Legendre Polynomials versus Linear Splines in the Canadian Test-Day Model

J. Bohmanova[1], J. Jamrozik[1], F. Miglior[23], I. Misztal[4], P.G. Sullivan[3]
[1]*CGIL, University of Guelph, Guelph, ON*
[2]*Agriculture and Agri-Food Canada, Sherbrooke, QC*
[3]*Canadian Dairy Network, Guelph, ON*
[4]*University of Georgia, Athens, GA, USA*

Introduction

The aim of this study was to identify the most appropriate function for modeling fixed and random regressions in the Canadian Test Day Model (**CTDM**) using test-day (**TD**) records up to 365 days in milk (**DIM**).

In the current CTDM only TD records recorded before 305 DIM are considered. Legendre polynomials of order four are fitted for both random and fixed regressions. High additive genetic variances at the extremes of lactation and negative correlations between the most distant test-days have been reported in random regression models (**RRM**) based on lactation curves and Legendre polynomials. The overestimation of variances at the edges of lactation is often explained by mathematical characteristics of polynomials. Splines have been recently advocated as a good alternative to Legendre polynomials (White *et al.,* 1999; Druet et. al., 2003; Silvestre *et al.,* 2006), mainly due to their limited sensitivity to the data (records influence only parts of function in their closeness) and direct interpretation of parameters (Misztal, 2006).

This study compared four RRM based on either Legendre polynomials or linear splines, using a broad range of model comparison criteria.

Material and Methods

Data

Variance components were estimated using Canadian Holstein data (**VC**) created by a random sampling of 50 herds (with >50 cows) from dataset used for routine genetic evaluation in August 2006. The data consisted of 96,756 TD milk, fat and protein yields, and somatic cell score (**SCS**) from the first three lactations recorded from 1988 to 2006. Only TD records with all traits present on a TD and DIM ≤365d were included. The pedigree file contained 18,178 animals.

In order to compare stability of estimated breeding values (**EBV**) between runs, two datasets based on test-day records recorded before August 2006 (**D06**) and August 2001 (**D01**) were used. Description of the data is given in Table 5.

Models

Four random regression multi-trait, multi-lactation models were compared. The general formula for all models was as follows:

$$y_{ijkmnt} = HTD_{jkn} + \sum_{l=1}^{q} \alpha_{jlmn} z_l(d) + \sum_{l=1}^{q} \beta_{ijln} z_l(d) + \sum_{l=1}^{q} \gamma_{ijln} z_l(d) + e_{ijkmnt}$$

All models included a fixed effect of herd-test-date (**HTD**), a fixed regression on DIM nested within age-season-region of calving class (α) and random regressions for additive genetic (β) and permanent environmental (γ) effects. Two seasons of calving and five regions within Canada were defined.

The fixed and random regressions were fitted either with Legendre polynomials of order four (**LEG**) or with linear splines with either four (**SPL4**), five (**SPL5**) or six (**SPL6**) knots. The location of knots is given in Table 1. Twelve classes of 30 DIM for residual variance were defined for every lactation.

Table 1. Description of random regression models.

Model	Function[§]	q[‡]	Position of knots
LEG	Legendre	5	-
SPL4	Splines	4	[5 65 245 365]
SPL5	Splines	5	[5 65 125 245 365]
SPL6	Splines	6	[5 65 125 245 305 365]

[§] - type of regression function
[‡] - order of polynomial in models with Legendre polynomials or number of knots in models with linear splines.

A Bayesian approach via Gibbs sampling was carried out in order to estimate model parameters. A single long chain of 100,000 samples was generated. The first 20,000 samples were discarded as a burn-in, and the remaining samples were used to compute posterior means of model parameters. Convergence of Gibbs chains was monitored by visual inspections of plots of samples.

Model comparison

The competing RRM were compared using the Deviance Information Criterion (**DIC**) defined by Spiegelhalter *et al.* (2002) as:

$$DIC = \bar{D} + p_D ,$$

where $\bar{D}$ is the posterior expectation of the Bayesian deviance (measure of the fit of the model), p_D is the effective number of parameters (penalty for increasing model complexity). The model with the smallest DIC is preferable.

Two genetic evaluations were carried for four RRM using D06 and D01 data with variance components previously estimated from the VC data. Mixed model equations were solved by iteration on data with Preconditioned Conjugate Gradient algorithm and a block diagonal preconditioner. Convergence criterion was defined as average relative difference between left and right hand side and was required to be less than 9.9×10^{-13}.

