

# Predição de valores genéticos utilizando inferência bayesiana e frequentista em dados simulados

José Marques Carneiro Júnior<sup>1\*</sup>, Giselle Mariano Lessa de Assis<sup>1</sup>, Ricardo Frederico Euclides<sup>2</sup>, Williane Maria de Oliveira Martins<sup>1</sup> e Priscila Ferreira Wolter<sup>1</sup>

<sup>1</sup>Empresa Brasileira de Pesquisa Agropecuária, Br 364, km 14, 69908-970, Rio Branco, Acre, Brasil. <sup>2</sup>Departamento de Zootecnia, Centro de Ciências Agrárias, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brasil. \*Autor para correspondência. E-mail: marques@cpafac.embrapa.br

**RESUMO.** Dados simulados foram utilizados para comparar as metodologias Eblup e Bayesiana, em dados com homogeneidade de variâncias, heterogeneidade de variância genética e heterogeneidade de variância genética e ambiental. Para obtenção dessas estruturas foram feitos descartes estratégicos dos valores genéticos aditivos e ambientais de acordo com o tipo de heterogeneidade e o nível de variabilidade desejada (alta, média ou baixa), sendo utilizados dois tamanhos de população (grande e pequena). Para a metodologia Bayesiana foram utilizados três níveis de informação *a priori*: não informativo, pouco informativo e informativo. A presença da heterogeneidade de variâncias causa problemas para a seleção dos melhores indivíduos, principalmente se a heterogeneidade estiver nos componentes de variância genética e ambiental, sendo os animais selecionados equivocadamente do ambiente mais variável. Os métodos comparados tiveram resultados semelhantes, quando distribuições *a priori* não informativas foram utilizadas, e as populações de tamanho grande, de modo geral, apresentaram melhores predições de valores genéticos. Foi observado, para a metodologia Bayesiana, que o aumento no nível de informação *a priori* influencia positivamente as predições dos valores genéticos, principalmente para as populações pequenas. O método Bayesiano é indicado para populações de tamanho pequeno quando há disponibilidade de distribuições *a priori* informativas.

**Palavras-chave:** heterogeneidade de variâncias, componentes de variância, simulação, informação *a priori*.

**ABSTRACT. Prediction of genetic values using bayesian inference and frequent on simulated data.** Simulated data were used to compare EBLUP and Bayesian methods in data with homogeneity of variance, heterogeneity of variance and genetic heterogeneity of genetic and environmental variance. For these structures were strategic disposal of additive genetic and environmental values in accordance with the type of heterogeneity and the desired level of variability: high, medium or low. We used two sizes of population: large and small. For the Bayesian methodology was used three levels of a priori information: no information, just information and informative. For verification of the introduction of different levels of information they were used the mistake percentage in relation to the true value of the variance components the Spearman correlation and the medium square of the mistake among the real genetic values and predicted them. The presence of heterogeneity of variances cause problems for the selection of the best individuals, especially if the heterogeneity is present in the components of genetic variance and environmental and animals are mistakenly selected the more variable environment. The methods presented similar results when compared not informative priors were used, and the populations of large size, in general, showed better prediction of breeding values. Was observed for the Bayesian methodology, the increase in the level of a priori information positively influences the predictions of genetic values, especially for small populations. The Bayesian method is preferred for populations of small size when there is availability of informative priors.

**Key words:** heterogeneity of variance, variance components, simulation, priori information.

## Introdução

A identificação e o acasalamento de animais geneticamente superiores são essenciais para o progresso genético de uma população. Entretanto, é necessário o conhecimento das fontes de variação

que compõem o fenótipo da característica de interesse. A parcela livre dos efeitos de ambiente corresponde aos valores genéticos dos indivíduos em comparação. A obtenção desses valores depende de métodos estatísticos sofisticados, capazes de indicar

quais animais devem ser selecionados, a partir da estrutura de dados disponíveis.

A metodologia de modelos mistos vem sendo amplamente empregada no melhoramento animal para obtenção do melhor preditor linear não viesado (Blup) dos valores genéticos. Porém, essa metodologia pressupõe que os componentes de variância sejam conhecidos, o que raramente ocorre na prática. Quando tais componentes são desconhecidos, utiliza-se como estratégia de análise a substituição dos valores paramétricos dos componentes de variância por estimativas de máxima verossimilhança restrita (Reml), recebendo esse método o nome de Blup Empírico (Eblup).

