Use of Information on Identified Genes to Reduce the Selection Bias on Genetic Evaluation

Ricardo da Fonseca

Viçosa, May 12, 2003
Ph.D. thesis
Federal University of Viçosa

Ricardo da Fonseca

Use of Information on Identified Genes to Reduce the Selection Bias on Genetic Evaluation

Thesis presented to the Genetics and Breeding Program in partial fulfillment of requirements for the degree of *Doctor Scientiae*.
Federal University of Viçosa.

Viçosa
2003

Ricardo da Fonseca

Use of Information on Identified Genes to Reduce the Selection Bias on Genetic Evaluation

Thesis presented to the Genetics and Breeding Program in partial fulfillment of requirements for the degree of *Doctor Scientiae*.
Federal University of Viçosa.

Paulo S. Lopes Comitee member	Robledo A. Torres Comitee member
Roberto A. A. Torres Júnior	Carmen Silva Pereira

Ricardo F. Euclydes
Advisor

Viçosa, May, 12
2003

Dedicatory

To,
My wife Lú,
My parents Mario and Solange,
My brothers Marcelo, Mario and Gustavo and,
My uncle Antônio and my aunt Sônia

Acknowledgments

I would like to thanks all the professors from Animal breeding(Bajá, Carminha, Paulo Sávio and Robledo) for their teachings, patience and friendship. I'm in debt with my friend Zozô for his useful comments and valuable help in this thesis. A special thanks I owe to my advisor for his unconditional confidence on my work and in my person.

I also offer my gratitude to CNPq and CAPES to have provided the funds for my studies in Brazil and USA respectively.

I'm grateful to all my friends from Animal Science Department, for their happiness and friendship which always aliviated me in the most difficult moments. To Adriana, Aldrin, Alexandre(Bodão), Amauri, Claudinho, Daniele, Elizângela, Fabiano, Fernanda, Fred, Guilherme(Zangado), Gustavo, Jane, Leandro, Lindenberg, Marcelo, Marcus Vinícius, Paulinho, Peloso, Policarpo, Raquel, Renata, Rodolphinho and Samuel my most true thanks. Furthermore, I also owe my gratitude to my several friends outside from Animal Science Department: Ana, Elisa, Guilherme, Lima, Maílson, Paulo Bonomo, Renata, Rita and Vevé.

Special thanks should also be sent to people from Iowa State University. Thanks to Aguimar, Antônio, Artur, Jô, Carla, Claudia, Cláudia Mendes, Claudio, Cristiano, Damião, David Henderson, Denise, Grace, Hauke, Luciene, Marcos, Patrícia, Peiq Chen(Pete), Petek, Radu, Raquel, Rogério, Sarah, Vicente and Viviane for help me to overcome all the difficulties that

eventually have appeared. My special thanks to Rohan L. Fernando for his extreme patience and friendship and to have put me in the right way.

Finally I am very grateful to the following people for the constant inspiration and support: My wife Lú for all his confidence and to have been on my side every moment, my parents Mario and Solange for all their support and love, my brothers Marcelo, Mario and Gustavo for their love and happiness and my uncle Antônio and my aunt Sônia for their confidence and support.

Biography

Ricardo da Fonseca, son of Mario Norberto da Fonseca and Solange Tinoco Galdi da Fonseca, was born in Campinas – São Paulo in June, 14 of 1973.

In March of 1992, he started the Animal Science course at Universidade Federal de Viçosa, where he was teaching assistant of the course of beggining statistics for three years.

In February of 1997, he got his degree in Animal Science at Universidade Federal de Viçosa.

In March of 1997, he started the graduate course in Genetics and Breeding at Universidade Federal de Viçosa and in March of 1999, he got his degree in *Magister Scientiae* under advisory of Dr. Ricardo Frederico Euclydes.

In April of 1999, he started the graduate course in Genetics and Breeding to obtain his degree in *Doctor Scientiae*. In October, 2000 he attended the "XI Curso Internacional sobre Mejora Genética Animal", spending one month in Madrid – Spain. In November, 2000 he spent 10 months at Iowa State University – Ames – USA under advisory of Dr. Rohan L. Fernando.

In May 12, 2003 submitted himself to the final exams.

da Fonseca, Ricardo, D.S., Universidade Federal de Viçosa. Maio, 2003. Use of Information on Identified Genes to Reduce the Selection Bias on Genetic Evaluation. Orientador: Ricardo Frederico Euclydes. Conselheiros: Paulo Sávio Lopes, Robledo de Almeida Torres.

Resumo

Os dados disponíveis para avaliação genética são invariavelmente originados de várias gerações de seleção. Entretanto, o BLUP, que é o método escolhido para predizer valores genéticos, assume que a seleção não ocorreu. Se as avaliações genéticas são conduzidas ignorando as mudanças nas médias e variâncias devido à seleção, as predições provavelmente serão viesadas e sem variância mínima e conseqüentemente, alterações no ordenamento dos animais podem ocorrer. A classificação errada dos animais reduz o ganho genético por geração e causa perdas econômicas para empresas de melhoramento e criadores. O problema da seleção na avaliação genética, é essencialmente, um problema de dados perdidos. Se todos os dados utilizados para tomar as decisões de seleção são incluídos na análise as inferências podem ser feitas ignorando a seleção, caso contrário, as predições não são obtidas da distribuição correta. Para verificar o impacto da inclusão da informação de genes para recuperar a informação perdida e diminuir os efeitos da seleção na avaliação genética, foi conduzido um estudo de simulação. Peso à desmama e peso aos 550 dias foram simulados. A primeira característica foi controlada por 20 genes e a segunda por 40 genes, em que os primeiros vinte genes eram compartilhados entre as duas características. As avaliações genéticas foram sempre realizadas para peso aos 550 dias sem o efeito de genes no modelo e com 20, 10, 5 e 2 genes incluídos como efeitos fixos. Cada análise foi conduzida na presença ou ausência de pre-seleção no peso à desmama. Os critérios usados para avaliar a técnica foram: ganho genético e viés. Os resultados mostraram que, de modo geral, a inclusão da informação de

genes identificados como efeitos fixos no modelo de avaliação genética não contribuíram para a redução do viés devido à seleção. Em algumas gerações o viés devido à seleção foi maior quando genes identificados foram usados na avaliação genética. Essa situação ocorreu nas gerações em que 20 e 10 genes foram incluídos como efeitos fixos no modelo, não sendo observada nos casos em que 5 e 2 genes foram utilizados. Para os dois últimos casos, não se observou efeito significativo da inclusão da informação genética para quase todas as gerações.

da Fonseca, Ricardo, D.S., Universidade Federal de Viçosa. May, 2003. Use of Information on Identified Genes to Reduce the Selection Bias on Genetic Evaluation. Advisor: Ricardo Frederico Euclydes. Comitee members: Paulo Sávio Lopes, Robledo de Almeida Torres.

Abstract

Data available for genetic evaluation are invariably originated from several generations of selection. However, BLUP, the usual method for predict genetic values, assumes that selection has not occurred. If genetic evaluations are performed ignoring changes in variances and means due to selection, predictions are likely to be biased and not minimum variance and ranking alterations may occur. Misclassification of animals reduce the genetic gain per generation and causes economic losses to breeding companies and farmers. The selection problem in genetic evaluations is, essentially, a problem of missing data. If all data used to take selection decisions are included in the analysis, selection can be ignored, otherwise, predictions are not obtained from the correct distributions. In order to verify the impact of including single genes information to recover missing data and to reduce selection bias in the genetic evaluation, a simulation study was carried out. Weaning weight and weight at 550 days were simulated. The first trait was controlled by 20 genes and the second by 40 genes where the first twenty were shared between the traits. Genetic evaluations were always performed for weight at 550 days without single genes included in the model and with 20, 10, 5 and 2 single genes included as fixed effects. Each analysis were carried out in presence and absence of pre-selection on weaning weight. The criteria used to evaluate the technique were: genetic gains and bias. The results showed, in general, that including of the information of the identified genes as fixed effects in the genetic evaluation model did not contribute to reduce the bias due to selection. In some generations, the selection bias was larger

when identified genes were used in the genetic evaluation. This situation occured in the setting which 20 and 10 identified genes were included as fixed effects in the model, not being observed in the cases which included 5 and 2 identified genes. Considering the two former cases, it was not observed significantly effect of including of the genetic information for nearly all generations.

Contents

1 Introduction

Data available for genetic evaluation are invariably originated from several generations of selection. If genetic evaluations are performed ignoring changes in variances and means due to selection, predictions are likely to be biased and not to have minimum variance resulting in suboptimal ranking. Misclassification of the animals reduces the genetic gain per generation and causes economic losses to breeding companies and farmers.

Bias on predictions of the additive genetic values, when the analyses do not account for selection, were shown analytically by Henderson et al. (1959), Henderson (1973, 1975, 1990) and by simulations in Pollak and Quaas (1981), Pollak et al. (1984) and Schenkel et al. (2002). Henderson (1975) and Gianola et al. (1988) developed new sets of equations to obtain BLUPs of the genetic additive values of the animals when selection has occurred. However, due to the large size of the set of equations and requirement of the knowledge of the selection history, they are not applicable in practice.

Alternatively, conditions where selection can be ignored were studied. Henderson (1982), assuming multinormality and selection based on available data, stated that inferences could be made ignoring selection if all data used in the selection decisions are included in the analysis and if selection is translation invariant($L'X = 0$). The matrix L represents the selection process history. Gianola and Fernando (1986), Fernando and Gianola (1990), Sorensen et al. (2001), under a bayesian setting, and Im et al. (1989), using missing data theory, showed that selection can be ignored if all data used to make selection decisions are included in the analysis. Multinormality and translation invariant selection are not required.

Thus, if data based on which selection decisions were taken is missing, predictions are likely to be biased and loss their optimum properties. In most developing countries and particularly in Brazil, mainly due to the farmer's level of instruction, selection problems in genetic evaluations are more serious, for the amount of missing data where selection decisions were taken is substantial. Besides, herds that are entering or re-entering in the genetic evaluation programs are also contributing with the missing data problem, since some important information prior to their admission might not have been recorded.

The obvious solution to the problem is try to recover the missing phenotypic data. However, this is usually not possible. Genetic molecular data, such as those from single genes of the traits in analysis, can provide a hope. If these data are included in the analysis, as fixed effects in the model, predictions may be more reliable and bias and other undesirable properties of the predictors, under a selection setting, should, at least, be diminished.

Therefore, the objective of this work is:

To verify if the inclusion of identified genes in the analises reduces the bias on predictions of the genetic gains under a situation where selection is not ignorable.

