J. Appl. Genet. 42(2), 2001, pp. 177-191

Application of the covariance function approach with an iterative two-stage algorithm to the estimation of parameters of a random regression test day model for dairy production traits

Joanna SZYDA

Department of Animal Genetics, Agricultural University of Wrocław, Wrocław, Poland

Abstract. The covariance function approach with an iterative two-stage algorithm of LIU et al. (2000) was applied to estimate parameters for the Polish Black-and-White dairy population based on a sample of 338 808 test day records for milk, fat, and protein yields. A multiple trait sire model was used to estimate covariances of lactation stages. A third-order Legendre polynomial was subsequently fitted to the estimated (co)variances to derive (co)variances of random regression coefficients for both additive genetic and permanent environment effects. Daily and 305-day heritability estimates obtained are consistent with several studies which used both fixed and random regression test day models. Genetic correlations between any two days in milk (DIM) of the same lactation as well as genetic correlations between the same DIM of two lactations were within a biologically acceptable range. It was shown that the applied estimation procedure can utilise very large data sets and give plausible estimates of (co)variance components.

Key words: covariance function, dairy cattle, random regression model, test day data.

Introduction

Estimating (co)variance parameters of a random regression test day model for dairy production traits is computationally very challenging. Among the main reasons are: the very large number of (co)variance parameters and effects to be estimated, high correlations between parameters, and a large size of the data set needed to provide reasonable accuracy of these estimates.

Received: November 8, 2000. Accepted: February 15, 2001.
Correspondence: J. SZYDA, Department of Animal Genetics, Agricultural University of Wrocław, ul. Kożuchowska 7, 51-631 Wrocław, Poland, e-mail: szyda@karnet.ar.wroc.pl

Up to now two general approaches have been applied, these are the one-step random regression model approach and the two-step model based on covariance functions (CF). In the one-step model animal genetic and permanent environmental effects are modelled through a lactation curve function. Random regression coefficients (RRC) of the function and their (co)variance components are estimated jointly with other effects, typically using Restricted Maximum Likelihood (REML, OLORI et al. 1999) or the Gibbs sampling algorithm (JAMROZIK, SCHAEFFER 1997; for a review see MISZTAL et al. 2000). The two-step approach estimates (co)variance parameters of genetic and residual effects of different lactation stages in the first step, and fits CF to the estimated (co)variance matrices to obtain (co)variances of RRC in the second step. The CF can be fitted using the generalised least square (GLS) inverse method (TIJANI et al. 1999), the expectation maximisation algorithm (MÄNTYSAARI 1999), modified GLS inverse (LIU et al. 2000), and weighted least squares (KIRKPATRICK et al. 1994, LIU et al. 2000). Although the one-step approach enables joint estimation of all model parameters, in practical application it is often not suitable for data sets large enough to provide accurate parameter estimates. This may result in estimates which are not biologically valid. The two-step approach enables to incorporate information from much larger data sets, and thus higher accuracy of estimates.

The objective of this study was to apply the covariance function approach with an iterative two-stage algorithm of LIU et al. (2000) to estimate (co)variance components of RRC of a random regression test day model for the first three lactations test day yields of the Polish dairy cattle population.

Material

A total of 338 808 test day records for milk, fat, and protein yields from the first three lactations were selected from the Polish Black-and-White dairy cattle population. The statistical description of the data set can be found in the earlier paper by SZYDA and LIU (1999). The following selection criteria were imposed: herd-test-date (HTD) classes with at least five records, supervised monthly testing with two times milking, and calving years beginning from 1993, 1994 and 1995, respectively, for the first, second and third lactations. In the case of duplicate test day records within each of the six lactation stages defined below, one record was randomly chosen. Only complete lactations were used for estimating the parameters. Sires with fewer than 30 daughters were excluded to achieve a better data structure. Table 1 shows the structure of the final test day data set and sire pedigree file used for parameter estimation. For each of the three lactations, 15 fixed lactation curves were fitted to data based on three calving seasons (January-March, April-August, September-December), and five classes of age at calving.

Method

Estimation of parameters on a daily basis

For the estimation of parameters of a random regression test day model, the covariance function approach with an iterative two-stage algorithm of LIU et al. (2000) was chosen. The iteration procedure is based on the iterated conditional expectation method as shown by ROYLE and BERLINER (1999).

Step 1

On the basis of the number of days in milk (DIM) each lactation was partitioned into six stages: 5-50, 51-105, 106-160, 161-215, 216-259, and 260-305 DIM. The (co)variance components for these six lactation stages were estimated using a multi-trait sire model applied to test day yields from the first three lactations:

$$y_{ijklmn} = \mu_{lm} + HTD_{il} + \sum_{p=1}^{5} \beta_{jlp} \nu_{pd} + s_{klm} + e_{ijklmn} ,$$

where: y_{ijklmn} is the test day yield in the m-th stage of the l-th lactation of cow n, μ_{lm} is the mean for stage m of lactation l, HTD_{il} is the i-th herd-test-date effect of lactation l, ν_{pd} is the p-th parameter of Ali-Schaeffer function (ALI, SCHAEFFER 1987) for d-th DIM, β_{jlp} is the p-th fixed regression coefficient for lactation l specific to age-season class j, s_{klm} is additive genetic effect of sire k for the m-th stage of the l-th lactation, and e_{ijklmn} is the residual effect. In the above model, different lactation stages are treated as correlated traits.

In the estimation procedure fixed effects of HTD and β are estimated by ordinary least squares (stage A). Then (co)variance components of sire effects ($\mathbf{G}_s$) and residual effects ($\mathbf{R}_s$) are estimated via restricted maximum likelihood (stage B):

stage A $$y_{ijklmn} - (\hat{\mu}_{lm}^{(r-1)} + \hat{s}_{klm}^{(r-1)}) = HTD_{il}^{(r)} + \sum_{p=1}^{5} \beta_{jlp}^{(r)} \nu_{pd} + \varepsilon_{ijklmn}^{(r)} ,$$

stage B $$y_{ijklmn} - (\hat{HTD}_{il}^{(r)} + \sum_{p=1}^{5} \hat{\beta}_{jlp}^{(r)} \nu_{pd}) = \mu_{lm}^{(r)} + s_{klm}^{(r)} + \xi_{ijklmn}^{(r)} ,$$

where superscript (r) denotes an iteration round, $\hat{s}_{klm}$ represents the estimated sire breeding value (EBV), ε_{ijklmn} and ξ_{ijklmn} are residual effects of models in stages A and B, respectively. Both stages are iteratively repeated until all (co)variance components and sire EBVs converge.

Parameters of the sire model were estimated using Fortran 90 programs and Unix shell scripts. Computations required for stage B were carried out by the VCE package (NEUMAIER, GROENEVELD 1998). Starting values for sire EBVs originated from a fixed regression test day model (REENTS et al. 1995), previously applied to the data. The total number of estimated components exceeded our computational feasibility so that the estimation task had to be divided into partial

analyses involving a lower number of traits. In each partial analysis (co)variance parameters were estimated for 9 traits: 6 lactation stages from one lactation plus 3 stages from another lactation. Consequently, in order to map 18 × 18 (co)variance matrix from the full model, seven 9-trait analyses had to be performed.

