Developing a Data Validation Tool Based on Mendelian Sampling Deviations

Filippo Biscarini, Stefano Biffani and Fabiola Canavesi
A.N.A.F.I. Italian Holstein Breeders Association
Via Bergamo, 292 – Cremona, ITALY

Abstract

A more comprehensive validation procedure for international genetic evaluation data is likely to be welcome in a scenario of growing exchange of livestock and semen. Mendelian sampling (MS) deviations are used in this paper to build a tool for validation of national genetic proofs before submission to the Interbull system. The regression analysis of the mean of MS terms proved to be more sensitive to small biases than the current validation methodology: it detected a bias that in the current situation would have a major influence on the international rankings (295% more Italian bulls for protein yield).

Introduction

The issue of data quality for the genetic evaluation of dairy cattle has become a great concern over time, for a number of reasons such as:

1. the increased international trade of livestock and cattle semen;
2. the related increasing importance of the international genetic evaluation of cattle carried out by the Interbull Centre;
3. the adoption, in many countries, of Test-day models for production traits, which restricted the possibility of use of current validation methods;
4. the evaluation of new traits, for which Weibull and threshold models are used, that make it more difficult to apply the current validation methods;
5. the increasing demand for highly accurate genetic information from the industry, breeders and farmers.

At present, Interbull checks for validation of genetic trend and sire standard deviation. The genetic trend validation test comprises three alternative methods: the first one compares first-lactation to multiple-lactation proofs; the second one analyses daughter yield deviations (DYD) over time; the last one is based on a regression analysis of official proofs through years.

These three methods all have some drawbacks: the first one is not suitable for genetic evaluation procedures that do not use a lactation repeatability model. DYDs are not easily estimated for Test-day models. The third method does not take into account the continuous updating of genetic evaluation techniques. The sire standard deviation test is not entirely appropriate for data validation purposes, since it focuses only on the degree of change in sire standard deviation across consecutive MACE runs (Miglior *et al.,* 2002).

Furthermore, as shown in figure 1, a small bias that could not be detected by the current validation procedure (purposely just under the threshold of detection) can lead to huge differences in the results of the international genetic evaluation; the proportion of Italian bulls in the top100 ranking for protein yield can be up to 19% before the bias is detected, a 295% increase compared to the official results (Biscarini *et al.,* 2006).

Thus, there are both need and scope for the development of additional tools to validate genetic evaluation data: as suggested by previous works (Van Doormal *et al.,* 1999; Van Doormal & Miglior, 2000; Miglior *et al.,* 2002) Mendelian sampling (MS) deviations can form the basis of a new data quality assessment tool.

MS terms can be used to validate both national data before submission to the MACE model and international proofs after Interbull's calculations.

The theoretical expectation is that trends in mean and variance of MS should remain constant over time.

Objective of this paper is to present results of the use of MS terms to validate national proofs before submission for the MACE model under official and biased circumstances.

Material and Methods

This study is based on the February 2005 Holstein bulls genetic proofs for milk, protein and fat yield of 8 countries (Canada, Germany, Denmark, France, Italy, The Netherlands, USA and UK), which constitute the input data of the MACE model.

Three sets of Italian input EBVs were used: the official February 2005 EBVs and two artificially biased sets:

- one with a large bias;
- one with a small bias.

The Italian biased data were compared to the official data from the other countries in reduced MACE runs.

Both biases were constructed so to accumulate over time, using the following quadratic function:

$$y_i = a_i^{n},$$

where:

y_i is the bias to be added to the i_{th} trait (milk, protein, or fat yield);
a_i is an "*ad hoc*" coefficient for each trait and each biased dataset (large or small bias);
n is the renumbered birth-year of bulls used as exponent of the function.

The coefficients for milk, fat and protein yield were empirically derived and were 1.31, 1.28, 1.26 and 1.40, 1.33, 1.33 for the small and large bias respectively. For fat and protein yields these coefficients were divided by 10, to account for the differences in magnitude of their standard deviations compared to that of milk yield.

The same biases were added to the hol040 files, with information on past genetic evaluations, needed for the official trend validation method 3, which analyses the variation of national proofs across evaluation runs (Boichard *et al*., 1994).

