Benefits and limits of increasingly sophisticated models for genetic evaluation: the example of pig breeding

Jean Pierre Bidanel

6th World Congress on Genetics Applied to Livestock Production, Armidale, January 1998, 25, 577-584

BENEFITS AND LIMITS OF INCREASINGLY SOPHISTICATED MODELS FOR GENETIC EVALUATION : THE EXAMPLE OF PIG BREEDING

J.P. Bidanel

INRA, Station de Génétique quantitative et appliquée, 78352 Jouy-en-Josas, France

SUMMARY

Genetic evaluation of most farm animals is performed using more and more complex statistical and genetic models. Benefits and problems associated with this increased sophistication are reviewed and illustrated using pig breeding examples. Model checking and validation methods are briefly introduced.

Keywords : genetic evaluation, mathematical models, model checking, pigs

INTRODUCTION

The use of mixed model methodology (MMM) has become a standard in most farm animal species to estimate genetic parameters through Restricted Maximum Likelihood and estimate breeding values through Best Linear Unbiased Prediction (BLUP). MMM has considerably evolved since the first publication of Henderson (1949) and early applications to the evaluation of dairy bulls. The increasing power of computers and major advances in computational methods have led to the replacement of the simple sire models used in early years by increasingly complex and realistic models which allow a more efficient use of the data available for selection decisions. The tendency towards the use of increasingly sophisticated models has been greatly facilitated over the last few years by the availability of flexible computing softwares for estimating breeding values and genetic parameters. In counterpart, the advantages associated with the use of more complex models may be offset by increased computational difficulties and a reduced robustness. Benefits and problems associated with the use of increasingly sophisticated models for estimating genetic parameters and breeding values are reviewed and illustrated using examples encountered in pig breeding. Some elements about model checking and validation are also given.

MODELS USED IN PIG BREEDING

Pig genetic evaluation models reflect the large variety of situations encountered in pig breeding, with two extreme cases. On one side, national breeding schemes often lead to large scale evaluations involving many participants and different testing environments (i.e. central testing stations, selection and multiplication herds) and often encounter problems such as genotype x environment interaction, small contemporary group size, immigration, preferential treatments, variance heterogeneity. On the other side, breeding companies generally own populations with a more limited size (one to several hundreds of females) distributed in a limited number of nucleus herds with a better control of testing environment and immigration, but have to cope with problems such as inbreeding and inaccuracies of genetic parameter estimates. Yet, all pig breeding schemes share some common features such as the use of multiple trait selection

objectives and criteria, the necessary use of an « animal model » to adequately describe the data and, because boars and gilts are selected on a weekly basis, the continuous aspect of the genetic evaluation process.

Mixed model methodology has been introduced more recently in pig breeding than in many other farm animal species, as it was not until 1985 that the first major application of BLUP was reported (Hudson & Kennedy, 1985). Other routine applications of BLUP only started at the end of the eighties (e.g. Bampton, 1992). In each case, single trait breeding values were computed and then combined in an aggregate index. The availability of easy-to-use and flexible computer softwares allowing for multiple trait models in the early nineties has considerably enhanced the use of BLUP - animal models and of REML estimation of genetic parameters. They have made it possible to adequately describe the sophisticated structure of pig breeding data, i.e. to combine traits from several testing environments with different fixed and random effects for some or all of the traits.

