# Effect of modelling unknown parent groups and metafounders on the historical genetic trend of fertility traits

***A. Legarra[1], P.M. VanRaden[2]***

*[1] Council on Dairy Cattle Breeding, 4201 Northview Drive, 20716 Bowie MD, USA*

*[2] U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705-2350, USA*

## Abstract

Unknown parent groups (UPG) allow modelling unobserved selection in unknown parents. Because UPG are defined at least partly by year of birth, biased estimates can also bias estimates of environmental trends like management. Different modelling of UPG can reduce biases and standard errors. The 4 fertility traits daughter pregnancy rate (DPR), cow conception rate (CCR), heifer conception rate (HCR) and early first calving (EFC) in US dairy cattle make a good study case, because those have been affected by selection on correlated traits such as milk yield and have greatly differing recording patterns. Traits DPR and CCR are strongly correlated but DPR was recorded since ~1960 and CCR was recorded since ~2000. For missing traits, current traditional evaluation compress UPG definitions for missing years to avoid solving for UPGs with no direct information, and treat UPG as correlated across traits but uncorrelated across years (RandomUPGs). New models included: Fixed UPGs; Metafounders fitting average coancestry across UPGs based on year of birth (MFDeltaF); or including expected magnitude of change due to selection on a correlated trait (MFDeltaG). The data set consisted in 94 million records with potentially large numbers of missing values depending on trait and year, a pedigree including 94 million animals. Genetic evaluations were by BLUP and results are presented for Holstein. In all cases UPGs are treated as "a priori" correlated across traits. Genetic trends resulted in all cases in a fast decrease of DPR from 1960 until 2000. For DPR, this descent was most pronounced with RandomUPGs, closely followed by MFDeltaF and MFDeltaG, which yielded slightly less change because the inclusion of average coancestry results in smaller a priori changes. Similar trends but with larger differences across methods were observed for the correlated trait CCR, where the trend is inferred from correlations because of absence of records. Trends from 2000 to 2020 for both CCR and DPR were positive, with MFDeltaF showing slightly faster increases. Solutions of UPGs/MFs were most noisy with FixedUPGs, followed by RandomUPGs, followed by MFDeltaF which was the smoothest. Overall, for traits with years of missing records and with selection due to correlated traits not included in the data, modelling UPGs as random, and possibly correlated across years, is useful for correct genetic trends.

**Key words:** Unknown parent groups, fertility traits, genetic trends, metafounders

## Introduction

Pedigrees are usually incomplete across all birth years in dairy cattle pedigrees and are classically modelled using Unknown Parent Groups (UPG). The theory of UPG (Masuda *et al.*, 2022) becomes more difficult in multiple trait situations with complex missing patterns. If fit as fixed, UPGs (**g**) are not *a priori* correlated to each other as shown in the pre-QP equations which include them as covariates of the form $(\mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{Q})\,\hat{\mathbf{g}} = \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}$ (Quaas, 1988) – however "contributions" from descendants of the groups do account for the covariance across traits. If UPGs are fit as random, usually $Var(\mathbf{g}) = \mathbf{I}\otimes\boldsymbol{G}_0$, which implies *a priori* covariance across traits ($\boldsymbol{G}_0$) but not across levels. However, in a population with steady genetic trend the average genetic value of missing parents is expected to evolve smoothly from one generation or year to the next. The situation in which levels of UPGs are

correlated across levels, $Var(\mathbf{g}) = \mathbf{\Sigma} \otimes \boldsymbol{G}_0$ (Masuda *et al.*, 2022) with $\mathbf{\Sigma}$ a (not diagonal) covariance matrix, has not really been studied. Equivalently, the notion of metafounders (MF) generalizes the use of UPGs to include changes in inbreeding due to (missing) relationships among (missing) parents, and allows to "refer" relationships to genotypes. MF is also conceived to better model across-breed relationships.

Fertility traits in US dairy cattle make a good study case for multiple traits and UPGs or MFs. Fertility traits evaluations are difficult in a multiple trait model with UPGs because of (1) low heritability with different lactation weights and data pattern for each trait; (2) correlated, negative genetic trends caused by selection for yield, which is not included in the multiple-trait evaluation; (3) correlations might change (natural mating vs. AI; hormonal treatments; heat detection) and (4) the latest UPG is also unstable because heifer fertility arrives before cow fertility.

This work analyses results of different modelling of UPGs and MFs on the genetic trends of fertility traits and compares the results with expectations based on genetic progress.

## Materials and Methods

Official data files from CDCB tri-annual all-breed BLUP evaluation of December, 2022 included 94 million records for four traits: daughter pregnancy rate (DPR) and early first calving (EFC), both recorded since 1960; cow conception rate (CCR) and heifer conception rate (HCR), both recorded since 2000. Our focus is on DPR and CCR with a high genetic correlation of 0.86. Missing records ranged from 4% for DPR to 87% for HCR.

