Clustering Herds Across Country Borders for International Genetic Evaluation

K. A. Weigel and R. Rekaya
Department of Dairy Science
University of Wisconsin, Madison 53706, USA

Abstract

International dairy sire evaluations are calculated in a two-step process. Lactation records from each country are used to predict national EBV, and these are transformed to the base, scale, and units of measurement of other countries using conversion equations or the MACE procedure. A major limitation of this approach is that traits are defined according to country borders. Herds located in small, neighboring countries may be more similar in management, climate, and genetic background than herds located far apart within a single large country. This study proposes international genetic evaluation using herd clusters. Data were 4.6 million lactation records from 46,000 herds in Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, Israel, Switzerland, and five regions of the US (Midwest, Northeast, Northwest, Southeast, and Southwest). Herds were grouped into clusters based on thirteen descriptive variables: herd size, calving interval, milking frequency, age at first calving, milk yield, month of calving, sire's PTA milk, sire's percent North American genes, latitude, altitude, temperature, rainfall, and percentage of arable land used for pasture. Five clusters were formed, and each cluster contained herds from 5-11 countries or regions. Genetic correlations between clusters were 0.81-0.97. The herd cluster model is intuitively appealing, because an animal's genetic merit is predicted for each unique environment or management system regardless of country borders, parsimonious (the number of traits was reduced from 13 to 5), and computationally feasible.

Introduction

Genetic evaluations of dairy sires are calculated by national organizations operating within country borders. Therefore, direct comparison of sire EBV between countries is not possible, due differences in the genetic base, scale, and units of measurement for each country. For many years, conversion equations were used to transform EBV from an exporting country to the base, scale, and units of an importing country. Equations were developed for each pair of countries using data from foreign bulls with imported semen (3, 11) or sets of full-sibs with progeny in both countries (6). More recently, the MACE procedure (2, 9) has become the method of choice for international dairy sire comparisons. The MACE procedure incorporates more data, because all bulls that are progeny tested in each country can be included, and it is operationally efficient, because data from all participating countries can be analyzed simultaneously.

Incorporation of foreign data via conversion equations or MACE has increased the efficiency of sire selection (7), but realized gains in accuracy of MACE EBV relative to conversion have been extremely small (10). International evaluation using daughter lactation records has been suggested (5), but progress in this area has been slow. Application of a single-trait model across countries precludes the possibility of genotype by environment interaction. On the other hand, a multiple-trait model would be computationally challenging for a large number of countries. Inadequate genetic links between countries can lead to erroneous covariance estimates, and this may negate any advantages one might otherwise achieve by using lactation records. Rekaya et al. (8) proposed a method for using management, climate, and genetic information to increase the precision of genetic parameter estimates in a structural model for covariances. The

next logical step is to consider the possibility of "borderless" genetic evaluations (5), in which traits or environments are not based on country boundaries, but rather on existing knowledge regarding the management, climate, and genetics of each herd.

The objectives of the present study were twofold. First, to develop a model for international dairy sire evaluation based on clustering herds across country borders using information about the climate, management, and genetic composition of each herd. Second, to demonstrate the intuitive appeal and computational feasibility of this model by applying it to a large international data set containing lactation records generated under a wide variety of production conditions.

Materials and Methods

Data included first lactation records of Holstein-sired cows in Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, Israel, Switzerland, and five regions of the US. The five regions and corresponding states were: Northeast = Pennsylvania; Midwest = Michigan; Northwest = Oregon and Washington; Southeast = Florida and Georgia, and Southwest = Arizona and New Mexico. Data were included for cows with first calving in 1979-1998 for Switzerland, 1982-1998 for Denmark, 1985-1998 for Israel, 1990-1998 for Estonia, and 1983-1998 for all other countries or regions. Only cows sired by AI bulls listed in the Interbull Holstein pedigree file were included. Additional data edits required lactation days in milk between 275-375 and dam's percentage of Holstein or Friesian genes (if known) ≥ 50%.