Em populações com tamanho suficiente para que apresentem normalidade e homogeneidade de variâncias, essa aproximação parece ser satisfatória. No entanto, estimativas dos componentes de variância e coeficientes de herdabilidade têm-se mostrado heterogêneas de acordo com diferentes níveis de produção e diferentes classes de desvio-padrão genético e/ou ambiental. Quando a heterogeneidade não é considerada, diferenças de variâncias dentro das subclasses podem resultar na predição de valores genéticos viesados, redução no progresso genético e desproporcional número de animais selecionados de ambientes com diferentes variâncias (WEIGEL; GIANOLLA, 1992).

Os métodos, que buscam eliminar a heterogeneidade de variâncias, consistem na transformação de dados ou na aplicação de fatores de ajustamento de forma que os dados transformados ou ajustados apresentem homogeneidade de variâncias. Diversos autores, entre eles Cardoso et al. (2001) relatam que as transformações de dados podem conduzir a pressuposições nem sempre realísticas e que a utilização de fatores multiplicativos, apesar da facilidade computacional, pode produzir avaliações genéticas viesadas.

Embora nas avaliações genéticas em que se utiliza a metodologia de modelos mistos seja possível considerar a heterogeneidade de variâncias, por meio da utilização de modelos multivariados, nem sempre o conjunto de dados é grande o suficiente para prover informações em cada nível de variabilidade. Segundo Blasco (1998), mesmo grandes conjuntos de dados com presença de heterogeneidade de variâncias podem apresentar pouca informação. Nessas situações, métodos que necessitam de justificativa assintótica, como o Eblup, podem produzir predições de valores genéticos não confiáveis dentro das subclasses.

Recentemente, métodos Bayesianos vêm sendo utilizados como uma opção para solucionar

problemas relacionados à avaliação de mérito genético em populações animais. A aplicação de métodos Markov Chain Monte Carlo (MCMC), dentre os quais se destaca a amostragem de Gibbs, pode ser utilizada de forma a propiciar uma inferência Bayesiana. Gibbs Sampling (GS) é uma técnica de integração numérica por simulação, muito usual em situações nas quais a integração analítica completa é impossível. O GS é aplicável à estimação de componentes de variância e predição dos valores genéticos, permitindo, por suas propriedades, a inferência Bayesiana.

Kachamn e Everett (1993) propõem a utilização de um modelo misto multiplicativo, utilizando a inferência Bayesiana, que ignora a relação linear entre variâncias e médias. Segundo os autores, o método proposto funciona bem quando a herdabilidade e a repetibilidade se mantêm constantes ao longo das classes de heterogeneidades de variâncias, pressuposição não realista para a maioria dos conjuntos de dados verificados no melhoramento genético animal.

Oliveira et al. (2000), trabalhando com dados reais de animais da raça Guzera, compararam a aplicação da análise Bayesiana na estimação de componentes de variância e covariância e na predição dos valores genéticos, com os resultados obtidos pela metodologia Reml. Segundo os autores, a análise Bayesiana apresenta grande flexibilidade tanto em relação aos modelos que podem ser utilizados nas análises quanto às inferências que podem ser realizadas a partir dos resultados.

Dessa forma, a heterogeneidade de variâncias representa um problema a ser resolvido. A solução mais comumente utilizada pela metodologia frequentista é o uso de um modelo misto multivariado. No entanto, essa medida pode resultar em outro problema, como por exemplo, a estimação em pequenos conjuntos de dados.

Nesse contexto, teoricamente a inferência Bayesiana apresenta-se como uma ferramenta promissora por levar em consideração a incerteza sobre os componentes de variâncias, permitir a inclusão de informação *a priori* e apresentar facilidade de implementação de distribuições mais robustas.

Dentro do contexto apresentado, os objetivos deste trabalho foram: verificar o efeito de diferentes tipos de heterogeneidade de variâncias e do tamanho da população na predição dos valores genéticos, bem como a influência do nível de informação *a priori* ao utilizar metodologia Bayesiana, para predição dos valores genéticos e comparar a metodologia frequentista com a Bayesiana na análise genética em dados com presença da heterogeneidade de variâncias.

**Material e métodos**

Os dados utilizados neste trabalho foram simulados por meio do sistema computacional Genesys (EUCLYDES, 1996). Desenvolvido em linguagem Fortran, esse sistema permite a simulação de genomas de certa complexidade para formação de populações de indivíduos que podem ser utilizadas em diversos estudos de avaliação genética.

