Auto-Regressive Versus Random Regression Test-Day Models for the Prediction of Milk Yields

Theo H.E. Meuwissen and Marco H. Pool
Institute Animal Science & Health, Box 65, 8200 AB Lelystad, Netherlands

Abstract

Character process models attempt to model the covariance function of test-day records, while random regression models attempt to model the production curves of individual animals. The simplest character process model is probably the auto-regressive (AR) model. We used AR here to model the covariance function, and to predict missing records, where the pattern of missing records resembled that of not yet completed lactations. Since the AR model gives about one equation per test-day (or test-week here), a new model was developed which only evaluated the covariance function at 9 knots along the lactation trajectory (ARrt). By choosing the knots at unequal time intervals, the assumption of a stationary process was alleviated. The AR and ARrt models were compared to a random regression model that fitted a 4-th order Legendre polynomial (LEG(4)). LEG(4), AR and ARrt gave quite different covariance functions, but all three the models seemed approximately equally well equipped to predict missing records from part lactations. LEG(4) fitted 16 variance components, and 4,755 random effects equations. For AR these figures were 4 and 42,795, respectively, and for ARrt they were 4 and 9,510, i.e. AR and ARrt fitted fewer variance components at the cost of more random effects equations. This may prove an advantage for AR and ARrt in small data sets, where the accurate estimation of many variance components is not possible. On the other hand, many equations may lead to computational problems in large-scale applications. The use of weights for the allocation of records to knots resulted in biased parameter estimates for the ARrt model.

Introduction

In recent years, test-day models were developed which predict breeding values for milk yields based on the recordings of individual yields at a number of test dates. The main goal was better correction for the effects of test dates, and the prediction of breeding values for the shape of the lactation curve, which makes selection for persistency possible. The most commonly used test-day model is probably the random regression model (RRM; Ptak and Schaeffer, 1993), where the curve of breeding values is predicted by a regression function whose coefficients are included as random effects and predicted per animal. Typically the lactation curves, that are used for the fixed prediction of average lactation curves, were also used as random regression curves, e.g. Wilmink's (Wilmink, 1987) and the Ali and Schaeffer curve (Ali and Schaeffer, 1987). However, also polynomials were used as random regression curves (e.g., Kirkpatrick et al., 1994).

The random regression curves are often very flexible in the fitting of individual curves, but they also implicitly define a (co)variance matrix of genetic effects along DIM. This (co)variance matrix may have an odd shape, e.g. highly increased variances at the beginning and end of the curves (Jamrozik et al., 1996). For example, a straight line may seem a simple and perhaps reasonable curve for the deviation of individual yields from the average lactation curve, but it leads to a quadratic curve for the variances, which seems often somewhat unrealistic. Also, random regression models can not model a correlation structure, that asymptotes to zero for days far apart (Jaffrezic and Pletcher, 2000). These shortcomings of polynomial RRM are alleviated by increasing their order of fit (Pool and Meuwissen, 1999; Pool, 2000),

but the number of parameters that need to be estimated is substantially increased.

Character process models do not attempt to model the production curve of an animal, but aim at modelling the (co)variance function (Jaffrezic et al., 2000). This contrasts with the RRMs, whose primary aim is to model the production curve, and the (co)variance function results from the estimates of the random regression curves. The simplest and computationally most feasible character process uses the covariance function:

$$C(t,s) = \sigma_t \sigma_s r^{|t-s|},$$

where C(t,s) = covariance between days in milk (DIM) t and s; σ_t (σ_s) standard deviation at DIM t (s); $r^{|t-s|}$ = the correlation between DIM t and s. When the data are collected at equally spaced intervals, the above character process equals the autoregressive model with autoregression correlation r (Carvalheira et al., 1998). The above character process is called stationary which means that the correlation between t and s depends only the absolute time difference |t-s| = ΔDIM. The latter assumption can be alleviated by a non-linear transformation of the time scale, and applying the assumption of stationarity on the transformed time scale (Nunez-Anton and Zimmerman, 2000).

A limitation of character process models is that the curve of breeding values of individuals is not explicitly estimated, and may be difficult to obtain. In the case of the autoregressive model, daily intervals may be used with missing records when an animal is not recorded at this DIM. This results in N_{dim}*N equations for the estimation of breeding values for every test-day, where N_{dim} = number of DIM, and N = total number of animals. Solving N_{dim}*N equations will be computationally infeasible in many practical data sets.

