

Chapter 1. General Introduction

Estimated breeding values (EBVs) derived using mixed model methodology (Henderson, 1975) for traits of economic importance are used for the genetic improvement of livestock. The estimation of breeding values is frequently undertaken using an animal model, which has become a model of choice across the world due to its useful and well-defined properties. The animal model takes into account the genetic relationships amongst all the animals through the numerator relationship matrix and assumes that all the animals with missing pedigree information are derived from a common base population with a mean breeding value of zero and a common additive genetic variance. The EBVs are expressed as a deviation from that of a common base population. The animal model incorporates all information on relatives with or without performance records. The fixed effects such as sex, age of animal, dam age, birth type, management groups and certain other effects depending on the trait and species are also taken into consideration to derive EBVs. In an animal model evaluation, selection is accounted for if all relevant information regarding performance data used for making selection decisions, and genetic relationships among animals are considered (Sorensen and Kennedy, 1984).

There are a range of important issues associated with the types of effects (fixed and random) that need to be considered for deriving EBVs with increased precision. For example, in almost all the cases, missing pedigree (Fikse, 2003; Casellas *et al.*, 2007) and performance data make it difficult to consider all the relevant information in deriving unbiased estimators of breeding values. This may affect rate of genetic improvement as it reduces the accuracy of genetic evaluation, and the selection intensity is reduced if animals with unknown parentage are not considered for selection (Perez-Enciso and Fernando, 1992; Cardoso and Tempelman, 2003). In addition, in certain cases, there will be animals with unknown parentage which may have been derived from a population with a different average genetic merit, thus leading to a deviation from the assumption that all animals have a common genetic base. For example, if sires with different genetic merit are imported from foreign countries and used for breeding

in the native population, it would affect the genetic mean of the local population. Therefore, the EBVs of all animals being evaluated on the same base would be biased to an extent depending on the level and intensity of selection exercised in the population from which the imported sires were obtained. Similarly, in the Australian Merino population, flocks have links to genetically diverse base populations and subpopulations which poses a challenge for Sheep Genetics (Brown *et al.*, 2006; 2007) evaluations.

The differences in the means of various base populations can be accounted for through an appropriate genetic grouping strategy (Robinson, 1986; Westell *et al.*, 1988; Quaas, 1988; Atkins *et al.*, 1998; Atkins *et al.*, 1999b; Fikse, 2003; Togashi *et al.*, 2004). However, the actual procedure required to assign animals to different genetic groups needs further research (Atkins *et al.*, 1998). Quaas (1988) argued that a complex grouping strategy must be avoided to overcome the potential problem of confounding with other fixed effects and hence the estimability of group effects using mixed model methodology. Additionally, Pollak *et al.* (1977) stated that confounding of non-genetic (environmental) and genetic group effects appears to impact the accuracy of group solutions. Fikse (2003) suggested that forming genetic groups by combining different time periods may lead to a bias as all phantom parents within a genetic group are assumed to be contemporaries.

If animals are to be selected on the basis of across-flock EBVs, then it would be important to estimate the actual genetic differences between flocks. The main advantage of across-flock evaluation is that it allows for selection from a large genetic base (Banks, 1997; Mathur *et al.*, 1998) and across age groups (Banks, 1997). Additionally, across-flock evaluations lead to the estimation of genetic trends for the entire population and for individual flocks over a given time period (Banks, 1997). The accuracy of an across-flock evaluation scheme will depend on the estimation of independent flock (environmental) and genetic group effects which in turn will require adequate levels of genetic linkage (proportion of progeny derived from common link sires) across flocks geographically isolated from each other (Kennedy, 1981; Kinghorn and Shepherd, 1990; Mathur *et al.*, 1998; Atkins *et al.*, 1998; Atkins *et al.*, 1999b). In the absence of genetic links among flocks, the EBVs of animals in a flock cannot be compared directly with the EBVs of animals in another flock. This is attributed to confounding of genetic and

environmental differences between flocks. Therefore, the amount of linkage required for an accurate dissociation of genetic and environmental differences between flocks needs investigation. Genetic links between flocks can be established either through the use of common sires or through transfer of breeding stock animals among flocks (Kuehn, 2005).

The lack of pedigree recording within flocks and suboptimal genetic linkage across flocks have been the main limitations to unbiased across-flock evaluations in the Australian Merino sheep industry in the past (Ashtiani and James, 1992; Atkins *et al.*, 1998; Casey and Atkins, 1999). However, previous studies by Huisman *et al.* (2005, 2006) involving industry data from Sheep Genetics have shown that adequate genetic links now exist for most flocks and traits to undertake across-flock evaluation. Depending on the proportion of linkage between flocks, the EBVs generated in OVIS (Brown *et al.*, 2000) in routine Sheep Genetics evaluations can be classified into two categories. The first category comprises flocks that are not sufficiently linked and have trait performance defined as flock breeding values (FBVs). The second category has flocks that are adequately linked and have trait performance described as Australian sheep breeding values (ASBVs) (Brown *et al.*, 2006). Flocks participating in Sheep Genetics evaluations must have effective minimum linkage with other flocks within their reporting group. This allows for an accurate comparison of animals across flocks.

In Sheep Genetics evaluations, different types of models are fitted using OVIS (Brown *et al.*, 2000) for deriving FBVs and ASBVs. Since the Australian Merino data are characterized by a high proportion of missing dam pedigree and trace back to multiple base populations and subpopulations, genetic groups are included in the evaluation model. Due to genetic differences between flocks, an effective strategy of defining genetic groups in Sheep Genetics evaluations would be grouping by flock of origin. Currently, Sheep Genetics utilizes the same grouping strategy for all traits which may not be appropriate due to the differences in selection history and data recording between traits. Therefore, alternative options for assigning genetic groups in a multi-trait evaluation need to be explored.

The main objective of this research is to identify the importance of genetic grouping and linkage in across-flock genetic evaluations pertaining to the Australian Merino sheep industry. This research was aimed at developing better models and techniques for allocating animals to genetic groups for Sheep Genetics evaluation schemes with specific emphasis on Merino sheep. This thesis will attempt to answer several questions arising from the issues of genetic grouping and connectedness in across-flock evaluations. Chapter 2 will deal with a review of literature to summarize the effect of linkage and genetic grouping on the accuracy of across-flock evaluations. Chapter 2 will also review different strategies used for incorporating genetic groups in the evaluation models. Additionally, the significance of connectedness in across-flock evaluations and different measures of connectedness proposed in the literature, and other related issues will be reviewed in Chapter 2.

The effect of linkage and genetic grouping on the accuracy of single-trait across-flock EBVs will be studied using simulation in Chapter 3. This chapter will also focus on the partitioning of genetic and environmental differences between flocks in across-flock evaluations. The proportion of linkage required for achieving this objective will be investigated. The type of data structure required for an accurate disentangling of genetic and environmental differences between flocks will be examined. The effect of the level of heritability and the number of progeny per link sire required for an accurate estimation of genetic group effects will also be investigated in Chapter 3.

The Merino population in Australia comprises different strains and bloodlines within strains, and has links to genetically diverse base populations and subpopulations. Since a large number of flocks are evaluated by Sheep Genetics, genetic groups are treated as random effects (Brown *et al.*, 2006; 2007). The effect of flock and genetic group size (number of animals) on the accuracy of estimated genetic group effects and genetic group variance in a single-trait across-flock evaluation will be examined in Chapter 4. This chapter will also make use of significance tests to evaluate the impact of linkage and flock size on the accuracy of estimated genetic group effects and genetic group variance.

Flocks with different sizes are jointly evaluated for a large number of traits of interest in the Australian Merino sheep industry. The accuracy of estimation of genetic group effects is likely to be reduced for smaller flocks. In Chapter 5, the feasibility of pooling smaller flocks together into one large group or cluster based on their estimated genetic group effects will be tested as an alternative option to overcome this problem without compromising the accuracy of evaluation procedure. The impact of different cluster sizes (number of clusters) on the accuracy of prediction of breeding values and various issues associated with clustering flocks in a single-trait across-flock evaluation will be examined in this chapter.

Data on fleece traits in the Australian Merino population are characterized by varying amounts of missing observations. Significant differences in selection history may necessitate assigning genetic groups in a different way for each trait (differential genetic groups) in a multi-trait evaluation. Therefore, different strategies of genetic grouping for different traits will be tested in Chapter 6 using numerical examples. The feasibility of differential genetic groups using mixed model methodology will be examined. Additionally, modified mixed model equations (MME) for an animal model evaluation with differential genetic groups will be presented.

The application of differential genetic grouping to multiple-trait evaluation using industry data from Sheep Genetics database will be investigated in Chapter 7. This chapter will compare the two strategies of genetic grouping namely, conventional and differential. The conventional strategy assigns each animal to the same genetic group for all traits whereas differential strategy allocates animals to genetic groups in a different manner for each trait. Chapter 7 will also evaluate how differences in the amount of data for each trait will affect multi-trait EBVs derived using the two strategies of genetic grouping.

The different issues regarding genetic groups reviewed in Chapter 2 and the results obtained in each experimental analysis (Chapters 3 to 7) will be discussed in Chapter 8. In Chapter 8, the practical application of the outcome of current research in the genetic evaluation of livestock will be outlined with special reference to the

challenges faced by the Australian Merino sheep breeders with regard to across-flock evaluations.

Finally, the concluding chapter of this thesis (Chapter 9) will briefly outline various conclusions drawn from the current research and their implications regarding across-flock evaluation schemes for the Merino sheep industry in Australia.

Chapter 2. Literature Review

Genetic evaluation of livestock is usually undertaken using a number of statistical models of which sire and animal models are the most widely used. Both sire and animal models make use of mixed model equations (Henderson, 1975; MME) which are solved for the simultaneous estimation of systematic fixed and random genetic effects (breeding values) for traits of interest. However, in most cases, industry data are characterized by varying amounts of missing records and pedigree information which can still affect the accuracy of evaluation scheme. Additionally, the assumption of all animals of interest being derived from a single base population cannot be true in most cases. Therefore, to account for missing pedigree information, genetic groups are generally included in the evaluation model (Pieramati and Van Vleck, 1993; Diaz *et al.*, 2002). The inclusion of genetic groups in the evaluation model could account for the differences in the means of base populations (Henderson, 1975; Van der Werf, 1990; Van der Werf and de Boer, 1990).

The aim of this chapter is to review the concept of genetic grouping and its application in the mixed model methodology to account for the differences in the means and variances of different base populations and subpopulations with special reference to the Australian Merino sheep industry. Genetic groups (based on flock) are generally included in routine Sheep Genetics evaluations to account for genetic differences between flocks. Additionally, the relative importance of genetic links (connectedness) in estimating genetic group effects, systematic fixed effects, and breeding values will be addressed. The properties of statistical models and their modification to include genetic groups in the MME will be discussed briefly in the following section.

2.1 The concept of genetic grouping

The animal model evaluation takes into account the relevant performance data and all genetic relationships among animals through the numerator relationship matrix (NRM) to derive breeding values for traits of economic importance. In order to obtain

unbiased estimates of breeding values, the pedigree links should extend to the true base population(s) which is rarely the case. Therefore, it is generally assumed that the base population comprises first generation of unrelated, unselected and non-inbred animals with records (Van der Werf, 1990). However, in almost all field datasets, varying amounts of pedigree information is missing (Casellas *et al.*, 2007) in addition to missing records on different traits. Furthermore, generally most field datasets used for genetic evaluation have performance records related to base animals derived from different selection paths and time periods (Van Vleck, 1990a). There is also a sequential increase in data and pedigree information with the passage of time. Therefore, animals with relatively more information (data and pedigree) have more emphasis in the BLUP (Best linear unbiased prediction) evaluation and have higher probability to be selected as parents of the future generation than animals with missing records and parentage. This may lead to a bias in the estimated genetic mean of the population, thereby leading to a downwards bias in the estimated heritability (Van Vleck, 1970) which in turn underestimates the selection response.

Van Vleck (1990a) stated that breeding values estimated by ignoring the genetic merit of missing animals are forced to regress towards the average group effect instead of the separate group effects. The NRM tends to account for most of the genetic trend by reducing the need for genetic grouping (Henderson, 1975; Thompson, 1979; Pollak and Quaas, 1983). However, Wiggans *et al.* (1988) and Fikse (2003) argued that genetic group effects need to be included in the evaluation model for representing differences in the genetic values of missing parent animals that might belong to different populations. The main objective of genetic grouping is to account for the genetic differences between subpopulations (Quaas and Pollak, 1981; Golden *et al.*, 1994) from which animals of interest have been derived. Therefore, the concept of using genetic groups in Merino sheep evaluations is very appropriate since the Australian Merino population has diverse links tracing back to multiple base populations and subpopulations. Genetic groups are generally assigned on the basis of certain criteria like flock of origin, selection path, generation number, sex of missing parents, year of birth, and country of origin.

Henderson (1975) was the first to propose the concept of including genetic groups in the evaluation model. The significance of including genetic group effects in the models used for genetic evaluation was later discussed by Thompson (1979) in the form of an accumulated groups model. The author reviewed procedures adopted for sire evaluation and stated that the conventional BLUP assumes that a group of sires (a set of sires from one stud entering service at one time) follow a normal distribution with a zero mean and that younger sires have superior genetic merit. The author argued that this assumption may not always be true. This argument of Thompson (1979) can be supported by the fact that there is a constant variation in the additive genetic variance with time period due to both natural and artificial selection. Therefore, the author proposed that the least squares estimates of breeding values of sires may be regressed to a group mean rather than zero in order to avoid a possible bias.

Westell (1984) and Westell *et al.* (1988) proposed the use of genetic grouping:

1. To account for the genetic merit of new migrants into a native population,
2. To account for the selection of animals without performance records,
3. To account for selection which is not accounted for by performance data on relatives and known genetic relationships where one or both parents of an animal are missing or not recorded, and
4. To account for the genetic trend.

All the above factors may vary with the data structure. In a majority of models used for genetic evaluation, only a single base population effect is assumed to account for various genetic levels of animals. In order to fit genetic groups to account for unknown selection, it is essential to know the missing information (missing pedigree and performance data) precisely (Westell *et al.*, 1988). The unknown or missing information in most cases is either the missing records or the “unknown parentage” of animals undergoing evaluation. The word “phantom parents” was introduced by Westell (1984) to denote groups of unrelated and unidentified parents to facilitate the solving of MME for deriving solutions. According to the author, the word “phantom” was used to symbolize such animals which were not of interest themselves.

2.2 Theoretical aspects of genetic grouping

2.2.1 Criteria for defining genetic groups in a model

Different methods and techniques have been proposed for defining genetic groups in the models for the estimation of breeding values. However, there are no criteria for defining genetic groups for a particular application that has universal acceptance (Quaas and Pollak, 1981; Robinson, 1986; Casellas *et al.*, 2007). Therefore, only arbitrary definitions have been used depending on criteria like bulls entering service at the same time from the same stud, their date of birth and geographic origin, and the amount of pedigree information available (Kennedy, 1981; Quaas and Pollak, 1981; Famula *et al.*, 1983).

Genetic groups can also be assigned on the basis of sex and birth year of parent(s) and sex of animal with missing parent(s). Genetic groups can be fitted in the model to replace missing parents to complement the relationship matrix (Wiggans *et al.*, 1988). Although missing parent animals could be replaced by phantom parents on the basis of the above-mentioned criteria, these are unlikely to hold well in most field datasets. The possibility of a phantom animal having no genetic relationship with other phantom or real animals in a population is also very low (Perez-Enciso and Fernando, 1992). Phantom parents could be estimated jointly in a single genetic group. Although one phantom parent may have only one progeny, grouping of phantom parents into a single genetic group will have large numbers of progeny so that many animals are related to the one unique genetic group.

Forming genetic groups on the basis of year of birth accounts for the genetic change over time (Theron *et al.*, 2002). Each selection path is defined according to the genetic selection differential for a phantom sire and a phantom dam, therefore, leading to two equivalent sets of groups each for phantom sires and phantom dams that are selected to be the parents of animals of interest (Westell *et al.*, 1988). Similarly, four sets of equivalent groups would represent phantom sires of sires, phantom sires of dams, phantom dams of sires, and phantom dams of dams, each corresponding to the four selection paths accounting for different selection intensity (Westell *et al.*, 1988).

Genetic groups may be cross-classified with year of birth depending on the sex of the animal and that of the phantom parent to account for the four different paths of selection (Westell *et al.*, 1988).

Defining genetic groups on the basis of time and sex of phantom parents accounts for different generation intervals if only two sets of groups (phantom sires and phantom dams) are defined (Westell *et al.*, 1988). However, with a large number of overlapping generations, this technique will be tedious. A phantom dam of a sire would not necessarily correspond to the similar time period group as a phantom dam of a dam even though the sire and the dam have the same date of birth. This is due to likely differences in the generation interval of dam of a sire and that of a dam (Westell and Van Vleck, 1987). The generation interval corresponding to the genetic group would then be defined by subtracting the average generation interval from the birth year of the animal with missing parent(s) to predict the year of birth of the missing parent (Westell, 1984; Westell and Van Vleck, 1987).

Sullivan (1995) argued if genetic groups are formed on the basis of breed and sex of phantom parents, the MME for missing sire and dam groups would be the same for different breeds with no single parent pedigrees, thus leading to confounding (singularity in MME) problems. Therefore, forming genetic groups based on arbitrary definitions must be avoided. Famula *et al.* (1983) suggested that a grouping structure should be derived based on its linear relationship to past selection practices in order to obtain unbiased estimates of breeding values.

Deriving appropriate grouping structures would be a tedious task for Australian Merinos because genetic differences between flocks are confounded with flock environmental differences and lack of genetic links between flocks does not allow for an accurate partitioning of these effects.

Although previous studies on pigs (Wolf *et al.*, 2002; Peskovicova *et al.*, 2004) and beef cattle (Phocas and Laloe, 2004) defined genetic groups on the basis of similar criteria as reviewed above, the task of defining genetic groups properly and accurately for the types of data generally available for animal breeding research would be quite

challenging. For example, with large datasets having varying amounts of missing pedigree, relating animals to a large number of phantom parent groups will make the evaluation procedure more complex. As reviewed above, genetic groups are always formed on the basis of arbitrary definitions which may not necessarily account for past selection when pedigree and data information is missing.

Expressed in the words of Famula *et al.* (1983), “to propose a general multipurpose definition for genetic groups is impossible”. This is attributed to the differences in data and pedigree recording, and also to either little or no information available regarding selection path through which animals with records have been derived from their base ancestors. In addition, when grouping is defined with sex as one of the criteria, there is a possibility of complete confounding between the parent genetic groups (both sire and dam) for animals with unknown parentage and other fixed effects like management groups, thus leading to problems in deriving the solutions (Wolf *et al.*, 2002). A sex-specific definition of groups would reduce the number of animals in individual genetic groups, which would lead to convergence problems in solving equations for the estimation of breeding values (Wolf *et al.*, 2002). Assigning missing parents to genetic groups on the basis of above-reviewed criteria may not necessarily account for the real genetic merit of those missing parents (Fikse, 2003). The problem associated with the prediction of real genetic merit of missing parents also gets aggravated by a sequential increase of animals with records with the passage of time (Pieramati and Van Vleck, 1993). There is no universally accepted grouping method available when genetic groups are included in sire or animal model evaluation.

2.2.2 Sire model with genetic grouping

Genetic group effects can be incorporated in sire model evaluation. A sire model with genetic grouping was proposed by Quaas and Pollak (1981) with the aim of accurately defining groups for evaluating sires which were genetically related. The authors stated that genetic groups can account for genetic trend when there is ongoing selection and relationships between sires are missing. A sire model with genetic groups in matrix notation can be written as:

$$\mathbf{y} = \mathbf{Xh} + \mathbf{ZQg} + \mathbf{Zs} + \mathbf{e}$$

where $\mathbf{h}$ is a vector of fixed herd-year-season effects, $\mathbf{s}$ is a vector of random sire effect, $\mathbf{g}$ is a vector of fixed genetic group effects and $\mathbf{e}$ is a vector of random environmental effects. $\mathbf{X}$ and $\mathbf{Z}$ are the known incidence matrices relating fixed effects in $\mathbf{h}$ and sire effects in $\mathbf{s}$ (and fixed genetic group effects in $\mathbf{g}$) to phenotypic observations in vector $\mathbf{y}$, respectively; and $\mathbf{Q}$ is a design matrix describing the proportion of genes contributed by each genetic group to the sires of interest. The variance for random effects in $\mathbf{s}$ and $\mathbf{e}$ can be described as $\text{var} \begin{bmatrix} \mathbf{s} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}/\alpha & 0 \\ 0 & \mathbf{I} \end{bmatrix} \sigma_e^2$ where $\mathbf{s}$ and $\mathbf{e}$ have zero means, α is the ratio of sire variance (σ_s^2) over residual variance (σ_e^2); $\mathbf{A}$ is the matrix of numerator relationships among sires, and $\mathbf{I}$ is an identity matrix. The incidence matrix for different sire genetic groups is represented by $\mathbf{ZQ}$ since sires are nested within genetic groups. The matrix $\mathbf{Q}$ has dimensions $a \times b$ where a and b are the number of sires and genetic groups, respectively. A random sire effect nested within a fixed group effect gives an estimated breeding value which is a function of a group solution plus a sire solution (Quaas and Pollak, 1981).

Thompson (1979) proposed a model that accounted for genetic group effects associated with sires that were being evaluated using performance data of their daughters. The author stated that the sire effect could be written as an accumulation of independent terms derived from its ancestors and itself. The author further added that these independent terms are associated with different genetic group effects.

