

The Interbull Audit Project Part II: Development of a Program for Auditing Breeding Value Estimation Programs

H. Täubert¹, H.H. Swalve² and H. Simianer¹

¹ *Institute of Animal Breeding and Genetics, University of Göttingen, Göttingen, Germany*

² *Institute of Animal Breeding and Husbandry, University of Halle, Halle/Saale, Germany*

Abstract

In this project a simulation program is developed to generate performance records in order to check national breeding value estimation programs. An algorithm developed by R. THOMPSON, 1997, is used to simulate breeding values for base animals and their offspring, that follows the rules of the mixed model equations. The program generates performance records based on simulated breeding values. Solving the mixed model equations leads exactly to the simulated fixed effects and breeding values.

Key words: Interbull audit group, simulation of breeding values.

Introduction

In 1998 the Interbull audit group decided in Rotorua / New Zealand, to found a project to audit national breeding value estimation programs. This project is to check the information about the data quality that is submitted to Interbull for the international breeding value estimation. The procedure is based on a paper of Robin Thompson, 1997. The project started in March 2001 with a work visit in Uppsala. It was continued in an international survey to collect different data structures. Meanwhile a program was developed to simulate estimated breeding values (EBVs) and performances under a single trait animal model.

Aim of the project

Breeding value estimation is altogether a multiple step procedure. Beginning with the data collection and data selection, continued with pre-adjustments and other steps (e.g. including foreign information), the BLUP estimation to calculate EBVs is only one step in the cascade of procedures. It is still followed by a number of post processing steps. The aim of the project is to check the breeding value estimation program. Of course, it will not be feasible to check all the

different steps in the whole process, so that the auditing project will only concentrate on the pure BLUP breeding value estimation procedure.

Influences on the BLUP estimations

The step of breeding value estimation is affected by several parameters that can vary. At first the model can be very different, beginning with a simple sire or animal model up to a highly developed random regression test day model. The project will concentrate first on a single trait animal model only. After the development of a simulation program for this model, more complicated models can be implemented quite easily.

Secondly the environmental effects are very important for the estimated breeding values. The distribution of the effects can have a high influence on the breeding values through the number of contemporary herd mates. To get an overview of the fixed effects used in practical estimation procedures an international survey was developed and is already analyzed (König *et al.*, 2002). Some of this information should be used in the program.

The last effect on the estimation is the genetic structure of the population. Information about this topic was also collected in the survey, to get an overview about the structure in real breeding programs

Procedure to check breeding value estimation programs

After the development of a program to check the breeding value estimations, the following procedure should be used in practical applications:

The country, whose breeding value estimation is to be checked, has to send information about the statistical model and the data structure to the Interbull audit unit. The audit unit simulates performance and pedigree data with the simulation program and sends those data back to the evaluating country. The country calculates EBVs from this data set and sends them back to the audit unit. From the simulation program, the true solutions are known to the audit unit and can be compared to the estimated values.

Theoretical background of the simulation program

The general strategy for the program is outlined in a paper by Robin Thompson, (1997).

Phenotypic data are calculated with a simple linear model $y_{ij} = b_i + a_j + e_{ij}$ with: y_{ij} = record of the j th animal, b_i = effect of the i th environment, a_j = breeding value of the j th animal and e_{ij} = residual effects.

Fixed effects are calculated using a GLS-estimator $X'X\hat{b} = X'y$. Performance data y_0 can be generated from fixed effects b_0 using $y_0 = Xb_0$. The solutions for \hat{b} are calculated with $\hat{b} = (X'X)^{-1}X'y_0$. If the algorithm is correct, $\hat{b} = b_0$. For a fixed effect model the solutions for \hat{b} always equal the simulated effects. The algorithm is very straightforward and produces always same results.