The aim was here to compare the Legendre polynomial random regression models with the autoregressive model for their ability to predict test-day records. Evaluating the breeding values of the animals at a limited number of DIM (so called knots) reduced the number of equations of the autoregressive model, and thus it's computational demands. The fuzzy classification approach of Strandberg and Grandinsson (1997) was used to assign the records, which are in between two knots, to their surrounding knots. By choosing the knots at unequal intervals, the time scale was non-linearly transformed, which alleviated the assumption of a stationary correlation.

Methods

Data

The data consisted of 951 lactations (536 first and 415 later parity from 605 Holstein Friesian cows), and 36,288 weekly recordings of test-day yields (on average 38 test-day yields per lactation, with a maximum of 44 weeks per lactation). Pool and Meuwissen (1999) gave a more detailed description of the data. In the following, 'missing records' will refer to records that are actually available but are excluded from the data set and are subsequently predicted by the models. About 50% of the lactations contain missing records, and the pattern of missing records mimics that of a lactation in progress, i.e. in lactations with missing records all records after DIM x are missing, where x is varied. The models will be compared for their ability to predict the missing records, which will be measured by the mean square error of prediction:

$$MSEP = \sum_{ij \in M} (y_{ij} - \hat{y}_{ij})^2 / n_M$$

where $\Sigma_{ij \in M}$ denotes summation over the set of missing records M; y_{ij} ($\hat{y}_{ij}$) = j-th weekly test-day record of lactation i (prediction); and n_M = the number of missing records in set M.

The complete Autoregressive model (AR)

In the complete autoregressive model, the data were modelled by:

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + a_i + b_{ij} + e_{ij} \qquad [1]$$

where $\boldsymbol{\beta}$ = vector of fixed effects for year-season of calving (3-monthly classes), age at calving (4-monthly age classes), cDIM = class of DIM * parity (weekly classes within first and later parities), and test date (date of milk recording); a_i = random cow * lactation effect; b_{ij} = random effect of j-th week within effect a_i; e_{ij} = random residual effect. Note that the fixed effect cDIM fits the average lactation curve. The covariance matrices of the random effects are:

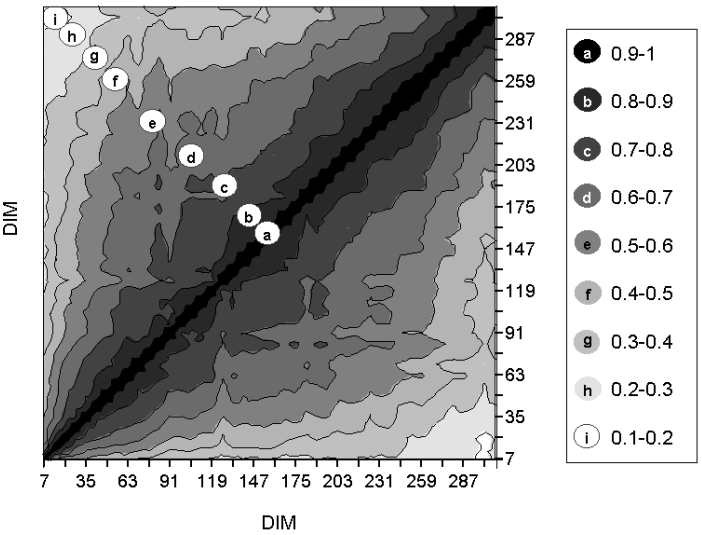
$$\text{Var}(\mathbf{a}) = \mathbf{I}_{951}\ \sigma^2_a,$$
$$\text{Var}(\mathbf{b_i}) = \mathbf{H}_{44}\ \sigma^2_b,$$
$$\text{Var}(\mathbf{e}) = \mathbf{I}_{36,288}\ \sigma^2_e,$$

where $\mathbf{I_n}$ = (n × n) identity matrix; $\mathbf{b_i}$ = vector of 44 weekly b_{ij} values; $\mathbf{H_{44}}$ = (44 × 44) weekly autoregression correlation matrix $H_{st} = r^{|s-t|}$, with r the auto-regression correlation and s and t are weeks since the beginning of the lactation. Note that if two test-dates in week 2 and 3 are actually only 1 day apart, they will still be assumed to be one week apart, i.e. fluctuations in the numbers of days between test-dates that are not expressed in week numbers are not accounted for by this autoregressive model. Residual Maximum Likelihood (REML) estimates for the parameters $\sigma^2_a, \sigma^2_b, \sigma^2_e$ and r were obtained from the complete data set (all 36,288 records) using the ASREML package (Gilmour et al., 2000). Hence, the total number of parameters estimated was 4 and the total number of random effects fitted was (1+44)*951 = 42,795. In the data sets with missing records, the missing records were predicted using the parameter estimates from the complete data set.

