Accounting for Genotype by Environment Interaction in Genomic Predictions for US Holstein Dairy Cattle

F. Tiezzi[1], K.L. Parker Gaddis[2], J.S. Clay[3] and C. Maltecca[1]
Department of Animal Science, North Carolina State University, Raleigh, NC, USA
Department of Animal Science, University of Florida, Gainesville, FL, USA
Dairy Records Management System, Raleigh, NC, USA
email: f_tiezzi@ncsu.edu

Abstract

Genotype by environment interaction (GxE) is known as a differential response to changes in environmental conditions for individuals with different genetic background. Accounting for this effect could help improve genomic prediction for several traits in the dairy industry. We obtained 11,747 intra-herd-year-season daughters-yield-deviation for milk yield, for a total 482 Holstein bulls. Bulls were genotyped with the Illumina 50k Beadchip. Different models were implemented in a Bayesian framework to estimate genomic, environment and GxE variance components. Environmental effect were defined as 1) the permanent environmental effect of herd-year-season, 2) a double covariate on the latitude and longitude of the farm location, 3) multiple covariates for average herd-year-season values of maximum, minimum and average daily temperatures, relative humidity, wind speed and atmospheric pressure, 4) a triple covariate for management parameters such as number of cows in the herd, percentage of Holstein cows, and number of milking times per day, 5) permanent environmental effect of the herd. Several models of increasing complexity were tested in a cross-validation scheme. Accuracy was measured as the correlation between predicted and observed phenotypic values. Models that fitted GxE often presented non-null estimates of variance components for this effect and improved predictive ability by 2 to 7%. Our study suggests that the inclusion of GxE would be beneficial for genomic predictions.

Key words: Genotype by environment interaction, genomic prediction.

Introduction

Selection programs in dairy cattle have lead to large improvement of productive performance over the past decades. Breeding values for selection candidates are based on phenotypes recorded across varying environmental conditions for productive, reproductive and type traits. The availability of low-cost genotyping techniques has led to a remarkable increase in the power of prediction of breeding values through genomic selection.

Genotype by environment interaction (GxE) is a component of phenotypic variation that is often neglected in (genomic) breeding value prediction models, although several authors demonstrated its presence in dairy cattle populations. In Australia, Haile-Mariam et al. (2008) found GxE for production, survival and fertility traits over environments defined by different management (herd-size and level of production and climatic variables). Hayes *et al*. (2009) discovered regions in the genome associated with heat stress performance reduction, followed by Dikmen *et al*. (2013) that performed the same analysis in a US Holstein cow population. These studies suggest that cows will have different tolerance to environmental stressors in a manner that is determined by their genetic background. In Europe, Windig *et al*. (2011) estimated GxE for somatic cell score over different productive levels in several strains of Irish dairy cows, whereas Streit *et al*. (2013) and Norberg *et al*. (2014) found GxE for protein yield in German Holstein and Danish Jersey, respectively. In the United States, Oseni *et al*. (2004) found a moderate GxE for days open in US Holstein reared in south-eastern states, and Ravagnolo and Misztal (2000) demonstrated that selection for reduced heat stress in dairy cattle is possible.

What emerges from the investigation of GxE is that different environments can often

be expressed as one or more continuous variables. These can be derived from field measures (Ravagnolo and Misztal, 2000) or can be inferred with specific algorithms (Su *et al*., 2009). Among all field measures, climatic parameters play a significant role in defining different environments (Bohmanova *et al*., 2008).

Accounting for GxE in prediction models is common in plant breeding (Jacquin *et al*., 2014; Lopez-Cruz *et al*., 2015). There is a particular need to develop lines that show high performance in specific environments; in other words, 'specialist' genotypes are needed (Kassen, 2002). The same need might arise in dairy cattle when some environmental parameters cannot be (completely) controlled (e.g., temperature and humidity). In this case, breeders could develop specialist genetic material to perform in extremely hot and humid environments. On the other hand, breeding companies could decide to develop robust 'generalist' individuals that are capable of maintaining constant performance over different environmental conditions. This is especially relevant in the face of climatic changes.

The objective of this study was to estimate variance components for genomic, environment and genotype by environment effects on milk yield of US Holstein cows when the 'environment' was defined according to different descriptors, as well as to test the predictive ability of the different models using cross-validation.

