COMBINING TEST DAY AND FULL LACTATION RECORDS IN PREDICTION OF BREEDING VALUES

E.A. Mäntysaari

MTT Agrifood Research Finland, Animal Production Research,
31600 Jokioinen, Finland

INTRODUCTION

In the Interbull evaluation of dairy production traits in November 2001 five countries (Canada, Estonia, Finland, Germany, and Switzerland) submitted evaluations based on test day (TD) data (Interbull, 2001a). The evaluations in Finland and Canada were based on random regression (RR) test day model (Lidauer *et al.*, 2000; Schaeffer *et al.*, 2000), while the other countries used so called fixed regression model (Interbull, 2000).

Recent "Guidelines for national and international genetic evaluation systems" by Interbull (2001b) recommended that genetic evaluations for production traits should be based on at least of 15 years of data. For TD evaluations this requirement is not easy to fill. In Finland and Canada the TD data starts from 1988, in Germany from 1990 (Interbull, 2000). The interest in costly retrieving of earlier production data is small, because already 10-12 years of data almost guarantees that cows still in production will have their first lactation in the data time span. On the other hand, the best long lasting cows can easily have dams that have had their first records earlier than 12 years ago. In Canadian TD evaluation the cows that have completed their third lactation before 1990 and are still producing, have received indices based on old 305d lactation model. Those are "blended" with TD evaluations so that their standard deviations and means would match the current genetic base (Schaeffer *et al.*, 2000). Another practical consequence of short data time span is a cosmetic problem in estimation of genetic trend. When the trend is plotted for progeny tested sires, the bulls born 4-5 years before the records are available do not seem to have much genetic progress. Those countries that are still on the move into TD model evaluations are likely to have still more problems in the search of old TD records, or in case of large countries, the data files arising from long time span become huge. For them a possiblity to combine TD observations and 305d records would be tempting.

The internationalization of dairy breeding has forced the breed associations administrating small populations to cooperate across national borders. As an example, Nordic countries Finland, Sweden, Denmark and Norway have signed an agreement that gurantees each member an access on the best progeny tested bulls. Such exchange of genetic material logically begs for joint evaluation that can facilitate reliable comparison of breeding values across populations. In Nordic consortium the animal evaluation practice in each population is different. Finland has evaluations based on TD, Denmark uses animal model for first three lactations, and Sweden and Norway are evaluating animals based on first lactation records only. If the joint evaluations are to be implemented now, the only possibility is to use 305d records on some countries and TD records on others.

Countries that have significant import of genetic material have developed procedures to return the foreign information of Interbull proofs back to their national evaluations (Bonaiti and Boichard, 1995). Typically this is done by including the daughter yield deviation information of imported bulls into domestic evaluations. This, however, has not been done in any of the TD evaluations. One reason is that the single trait Interbull evaluations are not found fashionable compared to multitrait multilactation models in TD evaluations, but the other is the complexity of including 305d yield deviations into TD data sets.

The objective of this work is to present a simple approach to combine TD and 305d records in animal evaluations. I do not try to be comprehensive on all types of test day models, but merely give an example how combining works for simple RR model. However, the method can be easily extended to more complex RR models. By a numerical example we give our compliments to pioneering presentation of Schaeffer and Dekkers (1994) in WCGALP 8 years ago; the presentation which truly started the TD evaluation era. The method presented here has been tested using large data sets. Those results are given elsewhere in this conference (Pösö and Mäntysaari, 2002).

