

Connectedness in Dairy Cattle Populations

Hossein Jorjani

Interbull Centre

Department of Animal Breeding and Genetics

Swedish University of Agricultural Sciences

S - 750 07, Uppsala, Sweden

Hossein.Jorjani@hgen.slu.se

Abstract

Good estimates of genetic correlations between countries are a pre-requisite of successful international bull evaluation. In this paper some ideas on how to identify well connected subsets of data and how to estimate genetic correlations in disconnected subsets are discussed.

Introduction

Implementation of an international breeding value evaluation of bulls with the method known as Multiple [trait] Across Country Evaluation (MACE), as is done in the Interbull Centre, requires good estimates of genetic correlations. However, estimation of genetic correlation between estimated breeding values of bulls in the various countries is problematic for several reasons. One of these reasons is what can be generally classified as poor connectedness. The aims of this paper are:

- 1) To report on the state of connectedness between various countries;
- 2) To present the preliminary results of a method for dividing countries into sub-groups; and
- 3) To examine the feasibility of a recently introduced concept, Genetic Covariance Structure, to estimate genetic correlations between countries.

Connectedness

As a first step to examine the connectedness in the international bull populations, a complete fixed effect model with no interaction was considered. Suppose we have a large a number of bulls, B_i ($i=1, \dots, n_b$), each of which has at least one proof in one of a number of countries, C_j ($j=1, \dots, n_c$). In each of the cells of the resulting bull-country grid (Figure 1) we will have number of daughters used in the

bull's evaluation(s). If a cell is filled, that is if a bull has a proof in a country, then the bull-country cell, BC_{ij} , can take a value of 1, and if the bull has no proof in that specific country, $BC_{ij}=0$. According to Searle (1971, 1987) if all linear contrasts between all levels of each of the fixed effects are estimable, then the data is connected and consequently mean of each cell, whether filled or empty, is estimable and has a BLUE, otherwise the data is disconnected.

Figure 1. Schematic representation of the g-connectedness method in a bull-country grid (see the text).

	C_j	1	2		N_c
B_i					
1		73			
2		23			
			57		29
N_b		11			47

For ascertaining if the international bull populations are connected or not the concept of geometric connectedness introduced by Weeks and Williams (1964) was implemented for a 2-way classification with an algorithm proposed by Fernando et al. (1983). The results showed that each of the world bull populations of Ayrshire, Brown Swiss, Guernsey, Holstein, Jersey, and Simmental, as are represented in the data submitted to the Interbull Centre,

actually consists of one single connected set of data. Because all bulls in each of the populations proved to be connected to each other, there is no reason for me to engage in a discussion of differences of the “connectedness concept” in fixed and random models. However, it should be mentioned that the final judgement on the geometric connectedness and the implemented algorithm should be that it is too crude and too qualitative. Furthermore, it is not an appropriate method for assessing connectedness in a sire model for international evaluations, because one single bull with proofs in two countries makes those two countries connected. As an example, the data in Table 1 shows number of bulls and number of proofs for each bull for milk and fat yield in the Ayrshire bull population, which is represented in eight countries. It is noteworthy that there is not a single bull with proofs in all eight countries and only nine bulls with proofs in more than half of the countries.

Table 1. Number of bulls and number of proofs per bull in the Ayrshire bull populations

Number of bulls	Number proofs / bull
13588	1
227	2
48	3
15	4
9	5
4	6
2	7

The low number of bulls with multiple proofs has its consequences in the form of high variances of estimated parameters and, also from a practical point of view, slow rates of convergence or lack thereof. The general picture one gets from the population structure is that there are few bulls common between countries. In a full bull-country grid

$$\sum_{i=1}^{nc} \sum_{j=1}^{nb} BC_{ij} / n_b \cdot n_c = 1,$$

and in a completely empty bull-country grid

$$\sum_{i=1}^{nc} \sum_{j=1}^{nb} BC_{ij} / n_b \cdot n_c = 0.$$

Table 2 shows the number of filled cells (in percentage) for the six bull populations.

