

The use of a cluster analysis in across herd genetic evaluation for beef cattle

F.W.C. Neser[#], G.J. Erasmus and M.M. Scholtz

Department of Animal, Wildlife and Grassland Sciences, UFS, PO Box 339, Bloemfontein 9300, South Africa

Abstract

To investigate the possibility of a genotype x environment interaction in Bonsmara cattle, a cluster analysis was performed on weaning weight records of 72 811 Bonsmara calves, the progeny of 1 434 sires and 24 186 dams in 35 herds. The following environmental factors were used to classify herds into clusters: solution for herd effects corrected for year-season, sex, age of dam and age at weaning (indicative of the management level in a herd), herd size and average temperature and rainfall. Two different genetic analyses were performed. Breeding values obtained in a univariate analysis were used as basis for comparison to breeding values obtained from a multivariate analysis where weaning weight in each cluster was considered as a separate trait. Direct additive, maternal additive, permanent maternal environment and Herd-Year-Season x Sire (HYSxS) interaction were included as random effects in both analyses. The direct genetic correlations between the clusters varied between 0.51 and 1.00. The low correlation estimates between some of the clusters indicate a possible genotype x environment interaction. Substantial reranking of sires between clusters did occur. However, further research is needed to identify and prioritize variables that can describe the genetics, management and climate of each herd more accurately.

Keywords: Bonsmara cattle, Genotype by environment interaction, weaning weight

[#] Corresponding author. E-mail: NeserFW.sci@ufs.ac.za

Introduction

Modern reproduction technologies such as artificial insemination (AI) and embryo transfer ease the movement of genetic material across countries. Sub-populations of traditional South African breeds, such as the Bonsmara and Dorper, exist in countries such as Brazil, Argentina and Australia. By combining these populations in a total breed analysis, selection progress can be enhanced and genetic evaluation costs reduced.

Genotype x environment interaction can be defined as different sets of genes that determine varying levels of expression in different environments (Bertrand *et al.*, 1987). Such genotype x environment interaction can reduce the accuracy of the genetic analysis (Neser *et al.*, 1996). If interactions exist between genotype and environment, the accuracy and reliability of such evaluations are overestimated (Ibi *et al.*, 2005).

In dairy cattle cluster analysis over different ecological regions and countries are used to overcome this problem. The herd-cluster model is appealing because an animal's genetic merit is predicted for each unique environment or management system, regardless of country borders (Weigel & Rekaya, 2000). A number of studies were done to investigate genotype environment interaction by using cluster analyses (Lin & Lin, 1994; Weigel & Rekaya, 2000; Fiske *et al.*, 2001; Zwald *et al.*, 2003a; b). These were mostly done on dairy cattle investigating borderless international genetic evaluations.

In South Africa large differences in animal performance exist between different biomes or ecological regions. These differences can also extend to management levels between different herds within a specific region. It is therefore possible, by using only South African data, to simulate a total breed analysis over different environments in order to investigate the feasibility of borderless genetic evaluations.

The aim of the study was to investigate the presence and consequences of genotype x environment interaction that could occur in borderless genetic evaluations. This was done by clustering the National Bonsmara dataset into different smaller sets, according to meteorological and performance data. Weaning weight in each cluster was considered as a separate trait in a multivariate genetic analysis.

Materials and Methods

Weaning weight records of 274 882 Bonsmara calves were available for the study. These records were obtained from the National Beef Cattle Improvement Scheme. After editing, weaning weight records on 72 811 calves, the progeny of 1 434 sires and 24 186 dams in 35 herds, were used in the analysis. The period covered was between five and 15 years per herd. The following edits were performed on the data: All incomplete records (records without both parents) and sires with less than five progeny as well as herds with less than five years data were deleted from the dataset. Only herds that were genetically linked to at least five other herds (via sires used) were used in the analysis. This was done to improve connectedness in the dataset. Connectedness was further enhanced by utilizing an extended pedigree file.

A stepwise regression analysis was done to assess the importance of different environmental factors on weaning weight. The following environmental factors were tested using SAS (1992) and were found to be significant: solutions for herd effects corrected for year-season, sex, age of dam and age at weaning (to give an indication of the management level within a herd), herd size, rainfall and temperature. Herds were allocated to five rainfall and six temperature zones using the mean annual rainfall chart of Dent *et al.* (1989) as presented in Tainton (1999) and a mean annual temperature chart of South Africa (Environmentek, CSIR, South Africa

http://www.environment.gov.za/Maps/PublishMaps/downloads/National/A4/Mean_annual_temperature.gif). These factors were used to classify herds into clusters.