The reduced and transformed Autoregressive model (ARrt)

The autoregressive and other stationary correlation models assume equal correlations at the k-th diagonal of the correlation matrix, where the 0-th diagonal is the main diagonal, and the 1st diagonal is first diagonal above the main diagonal and so on. In Figure 1, the iso-correlation lines did clearly not appear as parallel lines to the diagonal. The correlation between week 1 and 1+t reduced faster with increasing t, than ad mid-lactation, say, between week 21 and 21+t. Also, at the end of the lactation, this reduction in correlation seemed somewhat faster than at mid-lactation. Hence, the stationary correlation assumption is clearly violated.

Figure 1. Correlation structure of test-day records at different DIM obtained from bi-variate REML analyses (from Pool and Meuwissen, 1999).



In the following, we will reduce the dimensionality of the autoregressive model by only evaluating the b_{ij} values at a reduced number of weeks, which will be called knots. The knots will be chosen such that the stationarity assumption is as little as possible violated. Because we need a knot at the beginning of the lactation, the first knot is in week 1, i.e. k_1=1. Next, we have to find the week k_1+t, whose correla-tion with week k_1 is approximately 0.7, i.e. $k_2=k_1+t$ (where t>0). The latter step is repeated until the end of the lactation is reached, and also at the end of the lactation a knot was placed (week 44). The week $k_{i+1}=k_i+t$, whose correlation with week k_i is ~ 0.7, was found by visual examination of Figure 1. This resulted in 9 knots, namely week 1, 2, 4, 8, 16, 26, 35, 41, and 44.

The basic ARrt model is the same as the AR model, except that the number of weeks included in the $\mathbf{b_i}$ vector is reduced from 44 to 9 (i.e. the 9 knots), which reduces **H** to a (9 × 9) matrix. Furthermore, choosing the knots of Art at unequal interval alleviates the stationary assumption of

the AR model. A problem arises when we have a test-record in week 5 while the nearest knots are at weeks 4 and 8. This is resolved by the fuzzy classification approach of Strandberg and Grandinsson (1997). The test-record of week 5 is duplicated, where the first duplicate is allocated to week 4 and given a weight of (8-5)/(8-4) = 0.75; and the second duplicate is allocated to week 8 with a weight of 1-0.75=0.25. All the in between knots test-records are duplicated, allocated to their nearest knots and weighted according to their distance till the knots. Hence, the number of parameters estimates was 4 (same as the AR model) and the number of random effects fitted was (1+9)*951 = 9,510.

The RRM: LEG(4)

Pool and Meuwissen (1999) compared Legendre polynomial random regression models of order 0 to 7 (indicated by LEG(0) to LEG(7)). In this comparison, a 4-th order Legendre polynomial (LEG(4)) seemed to give a fair balance between fits, i.e. higher orders of fits yielded only marginal improvements in MSEP, and number of parameters that needed to be estimated. LEG(4) fits 5 random effects per lactation, which results in 5*951 = 4,755 random effects and (5*(5+1)/2)+1 = 16 estimates of variances and covariances. The same set of data and missing records (as in Pool and Meuwissen, 1999) was used in this study and results of the AR and ARrt models were compared to LEG(4).

Results

Table 1 yields the parameter estimates of the AR and ARrt models, and residual variance from LEG(4). Other parameters of the LEG(4) model can be found in Pool and Meuwissen (1999). The residual variance of AR was lower than that of the LEG(4) model, which indicates that the AR model explained a larger part of the variance of the data. ARrt had yet a smaller residual variance, but this may be affected by the weighing of the records. If a record has a weight of w, this is equivalent to assuming that its residual variance is σ_e^2/w, and this factor 1/w might cause that the true residual variance is underestimated by σ_e^2. The correlation between adjacent knots of the ARrt model is (0+23.23*.907)/(0+23.23 +2.68) = 0.8132, which is higher than 0.7, i.e. the aim when choosing the knots. This estimate of the correlation between adjacent knots may again be too high because the residual variance is underestimated, as explained above.

