Data Subsetting Strategies for Estimation of Across-Country Genetic Correlations

H. Jorjani, U. Emanuelson,* and W. F. Fikse

Interbull Centre, Department of Animal Breeding and Genetics,
Swedish University of Agricultural Sciences, Box 7023,
S-75007 Uppsala, Sweden

ABSTRACT

International genetic evaluation of dairy cattle requires estimation of genetic correlations among populations to account for genotype-environment interaction. Simultaneous estimation of across-country genetic correlations among all populations of a widespread breed, such as the Holstein breed is, however, hampered by connectedness problems and computational challenges. The purpose of this study was to examine the effects of using bulls with across-country, balanced distribution of daughters on estimates of genetic correlations. For this purpose, dairy cattle populations undergoing selection in 6 countries were simulated. Two population-size settings were used. In the small population-size setting (S-populations), the 6 simulated countries had 2000 cows and 20 young progeny testing bulls per generation. In the larger population-size setting (L-populations), the 6 simulated countries had between 2000 and 64,000 cows and 20 to 640 young progeny testing bulls per generation. The simulated (true) across-country genetic correlations, depending on the country combination, varied between 0.5 and 0.9. Simulations comprised a base population and 10 generations and were replicated 16 times. Results for the S-populations were not conclusive. For the L-populations, results indicated that by use of data from a relatively small subset of bulls with distribution of daughters balanced across countries, genetic correlations could be estimated with very small bias (overall average of absolute value of bias across replicates was 0.03 for the L-populations). The suggested bull subsetting strategy would allow simultaneous estimation of across-country genetic correlations to be computed for a larger number of countries and in a shorter window of time than was possible previously.

(**Key words:** international genetic evaluation, genetic correlation, data subsetting, dairy cattle)

Abbreviation key: **AN_EBV** = adjusted number of estimated breeding values, **BDD** = balanced daughter distribution, **CPU** = central processing unit, **EBV** = estimated breeding value, **EDC** = effective daughter contribution, **MACE** = multitrait across-country evaluation.

INTRODUCTION

International genetic evaluation of dairy cattle with the method proposed by Schaeffer (1994) and commonly known as multitrait across-country evaluation (**MACE**) requires estimation of genetic correlations (r_g) among populations to account for genotype-environment interaction (in this paper we use the words “country” and “population” interchangeably). Simultaneous estimation of genetic correlations among populations, however, is hampered by two obstacles. The first one is the computational demand of estimating (co)variance components in a multitrait model such as **MACE**. The second obstacle is poor connectedness among countries. The computational demand is concerned with both memory requirement and processing time. For example, in the current international genetic evaluations of dairy cattle computed at the Interbull Centre, simultaneous estimation of correlations among all Holstein populations (>25 populations, i.e., >25 traits and >100,000 bulls) for production and conformation traits is not feasible. Therefore, resorting to the use of only a subset of data seems inevitable.

For estimation of across-country genetic correlations, 2 different kinds of subsetting strategies can be used. The first subsetting strategy (country subsetting) is to divide populations into smaller groups and estimate correlations for each subset at a time. The other strategy (bull subsetting) is to use only a subset of bulls and estimate correlations based on the selected subset rather than using all bulls.

Use of only a subset of countries reduces the memory requirement, but is not an optimal solution for several reasons. First, the available data are not used opti-

Received February 21, 2004.

Accepted November 9, 2004.

Corresponding author: H. Jorjani; e-mail: Hossein.Jorjani@hgen.slu.se.

*Present address: Department of Epidemiology, Swedish University of Agricultural Sciences, Box 7019, S-75007 Uppsala, Sweden.

mally, i.e., genetic ties might be through bulls from countries not included in the country subset. Second, because of poor ties between certain country combinations, the connectedness problem will be accentuated, and the problem of nonestimability for some of the country subsets will arise. Third, to avoid the estimability issue, the so-called “link provider” countries must be included in all country subsets that contain countries with poor connectedness. Fourth, the computational demand in terms of processing time may actually increase because of the need to estimate correlations for many country subsets. Fifth, the multiple correlations estimated for different country subsets need to be combined with each other. As an example, for 25 countries and 2-country combinations, there are 300 separate 2-country correlation estimates to combine. In the majority of cases, the resulting matrix of all correlations also is not positive definite, and there is a need for bending, for example, by the method described by Jorjani et al. (2003).

International genetic evaluations, as computed at the Interbull Centre, use a combination of these 2 strategies for production and conformation traits (for theoretical considerations and simulation results see Sigurdsson et al., 1996; Klei and Weigel, 1998). The country subset comprises several countries at a time together with one or 2 of the “link provider” countries, usually the USA and Canada or Germany for the Holstein breed. The number of countries in a subset is most often <7. The bull subset comprises bulls from the country subset with national genetic evaluations based on domestic daughters in more than one country, i.e., the so-called “common bulls.” Occasionally, $\frac{3}{4}$ sib families with members in more than one country are used as well (using full-sib families does not lead to inclusion of many more bulls in comparison with “common bulls”). A combination of the 2 above-mentioned subsetting strategies is helpful in the estimation of correlations among a larger number of populations. However, estimation of correlations among all populations of a widespread breed, such as Holstein in one single analysis, is still not feasible. Therefore, it is of practical interest to see if a limited, but well-connected and well-balanced subset of data can be used to obtain empirically sensible estimates of genetic correlations.

Using only one large, well-connected subset of data for estimation of variance components has been an issue of interest and discussion in animal breeding for a long time (Schaeffer, 1975; Eccelston, 1978). Intuitively and theoretically (e.g., Eccelston, 1978), one expects the use of the data in its entirety to be preferred. Counter-intuitively, at least in the area of international genetic evaluation, results indicate that empirically sensible estimates may be obtained even though only a well-

connected subset of data (i.e., bulls) are being used (Sigurdsson et al., 1996; Klei and Weigel, 1998). Using only a well-connected subset of data may be interpreted as an extrapolation of Schaeffer's recommendation (1975) to discard disconnected parts of data from variance component estimation. Because disconnectedness is an extreme case of unbalancedness, one may conjecture that a well-balanced subset of bulls (with respect to their across-country daughter distribution) might be a good way for a more stringent bull subsetting strategy.

