to [24], but without the regression on year of use of the bull as generations were barely overlapping. The more complete model [32] was not considered as there were no repeated records. Figure 2 presents the resulting estimated genetic trends. When a country effect was included as the only fixed effect as in the current deregression procedures, the absence of DYD's correction for the dams EBV led to a gross overestimation of the genetic trend.

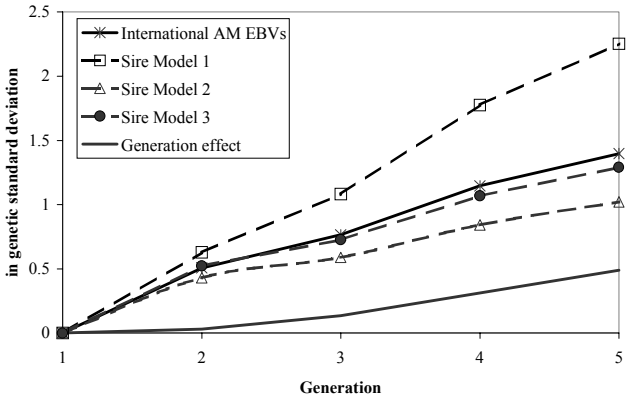

Figure 2: Average genetic trends and generation solutions in a simulated population of four countries and five generations when an improper sire model is used
(International AM EBV: as in Figure 1; Sire model: International sire model based on daughter yield deviations without correction for the dam EBV. Sire model 1: with a country effect and groups of unknown parents (GUP), Sire model 2: with a generation effect; Sire model 3: with a generation effect and GUP; generation solutions: for Sire model 3)

When the country effect was replaced by a country by generation effect equivalent to the φ_r's in [24], the genetic trend was much less biased, especially when groups of unknown parents were also included.

It was also found (results not shown here) that the REML estimation of genetic variances *and* correlations and residual variances with sire model 3 (with a generation effect and groups of unknown parents) gave virtually unbiased results (based on only 20 simulated data sets). This is in contrast with the results of, e.g., Madsen et al (2001), who found estimated genetic correlations biased downwards, even with DYD's instead of deregressed proofs, something they attributed to an inadequate estimation of the genetic trend.

4. Extension to more complex models

4.1 A simple rule for the computation of EDC's and DYD's

The use of deregression for traits described by complex, sometimes nonlinear models is questionable. A proper definition of EDC's for such traits has to be agreed upon as EDC's may greatly influence the right hand side values that will be calculated by deregression. When EDC's are incorrect, right hand sides are also incorrect and both the estimation of genetic parameters (genetic variances and correlations) and the international evaluation become dubious. Furthermore, there is no obvious extension of the genetic trend validation methods to apply to such traits.

The approach developed above may help to find a consensus on how to proceed: we started from the equations used in the national evaluation, considering an animal model (here). By "massaging" these equations, we were able to find an expression for EDC's and DYD's (equation [15]) suitable to run a *national* sire model with the same bull EBV's for bulls as for the animal model. These EDC's and DYD's can then be included in an international (multiple trait) sire model robust to shortcomings of national models, without requiring deregression nor strict trend validation (although the later is obviously still desirable).

Our conjecture is that in most cases, it is possible to "massage" the initial equations used in national evaluations to make apparent EDC's and DYD's that can be used in the international evaluation. **The critical point here is that the *national* sire model evaluation based on these EDC's and DYD's should lead to bull EBV's as close as possible to the ones obtained with the complex, possibly nonlinear (and possibly "animal") initial model.**

We will illustrate this approach with two examples:

4.2 EDC and DYD calculation for a survival analysis model

The approach was already described in Ducrocq (2001) and Ducrocq et al (2001). If length of productive life data are analysed using a Weibull proportional hazards model, the hazard function h(t) of a cow f is written:

$$h(t ; \mathbf{x}_f) = \rho t^{\rho-1} \exp\{ \mathbf{x}_f'\mathbf{b} + a_f\} \qquad [34]$$

Estimates of **b** and $\mathbf{a}=\{a_f\}$ are obtained maximising the logarithm of the joint posterior density of all location parameters (Ducrocq and Casella, 1996). In practice, a sire-maternal grandsire model is used but for the sake of simplicity, we will ignore this here (see Ducrocq, 2001 for details). At the joint mode of this joint posterior

density, its first derivative with respect to a_f is equal to 0, i.e.,

$$\delta_f - y_f^{\rho} e^{\mathbf{x}_f'\mathbf{b}+a_f} - \frac{1}{\sigma_a^2}\left(\mathbf{A}^{-1}\mathbf{a}\right)_f = 0 \qquad [35]$$

where y_f is the censored ($\delta_f = 0$) or uncensored ($\delta_f = 1$) length of life of cow f. The corresponding diagonal element of the information matrix (= minus the second derivatives of the joint posterior density with respect to a_f) is:

$$y_f^{\rho} e^{\mathbf{x}_f'\mathbf{b}+a_f} + \frac{d_f^{-1}}{\sigma_a^2} a_f \qquad [36]$$