Goodness of fit of all models was investigated by computing percentage of squared bias (**PSB**), correlation between observed and predicted records (**RHO**) and residual variance (**RV**) using D06 data.

The PSB (Ali and Schaeffer, 1987) for j[th] trait and n[th] lactation is defined as:

$$PSB_{jn} = \frac{\sum_{r=1}^{o} (y_{jnr} - \hat{y}_{jnr})^2}{\sum_{r=1}^{o} (y_{jnr})^2},$$

where y_{jnr} is the r^{th} observed record of j^{th} trait and n^{th} lactation, $\hat{y}_{jnr}$ is the r^{th} predicted record of j^{th} trait and n^{th} lactation and o is the number of records.

Stability of EBV of competing models was compared using an error of prediction (**ERP**) defined by Sullivan *et al.* (2005) as:

$$ERP = \sqrt{\frac{\sum_{i=1}^{n} (ebv06_i - pa01_i)^2}{n}},$$

where *ebv06* is EBV calculated from D06, *pa01* is parent average predicted from D01 and *n* is number of bulls with no daughters in D01 and at least 25 daughters in D06. Prior to computing the statistic, EBV from D06 were shifted by subtracting the average change in EBV from D01 to D06 for a set of bulls whose average EBV was not expected to change. The adjustment was based on 1,929 bulls with at least 25 daughters in D01, no new daughters and no more than 10 new granddaughters between D01 and D06.

Results

Posterior means of daily additive genetic variance for milk, fat, protein yield and SCS (Figure 2, 3, 4 and 5, respectively) of both Legendre and spline RRM increased with parity. Variances for LEG had a typical U-shape (highest variance at the beginning and at the end of the lactation and relatively constant trend in the mid-lactation). Variance function of linear spline models followed a parabolic shape among knots. The overall trend of the variance curve depends on number of knots and on correlations between knots. The higher the correlations the smoother curves were obtained.

The models with splines had smaller variances at extremes of lactation than the LEG model in production traits at second and third lactations. Variances of SCS from LEG model was higher along the whole lactation compared to models with splines.

In all models, residual variance was the highest at the beginning of lactation and gradually decreased with DIM (Figure 6). Smaller residual variance at the end of lactation was observed in LEG and SPL6 compared to SPL4 and SPL6.

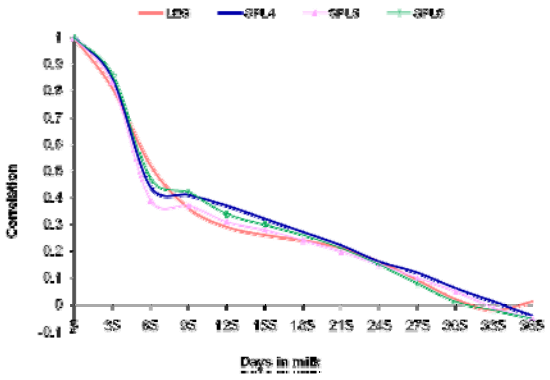
Although the pattern of daily heritabilities (Figure 7 to 10) was slightly different between models, posterior means of average daily heritabilities were similar across models (Table 2).

Table 2. Posterior mean estimates of average daily heritabilities for first lactation.

Model	Milk	Fat	Protein	SCS
LEG	0.44	0.34	0.41	0.21
SPL4	0.44	0.34	0.41	0.18
SPL5	0.45	0.35	0.42	0.19
SPL6	0.45	0.36	0.42	0.19

Genetic correlations between day 5 and the rest of lactation followed the same trend in all models, i.e. high to moderate correlations at the beginning of lactation and negative correlations at the end of lactation (Figure 1).

Figure 1. Genetic correlations between 5 DIM and the rest of lactation for milk yield in the first parity.



The best model based on DIC was the most complex model (SPL6). Both LEG and SPL5 provided similar DIC. The SPL4 ranked the last (Table 3).

Table 3. Estimates of Deviance Information Criterion (DIC) and rank of models (Rank).

Model	DIC	Rank
LEG	255,808	2
SPL4	274,532	4
SPL5	258,924	3
SPL6	236,646	1

Total CPU time and number of rounds required for running genetic evaluation with D06 is given in Table 4. The LEG model converged in the shortest total CPU time. Convergence rate of this model was better than convergence rate of SPL4 which has lower number of parameter than LEG. Slower convergence of RRM with linear splines compared to RRM with Legendre polynomials can be explained by higher correlations between knots compared to correlations between Legendre coefficients. This weakness of models with splines can be overcome by diagonalization of covariance matrix of random regression coefficients.