Several potential improvements of models have been investigated over the last few years. Alfonso (1995) compared the interest of a model considering each parity record of a sow as a different vs a standard repeatability model. The impact of maternal effects on genetic evaluation and selection response for litter size was thoroughly investigated by Roehe and Kennedy (1993a;b). Estany & Sorensen (1995) and Frey *et al.* (1997) reconsidered the way (fixed or random) to account for the effect of contemporary group when group size is limited. Other improvements of animal models which are currently developed or should be considered in the near future have been extensively reviewed by Simianer (1994) at the last WCGALP meeting. Like in most farm animal species, major improvements in pig genetic evaluation should come from further improvements of the statistical models used, particularly to account for heteroskedasticity and analyze traits which are not normally distributed or involve non additive mechanisms and, above all, from the use of more realistic genetic models. For instance, heterogeneous genetic and residual variances between herds or group of herds have been evidenced by e.g. Bidanel *et al.* (1994) or See (1994). Heteroskedasticity also generally occurs when several populations are considered in a single genetic evaluation. BLUP properly accounts for heterogeneous variances across environments, provided that the true variances in each environment are known (Gianola, 1986), but applications to pigs genetic evaluation are still very limited. Traits such as prenatal and preweaning survival or sow longevity cannot adequately be described by usual mixed models. Non linear models such as the threshold or survival models have been developed for such traits - see Ducrocq (1990) for a review -, but pig breeding applications have until now been limited.

With regard to genetic models, important non additive effects may exist for some traits in pigs. The large heterosis effects obtained for litter size and, to a lesser extent, growth traits, suggest that noticeable dominance and/or epistatic effects might exist for these traits. Under the infinitesimal model, mixed model methodology can easily account for non additive gene effects (Henderson, 1984), as well as for cytoplasmic or imprinting effects (Kennedy *et al.*, 1990) in non inbred populations, although the estimation of the corresponding variance components is

often a formidable task. Then, more and more evidence is accumulating showing that loci of medium to large effects exist for economically important traits (Le Roy *et al.*, 1990; Andersson *et al.*, 1994). Mixed models of inheritance, which assume one or several identified segregating loci, plus an additional polygenic component, have been developed. When genotypes at each identified locus are known, they can be appropriately treated as fixed effects in standard mixed model techniques (Kennedy *et al.*, 1992). When only genotypes at linked markers are known, the uncertainty due to unknown haplotypes and recombination events has to be taken into account (e.g. Fernando & Grossman, 1989).

BENEFITS FROM USING INCREASINGLY SOPHISTICATED MODELS

General considerations. The desirable properties of BLUP (Henderson, 1984) only hold when the model used appropriately describes the observed variation, i.e. that the distributions specified are correct and that all environmental and genetic components of variation are considered. Problems may also occur when populations undergo selection or non random mating, which may result in severe biases in estimates genetic parameters and breeding values. Properly taking into account these phenomena generally requires to include all the information related to the selection process in the analysis (Im *et al.*, 1989). In most cases, this implies to use animal models which exhaustively utilize available pedigree and data information and, when several traits are selected, the use of multiple trait models. The expected benefits from a more appropriate model specification will be detailed below.

Multivariate animal models. The advantage of animal models as compared to more simple models such as sire, sire-paternal grand sire or sire-dam models have been extensively detailed (e.g. see Schmidt, 1988 or Foulley & Molénat, 1994) and will not be considered here. As reviewed by Ducrocq (1994), further advantages can be drawn from the use of multivariate animal models. First, multivariate models are required to properly account for selection bias when several traits are selected. Severe biases in estimated breeding values and genetic trends may occur when univariate models are used in such situations, especially when genetic and residual correlations differ (e.g. see Pollak & Quaas, 1981; Sorensen & Johansson, 1992). Multivariate models also make it possible to use extra information on correlated traits or of direct information on related animals when one or several traits from the selection objective are difficult or impossible to measure on selection candidates. Obvious examples in pigs are the use of backfat thickness to improve carcass lean content and of meat quality measurements on slaughtered sibs for carcass and meat quality measurements. They also allow to cope with genotype x environment interactions by considering performance data measured in different environments like different traits (e. g. growth rate and backfat thickness measured on the farm and in testing stations). Data measured on crossbred animals might be incorporated in genetic evaluation in a similar way. Then improved accuracy and data structure can be obtained from multiple trait models. The gain in accuracy strongly depends on the genetic parameters of the traits considered. It increases when the genetic correlation between traits is high, when heritabilities or/and genetic and residual covariances differ, when more than one random factor is considered and when full or half sib family size is small (Ducrocq, 1994). Improved data

structure mainly comes from nonzero residual covariances which create more ties between animal and fixed effects (Thompson & Meyer, 1986).