Step 2

In the next step sire ($\mathbf{G}_s$) and residual ($\mathbf{R}_s$) (co)variances from step 1 were converted to an animal model using $\mathbf{G} = 4\mathbf{G}_s$ and $\mathbf{R} = \mathbf{R}_s - 3\mathbf{G}_s$. As **G** and **R** refer to the lactation stages, they have to be transformed into the (co)variances of RRC using CF modelled through the third-order normalised orthogonal Legendre polynomials. For that purpose the modified weighted least squares approach (LIU et al. 2000) was used. In contrast to the generalised least squares inverse approach (TIJANI et al. 1999), the weighted least squares method (KIRKPATRICK et al. 1994) incorporates information on the accuracy of parameter estimates through their sampling variances. The modification of LIU et al. (2000) comprises an iterative procedure to separate time-dependent permanent environmental effects from time-independent error effects. Derivation of (co)variances of RRC from the (co)variance estimates of the lactation stages was carried out by Maple V software.

Estimation of parameters on a lactation basis

The (co)variance estimates obtained on a daily basis were converted to lactation based estimates using:

$$h_L^2 = \frac{\sum_{i=D_{\min}}^{L}\sum_{j=D_{\min}}^{L}\sigma_{g(i,j)}}{\sum_{i=D_{\min}}^{L}\left[\sum_{j=D_{\min}}^{L}(\sigma_{g(i,j)}+\sigma_{p(i,j)})+\sigma_{e(i,i)}\right]},$$

where, h_L^2 is the heritability referring to the lactation of L days length, D_{min} denotes the value of DIM chosen as the beginning of lactation, $\sigma_{g(i,j)}$ and $\sigma_{p(i,j)}$ are respectively genetic and permanent environmental covariances between DIM i and j, $\sigma_{e(i,i)}$ is the error variance at DIM i. Based on daily estimates, the genetic correlation ($r_{g(n,m)}$) between two lactations can be derived as follows:

$$r_{g(n,m)} = \frac{\sum_{i_n=D_{min}}^{L}\sum_{j_m=D_{min}}^{L}\sigma_{g(i_n,j_m)}}{\sqrt{\left[\sum_{i_n=D_{min}}^{L}\sum_{j_n=D_{min}}^{L}\sigma_{g(i_n,j_n)}\right]\left[\sum_{i_m=D_{min}}^{L}\sum_{j_m=D_{min}}^{L}\sigma_{g(i_m,j_m)}\right]}},$$

where subscripts n and m correspond to two lactations.

Results

Heritability

Daily yield heritability estimates for milk, fat, and protein, based on the (co)variances of RRC, from the first three lactations are shown in Figures 1, 2 and 3. Estimated heritabilities change during the course of lactation, varying respectively for the first, the second, and the third lactation between: 0.12 and 0.23, 0.14 and 0.23, 0.12 and 0.17 for milk yield, 0.10 and 0.14, 0.09 and 0.17, 0.10 and 0.15 for fat yield, 0.08 and 0.16, 0.09 and 0.16, 0.09 and 0.15 for protein yield. The highest heritabilities among the three lactations were recorded for the first lactation, and among the studied traits for milk yield. For each of three traits considered it can be seen that the beginning and the end part of the first lactation show lower heritabilities than the middle part. For the two following lactations for daily milk and protein yield heritability estimates, a similar pattern is observed, but there is an increase of heritability at the very end of lactation.

Table 2 gives overall heritabilities derived for the 305-day lactation for milk, fat and protein yields based on daily parameter estimates. These estimates are higher than any of the daily heritabilities. As it is commonly seen in the majority of populations, milk yields have the highest heritabilities and protein yields have the lowest, later lactations show lower heritabilities.

Genetic correlation structure

Figures 4 and 5 show genetic correlations for daily milk and fat yields, respectively, between DIM 30, DIM 150, DIM 250 and the whole course of the first lactation (represented by yields at all other DIM). The three values of DIM were chosen to represent the beginning, the middle, and the end part of lactation. The genetic correlation between the end and the beginning of lactation, here expressed by $r_{g(30,305)}$ and $r_{g(5,250)}$, is in the range of 0.3. The pattern of daily genetic correlations remains the same for all three traits, but milk yields tend to be less correlated than fat and protein yields. Daily yields from the second and the third lactations are less correlated than from the first lactation. All the genetic correlations are positive.

Genetic correlations for milk, fat and protein yields between the same DIM from two lactations are presented in Figures 7, 8, and 9, respectively. All three traits present a similar level and a general pattern of daily genetic correlations. For milk yield the correlations vary from 0.50 for DIM 5 to 0.92 for DIM 110-165 for the first and the second lactation, from 0.86 for DIM 5-10 to 0.92 for DIM 205-235 for the second and the third lactation, and from 0.60 for DIM 5 to 0.89 for DIM 275-295 for the first and the third lactation. High daily correlations between the second and the third lactation indicate that they are genetically similar. The middle and end stages of lactation appear to be more genetically correlated