Table 4. Number of iterations, CPU time per round of iterations and total CPU time needed for predictions of EBV with D06 data set.

Model	Number of iterations	CPU time per iteration*	Total CPU time
LEG	667	1,076	8d 7h
SPL4	848	1,020	10d 0h
SPL5	804	1,345	12d 12h
SPL6	911	1,685	17d 18h

* 2.40 GHz processor

Similar values of PSB, RHO and RV were found in all competing models (Table 6). However, models with splines had the better goodness of fit in all traits and lactations than the LEG.

As shown in Table 6, the SPL6 model gave the smallest ERP from all of models. The difference between models were very small in production traits at first lactation but were significantly higher at later lactations and in SCS at all three lactations, where all models with splines had smaller ERP than the LEG model. The better performance of models with splines can be explained by smaller overestimation of additive genetic variance compared to the LEG model.

Conclusions

Both RRM with linear splines as well as RRM based on Legendre polynomials tended to overestimate additive genetic variances at extremes of lactation. However, this overestimation was smaller in models with splines. Similar goodness of fit was provided by both groups of models. Models with splines had more stable EBV. This fact was especially apparent in production traits at second and third lactations and in SCS at all three lactations. The model with six knots performed the best in all the statistical criteria used for model comparison. The drawback of this model was a slow convergence which was caused by high correlations between regression coefficients and higher number of parameters.

Acknowledgement

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: http://www.sharcnet.ca). Funding was provided by DairyGen Council of Canadian Dairy Network and NSERC of Canada.

References

Ali, T.E. & Schaeffer, L.R. 1987. Accounting for covariances among test day milk yields in dairy cows. *Can. J. Anim. Sci. 67,* 637–644.

Misztal, I. 2006. Properties of random regression models using linear splines. *J. Anim. Breed. Genet. 123,* 74-80.

Silvestre, A.M., Petim-Batista, F. & Colaço, J. 2005. Genetic parameter estimates of Portuguese dairy cows for milk, fat, and protein using a spline Test-day model. *J. Dairy Sci. 88,* 1225-1230.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64,* 583–639.

Sullivan, P.G., Wilton, J.W., Schaeffer, L.R., Jansen, G.J., Robinson, J.A.B. & Allen, O.B. 2005. Genetic evaluation strategies for multiple traits and countries. *Livest. Prod. Sci. 92(3),* 195-205.

White, I.M.S., Thompson, R. & Brotherstone, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *J. Dairy Sci. 82,* 632–638.

Table 5. Description of data sets.

	VC		D06		D01	
Number of TD records	96,756		45,120,202		26,832,479	
Number of cows	6,094		2,650,096		1,564,228	
Number of TD records per cow	16		17		17.1	
Number of HTD classes	3,915		3,593,917		2,503,244	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
DIM	161	95	161	95	162	95
Milk yield (kg, kg^2)	28.8	9.2	27.8	8.8	26.1	8.5
Fat yield (kg, kg^2)	1.04	0.33	1.02	0.34	0.96	0.32
Protein yield (kg, kg^2)	0.93	0.36	0.89	0.26	0.84	0.25
SCS	2.5	1.7	2.2	2.0	2.1	1.9

Table 6. Percentage of squared bias (PSB), correlation between observed and predicted data (RHO), residual variance (RV) and error of prediction (ERP) of models for milk yield and somatic cell score (SCS).

Trait	Model	Lactation 1				Lactation 2				Lactation 3			
		PSB	**RHO**	**RV**	**ERP***	**PSB**	**RHO**	**RV**	**ERP**	**PSB**	**RHO**	**RV**	**ERP**
Milk	**LEG**	12.5	0.86	13.8	705	11.6	0.90	18.9	834	10.5	0.91	20.6	782
	SPL4	12.2	0.83	15.8	701	11.6	0.89	21.4	814	10.3	0.89	24.0	765
	SPL5	12.6	0.86	13.7	709	12.0	0.91	18.5	825	10.8	0.91	20.0	773
	SPL6	12.1	0.86	13.3	699	11.4	0.91	18.2	801	10.2	0.91	19.7	759
SCS	**LEG**	22.5	0.82	1.22	0.43	23.8	0.86	1.35	0.48	19.9	0.85	1.32	0.58
	SPL4	24.1	0.80	1.30	0.42	24.5	0.85	1.41	0.47	20.9	0.84	1.42	0.56
	SPL5	22.5	0.82	1.21	0.41	22.6	0.87	1.33	0.47	19.1	0.86	1.31	0.55
	SPL6	22.4	0.82	1.19	0.40	23.2	0.87	1.31	0.45	19.5	0.86	1.29	0.52

*Error of prediction of 305 day breeding values for milk yield (Milk) and average daily breeding values for SCS of 1,984 sires with no daughters in D01 and at least 25 daughters in D06

Figure 2. Posterior mean estimates of additive genetic variance of daily milk yield in first, second and third lactation.

