The effect of phantom parent groups on genetic trend estimation

H.E. Theron[1#], F.H.J. Kanfer[2] and L. Rautenbach[1]
[1]ARC-Animal Improvement Institute, Private Bag X2, Irene 0062, South Africa
[2]Department of Statistics, University of Pretoria, South Africa

Abstract

In an animal model evaluation of breeding values it is assumed that the base animals are all at the same genetic level. However, in the South African Holstein population, animals of different genetic levels were imported from foreign countries, thus causing a deviation from this assumption. The effect of this deviation is considered using first lactation records from 393 458 Holstein cows. Genetic trend estimation is studied through a time trend analysis of within-Bull-yearly-Daughter Yield Deviations, or DYD's. Bias in the estimation of trend was reduced when phantom parent groups were taken into account. The 109 385 base animals were replaced by 64 phantom parent groups. Phantom parent groups were constructed by combining year of birth, country of birth and selection intensity of the phantom parents. In recent years, foreign sires have been more affected by the exclusion of phantom parents groups in the model, than local sires, although ranking coefficients for the 15 771 sires in the analysis were in excess of 90%. Ranking coefficients for cows were also high.

Keywords: Phantom parent groups, genetic trend, daughter-yield-deviations, Holstein
[#] Corresponding author. E-mail: helena@irene.agric.za

Introduction

In an animal model evaluation it is assumed that base animals (animals with unknown parents), are sampled from the same population with an average breeding value of zero and a common variance $\sigma^2_{a.}$ However, in certain circumstances, base animals are from different genetic merit. Such circumstances may exist when animals migrate into the population over time. Relating base animals to phantom parents makes it possible to correct for the different genetic levels (Westell *et al.,* 1988). The use of the term phantom is to emphasize that those animals are not themselves of interest (Westell *et al.,* 1988).

Phantom parent groups are constructed by combining year of birth, country (or region) of birth and selection intensity categories. The different genetic levels between countries motivate the country category. The year category allows for genetic improvement over time. A lack of pedigree information makes it difficult to link imported animals from a country to their common ancestors. Because of this limitation animals are linked to different base animals, which are at different genetic levels due to the improvement. It is, therefore, important to construct phantom parents to represent countries and different time intervals within the countries. Another contributing factor to the unequal base levels is selection intensity differences between breeding lines. Sires have a much higher selection intensity than dams, therefore necessitating the selection intensity category (Westell *et al.,* 1988).

As a lack of compliance to the above assumption in the animal model may cause bias in the estimation of the genetic trend, these aspects will be investigated for South African Holstein cattle. Large numbers of animals were imported into the South African Holstein population over time from different countries. It will be shown that linking of these animals to phantom parent groups will reduce bias in the estimation of genetic trend. This paper will study the possible bias in genetic trend estimation through a specific residual analysis according to the method described in Biochard *et al.* (1995). The effect on the prediction of breeding values will also be considered.

Materials and methods

The data set comprised of South African Holstein cows participating in the South African Dairy Cattle Performance Testing Scheme, and that had first lactation production records between January 1978 and July 2001. Pedigree information of all cows was included. The data set included the first lactation records of 393 458 cows and 834 972 pedigree records. Pedigrees were traced back as far as possible to identify base animals. When a parent was encountered of which no information was available, or provided no other pedigree links, it was considered a phantom parent and assigned to a phantom parent group.

Country of origin of the phantom parents was grouped into three levels, *viz*. South Africa, Europe and North America. A combination of the Netherlands, United Kingdom, Germany, France, Denmark and Israel constituted the category Europe. According to INTERBULL (2001), phantom parent groups should have a minimum size of 10-20 animals and different groups should be merged to attain reasonable size. As very few animals were imported from some European countries, these countries were lumped together. Too few animals were also imported from New Zealand / Australia to form a phantom parent group on their own. These animals were therefore included in the category Europe. The INTERBULL (2001) criterion could, however, still not be met in all instances. North America was a combination between the United States of America (USA) and Canada. Birth years of the phantom parents were categorized into 5-year groups, *viz.* pre-1975, 1976-1980, 1981-1985, 1986-1990 and 1991-1999. When year of birth of a phantom parent was unknown, it was assigned as five years previous to the birth of the oldest daughter. Selection intensities were categorized into four levels depending on the breeding line of the imported animal, *viz.* sire of a sire (most intense selection), sire of a dam, dam of a sire or dam of a dam (weakest selection).