Prior to genetic evaluation, herds from the thirteen countries or regions were grouped into clusters without regard to country borders. Thirteen descriptive variables related to management, climate, and genetic composition were used to cluster herds. Six variables related to herd management were considered: cows per herd = number of cows born from 1990-1995 (after data editing); calving interval = mean interval between first and second calving; milking frequency = mean number of milkings per day; age at calving = mean age of cows at first calving; lactation milk yield = mean first lactation yield, adjusted for age, days in milk, and milking frequency, and month of calving = mean month of first calving (an indicator of seasonal production). Two variables related to genetic composition of the cow population were used: sire PTA milk = weighted mean PTA milk on the US scale from the February 1999 Interbull evaluation for sires of cows in the herd, and sire's percent North American genes = weighted mean percentage of genes tracing back to the North American Holstein population for sires of cows in the herd. Five variables related to climate and land use were considered, and these variables were identical for all herds from a given country or region: latitude = mean latitude of major cities in the country or region; altitude = mean altitude of major cities in the country or region; July temperature = mean daily temperature in July for major cities in the country or region; July rainfall = mean rainfall total in July for major cities in the country or region, and land used as pasture = ratio of land used as pasture relative to total arable land.

Many of the descriptive herd variables were correlated (e.g., latitude and temperature). Therefore, these variables were standardized to zero mean and unit variance, and then transformed into principal components (4). The principal component analysis gave nine components with eigenvalues ≥ 0.50, and these were retained for the subsequent cluster analysis.

The cluster analysis was performed using the nearest centroid sorting method (1). This iterative method creates clusters by minimizing the sum of squared distances from the cluster means. For each herd, the standardized descriptive herd variables were combined into nine principal components using weights (eigenvectors) calculated in the principal component analysis. The cluster analysis was based on values of the nine principal components for each herd. Clusters of fewer than 300 herds were deleted, and at convergence, five clusters remained.

The genetic covariance matrix for milk yield in the five clusters was estimated using

a Bayesian implementation of a multiple-trait sire model via Gibbs sampling. A single long chain of 100,000 samples was employed, and the first 20,000 samples were discarded as a burn-in period. Weakly informative normal prior distributions were assumed for systematic effects and breeding values, and bounded uniform priors were assumed for covariance components. The following model was used for covariance estimation; all factors were nested within cluster:

$$y_{ijklmno} = \text{herd-year}_i + \text{season}_j + \text{age}_k + \text{frequency}_l + \beta * \text{DIM}_m + \text{sire}_n + \text{error}_{ijklmno}$$

where:

$y_{ijklmno}$ = first lactation milk yield,
herd-year_i = interaction of herd and year of calving,
season_j = season of calving (3-month seasons were used),
age_k = age of cow at calving,
frequency_l = number of times milked per day,
β = regression coefficient,
DIM_m = days in milk,
sire_n = sire of cow, and
$\text{error}_{ijklmno}$ = random error.

Only data from sires with ≥ 4 progeny were used in variance component estimation. After this edit was applied, data from 3,043,432 daughters of 38,619 sires remained for analysis.

Results and Discussion

A summary of the data is shown in Table 1. Data were requested from more than twenty countries, but some countries did not particiate due to commercial concerns or time constraints. In future studies, it would be desirable to obtain data from additional countries that have a high genetic level (e.g., France or Netherlands) and additional countries that differ widely in herd management practices (e.g., Australia or New Zealand).

In Tables 2A and 2B, means of the thirteen descriptive herd variables are shown for each country or region. The number of first lactation cows calving from 1990-1995 (after editing) spanned a wide range, from 17 in Finland to 1910 in the Southwest US. Pre-editing herd size would have been preferable, but some countries removed non AI-sired cows before sending the data. Calving interval was shortest in Finland and Denmark (cool climate) and longest in the Southeast US (hot and humid climate). Nearly all herds milked twice daily, except those in Israel and the US. Age at calving was youngest for Israel and Finland and oldest for Austria and Estonia. First lactation milk yield ranged from approximately 4000 kg in Estonia and Czech Republic to roughly 9000 kg in Israel and the Northwest US. Minimal variation was observed in month of calving; this variable would have been more useful if data had been available for countries like New Zealand and Ireland, where management intensive grazing is popular. The last five variables (latitude, altitude, July temperature, July rainfall, and land used as pasture) were fixed for all herds in a given country or region. It is likely that variation exists within each country or region regarding variables like altitude, temperature, and rainfall, so use of a single value for each country or region is an approximation. In addition, values shown in Table 2 were from meteorological data for major cities, and these cities may be located far from the major dairy production areas. Lastly, the proportion of arable land used for pasture provided only a rough estimate of the importance of grazing versus confinement in each country or region.