Table 2. Percent of filled cells in bull-country grid for the six bull populations

Breed	# of countries	# of Bulls	# of proofs	% of filled cells
AYS	8	13893	14329	12.89
BSW	10	8344	9337	11.19
GUE	4	1268	1362	26.85
HOL	23	85336	98515	5.02
JER	7	6439	7019	15.57
SIM	8	24568	25774	13.11

The results from implementation of Weeks and Williams (1964) method, however, are encouraging because of two reasons. The first reason is that the results proved that the problem we are facing is not un-estimability of parameters. The second reason is that the implemented algorithm had the nice property that it could be used in a screening process without the need to resort to the more sophisticated computations which are necessary for measurement of connectedness with the more recent methods (see, for example, Hanocq and Boichard, 1999).

A desired method of measuring connectedness to be used in the international bull evaluations should preferably have the following properties:

- 1) It should be quantitative, preferably bound between 0 and 1;
- 2) It should be able to yield a measure of connectedness for individual bulls (in contrast with connectedness between management groups, i.e. countries in our case);
- 3) It should be able to take into account some kind of weighting factor, examples of which are number of daughters, national reliabilities, nationally obtained (genetic or phenotypic) parameters or a combination of such factors;
- 4) It should be able to reflect genetic relationships between bulls by incorporating some information from the relationship matrix; and finally,
- 5) It should be able to avoid all other steps that are included in the international bull evaluations so that it could be used in the

screening process (included in the usual data checks).

Unfortunately, none of the existing methods of measuring connectedness possesses all of the above properties. Therefore, it seems that choosing one of the existing methods and modifying it would be a reasonable solution.

Genetic similarity

The concept of genetic similarity was introduced recently, although in a different framework, by Rekaya et al. (1999) as:

$$GS_{ij} = \frac{\sum_{k=1}^2 \sum_{l=1}^{CB_{ij}} ND_{lk}}{\sum_{k=1}^2 \sum_{l=1}^{TB_{ij}} ND_{lk}}$$

in which ND_{lk} is the number of daughters for each bull, CB_{ij} is the number of bulls with multiple proofs in countries i and j , and TB_{ij} is the total number of bulls in countries i and j .

Implementation of the genetic similarity concept to the Ayrshire bull populations leads to the following results (Table 3).

Table 3. Degree of genetic similarity between pairwise country combinations in Ayrshire bull populations

	F I N	N O R	S W E	U S A	N Z L	A U S	G B R
CAN	.01	.00	.01	.37	.16	.29	.18
FIN		.00	.05	.01	.00	.00	.00
NOR			.04	.00	.00	.02	.00
SWE				.01	.00	.06	.00
USA					.07	.10	.09
NZL						.28	.23
AUS							.27

Examination of the results shown in Table 3 indicates that perhaps these eight countries can be divided in two groups. One group consists of CAN, USA, NZL, AUS and GBR, and the other group of FIN, NOR and SWE. In order to see if this dividing has any effect on the estimates of genetic correlations, I estimated genetic correlations for these countries, first, for all of them together and

then, in the above mentioned two groups. The results are shown in Table 4.

Table 4. Estimates of genetic correlations between Ayrshire bull populations. Upper diagonals: all countries together, lower diagonals: countries divided in two parts as described in the text

	C A N	F I N	N O R	S W E	U S A	N Z L	A U S	G B R
CAN		.80	.70	.90	.85	.76	.79	.85
FIN			.75	.72	.76	.71	.75	.70
NOR		.78		.66	.67	.75	.75	.65
SWE		.77	.73		.73	.78	.75	.79
USA	.86					.74	.74	.76
NZL	.76				.76		.90	.73
AUS	.82				.78	.88		.85
GBR	.86				.78	.72	.85	

As can be seen in Table 4, and indicated by the use of bold-faced numbers, most of the estimates of genetic correlations either were improved or remained at same level when the countries were divided into two groups.

