Genomic simulation for the study of data transformation efficiency in the genetic evaluation of cattle in the presence of heterogeneity

Antonio Policarpo Souza Carneiro[1*], Paulo Luiz Souza Carneiro[2], Ricardo Frederico Euclydes[3], Sebastião Martins Filho[1], Fabyano Fonseca e Silva[1] and Paulo Roberto Cecon[1]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, sn, 36571-000, Viçosa, Minas Gerais, Brazil. [2]Departamento de Ciências Biológicas, Universidade Estadual do Sudoeste da Bahia, Jequié, Bahia, Brazil. [3]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: policarpo@ufv.br

ABSTRACT. Efficiency of data transformation strategies is evaluated to correct heterogeneity among cattle in the genetic evaluation of bulls, cows and progenies. Four data structures of bovine weaning weights were simulated: herds with and without heterogeneity for means and variances, and herds with and without genetic connectedness. In the genetic evaluations, data were used in the original scale and transformed (Logarithmic, Square root, standardization and ratio by phenotypic standard deviation). The evaluated data transformation strategies were not efficient in eliminating the negative effects of heterogeneity among cattle in the genetic evaluation of bulls, cows and progenies.

Keywords: genetic connectedness, rank correlation, Genesys, scale transformation.

Simulação genômica para estudo da eficiência da transformação de dados para avaliação genética de bovinos na presença de heterogeneidade

RESUMO. O objetivo deste estudo foi avaliar a eficiência de estratégias de transformação de dados para correção da heterogeneidade entre rebanhos na avaliação genética de touros, vacas e progênies. Quatro estruturas de dados relativos a peso à desmama de bovinos foram simuladas: rebanhos com e sem heterogeneidade para médias e variâncias, e rebanhos com e sem conexidade genética. Nas avaliações genéticas, utilizaram-se dados na escala original e transformados (Logarítmica, Raiz quadrada, Padronização e Razão pelo desvio-padrão fenotípico). As estratégias de transformação de dados avaliadas não foram eficientes para eliminar os efeitos negativos da heterogeneidade entre rebanhos sobre a avaliação genética de touros, vacas e progênies.

Palavras-chave: conexidade genética, correlação de ordem, Genesys, transformação de escala.

Introduction

Cattle herds undergoing genetic evaluations do not generally present the same means and the same genetic and phenotype variances. In other words, they present heterogeneity for such parameters. Heterogeneity among cattle herds reduce the efficiency of genetic evaluations and impair genetic progress.

Variance heterogeneity among cattle herds for production characteristics in dairy cattle and for performance in beef cattle is amply documented (CAMPELO et al., 2003; TORRES et al., 2000). Variance heterogeneity was even found within the same cattle herd with regard to fixed effects of gender, year and date of birth, age of cow at birth, pregnancy of the animal and randomized effect of steers for data on growth rate and scrotum circumference of Canchin breed animals (FREITAS, 2000). Carneiro et al. (2008), who

worked with simulated data with different heterogeneity structures, concluded that steers classification and mainly that of cows and progenies was highly affected by heterogeneity for genetic means among cattle herds.

Logarithmic transformation of data has been applied to correct the dependence between averages and variance (EVERETT; KEOWN, 1984). Garrick and Van Vleck (1987) stated that the strategy was adequate to make reports be close to normality. However, when environments with high phenotype means and variances also demonstrate greater heritability and the transformed data invert this relationship to the point that environments with low phenotype means and variances have highest heritability, the use of logarithmic transformation provides a decrease in selection efficiency.

According to Kachman and Everett (1993), the logarithmic transformation is a simple and adequate

computing method for the correction of data when variance is a simple function of product average which does not always occur. Weight observations by residual or phenotype standard deviation have also been applied. The method is proper when residual and individual variances are proportional among the environments. However, estimates of phenotype standard deviation for each environment may produce low preciseness.

Different strategies of data transformation in genetic evaluations to correct the heterogeneity of variance and production averages among cattle herds have been evaluated (CAMPELO et al., 2003; EVERETT; KEOWN, 1984; GARRICK; VAN VLECK, 1987; KACHMAN; EVERETT, 1993; TORRES et al., 2000). However, in these research works, the authors used real data which presented variance and / or average heterogeneity and verified that scale transformations eliminated or failed to eliminate heterogeneity. However, the effect of scale transformations on the preciseness of the animals' genetic evaluations was not quantified in any study. This was due to the fact that the true genetic rate of the animals remained unknown. Such procedure was now possible through data simulation.

