# Fast and flexible program for genetic evaluation in dairy cattle[1]

*Martin Lidauer and Ismo Strandén*
*Agricultural Research Centre - MTT, Animal Production Research, FIN-31600 Jokioinen, Finland*

## Abstract

A general purpose iteration on data BLUP-program (MiX99) was developed. Its fast performance makes continuous evaluation feasible, even when using a multiple-trait random regression test-day model. Reduction in computing time is due to four developments: preconditioned conjugate gradient as solving algorithm, a new technique for iteration on data, data compression, and ordering of equations by animal families. Three random regression test-day models with 0.25, 7.28, and 18.1 million unknowns in the mixed model equations were solved to compare MiX99 with the former available software. Total computing time (wall clock time) for first, second, and third model was 0.2, 3.3, and 9.2 hours, respectively, whereas corresponding values using the former software were 1.0, 19.7, and 172.4 hours. Results emphasize the superiority of the new implemented methods, especially when complexity of the model increases. The high performance was not impaired by the generality of the program, which allows a wide range of models.

## 1. Introduction

Upgrading genetic evaluations based on animal model to test-day model in dairy cattle leads to a manifold increase in computation. This is because test-day models include more effects, and also because number of records increases about ten times. In national evaluations number of test-day records and number of unknowns in the mixed model equations (MME) is likely to be beyond 10 million.

So far, iteration on data technique and algorithms, which have been found useful for animal models, were used for solving large test-day models (Reents et al. 1995, Jamrozik et al. 1997, Lidauer et al. 1998). These studies demonstrated that computations may last several days to obtain solution to the MME. Moreover, it might be difficult to assess the required number of iterations to meet convergence as indicated by Lidauer et al. (1999).

Recent studies introduced new developments suitable for iteration on data BLUP programs. Lidauer et al. (1999) advocated the use of preconditioned conjugate gradient (PCG) as solving algorithm. Strandén and Lidauer (1999) applied new techniques to enhance iteration on data procedure in association with PCG, and Lidauer and Strandén (1998) showed the usefulness of parallel computing. All these techniques reduce computing time considerably.

Aim of this work was to incorporate these new techniques into an iteration on data BLUP-program (MiX99). Furthermore, the program should be as flexible as its predecessor program DMUIOD (Lidauer et al. 1998). Performance of the new program is tested with three different random regression test-day models and compared with that of the former software.

[1] Presented at the international workshop on high performance computing and new statistical methods in dairy cattle breeding, Tuusula, Finland, March, 18-20, 1999

## 2. Computing methods

*Preconditioned conjugate gradient as solving algorithm*

The method of conjugate gradient solves the linear system **Ax**=**b** based on a geometric approach (Shewchuk 1994). In breeding value estimation **A** corresponds to the coefficient matrix of MME, **x** contains the solutions and **b** is the right-hand side of MME. When preconditioning, an equivalent system, $\mathbf{M}^{-1}\mathbf{Ax}=\mathbf{M}^{-1}\mathbf{b}$, is solved, where $\mathbf{M}^{-1}$ is a symmetric positive definite preconditioner matrix that approximates $\mathbf{A}^{-1}$. Together with a suitable preconditioner convergence rate is much better than that achieved by commonly used algorithms for solving MME (Lidauer et al. 1999). MiX99 creates a **M**-matrix which has diagonal blocks of the coefficient matrix **A**. Small fixed effects build one diagonal block of size equal to the number of fixed effect equations. For all other effects in the model for each level a diagonal block is built, which size is equal to number of equations in the level. Implementation of the PCG algorithm into an iteration on data program requires to keep four vectors, of size equal to the number of unknowns in the MME, in memory and to read the data and the preconditioner matrix once per iteration round. The algorithm does not require additional tuning parameters like relaxation factors.

*New technique for iteration on data*

The major computational task in PCG is the multiplication of the coefficient matrix with a vector each round of iteration. Therefore, when using iteration on data (IOD) technique, all data records must be read and processed. IOD technique requires for each record a certain amount (*N*) of floating point operations to calculate the record´s corresponding part of the product coefficient matrix times vector. Using standard IOD techniques, *N* follows an exponential function of the number of effects in the statistical model. For example for the DMUIOD program, $N=3f^2+2f-b$, where *f* is the number of effects in the model, and *b* is depending on the diagonal block structure and number of observations per animal. A new technique, introduced by Strandén and Lidauer (1999), reduces floating point operations considerably. Moreover, with increasing complexity of the statistical model *N* increases linearly only. Applying this technique, $N=2(2f+t^2)$, where *f* is as given and *t* is the number of traits in the model.

*Data compression*

Complex models with many effects may yield large iteration work files, which increases time consuming I/O-operations during the iteration process. Data compression is based on the concept of avoiding redundant information in the data files. Considering the structure of more complex statistical models and typical dairy cattle data, the following strategies were implemented. Pedigree information is stored as a separate iteration work file. If several effects in the model have the same class code, only one equation identification number is stored in the iteration file. This is possible by properly ordering the equations. Covariables, or a part of them, may be placed in a table rather than reading them from the iteration file. In case different traits are measured at different time (e.g. different lactation), observations of the traits may be grouped by the time component to avoid storing large amount of dummy variables for missing information. These techniques reduce the size of input data and iteration files considerably.

