THE EFFECT OF MILK SAMPLING PROCEDURES ON THE ROBUSTNESS OF GENETIC EVALUATION IN DAIRY CATTLE

R. van Dyk

Dissertation presented for the degree Magister Scientiae in the Faculty of Agriculture (Department of Animal Science) University of the Orange Free State

Supervisor: Dr. FWC Neser

Co Supervisor: Prof FH Kanfer

May 2000

INDEX

ACKNOWLEDGEMENTS

I would like to thank the following people:

My supervisor, Dr. F.W.C. Neser for his guidance, patience, motivation and helpfulness,

The University of the Orange Free State for the financial support during the time I worked for the University,

The Department of Animal Science for their support,

Prof. G.J. Erasmus for his input and guidance,

Prof. F.H.J. Kanfer from the University of Pretoria who as a friend and co-supervisor, made this study a possibility,

Luis and Christina Schwalbach for your kindness and given me a place to stay every time I went to Bloemfontein.

DEDICATION

TO MY FAMILY

- To my Heavenly Father, for His love and giving me the strength to complete my studies,

- To my parents, for all your love and guidance throughout my life, for all the support morally, financially and always being in my heart,

- To my lovely wife, Leen, for all your love, patience, and being a friend, wife and mother,

- To my lovely baby girl, Amanita, for your patience every evening when I worked and bringing so much love and joy into our lives,

- To my second baby girl, expected later this year, I know you will bring us just as much joy and happiness.

- Frans, Antionette and Annerie Kanfer, for your friendship.

CHAPTER 1

GENERAL INTRODUCTION

Milk recording entails measuring daily milk yield and composition and is used for herd management decisions and cow and sire evaluations. It was first implemented in South Africa in 1917, by the then Friesian Breed Society and subsequently taken over by the State in 1919 (Grobler & Loubscher, 1983). The Milk Recording Scheme was run entirely by the State and included on farm tests to check that participants followed the correct procedure. During 1975 it was decided to simplify the Scheme and allow farmers to participate more actively in the collection of multiple samples (morning and evening) and the sending to a central laboratory. In the laboratory the two samples were mixed and then analysed. Some errors did occur in the collection of the samples, but was in some way compensated for by what was thought to be a more accurate method of analysing in a central laboratory (Annual Report - ARC, 1996).

In order to obtain a more user-friendly Milk Recording Scheme, the Management Committee of the Dairy Cattle Performance Testing Scheme investigated the possibility of replacing composite sampling with a single sample (1995/6).

The following rules were laid down by the Management Committee :

(i) Composite samples will be replaced with a single sample.

(ii) The single sample will be taken at the first milking, after twelve noon, on the test day.

(iii) A formula, including a *calculated factor* based on hours between morning and afternoon milking, will be used to correct fat percentage only.

(iv) The protein percentage will be accepted without correction.

(iv) Milk weights will be recorded at each milking on the test day.

(vi) The time of each milking will be recorded.

This was implemented at the beginning of September 1995 (Annual Report - ARC, 1996).

Various reasons were presented for the simplification of the Scheme, the most important being obvious financial benefits and user-friendliness. The implementation of the new scheme was based on the fact that there were practically no differences between two test years averages (one year composite - and second year single sample) as presented in the National Dairy Cattle Performance Testing Scheme: South Africa's Annual Report 1996.

The practical reasons for changing to a single sample is obvious due to the following:

* Time is saved.

* There is no need to match the bottle from the previous sample in order to combine the sample.

* It is cheaper.
* Reduction in time and money leads to a more user friendly milk-recording scheme (de Waal, 1995).
* Overall improvement of logistics.

This comparison was, however, not scientifically valid, as there was no data available to make a scientific evaluation and compare the differences. In order to compare the two sampling methods, it is important to take both samples (single and composite) at the same time and day. A test day model for genetic evaluation can account for factors that are specific to each test day, such as management groups within a herd on test day, day of the year, and for each cow, days in milk, pregnancy status and number of times milked on test day. This clearly points to the problem of data collection. Many of these factors change from one test day to the next and would be difficult to model for 305 days yield (Janrozik *et al*, 1997). Errors are one third larger when estimates of lactation milk yield were based on one milking, one day each month rather than on both milkings (Smith *et al*, 1981).

Various factors influence the composition of milk for example:

(i) Stage of lactation : According to Stantan & Pearson (1992) the protein and butterfat production reach a low during peak lactation, after which it increase up to the end of lactation.

(ii) Variation in butterfat and protein percentage during milking: in the case of butterfat the most butterfat is produced during the last period of milking, while for protein this is the opposite and decreases during each period. When

comparing single and composite samples this clearly indicates why it should be taken at the same time and day.

(iii) Climate: According to Muller (1991) temperature has the biggest influence on milk yield and it will subsequently differ from one year to the next as well as from one milking to the next.

(iv) Nutrition: The different components of the ration and the combination in which they influence the composition of the milk. Single and composite samples taken when different rations are fed will thus be different in composition.

(v) Mastitis: Mastitis as well as any other diseases influences production and composition of milk.

All these factors indicate the importance of having a scientific evaluation with various records taken on the same day, time and year in order to prove the justification of changing from composite to single sampling.

To obtain information on the advantage and accuracy of having a single, instead of composite sample, records are needed. Suitable records are scarce since the collection is extremely time consuming and costly. De Waal (1995) did an investigation on the influence of sample frequency during milk recording on the reliability of performance testing in dairy cows. Single and composite samples were taken in four herds. A single sample was taken in the morning and evening at the same time as the two samples comprising the composite sample. Records were obtained from 236 cows. Because of the lack of "real" data the information from

his study was used to simulate data in order to investigate this problem. The information is used to quantify the relationship between a single and composite sample.

Only butterfat was used in this study in order to simplify the calculation. Butterfat is historically defined as the amount of butterfat content in a 36ml milk sample taken during milking over a period of 24 hours. This measurement was accurate enough to calculate the breeding value to a satisfactory degree (Rautenbach, 1999, personal communication). A single sample will change the definition of measurement to the amount of butterfat found in a sample taken after twelve noon on the test day. From the resulting measurements of butterfat regarding the two methods, the single sample will be calculated according to the differences between the two measurements. The question that now arises is, can accurate breeding values for butterfat still be calculated.

BLUP requires knowledge of the (co)variance components of the random effects included in the mixed linear model. In most cases breeders have to replace the unknown (co)variance components by estimations based on sample information. The resulting predictor of the random effects is not BLUP but estimated BLUP (EBLUP(Estimated BLUP)). This has an effect on the mean squared error (MSE) or standard error (SE). The common practise is to approximate the MSE (SE) of EBLUP by the formulas of the MSE (SE) of EBLUP using estimations instead of the (co)variance components (Tuchscherer & Herrendörfer, 1998). Prasad & Rao (1990) showed that this MSE approximation underestimates the exact MSE of EBLUP by an amount $E(BLUP - EBLUP)^2$. According to Tuchscherer &

Herrendörfer (1998), the relative loss of accuracy of EBLUP ranged between 10 and 90 percent of the accuracy of BLUP. The MSE of BLUP are very dependent on the experimental design and the variance components. Hagger (1991) showed that by including information on the selected animal's daughters the predicted error variance was reduced by up to 53%.