We computed expected decrease in fertility from 1960 to 2000 for DPR and CCR due to selection on milk yield, based on negative correlation of -0.34, estimated $\Delta G$ of 4.2 genetic s.d. for milk yield, and genetic s.d. of 4.9 for DPR and similarly for CCR.

Pedigree included 94 million animals and 417 UPGs defined by breed, year of birth and pathway (sex of the animal with missing parent, sex of its ancestor, and foreign/local origin). The Holstein breed had 219 UPGs across 5 different pathways, where four pathways had 39 to 56 UPGs (roughly, but not always, every year) and unknown parents of foreign bulls had 12 UPGs. Smaller breeds had far fewer groups combined across years and pathways. The mininum number of offspring (not necessarily with record) to create an UPG was 5000.

Models included animal and permanent effects, contemporary groups (different per trait), heterosis and inbreeding. Multiple trait MME included ~800 million equations which were solved in ~8h using blup90iod3 from University of Georgia, with the PCG algorithm. The different models for UPGs and MFs are detailed next.

### *Models for UPGs and MFs*

We first try a model with fixed UPGs. The second model was "RandomUPGs" with $Var(\mathbf{g}) = \mathbf{I} \otimes \boldsymbol{G}_0$. Then we run two models with metafounders. Thus the third model (MFDeltaF) with $Var(\mathbf{g}) = \mathbf{\Gamma} \otimes \boldsymbol{G}_0$ was inspired by (Sorensen and Kennedy, 1983) which describe that the means $\mu$ of each generation have the following covariance structure:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \bar{A}_0 & \bar{A}_0 & \bar{A}_0 & \dots \\ \bar{A}_0 & \bar{A}_1 & \bar{A}_1 & \dots \\ \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} \approx \begin{bmatrix} 0 & 0 & 0 & \dots \\ 0 & 2t_1\Delta F & 2t_1\Delta F & \dots \\ 0 & 2t_1\Delta F & 2t_2\Delta F & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}.$$

However, $\mathbf{\Gamma}$ obtained in this manner is not positive definite. A correct pseudo-inverse in this case is of the form, for invertible $\mathbf{B}$,

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}^{-} = \begin{bmatrix} \mathbf{1}'\mathbf{B}^{-1}\mathbf{1} & -\mathbf{1}'\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{1} & \mathbf{B}^{-1} \end{bmatrix}$$

But we did not attempt so. Instead, we used a version of **Γ** that is compatible with genomic relationships, i.e. (Wicki *et al.*, 2023)

$$\boldsymbol{\Gamma} = \begin{bmatrix} \Gamma_{1,1} & \Gamma_{1,1} & \Gamma_{1,1} & \dots \\ \Gamma_{1,1} & \Gamma_{1,1} + 2t_1\Delta F_\Gamma & \Gamma_{1,1} + 2t_1\Delta F_\Gamma & \dots \\ \Gamma_{1,1} & \Gamma_{1,1} + 2t_1\Delta F_\Gamma & \Gamma_{1,1} + 2t_2\Delta F_\Gamma & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

with $\Gamma_{1,1} = \frac{2}{nsnp}\left(2\boldsymbol{p} - \frac{\mathbf{1}}{2}\right)\left(2\boldsymbol{p} - \frac{\mathbf{1}}{2}\right)'$, $\boldsymbol{p}$ a row vector of base allele frequencies and $\Delta F_\Gamma = \Delta F_\gamma = \Delta F\left(1 + \frac{\Gamma_{11}}{2}\right)$. Matrix **Γ** was constructed within breed and pathway. The inverse of **Γ** is a tri-diagonal matrix linking each MF to its immediate neighbors. The value of $\Delta F$ was estimated to be 0.0014 per year. Inspection shows that this is almost identical to inverting a structure of the form $\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} + \mathbf{1}\mathbf{1}'k\right)$, with $k$ a large constant that gets confounded with the mean.

For the fourth model, MFDeltaG, given that trend of fertility traits was initially due to selection on the correlated trait milk yield, we also tried a version of the above that would include putative change $\boldsymbol{\Delta G}$, as follows:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & (\boldsymbol{t_1\Delta G})^2 & (\boldsymbol{t_1\Delta G})^2 & \dots \\ \mathbf{0} & (\boldsymbol{t_1\Delta G})^2 & (\boldsymbol{t_2\Delta G})^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

In this case, we used the regular Moore-Penrose $\boldsymbol{\Gamma}^-$ (not the optimal choice). The value of $\Delta G$ was estimated to be 0.034 genetic standard deviations/year, per the correlation of DPR with, and observed genetic trends for, milk yield.

## Results & Discussion

Estimated trends are presented in Table 1 and Figure 1. Genetic change in Table 1 was larger than expected, and all methods gave similar results for DPR (with actual 1960-2000 records) but not for CCR (no records in the period, inferred from genetic covariances). The expected genetic gain may have been underestimated because fertility also suffered from selection on "dairy form" (i.e. more angular) cows. In Figure 1 there is a genetic decrease in CCR followed by an increase. The trends differ at the beginning but as UPGs and MFs get more descendants and the database becomes larger, trends get closer to each other and are quite similar over the last 20 years when both traits have data. Note that for CCR the genetic trend 1960-2000 is entirely inferred from genetic covariances with DPR. The DPR phenotypic trend is partitioned into 53% genetic and 47% environmental trends (Figure 2), illustrating that deterioration of fertility is due to correlated response for selection for milk yield, but *also* to management changes (Lucy, 2001).

