

## A PIECEWISE WEIBULL MIXED MODEL FOR THE ANALYSIS OF LENGTH OF PRODUCTIVE LIFE OF DAIRY COWS

V. Ducrocq

INRA Station de Génétique Quantitative et Appliquée 78352 Jouy en Josas Cédex, France

### INTRODUCTION

In the current survival mixed models that are used to evaluate dairy bulls based on the length of productive life of their daughters, the hazard  $h(t)$  of a cow - i.e., her probability of being culled at time  $t$  given she is alive just before  $t$  - is described using a proportional hazards model (PHM). PHMs assume that  $h(t)$  is the product of a baseline hazard function  $h_0(t)$  describing the overall change in culling risk over time for the whole population and a cow specific stress dependent term  $\exp\{\mathbf{x}_m(t)' \mathbf{b}\}$ . Here,  $\mathbf{x}_m(t)$  represents a set of (possibly time-dependent) explanatory variables that modify the culling risk of cow  $m$  and  $\mathbf{b}$  is a vector of fixed and random parameters. Among these, a stage of lactation  $\times$  lactation number (SLLN) effect is usually included. The stage of lactation effect attempts to account for changes in culling policy during the lactation: culling is more intense at the end of the lactation, when production is lower, when it is known whether the cow is pregnant or not and when her carcass value is better. Stage of lactation have been defined somewhat arbitrarily in such a way that the SLLN indicator variables in  $\mathbf{x}_m(t)$  are piecewise constant functions of time. For example, the current French routine genetic evaluation assumes (Ducrocq, 1999a):

$$h(t) = h_0(t) \exp\{l_{jn}(t) + \sum_k f_k(t) + s_{\text{sire}} + 0.5s_{\text{mgs}}\} \quad [1]$$

where  $t$  is the number of days since first calving, the baseline  $h_0(t)$  is a Weibull hazard function;  $l_{jn}(t)$  is the effect at time  $t$  of the  $j^{\text{th}}$  stage of lactation for the  $n^{\text{th}}$  lactation, with changes of stage of lactation at time 0, 30, 60, 150, 240 days after each calving date (an extra stage being added for the dry period);  $f_k(t)$  is a long list of other environmental effects (year-season, variation in herd size, within herd-year milk, fat % and protein % production class, age at first calving, random herd-year-season); finally,  $s_{\text{sire}}$  and  $s_{\text{mgs}}$  are the additive genetic effects of the sire and of the maternal grand-sire of the cow.

A careful examination of the estimates obtained for the baseline and the  $l_{jn}(t)$ 's suggests that the first part of [1],  $\hat{h}_0(t) \exp\{\hat{l}_{jn}(t)\}$  presents a similar pattern for each successive lactation (see Ducrocq, 1999a, Table 2). This observation led Rönström et al. (2002) to study a reparameterisation of [1], for which the baseline hazard is defined on a lactation basis as:

$$h(t) = h_{0,n}(\tau) \exp\{l_{jn}(t) + \sum_k f_k(t) + s_{\text{sire}} + 0.5s_{\text{mgs}}\} \quad [2]$$

where  $\tau$  is the number of days since the previous calving date ( $\tau = t - t_n$ , and  $t_n$  is the value of  $t$  at the  $n^{\text{th}}$  calving date),  $h_{0,n}(\tau)$  is a Weibull hazard function specific to the  $n^{\text{th}}$  lactation. Results suggested a better overall fit, but showed that a (reduced) number of stage of lactation effects was still needed. The objective of this paper is to go one step further in the modelling of the change in culling risk within each lactation, defining piecewise Weibull baselines within lactation as well as across lactation.

## MATERIAL AND METHODS

**Data.** From the data set used for the routine longevity genetic evaluation of the Normande breed in October 2001, a subset of 671011 records was extracted. These cows had at least one calving date recorded between December 1 1989 and December 1 2000. They were daughters of 2820 AI bulls and all had their maternal grand sire known. The total pedigree file included 2967 sires. 182 groups of unknown parent groups were included for service sires with at least 200 daughters and 6 for other animals: defining an unknown parent group for a single sire is tantamount to considering it as fixed. Hence, estimated genetic parameters are not biased by selection, being based on information from young bulls only. For culled cows, end of life was set equal to the last known test day. 31.4% of the records were censored (i.e., still alive, sold to another herd or in a herd where all cows were simultaneously culled). 14% of the records were truncated, i.e., starting at the  $q^{\text{th}}$  ( $q > 1$ ) calving date.