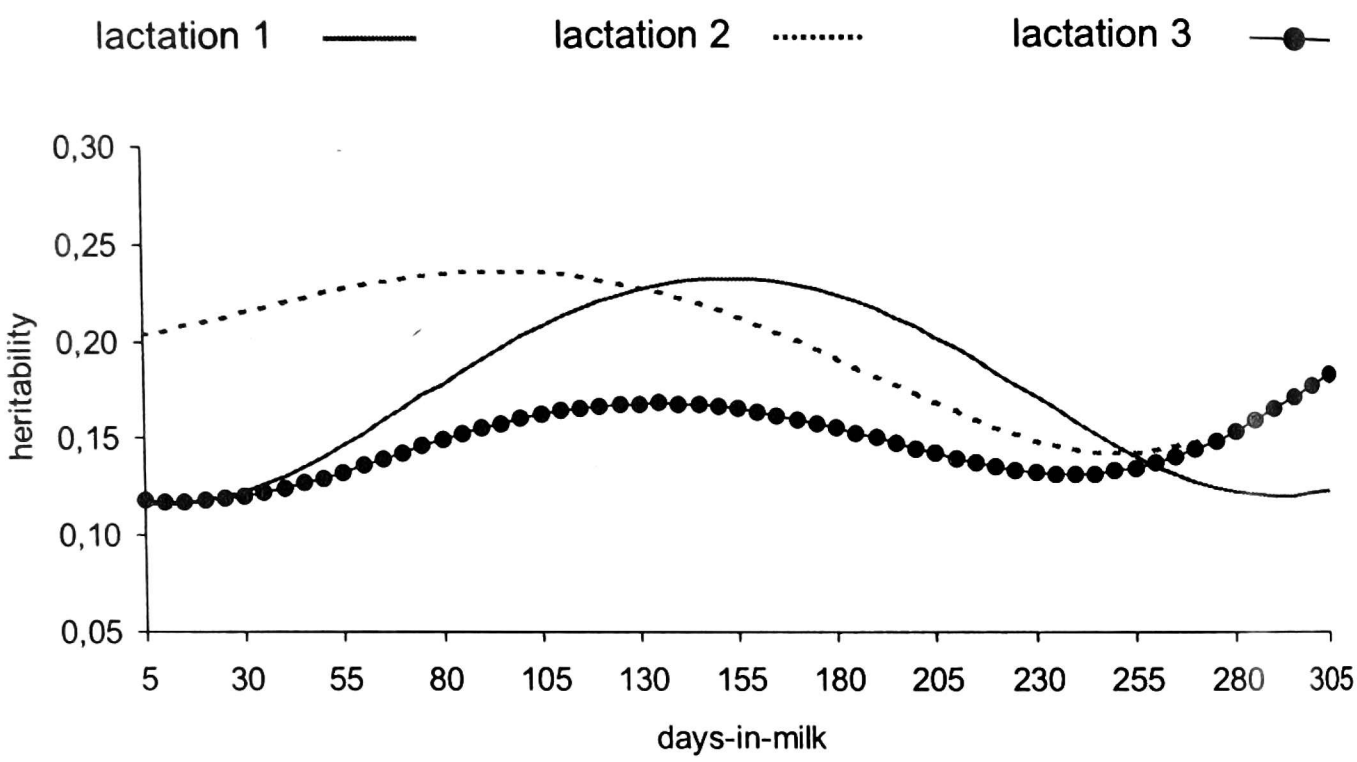

Figure 1. Heritability estimates for daily milk yield in the 1st, 2nd, and 3rd lactations

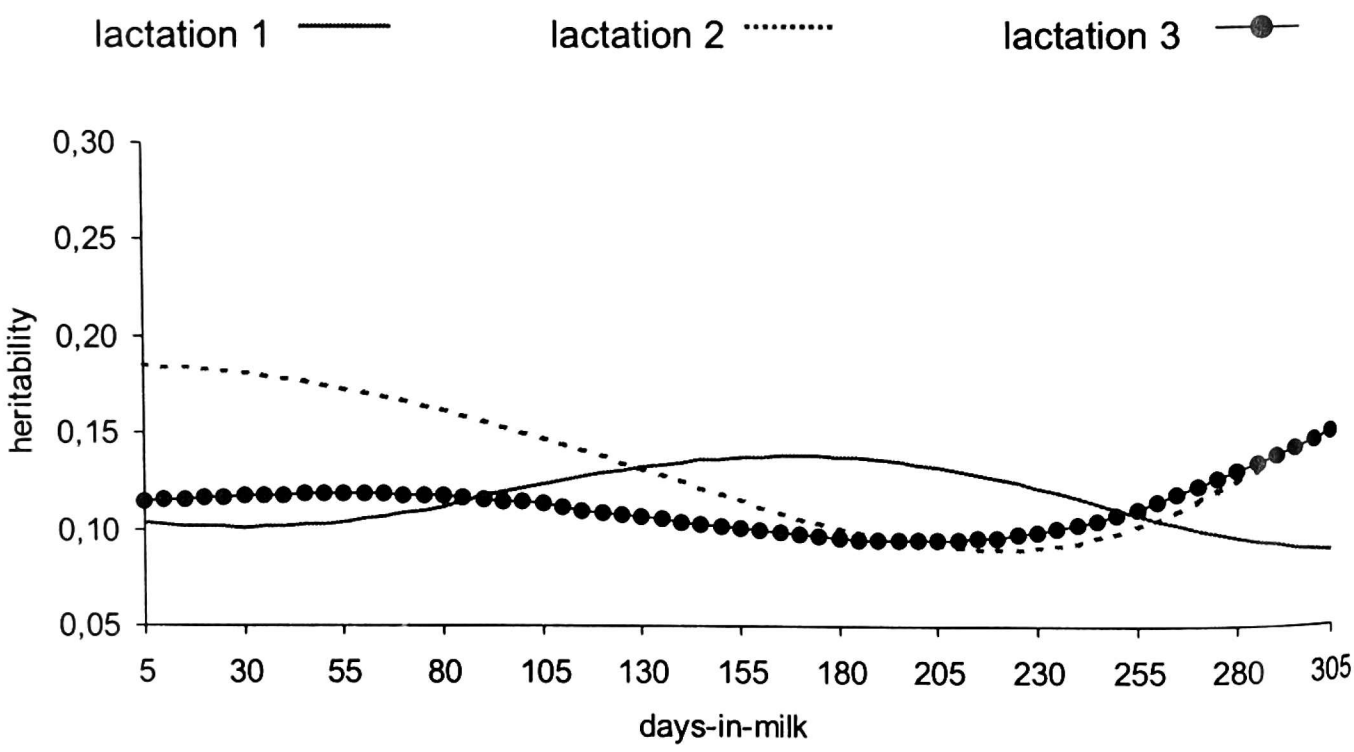

Figure 2. Heritability estimates for daily fat yield in the 1st, 2nd, and 3rd lactations