Data simulation is currently a highly useful tool to compare strategies of genetic evaluations. Several researchers (BRACCINI NETO et al., 2004; CARNEIRO et al., 2006, 2008; CUNHA et al., 2006; JANGARELLI et al., 2009) employed program Genesys – Genetic Simulation System (EUCLYDES, 1996) in studies related to animal genetic improvement. Simulation methodology used in the program reveals without error the true genetic rate of all simulated animals. Through simple procedures (order correlationship, coincidence of selection etc) one knows which evaluation methodology predicts genetic rates closest to the real ones.

Current paper simulates data structures for weight at weaning in cattle, taking into consideration different situations of genetic heterogeneity and connectedness among cattle herds for the evaluation of steers, cows and progenies.

Material and methods

Data used in this research were simulated by program Genesys (EUCLYDES, 1996) which presented the simulation of animal genome for different characteristics. It also simulated genes behavior during the reproduction process throughout several generations. The number of loci and alleles involved, the addition effect of each allele, the number of fixed effects and their greatness and the rates of hereditability, genetic averages, phenotype average and phenotype variance, population size, male: female proportion, number of female descendants[-1] and number of generations are defined in the Genesys simulation process. The rate of each simulated animal is obtained by the sum of addition effects of alleles in the animal's genome. Therefore, true genetic rates are known without error.

A genome of 2,000 centimorgans in length was simulated in current research for the characteristic determined by 200 loci with two alleles per locus. The alleles had only addition traits. The alleles' initial frequencies were simulated according to uniform distribution with an average of 0.50. Loci were distributed in 15 pairs of autosomal chromosomes of randomized sizes. The effect of environment with regard to 15 cattle herds and gender effect were also simulated. Randomized effects were simulated according to normal distribution. The sum of true genetic rate, gender, herd and randomized effects produced the phenotype rate of each simulated animal which, in turn, was used in the prediction of genetic rates. Greatnesses of simulated phenotype averages and variances for weight at weaning were determined by using results given by Alencar et al. (2005).

The genomes of the base-population consisting of 4,500 females and 4,500 males were simulated. Further, 75 males and 3,750 females from the above base population were randomly sampled. Each male was mated with 50 females producing an initial population of 37,500 progenies, distributed in 15 herds. The simulation process was repeated a thousand times to remove the genetic oscillation effect.

Two heterogeneity structures among herds were simulated: herds with heterogeneity for all parameters (phenotype and genetic variance and phenotype and genetic average) and herds without heterogeneity for the above parameters. Three levels of variability in terms of phenotype variance were simulated in the first data structure, or rather, herds with high, medium and low phenotype variance, with five herds per level. Herds with high phenotype variance also had high genetic variance, high phenotype average and high genetic average (Table 1).

Herds in the second data structure did not have heterogeneity for any parameter. Phenotype and genetic variances and phenotype and genetic averages were similar among the simulated herd groups (Table 2).

Besides the heterogeneity structures mentioned above, two genetic linkage standards were simulated among the herds: data without any genetic linkage in which steers with progenies in a certain herd did not have progenies in the other herds and data with full genetic linkage in which steers had progenies distributed in the 15 simulated cattle herds.

Table 1. Averages for phenotype and genetic parameters at three levels of phenotype variability for herds with heterogeneity in all parameters.

Parameters	Phenotype variability		
	High	Medium	Low
σ^2_p	637	342	204
σ^2_g	197	81	43
μ_p	285	219	150
μ_g	210	182	154

σ^2_p: Phenotype variance; σ^2_g: Genetic variance; μ_p: Phenotype average; μ_g: Genetic average.

Table 2. Averages for phenotype and genetic parameters in three herd groups for data structure without heterogeneity among the herds.

Parameters	Herd groups		
	Group 1	Group 2	Group 3
σ^2_p	725	714	727
σ^2_g	457	442	469
μ_p	228	223	223
μ_g	190	187	187

σ^2_p: Phenotype variance; σ^2_g: Genetic variance; μ_p: phenotype average; μ_g: genetic average.

The combination of two heterogeneity structures and two genetic linkage standards produced four conditions available for genetic evaluation. Simulated data were used to predict the genetic rate of steers, cows and progenies within five scales: Original, Logarithmic, Square Root, Standardized and Ratio for phenotype deviation standard of each variability level.