The purpose of this study is to make use of simulation techniques to investigate how well breeding values can be predicted from a single sample in comparison to a composite sample. Simulation models usually exercise themselves through the medium of computing, and are made up of algorithms comprising mixtures of text statements and mathematical equations (Whitemore, 1987). The advantage of computer simulation is that it is less time consuming and costly than working with an actual biological system.

Selection entails breeding from the best individuals, and this implies the ability to choose individuals with the best breeding values. The basic effect of selection is to change the gene frequencies (the response to selection.). The influence of selection on the animal model for use in a DFREML (Meyer, 1995) analysis and the responding variance estimates will be investigated in the study as well.

The study is divided into three separate sections. In the first section the influence of selection on genetic parameter estimates will be investigated by means of simulating a population including selection in the model and then simulating data without selection. In the second section the simulation process is validated by determining how well the breeding values and the resulting variance components

are predicted under DFREML procedures. This is done by the calculation of the fixed effect, variance components and product moment correlation between the estimated breeding values and the simulated breeding values. In the third section a multiple trait analysis is done in order to determine the genetic relationship between the single and composite sample. If the relationship between the two traits is high, it can be assumed that they are largely influenced by the same genes.

CHAPTER 2

SIMULATION PROCESS

2.1 Introduction

The simulation of biological models aid in the understanding of simpler models and in the building of more complex models. Often simulation is a first step in finding out what happens when the usual assumptions are not fulfilled (van Vleck, 1993). In this chapter the simulation process, the simulated herd and the model for analysis is described for both the single and composite sample.

2.2 Material and methods

In simulation the accepted "true" parameters are known, therefore comparisons between these values and their corresponding simulated values could be made for both fixed effects and variance components.

2.2.1 The Simulated Herd

For this study a dairy herd was simulated. In the first four years all females generated were retained, where after the herd operated as a normal dairy enterprise. Hundred AI sires, of the same genetic level (i_e with equal a_i), were simulated. In the first year ten bulls were used as sires, in the second year five new bulls were added while five bulls from the previous year were

retained. This pattern continued for twenty years and was done to establish strong genetic ties between years. This means that bulls were used on a random, non-selected basis. Two different scenarios were simulated *viz* with and without selection. This process was repeated on an annual basis for the remaining term. For each simulated cow in the herd the following information was recorded: cow number, fixed effect level at first lactation, production measurement at first lactation, genetic component of the measurement, error component of measurement, sire and dam. These animals were replaced either under the selection scenario or the no-selection scenario.

2.2.1.1. Selection scenario

Animals were ranked according to the genetic component of their butterfat measurement. The 25% poorest performing animals were replaced by the best performing progeny. Cows older than six years were also replaced by the remaining best performing progeny. The selection was done on a yearly basis.

2.2.1.2 No-Selection scenario

The 25% poorest performing animals were replaced with animals taken at random from the progeny, implying no attention to the performance of the progeny. This process exercised no selection and is illustrated in the fixed effect Figure 3.1.

Comparing results from these two scenarios describes the effects of selection on the estimation of the fixed effects, while the second scenario also helps to cross-evaluate the computer program. Measurements were simulated applying a model consisting of a fixed effect, random genetic and random error component. Only these components were included to keep the model simple. For the purpose of this study the fixed effect can be viewed as a year effect, or perhaps a herd-year effect.

2.3 Simulation process

The model used to simulate the composite sample can be written as:

$$y_{ij} = f_i + a_{ij} + e_{ij}$$

where f_i represents the fixed effect at level i,

a_{ij} represents the random genetic component of animal j under fixed effect i,

e_{ij} represents the random error component,

y_{ij} represents butterfat, and with the genetic and error components statistically independent.

The model used to simulate butterfat based on single sample measurements is:

$$y_{ij} = f_i + a_{ij} + d_{ij} + e_{ij}$$

with:

d_{ij} a random difference between composite and single sample as calculated from the data of de Waal (1995). The rest of the factors in the model stayed the same.

The random genetic component, (a_{ij}), and error component, (e_{ij}), has a normal distribution with average 0, variances σ_a^2 and σ_e^2 respectively ((a_{ij} ~ N(0, σ_a ~) , (e_{ij} ~ N(0, σ_a ~)).

This corresponds to the methods used by van Vleck (1993). He described the method of simulation as to obtain, in some way, pseudo-random values from a normal distribution with a mean, zero, and variance, one. This is similar to a Monte Carlo simulation. Tuchscherer & Herrendörfer (1998) and Canavesi & Miglor (1999) proposed similar models to evaluate estimated BLUP:

A standard Desktop PC, run under Linux (Red Hat, version 6.0) and programmed in Fortran was used for the simulations.

The fixed effect component is calculated by a prescribed formula defined as:

$$f_i = c + m(i)$$

with:

$m_i = 12.5$ and $c_i = 100$ if $1 \leq i \leq 8$,

$m_i = -4.5$ and $c_i = 201$ if $9 \leq i \leq 16$ and

$m_i = 10$ and $c_i = 10$ if $17 \leq i \leq 20$.

This equation estimated the "true level" of the fixed effect as is presented in Table 3.1. The genetic and error component were obtained by generating random numbers from a normal distribution with mean 0 and variances σ_a^2 and σ_e^2 respectively. The genetic- (σ_a^2 = 293) and error variance (σ_e^2 = 534) from the study of du Toit *et al*, (1998) was used in the simulation. This means that a heritability estimate of 0.35 was used.

The simulation needs to be repeated several times in order to describe the randomness in the system and to determine the long-term expected effect, which is used for comparisons. Because of logistic constrains, only fifty repetitions per scenario were simulated. The data, consisting of animal number, sire, dam, fixed effect level, and production measurement was analysed using REML procedures (Meyer, 1995). From this the estimated fixed effect, predicted breeding values and estimated variance components were determined. These results were then compared with the true values used in the simulation.

CHAPTER 3

THE EFFECT OF SELECTION ON GENETIC PARAMETER ESTIMATES

3.1 Introduction:

Genetic variance estimates generated from selected individuals can be quite different from those in the unselected base population. This was reported by Lynch and Walsh (1998) who stated that REML yields unbiased estimates of additive genetic variance in the base population, if the base population consists of unrelated, unselected, and non-inbred individuals and phenotypic data are available for all selected and unselected individuals.

Selection may cause problems in the estimation of genetic parameters and breeding values. Data available to animal breeders are invariably provided by herds in which artificial selection has been practised. Consequently, the usual assumption of random sampling invoked for estimation and prediction is no longer valid (Schaeffer *et al*, 1998). Simulation is the only apparent method for examining properties of estimators of variances and of their ratios (Henderson, 1977).