Take for example, semen of a bull with unknown parents that was imported into South Africa from the USA. Although the sire was unknown (or unlinked), it was born (or estimated to be born) in 1982. It will then be assigned to the following phantom parent group: North America x (1981-1985) x Sire-of-sire. Note that all pedigrees can be traced back to phantom parent groups and that every base animal was linked to a phantom parent group. Phantom parent groups with less than seven animals were merged with adjacent year groups.

In theory, mixed model methodology provides the best estimate of genetic trend, provided that all underlying assumptions are fulfilled. Boichard *et al.* (1995) has proposed three methods to validate the models used for genetic evaluation of dairy cattle. Genetic trend estimation in the current study was investigated through a residual analysis (Boichard *et al.*, 1995). The method uses 'within bull variation of Daughter Yield Deviations' or DYD values (VanRaden & Wiggans, 1991; Boichard *et al.,* 1995). The DYD analysis is quite general and can be used to detect problems in the model of analysis. This method investigates the non-genetic time trend over the considered period. The DYD values are obtained by removing all the estimated terms and non-contributing predicted terms from the phenotypic value. Averaging these values over the daughters of a sire reduces the stochastic component and represents a genetic component for the sire. Grouping daughters per year generates a trend for the sire's genetic component, which should be constant with only random variation about the true genetic value. A systematic deviation is indicative of bias in the estimated trend. In more detail, Daughter Yield Deviations were calculated as follows:

For each daughter k of sire j within year i:

$$YD_{kij} = y_{kij} - f - r - 0.5*a(dam).$$

where $\mathbf{y_{kij}}$ represents the phenotypic observations,
f the environmental fixed effects,
r the random genetic effects and
a(dam) the additive genetic component of the daughter's dam.

The value DYD_{ij} is then obtained by averaging over YD, for each daughter k of sire j, within year i. Usually the notation DYD is used to indicate DYD $_{ij}$ with no reference to the latent values j and i.

It is clear that YD_{kij} should be randomly distributed about the true genetic value of sire j, independent of the specific year. Thus, Daughter Yield Deviations of bulls are average performances of the bulls' daughters adjusted for the dam breeding value and for all the effects included in the model, except daughter breeding values (Boichard *et al.,* 1995). Expected DYD values should only depend on the bull and are theoretically independent of any environmental effect, particularly the birth year of the daughters. This property of the residuals may be used to validate the estimation of genetic trend, which is in this case the combined sire trends.

To test whether the genetic trend is over- or underestimated, a regression model is fitted to the DYD values of all bulls with daughters in at least 10 herds. The regression model is:

$$DYD_{ij} = sire_i + b * j + e_{ij}.$$

where DYD_{ij} is the DYD considering daughters of the i^th^ bull that first calved in the j^th^ year; by definition j=0 for the first year when at least 10 daughters of a bull calved for the first time; sire is the effect of the i^th^ bull.

When the estimate of genetic trend is unbiased, the year effect has a zero expectation and should not significantly differ from zero. If the term b*j is significant in the model it implies that the genetic trend estimation is biased. A positive value indicates an overestimation and a negative value an underestimation (Boichard *et al.,* 1995). The P values from standard regression analysis can be used to identify a significant b parameter. Other requirements can also be used. INTERBULL requires a b-value of less than $.01*\sigma_g$ with σ_g the genetic standard deviation of the trait under consideration (INTERBULL documentation).

Estimates of the additive genetic values, fixed effects and other random components were calculated using a standard multivariate animal mixed model written in generic notation as:

$$\mathbf{y} = \mathbf{f} + \mathbf{r} + \mathbf{a} + \mathbf{e}$$

where : **y** represents the phenotypic observed first lactation: kg milk, kg butterfat and kg protein,

f the environmental fixed effects: herd-year-status (21 254 levels), age-class (6 levels),

r the random non-additive genetic effect: sire by herd (101 410 levels),

a the additive genetic effect and

e the measurement error.

The VCE4 package (Groeneveld & García-Cortés, 1998) was used to estimate (co)variance components on a selection of the data. The estimates were calculated using PEST (Groeneveld & Kovac, 1990). These programmes accommodate phantom parent groups in the model through a code in the pedigree file, indicating whether the sire, dam or both/neither are assigned to phantom parent groups. The same (co)variance components were, however, used for the models when phantom parent groups were included or excluded, as the inclusion of phantom parent groups had a minor influence on the estimation of (co)variance components. Regression models were fitted with SAS (1996).