2.2.3 Animal model with genetic grouping

Genetic group effects can be incorporated in animal model in two different ways which results in equivalent solutions but require different computing algorithms. In the literature, they are commonly termed the $\mathbf{Q}$ - and $\mathbf{W}$ -matrix approach (e.g., Westell *et al.*, 1988). The single-trait animal model in matrix notation can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypic observations, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random additive genetic effects, and $\mathbf{e}$ is a vector of random residual effects associated with each observation. $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices relating fixed

effects in $\mathbf{b}$ and random animal effects in $\mathbf{u}$, respectively with the observations in $\mathbf{y}$. The expectation ($\mathbf{E}$) and variance ($\mathbf{V}$) for the fixed and random effects, respectively can be expressed as:

$$\mathbf{E} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{Xb} \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbf{V} \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_a^2 & 0 \\ 0 & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

where σ_a^2 and σ_e^2 are the additive genetic and residual variance, respectively; $\mathbf{A}$ is the numerator relationship matrix and $\mathbf{I}$ is an identity matrix. To include genetic groups, the above model and the corresponding MME can be modified according to the rules developed by Westell *et al.* (1988). Let us consider a vector $\mathbf{g}$ of genetic group effects, which are treated as additional fixed effects (besides other systematic effects) and a matrix $\mathbf{Q}$ ($\mathbf{Q} = \mathbf{T}\mathbf{G}$) that describes the proportion of genes derived by real animals from their respective genetic groups. $\mathbf{T}$ is a lower triangular matrix (describing the flow of genes from one generation to the other) obtained from $\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}'$ (Henderson, 1976) and has the same dimensions as $\mathbf{A}$, and $\mathbf{D}$ is a diagonal matrix. The association between phantom parents and various genetic groups is described by the $\mathbf{G}$ matrix. Each column of $\mathbf{G}$ denotes a genetic group such that the number of columns in $\mathbf{G}$ equals the number of genetic groups. The mixed model and the corresponding MME with genetic group effects (Westell and Van Vleck, 1987) can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za} + \mathbf{e} \text{ and}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{ZQ} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda & \mathbf{Z}'\mathbf{ZQ} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{X} & \mathbf{Q}'\mathbf{Z}'\mathbf{Z} & \mathbf{Q}'\mathbf{Z}'\mathbf{ZQ} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (2.1)$$

where λ is the ratio of residual to additive genetic variance. The elements of $\mathbf{Q}$ are the fractional coefficients relating the proportion of genes contributed by the genetic groups to the total genetic value of the animal (Westell *et al.*, 1988). The genetic group effects in vector $\hat{\mathbf{g}}$ estimated using the above model are then added to the animal effects in vector $\hat{\mathbf{a}}$ to obtain the EBVs. Hence, the breeding values are expressed as a deviation from combined base populations. The solution for each base population is determined by the proportion of genes contributed by it to the real animals (Westell *et al.*, 1988).

The QP-transformation (Quaas and Pollak, 1981; Pollak and Quaas, 1983) of the MME with genetic grouping above (2.1) leads to the direct estimation of breeding values without solving for $\hat{\mathbf{a}}$ and $\hat{\mathbf{g}}$ separately. These modified MME use a new matrix $\mathbf{W}^{-1}$ (Westell, 1984; Westell *et al.*, 1988) instead of $\mathbf{A}^{-1}$. The $\mathbf{W}$ -matrix is similar to the $\mathbf{A}$ -matrix except that it has genetic groups included in it, and has dimensions equal to the number of animals plus the number of genetic groups. The modified QP-transformed MME for a single-trait evaluation with genetic group effects can be represented as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{W}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (2.2)$$

and after partitioning the inverse of $\mathbf{W}$ -matrix and $\hat{\mathbf{u}}$, (2.2) can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & 0 \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{11}\lambda & \mathbf{A}^{12}\lambda \\ 0 & \mathbf{A}^{21}\lambda & \mathbf{A}^{22}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} + \mathbf{Q}\hat{\mathbf{g}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ 0 \end{bmatrix} \quad (2.3)$$

where matrices $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{Q}$ and vectors $\mathbf{y}$ and $\hat{\mathbf{b}}$ are as above; $\mathbf{A}^{11}$ (Animal-Animal), $\mathbf{A}^{12}$ (Animal-Group), $\mathbf{A}^{21}$ (Group-Animal), and $\mathbf{A}^{22}$ (Group-Group) are the sub-matrices from $\mathbf{W}^{-1}$, λ is the ratio of residual to additive genetic variance, and superscripts 1 and 2 correspond to animal and group, respectively. The QP-transformation (Quaas and Pollak, 1981; Pollak and Quaas, 1983) of the MME with genetic grouping above (2.1) leads to a term $\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$ which is equivalent to the $\mathbf{A}^{22}$ submatrix of $\mathbf{W}^{-1}$. The vector $\hat{\mathbf{u}}$ includes both $\hat{\mathbf{g}}$ (genetic group effects) and $\hat{\mathbf{a}}$ (random animal effects) so that $\hat{\mathbf{u}} = \hat{\mathbf{a}} + \mathbf{Q}\hat{\mathbf{g}}$ from (2.1), and $\mathbf{Z}$ has null columns corresponding to genetic group effects (Westell, 1984). Likewise, the above MME with genetic groups (via $\mathbf{Q}$ - and $\mathbf{W}$ -matrices) for a single-trait evaluation (2.1 and 2.2) can be modified to accommodate multiple traits. There are several advantages of these modified MME (2.2 and 2.3), the main being that the rules for building the inverse of $\mathbf{A}$ (Henderson, 1976) could be used to build genetic group equations. Another benefit of using the modified MME is that genetic groups can have various and complex contributions to further descendants of base animals. However, cited by Diaz *et al.* (2002), when genetic group effects need to be assigned separately for direct and maternal genetic components, the modified QP-transformed

MME (2.2) cannot be used because the vectors for direct and maternal effects do not necessarily contain the same animals (Van Vleck, 1990a).

Westell *et al.* (1988) presented the following rules for computing the inverse of **W**-matrix:

1. Computing $\mathbf{A}^{11}$ (*Animal-Animal*) submatrix: The terms allocated to $\mathbf{A}^{11}$ are identical to those that would be computed for conventional $\mathbf{A}^{-1}$ following the rules of Henderson (1976).

2. Computing $\mathbf{A}^{12}$ (*Animal-Group*) submatrix: $\mathbf{A}^{12}$ associates animals to genetic groups and can be calculated using the rules of Henderson (1976) for computing $\mathbf{A}^{-1}$. For animal *i*, an offspring of *j* and *m*, when both parents *j* and *m* are phantom, add the following to the specified coefficients of $\mathbf{A}^{12}$,

add -0.5 to (*i*, group of *j*)

add -0.5 to (*i*, group of *m*).

If only a single parent, *m*, is phantom,

add -2/3 to (*i*, group of *m*),

and add 1/3 to (*j*, group of *m*).

3. Computing $\mathbf{A}^{22}$ (*Group-Group*) submatrix: When both parents, *j* and *m* are phantom, then

add 0.25 to the diagonal coefficient corresponding to the group of *j*.

add 0.25 to the diagonal coefficient corresponding to the group of *m*.

add 0.25 to the (*j*, *m*) group coefficient.

add 0.25 to the (*m*, *j*) group coefficient.

If only a single parent *m* is phantom,

add 1/3 to the diagonal coefficient corresponding to the group of *m*.

The same rules as for $\mathbf{A}^{12}$ (*Animal-Group*) above can be used to construct $\mathbf{A}^{21}$ (*Group-Animal*) as they are transpose of one another. The resulting **W**-matrix is such that any column of **W** will sum to zero. Genetic groups can be treated as fixed or random effects

in animal model evaluations. When random genetic groups are assumed, a value is added to the diagonal elements of $\mathbf{A}^{22}$ (Group-Group) submatrix of $\mathbf{W}^{-1}$.

2.2.4 Fixed versus random genetic group effects

When genetic group effects are treated as fixed, no value is added to the diagonals of $\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$ ($=\mathbf{A}^{22}$ or Group-Group block of $\mathbf{W}^{-1}$) in the MME, which leads to $\mathbf{Q}'\mathbf{A}^{-1}\hat{\mathbf{u}} = 0$, where $\hat{\mathbf{u}}$ is a vector of random animal effects, as described earlier. Alternatively, when genetic group effects are treated as random, a value is added to the corresponding diagonal elements of the coefficient matrix ($\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$). This implies that $\mathbf{Q}'\mathbf{A}^{-1}\hat{\mathbf{u}} = \hat{\mathbf{g}}$, where $\hat{\mathbf{g}}$ is a vector of genetic group effects (Schaeffer, 1994). In the latter, the implicit assumption is that the genetic group effects have an expectation of zero (Schaeffer, 1994) and have an associated variance which determines the value added to the diagonal elements of $\mathbf{A}^{22}$ (Group-Group) submatrix of $\mathbf{W}^{-1}$.

The estimation of genetic group effects and their associated variance (genetic group variance) in across-flock evaluations may require certain assumptions regarding their covariances and prior distributions. The genetic group variance could differ from the additive genetic variance depending on the differences in selection history between groups. Therefore, the question arises whether genetic group variance could be estimated independent of the additive variance, and possibly if there is a covariance structure between the additive and genetic group variances. For example, when fitting breeds as genetic group effects in across-breed evaluations, genetic group variance could be attributed to the differences between breeds. Similarly, considering the wide variation between genetically different strains and bloodlines for economically important traits in the Australian Merino population (Mortimer and Atkins, 1989; Pietsch *et al.*, 1997; Mortimer and Atkins, 2001; Mortimer *et al.*, 2006), various strains could be treated as separate genetic groups and the differences estimated using a model which fits strain as a group effect.

Sullivan and Schaeffer (1994) analyzed simulated data to investigate the alternative of fitting genetic groups as fixed or random effects. All the dams were assumed to be known and missing sires were grouped on the basis of progeny birth generation. Random genetic groups were assumed by adding the residual to animal

genetic variance ratio (σ^2_e/σ^2_a) to the diagonal element of the MME corresponding to each genetic group. In their study, the random groups model either improved or had no impact on the accuracy of evaluation. The authors suggested that the advantage of treating genetic groups as random effects was greater especially with prior random selection of parent animals and when the heritability of a trait is low. Their results also showed that the random genetic groups model has consistent advantages over fixed genetic groups model with regard to the computation time. Previous simulation studies by Redman *et al.* (1993) have shown that more computation time is required for reaching convergence when genetic groups are fitted as fixed rather than random effects in the model. The increase in the time required for convergence could be an “indicator of apparent tendency to singularity in the MME” (Kinghorn *et al.*, 1993). Thus, with respect to computing requirements, fitting random genetic groups in the model appears to be more beneficial than assuming genetic group effects as fixed. However, treating genetic groups as random effects could be regarded as a biased estimation of the fixed effects of phantom parent groups as these effects contain accumulated effects of past selection and therefore should not have an expectation of zero (Schaeffer, 1994). This argument appears to be reasonable only if few genetic groups are defined. However, in Sheep Genetics evaluations using OVIS (Brown *et al.*, 2000), a large number of genetic groups are formed (Brown *et al.*, 2007) for different traits based on different combinations of criteria (e.g., flock, period, and numbers of records) thereby implying that group effects could be treated as random.

In the study of Sullivan and Schaeffer (1994), when truncation selection was applied, the advantage of assuming genetic groups to be random effects still remained considerable. They concluded that fitting random genetic groups eliminates the potential problem of singularity (confounding with other fixed effects resulting in non-estimability) related with fixed genetic groups model. Similar findings were reported by Sullivan (1995). However, in the study of Sullivan and Schaeffer (1994), a simple grouping strategy was followed by grouping sires on the basis of progeny birth generation and all dams were assumed to be known. Considering significant amount of missing dam pedigree in the Australian Merino industry (Casey and Hygate, 1992; Atkins *et al.*, 1998; Brash, 1999) and other issues like the type of data generally

available and the range of heritability of fleece traits, the conclusions of Sullivan and Schaeffer (1994) may not be valid.

The underlying assumption of different grouping strategies that groups are genetically and statistically independent (Sullivan and Schaeffer, 1994) will only be true for a single trait evaluation. However, when genetic relationships between animals from varying breeding studs exist, as has been shown in the context of sire evaluation in the United States and Canada (Kennedy and Moxley, 1975) due to widespread use of AI, the above assumption will not be true. This assumption will also not be true in a multi-trait evaluation involving correlated traits where group effects for different traits could have a covariance structure. Similar argument can be made for Merino sire evaluation in Australia. This can be justified by the fact that previous research by Huisman *et al.* (2005, 2006) has shown that adequate genetic linkage now exists for most flocks in the Australian Merino population. In such cases, random genetic groups in conjunction with relationships between animals from varying groups and the covariance between genetic groups need to be accounted for in the evaluation model. However, for some flocks without adequate linkage and/or where the between-population genetic group effects are not distributed normally (due to differences in data recording and selection history), which is likely to be the case for some traits like worm egg count (WEC), treating genetic groups as random effects would still not account accurately for the genetic differences between base populations. Therefore, for such scenarios and data structure, genetic groups need to be treated as fixed effects. Furthermore, Ashtiani and James (1993) argued that if the number of genetic groups defined is few, they may be treated as fixed effects. They emphasized that in the Australian Merino industry, there are a large number of studs, therefore, it would be realistic to treat their mean breeding values as random effects.

With special reference to the French AI beef sire evaluation, Phocas and Laloe (2004) used simulation to examine whether fixed or random genetic groups need to be included in the model given a priori genetic differences exist between subpopulations of base animals. The criteria used were the mean square error (MSE) of some extreme contrasts between animals, selection response when animals were selected across different groups, and the selection bias which was defined as the difference between

true and estimated selection responses. Fixed or random genetic groups were formed on the basis of birth year. Under BLUP selection across genetic groups, the authors found that including genetic groups as fixed effects in the model reduced the accuracy of estimation of selection response. This could be due to a fixed groups model not accounting accurately for the between-genetic group variance. However, the differences in the means of different groups were estimated unbiased when genetic groups were treated as fixed effects.

In the study of Phocas and Laloe (2004), when connectedness (proportion of progeny in each genetic group derived from common link sires) among genetic groups, genetic trend and the heritability were increased, there was a reduction in the bias in selection response. This was due to a reduction in the variance of EBVs resulting from an increase in the proportion of link progeny from common sires. In addition, the variance of estimation of differences between groups was large when genetic groups were fitted as fixed effects. Including random genetic group effects was found to only minimize the MSE between different groups with significantly varying genetic means in the presence of sufficient connectedness. This indicates that the choice of fitting fixed or random genetic group effects is dependent on the heritability and the level of connectedness across groups. If connectedness between groups is not adequate, it will be appropriate to treat genetic groups as random effects.

In a simulation study involving only two flocks or genetic groups, Kinghorn and Shepherd (1990) applied mixed model methodology to evaluate the accuracy of prediction of differences between base flock genetic means for fixed and random genetic grouping strategies for clean fleece weight and fibre diameter. For varying flock sizes and generations, fitting random genetic groups led to a slight increase in the accuracy of prediction when compared to fixed genetic grouping strategy. However, this difference in the accuracy was not significant.

Foulley *et al.* (1992) developed a criterion to assess the magnitude of differences between groups based on data structure and the degree of connectedness across groups prior to including genetic groups as fixed effects in the genetic evaluation model. If the differences between groups are not large and limited amount of data and pedigree are

available, random genetic group effects can be fitted as suggested by Foulley *et al.* (1990b). Foulley *et al.* (1992) stated that it would be advantageous to assume genetic groups as random effects when small group sizes can not be avoided. When genetic groups need to be included in the evaluation procedure for very large datasets, treating group effects as fixed or random must not make any significant difference in the ranking of animals within $\mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{a}}$ although the accuracy of genetic group solutions may be reduced for smaller groups. However, when the number of animals in certain groups is relatively smaller than the others, a characteristic of field datasets, this would lead to inaccurate (large standard errors) estimates of the genetic group solutions (Schaeffer, 1994) for smaller groups. This is a serious problem that warrants attention with special focus on the Merino evaluation schemes in Australia where groups formed vary significantly in their size. The problem regarding the accuracy of estimation for smaller flocks could be solved by pooling similar flocks with regard to their estimated genetic group effects into one bigger group (or cluster) using clustering procedures.

Woolaston (1997) examined the statistical power of detecting differences between genetic groups with reference to FEC (Faecal egg count) infection in sheep. The author reported that if the real difference between groups was three times the phenotypic standard deviation, the number of animals required per genetic group would be 100. The author also concluded that 400 animals per genetic group are required if the difference is two times the phenotypic standard deviation implying that more animals per group are needed if the difference is smaller and vice-versa. However, these findings cannot be practically implemented when large numbers of genetic groups need to be fitted in the model because estimating differences between different combinations of genetic groups would be a tedious task.

The PEV of fixed genetic group effects has no upper limit unlike the random within-group genetic deviations (estimated in a model without genetic group effects) for which the PEV has to be less than the additive genetic variance (Hanocq *et al.*, 1996; Kuehn, 2005; Kuehn *et al.*, 2007). Since the EBVs from a genetic groups model are expressed as a function of random genetic deviations and fixed genetic group effects, an increase in the error variance of the fixed genetic group effects would subsequently increase the PEV of the EBVs from a genetic groups model (Tong *et al.*, 1980; Kuehn,

2005; Kuehn *et al.*, 2007). Therefore, the accuracy of genetic evaluation using fixed genetic groups model is strongly dependent on the size of the groups (number of animals) and the degree of connectedness (Hanocq *et al.*, 1996; Kuehn, 2005; Kuehn *et al.*, 2007). If the flock or group size is small, estimated genetic group solutions will have large standard errors irrespective of the degree of genetic links between flocks. Therefore, a small proportion of bias in genetic group solutions is acceptable when assuming random genetic groups, which leads to a reduction in the associated mean square error (Schaeffer, 1994).

On the contrary, if limited linkage exists between some flocks and the number of groups defined in an evaluation scheme is large with varying group sizes, it would be suitable to treat group effects as random. The latter scenario will be more applicable to the Merino population in Australia which is characterized by significant differences between strains and bloodlines within strains. Additionally, in a multiple-trait evaluation scheme, a model with random genetic group effects needs information on the between-group variance for each trait and the covariances between traits. To use within-group (co)variances might be an appealing approach where genetic groups are used to cover for “gaps” in the pedigree back to a likely single base population, a scenario common to dairy cattle industry. However, it is likely to be a different scenario when genetic groups are used to account for different genetic means of multiple base populations and subpopulations which are known to have a very different variance and covariance structure for between-flock versus within-flock variation. The latter scenario is a common feature in Australian Merino data used for genetic evaluation by Sheep Genetics.

2.2.5 Estimation of breeding values with maternal genetic groups

Maternal effects are usually included in an animal model for genetic evaluation of a number of traits in meat and wool sheep industries. Maternal effects can be partitioned into dam genetic and dam permanent environmental effects. As reviewed in this chapter, genetic groups are generally used to account for prior selection on base animals that have missing records implying that genetic grouping corrects for the bias in

the EBVs. However, when maternal effects are significant for a trait and the female ancestors in the pedigree have missing performance records, the EBVs would contain a component of the dam's maternal genetic or permanent environmental effect, thus reducing the accuracy of evaluation procedure. Therefore, a strategy is required to assign genetic groups for maternal genetic effects in the model.

Van Vleck (1990a, b) proposed a model to incorporate genetic groups for maternal genetic effects besides the random additive genetic effects on the basis of similar criteria. According to this approach, the parent animals of dams with missing observations can be assigned to the same genetic group as the dam without a record would have been allocated for models without maternal effects, that is, for additive genetic models only. This approach can also handle non-genetic (environmental) maternal effects depending on the data structure. However, the potential problem of interdependencies among the rows (and columns) of the coefficient matrix obtained under this technique cannot be avoided in some cases (Van Vleck 1990a) where group effects are treated as fixed. The problem of interdependency in the coefficient matrix could be solved by assuming genetic groups to be random effects. Alternatively, depending on the data structure and selection history of traits, genetic groups can be assigned differentially for additive genetic and maternal genetic effects in an animal model evaluation.

