Breeding Value Estimation for Environmental Sensitivity on a Large Dairy Cattle Data Set

L. Rönnegård[1], W.F. Fikse[1], H. A. Mulder[2,3], E. Strandberg[1]
[1]Dep Animal Breeding and Genetics, SLU, 750 07 Uppsala, Sweden.
[2]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, The Netherlands.
[3]Animal Breeding and Genomics Centre, Wageningen University, The Netherlands.

Abstract
Animal robustness, or environmental sensitivity, may be studied through individual differences in residual variance. These differences appear to be heritable, and there is therefore a need to fit models having breeding values explaining differences in residual variance. The aim of this report is to study whether breeding value estimation for environmental sensitivity (vEBV) can be performed on a large dairy cattle data set having around 1.6 million records. Two traits were analyzed separately, somatic cell score and milk yield. Estimation of variance components, ordinary breeding values and vEBVs was performed using standard variance component estimation software (ASReml), applying the methodology for double hierarchical generalized linear models. Converged estimates were obtained by running ASReml iteratively 20 times, which took less than 10 days on a Linux server. The genetic coefficients of variation for environmental variance were 0.45 and 0.52, for somatic cell score and milk yield, respectively, which indicate a substantial genetic variance for environmental variance. This study shows that estimation of variance components, EBVs and vEBVs, is feasible for large dairy cattle data sets using standard variance component estimation software.

Introduction
Differences between animals in robustness for a certain trait may be described in terms of differences in residual variance. For example, for some bulls there is considerable variation in performance within their daughter group whereas offspring of other bulls show relatively little variation. Models for micro-environmental sensitivity include breeding values explaining differences in residual variance (also referred to as genetic heterogeneity), and selection for robustness can be performed by selecting animals based on these breeding values.

Estimating the breeding values for residual variance (vEBV) and their associated variance components has not, to our knowledge, been performed on large scale dairy cattle data before. Previous studies have analyzed data including at most 10 thousand observations, where different Bayesian models have been fitted using MCMC methods (e.g. Sorensen and Waagepeterson 2003). These methods are computationally time consuming and not feasible to apply on large data sets.

A non-Bayesian method, based on hierarchical generalized linear models (Lee and Nelder 1996), was suggested for genetic heterogeneity models by Rönnegård et al. (2010). They showed that a model for genetic heterogeneity can be described as a double hierarchical generalized linear model (DHGLM; Lee and Nelder 2006) and that it can be fitted using standard variance component estimation programmes such as ASReml.

The aim of this report is to study whether breeding value estimation for environmental sensitivity (ie vEBVs) can be performed on a large dairy cattle data set having more than 1.6 million records.

Material and Methods
Data description
Data included 1.6 million test-day records on somatic cell score (SCS) and milk yield for nearly 180 thousand Swedish Holstein cows (Table 1, Figure 1). Data included information from first lactation only, and each cow had on average 9.5 recorded test-days. Pedigree was traced back such that sires of all cows with records had at least two generations of male ancestors known.

Statistical model
The fitted model consists of two parts, the mean and the residual variance. The model describing the mean includes fixed effects β , a random

animal effect a, and a random permanent environmental effect u:

$$y = \mathbf{X}\beta + \mathbf{Z}a + \mathbf{W}u + e$$

The animal effects are $a \sim N(0, \mathbf{A}\sigma_a^2)$ and the permanent environmental effects are $u \sim N(0, \mathbf{I}\sigma_u^2)$.

The residuals *e* are also assumed to be normally distributed but with different variances for each observation. The model for the residual variance is:

$$V(e) = \exp(\mathbf{X}_d\beta_d + \mathbf{Z}a_d + \mathbf{W}u_d)$$

where β_d are the fixed effects in the model for the residual variance, and a_d and u_d are the animal and permanent environmental effects, respectively, in the model for the residual variance. We assume $a_d \sim N(0, \mathbf{A}\sigma_{a_d}^2)$ and $u_d \sim N(0, \mathbf{I}\sigma_{u_d}^2)$. In the current paper, we compute the breeding values for the mean $\hat{a}$ (EBV) and the residual variance $\hat{a}_d$ (vEBV) assuming independence between a and a_d.

