How to Summarize Historic Test-Day Record Information?

Vincent Ducrocq and Maria del Pilar Schneider
UR337 Station de Génétique Quantitative et Appliquée, Département de Génétique Animale, Institut National de la Recherche Agronomique, 78352 Jouy en Josas, France
email: vincent.ducrocq@dga.jouy.inra.fr

Abstract

A method to reduce the size of test-day (**TD**) data used for genetic evaluations for production is proposed. The approach "summarizes" TD performances recorded before a given limit date ("historic records") into a reduced number of pseudo TD records, and then combines the summarized information with more recent data. The computed pseudo TD records and their associated set of random regression coefficients are included in Mixed Model Equations (**MME**) as any other record. The method was applied to test-day milk records of Montbéliarde cows calving from 1988 to 2005. Different scenarios were studied. MME animal solutions obtained using the full dataset were compared with solutions obtained from the transformed data. When all data were considered as historic and were summarized, solutions based on summarized records were identical to initial solutions and the dataset was 5 times smaller. For more realistic situations, reduction in size was moderate (57% when the limit date was January 1, 2000). Correlations among estimated random regression genetic effects were high (>0.998) and even higher (>0.9996) for lactation values. Another benefit of the approach is a faster convergence.

1. Introduction

In France, genetic evaluations for production traits are run three times a year for official purposes and monthly for management purposes (milk recording schemes). Test day (TD) records have been stored in the national database since 1988. About 300 million individual records are available for the largest breed (Holstein). Implementing a TD genetic evaluation in this context is therefore a real computing challenge and complete evaluations will no longer be possible with the same frequency as now. On the other hand, the effect of adding new data between two evaluations is likely to have a negligible impact on proofs of old animals, say, two generations back or more. If the size of the initial dataset could be reduced, a lot of time, computing capacity and costs would be saved.

The objective of the study was to develop a method to reduce the size of the test-day dataset to be used in TD genetic evaluations by summarizing "historic" test-day records while limiting as much as possible the loss of information, and by combining summarized data with new test-day records.

2. Method

First, it is necessary to define what historic data is. The complete TD dataset is partitioned into two subsets by setting a limit date. TD records previous to this limit date constitute "historic information" and TD records posterior to the limit date represent "recent information". Consequently, three types of cows can be defined: a) cows which have only historic TD records (type H), b) cows which have only recent TD records (type R), and c) cows which have a combination of both historic and recent TD records (type HR), i.e., TD records before and after the limit date. TD from these distinct types of cows will be treated differently.
For illustration, consider a single trait random regression animal model. In matrix notation:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wp} + \mathbf{e} \qquad [1]$$

where **y** is the vector of TD records, **b** is the vector of fixed effects, **a** is the vector of N_a additive genetic random regression coefficients, **p** is the vector of permanent environmental random regression coefficients, and **e** is the vector of random residual effects. **Z** and **W** are random regression coefficients matrices and **X** is the incidence matrix for fixed effects. The BLUP mixed model equations (MME) are:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} & \mathbf{X'R^{-1}W} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1} \otimes A^{-1}} & \mathbf{Z'R^{-1}W} \\ \mathbf{W'R^{-1}X} & \mathbf{W'R^{-1}Z} & \mathbf{W'R^{-1}W + P^{-1} \otimes I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \\ \mathbf{W'R^{-1}y} \end{bmatrix} \qquad [2]$$

where **A** is the additive genetic relationship matrix, and **G**, **P** and **R** are the (co)variance matrices for the random effects **a**, **p** and **e**.

2.1. Animals with historic data only

The historic TD records of an animal i will be summarized into a reduced number of n ($n \leq N_a$) pseudo test-day records ys_{ij}, j=1, n. Together with each ys_{ij}, an associated set of random regression coefficients can be derived such that the model:

$$\mathbf{ys} = \mathbf{Z}^*\mathbf{a} + \mathbf{e} \qquad [3]$$

leads to the same MME $\hat{\mathbf{a}}$ solutions for all animals. In [3], **ys** is the vector of $\mathbf{ys_i}=[ys_{i1} \ldots ys_{in}]'$ sub-vectors of pseudo-records. $\mathbf{Z}^*$ is a new incidence matrix equal to the direct sum of individual sub-matrices $\mathbf{Z_i}^*$ containing new sets of random regression coefficients associated to $\mathbf{ys_i}$.