There has been no claim that genetic similarity is a measure of connectedness. However, the proportion of cows with common sire between two countries, not only is an indication of the degree of connectedness between two countries, but it can also be considered as an indication of how these two countries are genetically related to each other. And by judging from the results in Table 4 one can intuitively say that these results are promising. As was mentioned in the previous section, the search for a measure of connectedness with the desired properties should continue. However, it seems that the concept of genetic similarity is *a* likely candidate to be subjected to further modifications in order to make it *the* desired measure of connectedness for international bull evaluations.

Although finding a good measure of connectedness helps us in obtaining better estimates of genetic correlations between well connected sub-sets of data, i.e. different country combinations, however, it is of little use for those country combinations that have little or no common bulls between them. As can be seen from Table 1, there are only 305 bulls with multiple proofs in this particular

data set from Ayrshire breed. (The choice of Ayrshire bull populations as example in this paper is because of its average size, both with respect to the number of bulls and the number of countries and does not restrict generality of the conclusions in any way). The proportionate number of such bulls is not markedly different in other breeds, as can be seen from Table 2. These numbers are indeed very low and do not provide an optimum condition for estimation of genetic correlations. The ultimate solution to the problem of estimating genetic correlations is more simultaneous progeny testing of young bulls in several countries. In the meanwhile, one must think of a temporary solution to the problem.

Genetic Covariance Structure

Rekaya et al. (1999) proposed a linear model to describe genetic correlations between five traits. In order to see if their model can be applied to correlation of a bull's breeding values in different countries a number of different models containing one or several of the following factors were used to analyze the genetic correlations between different countries for Holstein breed.

- 1) Genetic similarity between countries, GS, as was described before;
- 2) Regional location of the countries, RL, dividing them into six groups: North American countries, Nordic countries, North European countries, East European countries, South European countries and Southern Hemisphere countries.
- 3) Genetic evaluation model practiced in the various countries, EM, dividing them into four categories: single-trait, BLUP, sire model; single-trait, BLUP, repeatability, animal model; multiple-trait, BLUP, animal model; and multiple-trait, BLUP, test-day, animal model.
- 4) Climatic conditions in different countries, CC, dividing them into two groups, Northern Hemisphere and Southern Hemisphere.

Obviously, the division of countries in different groups was highly subjective and can be criticized for a number of reasons. Furthermore, sometimes these factors were statistically treated in a way that was not

consistent with their nature (e.g. fixed effects treated as random and visa versa). However, at this exploratory stage, I decided to toy with the idea to see if, given objectively chosen factors and quantitatively defined parameters, the results can make any biological sense. The results indicated that with the exception of climatic conditions, that was defined in an extremely inaccurate manner, the rest of factors, that is genetic similarity, regional location and genetic evaluation model were potentially very informative. For example, RL and EM, when treated as fixed effects, were highly significant ($p < 0.0001$) and the model was able to describe more than 91% of the variation in the observations. Adding GS to the model as fixed effect, a dubious thing to do, increased R^2 values to 98%, and even higher values were obtained when variances were excluded from the covariance matrix.

Acknowledgements

I would like to thank Freddy Fikse for many fruitful discussions. Financial support from USDA/NAAB is gratefully acknowledged.

References

- Fernando, R.L., Gianola, D. & Grossman, M. 1983. Identifying all connected subsets in a two-way classification without interaction. *J. Dairy Sci.* 66, 1399-1402.
- Hanocq, E. & Boichard, D. 1999. Connectedness in the French Holstein cattle populations. *Genet. Sel. Evol.* 31, 163-176.
- Rekaya, R., Weigel, K.A. & Gianola, D. 1999. Bayesian estimation of a structural model for genetic covariances for milk yield in five regions of the USA. 50th Meeting of the EAAP, Zurich, Switzerland, 22-26 August 1999.
- Searle, S.R. 1971. *Linear Models*. John Wiley & Sons, New York.
- Searle, S.R. 1987. *Linear Models for Unbalanced Data*. John Wiley & Sons, New York.
- Weeks, D.L. & Williams, D.R. 1964. A note on the determination of connectedness in an N-way cross classification. *Technometrics* 6, 319-324.