2.2.5.1 Differential genetic grouping for direct additive and maternal genetic effects

Cantet *et al.* (1992a, b) proposed a reduced animal model (RAM; Quaas and Pollak, 1980) with differential criteria for including genetic groups for additive genetic and maternal effects. They derived MME that included missing dams of animals with performance records. The modified MME facilitated the use of differential criteria of assigning genetic groups (differential genetic groups) for direct and maternal effects. In general, an animal model with genetic groups for direct additive and maternal effects (defined on the basis of similar criteria) as described by Van Vleck (1990a, b) can be written as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_o\boldsymbol{\alpha}_o + \mathbf{Z}_m\boldsymbol{\alpha}_m + \mathbf{E}_m\mathbf{e}_m + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of n observations, and $\mathbf{X}$ is an $n \times p$ incidence matrix that associates observations in $\mathbf{y}$ with p location parameters (fixed effects) in vector $\mathbf{b}$. The incidence matrices $\mathbf{Z}_o$, $\mathbf{Z}_m$ and $\mathbf{E}_m$ relate the random additive genetic, maternal genetic and maternal environmental effects in vectors $\boldsymbol{\alpha}_o$, $\boldsymbol{\alpha}_m$ and $\mathbf{e}_m$, respectively with observations in $\mathbf{y}$, and $\boldsymbol{\varepsilon}$ is a vector of random residuals. The number of animals (a) in $\boldsymbol{\alpha}_o$ and $\boldsymbol{\alpha}_m$ are the same and contain their direct and maternal breeding values, respectively. The $a \times 1$ random vectors of direct ($\boldsymbol{\alpha}_o$) and maternal ($\boldsymbol{\alpha}_m$) effects can be expressed as:

$$\begin{bmatrix} \boldsymbol{\alpha}_o \\ \boldsymbol{\alpha}_m \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_o \mathbf{g}_o \\ \mathbf{Q}_m \mathbf{g}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\alpha}_o^* \\ \boldsymbol{\alpha}_m^* \end{bmatrix}$$

where $\mathbf{Q}_o(a \times n_{go})$ and $\mathbf{Q}_m(a \times n_{gm})$ are the known incidence matrices associating breeding values (direct and maternal) with their corresponding means, that is, the n_{go} number of elements of vector $\mathbf{g}_o$ (containing additive genetic group effects) and the n_{gm} number of elements of vector $\mathbf{g}_m$ (containing maternal genetic group effects). The dimensions of $\mathbf{Q}_o$ and $\mathbf{Q}_m$ may be the same if similar criteria are used for assigning genetic groups for additive genetic and maternal effects [as proposed by Van Vleck (1990a, b)] or unequal when differential genetic groups are assigned for additive genetic and maternal effects as proposed by Cantet *et al.* (1992a, b). The vectors of random effects, $\boldsymbol{\alpha}_o^*$ and $\boldsymbol{\alpha}_m^*$ are such that;

$$E \begin{bmatrix} \boldsymbol{\alpha}_o^* \\ \boldsymbol{\alpha}_m^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \text{Var} \begin{bmatrix} \boldsymbol{\alpha}_o^* \\ \boldsymbol{\alpha}_m^* \end{bmatrix} = \begin{bmatrix} \sigma_{Ao}^2 & \sigma_{AoAm} \\ \sigma_{AmAo} & \sigma_{Am}^2 \end{bmatrix} \otimes \mathbf{A}$$

where $\mathbf{A}$ is the numerator relationship matrix of dimensions $a \times a$, and σ_{Ao}^2 , σ_{Am}^2 and σ_{AoAm} are the variances of direct additive and maternal genetic effects and the covariance between the two effects, respectively and $\otimes$ is the Kronecker product.

A common problem encountered while specifying the variance of observations in vector $\mathbf{y}$ is the mis-specification of the additive part of the variances or the covariances between individual animals when there is missing information on their

dams. This may lead to confounding of direct and maternal genetic group effects (Cantet *et al.*, 1992a). In certain cases, this problem can be solved by including “phantom” dams in α_o and α_m . However, the inclusion of “phantom” dams in the pedigree would increase the computation burden (Cantet *et al.*, 1992a). Nevertheless, the strategy of differential genetic grouping for direct and maternal effects will lead to the estimation of separate additive and maternal effects with higher accuracy for traits which are influenced significantly by maternal characteristics. In addition, animals with missing records and sires but with known dams will get an EBV based on the estimated maternal effects and the covariance between the maternal genetic and additive genetic effects. However, the approach of Cantet *et al.* (1992a, b) will be more relevant to the evaluation method when maternal effects are significant for some trait(s).

In a study involving field data, Cantet *et al.* (1993) estimated variance components using an animal model with genetic groups for direct and maternal effects based on similar criteria. They found difference in the estimates of direct and maternal genetic trends and concluded that differential criteria must be used for assigning genetic groups for direct additive and maternal effects in the model. However, differences between genetic groups for direct effects could be smaller in magnitude than those for the maternal effects in which case differential grouping might not be of any significance. Nevertheless, further research should examine whether it would be feasible for REML methods to have different grouping strategies for additive and maternal genetic effects.

2.2.6 Incorporating genetic groups in a multi-trait animal model evaluation

As reviewed in the previous sections of this chapter, several authors have proposed and discussed the principles and methodology of including genetic groups in the MME. Most strategies proposed did not deal specifically with the issue of fitting genetic groups in a multi-trait animal model evaluation. Da and Grossman (1991) presented the modified MME to accommodate for genetic groups with respect to

multiple trait animal model evaluation. The following assumptions were considered for this purpose:

1. The additive genetic values of each trait have a non-zero mean which is defined as a linear function of genetic group (fixed) effects of missing parents of known animals. This assumption would account for varying genetic means of base populations for each trait.
2. The additive genetic relationship between an animal and groups of unknown parents is the same for each trait.

Da and Grossman (1991) argued that the latter assumption is a reasonable one if equal selection pressure is applied to each trait and when autosomal inheritance is considered. The same assumption would hold if groups are replaced by breeds. A multiple-trait mixed model with genetic groups can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{ZA}_m\mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ ($N \times 1$) is a vector of N records on t traits; $\mathbf{X}$ and $\mathbf{Z}$ are the known incidence matrices relating systematic effects other than the genetic group effects in $\mathbf{b}$ ($pt \times 1$) and random genetic effects in $\mathbf{a}$ ($qt \times 1$) with observations in $\mathbf{y}$, respectively; $\mathbf{A}_m$ ($qt \times nt$) is an additive genetic relationships matrix that associates each of t traits on q animals with n genetic groups of missing parents, and is the same as matrix $\mathbf{Q}$ in Equation 2.1; $\mathbf{g}$ is a vector of fixed genetic group effects; p is the number of fixed effects in $\mathbf{b}$; q is the number of additive genetic effects in $\mathbf{a}$, and $\mathbf{e}$ ($N \times 1$) is a vector of random residual errors associated with each observation. Additionally, it was also assumed that $N = qt$ if one record is available per trait and there are no missing observations. The MME for the above model can be expressed as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{ZA}_m \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZA}_m \\ \mathbf{A}_m'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{A}_m'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{A}_m'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZA}_m \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{A}_m'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where $\mathbf{G} = \text{var}(\mathbf{a}) = \mathbf{G}_0 \otimes \mathbf{A}$; $\mathbf{R} = \text{var}(\mathbf{e}) = \mathbf{R}_0 \otimes \mathbf{I}_q$, for one record on each trait assuming no missing data; $\mathbf{A}$ ($q \times q$) is the numerator relationship matrix; $\mathbf{I}_q$ is the identity matrix of order q ; $\mathbf{G}_0$ ($t \times t$) is the covariance matrix of additive genetic values

for the t traits, and $\mathbf{R}_o$ ($t \times t$) is the covariance matrix of random residual errors associated with the t traits. The above MME are the same as those in Equations 2.1 and 2.3 for a single trait model with genetic grouping.

As stated by Da and Grossman (1991), the MME presented above is not feasible computationally since $\mathbf{A}_m$ is difficult to compute. Computing of one element of $\mathbf{A}_m$ might require a search through the entire pedigree. Therefore, the authors added that the approach of Quaas (1988), Westell *et al.* (1988) and Wiggans *et al.* (1988) for a uni-trait animal model and later complemented by QP transformation (Quaas and Pollak, 1981) can be followed to derive computationally feasible MME. The modified MME for multi-trait evaluation can then be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^-\mathbf{X} & \mathbf{X}'\mathbf{R}^-\mathbf{Z} & 0 \\ \mathbf{Z}'\mathbf{R}^-\mathbf{X} & \mathbf{Z}'\mathbf{R}^-\mathbf{Z} + \mathbf{G}_o^{-1} \otimes \mathbf{T}_{uu} & -\mathbf{G}_o^{-1} \otimes \mathbf{T}_{ug} \\ 0 & -\mathbf{G}_o^{-1} \otimes \mathbf{T}_{gu} & \mathbf{G}_o^{-1} \otimes \mathbf{T}_{gg} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^-\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^-\mathbf{y} \\ 0 \end{bmatrix}$$

where $\mathbf{R}^-$ is the zero-type generalized inverse of $\mathbf{R}$, and $\mathbf{T}$ is defined as follows:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{uu} & \mathbf{T}_{ug} \\ \mathbf{T}_{gu} & \mathbf{T}_{gg} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{A}_g \\ -\mathbf{A}_g'\mathbf{A}^{-1} & \mathbf{A}_g'\mathbf{A}^{-1}\mathbf{A}_g \end{bmatrix}$$

This strategy of replacing $\mathbf{R}^{-1}$ with $\mathbf{R}^-$ was suggested by Henderson and Quaas (1976) under which zero elements are set that correspond to missing values in $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{R}$, in order to keep $\mathbf{G}$ unchanged. Da and Grossman (1991) used a similar strategy by setting to zero elements corresponding to missing data in $\mathbf{Z}\mathbf{A}_m$ so that $\mathbf{A}_m$ remains unaltered. In addition, as pointed by the authors, this approach does not require modifying $\mathbf{A}_m$ for the purpose of setting to zero elements in $\mathbf{Z}\mathbf{A}_m$ because when a particular row in $\mathbf{Z}_m$ is set to zero, the same row in $\mathbf{Z}\mathbf{A}_m$ is set to zero automatically. The matrix $\mathbf{A}_m$ is exactly the same as $\mathbf{Q}$ -matrix described previously.

The above MME can be used to estimate differences between genetic groups specific to each trait. It is important to note that $\mathbf{u} = \mathbf{a} + \mathbf{A}_m\mathbf{g}$ is a non-estimable function (Van Vleck, 1990a), and only pair-wise differences between the elements of $\mathbf{u}$ are usually estimable (Da and Grossman, 1991). Therefore, unique genetic group solutions specific to a trait cannot be derived using the MME mentioned above. However,

estimability of genetic group effects needs to be investigated separately for each trait in case there is missing data (Da and Grossman, 1991). In addition to the estimability of genetic group effects, it is equally important to examine the accuracy of estimated genetic group effects separately for each trait. Depending on the breeding objectives, the numbers of missing records could vary for different traits in industry datasets. For example, breeders in low rainfall regions may be less interested in recording worm egg counts (WEC) in comparison to those in high rainfall areas. Therefore, in a multi-trait animal model evaluation including genetic groups, differences in data recording and selection history between traits would affect the accuracy of estimated genetic group solutions differently for each trait.

2.3 Considerations regarding estimation of genetic group effects

2.3.1 Connectedness in across-flock evaluation

Across-flock genetic evaluation provides a basis for more efficiently designed multiple-flock breeding programs (Kinghorn and Shepherd, 1990) and is gaining popularity amongst the Australian Merino ram breeders (Banks, 1997; Atkins *et al.*, 1998) to maximize genetic gain for a combination of traits of economic importance. However, across-flock evaluation requires genetic connectedness in order to compare animals between flocks (Kinghorn and Shepherd, 1990; Foulley *et al.*, 1990b; Lewis *et al.*, 1999; Mathur *et al.*, 2002; Kuehn, 2005). Connectedness is best established through the use of common sires in different management units (flocks) in order to create genetic relationships across units. This provides a basis through which genetically-unrelated animals can be compared (Kuehn, 2005; Kuehn *et al.*, 2007). In the absence of genetic linkage, animals from one flock cannot be fairly compared with animals from another flock due to confounding of genetic and environmental effects (Schaeffer, 1975; Foulley *et al.*, 1992; Kuehn, 2005; Kuehn *et al.*, 2007).

Data structure is an important issue (Tosh and Wilton, 1994; Clement *et al.*, 2001) that needs to be considered prior to comparing animals across flocks. For example, different types of effects (e.g., sire, flock, sex, etc.) included in mixed linear

models of genetic evaluation need to be connected for the accurate estimation of pairwise differences between varying levels of those effects. In the absence of proper (quality) and sufficient (degree) connectedness across different levels of fixed effects, these differences are inestimable and the data are said to be disconnected (Kennedy and Trus, 1993). Therefore, it is important to determine the level of connectedness in across-flock evaluations.

2.3.1.1 Measures of connectedness

Several authors have proposed different measures that can be used to assess the degree of connectedness across subpopulations or management units (flocks or herds). Tosh and Wilton (1990) defined the degree of connectedness as “the probability of making valid pairwise comparisons among factor (any fixed effect) levels when analyzing all data together” which for the i -th factor can be written as:

$$c_i = n_{pi} / [n_i (n_i - 1)/2]$$

where n_{pi} is the number of unique pairs across all groups and n_i is the total number of levels of the i -th factor such that $n_i \geq 2$. For the complete design considering all factors or effects, the total degree of connectedness was defined as:

$$c_T = (\sum n_{pi}) / (\sum [n_i (n_i - 1) / 2])$$

According to the above definition, the values of degree of connectedness range from 0 (all levels of the factors are disconnected) to 1 (complete connectedness). However, the application of this measure of connectedness to complex multi-trait models (e.g., random regression) including a large number of fixed and random genetic effects would make comparison between varying levels of fixed effects a tedious task.

Kennedy and Trus (1993) proposed that the prediction error variance (PEV) of differences in the EBVs of animals from different management units could be used as an indicator of connectedness. They applied the PEV method to a sire model which is an example of a simple breeding structure where connections are only through paternal half-sibs and later extended to an animal model. Under an animal model, the PEV can be expressed as follows:

$$\text{var}(\hat{a} - a) = (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}^{-1}\lambda)^{-1} \sigma_e^2$$

where $\hat{a}$ and a are the predicted and true genetic values, respectively; $\mathbf{Z}$ is the known incidence matrix relating observations with animals, $\mathbf{M}$ is the absorption matrix for the fixed effects in $\mathbf{X}$ and is written as; $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{I}$ is an identity matrix and λ is the ratio of residual to additive genetic variance. For biased evaluations (with poor or no connectedness), the MSE (Mean Square Error) of prediction was suggested to be equivalent to PEV.

Kennedy and Trus (1993) argued that the PEV of an individual animal (diagonal element of the inverse of the coefficient matrix of MME) remains constant provided the same numbers of animals are included in the evaluation scheme. However, they pointed that the PEV of difference in EBVs depends on whether animals are compared within or across management units. For example, when animals are evaluated within flock, the PEV is a function of residual variance within flock. On the other hand, when animals from one flock are compared with those from another flock, the PEV is a function of both residual variance within and that across flocks. Therefore, it was suggested that the accuracy of prediction of fixed effects of management units is an indicator of the degree of connectedness. However, as suggested by Roso *et al.* (2002), Mathur *et al.* (2002) and Roso *et al.* (2004), the practical implication of this statistic (PEV) of connectedness is limited as it is extremely time consuming and not suitable for routine evaluations.

The three other alternative measures for assessing connectedness across management units proposed by Kennedy and Trus (1993) are the gene-flow method, genetic drift variance and variance of estimates of management unit effects. These three methods were proposed as an alternative to computing PEV because computation of the PEV matrix is cumbersome or impossible for extremely large datasets (Kennedy and Trus, 1993). Kennedy and Trus (1993) estimated correlations between the PEV and the other three alternative measures of connectedness. They found correlations of 0.99, 0.92 and 0.67 with the variance of management unit effects, genetic drift variance and gene-flow method, respectively indicating the relative accuracy of each alternative measure with respect to the PEV. The high correlations suggest that the estimation of variance of management unit effects is a fairly reliable method for assessing the level of

connectedness. Since sire pedigree is generally well recorded by most breeders, the variance of estimated management unit effects will be especially useful where flocks have some proportion of genetic links established through the use of both common sires and dams, and where dam pedigree is missing.

The minimization of PEV does not necessarily lead to increased rate of genetic progress because it may lead to reduced selection intensity as a result of selecting genetically related animals (Kennedy and Trus, 1993). A similar argument was put forth and proved mathematically by Laloe *et al.* (1996) who stated that an increase in genetic relationships decreases the PEV with a simultaneous reduction in the selection intensity and the genetic variability. Selecting genetically related animals could also lead to a higher inbreeding level in the later generations and subsequently lead to a reduction in the genetic trend. Therefore, an equilibrium between the level of connectedness and selection intensity must be maintained for maximizing genetic gain (Roso *et al.*, 2004) in a population. The measures of connectedness proposed by Kennedy and Trus (1993) could be used depending on the amount and type of data available and the computational requirements. However, Kennedy and Trus (1993) did not present an overall criterion for assessing the precision of a genetic evaluation scheme (Laloe *et al.*, 1996).

Mathur *et al.* (1998, 2002) stated that the PEV of the difference between management unit effects as proposed by Kennedy and Trus (1993) is dependent on the size and structure of management units and the type of connections between them. Mathur *et al.* (1998, 2002) proposed a new statistic, connectedness rating (CR) to disentangle the concept of connectedness from the effects of herd size and structure. CR is defined as a correlation between estimated effects of two management units and can be expressed as:

$$CR = \frac{\text{cov}(\hat{b}_1, \hat{b}_2)}{\sqrt{\text{var}(\hat{b}_1) \text{var}(\hat{b}_2)}}$$

where $\hat{b}_1$ and $\hat{b}_2$ are the predicted herd effects for herds 1 and 2, respectively. Mathur *et al.* (2002) suggested that using CR would be more beneficial to distinguish between reduction in accuracy due to poor connectedness and that associated with small

management unit size. Nevertheless, the authors added that connectedness measured using CR and management unit size have an important impact on the accuracy of across-management unit evaluations which would be reflected in the variance of estimated management unit effects or the PEV. Therefore, they further suggested that increasing management unit size will improve accuracy more than increasing connectedness.

Laloe (1993) and Laloe *et al.* (1996) proposed a criterion, the generalized coefficient of determination (CD) for evaluating the quality (quality of the prediction of a contrast) of genetic evaluation procedure by assessing the degree of disconnectedness between random effects. It was defined as the CD of a linear combination of genetic values ($\mathbf{x}'\mathbf{u}$) corresponding to the square of correlation between the estimated and true difference in breeding values. The accuracy of comparison is greater when the value of CD gets closer to 1. Laloe (1993) argued that a design with a random effect could be disconnected if the CD of a contrast between different levels of a random effect is zero. The generalized CD proposed by Laloe (1993) can be written as:

$$CD(\mathbf{x}) = 1 - (\text{var}(\mathbf{x}'\mathbf{u} \mid \hat{\mathbf{u}}) / \text{var}(\mathbf{x}'\mathbf{u})) = \mathbf{x}' (\mathbf{A} - \lambda \mathbf{C}^{uu}) \mathbf{x} / \mathbf{x}' \mathbf{A} \mathbf{x}$$

where $\mathbf{x}$ is a vector which defines a contrast and whose elements sum to zero, $\mathbf{u}$ is a vector of true breeding values (random factors), $\hat{\mathbf{u}}$ is the BLUP of $\mathbf{u}$, $\mathbf{A}$ is the additive genetic relationship matrix, λ is the ratio of residual to genetic variance, and $\mathbf{C}^{uu}$ (PEV matrix) = $(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \lambda\mathbf{A}^{-1})^{-1}$; $\mathbf{M}$, $\mathbf{X}$ and $\mathbf{Z}$ are as previously defined.

Laloe (1993) stated that “CD measures the amount of information supplied by the data that has contributed to the prediction of $\mathbf{u}$ ”. The author also stated that a positive CD implies that relevant information is supplied by the data for a contrast between different levels of random effects included in the model. Although contrasts among different levels of random effects are always estimable as has been shown by Foulley *et al.* (1990b, 1992), it is important to note that a prior information about data structure (design) using different measures of connectedness reviewed above is not sufficient (Laloe *et al.*, 1996) in terms of estimating the relative accuracies of genetic evaluation procedures. In the words of Laloe *et al.* (1996), “an orthogonal (complete connectedness) design could not lead to any genetic progress”. Additionally, a

disconnected design could have significant impact on the CDs of contrasts and matrix ranks, thus affecting the estimates of breeding values (Laloe, 1993). The degree of connectedness has similar consequences in terms of a decrease in the rank of a coefficient matrix (e.g., Foulley *et al.*, 1990b) for both fixed and random factors (Laloe, 1993; Laloe *et al.*, 1996). Laloe *et al.* (1996) recommended the use of CD for assessing the precision of genetic evaluation or for the optimization of corresponding designs because CD takes into consideration the data structure and the amount of information plus it accounts for the PEV and genetic variability. Another advantage of CD as stated by Laloe (1993) is its application in judging the precision of comparison between animals born in different years and in cluster analysis to form groups (clusters) of animals that are at a comparable genetic level.

Hanocq and Boichard (1999) evaluated connectedness in the French Holstein population using the generalized CD (Laloe, 1993) using a full model (with herd effect) and a reduced model (without herd effect). Hanocq and Boichard (1999) concluded that the CD provided information regarding the amount of data and the connectedness of the design when a full model was fitted. However, when a reduced model was used, the CD was affected only by the amount of data when the design was fully connected (orthogonality). The authors stated that the actual effect of connectedness on individual CDs is limited when a large number of bulls are sampled per stud and each bull is compared to a large subpopulation. Thus, they concluded that the individual CD is not a useful measure for estimating the level of connectedness. Furthermore, computing CD could be cumbersome because of the storage and processing time associated with the estimation of the inverse of the coefficient matrix and the (non-inverted) numerator relationship matrix (Kuehn *et al.*, 2007).

A different measure for assessing connectedness between varying levels of fixed and random animal effects (breeding values) in an animal model evaluation has been defined by Foulley *et al.* (1992):

$$\gamma = (|C_R| / |C_F|)^{1/n}$$

where $|C_R|$ and $|C_F|$ are the determinants of the coefficient matrices under the full (F) model (with environmental effects) and reduced (R) model (without environmental

effects), respectively and n is the number of animals. The value for γ varies from 0 (complete disconnectedness) to 1 (perfect connectedness or orthogonality). Although perfect connectedness across sub-populations will allow for the accurate estimation of differences between varying levels of different fixed effects included in the model, it should be borne in mind that an orthogonal design may not necessarily lead to any genetic improvement as pointed by Laloe *et al.* (1996). The authors stressed that a prior knowledge of the data structure regarding connectedness is not sufficient for assessing the accuracy of evaluation procedure. However, a prior knowledge of the data structure would aid in ascertaining whether across-management unit evaluation could be undertaken.