Table 1. Parameter estimates of the AR, ARrt and LEG(4) models

Parameter[1]	AR	ARrt	LEG(4)[2]
$Var(a_i)$	2.57	0	
$Var(b_{ij})$	21.64	23.23	
Auto-correlation	0.97	0.91	
$Var(e_{ij})$	3.02	2.68	4.38

[1] effects a_i, b_{ij}, and e_{ij} denote the lactation effect, the effect of week (AR) or knot (ARrt) within lactation, and the residual, respectively (Equation [1]).

[2] For LEG(4) only $Var(e_{ij})$ is given, because only this parameter has a comparable parameter in the AR and ARrt models. For other parameters of LEG(4) see Pool and Meuwissen (1999).

Figure 2 shows the predicted correlation structures of the models AR, ARrt, and LEG(4). As expected, the iso-correlation lines for the AR model are parallel to the diagonal, i.e. AR does not account for a faster reduction in correlation with increasing ΔDIM earlier (or later) in the lactation compared to at mid-lactation. The iso-correlation curves from the models ARrt and LEG(4) do show such a curvature were the correlations reduce faster with ΔDIM in early or late lactation compared to mid-lactation. Note however that for ARrt, the curvature in the iso-lactation lines is not estimated by the ARrt model but entirely defined by the choice of the knot points. If the correlations were plotted against the knot points instead of actual DIM, the iso-correlations would also be parallel to the diagonal for the ARrt model. Visual comparison between Figures 1 and 2b suggests that the correlation at the end of the lactation decreases too quickly with ΔDIM in Figure 2b, which could be remedied by choosing the knot-points at the end of the lactation further apart.

When comparing AR (Figure 2a) against the bi-variate correlations (Figure 1), it seems that the correlations at high ΔDIM, i.e. the lowest correlations, are somewhat too high. This suggests that the correlations in the first order auto-regressive model do not drop off sufficiently quick. In the ARrt model, the correlations seem rather high at combinations of DIM. The LEG(4)

model seems to give the best correlation structure. This is probably due to the large number of parameters involved in the LEG(4) model, which gives it an increased flexibility.

Figure 2. Correlation structure of test-day records at different DIM obtained from the different models. Above: figure ***A***: AR (upper triangle) and, figure ***B***: ARrt (lower triangle). Below: figure ***C***: LEG(4)[1] model.

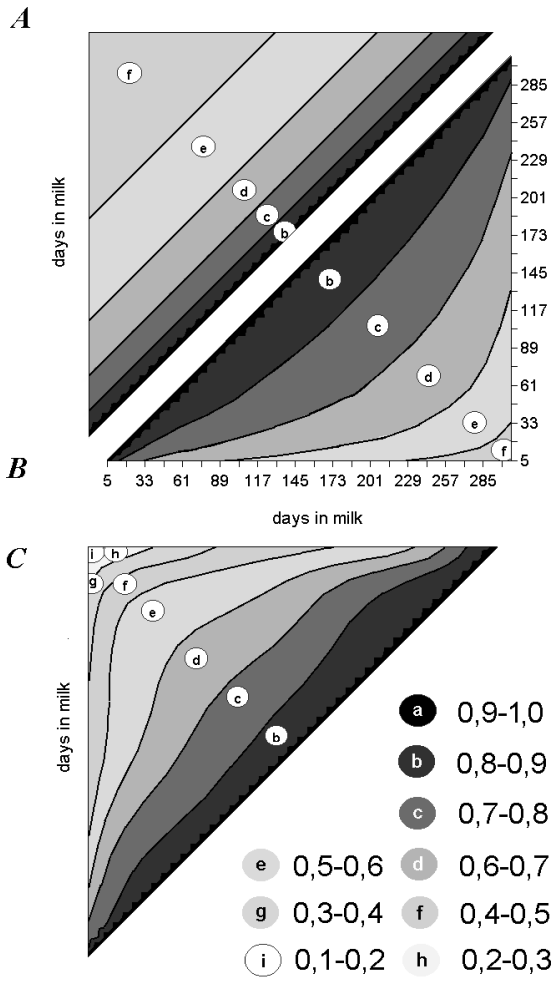
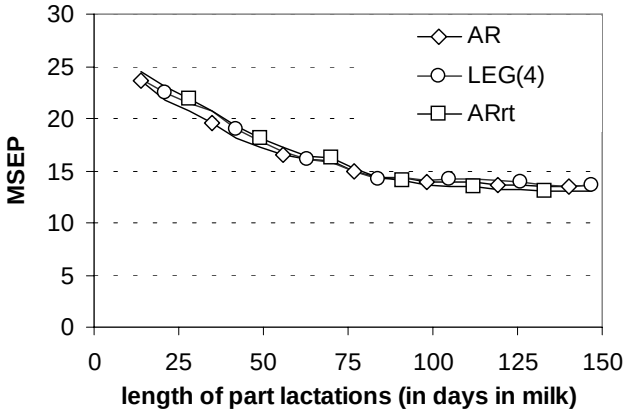