The DHGLM method by Rönnegård et al. (2010) was used for estimation of variance components and breeding values. The estimation method iterates between several rounds of ASReml runs by fitting a weighted animal model for the mean part and fitting the adjusted squared residuals from the mean model using a generalized linear mixed model to obtain new weights for the mean model. The final variance components estimates from the two models give $\hat{\sigma}_a^2$, $\hat{\sigma}_u^2$, $\hat{\sigma}_{a_d}^2$ and $\hat{\sigma}_{u_d}^2$, and the BLUP from the two models produces the estimated breeding values $\hat{a}$ and $\hat{a}_d$.

The following fixed effects were considered: year-season of calving (ys), herd-testday (htd), age-at-calving (AgeatC), and days-in-milk (DinM). Four seasons were defined: Jan-Mar, Apr-Jun, Jul-Sep and Oct-Dec. Adjacent herd-test-days were merged to ensure sufficient number of observations, using the algorithm by Crump et al. (1997).

Fixed effects included in the mean model were: ys, htd, AgeatC, $(\text{AgeatC})^2$, $(\text{AgeatC})^3$, DinM, exp(-0.05*DinM). Fixed effects included in the residual variance model were: ys, AgeatC, $(\text{AgeatC})^2$, DinM, $(\text{DinM})^2$.

Results

The estimation was performed by iterating between 20 ASReml runs. The variance component estimates changed by less than 10^{-4} between the last ASReml runs. In total the estimation took 10 days per trait on a Linux server.

Estimates for Somatic Cell Score

The variance component estimates were $\hat{\sigma}_a^2$ = 0.28, $\hat{\sigma}_u^2$ = 1.02, $\hat{\sigma}_{a_d}^2$ = 0.20 and $\hat{\sigma}_{u_d}^2$ = 0.58. Hence, estimated variance for the permanent environmental effects were slightly larger than for the animal effects both in the mean and variance parts of the model. As a reference, estimated residual variance from a model with constant residual variance was 1.35.

Estimates for Milk Yield

The VCEs were $\hat{\sigma}_a^2$ = 8.79, $\hat{\sigma}_u^2$ = 12.42, $\hat{\sigma}_{a_d}^2$ = 0.27 and $\hat{\sigma}_{u_d}^2$ = 0.30. Also here, VCEs for the permanent environmental effects were slightly larger than the VCEs for the animal effects both in the mean and variance parts of the model. Estimated residual variance from a model with constant residual variance was 10.5.

Discussion

For the first time we have shown that fitting a model for genetic heterogeneity is possible for large dairy data sets using standard VCE software.

Results indicated large genetic variance in residual variance for both milk yield and SCS. The genetic variance in residual variance $\hat{\sigma}_{a_d}^2$ is roughly the squared value of the genetic coefficient of variation for environmental variance (Mulder et al., 2007). The square root values of $\hat{\sigma}_{a_d}^2$ are 0.45 and 0.52, for somatic cell score and milk yield, respectively, which indicate a substantial genetic variance for environmental variance (Hill and Mulder, 2010). The genetic coefficients of variation for environmental variance in this study are in the range what has been found across traits (e.g. body weight and litter size predominantly) in different species (pigs, chickens, rabbits, mice), but are higher than the median reported in Hill and Mulder (2010)

across species and traits. The large genetic coefficients of variation indicate that, changing micro-environmental sensitivity by selection seems feasible.

A few possibilities for future development deserve to be mentioned. The model fitted in this study did not include a correlation between the random animal effects in the mean and residual variance parts of the model, which is a parameter of interest (see eg Sorensen and Waagepetersen, 2003), since for instance a positive correlation would imply that selection on high EBVs would also give high vEBVs and thereby increase the residual variance.

The distribution of SCS is skewed (Figure 1) and the estimates might be affected if Box-Cox transformed somatic cell counts are used instead (see Yang et al. 2011). The sensitivity of the estimates depending on the transformation of the trait values needs to be assessed in the future.

Estimation was performed as in Rönnegård et al. (2010) by iterating between several runs of ASReml. Recent developments in ASReml allow direct implementation of the algorithm. Hence, it will not be necessary to iterate between several runs of ASReml in the future and a dramatic decrease in computation time is expected.