The aim of this chapter was to compare the fixed effect and genetic parameter estimates for both a selection and no-selection scenario utilising simulated data.

3.2 Material and Methods

The model described in Chapter 2 was used to simulate 50 repetitions of a composite sample under both a selection and non-selection scenario. REML (Meyer, 1995) procedures were used to estimate genetic parameters.

The following model was fitted:

$$Y = X\beta + Za + e$$

Where **y** is a **n×1** vector of records, **X** is a **n×p** incidence matrix that relates data to the unknown vector of location parameters β. The vector β contained year as fixed effect. The incidence matrix **Z** relate the unknown random vectors of the direct breeding value (**a**) to **y.** The unknown vector **e** contains the random residuals due to environmental effects peculiar to individual records.

3.3 Results and discussion

In Table 3.1 and Table 3.2 the true-, the average-estimated, as well as the 99% confidence interval for the true fixed effect levels, as calculated from the fifty simulation repetitions respectively, for both the selection and no-selection scenarios, are presented.

Table 3.1: True fixed effect value, average estimated fixed effect levels and 99% confidence interval for the true fixed effect level. No-selection scenario.

NO-SELECTION SCENARIO				
			99% Confidence Interval of estimated value	
Fixed effect level	**True value**	**Average estimated value**	**Lower border**	**Upper border**
1	112.5	112.2	111.5	112.9
2	125.0	125.0	124.4	125.7
3	137.5	137.4	136.8	138.0
4	150.0	150.3	149.6	151.0
5	162.5	154.4	153.3	155.4
6	175.0	167.0	166.0	168.0
7	187.5	185.2	183.9	186.5
8	200.0	197.3	195.9	198.6
9	160.5	157.8	156.6	158.9
10	156.0	152.9	151.7	154.2
11	151.5	148.8	147.8	149.9
12	147.0	145.0	143.9	146.1
13	142.5	141.0	139.9	142.1
14	138.0	134.0	135.7	138.2
15	133.5	132.7	131.4	134.0
16	129.0	127.7	126.3	129.1
17	180.0	179.3	178.1	180.6
18	190.0	189.3	188.2	190.3
19	200.0	199.4	198.2	200.6
20	210.0	209.6	208.4	210.8

Table 3.2: True fixed effect values, average estimated fixed effect and 99% confidence interval for the true fixed effect level. Selection scenario.

SELECTION SCENARIO				
			99% Confidence Interval of estimated value	
Fixed effect level	**True value**	**Average estimated value**	**Lower border**	**Upper border**
1	112.5	112.6	111.7	113.3
2	125.0	125.3	124.4	126.2
3	137.5	137.3	136.5	138.1
4	150.0	149.9	149.0	150.6
5	162.5	168.0	166.9	169.2
6	175.0	183.7	182.6	184.9
7	187.5	201.5	200.2	202.7
8	200.0	214.2	212.7	215.8
9	160.5	176.0	174.7	177.2
10	156.0	171.8	170.4	173.1
11	151.5	166.0	164.8	167.2
12	147.0	162.9	161.4	164.4
13	142.5	159.3	157.7	160.9
14	138.0	154.8	152.9	156.8
15	133.5	151.8	150.3	153.2
16	129.0	147.2	145.4	148.9
17	180.0	198.2	196.4	200.1
18	190.0	208.5	206.9	210.1
19	200.0	218.7	217.1	220.3
20	210.0	228.7	227.3	230.1

It is interesting to note that from year one up to four the estimated fixed effects are included in the 99% confidence interval. This could be ascribed to the fact that there was no selection exercised in this period as to obtain and build a herd. The estimated fixed effect for the remaining period were substantially over estimated under the selection scenario. In the no-selection scenario the fixed effect levels remained close to the true levels used in the simulation (see Figure 3.1).

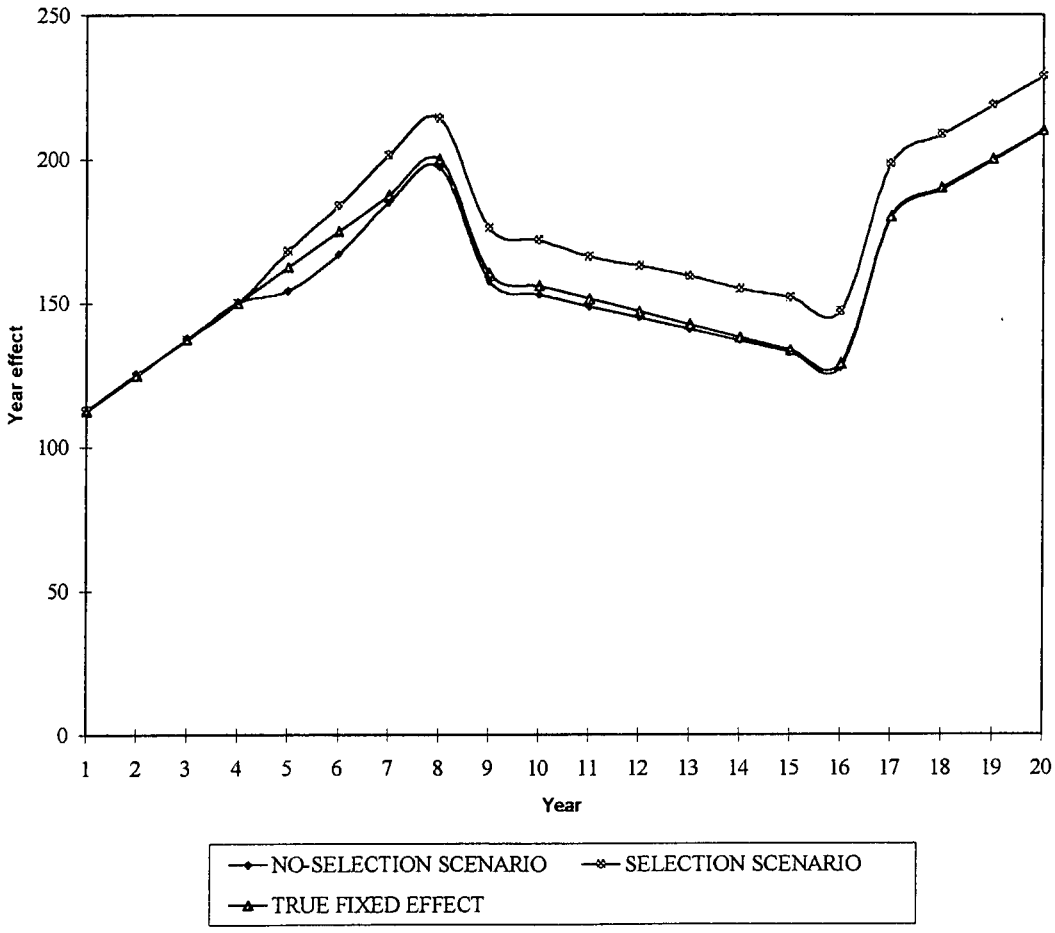

Figure 3.1: Estimated and true fixed effect levels

As depicted in Table 3.2 and Figure 3.1 there is a considerable overestimation of the fixed effect for the model under selection. When selection was not included in

the simulation the estimation of the fixed effect was much closer to the true fixed effect value.

