J. Dairy Sci. 91:3179–3183
doi:10.3168/jds.2007-0972

A General Method to Validate Breeding Value Prediction Software

H. Leclerc,*[1,2] M. Wensch-Dorendorf,†[2] J. Wensch,‡ V. Ducrocq,* and H. H. Swalve†
*Institut National de la Recherche Agronomique, UR337, Station de Génétique Quantitative et Appliquée, F-78352 Jouy-en-Josas, France
†Institute of Agricultural and Nutritional Sciences, Martin-Luther University, Halle-Wittenberg, 06099 Halle, Germany
‡Institute of Scientific Computing, Technical University Dresden, 01062 Dresden, Germany

ABSTRACT

The validity of national genetic evaluations depends on the quality of input data, on the model of analysis, and on the correctness of genetic evaluation software. A general strategy was developed to validate national breeding value prediction software: performances from a real data file were replaced with simulated ones, created from simulated fixed and random effects and residuals in such a way that BLUP estimates from the evaluation software must be equal to the simulated effects. This approach was implemented for a multiple-trait model and a random regression test-day model. An example was presented on test-day observations analyzed with a random regression animal model including a lactation curve described as a sum of fixed polynomial regression and fixed spline regression on days in milk, and with genetic and permanent environmental effects modeled by using Legendre polynomials of order 2. Residuals had heterogeneous variances, and phantom parent groups were included. This method can be easily extended to other linear models. The comparison of genetic evaluation results with simulated true effects is used to demonstrate the great efficiency and usefulness of the proposed method.
Key words: genetic evaluation, validation, best linear unbiased predictor, random regression

INTRODUCTION

At the national level, the reliability of genetic evaluation depends on a large number of factors, such as 1) the quality of raw data, 2) the data edits, 3) the correctness of the evaluation model and of the software used, and 4) the postprocessing steps. A large number of tests are done at the national level by genetic evaluation centers to ensure the quality of their EBV. For instance, the results from 2 consecutive evaluations are systematically compared. More and more countries, such as France, Germany, or the Netherlands, have developed or are developing a quality management system based on ISO 9001 standards. At the worldwide level, Interbull has been providing international predicted breeding values of dairy bulls since 1995 on the scale of each participating country by using national genetic evaluation results. Input data quality is a crucial issue in international genetic evaluations, because the validity of results from complex genetic and statistical analyses depends on it. Therefore, monitoring and validation of input data are essential for Interbull (Fikse, 2004): data included in international evaluations have to pass a series of stringent tests before acceptance. The consistency of evaluations is assessed by the comparison of breeding values from consecutive evaluations to identify changes larger than expected based on statistical properties of the breeding values (Klei et al., 2002). Genetic trends are estimated to check that national breeding values are unbiased (Boichard et al., 1995). These checks are also used at the domestic level to guarantee the quality of national genetic evaluations and to keep customers satisfied. However, these tests do not guarantee correctness of computation. The diversity and complexity of models used in the various countries to analyze different traits have led to a situation in which new methods of validation of national data, models, or both are increasingly needed. One research project identified by Interbull in 2002 was the development of a general simulation tool to validate national genetic evaluation systems, especially the development of a simulation environment to test breeding value prediction software. With this aim, a program was developed from a strategy described by R. Thompson (Rothamsted Research, Harpenden, UK; personal communication) to simulate data with known breeding values and phenotypes for a single-trait animal model (Täubert et al., 2002) in such a way that BLUP solutions for breeding values should be equal to the simulated ones. This strategy assumes that residuals are zero but with constrained fixed and random effects. This method was later extended to a multiple-trait animal model (Wensch-Dorendorf et al., 2005).

Received December 21, 2007.
Accepted March 26, 2008.
[1]Corresponding author: helene.leclerc@jouy.inra.fr
[2]Equal contribution.

Making use of a simulation tool is not the only option to validate new genetic evaluation software. A simpler alternative may be preferred. For instance, the direct inversion of mixed-model equations (**MME**) for a small data example could verify that the iterative EBV match those obtained from direct inversion. Nevertheless, when the genetic evaluation model is more sophisticated, this does not guarantee that the MME are properly set up, and numerical problems are often detected only on larger data sets. Another option is to compare results from the new software with the ones obtained from reference software. This method is widely used when it is technically possible.