Mixed model solutions are needed before summarizing the historic data of type H cows. Once they were obtained, the following steps are applied to compute first $\mathbf{Z_i}^*$ and then $\mathbf{ys_i}$:

Step 1: The contributions from own performances for each cow i are calculated and simultaneously cumulated, reading the initial data set :

$$\begin{bmatrix} \mathbf{Z_i'R_i^{-1}Z_i} & \mathbf{Z_i'R_i^{-1}W_i} \\ \mathbf{W_i'R_i^{-1}Z_i} & \mathbf{W_i'R_i^{-1}W_i + P^{-1}} \end{bmatrix} \qquad [4]$$

Step 2: Once this is done for all cows of type H, the permanent environment part is absorbed, resulting in a reduced matrix $\mathbf{S_i}$ of order $n \leq N_a$:

$$\mathbf{S_i} = \mathbf{Z_i'}\left(\mathbf{R_i^{-1}} - \mathbf{R_i^{-1}W_i}\left(\mathbf{W_i'R_i^{-1}W_i + P^{-1}}\right)^{-1}\mathbf{W_i'R_i^{-1}}\right)\mathbf{Z_i} \qquad [5]$$

Step 3: After absorption, matrix $\mathbf{S_i}$ is decomposed into a sum of n products of vectors $\mathbf{u_j}$ by $\mathbf{u_j'}$.

$$\mathbf{S_i} = \sigma_e^{-2} \sum_{j=1}^{n \leq N_a} \mathbf{u_j u'_j} \qquad [6]$$

This decomposition is described in the Appendix.

Take $\mathbf{Z_i}^* = \begin{bmatrix} \mathbf{u'}_1 \\ .. \\ \mathbf{u'}_n \end{bmatrix}$

The number n of vectors $\mathbf{u_j}$ in $\mathbf{Z_i}^*$ is less than N_a when the rank of $\mathbf{Z_i}$ is less than its size. In the French model, this is the case for cows with TD only infirst lactation (n=2, N_a=4). With this decomposition, we have:

$$\mathbf{S_i} = \mathbf{Z_i^{*'}R_i^{-1}Z_i^*} \qquad [7]$$

Step 4: Right hand sides ($\mathbf{rhs_i}$) for equations associated to each animal i are then calculated. Based on equation [2], the vector of all right-hand sides $\mathbf{rhs_i}$ can be obtained after correction for fixed effects and absorption of permanent effects. In fact, the same quantity is obtained much more easily as:

$$\mathbf{rhs} = (\mathbf{S} + \mathbf{G^{-1}} \otimes \mathbf{A^{-1}})\,\hat{\mathbf{a}} \qquad [8]$$

Using [7], this right hand side is equal to:

$$(\mathbf{Z^{*'}R^{-1}Z^*} + \mathbf{G^{-1}} \otimes \mathbf{A^{-1}})\,\hat{\mathbf{a}}\,.$$

Step 5: If **ys** represents the vector of pseudo-TD records associated to the coefficients in the rows of matrix **Z***, we must have:

$$\mathbf{Z^{*'}R^{-1}ys} = (\mathbf{Z^{*'}R^{-1}Z^*} + \mathbf{G^{-1}} \otimes \mathbf{A^{-1}})\,\hat{\mathbf{a}} \qquad [9]$$

Therefore, the pseudo test-day records $\mathbf{ys_i}$ for animal i can be obtained as:

$$\mathbf{ys_i} = (\mathbf{Z_i^{*'}R^{-1}})^{-}\,\mathbf{rhs_i} \qquad [10]$$

where the number of pseudo-records depends on the rank of $\mathbf{Z_i}^*$.

2.2. Animals with recent data

TD performances recorded after the limit date from cows of type R and HR are kept without any modification nor transformation. These records are referred as recent TD records (denoted **yr**).