The aim of the present study was to examine whether the bull subsetting strategy can be developed further, so that fewer bulls and a larger number of countries, perhaps all of them, can be considered simultaneously in one single analysis with reasonable computational demands.

MATERIALS AND METHODS

Data Simulation

Population structure. A progeny testing scheme was simulated for 6 countries and replicated 16 times. Two population-size settings were used. In one setting, S-populations, all cow populations were small and of equal size (2000 cows per generation). In the other setting, L-populations, cow populations were of varying size (2000 to 64,000 cows per generation). In each country, a base population and 10 further generations of animals were simulated. Population structure parameters, including cow and bull population sizes for each country, are shown in Table 1.

Genetic and phenotypic values. For each animal, 6 breeding values were simulated according to the following model:

$$a|A,G \sim MVN(0, A \otimes G) \quad [1]$$

where a is the matrix of breeding values for all animals in the 6 simulated countries, A is the relationship matrix among all animals, and G is the covariance structure among the 6 countries (Table 2). In other words, breeding values of each animal were simulated as

$$a_{ij} = \frac{1}{2} a_{s_{ij}} + \frac{1}{2} a_{d_{ij}} + \gamma_{ij} \quad [2]$$

where a_{ij} , $a_{s_{ij}}$ and $a_{d_{ij}}$ are the breeding values of the animal i and its sire and dam in country j , respectively, and γ_{ij} is the Mendelian sampling term of animal i in country j . The Mendelian sampling terms (γ) for the base population animals were simulated as

$$\gamma \sim MVN(0, G), \quad [3]$$

Table 1. Number of animals per generation and mating ratios in the 6 simulated countries under the 2 population-size settings.

Country	Population-size setting (number of animals/generation)											
	Small						Large					
	1	2	3	4	5	6	1	2	3	4	5	6
Cow	2000	2000	2000	2000	2000	2000	2000	4000	8000	16,000	32,000	64,000
Young bull	20	20	20	20	20	20	20	40	80	160	320	640
Proven bull	10	10	10	10	10	10	10	20	40	80	160	320
Bull sire	10	10	10	5	5	5	10	10	20	40	80	128
Imported bull ¹	2	2	2	2	2	2	2	4	8	16	32	64
Cow/young bull ²	50	50	50	25	25	25	60	50	40	30	20	10
Cow/proven bull ²	100	100	100	150	150	150	80	100	120	140	160	180
Son/bull sire	2	2	2	4	4	4	2	4	4	4	4	5

¹Importation started at generation 2.

²Mean value of number of cows per bull. Actual numbers were normally distributed with a standard deviation equal to 15% of the mean.

and for animals of later generations as

$$\gamma \sim MVN(0, (0.50 - 0.25(F_s + F_d))G). \quad [4]$$

The symbols F_s and F_d refer to the inbreeding coefficients of sire and dam of each animal, respectively.

To create phenotypic values, the cow population of each country was randomly distributed across 500 herds of equal size (4 cows per herd for all countries in the S-populations and 4, 8, 16, 32, 64, and 128 cows per herd for countries 1 to 6, respectively, in the L-populations). For each cow, only one phenotypic record in the country of birth was simulated. Phenotypic records were the sum of breeding value in the country of birth and values for the herd, country, and residual effects. Herd, country, and residual effects were uncorrelated within and across countries and were normally distributed with mean zero and standard deviations shown in Table 3. The resulting heritability values were 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40 for countries 1 to 6, respectively.

Progeny testing and bull selection. In the base population, bulls were randomly assigned to 2 groups. Depending on the population size, the larger group contained between 20 and 640 bulls, and the smaller group contained between 10 and 320 bulls (Table 1). Part of the bulls from the smaller group, between 5 and 128 bulls, were each mated to between 2 and 5 cows to breed

the next generation of young bulls. Numbers of bulls in the larger and smaller groups in the base population correspond to the numbers of young and proven bulls in later generations. Each mating resulted in only one male offspring, which led to the production of 20 to 640 young bulls for the next generation. Then, bulls from both groups were each mated to an average of 10 to 180 cows, depending on population size and bull category, to breed the next generation of cows. Standard deviation of number of cows per bull was 15% of the average. Each mating resulted in only one female offspring, which led to the production of 2000 to 64,000 cows for the next generation.

In all generations, including the base population, bulls were ranked based on the average simulated breeding values of their daughters. At the time of mating and reproduction, the bulls in the base population were still not selected. When breeding values of their daughters were established, bulls of the base population were subjected to selection, and only a part of them could continue to the next generation. The top-ranking bulls, assigned as proven bulls (including bull sires), together with the newly produced young bulls, were included in the pool of bulls that were allowed to continue into the next generation. The top-ranking bulls among proven bulls were used as bull sires. Number of young and proven bulls (including bull sires) and mating ratios for each of them (Table 1) were the same for

Table 2. Genetic correlation structure among the 6 simulated countries. Variances on diagonals, covariances on lower diagonals, correlation on upper diagonals.

Country	1	2	3	4	5	6
1	90,000	0.90	0.80	0.70	0.60	0.50
2	108,000	160,000	0.90	0.80	0.70	0.60
3	126,000	189,000	275,625	0.90	0.80	0.70
4	136,500	208,000	307,125	422,500	0.90	0.80
5	144,000	224,000	336,000	468,000	640,000	0.90
6	142,500	228,000	349,125	494,000	684,000	902,500

Table 3. Means and SD of herd, country, residual, and additive effects used in simulation of animals in the 6 simulated countries and the resulting phenotypic SD and heritability values.