In actuality, a large number of BLUP software programs have been developed worldwide, fulfilling specific needs. For instance, in France, BLUP software (GeneKit, V. Ducrocq, personal communication) was developed to deal with test-day models (**TDM**). Indeed, national genetic evaluation models for dairy traits are increasingly based on TDM instead of 305-d lactation models. A large variety of models have been proposed, differing in 1) how the lactation curve is modeled as a function of DIM [with fixed classes, parametric curves, or semiparametric (spline) curves; White et al., 1999], 2) how the genetic and permanent environmental components are described (fixed or random regression using Legendre or other polynomials), and 3) how heterogeneous residual variances are accounted for. Unfortunately, no general evaluation software including all possible models is available for TDM with very large data sets. Therefore, countries have developed custom software to perform routine genetic evaluations for their own population using TDM. The lack of reference software makes the software validation step complex. Furthermore, the routine genetic evaluation software usually relies on iterative solving algorithms, which makes it even more difficult to debug them and makes the results complicated to verify. For TDM needs, extension from R. Thompson's strategy to random regression situations is not straightforward.

The objective of this paper was to present a general and flexible strategy that could be used to validate the correctness of newly developed genetic evaluation software. Consistent phenotypic data were generated in such a way that BLUP estimates from correct evaluation software were mathematically equal to the simulated effects. This methodology can be considered a helpful tool in the development and further refinement of BLUP software.

MATERIALS AND METHODS

Outline of the Procedure

The starting point is a pedigree file and a data file containing, for each record, the relevant levels, variables, or a combination of both for all effects, the animal's recoded number and permanent environmental effect level, and all other pertinent pieces of information [elements required to compute random regression coefficients; the weight of records; the genetic, permanent environment, and residual (co)variance matrices, etc.]. These files can be real data sets. The procedure to check genetic evaluation software can then be divided into 3 steps:

1. For each effect as well as for one residual per observation, simulate values following the approach described below, leading to a simulated performance record for each record in the data file.
2. Include as input data these simulated performance records in the national genetic evaluation software. Estimates are obtained for all effects included in the model.
3. Compare estimates of fixed effects and predicted random effects from the national genetic evaluation software with the true (simulated) ones. If the resulting breeding values, permanent environmental effects, and all estimable contrasts of fixed effects are identical to the true ones, then the genetic evaluation software can be considered as correct.

Estimation Method

The following multiple-trait linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$ is considered to describe the derivation of constraints enforcing the simulated performance to fulfill the required properties. $\mathbf{y}$ is the vector of observations; $\mathbf{b}$ is a vector of fixed effects; $\mathbf{a}$ is the vector of breeding values following a normal distribution, with $E[\mathbf{a}] = 0$ and $Var[\mathbf{a}] = \mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$; and $\mathbf{e}$ is the vector of random residuals following a normal distribution, with $E[\mathbf{e}] = 0$ and $var[\mathbf{e}] = \mathbf{R} = \mathbf{R}_0 \otimes \mathbf{I}$. $\mathbf{X}$ and $\mathbf{Z}$ are matrices relating $\mathbf{y}$ to the appropriate fixed and genetic effects. $\mathbf{G}_0$ is the covariance matrix for the genetic effects, and $\mathbf{A}$ is the additive genetic relationship matrix.

The MME corresponding to this model are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \quad [1]$$

The MME are analogous to the normal equation in the standard linear model. Indeed, MME were initially obtained by maximizing the posterior distribution of a given $\mathbf{y}$ (Henderson et al., 1959). System [1] can also be rewritten as $\mathbf{F}'\mathbf{F}\mathbf{x} = \mathbf{F}'\mathbf{c}$, with $\mathbf{F} = \begin{bmatrix} \mathbf{R}^{-1/2}\mathbf{X} & \mathbf{R}^{-1/2}\mathbf{Z} \\ \mathbf{0} & \mathbf{G}^{-1/2} \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix}$, and $\mathbf{c} = \begin{bmatrix} \mathbf{R}^{-1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}$, where $\mathbf{M}^{1/2}$ is, for a symmetric general positive definite matrix $\mathbf{M}$, the unique positive