Hanocq *et al.* (1996) estimated connectedness on the basis of γ described above and found that including fixed environmental effect in the model (Full model) leads to relatively higher loss in accuracy when the degree of connectedness is high. On the contrary, they stated when low connectedness exists, the accuracy is already low using a reduced model (without environmental effect) thereby limiting the relative loss in accuracy when adding the environmental effect to the model. Therefore, Hanocq *et al.* (1996) concluded that γ is an indicator of the quality of connectedness in a design rather than an absolute measure. The quality of connectedness may be one-directional or two-directional depending on the use of link sires from one population into another for generating link progeny.

Another measure of connectedness proposed by Foulley *et al.* (1992) is defined as the ratio of prediction error variance from a reduced (R) model (without environmental effects) to that from the full (F) model (with environmental effects). The authors presented an example to evaluate the degree of connectedness (due to the incidence of other fixed effects) between different genetic group levels. They stated that the measure of connectedness can be used to test whether differences among groups if any can be precisely captured by the data structure prior to estimating differences between genetic groups (Foulley *et al.*, 1992).

Roso *et al.* (2002) evaluated the degree of connectedness between groups of station-tested beef bulls in Canada using different measures of connectedness for their

relative accuracy. The measures used were CR (Mathur *et al.*, 2002), PEV of the differences between EBVs across groups (Kennedy and Trus, 1993) and GLt which was defined by Fries (1998) as the number of direct genetic links established between groups as a result of common sires and dams. Roso *et al.* (2002) concluded that the computational burden in estimating CR for large datasets is quite substantial in comparison to GLt. They found CR to be highly correlated with the group size (correlation of 0.95), an argument against that of Mathur *et al.* (2002). Additionally, Roso *et al.* (2002) stated that differentiation of completely disconnected groups from connected ones on the basis of CR was not always possible since large disconnected groups had CR estimates comparable with those of groups with connections. They further stated that GLt was more efficient in identifying groups without any genetic links. However, following Roso *et al.* (2004), there is a possibility that some of the groups lacking direct genetic links may be indirectly connected and therefore, may have a higher degree of accuracy of comparison of EBVs across them. Thus, Roso *et al.* (2002) stressed that GLt is not a precise indicator of the degree of connectedness between different groups. They also stated that the arithmetic mean of GLt related with each pair of groups is not a sufficient measure since pairs of groups having similar arithmetic means can in fact have varying levels of connectedness between them. The authors therefore, recommended the use of harmonic mean of GLt rather than the arithmetic mean.

Lewis *et al.* (1999) assessed the relationship between connectedness and errors of prediction of differences in EBV between pairs of animals in different flocks. Connectedness was defined as the average prediction error correlation (PEC) between flock genetic means. They concluded that the prediction error correlation (averaged over flocks) decreases with increasing connectedness across flocks. The measure of connectedness (PEC) proposed by Lewis *et al.* (1999) will be useful for small flock sizes only where individual comparisons can be made in a relatively simple manner. Kuehn *et al.* (1999) measured connectedness in a group breeding scheme with reference to sheep industry using PEC. They found similar results as those of Lewis *et al.* (1999). According to Huisman *et al.* (2006), however, it is virtually impossible to compare these correlations (PEC) for all animals undergoing evaluation.

Most measures of connectedness cannot be implemented when the evaluation scheme involves a large number of animals (Roso *et al.*, 2004) which is a common feature with most industry datasets. Most studies examined the significance of connectedness in relation with the comparison of animals across management units. Further studies are therefore, required to examine the significance of connectedness across genetic groups with special focus on across-flock evaluation in the Australian Merino industry. Different definitions of connectedness have been used by different authors and some of the proposed methods are computationally demanding (Roso *et al.*, 2004; Huisman *et al.*, 2006), thus making their application in field datasets quite cumbersome (Mathur *et al.*, 2002; Huisman *et al.*, 2006). The factors that need to be considered prior to assessment of genetic ties between groups of animals using different criteria are the data structure (e.g., Roso *et al.*, 2004), group size and computational requirements. The measure of connectedness should be able to provide precise information regarding the magnitude and quality (e.g., Huisman *et al.*, 2006) of genetic ties among groups being evaluated. In addition, it is important to note that random factors (breeding values) have larger numbers of levels than the fixed factors, thereby affecting the degree of connectedness of the complete design (Tosh and Wilton, 1990). Therefore, a prior understanding of the underlying principles of connectedness among random effects is required for deriving across-flock EBVs with higher accuracy.

The choice of criteria for estimating connectedness depends on the objective of study and the structure of the population (Hanocq and Boichard, 1999; Roso *et al.*, 2004). There is no standard method in the literature for deciding on the measure of connectedness for a given data structure and vice versa (Fries, 1998; Kuehn *et al.*, 2007). As stated by Mathur *et al.* (2002), “connectedness is a statistical measurement rather than one of genetic relationship”. Fernando *et al.* (1983) concluded that estimating connectedness is a statistical requirement for certain applications and computational requirement for others. Mathur *et al.* (2002) showed that the methods for measuring connectedness used by Kennedy and Trus (1993) were based on the degree of genetic relationships among animals rather than on a statistical measure (e.g., based on gene flow or on additive genetic relationships) and therefore do not provide accurate results in most cases. Fries (1998) stressed that the mixed model evaluation always

provides a solution irrespective of the data structure and without any minimal demand regarding the level of connectedness. This can lead to biased estimates of EBVs when animals are compared across unconnected management units. This argument of Fries (1998) will be especially true for Australian Merino industry data where some flocks may have adequate genetic links whereas others may either lack sufficient linkage or may be unlinked. The lack of genetic linkage between some flocks will be a constraint to an unbiased evaluation of animals across flocks. Additionally, in the statistical sense, the variation in flock size (number of animals) in Australian Merino population is likely to affect the accuracy (standard errors) of estimated genetic group effects. Therefore, the effect of linkage and flock size on the accuracy of across-flock evaluation in Merinos needs attention. Furthermore, the type of data structure required for an accurate estimation of genetic group effects in across-flock evaluations requires further research.

2.3.1.2 Factors affecting the accuracy of across-flock genetic evaluation

The widespread use of artificial insemination (AI) in livestock industry has established genetic links across herds and therefore an unbiased evaluation of animals across a range of climatic and production environments is possible. However, the efficient use of AI is dependent on several factors like species, production and management environment, and the degree of cooperative participation by breeders in well organized selection programs (Foulley *et al.*, 1990b). Despite the use of AI and natural service in farm livestock like beef cattle and sheep (Foulley *et al.*, 1990b), herds or flocks are generally isolated geographically (Huisman *et al.*, 2006; Kuehn *et al.*, 2007). Therefore, genetic links in the beef or sheep industry may not be adequate to accurately estimate genetic group effects and genetic trend across flocks. However, having good linkage across years by using the same sires within a flock over a time period will improve the prediction of genetic trend.

In a simulation study, Hanocq *et al.* (1996) investigated the influence of connectedness on genetic trend. They concluded that the establishment of genetic ties between subpopulations improves the genetic trend of the entire population and introduces heterogeneity of response to selection across subpopulations. The researchers

attributed the increase in genetic trend to gene flow from subpopulations with a high genetic mean to those with a low genetic level as a result of connectedness, thereby gradually reducing the differences in the average genetic merit between subpopulations. However, the authors concluded that the improvement in genetic trend due to connections would result only when there are significant differences in the genetic means of subpopulations. Poor connectedness limits the overall genetic trend and introduces a large bias in across-flock evaluation as the differences between subpopulations are inestimable and are assumed to be zero. Keeping in view the existing genetic differences between flocks, and across strains and bloodlines in the Australian Merino population, the establishment of linkage would lead to the overall genetic improvement.

As discussed in this section, there are several advantages of establishing connectedness across subpopulations and management units. Commercial sheep producers and breeders can participate in cooperative breeding schemes using reference sires for rapidly improving genetic links among flocks (Kuehn *et al.*, 1999; Kuehn *et al.*, 2007) through AI. Sire referencing schemes permit increased selection intensity which leads to increased rates of genetic gain (Simm *et al.*, 2001). Such schemes have led to the genetic improvement of many specialized meat sheep breeds in Britain (Simm *et al.*, 2001). In recent times, there has been some market driven exchange of sires and dams to provide linkage between small groups of flocks (Huisman *et al.*, 2006). The use of common sires with superior genetic merit would be beneficial for improving connectedness across groups (Roso *et al.*, 2004). Therefore, individual ram breeders and artificial breeding companies should encourage increased use of link sires across flocks on a national basis to minimize the bias in across-flock evaluations due to regional differences. Currently, in order to express the level of linkage between flocks, Sheep Genetics publishes Australian sheep breeding values (ASBV) versus Flock breeding values (FBV), and the distinction between the two are determined by the level of linkage for each flock.

Besides linkage, another important issue associated with an across-flock evaluation scheme that needs to be considered is the possibility of variation in the performance of a sire's progeny tested in different locations which implies sire x

environment interactions (Cottle and James, 1997; James and Cottle, 1997; Atkins *et al.*, 1999a; Atkins *et al.*, 1999b). Sire x flock interactions could be significant depending on the number of progeny per sire and the genetic merit of the ewes to which the sire is mated (Atkins *et al.*, 1999a). These interactions can significantly bias the estimates of additive genetic values of sires in across-flock evaluations which in turn can also bias the estimates of genetic trend over a time period. In a study involving field data on South African Holstein population, Theron *et al.* (2002) showed that the bias in the estimation of genetic trend reduces when phantom parent groups are included in the MME. Atkins *et al.* (1998, 1999a) found that including genetic groups (based on ram sources) in the model substantially reduced the apparent sire x environment interactions. This could be attributed to a reduction in between-flock variance by including genetic groups in the analysis. Therefore, establishing adequate linkage between flocks and accounting for any sire x environment interactions in the model will improve the accuracy of across-flock evaluations.

2.3.2 Estimability of genetic group effects using mixed model methodology

The random effects fitted in a linear mixed model are usually always predictable (Fernando *et al.*, 1983; Foulley *et al.*, 1990b). When genetic groups are not included in a model, a unique solution to random effects is always obtained irrespective of the degree of connectedness between sires and herds (Foulley *et al.*, 1990b). However, when a number of fixed effects (sex, flock, birth type, rearing type, age of animal, and age of dam) are needed to be estimated simultaneously with fixed genetic group effects using the mixed model methodology, the potential problem of “estimability” is frequently encountered (Thompson, 1979; Quaas, 1988; Wiggans *et al.*, 1988; Kennedy and Trus, 1993; Sullivan and Schaeffer, 1994; Hanocq *et al.*, 1996; Wolf *et al.*, 2002). The introduction of genetic groups as extra fixed effects in the MME needs prior knowledge of criteria to be used for defining genetic groups. This can be quite tedious in the case of beef cattle and sheep where both AI and natural mating are used (Foulley *et al.*, 1990b).

Sullivan and Schaeffer (1994) stated that confounding among genetic group effects can also occur when defined groups are not statistically independent. These

authors added that this type of problem arises if missing sires and dams are grouped separately based on different criteria and single parent pedigrees are available. In such a situation, the authors stressed that fitting unknown sire and dam groups for the same set of observations may result in identical equations for both types of groups. The researchers further stated that simple group definitions satisfying the assumption of genetically independent groups may lead to singularity resulting in estimates of breeding values which are difficult to interpret. Therefore, the estimability of genetic group effects is an important issue that requires a proper and thorough understanding as it is dependent on the data structure and the criteria on the basis of which genetic groups are formed. However, this would be cumbersome when assigning genetic groups in large industry datasets because certain flocks may have adequate genetic links while others may have suboptimal links. Therefore, the estimated genetic group effects would be more accurate for linked flocks than for flocks without sufficient linkage.

The problem of dependency due to mixing of performance records and/or other fixed effects could be overcome by the use of an algorithm proposed by Fernando *et al.* (1983). The algorithm assesses the level of connectedness between flocks in order to determine the estimability of genetic group effects. However, the degree of connectedness can vary considerably depending on the structure of the data. Therefore, if the differences between genetic groups are not really significant or if the state of connectedness does not allow for the estimation of such differences with greater accuracy, then the group effects should not be included in the model (McClintock and Taylor, 1982; Foulley *et al.*, 1990b). Further studies are required to investigate the effect of connectedness on estimable linear functions of equations and variance components (Fernando *et al.*, 1983; Tosh and Wilton, 1990) to overcome the problem of lack of connectedness between flocks and therefore, the estimability of genetic group effects to derive unbiased predictors of breeding values (Foulley *et al.*, 1990b). Hence, the estimability of genetic group effects will affect the accuracy of comparing animals across flocks.

2.4 Application of genetic grouping

2.4.1 Effect of genetic grouping on estimates of variance components

The estimation of variance components with performance records obtained through recording schemes is required for the prediction of breeding values for traits of interest. Several studies in the past have investigated the effect of genetic grouping on the estimates of variance components. Some of these were more theoretical work while others focused on specific livestock species. Pieramati and Van Vleck (1993) used simulation to examine the effect of genetic grouping on the REML estimates of additive genetic variance obtained from different amounts of data. The simulated genetic and residual variances were such that the initial heritability was 0.5. The data were analyzed using models with and without genetic group effects which were based on sex and generation of the animals. The results indicate that the models with fixed genetic group and generation effects resulted in estimates of genetic variance that were in general, lower than those from the models without either effect. The main purpose of including generation effect in the model was to examine whether it accounted for selection effects or not. The standard errors of estimates of genetic variance were much larger than those obtained without fitting any group or generation effect in the model.

Kuehn (2005) presented the following explanation for the reduced estimates of additive genetic variance using a model with genetic group effects. The additive genetic variance can be partitioned into two components: variance within base population and variance between base populations. Since management units are generally fitted as fixed environmental effects, additive genetic variance between base populations is untraceable unless sufficient genetic links exist between different base populations. When genetic groups are included in the model, the variance between base populations is described by the genetic group solutions and estimated additive genetic variance only describes the variance within base populations. Thus, artificially low estimates of additive variance are obtained when genetic groups are fitted although the variance between base populations can still be exploited in breeding programs. Similar conclusions were made by Ashtiani and James (1993), Hanocq *et al.* (1996), Atkins *et*

al. (1999b), and Diaz *et al.* (2002) with reference to the differences between within- and across-genetic group variance.

Pieramati and Van Vleck (1993) found the difference in the estimates of genetic variance to be more evident for selected populations when either genetic group or generation effect was fitted in the model. The authors also stated that in situations when only a small proportion of data on males were used, some of the base animals with relatively low genetic values and included in the pedigree of selected populations might have added to an increase in the genetic variance. In addition, an increase in the standard errors of the estimates was more marked in the case of selected populations and when generation effect was included in the model. Therefore, Pieramati and Van Vleck (1993) concluded that in the case of selective matings, it is important to include selection records in the evaluation process otherwise the estimates of additive genetic variance would be biased downwards. The authors further added that including such records would reduce the need for including genetic group effects in the model. It was also suggested that if there is no information available on the genetic relationships amongst the base animals, including a generation effect (assumed to be fixed) could reduce the bias due to selection in the estimated genetic variance.

The results of Pieramati and Van Vleck (1993) indicate that selection of base animals leads to an underestimation of additive genetic variance when genetic group effects are included in the model. The authors concluded that the effect of genetic grouping varies depending on the amount and the type of information that is missing. However, in their study, the estimates of genetic variance were unaffected, when genetic groups were treated as random. Further analytical and simulation studies are required to investigate the effect of random genetic groups on the estimates of variance components. In practice, all data since the start of selection are generally not available and not using such data on selected ancestors leads to reduced estimates of additive genetic variance due to not accounting for all gametic disequilibrium (Van der Werf and de Boer, 1990). Therefore, it would be essential to include genetic group effects for evaluating animals whose ancestors are most likely a result of genetic selection in the past. Additionally, in most species of domestic livestock, information on genetic relationships between base animals is either missing or not complete. Hence, it is

important to note that the estimates of additive genetic variance are biased by covariances, inbreeding and gametic disequilibrium to a lesser extent when additional relationships between assumed base animals are accounted for in the model (Van der Werf and de Boer, 1990).

Lopes de Oliveira *et al.* (2001) investigated the occurrence of heterogeneity of variance among genetic groups and its probable effect on the ranking of animals in the Canchim breed of beef cattle. Variance components were estimated using an approach that assumed the expression of a trait in each genetic group as a different characteristic. The authors found that the variance components were different for the traits analyzed and the models fitted. They concluded that heterogeneous (not uniform) variances existed across genetic groups used during the development of Canchim breed.

Van Vleck *et al.* (2003) examined the effect of genetic grouping to account for selection on the estimates of genetic parameters for a line of Hereford cattle. Two types of analyses were tested. In the first analysis, pedigree information on an animal's sire was included. In the second analysis, groups were assigned arbitrarily for the known sires that were assumed missing. Different numbers of groups were formed to replace varying numbers of sires and maternal (genetic and permanent environmental) effects were included in the model. The authors stated "the effect of groups in the model on estimates of variance components seems to be unpredictable". This argument can be justified by the fact that missing sires were assigned to the same and different genetic groups based on certain prior assumptions. No account was taken of the possible genetic relationships that could have existed among sires allocated to different groups, a possibility which cannot be ruled out completely due to widespread use of AI especially in the dairy and beef cattle industries.

Diaz *et al.* (2002) studied the effect of genetic grouping on the estimates of variance components for a maternally influenced trait in the Spanish Limousine population. Animal models without and with genetic group effects for both direct and maternal components were fitted as covariates. Genetic groups were formed on the basis of sex of parent animals and generation in two different approaches depending on whether complete or partial pedigree information was included. They found that the

estimates for direct additive variance and those for genetic covariance between direct and maternal genetic effects were reduced significantly when genetic groups were included in the model in comparison to a no-groups model when complete pedigree information was used. These findings are in agreement with those of Pieramati and Van Vleck (1993), and could be described using the same explanation as that presented by Kuehn (2005) as reviewed above. Diaz *et al.* (2002) stated that in a majority of datasets, the pedigree does not trace back to the original base population, but to a base population which is the result of selection. Therefore, the authors suggested that the evaluation models should account for the reduction in variance resulting from selection in the operational base population.

In the study of Diaz *et al.* (2002), the estimates for maternal genetic variance were only slightly reduced when genetic groups were included in the model and complete pedigree was considered. The authors attributed these results to less variation across maternal genetic groups because of lower selection intensity associated with the maternal component. When partial pedigree was considered in a no-groups model, the impact on variance components was not significantly large. However, if selection is practiced for a maternally influenced trait like birth weight, then it would result in indirect selection of a correlated trait which is also affected by maternal genetic component like weaning weight. In such a case, the impact of genetic grouping on estimated maternal genetic variance may be different to that reported by Diaz *et al.* (2002).

Cantet *et al.* (1993) estimated variance components for direct and maternal effects using REML on field data on weaning weight in Angus calves using an animal model with and without genetic groups. Similar grouping criteria (origin and time) were used for direct and maternal genetic effects where genetic groups were considered. The authors found that the estimates of variance components from a genetic groups model did not differ significantly from those obtained using a model without genetic groups. These results differ from those of Van Vleck *et al.* (2003) probably because of the number of effects included in the model (with and without maternal permanent environment effects). Therefore, the effects of genetic grouping on variance components are varying depending on the type of model assumed (finite locus or polygenic),

grouping criteria, pedigree information taken into account and the type and amount of data. The results reported by Pieramati and Van Vleck (1993) and Lopes de Oliveira *et al.* (2001) indicate that the variance within- and across-base populations is different for different traits. This implies that the within-population co(variance) structure (between different traits) varies across base populations, which is a common characteristic in dairy cattle. Therefore, to account for these differences or gaps in the pedigree in the evaluation of dairy cattle, genetic groups are usually included in the model. On the other hand, different base populations with different co(variance) structures for between-flock versus within-flock variation reinstates the use of genetic groups in across-flock evaluation of Australian Merino sheep.

2.4.2 Effect of genetic grouping on the accuracy of genetic evaluation

Genetic evaluation using models without and with genetic groups will affect the accuracy of EBVs depending on factors like the data structure, the proportion of linkage between flocks, the comparison of animals within or across flocks, the magnitude of genetic differences between flocks, the numbers of animals in each flock (flock size), and treating genetic groups as fixed or random. When animals are evaluated across breeds, genetic groups could be assigned based on breed of origin. In a study involving field data on purebred and crossbred pigs, Wolf *et al.* (2002) evaluated the effect of genetic grouping and crossbred information on the prediction of breeding values for commonly used sire breeds in the Czech Republic. The purebred animals were from six different breeds and the crossbreds were any combination between the six breeds. Genetic groups were defined on the basis of year of birth, origin and breed or breed group of the animal, and the rules of Quaas (1988) and Westell *et al.* (1988) were followed to assign phantom parent groups to account for the missing pedigree information. Four different animal models depending on the inclusion of crossbred information and genetic grouping were used to evaluate local and imported sires. The authors found that the ranking of boars to a large extent was more significantly affected by including genetic groups in the model than by accounting for the crossbred information. It was observed that the inclusion of crossbred information had either little

or no effect on the variability of EBVs. In the study of Wolf *et al.* (2002), the EBVs of imported boars had higher accuracy when genetic groups were included in the model. In addition, including genetic groups in the model led to an increase in the variation of EBVs, which was manifested in their higher standard deviations and larger ranges. The authors estimated correlations between breeding values for all four traits estimated using different models and the total breeding value with respect to all the traits combined. Although different models explained the variation in each trait, they recommended a genetic groups model when total breeding value was considered.