MATERIAL AND METHODS

Basic models. Based on a simple RR model the TD trait y_{ijk} of cow k can be considered as a sum of herd test day (htd_i), fixed function of stage of lactation ($\dim_j$), and RR functions of breeding value and non-genetic cow effects:

$$y_{ijk} = htd_i + \sum_{r=0}^{3} \beta_r \phi(\dim_j)_r + \sum_{r=0}^{3} a_{kr} \phi(\dim_j)_r + \sum_{r=0}^{3} p_{kr} \phi(\dim)_r + \varepsilon_{ijk} . \qquad [1]$$

Schaeffer and Dekkers (1994) used functions with covariables for intercept, linear and logarithmic terms. I use here 4 orthogonal polynomials because they will lead into lower dependencies among regression coefficients. Equation [1] can be written in matrix form:

$$\mathbf{y}_k = \mathbf{h}_k + \mathbf{\Phi}_k \boldsymbol{\beta}_k + \mathbf{\Phi}_k \mathbf{a}_k + \mathbf{\Phi}_k \mathbf{p}_k + \boldsymbol{\varepsilon}_k , \qquad [2]$$

where $\mathbf{y}_k$ is a vector of observations of a cow k, Φ_k is matrix of covariables, $\mathbf{h}_k$, $\boldsymbol{\beta}_k$, $\mathbf{a}_k$ and $\mathbf{p}_k$ are vectors of appropriate *htd* effects, fixed regression coefficients, and RR coefficients, respectively. Denote

$$\text{var}(\Phi_k \mathbf{a}_k) = \Phi_k \text{ var}(\mathbf{a}_k) \Phi_k^T = \Phi_k \mathbf{K}_a \Phi_k^T,$$
$$\text{var}(\Phi_k \mathbf{p}_k) = \Phi_k \text{ var}(\mathbf{p}_k) \Phi_k^T = \Phi_k \mathbf{K}_p \Phi_k^T$$

and

$$\text{var}(\boldsymbol{\varepsilon}_k) = \mathbf{I} \sigma_e^2.$$

Now assume that 305 d yield is composed from daily yields of a cow with 10 TD records with standard dims being 15, 45, 75,…,285. Thus, it can be assumed to have a model:

$$y_{305} = 30\mathbf{1}^T\mathbf{y}_l = 30(\mathbf{1}^T\mathbf{h}_l + \mathbf{1}^T\mathbf{\Phi}_l\,\mathbf{a}_l + \mathbf{1}^T\mathbf{\Phi}_l\,\mathbf{p}_l + \mathbf{1}^T\boldsymbol{\varepsilon}_l) = hys_l + \mathbf{\Phi}_{305}\,\mathbf{a}_l + e_l \qquad [3]$$

where hys_l is a "mean" of the *htd* effects in [1], and $\mathbf{\Phi}_{305}$ is a summed vector of covariables of corresponding standard dims, *i. e.*, rows of $\mathbf{\Phi}_l$ multiplied by the average length of the test periods. The non-genetic cow effect ($\mathbf{\Phi}_{305}\mathbf{p}_l$) and measurement error ($30\,\mathbf{1}^T\boldsymbol{\varepsilon}_l$) can be summed to a single residual e_l. The model leads into two alternative methods for describing 305d records by TD model. In a covariable summing approach the 305d record is seen as a special case of TD record, a kind of *unified trait,* [4], and in the *correlated trait* approach the models are written jointly for TD yields and for 305d yields [5]:

$$\begin{bmatrix}\mathbf{y}_k\\ y_{305k}\end{bmatrix} = \begin{bmatrix}\mathbf{h}_k\\ hys_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k\,\boldsymbol{\beta}_k\\ \end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k\\ \mathbf{\Phi}_{305}\end{bmatrix}\begin{bmatrix}\mathbf{a}_k\\ \end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k\,\mathbf{p}_k\\ \end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_k\\ e_k\end{bmatrix} \qquad [4]$$

and

$$\begin{bmatrix}\mathbf{y}_k\\ y_{305k}\end{bmatrix} = \begin{bmatrix}\mathbf{h}_k\\ hys_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k\,\boldsymbol{\beta}_k\\ \end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathbf{a}_k\\ a_{305_l}\end{bmatrix} + \begin{bmatrix}\mathbf{\Phi}_k\,\mathbf{p}_k\\ \end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_k\\ e_k\end{bmatrix} \qquad [5]$$

In both the approaches the $var(e_k) = \mathbf{\Phi}_{305}\mathbf{K}_p\mathbf{\Phi}_{305}^T + 10*30^2\sigma_\varepsilon^2$ and $cov(\varepsilon_k, e_k) = 0$. For multitrait model the $var(a_{305_l}) = \mathbf{\Phi}_{305}\mathbf{K}_a\mathbf{\Phi}_{305}^T$ and $cov(a_{305_l}, \mathbf{a}_k) = \mathbf{\Phi}_{305}\mathbf{K}_a$.