Country	1		2		3		4		5		6	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Herd effect	0	350	0	400	0	450	0	500	0	500	0	575
Country effect	0	350	0	400	0	450	0	500	0	500	0	575
Residual effect	0	500	0	575	0	650	0	700	0	800	0	825
Additive effect	0	300	0	400	0	525	0	650	0	800	0	950
Phenotype	0	750	0	900	0	1050	0	1200	0	1350	0	1500
Heritability (%)	15		20		25		30		35		40	

all generations. In the ensuing generations, the ranking of the proven bulls was based on the simulated breeding values of their new batch of daughters. Consequently, a proven bull could continue to produce new offspring in the ensuing generations as long as he was among the top-ranking bulls.

Across-country exchange of bulls. Starting from generation 2, semen from a number of bulls (between 2 and 64, depending on the population size) was imported from other countries (Table 1). Country 1 was a pure importer, but countries 2 to 6 were both importers and exporters. The choice of the export country was at random. The probability of being chosen as an export country was equal to 0.10, 0.15, 0.20, 0.25, and 0.30 for countries 2 to 6, respectively. Within each exporting country, the highest-ranking bull was selected for the export. Therefore, it was possible for a top-ranking bull to be exported to up to 5 countries, i.e., to have daughters in all countries.

National Genetic Evaluation. At the end of generation 10, phenotypic records of all cows in each country were used to estimate national breeding values with an animal model. The pedigree information was traced back to the base population. However, parents of the bulls with imported semen were assigned to phantom parent groups based on the birth generation of these bulls and their country of origin. The fixed effect of herd was also included in the national model. For analysis of the national data, the DMU package (Jensen and Madsen, 1994) was used. The simulated genetic and residual (co)variances were used in the analyses to compute estimated breeding values (EBV).

International genetic evaluations. Estimated breeding values of all bulls from each country together with their effective daughter contributions (EDC; Fikse and Banos, 2001), number of herds with daughters, and their pedigree information were used for estimation of across-country genetic correlations. In international genetic evaluation, number of herds is traditionally used as the inclusion-exclusion criterion to ensure that only AI progeny-tested bulls, with a widespread distribution of daughters across herds, are included in the

evaluations. The minimum required value for number of herds depends on the bull category (young bulls with only one batch of daughters, imported bulls or semen with a home country EBV, etc.). For young bulls, the required minimum number of herds is equal to 10.

Estimation of genetic correlations. For estimation of genetic correlations, the method described by Klei and Weigel (1998) with a reduced set of equations was used. The estimation process continued for 10,000 iterations or until the criteria for either one of the 2 convergence measures were met. The 2 convergence measures were $\Delta_{jk} = G_{jk}^n - G_{jk}^{n-1}$ and $\Lambda_{jk} = (\lambda_{jk}^n - \lambda_{jk}^{n-1}) / \lambda_{jk}^{n-1}$, where $\lambda_{jk} = G_{jk} / \sqrt{R_j R_k}$, G and R refer to genetic and residual covariance matrices, respectively; subscripts j and k to the countries; and superscripts n and $n - 1$ to the iteration number. Values of 10^{-6} and 10^{-12} were used for the 2 convergence measures Δ_{jk} and Λ_{jk} , respectively. The starting value of 0.85 was used for all correlations. Based on the structure of data, and because there was no doubt about random distribution of a bull's daughters across herds in the simulations, all bulls with daughters in at least 9 herds were included in the international genetic evaluation, which is very close to the common practice at the Interbull Centre.

Bull Selection Criteria

Based on the arguments put forward in the Introduction, it was postulated that a bull subsetting based on a measurement of balanced daughter distribution (BDD) may result in empirically reasonable estimates. For this purpose, the following measurement was used:

$$BDD_i = 1 - \frac{\sum_{j=1}^{N_c} |n_{ij} - \bar{n}_i|}{2n_i} \quad [5]$$

where BDD_i is the BDD measurement for bull i , n_{ij} is the number of daughters of bull i in country j , n_i is the sum of number of daughters of bull i in all countries, $\bar{n}_i$ is the average number of daughters of bull i in all

Table 4. Number of daughters and the resulting values for balanced daughter distribution (BDD) and adjusted number of estimated breeding values (AN_EBV) for 10 hypothetical bulls.

Bull number	Number of daughters						BDD	AN_EBV
	Country 1	Country 2	Country 3	Country 4	Country 5	Country 6		
1	50	0	0	0	0	0	0.17	1.00
2	1000	0	0	0	0	0	0.17	1.00
3	1000	50	0	0	0	0	0.21	1.29
4	50	50	0	0	0	0	0.33	2.00
5	1000	1000	0	0	0	0	0.33	2.00
6	1000	50	50	50	50	50	0.37	2.20
7	1000	1000	1000	50	50	50	0.55	3.29
8	50	50	50	50	0	0	0.67	4.00
9	50	50	50	50	50	50	1.00	6.00
10	1000	1000	1000	1000	1000	1000	1.00	6.00

countries, and N_c is the number of countries in the evaluation. Preferably, number of daughters could be replaced by the weighting factors used in the international genetic evaluation, i.e., by EDC.

To make the interpretation of BDD values easier, it is possible to multiply BDD value by the total number of countries involved in the evaluation. The resulting measurement, adjusted number of EBV (**AN_EBV**), informally, can be considered as a BDD-adjusted value of the number of national EBV that a bull has, i.e.,

$$AN_EBV_i = N_c * BDD_i. \quad [6]$$

For estimation of genetic correlations in the simulated populations, 3 levels of AN_EBV values were used: AN_EBV ≥ 1.0 (all bulls); AN_EBV > 1.0 (all common bulls), which is equivalent to the current practice at the Interbull Centre; and AN_EBV > 2.0 (common bulls with an across-country balanced number of daughters).

To assess the results, estimated genetic correlations with different bull subsetting strategies were compared with their corresponding true (simulated) values.

RESULTS AND DISCUSSION

Results of using equation [5] for calculation of BDD for some hypothetical bulls are presented in Table 4. From the above equation, one can verify that for a bull with only one national EBV based on domestic daughters (e.g., bulls 1 and 2 in Table 4), irrespective of the bull's number of daughters, the BDD value is equal to $1/N_c$. For bulls with national EBV from more than one country, the maximum possible value of BDD is 1.0. The maximum BDD value can be achieved when there is a completely balanced distribution of daughters across countries (e.g., bulls 9 and 10 in Table 4).