The effect of the inclusion and exclusion of phantom parent groups in the model on genetic trend for the population was examined, as well as the bias on the breeding values of local and foreign sires. The effect on the ranking of breeding values of cows and sires was also investigated.

Results

The data set included the first lactation records of 393 458 cows and 834 972 pedigree records. The first analysis (not taking phantom parent groups into consideration), had 109 385 base animals, which was reduced to 64 phantom parent groups in the second analysis. Sires numbered 8 324 and dams 426 666.

The total possible number of phantom parent groups equaled 72, three country groups, six year-groups and four selection lines. Only 64 groups contained animals (Table 1). The phantom parents encompassing sire of dam and dam of dam and which originated in South Africa, were linked to most animals (Table 1). This is probably due to cows with unrecorded pedigrees in herds that entered the Dairy Cattle Performance Testing Scheme over time. Foreign genetic material mostly originated in the USA.

The effect that including and excluding phantom parent groups has on genetic trend, is illustrated graphically in Figure 1 for milk production. As can be seen in Figure 1, taking phantom parent groups into account has had a drastic effect on the estimated genetic trend for South African Holstein cattle. A much steeper trend was estimated when phantom parents were included *vs.* excluded in the analysis: 64.0 kg *vs.* 35.6 kg per year for milk yield; 1.78 kg *vs.* 0.98 kg per year for protein yield and 1.79 kg *vs.* 0.96 kg per year for butterfat yield, respectively. The estimated breeding values of older animals were generally lowered, while the breeding values of younger animals tended to be higher when phantom parent groups were included in the model.

Table 1 Number of base animals per phantom group

		Country		
Selection line	**Year of birth**	**RSA**	**Europe***	**North America****
Sire of sire	Pre 1970	513	112	56
	1970-1974	365	18	18
	1975-1979	185	7	9
	1980-1984	135	24	45
	1985-1989	100	101	203
	1990-	0	0	22
Dam of sire	Pre 1970	539	115	57
	1970-1974	245	17	18
	1975-1979	178	7	12
	1980-1984	135	24	47
	1985-1989	55	104	221
	1990-	0	0	22
Sire of dam	Pre 1970	32168	156	212
	1970-1974	25370	40	128
	1975-1979	25052	39	173
	1980-1984	19157	11	138
	1985-1989	33742	23	60
	1990-	48184	0	0
Dam of dam	Pre 1970	26109	156	207
	1970-1974	31482	41	127
	1975-1979	31788	40	194
	1980-1984	29051	17	193
	1985-1989	47842	24	69
	1990-	24329	0	0

* Europe: Netherlands, United Kingdom, Germany, France, Denmark, Israel and New Zealand
** North America: United States of America, Canada