Eigenvalues of the correlation matrix for descriptive herd variables ranged from 0.003 to 0.293. Some variables, such as latitude and daily temperature, were highly correlated, and principal component analysis was used to develop unique contrasts of these variables. Nine eigenvalues were larger than the pre-selected minimum of 0.50, and these eigenvalues explained more than 93% of the variation in the descriptive herd variables. Eigenvectors (weights) for the standardized descriptive herd variables were calculated for each principal component. These weights were used to calculate values of the nine

principal components for each herd. Subsequently, a cluster analysis based on values of the nine principal components resulted in five herd clusters, each containing data from 700-26,000 herds.

Table 3 shows means of the descriptive herd variables for each of the five clusters, and Table 4 shows the number of herds from each country or region in each cluster. Cluster 1 consisted primarily of medium-sized herds in the Midwest, Northeast, and Southeast US that milked twice daily, had a high genetic level, and had relatively high summer temperature and rainfall. Cluster 2 contained mainly large herds from Israel and several of the US regions; these herds were characterized by early age at first calving, high milk yield and sire PTA milk, high temperature, and three times daily milking. Cluster 3 was primarily made up of small herds in Austria, Czech Republic, Estonia, and Switzerland; these herds typically had late age at first calving, low milk yield, low sire PTA milk, high pasture usage, and a very favorable climate. Cluster 4 largely consisted of herds in Israel and the Northwest and Southwest US; these herds had twice daily milking, high milk yield and PTA milk, and high daily temperatures with little rainfall. Cluster 5 consisted primarily of small herds in Belgium, Denmark and Finland; these herds had twice daily milking, a short calving interval, intermediate production, and a favorable climate.

Genetic parameters for milk production in the five herd clusters are shown in Table 5. Heritability of milk yield ranged from 0.28-0.39. Genetic correlations between clusters 1, 2, and 4 were extremely high (0.96-0.97). These clusters consisted primarily of herds from the US and Israel with high milk production and a high genetic level. These clusters also shared numerous genetic ties (due to a large number of US herds in each cluster). Correlations between these regions and regions 3 and 5, which contained mainly herds from Europe, ranged from 0.81-0.89. Lastly, the estimated genetic correlation between clusters 3 and 5 was 0.89. As shown by the 95% highest posterior density regions for the estimated genetic parameters, precision of estimates depended on the number of observations per cluster and on genetic ties between clusters.

Conclusions

A multiple-trait herd cluster model is proposed for international genetic evaluation. Information regarding management, climate, and genetic composition of each herd is used to form herd clusters, regardless of country borders, and each of these clusters represents a separate trait in a multiple-trait BLUP analysis. Herd clusters were not developed directly from the descriptive herd variables, but rather from their corresponding principal components, due to concerns about correlations between descriptive variables. Further research is needed to identify and prioritize variables that can describe the genetics, management, and climate of each herd. The model is flexible, in terms of the number and type of variables that can be included.

Lack of parsimony, i.e., too many genetic parameters, is a major problem in current international dairy sire evaluations. Using the herd cluster model, the number of traits was reduced from 13 to 5, and the number of covariance parameters was reduced from 91 to 15. Precision of estimated genetic parameters depended more heavily on the level of genetic ties between clusters than the number of observations per cluster. For example, the estimated genetic correlation between one pair of clusters that contained more than 1 million cows still had a 95% highest posterior density region of 0.76-0.91. The problem of genetic ties between traits is reduced in the herd cluster model, but there is still a lack of ties between herds within a cluster. This problem can only be solved by continued diffusion of semen across countries.