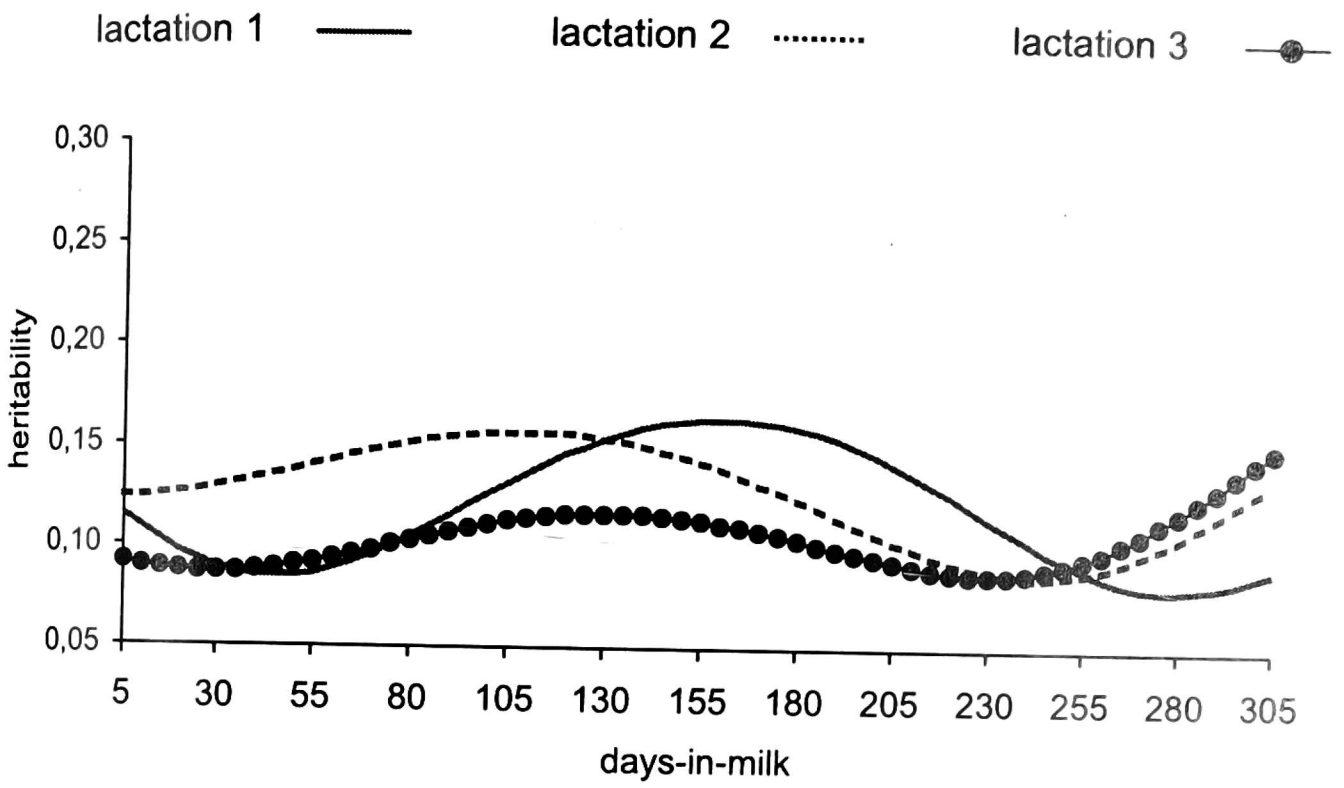

Figure 3. Heritability estimates for daily protein yield in the 1st, 2nd, and 3rd lactations

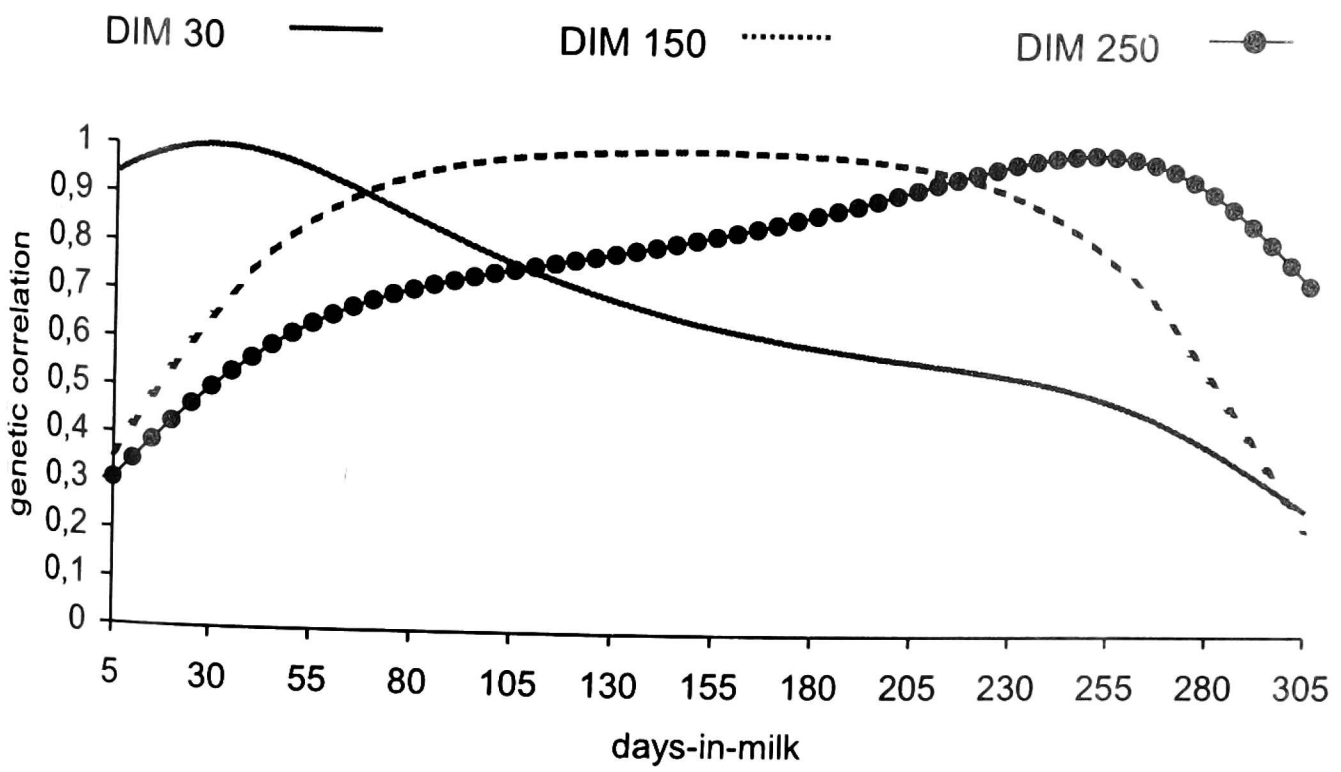

Figure 4. Genetic correlation estimates between daily milk yields and a given DIM in the 1st lactation