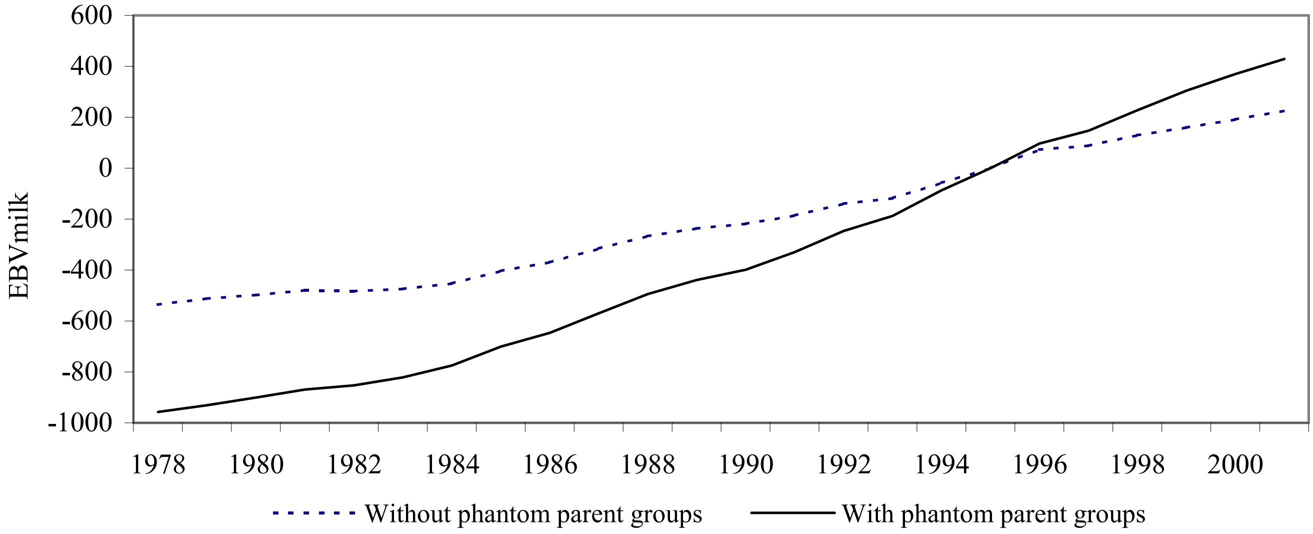

Figure 1 The effect of including or excluding phantom parent groups on the genetic trend for milk yield of the South African Holstein population. EBV – estimated breeding value

The DYD test of Boichard *et al.* (1995) indicated a significant underestimation in the genetic trend when phantom parent groups were not included in the model. The values of 01*σ_g for milk, butterfat and protein were 5.97, 0.20 and 0.17 respectively. The values of |b| in the DYD-test, when phantom parent groups were excluded, were 10.82, 0.37 and 0.33 for milk, butterfat and protein, respectively. These values were negative, indicating an underestimation, and are outside the range of 01*σ_g values, indicating significant underestimation. The values of b were -2.26, -0.07 and -0.07 for the traits respectively, when phantom parent groups were included in the model. As these values fall well into the range of 01*σ_g values, it indicates that there is no significant bias in the estimation of the genetic trend. It is, therefore, clear that phantom parent groups should be included in the model.

The genetic trends for daughters from foreign sires and daughters from local sires are shown for milk yield in Figure 2. It can be seen from the figure that, in recent years, foreign sires have tended to be more severely affected by the exclusion of phantom parent groups in the genetic model than local sires.

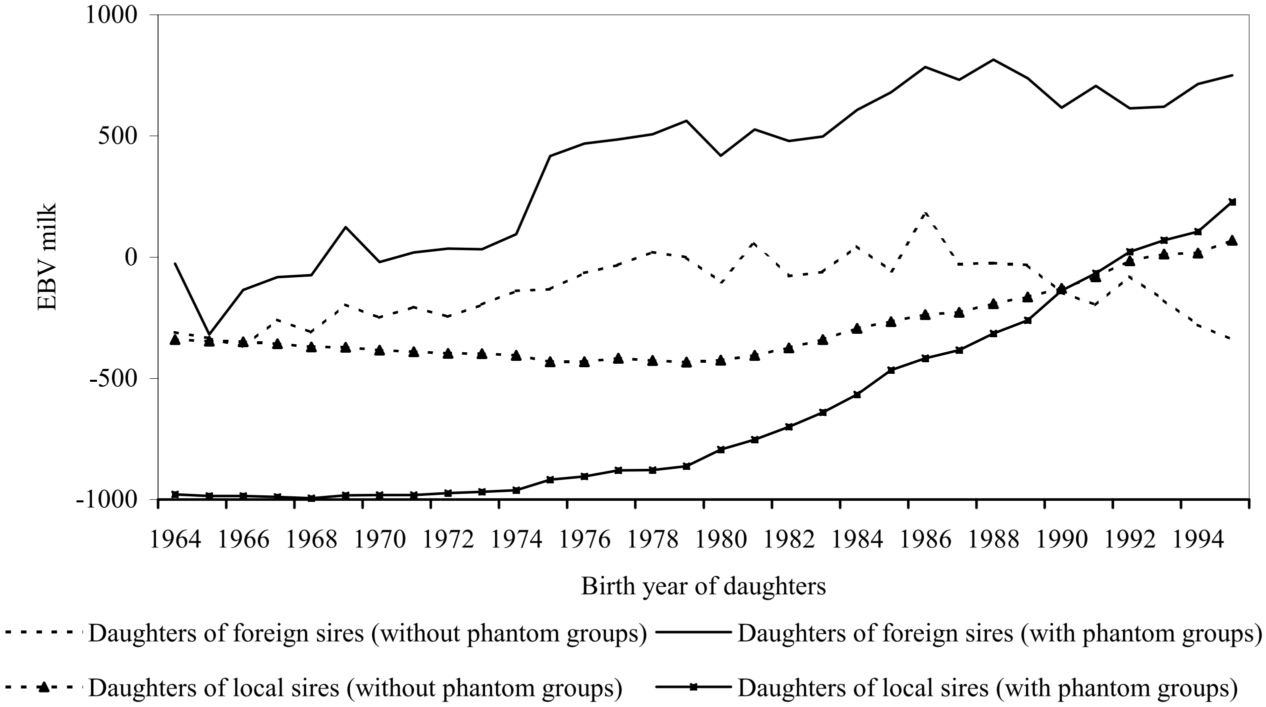

Figure 2 The effect of the inclusion of phantom parent groups in the genetic model on the genetic trends of daughters of local and foreign sires. EBV – estimated breeding value