Theron *et al.* (2002) evaluated the effect of phantom parent grouping on the estimation of genetic trend in South African Holstein population. The authors stated that because of regular importation of animals from foreign countries into the native population, the implicit assumption of an animal model that all base animals have the same genetic merit cannot be true. The effect of this deviation was examined using the first lactation records on Holstein cows. The base animals were grouped on the basis of birth year, country of birth and selection intensity of the phantom parents. The authors found that the inclusion of phantom parent groups in the analysis reduced the bias in the estimation of the genetic trend. In addition, it was further concluded that the exclusion of phantom parent groups in the model affected ranking of imported sires much more than the local sires. Nevertheless, the authors also stated although accounting for phantom parent groups in the model affected estimation of genetic trend to a significant extent, it did not influence ranking of animals considerably.

Different populations of animals in different geographical regions will have variable allele frequencies of a particular gene at a specific locus, which is attributed to either natural or artificial selection or both in previous generations. When animals across regions are evaluated, it is imperative to account for this variation in allele frequencies depending on certain criteria like place of origin, year of birth and possibly path of selection if known. In a study involving field data that included records on imported animals with incomplete pedigree information, Culbertson *et al.* (1998) examined the influence of genetic groups (based on birth year) on genetic parameters and inbreeding depression in pigs. They found that the accurate estimation of inbreeding depression was not possible due to confounding with the missing parent groups. The

authors concluded that the genetic evaluation models must account for unknown parent groups to overcome the problem of confounding which decreases the accuracy of prediction of breeding values.

In a simulation study involving selection, missing records and non-overlapping multiple generations, Fikse (2003) compared two different approaches of assigning phantom parents to genetic groups: distinct and fuzzy classification. In the first approach, two generations were combined into one group. In the latter approach, each phantom parent was assigned equal proportions of two groups relating to the birth generation and the generation after birth. The author found that the representation of animals (with and without associated phantom parents) in the lists of genetically superior animals was more accurate under the latter approach (fuzzy classification). The author also warned against formation of genetic groups with small proportions of phantom parents to avoid making the MME less diagonally dominant (due to very small values being added to the diagonal elements). Similar conclusions were made by Quaas (1988).

Phocas and Laloe (2004) stated that accounting for genetic groups in a model should be considered when the number of animals per group is sufficiently large, the level of connectedness between groups and accuracy of selection (heritability and available data and pedigree information) are high, and significant genetic differences exist between groups. In general or in terms of connectedness and contrasts across groups, similar findings were reported by Foulley *et al.* (1990b). Foulley *et al.* (1983) proposed that 10 to 40% of progeny should be generated from link sires for making accurate across-genetic group evaluations. They stated that the accuracy of estimation of genetic group differences is dependent on the number of link sire progeny. Tong *et al.* (1980) used simulation to examine the joint effect of semen exchange (linkage) and the stud of origin of sires on estimating genetic differences between semi-isolated populations of two regions. The model accounted for the fixed effects of herd and region of stud (source of origin) and random effects of sires within region of stud. The authors concluded that increasing the level of semen exchange between populations estimated the mean genetic difference (between the two populations) with greater accuracy by reducing the variance of the estimated mean genetic difference. They also

stated that reciprocal exchange of genes across the two populations was more desirable than the one-directional exchange. Although the researchers tested varying levels of linkage (5, 10, 25, and 50%), they could not determine the optimum proportion of linkage that would be required between the two populations to facilitate sire evaluation across regions with sufficient accuracy.

Kennedy (1981) found that a model without genetic groups was more accurate than a model with genetic groups when the real differences between groups are not significantly large. Substantial amount of between-strain genetic variation in the Australian Merino sheep for traits of interest is well documented (Ashtiani and James, 1993; Atkins *et al.*, 1999b) and can be exploited for making faster genetic progress. To account for these genetic differences between strains in across-flock EBVs, genetic groups must be included in the model (Atkins *et al.*, 1999b).

Kennedy and Moxley (1975) compared the method of accounting for genetic groups and that of including the genetic relationships among animals for mixed model methodology used for sire evaluation. Under the former method, sires were grouped according to the time of entry into service. Therefore, fixed genetic groups were considered. For the latter method, pedigrees of all sires were traced back as far as the grandparents from both the maternal and paternal sides, and numerator relationship matrix was obtained. The authors' general conclusion was that the relationship method is advantageous over the grouping method when complete pedigree information is available. The authors also stated that the need for genetic grouping is desirable only when sufficient genetic links exist across different groups and pedigree information is missing and vice versa. However, in the absence of increased genetic ties between sires in a large population, the authors proposed a method which would be a combination of both relationship and genetic grouping methods with a differential grouping criterion for related (genetic grouping based on family) and unrelated sires (genetic grouping based on year of birth and stud of origin).

The inclusion of all known relationships amongst animals through numerator relationship matrix can account for selection over generations, thereby reducing the need for genetic grouping (Sorensen and Kennedy, 1984; Van der Werf and de Boer,

1990). Pollak and Quaas (1983) stated that accounting for more information in the numerator relationship matrix reduces the differences between genetic groups, thereby minimizing the role of genetic groups. However, Kennedy (1981) and Robinson (1986) have shown that including known genetic relationships among sires in the evaluation procedure does not completely account for the differences in the mean and genetic variance of base animals. This implies that a fraction of the bias still remains especially when the differences between groups are large and the pedigree of imported animals is either missing or incomplete. Kennedy (1981) and Robinson (1986) concluded that besides the need for genetic grouping when pedigree information is incomplete, the optimal use of link sires across genetic groups is required. However, the use of link sires across genetic groups merits further examination of a possible sire x environment interaction (Cottle and James, 1997; James and Cottle, 1997) being a significant source of variation.

Atkins *et al.* (1998) analyzed field data of fleece weight and fibre diameter using a model with genetic group effects and found that including genetic groups improved the models by reducing the contribution of sire x management group interaction. The authors used sire x management group interactions as a measure of the utility of genetic groups to account for flock differences and called this interaction variance “apparent”. They concluded that genetic grouping must be accounted for in across-flock evaluation schemes for the key Merino traits of fleece weight and fibre diameter to improve accuracy. Similar findings were reported by Atkins *et al.* (1999b) who stated that “the main aim of grouping is to reduce the between-sire (within-group) variance to a value that is in line with the assumed within-flock sire variance”.

In a simulation study, Ashtiani and James (1993) evaluated the design of progeny tests with the objective of identifying the best sires with reference to Australian Merino population with a large number of genetically different strains. Different strains were treated as different genetic groups and were fitted as random effects. They stated that the accuracy of across-flock EBVs would be compromised if genetic groups are ignored in the model. This was attributed to the fact that not accounting for genetic grouping would ignore the relationships between animals derived from the same strain. The authors also added that this type of problem associated with grouping is faced by

the Merino industry in Australia unlike the cattle populations where there are no distinct strains and pedigree information can be used to account for the genetic relationships among animals.

Famula (1985) investigated the equivalence between a restricted selection model and a genetic groups model, and how genetic grouping accounts for selection not accounted by the numerator relationship matrix. In the words of Famula (1985), “the goal of restricted selection is to maximize response to selection in traits of interest while restricting response in other correlated traits to zero”. The vector of unknown genetic group effects, $\mathbf{g}$ usually incorporated in the MME with genetic grouping as proposed by Quaas and Pollak (1981) and Westell *et al.* (1988) and reviewed previously in this chapter, is treated as a fixed effect and can be expressed (Pollak and Quaas, 1983) as:

$$\mathbf{g} = (\mathbf{Q}'\mathbf{G}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{G}^{-1}\hat{\mathbf{e}}$$

where $\mathbf{Q}$ is a known design matrix describing the proportion of genes derived by real animals from different genetic groups, $\mathbf{G}$ is the genetic co(variance) matrix, and $\hat{\mathbf{e}}$ is a random vector of individuals' deviations from their respective group means. According to Famula (1985), this definition of $\mathbf{g}$ can be explained based on $\mathbf{Q}$ and $\mathbf{G}$, as functions of selection differentials, which in turn, are functions of random variables. Pollak and Quaas (1983) argued that $\mathbf{Q}'\mathbf{G}^{-1}$ is a function of random selection differentials which accounts for the real group effects. $\mathbf{Q}'\mathbf{G}^{-1}\hat{\mathbf{e}}$ is stated to be a random conditional variable on which selection is assumed to be based (Famula *et al.*, 1983; Famula, 1985; Alfonso and Estany, 1999) and when selection occurs, $\mathbf{Q}'\mathbf{G}^{-1}\hat{\mathbf{e}}$ has a different mean and variance from those the same variable would have in the absence of selection (Famula *et al.*, 1983).

Famula (1985) emphasized that the predicted breeding value ($\hat{\mathbf{u}}$) of an animal is a function of the estimated genetic group effects ($\mathbf{Q}\hat{\mathbf{g}}$) plus the individual deviation from the group mean ($\hat{\mathbf{s}}$) expressed as $\hat{\mathbf{u}} = \mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{s}}$. The author added that the covariance between estimated and true genetic group effects is non-zero. This argument can be considered to be true only when genetic group effects are treated as random because assuming $\mathbf{g}$ as a fixed effect nullifies the fact that covariances exist between different groups. Another conclusion made by the author is that genetic grouping accounts for

prior selection by removing the covariance between predictors of breeding values and the changes in the population mean resulting from selection. Similar findings were reported by Pollak and Quaas (1983). Therefore, genetic groups must be formed depending on the type of selection practiced (Famula, 1985). However, selection history is not always known as in Merino populations characterized by significant genetic differences between strains. These results lead to the conclusion that selection decisions made in the present are independent of selection decisions made historically (Famula, 1985; Alfonso and Estany, 1999). Nevertheless, the above conclusion will not be applicable in case descendants of base animals are also selected (Alfonso and Estany, 1999), a condition which cannot be ruled out for most livestock industry datasets.

The review in this section shows that including all known relationships amongst animals reduces the need for grouping animals. However, when information on pedigree or performance data are missing and/or when there are likely differences between multiple base populations, grouping of animals is required. Depending on the criteria on the basis of which groups are formed, different strategies of genetic grouping could be developed and implemented in evaluation schemes in Merino industry. If the differences between flocks with respect to their phenotypic means are likely to be significant and sufficient linkage exist between flocks, then genetic groups must be fitted on the basis of flock of origin. However, when the differences between flocks are not significantly large, genetic groups need not be included in the model as shown in previous studies by Kennedy and Moxley (1975) and Phocas and Laloe (2004).

The approach of Kennedy (1981) could be used to determine whether genetic groups need to be fitted or not in the model. The author derived the expectation of the bias in genetic evaluation when genetic groups are not fitted in the model given priori genetic differences exist between base populations. The author suggested to obtain the mean square error (MSE) of prediction for the model without genetic groups as the sum of the squared bias in the predicted difference between genetic groups and the PEV of EBVs of animals in different flocks. The author stated that the MSE from the two models (without and with genetic groups) could then be compared to verify the model of choice. The author further stated that if only two genetic groups are considered, the non-groups model would have lower MSE than the groups model when the true genetic

difference between the two genetic groups is less than or equal to the associated SED from a genetic groups model. However, as pointed by the author, calculation of SED is less tedious than the MSE. Therefore, the author also concluded that the point when the MSE from the non-groups model is less than or equal to the PEV from a groups model is a critical one because at that point the true difference between the two genetic groups is less than or equal to the SED associated with the estimated difference. However, as further pointed by the author, a limitation of this technique is that it cannot be implemented for models involving more than two genetic groups. A similar method proposed by Famula *et al.* (1983) could be used for ascertaining the need for including or excluding genetic groups in the model.

Casellas *et al.* (2007) developed a procedure based on Bayes factor for a prior analysis to decide whether inclusion of genetic groups in the model is required and what structure of random genetic groups would account for the differences between base animals. However, as pointed by the authors, the main disadvantage of the Bayesian approach is its dependence on prior assumptions regarding distributions which may or may not necessarily be valid for a wide range of data on different traits recorded in the sheep industry. For example, for traits like worm egg count (WEC), the data generally has a skewed distribution. However, the distribution of data may be different depending on the level of resistance exhibited by animals and also on the species of internal parasites that are more prevalent in a given production environment. Additionally, the procedure developed by Casellas *et al.* (2007) cannot be implemented when genetic groups are treated as fixed effects.

The approach of Kennedy (1981) would be relatively easier and straightforward to apply on small datasets where only a few genetic groups are formed so that all possible pairwise comparisons between any two groups could be undertaken. However, in routine industry evaluations such as in the Australian Merinos where large numbers of genetic groups are defined, this approach would not be suitable for testing the need for grouping because differences in the MSE could be significant for some combinations and non-significant for other combinations of groups. Another important aspect that needs to be taken account of when deciding upon the choice of groups versus non-groups model is the size of flocks or genetic groups (number of animals). The

accuracy of prediction using a model with fixed genetic group effects will be strongly dependent on the size of flocks or genetic groups (Kuehn, 2005). The impact of the size of genetic groups on the accuracy of estimation when treated as random effects could be explored in further research. If there are numerous flocks with small numbers of animals, it is likely that treating such flocks as distinct genetic groups would not be suitable due to the inaccuracy of their respective group solutions. Further research could examine the possibility of combining smaller flocks into one large genetic group or cluster based on similarity in their genetic group means so as to overcome this type of problem. In addition, industry datasets in Australian Merinos are characterized by large amounts of missing dam pedigree and data information besides presumed differences in the means of base populations and subpopulations. Therefore, it would be beneficial to examine the impact of genetic grouping on the accuracy of across-flock EBVs when pedigree and data information is missing.

2.4.3 Inclusion of genetic groups in models for estimating QTL effects

A number of studies in the world are currently focusing on the identification of quantitative trait loci (QTL) and mapping of genes for traits of economic value in different species of farm animals. The information on EBVs and marker-QTL linkage can be combined to improve the efficiency of selection in livestock breeding programs (Machado *et al.*, 2003). The identification of QTL could provide vital information on genetic markers which could be included in the MME to implement marker-assisted selection (MAS) for the prediction of breeding values, where QTL effects are assumed to be random (Hanford *et al.*, 2004). However, with repeated MAS over a period of time, there is a likelihood of QTL becoming fixed in the later generations of a population. Additionally, the distribution of alleles with QTL effects is likely to vary within different subpopulations and specifically for some families (Hanford *et al.*, 2004). If MAS was carried across populations, there could be a QTL effect specific to each population. Therefore, it would be essential to link these different QTL effects to their respective base ancestors through genetic grouping to account for unknown selection.

Hanford *et al.* (2004) developed a technique for accounting for genetic grouping in the MME for QTL effects exploited in MAS. A polygenic animal model was used to develop the procedure for accounting for genetic group effects that were incorporated in the MME by adding a single equation corresponding to each group. The researchers stated that with slight modifications in the rules for obtaining the inverse of the NRM without group effects could be used to derive the inverse of the QTL coefficient matrix. The QTL coefficient matrix is obtained after absorbing the equations for the phantom animals in the MME. The authors further proposed that genetic group effects can be incorporated in the MME by augmenting the MME by the number of genetic groups used in models without marker information to obtain genetic evaluations for MAS.

Machado *et al.* (2003) investigated the effect of QTL for liveweight in two different genetic groups of Canchim breed of beef cattle. The genetic groups were based on the proportion of Charolais and Zebu blood. The authors found significant differences in allele distribution for the identified markers between animals from the two genetic groups which were attributed to varying factors like selection pressure, different allele frequencies in the base animals and random genetic drift. They stated that the QTL controlling a certain trait in a given population could have a totally different effect in another population due to epistatic interactions between the QTL and the genetic background. This indicates that the statistical methods used for identifying QTL must account for the differences between genetic groups. In addition, different genetic groups formed must be equally represented (balanced design) otherwise the power of QTL detection would be reduced (Machado *et al.*, 2003). Similar conclusions were made by Mackowski *et al.* (2005) in a study involving six genetic groups with varying proportions from Duroc and Polish Large White breeds of swine. The authors found significant variation in the frequency of a specific genotype in different genetic groups.

Pereira *et al.* (2005) examined the genotype (growth hormone and insulin-like growth factor polymorphisms) effects on EBVs for birth, weaning and yearling weights in two genetic groups of Canchim breed. The genetic groups were fitted as fixed effects in the model and the authors found significant ($P < 0.01$) effects of genetic groups on birth and yearling weights. They concluded that these differences in the genotype

effects across populations may indicate the effect of linked QTL. In other comparable studies focusing on the effect of different genotypes on production traits in the Canchim breed, similar results have been reported by Curi *et al.* (2005a, b). The results reviewed in this section show that inclusion of genetic groups is essential in the statistical analyses undertaken to detect QTL in animal populations having differences in selection history. The inclusion of genetic group or breed effects in the model will account for the genetic differences in the means of different populations that are attributed to QTL effects. Although no published literature focusing on the identification of QTL in Merino sheep could be found, future research in this area may provide better genetic options for manipulating traits of interest.

2.4.4 Effect of genetic grouping and deregression on international evaluation

Genetic groups are generally included in the model used for evaluation of animals across countries in order to account for differences in the means and variance of genetically diverse base populations. Genetic evaluation on an international level has become very popular especially in dairy cattle. There are several constraints to across-country evaluation like inadequate genetic links between countries (Schaeffer, 1985; Wilmink *et al.*, 1986; Rozzi *et al.*, 1990), variation between countries in data recording schemes (Wilmink *et al.*, 1986), and accounting for past selection within each country (Rozzi *et al.*, 1990). Several studies in the past have focused on developing means and methods for an accurate evaluation of sires from different countries. The following section briefly reviews some of the results.

Schaeffer (1985) proposed a linear model for international evaluation of dairy sires based on progeny tests from different countries. The following were stated to be the main characteristics of data from each participating country:

1. Within country EBV(s) for a trait(s)
2. The number of daughters on the basis of which sire EBVs are estimated.
3. The year and country of birth, breed category, and sire and maternal grandsire information for each bull to be evaluated. This information would be used as criteria for forming genetic groups.

The estimated genetic group effects representing differences across countries are added to sire solutions to obtain international sire EBVs. As stated by Schaeffer (1985), the problems associated with this approach are the lack of connectedness across countries and defining genetic groups based on country and year of birth. Another disadvantage associated with the approach proposed by Schaeffer (1985) is that matrix **Q** (proportion of genes contributed by different genetic groups to each sire) could get too complicated due to fractional coefficients if a large number of genetic groups are assigned. However, the **W**-matrix approach (Westell *et al.*, 1988) of genetic grouping could be employed to overcome this problem. In addition, Schaeffer (1985) also stated that including additive genetic relationships between bulls from different countries would improve the accuracy of international evaluation. Genotype x environment interaction was not accounted for in the model which in some cases could be a significant source of variation in the performance of link sires used in different countries depending on the climate and management conditions. This would be especially true when animals reared in tropical countries are jointly evaluated with those raised in temperate countries. Amores *et al.* (1999) reported significant genotype x environment interactions for lamb traits (birth weight, weaning weight and growth rate) in Australian Merinos. Smith and Banos (1991) stated that significant genotype x environment interactions may cause the genetic correlations between breeding objectives formulated in different countries to be less than unity in which case the advantages of combined selection across populations could be reduced. Another drawback of the model proposed by Schaeffer (1985) is that it is based on a number of assumptions like a genetic correlation of unity between traits (same traits recorded in different countries), same sire variance and units for measuring daughter deviations across countries. All of these assumptions may not necessarily be valid for industry datasets. Other disadvantages include inability to handle large datasets and increased computing cost (Rozzi *et al.*, 1990).

Multiple-trait Across Country Evaluations (MACE) was proposed by Schaeffer (1994) for international evaluation of dairy sires. National evaluation results from different countries are deregressed to derive dependent variables, which in turn are used in MACE to obtain across-country EBVs for dairy sires. In a Swedish study, the

properties of deregression for clinical mastitis (heritability of 0.02) and effects of using different definitions of genetic groups in within-country deregression versus the across-country analysis were examined by Mark *et al.* (2002). Mark *et al.* (2002) stated that the importance of pedigree information and that of genetic group effects relative to daughter information increases for traits having a low heritability with respect to national evaluation results. Similar findings were reported by Famula (1993), Sullivan (1995) and Cardoso and Tempelman (2003).

Data on Holstein bulls with a Swedish national evaluation result for clinical mastitis were included in the study of Mark *et al.* (2002) along with the national evaluation results on clinical mastitis from two other Nordic countries. In addition, milk somatic cell count results from nine non-Nordic countries were also included in an across-country analysis. A single-trait sire model was used to deregress national breeding values from each within-country analysis and the pedigree was traced back to two generations with genetic groups formed on the basis of birth year, country of first registration, and selection path. Each genetic group was required to have a minimum of 30 phantom parents. Deregressed national breeding values were used as dependent variables in three different evaluations designated by I_1 , I_2 and I_3 , respectively (Table 2.1).

Table 2.1. Comparison of national genetic evaluation results (X) with results from MACE applied to only mastitis in Sweden (I_1), with MACE results using genetic correlations set to zero (I_2) and with MACE results using estimated genetic correlations (I_3)*

	Mean ¹	SD	Difference (I-X)		SD	Correlation (X, I)
			Mean	Max absolute value		
I_1	10.00	1.00	0.000	0.000	0.000	1.000
I_2	10.00	1.00	-0.002	0.172	0.037	0.999
I_3	9.99	1.01	0.011	1.752	0.241	0.971

¹ Results were standardized corresponding to mean 10 and standard deviation 1 for X

* Adapted from Mark *et al.* (2002)

The MACE procedure was applied to only clinical mastitis in Sweden (single-trait) in the first analysis (I_1) with identical genetic groups for I_1 and deregression

implying that an equivalent set of MME were considered in both the cases. The objective of this analysis was to investigate whether the technique of deregression worked as per the theory and whether it was affected by the low heritability of a trait. In the second (I_2) and third analyses (I_3), MACE was applied to deregressed evaluation results from 12 countries for milk somatic cell count or clinical mastitis. In the second analysis (I_2), genetic correlations between countries were assumed to be zero. The objective of this analysis was to examine the effect of differences in phantom parent grouping for foreign bulls in an across-country evaluation when compared to deregression. The third analysis (I_3) involved MACE with respect to clinical mastitis and milk somatic cell count deregressed evaluation results from 12 different countries assuming different but correlated traits (genetic correlation = 0.37 to 0.97). The purpose of this analysis was to study the effect of correlated information regarding bulls evaluated in countries besides Sweden on clinical mastitis in Sweden. Genetic groups were assigned for the three analyses on the basis of similar criteria described above. Mark *et al.* (2002) found that the deregression procedure was exactly the reverse of MACE (Table 2.1). The authors concluded that the effects of different phantom parent grouping strategies in a within-country deregression in comparison with an across-country evaluation were negligible. They also stated that further studies are required to examine the effect of phantom parent grouping and properties of deregression for production traits, which have moderate heritabilities.