2 Literature review

This literature review was divided in two subsections, namely, selection models and simulation experiments. The first one has the objective of showing the development of the models to deal with selected data and the differences between them. The second subsection shows practical results and their agreement with the theory showed in the first subsection. Further, because this thesis is based in a simulation experiment, features of the experiments were also highlighted.

An attempt was made to write the review in chronological order, although sometimes, in order to link ideas, it was not possible.

2.1 Selection models

One of the first attempts to deal with selected data was Henderson et al. (1959). It was pointed out that least squares procedure to estimate year effects lead to biased estimates. The authors, in order to show the origin of bias, used a simple example where cows were culled in the first year according to their production, and only the selected animals were allowed to have a second lactation. Henderson showed that the developed mixed model equations eliminated the bias in that situation. A maximum likelihood approach, developed by Kempthorn and von Krosig, was also shown to eliminate the bias in the culling cows scenario. However, conditions required for unbiasedness were not stated for neither of the methods.

Henderson (1973) showed his mixed model equations modified to accommodate three types of selections: selection based on data($L'\boldsymbol{y}$), selection based on genetic values($L'\boldsymbol{u}$) and selection based on environment($L'\boldsymbol{e}$); L is the matrix which describes the selection process(how selection was performed and the proportion of animals selected). Conditions required for unbiasedness were stated. Specifically to the selection based on data, if the term $L'X$, in the coefficient matrix of the modified equations, is null, the solutions are the same as those in the no selection case. According to Henderson, the condition $L'X$ null, implies selection within levels of fixed effects.

The development of the results presented in the above papers was showed

in Henderson (1975). A new vector variable $\boldsymbol{w}$ was defined, which was correlated with the others variables in the model and where selection process was described(under selection based on data, $\boldsymbol{w} = \boldsymbol{L}'\boldsymbol{y}$). Assuming normality and using results due to Pearson (1903), a conditional model on $\boldsymbol{w}$ was written, upon which the modified mixed model equations were obtained. Unbiased predictors were the major concern in the derivation. In the selection case, $\hat{\boldsymbol{u}}$ is not necessarily predictable as in the no selection case and the knowledge of the selection process(to construct $\boldsymbol{L}$) is required. Finally, Henderson wrote that the results obtained could not be valid for more than one cycle of selection, since the normality required in the derivation was not guaranteed.

Alternative solutions were also tried at that time. In dairy cattle, mainly to avoid the selection bias, sire evaluations were based on records from first lactation only, ignoring records from later lactations despite of their economic importance (Cassel and McDaniel, 1983). Recognizing that missing data for cows' first lactation could cause serious bias in sire evaluation when records from several lactations were employed, Keown et al. (1976) suggested corrections factors for female selection to be used before data were incorporated in the usual mixed model equations. Eriksson (1982), as in Keown et al. (1976), suggested some prior adjustment on data to minimize the effect of selection. Working with first and second lactation records, he proposed three adjustment methods for second lactation records. These adjusted data were used in a single trait analysis for purposes of sire evaluation. The alternative methods of evaluation did not eliminate the bias from predictions which increased along with the selection intensity. Because all these methods either did not use all the economic important data available or fail to eliminate bias, efforts concentrated on the development of models to handle selected data(such as the Henderson's model) and in specifying conditions upon which the usual mixed model equations would yield unbiased predictions.

Thompson (1979) pointed out two limitations of the model proposed in Henderson (1975). Based on Kempthorne and von Krosig's work in Henderson et al. (1959) it was asserted that the model conditional on selection uses only part of information available and $\boldsymbol{L}$ should not be considered as fixed, for it was supposed to vary in reapeted sampling. Because $\boldsymbol{L}$ matrix is considered fixed, always the same cow is selected over reapeted sampling,

i.e.,always cow 1 has a higher record than cow 2 for instance. Because of such assumptions, this situation can lead to a loss of statistical information on parameters (Gianola et al., 1988).

Besides the theoretical problems, the use of the mixed model equations under a selection model is difficult in real world situations. Because the number of equations is extremely high, computational difficulties arise and estimability problems are serious regarding the additive genetic effects. Further, writing the L matrix is troublesome and, usually, information required to set it up is not available. Finally to apply the equations for more than one cycle of selection normality should be a reasonable assumption. It turns out that the studies defining situations in which the usual mixed model equations could be used ignoring the selection process became valuable.

That was the main goal in Henderson (1982). It was reinforced that all data related with the selection process should be used in the analysis and proved that assuming multinormality and $L'X = 0$, BLUP ignoring selection is also BLUP under the selection model. Situations where $L'X = 0$ were stated. First, selection should occur within levels of fixed effects, which is not a very realistic situation. Second, data employed to evaluate animals should be corrected for fixed effects using an unbiased estimator thereof. Therefore, all selection functions($L'\boldsymbol{y}$) would have the same mean, namely zero, and $L'X = 0$. Finally, it was stated that a general condition for $L'X = 0$ is that $L'\boldsymbol{y}$ was invariant to the value of the fixed effects.

Goffinet (1983) presented some general results to handle selected data to predict genetic values. A predictor which maximizes the mean breeding values of a fixed number of selected sires($\mathrm{E}[a_1 s_1 + a_2 s_2]$ where, $a_1 + a_2 =$ number of sires and $s_i =$ sire i) was derived by minimizing the average squared risk$\big(\mathrm{E}[s_i - \hat{s}_i]^2\big)$ restricted to translation invariance. For a specific breeding plan, the predictor obtained was the expectation of true breeding values given all the data related to the selection process corrected to the fixed effects$\big(\mathrm{E}[\boldsymbol{s}/\boldsymbol{y_o}, \ldots, \boldsymbol{y_n}]$ where, $\boldsymbol{y_i} =$ data on i^{it} generation$\big)$. Goffinet's predictor, is more general than Henderson's (1975) predictor since the former is not necessarily linear and does not assumes normality, although a known distribution is required. According to the author, in the case of multinormality every unbiased linear estimator of $\boldsymbol{s}$ is a linear function of the data corrected to the fixed effects. In the class of the unbiased estimators, the

conditional expectation minimizes the average squared risk, so $\hat{\boldsymbol{s}}$ is BLUP. These results are in agreement with those obtained in Henderson (1975, 1982) and Pollak and Quaas (1981), i.e., when prior selection is translation invariant(e.g. $L'X = 0$) and all data have been used, the bias was eliminated from predictions.

Fernando and Gianola (1986), assuming that all data related to the selection process was used, proved that selection based on the conditional mean of the merits(Goffinet's (1983) predictor) maximizes the mean of the selected candidates when a fixed number of individuals are selected. However, it is not always true that for a truncation selection(variable number of individuals) that the conditional mean maximize the expected merit of the selected individuals; neither that the selection based on conditional means will always be more efficient upon truncation than upon fixed number of selected individuals. Further, it was shown that ordering candidates by this type of predictor does not, in general, maximize the probability of correctly ordering the true breeding values of the selected animals. Therefore, assuming multivariate normality and translation invariance, BLUP has such properties.

Under the conditions stated by Henderson (1982) for ignorability of selection process, Sorensen and Kennedy (1984) argued that if one is assuming an infinitesimal model as described in Bulmer (1971), normality still holds for several generations of selection. Therefore, the relation $A\sigma_a^2$ applies for more than one cycle of selection. Further, if unbiased predictors are to be obtained, one should use the complete relationship matrix of all animals involved in the selection process. As a consequence, an animal model, which uses the complete relationships between animals, should be preferred over a sire model, for instance. Schenkel (1998), referring to a paper by Woolians and Thompson, wrote that if all genetic relationships to an unselected, unrelated base population are available, then breeding values of all animals can be expressed as the sum of their own mendelian sampling effect plus the mendelian sampling effects of their ancestor going back to the base animals, whose expected breeding values are equal to mendelian sampling effects. Therefore, under an infinitesimal model the change in expected value of $\boldsymbol{u}$ is accommodated through a complete and correct A matrix. It is also stressed that the estimators of the genetic means, although unbiased are not minimum variance. Such a predictor requires the knowledge of the L matrix.

Working under a bayesian perspective, Gianola and Fernando (1986) derived a predictor considering selection had occurred. The results showed that if data prior to selection is included in the analysis, then inferences about breeding values could be made from posterior density as if selection had not occurred, i.e., $f(\boldsymbol{\theta}/\boldsymbol{y_0}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$, where $\boldsymbol{\theta}$ represents the parameters to be inferred. Since this result does not depend on normality it is a generalization of Henderson (1975) results. In the bayesian setting, the requirement of translation invariance or equivalently $L'X = 0$ does not arise, the reason being that in the bayesian approach there are no fixed effects, all of them are regarded as random variables. In Henderson (1975) treatment of selection, linear selection decision were required. However, the derivation of bayesian predictor does not impose the restriction that selection and mating decisions based on data need to be linear, even if normality is assumed. If one argues from a classical view-point, however, where there are fixed effects then, as pointed out by Goffinet (1983), these nonlinear selection decisions would need to be translation invariant (Gianola and Fernando, 1986).

Gianola et al. (1988) used Pearson's selection model to derive predictors maximizing the joint density of data and additive genetic values after selection, without restriction to unbiasedness. According to the authors, such a predictor could be better in the sense of maximizing the genetic progress. In order to solve the resulting set of equations, knowledge of the selection process is required as in Henderson (1975). Regarding selection based on data, the conditions for ignorability of selection process was that $L'X = 0$, i.e, selection must be based on linear or non linear translation invariant functions. If selection was based on group solutions, then selection is ignorable if it was performed within groups. The results are similar to those in Henderson (1975).

Im et al. (1989) demonstrated a general likelihood based approach for inferences using selected data. The data subject to selection is viewed as data with missing values, selection being the process that causes missing data. A vector variable R which represents the data patterns is described and inferences should be made from a joint likelihood function of observed data and R. For example, if two cows are recorded and one should be selected based on their phenotypic values, then two data patterns are possible: (1) (1 1 1 0), $R = 0$ if cow 1 is selected and (2) (1 1 0 1), $R = 1$ if cow two is selected. The ones representing the observed data and zeros the missing

data. This model is different from that applied by Henderson (1975), where due to a fixed L matrix, only one pattern of data after selection is possible, for instance (1). In general, selection can be ignored if all data related with the selection process is used in the likelihood approach. If some external variable is included in the process, selection is ignorable only if data and the external variables are independent. Because a likelihood approach was used, translation invariance, required in Henderson (1975) does not apply.