Cluster analysis is an exploratory technique designed to classify data into subgroups, which share similar characteristics (Lin & Lin, 1994). The Fastclus clustering procedure in SAS (1992) was used for herd clustering. This procedure performs a disjoint cluster analysis based on Euclidean distances. This iterative method guarantees that the distances between all observations in the same cluster will be less than the distances between observations in different clusters (Zwald *et al.*, 2003a).

Two genetic analyses were carried out. In the first analysis a univariate animal model was fitted using all weaning weight records without clustering. In the second analysis a multivariate animal model, treating weaning weight as a different trait in the different clusters was fitted.

Estimates of the (co)variance components were obtained using the ASREML program developed by Gilmour *et al.* (2002). Direct additive, maternal additive, permanent maternal environment and Herd-Year-Season x Sire (HYSxS) interaction were included as random factors in both the univariate and multivariate analysis. The covariance between direct and maternal effects was excluded from the model as it causes convergence problems in the multivariate analysis. Herd-year-season and sex were included as fixed effects. Age of dam and age at weaning were fitted as linear regressors.

The following model was used for all the different analyses:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{c}_1 + \mathbf{Z}_4\mathbf{c}_2 + \mathbf{e}$$

Where \mathbf{y} is a $n \times 1$ vector of records, \mathbf{X} is a $n \times p$ incidence matrix that relates data to the unknown vector of location parameters $\boldsymbol{\beta}$. The vector $\boldsymbol{\beta}$ contained HYS and sex as fixed effects and age of dam and age at weaning as random regressors. The incidence matrices \mathbf{Z}_1 and \mathbf{Z}_2 relate the unknown random vectors of direct breeding value (\mathbf{a}) and maternal breeding value (\mathbf{m}), respectively, to \mathbf{y} . The incidence matrices \mathbf{Z}_3 and \mathbf{Z}_4 relate the unknown additional random vectors, permanent maternal environment (\mathbf{c}_1) and HYSxS interaction (\mathbf{c}_2), to \mathbf{y} . The unknown vector \mathbf{e} contains the random residuals due to environmental effects peculiar to individual records. The assumed (co)variance structure of the model is as follows:

$$\mathbf{V} \begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{c_1}\sigma_{c_1}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{c_2}\sigma_{c_2}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_n\sigma_e^2 \end{pmatrix}$$

where \mathbf{A} is the numerator relationship matrix, \mathbf{I}_{c_1} is an identity matrix with order number of cows, \mathbf{I}_{c_2} is an identity matrix with order number of HYSxS interaction levels and \mathbf{I}_n is an identity matrix with order number of records.

Weaning weight in each cluster was considered as a separate trait. The results of the single trait analysis were used as starting values for the multivariate analysis. The multivariate analysis was used to obtain correlation estimates between the clusters. Data from progeny and other relatives in different clusters can be used to predict the performance of each sire in each production environment (Zwald *et al.*, 2003a).

Results and Discussion

Four clusters were formed. Solutions for herd effects were by far the most important factor, accounting for about 82% of the variation in weaning weight. Ten percent of the variation was accounted for by rainfall while the rest of the factors made up less than 5%. The large contribution of the herd effects is unfortunate as it totally dominates the whole clustering process. Research to find better descriptors of the climate and management levels in the different herds should form part of further studies.

In Table 1 the basic statistics for each cluster are presented. One of the clusters (Cluster 2) had less than 3500 animals. The distribution of records between Clusters 1, 3 and 4 was quite even. Unfortunately only 22 sires were used in all four clusters, while 95 were used in at least three clusters and 207 in only two clusters.

Table 1 Summary statistics for weaning weight in the four different clusters

Cluster	Number animals	Number herds	Mean (kg)	Minimum (kg)	Maximum (kg)	Standard deviation (kg)	Coef. of variation (%)
1	22180	22	200.7	126.0	343.0	29.7	15
2	3188	8	184.5	100.0	290.0	29.8	16
3	27586	25	221.7	100.0	365.0	34.2	15
4	19857	14	235.3	103.0	370.0	36.7	16

Table 2 Distribution of common sires across clusters

	1	2	3	4
1		28	44	34
2			30	25
3				46

Cluster 2 had the least number of sires in common with the other clusters (Table 2). This was to be expected as it comprises only 3 188 weaning weight records. The restricted use of AI is, however, a common problem in all beef cattle herds in South Africa and this could, unfortunately, adversely affect the genetic correlation estimates between clusters. This is supported by Weigel & Rekaya (2000) who stated that inadequate genetic ties might lead to erroneous covariance estimates.

