New Genetic Parameters for National Evaluations of Production Traits in Spanish Holsteins Excluding Selected base Animals from the Estimation of Genetic Variance

J.Pena[1], M.A. Ibañez[2], M.J. Carabaño[3], L. L.G. Janss[4]
[1]*CONAFE, Madrid, Spain.*
[2]*ETSIA, Universidad Politécnica, Madrid, Spain.*
[3]*Instituto Nacional de Investigación y technologia agraria y alimentaria, Madrid, Spain*
[4]*ID-Lelystad. Institute for Animal Science and Health, The Netherlands.*

1. Introduction

Accurate estimates of genetic parameters at a national level are needed for national genetic evaluations to obtain more reliable breeding values, but those parameters are also a key point in international evaluations, as best bulls from countries with higher heritabilities have more chance of better rankings, because information of daughters in their country are weighed more heavily.

In Spain, national evaluations for production traits are calculated with a repeatibility animal model assuming heritability 0.25 and repeatability 0.4, while some other countries assume higher values for these parameters (France and Italy use heritability 0.30 and repeatability 0.50, USA 0.30 (average) and 0.55, and The Netherlands 0.35 and 0.55). Various countries had raised heritabilities used in their national evaluations based on recent estimates, as reported by Bagnato et al. (1996) and Van Tassell et al. (1997) for first lactations data and by Janss and de Jong (1999) for first three lactations.

Several estimates of heritability have been obtained recently based on Spanish data. Rekaya (1997), applying a multiple trait animal model to a subset of North of Spain reported decreasing heritabilities for first three lactations, 0.28, 0.24 and 0.23 for kg milk and 0.25 0.22 and 0.22 for kg protein. Charfeddine (1998) analized first lactation data with a multiple trait animal model between production, type and longevity, reporting heritabilities of 0.31 for kg milk and 0.28 for kg protein. Ibañez et al. (1999), analysing first lactation data from all Spain with a sire model, found a much lower value for heritability of kg milk, 0.26, and, also, a great heterogeneity of heritabilities between regions, periods of time and herd production level.

Hernandez et al. (1998) applied a Gibbs sampling squeme for estimating variance components (VC) with the same data and model used in Spanish national evaluation of January 1998, considering up to 10 lactations per cow. They obtained heritability of 0.23 and repeatability of 0.46 for kg milk and values of 0.22 and 0.46 for kg protein, much lower heritabilities than the ones found in previous studies. Based on the same data set, Hernandez et al. (1999) estimated genetic parameters by region trying to verify if regions with a big increase in milk recorded animals were contributing to lower national estimate of heritability and if there was heterogeneity of heritabilities between regions, as found by Ibañez et al. (1999), but they couldn't confirm this hypothesis. They suggested, following Pieramati and Van Vleck (1993), that lower estimates of heritabilities compared to previous results with reduced data sets could be caused by genetic groups reducing estimates of additive genetic variance, as earlier analysis did not include phantom groups in the model.

Selected base animals has been reported to be a source of bias in the estimation of genetic variance (Kennedy et al., 1988; Van der Werf and De Boer, 1990). As reported by Jeyaruban and Gibson (1996), animal model estimates are subject to large and essentially unpredictable bias (-45% to –10%) when there has been selection and pedigrees do not trace back to an unselected base generation, what could be of special importance in the Spanish Holstein population with a big increase in recorded animals in the last 15 years. As reported by

Koots and Gibson (1996), data structure and selection history among populations explain much of the residual variation in estimates of heritabilities among populations that is unrelated to sample size.

Graser et al. (1987) suggested that, in presence of selected base animals, genetic variance prior to selection could be estimated considering base animals as fixed. The basis for it is that estimation of genetic variance would be based on variance due to mendelian sampling variance and, under the infinitesimal model, this variance is assumed not to be affected by selection (Bulmer, 1971). But Van der Werf (1992) found that, when treating base animals as fixed, bias still exist if progeny of these base animals was also selected, although this bias was reduced as more generations of data were include in the estimation. Van der Werf (1992) also concluded that considering base animals as fixed is equivalent to consider a different phantom group for each base animal. Inclusion of phantom groups in the model takes into account the effect of selection on the mean of base animals and it can reduce some bias in the estimation of genetic variance due to selected base animals (Pieramati and Van Vleck, 1993), but not completely. Van der Werf and Thompson (1992) separated estimation of genetic variance between selected base animals and non base animals for simple designs and described biases that may arise if assuming wrong reductions in the genetic variance of selected base animals.