The effect of including phantom parent groups on the ranking of breeding values of animals was compared using Spearman correlation coefficients (Table 2). The coefficients were generally high, ranging between 0.882 and 0.957. Despite the fact that the inclusion of phantom groups causes a significant change in the genetic trend estimation, it does not seem to have a major effect on the ranking of animals.

Table 2 Spearman correlation coefficients between breeding values estimated from an analysis where phantom parent groups were included *vs*. excluded

Trait	Cows with records (n = 393 457)	All cows (n = 819 125)	Bulls (n = 15 771)
Milk	0.949	0.885	0.923
Butterfat	0.948	0.883	0.904
Protein	0.947	0.882	0.910

Discussion

The usefulness of a statistical procedure is based on the reasonability of the assumptions of the procedure. The techniques used for predicting breeding values assume that all the animals of which the ancestry is unknown come from populations with the same average genetic level. This is only true if no animals enter the population over time. However, in the South African Holstein population, animals of superior genetic value were imported from other countries and entire herds entered performance testing. Some of these animals could have extensive pedigrees, but do not have recognized parents that contribute ties and records to the data. This caused deviation from one of the underlying assumptions of the animal model, which led to biased estimation of genetic trend. Taking the various subpopulations' genetic levels into account within the statistical model can reduce bias in the genetic trend. In the case of an unknown sire, the available information, such as country of birth, year of birth, etc. can be included on the basis of grouping the unknown parents (phantom parents) into phantom parent groups. Group effects can be thought of as accounting for selection not accounted for by records of relatives. Under this concept, groups would be assigned only if animals were missing genetic relationships. The genetic merit of all descendants of any animal that has a missing parent would then include a function of the phantom parent group of the missing ancestor (Westell *et al*., 1988). The phantom parent group is then used as a substitute for the parent in the genetic analysis. By doing this the genetic merit of each group can be predicted and used to correct the breeding value of each animal on the basis of its relatedness to the phantom group.

Phantom parent groups are assigned to unknown parents so that a phantom parent group replaces each unknown parent. The total genetic value for every animal includes a function of genetic groups. The function of genetic groups is specific for each individual animal and depends on the number of generations to the base phantom ancestors and on the genetic groups to which those phantom ancestors are assigned (Westell *et al*., 1988). The genetic merit of each group was predicted and used to correct the breeding value of each animal on the basis of its relatedness to the phantom group.

Conclusion

It has been shown that the assumption of the animal model that all base animals are sampled from the same population, has not been met in South African Holstein cattle. This is due to the extensive use of imported semen and herds continually entering performance testing. The violation of this assumption has led to significant bias in the genetic trend for milk, butterfat and protein yield. By replacing base and migrant animals with phantom parent groups in the analysis, the genetic level of migrant animals is separated from that of base animals. Inclusion of phantom parent groups in the model has reduced the bias in the genetic trend to insignificance. In recent years, foreign sires have been more affected by the exclusion of phantom parents groups in the model, than local sires, although ranking coefficients for the 15 771 sires in the analysis were in excess of 90%. Ranking coefficients for cows were also high.

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S., 1995. Three methods to validate the estimation of genetic trend for dairy cattle. J. Dairy Sci. 78, 431-437.

Groeneveld, E. & García-Cortés, A., 1998. VCE 4.0, a (co)variance component package for frequentists and Bayesians. In: 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia. 27, 455-458.

Groeneveld, E. & Kovac, M., 1990. A generalised computing procedure for setting up and solving mixed linear models. J. Dairy Sci. 73, 513-531.

INTERBULL, 2001. Interbull guidelines for national and international genetic evaluation systems in dairy cattle with focus on production traits. Bulletin No 28, Uppsala, Sweden.

SAS, 1996. SAS user's guide: Statistics, Release 6.12. SAS Institute Inc., Cary, North Carolina, USA.

VanRaden, P.M. & Wiggans, G.R., 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74, 2737-2746.

Westell, R.A., Quaas, R.L. & Van Vleck, L.D., 1988. Genetic groups in an animal model. J. Dairy Sci. 71, 1310-1318.