Materials and Methods

Data

Milk yield records

Production records were extracted using Format 4 datafiles from the DRMS (Dairy Records Management System, Raleigh, NC, USA) database. The dataset included test-day records for each cow, for a total of 22,593,022 records on over 1,036,040 cows. Milk yield (expressed as kilograms of milk produced per day) was analyzed with the following model:

$$y_{ijklm} = \mu + parsolmf_i + hys_j + cowlact_{kl} + addgen_l + e_{ijklm}$$

where y_{ijklm} is the milk yield measure, μ is the overall mean, $parsolmf_i$ is the fixed effect of the *i*-th class defined as parity (4 classes) by stage of lactation (40 15-days classes) by milking frequency (2 or 3 times per day), hys_j is the fixed effect of the *j*-th herd-year-season class (HYS, where seasons are defined as 3 months periods: January to March, April to June, July to September, October to December), $cowlact_{kl}$ is the random effect of the *k*-th lactation of the *l*-th cow, $addgen_l$ is the additive genetic effect of the *l*-th cow and e_{ijklm} is the random residual. In order to complete the GxE analysis, data had to be organized into herd-year-season-daughter-yield-deviations (HYS-DYD) expressing the average and adjusted performance of a bull's daughters in a given herd-year-season class. HYS-DYD were defined as

$$HYS\text{-}DYD_{nj} = hys_j + addgen_l + e_{ijklm}$$

and then weighted for the effective daughters contributions. After editing, there were 11,747 HYS-DYD records for 482 bulls and 1,314 HYS classes from 103 herds. Bulls were genotyped with the Illumina 50k Beadchip and markers were edited for call rate and minor allele frequency.

Historical climate data were downloaded from the National Climatic Data Center (**NCDC**) at the National Oceanic and Atmospheric Administration (**NOAA**). Station-year-season summaries were created after quality control and merged to the HYS-DYD records using geographical coordinates approximated for each herd based on zip codes using the R (R Development Core Team, 2014) packages "zipcode" (Breen, 2012) and "geosphere" (Hijmans *et al*., 2012). The following variables were retained in order to characterize the climatic condition of each HYS class: average daily temperature, maximum daily temperature, minimum daily temperature, average daily relative humidity, average daily atmospheric pressure, average daily wind speed and sum of monthly rainfall.

Herd management profile information was extracted from Format 4 and consisted of the following variables: average number of heads in the herd in the year-season period, percentage of Holstein cows and number of milking times per day (two or three).

Statistical models

Data were analyzed with different models and variance components were estimated in order to assess the impact of each effect. The first model (G) accounted for only the genetic effect. The model was

$$y_{ijk} = \mu + g_i + e_{ijk}$$

where y_{ijk} is the HYS-DYD, μ is the overall mean, g_i is the additive genetic effect of the *i*-th bull, and e_{ijk} is random residual. Effects were so defined

$$e_{ijk} \sim N(0, \mathbf{I}\sigma^2_e)$$
$$g_i \sim N(0, \mathbf{G}\sigma^2_g)$$

where **G** is the genomic relationship matrix built on marker information (VanRaden, 2008).

The second (GL) and third (GLx) models accounted for the environmental effect defined by the covariates of latitude and longitude on the herd and its interaction with the genotype. Models were defined as

$$y_{ijk} = \mu + g_i + l_j + e_{ijk}$$

and

$$y_{ijk} = \mu + g_i + l_j + gl_{ij} + e_{ijk}$$

respectively, where l_j is the effect of latitude and longitude and gl_{ij} is their interaction with genotype, defined as

$$l_j \sim N(0, \mathbf{LL'}\sigma^2_l)$$
$$gl_{ij} \sim N(0, [\mathbf{G}\circ\mathbf{LL'}]\sigma^2_{gl})$$

where **L** is a matrix reporting latitude and longitude of the herds and ° indicates the Hadamard product of the two matrices.

The fourth (GW) and fifth (GWx) models accounted for the environmental effects defined by all climate covariates and their interaction with the genotype. Models were defined as

$$y_{ijk} = \mu + g_i + w_j + e_{ijk}$$

and

$$y_{ijk} = \mu + g_i + w_j + gw_{ij} + e_{ijk}$$

where w_j is the effect of climate variables and gw_{ij} is their interaction with genotype, defined as

$$w_j \sim N(0, \mathbf{WW'}\sigma^2_w)$$
$$gw_{ij} \sim N(0, [\mathbf{G}\circ\mathbf{WW'}]\sigma^2_{gw})$$

respectively, where **W** is a matrix reporting the climate variables.