Structure of dispersion parameters. Ignoring the heterogeneity of variances may result in substantial losses in response to selection, in particular due to selecting too many individuals from the most variable environments (e.g. Hill, 1984). As indicated above, BLUP accounts for heterogeneous variances when these variances are known. However, the estimation of heterogeneous variances may be a very difficult task, especially when environmental cells are numerous and have a limited size. In such cases, the structural mixed linear model on log-variance components recently developed (e.g. Foulley & Quaas 1994) provides an appealing theoretical framework for identifying meaningful sources of variation of variance components. The consequences of ignoring other non additive effects such as epistatic, cytoplasmic or imprinting effects have not yet been thoroughly investigated, but ignoring them may also bias prediction of direct effects, in particular because these effects are partly transmitted through successive generations.

Non linear models. Although BLUP techniques are rather robust to departures from normality, non linear models when traits are not normally distributed have been shown to more adequately describe observed data and result in higher responses to selection in various situations (e.g. see Meijering & Gianola, 1985). In other instances, such as survival data, no linear model can adequately and exhaustively describe available data (Ducrocq, 1990). Inference in non linear models is often based on bayesian methodology, which offers a global framework for the estimation of both location and dispersion parameters. Until recently, available algorithms were based on asymptotic approximations of posterior distributions (Ducrocq, 1990). However, things have dramatically changed during the last few years with the development of new algorithms based on Markov chain Monte Carlo methods, which can compute entire posterior distributions (e.g. Sorensen *et al.*, 1995).

Genetic models. Under the infinitesimal model, Henderson (1975) algebraically showed that ignoring some random effects in genetic evaluation still yields unbiased predictions, but with an increased prediction error variance and, as a consequence, lower responses to selection. Moreover, predictions do not remain unbiased in any situation. In the multivariate case, using a wrong model to estimate (co)variance components in selected populations results in incorrect parameters, which do not allow to properly account for selection bias (Ducrocq, 1994). Ignoring maternal effects also results in biased estimates of direct effects, even when direct heritability is correct and the correlation between direct and maternal effects is null (Roehe & Kennedy, 1993b). Predictions of additive affects are also biased in presence of dominance and inbreeding when these effects are ignored, but remain unbiased when they are properly accounted for (e.g. De Boer & Van Arendunk, 1992).

Extra genetic gain is usually expected from including information on genes with medium to large effects in the genetic evaluation process. Numerous studies have investigated this problem in recent years, though none of them has specifically addressed the case of pig breeding

schemes. Results are not always comparable, because selection criteria differed between studies (i.e. from an index based on individual information to animal models), but they all indicate that the knowledge of genotypes at quantitative trait loci generally improves short term response to selection (Larzul *et al.*, 1997). Conversely, some discrepancies have been obtained for long term response to selection (e.g. see Larzul *et al.*, 1997 for a discussion). In the less favourable situation where only genotypes at linked markers are known, results largely depend on the situation considered. Large gains can be expected in the most favourable situations, i.e. when linkage desequilibrium exists at the population level (Lande & Thompson, 1990) and when traits are expressed lately, are sex limited or are difficult to measure (e.g. Ruane & Colleau, 1996). In other cases, the advantage of marker assisted selection may be questionable (e.g. see Ruane & Colleau, 1996).

