On the Dilution of Information

Hossein Jorjani
Interbull Centre, Department of Animal breeding & Genetics
Swedish University of Agricultural Sciences, Box 7023, S-750 07, Uppsala, Sweden
Hossein.Jorjani@hgen.slu.se

Abstract

In this study the effect of number of countries included in an analysis on the estimated genetic correlations among countries was examined. National genetic evaluation results submitted to the Interbull Centre for the test evaluation of female fertility traits in September 2007 were used. Results indicate that higher number of countries has a small (negligible) negative effect on the correlations among strongly connected countries, but a large positive effect on the correlations among weakly connected countries.

Introduction

For several reasons estimation of across country genetic correlations at the Interbull Centre (and Holstein Association, USA) is based only on a sub-set of data. First, because of heavy computational demands; using the whole data in estimation of correlations is in need of prohibitively expensive computers and impossible to perform in a reasonable amount of time with ordinary computers. Second, results of simulation studies (Klei and Weigel, 1998; Jorjani *et al.*, 2005) have shown that unbiased estimates of genetic correlations can be obtained by various, even drastic, methods of sub-setting. Third, results of other simulation studies (Sigurdsson *et al.*, 1996) have shown that using the whole data may actually lead to biased estimates. The latter reason must seem counter-intuitive to animal breeders. The reason is that, given the levels of connectedness among the world's dairy cattle populations (Jorjani, 1999; Fouilloux, 2008) and more importantly, the "non-fixed" models that we use in animal breeding, the recommended strategy is to use the whole data (Schaeffer, 1975) rather than a sub-set of it. So, why are we considering the use of sub-setting as non-controversial and natural?

The explanation lies in the speculation that there are two sorts of "data dilution" or "dilution of information" in the international genetic evaluation (Jorjani *et al.*, 2005), both of which have been conjectured to be connected to the process of "data augmentation" and going back and forth between "incomplete data" and "complete data" in different iterations of the EM-REML algorithm. If all bulls are used, then the "complete data" would be more than 20 times larger than the "incomplete data".

From a practical point of view, this is a re-current problem in the international genetic evaluations conducted at the Interbull Centre, which manifests itself in different estimates of correlations between any two country-traits when they are estimated in different constellation of traits. An example is the correlation estimated between the cow conception traits from Czech Republic and France in the Holstein breed when they have been estimated in 5-variate and 16-variate analyses. There are numerous other examples of such correlations in all breeds and trait groups, listing of which would be exhaustive. In any case, we are confronted with an "unconfirmed hunch" that correlations estimated with lower number of countries are generally higher than the correlations with higher number of countries.

The purpose of this study was to examine the effect of level of connectedness and number of observations per number of estimated parameters. For this purpose the data from 16 countries submitted for the test evaluation of female fertility traits in September 2007 was used. Estimated

correlations in different constellations of 4-, 8-, 12, and 16-variate analyses were compared to see if any general trend in the level of estimated correlations can be observed.

Material and Methods

Data used in this study was the data submitted to the Interbull Centre for Trait 4 (cow conception) of the female fertility test-evaluation in September 2007. Countries submitting data, trait definitions, reported heritability values, number of qualifying records, and average number of common bulls (CB) with all other countries are shown in Table 1. Based on the average number of common bulls with all other countries, four country blocks were formed (Table 2):

Table 1. Description of data used in this study including abbreviation of country names, trait definition, reported heritability values, number of submitted records and the average number of common bulls between each country and all other countries in this study.

Country	Trait definition	h^2	# of records	CB
BEL	PR=Pregnancy Rate (=[21/(DO-45+11)]*100, with DO=days open)	0.040	1497	396
CAN	FC=Interval first insemination-conception in cows	0.077	3589	346
CHE	NR=Non Return Rate after 56 Days (NRR), %	0.010	1135	242
CHR	NR=Cows' Non Return Rate after 56 Days (NRR), binary	0.013	1208	163
CZE	CR=Cows' Conception rate (pregnant or not after 3 months)	0.030	1661	180
DEU/AUS	FL=Interval from first to last insemination cows (days)	0.010	15158	614
DFS	FL=Interval from first to last insemination cows (days)	0.020	12325	413
ESP	DO=Days open	0.045	3614	629
FRA	CR=Cows' Conception rate (binary trait) for cows	0.020	10738	522
GBR	CI=days between 1st and 2nd calvings	0.033	4438	574
IRL	CI=Calving interval	0.037	2257	368
ISR	CR=Inverse of the number of insemination to conception (%)	0.024	1009	39
ITA	CI=Calving Interval (days)	0.038	6249	473
NLD	CI=Calving Interval (days)	0.145	10506	539
NZL	CM=Lactating cow's ability to conceive (CR42)	0.030	5234	330
USA	DP=Daughter Pregnancy Rate	0.040	35125	694

Table 2. Country blocks used in this study and number of bulls with more than 1 record within block.