Define $w_f = y_f^{\rho} e^{\mathbf{x}_f'\hat{\mathbf{b}}+\hat{a}_f}$ and

$$y_f^* = \frac{\delta_f}{y_f^{\rho} e^{\mathbf{x}_f'\hat{\mathbf{b}}+\hat{a}_f}} + \hat{a}_f - 1 .$$

Then [35] can be written as:

$$w_f\ \hat{a}_f - \frac{1}{\sigma_a^2}\left(\mathbf{A}^{-1}\mathbf{a}\right)_f = w_f\ y_f^* \qquad [37]$$

The definitions of w_f and y_f^* are slightly different from the ones proposed in Ducrocq (2001). They are more consistent as they are related to the information matrix but for practical purposes, these differences are really minor. Using the Normande data set (more than a million records), the sire EBVs based on a Weibull analysis and on a BLUP sire model using the above definitions of w_f and y_f^* had a correlation of 0.994.

Summing the w_f's and y_f^*'s over sires and year of first calving leads to definitions of EDC's and DYD's that can be used in international evaluations with model [24].

4.2 EDC and DYD calculation for a threshold model

Again, in a Bayesian analysis of discrete data with a threshold model, the thresholds, fixed effects and breeding values are traditionally estimated maximising the logarithm of the joint posterior density (Gianola and Foulley, 1983; Foulley and Gianola, 1996).

The Fisher scoring algorithm used for the maximis-ation involves the iterative solution of the system (Gianola and Foulley, 1983):

$$\begin{bmatrix} \mathbf{T}^{[i-1]} & \mathbf{L}'^{[i-1]}\mathbf{X} & \mathbf{L}'^{[i-1]}\mathbf{Z} \\ \mathbf{X}'\mathbf{L}^{[i-1]} & \mathbf{X}'\mathbf{W}^{[i-1]}\mathbf{X} & \mathbf{X}'\mathbf{W}^{[i-1]}\mathbf{Z} \\ \mathbf{Z}'\mathbf{L}^{[i-1]} & \mathbf{Z}'\mathbf{W}^{[i-1]}\mathbf{X} & \mathbf{Z}'\mathbf{W}^{[i-1]}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \xi^{[i]} - \xi^{[i-1]} \\ \mathbf{b}^{[i]} - \mathbf{b}^{[i-1]} \\ \mathbf{a}^{[i]} - \mathbf{a}^{[i-1]} \end{bmatrix} = \begin{bmatrix} \mathbf{p}^{[i-1]} \\ \mathbf{X}'\mathbf{v}^{[i-1]} \\ \mathbf{Z}'\mathbf{v}^{[i-1]} - \mathbf{G}^{-1}\mathbf{a}^{[i-1]} \end{bmatrix} \qquad |38]$$

at iteration i, where ξ refers to the thresholds; $\mathbf{W}$ and $\mathbf{v}$ are respectively a diagonal matrix and a vector with standard elements functions of the thresholds, the terms $\mathbf{x}_f'\mathbf{b}+a_f$ and the probability of response in each discrete category. The detailed expressions for the matrices $\mathbf{W}$, $\mathbf{T}$ and $\mathbf{L}$ and the vectors $\mathbf{v}$ and $\mathbf{p}$ are given in Gianola and Foulley (1983). Concentrating on the lower part of system [38], we get, at convergence (at iteration i="*"):

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{X}'\mathbf{W}^*\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}^*\mathbf{X} & \mathbf{Z}'\mathbf{W}^*\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{y}^* \\ \mathbf{Z}'\mathbf{W}^*\mathbf{y}^* \end{bmatrix} \qquad [39]$$

where $\mathbf{y}^*$ is a vector of working variables y_f^* such that, for animal f:

$$y_f^* = \mathbf{x}'_f\ \hat{\mathbf{b}} + \hat{a}_f + (w_f^{-1})\,\hat{v}_f \qquad [40]$$

In [40], w_f and v_f are the elements of $\mathbf{W}^*$ and $\mathbf{v}$ relevant for animal f. Once more, summing the w_f's (or even better: the w_f's obtained after absorption of (some of) the fixed effect equations in [40] into the genetic effect equations) and the y_f^*'s over sires and year of performance leads to new definitions of EDC's and DYD's to use in international evaluations.

5. Discussion

The strategy described in this paper has a clear underlying objective: getting rid of the deregression step in international evaluations.

There has been a debate on whether deregression followed by MACE is a reversible process, at the national level (Madsen et al., 2001; Madsen and Mark, 2002). The conclusion was positive *"as long as the same pedigree information is used in the two steps"*. This is almost never the case, in particular when animal models or complex sire-maternal grand-sire (nonlinear) models are used at the national level.