**Statistical analyses.** From the raw data and for each lactation (1, 2, ..5, 6 and above), a non parametric estimate of the hazard function was obtained as a by-product of the Kaplan-Meier estimate  $\hat{S}_n(\tau)$  of the survivor function (Klein and Moeschberger, 1997). A graphical test was performed plotting  $\log [-\log \hat{S}_n(\tau)]$  against  $\log \tau$ . If the plot displays a straight line, the Weibull assumption can be accepted. An examination of a smoothed version of the estimated hazard function and of the straight segments of the plot of  $\log [-\log \hat{S}_n(\tau)]$  against  $\log \tau$  led to a choice of a number of « candidate » periods within which the Weibull assumption seemed to hold. The bounds of these periods were further refined by comparing the likelihood values obtained from piecewise Weibull models relying on the candidate periods or on slight variations (+ or -20 days) of their bounds.

The number and bounds of the periods being fixed, a complete piecewise Weibull mixed model was applied after adding the same fixed and random effects (except the SLLN effect) as for the current routine genetic evaluation :

$$h(t) = h_{0,jn}(\tau) \exp \{ \sum_k f_k(t) + s_{\text{sire}} + 0.5 s_{\text{mgs}} \} \quad [3]$$

where the baseline hazard function  $h_{0,jn}(\tau)$  for the  $j^{\text{th}}$  period of the  $n^{\text{th}}$  lactation follows a specific Weibull hazard distribution. Then, genetic parameters and EBVs were estimated and compared with those of the current model [1].

**Practical implementation.** All analyses were performed using the current version (V3.12) of the Survival Kit ( <http://www.boku.ac.at/nuwi/software/softskit.htm> ; Ducrocq and Sölkner, 1998). The use of this software for piecewise Weibull hazards is based on the following considerations (see Rönxström et al, 2002 for details): the contribution to the log-likelihood function of a record of a cow  $m$  culled ( $\delta_m=1$ ) or censored ( $\delta_m=0$ ) at time  $y_m$  and truncated at time  $B$  ( $B=0$  if not truncated) is:

$$\delta_m \left[ \log h(y_m) \right] - \int_B^{y_m} h(u) du \quad [4]$$

Partition the time axis at  $B=b_0 < b_1 < .. b_j < .. < b_j = y_m$  where  $b_{j-1}$  and  $b_j$  are the lower and upper bounds of the  $j^{\text{th}}$  period. The first part of [4] appears at most once for each cow, when she is culled. The second part can be decomposed into the sum of period-specific

contributions :  $\int_B^{y_m} h(u) du = \sum_j \int_{b_{j-1}}^{b_j} h(u) du$ . It follows that the full log likelihood is the

same whether one considers the whole herd life of cow  $m$  or her  $J$  conceptual « partial » records, truncated at  $b_{j-1}$  and censored at  $b_j$  (with  $j=1, \dots, J$  and except for the last one which is possibly uncensored). Each partial record is processed as a record from a different cow, with its specific Weibull baseline.

### RESULTS AND DISCUSSION

The non parametric estimates of the hazard functions for each lactation (figure 1) and the plot of  $\log [-\log \hat{S}_n(\tau)]$  against  $\log \tau$  (figure 2) suggest for each lactation the definition of 4 periods for which a distinct Weibull hazard is acceptable : before day 30, between day 30 and 270, between day 270 and 380 and after day 380. The first and third periods are characterised by a fast increase in risk (estimated parameter  $\rho$  of the Weibull distribution between 3.5 and 4.2). The increase is much slower between 30 and 270 days ( $\hat{\rho}$  between 1.70 and 1.95, except in lactation 1,  $\hat{\rho}=1.27$ ) and after 380 days ( $\hat{\rho} \sim 1.5$ ). The current evaluation model overestimates culling risk in early lactation (before 270 days) and at the end of very long lactations (>380 days). The overestimation of the culling risk in early lactation *may* be one of the reasons for the frequent temporary increase of EBVs of proven sires after the inclusion of a large number of young second crop daughters (see Ducrocq, 1999, section 4.4) : for these bulls, much fewer cullings are observed than what the current model predicts.

The log-likelihood value for the piecewise Weibull model is much better than for models with SLLN effects (table 1). With the current model, the time partition proposed (at  $t=30, 270$  and  $380$ ) is already better than the one currently used (with 6 periods instead of 4). Changing the period bounds from 270 d to 250 d or 290 d, or from 380 d to 360 d or 400 d or deleting the partition at 30 days always leads to a poorer fit.