For trend validation, Interbull method 3 and the regression analysis of MS terms have been compared for efficacy. According to the Interbull Code of Practice, the "t" parameter of method 3 must be lower than 2% of the standard deviation of the trait considered for the data to be accepted (Interbull, 2006); the same criterion has been used for the regression coefficients of MS trends.

Within-country sire variances for each one of the 3 above mentioned Italian datasets, were calculated using a copy of the official MACE programs provided by the Interbull Centre.

SAS® statistical procedures have been used to generate the biased datasets, and to validate the genetic trend using either official method 3 or MS deviations trends.

Results

First of all, the impact of biases on sire variance estimation has been considered: results are shown in table 1. While in the case of the large bias the variation far exceeds the 5% limit set by the Interbull Centre and data are therefore easily discarded, when the bias is small the sire standard deviations are well within the boundaries (0,85, 1,30 and 0,81% for milk, fat and protein respectively), data are accepted and have quite an effect on final MACE results.

Interbull method 3 for trend validation (table 2) can't detect a small but not negligible bias either, both in the scenario of a recently introduced bias (official hol040 file) and of a long established one (biased hol040 file).

When looking at the regression analysis of MS terms in table 3, it can be noticed that their average trend through years is more sensitive to biases than method 3, being able to detect also the small bias that has been introduced in the Italian data: regression coefficients are well beyond the limit of the 2% of the standard deviation of the trait even in this case. This can be deduced also visually, looking at the graph in figure 2.

Contrariwise, the regression analysis of the variance of MS deviations does not seem very useful in assessing the validity of national genetic evaluations: trends look flat and regression coefficients remain always within the 2% threshold of acceptance. However, this is likely due to the additive nature of the bias.

Conclusions

Current Interbull official validation procedures can hypothetically allow for the introduction of biases with potentially considerable effects on MACE results. The development of a new tool, based on MS terms, can provide a more sensitive method to ensure a better data quality. The regression analysis of the average MS deviations, combined with the other currently available methods, can help identify also small, but not negligible, biases in national genetic evaluations. MS variance does not seem as effective.

The analysis of MS of also international proofs, would properly supplement this validation tool and can be the objective of a further study.

References

Biscarini, F., Biffani, S. & Canavesi, F. 2006. The consequences of biases in the international genetic evalaution. *Proceedings of WCGALP 8* (submitted).

Miglior, F., Sullivan, P. & Van Doormaal, B. 2002. Preliminary analysis of Mendelian sampling terms for genetic evaluation validation. *Interbull Bulletin 29,* 183-187.

Van Doormaal, B. & Miglior, F. 2000. Trends in sire variance estimates by birth year. *Interbull Bulletin 25,* 70-73.

Van Doormaal, B., Kistemaker, G. & Sullivan, P. 1999. Heterogeneous variances of Canadian Bull EBVs over time. *Interbull Bulletin 22,* 141-148.

Bonaiti, B., Boichard, D., Barbat, A. & Mattalia, S. 1994. Three methods to validate the estimation of genetic trend in dairy cattle. *Interbull Bulletin 10,* 9 pp.

Interbull code of practice, 2006. - http://www-interbull.slu.se/service_documentation/General/Code_of_practice/framesida-code.htm. Acc.29/05/06.

SAS® 1982. *User's Guide: Statistics.* Version 5.18. SAS Inst., Inc., Cary, NC.

Table 1. Italian sire standard deviation.

	uff	small bias	large bias
milk	355	358	443
fat	13,08	13,25	14,47
protein	11,15	11,24	12,83

Table 2. Results of trend validation by means of Interbull method 3.

		official		large bias		small bias		reference	
	Trait	t	std err	t	std err	t	std err	dev std	2%
milk	hol040_off	1,517	12,93	-68,01	24,31	-14,37	12,7		
	hol040_b4	-	-	-67,66	24,18	-	-	875,396	17,508
	hol040_b5	-	-	-	-	-14,11	12,7		
fat	hol040_off	0,347	0,53	-1,96	0,84	-0,62	0,55		
	hol040_b4	-	-	-1,71	0,76	-	-	36,905	0,738
	hol040_b5	-	-	-	-	-0,42	0,53		
protein	hol040_off	0,334	0,47	-2,35	0,79	-0,45	0,46		
	hol040_b4	-	-	-2,07	0,7	-	-	27,139	0,543
	hol040_b5	-	-	-	-	-0,21	0,45		

Table 3. Regression analysis of average and standard deviation of MS terms.