The deregression suffers two fundamental drawbacks:

- First, it is heavily dependent on the input parameters that are used: if one chooses wrong variance ratios and/or wrong EDC's, the right hand sides obtained by deregression will be inadequate. What we mean by "inadequate" is that, if used with the *correct* EDC's, these deregressed right-hand sides in a univariate sire model will *not* give back the initial (national) sire EBV's, *especially when the model used at the national level .is not a sire model but an animal model or a complex sire-maternal grand-sire model.*. However, as a deregressed right hand side is always easily obtained, the consequences of the choice of incorrect EDC's on the rest of the international procedure is easily overlooked. Taking an extreme (stupid) example, if ones add a (different) random value between 0 and 1000 to each EDC, deregression of sire proofs will still give a plausible result. Such a huge error is fortunately not common for usual linear traits. **However, for nonlinear traits such as those analysed via survival analysis or threshold models, the proper EDCs are functions of, respectively, cumulated risks or probabilities of response that can greatly differ in magnitude from simple record counts.**
- Deregression produces one single figure per sire. Consequently, there is no possibility to have any insight at potential systematic biases, no possibility to disentangle the underlying phenomena that may lead to such biases. A single safeguard has been proposed: the validation tests of national genetic trends. With a simple example, it has been shown that, although desirable, valid national trends are not always synonymous with exact genetic trends. After all, selection taking place abroad is accounted for in national evaluations only through approximations (usually with groups of unknown parents).

Two directions for improvement have been proposed here:

- First, the systematic supply of EDC's and DYD's within sire, year of performance - and possibly, lactation - is a step towards more complex international sire models useful to unravel national particularities or limitations. Admittedly, this requires more preparation than just sending proofs. But the ingredients are almost the same as the ones needed for the current trend validation procedures anyway. Perhaps more importantly, from a "political" perspective, trend validations would no longer be felt as an hurdle: if a country does not "pass" the trend validation tests, its data would not have to be necessarily excluded. Idealistically, it would also alleviate a possible climate of suspicion, as under- or over-estimation of national genetic trends would no longer influence international rankings. Finally, the more complex international sire model could also be used for a better estimation of genetic correlations between countries.
- Secondly, a general strategy has been proposed for the computation of EDC's and DYD's, by means of a rule based on common sense:
With the national information sent to Interbull, one should be able to get back one's own national sire proofs, using a simple univariate sire model. This rule could be the basis of a new kind of "national" validation, required before sending data to Interbull. An extreme consequence could be a non uniform definition of EDC's across country, adapted to each national situation. The strategy could also be applied to situations not considered here, such as for traits described by random regression models (following the work of Mrode and Swanson, 2002) or in multivariate settings (as in Liu et al., 2003) or including maternal effects.

Acknowledgement

Useful remarks and comments from Z. Liu, F. Canavesi and F. Fikse are gratefully acknowledged.

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci. 78,* 431-437.

Bonaiti, B., Boichard, D., Barbat, A. & Mattalia, S. 1993. Problems arising with genetic trend estimation in dairy cattle. *Interbull Bulletin 8*.

Canavesi, F., Boichard, D., Ducrocq, V., Gengler, N., de Jong, G. & Liu, Z. 2002. An alternative procedure for international evaluations: production traits European joint evaluation (PROTEJE). *Proc. 7th WCGALP*, Communication N° 01-59.

Delaunay, I., Ducrocq, V. & Boichard, D. 2002. A structural model for the matrix of genetic correlations between countries in international evaluations. *Proc. 7th WCGALP,* Communication N° 01-14.

Ducrocq, V. 2001. A two –step procedure to get animal model solutions in Weibull survival models used for genetic evaluations on length of productive life. *Interbull Bulletin 27*, 147-152.

Ducrocq, V., Boichard, D., Barbat, A. & Larroque, H. 2001. Implementation of an approximate multitrait BLUP evaluation to combine production traits and functional traits into a total merit index. *52nd EAAP Annual Meeting.* Budapest, Hungary, 7 (Abstract).

Ducrocq, V. & Casella, G. 1996. A Bayesian analysis of mixed survival models. *Genet. Sel. Evol. 28,* 505-529.

Foulley, J.L. & Gianola, D. 1996. Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model. *Genet. Sel. Evol. 28,* 249-273.

Gianola, D. & Foulley, J.L. 1984. Sire evaluation for ordered categorical data with a threshold model. *Génét. Sél. Evol. 15,* 201-224.

Liu, Z., Reinhardt, F. & Reents, R. 2003. Calculation and use of daughter yield deviations and associated reliabilities of bulls under multiple trait models. *Interbull forum (*http://www-interbull.slu.se/w-agora */index.php?bn=interbullforum_internationalges)*

Madsen, P., Sorensen, M.K. & Mark, T. 2001. Validation and comparison of methods to estimate (co)variance components for MACE. *Interbull Bulletin 27,* 73-79.

Madsen, P. & Mark, T. 2002. Estimation of across country genetic parameters for MACE based on DYD's or deregressed proofs. *Interbull Bulletin 29,* 28-31.

Mrode, R.A. & Swanson, G.J.T. 2002. The calculation of cow and daughter yield deviations and partitioning of genetic evaluations when using a random regression. *Proc 7th WCGALP*, Communication N° 01-04.

VanRaden, P.M. & Wiggans, G.R. 1991. Derivation, calculation and use of national animal model information. *J. Dairy Sci. 74,* 2737-2746.