Figure 3.2 shows the genetic trend for the selection and the no-selection scenario, respectively. From the figure it can be seen that genetic progress was made under the selection scenario as expected, while the no-selection scenario made little progress.

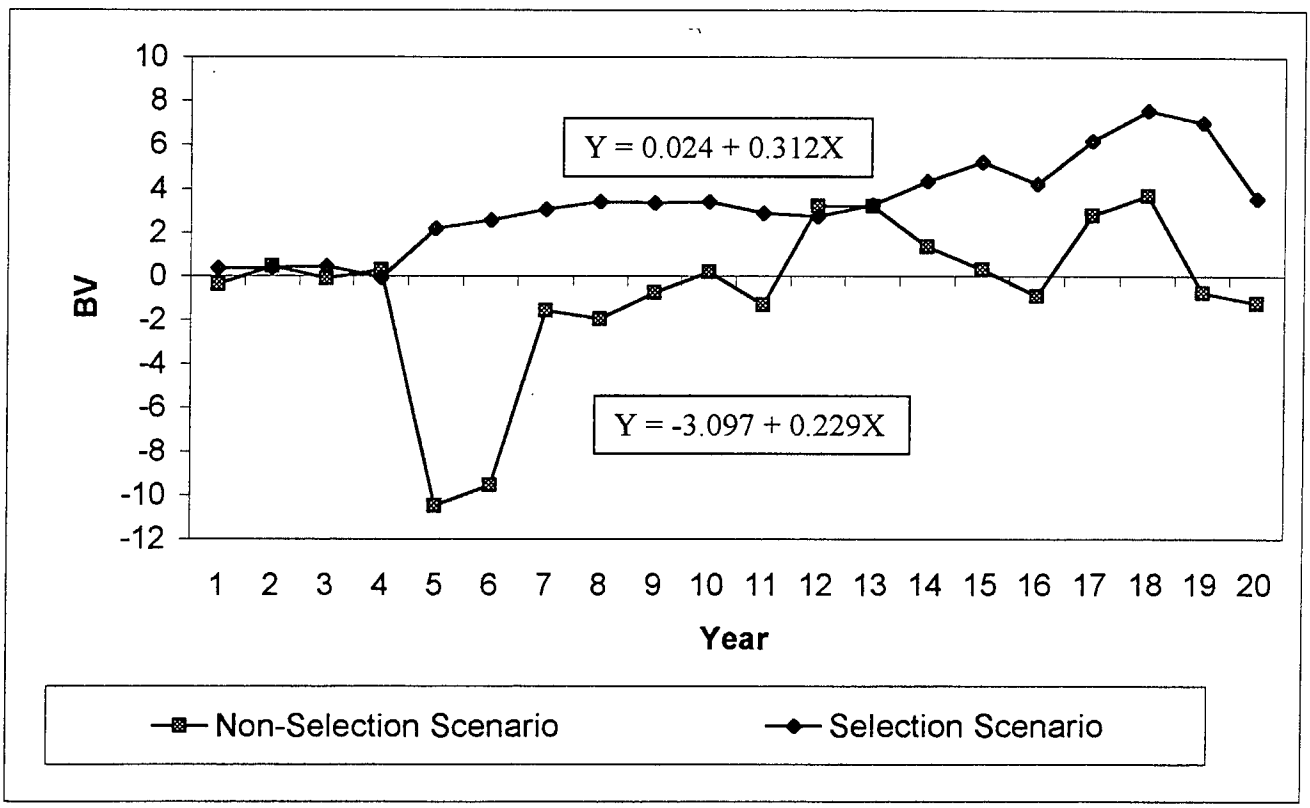

Figure 3.2 Genetic trends for the selection and non-selection scenario.

The trend under selection scenario showed an increase of 312g per year, which was higher than the 229g of the non-selection scenario. The progress made under the no-selection scenario could be ascribed to random selection. This is an indication that the procedures used to simulate both scenarios were successful.

In Table 3.3 and Table 3.4 the estimated genetic and the error variance for the fifty repetitions for the two scenarios are presented. The tables also include the true genetic and error variances used in the simulation as well as the 99% confidence interval.

Table 3.3: Estimated variance components (No-selection scenario)

Simulation Round	Genetic Variance	Error Variance	Simulation Round	Genetic Variance	Error Variance
1	240.89	430.54	26	201.68	447.10
2	197.59	458.28	27	232.17	439.86
3	219.57	447.40	28	199.86	461.33
4	202.51	448.32	29	214.87	448.47
5	206.52	451.07	30	222.12	424.47
6	225.60	459.77	31	193.02	456.71
7	245.35	447.73	32	239.29	437.99
8	258.21	428.68	33	209.50	427.49
9	224.98	448.42	34	239.79	433.58
10	218.44	457.80	35	202.57	440.47
11	227.97	452.67	36	197.63	478.44
12	228.74	443.78	37	220.27	442.41
13	214.56	455.47	38	220.11	447.87
14	213.33	450.42	39	214.58	442.20
15	202.33	447.70	40	188.99	460.51
16	244.10	425.71	41	195.70	463.34
17	218.63	437.06	42	207.63	439.25
18	268.35	395.75	43	225.45	430.90
19	197.16	463.01	44	207.95	430.93
20	208.96	460.99	45	224.74	449.51
21	210.52	444.77	46	238.18	435.82
22	205.52	443.53	47	217.04	430.18
23	234.23	429.12	48	180.23	475.41
24	232.84	433.34	49	224.64	441.25
25	251.74	434.91	50	194.99	421.01
			AVG	219.13	444.32
			STDEV	18.58	14.37
			99% confidence interval	211.70 ; 225.60	438.48 ; 449.86

Table 3.4: Estimated variance components (Selection scenario).