Model parameters.

Van der Werf *et al*. (1998) applied the Kirkpatrick *et al*. (1990) method to derive the variance parameters for RR model. To differentiate their approach from direct fit of covariance functions via RR REML (*e. g*. Jamrozik and Schaeffer, 1997; Kettunen *et al*., 2000) they started to call it 2-step approach. In the first step, a multitrait covariance matrices (*e. g*. **G**), that represent distant stages of lactation are assumed (or estimated). In the second step, covariance functions:

$$\mathbf{G} = \mathbf{\Phi}\,\mathbf{K}_a\mathbf{\Phi}^T \quad \text{and} \quad \mathbf{R} = \mathbf{\Phi}\mathbf{K}_p\mathbf{\Phi}^T + \mathbf{I}\sigma_\varepsilon^2 \qquad [6]$$

are fitted to the covariance matrices, and the coefficients $\mathbf{K}_a$ and $\mathbf{K}_p$ of the functions are taken as variance parameters for RR functions. We fitted the covariance functions on covariance matrices of first lactation TD milk yields estimated on 5 different lactation periods (Lidauer *et al.*, 2000). The lengths of the periods were 15 to 30 days, and the mean dim of the periods were 12, 46, 146, 226, and 315. The covariance estimates were:

$$\mathbf{G} = \begin{bmatrix} 2.414 & & & & \\ 2.300 & 3.306 & & & \\ 2.071 & 3.103 & 4.031 & & \\ 1.768 & 2.662 & 3.729 & 3.750 & \\ 1.149 & 1.361 & 2.257 & 2.435 & 2.813 \end{bmatrix},\ \mathbf{R} = \begin{bmatrix} 9.921 & & & & \\ 4.748 & 8.912 & & & \\ 3.447 & 4.423 & 8.379 & & \\ 2.552 & 3.215 & 4.559 & 8.357 & \\ 1.810 & 2.040 & 2.860 & 3.844 & 10.515 \end{bmatrix} \qquad [7]$$

As the covariance functions are specific to the covariables, the form of $\mathbf{\Phi}$ affects the characteristics of the RR functions. Commonly used orthogonal Legendre polynomials are usually derived for the interval where the data comes from (Kirkpatrick *et al*., 1990). Therefore the linear term was scaled to start from –1.0 at dim=5 and end to +1.0 at dim=350. These covariables, however, are not orthogonal on the interval where the information for 305d yield comes from. To retain the orthogonality the Legendre polynomials $\mathbf{\Phi}_L$ were transformed as $\mathbf{\Phi} = \mathbf{\Phi}_L \mathbf{R}^{-1}$, where $\mathbf{R}^{-1}$ contains the first 4 rows of the QR decomposition of the 10×4 matrix $\mathbf{\Phi}_l$ in [3]. In $\mathbf{QR} = \mathbf{\Phi}_l$ the matrix $\mathbf{Q}$ is orthogonal, and therefore whenever $\mathbf{\Phi}_L$ is $\mathbf{\Phi}_l$ the crossproducts of covariables in $\mathbf{\Phi} = \mathbf{\Phi}_L \mathbf{R}^{-1}$ sum to zero. The covariance functions in [6] were fitted as explained in Mäntysaari (1999). This yielded on following estimates of the covariances:

$$\mathbf{K}_a = \begin{bmatrix} 30.63 & & & \\ -1.146 & 2.76 & & \\ 4.44 & -.251 & 1.60 & \\ .335 & .220 & .2342 & .236 \end{bmatrix}, \mathbf{K}_p = \begin{bmatrix} 40.67 & & & \\ .370 & 6.46 & & \\ 2.95 & -.389 & 2.51 & \\ .219 & .0625 & .2606 & .4569 \end{bmatrix} \text{ and } \sigma_e^2 = 3.721 \qquad [8a]$$