For bulls with only one national EBV, irrespective of the bull's number of daughters, AN_EBV (equation [6])

has a value equal to 1.0. For bulls with national EBV from more than one country, the maximum possible value of AN_EBV is N_c . The more unbalanced the number of daughters in different countries is, the higher becomes the difference between AN_EBV and actual number of EBV. The AN_EBV value will be higher for bulls that have a more balanced distribution of daughters in different countries (e.g., bulls 4 and 5 compared with bull 3, or bull 8 compared with bulls 6 and 7, all in Table 4).

Range of number of nationally estimated bull EBV in the present study for the L-populations (Table 5, 6, and 7) was between 209.9 ± 0.1 and 5971.1 ± 7.3 for countries 1 and 6, respectively. Among the 27 Holstein populations that are currently participating in the international genetic evaluations computed at the Interbull Centre, 20 populations have a number of bull EBV contained in this range. Further, there are 3 populations that participate with < 209.9 bull EBV and 4 populations that participate with > 5971.1 bull EBV. The extent of overlap between population sizes in the actual Holstein populations and in the present study can be interpreted as a sign of usefulness of the present results for the L-populations and their applicability in the real Holstein population analysis. Similarly, there is an overlap between population sizes in the S-populations and some of the smaller breeds, e.g., real Guernsey populations.

Comparison of the number of records (bull EBV) reveals that the more stringent bull subsetting strategy (AN_EBV > 2.0) has reduced the number of common bulls among larger countries. At the same time, the number of common bulls among smaller countries, or between smaller and larger countries, is virtually unaffected (Table 5, 6, and 7). For example, the average number of common bulls between countries 5 and 6 in the L-populations has been reduced from 461.0 ± 3.4 (average of 16 replicates $\pm$ empirical standard errors; Table 6) to 208.3 ± 2.3 (Table 7). The corresponding

Table 5. Average and SE¹ of 16 replicates for number of records (bulls) available from the 6 simulated countries for the small (above diagonal) and large (lower diagonal) population-size settings. Results shown are from a subset of bulls with an adjusted number of estimated breeding values (AN_EBV) ≥1.0.

			Country 1		Country 2		Country 3		Country 4		Country 5		Country 6		Country
Country			Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
1	209.9	0.1	210.0	0.0	5.0	0.0	11.0	0.0	11.0	0.0	11.0	0.0	15.0	0.0	1
2	12.8	0.4	419.6	0.1	210.0	0.0	12.0	0.0	10.0	0.0	14.0	0.0	14.0	0.0	2
3	16.1	0.5	38.7	0.7	839.9	0.1	210.0	0.0	15.0	0.0	16.0	0.0	18.0	0.0	3
4	17.6	0.1	52.3	1.0	95.5	0.9	1679.8	0.1	210.0	0.0	14.0	0.0	19.0	0.0	4
5	17.9	0.1	74.3	1.5	127.6	1.5	215.4	1.4	3359.5	0.2	210.0	0.0	16.0	0.0	5
6	17.2	0.3	104.3	1.2	174.4	1.9	289.4	2.6	461.0	3.4	5971.1	7.3	210.0	0.0	6
Mean SE			Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE			
Country 1			Country 2		Country 3		Country 4		Country 5		Country 6				

¹A combination of small population size and rounding of the numbers resulted in virtually no variation in the number of bulls in the small population size setting.

Table 6. Average and SE¹ of 16 replicates for number of records (bulls) available from the 6 simulated countries for the small (above diagonal) and large (lower diagonal) population-size settings. Data are shown for a subset of bulls with an adjusted number of estimated breeding values (AN_EBV) >1.0.

			Country 1		Country 2		Country 3		Country 4		Country 5		Country 6		Country
Country			Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
1	17.9	0.1	18.0	0.0	5.0	0.0	11.0	0.0	11.0	0.0	11.0	0.0	15.0	0.0	1
2	12.8	0.4	110.5	1.1	24.0	0.0	12.0	0.0	10.0	0.0	14.0	0.0	14.0	0.0	2
3	16.1	0.5	38.7	0.7	183.4	1.8	29.0	0.0	15.0	0.0	16.0	0.0	18.0	0.0	3
4	17.6	0.1	52.3	1.0	95.5	0.9	26.0	0.0	14.0	0.0	29.0	0.0	19.0	0.0	4
5	17.9	0.1	74.3	1.5	127.6	1.5	301.2	2.6	29.0	0.0	16.0	0.0	16.0	0.0	5
6	17.2	0.3	104.3	1.2	174.4	1.9	215.4	1.4	479.4	3.1	33.0	0.0	33.0	0.0	6
Mean SE			Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE			
Country 1			Country 2		Country 3		Country 4		Country 5		Country 6				

¹A combination of small population size and rounding of the numbers resulted in virtually no variation in the number of bulls in the small population size setting.

Table 7. Average and SE¹ of 16 replicates for number of records (bulls) available from the 6 simulated countries for the small (above diagonal) and large (lower diagonal) population-size settings. Data are shown for a subset of bulls with an adjusted number of estimated breeding values (AN_EBV) >2.0.

			Country 1		Country 2		Country 3		Country 4		Country 5		Country 6		Country
Country			Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
1	17.8	0.1	16.0	0.0	4.0	0.0	11.0	0.0	11.0	0.0	11.0	0.0	13.0	0.0	1
2	12.8	0.4	63.4	1.2	16.0	0.0	10.0	0.0	10.0	0.0	10.0	0.0	12.0	0.0	2
3	16.1	0.5	38.5	0.7	111.7	1.2	21.0	0.0	13.0	0.0	14.0	0.0	16.0	0.0	3
4	17.6	0.1	50.9	1.0	92.7	1.1	21.0	0.0	21.0	0.0	14.0	0.0	16.0	0.0	4
5	17.8	0.1	62.0	1.3	110.1	1.2	187.3	1.2	22.0	0.0	22.0	0.0	15.0	0.0	5
6	17.1	0.3	60.9	1.2	106.8	1.2	184.6	1.4	213.8	2.3	24.0	0.0	24.0	0.0	6
Mean SE			Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE	Mean SE			
Country 1			Country 2		Country 3		Country 4		Country 5		Country 6				

¹A combination of small population size and rounding of the numbers resulted in virtually no variation in the number of bulls in the small population size setting.