The objective of this study was to estimate **genetic parameters** for kg of protein with an animal model applying Gibbs sampling to a new model and with renewed data requirements, analysing the effect of including phantom groups in the model and the effect of excluding selected base animals when estimating the additive genetic variance.

2. Materials and Methods

2.1. Data

Data used were 904851 lactations from calvings between year 1984 to 1998, with only 10% of the data from calvings before 1991. Edits were the same as in national evaluations, but all records were projected to 305 days with the method used in official evaluations. A minimum of 215 days in milk was asked for data to be included in the analysis and up to five lactations were used by cow. All animals with data were required to have at least an informative parent. No preadjustment for heterogeneity of variance was done. Pedigree file included 588203 animals.

2.2. Model

Factors considered in the model were as in the new model developed for national genetic evaluations:

- Herd-Year-Imported-Season-Parity (86317 levels), built with a flexible strategy in order to achieve a minimum of five observations per management group when splitting herd-year into imported cow or not, season and parity.
- Age within lactation (1,2,3,4,5) nested to production system (444 levels)
- Month of the year within lactation (1 or later) nested to production system (288 levels)
- Permanent environmental effect to model repeated lactation records (420851 levels)
- Additive genetic effect
- Phantom groups

Production system was defined as level of production on first lactation, region and period of time. Two regions (Cantabric Cornice and Rest of Spain) and two periods of time (before 1995 and from 1995) were considered. Three levels of production were defined within region and period of time.

Phantom groups were defined by sex of animal with unknown parent(s), sex of missing parent, periods of three years, country of origin (USA, CAN, NLD,FRA,ITA, DNK, AUT and rest of the countries) and region of origin in Spanish animals. Animals with both parents unknown were put into the same group. Because of problems in convergence of the Markov chain, two alternative strategies were used. In strategy A, 200 groups were considered, with a minimum of 100 observations per group. With strategy B, groups were merged so that a minimum of 200 animals was achieved and male and female parents were merged in the same group, resulting in 129 groups. Distribution of base animals through phantom group sizes is shown in Table 1.

Table 1. Distribution of base animals through phantom group sizes, considering two different strategies.

Number of observations Per group	Percentage of animals assigned to each group	
	Strategy A (200 groups)	Strategy B (129 groups)
< 200	20%	0
200-500	35%	25%
500-1000	20%	50%
> 1000	25%	25%

2.3. VC Estimation

With a so big data set, estimation of genetic parameter was realised with a Bayesian analysis implemented via Gibbs sampling with software Maggic (Janss, 1998). A requirement for using it is animals must have both parents known or both unknown and, for achieving it, fictitious parents are added when there is only one parent known (138421 animals). Final number of base animals were 168090, that represents 23% of total animals in the final pedigree used in the calculations

Scaled inverted chi-squared distributions were assumed as prior distributions for variance components. Flat priors were avoided for variances in order to prevent improper posterior distributions of these parameters. So, vague priors were considered as in Hernandez et al. (1998), with 4 degrees of belief (*ν*) and scaling factors (***H***) 237 for genetic variance, 207 for permanent environment variance and 501 for error variance.

When sampling a new realization of genetic variance, scaling factor of the posterior conditional distribution, that is also an scaled inverted chi-squared distribution, is calculated as:

$$s_a^2 = \frac{u'A^{-1}u + H}{n + \nu}$$

being *n* number of animals considered. Quadratic form $\mathbf{u'A^{-1}u}$ is calculated as the sum of squared deviations of breeding values from average phantom parents solutions for all *base animals* plus twice the sum of squared deviations from parental averages for *non-base animals*. If not including phantom groups in the model, breeding values of base animals are deviated from cero.