Simulation Round	Genetic Variance	Error Variance	Simulation Round	Genetic Variance	Error Variance
1	95.14	490.64	**26**	102.90	527.18
2	86.13	517.58	**27**	117.14	509.07
3	50.12	548.43	**28**	63.46	541.97
4	80.67	527.43	**29**	107.51	526.57
5	85.70	521.77	**30**	85.70	521.77
6	86.07	515.36	**31**	85.70	521.77
7	74.91	538.06	**32**	85.70	521.77
8	71.20	552.19	**33**	119.86	499.06
9	65.97	528.25	**34**	116.14	500.21
10	72.48	531.88	**35**	87.25	507.83
11	67.81	533.89	**36**	103.69	497.18
12	121.84	488.95	**37**	127.79	507.26
13	105.33	497.60	**38**	92.09	511.17
14	85.00	508.44	**39**	64.56	529.08
15	99.29	505.05	**40**	100.03	521.27
16	59.36	549.51	**41**	88.86	519.32
17	68.66	503.53	**42**	55.83	548.71
18	78.62	511.09	**43**	85.70	521.77
19	54.10	530.88	**44**	78.70	523.86
20	115.08	504.64	**45**	55.63	539.76
21	106.39	509.92	**46**	114.23	495.84
22	83.28	522.35	**47**	68.75	516.85
23	119.85	493.32	**48**	91.99	497.90
24	83.12	597.68	**49**	77.96	528.22
25	73.19	577.09	**50**	54.37	547.46
			AVG	85.70	521.77
			STDEV	21.43	21.45
			99% Confidence interval	76.21 ; 92.74	513.48 ; 529.46

The true genetic variance obtained from the study of du Toit *et al*, (1998) is equal to 293 and it seems that this value is underestimated under the selection scenario (average of 85.7). Under the no-selection scenario the underestimation is slight (average of 219.13). The error variance doesn't change much. The true error variance used in the simulation is equal to 534. Although the estimated error variance in both the scenarios was underestimated, the underestimation in the selection scenario was only slight (average of 521.77). A substantial underestimation occurred in the non-selection scenario (average of 444.3).

In Table 3.5 the heritability estimates of each simulation round for both the selection and non-selection scenarios are presented.

Table 3.5 Heritability estimates (in %) for the selection and no- Selection scenarios, respectively for 50 simulation rounds.

Simulation Round	Selection	No-Selection	Simulation Round	Selection	No-Selection
1	16.24	35.88	26	16.33	35.60
2	14.27	30.13	27	18.71	31.09
3	8.37	32.92	28	10.48	34.55
4	13.27	31.12	29	16.95	30.23
5	14.11	31.41	30	14.11	32.39
6	14.31	32.92	31	14.11	34.35
7	12.22	35.40	32	14.11	29.71
8	11.42	37.59	33	19.37	35.33
9	11.10	33.41	34	18.84	32.89
10	11.99	32.30	35	14.66	35.61
11	11.27	33.49	36	17.26	31.50
12	19.95	34.01	37	20.12	29.23
13	17.47	32.02	38	15.27	33.24
14	14.32	32.14	39	10.87	32.95
15	16.43	31.13	40	16.10	32.67
16	9.75	36.44	41	14.61	29.10
17	12.00	33.34	42	9.23	29.69
18	13.33	40.41	43	14.11	32.10
19	9.25	29.86	44	13.06	34.35
20	18.57	31.19	45	9.34	32.55
21	17.26	32.13	46	18.72	33.33
22	13.75	31.66	47	11.74	35.34
23	19.55	35.31	48	15.60	33.53
24	12.21	34.95	49	11.75	32.12
25	6.05	36.66	50	12.86	27.49
Average				14.08	33.01

A heritability of 35% was used in the simulation process. As in the estimated additive variance, a substantial underestimation of the heritability occurred in the selection scenario. The heritability estimates in the non-selection scenario were more in line with the true value used in the simulation process.

In Table 3.6 product moment correlation estimates between the simulated breeding value and the estimated breeding value are presented. The correlation range from 0.47 to 0.71 with an average value of 0.61 for the fifty repetitions under the selection scenario, while the values range from 0.75 to 0.82 with an average value of 0.79 under the no-selection scenario. Van der Werf (1990) also showed that after five generations of selection with data used for estimation of genetic parameters, which did not include all the relationships and data, the additive genetic variance decreased due to inbreeding, gametic phase disequilibruim ("Bulmer effect"-Bulmer, 1971) and covariance among animals. This is also supported by Satoh *et al.*, (1992) who showed that the correlation increased as the number of generations increased. Hagger (1991) showed that by including more information on the daughters during a selection experiment, an increase in the correlation between the true and predicted breeding value of up to 55% can occur.

Table 3.6: The product moment correlation estimates between the simulated breeding values and the estimated breeding value derived from DFREML analyses.

Simulation Round	Correlation with selection	Correlation without selection	Simulation Round	Correlation with selection	Correlation without selection
1	0.66	0.82	**26**	0.64	0.78
2	0.67	0.81	**27**	0.65	0.80
3	0.55	0.77	**28**	0.58	0.80
4	0.62	0.80	**29**	0.66	0.78
5	0.62	0.78	**30**	0.66	0.78
6	0.70	0.78	**31**	0.66	0.77
7	0.56	0.81	**32**	0.60	0.81
8	0.65	0.79	**33**	0.68	0.81
9	0.59	0.82	**34**	0.60	0.77
10	0.58	0.78	**35**	0.59	0.80
11	0.62	0.79	**36**	0.66	0.78
12	0.71	0.75	**37**	0.64	0.76
13	0.69	0.76	**38**	0.60	0.79
14	0.61	0.80	**39**	0.55	0.78
15	0.57	0.78	**40**	0.66	0.77
16	0.56	0.78	**41**	0.69	0.76
17	0.55	0.81	**42**	0.55	0.78
18	0.51	0.79	**43**	0.55	0.77
19	0.53	0.82	**44**	0.65	0.78
20	0.67	0.78	**45**	0.61	0.80
21	0.64	0.79	**46**	0.71	0.80
22	0.64	0.78	**47**	0.59	0.77
23	0.60	0.80	**48**	0.63	0.78
24	0.64	0.80	**49**	0.56	0.76
25	0.47	0.79	**50**	0.51	0.76
Average				0.61	0.79

To account for parental selection the following should be available

(a) complete pedigrees back to a base population of non-selected, non-related and non-inbred animals (Sorensen & Kennedy, 1984).

(b) records on all candidates available for selection (Henderson, 1975; Goffinet, 1983).

(c) knowledge of the selection process and distribution of selection criteria (Henderson, 1975; Im, 1989; Fernando & Gianola, 1990).

The first two conditions guarantee that likelihood based inferences not accounting for selection, are the same as those obtained considering selection, regardless of translation invariance of the selection criterion or its form (linear or non linear) (Gianola & Fernando, 1986; Im, 1989; Fernando & Gianola, 1990). Even if the first and second conditions are met, an additional condition of translation invariance of selection criteria must be verified for Henderson's Mixed Model Equations (HMME) to yield BLUE and BLUP, otherwise unbiasedness does not hold (Schaeffer *et al.*, 1998). The third condition is generally needed when data are missing. Im (1989) showed that if data are missing at random, inferences could be made using the likelihood function without accounting for the missing data process. Otherwise, the missing data process has to be described and included in the likelihood function (Schaeffer *et al.*, 1998).

Sorenson & Kennedy (1984) and van der Werf & de Boer (1990) supports these results and showed that estimates of genetic variance based on simulated data have shown to be unaffected by selection over generations if all data and all genetic relationships since the beginning of selection are included in the analysis. This

simulation showed that under certain experimental conditions the REML variance-component estimates are influence by selection. In this study as in a normal dairy enterprise animals were culled before their records could be included in the analysis.