Table 8. Bias and SE of simulated genetic correlations (estimated-true) for the 16 replicates among the 6 simulated populations. Upper diagonal values for the small and the lower diagonal values for the large population-size settings. Results shown are from a subset of bulls with adjusted number of estimated breeding values (AN_EBV) ≥ 1.0 .

Country	Country 1		Country 2		Country 3		Country 4		Country 5		Country 6	
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1			0.07	0.01	0.15	0.01	0.12	0.03	0.14	0.03	0.05	0.04
2	0.06	0.00			0.08	0.00	0.10	0.01	0.13	0.01	0.05	0.02
3	0.12	0.01	0.06	0.00			0.05	0.01	0.10	0.01	0.05	0.01
4	0.14	0.01	0.08	0.01	0.06	0.00			0.09	0.00	0.12	0.00
5	0.16	0.02	0.10	0.01	0.09	0.00	0.06	0.00			0.06	0.00
6	0.19	0.02	0.12	0.01	0.13	0.00	0.11	0.00	0.07	0.00		

numbers for countries 1 and 6 show little difference (from 17.2 ± 0.3 to 17.1 ± 0.3 ; Tables 6 and 7). On average, number of common bulls among the 4 smallest countries has decreased only by 3%, while the corresponding figure for country pair 5 and 6 is $>50\%$. Reduction of number of common bulls among larger countries is deemed to have negligible effects because even after reduction, there are still enough bulls present in the data for these countries.

Comparison of the genetic correlations (Tables 8, 9, and 10), expressed as the average difference between estimated and simulated values (bias), showed several interesting, though complex, patterns. Using all bulls (AN_EBV ≥ 1.0) in the S-populations (upper diagonal values in Table 8) resulted in both underestimation and overestimation of simulated genetic correlations. Estimates between countries 1 and 2 with country 6 (average of 16 replicates $\pm$ empirical standard errors: 0.05 ± 0.04 and 0.05 ± 0.02 , respectively) were both underestimated and overestimated in individual replicates. Correlations among all other country combinations were overestimated (by up to 0.15 ± 0.01 for countries 1 and 3). In the L-populations (lower diagonal values in Table 8), correlations were invariably overestimated (by up to 0.19 ± 0.02 for countries 1 and 6). For AN_EBV ≥ 1.0 , the overall average of absolute differences between estimated and simulated correlations

across all country combinations was 0.09 and 0.10 for the S- and L-populations, respectively. The values of bias were usually smaller for countries with high correlation ($0.80 < r_g < 0.90$) and larger for countries with moderate correlations ($0.50 < r_g < 0.70$).

When only a subset of bulls was used (AN_EBV > 1.0 and > 2.0 , Tables 9 and 10, respectively), both underestimation and overestimation of genetic correlations could be observed. The maximum underestimation was -0.13 ± 0.02 for countries 2 and 6 in the L-populations (lower diagonal values in Table 9). The maximum overestimation was 0.21 ± 0.01 for countries 1 and 5 in the S-populations (upper diagonal values in Table 10). The overall averages of absolute differences between estimated and simulated correlations across all country combinations were 0.06 and 0.08 for the S-populations for AN_EBV > 1.0 and > 2.0 , respectively. The corresponding values for the L-populations were 0.05 and 0.03 for AN_EBV > 1.0 and > 2.0 , respectively.

Tables 8, 9, and 10 also show, especially in the L-populations, that the more stringent bull subsetting strategy has resulted in smaller bias compared with other bull subsetting strategies. Average absolute bias across all country pairs for the L-populations was 0.10, 0.05, and 0.03 for AN_EBV ≥ 1.0 , > 1.0 , and > 2.0 , respectively. For the countries with > 800 national EBV and $r_g \geq 0.70$, i.e., countries 3 to 6, the more stringent bull

Table 9. Bias and standard error (SE) of simulated genetic correlations (estimated-true) for the 16 replicates among the six simulated populations. Upper diagonal values for the small and the lower diagonal values for the large population-size settings. Results shown are from a subset of bulls with adjusted number of estimated breeding values (AN_EBV) > 1.0 .

Country	Country 1		Country 2		Country 3		Country 4		Country 5		Country 6	
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1			-0.01	0.02	0.09	0.01	0.12	0.02	0.17	0.01	0.10	0.02
2	0.05	0.01			0.02	0.01	0.05	0.01	-0.02	0.02	-0.14	0.02
3	0.05	0.02	0.00	0.01			0.05	0.00	0.04	0.01	-0.05	0.01
4	-0.02	0.03	-0.06	0.01	-0.02	0.01			0.04	0.00	0.02	0.01
5	-0.05	0.03	-0.10	0.01	-0.07	0.01	-0.02	0.01			0.05	0.00
6	-0.06	0.03	-0.13	0.02	-0.09	0.02	-0.03	0.01	0.02	0.01		

Table 10. Bias and standard error (SE) of simulated genetic correlations (estimated-true) for the 16 replicates among the six simulated populations. Upper diagonal values for the small and the lower diagonal values for the large population-size settings. Results shown are from a subset of bulls with adjusted number of estimated breeding values (AN_EBV) >2.0.