Henderson (1990), referring to results in Henderson (1975), stated that unbiased estimates and predictions would be obtained by usual mixed model equations if selection is translation invariant($L'X = 0$) and all data is used. Because of the estimability problems and difficulties in writing the L matrix, the modified form should be avoided. When it is necessary to modify the equations to obtain unbiased predictors, it is not certain that the modified solution is a better predictor than the unmodified one in the sense of minimizing mean squared errors of prediction. It was shown also that selection based on variables not related with the variables in the model does not bias the predictions. Examples of $L'\boldsymbol{u}$ and $L'\boldsymbol{e}$ to account for association between sire values and herd merits and to preferential treatment respectively, were provided. Finally, it was showed that the problem of assortative mating(without and with selection) causes no problem if the above conditions to ignore selection holds.

Fernando and Gianola (1990), using a bayesian setting, concluded as in Gianola and Fernando (1986) that when the information used to make breeding decisions is available, joint and marginal posterior densities constructed taking into consideration non-random mating and selection are identical to those constructed ignoring selection. Thus if inferences are based on posterior densities and all the information used to make breeding decisions is used to construct such densities, inferences can be made ignoring the complications due to non-random mating and selection. This will hold for any distribution, any type of selection or non-random mating based on the information available and any method of inference based on a posterior density. It was also shown that if selection decisions were based in some missing information, selection is ignorable only if the missing information and the breeding values are independent, given the data, the posterior density under non-random mating and selection is identical to the posterior density in the absence of selection. However, in general, when some of the infor-

mation used to make selection decisions is not available, calculation of the posterior density requires knowledge of the selection process. Besides that, the authors stated that under normality, BLUP minimizes prediction error variance among all linear unbiased predictors. This is so, provided that breeding decisions are based on translation invariant functions of the data. Further, because under normality BLUP is a conditional mean (Goffinet, 1983; Fernando and Gianola, 1986), selection based on BLUP will maximize the mean of the selected individuals among all translation invariant selection rules, what is an extension of Henderson (1975) results, which relied only on functions of the type $L'\boldsymbol{y}$ with $L'X = 0$. The model employed by the authors is similar to that in Im et al. (1989). Because the likelihood is the main part of the posterior densities, results of the latter paper also apply to a bayesian setting.

Goddard (1990) commented some points about the Henderson (1990) and Fernando and Gianola (1990): (1) for BLUP prediction of breeding values the usual mixed model equations must include the whole selection history of the population(via the A matrix) back to an unselected base population, and estimates of G and R on the base population are required, (2) even if many years of records on a population undergoing selection are available, it may be desirable for practical reasons to include only the more recent data in the analysis. This can be done if the correct probability distribution of the observations at the start of the included data is used, (3) when selection is carried out in two stages but the first stage selection is based on a categorical trait, we do not know how well the mixed model equations would correct for the the selection practiced, (4) The need to modify the mixed model equations when selection is not based on a translation invariant function ($L'X = 0$) appears to contradict the likelihood principle. This principle states that inferences should be based only on the data.

Another attempt to handling selected data was made by Fries and Schenkel (1993). It was shown that the mixed model solutions to the fixed effects contains a function of the additive genetic effect solutions. Bias does not appear in these solutions if $\mathrm{E}(\boldsymbol{a}) = 0$, what probably does not occur with field collected data. Therefore it might not be possible to make comparisons among animals in different levels of fixed effects if the solutions thereof contain large contamination of genetic effects. Further, part of the additive genetic variability could be shifted to the sum of squares associated with fixed ef-

fects what causes a substantial decrease on the variability of the predictions of the additive genetic effects (Schenkel, 1998). A reparametrization of the GLS equations to produce BLUP solutions under selection was proposed by Fries and Schenkel (1993). This set of equations were called Lush's mixed model equations. However, Schenkel (1998) cite that these equations poses computational difficulties because the coefficient matrix is not symmetric and have rank deficiencies equal to GLS. In addition, under an individual animal model, with one record per animal, solutions for the fixed effects can be shown to be zero.

Finally, Sorensen et al. (2001) provided a more general derivation related to conditions for ignorability of selection. They found that for a bayesian inference, only two conditions are sufficient to allow selection to be ignored: (1) the priori distribution of the parameters to be estimated and the parameters associated with the selection variable are independent, (2) conditionally on the observed data, the probability of choosing a specific data pattern does not depend on the parameters to be inferred. This latter condition is satisfied when: (1) selection is random, (2) selection is based on data and all data relevant in selection decision is included in the analysis and (3) selection is based on a variable w not included in the data, then the distribution of w given the observed data is independent of the parameters to be inferred.

One should note that all the procedure employed by Henderson (1975, 1982, 1990) to model selection was based on unbiasedness. In order to obtain BLUP ignoring the selection process, it might be reasonable to assume multinormality, work with translation invariant functions of the data($L'X = 0$) and all the records used to take selections decisions should be used in the analysis. In the model employed by Goffinet (1983), Gianola and Fernando (1986), Im et al. (1989), Fernando and Gianola (1990) and Sorensen et al. (2001) multinormality is not necessary because unbiasedness is not a concern and/or bayesian and likelihood based methods are used, translation invariance is not also an issue. Upon this model the only condition necessary for ignorability to the selection process is that all data related with the selection decisions be included in the analysis. Therefore, because the models employed were different, results and conditions to ignore selection were also different in some cases (Henderson in (Im et al., 1989)).

2.2 Simulation experiments

Aiming to study the predictors of breeding values for beef growth traits (pos-weaning gain and yearling weight), Pollak and Quaas (1981) proposed a simulation study, where two analysis were carried out: (1) multiple trait mixed model evaluation using weaning weight records for all animals and (2) multiple trait mixed model evaluation ignoring weaning weight data for animals which were culled at weaning. The statistical model used comprised the fixed effect of the overall mean and the additive random effect of the animal, what is equivalent to selection was within fixed effects($L'X = 0$). One cycle of selection and three levels of truncation were used. To reduce the influence of accuracy on these comparisons the sires were categorized by the amount of available information. Bias and mean squared error were estimated to study the predictors. Biased predictors were obtained for yearling weight when analysis was conducted including records of selected animals only. Bias did not occur when all records were included in the analysis. A decrease in the mean of the real breeding value of the selected sires was also verified when analysis did not include all records. As cited by the authors, this reflects the effects of bias and loss of accuracy. When culling was performed in a trait with lower correlation with yearling weight, bias in prediction was minimum, even in the most intense selection.

Because all data should be included in the analysis to obtain BLUP, single trait evaluations poses a problem since they usually don't include data upon which a previous selection was based (Pollak et al., 1984). An overview of the benefits of multiple trait analysis to eliminate selection bias was provided in the latter paper. First, a sequential selection situation was simulated as in Pollak and Quaas (1981). Multiple trait evaluations, excluding and including data from animals culled at weaning, were performed for one cycle of selection. Evaluation of bulls using only selected data was biased by selection at weaning. The tendency was to over-predict the worst bulls and under-predict the best bulls. According to the authors, this could result in misranking bulls solely because of differential intensity of selection. Bias was eliminated in the multiple trait evaluation that incorporated all weaning weight records. Second, two traits, where the first was selected directly and the second responded as a correlated trait were simulated. Three non-overllaping generations were used to generate the data set, and evalu-

ations were compared only for the young bulls. Single trait analyses and a multiple trait analysis were performed. As expected, the single trait analysis of the second trait resulted in biased predictions of the breeding values. Multiple trait evaluation, removed the bias from predictions of breeding values. All previous selections were carried out according to translation invariant criteria.

To verify the bias in estimates of the genetic mean a simulation study was conducted by Sorensen and Kennedy (1984). It was assumed that the genetic variance resulted from a large number of unlinked loci. Three non overlapping generations were simulated and the genetic evaluation was then performed and the estimates of the genetic means calculated. The statistical model used comprised one fixed effect(overall mean), the random additive genetic effect and the random error effect. Since under this model selection is invariant to translation and all data was used, the estimates of genetic mean after two cycles of selection were not biased.

Hudson and Schaeffer (1984) carried out a simulation study with dairy cattle. A 2000-cow population with overlapping generations was designed. First and later lactations were used in analysis. Four models were employed to analyze data from the selection program: (1) animal model, (2) sire model, (3) sire-maternal grandsire model and (4) sire model with the predicted transmitting ability of the cow as a covariate. Bias was measured to compare the different methods. The animal model elimated the bias from the predictions while all other models biased the estimates, since they did not include all the relationship information available about the animals. These results corroborates the assertions regarding the relationship matrix in Sorensen and Kennedy (1984).

Fernando and Gianola (1990), suggested that unbiasedness could not be a good optimality criterion to obtain predictors in all situations. In a small scale simulation the modified mixed model equations and the usual mixed model equations were compared in a group + sire selection type($L'X \neq 0$) regarding their predictions. The usual mixed model equations yielded biased estimators and the modified equations unbiased ones. However, in that specific setting, the usual mixed model equations were better in the sense of mean squared error and the mean of the highest ranking sire.

Schenkel et al. (2002), studied the performance of bayesian estimators

and bayesian predictors against the frequentist ones when data were subject to selection. Data were generated for six non-overlapping generations. A univariate animal model was assumed and selection was performed only on males. Two models regarding the fixed effects were used: (1) the mean was the only fixed effect and (2) Besides the mean a contemporary group effect was simulated. Methods were compared with respect to the amount of bias, mean square error and Spearman's rank correlation regarding to the true and predicted breeding values. Two scenarios were also generated according to the pedigree information: (1) full pedigree information and (2) randomly missing pedigree information. No differences in Spearman's correlation were found between methods but the correlations decreased when selection was performed in the presence of missing pedigree data. Small differences were found between methods regarding the amount of bias. The scenario using the model with contemporary groups and missing pedigree presented the largest amount of bias. Further, bias increased over the generations. Similar results were found when mean square error was considered. Selection associated with missing pedigree greatly increased mean square error over generations. The results of Schenkel et al. (2002) agree with results presented in Sorensen and Kennedy (1984).

3 Material and Methods

3.1 Simulation

3.1.1 Populations' genetic structure

In order to study the effect of selection on genetic evaluation a sequential selection situation was simulated. Weaning weight(WW) and weight at 550 days(W550) were chosen as illustrative traits. An additive genetic model was used to simulate two correlated traits according to the genetic parameters in Table 1.