3.5 Conclusion

It is notable from the results that deviations exist between the true and estimated values. It seems that this can be contributed to the process of selection and therefore deviations caused by the underlying assumptions of the mixed linear model. An assumption of Henderson's Mixed Model Equations is that the expected value of every element of **a** (vector of breeding values) is 0. If animals result from a long-term selection program, then the expected value of breeding values in later generations could be different from 0. Furthermore, for the records of selected individuals, all variances are reduced and non-zero covariances are generated between previously uncorrelated effects, such as between **a** and **e** (Schaeffer *et al*, 1998)..

A model with conditioning on selected base animals cannot be used to obtain unbiased estimates of variance components (as with this study). Conditioning on selected parents requires knowledge about the regression of parents on offspring, and therefore inference has to be made concerning the dispersion of base animals in any case of estimating variances based on their progeny (van der Werf & Thompson, 1992).

CHAPTER 4

A COMPARISON BETWEEN THE USE OF SINGLE AND COMPOSITE SAMPLES FOR ANIMAL EVALUATION

4.1 Introduction:

The gain in accuracy in prediction for one trait from using other correlated traits is partitioned into a direct gain from measuring other traits and gain because fixed effects are estimated more precisely (Thompson & Meyer, 1986). Even slight improvement of accuracy can have dramatic economic effects in large populations (Pollak *et al.*, 1984). The aim of this chapter was to determine whether the decision to replace the composite sample with a single sample could be justified. This will be done by comparing the composite and the single sample under a multiple trait analysis.

4.2 Materials and methods

It was decided to simulate a single and composite sample under the selection scenario. The effect of selection was discussed in length in the previous chapter.

Twenty five repetitions for each scenario was simulated in order to keep the analysis as simple as possible due to time constraints.

The model used to simulate the composite sample was as follows:

$$y_{ij} = f_i + a_{ij} + e_{ij}$$

where f_i represents the fixed effect at level i,

a_{ij} represents the random genetic component of animal j under fixed effect i,

e_{ij} represents the random error component,

y_{ij} represents butterfat, and with the genetic and error components statistically independent.

The single sample were simulated by the following model:

$$y_{ij} = f_i + a_{ij} + d_{ij} + e_{ij}$$

with f_i represents the fixed effect at level i,

a_{ij} represents the random genetic component of animal j under fixed effect i,

e_{ij} represents the random error component,

y_{ij} represents butterfat, and with the genetic and error components statistically independent.

d_{ij} a random difference between composite and single sample as calculated from the data of de Waal (1995).

Multiple trait analyses using REML (Meyer, 1998) procedures were done for all 25 simulation rounds to determine the (co)variance components and subsequent genetic correlation estimate between the single and composite samples.

The following model were used in the analysis:

$$Y = X\beta + Za + e$$

Where **y** is a **n×1** vector of records, **X** is a **n×p** incidence matrix that relates data to the unknown vector of location parameters β. The vector β contained year as fixed effect. The incidence matrice **Z** relate the unknown random vectors of the direct breeding value (**a**) to **y**. The unknown vector **e** contains the random residuals due to environmental effects peculiar to individual records.

Starting values from the single trait analysis were used for the analysis of the first simulation round. Thereafter results of the first simulation were used as starting values for the remaining simulation rounds. Because of the high genetic correlation that exists between the two traits it was difficult to reach conversion using the simplex search method. Therefore, the Powell (1965) search method was used.

4.3 Results and discussion

In Table 4.1 and Table 4.2 the true-, the average-, as well as the 99% confidence interval for the true fixed effect levels as estimated from the twenty five simulation rounds, for the single and composite sample respectively, are presented.

Table 4.1: True fixed effects, average estimated fixed effect levels and 99% confidence interval for the true fixed effects level. (Composite sample).

COMPOSITE SAMPLE				
		99% Confidence Interval of Estimated value		
Fixed effect level	**Average estimated value**	**Lower border**	**Upper border**	**True value**
1	106.5	107.5	114.0	112.5
2	118.4	122.8	124.6	125.0
3	130.9	134.6	136.7	137.5
4	143.8	148.4	147.5	150.0
5	163.8	175.2	170.1	162.5
6	179.1	188.4	187.5	175.0
7	193.3	198.3	204.0	187.5
8	206.2	217.0	215.3	200.0
9	169.7	186.5	173.9	160.5
10	165.2	180.8	172.9	156.0
11	161.0	175.6	166.8	151.5
12	157.0	169.5	162.6	147.0
13	153.6	158.7	164.3	142.5
14	148.1	157.0	159.0	138.0
15	144.8	153.6	153.4	133.5
16	139.7	145.8	146.5	129.0
17	189.2	202.0	195.4	180.0
18	199.2	208.1	206.8	190.0
19	209.1	216.0	219.2	200.0
20	218.7	223.0	233.9	210.0

Table 4.2: True fixed effects, average estimated fixed effect levels and 99% confidence interval for the true fixed effects. (Single sample).

SINGLE SAMPLE				
		99% Confidence Interval of Estimated value		
Fixed effect level	**Average estimated value**	**Lower border**	**Upper border**	**True value**
1	107.6	108.6	114.8	112.5
2	119.8	123.6	125.6	125.0
3	132.1	135.8	137.1	137.5
4	145.0	148.9	148.8	150.0
5	164.9	175.9	170.8	162.5
6	180.1	189.4	187.9	175.0
7	194.7	199.6	205.0	187.5
8	207.6	218.1	215.5	200.0
9	171.5	187.1	175.0	160.5
10	167.1	182.0	174.0	156.0
11	162.6	176.9	167.3	151.5
12	158.9	170.7	163.4	147.0
13	155.3	159.8	165.3	142.5
14	149.4	158.1	160.2	138.0
15	146.0	154.5	154.1	133.5
16	141.6	146.9	147.0	129.0
17	191.0	202.3	197.0	180.0
18	200.8	209.5	207.8	190.0
19	210.8	216.6	220.4	200.0
20	219.4	223.6	234.3	210.0

The 99% confidence interval does not always include the true value. This was explained and discussed in the previous chapter. As expected, the selection included in the model causes an overestimation in the composite and single sample fixed effects level.

In Figure 4.1 the true fixed effect levels, fixed effects when simulated for single sample and the fixed effect when simulated for composite sample are presented.