Country	Country 1		Country 2		Country 3		Country 4		Country 5		Country 6	
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1			−0.07	0.02	0.08	0.02	0.16	0.01	0.21	0.01	0.21	0.01
2	0.06	0.01			0.08	0.00	0.06	0.01	−0.02	0.01	−0.10	0.02
3	0.07	0.01	0.01	0.01			0.04	0.00	0.01	0.01	−0.04	0.01
4	0.02	0.02	−0.02	0.02	0.01	0.01			0.05	0.00	0.07	0.01
5	0.02	0.03	−0.05	0.02	−0.01	0.02	0.01	0.01			0.07	0.00
6	0.00	0.03	−0.09	0.03	−0.02	0.02	0.01	0.01	0.02	0.01		

subsetting strategy (AN_EBV >2.0) has missed the simulated correlation values by only 0.01 to 0.02 (lower diagonal values in Table 10). In contrast, using all common bulls (AN_EBV >1.0, equivalent to the current practice at the Interbull Centre) has led to a general underestimation of simulated values by up to -0.09 ± 0.02 (between countries 3 and 6, lower diagonal values in Table 9). This is of special interest because in the Holstein populations, which are the main concern of this study, more than 2/3 of countries contribute with 800 or more national EBV for estimation of correlations and point estimates of correlations among all Holstein populations are generally >0.70.

The reason for the low levels of bias in the most stringent bull subsetting strategy (AN_EBV >2.0) probably lies in how the bulls that are common only to the large countries are treated (Tables 5, 6, 7, and 11). For example, number of bulls with 4 to 6 national EBV (Table 11) is virtually the same for the 3 bull subsetting

strategies. However, bulls with only 1 to 3 national EBV from the large countries are prone to be discarded in the more stringent bull subsetting strategy. This can be seen in comparison of bulls with 3 EBV in the S- and L-populations. In the S-populations, the number of bulls with 3 EBV decreased from 15.2 ± 0.9 only slightly to 14.5 ± 0.9 . In the L-populations, the corresponding reduction was from 161.8 ± 2.0 to 113.1 ± 2.2 , indicating that the elimination of bulls is restricted to the bulls that are common only to the large countries. In other words, it seems that the estimation process utilizes the information from 2 groups of bulls. The first group consists of bulls that have nationally computed EBV in small countries or in a combination of small and large countries. The second group consists of bulls that have nationally computed EBV only in large countries. By removing the bulls with a low BDD value, which seem to be the bulls with evaluations in only large populations, the priority is given to the bulls with

Table 11. Number of bulls used in estimation of across-country genetic correlations in the small (S-) and large (L-) population-size settings for different bull subsetting strategies and some indicators of the computational demand (means and SE for the 16 replicates). AN_EBV = Adjusted number of estimated breeding values.

EBV/bull, no.	S-Population						L-Population					
	AN_EBV ≥1.0		AN_EBV >1.0		AN_EBV >2.0		AN_EBV ≥1.0		AN_EBV >1.0		AN_EBV >2.0	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
	(No. of bulls)											
1	1099.8	0.6	0.0	0.0	0.0	0.0	10,718.4	8.8	0.0	0.0	0.0	0.0
2	20.5	0.8	20.5	0.8	0.0	0.0	414.8	3.3	414.8	3.3	0.0	0.0
3	15.2	0.9	15.2	0.9	14.5	0.9	161.8	2.0	161.8	2.0	113.1	2.2
4	10.4	0.7	10.4	0.7	10.4	0.7	68.6	1.5	68.6	1.5	65.6	1.5
5	4.9	0.4	4.9	0.4	4.9	0.4	31.4	1.0	31.4	1.0	31.3	1.0
6	1.3	0.3	1.3	0.3	1.3	0.3	11.1	0.5	11.1	0.5	11.1	0.5
Total bulls, no.	1152.0	0.0	52.2	0.6	31.1	0.4	11,406.2	7.9	687.8	2.1	221.1	2.0
Total EBV, no.	1259.8	0.2	159.9	0.7	116.9	1.1	12,531.8	7.5	1813.4	2.7	824.8	6.2
Average EBV/bull	1.1		3.1		3.8		1.1		2.6		3.7	
Countries, ¹ no.	6.0		5.0		3.0		6.0		5.0		4.0	
Iterations, no.	323.4	70.3	594.9	98.5	731.8	118.7	138.4	16.4	323.2	54.5	314.6	66.5
CPU time/iteration ²	0.254	0.011	0.020	0.001	0.016	0.002	21.801	1.022	1.455	0.053	0.409	0.048

¹Number of countries that contributed to the international genetic evaluations with their domestically born bulls.

²Central processing unit (CPU) time (in s) per iteration.

EBV in smaller countries. If we assume that an unbiased estimate is more likely to come from an unbiased sample of animals, then the question is which of the two groups of bulls constitute a less biased sample of the populations under consideration. It seems that removing those bulls that are common only to the large countries leads to a more balanced subset of data, a subset that is closer to a random sample, and the result is correlation estimates that are closer to the simulated (true) values. If disconnectedness is interpreted as an extreme case of unbalancedness, these results seem to lend some support to the suggestion put forward by Schaeffer (1975) to discard disconnected parts of data.

Table 11 also shows some parameters that are helpful in judging the computational demand of different bull subsetting strategies. For the S-populations, the number of iterations to reach convergence increased with the more stringent bull subsetting strategy, as if there has not been enough information to locate the maximum likelihood. In the L-populations, there was a plateau beyond which it was neither harder nor easier to achieve convergence. In the latter case, near equal number of iterations combined with lower numbers of bulls, and EBV would mean that a larger number of countries can be included in the estimation process and consequently better use of hardware and/or faster estimation can be achieved. This can be seen from the central processing unit (CPU) time needed for computation of each iteration (Table 11).

Results from using a stringent bull subsetting strategy ($AN_EBV > 2.0$) raises the question of whether there is a lower limit for the number of bulls and/or countries necessary for simultaneous estimation of all genetic correlations. In other words, is there any need for all countries to contribute domestically born bulls or is it enough to contribute bull EBV (either from domestic or imported bulls or semen); if this is the case, then how many EBV per bull are required. This is an important issue because, under prevailing conditions in the world's dairy cattle populations, there is a common belief that there might be a difference between those bulls that are most suitable for estimation of within-country variances and those bulls that are actually used for estimation of across-country covariances. The distinction is related to the degree of randomness in the sampling of young progeny testing bulls and in the sampling of bulls with imported semen and how daughters of these 2 groups of bulls are treated by farmers.