Figure 3 shows the MSEP of later test-day records using the information of running lactations. The MSEP profiles are remarkable similar showing that the models AR, ARrt, and LEG(4) are approximately equally able to predict missing records from part lactations. The reduction in number of random effects equations by using ARrt instead of AR (9,510 vs. 42,795 equations) did clearly not result in a poorer MSEP for the ARrt model.

Figure 3. MSEP of missing records of part lactations when the records up to DIM are available.



Discussion

Three models, AR, ARrt, and LEG(4) were mainly compared for their predicted correlation structures and MSEP of missing rec??ords of part lactations. Low MSEP of missing records, i.e. test-day records after DIM t, suggests that a) the model explains a large part of the variation of the milk yields, and b) the predictions will show little changes when the records after DIM t become available (because the future records are inline with the expectations of the model). These are desirable properties for breeding value estimation, which is often based on a substantial number of part lactation records.

It is remarkable that AR, ARrt, and LEG(4) have fairly different correlation structures, but very similar MSEP (Figures 2 and 3). It seems that accurate prediction of correlation structures is not critical for the MSEP of missing observations. This suggests that we can easily reduce the dimensionality of the model for the correlation structure, as long as the shape of the correlation structure is not too different from the bi-variate correlations. The latter is for instance the case for the LEG(1) model, which yields concave instead of the convex iso-correlation lines (Figures 2b-c) (Pool and Meuwissen, 1999). The AR and ARrt models have each 4 parameters to describe the covariance structure, and were thus quite extreme in reducing the dimensionality of the covariance

[1] Result from model LEG(4) are from Pool and Meuwissen (1999).

model. When the covariance function is estimated from a small data set, the estimates of the 16 parameters of the LEG(4) model might show considerable estimation error while the estimates of the 4 parameters of AR and ARrt might still be reasonably accurate. These estimation errors will be further increased when genetic and environmental covariance functions are estimated. Hence, the AR and ARrt models may be preferred over LEG(4) when the data set is relatively small and/or genetic and environmental covariance functions are estimated.

It may be noted that LEG(4) models the complete covariance function, while AR and ARrt only model the correlations, and assume a constant variance across the lactation. This assumption is reasonable in the current data set (see Pool and Meuwissen, 1999), but will not hold in general. Hence, in many situations, also variances along DIM need to be modelled which will require an increased number of parameter estimates (Jaffrezic et al., 2000). Generally it is envisaged as an advantage that the variances and correlations are modelled by different curves, which makes it easier to choose a suitable curve for the variances and for the correlations (Jaffrezic and Pletcher, 2000).

The fuzzy classification implies that Cov(y(DIM_1); y(DIM_2)) is obtained by linear interpolation from the known covariances at the knots, i.e. Cov(y(L_1);y(L_2)), Cov(y(U_1),y(L_2)), Cov(y(L_1),y(U_2)), and Cov(y(U_1),y(U_2)), where y(DIM_1) = milk yield at day in milk DIM_1, and L_1 (L_2) is the knot just before DIM_1 (DIM_2); and U_1 (U_2) is knot just after DIM_1 (DIM_2). This linear interpolation works reasonable well in areas where the covariance function is approximately linear. But it will be poor around the top of the curve (i.e. where the variances are modelled). For example, let DIM_1=DIM_2 be in week 3, L_1=L_2 is in week 2, and U_1=U_2 is in week 4, then Var(y(3)) = .25*[Var(y(2)) + Var(y(4)) + 2*Cov(y(2);y(4))], i.e. the average of 4 terms of which two terms are covariances instead of variances. The covariance terms are lower than the variances, such that the model expects Var(y(3)) to be substantially smaller than Var(y(2)) and Var(y(4)). In the data set, probably all three variances were approximately equal, which may have been accommodated by the model by overestimating the autoregression coefficient somewhat (i.e. increase Cov(y(2);y(3)), and by overestimating Var(b_{ij}) (i.e. increase Var(y(2))+Var(y(4))). Both overestimations seem to have happened in Table 1, i.e. Var(b_{ij}) is increased relative to the AR model, and the correlation between knots is higher than expected (0.8132 vs. 0.7; see Results section). An alternative to the current fuzzy logic approach, where the records are weighted, is to include the weights in the design matrix of b_{ij}, i.e. [1] may be replaced by