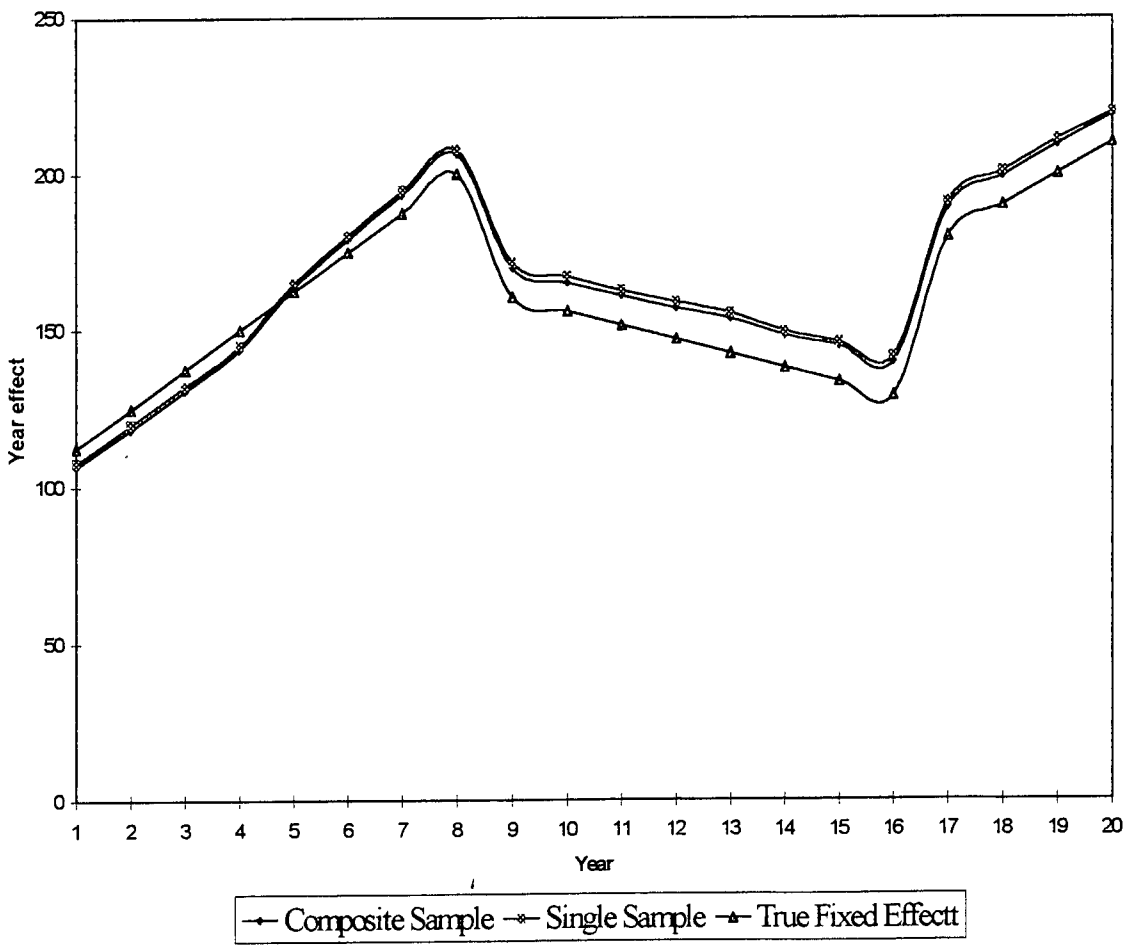

Figure 4.1: Estimated single and composite sample effects and true fixed effect.

In the first four years an underestimation of the fixed effect in both traits occurred. No selection took place during this period and all animals were retained. After this period (year 5-20) an overestimation of the fixed effect occurred. This could be ascribed to the selection that took place during that period. It is interesting to note that there is virtually no difference between the estimated fixed effect of the two traits.

In Table 4.5 the variance component estimates for the single and composite sample are presented.

Table 4.5 Estimated variance components for single and Composite sample (Multiple Trait).

Simulation Round	Single Sample Heritability	Single sample Additive Variance	Single sample Error Variance	Composite Sample Heritability	Composite sample Additive Variance	Composite sample Error Variance
1	16.62	81.49	408.94	19.21	99.97	420.54
2	16.96	81.37	398.38	19.58	100.02	410.73
3	17.21	81.37	391.57	19.67	100.02	408.47
4	16.68	81.37	406.51	19.08	100.02	424.16
5	17.19	81.37	391.94	19.74	100.02	406.58
6	16.74	81.37	412.58	19.03	100.02	425.45
7	16.71	81.37	405.48	19.21	100.02	420.54
8	16.56	81.37	410.10	19.16	100.02	422.10
9	16.15	81.37	428.96	18.69	100.02	418.34
10	16.76	81.37	422.50	19.43	100.02	435.13
11	17.04	81.37	396.17	19.44	100.02	414.36
12	16.60	81.37	408.93	18.83	100.02	431.05
13	16.70	81.37	405.86	18.98	100.02	426.92
14	16.94	81.37	399.14	19.50	100.02	412.94
15	16.76	81.37	404.24	19.43	100.02	414.76
16	16.74	81.37	404.62	19.31	100.02	418.04
17	17.21	81.37	391.47	19.69	100.02	407.95
18	16.32	81.37	417.14	18.74	100.02	433.82
19	17.14	81.37	393.35	19.71	100.02	407.52
20	16.80	81.37	403.11	19.31	100.02	417.87
21	16.34	81.37	416.74	18.66	100.02	436.01
22	16.29	81.37	418.20	18.50	100.02	440.57
23	16.19	81.37	421.38	18.42	100.02	443.00
24	16.21	81.37	420.62	18.66	100.02	436.03
25	16.71	81.37	405.63	19.07	100.02	424.50
Average	16.06	78.24	391.68	18.43	96.17	406.05

Both the additive and error variance in the composite samples was higher than those obtained in the single samples. This also reflected in the heritability estimates. The higher heritability estimates in the composite sample could be ascribed to the higher accuracy of the composite sample in predicting the BV of an animal. It should be noted that the simulated breeding value of an animal was used as basis for the simulation of the single and composite sample and this will explain the similarity of the additive variance between the different simulation rounds. In each case the genetic correlation between the two traits were estimated as one.

4.4 Conclusion

The results from the multiple trait analysis showed that a perfect positive (one) genetic correlation exist between the single and composite sample. This indicates that the same genes are responsible for both the single and composite sample and could therefor be considered as the same trait. This study showed by means of a simulation model that it is possible, in practise, to make use of a single sample instead of a composite sample in order to determine breeding values for dairy cattle. This is important as described in Chapter 1 in order to have a simplified scheme, which is easy to manage and saves money.

CHAPTER 5

SUMMARY

The study was divided into three separate sections. Firstly the influence of selection is investigated by means of simulating a population including selection in the model and then simulating data with out selection. In the second section the simulation process was validated by determining how well the breeding values and the resulting variance components were predicted under DFREML procedures. This was done by the calculation of the fixed effect, variance components and product moment correlation between the estimated breeding values and the simulated breeding values. In the third section a multiple trait analysis was done in order to determine the relationship between the single and composite sample.

The effect of selection showed a substantial overestimation of the fixed effect and the resulting variance component estimates. Various authors (Hagger, 1991; Tuchsceherer & Herrendörfer, 1998, van der Werf, 1990) supported these results and this effect can be attributed to fact that all information on all animals in the analysis should be available to compensate for selection.