Animals of type HR have TD records before and after the limit date. The strategy used to summarize records implies the absorption of permanent environmental effects (steps 1 and 2, equations [4] and [5]). But the analysis of recent TD records of these animals explicitly requires the incidence matrix for permanent environment effects and the addition of the $\mathbf{P^{-1}}$ matrix in the MME. Direct combination of summarized and unchanged TD would lead to double counting $\mathbf{P^{-1}}$ and to poor and biased estimation of the permanent environment

effects based only on the most recent records. Hence, historic records from type HR cows must not be summarized. But they cannot be kept as the recent records either: in such a case, the contemporary group effects corresponding to records before the limit date would be poorly estimated. They would be based on records of type HR cows only. while summarized records of type H cows are free of contemporary group effects (equation [3]). The easiest alternative consistent with our objective is to correct historic records from type HR cows for all fixed effects:

$$yc_{ij} = y_{ij} - \mathbf{x_{ij}}\hat{\mathbf{b}} \qquad [11]$$

where yc_{ij} is the j^{th} test-day record of animal i with the corresponding incidence vector $\mathbf{x_{ij}}$.

Numerical application

Data. The method was applied to test-day milk records of Montbéliarde cows calving from 1988 to 2005. Data included only records from the French administrative region of Jura. First to third lactations with at least 3 TD records were included in the analysis, and stage of lactation had to be between 7 and 335 days. Here the limit date to summarize the data was arbitrarily set to January 1, 2000. Thus, before this date TD records were considered as historic information.

After editing, the data had 2,648,673 TD records from 148,642 cows in 1,254 herds. The pedigree file included a total of 210,199 animals (7,087 sires and 203,032 cows). Six groups of phantom groups were included.

Model. The model used was a simplified version of the one being implemented in France (Leclerc *et al.*, 2008). It included the fixed effects of herd by test-date (149,613 combinations), month of calving x year x lactation number (588 combinations), and 4 random genetic and 4 permanent environ-ment effects. The (co)variance matrices were derived by Druet *et al.* (2003, 2005). The residual variance was a spline function of days in milk.

Implementation. The internal software used for BLUP evaluations of dairy cattle at INRA, called *Genekit*, was extended to include subroutines to summarize historic information and to correct test-day records for fixed effects. The method can be implemented with only one run of the software. For this study, a second run was done to compare exact solutions from the full data set with those obtained after summarizing historic data, hereafter called "reference solutions". A preconditioned conjugate gradient iterative algorithm was used and convergence was reached when the Euclidian norm of the difference between the solutions from two consecutive runs was less than 10^{-7}.

Different scenarios were considered: s*cenario 0* only aimed at checking the software and consisted in, first, summarizing all TD of all cows and second, applying the reduced model to the summarized data (equation [3]).

In *scenario 1*, TD records before January 1, 2000 were summarized. Thus, summarized (**ys**) and corrected (**yc**) historic records were computed and a dummy class for each fixed effect was defined for them. Then recent (**yr** unchanged) records from 2000 to 2005 were added without modification and the resulting data set was analyzed again.

To check what happens in practice when new data is added to the summarized one, *scenario 2* was considered in which the initial data set was split into two parts. The first part included TD from 1988 to 2004, and the limit date was again set to January 1, 2000. The second part included only TD records from 2005. The method was applied to the first part of the data as in the first run of scenario 1. Then, the 2005 TD records were added and a new evaluation was run. Solutions for each genetic effect and their combinations to get lactation EBV were compared for the different scenarios. For animal i, the later ones were obtained as:

$$EBV_{ik} = \sum_{j=1}^{N_a} \left(\sum_{l=7}^{l=305} c_{jlk} \right) \widehat{a}_{ji} \qquad [12]$$

where k is the lactation number (k=1 to 3), l is day in milk, c_{jlk} is the animal genetic random regression coefficient j at l days in milk in lactation k, and $\widehat{a}_{ji}$ is the solution for the j^{th} genetic effect of animal i. If the method works properly, solutions of second runs of scenarios 0 to 2 should be similar to the reference ones.