Rozzi *et al.* (1990) proposed a model which accounted for the fixed genetic group effects of sires based on country of origin and year of birth. Daughter averages to be used as observations in the international model were estimated by deregressing sire solutions. The model assumed same sire variance, genetic correlation of unity between traits (same traits recorded in different countries) and same units of measurement for daughter deviations across countries. Miglior *et al.* (2003) evaluated the effect of different time-edits (data collected on bulls born in different years) on performance data on the estimates of genetic group solutions in MACE evaluations of dairy bulls. National Holstein bull EBVs for milk protein from ten dairy countries including Australia were analyzed. Different analyses were undertaken based on different time-edits for proofs included in MACE and sires used for the estimation of sire variance.

These authors reported that the sire variance estimation using different time-edits on data had a small effect on the solutions of estimated genetic group effects and the ranking of bulls. They also stated that the differences in estimated genetic group effects from varying time-edits on data were more evident when plotted versus time. This can be attributed to differences in the genetic trends over different time periods.

An international genetic evaluation system in sheep could be established based on the same principles as MACE for dairy sire evaluation. This would allow for selection across countries leading to faster genetic progress and for utilizing information across countries (or flocks or breeds). The principles of MACE technique would be more beneficial when relationships between genetic groups are sparse. However, establishing connectedness across countries would be a critical requirement for international evaluation of sires in the sheep industry. Additionally, various types of effects such as genotype x environment interaction and heterosis would need to be included in the international model. National breeding values of participating nations could be deregressed to derive dependent variables that could be included in the international model for across-country evaluations of sires for meat and wool sheep industries.

2.4.5 Impact of genetic grouping on the accuracy of across-flock Merino sire evaluation in Australia

The Merino sheep industry is a major livestock industry in Australia that contributes significantly to the national economy. The breed has been developed for wool and meat purposes. There are different strains and bloodlines within the Merino breed that can be broadly classified into three major categories, namely broad wool, medium wool and fine wool. The Merino sheep are also used extensively for crossbreeding to generate first-cross dams which are later mated with a terminal sire breed to produce prime lambs. Central Test Sire Evaluation (CTSE) was developed in 1987 to improve the accuracy of across-flock evaluation of sires throughout Australia. The evaluation results provide key information to ram breeders to compare their home-bred sires with elite sires from different parts of Australia (Cottle and Hickson, 1996; Cottle and James, 1997; Hickson, 1997). CTSE has been beneficial to the Australian

Merino industry by providing on-farm evaluation guidelines to breeders and genetic linkage necessary for across-flock evaluation (Atkins *et al.*, 1999a). The establishment of links between CTSE and on-farm has led to data recording from both sources for a large number of flocks which are geographically isolated (Russell and Cottle, 1992; Atkins *et al.*, 1998). Due to increased genetic links between flocks, breeders are selecting from a large number of potentially superior sires implying significant increase in selection differential, thus leading to rapid genetic gains in the wool industry (Hickson, 1997). Additionally, commercial breeders have benefited enormously by making rapid genetic gain from the use of proven high performing sires and by selling semen from sires which had performed well in CTSE (Casey and Atkins, 1999).

Cottle and James (1997) investigated the impact of linkage between CTSE and on-farm on the accuracy of CTSE schemes in Australia. The average PEVs of differences between EBVs of sires in different genetic groups were estimated using the approach of Kennedy and Trus (1993) as reviewed under 2.3.1.1. Genetic groups were formed using two different criteria, namely the site and year in which the sire was first progeny tested and secondly on the basis of bloodline (fine and medium wool). Three different analyses were conducted depending on the level of linkage and year of data recording, and two levels of heritability (0.3 and 0.5) were tested. The PEV was significantly lower when genetic groups were formed on the basis of bloodline in comparison to when grouping was based on site/year of first use of sire (Table 2.2).

Table 2.2. The average PEV of EBV differences between sires (genetic variance units) for the medium wool schemes (1989-2000) with and without elite link schemes in 1996 and 2000 for no genetic groups, sire x year groups and bloodline groups. The average PEV of bloodline mean differences and the variance proportion compared with no elite link schemes (in brackets) are also presented*

Heritability	<u>Link schemes</u>			
	-	1996+2000	1996	2000
No genetic groups				
0.3	0.1605	0.1583 (0.986)	0.1593 (0.993)	0.1596 (0.994)
0.5	0.1106	0.1086 (0.982)	0.1095 (0.991)	0.1097 (0.992)
Site/year groups				
0.3	0.2986	0.2835 (0.949)	0.2847 (0.954)	0.2964 (0.993)
0.5	0.2217	0.2132 (0.962)	0.2139 (0.965)	0.2205 (0.994)
Bloodline groups				
0.3	0.1655	0.1628 (0.983)	0.1641 (0.991)	0.1642 (0.992)
0.5	0.1133	0.1109 (0.979)	0.1121 (0.989)	0.1121 (0.990)
Bloodline means				
0.3	0.0251	0.0250 (0.994)	0.0250 (0.997)	0.0250 (0.996)
0.5	0.0380	0.0377 (0.994)	0.0379 (0.997)	0.0378 (0.996)

* Adapted from Cottle and James (1997)

In the study of Cottle and James (1997), the PEV with bloodline groups was the same as with no groups in the model (Table 2.2). They stated that this result (PEV with bloodline groups) was because of the following two reasons. Firstly, when grouping is based on bloodline implying smaller number of genetic groups (large group size), a larger proportion of comparisons between sires is within groups which leads to larger standard errors of genetic group solutions. However, since sires are compared within groups, the errors associated with estimated genetic group effects do not affect the PEV. Secondly, the accuracy of estimated genetic group effects is increased because of better genetic links between groups. Therefore, the authors concluded that where genetic groups could be reasonably defined and sufficient genetic links exist, it would be advantageous to group sires based on bloodline rather than site/year of first use. The average PEV was lower for a heritability of 0.5 than for a heritability of 0.3 whether grouping was based on site/year or bloodline or when groups were not fitted in the model (Table 2.2). These results indicate that genetic groups should be based on factors which can account for real genetic differences between groups of animals like bloodline

and flock of origin rather than factors like year of first use/site. Although defining groups based on year of first use will account for the genetic trend to a certain extent, forming groups based on bloodline will account for the actual genetic differences between flocks with links to diverse base populations.

Atkins *et al.* (1998, 1999b) investigated the effect of genetic grouping on improving the accuracy of across-flock Merino sire evaluation. In a preliminary study, Atkins *et al.* (1998) analyzed field data on fleece weight (FWT) and fibre diameter (FD) at different levels of linkage (due to sires only). The two traits were recorded at yearling (Y; 10-13 months), hogget (H; 14-19 months) and adult (A; over 19 months) ages to adjust for the impact of age on genetic parameters. They found varying effects of sire x management group interactions on both traits (Table 2.3). This was attributed to the wide variation across different flocks (belonging to different genetic groups) indicating that the performance of sires was dependent on the mean genetic merit of the flocks in which they were used (Atkins *et al.*, 1999a). Subsequently, genetic groups were defined on the basis of stud of origin of sire and bloodline (fine-medium and medium wool), and fitted as fixed effects in the model. The authors found that including genetic groups reduced the contribution of sire x management group interactions to a significant extent and thus, improved the goodness of fit (coefficient of determination or R^2) for the models tested. This may be attributed to the fact that accounting for genetic groups in the model has explained to some extent the differences between management groups in their means and variance. The inclusion of genetic groups for fibre diameter led to a reduction in the sire x management group component with little impact on the sire component. For fleece weight, however, both the sire and sire x management group components were reduced when genetic groups were accounted for in the model. These results imply that genetic grouping has varying effect (difference in variation explained) on different traits depending on the genetic variance. However, the impact of including genetic groups in the model on the accuracy of across-flock evaluation cannot be discussed irrespective of the proportion of linkage between flocks. Atkins *et al.* (1998) concluded that accounting for genetic groups based on stud of origin of sires would improve the accuracy of across-flock EBVs by reducing sire x management group interactions.

Table 2.3. Estimates for sire and sire x management group (sire.mg) components of variance (estimated as a fraction of residual variance) for models without and with genetic group effects*

Trait ¹	<u>Without genetic groups</u>		<u>With genetic groups</u>	
	Sire	Sire.mg	Sire	Sire.mg
YFWT	0.068	0.075	0.051	0.059
HFWT	0.114	0.017	0.099	0.011
AFWT	0.105	0.082	0.086	0.049
YFD	0.134	0.032	0.141	0.019
HFD	0.221	0.168	0.207	0.003
AFD	0.129	0.051	0.127	0.036

* Adapted from Atkins *et al.* (1998); ¹ YFWT-Yearling fleece weight; HFWT-Hogget fleece weight; AFWT-Adult fleece weight; YFD-Yearling fibre diameter; HFD-Hogget fibre diameter; AFD-Adult fibre diameter

Atkins *et al.* (1999b) investigated further the influence of genetic differences among sources of origin on Merino sire evaluation by estimating sire x dam flock interactions and including varying numbers of genetic groups (based on bloodline) that accounted for the differences in the source of rams (Table 2.4). The traits analyzed were greasy fleece weight (GFWT), yield (Y), clean fleece weight (CFWT), fibre diameter (FD) and off-shears body weight (OSBWT). Four different types of models were tested namely Full Model, Model A, Model B, and Model C fitting 8, 3, 2, and 1 genetic groups, respectively. Model C with one genetic group implied no grouping as all sires were assumed to be from the same flock. The models were fitted without and with sire x dam-source interaction to correct for any significant non-additive effects due to sire usage across-bloodlines. When sire-source x dam-source was not fitted, the estimates for sire x dam-source variances were relatively small for all traits and across different models. However, when sire-source x dam-source was included in the model, the estimates for sire x dam-source variance were significantly reduced for all models and traits (Table 2.4). These results signify the relative importance of accounting for genetic groups based on the source of rams. These findings also indicate that when genetic differences exist between dams, genetic groups should also be defined on the basis of

source of dams. For a given trait, the error variances were the same across models (Table 2.4).

Table 2.4. Estimates of variances for sires, sire x dam source and error from models with varying numbers of genetic groups*

Trait¹	Full model (8 GG²)	Model A (3 GG)	Model B (2 GG)	Model C (1 GG)
GFWT – sires	0.0331	0.0378	0.0458	0.0839
Sire x dam-source	0.0032	0.0041	0.0044	0.0040
Error	0.4507	0.4500	0.4498	0.4507
Y – sires	1.54	1.58	1.70	2.28
Sire x dam-source	0.00	0.00	0.00	0.00
Error	21.18	21.19	21.19	21.20
CFWT – sires	0.0175	0.0195	0.0264	0.0569
Sire x dam-source	0.0015	0.0020	0.0018	0.0015
Error	0.2375	0.2370	0.2374	0.2379
FD – sires	0.2644	0.2887	0.3361	0.6094
Sire x dam-source	0.0036	0.0043	0.0026	0.0008
Error	1.4820	1.4810	1.4818	1.4819
OSBWT – sires	2.13	2.04	2.53	2.79
Sire x dam-source	1.09	1.08	1.09	1.09
Error	22.62	22.63	22.62	22.63

* Adapted from Atkins *et al.* (1999b); ¹ GFWT-Greasy fleece weight; Y-Yield; CFWT-Clean fleece weight; FD-Fibre diameter; OSBWT-Off shears body weight; ² Genetic groups

In the study of Atkins *et al.* (1999b), when the source of sire is not accounted for in the model (Model C), or not fully accounted for (Model B), the estimates for variance components were biased (Table 2.4). The apparent increase in estimated sire variance resulted because of the between-flock variance that gets added to the within-flock sire variance (Atkins *et al.*, 1999b) when grouping was incomplete. Therefore, the authors concluded that the accuracy of an across-flock evaluation would be reduced if genetic groups (to account for ram source) are not included in the model. To support their

argument, the authors stated that grouping would reduce the between-flock sire variance to a value that is in line with the assumed within-flock sire variance. It was also stated that this strategy could be used to test the effectiveness of the genetic groups so defined. The strategy proposed by Atkins *et al.* (1999b) could be advantageous in making a fairly accurate assumption regarding genetic group variance. However, they emphasized that grouping of highly selected sires from different locations would not be feasible without taking into consideration extra information. For example, if highly selected sires from outside sources are used for establishing genetic links across flocks, additional information on the selection history and possible genetic relationships between those sires needs to be considered to account for inbreeding.

Atkins *et al.* (1999b) also explored the effect of using the appropriate genetic parameters with an inappropriate (incomplete) statistical model. For this, BLUP solutions were obtained using the sire variances estimated from a full model. The incomplete models accounted for genetic group (based on sire source) and sire effects in addition to fixed flock effects estimated from a full model (with 8 genetic groups). The sire solutions obtained from incomplete models were later compared with those derived from full model analyses. The statistical measures used for making such comparisons were regressions and correlations. They found that the correlations between sire solutions derived from different models were all very high for different traits (Table 2.5) due to the fact that the same sire variance was assumed and the within-sire bloodline group solutions would be similar irrespective of whether all genetic groups are fitted (Full Model) or not (Models A, B, and C). However, the regressions of the biased estimates (from inappropriate model) on the true estimates (BLUP solutions from a full model) for all traits were significantly different when full or incomplete models were used (Table 2.5). These results indicate that when genetic grouping is not fully accounted for in the model, the estimates for variance components are likely to be biased. The magnitude of the bias in the estimated variance components would depend on the accuracy of grouping strategy adopted.

Table 2.5. Effect of using sire variances from the full model on BLUP solutions from the incomplete models. Results are expressed as standard deviations (SD) among sires solutions and the regressions (and correlations, r) of sire solutions from the incomplete models versus the full model*

Trait¹	Full model (8 GG²)	Model A (3 GG)	Model B (2 GG)	Model C (1 GG)
GFWT – SD	0.274	0.266	0.255	0.205
Regression (r)	-	0.96 (0.99)	0.92 (0.98)	0.72 (0.96)
Y – SD	1.33	1.30	1.24	1.09
Regression (r)	-	0.96 (0.98)	0.91 (0.98)	0.79 (0.96)
CFWT – SD	0.228	0.224	0.213	0.162
Regression (r)	-	0.98 (0.99)	0.92 (0.98)	0.68 (0.96)
FD – SD	0.75	0.72	0.71	0.65
Regression (r)	-	0.87 (0.91)	0.87 (0.92)	0.77 (0.90)
OSBWT – SD	1.48	1.46	1.37	1.30
Regression (r)	-	0.98 (0.99)	0.91 (0.98)	0.85 (0.97)

* Adapted from Atkins *et al.* (1999b); ¹ GFWT-Greasy fleece weight; Y-Yield; CFWT-Clean fleece weight; FD-Fibre diameter; OSBWT-Off shears body weight; ² Genetic groups

In another study, Casey and Atkins (1999) reviewed and expanded upon the findings reported by Atkins *et al.* (1998) from an analysis of Merino Benchmark data. Casey and Atkins (1999) found that when genetic groups were included in the model of analysis, the average progeny values for sires used in each stud had a range of 12% and 0.9 micron for CFWT and FD, respectively. However, when genetic groups were excluded from the analysis, the differences in average progeny values shrank to 4.5% and 0.45 micron for CFWT and FD, respectively. This indicates that in the absence of genetic groups, the between-group variance was not accounted for thereby affecting the across-group ranking of animals.

Huisman and Brown (2007) used Bayesian approach to examine the genetic group structure required for worm egg count (WEC) data on Australian Merino population. Different strategies of genetic grouping were compared using data extracted

from Sheep Genetics to investigate whether the genetic group variance was significantly different from zero. Animals were grouped based on average fibre diameter (μ), flock of origin (F), flock-period (FP), and flock-year (FY). Furthermore, two additional techniques of genetic grouping were also tested based on combined flock-period ($FP_{0.75}$) and flock-year ($FY_{0.75}$) with an assumed autocorrelation (0.75) structure across different groups within flock. They found similar estimates for additive genetic and residual variance across different strategies of genetic grouping (Table 2.6). Varying estimates of genetic group variance were observed for different strategies of genetic grouping. The ratio for Bayes factor (RBF) was maximum for FP and FY strategies indicating strong evidence against the null hypothesis of zero genetic group variance (Table 2.6). The RBF value was the least when grouping was based on average fibre diameter implying that this strategy is not the best option. This conclusion is supported by the least estimate of genetic group variance (Table 2.6). The authors concluded that accounting for autocorrelation between genetic groups within flock had a large impact on the estimation of genetic group variance. Although these findings indicate that different genetic groups could be defined within a flock, the adoption of this strategy in large scale Sheep Genetics evaluations may lead to inestimability problems or may lead to inaccurate group solutions for some flocks depending on their size (number of animals) and the level of linkage. Alternatively, grouping animals by flock appears to be the most appropriate option for Merino data as this would account for the genetic differences between flocks.

Table 2.6. Estimates for additive genetic (σ^2_a), genetic group (σ^2_g), and residual (σ^2_e) variance components, σ^2_g/σ^2_e (ρ^2_g) and the ratio for Bayes factor (RBF) for yearling WEC (standard errors in parentheses) using different grouping strategies *

Grouping method ¹	Number of GGs ²	σ^2_a	σ^2_g	σ^2_e	ρ^2_g	RBF
μ	6	1.91(0.15)	0.01(0.33)	5.90(0.13)	0.00(0.04)	0.105
F	110	1.82(0.15)	0.54(0.28)	5.98(0.13)	0.09(0.04)	2.088
FP	161	1.84(0.15)	0.34(0.18)	5.95(0.13)	0.05(0.03)	0.555
FY	311	1.82(0.15)	0.31(0.11)	5.95(0.13)	0.05(0.02)	3.440
FP _{0.75}	161	1.85(0.16)	1.44(0.47)	5.96(0.14)	0.20(0.05)	700.5
FY _{0.75}	311	1.80(0.15)	1.37(0.34)	5.99(0.13)	0.19(0.04)	150598

* Adapted from Huisman and Brown (2007); ¹ μ -Average fibre diameter; F-Flock of origin; FP-Flock-period; FY-Flock-year; FP_{0.75}-Flock-period with autocorrelation of 0.75; FY_{0.75}-Flock-year with autocorrelation of 0.75; ² Genetic groups

Huisman and Brown (2007) indicated that genetic groups may have an associated variance especially when the pedigree traces back to a large number of genetically diverse base populations and subpopulations. This argument appears to be applicable to Sheep Genetics evaluations where large numbers of genetic groups are defined. However, the effectiveness of FP and FY strategies in accurately estimating genetic group variance will depend on the number of observations (n) in each genetic group within flock. In case, small groups are formed within flock, the accuracy of genetic group solutions and estimated genetic group variance may get affected. Therefore, further studies should examine the impact of flock or genetic group size on the accuracy of estimated genetic group effects and genetic group variance. Further research can also examine how to model the covariance structure between genetic groups of different traits for additive and maternal genetic effects in a multi-trait evaluation.

2.5 Alternative Bayesian approaches to genetic grouping for evaluation with uncertain parentage

As reviewed above, missing or unknown parents are a common feature in most livestock industry datasets. This is mainly due to the use of multiple-sire matings common in pastoral production systems which lead to uncertainty of parentage (Famula, 1992, 1993; Cardoso and Tempelman, 2003). A number of different approaches have been proposed to account for missing or unknown parents in genetic evaluation. The most common one is the strategy of “phantom parent grouping” suggested by Westell *et al.* (1988) and Quaas (1988) with an implied assumption that all phantom animals are non-inbred and unrelated to other animals (real and phantom) in a given population. However, the above assumption can not be considered to be true in a majority of farm animal populations (Perez-Enciso and Fernando, 1992) where animals of interest are derived from parents belonging to different generations (overlapping generations). The possibility of real and phantom animals being genetically related cannot be ruled out completely especially when significant amount of pedigree information is missing as in the Merino sheep industry in Australia.

Several authors (Foulley *et al.*, 1987, 1990a; Henderson, 1988b; Dempfle, 1990; Famula, 1992, 1993; Perez-Enciso and Fernando, 1992; Sullivan, 1995; Cardoso and Tempelman, 2003) suggested alternative methods most of which are based on Bayesian statistical procedures for genetic evaluation of data with uncertain parentage. Under this approach, probabilities of parentage are assigned to a known number of potential candidate animals. Foulley *et al.* (1987, 1990a) proposed a method for univariate sire evaluation which requires specification of prior probabilities p_{ij} that a given progeny animal i with uncertain paternity has been derived from a sire j . They also stated that the Bayesian methods provide a better means of accounting for uncertain paternity than the conventional technique of genetic grouping. The authors further added that the Bayesian method can be easily extended to multivariate evaluation. However, the models proposed by Foulley *et al.* (1987, 1990a) do not account for the variances due to Mendelian sampling effects which are dependent on the probabilities assigned for

parentage, and this would affect the structure of the residual covariance matrix of the MME (Famula, 1992).