$$\text{var}\begin{bmatrix} \mathbf{a}_k \\ a_{305_k} \end{bmatrix} = \begin{bmatrix} 30.63 & & & & \\ -1.146 & 2.76 & & & \\ 4.44 & -.251 & 1.60 & & \\ .335 & .220 & .2342 & .236 & \\ 2905.7 & -108.72 & 421.03 & 31.759 & 2756559 \end{bmatrix} \text{ and } \text{var}\begin{bmatrix} \varepsilon_k \\ e_k \end{bmatrix} = \begin{bmatrix} 3.721 & \\ 0. & 399509. \end{bmatrix} \qquad [8b]$$

Numerical example. Consider the example data for first lactation cows originally presented by Schaeffer and Dekkers (1994). Their data consisted of 8 cows and 3 sires, and one fixed effect. To illustrate the inclusion of lactation records we added two more animals: parents for one of the parental females, and correspondingly three 305 d records on 3 dams (table 1).

First the data was analyzed using TD records only. The covariables ϕ_1, ϕ_2, and ϕ_3 were fitted as fixed regressions. In addition to them, the mixed model equations (MME) included 6 herd test day and, 52 breeding value and 24 non-genetic animal equations. That yielded the solutions for breeding values listed in table 2. Because the breeding value for 305d was defined as a sum of breeding values of days 15, 45, 75,…, 285, and not as a direct sum of first 305 days of lactation, the first breeding value coefficient is directly related to the 305d breeding value. Actually the 305d EBV is obtained by multiplication of a_o by 94.87. Schaeffer and Dekkers (1994) fitted a model with functions for breeding values only. Also, the genetic parameters they used were different. Although the animals rank almost the same as in their paper, the differences in breeding values between animals are much less. The solutions for

fixed regressions were 27.61, -6.26 and –2.94, and the solutions for hys effects were 18.02, 18.90, 18.54, 16.32, 15.90, and 15.74.

Table 1. Example on test day data of nine cows and three sires (adapted from Schaeffer and Dekkers, 1994)

Htd / Hys	Animal	Sire	Dam	Dim	Regression covariables $\phi(dim)_0$	$\phi(dim)_1$	$\phi(dim)_2$	$\phi(dim)_3$	TD Milk	305d Milk
1	1	9	7	73	0.32	0.28	0.07	0.37	26	-
2	1			123	0.32	0.10	0.32	0.22	23	-
3	1			178	0.32	-0.10	0.32	-0.23	21	-
1	2	10	8	34	0.32	0.43	-0.29	-0.02	29	-
2	2			84	0.32	0.24	0.15	0.39	18	-
3	2			139	0.32	0.04	0.35	0.10	8	-
4	2			184	0.32	-0.12	0.30	-0.27	1	-
1	3	9	2	8	0.32	0.52	-0.62	-0.66	37	-
2	3			58	0.32	0.34	-0.05	0.29	25	-
3	3			113	0.32	0.14	0.29	0.29	19	-
4	3			158	0.32	-0.03	0.36	-0.07	15	-
5	3			218	0.32	-0.25	0.14	-0.39	11	-
6	3			268	0.32	-0.43	-0.31	0.06	7	-
2	4	10	8	5	0.32	0.53	-0.66	-0.76	44	-
3	4			60	0.32	0.33	-0.03	0.30	29	-
4	4			105	0.32	0.17	0.26	0.34	22	-
5	4			165	0.32	-0.06	0.35	-0.13	14	-
6	4			215	0.32	-0.24	0.15	-0.39	8	-
4	5	11	7	14	0.32	0.50	-0.54	-0.48	35	-
5	5			74	0.32	0.28	0.08	0.38	23	-
6	5			124	0.32	0.10	0.33	0.22	17	-
5	6	11	1	31	0.32	0.44	-0.33	-0.08	28	-
6	6			81	0.32	0.25	0.13	0.39	22	-
1	7			0	94.87	0.00	0.00	0.00	-	8750
1	8	13	12	0	94.87	0.00	0.00	0.00	-	5750
1	12			0	94.87	0.00	0.00	0.00	-	6750