Results and Discussion

Table 1 shows the initial number of TD records analyzed for each scenario and the number of historic, corrected and recent records in the final dataset. The size of the data was divided by 5 when all data were summarized (scenario 0). For scenarios 1 and 2, the reduction was of the same order for historic animals but recent and corrected TD lead to an overall reduction which is only moderate (-43%).

Table 1. Number of initial test-day records read and pseudo test-day (ys), corrected test-day (yc) and recent test-day (y) records obtained after the application of the method.

Records	Scenario 0	Scenario 1	Scenario 2
Initial test-day	2,648,673	2,648,673	2,485,544
Summarized (**ys**)	508,193	264,072	264,072
Unchanged (**yr**)	-	1,061,186	898,057
Corrected (**yc**)	-	192,188	192,188
New data 2005	-	-	163,129
Total	508,193	1,517,446	1,517,446
Total/initial	19%	57%	57%

Table 2 illustrates another benefit from using summarized data: not only the data file is smaller, but convergence is significantly faster, with a number of iterations divided by 9 in scenario 0 and by about 2 in the two more realistic scenarios.

Table 2. Computational requirements for the three scenarios (time in seconds).

	Scenario		
	0	1	2
Iterations (run 1)	690	690	673
Time to solve (run 1)	1354	1354	925
Time to summarize (run 1)	649	654	653
Iterations (run 2)	76	294	340
Time to solve (run 2)	35	376	453

Table 3. Overall correlations between estimated random regression (RR) genetic effects and lactation estimated breeding values (EBV).

	Scenario 0	Scenario 1	Scenario 2
1st RR	1.00000	0.99981	0.99981
2nd RR	1.00000	0.99977	0.99982
3rd RR	1.00000	0.99933	0.99826
4th RR	1.00000	0.99869	0.99862
1st lactation	1.00000	0.99966	0.99969
2nd lactation	1.00000	0.99976	0.99977
3rd lactation	1.00000	0.99983	0.99984

Table 3 shows the overall correlations between estimated animal random regression genetic effects and lactation EBV and reference solutions. Correlations were all equal to 1 for scenario 0. Hence, it is possible to summarize the information of a cow with, say, 30 TD historic records over three lactations into 4 pseudo test-day records without loosing any information. Scenarios 1 and 2 gave very similar results, showing that the addition of one year of data had a very limited impact on EBV of "historic" animals. For both scenarios, solutions for the first two RR genetic effects exhibited a correlation larger than 0.9997 with the reference solutions. Correlations were somewhat lower (but still >0.998) for the third and fourth coefficients, but their contribution to the total genetic variance is limited. As a consequence, lactation EBV were very similar to the reference ones (overall correlation >0.9996).

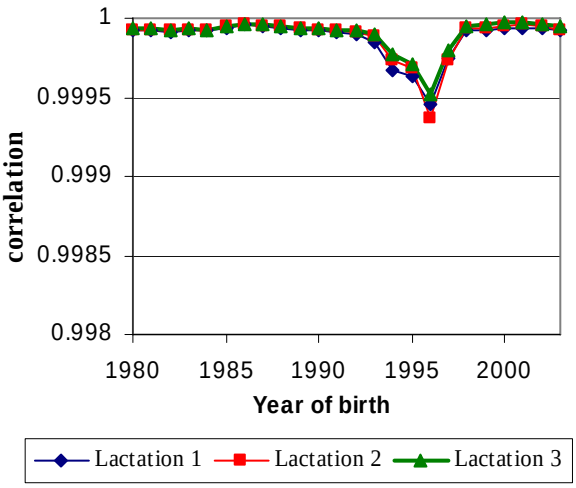

Figure 1. Within year correlations between cow EBV obtained with scenario 2 and reference solutions.

For scenarios 1 and 2, a small decrease in within year correlations was observed when animals having combined information (type HR cows) were born (figure 1). But even for these animals, correlations were larger than 0.9994 for cows and even higher for sires. Furthermore, genetic trends for both cows and sires were the same for all scenarios (results not shown).

The advantage of applying the proposed method is that once data have been summarized, there is no need to re-do it at each evaluation. Scenario 2 mimics a real case when new data is added every month or trimester or year. The slight decrease in correlations can be considered insignificant in the light of data size reduction and computing time saving. Thus, new data can be added successively nearly without consequences on old EBV.