Data simulation by Genesys shows, without any error, the real genetic rate of each simulated animal. Order correlation (Spearman's correlation) was calculated between the predicted and true genetic rates separately for steers, cows and progenies.

Estimates of variance components and the prediction of genetic rates of simulated animals were obtained by Multiple Trait Derivative-Free Restricted Maximum Likelihood (MTDFREML) developed by Boldman et al. (1995). Statistic package SAS 9.1 (SAS, 2010) was used for the preparation of data files for analyses, for data transformation and for the calculation of Spearman's correlation between the predicted and true genetic rates.

Results and discussion

According to Crews and Franke (1998), order correlation rates below 70% may indicate important changes in the classification of animals which may jeopardize genetic selection and progress. When there is no heterogeneity among the herds and for simulated data without any genetic linkage, order correlation among genetic rates was close to 1 and varied between 82 and 96% (Table 3). However, regardless of the scale used, heterogeneity among the herds had a great influence on the genetic evaluation of steers, cows and progenies. Order correlation in such data structure between genetic rates was very low and varied between -3 and 10% (Table 3). This fact showed that animal classification based on predicted genetic rates produced a low relationship with the true classification of animals.

Order correlations among genetic rates obtained from data in the original scale and with scale transformation for data structures without and with heterogeneity among the cattle herds with and without genetic linkage were very similar in all animal categories. Order correlations among genetic rates in the presence of heterogeneity among herds and with genetic linkage were low for cows and progenies and high for steers, regardless of the transformation strategy of the applied data (Tables 3 and 4). Results show that these strategies were totally inefficient for the correction of heterogeneity among cattle herds.

Table 3. Order correlation (%) between true and predicted rates for steers, cows and progenies by different data scale transformations to correct variance heterogeneity among herds without any genetic linkage.

	Transformation				
	Original	Logarithmic	Square root	Standardized	Ratio SD[1]
	Without heterogeneity among herds				
Steers	95.75	96.32	95.87	96.45	96.25
Cows	82.41	81.89	82.28	82.16	81.87
Progenies	82.37	82.43	82.17	82.13	82.43
	With heterogeneity among herds				
Steers	2.74	2.52	-3.12	-2.57	-2.89
Cows	1.45	1.56	1.85	1.78	1.26
Progenies	10.18	10.25	7.57	8.26	8.57

[1]Ratio by phenotype deviation standard.

Table 4. Order correlation (%) between true and predicted genetic rates for steers, cows and progenies by different data scale transformations to correct variance heterogeneity among herds with genetic linkage.

	Transformation				
	Original	Logaríthmic	Square root	standardized	Ratio SD[1]
	Without heterogeneity among herds				
Steers	98.41	97.98	98.26	98.23	98.15
Cows	82.35	82.24	81.78	82.45	82.19
Progenies	83.57	83.41	83.47	83.46	83.27
	With heterogeneity among herds				
Steers	94.23	94.12	93.85	92.23	92.12
Cows	17.45	17.26	18.43	18.18	18.33
Progenies	42.14	42.23	42.41	41.35	41.28

[1]Ratio by phenotype deviation standard.

In the case of data on milk production in Dutch cows, Torres et al. (2000) also concluded that data transformation strategies: Logarithmic, Square Root, Standardization and Ratio by deviation standard did not adequately correct the differences for genetic and residual variances among the classes of phenotype variances.

When steers had progenies in all herds, or rather, for simulated data with full genetic linkage, the effect of heterogeneity varied according to the animal category but was not affected by data transformation (Table 4). Heterogeneity among herds did not affect the genetic evaluation of steers. Order correlation between true and predicted genetic rates was high and varied approximately between 92 and 94%, slightly lower than rates obtained for data without heterogeneity. However, the determination of the genetic rates of cows and progenies was jeopardized by heterogeneity even for data with full genetic linkage. Order correlations increased considerably with regard to those obtained with herds without any genetic linkage even though they did not exceed 19% for cows and 43% for progenies (Table 4).

Conclusion

Data transformation strategies (Logarithmic; Square root; Standardized and Ratio by phenotype standard deviation) were not efficient to eliminate the effects of heterogeneity among herds with regard to genetic evaluation of steers, cows and progenies, regardless of genetic linkage. New strategies for heterogeneity data correction among herds should be proposed and evaluated.

In the case of high genetic linkage, heterogeneity among herds jeopardized only the genetic evaluation of cows and progenies. In fact, data transformation failed to be a contribution towards the improvement of selection efficiency of the categories.