Foi simulado um genoma de 3.000 centimorgans (não encontramos no dicionário) de comprimento, considerando uma única característica quantitativa, governada por 800 locos com dois alelos cada um. Os locos quantitativos foram simulados segundo a distribuição normal, com média zero e desvio-padrão de 20, sendo distribuídos em 20 pares de cromossomos de tamanhos aleatórios.

A função de distribuição de probabilidade, usada para simular as frequências iniciais dos locos quantitativos nas populações, foi a uniforme, com média de 0,50. Foi utilizado um efeito sistemático de rebanho com 6 níveis e os efeitos de ambiente não controláveis foram simulados conforme a distribuição normal com média zero e desvio-padrão 30, sendo o processo de simulação repetido 100 vezes.

Segundo a estrutura genômica proposta para formar a população-base, foram simulados 1.500 machos e 1.500 fêmeas. Essa população foi utilizada para formar duas populações iniciais, uma grande e outra pequena. Para a primeira foram escolhidos aleatoriamente 50 machos e 1.500 fêmeas, que deram origem a 3.000 descendentes (dois descendentes por fêmea). Essa população foi submetida a uma geração de seleção ao acaso para formar a população inicial grande com 6.000 descendentes. Na formação da população inicial pequena foram escolhidos aleatoriamente

4 machos e 120 fêmeas, que deram origem a 240 descendentes (dois descendentes por fêmea). Essa população foi submetida a uma geração de seleção ao acaso para formar a população inicial pequena com 480 descendentes.

Visando avaliar o efeito dos diferentes tipos de heterogeneidade de variâncias, foram realizados descartes estratégicos, dos valores máximos e mínimos genéticos e/ou ambientais, a fim de se obter a heterogeneidade e o nível de variabilidade desejada. Esse descarte provocou uma redução no número de indivíduos, tanto para a população de tamanho grande quanto para a pequena. Na Tabela 1 constam os códigos utilizados para representar as populações com diferentes tipos de heterogeneidade de variâncias, níveis de variabilidade e tamanhos de população.

Dois conjuntos de análises foram realizados para as metodologias frequentista e Bayesiana. No primeiro, desconsiderou-se a presença de heterogeneidade de variâncias admitindo-se que todas as subpopulações pertenciam à mesma população, sendo as análises realizadas com a população total.

No segundo conjunto, as análises foram realizadas isoladamente para cada subnível de variabilidade. O seguinte modelo linear misto unicaracterístico foi empregado nos dois conjuntos de análises:

$$Y_{ij} = + EF_i + a_{ij} + e_{ij}$$

em que:  $y_{ij}$  é a observação fenotípica do  $j$ -ésimo animal no  $i$ -ésimo rebanho EF;  $\mu$  é uma constante inerente a todas as observações;  $EF_i$  é o efeito do  $i$ -ésimo rebanho;  $a_{ij}$  efeito aleatório genético aditivo do  $j$ -ésimo animal no  $i$ -ésimo rebanho EF;  $e_{ij}$  é o efeito aleatório ambiental.

**Tabela 1.** Código utilizado para as populações com diferentes tipos de heterogeneidade de variâncias, níveis de variabilidade e tamanhos de população.

Estrutura dos dados	Nível de variabilidade	Tamanho da população		Número de indivíduos	
		População pequena (P)	População grande (G)	(P)	(G)
Homogeneidade de variâncias (HO)	População 1 (1)	HO-P-1	HO-G-1	100	1.600
	População 2 (2)	HO-P-2	HO-G-2	100	1.600
	População 3 (3)	HO-P-3	HO-G-3	100	1.600
	Total (T)	HO-P-T	HO-G-T	300	4.800
Heterogeneidade de variância genética (HG)	Alta (A)	HG-P-A	HG-G-A	100	1.600
	Média (M)	HG-P-M	HG-G-M	100	1.600
	Baixa (B)	HG-P-B	HG-G-B	100	1.600
	Todas (T)	HG-P-T	HG-G-T	300	4.800
Heterogeneidade de variância genética e ambiental (HGA)	Alta (A)	HGA-P-A	HGA-G-A	100	1.600
	Média (M)	HGA-P-M	HGA-G-M	100	1.600
	Baixa (B)	HGA-P-B	HGA-G-B	100	1.600
	Todas (T)	HGA-P-T	HGA-G-T	300	4.800

T = (A+M+B).