The actual magnitude of reduction in CPU time needs to be assessed in routine situations: here, all evaluations were started from 0 and the proportion of records to summarize in the first run of each scenario was high. In practice, results from previous evaluations will be used, both for "already summarized" historic records and for starting values for MME solutions. The actual number of records to summarize and of iterations to reach convergence will be reduced.

Of course, there is a trade-off when choosing the limit date: an old date leads to limited time and size savings while correlations with reference solutions may be lower when parents of animals with newly added TD have summarized records. An optimal limit date should be determined.

Conclusions and Perspectives

Results showed that the method proposed to summarize historic information works properly. The same MME genetic solutions were obtained. It can be easily implemented and the size of the data and computing time can be significantly reduced.

It should be noted that there are other instances when it is important to summarize TD evaluation results without loosing information. For example, multiple trait analogues of daughter yield deviation (DYD) and equivalent daughter contributions (EDC) are needed for MT-MACE evaluations (Liu *et al.*, 2004). Handling DYDs and EDCs in a way similar to the one proposed here guarantees a very limited loss of information and an easier inclusion into existing software, as they are expressed in terms of pseudo-records and associated RR coefficients (instead of, e.g., EDC matrices).

References

Druet, T., Jaffrézic, F., Boichard, D. & Ducrocq, V. 2003. Estimation of genetic parameters for first parity lactation curves of French Holstein cows. *J. Dairy Sci. 86,* 2480-2490.

Druet, T., Jaffrézic, F. & Ducrocq, V. 2005. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. *Genet. Sel. Evol. 37,* 257-271.

Leclerc, H., Duclos, D., Barbat, A., Druet, D. & Ducrocq, V. 2008. Environmental effects on lactation curves included in a test-day model genetic evaluation. *Animal 2,* 344-353.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. Derivation and calculation of approximate reliabilities and daughter yield deviations of a random regression test-day model for genetic evaluation of dairy cattle. *J. Dairy Sci. 87,* 1896-1907.

APPENDIX
Decomposition of matrix $\mathbf{S_i}$ (equation [6])

For illustration, consider N_a=4 random regression coefficients for the additive animal and permanent environment effects in the model. Let the matrix $\mathbf{S_i}$ be:

$$\mathbf{S_i} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Define:

$$\mathbf{u_1'} = \sigma_e \begin{bmatrix} \sqrt{a_{11}} & \frac{a_{21}}{\sqrt{a_{11}}} & \frac{a_{31}}{\sqrt{a_{11}}} & \frac{a_{41}}{\sqrt{a_{11}}} \end{bmatrix}$$

Then, we have:

$$\mathbf{S_i} - \frac{1}{\sigma_e^2}\mathbf{u_1}\mathbf{u_1'} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & b_{22} & b_{23} & b_{24} \\ 0 & b_{32} & b_{33} & b_{34} \\ 0 & b_{42} & b_{43} & b_{44} \end{bmatrix}$$

This approach can be repeated on the lower right block to successively get vectors $\mathbf{u_2}$, $\mathbf{u_3}$ and $\mathbf{u_4}$ such that:

$$\mathbf{S_i} - \frac{1}{\sigma_e^2}\mathbf{u_1}\mathbf{u_1'} - \frac{1}{\sigma_e^2}\mathbf{u_2}\mathbf{u_2'} - \frac{1}{\sigma_e^2}\mathbf{u_3}\mathbf{u_3'} - \frac{1}{\sigma_e^2}\mathbf{u_4}\mathbf{u_4'} = \mathbf{0}$$

Therefore:

$$\mathbf{S_i} = \frac{1}{\sigma_e^2}(\mathbf{u_1}\mathbf{u_1'} + \mathbf{u_2}\mathbf{u_2'} + \mathbf{u_3}\mathbf{u_3'} + \mathbf{u_4}\mathbf{u_4'}) = \mathbf{Z_i^{*'}}\mathbf{R^{-1}}\mathbf{Z_i^{*}}$$