For mixed model equations, including fixed and random effects, generating y_0 is not as easy as with fixed effects only. If the equation $y_0 = Xb_0$ is

simply enlarged with the random effects to $y_0 = Xb_0 + Zu_0$, it does not have the solution $\hat{b} = b_0, \hat{u} = u_0$. The variance- covariance-matrix of the random effects is A^{-1} and the mixed model equations are:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

It is well known (Henderson, 1975), that the mixed model solutions for the breeding values have the property:

$$1' A^{-1} \hat{u} = 0$$

This, however, is a necessary, but not sufficient condition. For a single trait animal model with one observation for each animal, $Z = I$ and the MME are

$$\begin{bmatrix} X'X & X' \\ X & I + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ y \end{bmatrix}.$$

Premultiplying the left and right hand side with

$$\begin{bmatrix} I & 0 \\ 0 & X' \end{bmatrix} \text{ leads to}$$

$$\begin{bmatrix} X'X & X' \\ X'X & X' + \alpha X'A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ X'y \end{bmatrix}.$$

Subtracting the first from the second row yields

$$\alpha X'A^{-1}\hat{u} = 0$$

which leads to the additional condition

$$X'A^{-1}\hat{u} = 0.$$

All base animals were simulate to be in one 'dummy' category of the fixed effect. Offspring were always generated without mendelian sampling, i.e. the breeding value of a progeny always was exactly the average of the parents.

Fixed effects, but no random error term were added to the simulated breeding value to yield a simulated performance. Solving the MME leads

exactly to the simulated fixed effects and breeding values, so the exact MME solutions are known before any numeric solution.

Missing observations, an unresolved problem

This algorithm is only useful, if all animals have a performance record. In dairy cow breeding, the bulls are an important part in the pedigree and get a breeding value, but they do not have a performance record. In MME the animals without performance record get a zero in the Z-matrix. For the simulation program this problem is not trivial to solve and takes some more time for development of the algorithm.

Development of a program to check mixed model equations

A FORTRAN program was developed to check mixed model equations. At first base animals and their breeding values are simulated. Offspring is generated by randomly selecting male and female base animals with breeding values calculated as shown before. Several generations can be produced. After this step the mixed model equations are set up for all animals using randomly chosen classes of one fixed effect. Base animals are all in one big 'dummy' class.

The algorithm was tested in setting up the whole equation system for 3 generations with 50'000 animal per generation. The simulated data set was checked in calculating breeding values with the program PEST (Groeneveld, 1990) using the same model. The solutions for fixed and random effects matched exactly the simulated values.

The program structure allows to simulate fixed and random effects in two different steps. Input parameters allow to define the genetic and the environmental structure in a flexible way.

The simulation program gives the audit group one big advantage. The performance records are calculated in a way, that the true results of the breeding value estimation are already known. If the audited country is using the correct model and a correctly working algorithm, they produce breeding values, that match the true results. All

different types of mistakes, e.g. a too small number of iteration rounds, can be seen immediately in wrong results.

Next development steps of the program

The international survey showed varying distributions of fixed effects in different breeding schemes all over the world. The program to simulate one fixed effect (equally distributed) and one random effect is already finished. It should not be very difficult to extend the program to several fixed effects with different distributions according to the results of the international survey, which will be the next phase of development.

The problem of animals without performance records needs to be solved in future. A pragmatic solution would be to assign 'dummy observations' to all non-performing animals. However, an exact solution will be developed.

The program does random mating in every generation. The number of male and female animals can be adjusted already. The program should take different genetic structures into account. The international survey gives lots of examples for different genetic structures in dairy cow populations. The first simulation of a realistic population is planned for this fall.

References

- Groeneveld, E. 1990. *PEST User's manual*.
- Henderson, C.R. 1975. Use of relationships among sires to increase accuracy of sire evaluation. *J. Dairy Sci.* 58, 1731-1738.
- König, S. *et al.* 2002. The Interbull audit project part I: Results of an international Survey *Interbull meeting*, Interlaaken.
- Thompson, R. 1997. Generating data to check mixed model equations. *Personal communication*.