Four different estimates of VC were run:

- Without phantom groups in the model
- With strategy A of phantom groups
- With strategy B of phantom groups
- With strategy B of phantom groups and excluding selected base animals from the estimation of genetic variance.

For eliminating the effect of selected base animals on the estimation of genetic variance, squared deviations of breeding values from average phantom parents solutions for *base animals* are not included when calculing $\mathbf{u'A^{-1}u}$, which means that genetic variance estimate would be based on Mendelian sampling variance. This is possible because, as described before, all animals have both parents known or both unknown. Only slight modification of the software is needed for implementing this approach.

For each analysis a unique long chain of 100,000 samples was implemented. The first 20,000 samples were discarded as burn-in and the remaining samples were used in the computation of summaries from the posterior distributions of genetic parameters.

3. Results and Discussion

3.1. Convergence of Markov chains

Estimates with strategy A for phantom groups did not converge at all. When 40000 samples were obtained, heritability had increased with each sample, from an initial value of 2 to 80% while permanent environment variance over total phenotipic variance decreased from 46% to 0% and repeatability had gone from 48% to 80%. Residual variance remained fairly constant. Similar results were obtained with different starting values. Generating a lot more random numbers in each sampling of variances and phantom parents (Janss, personal comunication, 1998) didn't have any positive result. Hernández et al. (1998) did not report convergence problems in the analysis, even though they had used phantom group strategy as defined for national evaluation, which is close to strategy A. An explanation might be they did use different software and no fictitious animals needed to be added to the pedigree. Too many small phantom groups combined with big amount of fictitious animals added to the pedigree might be a reason for non convergence.

With strategy B, excluding or including base animals, or when phantoms groups were not considered, convergence was achieved around 2000 cycles, based on visual inspection of a plot.

3.2. Posterior means and standard deviations of the parameters

In Table 2 are shown the posterior means and standard deviations of the parameters of the repeatability model for the three analysis that did converge without problems.

Table 2. Posterior means and standard deviations (into brackets) for genetic parameters

Genetic Parameters	I (No PG^{a})	II (PG^{a}-B)	III (PG^{a}-B + MS^{b})
Error Variance	584,97 (1,27)	584,99 (1,27)	584,87 (1,27)
Permanent Enviromental Variance	287,64 (3,48)	284,82 (3,61)	253,07 (3,59)
Genetic Variance	273,02 (4,75)	278,11 (4,93)	331,46 (5,38)
Phenotypic variance	1145,64 (2,61)	1147,92 (2,70)	1169,40 (2,98)
Heritability	0.2383 (0.0038)	0.2423 (0.0039)	0.2834 (0.0041)
Repeatability	0.4894 (0.0014)	0.4904 (0.0014)	0.4999 (0.0015)

[a]PG: Estimates including Phantom Groups
[b]MS: Estimates based on Mendelian Sampling (excluding selected base animals)

3.2.1. Phantom groups effect

Contrary to Pieremati and Van Vleck (1993) that reported an important reduction on genetic variance estimates when including phantom groups in the analysis, slight increases in genetic variance and heritabilities were found when including or not phantom groups in the model. Li and Kennedy (1994) also reported increase in heritabilties, but of a higher magnitude than the ones found here. So, differences in genetic parameters due to inclusion of phantom groups in the model might depend on structure of the data set used in the analysis.

3.2.2. Effect of selected base populations on estimate of genetic variance

Big changes were found in genetic variance estimates when excluding base animals. Inclusion of base animals in the estimation makes a decrease of 16% in genetic variance estimates and an increase of permanent environment variance. So, as it seems logical, permanent environment variance was overestimated when genetic variance was underestimated. This increase in genetic variance is the reason for increase in heritability, while repeatability increased only +0.01. Heritability equal to 0.28 is more in line with values assumed in national evaluation of main countries.