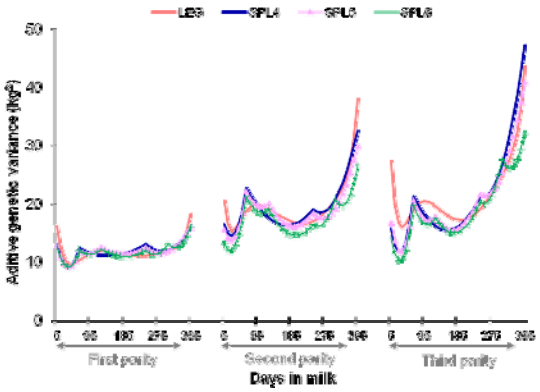


Figure 3. Posterior mean estimates of additive genetic variance of daily fat yield in first, second and third lactation.

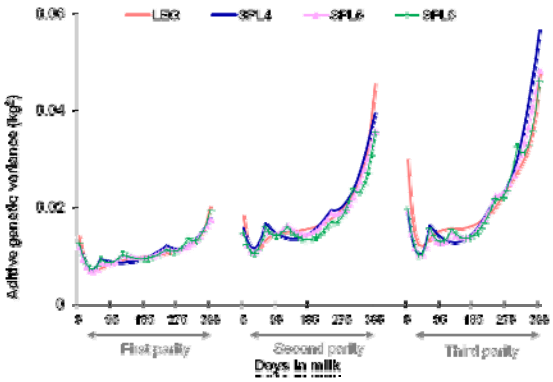


Figure 4. Posterior mean estimates of additive genetic variance of daily protein yield in first, second and third lactation.

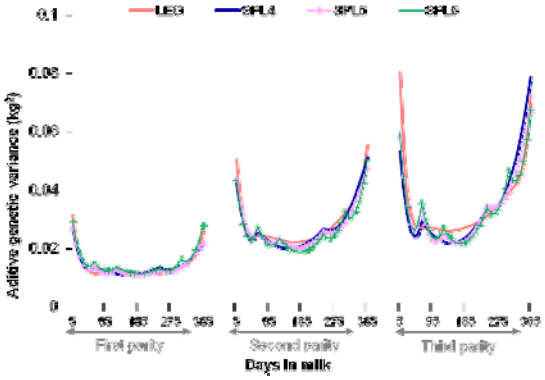


Figure 5. Posterior mean estimates of additive genetic variance of daily SCS in first, second and third lactation.

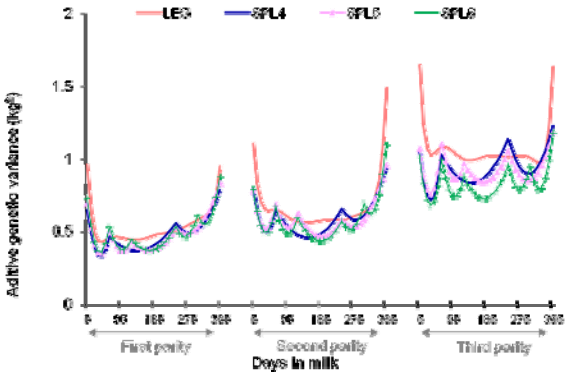


Figure 6. Posterior mean estimates of residual variance of daily milk yield in first, second and third lactation.