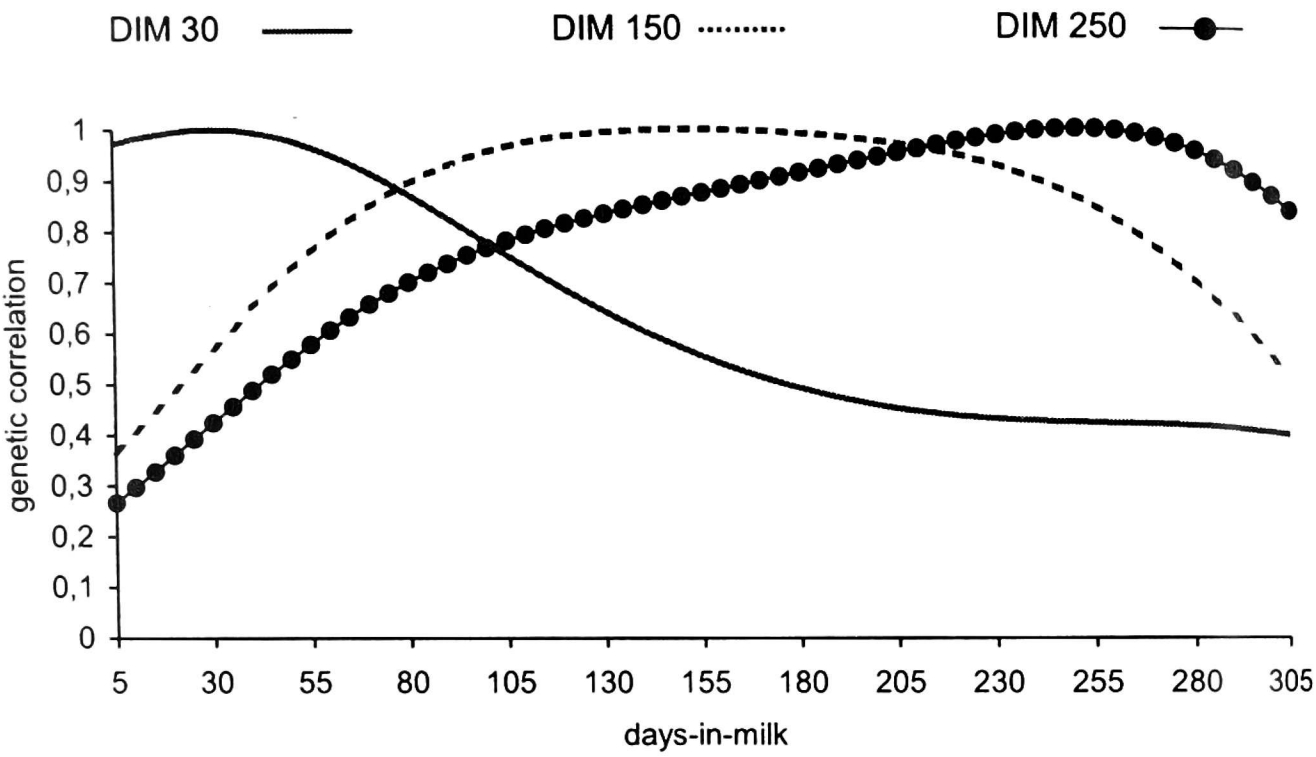

Figure 5. Genetic correlation estimates between daily fat yields and a given DIM in the 1st lactation

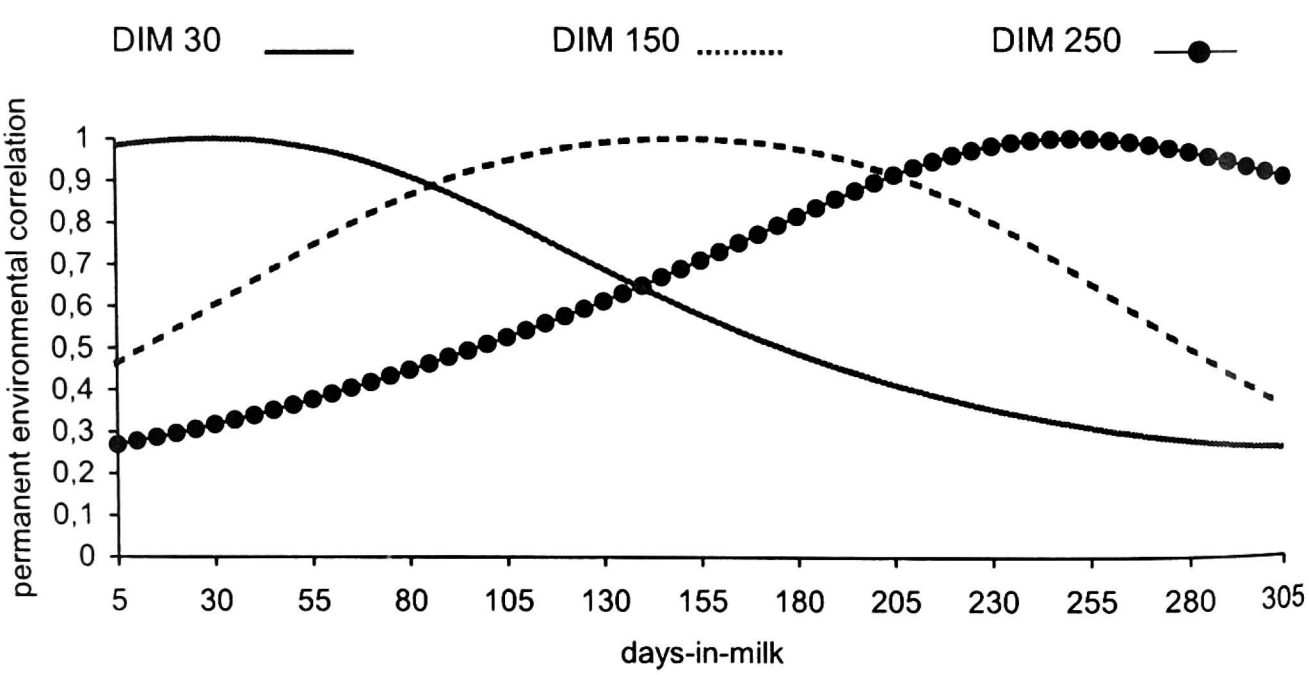

Figure 6. Permanent environmental correlation estimates between daily milk yields and a given DIM in the 1st lactation

Table 1. Number of individuals and subclasses of final data set used for parameter estimation

Cows	Sires of cows	Animals in sire pedigree file	Test day records	Herd-test-day subclasses	Fixed lactation curves
50 096	2 239	3 255	338 808	52 790	45

Table 2. Heritability estimates for the 305-day lactation

Lactation	Milk yield	Fat yield	Protein yield
1	0.31	0.26	0.22
2	0.34	0.25	0.22
3	0.23	0.20	0.16

Table 3. Estimates of genetic correlations between 305-day lactations

Lactations	Milk yield	Fat yield	Protein yield
1 and 2	0.87	0.91	0.95
2 and 3	0.92	0.83	0.79
1 and 3	0.82	0.76	0.80

than the beginning of lactation. This is especially profound for the correlations involving the first lactation.

Genetic correlations calculated on the 305-day lactation basis are presented in Table 3. For milk yield the highest correlation of 0.92 is between the second and the third lactation, while for fat and protein yields between the first and the second lactation, respectively 0.91 and 0.95. These values are higher than the correlations averaged over all DIM, as they account for the positive covariances between DIM of the same lactation.

Permanent environmental correlation structure

Looking at the estimates of the correlation of permanent environmental effects of milk yield of a given DIM with the rest of lactation (Figure 6) one observes that the correlations between the neighbouring DIM are as high as respective genetic correlations (Figure 4), amounting to 0.99 for DIM located approximately every 20 days at the beginning and the middle stage of lactation, and every 15 days at the end stage. However, comparing to genetic correlations, the decrease in correlation for more distant DIM is faster. Figure 10 shows correlations of daily permanent environmental effects for milk yield between two lactations. Compared to

