UDK 636.082.13;636.4

Izlaganje sa znanstvenog skupa
Conference paper

# DATA STRUCTURE AND MIXED MODEL TECHNOLOGY IN PIG BREEDING PROGRAMS

**E. Groeneveld, J. Spilke**

**Summary**

The effect of data structure on the BLUPs is investigated using a sample data set and simulation. Two herd sizes are chosen with 250 sows in 4 herds and 25 sows in 40 herds with 3 levels of across herd links. Same and different genetic bases were assumed. Genetic links increase selection response. BLUP does recover higher genetic values simulated in some of the base herds very quickly. It is concluded that BLUB can be viewed as a self optimizing system when selection is on BLUPs.

Keywords: BLUP, simulation, data structure, genetic ties.

## *Introduction*

While mixed model methodology has been becoming the state of the art in genetic evaluation in farm animals deficiencies in data structure continue to be a source of concern. Once a unique identification of current and historic animals is in place, BLUPs can be computed rather easily for a variety of models. Even if the latter are inadequate the printout of the BLUPs command the same apparent authority as models and runs that do not exhibit these deficiencies. Therefore, concern about insufficiencies in data structure (and modeling - which is not the objective of this paper) is indeed very appropriate. It is the objective of this paper to investigate the implications of data structure relative to one of the main objectives in animal breeding: achieving maximum genetic changes for a defined set of objectives.

Rad je priopćen na "6th World Congress on Genetics Applied to Livestock Production", Armidale, 1998.

E. Groeneveld, Institut für Tierzucht und Tierverhalten, FAL, D-31535 Neustadt, Germany,
J. Spilke, University of Halle, Ludwig-Wucherer-Straße 2-85, D-06108 Halle/Saale.

*The effect of bad data structure*

According to theory, BLUPs are comparable across time and space in a breeding population which is evaluated as a whole by mixed model methodology. It is this feature which makes BLUP so appealing and useful in breeding programs, as it holds the promise of maximizing the rate of genetic change if selection is on BLUPs. This applies to both the initial selection of young animals as well as the culling process. At this point where doubts about the comparability of the BLUPs arise, should the data structure be in fact insufficient. If this concern is well founded using BLUPs on inadequate data could possibly do more harm than not using it at all. Two areas of problems related to 'bad' structure can be identified. Firstly, with really bad data structure we may not be able to estimate (co) variance matrices because of some hidden dependencies in the mixed model equations or - more generally - because the mixed model equations are ill conditioned. This does not happen often, but it does. The second problem is potentially more severe because it can go undetected: estimates and predictions may be meaningless or invalid because of bad structure. It is this group of problems that we shall address in the following.

Clearly, *lack of genetic* ties ranks high on the list of potential deficiencies in data structure. Some attention has been focused on measures of connectedness (Foulley et al., 1990; Foulle et al., 1992; Kennedy and Trus 1993). Sometimes it is even assumed that BLUP can only be used in conjunction with artificial insemination (AI). Other areas of potential problems are few observations within contemporary group, few or distant links between field and station estimating fixed effects with no genetic ties between units.

*The model*

In the following we shall look at a simple statistical model using a sample data set to describe the relation of between BLUPs and the data. Consider the following model:

$$y_{ij} = herd_j + animal_i + e_{ij} \qquad [1]$$

*herd* is taken fixed while *animal* is random with the normal genetic covariance structure derived from the pedigree. And

$$animal_i = 1/2\ (u_{sire} + u_{dam}) + a_i \qquad [2]$$

with $u$ being the breeding values of the parents and $a_i$ the Mendelian sampling of the animal i. Thus

$$y_{ji} = herd_j + 1/2\ (u_{sire} + u_{dam}) + a_i + e_{ij} \qquad [3]$$

The sample data set is given in Table 1. Running an animal model with a heritability of .25 on this dataset gives the solutions listed in Table 2. In the following we assume that selection is for animals with the highest BLUPs. Thus, conclusion pertain to this situation only.