Table 1: Values of genetic parameters used to simulate the populations

Weaning weight							**Weight at 550 days**		
$h^{2\,a}$	$\sigma_a^{2\,b}$	$\sigma_e^{2\,c}$					h^2	σ_a^2	σ_e^2
0.28	114.48	294.37					0.41	453.03	651.93
			$r_a{}^d$	$\sigma_a{}^e$	$r_e{}^f$	$\sigma_e{}^g$			
			0.82	187.88	0.40	175.23			

[a]Heritability
[b]Genetic additive variance
[c]Environment variance
[d]Genetic additive correlation
[e]Genetic additive covariance
[f]Environmental correlation
[g]Environmental covariance

It was assumed that weaning weight was controlled by 20 loci all having the same effect and the weight at 550 days was controlled by 40 loci, which were divided in two groups of 20 alleles. Within each group every locus had the same effect for each of three possible genotypes. The first 20 loci were common to the two traits, although their effects were different for both, weaning weight and weight at 550 days. All loci had two alleles with initial frequencies of 0.5. The loci were mutually independent. Table 2 show the values for the simulated genotypes.

Table 2: Simulated effects for each genotype within groups of loci for each trait

	WW			**W550**		
Loci	*AA*	*Aa*	*aa*	*AA*	*Aa*	*aa*
1–20	0.0000	3.3835	6.7669	0.0000	5.5526	11.1059
21–40	–	–	–	0.0000	3.8042	7.6084

3.1.2 Data structure

Two types of data structure were generated: data from a scheme where no pre-selection have been performed(WP) and data from a scheme where pre-selection have been performed(PS). For both schemes the base population comprised 600 cows and 30 bulls that were unrelated and unselected. The animals were mated at random at the proportion of 20 females per male to produce 300 females and 300 males.

For (WP), all males and females offspring were kept in the herd until the measure of weight at 550 days could be performed. For (PS), 60 females and 210 males were culled based on their weaning weight performance. Thus, only 240 females and 90 males were allowed to have a measurement for weight at 550 days.

From adults cows and young cows with data for weight at 550 days, 600 females were selected based on the BLUP of their breeding values and kept in the herd. From adults bulls and young bulls with data for weight at 550 days, 30 males were selected based on the BLUP of their breeding values and kept in the herd. This completed one cycle of selection, which correspond to approximately two years in Nelore cattle herd. In order to generate the offspring selected bulls and cows were mated at random and all the process described above was repeated. The two schemes to generate data are shown in Figure1 and in Figure2.

3.2 Genetic evaluation

BLUP was employed to predict genetic values for selection purposes. Mixed Model Methodology(Henderson, 1973) was used to get the BLUP for the simulated animals according to the following general model:

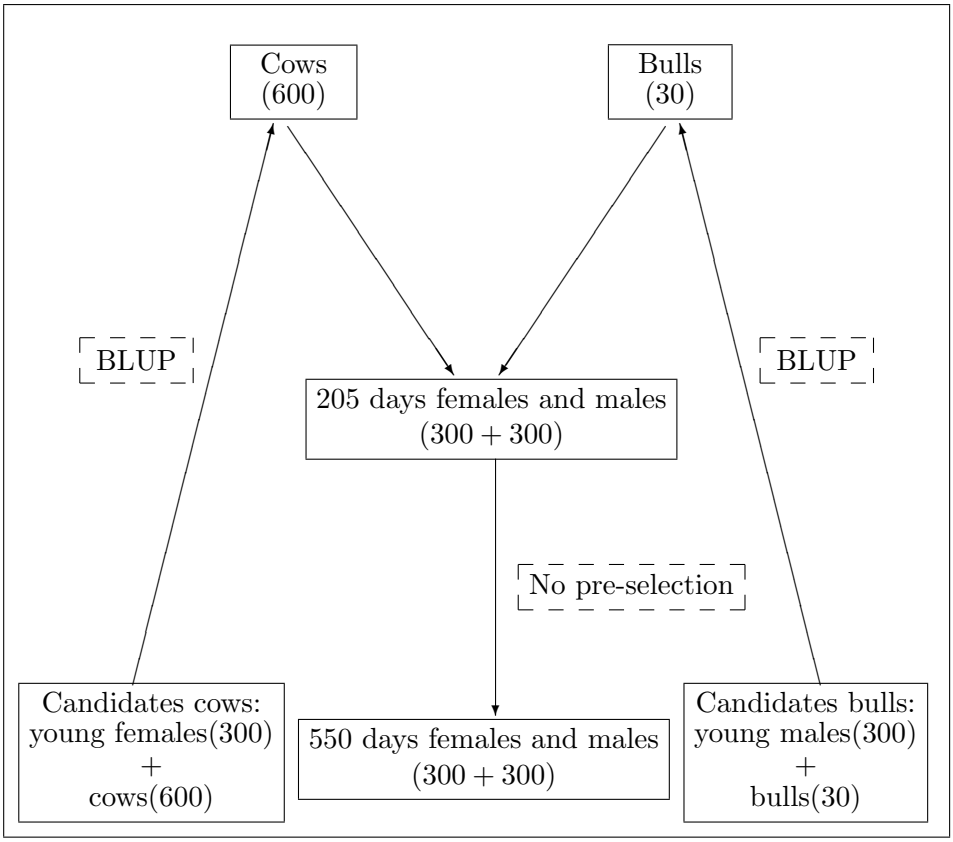

Figure 1: Scheme and number of animals used to generate the data structure without pre-selection

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{a} + \boldsymbol{e} \tag{1}$$

where,

$\boldsymbol{y}$ = vector of observations,

X = incidence matrix of fixed effects,

Z = incidence matrix of genetic additive random effects,

$\boldsymbol{\beta}$ = vector of fixed effects,

$\boldsymbol{a}$ = vector of the genetic additive effect of the animals,

$\boldsymbol{e}$ = vector of the environmental effect.

The mixed model equations are represented as follows:

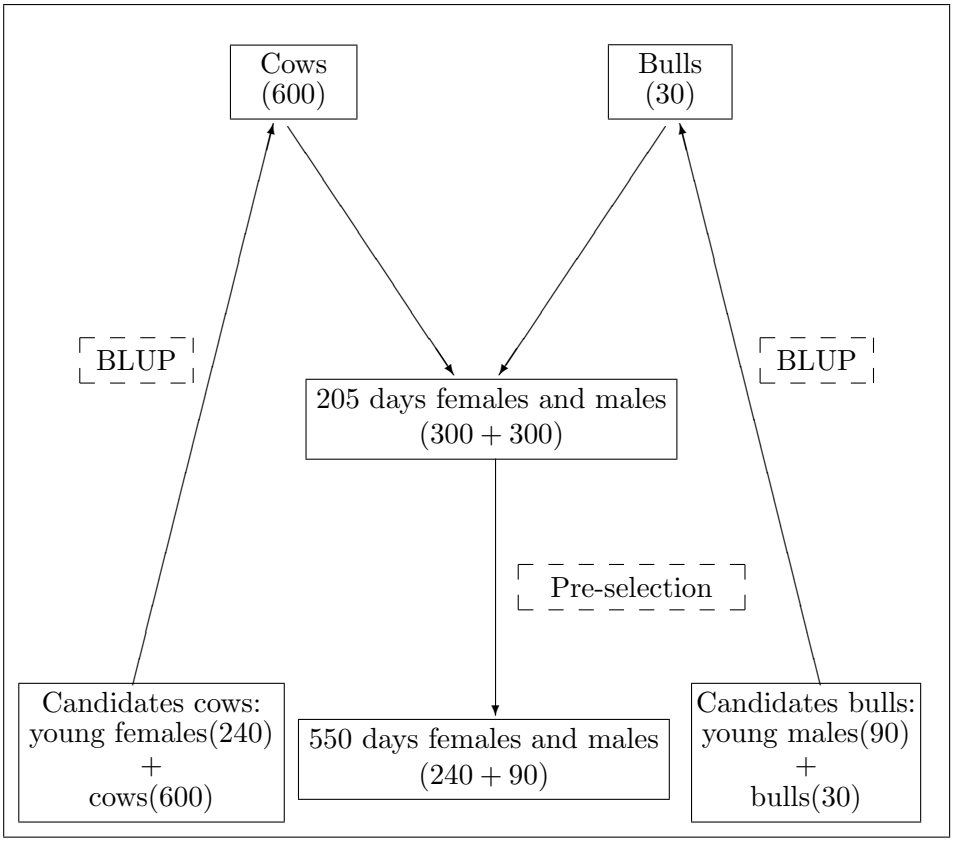

Figure 2: Scheme and number of animals used to generate the data structure with pre-selection

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{\circ} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} X'R^{-1}\boldsymbol{y} \\ Z'R^{-1}\boldsymbol{y} \end{bmatrix} \tag{2}$$

where,

$\boldsymbol{y}$, X and Z = are defined as in (1),

R = variance-covariance matrix of residuals effects,

G = variance-covariance matrix of the additive genetic effects,

$\boldsymbol{\beta}^{\circ}$ = vector of the estimates of the fixed effects,

$\hat{\boldsymbol{a}}$ = vector of the predictions of the additive genetic effects of the animals,

The variances and covariances needed to perform the analysis were assumed known.

Specific models were used depending if the identified genes were or not included in the analisys. Those models will be described in the next two topics.

3.2.1 Evaluations by models that do not include gene effects

For objectivity and simplicity reasons, the same model was considered for both traits, which contained one fixed effect term, the polygenic effect term and the random error term. A more detailed description follows:

$$y_{it} = \mu + a_i + e_{it} \tag{3}$$

where,

$y_{it} = i^{\text{th}}$ observation on trait t for the i^{th} animal,

μ = fixed overall mean,

$a_i = i^{\text{th}}$ random additive genetic effect of the $i^{\text{th}}th$ animal,

e_{it} = random error effect associated with the observation on trait t for the i^{th} animal.

A single-trait analysis for weight at 550 days was carried out for each data structure(WS and PS) according to (3) and (2). Animals were selected by their BLUP for the polygenic additive genetic effect($\hat{a}_i$).

3.2.2 Evaluations by models that include gene effects

The model used was the same as in (3) except for including a fixed term representing the additive genetic effect of a previously identified gene. It is the same model presented by Kennedy et al. (1992) modified to accommodate several single genes effects:

$$y_i = \mu + \boldsymbol{k_i'g} + a_i + e_i \tag{4}$$

where,

y_i = observation on animal i,

μ = fixed overall mean,

$\boldsymbol{k_i}$ = vector of 0's and 1's referring to animal i,

$\boldsymbol{g}$ = vector of fixed additive genetic effects of the identified genes,

a_i = remaining random additive effect of the i^{th} animal, after exclusion of the additive effects for the identified genes,

e_i = random error effect associate with the observation for the i^{th} animal..

Single trait analyses for weight at 550 days were performed based on (4). It was assumed that all individuals with observations were genotyped for all the single genes and there were no missing genotypes. Since the additive genetic effect and the genotypes for each identified locus were known the total genetic variance was calculated by:

$$\sigma_g^2 = \sum_{j=1}^{n_g} (2pqa_i^2) \tag{5}$$

where,

σ_g^2 = total variance due to the identified genes,

n_g = number of identified genes,

p and q = allelic frequencies for the identified genes,

a_i = mean additive genetic value of the locus i.