Environmental effects. Environmental effects include intrinsic effects such as age, breed or genetic type, parity and sex as well as contemporary or management group (CG), type of mating ... and their interactions. They are usually considered as fixed effects. As shown by Henderson (1975), ignoring or incorrectly specifying an important fixed effect leads to biased estimates of breeding values. Correctly specifying environmental effects is often a difficult task, mainly because sources of variation are not always available (e.g. different buildings in the same herd, preferential treatments,...) or may be poorly estimated. Small CG size in pigs mainly occur for meat quality (CG = animals sent together to the abattoir) or reproductive traits (CG is a complex interaction between herd, year, season and type and/or number of matings). In such a case treating CG as a random effect reduces PEV, but may result in biased predictions in case of non-random associations between sires and CG (this might have occur for instance if a high breeding value for growth rate was associated with poor meat quality). Another possibility consist in grouping successive small CG using clustering techniques.

PROBLEMS AND LIMITS OF COMPLEX MODELS

Main difficulties associated with the use of increasingly complex model are lack of robustness and computational problems. Other aspects, such as higher reduction of genetic variability and increase in inbreeding should not be underestimated, but are not considered here.

Robustness. Best linear unbiased predictions of breeding values are computed assuming that population dispersion parameters are perfectly known. More complex models often involve more dispersion parameters than simpler ones (e.g. genetic, litter and residual correlations between traits in multivariate analyses, direct and maternal heritabilities, plus a correlation between direct and maternal effects for maternal effect models,...). This increased number of parameters has several consequences. First, parameters are usually less accurately estimated for a given amount of data, because they often involve covariances, which are less accurately estimated than variances, but also because variances have a reduced accuracy. For instance, standard errors of heritabilities can be 3 to 5 times larger with a maternal effect model as compared to a model involving only direct effects (Thompson, 1976). Similar conclusions can be drawn with models involving non additive effects (e.g. see Misztal, 1997). Then, because more parameters are involved, complex models are often more sensitive to a given variation in

parameter values. As a consequence, gains from using a more sophisticated and appropriate model can be reduced or annihilated by poor parameter estimates, as shown by several recent studies. Alfonso (1995) compared the efficiency of a multiple trait vs a repeatability animal model for genetic evaluation of litter size when genetic are inaccurately known and found, using a decision theory approach, that the simple repeatability model should be preferred to a multivariate model. Ruane & Colleau (1996) showed that the interest of marker assisted selection over conventional BLUP-animal model is also reduced or even annihilated when the QTL variance has been incorrectly estimated. Schaeffer (1984) compared the efficiency of multiple trait (MT) vs single trait (ST) BLUP when using incorrect residual and genetic correlations. MT-BLUP sometimes had a lower efficiency of ST-BLUP but, as emphasized by Ducrocq (1994), ST-BLUP implicitly assumes null correlations and may have a low efficiency when true parameters differ from zero.

Poor estimates of fixed effects may also reduce the interest of complex models. For instance, the efficiency of multivariate models for litter size in pigs is reduced due to the smaller size of contemporary groups, which result in less accurate estimates of their effects. Overparameterized models for fixed effects also unnecessarily increase prediction error variance (Henderson, 1975) decreases connectedness and may result in computational problems. For instance, using a herd * year * season * type of mating interaction in genetic evaluation for litter size in France has been shown to give a lower predictive ability than a more parsimonious model involving a herd * year * type of mating effect and an additive effect of farrowing month (Bidanel, unpublished results). A similar situation was found when too many unknown parent groups were defined.

The lower robustness of complex models has important consequences when population size is limited, which is the case many pig strains, as variance components cannot be accurately estimated. Yet, the same problem may occur in large populations, as available computing resources still often limit the size of the system of equations involved. As a consequence, complex models involving several additional variance components are likely to give inaccurate dispersion parameter estimates. Moreover, variance components are generally estimated from subsets of the data used for genetic evaluation. A good sampling strategy is then fundamental to obtain unbiased estimates of dispersion parameters, in particular to avoid selection bias.