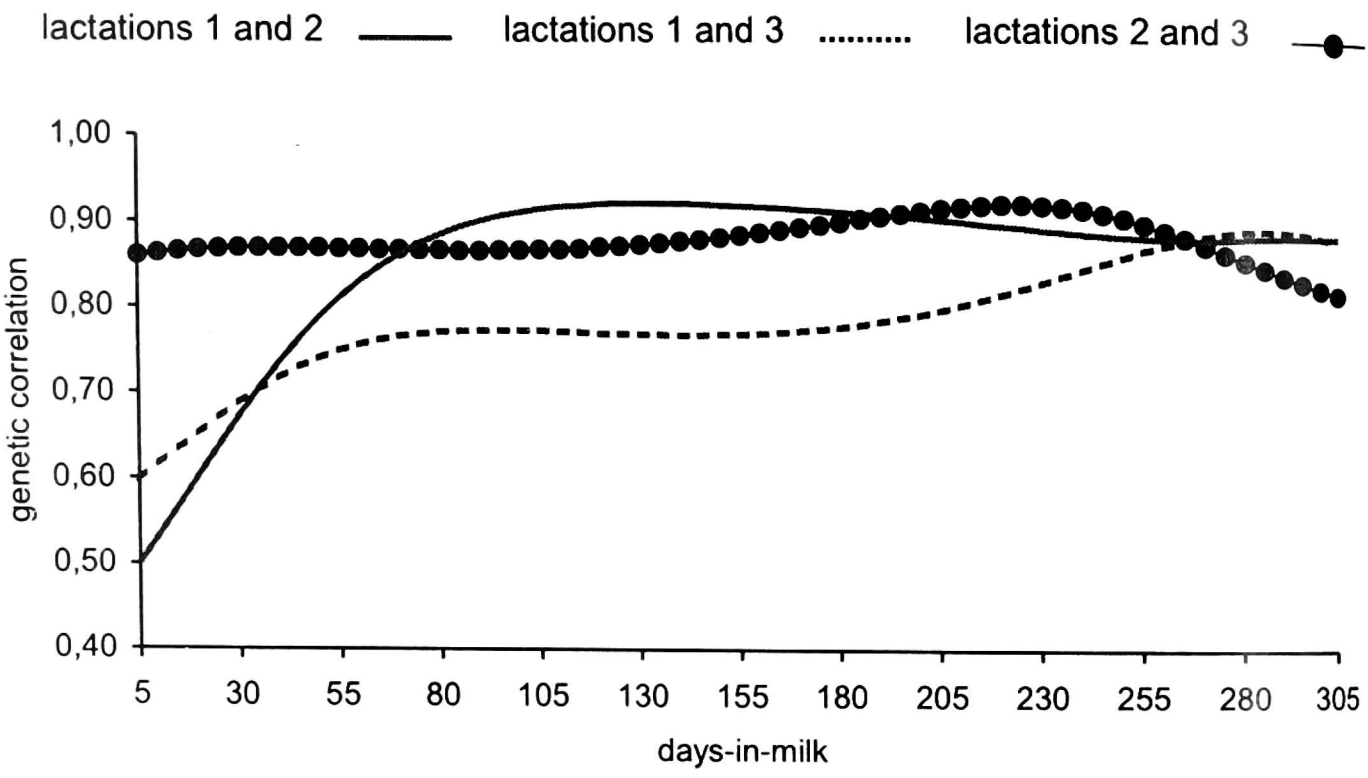

Figure 7. Genetic correlation estimates between the same DIM of two lactations for daily milk yields

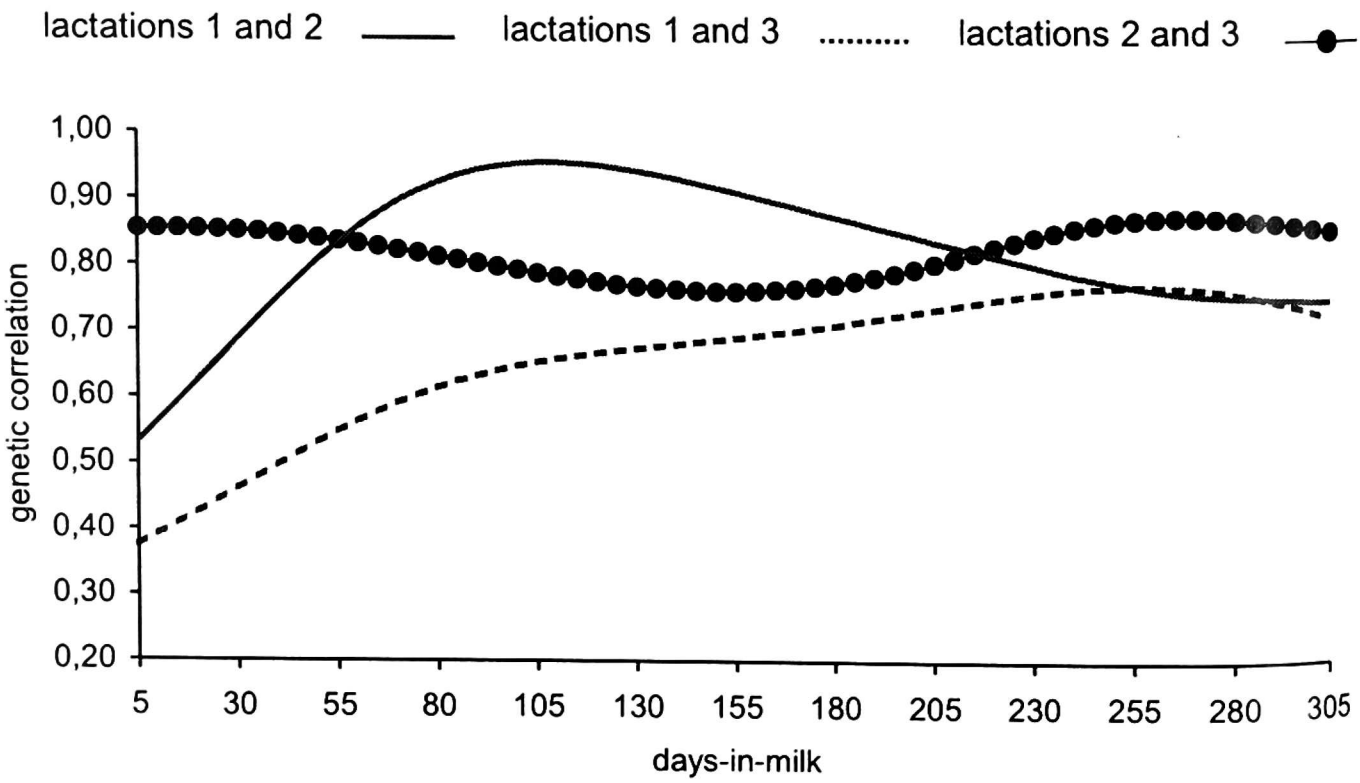

Figure 8. Genetic correlation estimates between the same DIM of two lactations for daily fat yields