		Official feb 05			small bias			large bias		
		Ms_milk	ms_fat	ms_prot	ms_milk	ms_fat	ms_prot	ms_milk	ms_fat	ms_prot
mean	B	-14,24	-0,58	-0,59	28,55	1,89	0,99	233,94	6,12	6,06
	std err	1,55	0,07	0,04	6,37	0,31	0,2	35,19	0,85	0,83
ref	std dev	milk	875,396		fat	36,905		protein	27,139	
	2%		17,508			0,738			0,543	
std dev	std err	1	0,055	0,032	1,11	0,05	0,03	7,5	0,14	0,14
	B	-1,78	0,023	-0,021	-0,55	0,006	0,005	23,31	0,38	0,42
		Ms_stdm	ms_stdf	ms_stdp	ms_stdm	ms_stdf	ms_stdp	ms_stdm	ms_stdf	ms_stdp

Figure 1. Effects of the size of an additive bias on international protein yield rankings.

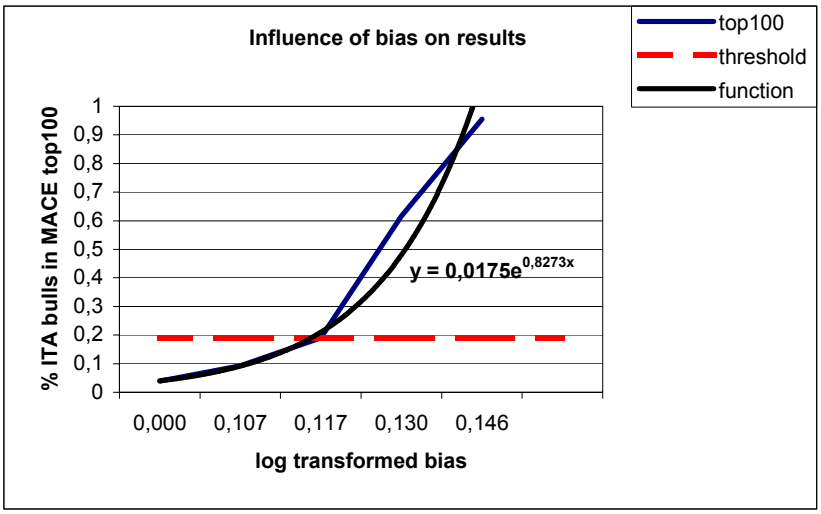

Figure 2. Trend of average MS for protein yield.

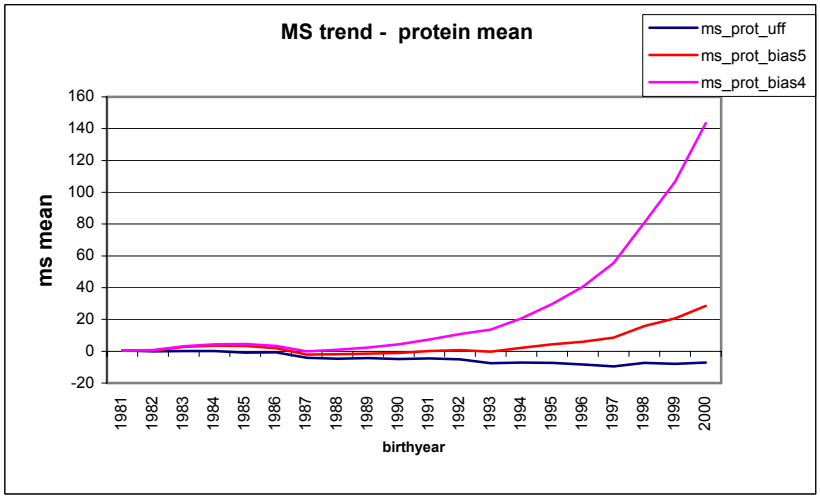

Figure 3. Trend of the standard deviation of MS for protein yield.