The sixth (GM) and seventh (GMx) models accounted for the environmental effect defined by the herd management covariates and their interaction with the genotype. Models were defined as

$$y_{ijk} = \mu + g_i + m_j + e_{ijk}$$

and

$$y_{ijk} = \mu + g_i + m_j + gm_{ij} + e_{ijk}$$

respectively, where m_j is the effect of the management variables and gm_{ij} is their interaction with genotype, defined as

$$m_j \sim N(0, \mathbf{MM'}\sigma^2_m)$$
$$gm_{ij} \sim N(0, [\mathbf{G}\circ\mathbf{MM'}]\sigma^2_{gm})$$

where **M** is a matrix reporting the management variables.

The eighth (GH) and ninth (GHx) models accounted for the permanent environmental effect of the herd and its interaction with the genotype. Models were defined as

$$y_{ijk} = \mu + g_i + h_j + e_{ijk}$$

and

$$y_{ijk} = \mu + g_i + h_j + gh_{ij} + e_{ijk}$$

respectively, where h_j is the permanent environmental effect of the herd and gh_{ij} is the interaction with genotype, defined as

$$h_j \sim N(0, HH'\sigma^2_h)$$

$$gh_{ij} \sim N(0, [G°HH']\sigma^2_{gh})$$

where **H** is the matrix for the herd effect (103 levels).

Analyses were performed using the R package BGLR (Perez and de los Campos, 2014, Jarquin *et al*., 2014) that implements a Gibbs sampler over the eigenvalue decomposition of the environmental and genomic relationship matrices (Janss *et al*., 2012). A unique chain per model was run, which included a total of 62,000 iterations with 2,000 iterations discarded as burn-in and thinning every 10 iterations. Convergence of the models was assessed by visual inspection of trace plots and running post Gibbs analyses using the 'coda' R package (Plummer *et al*., 2006).

Cross-Validation

In order to evaluate the predictive ability of the models, a cross-validation scheme was designed. First, the data were assigned to five different macro-regions within the US: Mid-West, South-West, South-East, North-East I and North-East II. The last 2 regions were created because a large amount of data came from that region. Data were masked according to the three following criteria.

New bulls

Bulls that had at least 50 HYS-DYD were randomly assigned to 4 folds and masked successively, simulating bulls that were not progeny tested in the US. Models were re-run four more times, masking one fold for each run.

Incomplete progeny testing

Bulls that had HYS-DYD in at least 4 regions were selected and randomly assigned to 4 folds. One region per bull was masked, simulating bulls that were progeny tested but with daughter information missing for some region.

Missing region

Data from each of the five regions were alternatively masked. This aimed at simulating a validation set where some environmental conditions were not found in the training set.

Results and Discussion

Variance components

Descriptive statistics for HYS-DYD and environmental/management covariates are reported in table 1; variance components estimates are reported in table 2.

Table 1. Descriptive statistics for milk yield and covariates used in the study.

	Mean	**SD**
Milk yield HYS-DYD, kg	0.0	0.69
Maximum temperature, °C	19.5	9.0
Minimum temperature, °C	7.8	8.4
Average temperature, °C	13.8	8.7
Relative Humidity, %	65.9	7.0
Pressure, mmHg	744.9	21.9
Wind Speed, km/h	9.9	3.00
Rainfall, mm	251.0	103.2
Number of heads, n	940	654
Percentage Holstein cows	99.4	9.0
Milking times per day, n	2.76	0.4

Genomic effects accounted for 19.2% to 56.8% of total variance. The highest values were reached in model G when no environmental effects were fit, and the lowest values were reached when strong environmental and GxE effects were fit (GL and GLx). Geographical coordinates showed small impact (5.4% and 6.2% in GL and GLx) but their interaction with the additive genetic effect was strong (37.5% in GLx). Climate covariates also had a small impact (9.8% and 6.1% in GW and GWx), while their interaction was of larger magnitude (16.8% in GWx). Management variables also had a moderate effect (3.7% and 3.6% in GM and GMx), yet

the interaction with additive effects was strong (28.3% in GMx). Herd permanent environmental effects had the strongest effects (35.4% and 35.2% in GH and GHx) but null effect for the interaction with genotype (0.4% in GHx). The herd effect accounted for a large amount of phenotypic variance, confirming milk yield as a trait mainly driven by environmental conditions in general. The effect of latitude and longitude, which accounts for spatial permanent variation between herds, was moderate compared to the herd permanent environmental effect, suggesting that herd location explains little about the "environment" that it provides. Climate variables explained slightly more variance than geographical coordinates. The Management variables accounted for little amount of variance. The strongest GxE effect was shown with the geographical coordinates, followed by management and climate variables. Interaction between genotype and herd permanent effect was almost null.