The direct heritability estimates in Clusters 1 and 2 were similar, while the estimates for Clusters 3 and 4 were closer to those obtained in the univariate analysis. The direct heritability estimates obtained in Clusters 1, 2 and 3 are similar to those obtained by Nesar *et al.* (1996) and Nephawe *et al.* (1999) in the same breed.

The estimates for maternal heritability were quite similar in both the clusters and the univariate analysis. The estimates are, however, lower than those obtained by Nesar *et al.* (1996) in the same breed. In general all heritability estimates were within the range (0.068 - 0.66) obtained by Meyer (1992) and Koots *et al.* (1994).

The estimates for permanent maternal environment were in general higher than the maternal heritability estimates. The estimates of HYSxS as a proportion of the total variance correspond to results obtained by Nesar *et al.* (1996) and Nephawe *et al.* (1999) in the same breed.

Table 3 Estimates for (co)variance components (kg^2), genetic parameters and standard errors obtained in the univariate and multivariate analysis

	Uni- Variate	Cluster			
		1	2	3	4
Direct heritability	0.18	0.11	0.11	0.15	0.21
s.e.	0.01	0.02	0.06	0.08	0.02
Maternal heritability	0.07	0.10	0.06	0.08	0.06
s.e.	0.01	0.01	0.04	0.01	0.01
Permanent maternal environment	0.13	0.10	0.16	0.13	0.13
s.e.	0.06	0.01	0.04	0.01	0.01
HYSxS as proportion of total variance	0.12	0.11	0.06	0.09	0.16
s.e.	0.00	0.01	0.02	0.01	0.01
Direct additive variance	93.17	49.47	109.41	91.27	129.78
Direct maternal variance	38.65	43.11	29.05	42.97	39.59
Variance due to permanent maternal environment	65.99	43.74	79.78	69.93	77.96
Variance due to HYSxS interaction	61.75	47.64	29.03	46.52	98.56
Error variance	270.3	253.5	241.0	280.6	275.1
Phenotypic variance	529.9	437.5	488.3	531.2	621.0

HYSxS - Herd-Year-Season x Sire interaction.

Table 4 Direct genetic correlation estimates (\pm s.e.) for weaning weight between the different clusters

	1	2	3	4
1		1.00 (0.16)	0.88 (0.21)	1.00 (0.22)
2			0.54 (0.26)	0.65 (0.35)
3				0.51 (0.27)

The direct genetic correlation estimates between the clusters varied between 0.51 and 1.00 (Table 4). In a preliminary study on a sub data set of the Bonsmara breed, in which AI was used extensively, Naser (2002) obtained genetic correlations that varied between -0.04 and 0.91 between six different clusters. This is both higher and lower than the present study. Robertson (1959) indicated that the genetic correlation gives a measure of the practical rather than statistical significance of the traits and suggested that an estimated genetic correlation appreciably lower than one (<0.80) would indicate changes in the ranking of genotypes in the two environments. Although the genetic correlation between some of the clusters was below 0.80 indicating a possible genotype x environment interaction, the high standard error in some of the estimates renders this result doubtful. The low correlations between some of the clusters indicate that a re-ranking of animals could take place. This would mean that different animals will be selected for different clusters (management levels). Similar results were obtained by Bradfield *et al.* (1997) and Nephawe *et al.* (1999), when investigating the possibility of a genotype x environment interaction between different regions.

A further analysis was conducted to assess the impact of genotype x environment interaction on the ranking of sires using estimated breeding values of sires from both the univariate and multivariate analyses. All sires with more than 100 progeny were sorted according to their estimated breeding values in the univariate analysis. The top 20 sires out of a possible 201 are presented in Table 5.

The bull that performed the best in all of the clusters ranked third in the univariate analysis (Table 5), while the best bull in the univariate analysis ranked 2nd, 7th, 21st and 2nd in the different clusters. There was, however, a substantial re-ranking of sires in the different clusters. Bull no 1143 for instance ranked 10th in the univariate analysis, but 22nd, 48th, 63rd, and 14th in the different clusters while bull no 153 556 ranked 29th in the univariate analysis, but 73rd, 74th, 10th and 99th in the different clusters. This re-ranking occurred despite the high genetic correlations that exist between the clusters.