The values calculated by (5) were subtracted from the true additive genetic variances of the polygenic effect. The corrected values of variance were then used in (2).

The animals were selected based on their BLUP of the polygenic effect plus the sum of the estimates for the fixed effects. The general mean was included for the sum of the effects of the genotypes of the identified genes was not estimable. Therefore, the selection criterion was:

$$\text{sc} = \mu + \boldsymbol{k_i'g}^{\circ} + \hat{a}_i \tag{6}$$

where,

sc = selection criterion,

μ = general mean,

$\boldsymbol{k}_i'\boldsymbol{g}^\circ$ = sum of the estimated effects for the identified genes[1],

$\hat{a}_i$ = BLUP of the remaining polygenic additive effect

Four analysis were performed based on the number of identified genes shared between the two traits:

1. 20 genes were considered known(approximately 43.83% of the total genetic variance).
2. 10 genes were considered known(approximately 21.91% of the total genetic variance).
3. 5 genes were considered known(approximately 10.96% of the total genetic variance).
4. 2 genes were considered known(approximately 4.38% of the total genetic variance).

* * *

Each type of analysis received a code to be referred later, which can be seen in Table3.

A general view of the type and amount of information available for each setting is shown in Table 4 and Table 5, following Im's et al. (1989) representation with some modifications. In these tables the code *nu* (not used) was used to represent data that was not required by the particular simulation scheme, as was the case in the univariate analysis and in those that did not included genotypic information. The code *na* (not available) was used to represent data that could not be collected due to culling of animals.

For each of the combinations shown in the Table 3, 10 cycles of selection were simulated(approximately 20 years) and replicated 100 times.

[1]The superscript $\circ$ means that only some functions of the elements of the vector $\boldsymbol{g}$ are estimable

Table 3: Codes for the combinations of data structure, number of traits and model used.

Data structure	Model[a]	Number of genes	Code
WS	regular	—	R.WS
PS	regular	—	R.PS
WS	fixed genes	20	G20.WS
WS	fixed genes	10	G10.WS
WS	fixed genes	5	G05.WS
WS	fixed genes	2	G02.WS
PS	fixed genes	20	G20.PS
PS	fixed genes	10	G10.PS
PS	fixed genes	5	G05.PS
PS	fixed genes	2	G02.PS

[a]The model without fixed gene effects represented by (3) was called regular and the one with fixed genes effects represented by (4) was called fixed genes

3.3 Criteria used on comparison of the methods

1. *Genetic gain in one generation*

$$\hat{\bar{G}}_t = \hat{\bar{g}}_t - \hat{\bar{g}}_{t-1} = \frac{1}{r}\sum_{i=1}^{r}\hat{g}_t - \frac{1}{r}\sum_{i=1}^{r}\hat{g}_{t-1} = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n_t}\frac{\hat{a}_{i_t}}{n_t} - \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n_{t-1}}\frac{\hat{a}_{i_{t-1}}}{n_{t-1}} \tag{7}$$

where,

$\hat{\bar{G}}_t$ = estimated genetic gain on generation t,

$\hat{\bar{g}}_t$ = genetic mean of generation t over all replicates,

r = number of replicates,

$\bar{g}_t$ = genetic mean of generation t for a particular replicate,

$\hat{a}_{i_t}$ = estimate of the additive genetic value of the animal i on generation t,

n_t = number of animals within generation t.

Table 4: Type and amount of information available in the traits for the simulations settings without pre-selection

	Code[a]			**Code**		
	R.WS			G20.WS, G10.WS, G05.WS, G02.WS		
	P[b]		**G**[c]	**P**		**G**
ID[d]	**WW**	**W550**		**WW**	**W550**	
1	nu	*	nu	nu	*	+
2	nu	*	nu	nu	*	+
3	nu	*	nu	nu	*	+
⋮						
k	nu	*	nu	nu	*	+
$k+1$	nu	*	nu	nu	*	+
$k+2$	nu	*	nu	nu	*	+
⋮						
m	nu	*	nu	nu	*	+

[a]Codes from Table 3
[b]Phenotypic information: present(*)
[c]Genotypic information: present(+), not used(nu)
[d]Animal's identification number

2. *Average genetic gain*

$$\hat{\bar{G}} = \frac{1}{f}\sum_{t=1}^{f}\hat{\bar{G}}_t \tag{8}$$

where,

$\hat{\bar{G}}$ = estimated total genetic gain,

f = total number of generations,

$\hat{\bar{G}}_t$ = defined as in (7).

3. *Bias on the estimates of the genetic gain in one generation*

$$\hat{\bar{B}}_{G_t} = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{G}_t - G_t\right) = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n_t}\frac{\hat{a}_{i_t}}{n_t} - \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n_t}\frac{a_{i_t}}{n_t} \tag{9}$$

where,

$\hat{\bar{B}}_{G_t}$ = bias on estimates of the genetic gain in generation t,

Table 5: Type and amount of information available in the traits for the simulations settings with pre-selection

	Code[a]			**Code**		
	R.PS			G20.PS, G10.PS, G05.PS, G02.PS		
	P[b]		**G**[c]	**P**		**G**
ID[d]	**WW**	**W550**		**WW**	**W550**	
1	–	*	nu	–	*	+
2	–	*	nu	–	*	+
3	–	*	nu	–	*	+
⋮						
k[e]	–	*	nu	–	*	+
$k+1$	–	–	nu	–	–	na
$k+2$	–	–	nu	–	–	na
⋮						
m[f]	–	–	nu	–	–	na

[a]Codes from Table 3
[b]Phenotypic information: present(*), missing(–)
[c]Genotypic information: present(+), not available(na), not used(nu)
[d]Animal's identification number
[e]last pre-selected animal
[f]last culled animal

G_t = true genetic gain on generation t

r, n_t and $\hat{a}_{i_t}$ = defined as in (7),

a_{i_t} = true genetic additive value of the animal i on generation t.

4. *Bias on the estimates of the average genetic gain*

$$\hat{\bar{B}}_{G_T} = \frac{1}{r}\sum_{i=1}^{r}\hat{G}_T - \frac{1}{r}\sum_{i=1}^{r}G_T \tag{10}$$

where,

$\hat{\bar{B}}_G$ = bias on estimate of the total genetic gain,

t and $\hat{\bar{G}}$ = defined as in (7) and (8),

G = actual average genetic gain.

5. *Difference of responses*

   To verify the magnitude of the bias of the predictions from the two models in equation (3) and equation (4), a difference based on the following rationale was calculated:

   Let the response in the genetic gain from the R.WS setting be:

   A = response from single stage selection;

   Let the response in the genetic gain from the R.PS setting be:

   B = A + P – b = response from single stage selection + extra response due to pre-selection – selection bias;

   Let the response in the genetic gain from the Gx.WS (x = 02, 05, 10 and 20) setting be:

   C = A + G = response from single stage selection + effect of the identified genes;

   Let the response in the genetic gain from the Gx.PS (x = 02, 05, 10 and 20) setting be:

   D = A + G + P – c = response from single stage selection + effect of the identified genes + extra response due to pre-selection – selection bias.

   One wants to know if c is smaller than b. To achieve that the following calculation is performed:

$$\text{Diff} = (\text{B} - \text{A}) - (\text{D} - \text{C}) = -\text{b} + \text{c}. \tag{11}$$

   Thus, if the difference is null, there is no effect in reducing bias due to include identified genes in the analises. If the difference is positive the resuting bias is larger when the identified genes are included as fixed effects in the model. A negative result indicates that identified genes can aid to reduce selection bias in the genetic evaluation.

4 Results and Discussion

4.1 Evaluations by models that do not include gene effects

The results for the R.WS setting are presented in Table 6. It was simulated a population without pre-selection and a single trait analysis was performed to W550. In this situation selection poses no problem to estimates and predictions (Gianola and Fernando, 1986; Im et al., 1989; Fernando and Gianola, 1990; Sorensen et al., 2001) .

Table 6: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting without pre-selection and evaluated by a model not including identified genes as fixed effects (R.WS).

Generation	**TGG**	**PGG**	**B**
1	4.6240	4.5809	$-0.0431^{\text{n.s.}}$ [a]
2	10.8943	10.9942	$0.0999^{\text{n.s.}}$
3	6.2824	6.2288	$-0.0536^{\text{n.s.}}$
4	7.4058	7.5042	$0.0984^{\text{n.s.}}$
5	6.2648	6.3127	$0.0480^{\text{n.s.}}$
6	6.6247	6.6588	$0.0341^{\text{n.s.}}$
7	6.3367	6.2734	$-0.0633^{\text{n.s.}}$
8	6.1965	6.2483	$0.0518^{\text{n.s.}}$
9	6.2318	6.2720	$0.0402^{\text{n.s.}}$
10	5.8500	5.9197	$0.0697^{\text{n.s.}}$
Total(3) [b]	21.8007	21.8039	0.0032
Average(3) [c]	7.2669	7.2680	$0.0106^{\text{n.s.}}$
Total(10) [d]	66.7109	66.9936	0.2821
Average(10) [e]	6.6711	6.6993	$0.0282^{\text{n.s.}}$

[a] Not significantly different from 0 by the t test.
[b] Totals obtained from the first three generations
[c] Average value obtained from the first three generations
[d] Totals obtained from ten generations
[e] Average value obtained from ten generations

The true and predicted genetic gains were maximum at generation 2, thereafter the gains were smaller and nearly constant until generation 10. This was observed for all simulated populations. The genetic means were

calculated for each generation including cows, bulls, and 550 days males and females. In generation 1, the true and predicted genetic gains were calculated as the difference between the genetic mean of the base generation, which is zero, and the genetic mean of the generation 1. The genetic mean of generation 1 included selected cows and bulls and unselected 550 days males and females, who were sons of animals from base population. The gains in generation 2 similarly, were calculated as the difference between the genetic means in generation 2 and generation 1. The genetic mean of generation 2 included selected cows and bulls and selected 550 days males and females. Because all the animals were selected in this generation, the difference between generation 1 and 2 was large, generating a large genetic gain. Thereafter, all the differences were between genetic means that included selected animals, generating gains smaller than that in the generation 2. Figure 1 helps to understand the dynamic of the population for each cycle of selection.

After generation 3 the genetic gains are nearly constant oscillating in small amount between generations. This is probably due to the increased amount of information accumulated by the relationship matrix in each generation. As the information is added more reliable estimates are obtained.

The table also show the short term response and long term response after three and ten cycles of selection respectively. No bias was detected for R.WS setting over 3 and 10 generations.