About correctness of excluding selected base animals for estimating genetic variance, it seems it reduces bias in this estimate, but the problem is not completely solved, as selected base animals would be assigned higher genetic variance than the one that correspond to them, and that holds for the estimation procedure and for its application in national evaluation. About this last argument, it would be the same if genetic parameters had been estimated with a well structured data set in which pedigrees trace back better to the unselected base population. A better approach would be to estimate different genetic variance between base and non base animals, as suggested by Van der Werf and Thompson (1992). As a more general approach, Alfonso and Estany (1999) made proposals for inclusion of different genetic variances for different groups of base animals into genetic evaluations, once these variances are known. An appropiate procedure for estimating those different genetic variances could be developed. Problems could be accuracy of the estimates and at least phantom groups should be grouped into "super-phantom groups" for estimating separate genetic variance for them.

Janss and de Jong (1999) with Dutch data found much higher heritabilties (0.360) when applying a repeatability animal model for kg protein. Some reasons that may explain these differences are structure of lactational data, herd production level and more homogeneous environment in The Netherlands than in Spain, but no bias or little is expected from selected base animals in data set used by Janss and de Jong (1999), as they included only records from herds with 18 years of data and, so, pedigree of animals included in the analysis trace back better to the unselected base population.

3.2.3. Other factors affecting VC estimates

Besides data structure, an important factor for explaining some of the differences of heritability estimates between results shown here and those of Hernandez et al. (1998), Charfeddine (1998) and Rekaya (1997) in different sets of Spanish data could be **number of lactations per animal.** Estimates on a subset of North of Spain with calvings before 1994 and high level of production (Pena, unpublished results, 1999), showed how heritabilities for kg milk drop down sustancially, from 0.31 to 0.256 when including one versus first three lactations, and drop to 0.244 when first 10 lactations were considered. This trend is also shown by Janss and de Jong (1999), as they found higher heritabilities for first lactations than for first three lactations (increases of 0.03 for kg protein and 0.06 for kg milk). Similar trends were reported by Mrode and Swanson (1994) and Rekaya (1997).

Other factor that may explain some differences in heritability estimates is the use of a **sire model** versus an animal model. With the same data than the one reported in previous paragraph, Pena (unpublished results, 1999), found heritability of 0.358 when applying a sire model for first lactations, higher than the previously reported value of 0.31 for an animal model. Including only first lactations and a sire model could explain the higher heritabilities obtained by Ibañez et al. (1999) against ours without including phantom groups and including base animals.

Preliminar analysis with Gibbs sampling on similar data sets than used in this work showed that **modifications on data and fixed effects** considered in the analysis do influence heritability estimates. When projecting all lactations to 305 days heritability increased +0.006 (2.7%) and repeatability +0.021 (4.6%). Elimination of data from both parents unknown and none accurate birth date increased heritability in +0.01 (4.5%). When parity was defined within herd-year-imported-season, heritability increased in +0.01 (4.5%) and repeatability +0.004 (0.8%). If lactations were pre-corrected by **heterogeneity of variance** as presently done in Spanish national evaluation following Ibañez et al. (1996), heritability increased +0.012 (5.9%) and repeatability +0.005 (1.1%). Reverter et al. (1997) also reported increases in heritabilities for different traits (4.2% in average) when applying the multiplicative model of Meuwissen et al. (1996). Althought all this changes in heritabilities and repeatabilities are sligh and some of them may not be significative, they do show a trend.

Data included in the analysis and fixed effects considered in the model may explain the slightly higher heritability found in our analysis including phantom groups, 0.2423, against results of Hernández et al. (1998), 0.2215, that also included genetic groups in the model. Some differences between both data sets are those due to modifications in data planned for next national evaluation: lactations from animals with both parents unknown and none accurate birth date are eliminated and all lactations were projected to 305 days, without considering some old lactations without test day data. Also only up to first five lactations per animal would be included.

Heterogeneity of heritabilities within a country has been described by several authors (Dong and Mao, 1990; Hill et al., 1983; De Veer and Van Vleck, 1987; Ibañez et al., 1999; Dodenhoff and Swalve, 1999). Although some different factors as data structure and effect of selected base animals could affect those reported heterogeneous heritabilities, caution should be taken about it. If assuming constant heritability and inclusion of all available data is not affordable for VC estimation, heterogeneity of heritabilities within a country could affect VC estimates depending on the criteria for selecting the subset(s) of data to be used.