Esse modelo pode ser descrito sob a forma matricial a seguir, pressupondo-se que  $a \sim N(0, A\sigma_a^2)$  e  $e \sim N(0, I\sigma_e^2)$ :

$$y = X + Za + e$$

em que:  $y$  é o vetor das observações dos indivíduos;  $X$  é a matriz de incidência de efeitos fixos;  $\beta$  é o vetor de efeitos fixos (rebanho);  $Z$  é a matriz de incidência dos efeitos aleatórios;  $a$  é o vetor de efeitos aleatórios (valores genéticos dos indivíduos) e  $e$  refere-se ao vetor de erros aleatórios.

As estimativas dos componentes de variâncias e predição dos valores genéticos, pelo método frequentista, foram obtidas utilizando-se o programa Multiple Trait Derivative-Free Restricted Maximum Likelihood (MTDFREML), descrito por Boldman et al. (1995).

Para o método Bayesiano, as estimativas foram obtidas utilizando-se o programa Multiple Trait Gibbs Sampling in Animal Models (MTGSAM), descrito por Van Tassel e Van Vleck (1995). Nas análises realizadas pela metodologia Bayesiana, foram utilizados três níveis de informações *a priori*: não informativas (NI), pouco informativas (PI) e informativas (I).

No nível não informativo, esses valores resultaram em distribuição uniforme para as variâncias. Nos níveis pouco informativos e informativos, os valores do hiperparâmetro  $S_i^2$  corresponderam aos valores reais dos componentes de variância e os do hiperparâmetro  $V_i$  foram definidos fixando-se o valor de  $S_i^2$  desejado e alterando-se os de  $V_i$ , de modo que maior ou menor grau de confiança fosse atribuído às variâncias informadas. Esse procedimento foi realizado por meio da análise gráfica das distribuições de qui-quadrado invertidas escalonadas.

O critério de convergência e as análises das cadeias de Gibbs foram baseados na inspeção visual dos gráficos do valor de cada amostra e no programa GIBANAL (VAN KAAM, 1998).

#### Correlações entre os valores genéticos

As correlações de ordem entre valores genéticos preditos e os reais foram obtidas pelo pacote estatístico Statistical Analysis System® (SAS, 1990). Por meio da correlação não paramétrica de Spearman foi possível verificar o efeito do tipo de heterogeneidade, do nível de variabilidade, do método estudado e do nível de informação *a priori* na classificação dos indivíduos.

#### Quadrado médio do erro (QME)

Os valores genéticos preditos foram comparados com os valores genéticos reais, por meio da média do erro ao quadrado, como em Harvey (1979).

$$QME = \sum_{i=1}^n \frac{1}{n} (\hat{a}_i - a_i)^2$$

em que: QME é o quadrado médio do erro,  $n$  é o número de indivíduos,  $\hat{a}_i$  refere-se ao valor genético predito do  $i$ -ésimo animal e  $a_i$  ao valor genético real. Quanto mais próximo de zero maior é a acurácia da predição dos valores genéticos.

#### Resultados e discussão

##### Correlação de Spearman

Na Tabela 2 são apresentadas as correlações não paramétricas de Spearman para as populações totais e subpopulações. Quando a presença da heterogeneidade de variâncias foi desconsiderada, as análises foram conduzidas com as populações totais. A população de tamanho grande com homogeneidade de variâncias apresentou correlação de ordem de 0,78, sendo os métodos similares entre si, e o aumento no nível de informação não resultou em melhor classificação dos indivíduos. Entretanto, para a população de tamanho pequeno, com homogeneidade de variâncias, o aumento no nível de informação elevou a correlação de 0,61, com distribuições não informativas, para 0,66 no nível informativo. Pode-se verificar ainda que foram obtidas menores correlações para as demais estruturas de população.

Na população de tamanho pequeno com heterogeneidade genética, o aumento no nível de informação também ocasionou melhor classificação dos indivíduos, sendo o valor obtido pelo método Bayesiano, com distribuições *a priori* não informativas de 0,60, superior ao obtido pelo método Eblup de 0,57, contudo esse resultado não é suficiente para maiores conclusões.