Acknowledgements

References

ALENCAR, M. M.; MASCIOLI, A. S.; FREITAS, A. R. Evidências de interação genótipo x ambiente sobre características de crescimento em bovinos de corte. **Revista Brasileira de Zootecnia**, v. 34, n. 2, p. 489-489, 2005.

BOLDMAN, K. G.; KRIESE, L. A.; VAN VLECK, L. D.; VAN TASSEL, C. P.; KACHMEN, S. D. **A manual for use of MTDFREML**: a set of programs to obtain estimates of variances and covariances (DRAFT). Lincoln: Department of Agriculture/Agricultural Research Service, 1995.

BRACCINI NETO, J.; SILVA, M. A.; EUCLYDES, R. F.; LOPES, P. S.; TORRES, R. A.; REGAZZI, A. J. Comparação de metodologias de predição de valores genéticos utilizando dados simulados. **Revista Brasileira de Zootecnia**, v. 33, n. 6, p. 1683-1688, 2004.

CAMPELO, J. E. G.; LOPES, P. S.; TORRES, R. A.; SILVA, L. O. C.; EUCLYDES, R. F.; ARAÚJO, C. V.; PEREIRA, C. S. Influência da heterogeneidade de variâncias na avaliação genética de bovinos de corte da raça Tabapuã. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 6, p. 685-693, 2003.

CARNEIRO, A. P. S.; TORRES, R. A.; LOPES, P. S.; EUCLYDES, R. F.; CARNEIRO, P. L. S.; CUNHA, E. E. Erros na classificação de touros, vacas e touros jovens geneticamente superiores avaliados na presença de heterogeneidade. **Revista Brasileira de Zootecnia**, v. 35, n. 5, p. 1907-1914, 2006.

CARNEIRO, A. P. S.; TORRES, R. A.; LOPES, P. S.; EUCLYDES, R. F.; CARNEIRO, P. L. S.; SILVA, F. F. Heterogeneidade e avaliação genética em bovinos, estudo utilizando dados simulados. **Acta Scientiarum. Animal Sciences**, v. 30, n. 1, p. 113-119, 2008.

CREWS, D. H.; FRANKE, D. E. Heterogeneity of variances for carcass traits by percentage Brahman inheritance. **Journal of Animal Science**, v. 76, n. 7, p. 1803-1809, 1998.

CUNHA, E. E.; EUCLYDES, R. F.; TORRES, R. A.; CARNEIRO, P. L. S. Simulação de dados para avaliação genética de rebanhos de gado de corte. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 58, n. 3, p. 381-387, 2006.

EUCLYDES, R. F. **Genesys**: sistema de simulação genética. Viçosa: UFV, 1996.

EVERETT, R. W.; KEOWN, J. F. Mixed model sire evaluation with dairy cattle – Experience and genetic gain. **Journal of Animal Science**, v. 59, n. 2, p. 529-541, 1984.

FREITAS, A. R. Avaliação de procedimentos na estimação de parâmetros genéticos em bovinos de corte. **Revista Brasileira de Zootecnia**, v. 29, n. 1, p. 94-102, 2000.

GARRICK, D. J.; VAN VLECK, L. D. Aspects of selection for performance in several environments with heterogeneous variances. **Journal of Animal Science**, v. 65, n. 2, p. 409-421, 1987.

JANGARELLI, M.; EUCLYDES, R. F.; NOGUEIRA, A. P. O. Avaliação de diferentes níveis de significância na identificação de marcadores moleculares no melhoramento genômico. **Revista Ceres**, v. 56, n. 2, p. 134-139, 2009.

KACHMAN, S. D.; EVERETT, R. W. A multiplicative mixed model when the variances are heterogeneous. **Journal of Dairy Science**, v. 76, n. 3, p. 859-867, 1993.

SAS-Statistical Analysis System. **User's guide**. Versão 9.1. Cary: SAS Institute Inc., 2010.

TORRES, R. A.; BERGMANN, J. A. G.; COSTA, C. N.; PEREIRA, C. S.; VALENTE, J.; PENNA, V. M.; TORRES FILHO, R. A.; ARAÚJO, C. V. Heterogeneidade de variância e avaliação genética de bovinos da raça Holandesa no Brasil. **Revista Brasileira de Zootecnia**, v. 29, n. 4, p. 1050-1059, 2000.