Table 1. - SAMPLE DATA SET WITH PEDIGREE

| Animal | Herd | Backfat | Pedigree | | |
|---|---|---|---|---|---|
| a1 | h1 | 12 | a1 | s1 | d1 |
| a2 | h1 | 11 | a2 | s4 | d3 |
| a3 | h1 | 13 | a3 | s5 | d4 |
| a4 | h2 | 11 | a4 | s1 | d2 |
| a5 | h2 | 12 | a5 | s3 | d5 |
| a6 | h3 | 16 | a6 | 0 | 0 |
| | | | d5 | s1 | d1 |

*Observations on small contemporary groups*

Given the current model, BLUEs for the herds are simple arithmetic averages of the observations pertaining to that fixed effect. The BLUPs of the animals are functions in the deviations of the animal from the class mean which, of course, is an outcome of equation 1. Thus, with small numbers of observations within contemporary groups we have two tendencies: firstly, with one observation only the BLUP will always be zero, as can be seen from animal a6 which is the sole animal in herd h3. Thus, no matter how good that animals is, it will never get selected (provided the truncation point is above the average). Secondly, with increasing subclass numbers the deviation of an animal from the subclass mean will increase (with two observations the average is determined the candidate to an extent of 50%). This, in turn will lead to larger BLUPs for a given observation.

Animals a1, a2 and a3 originate in herd h1, while a4 and a5 are from herd h2. Although the difference in backfat is the same between animals a2 and a1 compared to a4 and a5 (1 mm) the differences in backfat is the same be BLUPs are .25 versus .2258. In the case of less information, BLUPs are regressed toward the mean. Thus, having more animals in a subgroup leads to a'spread' in the BLUPs. This is also expressed by the standard deviation of the estimate which is .91287 for all three animals in herd h1 and .94382 for the two animals in herd h2. (This shows nicely that accuracy of the estimate or prediction is automatically reflected in the BLUPs and underlines the theoretical result that consideration of accuracy cannot improve selection response). Therefore, small group sizes will have the effect of yielding BLUPs which are closer to

the average. If the data structure is mixed with regard to group size we can expect to select more animals from larger groups. Thus, including data with small contemporary groups should not lead to selecting the 'wrong' animals. Instead, these animals will simply be ignored.

Table 2. - BLUEs AND BLUPs AND STANDARD ERRORS OF THE ESTIMATES FOR THE SAMPLE DATASET (h2 = .25)

| Herd | BLUEs | $\hat{\sigma}$ BLUEs |
|---|---|---|
| h1 | 12.0000 | 1.1547 |
| h2 | 11.5000 | 1.4325 |
| h3 | 16.0000 | 2.000 |
| Animal | BLUPs | $\hat{\sigma}$ BLUPS |
| a1 | 0.00000 | 0.91287 |
| a2 | -.25000 | 0.91287 |
| a3 | 0.25000 | 0.91287 |
| a4 | -.11390 | 0.94382 |
| a5 | 0.11290 | 0.94382 |
| d5 | 0.03226 | 0.97482 |
| a6* | 0.00000 | 1.00000 |
| d1* | 0.03226 | 0.97482 |
| d2* | -.06452 | 0.98374 |
| d3* | -.12500 | 0.97895 |
| d4* | 0.12500 | 0.97895 |
| s1* | -.03226 | 0.97482 |
| s3* | 0.06452 | 0.98374 |
| s4* | -.12500 | 0.97895 |
| s5* | 0.12500 | 0.97895 |

*Observations on lack of genetic ties between herds*

As mentioned before lack of genetic ties across herds are often cited as a main obstacle to using BLUP in a population. This problem can be viewed from two sides. There is first intrinsic logic in this statement: if there are no genetic ties across herds we are apparently not dealing with one but with two or more populations. Lack of ties can thus be viewed as an expression of disinterest of breeders to use stock from outside. If this is really the case, there is clearly no point in running an across herd evaluation. Along the same line of argument: if selection would be done on the basis of an across herd evaluation (implying now across he use of stocks) we would have ties in the following

generation which seems to make the whole argument less threatening. We shall now investigate the error that might arise in an across herd genetic evaluation when basing across herd selection on those BLUPs with little or no ties among the herds.