The R.PS setting in Table 7 simulates a scenario where pre-selection have been practiced in WW and a single trait analysis was carried out for W550. Selection has undesirable effects for this population (Gianola and Fernando, 1986; Im et al., 1989; Fernando and Gianola, 1990; Sorensen et al., 2001).

The genetic gains for this setting followed the same pattern of those in the former setting, which was explained before. The dynamic of this population can be found in Figure 2.

Considering the first three generations, bias was not detected in the second one, although, in average, a significant value ($p < 0.05$) was found (Average(3) field). The values on the first and third generations were significantly different from zero ($p < 0.01$). Bias was, in average, 50.00% larger than the respective value in R.WS. The results in Tables 6 and 7 agree with those found in Pollak and Quaas (1981) and Pollak et al. (1984).

Table 7: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) for a setting with pre-selection and evaluated by a model not including identified genes as fixed effects (R.PS).

Generation	**TGG**	**PGG**	**B**
1	5.4877	5.8138	0.3261^{**} [a]
2	8.1845	8.2808	$0.0963^{\text{n.s.}}$ [b]
3	6.2303	6.4471	0.2168^{**}
4	6.7175	6.9558	0.2383^{**}
5	6.2254	6.4466	0.2211^{**}
6	6.3239	6.5037	0.1798^{**}
7	6.0364	6.1913	0.1549^{**}
8	5.9153	6.0471	0.1318^{**}
9	5.8299	5.8796	$0.0496^{\text{n.s.}}$
10	5.6222	5.7799	0.1577^{**}
Total(3) [c]	19.9025	20.5417	0.6392
Average(3) [d]	6.6342	6.8472	0.2131^{*} [e]
Total(10) [f]	62.5732	64.3456	1.7724
Average(10) [g]	6.2573	6.3456	0.1772^{**}

[a] Significantly different from 0 by the t test considering a significance level of 1%
[b] Not significantly different from 0 by the t test.
[c] Totals obtained from the first three generations
[d] Average value obtained from the first three generations
[e] Significantly different from 0 by the t test considering a significance level of 5%
[f] Totals obtained from the first three generations
[g] Average value obtained from the first three generations

The bias was not significantly different from zero in the ninth generation. Bias accumulated after ten cycles of selection was 84.00% higher than the one observed in Table 6 (Total(10) field).

Selection does not cause problem in the R.WS setting since there is no pre-selection and all the information used to select animals for W550 are included in the analysis. The opposite situation occurs in the R.PS setting. The animals are pre-selected for WW but this information is not available for selections decisions. In this case, the distribution used to make inferences are not the correct one (Im et al., 1989). Thus, results obtained from a wrong distribution are likely to be biased and non-accurate. This explains

the results found in Table 7.

4.2 Evaluations by models that include gene effects

Results from Table 8 to 11 are associated with single-trait analysis including increasing number of identified genes as fixed effects in the model as described in Table 3. Selection does not have harmful effects in these populations (Gianola and Fernando, 1986; Im et al., 1989; Fernando and Gianola, 1990; Sorensen et al., 2001).

All 20 shared genes(or 43.83% of the additive genetic variance of the W550) between the two traits were controlled in the G20.WS setting(Table 8).

Regarding the first three generations one can notice that the actual and predicted total genetic gains were larger than those observed in Table 6, providing 14.27% and 14.94%more improvement than the situation where no genotypes were included. An unexpected bias, with a high significant value ($p < 0.01$), appeared in the first generation. In the second and third ones, no bias was detected.

The total predicted and true genetic gains in long term selection were approximately 14.10% and 13.97% larger than those on Table 6. No bias was noticed over ten generations, except the first one.

The effect of controlling 10 shared genes(or 21.91% of the genetic additive variance of the W550) in a situation without pre-selection is shown in Table 9.

The response after three cycles of selection were 8.38% and 8.22% less than those for the G20.WS setting for predicted and actual genetic gains respectively (Average(3) fields). The gains were larger than those in Table 6. A bias not significantly different from the bias found in the former setting was also observed in the first generation. Bias was also found on the second generation and, similarly to the G20.WS setting, it was not expected to appear.

Regarding ten generations, the actual genetic gain was 6.66% larger than that in Table 6. The difference in the predicted gain was 6.72%. Comparison with the G20.WS setting shows that the genetic gains after ten generations

Table 8: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) 10 generations for a setting without pre-selection and evaluated by a model including 20 identified genes as fixed effects (G20.WS).

Generation	**TGG**	**PGG**	**B**
1	5.6404	5.8132	0.1728^{**} [a]
2	12.7899	12.8314	$0.0415^{\text{n.s.}}$ [b]
3	6.9985	6.9880	$-0.0106^{\text{n.s.}}$
4	8.7884	8.6997	$-0.0887^{\text{n.s.}}$
5	7.3822	7.3975	$0.0153^{\text{n.s.}}$
6	7.8479	7.8368	$-0.0111^{\text{n.s.}}$
7	7.2500	7.2706	$0.0205^{\text{n.s.}}$
8	7.2558	7.2330	$-0.0227^{\text{n.s.}}$
9	6.9090	6.9934	$0.0844^{\text{n.s.}}$
10	6.8011	6.8177	$0.0167^{\text{n.s.}}$
Total(3) [c]	25.4288	25.6326	0.2249
Average(3) [d]	8.4763	8.5442	$0.0750^{\text{n.s.}}$
Total(10) [e]	77.6632	77.8812	0.2180
Average(10) [f]	7.7663	7.7881	$0.0218^{\text{n.s.}}$

[a] Significantly different from 0 by the t test considering a significance level of 1%
[b] Not significantly different from 0 by the t test.
[c] Totals obtained from the first three generations
[d] Average value obtained from the first three generations
[e] Totals obtained from ten generations
[f] Average value obtained from ten generations

were 7.97% and 7.77% respectively for the actual and predicted genetic gains. In average, no bias was detected for this setting over three and ten generations, similar to the results in the G20.WS setting. Although the average value for the first generations should be interpret with caution, since bias were detected for one or two generations.

The results in Table 10 are associated with 10.96% of control of the additive genetic variance of the W550 or equivalently 5 identified shared genes.

Calculations up to three generations showed that the predicted and the true genetic gain were around 4.92% and 3.68% smaller than that observed

Table 9: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting without pre-selection and evaluated by a model including 10 identified genes as fixed effects (G10.WS).

Generation	**TGG**	**PGG**	**B**
1	4.9464	5.0746	0.1283^{*} [a]
2	11.7777	11.9367	0.1590^{**} [b]
3	6.5734	6.5138	$-0.0596^{\text{n.s.}}$ [c]
4	7.9002	7.9078	$0.0076^{\text{n.s.}}$
5	6.8995	6.8846	$-0.0148^{\text{n.s.}}$
6	7.1563	7.1930	$0.0367^{\text{n.s.}}$
7	6.7236	6.7608	$0.0372^{\text{n.s.}}$
8	6.7423	6.7384	$-0.0039^{\text{n.s.}}$
9	6.5045	6.5439	$0.0394^{\text{n.s.}}$
10	6.2496	6.2730	$0.0234^{\text{n.s.}}$
Total(3) [d]	23.2975	23.5251	0.2277
Average(3) [e]	7.7658	7.8417	$0.0759^{\text{n.s.}}$
Total(10) [f]	71.4735	71.8266	0.3531
Average(10) [g]	7.1474	7.1827	$0.0353^{\text{n.s.}}$

[a] Significantly different from 0 by the t test considering a significance level of 5%
[b] Significantly different from 0 by the t test considering a significance level of 1%
[c] Not significantly different from 0 by the t test.
[d] Totals obtained from the first three generations
[e] Average value obtained from the first three generations
[f] Totals obtained from ten generations
[g] Average value obtained from ten generations

in G10.WS setting (Table 9) and were 12.74% and 11.76% smaller than those in G20.WS setting (Table 8). Compared to the R.WS situation the gains were 2.85% and 2.52% larger for the true and predicted respectively. No bias was observed after the first three cycles of selection, contradicting the results for the G20.WS and G10.WS.

The same trend was observed when considering the ten cycles of selection. The predict and actual genetic gains were smaller than those observed in Table 8 and Table 9. The gains over ten generations were larger than those presented in Table 6 for R.WS setting. No biases were noticed over

Table 10: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting without pre-selection and evaluated by a model including 5 identified genes as fixed effects (G05.WS).

Generation	**TGG**	**PGG**	**B**
1	4.8115	4.8365	$0.0250^{\text{n.s.}}$ [a]
2	11.2658	11.2489	$-0.0169^{\text{n.s.}}$
3	6.3620	6.2819	$-0.0802^{\text{n.s.}}$
4	7.8198	7.9097	$0.0900^{\text{n.s.}}$
5	6.5025	6.5061	$0.0036^{\text{n.s.}}$
6	6.8078	6.7818	$-0.0260^{\text{n.s.}}$
7	6.3684	6.4183	$0.0498^{\text{n.s.}}$
8	6.4050	6.4136	$0.0086^{\text{n.s.}}$
9	6.1026	6.1495	$0.0469^{\text{n.s.}}$
10	6.0663	6.0992	$0.0329^{\text{n.s.}}$
Total(3) [b]	22.4393	22.3673	-0.0721
Average(3) [c]	7.4798	7.4558	$-0.0240^{\text{n.s.}}$
Total(10) [d]	68.5118	68.6454	0.1336
Average(10) [e]	6.8512	6.8645	$0.0134^{\text{n.s.}}$

[a] Not significantly different from 0 by the t test.
[b] Totals obtained from the first three generations
[c] Average value obtained from the first three generations
[d] Totals obtained from ten generations
[e] Average value obtained from ten generations

the ten generations.

The results in Table 11 are those associated with 4.38% of control of the additive genetic variance of the W550 or equivalently 2 shared genes.

A short term response in the genetic gains shows that the actual and predicted genetic gains were smaller than those in the former setting (Table 10). The gains were also larger than those obtained in R.WS (Table 6). No bias was detected in the first three generations, agreeing with the results of the former setting.

The actual genetic gains were approximately 11.78%, 5.07% and 0.97% less than G20.WS, G10.WS and G05.WS respectively but it was around

Table 11: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting without pre-selection and evaluated by a model including 2 identified genes as fixed effects (G02.WS).