Computational problems. Genetic evaluations often require to solve huge linear systems using iterative methods. The number of equations involved often increases when increasingly sophisticated models are used. For instance, the number of equations is multiplied by the number of traits in multivariate analyses, by almost 2 when dominance effects are included and by almost 3 for marker assisted evaluation with one QTL using the model of Fernando & Grossman (1989) . The increase in computing time is much more important due to a higher computing time per iteration and a slower convergence. Several thousands of iterations may be required in presence of selection , when groups of unknown parents or more than one random effects are included and in case of poor connectedness (Ducrocq, 1994). Non linear models are also computationally very demanding, because of a low convergence rate and the higher

complexity of the system of equations to be solved. As emphasized by Ducrocq (1994), this may be problematic when a fixed number of iterations is used as a stopping criterion. Computing time can be drastically reduced in many instances by exploiting the structure of the system of equations, using specific algorithms or computational tricks - see Ducrocq (1994) for a review -. Unfortunately, the integration of these improved algorithms in general purpose packages, which would greatly favour their use, often remains to be done.

CHOICE AND VALIDATION OF A MODEL

As emphasized by Gelman *et al.* (1997), « a good [Bayesian] analysis should include at least some check of the adequacy of the fit of the model to the data and the plausibility of the model for the purpose for which the model will be used ». Though it may appear as common sense, this crucial step is not always satisfactorily achieved when elaborating genetic evaluation models. The first point, i.e. goodness of fit, can be evaluated using mean square error when dispersion parameters are known. In that case, fixed effects can be tested using an appropriate F tests (Henderson, 1984). When variances are unknown, nested models can be compared on the basis of their likelihood, and likelihood ratio tests (LRT) can be used to select the appropriate model. For instance, Robert *et al.* (1995) used LRT in the case of heteroskedastic linear mixed models to test the homogeneity of a set of genetic correlations. With non nested models, criteria such as the Akaike's Information Criterion can be used to select the best fit model among alternative models (see Wada & Kashiwagi, 1990, for an application to animal breeding problems). In the case of genetic evaluation problems, the second point mainly concern model's predictive ability. Crossvalidation techniques, in which observed data are partitioned and each data subset compared to its predictions conditional on the model and the rest of the data, have recently been used to compare linear vs non linear models (e.g. Perez-Enciso *et al.*, 1996) or models with and without groups of unknown parents (Estany & Sorensen, 1995). Yet, as noted by Frey *et al.* (1997), better ways than simply dividing the data into two equivalent subsets may be found to compare predictive ability of genetic evaluation models (e.g. prediction of progeny performance with parameters estimated from parental data). A somewhat related strategy for model checking in the bayesian framework is to compare posterior predictive distribution of future observations to real data or substantive knowledge. A review of bayesian model checking methods can be found in Gelman *et al.* (1997). Finally, simulation studies are also very helpful to compare model efficiency on the basis of long term response to selection (e.g. Roehe & Kennedy, 1993b; Ruane & Colleau, 1996).

CONCLUSION

Increasingly sophisticated genetic evaluation models may undoubtedly contribute to increase the efficiency of animal breeding plans. Recent advances have in particular been made to more adequately describe the structure of dispersion parameters and consider more realistic genetic models. However, a careful model checking and validation is a necessary prior step to insure that the proposed model is fully justified. A lot of work still remains to be done in this area.

REFERENCES

Alfonso, L.A. (1995) Doctoral Thesis, University of Lleida, Spain.