Block	Countries	Number of bulls with more than 1 record within Block		
		2 records	3 records	4 records
I	USA, ESP, DEU, GBR	1843	608	400
II	NLD, FRA, ITA, DFS	822	273	286
III	BEL, IRL, CAN, NZL	544	175	121
IV	CHE, CZE, CHR, ISR	295	49	3

Standard software (MACE package, Bert Klei, formerly at Holstein Association, USA) for estimation of across country genetic correlation (Klei and Weigel, 1998) was used to estimate genetic correlations using all possible combination of country blocks I-IV. Consequently, there were 15 independent unique runs including 4, 8, 12 or 16 countries in each run (as described in bold faced fonts in Table 3).

Table 3. Description of 15 independent analyses performed to estimate across country genetic correlations among the 16 countries considered in this study.

BLOCK	4-variate	8-variate	12-variate	16-variate
I USA-ESP-DEU-GBR	**I**	**I+II** **I+III** **I+IV**	**I+II+III** **I+II+IV** **I+III+IV**	**I+II+III+IV**
II NLD-FRA-ITA-DFS	**II**	II+I **II+III** **II+IV**	II+I+III II+I+IV **II+III+IV**	I+II+III+IV
III BEL-IRL-CAN-NZL	**III**	III+I III+II **III+IV**	III+I+II III+I+IV III+II+IV	I+II+III+IV
IV CHE-CZE-CHR-ISR	**IV**	IV+I IV+II IV+III	IV+I+II IV+I+III IV+II+III	I+II+III+IV

Results and Discussion

Generally speaking, connectedness in the material used in this study seems to be very strong compared to the previous studies using simulated data (e.g. Jorjani *et al.*, 2005) or field data (e.g. Jorjani, 2001). In comparison with the previous studies using field data, several arguments can be use to explain the differences. First, it might be the case that the connectedness among countries has improved during the past eight years. Second, some of the poorly connected country combinations, e.g. Finland and Israel, are now better connected through joint evaluation in the Nordic countries. Third, some of the other poorly connected country combinations, e.g. Estonia and Hungary, are totally absent in this study. In any case, the final result is that in the most poorly connected block (Block IV: CHE-CZE-CHR-ISR) there are 295, 49 and 3 bulls with 2, 3 and 4 records, respectively, which provide reasonable number of observations for estimation of covariances (Table 2).

Numbers of (co)variance components to estimate were 10, 36, 78, and 136 in the 4-, 8, 12-, and 16-variate analyses, respectively. Ratio of number of records / number of (co)variance components to be estimated in each analysis is shown in Table 4 (which corresponds to the description of analyses in Table 3).

Table 4. Ratio of total number of bull records to the number of (co)variance components in different analyses performed in this study (for description of different analyses see Table 3).

BLOCK	4-variate	8-variate	12-variate	16-variate
I USA-ESP-DEU-GBR	2761	1517 939 833	766 724 460	452
II NLD-FRA-ITA-DFS	1984	1517 693 606	766 724 344	452
III BEL-IRL-CAN-NZL	361	939 693 148	766 460 344	452
IV CHE-CZE-CHR-ISR	103	939 606 148	724 460 344	452