The multiple-trait herd cluster model is more intuitively appealing than a multiple-trait model based on country boundaries. Clearly herds located in small, neighboring countries like Belgium and The Netherlands or Austria and Switzerland share many more similarities in management, climate, and genetic composition than herds located in the extremes of a large country like the US. With the herd cluster model, herds are grouped

based on likeness, rather than location. This model may lead to greater reliability and credibility of international genetic evaluations, because young bulls tested in a certain climate and management system will have their performance estimated precisely for other herds with similar conditions. Under this model, different sire EBV are appropriate for different herds within a country (i.e., no national ranking list), and distribution of results could be facilitated by using Internet or the Web. Lastly, this study showed that the multiple-trait herd cluster model is computationally feasible for large data sets.

Acknowledgements

Data were generously provided by the Federation of Austrian Cattle Breeders, Vlaamse Rundveeteelt Vereniging, the Czech-Moravian Breeders Corporation, the Danish Agricultural Advisory Centre, the Estonian Animal Recording Centre, the Finnish Animal Breeding Association, A. R. O. The Volcani Center, the Holstein Association of Switzerland, the Interbull Centre, and USDA Animal Improvement Programs Laboratory. Financial support was provided by the National Association of Animal Breeders and World-Wide Sires, Inc.

References

1 Anderberg, M.R. 1973. Cluster Analysis for Applications. Academic Press, New York, NY.

2 Banos, G. & Sigurdsson. A. 1996. Application of contemporary methods for the use of international data in national genetic evaluations. J. Dairy Sci. 79, 1117-1125.

3 Goddard, M.E. 1985. A method of comparing sires evaluated in different countries. Livest. Prod. Sci. 13, 321-331.

4 Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psych. 24, 417-441.

5 Lohuis, M.M. & Dekkers, J.C.M. 1998. Merits of borderless evaluations. Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia XXVI, 169-172.

6 Mattalia, S. & Bonaiti, B. 1993. Use of full sib families to estimate the 'a' coefficients of conversion formulas between countries. Pages 73-74 *in* Proc. Interbull Annu. Mtg., Aarhus, Denmark. Bull. No. 8, Interbull, Uppsala, Sweden.

7 Powell, R.L. & Norman, H.D. 1998. Use of multinational data to improve national evaluations of Holstein bulls. J. Dairy Sci. 81, 2257-2263.

8 Rekaya, R., Weigel, K.A. & Gianola, D. 1999. Bayesian estimation of parameters of a structural model for genetic covariances between milk yield in five regions of the USA. Proc. 50th Annu. Mtg. EAAP, Zurich, Switzerland, August 23-26.

9 Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. J. Dairy Sci. 77, 26712678.

10 Weigel, K.A. & Powell, R.L. 1999. Retrospective analysis of the accuracy of conversion equations and multiple-trait across country evaluations of Holstein bulls used internationally. J. Dairy Sci. (submitted).

11 Wilmink, J.B.M., Meijering, A. & Engel, B. 1986. Conversion of breeding values for foreign populations. Livest. Prod. Sci. 14, 223-229.

Table 1. Summary of data used in the present study.

Country or Region	No. Herds	No. Cows
Austria	1698	43,011
Belgium	3365	199,534
Czech Republic	5688	439,267
Denmark	11,176	1,698,394
Estonia	1111	102,331
Finland	11,343	174,590
Israel	975	248,032
Switzerland	2935	186,700
US – Midwest	1392	327,638
US – Northeast	5172	714,607
US – Northwest	669	282,675
US – Southeast	328	102,137
US – Southwest	84	108,269
Total	45,936	4,627,185

Table 2A. Summary of descriptive variables used to classify herds into clusters

	Country or Region							
Descriptive Herd Variable	AUT	BEL	CSK	DNK	EST	FIN	ISR	CHE
Cows per herd (1990-1995)	21	61	204	134	338	17	472	36
Calving interval (mo)	13.5	13.1	13.1	12.8	13.2	12.6	13.0	13.0
Milking frequency	2.00	2.00	2.00	2.00	2.00	2.00	2.76	2.00
Age at calving (mo)	30.7	27.9	27.8	28.3	30.6	25.3	24.2	29.8
Lactation milk yield (kg)	5497	6556	4316	6414	3943	5842	9239	5584
Month of calving	6.98	7.60	6.45	7.21	6.22	6.91	7.10	7.26
Sire PTA milk (kg)	-165	110	-166	-15	-335	-163	296	-174
Sire's % N. American genes	79	91	62	76	38	35	32	85
Latitude (°N)	47	50	49	56	59	64	31	47
Altitude (m)	390	50	200	30	10	10	270	490
July temperature (°C)	18	17	18	16	17	17	26	18
July rainfall (mm)	115	80	70	60	65	70	0	120
Land used as pasture (%)	57	46	21	12	21	4	25	73