**Table 1.** Genetic change 1960-2000 for DPR and CCR

| | DPR | CCR |
|---|---|---|
| Phenotypic | -16.00 | - |
| Expected genetic | -6.86 | -5.17 |
| FixedUPGs | -9.23 | -9.03 |
| RandomUPGs | -8.56 | -8.33 |
| MetafoundersDeltaF | -8.21 | -7.93 |
| MetafoundersDeltaG | -8.38 | -6.49 |

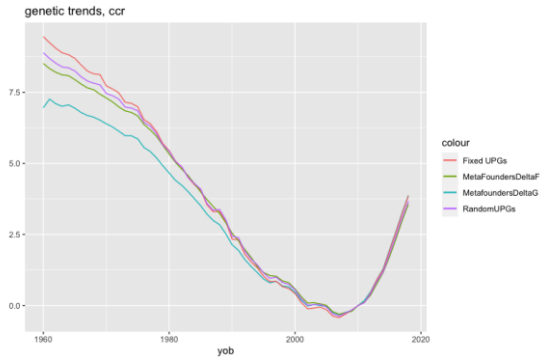


**Figure 1.** Estimated genetic trends for CCR under different models

Finally, Figure 3 shows estimates of UPGs/MFs for CCR and different models in pathway "21" (unknown sires of foreign dams) in Holstein. The estimates align well with those for DPR, even if CCR recording started in 2000. All models capture correctly the overall trend,

but Fixed UPG is very noisy, Random UPG is smoother and MFDeltaF gives a smooth, continuous line that provides the same long-term genetic trend

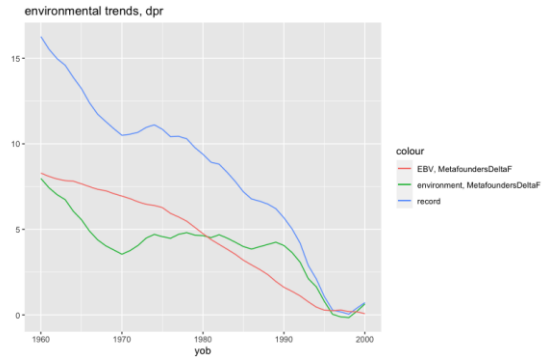


**Figure 2.** Decomposition of phenotypic into genetic and environmental trend, DPR.

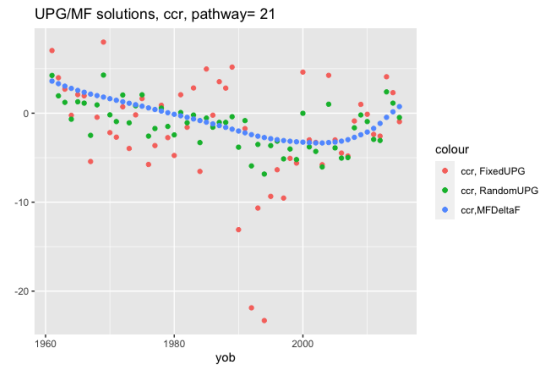


**Figure 3.** Different solutions of UPGs/MFs for the unknown sires of foreign dams pathway along time, CCR.

## Conclusions

Modelling differently UPGs and MFs does result in different genetic trends, although the effect was small in this large data set. More research is needed to ascertain the effect of this modelling in smaller data sets with unequal recording across traits. Fixed or Random UPGs or MFDeltaF provided meaningful results and are computationally easy. MFDeltaG gives less noisy solutions when there is not enough data. MFDeltaG is not recommended as the value for $\Delta G$ is trait dependent. MFDeltaG values for ΔG could in theory be estimated from each trait's covariance with the index but is less practical for most uses.

## Acknowledgments

AGIL staff and USDA funding of project 8042-31000-113-000-D, "Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals" and AL thanks Ezequiel Nicolazzi for lots of practical advice.

## References

Lucy, M.C. 2001. Reproductive Loss in High-Producing Dairy Cattle: Where Will It End? Journal of Dairy Science, 84: 1277–1293.

Masuda, Y., VanRaden, P.M., Tsuruta, S., Lourenco, D.A.L. and Misztal, I. 2022. Invited review: Unknown-parent groups and metafounders in single-step genomic BLUP. Journal of Dairy Science, 105: 923–939.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science, 71: 1338–1345.

Sorensen, D.A. and Kennedy, B.W. 1983. The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. Theoret. Appl. Genetics, 66: 217–220.

Wicki, M., Raoul, J. and Legarra, A. 2023. Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction. Journal of Dairy Science, 106: 5570–5581.