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + a_i + wb_{ij} + (1\text{-}w)\, b_{i(j+1)} + e_{ij},$$

where w = the weight at the knot of b_{ij}. This does not avoid the above averaging of the 4 (co)variance terms, but it may avoid the biased estimation of σ_e^2.

Another problem with the fuzzy logic classification occurs when we extent the current model to a genetic model. The straightforward extension involves an autoregressive correlation matrix for the genetic and for the permanent environmental effects, i.e. it involves genetic and permanent environmental autocorrelations and genetic and permanent environmental knots. However, we can not assign the genetic part of a record to some knots and the environmental part to other knots since we do not know which part of the record is genetic and which part is environmental (although we could perhaps use their estimates, but still the model would become much more complicated). Hence, extending the ARrt model to a genetic model is much easier if the same knots could be used for both the genetic and the permanent environmental curves. Perhaps, in practice, we will find that the genetic covariance function and the permanent environmental covariance function change more rapidly and more slowly at approximately the same DIMs, such that using the same knots for both curves gives a reasonable approximation. Note that the extension of the AR model to a genetic model is conceptually straightforward: it simply involves having genetic and permanent environmental a_i and b_{ij} effects. However, this model may give computational difficulties, because the number of equations may become very large.

Conclusion

The LEG(4), AR and ARrt models proved approximately equally well equipped to predict the missing records of part lactations, despite that they are very different models with quite different covariance structures. The AR and ARrt model may

be especially useful in small data sets, because they require the estimation of only a few variance components. In large scale applications, LEG(4) may be computationally easier because LEG(4) requires much fewer equations than ARrt and especially than AR. Further research is needed to reduce the computational problems of AR and ARrt models, and some estimation biases due to the assignment of records to knots in the ARrt model.

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.

Carvalheira, J.G.V., Blake, R.W., Pollak, E.J., Quaas, R.L. & Duran-Castr, C.V. 1998. Application of an autoregressive process to estimate genetic parameters and breeding values for daily milk yield in a tropical herd of Lucerna cattle and in United States Holstein herds. *J. Dairy Sci. 81,* 2738-2751.

Gilmour, A.R., Cullis, B.R., Welham, S.J. & Thompson, R. 2000. ASREML reference manual version 2000. *New South Wales Agriculture,* Orange, Australia.

Jaffrezic, F., White, I.M.S., Thompson, R. & Hill, W.G. 2000. A link function approach to model heterogeneity of residual variances over time in lactation curves. *J. Dairy Sci. 83,* 1089-1093.

Jaffrezic, F. & Pletcher, S.D. 2000. Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics 156,* 913-922.

Jamrozik, J., Schaeffer, L.R. & Dekkers, J.C.M. 1996. Random regression models for production traits in Canadian Holsteins. *Interbull Bulletin 14,* 124-134.

Nunez-Anton, V. & Zimmermann, D.L. 2000. Modeling non-stationary longitudinal data. *Biometrics 56,* 699-705.

Pool, M.H. & Meuwissen, T.H.E. 1999. Prediction of daily milk yields from a limited number of test-days using test-day models. *J. Dairy Sci. 82,* 1555-1564.

Pool, M.H. 2000. Test-day models, breeding value estimation based on individual test-day records. *Thesis,* Wageningen University, Wageningen.

Ptak, E. & Schaeffer, L.R. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. *Livest. Prod. Sci. 34,* 23-34.

Kirkpatrick, M., Hill, W.G. & Thompson, R. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genetical Research Cambridge. 64,* 57-69.

Strandberg, E. & Grandinsson, K. 1997. Adjusting for seasonal effects in animal model using fuzzy classification. *Interbull Bulletin 16,* 100-103.

Wilmink, J.B.M. 1987. Comparison of different methods of predicting 305-day milk yield using means calculated from within-herd lactation curves. *Livest. Prod. Sci. 17,* 1-17.