Table 2 lists also the breeding value estimates of the animals from an analysis considering also 305d records. In the unified trait covariable summing the MME had one more equation for the hys effect of the 305d lactations. Setting up the MME for the correlated trait approach cannot be done directly with the combined covariance matrices [8b] because the 5×5 matrix for genetic covariances is not full rank. The problem was avoided by adding 0.1 to the genetic variance of 305d yield trait. In the correlated trait approach the number of equations was larger than in the unified trait approach, because of extra 13 breeding values fitted. The solutions for both the methods were the same up to whole output precision. The correlated trait model had a pleasant feature to produce the 305d breeding value estimates for all animals automatically, when the unified approach required multiplication of intercept term.

The 305d records added were for the dams and for the grand dam of the cows 2 and 4 and for the dam of the cow 5. The new records were expected to give more information about the early relatives of the cows, and thus mates of the sires. As seen by results, cow 7 which ranked phenotypically highest (within its 305d comparison group) improved its index, while the bulls she was mated with lost their values slightly.

Table 2. Solutions for the breeding value RR coefficients from the example data

Animal	Solutions using TD data only				Solutions using combined data			
	$94.87 \cdot \hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$94.87 \cdot \hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
1	495.4	-1.06	1.06	-.003	648.6	-1.12	1.26	.009
2	-738.3	1.09	-1.57	-.074	-880.3	1.19	-1.76	-.080
3	-130.0	-.383	-.279	-.103	-208.2	-.310	-.390	-.105
4	-127.4	1.11	-.463	.055	-266.4	1.18	-.654	.044
5	175.4	-.309	.442	.041	356.5	-.433	.708	.053
6	237.2	-.744	.680	.031	300.1	-.785	.763	.035
7	254.8	-.422	.550	.021	719.1	-.631	1.24	.068
8	-288.5	.733	-.676	-.006	-737.3	.930	-1.35	-.053
9	303.5	-.888	.644	-.037	260.5	-.858	.567	-.045
10	-288.5	.733	-.676	-.006	-204.7	.720	-.534	.008
11	18.75	-.157	.159	.031	-13.64	-.170	.109	.025
12	-144.3	.367	-.338	-.003	-404.7	.479	-.729	-.031
13	-144.3	.367	-.338	-.003	-356.7	.460	-.655	-.025

DISCUSSION

The derived and illustrated methods gave logical results. Both methods have also been tested on real data sets. Pösö and Mäntysaari (2002) applied unified trait approach on joint evaluation of Finnish and Swedish Red and White cattle. Villumsen *et al*. (2002) tested correlated trait method on breeding value estimation of 117,657 Danish Holstein cows.

On the small example the convergence characteristics of the unified trait approach (50 iterations) were superior to the convergency of the correlated trait approach (375 iterations). These conclusions might not be extended into real world sized problems. Nevertheless, it is likely that the near singularity of genetic covariance matrix in the correlated trait method will cause problems. Larger tests already applied using the unified trait approach (Pösö and Mäntysaari, 2002) have not indicated any worse convergence characteristics than the same equations without using 305d records. However, the problems in convergence might rise because of the data structures the techniques become applied on. Pösö and Mäntysaari (2002) used the method in an across country evaluation where another population had no TD records available.

We defined the regression covariables which were orthogonal over days in milk of our definition of 305d summed records. This leads into an easy post processing of breeding value