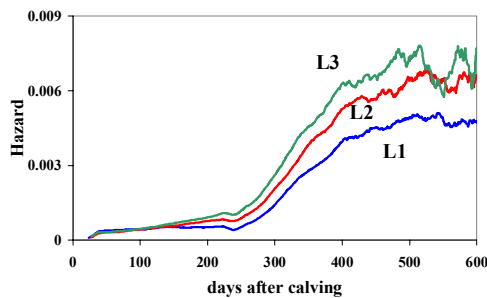


Figure 1. Nonparametric estimate of the hazard function for the first three lactations

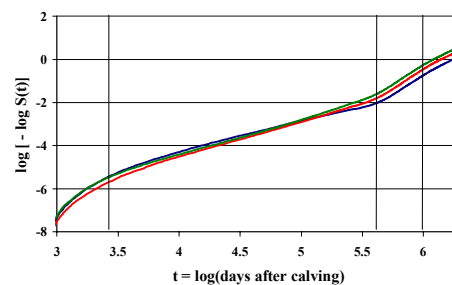


Figure 2. Plot of  $\log [-\log \hat{S}_n(t)]$  against  $\log(t)$ . On the left, from bottom to top : lactations 1, 2 and 3

**Table 1. Likelihood comparisons**

Modelling unit / model	Stage of lactation effect with changes at:	Piecewise Weibull with changes at:	Parameters	-2 Log-likelihood
Herdlife [1]	30, 60, 150, 240, dry		31	82214
Lactation [2]	30, 60, 150, 240, dry		31	80814
Lactation [2]	30, 270, 380		25	78200
Lactation [3]		30, 270, 380	48	<b>57444</b>

**Table 2. Estimated genetic parameters**

Modelling unit / model	sire variance	herd-year-season variance	h <sup>2</sup> (%) <sup>A</sup>	Expected reliability after 400 days <sup>A,B</sup>
Herdlife [1]	0.0371 +/- 0.0023	0.302	14.3	0.46
Lactation [3]	0.0359 +/- 0.0023	0.345	13.9	0.45
Early lactation only [3]	0.0518 +/- 0.0037	0.531	19.7	0.37
Late lactation only [3]	0.0209 +/- 0.0027	0.298	8.2	0.19

<sup>A</sup> as in Yazdi et al (2002) <sup>B</sup> for a sire with 100 daughters, using progeny information only

The sharp increase in culling risk after day 270 is another illustration that cullings in early and late lactation are different events, corresponding to different culling reasons. Considering them as identical traits with the same genetic determinism as it is assumed in the current evaluation models is probably an unsatisfactory approximation. The clear-cut 270 day limit (at least in the present data set) and the possibility to analyse independently “left-truncated, right-(un)censored” partial records led us to also analyse separately early and late lactation records. Genetic parameters were estimated using Laplacian integration techniques (Ducrocq and Casella, 1996). Clearly, the use of a piecewise Weibull baseline hazard did not change much the genetic variance, but slightly increased the herd-year-season variance (table 2). In contrast, genetic parameters estimated when the early and late parts of each lactation were analysed separately were extremely different: in early lactation, herd-year-season and sire effects were much more variable. Correlations between EBVs for longevity analysed on a herdlife basis or on a lactation basis were high (0.972) but they were lower between the analyses for the full lactations and the early (0.875) and late (0.790) parts of lactation. The low correlation between EBVs in early and late lactation (0.50) illustrates that the ability to stay alive during these two periods probably corresponds to genetically different traits. This should be recognised in genetic evaluations, e.g., by including time-dependent sire effects (Ducrocq, 1999b))

## REFERENCES

- Ducrocq, V. (1999a) *Interbull Bulletin* **20** : 60-67.  
 Ducrocq, V. (1999b) *Interbull Bulletin* **20** : 181-189.  
 Ducrocq, V. and Casella, G. (1996) *Genet. Sel. Evol.* **28** : 505-529.  
 Ducrocq, V. and Sölkner, J. (1998) *Proc. 6<sup>th</sup> WCGALP* **27** : 447-448  
 Klein, J. and Moescheberger, M. (1997) *Survival analysis*. Wiley and sons, New-York, USA.  
 Rösström, A., Ducrocq, V. and Strandberg E. (2002) *Genet. Sel. Evol.* **34** (submitted).  
 Yazdi, M.H., Visscher, P. M., Ducrocq, V. and Thompson, R. (2002) *J. Dairy Sci.* **85** (in press).