*Assuming same genetic base*

In situations with little but some exchange of genetic material we can assume that the genetic level of the herds is not too different. However, genetic links among herds may still be few perhaps also because pedigrees beyond parents may not be available. This implies that the only information available originates from within the herd. Accordingly, with small groups sizes accuracy of prediction will be rather low. If on the other hand genetic links to other herds exist perhaps via AI or paternal halfsibs (and implying a larger number of offspring per sire), we can expect to increase the accuracy of prediction, because the animals evaluated will use testdata from relatives thereby increasing the accuracy and thus the variance of the BLUPs. Dataset 2 (Table 3) is intended to demonstrate this effect. We assume two herds $h_1$ and $h_2$. In each herd we have four animals from 2 dams. To assess the effect of genetic ties the four dams are mated to the two sires in two ways as indicated by the two last columns: the last has one sire per herd (for no ties) while the second last column uses both sires on both herds, thus connecting the two herds.

Table 3. - SAMPLE DATA SET 2 WITH PEDIGREE

| Animal | Herd | Backfat | Pedigree 1 and 2 | | |
|---|---|---|---|---|---|
| a1 | h1 | 10 | a1 | d1 | s1 |
| a2 | h1 | 11 | a2 | d1 | s2 |
| a3 | h1 | 10 | a3 | d2 | s1 |
| a4 | h1 | 11 | a4 | d2 | s2 |
| a5 | h2 | 10 | a5 | d3 | s1 |
| a6 | h2 | 11 | a6 | d3 | s2 |
| a7 | h2 | 10 | a7 | d4 | s1 |
| a8 | h2 | 11 | a8 | d4 | s2 |

The solutions of the two models are given in Table 4. As can be seen with genetic ties standard error of BLUPs is smaller compared to the no tie situation. Notice, that the total amount of information is the same for both situations. Creating ties across animals results in making test information available which has been generated (and paid for) elsewhere for ones own

purpose. It can thus be viewed as a way to increase ones own test capacity without paying for it. The resulting effect will be an increase in the variance of the BLUPs thereby increasing the proportion of animals above the truncation point of selection which in turn will increase the selection response.

Interestingly, also the standard error of the BLUEs, i.e. the fixed effects are reduced by creating links in the data set.

Table 4. - SOLUTIONS FOR MODEL WITH AND WITHOUT GENETIC TIES

| Herd | BLUEs | $\hat{\sigma}$ BLUEs | BLUEs | $\hat{\sigma}$ BLUEs |
|---|---|---|---|---|
| h1 | 10.50000 | 1.06066 | 10.50000 | 1.11803 |
| h2 | 10.50000 | 1.06066 | 10.50000 | 1.11803 |
| Animal | BLUEs | $\hat{\sigma}$ BLUPs | BLUPs | $\hat{\sigma}$ BLUEs |
| a1 | -.166667 | .920986 | -.071428 | .94966 |
| a2 | .166667 | .920986 | .071428 | .94966 |
| a3 | -.166667 | .920986 | -.071428 | .94966 |
| a4 | .166667 | .920986 | .071428 | .94966 |
| a5 | -.166667 | .920986 | -.071428 | .94966 |
| a6 | .166667 | .920986 | .071428 | .94966 |
| a7 | -.166667 | .920986 | -.071428 | .94966 |
| a8 | .166667 | .920986 | .071428 | .94966 |
| d1* | .000000 | .968242 | .000000 | .96822 |
| d2* | .000000 | .968242 | .000000 | .96822 |
| d3* | .000000 | .968242 | .000000 | .96822 |
| d4* | .000000 | .968242 | .000000 | .96822 |
| s1* | -.222222 | .942802 | .000000 | 1.00000 |
| s2* | .222222 | .942802 | .000000 | 1.00000 |

*Assuming same and different genetic base: a simulation*

In a genetic evaluation we usually assume that the base generation has the same genetic value with a known additive genetic covariance structure. If populations have developed separately with a different selection strategy in the past, base animals from different herds may certainly have a different genetic level.