Finally, as reported by Rekaya et al. (1999), correlations less than one within country is another issue and should be considered.

4. Conclusions and Suggestions

Based on results presented here heritabilities and repetabilities of kg protein should be set to 0.28 and 0.50, respectively, in Spanish national evaluation. As multiplicative model of Meuwissen et al. (1996) will be applied in Spanish national evaluation of production traits, genetic parameters should be updated considering it. Parameters for kg milk and kg fat should be estimated with the same approach.

Given the importance and impact of accurate genetic parameters assumed in national evaluations, some guidelines should be stablished by Interbull for estimating them in countries participating in international evaluations (Bagnato et al., 1996). Some suggestions that may contribute to them could be the following some of them already followed in many VC estimations:

1. Phantom groups should be included also in the estimation of variance components.

2. Effect of selected base populations in stimates of genetic variance should be considered. If selecting the data set, data and, at least, pedigrees should trace back as much as possible to better represent the unselected base population. A simple strategy could be to select herds with many years of data. Otherwise partitioning genetic variance between base and non base animals could be an alternate approach for estimation of VC and routine genetic evaluations.

3. Model used for estimating variance components should be the same than the one used in national evaluation. In example, if national evaluation uses all lactations, variance component estimation should include them also.

4. Genetic parameters should be updated when modifications of data quality or model are to be implemented in national evaluations, in example if starting to project all lactations to 305 days, modified edits or changing definition of comparison groups. Some changes would not affect heritabilties but could affect other parameters as repeatability.

5. Adjustment for heterogeneity of variances should be considered when estimating genetic parameters.

6. Because of possible heterogeneity of heritabilties, if selecting the data set an additional criteria is that it should be representative of different environments or, better, different estimates be run for different environments.

7. Movement towards test-day models with really complex variance-covariance structures may fit better the biology of the observations, but should not implicate to forgot evidence of heterogeneity of heritabilties and possible genetic correlations of less than 1 within country.

Finally, the effect of selected base populations could be also an important issue in international evaluations with MACE, but are of special importance when applying a Global Animal Model based on all available data from each country, both for estimating VC and for genetic evaluation.

Acknowledgements

The authors acknowledge Sofía Alday (CONAFE) for projecting incomplete lactations to 305 days.

References

Alfonso, L. & Estany, J. 1999. An expression of mixed animal model equations to account for different means and variances in the base population. *Gen. Sel. Evol.. 31,* 105-113.

Bagnato, A., Carnier, P., Canavesi, F., Cassandro, M. & Dadati, E. 1996. Change of Genetic Parameters for National Evaluations of Italian Holstein and Effect on International Proofs. *Interbull bulletin Nº 14.*

Bulmer, M.G. 1971. The effect of selection on genetic variability. *Am. Nat. 105,* 201.

Charfeddine, N. 1998. Selección por mérito económico global en el vacuno Frisón en España. *Tesis Doctoral. ETSI Agrónomos. Universidad Politécnica de Madrid.*

Dodenhof, J. & Swalve, H.H. 1999. Heterogeneity of variances across regions of northern Germany and adjustment in genetic evaluation. *Livest. Prod. Sci. 53,* 225-236.

Ibañez, M.A., Carabaño, M.J., Foulley, J.L. & Alenda, R. 1996. Heterogeneity of herd-period phenotypic variances in the Spanish Holstein-Friesian cattle: Sources of heterogeneity and genetic evaluation. *Livest. Prod. Sci. 45,* 137-147.

Ibañez , M.A., Carabaño, M.J. & Alenda, R. 1999. Identification os sources of heterogeneous residual and genetic variances in milk yield data from Spanish Holstein-Friesian population and impact on genetic evaluation. *Livest. Prod. Sci. 59,* 33-49.

Graser,H.-U, Smith, S.P. & Tier, B. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *J.Anim.Sci. 64,* 1362.