In the present study, and for the most stringent bull subsetting strategy ($AN_EBV > 2.0$), there was a need for an average of 31.1 and 221.1 bulls or 116.9 and 824.8 EBV for the simultaneous estimation of all correlations in the S- and L-populations, respectively (Table 11). The resulting average number of EBV per bull was

3.8 and 3.7 for the S- and L-populations, respectively. The average number of bulls per country was 5.2 and 36.9. Considering that there were 15 correlations to estimate, 7.8 and 55.0 EBV per correlation were needed for the S- and L-populations, respectively. Further reduction of data volume was not possible, i.e., nonpositive definite covariance matrices were observed, and nonestimability problems were encountered. There may be an optimum value for the number of countries, bulls, and EBV per bull that ensures estimates of genetic correlations that are as close to the real values as possible. If there are too many bulls with only one EBV ($AN_EBV \geq 1.0$), there is a "dilution" of information from common bulls, probably because of the need for data augmentation. Further, if there are too few EBV per common bull (e.g., 2.6 vs. 3.7 for $AN_EBV > 1.0$ vs. $AN_EBV > 2.0$ in the L-populations; Table 11), there is another sort of "dilution" of information because of the need to estimate too many correlations from too many weakly connected bulls.

Determining a minimum required number of EBV per correlation is of special interest because the direct links (i.e., common bulls) between some country pairs are very weak. In such cases, across-country genetic correlations may be entirely based on indirect links, e.g., pedigree animals, which is speculated to lead to underestimation of genetic correlations (Sigurdsson et al. 1996; Klei and Weigel, 1998) and probably to high standard errors of estimates. In any case, if there is a need for a minimum required number of EBV per correlation, and if the requirement is not fulfilled, then excluding such countries from the analysis might be a wise decision. An international genetic evaluation for such countries may still be carried out in the following way. Across-country genetic correlations among strongly linked and connected countries can be estimated using standard theory (e.g., Klei and Weigel, 1998). All other genetic correlations can be set to some biologically sensible values (for an example see Mark et al., 2003). These two groups of correlations can be combined with each other using a weighted bending approach (Jorjani et al., 2003). The resulting positive-definite matrix can then be used in MACE to estimate international EBV for all bulls in all countries.

In determining a minimum required number of EBV per correlation, the 2 issues of estimability of the parameters, i.e., genetic correlations, and the sensibility of the estimated correlations should be addressed separately, because being able to numerically estimate a correlation does not mean that the estimated correlation is biologically sensible. For estimability it is important that, for each country pair (equivalent to a cell in a 2-dimensional grid), there are some data points to provide some information about the covariance between

those 2 countries. For the estimated covariance to be sensible, e.g., to have low variance and/or bias, probably a large number of randomly sampled data points is needed.

In regard to the estimability issue, the lowest number of EBV per correlation in the present study was the value of 7.8 EBV per correlation obtained in the S-populations (116.9 EBV/15 correlations; Table 11). Ignoring the problems associated with the extrapolation of the results from the S-populations to the real Holstein populations, for simultaneous estimation of all genetic correlations in an analysis with 27 countries, there is a need for 2738 EBV (351 correlations $\times$ 7.8 EBV per correlation) from 721 bulls (2738 EBV/3.8 EBV per bull). Incidentally, this number is in agreement with the findings of Jorjani (2001), who used measurements of the AN_EBV for bull subsetting in simultaneous estimation of genetic correlations among 27 Holstein populations. Jorjani (2001) used AN_EBV values between 2.0 and 8.5, depending on the population size, for bull subsetting and could reduce the number of bulls to 716 bulls with an average of 4.2 EBV per bull (8.6 EBV per correlation). The resulting correlations were not generally different from the usual point estimates obtained at the Interbull Centre, except for the occasional very low correlations between certain country pairs. Further gradual reduction of number of bulls resulted in the selection of 436 bulls with an average of 3.8 EBV per bull (4.8 EBV per correlation). However, using 436 bulls resulted in a larger number of very low correlations. Further reduction of number of bulls to values <436 was not possible because of the nonestimability problem. Therefore, it seems that a minimum of about 5 EBV per correlation are needed to avoid the nonestimability problem. With 5 to 8 EBV per correlation estimation is possible, but we are in danger of serious over- and especially underestimation of genetic correlations.

In regard to the sensibility of the estimated genetic correlations, the results from the L-populations indicate that, to be sure that reasonable results are obtained, there is a need for 55.0 EBV per correlation (824.8 EBV/15 correlations; Table 11). Extrapolating the results from the L-populations to the real Holstein populations, for a 27-country combination, there is a need for 19,300 EBV (351 correlations $\times$ 55.0 EBV per correlation) from 5174 bulls (19,300 EBV/3.7 EBV per bull) for simultaneous estimation of all genetic correlations. This number may be considered as an overestimation, because, in the L-populations, the daughter distribution is extremely unbalanced. The minimum required number of EBV per correlation that ensures both the estimability of the genetic correlations and sensibility of the estimated correlations probably lies some-

where between the values discussed above for estimability (5 to 8) and sensibility (55). This number might actually be closer to 8 than to 55, because, as mentioned before, Jorjani (2001) was able to obtain point estimates similar to the usual estimates obtained at the Interbull Centre for 27 populations in one single run with the use of 8.6 EBV per correlation. Even with numbers as high as 15 to 20 EBV per correlation, simultaneous estimation of correlations among 30 countries with reasonable computational costs would be feasible with the stringent bull subsetting strategy proposed here. Despite some agreement between present results and the results by Jorjani (2001), we should strongly warn that extrapolating to actual Holstein populations must be interpreted with extreme caution.

Observation of both under- and overestimation in the present study is in contrast to the results from the influential study by Sigurdsson et al. (1996) who only observed underestimation of genetic correlations (by up to -0.08 for the simulated values of 0.90). However, the results of the present study add some weight to the previous evidence provided by Klei and Weigel (1998) who observed both under- and overestimation of genetic correlations (by up to ± 0.04 for the simulated correlations of 0.70 and 0.95). Based on the present results and results of Klei and Weigel (1998), one may speculate that the failure of Sigurdsson et al. (1996) to observe overestimation of across-country genetic correlations might have resulted from some particular combination of circumstances in their simulations rather than a genuine outcome of the weak connectedness across countries and "dilution of information."