Table 5 Estimated breeding values and ranking of the top 20 sires out of a possible 201 sires with more than 100 progeny according to the univariate analysis as well as their corresponding ranking in the different clusters

Sire	N ¹	Breeding value					Uni. var.	Ranking				Uni. Var.
		1	2	3	4	1		2	3	4		
45232	130	18.35	15.58	11.33	26.24	25.78	2	7	21	2	1	
62981	135	17.11	14.41	8.38	25.64	24.75	3	9	38	3	2	
3775	247	18.67	29.98	22.26	34.90	20.74	1	1	1	1	3	
145652	149	13.90	16.76	12.57	22.62	19.78	4	5	11	4	4	
4492	129	13.43	11.40	8.29	19.20	19.36	5	14	40	7	5	
56107	342	12.26	11.22	8.70	17.01	18.19	8	15	35	10	6	
112681	113	11.93	10.62	7.36	17.18	17.41	11	18	50	9	7	
37713	207	11.98	10.26	8.06	16.84	17.36	9	23	41	11	8	
177444	203	12.68	11.21	14.78	14.40	16.09	7	16	5	18	9	
1143	143	9.39	6.74	5.92	15.57	15.73	22	48	63	14	10	
52854	279	10.21	9.14	15.62	9.59	15.11	18	32	4	37	11	
32215	101	11.58	19.11	11.91	19.23	14.17	12	3	17	6	12	
123074	114	11.40	9.81	9.58	14.93	14.14	14	28	28	16	13	
63328	116	7.98	8.99	13.62	8.03	14.06	33	33	6	44	14	
5144	194	9.17	9.15	5.84	14.41	13.95	23	31	64	17	15	
31711	196	11.96	10.42	12.48	14.37	13.85	10	20	12	19	16	
197313	107	11.34	14.83	12.42	15.80	13.62	15	8	14	12	17	
3529	116	11.55	10.25	15.73	11.90	13.57	13	24	3	26	18	
135013	201	8.60	7.70	13.16	8.08	12.97	27	40	9	43	19	
32157	132	9.16	7.77	5.55	13.16	12.50	24	39	69	22	20	

¹number of progeny per sire.

Bulls in the top third are more likely to be selected for breeding purposes. It was therefore decided to estimate the product moment correlations as well as the Pearson correlation for animals in the top third, ranked according to their estimated breeding values in the univariate analysis (Table 6).

If Cluster 3 is ignored, the correlations indicate no genotype x environment interaction (product moment correlations above 0.80) and only a small to moderate re-ranking between the clusters and the univariate analysis. In the case of these clusters the product moment correlations were between 0.67 and 0.86. Between clusters the correlation was as high as 0.94. In contrast, re-ranking in Cluster 3 was the most severe. For this cluster all the product moment correlations were below 0.40. The reasons for the peculiar values obtained whenever cluster 3 is involved needs to be investigated in future studies.

Table 6 The product moment (above the diagonal) and Pearson correlation estimates (below the diagonal) for the top third sires, ranked according to their breeding values in the univariate analysis

	Clus 1	Clus 2	Clus 3	Clus 4	Uni- var
Cluster 1		0.85	0.25	0.85	0.86
Cluster 2	0.88		0.29	0.94	0.67
Cluster 3	0.16	0.18		0.20	0.35
Cluster 4	0.74	0.89	0.06		0.75
Univariate	0.74	0.65	0.50	0.58	

Conclusions

The results of this study give an indication of some of the problems that could occur in a cross country genetic evaluation when genotype x environment interaction is present. The most important problem in the beef industry in South Africa is the limited use of AI and the lack of ties between clusters. The precision of the estimated genetic parameters depended more heavily on the level of genetic ties between clusters than the number of observations per cluster (Weigel & Rekaya, 2000). Ties between beef herds in South Africa are much weaker than in dairy herds. Furthermore, more complex models are used in the beef industry compared to the dairy industry, which enhances this problem. The only way to overcome this problem is to actively promote the use of AI in the beef industry.

With the herd-cluster model, herds are grouped based on likeness, rather than location (Weigel & Rekaya, 2000). Animals can therefore obtain breeding values for different management levels. This would simplify selection substantially as the best-adapted animal for a specific environment or management level could be selected. In theory an animal could thus get four breeding values, one for each cluster. The different breeding values would then give an indication of environmental sensitivity. Breeding values of animals that are more sensitive to the environment should vary more between the different clusters. This could facilitate the selection of better adapted animals over a wider variety of management levels as well as environments.

The increasing use of large scale genetic evaluations also raises concern about possible occurrence of large differences between the environment of the seed stock producer and that of the commercial producer. With the current information available the best option for commercial producers seems to be to buy bulls from seed stock producers that have similar environments and production systems.