As menores correlações foram encontradas quando a heterogeneidade estava presente nos componentes genéticos e ambientais, sendo de 0,43 para a população pequena e de 0,54 para a grande.

Ressalta-se que a presença da heterogeneidade de variância reduz a eficiência da classificação dos indivíduos, principalmente se for genética e ambiental, indicando que maiores cuidados devem ser tomados na padronização de práticas de manejo, nutrição dos animais e aspectos sanitários dos rebanhos. De forma geral, as correlações foram menores para as populações pequenas, mantendo-se

sempre abaixo do valor de 0,70, indicando grandes alterações na classificação dos animais.

**Tabela 2.** Correlações não paramétricas entre os valores genéticos reais e preditos, obtidos pela metodologia Eblup e Bayesiana sob três níveis de informação, para as populações totais e subpopulações com três níveis de variabilidade e dois tamanhos nas diferentes estruturas de variância.

Populações	Eblup	Bayesiana		
		NI	PI	I
Homogeneidade de variâncias				
HO-P-1	FC	0,48	0,50	0,50
HO-P-2	0,60	0,60	0,59	0,59
HO-P-3	0,69	0,68	0,69	0,68
HO-P-T	0,59	0,61	0,64	0,66
HO-G-1	0,75	0,75	0,75	0,75
HO-G-2	0,78	0,78	0,78	0,78
HO-G-3	0,75	0,75	0,75	0,75
HO-G-T	0,78	0,78	0,78	0,78
Heterogeneidade de variância genética				
HG-P-A	0,69	0,69	0,71	0,70
HG-P-M	0,61	0,60	0,61	0,61
HG-P-B	FC	0,38	0,38	0,38
HG-P-T	0,57	0,60	0,64	0,66
HG-G-A	0,75	0,75	0,75	0,75
HG-G-M	0,68	0,69	0,69	0,70
HG-G-B	0,58	0,58	0,58	0,59
HG-G-T	0,71	0,72	0,72	0,72
Heterogeneidade de variância genética e ambiental				
HGA-P-A	0,41	0,38	0,40	0,40
HGA-P-M	FC	0,40	0,41	0,41
HGA-P-B	0,36	0,34	0,35	0,35
HGA-P-T	0,42	0,43	0,43	0,44
HGA-G-A	0,52	0,52	0,52	0,52
HGA-G-M	0,49	0,49	0,49	0,49
HGA-G-B	0,45	0,46	0,46	0,46
HGA-G-T	0,54	0,54	0,54	0,54

NI = não informativo; PI = pouco informativo; I = informativo; FC = falhou a convergência.

Quando as análises foram realizadas dentro das subclasses de variâncias, de modo geral, obtiveram-se estimativas de correlações baixas para os tipos de heterogeneidade estudados, em ambos os tamanhos de população. Os métodos Eblup e Bayesiano foram similares na classificação dos indivíduos, e o aumento no nível de informação somente em alguns casos produziu efeito positivo na classificação dos indivíduos.

As populações com alta variabilidade genética aditiva apresentaram correlações superiores quando comparadas às de baixa variabilidade. Uma possível explicação para esse fato é que nas populações de maior variabilidade genética as diferenças entre os valores genéticos dos indivíduos são mais pronunciadas, facilitando aos métodos uma classificação mais acurada dos indivíduos. Os valores mais altos de correlação foram observados para as subpopulações grandes, oriundas da população com homogeneidade de variâncias, independentemente do método empregado ou do nível de informação utilizado, sendo 0,75; 0,78 e 0,75 para os níveis de alta, média e baixa variabilidade, respectivamente.

Nas subpopulações de tamanho pequeno, os valores encontrados foram os mais problemáticos,

indicando grandes alterações na classificação dos indivíduos e baixa acurácia na predição dos valores genéticos. Provavelmente, esse fato ocorreu porque as subpopulações formadas foram muito pequenas, apresentando desvios na normalidade dos dados.

Em algumas situações, as análises realizadas nos subníveis apresentaram valores mais altos de correlação de Spearman, em relação àquelas com as populações totais. Por exemplo, quando a presença de heterogeneidade de variância genética aditiva foi desconsiderada, utilizando-se a metodologia Eblup, o valor da correlação obtido para a população total foi de 0,71, enquanto as análises realizadas nos subníveis formados apresentaram resultados de 0,75; 0,68 e 0,58 para alta, média e baixa variabilidade. Portanto, desconsiderar a presença de heterogeneidade de variâncias genética pode causar prejuízos para os subníveis de variabilidade formados, sendo preferíveis as análises realizadas isoladamente em cada estratificação de variância.