Generation	**TGG**	**PGG**	**B**
1	4.5774	4.5682	$-0.0092^{\text{n.s.}}$ [a]
2	11.0761	11.0432	$-0.0329^{\text{n.s.}}$
3	6.1014	6.0875	$-0.0139^{\text{n.s.}}$
4	7.7239	7.7388	$0.0149^{\text{n.s.}}$
5	6.3954	6.4471	$0.0518^{\text{n.s.}}$
6	6.9113	6.8501	$-0.0613^{\text{n.s.}}$
7	6.2624	6.2935	$0.0311^{\text{n.s.}}$
8	6.4555	6.4896	$0.0341^{\text{n.s.}}$
9	6.1755	6.1698	$-0.0057^{\text{n.s.}}$
10	6.1692	6.2871	0.1178^{*} [b]
Total(3) [c]	21.7549	21.6989	−0.0560
Average(3) [d]	7.2516	7.2330	$-0.0187^{\text{n.s.}}$
Total(10) [e]	67.8481	67.9748	0.1268
Average(10) [f]	6.7848	6.7975	$0.0127^{\text{n.s.}}$

[a] Not significantly different from 0 by the t test.
[b] Significantly different from 0 by the t test considering a significance level of 5%
[c] Totals obtained from the first three generations
[d] Average value obtained from the first three generations
[e] Totals obtained from ten generations
[f] Average value obtained from ten generations

1.68% larger than R.WS setting (Table 6) for ten generations. The predicted genetic gain presented the same behavior.

The differences in the genetic gains in the settings shown before compared to the R.WS setting are mainly due to the extra information of the identified genes included in the statistical model. One can notice that basically the same response in the genetic gains were achieved when five and two identified genes were included as fixed effects in the model.

By the same reasons explained before for R.WS, selection does not have undesirable effects in the G20.WS, G10.WS, G05.WS and G02.WS settings. However, an unexpected bias occurred in the first generation for G20.WS

and in the first and second generations for the G10.WS setting. This was probably due to identified genes effects included as fixed effects in the model, since there is a trend of increasing the value of the bias as more fixed effects of identified genes are included in the model (Tables 8 to 11). For instance, in G20.WS, 20 fixed effects each with three levels (genotypes) were included in the model, while 10 fixed effects with three levels each were included in the model of the G10.WS setting. The biases are probably due to a small amount of animals in early generations, which caused a poor estimation of the fixed gene effects and since these fixed effects are used to calculate the genetic gains, thoseare biased by the small amount of information on early generations. Thus, the larger the number of genotypes included in the model, the larger should be the population to obtain reliable estimates of the genetic fixed effects.

Despite of the bias in the first generation presented on Tables 8 and 9 the results agree with the theory presented in Im et al. (1989); Fernando and Gianola (1990) and Sorensen et al. (2001).

The results presented in Table 12 to Table 15 are associated with situations where pre-selection was present and identified gene effects were accounted for in the statistical model. Selection has undesirable effects on the prediction in these situations. The results show the effect of including the identified genes in the model when pre-selection is performed.

The predicted genetic gain for G20.PS was the largest among all settings with pre-selection practice, including R.PS. Considering, three generations, It was possible to obtain an actual gain nearly 12.77% larger than the one obtained in the R.PS setting (Table 7). Biases were detected in the first three generations and they were significantly larger than those in Table 7 ($p < 0.01$ for the first two generations and $p < 0.05$ for the third generation).

After ten generations the TGG and the PGG were 11.62% and 12.77% larger than the respective gains R.PS. The values of the bias not significantly different from zero were found in generation 7, 8 and 10. The mean bias was not significantly different from that of the R.PS setting.

Table 13 shows that the TGG and PGG were nearly 7.27% and 6.91% smaller than G20.PS after the first three generations. The gains in G10.PS were larger than R.PS approximately by 7.37% and 12.37%. Biases were

Table 12: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting with pre-selection and evaluated by a model including 20 identified genes as fixed effects (G20.PS).

Generation	**TGG**	**PGG**	**B**
1	6.6656	7.6956	1.0300**[a]
2	9.4555	10.0225	0.5669**
3	7.0508	7.4652	0.4143**
4	7.5832	7.8924	0.3093**
5	7.0268	7.2629	0.2361**
6	7.0531	7.1918	0.1387**
7	6.8130	6.8375	0.0245[n.s.]
8	6.5611	6.6217	0.0607[n.s.]
9	6.3917	6.4932	0.1015*[b]
10	6.2007	6.2849	0.0842[n.s.]
Total(3)[c]	23.1719	25.1833	2.0112
Average(3)[d]	7.7239	8.3944	0.6704**
Total(10)[e]	70.8014	73.7676	2.9663
Average(10)[f]	7.0801	7.3768	0.2966**

[a]Significantly different from 0 by the t test considering a significance level of 1%
[b]Significantly different from 0 by the t test considering a significance level of 5%
[c]Totals obtained from the first three generations
[d]Average value obtained from the first three generations
[e]Totals obtained from ten generations
[f]Average value obtained from ten generations

observed in generations one, two and three, and in average, the mean bias was larger than that for R.PS ($p < 0.01$).

After ten cycles of selection the gains were around 6.89% and 6.30% smaller than those for G20.PS, but they were 5.08% and 6.91% larger than those found for R.PS (Table 7). Bias was not detect only for generation 10.

No significantly differences were found between biases yielded in G20.PS and G10.PS settings. The same result was detected when comparisons were between G10.PS and R.PS. The average bias after ten cycles was 0.0901 while the value 0.1235 was observed in the R.PS setting; the difference was not significantly different from zero.

Table 13: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting with pre-selection and evaluated by a model including 10 identified genes as fixed effects (G10.PS).

Generation	**TGG**	**PGG**	**B**
1	6.1921	7.0737	0.8816**[a]
2	8.7340	9.2607	0.5267**
3	6.5618	7.1081	0.5463**
4	7.1128	7.4834	0.3706**
5	6.3726	6.7100	0.3374**
6	6.5808	6.7117	0.1310**
7	6.2020	6.3349	0.1329**
8	6.2276	6.3390	0.1114**
9	5.9908	6.1534	0.1626**
10	5.9452	5.9473	0.0021[n.s.][b]
Total(3)[c]	21.4879	23.4425	1.9546
Average(3)[d]	7.1627	7.8142	0.6515**
Total(10)[e]	65.9197	69.1223	3.2026
Average(10)[f]	6.5920	6.9122	0.3203**

[a]Significantly different from 0 by the t test considering a significance level of 1%
[b]Not significantly different from 0 by the t test.
[c]Totals obtained from the first three generations
[d]Average value obtained from the first three generations
[e]Totals obtained from ten generations
[f]Average value obtained from ten generations

The G05.PS setting(Table 14) includes 5 shared genes in the model as fixed effects, which corresponds to 10.96% of the additive genetic variance of the W550.

In short term selection, TGG and PGG were 4.21% and 3.46% smaller than in G10.PS. No difference on biases were detected by the t test over the generations between G05.PS and G10.PS. Comparison with R.PS showed that the gains were 3.31% and 9.24% larger respectively to TGG and PGG. Regarding R.PS, the average bias was 68.79% larger ($p < 0.01$).

In long term, the percentage of loss of the gains were 2.97 and 2.99 respectively for TGG and PGG when compared with G10.PS. Compared with

Table 14: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting with pre-selection and evaluated by a model including 5 identified genes as fixed effects (G05.PS).

Generation	**TGG**	**PGG**	**B**
1	5.7982	6.6864	0.8883^{**} [a]
2	8.4765	9.1339	0.6575^{**}
3	6.3094	6.8119	0.5025^{**}
4	6.8137	7.1053	0.2916^{**}
5	6.3098	6.6105	0.3008^{**}
6	6.2933	6.4634	0.1700^{**}
7	6.1808	6.2849	0.1041^{*} [b]
8	6.0974	6.1790	$0.0816^{\text{n.s.}}$ [c]
9	5.9608	5.9762	$0.0154^{\text{n.s.}}$
10	5.7198	5.8020	$0.0822^{\text{n.s.}}$
Total(3) [d]	20.5841	22.6322	2.0483
Average(3) [e]	6.8614	7.5441	0.6828^{**}
Total(10) [f]	63.9595	67.0534	3.0939
Average(10) [g]	6.3960	6.7053	0.3094^{**}

[a] Significantly different from 0 by the t test considering a significance level of 1%
[b] Significantly different from 0 by the t test considering a significance level of 5%
[c] Not significantly different from 0 by the t test.
[d] Totals obtained from the first three generations
[e] Average value obtained from the first three generations
[f] Totals obtained from ten generations
[g] Average value obtained from ten generations

R.PS the gains were 2.17% and 4.04% larger for TGG and PGG respectively. No biases were detected for the first three generations. The average bias after ten generations was 0.3094, which was not significantly different from that found for the R.PS and G10.PS.

In the G02.PS setting, 4.38% of the additive genetic variance of the W550 was under control(or 2 shared genes).

The genetic gains obtained were the smallest ones in short and long term (Table 15) among those for the settings with pre-selection and identified genes effects included in the model. The predicted genetic gains were

Table 15: Average value over 100 replicates of the true genetic gain(TGG), predicted genetic gain(PGG) and bias(B) over 10 generations for a setting with pre-selection and evaluated by a model including 2 identified genes as fixed effects (G02.PS).

Generation	**TGG**	**PGG**	**B**
1	5.6476	6.4617	0.8141**[a]
2	8.1662	8.7035	0.5373**
3	6.1674	6.6041	0.4367**
4	6.8075	7.1521	0.3446**
5	6.3312	6.5714	0.2402**
6	6.3255	6.5698	0.2444**
7	6.0721	6.1980	0.1259**
8	6.0499	6.0959	0.0460[n.s.][b]
9	5.7433	5.8755	0.1322**
10	5.8171	5.8522	0.0351[n.s.]
Total(3)[c]	19.9812	21.7693	1.7881
Average(3)[d]	6.6604	7.2564	0.5960**
Total(10)[e]	63.1277	66.0841	2.9564
Average(10)[f]	6.3128	6.6084	0.2956**

[a]Significantly different from 0 by the t test considering a significance level of 1%
[b]Not significantly different from 0 by the t test.
[c]Totals obtained from the first three generations
[d]Average value obtained from the first three generations
[e]Totals obtained from ten generations
[f]Average value obtained from ten generations

approximately 13.55%, 7.14% and 3.81% smaller than those for G20.PS, G10.PS and G05.PS respectively considering the first three generations. The PGG in this setting was 5.63% larger than that in Table 7. Biases were observed in the three generations. The mean value of the bias was 0.5960, while the values 0.6704, 0.6515, 0.6828 and 0.2131 were found for G20.PS, G10.PS, G05.PS and R.PS. No significant differences were found in comparisons between G02.PS and the other settings that included identified genes in the analyses, but a actual difference was found between this setting and the R.PS ($p < 0.01$).