Hernández, D., Carabaño, M.J. & Rekaya, R. 1998. Estima de parametros genéticos para poblaciones de vacuno lechero mediante metodología bayesiana. *ITEA 19998. Vol 94ª N.º 3,*305-315.

Hernández, D., Lopez, P., Carabaño, M.J. & Alenda, R. 1999. Aplicación de la metodología bayesiana en la estima de parámetros genéticos para la producción de leche. *ITEA Vol 20, Tomo1* 318-320.

Janss, L. 1998. Maggic. A package of subroutines for genetic analyses with Gibbs sampling. *ID-Lelystad. Institute for Animal Science and Health.*

Janss, L. & de Jong, G. 1999. MCMC based estimation of variance components in a very large dairy cattle data set. *Proceedings of the Computational Cattle Breeding'99 Workshop. Tuusula, Finland. March 18-20.*

Jeyaruban, M.G. & Gibson, J.P. 1996. Estimation of additive genetic variance in commercial layer poultry and simulated populations under selection. *Theor. Appl. Genet. 92,* 483-491.

Kennedy, B.W., Schaeffer, L.R. & Sorensen, D.A. 1988. Genetic properties of animal models. *J. Dairy Sci. 71 suppl 2,* 17-26.

Koots, K.R. & Gibson, J.P. 1996. Realized sampling variances of estimates of genetic parameters and the difference between genetic and phenotipic correlations. *Genetics 143,* 1409-1416.

Li, X. & Kennedy, B.W. 1994. Comparison of genetic parameters estimates for growth rate and backfat from single and multiple trait models with and without genetic groups. *Proc. 5th World Congr. Genet. Appl. Livest. Prod., Guelph. Vol 18. Pg* 418-421.

Meuwissen, T.H.E., de Jong, G. & Engel, B. 1996. Joint estimation of breeding values and heterogeneous variances of large data files. *J. Dairy Sci. 79,* 310.

Mrode R.A. & Swanson, G.J.T. 1994. Animal model estimates of sire-herd interactions for production traits for the major dairy breeds in the United Kingdom. *Proc. 5th World Congr. Genet. Appl. Livest. Prod., Guelph. Vol 17. Pg* 19-22.

Pieramati, Van Vleck & Van der Werf. 1993. Effect of genetic groups on estimates of additive genetic variance. *J. Anim. Sci. 1993. 71,* 66-70.

Rekaya, R. 1997. Analisis Bayesianode datos de producción en los días del control para la selección de caracteres lecheros. *Tesis Doctoral. ETSI Agrónomos. Universidad Politécnica de Madrid.*

Rekaya, R., Weigel, K.A. & Gianola, D. 1999. Bayesian estimation of parameters of a structural model for genetic covariances for milk yield in five regions of USA. *Annual meeting of the European Association for Animal Production. Zurich. Switzerland. August 1999.*

Reverter, A., Tier, B., Johnston, D.J. & Graser, H.-U. 1997. Assesing th efficiency of Multiplicative Mixed Model Equations to account for heterogeneous variance across herds in carcass scan traits from beef cattle. *J. Anim. Sci. 75:1477-1485.*

Reverter, A. & Kaiser, C.J. 1997. The role of different pedigree structures on the sampling variance of heritability estimates. *J. Anim. Sci. 75,* 2355-2361.

Van Tasell, C.P., Wiggans, G.R. & Norman, H.D. 1999. Method R estimates of heritability for Milk, Fat and protein Yields of united States Dairy Cattle. *J. Dairy Sci. 82,* 2231-2237.

Van der Werf, J.H.J. & de Boer, I.J.M. 1990. Estimation of additive genetic variance when base populations are selected. *J. Anim. Sci. 1990. 68,* 3124-3132.

Van der Werf. 1992. Restricted maximum likelihood estimation of additive genetic variance when selected base animals are considered fixed. *J. Anim. Sci. 1992. 70,* 1068-1076.

Van der Werf, J.H.J. & Thompson, R. 1992. Variance decomposition in the estimation of genetic variance with selected data. *J. Anim. Sci. 1992. 70,* 2975-2985.