#### Quadrado médio do erro

Na Tabela 3 são apresentados os quadrados médios dos erros (QME), entre os valores genéticos reais e preditos, obtidos pela metodologia Eblup e Bayesiana, sob três diferentes níveis de informação *a priori*.

De forma geral, observou-se que os QMEs obtidos para as populações totais foram menores nas populações de tamanho grande e maiores nas de tamanho pequeno, indicando menor acurácia das predições dessas populações, como já havia sido constatado anteriormente. Os resultados foram similares para as duas metodologias. Porém, a utilização de distribuições *a priori* mais informativas pela metodologia Bayesiana, em todas as situações, conduziu à predição de valores genéticos ligeiramente mais acurados, principalmente nas populações pequenas. Para a estrutura com homogeneidade de variâncias e de tamanho pequeno, os valores obtidos foram de 137,45; 122,98 e 110,35, para os níveis NI, PI e I, respectivamente. Quando a heterogeneidade genética foi desconsiderada, os valores obtidos foram de 131,14; 112,91 e 99,66 para os mesmos níveis. Esses resultados demonstram como informações adicionais *a priori* podem contribuir para o aumento da acurácia na predição dos valores genéticos. Entretanto, é possível observar que para as populações de tamanho grande o aumento da acurácia, proporcionado pela utilização de níveis mais informativos, foi pouco expressivo demonstrando a maior importância da informação *a priori* para as populações de tamanho pequeno.

**Tabela 3.** QME entre os valores genéticos reais e os preditos pela metodologia Eblup e Bayesiana sob três níveis de informação *a priori*, para as populações totais e subpopulações com três níveis de variabilidade e dois tamanhos nas diferentes estruturas de variância.

Populações	Eblup	Bayesiana		
		NI	PI	I
Homogeneidade de variâncias				
HO-P-1	FC	167,63	126,82	118,11
HO-P-2	144,96	139,55	127,44	128,14
HO-P-3	100,79	168,84	100,96	101,28
HO-P-T	139,93	137,45	122,98	110,35
HO-G-1	126,51	125,27	125,45	124,59
HO-G-2	113,87	113,36	113,40	113,23
HO-G-3	128,44	129,58	127,41	125,42
HO-G-T	113,19	113,40	113,33	112,32
Heterogeneidade de variância genética				
HG-P-A	188,76	202,12	109,66	109,88
HG-P-M	106,47	102,85	105,37	106,12
HG-P-B	FC	114,55	105,35	103,88
HG-P-T	133,01	131,14	112,91	99,66
HG-G-A	126,53	125,52	125,42	124,65
HG-G-M	123,33	117,02	115,00	102,83
HG-G-B	95,06	92,72	92,76	91,16
HG-G-T	100,40	99,74	99,74	99,74
Heterogeneidade de variância genética e ambiental				
HGA-P-A	233,91	222,70	214,54	213,67
HGA-P-M	FC	126,85	116,67	113,78
HGA-P-B	82,72	84,18	82,23	82,43
HGA-P-T	153,50	146,32	134,05	129,82
HGA-G-A	248,09	243,39	244,15	242,33
HGA-G-M	148,55	145,98	146,46	146,19
HGA-G-B	100,43	97,06	97,12	96,13
HGA-G-T	154,40	152,62	152,82	151,97

NI = não informativo; PI = pouco informativo; I = informativo; FC = fálhou a convergência.

Ainda na Tabela 3 são apresentados os QMEs quando as análises foram realizadas dentro de cada subnível de variabilidade. Pode-se observar que as duas metodologias foram semelhantes entre si. Contudo, uma das vantagens apresentadas pela metodologia Bayesiana é que mesmo nas situações nas quais a metodologia REML não obteve convergência, o QME obtido com essa metodologia foi similar aos das demais subpopulações do grupo estudado e o nível de informação *a priori* proporcionou uma melhoria considerável na acurácia da predição.