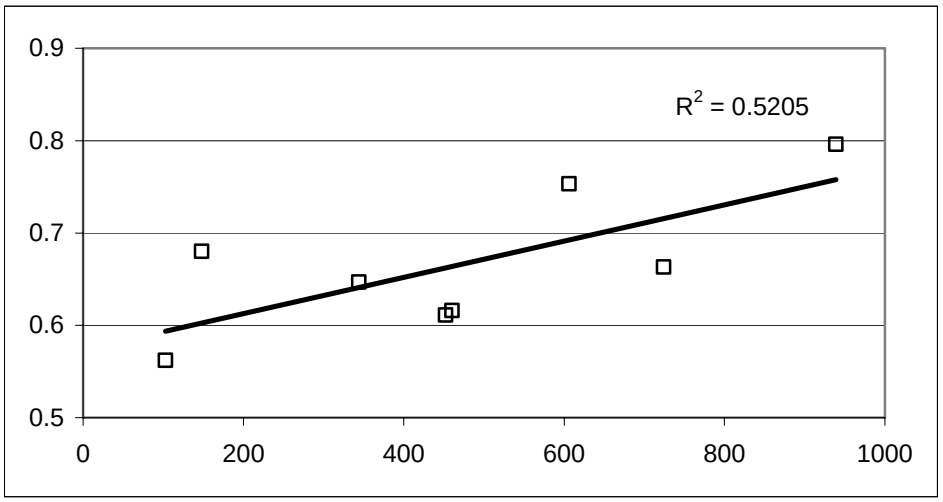
In case of 4-variate analyses (Column 1 in Table 4) a distinct decreasing trend in the available amount of information is observed (from 2761 to 103), i.e. less information to estimate any (co)variance component. The same kind of decreasing trend can be observed for the country Block I (Row 1), when the information from countries of Block I is used to estimate more and more components. Even within each cell of the first row we can observe progressively less information when the data from the strongly linked countries of Block I are mixed with the data from more poorly linked countries of Block II, III and IV (lines 1, 2, and 3, within each cell of Row 1). However, the trend observed for Block I (Row 1) starts to dissipate for Block II, and disappears completely in Block III and IV.

Table 5. Average of within block across country genetic correlations estimated in different analyses performed in this study (for description of different analyses see Table 3).

	BLOCK	4-variate	8-variate	12-variate	16-variate
I	USA-ESP-DEU-GBR	0.860	.841 .861 .864	.842 .843 .862	0.844
II	NLD-FRA-ITA-DFS	0.691	.709 .691 .699	.709 .711 .696	0.710
III	BEL-IRL-CAN-NZL	0.669	.664 .585 .678	.623 .652 .584	0.617
IV	CHE-CZE-CHR-ISR	0.562	.796 .753 .680	.663 .616 .647	0.611

Only average of within block correlations are reported here (Table 5). Among the 4-variate analyses, Block I shows the highest average correlation and a trend toward lower average correlations can be observed when we move from Block I to Block IV. However, this decreasing trend cannot be explained only by dilution of information, because even the nature of traits is different in different Blocks. Of the five countries submitting “rate traits”, except for France, all of them are included in Block IV (“rate traits” have generally lower h^2 values and correlations among “rate traits” are generally lower than among “interval traits”). Therefore, the trend should be sought in Rows. One must bear in mind that all correlations presented in a row belong to the same Block. For example, each of the eight correlations reported in Row 1 of Table 5 are the average correlations of the four countries of Block I.

Close look at Table 5 shows two trends. The first trend can be observed in Row 1, where dilution of information has resulted in a general decrease (though erratic) of average genetic correlation from 0.860 in the 4-variate analysis to 0.844 in the 16-variate analysis. The more pronounced trend can be seen for Row 4 in which moving away from the 4-variate analysis counteracts data dilution and there is a correlation of 0.72 between the ratio of number of records per number of (co)variance components and the estimated correlations.



Conclusions

Within the range of connectedness values found in data used in the present study dilution of information for strongly connected countries had small (and probably negligible) negative effects. In other words adding to the number of countries in an analysis has marginal effects on the correlations among strongly connected countries. Further, adding to the number of countries in an analysis has large positive effect on the correlations among poorly connected countries.

References

Fouilloux, M.N., Dassonneville, R., Minéry, S., Mattalia, S., Laloë, D. & Fikse, W.F. 2008. To be connected or not? Answers for dairy cattle international genetic evaluations. Proc. of Interbull Open Meeting, Niagara Falls, USA, June 16-18, 2008. *Interbull Bulletin 38*, In press.

Jorjani, H. 1999. Connectedness in dairy cattle populations. Proc. of Interbull Open Meeting, Zurich, Switzerland, August 26-27, 1999. *Interbull Bulletin 22*, 21-24.

Jorjani, H. 2001. Simultaneous estimation of genetic correlations for milk yield among 27 Holstein populations. Proc. of the Interbull Open Meeting, Budapest, Hungary. August 29-31 2001. *Interbull Bulletin 27*, 80-83.

Klei, L. & Weigel, K.A. 1998. A method to estimate correlations among traits in different countries using data on all bulls. Proc. of Interbull Open Meeting, Rotorua, New Zealand, January 18-19, 1998. *Interbull Bulletin 17,* 8-14.

Schaeffer, L.R. 1975. Disconnectedness and variance component estimation. *Biometrics 31,* 126-135.

Sigurdsson, A., Banos, G. & Philipsson, J. 1996. Estimationof genetic (co)variance components for international evaluation of dairy bulls. *Acta Agric. Scand., Sect. Anim. Sci. 46,* 129-136.