In the studies by Sigurdsson et al. (1996) and Klei and Weigel (1998) only 2 populations were simulated. For the L-populations the cow and bull population sizes in countries 2 and 3 (44,000 and 88,000 cows in total, respectively) can be compared with the population size used in the study by Sigurdsson et al. (1996), who used 66,000 cows and 600 bulls. Population sizes for countries 4 and 5 (16,000 and 32,000 cows per generation, respectively) can be compared with the population size used by Klei and Weigel (1998), i.e., 24,000 cows per generation. Across-country exchange of bulls, however, was much weaker in the present study compared with the 2 previously mentioned studies. The average number of foreign bulls in each country in the present study was about 8% of the bulls (between 2 to 13% depending on the country). The corresponding value for the study by Sigurdsson et al. (1996) was 50% and for the study Klei and Weigel (1998) was 15 or 50%. The reason for choosing a weaker connection among countries in the present study was the desire to get closer to the prevailing conditions among the countries participating in international evaluations computed at the Interbull Cen-

tre. In regard to the size of under- and overestimations and their interpretation, it is useful to bear in mind that the range of correlations used in the present study (0.50 to 0.90) was larger than the values 0.90 and 0.70 to 0.90 that were used by Sigurdsson et al. (1996) and Klei and Weigel (1998), respectively.

Using the more stringent bull subsetting strategy proposed here (using bulls with AN_EBV >2.0) would most probably lead to more similar bull ranking lists for different countries. The reason is that the current practice at the Interbull Centre (equivalent to using bulls with AN_EBV >1.0) may actually lead to a general underestimation of across-country genetic correlations, especially between small and large countries (lower diagonal values in Table 9).

International exchange of superior genetic material (e.g., frozen semen and embryos) in dairy cattle populations is a very competitive business. New lists of bull rankings are published quite often. The new lists rely on the international genetic evaluations of dairy cattle populations computed at the Interbull Centre. Evaluations, in turn, rely on the across-country genetic correlations, which logistically are expected to be estimated within a relatively short window of time. In the long run, efforts should be directed at the adoption of estimation methods and algorithms that can use all data within the confines of prevailing hardware limitations. However, short-term (and possibly short cut) solutions are also needed. The results presented here indicate that using only a subset of bulls with a well-balanced distribution of daughters may have 2 advantages. The first advantage is that we may actually obtain better estimates of across-country genetic correlations. The second advantage is that we certainly will be able to reduce the computational demand, in terms of memory and CPU time, so that either more countries can be included in any multicountry estimate of genetic correlations or estimation of genetic correlations among the same number of countries can be computed in a shorter window of time.

CONCLUSIONS

The main purpose of this study was to see how far one can discard the data and still obtain reasonable estimates of across-country genetic correlations for a larger number of countries. Results indicate that a shift from the current bull subsetting strategy (use of all "common bulls") to a more stringent bull subsetting strategy, i.e., using fewer bulls and all countries in one single analysis, is feasible. However, the bulls used in the more stringent bull subsetting strategy must have

a more balanced daughter distribution than the average "common bull." Further, it may be concluded that the more stringent bull subsetting strategy not only may lead to better estimates of across-country genetic correlations, but also the estimates can be computed relatively faster. Judging from the small differences between simulated and estimated values (especially for larger populations), it can be recommended that the bull subsetting strategy based on BDD and AN_EBV presented here is adopted for routine estimation of across-country genetic correlations. The exact value of AN_EBV to adopt very much depends on the number and structure of the populations involved.

ACKNOWLEDGMENTS

The authors thank Bert Klei (formerly at the Holstein USA) and our colleagues at the Interbull Centre team, especially Thomas Mark, for much help, fruitful discussions, and valuable feedbacks. Constructive suggestions by 2 anonymous reviewers are also acknowledged. H. Jorjani and W. F. Fikse also acknowledge financial support provided by a USDA/National Association of Animal Breeders (NAAB) research grant during initial stages of this study.

REFERENCES

- Eccelston, J. A. 1978. Variance components and disconnected data. *Biometrics* 34:479–481.
- Fikse, W. F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. *J. Dairy Sci.* 84:1759–1767.
- Jensen, J., and P. Madsen. 1994. DMU: A package for the analysis of multivariate mixed models. In *Proceedings of the 5th World Congr. on Genetics Appl. to Livest. Prod.*, University of Guelph, Ontario, Canada 22:45–46.
- Jorjani, H. 2001. Simultaneous estimation of genetic correlations for milk yield among 27 Holstein populations. *Proc. of the Interbull Open Mtg.*, Budapest, Hungary. August 30–31, 2001. *Interbull Bull.* 27:80–83.
- Jorjani, H., L. Klei, and U. Emanuelson. 2003. A simple method for weighted bending of genetic (co)variance matrices. *J. Dairy Sci.* 86:677–679.
- Klei, L., and K. A. Weigel. 1998. A method to estimate correlations among traits in different countries using data on all bulls. *Proc. of the Interbull Open Mtg.*, Rotorua, New Zealand. January 18–19, 1998. *Interbull Bull.* 17:8–14.
- Mark, T., P. Madsen, J. Jensen, and F. Fikse. 2003. MACE for Ayrshire conformation: Impact of different uses of prior genetic correlations. *Proc. Interbull Tech. Worksh.*, Beltsville, MD. March 2–3, 2004. *Interbull Bull.* 30:126–135.
- Schaeffer, L. R. 1975. Disconnectedness and variance component estimation. *Biometrics* 31:969–977.
- Schaeffer, L. R. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci.* 77:2671–2678.
- Sigurdsson, A., G. Banos, and J. Philipsson. 1996. Estimation of genetic (co)variance components for international evaluation of dairy bulls. *Acta Agric. Scand., Sect. Anim. Sci.* 46:129–136.