*Animal family blocks*

Equations of the MME are ordered by animal families. An animal family block comprises of all equations of the MME which are closely linked to each other. For instance in dairy cattle all fixed and random effects which belong to the same herd built a block of

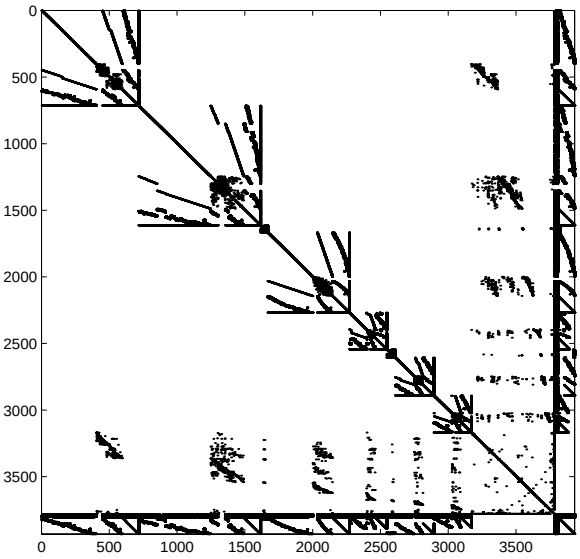


Figure 1. Example of the structure of the coefficient matrix when sorting equations of the mixed model equations by animal families. Each dot is a non-zero element in the coefficient matrix.

equations. Fixed and random effects, which are present in different herds, e.g., age effect or sire effect, are combined in common blocks of equations. This ordering gives data locality during the iteration process (Figure 1), which enhances computing speed, and which is essential when using parallel computing.

*Flexibility of the program*

MiX99 supports all those models that were in the DMUIOD program (Lidauer et al. 1998) plus some new ones. It allows multiple trait models. For each trait the statistical model can be defined separately. There is no limit on the number of following effects: fixed effects, regression effects, regression effects associated with fixed effects, random effects, regression effects associated with random effects, maternal effects, and paternal effects. MiX99 accommodates sire and animal models, and allows grouping of unknown parents by phantom parent groups in case of an animal model. It supports repeatability models and, as a new feature for multiple-trait models, it allows defining for each effect whether or not observations are treated as repeated observations or observations of different traits. This feature is useful for certain types of reduced rank random regression test-day models.

## 3. Application

Three random regression test-day models were used to compare performance of MiX99 with the former software DMUIOD (Lidauer et al. 1998), which uses block Gauss-Seidel iteration for the herd effect and block 2$^{nd}$ order Jacobi iteration for all other effects. Model 1 (M1) and model 2 (M2) were the same single trait random regression test-day models for milk yield as described by Lidauer et al. (1999) including 0.24 million records and 38,254 animals, and 6.7 million records and 1,099,730 animals, respectively. Model 3 (M3) was a multiple-trait random regression test-day model for first lactation milk, protein, and fat yield with 8.4 million records and 1,343,337 animals and had the following form:

$$y_{tijklmnop} =$$

$$herd_{ti} + ym_{tj} + \sum_{w=1}^{4} \mathbf{v}_{DIM}(w)\mathbf{s}_{tk}(w) + age_{tl} + dcc_{tm} + htm_{tn} + \phi'\pi_{DIMt}\mathbf{p}_o + \phi'\alpha_{DIMt}\mathbf{a}_o + e_{tijklmnop},$$

where $y_{tijklmnop}$ is record p made on cow o in herd i on days in milk DIM for trait t; $ym_{tj}$ is fixed year-month effect, **v** is a vector with four covariables for days in milk DIM, **s** contains four regression coefficient for season subclass k, $age_{tl}$ is fixed calving age effect, $dcc_{tm}$ is fixed days carried calf effect, $htm_{tn}$ is random herd-test-month effect, $\phi'\pi_{DIMt}$ is a vector with five covariables associated with permanent environment effect, for days in milk DIM and trait t. $\mathbf{p}_o$ is a vector with five random regression coefficient describing the permanent environment effect for cow o, $\phi'\alpha_{DIMt}$ is a vector with five covariables associated with animal effect, for days in milk DIM and trait t. $\mathbf{a}_o$ is a vector with five random regression coefficient describing the animal effect of cow o, and $e_{tijklmnop}$ is the residual. Note that $\mathbf{p}_o$ and $\mathbf{a}_o$ are the same for all three traits. Model M1, M2, and M3 included 16, 16, and 57 different effects and had 0.25 million, 7.3 million, and 18.1 million equations in the MME, respectively.

Comparison of the programs was based on wall clock time to prepare data for the solver, wall clock time of the solver until convergence, number of iterations until convergence, and size of data files when solving M3. Convergence criteria were the same as found for M2 by Lidauer et al. (1999). These were relative average difference between right-hand and left-hand side being smaller than $10^{-14}$ for MiX99, and relative average difference between solutions of consecutive iterations being smaller than $1.7^{-10}$ for DMUIOD. Comparison was carried out on a Cycle Ultra AXmp computer with two GB RAM.