In this study, the average rainfall and temperature levels of each region *inter alia* were used to classify the herds in the different clusters. In future studies climatic factors specific to each herd should be used. Further research is needed to identify and prioritize variables that can describe the genetics and management levels of each herd more accurately before across country analysis could be performed.

Acknowledgments

The authors would like to thank the Bonsmara Breeder's Society as well as the ARC – Livestock Business Division for providing the data for this study.

References

- Bertrand, J.K., Hough, J.D. & Benyshek, L.L., 1987. Sire x environment interactions and genetic correlations of sire progeny performance across regions in dam-adjusted field data. *J. Anim. Sci.* 64, 77-82.
- Bradfield, M., Graser, H-U. & Johnston, D.J., 1997. Investigation of genotype x production environment interaction for weaning weight in the Santa Gertrudis Breed in Australia. *Aust. J. Agric. Res.* 48, 1-5.
- Dent, M.C., Lynch, S.D. & Schulze, R.E., 1989. Mapping mean annual and other rainfall statistics over Southern Africa, Report No. 109/1/89, Water Research Commission, Pretoria. (Cited by Tainton, 1999).
- Environmentek, CSIR, South Africa:
http://www.environment.gov.za/Maps/PublishMaps/downloads/National/A4/Mean_annual_temperature.gif
- Fikse, F., Rekaya, R. & Weigel, K.A., 2001. Genotype by environment interaction for milk production traits in Guernsey cattle. *Proc. Interbull Bull.* 27, 9-12.
- Gilmour, A.R., Cullis, B.R., Welham, S.J. & Thompson, R., 2002. ASREML user guide. NSW Agriculture Biometric Bulletin No. 3. NSW Agriculture, Orange Agricultural Institute, Forest Road, Orange 2800, NSW, Australia.
- Ibi, T., Hirooka, H., Kahi, A.K., Sasae, Y. & Sasaki, Y., 2005. Genotype x environment interaction effects on carcass traits in Japanese Black cattle. *J. Anim. Sci.* 83, 1503-1510.
- Koots, K.R., Gibson, J.P., Smith, C. & Wilton, J.W., 1994. Analyses of published genetic parameter estimates for beef production traits. 1 Heritability. *Anim. Breed. Abstr.* 62, 309-338.
- Lin, C.Y. & Lin, C.S., 1994. Investigation of genotype-environment interaction by cluster analysis in animal experiments. *Can. J. Anim. Sci.* 74, 607-612.

- Meyer, K., 1992. Variance components due to direct and maternal effects for growth traits of Australian beef cattle. *Livest. Prod. Sci.* 31, 179-204.
- Nephawe, K.A., Naser, F.W.C., Roux, C.Z., Theron, H.E., Van der Westhuizen, J. & Erasmus, G.J., 1999. Sire x ecological region interactions in Bonsmara cattle. *S. Afr. J. Anim. Sci.* 29, 189-201.
- Naser, F.W.C., 2002. A preliminary investigation into the use of cluster analyses in genotype x environment interaction studies in beef cattle. *Proc. 7th World Congr. Gen. Appl. to Livestock Prod. Montpellier, France.* 32, 391-393.
- Naser, F.W.C., Konstantinov, K.V. & Erasmus, G.J., 1996. The inclusion of herd-year-season by sire interaction in the estimation of genetic parameters in Bonsmara cattle. *S. Afr. J. Anim. Sci.* 26, 75-78.
- Robertson, A., 1959. The sample variance of the genetic correlation coefficient. *Biometrics* 15, 469-475.
- SAS, 1992. *Statistical Analysis Systems user's guide (Ver 6.03)*. SAS Institute Inc., Cary, North Carolina, USA.
- Tainton, N., 1999. *Veld Management in South Africa*. University of Natal Press, Scottsville, South Africa.
- Weigel, K.A. & Rekaya, R., 2000. A multiple – trait herd cluster model for international dairy sire evaluation. *J. Dairy Sci.* 83, 815-821.
- Zwald, N.R., Weigel, K.A., Fikse, W.F. & Rekaya, R., 2003a. Application of a multiple trait herd cluster model for genetic evaluation of dairy sires from seventeen countries. *J. Dairy Sci.* 86, 376-382.
- Zwald, N.R., Weigel, K.A., Fikse, W.F. & Rekaya, R., 2003b. Identification of factors that cause genotype by environment interaction between herds of Holstein cattle in seventeen countries. *J. Dairy Sci.* 86, 1009-1018.