coefficients as the index for 305d breeding value depends only on the intercept for the trait. It should be noted that orthogonalization can be done for other functions, like Wilmink (1987) or Ali and Schaeffer (1984) functions, as well. In Pösö and Mäntysaari (2002) the unified trait covariables were obtained directly from multitrait reduced rank covariables (Lidauer *et al.*, 2000), and thus did not fullfill the orthogonality on standard sampling days for 305d sum. Therefore, each 305d record contributed information on all breeding value coefficients of the animal. It is unclear which covariables lead into better model convergence characteristics. When the breeding value of the total yield is independent on the coefficients defining lactation curve shape, it would be practical to recognize branches of pedigree having 305d records only, and to absorb coefficients of the shape of the lactation curve into coefficients of intercept.
Strict derivation of 305d record assumed that there were no environmental effects in the TD model that would not be accounted by 305d record model. Therefore the heritability of the 305d record might have been overestimated. A natural modification to account for lower accuracy of 305d prediction would be to inflate the var(e_k) to include the uncounted environmental variation. The example here did not illustrate the derivation of var(e_k) when the data includes several lactations per animal. If the lactations are treated as different traits (*e. g*. Schaeffer *et al.*, 2000), the equation [3] has to be modified to account for covariances among residuals of different traits. If the lactations, or some of the lactations, are modeled as repeated observations (Lidauer *et al*., 2000), it is best to form a cow-wise permanent environment effect from the across-lactation non-genetic cow effects ($\mathbf{\Phi}_{305} p_l$ in [3]).

In a typical cases the data might include cows that have test day records only on later lactations. Then the inclusion of early 305d records would help in accounting for the selection and would allow to estimate the true genetic trend also at the beginning of the data collection period. In some cases cows could have missing TD records during their production years. Methods here are relatively easy to modify to properly account for the covariances among non-genetic effects across different lactations. In unified trait model, it would be tempting to simply overparameterize each 305d permanent environment effect, and model them by the same number of covariables as is used for TD records. If the data would include both the TD records and 305d records from the same lactations, the covariances among measurement errors would become hard to trace. In an extreme case, a cow could have all the TD records contributing to 305d yield, and thus the covariance matrix for measurement errors would become singular.

Both the methods presented can be used to model the daughter yield deviations of 305d records. This leads into an easy way to blend foreign information, like Interbull proofs, back to national evaluations.

CONCLUSION

The presented results show that 305d records can be easily combined into genetic evaluations based on TD data. The example here was for simple random regression model, but the methods can be easily extended into much more complicated multiple trait and repeatability models.

REFERENCES

Ali, T. E. and Schaeffer, L. R. (1987) *Can J. Anim. Sci*. **67** : 637-646.

Bonaiti, B. and Boichard, D. (1995**)** *International Bull Evaluation Service* **11.** Uppsala, Sweden.
Interbull, (2000) *International Bull Evaluation Service* **24**. Uppsala, Sweden.
Interbull, (2001a) http://www-interbull.slu.se/eval/framesida-prod.htm
Interbull, (2001b) *International Bull Evaluation Service* **28** . Uppsala, Sweden.
Jamrozik, J. and Schaeffer, L. R. (1997) *J. Dairy Sci*. **80** : 762-770.
Kirkpatrick, M., Lofsvold, D. and Bulmer, M. (1990) *Genetics* **124** : 979-993.
Kettunen, A., Mäntysaari, E. A. and Pösö, J. (2000) *Livest. Prod. Sci*. **66** : 251-261.
Lidauer, M., Mäntysaari, E. A., Strandén I. and Pösö, J. (2000) *International Bull Evaluation Service* **25** : 81-84. Uppsala, Sweden.
Mäntysaari, E. A. (1999) *Proc. 50th EAAP*, Book of Abstracts **5** : 26.
Pösö, J. and Mäntysaari, E. A. (2002) *Proc. 7th WCGALP*.
Schaeffer, L. R. and Dekkers, J. C. M. (1994) *Proc. 5th WCGALP* **18** : 47-48.
Schaeffer, L. R., Jamrozik, J., Kistemaker, G. J. and Van Doormaal, B. J. (2000) *J. Dairy Sci*. **83** : 1135-1144.
Van Der Werf, J. H. J., Goddard, M. E. and Meyer, K. (1998) *J. Dairy Sci*. **81** : 3300-3308.
Villumsen, T., Madsen, P., Jensen, J. and Jakobsen, J. H. (2002) *Proc. 7th WCGALP*.
Wilmink, J. B. M. (1987) *Livest. Prod. Sci*. **16** : 335-348.