O nível de informação empregado para a metodologia Bayesiana, principalmente nas populações pequenas, proporcionou uma queda no QME em todas as análises realizadas. Apesar desse avanço na acurácia da predição dos valores genéticos não ter sido suficiente para aperfeiçoar a classificação dos animais, poderia ter sido mais importante em populações cujos animais apresentassem méritos genéticos mais próximos entre si, de modo que uma pequena melhora na acurácia fosse suficiente para alterar a classificação dos indivíduos. Outro fator a ser considerado é que melhor acurácia de valores genéticos pode resultar em melhores predições de ganhos genéticos.

As análises realizadas dentro dos subníveis de variabilidade conduziram a maiores valores de

QME, quando comparadas aos QMEs resultantes das análises com as populações totais. Entretanto, esses resultados foram diferentes daqueles encontrados para as correlações, sendo os valores obtidos maiores para as análises realizadas dentro dos subníveis. Uma provável explicação para esses resultados é que os métodos podem estar conduzindo a erros de proporções semelhantes para todos os indivíduos, não prejudicando a classificação, apesar de provocar aumentos no QME.

Para a estrutura com heterogeneidade de variância genética aditiva, a subpopulação pequena, com alta variância, apresentou valor de QME igual a 188,76, sendo esse o maior valor encontrado por meio da metodologia frequentista. Por outro lado, apresentou correlação igual a 0,69, sendo a maior dentre as subpopulações pequenas, indicando que apesar dessa situação de alta variabilidade ter resultado em uma predição menos acurada, conduziu a uma classificação mais correta.

Dessa forma, para efeito de comparação, preferem-se os resultados obtidos quanto às correlações de Spearman, em que as análises realizadas dentro das subclasses foram indicadas como as mais adequadas.

#### Porcentagens de indivíduos selecionados em comum

Com o objetivo de verificar o efeito da heterogeneidade de variâncias, dos níveis de informação *a priori* e do tamanho de população sobre a seleção, calcularam-se as porcentagens de indivíduos comuns para 15% dos melhores indivíduos oriundos das análises realizadas com as populações totais (Tabela 4).

A porcentagem de indivíduos em comum, obtidos para a estrutura HO-G-T, foi de 62% para ambas as metodologias, sendo o maior valor encontrado nas diferentes estruturas de heterogeneidade de variâncias. Para a estrutura HO-P-T, considerando homogeneidade em análise Bayesiana, foi de 42, 44 e 47% para os níveis NI, PI e I, respectivamente, sendo levemente melhor no nível mais informativo. A introdução de níveis mais informativos apresentou, em apenas alguns casos, resultados mais acurados, indicando que uma pequena melhora pode ocorrer na acurácia quando maiores níveis de informação a respeito dos componentes de variâncias estão disponíveis.

Na estrutura HG-G-T, o valor encontrado foi de 56% para as metodologias e níveis de informação, demonstrando uma queda da porcentagem relacionada à situação com homogeneidade de variâncias. Além disso, os valores mais baixos foram encontrados na estrutura com heterogeneidade genética e ambiental, sendo de 42% para a população

grande e de 40% para a pequena, confirmando, portanto, que ao se desconsiderar esse tipo de heterogeneidade obtêm-se os piores resultados tanto para a correlação quanto para o QME.

**Tabela 4.** Porcentagem de indivíduos comum, entre os 15% melhores valores genéticos reais e os preditos pelas metodologias Eblup e Bayesiana sob três níveis de informação, para as populações totais e subpopulações com três níveis de variabilidade em dois tamanhos nas diferentes estruturas de população.

Populações	Eblup	Bayesiana		
		NI	PI	I
Homogeneidade de variâncias				
HO-P-1	FC	60	53	60
HO-P-2	47	47	47	47
HO-P-3	53	60	53	53
HO-P-T	42	42	44	47
HO-G-1	61	62	62	62
HO-G-2	59	60	60	60
HO-G-3	58	58	58	58
HO-G-T	62	62	62	62
Heterogeneidade de variância genética				
HG-P-A	53	53	53	53
HG-P-M	40	40	40	40
HG-P-B	FC	27	27	27
HG-P-T	44	47	47	47
HG-G-A	69	62	62	62
HG-G-M	49	49	48	49
HG-G-B	42	43	43	44
HG-G-T	56	56	56	56
Heterogeneidade de variância genética e ambiental				
HGA-P-A	27	27	27	20
HGA-P-M	FC	27	27	27
HGA-P-B	40	40	40	40
HGA-P-T	38	40	40	40
HGA-G-A	42	42	42	42
HGA-G-M	38	37	37	37
HGA-G-B	35	36	35	36
HGA-G-T	42	42	42	42

NI= não informativo; PI = pouco informativo; I = informativo; FC = falhou a convergência.