The geographical, climate and management parameters we considered could not explain all the variability that exists between herds. There are probably other characteristics (peculiar of each herd) that could better describe the environmental variation. It was also observed that the impact of GxE was inversely proportional to the direct effect of the environment. For instance, the permanent environmental effect of the herd has a strong impact on the overall performance of cows in a given herd and there is no genetic control in the within-environment variation of the bulls' daughters. On the other hand, climate has moderate impact on the overall performance in a given HYS, but cows have different capability of coping with different weather conditions and this appears to be under genetic control. A further speculation could be that some environmental conditions accounted for by the models GH and GHx would reduce GxE, since the environments themselves tend to keep all cows at the same productive level with no expression of genetic potential to react to environmental conditions, e.g., farmers adapt diet to the genetic potential of the cow. In other words, considering herds as separate blocks (i.e. unique combinations of unknown characteristics) also accounts for the ability of the farmers to keep all cows at the expected productive level, while stratifying herds for their climatic conditions does not keep all cows at the expected productive level. Farmers are therefore more capable of managing cows individually once management parameters have been set, rather than providing climate conditions that meet specific cow requirement. Under these conditions, cows must express their genetic potential for coping with climatic stressors (e.g. summer heat), but not for different management practices across farms.

Predictive ability

The number of records masked in the CV scheme is reported in table 3, while predictive ability of the models is reported in table 4.

Table 3. Number of records masked and their proportion on total dataset for each split and fold used in the cross-validation (fold 5 for the 'missing region' split, containing 1,993 (17%) records, is not shown).

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**
New bulls	1,021 (9%)	1,039 (9%)	1,905 (16%)	1,373 (12%)
Incomplete progeny test	522 (5%)	575 (5%)	668 (6%)	663 (6%)
Missing region	2,730 (23%)	1,757 (15%)	3,064 (26%)	1,152 (10%)

When data from new incoming bulls were masked, the models that performed best were those including GL and GLx, with an advantage of the latter (accuracy of 0.293 vs. 0.264). The models including GH and GHx performed slightly worse than the others (0.24 and 0.23, respectively), followed by models incorporating GM and GMx (0.15 and 0.156) and GW and GWx (0.138 and 0.198). The G model performed worst (0.065), suggesting that environmental conditions are to be taken into account if their effect is not removed.

In the 'incomplete progeny test' CV scheme, models including GH and GHx performed best (0.275 for both models), followed by model with GL (0.252) and with GM (0.158), but their performance declined when GxE was included (0.192 and 0.087 GLx

AND GMx respectively). Models GW and GWx performed moderately (0.148 and 0.205) but this was the only case where GxE brought an advantage to the performance of the models. Again, G showed low performance (0.098).

Table 4. Accuracy (average of the correlation between predicted and observed values over the folds) of prediction for the different models over the different splits used in the cross-validation.

Model	New bulls	Incomplete progeny test	Missing region
G	0.065	0.098	0.124
GL	0.264	0.252	0.142
GLx	0.293	0.192	0.06
GW	0.138	0.148	0.097
GWx	0.198	0.205	0.106
GM	0.15	0.158	0.137
GMx	0.156	0.087	0.09
GH	0.24	0.275	0.106
GHx	0.23	0.275	0.097

In the 'missing region' CV scheme, again all models with GxE performed worse than their environment-only counterpart except for models with climate variables (0.097 and 0.106 for GW and GWx, 0.142 and 0.060 for GL and GLx, 0.137 and 0.090 for GM and GMx, 0.106 and 0.097 for GH and GHx). Best performance was therefore achieved with models GL, GM and G. The reason of the poor performance for the models that incorporate GxE could be that once the model is trained on a restricted part of the country, coefficients for this effect are specific for the regions used in the training set and do not work for a different region in the validation set. The exception of the climate variables could confirm this, because these variables provide a stronger link between the different regions of the country.