Perez-Enciso and Fernando (1992) compared the method of Westell *et al.* (1988) which incorporates genetic groups as extra fixed effects in the MME with the Bayesian approach that assigns probabilities of parentage to a finite number of possible individual animals. A population was simulated and both males and females were selected on the basis of two evaluation procedures, namely BLUP using genetic groups (BLUP-G) and BLUP using probabilities of identifying the missing parents (BLUP-A). Perez-Enciso and Fernando (1992) found that greater response to selection was obtained using BLUP-A than using BLUP-G. Similar findings were reported by Henderson (1988a). However, Perez-Enciso and Fernando (1992) pointed out that this advantage of BLUP-A over BLUP-G would decrease with an increase in the percentage of animals with unknown parents and the number of potential parents. This implies that BLUP-A may not be an accurate evaluation procedure and therefore, may not provide accurate predicted response to selection if there are an increasing number of animals in the pedigree with missing parentage. Nevertheless, still BLUP-A seems to be a better procedure in comparison to BLUP-G as the latter treats genetic group effects as fixed, therefore, it is very likely that the correct proportion of selection differential from the phantom parents (replacing missing parents) would not be transmitted to the real animals (Tier, personal communication).

Perez-Enciso and Fernando (1992) stated that the response to selection was biased using BLUP-G whereas that obtained using BLUP-A was not. Another conclusion made by the authors was that the numerator relationship matrix (NRM) proposed by Henderson (1988a) for accounting for uncertain parentage is the correct covariance matrix of additive genetic effects. Perez-Enciso and Fernando (1992) further added that using the NRM in the MME, unbiased EBVs could be obtained through BLUP-A. However, in the case of a large number of potential individual animals to be assigned probability of being true parents, the EBVs are likely to be biased. Additionally, the estimation of genetic trend using BLUP-G will lead to biased estimates over a period of time since inbreeding is not accounted for while computing the NRM. Furthermore, if selection has been practiced for a number of generations, then

animals with missing parents are basically excluded from the breeding program (Dempfle, 1990), thus leading to a reduction in the selection intensity and thereby leading to reduced selection response. Therefore, it is necessary to account for the missing parentage in the NRM. Dempfle (1990) proposed modifications in the NRM to account for some types of unknown parentage.

In a simulation study, Sullivan (1995) compared the method of accounting for uncertain paternity using genetic groups in the MME with other alternative methods based on assigning probabilities. The author found that the differences in predicted selection response for different methods were largest when the level of heritability was low and the degree of paternal uncertainty was high. For all the methods evaluated, selection response was higher in the probability methods than those with the genetic groups. However, all the methods (based on genetic groups and assigning probabilities) provided unbiased estimates of genetic trend at varying levels of heritability. Nevertheless, estimating genetic trend with accuracy will depend on the type of data structure (for example, genetic links across years) and the criteria used for genetic grouping over a period of time. Since there is a constant addition of new data with the passage of time, estimation of genetic trend with reasonable accuracy could be a tedious task. Quaas *et al.* (1998) presented a Bayesian procedure to overcome this problem by manipulating the coefficient matrix of the MME. Famula (1993) evaluated the contribution of animals with uncertain paternity to the accuracy of sire evaluation and found that the effect of including performance data on animals with missing sire pedigree is greatest for sires with few certainly identified progeny. The author stated that this results in young sires with less progeny getting greater benefit than proven sires.

In studies simulating multiple-sire matings, Cardoso and Tempelman (2003, 2004) extended the Bayesian method of Foulley *et al.* (1987) by combining performance data and prior information to estimate posterior probabilities of sire assignments. Cardoso and Tempelman (2003, 2004) compared a hierarchical Bayes (HIER) model to an animal model based on the use of average numerator relationship matrix (ANRM) as proposed by Henderson (1988b) for evaluating data with uncertain paternity. Under the HIER model, prior information was used to determine posterior

probabilities of sire assignments to animals with missing sire pedigree. For the ANRM model, equal and fixed probabilities were assigned to sires for animals having unknown sire pedigree. Cardoso and Tempelman (2003, 2004) concluded that the HIER model provides accurate estimates of variance components when some animals in the dataset have missing sire pedigree. They found a better statistical fit to simulated data on animals with uncertain paternity for HIER than the ANRM model. However, in the absence of prior information (e.g., genetic markers) especially for large datasets, the authors stated that the ANRM may be preferable due to its reduced computational time.

Genetic markers and other physiological markers could be used to determine the posterior probabilities of assigning sires to animals with missing sire pedigree. According to Cardoso and Tempelman (2003, 2004), the HIER model is preferable when prior information on assigning sires to their progeny is available through genetic markers. Foulley *et al.* (1987) suggested that information from red blood cell types, major histocompatibility markers or accurate records on breeding period and gestation length can be incorporated in the evaluation procedure for identifying missing sires with increased precision. The recent advances in the field of molecular genetics (for example, DNA fingerprinting) will be immensely useful in overcoming the problem of missing pedigree information in livestock production systems characterized by multiple-sire matings. However, the cost associated with these information sources need to be considered prior to their use in breeding programs.

Different methods can be used to account for uncertain parentage in genetic evaluation systems. These can be broadly classified into two categories, namely by introducing phantom parent groups in the MME or through Bayesian statistical methods. However, none of these methods has universal acceptance. According to Famula (1992), “adapting to the problem of parental uncertainty with new statistical techniques may never be as satisfying as determining the true identity of the parents with other analytical procedures”. Additionally, most research in the past on genetic evaluations with uncertain pedigree information has been conducted employing simulation and analytical techniques. Therefore, further studies on the application of different methods on industry data are required.

2.6 Characteristics of industry data used for genetic evaluation of Merino sheep in Australia

2.6.1 OVIS software package

The routine genetic evaluation of sheep in Australia is undertaken using OVIS software package (Brown *et al.*, 2000) which generates multiple-trait BLUP EBVs using an animal model for up to 46 traits of economic value for both wool and meat sheep industries. OVIS adjusts input data for various kinds of systematic effects like year, flock, sex and management group. Different types of models can be fitted in OVIS depending on the data structure including a maternal effects (genetic and permanent environment) model and a genetic groups model. The latter accounts for the differences in the genetic merit of multiple base populations.

Field data in the Australian Merino sheep industry are collected across a wide range of geographical and climatic regions wherein flocks have been derived from different base animals (subpopulations) with varying genetic means. In case performance data and pedigree recording begins after differentiation among subpopulations, and the level of genetic ties between flocks is insufficient, these genetic differences will not be reflected in the across-flock EBVs (Kuehn, 2005). Additionally, previous studies by Atkins *et al.* (1998) and Brash (1999) have reported that missing dam pedigree is a common feature in the Australian Merino industry. Furthermore, approximately 80 percent of Merino stud breeders in Australia record sire pedigree only (Casey and Hygate, 1992). The survey by Casey and Hygate (1992) also revealed that only 6 percent of animals had both sire and dam pedigrees recorded. Dam identities are missing in varying amounts (0 to 100% missing dams in the pedigree) in most flocks, and in some cases, sire identities are missing as well (Atkins *et al.*, 1998).

Besides genetic differences and limited linkage between Merino flocks, the lack of dam pedigree information has been a major constraint to an accurate across-flock evaluation in the Australian wool industry. Clement *et al.* (2001) emphasized that accounting for complete pedigree information through the numerator relationship matrix allows for a more accurate dissociation of genetic and environmental effects. In certain

situations, the authors suggested that a bias of about 10% in genetic parameters can be ignored despite its influence on the accuracy of prediction of breeding values. The authors recommended application of DNA fingerprinting under such situations for determining the pedigree. Barnett *et al.* (1997) and Atkins *et al.* (1998) stated that the option of DNA fingerprinting could be explored to account for missing pedigree information in across-flock evaluations in the Australian Merino industry. Barnett *et al.* (1997) argued that the technology of DNA pedigreering will provide ram breeders with accurate pedigree information that could be used to estimate breeding values with greater accuracy in comparison to those estimated using phenotypic data only. Although DNA fingerprinting could serve as a reliable means for ascertaining the pedigree, its widespread use in the Australian Merino industry is less likely to be a practical option due to the cost associated with its implementation. Barnett *et al.* (1997) and Clement *et al.* (2001) argued that the economic aspect of DNA fingerprinting will need to be assessed prior to its use in genetic evaluation systems. Alternatively, the method proposed by Richards *et al.* (2006) based on electronic matching of lambs and dams could be adopted for determining the pedigree in sheep flocks. The accurate recording of pedigree information will improve the accuracy of evaluation schemes and simultaneously reduce the need for genetic grouping. For the overall improvement of genetic evaluation systems in Australian sheep industry, Sheep Genetics has been established.

2.6.2 Sheep Genetics

The Australian Merino sheep industry is one of the largest wool growing and exporting industries in the world and relies on a number of different genetic evaluation systems for maximizing production efficiency. Performance data on fleece traits are collected across different geographical and production environments, thereby making it difficult for stud breeders to compare animals from different flocks based on their genetic merit. To overcome this problem, Sheep Genetics database has been established which is a joint project of Australian Wool Innovation Limited (AWI) and Meat and Livestock Australia (MLA). The Sheep Genetics database has combined genetic information on trait measurements for Merino sheep industry into a single database. Commercial sheep breeders and farmers can access Sheep Genetics database to identify

animals with superior genetic merit across flocks from different parts of Australia. The Sheep Genetics database provides information on a range of traits of interest like growth, wool quantity and quality, carcass, reproduction, and growth (www.sheepgenetics.org.au).

The operation of Sheep Genetics database is based on the same principles as other livestock evaluation systems in Australia. The database is linked to sheep databases in New Zealand. As reviewed in this chapter, variation exists in the mean EBVs of flocks belonging to different genetic groups. The Sheep Genetics database will enable sheep breeders to use EBVs within each genetic group to assess their own breeding direction and combine different traits in a selection index depending on their economic weightings and market requirements. The database has been designed to operate in conjunction with an extension program which provides guidelines regarding the use of genetic information generated for sheep farmers.

The Sheep Genetics database allows for a comparison of the following three major categories of sheep:

1. The wool breeds (MERINOSELECT), mainly Merinos where animals are directly compared.
2. The LAMBPLAN maternal sire breeds which involves comparison of breeds like Border Leicester, East Friesian, Coopworth, Corriedale and some other composites for their maternal characteristics.
3. The LAMBPLAN terminal sire breeds which compares breeds like Poll Dorset, Suffolk, White Suffolk and Texel for use as terminal sires for the Australian prime lamb industry (www.sheepgenetics.org.au).

Depending on the level of linkage between flocks, the EBVs generated in OVIS (Brown *et al.*, 2000) can be classified into two categories. The first category comprises flocks that do not have sufficient genetic links and have trait performance described as flock breeding values (FBVs). The second category has flocks that are adequately linked and have trait performance described as Australian sheep breeding values (ASBVs) (Brown *et al.*, 2006). Flocks participating in Sheep Genetics database must have effective minimum linkage with other flocks within their reporting group. This

allows for an accurate comparison of animals across flocks. In Sheep Genetics evaluations, different types of models are fitted using OVIS (Brown *et al.*, 2000) for deriving FBVs and ASBVs. Since the Australian Merino data are characterized by a high proportion of missing dam pedigree and trace back to multiple base populations and subpopulations, genetic groups are included in the evaluation model. In MERINOSELECT evaluations, genetic groups are formed on the basis of flock and year of data recording. Since a large number of animals in the LAMBPLAN database are derived from different types of structured and unstructured crossbreeding systems, genetic grouping in LAMBPLAN evaluations is based on breed and year (Brown *et al.*, 2006). A large number of flocks participate in Sheep Genetics evaluations, therefore the estimated genetic group solutions for smaller flocks may have large standard errors especially when genetic groups are treated as fixed effects. To overcome this type of problem, smaller flocks could be clustered together into one large genetic group (or cluster) on the basis of their estimated genetic group effects.

2.7 Clustering methods and algorithms

Clustering can be used to group similar flocks with regard to their estimated genetic group solutions in across-flock evaluations. Clustering can be defined as the partitioning of a dataset into different subsets or clusters in such a manner that the data in each cluster exhibit a common characteristic based on proximity according to a predefined criterion. Cluster analysis is a data exploratory tool which is used for combining similar observations into a single and unique cluster. The data points that are not significantly different are allocated to the same cluster. Each cluster is characterized by a unique mean and variance depending on the number of observations. Clustering is a common technique used for statistical analysis of data originating in the fields of bioinformatics, computational biology and quantitative genetics applied to animal breeding. This helps in investigating the pattern and association between observations in a dataset. The technique has been used extensively in mining gene expression data to investigate underlying genetic and biological mechanisms (Hu and Yoo, 2004).

Previous research (Amar *et al.*, 1998; Weigel and Rekaya, 2000; Zwald *et al.*, 2003; Fouilloux *et al.*, 2008; Vasconcelos *et al.*, 2008) examined the application of

cluster analysis in genetic evaluation. The objective of using clustering techniques ranged from investigating differences among genetically diverse groups of wool sheep for crossbreeding to condensing large numbers of connected genetic groups into compact clusters to estimate connectedness among herds of beef cattle. There are various types of clustering mechanisms and algorithms which are reviewed in the following section.

1. **Hierarchical clustering:** Hierarchical clustering yields a dendrogram (tree of clusters) which represents the pattern exhibiting the relationship between clusters. There are two types of hierarchical clustering namely, agglomerative and divisive.

(a). **Agglomerative clustering:** Agglomerative clustering initially starts with each data point as a single and individual cluster and then subsequently forms bigger and bigger clusters by grouping similar data points together until the whole dataset is encapsulated into one final cluster. This type of clustering is commonly used in microarray analyses. However, this type of clustering method is not suitable for handling very large problems (Fouilloux *et al.*, 2008).

(b). **Divisive clustering:** Divisive clustering first considers the entire dataset to be a single cluster and then divides it into smaller and smaller subsets until each subset contains only a single data point (www.en.wikipedia.org/wiki/Data_cluster).

2. **Partitioning clustering:** Partitioning clustering breaks the data points into distinct and nonoverlapping groups or clusters. Partitioning clustering can be classified into three categories, namely K -means, Fuzzy c-means, and Quality threshold (QT) clustering. Under K -means clustering, the numbers of clusters required are pre-specified. The K -means algorithm allocates each data point to a cluster whose centre is the closest. The centre or centroid of a cluster is the average of all the points within a cluster, that is, the coordinates of a centroid are the arithmetic mean for each dimension separately over all the data points within that cluster (www.en.wikipedia.org/wiki/Data_cluster). The K -means algorithm steps (MacQueen, 1967) are summarized below:

- Specifying the number of clusters, k
- Random generation of k clusters and determining the cluster centroids

- Assigning each data point to the nearest cluster centroid
- Recomputation of new cluster centroids
- Repetition of the two previous steps until some convergence criterion (e.g., the assignment of data points to different clusters has not changed) is fulfilled

Under fuzzy c -means clustering, each data point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one unique cluster. Therefore, the data points on the edge of a cluster may be in the cluster to a lesser degree than the data points close to the centre of that cluster. This type of clustering allows a data point to be assigned to two or more clusters. The fuzzy c -means algorithm steps (Dunn, 1973) are as follows:

- Choosing the number of clusters, c
- Random assigning to each data point the coefficients for being in c clusters
- Repetition until the algorithm converges, that is, the coefficients' change between any two successive iterations is not greater than the given sensitivity threshold
- Computing the centroid for each cluster
- Estimation of coefficients of being in different clusters for each data point

The QT clustering is an alternative technique of partitioning the data which was invented for gene clustering. This type of clustering does not require specifying the number of clusters (www.en.wikipedia.org/wiki/Data_cluster). The steps of QT algorithm (Heyer *et al.*, 1999) are as below:

- Specifying a maximum diameter for clusters
- Forming a candidate cluster for each data point by including the closest data point, and continuing with this process until the diameter (maximum distance among the points of a cluster) of the cluster exceeds the specified threshold
- Saving the candidate cluster having the most data points as the initial true cluster, and subsequently removing all points in the cluster from further consideration
- Recursing with the reduced set of data points

The mechanism of clustering is based on similarity between data points which is determined by the distance measure specified.

Distance measure

A key step in any clustering mechanism is the selection of a measure of distance, which will determine how the similarity between any two data points is estimated. This will affect the shape of the clusters formed as certain data points may be closer to one another according to one measure of distance and further away according to another (www.en.wikipedia.org/wiki/Data_cluster). Euclidean distance, Manhattan distance, Mahalanobis distance, and Hamming distance are some of the commonly used measures of distance:

1. *Euclidean distance*: The straight line distance between any two points in a plane is known as euclidean distance. Clustering mechanism in the FASTCLUS procedure of SAS (1990) is based on euclidean distance between data points.
2. *Manhattan distance*: The Manhattan distance between two points in a plane is the sum of the (absolute) differences of their coordinates
3. *Mahalanobis distance*: The mahalanobis distance takes into consideration the covariance among the variables in calculating distances. The use of mahalanobis distance overcomes the problems of scale and correlations inherent in the euclidean distance. The mahalanobis distance is widely used in cluster analysis and other classification techniques. It also finds application in detecting outliers especially in the development of linear regression models.
4. *Hamming or edit distance*: The hamming distance measures the minimum number of substitutions required to change one data point to another.

There are different ways in which data could be clustered. Depending on the type of data, different measures of distance can be used for determining the similarity between data points. Each method of clustering has its advantages and drawbacks depending on its application. The main advantages of K -means clustering are that we do not have to calculate the distance measures between different pairs of data points and the speed which allows it to handle large datasets. However, K -clustering does not provide the same pattern of clusters with each run because the resulting clusters are

dependent on the initial random assignment of data points. This implies that the results obtained are dependent on the value of k . This type of effect can be reduced by multiple runs. Although K -clustering minimizes the intra-cluster (within a cluster) variance, it does not ensure that the resulting clusters have a global minimum of variance (across-cluster variance). In other words, the algorithm does not necessarily find the most optimal configuration corresponding to the global objective function minimum (www.home.dei.polimi.it/matteucc/Clustering). Similar disadvantages are also associated with fuzzy c -means clustering, the results of which are also dependent on the initial assignment of cluster seeds.

Divisive clustering is more computationally expensive when it comes to making decisions in dividing a cluster in two given all possible options. However, with this type of clustering, the overall structure of the data is retained which means that the “upper” levels of the dendrogram are highly representative of the original structure of the data. Hierarchical clustering (agglomerative and divisive) as a whole should not be used if there is a significant amount of noise expected to be present in the dataset. Both agglomerative and divisive clustering generate highly compact clusters and under either method, outliers are given more emphasis in the cluster forming decision (www.ucl.ac.uk/oncology/MicroCore). Different clustering algorithms may generate varying clustering results due to their bias and assumptions (Hu and Yoo, 2004) which in turn is likely to affect their interpretation. There is no particular clustering algorithm that performs best across a range of datasets (Hu and Yoo, 2004). Therefore, the choice of clustering method in genetic evaluation will depend mainly on the data structure.

2.7.1 Application of cluster analysis in genetic evaluation

Clustering has been used in previous research focusing on the significance of genetic groups in genetic evaluation of livestock. Amar *et al.* (1998) investigated genetic divergence for different traits between Nali and different grades of Nali sheep having exotic inheritance of Russian Merino and Corriedale. Different combinations of genetic groups were clustered and they concluded that cluster formation was necessary to identify groups that exhibit significant genetic diversity so that future parents could

be selected for crossbreeding programs. Weigel and Rekaya (2000) presented a multiple-trait herd cluster model for genetic evaluation of dairy sires. The model was later extended by Zwald *et al.* (2003). The procedure involved clustering herds from different countries according to the variables related to management, climatic and genetic factors. Therefore, each cluster represented a unique production environment. The variables were subsequently weighted by their relative significance in accounting for genotype by environment interactions between herds. A multi-trait sire model was used for analyzing the data and production in each cluster was considered to be a separate trait. The EBVs for sires were estimated in each cluster and ranking of sires across clusters was also investigated. The proposed model could predict the performance of dairy sires based on the production environment in which their progeny will perform, rather than the country where they will be located. However, as pointed by Weigel and Rekaya (2000), inadequate linkage between countries may lead to inaccurate estimates of covariances between same traits evaluated in different countries.

Fouilloux *et al.* (2008) developed a clustering algorithm (Caco or Criterion of Admission to the group of Connected herds) alternative to hierarchical agglomerative clustering method to identify well-connected herds or genetic groups in large datasets. The authors claimed that the Caco algorithm accounts for known pedigree information and data structure when measuring connectedness between herds. This method was stated to be not suitable for handling large-sized evaluations.

As part of Sheep Genetics research and development, Brown (Personal communication) investigated the use of cluster analysis to reduce the number of genetic groups that need to be fitted in a model. Data from Sheep Genetics database were used for comparing different numbers of clusters. Each flock was treated as an individual genetic group and the traits analyzed were body weight, greasy fleece weight, fibre diameter, and fibre diameter coefficient of variation at yearling, hogget and adult ages. Genetic group effects were fitted as random effects in OVIS (Brown *et al.*, 2000) and genetic group variance was assumed to be the same as the additive genetic variance. Flocks (or groups) were clustered on the basis of their estimated genetic group effects using the FASTCLUS procedure of SAS (1990). Three different types of estimated genetic group effects were used as input for the clustering processes, namely raw (direct

from OVIS), standardized (having a zero mean and a standard deviation equal to one), and corrected (ratio of genetic group solution for a given trait and flock to the genetic standard deviation for that trait).