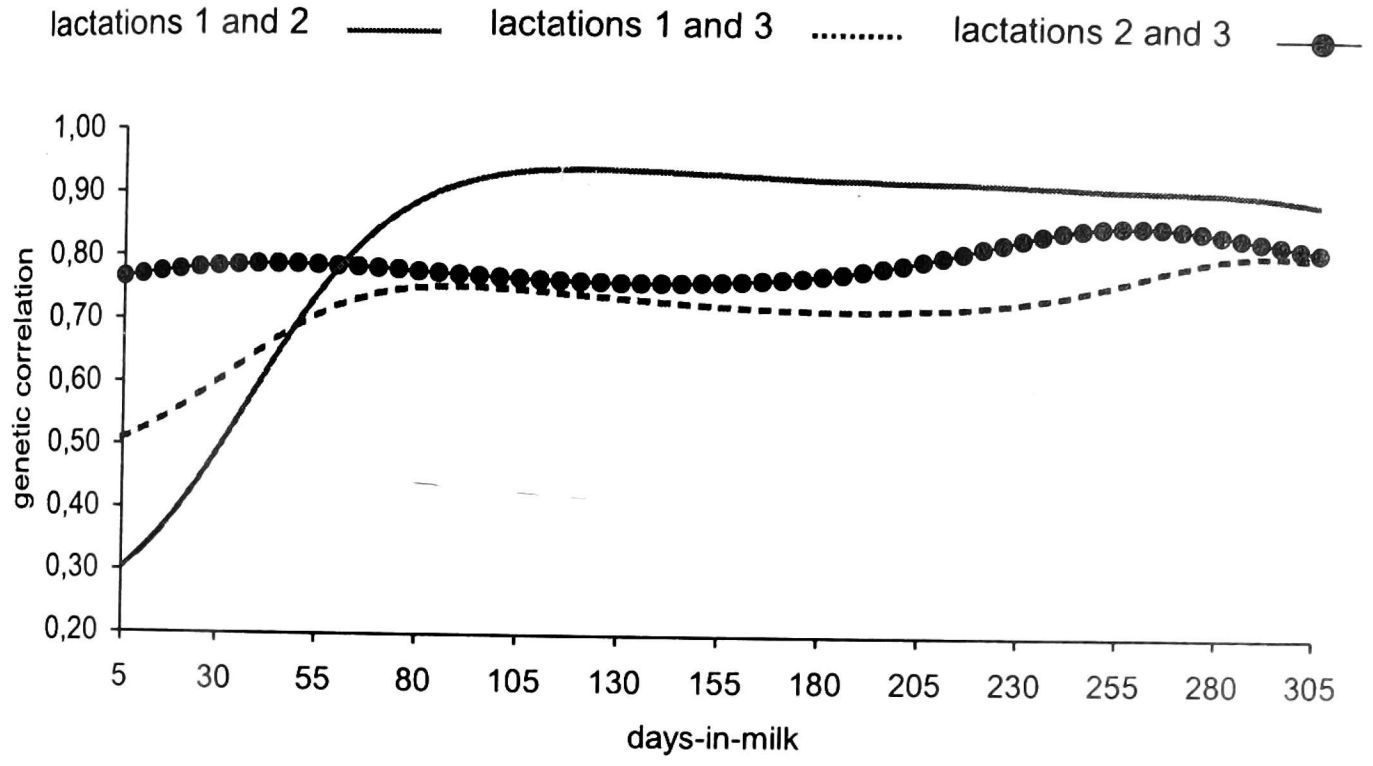

Figure 9. Genetic correlation estimates between the same DIM of two lactations for daily protein yields

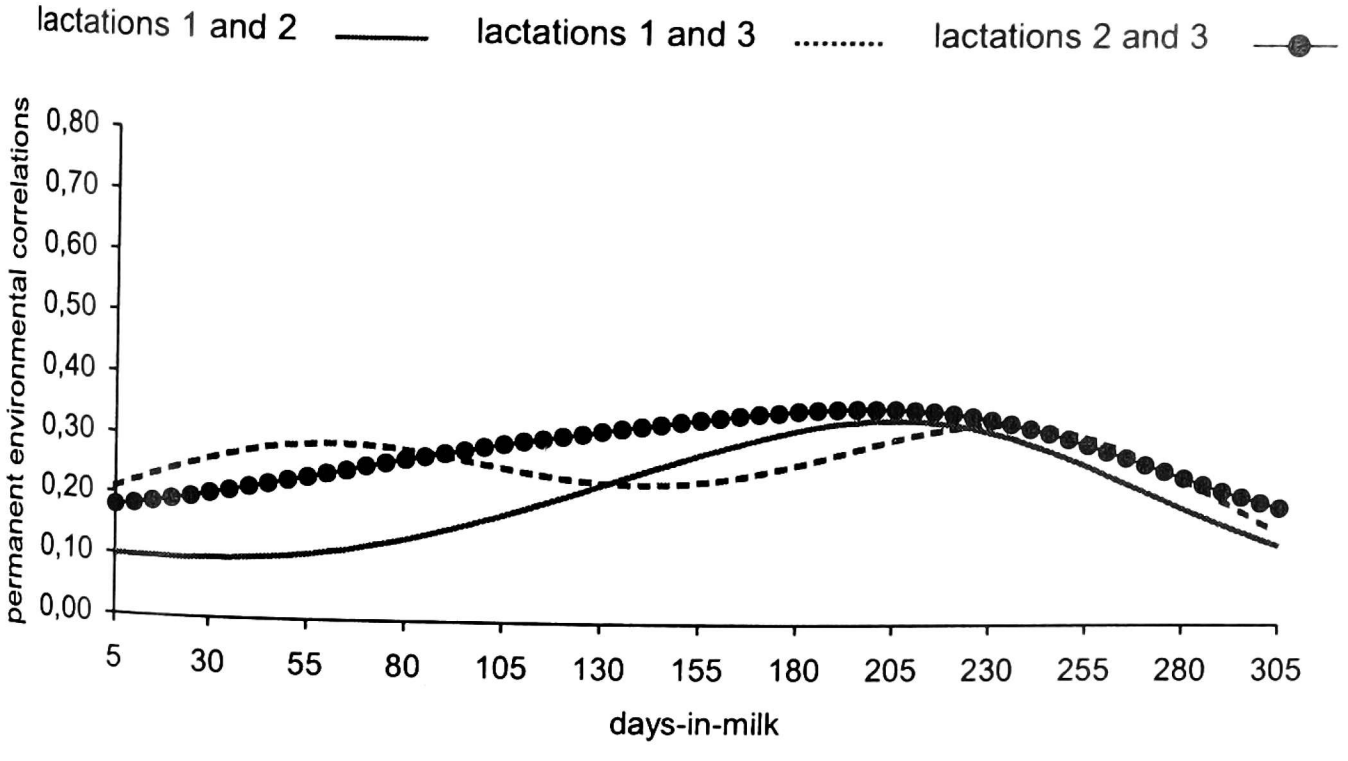

Figure 10. Permanent environmental correlation estimates between the same DIM of two lactations for daily milk yields

daily genetic correlations (Figure 7), permanent environmental effects have much lower correlations, ranging from 0.10 between DIM 5 and DIM 50 for the first and the second lactations, to 0.35 between DIM 170 and DIM 220 for the second and third lactations.

Discussion

The estimation of parameters of a random regression model has shown to be a computationally and methodologically difficult task (STRABEL, MISZTAL 1999). The known drawbacks include low heritability estimates as obtained by MÄNTYSAARI (1999), and TIJANI et al. (1999), high heritability estimates as in JAMROZIK and SCHAEFFER (1997), POOL et al. (2000), and low or even negative genetic correlation between early and late lactation stages (JAMROZIK, SCHAEFFER 1997). The covariance function approach with an iterative two-stage algorithm of LIU et al. (2000) used in this study seems to be robust towards the structure of the data (i.e. a large number of highly correlated parameters) and the computational difficulties (i.e. a large number of records required for the accurate estimation). Originally, it has been applied to the estimation of a random regression test day model parameters for a German Holstein population. Here parameters are estimated for the population of Polish Black-and-White dairy cattle. The values appear to follow the empirical expectations, thus giving credibility to the results obtained. Later lactations have somewhat lower heritabilities than the first one, and no evident difference in heritability was observed between the second and the third lactation. Among the three production traits, milk yield has the highest heritability. Compared to a one-step approach, the covariance function approach does not model each DIM as a separate trait, so that the parameter estimates reflect averages over all DIM from the same lactation stage. Thus, curves estimated in this way are likely to exhibit lower variation than curves estimated under a one step approach. Having in mind the difficulties in obtaining accurate parameter estimates while applying a one-step approach to large data sets, it seems reasonable to reduce the number of estimated parameters by averaging over closely neighbouring DIM.