Conclusions

The inclusion of GxE in genomic prediction models can be advantageous, but careful investigation of the covariates used is needed. The permanent environmental effect of the herd shows strong effect but null interaction with genotype, and this interaction gives null or negative advantage when included in genomic prediction models. On the other hand, other covariates, such as geographical positioning and climate and management variables, show interaction with the genotype. All covariates bring an advantage in prediction only when their entire range is included in the training set for each bull. Therefore climate variables seem the most promising as they always confer an advantage when GxE is included in the model. Further research will consider different sets of covariates, different cross-validation schemes and new traits to be included in the analyses.

Acknowledgements

The authors wish to thank the Animal Genomics and Improvement Laboratory, USDA-ARS for providing data and genotypes.

References

Bohmanova, J., Misztal, I., Tsuruta, S., Norman, H.D. & Lawlor, T.J. 2008. Short communication: Genotype by environment interaction due to heat stress. *Journal of Dairy Science 91:2,* 840-846.

Breen, J. 2012. zipcode: U.S. ZIP Code database for geocoding.

Dikmen, S., Cole. J.B., Null, D.J. & Hansen, P.J. 2013. Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. *PLoS ONE 8: e69202.*

Haile-Mariam, M., Carrick, M.J. & Goddard, M.E. 2008. Genotype by environment interaction for fertility, survival, and milk production traits in Australian dairy cattle. *Journal of Dairy Science 91:12,* 4840-4853.

Hayes, B.J., Bowman, P.J., Chamberlain, A. J., Savin, K., Van Tassell, C.P., Sonstegard, T. S. & Goddard, M.E. 2009. A validated genome wide association study to breed cattle adapted to an environment altered by climate change. *PLoS One 4:8,* e6676.

Hijmans, R.J., Williams, E. & Vennes, C. 2012. geosphere: Spherical Trigonometry.

Janss, L., de Los Campos, G., Sheehan, N. & Sorensen, D. 2012. Inferences from genomic models in stratified populations. *Genetics 192:2,* 693-704.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J. & de los Campos, G. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics 127:3*, 595-607.

Kassen, R. 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. *Journal of Evolutionary Biology 15*, 173-190.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.L. & de los Campos, G. 2015. Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker× Environment Interaction Genomic Selection Model. *G3: Genes| Genomes| Genetics*, g3-114.

Norberg, E., Madsen, P., Su, G., Pryce, J.E., Jensen, J. & Kargo, M. 2014. Short communication: Heterosis by environment and genotype by environment interactions for protein yield in Danish Jerseys. *Journal of Dairy Science 97:7,* 4557-4561.

Oseni, S., Misztal, I., Tsuruta, S. & Rekaya, R. 2004. Genetic components of days open under heat stress. *Journal of Dairy Science 87:9,* 3022-3028.

Plummer, M., Best. N., Cowles, K. & Vines, K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News 6,* 7-11.

Ravagnolo, O. & Misztal, I. 2000. Genetic component of heat stress in dairy cattle, parameter estimation. *Journal of Dairy Science 83:9,* 2126-2130.

Streit, M., Reinhardt, F., Thaller, G. & Bennewitz, J. 2013. Genome-wide association analysis to identify genotype by environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins. *Journal of Dairy Science 96:11,* 7318-7324.

Su, G., Madsen, P. & Lund, M.S. 2009. Reaction norm model with unknown environmental covariate to analyze heterosis by environment interaction. *Journal of Dairy Science 92:5*, 2204-2213.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science 91,* 4414–4423.

Windig, J.J., Mulder, H.A., Bohthe-Wilhelmus, D.I. & Veerkamp, R.F. 2011. Simultaneous estimation of genotype by environment interaction accounting for discrete and continuous environmental descriptors in Irish dairy cattle. *Journal of Dairy Science 94:6*, 3137-3147.

Table 2. Posterior mean of the ratio of variance absorbed by each effect over the total phenotypic variance for each model tested on the entire dataset (11,747 herd-year-season-DYDs).

Model	G	L	GL	W	GW	M	GM	H	GH
G	56.8	.	.	.	.	.	.	.	.
GL	52.1	5.4	.	.	.	.	.	.	.
GLx	19.2	6.2	37.6	.	.	.	.	.	.
GW	52.6	.	.	9.8	.	.	.	.	.
GWx	38.2	.	.	6.1	16.8	.	.	.	.
GM	44.9	.	.	.	.	3.7	.	.	.
GMx	27.7	.	.	.	.	3.6	28.3	.	.
GH	37.2	.	.	.	.	.	.	35.4	.
GHx	37.2	.	.	.	.	.	.	35.2	0.4