The product moment correlation between the true genetic component and the breeding values determined by DFREML as well as the estimated variance

components confirmed that the simulation process and the model used to simulate data can be applied to simulate data. These results were used in a multiple trait analysis to evaluate the difference between a single and composite sample analysis. Results from multiple trait analysis showed a high genetic correlation between the breeding values of the composite and single sample. This indicates that the same genes are responsible for both traits and that it is indeed possible to make use of a single sample instead of a composite sample.

OPSOMMING

Die studie was in drie afsonderlike seksies opgedeel. In die eerste afdeling is die effek van seleksie ondersoek deur middel van die simulasie van data met en sonder seleksie. In die tweede afdeling word die simulasie proses ondersoek deur gebruik te maak van die teelwaardes en die variansie komponente soos bepaal deur DFREML. Dit is gedoen om die simulasie te verifiseer. Tydens die derde en finale afdeling word 'n veelvoudige eienskap analise gedoen om die verwantskap tussen die enkel en saamgestelde monster te bepaal.

Die effek van seleksie op 'n ontleding vir die bepaling van teelwaardes word duidellik aangetoon in die eerste afdeling. 'n Aansienlike oorberaming van die vaste effek word verkry en word toegeskryf aan die feit dat alle inligting aangaande elke dier wat in die analise gebruik word, beskikbaar moet wees vir ontleding. Hierdie bevinding word deur verskeie outeurs in die literatuur

ondersteun (Hagger, 1991; Tuchsceherer & Herrendörfer, 1998, van der Werf, 1990).

Die korrelasie tussen die teeltwaardes van die gesimuleerde data en die teeltwaardes soos beplaal deur DFREML (Meyer, 1995) asook die beraamde variansie komponente het getoon dat die simulasie proses en die model wat gebruik is korrek is en nou gebruik kon word vir die vergelyking tussen 'n enkel en dubbel monster. 'n Veelvoudige eienskap analise tussen die enkel en dubbel monster het 'n hoe genetiese korrelsie tussen die twee eienskappe getoon. Dit impliseer dat dieselfe gene verantwoordelik is vir die bepaling van beide eienskappe en dat dit wel moontlik is om die dubbel monster met die enkel monster te vervang.

CHAPTER 6

REFERENCES

Annual Report, 1996. National Dairy Cattle Performance Testing Scheme: South Africa. *Agriculture Research Council, Irene, South Africa.*

Bulmer, M.G., 1971. The effect of selection on genetic variability. *The American Naturalist.* 105:201.

Canavesi, F. & Miglor, F., 1999. Effect of de-regression on lactation number on MACE Evaluations: A Simulation study. Bulliten. 22:49

De Waal, H.C., 1995. 'n Ondersoek na die invloed van monsterfrekwensie tydens melk aantekening op die betroubaarheid van prestasietoetsing by melkkoeie. Univerity of Stellenbosh. M.Sc dissertation.

Du Toit, J., van Wyk, J.B. & van der Westhuizen, J., 1998. Genetic parameter estimates in the South African Jersey breed. *J. Anim. Sci.* 28: 146

Fernando, R. L. & Gianola, D., 1990. In: *Advances in Statistical Methods for Genetic Improvement of Livestock.* D. Gianola & K. Hammond (ed.). 437. Springer-Verlag, New York.

Gianola, D. & Fernando, R. L., 1986. Baysian methods in animal breeding theory. *J. Anim. Sci.* 63: 217.

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

Grobler, B.R. & Loubser, L.F.B., 1983. Die ontwikkeling van melkprestasie toetsing in die RSA. Nasionale Melkprestasie-en-Nageslag Toetsskema. Jaarverslag: 3.

Hagger, C., 1991. Changes in (co)variance derived properties of animal model breeding values when offspring information became available in a selection experiment. *J. Anim. Breed. Genet.* 108:1.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. *Biom.* 31:423.

Henderson, C.R., 1977. Simulation to examine distribution of estimators of variances and ratios of variances. *J. Dairy Sci.* 61: 267

Im, S., 1989. A note on sire evaluation with uncertain paternity. *Biom.* 31:749

Jamrozik, J., Schaeffer, L.R., & Dekkers, J.C.M., 1997. Genetic evaluation of dairy cattle using test day yields and Random Regression Model. *J. Dairy Sci* 80: 1217.

Lynch, M. & Walsh, B., 1998. Genetics and analysis of quantitative traits. Massachusetts, U.S.A.

Meyer. K., 1995. DFREML programs to estimate variance components restricted maximum likelihood using a derivative free algorithm. User notes.

Meyer, K., 1998. DFREML version 3.0. User notes.

Muller, C.J.C., 1991. The effect of shade during summer and confinement area during winter on milk production, milk composition, physiological responses and behaviour of Friesian cows in a temperate climate. MSc thesis, University of Stellenbosch, Stellenbosch.

Pollak, E.J., van der Werf, J., & Quaas, R.L., 1984. Selection bias and multiple trait evaluation. *J. Dairy Sci.* 67: 1590

Powell, M.J.D., 1965. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comp. J.* 7:155

Prasad, N.G.N. & Rao, J.N.K., 1990. Estimation of the mean square error of small area estimators. *J. Amer. Stat. Assoc.* 85: 163

Satoh, M., Nishida, A. & Furukawa, T., 1992. The influence of variation in the generation effect on the accuracy of predicting breeding values – a computer simulation on a closed herd of swine. *Anim. Sci. & Tech.* 63: 457.

Schaeffer, L.R., Schenkel, F.S. & Fries, L.A., 1998. Selection bias on animal model evaluation. *Proc. 6th World Congr. Genet. Appl. Livest. Prod.* 25:501. Armidale Australie.

Sorensen, D.A. & Kennedy, B.W., 1984. Estimation of genetic variance from unselected and selected populations. *J. Anim. Sci.* 59: 1213.

Smith J.W. & Pearson, R.E., 1981. Development and evaluation of Alternate testing procedures for Official Records. *J. Dairy Sci* 64: 466.

Stantan, T.L. Jones, L.R., Everett, R.W. & Kachman, S.D., 1992. Estimating milk, fat and protein lactation curves with a test day model. *J. Dairy Sci.* 75: 1691.

Thompson, R., & Meyer, K., 1986. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livest. Prod. Sci.* 15: 299.

Tuchsceherer, A., & Herrendörfer, G., 1998. Evaluation of estimated BLUP in mixed linear models by a designed computer simulation. *Proc 6th World Congr. Genet. Appl. Livest.* 25: 585. Armidale Australia.

Van der Werf, J.H.J., 1990. A note on the use of conditional models to estimate additive genetic variance in selected populations. *Proc. 4th World Congr. Genet. Appl. Livest. Prod.* 476. Edinburgh Scotland.

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

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

Van Vleck, L.D., 1993. Selection index and introduction to mixed model methods. CRC Press Inc. Florida. 399.

Whitemore, C.T., 1987. Elements of pig science. Longman Handbooks in Agriculture, 147. Harlow, Essex: Longman.