When ten generations were taken into account, the predicted gains were

around 10.42%, 4.39%, and 1.44% smaller than those for G20.PS, G10.PS and G05.PS, respectively. However, it was 3.97% larger than that for R.PS setting. Biases were not detected in generations 8 and 10. In average, no differences between bias was found in the settings which pre-selection have been performed.

In general significant differences on bias were found only in the first three generations among the settings which pre-selection have been performed. When ten generations were considered, in average, no significantly differences were observed in the bias. Besides, the last generations tended to do not show significant values of the bias (Tables 12 to 15.

The results obtained from these tables are due to the including of identified genes effects and due to the selection effect. Because these effects are confounded a method to assess the reduction of bias is shown below.

* * *

In order to verify if the effect of including genotypes have a direct effect in minimizing the selection bias, the a difference was calculated as indicated in equation (11).

In Table 16 the results for the differences for each amount of single genes controlled in the study are shown.

For Diff.1, the values obtained in the generations 2, 3, 4, 5 and 9 were not significantly different from zero, indicating that there was no effect of the identified genes in reducing selection bias. A larger bias than that obtained in the R.PS was detected in generations 6, 7, 8 and 10.

For Diff.2, an effective reduction of selection bias was also found in generation 3. No significant values were found in generations 2, 4, 5, 8, 9 and 10. Larger biases than those obtained in Table 7 were found in generations 6 and 7.

For Diff.3, real reductions were obtained in generations 1 and 2 and no reduction was observed for the later generations.

For Diff.4, reduction was found only in the first generations, and similarly to Diff.3, no reduction was observed for later generations.

Table 16: Average values over 100 replicates of the differences of the responses (A[a], B[b], C[c] and D[d])considering the true genetic gains(TGG).

Generation	**Diff.1**[e]	**Diff.2**[f]	**Diff.3**[g]	**Diff.4**[h]
1	-0.6494^{**}[i]	-0.7662^{**}	-0.6170^{**}	-0.6606^{**}
2	$0.0955^{\text{n.s.}}$[j]	$-0.074^{\text{n.s.}}$	-0.5984^{**}	$-0.3737^{\text{n.s.}}$
3	$-0.2588^{\text{n.s.}}$	-0.3759^{*}	$-0.3116^{\text{n.s.}}$	$-0.2982^{\text{n.s.}}$
4	$0.2589^{\text{n.s.}}$	$-0.1239^{\text{n.s.}}$	$0.2561^{\text{n.s.}}$	$0.0384^{\text{n.s.}}$
5	$0.2684^{\text{n.s.}}$	$0.3084^{\text{n.s.}}$	$0.0294^{\text{n.s.}}$	$-0.0096^{\text{n.s.}}$
6	0.4899^{**}	0.3261^{*}	$0.1633^{\text{n.s.}}$	$0.1251^{\text{n.s.}}$
7	0.3510^{*}[k]	0.3438^{*}	$0.0512^{\text{n.s.}}$	$0.0134^{\text{n.s.}}$
8	0.4101^{**}	$0.1981^{\text{n.s.}}$	$0.0334^{\text{n.s.}}$	$0.1925^{\text{n.s.}}$
9	$0.1078^{\text{n.s.}}$	$-0.0019^{\text{n.s.}}$	$-0.2190^{\text{n.s.}}$	$-0.0980^{\text{n.s.}}$
10	0.3929^{**}	$0.1858^{\text{n.s.}}$	$0.1573^{\text{n.s.}}$	$0.2949^{\text{n.s.}}$

[a]response from a single stage selection
[b]A+P–b = response from a single stage selection + extra response due to pre-selection – selection bias
[c]A+G = response from single stage selection + extra response due to the identified genes
[d]A+P+G–c = response from a single stage selection + extra response due to pre-selection + extra response due to the identified genes – selection bias
[e]Difference 1 = (B – A) – (D – C) for 20 identified genes included in the model
[f]Difference 2 = (B – A) – (D – C) for 10 identified genes included in the model
[g]Difference 3 = (B – A) – (D – C) for 05 identified genes included in the model
[h]Difference 4 = (B – A) – (D – C) for 02 identified genes included in the model
[i]Significantly different from 0 by the t test considering a significance level of 1%
[j]Not significantly different from 0 by the t test
[k]Significantly different from 0 by the t test considering a significance level of 5%

The results also shows that a reduction of bias was achieved in the first generation for all settings which included identified genes compared to the setting where no fixed effects of genes were included in the model (Differences 1, 2, 3 and 4). However, the values were not significantly different among the settings, suggesting that the bias reduction is not proportional to the amount of identified genes included in the analyses. Generation 1 still had animals from the base population (Figure 2). Because of that the shift in the means and variance due to pre-selection was not large and the extra information on the identified genes (even in small number) should have improved the predictions and consequently reduced the bias in the genetic gain.

In general, no effect of reduction of the selection bias was achieved by

including identified genes as fixed effects in the model for genetic evaluations. Besides, in some generations the bias on predictions of the genetic gains were larger when information of the identified genes were used.

The reduction on selection bias occurred because the effect of identified genes did not recover the missing information created in the sequential selection process. Although the inclusion of the identified genes in analyses tends to create more reliable estimates, it was not able to provide the enough information about the correct distribution, which should be used to make inferences. Thus, despite of the presence of the identified genes, inferences were performed under a incorrect probability distribution.

The larger bias when identified genes were used in genetic evaluations were found in some generations and are probably due to a poor estimation of the fixed effects due to the amount of levels of fixed effects in the model. It is suggested by the presence of these results in the settings which included 20 and 10 identified genes as fixed effects and not in those including 5 and 2 genes.

5 Implications

The information on identified genes should not be used if one intends to minimize the selection bias in genetic evaluations. The high costs of genotyping animals the requirement of large population for reliable estimates of the fixed effects and the absence of effect in reduction of selection bias turns this method of little practical application for the selection problem in animal genetic evaluation.

References

Bulmer, M. G. (1971). The effect of selection on genetic variability. *Amer. Nat.*, 105:201–211.

Cassel, B. G. and McDaniel, B. T. (1983). Use of later records in dairy sire evaluation: A review. *J.Dairy.Sci.*, 66:1–10.

Eriksson, J.-Å. (1982). Estimating sire's genetic value for milk yield in first and second lactation by different mixed model procedures, using selected second-lactation records–a simulation study. *Ac. Agric. Scand.*, 32:193–206.

Fernando, R. L. and Gianola, D. (1986). Optimal properties of the conditional mean as a selection criterion. *Theor. Appl. Genet.*, 72:822–825.

Fernando, R. L. and Gianola, D. (1990). Statistical inferences in populations undergoing selection or non-random mating. In: *Advances in Statistical Methods for Generic Improvement of Livestock* (D. Gianola and K. Hammond, eds.), Springer-Verlag, Berlin. 437–453.

Fries, L. A. and Schenkel, F. S. (1993). Estimation and prediction under a selection model. In: *Anais da 30ª. Reunião da Sociedade Brasileira de Zootecnia*. Rio de Janeiro–Brasil, 1–22.

Gianola, D. and Fernando, R. L. (1986). Bayesian methods in animal breeding. *J.Anim.Sci.*, 63:217–244.

Gianola, D., Im, S. and Fernando, R. L. (1988). Prediction of breeding value under hendersons's selection model: A revisitation. *J.Dairy.Sci.*, 71:2790–2798.

Goddard, M. (1990). Selection and non-random mating–discussion summary. In: *Advances in Statistical Methods for Generic Improvement of Livestock* (D. Gianola and K. Hammond, eds.), Springer-Verlag, Berlin. 474–475.

Goffinet, B. (1983). Selection on selected records. *Genet. Sel. Evol.*, 15:91–98.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In: *Proc. Anim. Breed. and Genet. Symp. in Honor of Dr. J.L. Lush*. Am. Soc. Anim. Sci., Champaign, IL, 10–41.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Henderson, C. R. (1982). Best linear unbiased prediction in populations that have undergone selection. In: *Proc. World Congr. Sheep Beef Cattle Breed.* (R. A. Barton and W. C. Smith, eds.). Dunmore Press, Palmerston North, vol. 1, 191–200.

Henderson, C. R. (1990). Accounting for selection mating biases in genetic evaluation. In: *Advances in Statistical Methods for Generic Improvement of Livestock* (D. Gianola and K. Hammond, eds.), Springer-Verlag, Berlin. 413–436.

Henderson, C. R., Kempthorne, O., Searle, S. R. and von Krosig, C. M. (1959). The estimation of environmental and genetics trends from records subject to culling. *Biometrics*, 15:192–218.

Hudson, G. F. S. and Schaeffer, L. R. (1984). Monte Carlo comparison of sire evaluation models in populations subject to selection and nornrandom mating. *J.Dairy.Sci.*, 67:1264–1272.

Im, S., Fernando, R. L. and Gianola, D. (1989). Likelihood inferences in animal breeding under selection: A missing-data theory view point. *Genet. Sel. Evol.*, 21:399–414.

Kennedy, B. W., Quinton, M. and van Arendonk, J. A. M. (1992). Estimation of effects of single genes on quantitative traits. *J.Anim.Sci.*, 70:2000–2012.

Keown, J. F., Norman, H. D. and L, P. R. (1976). Effects of selection bias on sire evaluation procedures. *J.Dairy.Sci.*, 59:1808–1816.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. xi. on the influence of natural selection on the variability of organs. *Phil. Trans. of the R. Soc. London*, A 200:1–66.

Pollak, E. J. and Quaas, R. L. (1981). Monte Carlo study of genetic evaluations using sequentially selected records. *J.Anim.Sci.*, 52:257–264.

Pollak, E. J., van der Werf, J. and Quaas, R. L. (1984). Selecion bias and multiple trait evaluation. *J.Anim.Sci.*:1591–1595.

Schenkel, F. S. (1998). *Studies on Effects of Parental Selection on Estimation of Genetic Parameters and Breeding Values of Metric Traits*. Ph.D. thesis, University of Guelph.

Schenkel, F. S., Schaeffer, L. R. and Boettcher, P. J. (2002). Comparison between estimation of breeding values and fixed effects using bayesian and empirical BLUP estimation under selection on parents and missing pedigree information. *Genet. Sel. Evol.*, 34:41–59.

Sorensen, D., Fernando, R. L. and Gianola, D. (2001). Inferring the trajectory of genetic variance in the course of artificial selection. *Genet. Res.*, 77:83–94.

Sorensen, D. A. and Kennedy, B. W. (1984). Estimation of response to selection using least-squares and mixed model methodology. *J.Anim.Sci.*, 58:1097–1106.

Thompson, R. (1979). Sire evaluation. *Biometrics*, 35:339–353.