Acknowledgement

This project was financed by the RobustMilk project, which is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. The content of this paper is the sole responsibility of the authors, and it does not necessarily represent the views of the Commission or its services.

References

Crump, R.E., Haley, C.S., Thompson, R., Mercer, J., 1997, Assigning pedigree beef performance records to contemporary groups taking account of within-herd calving patterns. Animal Science 65:193-198.

Hill, W. G., Mulder H. A., 2010, Genetic analysis of environmental variation. Genetics Research 92:381-395.

Lee, Y., Nelder, J.A., 1996, Hierarchical generalized linear models (with Discussion). J. R. Statist. Soc. B. 58:619-678.

Lee, Y., Nelder, J.A., 2006, Double hierarchical generalized linear models. App. Stat. 55: 139-185.

Mulder, H. A., Bijma, P., Hill, W. G., 2007. Prediction of breeding values and selection responses with genetic heterogeneity of environmental variance. Genetics 175:1895-1910.

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H., Strandberg, E., 2010, Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. Genetics Selection Evolution 42:8.

Sorensen, D., Waagepeterson, R., 2003, Normal linear models with genetically structured residual variance heterogeneity: a case study. Genetical Research 82: 207-222.

Yang, Y, Christensen, O.F., Sorensen, D., 2011, Analysis of a genetically structured variance heterogeneity model using the Box-Cox transformation. Genetics Research 93:33-46.

Table 1 Description of the Swedish Holstein data

No. of records	1,693,154	
No. of animals	177,411	
Years	2002-2009	
No. of months	96	
No. of herds	1,759	
No. of herd-test-days	21,570	
Mean age at calving	838 days	
Traits		
Somatic Cell Score	Mean: 2.36	Median: 2.05
Milk yield (l/day)	Mean: 29.13	Median: 29.20

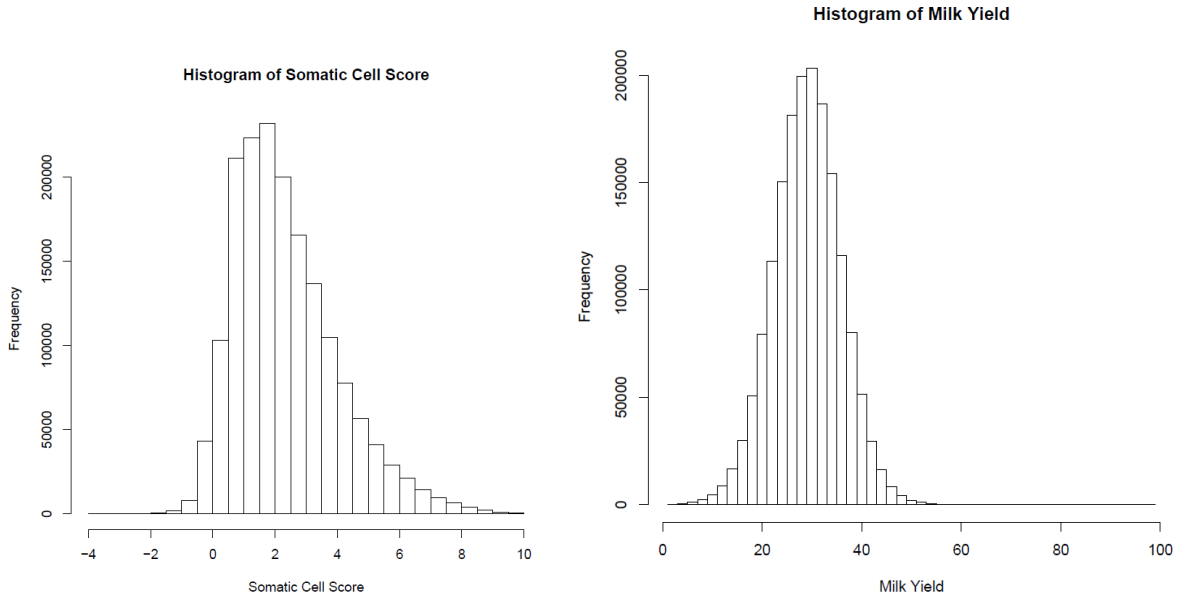

Figure 1 Histograms for the studied traits somatic cell scores and milk yield (l/day)