Table 2B. Summary of descriptive variables used to classify herds into clusters

	Country or Region				
Descriptive Herd Variable	MW	NE	NW	SE	SW
Cows per herd (1990-1995)	344	144	642	826	1910
Calving interval (mo)	13.2	13.3	13.3	13.8	13.2
Milking frequency	2.17	2.04	2.32	2.39	2.36
Age at calving (mo)	26.9	26.8	26.4	27.2	26.4
Lactation milk yield (kg)	7926	7413	8703	7337	8325
Month of calving	6.54	6.77	6.43	6.91	6.40
Sire PTA milk (kg)	368	332	370	350	360
Sire's % N. American genes	100	100	100	100	100
Latitude (°N)	43	40	47	30	34
Altitude (m)	220	170	350	100	330
July temperature (°C)	22	24	18	27	33
July rainfall (mm)	70	105	15	185	20
Land used as pasture (%)	6	14	11	21	15

Table 3. Means of descriptive herd variables corresponding to each herd cluster.

	Herd Cluster				
Descriptive Herd Variable	1	2	3	4	5
Cows per herd (1990-1995)	86	373	45	113	44
Calving interval (mo)	13.3	13.3	13.2	13.1	12.8
Milking frequency	2.02	2.86	2.00	2.01	2.00
Age at calving (mo)	27.1	25.6	29.3	25.8	26.9
Lactation milk yield (kg)	7183	8799	4682	8223	6099
Month of calving	6.69	6.72	6.67	6.66	7.10
Sire PTA milk (kg)	320	315	-205	329	-86
Sire % N. American genes	100	78	67	62	59
Latitude (°N)	40	38	49	38	59
Altitude (m)	177	243	283	304	26
July temperature (°C)	23.7	24.1	17.9	22.9	16.6
July rainfall (mm)	101	53	90	7	67
Land used as pasture (%)	13	17	41	19	13

Table 4. Number herds from each country or region corresponding to each herd cluster.

	Herd Cluster					
Country or Region	1	2	3	4	5	
Austria	0	4	1693	0	1	
Belgium	0	0	454	0	2911	
Czech Republic	14	0	5346	0	328	
Denmark	0	0	26	0	11,150	
Estonia	0	0	818	0	293	
Finland	0	0	31	0	11,312	
Israel	0	231	0	744	0	
Switzerland	0	0	2935	0	0	
US – Midwest	1255	121	7	0	9	
US – Northeast	5029	141	2	0	0	
US – Northwest	0	115	3	535	16	
US – Southeast	263	65	0	0	0	
US – Southwest	7	39	1	37	0	
Total herds	6568	716	11,316	1315	26,020	
Total cows (x1000)	930.6	234.6	101.9	179.0	1597.3	
Total sires	15,685	6093	2245	3992	17,888	

Table 5. Heritabilities and genetic correlations for milk yield between clusters (95% highest posterior density interval in italic print)

	Herd Cluster				
Herd Cluster	1	2	3	4	5
1	0.29	0.97	0.84	0.97	0.86
	0.28-0.31	*0.96-0.98*	*0.76-0.91*	*0.96-0.98*	*0.83-0.90*
2		0.28	0.81	0.96	0.87
		0.26-0.29	*0.72-0.88*	*0.94-0.97*	*0.83-0.91*
3			0.39	0.87	0.89
			0.35-0.43	*0.79-0.93*	*0.83-0.94*
4				0.31	0.89
				0.29-0.33	*0.86-0.93*
5					0.35
					0.34-0.37