Andersson, L. *et al.* (1994) *Science* **263** : 1771-1774.
Bampton, P.R. (1992) *Pig News and Information* **13** : 125N-129N.
Bidanel, J.P., Ducos, A., Guéblez, R. and Labroue, F. (1994) *Livest. Prod. Sci.* **40** :291-301.
De Boer, I.J.M. and Van Arendunk, J.A.M. (1992) *Theor Appl. Genet.* **84** :451-459.
Ducrocq, V. (1990) *Proc. 4th WCGALP* **XIII** : 419-428.
Ducrocq, V. (1994) *Proc. 5th WCGALP* **18** : 455-462.
Estany, J. and Sorensen, D. (1995) *Anim. Sci.* **60** : 315-324.
Fernando, R.L. and Grossman, M. (1989. *Gen. Sel. Evol.* **21** : 467-477.
Foulley, J.L. and Molénat M. (ed.) (1994) Séminaire modèle animal, INRA, France, 157 p.
Foulley, J.L. and Quaas, R.L. (1994) *Proc. 5th WCGALP* **18** : 341-348.
Frey, M., Hofer, A. and Künzi, N. (1997) *Livest. Prod. Sci.* **48** : 135-141.
Garrick, D.J. and Van Vleck, L.D. (1987*) J. Anim. Sci.* **65** : 409-421.
Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1997) « Bayesian data analysis ». Chapman & Hall, London.
Gianola, D. (1986) *Theor. Appl. Genet.* **72** : 671-677.
Goddard, M.E. (1992) *Theor. Appl. Genet*. **83** : 878-886.
Henderson, C.R. (1949) *J. Dairy Sci.* **32** : 709 (abstract).
Henderson, C.R. (1975) *J. Anim. Sci.* **41** :760-770
Henderson, C.R. (1984) *Applications of linear models in animal breeding*. U. of Guelph, 462p.
Hill, W.G. (1984) *Anim. Prod.* **39** : 473-477
Hudson, G.F.S. and Kennedy, B.W. (1985) *J. Anim. Sci.* **61** : 83-91.
Im, S., Fernando, R.L. and Gianola, D. (1989) *Genet. Sel. Evol*. **21** : 399-414
Kennedy, B.W. and Schaeffer, L.R. (1990) In : D. Gianola and K. Hammond (ed.) *Advances in statistical methods for genetic improvement of livestock*, Springer-Verlag, 507-532.
Kennedy B.W., Quinton M. and Van Arendunk J.A.M. (1992) *J. Anim. Sci.* **70** : 2000-2012.
Larzul, C., Manfredi, E. and Elsen, J.M. (1997) *Genet. Sel. Evol*. **29**: 161-184.
Le Roy, P., Naveau, J., Elsen, J.M. and Sellier, P. (1990) *Genet Res* **55**: 33-40.
Meijering, A. and Gianola, D. (1985) *Genet. Sel. Evol.* **17** :115-132.
Misztal, I. (1997) *J. Dairy Sci.* **80** : 965-974.
Perez-Enciso, M., Tempelman, R.J. and Gianola, D. (1993) *Livest. Prod. Sci.* **35** : 303-316.
Pollak, E.J. and Quaas, R.L. (1981) *J. Dairy Sci.* **52**: 257-264.
Robert, C., Foulley, J.L. and Ducrocq, V. (1995) *Genet. Sel. Evol.* **27** :111-123.
Roehe, R. and Kennedy, B.W. (1993a) *J. Anim. Sci.* **71** : 2891-2904.
Roehe, R. and Kennedy, B.W. (1993b) *J. Anim. Sci.* **71** : 3251-3260.
Schaeffer, L.R. (1984) *J. Dairy Sci.* **67** : 1567-1580.
Schmidt, G.H. (ed.) (1988) Proc. Animal Model Workshop. *J.Dairy Sci.* **71** (suppl.2), 125 p.
Simianer, H. (1994) *Proc. 5th WCGALP* **18** : 435-442.
Sorensen, D. and Johansson, K. (1992) *J.Anim. Sci.* **70**: 2038-2044.
Sorensen, D., Andersen, S., Gianola, D. and Korsgaard, I. (1995) *Genet. Sel. Evol.* **27**:229-249.
Thompson, R. (1976) *Biometrics* **32** : 903-917.
Thompson, R. and Meyer, K. (1986) *Livest. Prod. Sci.* **15**: 299-313.
Wada, Y. and Kashiwagi, N. (1990) *J. Dairy Sci.* **73** : 3575-3582.