## 4. Results

Time to prepare data for the solver was same or shorter with MiX99 for all three models (Table 1). Advantage of MiX99 over DMUIOD increased with increasing size of data and model. The shorter preparation phase of MiX99 was due to ordering of equations by animal families, which reduced sorting work of coefficients of the diagonal block matrix, and because of the smaller data files, which reduced I/O-operations.

Table 1: Wall clock time (in minutes) to solve three mixed model equations of different size (M1, M2, M3), and number of iteration until convergence, by different computing software.

| | MiX99 | | | DMUIOD | | |
|---|---|---|---|---|---|---|
| | M1 | M2 | M3 | M1 | M2 | M3 |
| Time to prepare data for solver | 2 | 36 | 62 | 2 | 66 | 189 |
| Time for solving until convergence | 7 | 161 | 490 | 55 | 1116 | 10152 |
| Number of iterations | 212 | 149 | 167 | 438 | 305 | 380 |

Table 2: Number of floating point operations to calculate a record's corresponding part of the matrix multiplication coefficient matrix times vector by different computing software.

| Model | Number of Traits | Effects in Model | MiX99 | DMUIOD |
|---|---|---|---|---|
| M1, M2 | 1 | 16 | 66 | 573 |
| M3 | 3 | 57 | 246 | 8487 |

Table 3: Size of files (in megabytes) when analyzing 8.4 million test-day records by a multiple-trait random regression test-day model (M3) with 18.1 million unknowns in the mixed model equations by different computing software.

| | MiX99 | DMUIOD |
|---|---|---|
| Pedigree file | 50 | 50 |
| Data file | 436 | 1576 |
| Iteration work files | 811 | 2747 |
| Other work files | 95 | 85 |
| Solution files | 298 | 298 |

Time of solver until convergence was clearly smaller for MiX99 (Table 1). Reduction in computing time was largest for the complex model M3. This was mainly due to the new IOD technique, which required only 0.3% of floating point operations of that required by DMUIOD to process one record (Table 2). Also data compressing had important influence on execution time. For M3, size of iteration work files for MiX99 were 30% of the size when using DMUIOD (Table 3). The reduction in size was achieved by a more efficient storage of equation identification numbers for small fixed effect and by the possibility to store covariables in a table rather than reading them from the iteration work files.

Convergence of solutions was reached with about half amount of iterations compared to the former software DMUIOD (Table 1). Fast convergence of PCG algorithm was also reported by Carabaño et al. (1989). Quick convergence makes the algorithm an attractive alternative for solving MME.

## 5. Conclusion

Implementation of new techniques into an iteration on data BLUP program lead to a manifold reduction in computing time. Advantage was largest for complex models. Solving a multiple-trait random regression test-day model with 57 effects in the model and 18.1 million unknowns in the MME could be accomplished in 9 hours of calculations. The developed software makes continuous estimation for Finnish dairy cattle based on a multiple-trait random regression test-day model possible.

## Acknowledgment

The development of the iteration on data program MiX99 was supported by the European Commission (Esprit project 23770). The project was carried out in co-operation with MTT, CSC (Centre for Scientific Computing at Espoo, Finland), FABA (Finnish Animal Breeding Association) and the Agricultural Data Processing Centre.

**References**


Carabaño, M.J., Najari, S. and Jurado, J.J. 1992. Solving iteratively the mixed model equations. Genetic and numerical criteria. *Proc. 43th Ann. Meeting EAAP*, Madrid, Spain. pp 258-259.

Jamrozik, J., Schaeffer, L. R., Liu, Z. and Jansen, G. 1997. Multiple trait random regression TD model for poduction traits. *INTERBULL Bulletin* No. 16, 43.

Lidauer, M., Mäntysaari, E.A., Strandén, I., Kettunen, A. and Pösö, J. 1998. DMUIOD: A multitrait BLUP program suitable for random regression testday models. *Proc. 6th World Congr. Genet. Appl. Livest. Prod.*, Armidale, NSW, Australia, XXVII:463-464.

Lidauer, M. and Strandén, I. 1998. Experiences in using parallel computing to solve large test-day models. *Proc. 49th Ann. Meeting EAAP*, Warsaw, Poland. No 4: pp 46.

Lidauer, M., Strandén, I., Mäntysaari, E.A., Pösö, J. and Kettunen, A. 1999. Improving convergence of iteration on data by preconditioned conjugate gradient applied on large test-day models. Submitted.

Reents, R., Dekker, J.C.M. and Schaeffer, L.R. 1995. Genetic evaluation for somatic cell score with a TD model for multiple lactations. *J. Dairy Sci*. 77:2671.

Shewchuk, J.R. 1994. An introduction to the conjugate gradient methods without the agonizing pain. Tech. Rep. CMU-CS-94-125. Carnegie Mellon University, Pittsburgh.

Strandén, I. and Lidauer, M. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. Submitted.