Lack of ties among herds and together with different genetic levels among them are often considered obstacles of using BLUP for across herd evaluation in breeding programs. While this seems to be a contradiction in terms (one breeding program and no exchange of animals?) a simulation was performed to

assess the effect of doing BLUP in such a situation. The population size was 960 sows per generation with 80 boars. Each sow produced a maximum of 4 male and 4 female offspring with probabilities 0.0355, 0.1767, 0.4176, 0.3702 for 1 to 4 male or female offspring. This resulted in a first generation with 0 for the average of the genetic effects. The traits considered are given in Table 5.

Table 5. - TRAITS USED IN SIMULATION

| Traits | | Econ. weight | Genetic covariance matrix | | | | |
|---|---|---|---|---|---|---|---|
| Daily Gain Field | (DGF) | 0.06 | .26 | .15 | -.19 | 0 | 0 |
| Backfat Field | (BFF) | 4.72 | | .43 | .25 | .87 | 0 |
| Feed consumption | (FST) | -29.88 | | | .39 | .25 | .60 |
| Lean-Fat Ration | (LFST) | -157.37 | | | | .42 | .25 |
| pH Station | (pHST) | 85.48 | | | | | .21 |

Base animals are distributed randomly across herds. Here, multivariate BLUPs were computed using PEST and the covariance matrix given in Table 4. From the 3000 male and 3000 female offspring 960 females and 80 males were selected. Computations were done with the program by Mielenz (1994).

Two herd sizes were simulated: setup A had a represented a rather good herd size with close 250 sows (four herds) while setup B represented a much worse structure with only 25 sows for each of the 40 herds (see Table 6 for the design of the simulation). Three selection schemes were used across the herd structures: in I, ranking and selection was across herds, in II, ranking and selection was restricted: 50% of the animals were selected within herds. In version III selection was only done within herds.

In a second setup d, we simulated the effect of a different genetic level of the base animals: here, in half of the herds each female gets the selection response from 5 generations resulting from IA (DGF: 35g, BFF: -l.3mm, FST: -0.14kg, LFST: -0.1, PHST: 0.14).

Sires are assumed to be used in AI (except for version III which is within herd selection) which implies that only the base generation has no genetic ties. For the d version with different genetic bases it would have been ideal, to use only within herd selection in the first generation. Due to constraints in the simulation program this could not be done. However, we feel that is not a large deficiency: if it is decided to do an across herd genetic evaluation the intention has to be to also select animals across herds. If this is not the case, there is no point in performing an across herd evaluation. This implies that on the basis of the first across herd evaluation animals will also be selected across herds,

producing ties in the following generation, as our simulation assumes from the first generation on.

For each of the variants as given in Table 6. 5 generations are generated (i.e. 4 selection steps) with 100 replicates and 200 for the version IIIB.

*Results for equal base.* The effect of herd size on selection response depends obviously on selection scheme: if selection is across herds (I) the rather bad structure with 960 sows on 40 herds is nearly as good as having 4 herds only. However, when selection is within herds only, the total selection response drops to 86% of the best possible. Conversely, with large herds the effect of selecting across herds is much less pronounced. This indicates that more can gained from BLUP across herd evaluation in populations with small herds.

As to be expected, inbreeding is clearly largest with small herd size and selection within. With an inbreeding coefficient of nearly 35% for selection with small herds this value can be reduced to only about 8% when selection is done on BLUPs across herds, a value that is only insignificantly larger than the across large herd selection.

Interestingly, the variance of the selection response (over replicates) skyrockets in the within small herd selection scheme. With a value of more than 100 it is 10 times larger as compared to selection across herds. This means that the predictability of the selection response increase as across herd selection is extended.