Comparison of daily estimates

Random regression model parameters for a very similar, but smaller population (the same breed and breeding region, overlapping cow birth years) were recently estimated by STRABEL and MISZTAL (1999) using a one-step animal model approach.

For these two populations no marked differences in the level of daily heritabilities is observed. However, there are some differences in the shape of daily heritability estimates curve. The unexpected pattern found in both studies concerns highly increased heritabilities at the very early and/or the very late stage

of lactation. This is especially profound for the one-parity model of STRABEL and MISZTAL (1999). The least increase is observed in our study. A possible explanation for such results is the lack of phenotypic information necessary for an accurate estimation of (co)variances at the beginning and end of lactation. A one-parity model of STRABEL and MISZTAL (1999) utilises 96 798 test day records, their two-parity model 134 871 test day records, and our three-parity model 338 808 test day records. Moreover, the characterisation of 1 246 622 test day records from the population available for our study indicates that there are less tests available for the end part of a lactation, i.e. approximately after DIM 245 (see Figure 3 in SZYDA, LIU 1999). This is a similar stage of lactation at which we also observe an increase in daily heritability estimates. Other putative reasons for the phenomenon are constraining residual variance or properties of polynomial regression, as pointed by OLORI et al. (1999).

Comparing daily estimates of genetic correlations between the first and the second lactation, our results show slightly more variation in the range of estimates as in STRABEL and MISZTAL (1999), but again a very similar pattern of correlations is observed in both studies. Also curves describing genetic correlations of DIM 30, DIM 150, and DIM 250, respectively, and the remaining part of the first lactation's milk yield are almost the same for both populations. An unexpected pattern is observed for daily permanent environmental correlations, in which in the first half of lactation correlations between the first and third parities are generally higher than between the first and second parity.

In conclusion, the above comparison of the results of the one-step animal model approach of STRABEL and MISZTAL (1999) with the two-step sire model approach of LIU et al. (2000) shows that both methods stay in close agreement in terms of estimating daily (co)variance parameters. Using an iterative two-stage algorithm and a sire model instead of an animal model allows for analysis of much more data (see LIU et al. 2000 for results based on 17 161 866 test day records for a German Holstein population), which appears to be important for improvement of the accuracy of modelling of the (co)variances at the beginning and end of a lactation.

Comparison of 305-day estimates

The latest 305-day lactation estimates for the Polish Black-and-White population based on the multivariate animal lactation model are available through the INTERBULL (2000). Comparison of the heritability estimated by both methods reveals close similarities: milk yield has the highest heritability and protein yield has the lowest among the traits, the first lactation has the highest heritability and the third has the lowest. In general, heritabilities resulting from a lactation-based model are lower than from a test day model, because the latter one accounts for environmental effects such as feeding specific to each test day, while the lactation model can only consider effects averaged for the whole course of lactation.

Considering genetic correlations between parities, as expected, for all three traits correlation between the first and the third parity is the lowest under lactation and test day model, and both groups of estimates are very similar. In contrast to the test day model, the lactation model estimates no large differences in correlations between the subsequent parities (i.e. 1&2, 2&3).

In conclusion, the covariance function approach with an iterative two-stage algorithm (LIU et al. 2000) is an efficient method for the estimation of parameters of a random regression test day model. Among the most important advantages of this approach is its ability to utilise large data sets, which is a prerequisite for accurate estimation, especially at the beginning and end stages of lactation.

Acknowledgements. Zengting LIU (VIT, Verden, Germany) is gratefully acknowledged for technical assistance and discussions. I thank both reviewers for helpful comments. The study was financially supported by the State Committee for Scientific Research grant no. 5 P06D 033 14.

REFERENCES

ALI T.E., SCHAEFFER L.R. (1987). Accounting for covariances among milk yields in dairy cows. Can. J. Anim. Sci. 67: 637-644.

INTERBULL (2000). National genetic evaluation programmes for dairy production traits practised in Interbull member countries 1999-2000. Bulletin 24: 78-79.

JAMROZIK J., SCHAEFFER L.R. (1997). Estimates of genetic parameters for a test day model with random regression for yield traits of first lactation Holsteins. J. Dairy Sci. 80: 762-770.

KIRKPATRICK M., HILL W.G., THOMPSON R. (1994). Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. Genet. Res. Camb. 64: 57-69.

LIU Z., REINHARDT F., REENTS R. (2000). Estimating parameters of a random regression test day model for first three lactation milk production traits using the covariance function approach. Interbull Bulletin 25: 74-80.

MÄNTYSAARI E.A. (1999). Derivation of multiple trait reduced rank random regression model for the first lactation test day records of milk, protein and fat. 50th Annual Meeting of EAAP, Zurich, August 22-26.

MISZTAL I., STRABEL T., JAMROZIK J., MÄNTYSAARI E.A., MEUWISSEN T.H.E. (2000). Strategies for estimating the parameters needed for different test-day models. J. Dairy Sci. 83: 1125-1134.

NEUMAIER A., GROENEVELD E. (1998). Restricted maximum likelihood estimation of covariances in sparse linear models. Genet. Sel. Evol. 30: 3-26.

OLORI V.E., HILL W.G., McGUIRK B.J., BROTHERSTONE S. (1999). Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. Livest. Prod. Sci. 61: 53-63.

POOL M.H., JANSS L.L.G., MEUWISSEN T.H.E. (2000). Genetic parameters of Legendre polynomials for first parity lactation curves. J. Dairy Sci. 83: 2640-2649.

REENTS R., JAMROZIK J., SCHAEFFER L.R., DEKKERS J.C.M. (1995). Estimation of genetic parameters for test day records of somatic cell score. J. Dairy Sci. 78: 2847-2857.

ROYLE J.A., BERLINER L.M. (1999) A hierarchical approach to multivariate spatial modelling and prediction. J. Agric. Biol. Env. Stat. 1: 29-56.

STRABEL T., MISZTAL I. (1999). Genetic parameters for first and second lactation milk yields of Polish Black and White cattle with random regression test-day models. J. Dairy Sci. 82: 2805-2810.

SZYDA J., LIU Z. (1999) Modelling test day data from dairy cattle. J. Appl. Genet. 40: 103-116.

TIJANI A., WIGGANS G.R., VANTASSEL C.P., PHILPOT J.C., JENGLER N. (1999) Use of (co)variance functions to describe (co)variances for test day yield. J. Dairy Sci. 82(Jan.). Online: 225.