Na Tabela 4 são apresentadas as porcentagens de indivíduos em comum quando as análises foram realizadas dentro dos subníveis de variabilidade. Pode-se verificar que os resultados obtidos para as subpopulações de tamanho grande continuaram sendo maiores que os das subpopulações de tamanho pequeno, sendo as duas metodologias avaliadas semelhantes entre si. Além disso, o aumento do nível de informação *a priori* não produziu melhores resultados, indicando que a melhoria da acurácia, constatada em análises anteriores para os componentes de variância, não foi refletida na mesma proporção para a seleção dos melhores indivíduos.

Para a estrutura de dados com heterogeneidade de variâncias genéticas (HG), os ambientes mais variáveis apresentaram maiores porcentagens de indivíduos em comum, sendo de 62, 49 e 43% para os subníveis de alta, média e baixa variabilidade, respectivamente, em análise Bayesiana NI. Essa tendência também foi verificada para a estrutura com heterogeneidade

de variâncias genéticas e ambientais (HGA). Segundo Winkelman e Schaeffer (1988), sob seleção intensa e variâncias heterogêneas entre rebanhos, a escolha de touros e vacas poderia resultar na seleção de animais que não são geneticamente os melhores. Segundo esses autores, quando a heterogeneidade é causada por fatores ambientais, a avaliação genética dos animais poderia refletir mais o ambiente do que o seu potencial genético.

### Conclusão

A predição dos valores genéticos dos animais é mais acurada quando maiores níveis de informação são utilizados. Porém, essa maior acurácia não é refletida, na mesma proporção, para a classificação dos animais.

A presença da heterogeneidade de variâncias causa problemas para a seleção dos melhores indivíduos, principalmente se estiver presente nos componentes de variância genética e ambiental, sendo os animais selecionados equivocadamente do ambiente mais variável.

O método Bayesiano é preferido para populações de tamanho pequeno quando há disponibilidade de informação *a priori*.

### Referências

BLASCO, A.; SORENSEN, D.; BIDANEL, J. P. Bayesian inference of genetic parameters and selection response for litter size components in pigs. **Genetics**, v. 149, n. 1, p. 301-306, 1998.

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

CARDOSO, F. F.; CARDELLINO, R. A.; CAMPOS, L. T. Fatores ambientais sobre escores de avaliação visual à desmama em bezerros Angus criados no Rio Grande do Sul. **Revista Brasileira de Zootecnia**, v. 30, n. 2, p. 318-325, 2001.

EUCLYDES, R. F. **Sistema de simulação genética Genesys**: genetic system simulation. Vigosa: UFV, 1996.

HARVEY, W. R. **LSMLMW-Mixed model least squares and maximum likelihood computer program**. Ohio: Ohio State University, 1979.

KACHAMN, 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.

OLIVEIRA, H. N.; LÔBO, R. B.; PEREIRA, C. S. Comparação de modelos não lineares para descrever o crescimento de fêmeas da raça Guzerá. **Pesquisa Agropecuária Brasileira**, v. 35, p. 1843-1851, 2000.

SAS-Statistical Analysis System. **Statistical analysis system user's guide**: statistic. Cary: Statistical Analysis System Institute, 1990.

VAN KAAM, J. B. C. H. M. **Gibanal**. Analyzing program for Markov Chain Monte Carlo Sequences (Version 2.10). Wageningen: Wageningen Agricultural University, 1998. (Manual).

VAN TASSELL, C. P.; VAN VLECK, L. D. **A manual for use of MTGSAM**. A set of Fortran programs to apply Gibbs Sampling to animal models for variance components estimation. Lincoln: Department of Agriculture; Agricultural Research Service, 1995. (Manual).

WEIGEL, D. A.; GIANOLA, D. Estimation of heterogeneous within-herd variance components using

empirical Bayes Methods: a simulation study. **Journal of Dairy Science**, v. 75, n. 10, p. 2824-2833, 1992.

WINKELMAN, A.; SCHAEFFER, S. R. Effect of heterogeneity of variance on dairy sire evaluation. **Journal of Dairy Science**, v. 71, n. 11, p. 3033-3039, 1988.

*Received on August 3, 2009.*

*Accepted on May 11, 2010.*

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.