Table 6. - RESULTS FROM SIMULATION AT GENERATION 5

| | Same genetic base | | | | | |
|---|---|---|---|---|---|---|
| | Inbreeding | | Aggr. genotype | | Variance of AG | |
| | A | B | A | B | A | B |
| I | 7.05 | 7.67 | 27.41 | 27.01 | 9.24 | 9.63 |
| II | 5.53 | 7.70 | 26.02 | 26.07 | 8.60 | 9.08 |
| III | 11.42 | 34.84 | 23.57 | 14.74 | 20.70 | 103.37 |
| | Different genetic base | | | | | |
| I | 5.00 | 5.11 | 36.80 | 36.50 | 7.18 | 9.88 |
| II | 5.82 | 7.90 | 36.16 | 36.94 | 9.96 | 9.58 |

*Results for different base.* The total selection response is much higher for this group. This indicated that BLUP was capable of retrieving the higher genetic level of half of the herds. On the basis of the whole population the superiority of the d version is $\frac{27.41}{4}$.

Had BLUP identified the genetic base completely we would expect the final selection response to be $\frac{27.41}{4}$=13.55 above the response from equal bases. The current superiority which is close to 10 at generation 5 indicates, that even with different bases across herd BLUP evaluation is capable of recovering the true genetic value to a large degree. Furthermore, the genetic response per generation is higher for the d version, indicating that after a few more generations the maximum attainable response from the higher genetic base will be reached. The differences between the selection schemes with different genetic bases (and ignoring them in the genetic evaluation) are rather small.

In general, it can be concluded that different genetic basis are no obstacle to using BLUP provided that selection is from there on across herds.

*Implications for the design of breeding programs*

Bad data structure rarely makes mixed model genetic evaluation of populations impossible. Thus, no minimum requirements can be set from the structure point of view. Bad data structure does, however, reduce the realized selection by resulting in a lower variance of the BLUPs than would be possible under a better structure.

*Herdbook type structure*

In Herdbook type breeding programs can this fact be used to optimize the structure of the population if selection is for the BLUPs well above the average. Herds with a better testing program and larger contemporary groups will-on average-exhibit a smaller prediction error variance and consequently a larger variance in the BLUPs. For a fixed truncation point at selection a larger proportion of animals from these herds will be selected. Thus, it seems most promising to establish a defined selection policy in a cooperative breeding program that puts the cut off point well above the population mean, instead of trying to reach an agreement on minimum herd size, minimum number of animals tested. While including 'badly' structured data in the genetic evaluation will bloat the dataset and therefore the computing time, noting can really be gained by excluding them. Thus, mixed model genetic evaluation can be considered a self optimizing system with regard to data structure.

*Commercial breeding program*

One major feature of commercial breeding programs can be seen in a centralized decision regarding selection decisions. Accordingly, no self optimization can operate. Instead, optimum structures have to be determined

explicitly. Clearly, use of AI with balanced usage of boars is very desirable to improve use of information via links across testing units. Also, putting more animals into expensive testing programs will increase the accuracy for the candidates to be evaluated. However, the proportion of this category can only be determined after a cost/benefit analysis, which goes well beyond the scope of this paper.

## REFERENCES

1. Foulley, J., J. Bopuix, B. Goffinet, J. Elsen (1990): In D. Gianola and K. Hammond, Editors, Advances in Statistical Methods for Genetic Improvement of Livestock, 277, Springer-Verlag Berlin.
2. Foulley, J. E. Hanocq, D. Boichard (1992): Genetic Sel. Evol. 24:315.
3. Kennedy, B. W., D. Trus (1993): J. Animal Science, 71. 2341-2352.
4. Mielenz, N. (1994): Programm Documentation.

## STRUKTURA PODATAKA I TEHNOLOGIJA MJEŠOVITOG MODELA U PROGRAMIMA UZGOJA SVINJA

### Sažetak

Istražuje se djelovanje strukture podataka na BLUP primjenom jednostavnog skupa podataka i simulacije. Izabrana su krda dviju veličina s 250 krmača u 4 krda i 25 krmača u 40 krda s vezom kroz krda na 3 razine. Pretpostavljene su iste i različite genetske osnove. Genetske veze povećavaju selekcijski odgovor / reakciju. BLUP vrlo brzo obnavlja više genetske vrijednosti simulirane u nekim osnovnim krdima. Zaključuje se da se BLUP može promatrati kao sustav koji se sam usavršava kad je selekcija na BLUP

Ključne riječi: BLUP, simulacija, struktura podataka, genetske veze.

Primljeno: 20. 10. 1998.