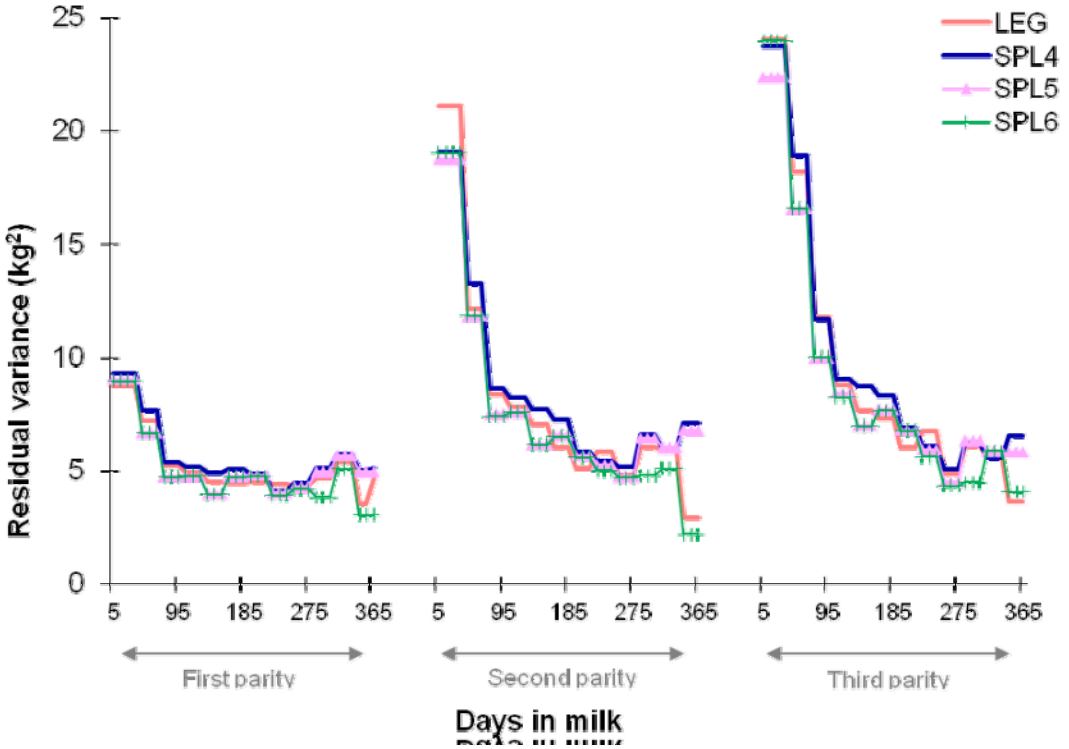


Figure 7. Posterior mean estimates of heritabilities for daily milk yield in first, second and third lactation.

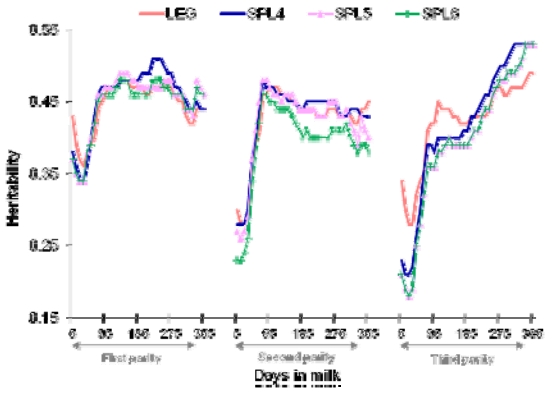


Figure 8. Posterior mean estimates of heritabilities for daily fat yield in first, second and third lactation.

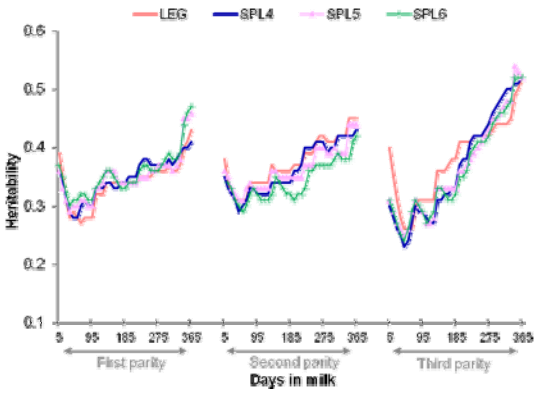


Figure 9. Posterior mean estimates of heritabilities for daily protein yield in first, second and third lactation.

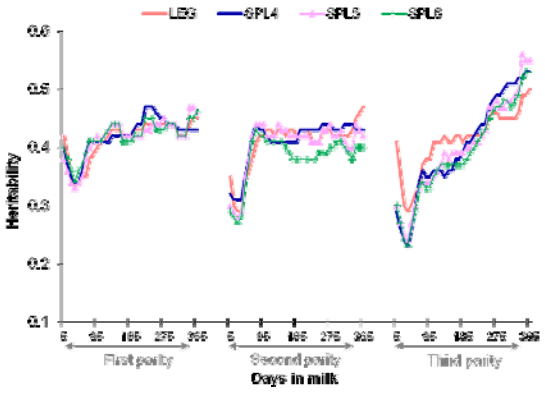


Figure 10. Posterior mean estimates of heritabilities for daily SCS in first, second and third lactation.