The purpose of standardizing estimated genetic group solutions was to avoid a bias in the clustering mechanism which gets dominated by a trait with a large genetic variance. The distance between the means of any two flocks for a given trait was constrained to one which aids in manipulating the number of clusters formed. The way the estimated genetic group effects were expressed prior to clustering flocks significantly affected the number of clusters formed and hence the final EBVs. This shows that clustering had some measurable impact on estimability and accuracy of cluster solutions (estimated genetic group effects from clustered flocks). Clustering led to an increase in the variance between clusters (or genetic groups) when individual genetic groups of different clusters were treated as random effects. This is a characteristic of *K*-means clustering, which results in an increased variance between clusters formed, as reviewed above. Brown (Personal communication) estimated regression coefficients by regressing the adjusted progeny phenotypes on the sire EBVs to evaluate the accuracies of different numbers of clusters. However, the optimal configuration of clusters could not be identified because the underlying mechanism of clustering is dependent on the magnitude of genetic group effects.

In a multiple-trait across-flock evaluation, the underlying mechanism of conventional clustering (animal is allocated to the same cluster for all traits) is dominated by the trait with a large mean and phenotypic variance. This may result in a biased clustering of flocks for the less variable trait(s). Alternatively, if the differences in the phenotypic variance of traits are large, flocks can be clustered separately for each trait using differential clustering. The term “differential” means that each animal may be allocated to different clusters for different traits.

2.8 Conclusions

Missing pedigree and performance data information is a common feature in most livestock industry datasets used for routine genetic evaluations. Therefore, the underlying assumption of an animal model evaluation that all base animals are

unselected and have the same genetic merit cannot be true in most cases. A number of methods have been proposed by different researchers to account for missing parentage and previous selection on base ancestors to increase the accuracy of the evaluation procedure. These methods can be broadly classified into two categories. The first approach comprises assigning “phantom parents” and therefore introducing genetic groups in the mixed model equations. Genetic groups can be treated as fixed effects as each group is weighted by its genetic contribution to the real animals (animals with data). The second strategy involves identification of potential parents and subsequently assigning them a probability of being the true parent. Therefore, genetic groups are not introduced in the mixed model equations using this approach. The former approach has been more widely used than the latter.

Genetic groups can be assigned on the basis of a number of criteria like year of birth, country of origin, sex and selection path depending on the information available. However, no method of genetic grouping has universal acceptance. Genetic grouping can account for both direct and maternal genetic effects in sire and animal model evaluations. Computational requirements for a sire model with genetic groups are much lower than those for an animal model because of the difference in the number of equations solved. Furthermore, genetic groups can be included in the model either as fixed or random effects depending on factors like the group size, data structure, heritability, amount of missing pedigree information, magnitude of genetic differences between base populations, and the level of connectedness across genetic groups.

When groups have a large size (number of animals) they may be treated as fixed or random effects. The accuracy of estimation of genetic group effects when treated as fixed will be reduced for smaller groups. However, when the size of genetic groups is not large, it is preferable to assume genetic groups as random effects. Varying effects of genetic grouping on the estimates of across-flock variance components have been reported in the literature. The estimability of genetic group effects depends on the structure of the data, the level of genetic links between flocks and the criteria on the basis of which groups are formed. The simultaneous estimation of genetic group and other fixed effects (for example, flock) using mixed model methodology cannot be

accomplished in all the cases due to the potential problem of confounding of these effects.

Until recently, Australian sheep farmers and ram breeders used to rely on a number of databases for comparing animals on the basis of their EBVs. The establishment of Sheep Genetics database is providing single and unique Australian EBVs for both wool and carcass traits, and will aid sheep farmers to breed animals as per their breeding objectives and market requirements. The Sheep Genetics database contains performance records on a large number of animals reared in different parts of Australia with significant differences between flocks. Additionally, a large number of Merino flocks are not linked genetically, thereby leading to the problem of confounding of genetic group and fixed flock effects. Therefore, establishing adequate linkage between flocks will lead to an accurate estimation of genetic group effects and across-flock EBVs. The minimum level of linkage required to achieve this objective needs investigation.

Currently, Sheep Genetics evaluations utilize the same genetic grouping strategy based on flock of origin for all traits. The conventional strategy of grouping base animals may not be effective when grouping is accounted for in a multi-trait evaluation for Merino industry because datasets have varying amounts of missing data on different traits. Additionally, traits are characterized by differences in selection history. For example, greasy fleece weight is characterized by long selection history and hence generally well recorded. On the other hand, worm egg count (WEC) has a recent selection history and is not recorded by all breeders. Therefore, efficient strategies for genetic grouping need to be developed and incorporated in the models fitted in OVIS for improving the accuracy of across-flock EBVs for fleece traits in Merinos. Additionally, key Merino traits like fleece weight, fibre diameter and body weight have varying levels of heritability and flocks have different sizes (number of animals). If the trait is highly heritable, it is likely that the impact of genetic grouping on across-flock EBVs would be small and vice-versa. Furthermore, estimated genetic group solutions for smaller flocks are likely to be less accurate especially when group effects are treated as fixed. Therefore, certain factors like the size of genetic groups, the level of

heritability and the amount of missing data on each trait need to be considered prior to including genetic groups in the evaluation model.

Recent advances in the field of molecular genetics have led to the identification of different types of genetic markers associated with QTL influencing traits of interest. The information on genetic markers is used in conjunction with phenotypic data for the genetic improvement of livestock. Genetic groups can be assigned depending on the recombination rate and the level of fixation of the QTL in different populations. No published work on genetic grouping based on QTL in Merinos could be found in the literature. Future studies could examine whether key Merino traits like greasy fleece weight and fibre diameter are influenced by QTL and what genetic group structure will account for genetic differences (due to QTL effects) between base populations. Further research could also explore the possibility of defining genetic groups based on the molecular genetic information.

In the Australian Merino population, genetic differences and limited genetic links between flocks are the major constraints to an unbiased across-flock evaluation. The industry data on fleece traits in Merinos are collected across a range of climatic and production environments. Flocks vary widely in their phenotypic means and most likely, the mean genetic merit depending on the magnitude of environmental differences. Genetic linkage across flocks is a prerequisite in conjunction with genetic grouping for an accurate partitioning of flock genetic and fixed flock effects. Therefore, the amount of linkage and the type of data structure required to achieve this objective needs investigation. Additionally, within a breed, for example, in Merino sheep, there are different strains (Peppin, Saxon, South Australian and Spanish) and within strains are different bloodlines (fine wool, medium wool and coarse wool), which exhibit wide variation in traits both within- and across-flocks. In a majority of flocks, it is difficult to determine the exact selection path because of their diverse links to populations and subpopulations with differences in their phenotypic means. Thus, it will be imperative to include breed or strain as one of the criteria for defining genetic groups for deriving unbiased across-flock and across-breed EBVs.

The inclusion of genetic groups in the evaluation procedure will account for the differences in the genetic means of base populations and subpopulations and hence, reduce the bias in across-flock EBVs. However, prior to the actual application of this type of model for industry evaluations, certain issues need clarification. Further theoretical research should focus on the development of models and strategies for fitting strain or bloodline effects to account for different types of past selection in different strains of Australian Merinos. In addition, future research could focus on the development of novel techniques and algorithms that would facilitate combined analysis of quantitative and molecular data, the latter could provide useful information in ascertaining the missing pedigree. Previous studies have shown that accounting for genetic groups in the model improves the accuracy of Merino sire evaluation. However, past studies in this area did not evaluate the impact of flock and genetic group size on the accuracy of estimated genetic group effects and genetic group variance in across-flock evaluation. However, flocks vary considerably in their size (number of animals) and hence, for smaller flocks, genetic group effects may not be estimated with acceptable accuracy. Thus, further analytical and theoretical research should examine this issue in more detail.

Besides genetic grouping, linkage or connectedness is required for an accurate partitioning of genetic and environmental differences in across-flock evaluations. The magnitude and quality of genetic links across flocks can significantly impact the accuracy of across-flock evaluation. A number of criteria for measuring the degree of connectedness in industry datasets have been proposed by different authors. However, the actual choice of measure of connectedness depends on the objective of study, data structure, genetic group size (number of animals per group) and computational requirements. Connectedness is more of a statistical measure rather than that of genetic relationships between different genetic groups. There is no measure of connectedness with universal acceptance.

Cluster analysis has been used in previous research for genetic evaluation of livestock. Different methods and algorithms reported in the literature could be used for clustering of data. However, each algorithm is suited to work well for a specific type of dataset. When multiple generations are involved and genetic groups need to be assigned

for both additive and maternal genetic components using different criteria, the number of genetic groups could be reduced by grouping generations as was undertaken by Diaz *et al.* (2002). Clustering flocks on the basis of their estimated genetic group effects should be tested on simulated and industry data for its accuracy as a tool for reducing the number of genetic groups fitted in the model for across-flock evaluations in the Merino industry.

Chapter 3. The effect of linkage and genetic grouping on the partitioning of genetic and environmental differences between flocks

3.1 Introduction

Across-flock evaluations allow for selection of genetically superior animals from a large genetic base (Banks, 1997; Mathur *et al.*, 1998). The inclusion of genetic group effects in an across-flock evaluation model accounts for the bias due to varying genetic means of base populations (Quaas, 1988). The models with genetic groups are often used when the records on selected animals are either not available and (or) the pedigree information is missing (Kennedy and Moxley, 1975; Diaz *et al.*, 2002). Additionally, in the Australian Merino industry, flock environmental and genetic group effects are likely to be confounded. However, simultaneous estimation of genetic group and systematic environmental (flock) effects requires an adequate proportion of genetic linkage among flocks (Kennedy, 1981). Therefore, the proportion of linkage required for an accurate partitioning of environmental and genetic differences between flocks needs investigation. Furthermore, the estimability of genetic group solutions in an across-flock evaluation depends on the data structure and the criteria on the basis of which genetic groups are defined (Foulley *et al.*, 1990b; Laloe, 1993; Laloe and Phocas, 2003). With across-flock evaluations gaining popularity in the Australian Merino industry, there has been some exchange of genes (through rams) amongst different flocks. However, limited genetic links and lack of pedigree recording still remain a major constraint to across-flock evaluation in the Australian Merinos (Atkins *et al.*, 1998). Huisman *et al.* (2005, 2006) analyzed industry data from Sheep Genetics and reported that adequate genetic links now exist for most flocks and traits to evaluate animals across flocks. If genetic group effects are not estimated with precision in industry evaluations, the across-flock EBVs are likely to be inaccurate.

The objective of this study is to use simulation to explore the joint effect of genetic grouping and linkage on the accuracy of single-trait across-flock EBVs with special reference to the Australian Merino industry. The Sheep Genetics Merino database is characterized by a high level of incomplete dam pedigree with many breeders recording only the sire for a large proportion of progeny (Huisman and Brown, 2007). The effect of genetic grouping on the accuracy of across-flock EBVs when dam pedigree is missing will also be examined. The proportion of linkage required for estimating genetic group effects and across-flock EBVs at 5% level of significance ($P < 0.05$) will be explored. Since Australian Merino sheep have variable data especially regarding pedigree information, different simulated scenarios will be tested to investigate the proportion of linkage and the type of data structure required for an accurate disentangling of genetic and environmental differences within and between flocks.

3.2 Materials and Methods

In Scenario I, two flocks were simulated and each flock was treated as a separate genetic group. Under Scenario II, four flocks were simulated and each genetic group comprised two flocks. The assumptions used in simulating different scenarios and the values chosen for the magnitude of genetic and environmental differences between flocks, the number of sires used for generating progeny in each flock, and the number of progeny derived from each sire resembled those observed in field data (for example, Brown *et al.*, 2001b; 2007) evaluated by Sheep Genetics.

Data simulation

Scenario I

Pedigree and performance data on a single trait were simulated for two flocks of 1000 progeny each involving only one generation. Each flock was simulated to have a different genetic mean so as to create two base populations. The progeny in both flocks descended from sires of three different origins (Figure 3.1). The pedigree of all sires was assumed to be unknown. The following notation and definitions will be used for describing the sires used:

1. Xsires: The external sires used in both flocks but derived from a third flock, and do not have any performance records. Therefore, a mean breeding value of zero was assumed.

2. WFsires: internal sires bred in one flock and used as sires in both flocks.

3. WNsires: Within-flock sires of the two flocks and used only in their own flocks.

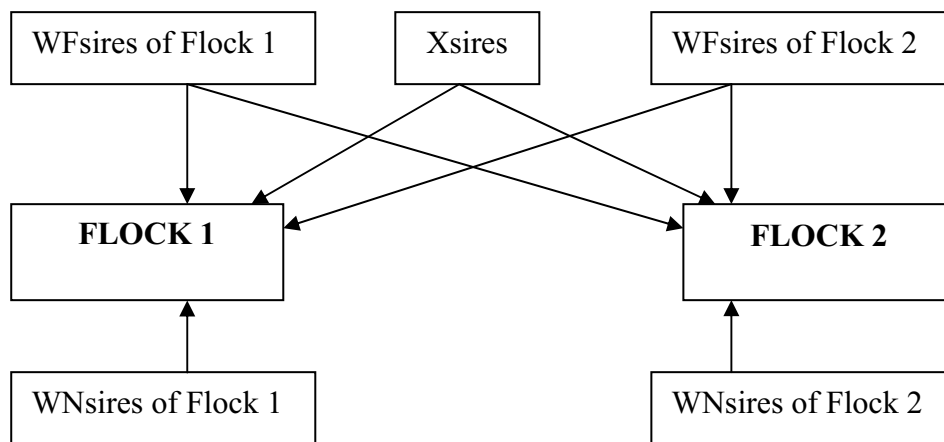


Figure 3.1. Diagrammatic illustration of the use of sires with different origin

Simulation of flocks

The purpose of this simulation was to create a simple data structure for examining the variation in the genetic group effects of the two flocks at varying levels of linkage and using models without and with genetic groups. Each flock was assumed to be a separate genetic group and a third genetic group was made up of Xsires. Thus, there were three genetic groups with different genetic means. 50 sires from all three genetic groups were randomly mated to unknown dams to generate 20 progeny each in each flock. All the dams were assumed to be indigenous, that is, they originated in the same genetic group in which progeny were generated. The simulated TBVs (True breeding value) of the dams included the genetic group effect plus the additive genetic effect. The TBVs of progeny in each flock were simulated as:

$$TBV_i = \frac{1}{2} (TBV_{sire} + TBV_{dam}) + \Phi$$

where TBV_i is the true breeding value of the i th animal, TBV_{sire} and TBV_{dam} are the TBVs of the sire and the dam, respectively of the i th animal, and Φ is a factor for the Mendelian sampling effect which was simulated with mean zero and half of the additive genetic variance. The TBVs of sires (Xsires, WFsires and WNsires) were simulated using a normal distribution with zero mean and additive genetic variance (σ_a^2) of 1.5 micron², and included the effects of their respective genetic groups. The genetic group effect for Xsires was simulated to be zero. For each progeny, a phenotypic observation (P_i) for fibre diameter (FD) was simulated around a global mean (μ) of 19 micron with a within-flock heritability (h^2) of 0.43 as:

$$P_i = \mu + TBV_i + F_j + e_{ijk}$$

where F_j and e_{ijk} are the environmental flock and random residual effects, respectively associated with each phenotypic record. The h^2 of 0.43 was chosen based on the parameter used for FD in Sheep Genetics evaluations. The environmental flock effects for flocks 1 and 2 were set to be -1 and +1 micron, respectively. The residual effects were sampled from a normal distribution with a mean of zero and variance (σ_e^2) of 2 micron².

A genetic difference of 2 micron (genetic effects of -1 and 1 micron for flocks 1 and 2, respectively) was simulated in the mean FD of the two genetic groups. The proportion of linkage due to either Xsires or WFsires between the two flocks was varied over 6 levels from 0, 10, 20, 30, 40, to 50% with the other sire group at zero, respectively. Both forms of linkage were tested independent of one another. Unless specifically stated, all results shown are averages from 50 replicates.

Scenario II

The aim of this scenario was to create a data structure for examining the variation in the genetic group and fixed effects of the four flocks at different levels of h^2 , linkage, number of progeny per link sire, and using models without and with genetic group effects.

Simulation of flocks

Pedigree and phenotypic data for FD were simulated for four flocks with 1000 animals each from a single generation. The h^2 levels of 0.1, 0.3 and 0.43 were tested. The h^2 of 0.1 and 0.3 were tested to cover the range for most production traits of economic significance. The phenotypic variance (σ_p^2) was constant (3.5 micron²) across all levels of h^2 . The types of sires and dams used in this scenario were the same as described in Scenario I. Each sire was mated with unknown dams to generate 10 progeny. Additional testing was performed (using h^2 of 0.43) by changing the number of progeny per sire to 5 and 2. The simulation of performance data and TBVs for each animal and dams were similar to Scenario I. Two flocks were assigned to a genetic group, that is, flocks 1 and 2 in genetic group 1 and flocks 3 and 4 in genetic group 2. A third genetic group (assumed to be of mean breeding value of zero) was the source of Xsires (external sires) which were used for creating genetic links between the two genetic groups.

Flocks 1, 2, 3, and 4 were simulated to have environmental effects of -0.5, -0.7, +0.5, and +0.3 micron, respectively. These flock effects were simulated such that the difference between flocks across-genetic groups was half of the residual variance (σ_e^2) of 2 micron². A genetic difference of 0.8 micron (-0.4 and +0.4 micron for genetic groups 1 and 2, respectively) was simulated in the mean FD of the two genetic groups. Similarly, the two genetic group effects were simulated such that the difference between the two was approximately half of the additive genetic variance (σ_a^2) of 1.5 micron². The level of linkage due to Xsires between the two genetic groups was varied over 13 levels from 0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, to 50%. The amount of linkage due to WFsires between the two genetic groups was varied over 13 levels from 0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, to 50%. The two types of linkage were tested independently between the two genetic groups due to Xsires and WFsires. The differences between the two scenarios are presented in Table 3.1.

Table 3.1. Differences between the two scenarios

Variable	Scenario I	Scenario II
1. Number of flocks	2	4
2. Number of flocks per genetic group	1	2
3. Progeny per sire generated	20	10
4. Genetic difference between two genetic groups (micron)	2	0.8
5. Flock effects (micron)	-1, +1	-0.5, -0.7, +0.5, +0.3
6. Levels of linkage tested	6	13

Genetic analysis

Genetic group effects, flock effects, and the EBVs of animals were estimated from each replicate in univariate analyses using an animal model in the ASReml package (Gilmour *et al.*, 2006). Two different models were fitted to examine the effect of excluding (Model 1) and including (Model 2) genetic group effects on the accuracy of partitioning of genetic and environmental differences between flocks. The models fitted were of the following form:

$$\text{(Model 1)} \quad \mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

$$\text{(Model 2)} \quad \mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypic observations on the simulated trait, $\mathbf{b}$ is a vector of fixed effects attributed to each flock, $\mathbf{u}$ is a vector of random additive genetic effects of the animal, $\mathbf{g}$ is a vector of genetic group effects, $\mathbf{e}$ is a vector of random residuals; $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices associating observations in $\mathbf{y}$ with the flock and random animal effects in $\mathbf{b}$ and $\mathbf{u}$, respectively and $\mathbf{Q}$ is a design matrix describing the proportion of genes contributed by each genetic group to the animals with records.

Genetic groups were defined on the basis of source of origin (flock) and were fitted in the model through the pedigree. Genetic group effects were treated as fixed effects. The genetic variance was estimated and for each progeny, the sire was known and the dam was indigenous, that is, the dam was from the same genetic group as the progeny. The accuracy of prediction was determined using the correlation (across-flock

and within-flock) between the simulated TBVs and the EBVs. The within-flock correlation was calculated and averaged over all animals within a flock. The across-flock correlation was calculated and averaged over all animals across flocks and was therefore a mixture of between- and within-flock correlations, dominated by the within-flock relationships. Each analysis was repeated by excluding and including genetic group effects in the pedigree and with varying levels of each type of sire linkage.

Test of significance for determining the minimum level of linkage, choice of models, and accuracy of across-flock EBVs

The minimum level of linkage required for estimating genetic group effects and across-flock EBVs with accuracy at 5% level of significance was determined using a t-test. The average difference in the genetic means of two genetic groups (genetic groups 1 and 2) for h^2 of 0.1, 0.3 and 0.43 was estimated along with the standard error of difference (SED) at varying levels of external and across-flock sire linkage. The SED can be defined as the standard error associated with the estimated difference between the means of two genetic groups. The level of significance of the difference between the means of two genetic groups was determined using a t-statistic (ratio test for determining the level of significance) where the product of 1.96 x SED needed to be smaller than the difference. The t-test provides a discrete statistical measure for explaining the difference in the means of the two genetic groups relative to the spread or variability. The formula for the t-statistic is:

$$t = \frac{\overline{GG1} - \overline{GG2}}{SED}$$

where $\overline{GG1}$ and $\overline{GG2}$ are the means for genetic groups 1 and 2, respectively.

The estimated SED was also used to assess the significance of fitting genetic groups to Model 2 for the difference (between genetic groups 1 and 2) and data structure simulated in this study. At a given level of linkage, if the difference between the means of two genetic groups is equal to or less than the associated SED, then genetic groups could be ignored in the model so as to minimize the MSE (Mean Square Error) of prediction (Kennedy, 1981). However, Kennedy (1981) suggested that if the difference between the means of two genetic groups is greater than the associated SED,

then genetic groups must be included in the model. To assess the SED associated with the estimated difference in the means of two genetic groups (genetic groups 1 and 2), the SED was compared with the genetic standard deviation (σ_a) in order to evaluate the accuracy of estimated genetic group solutions. The acceptable threshold for the accuracy of EBVs are often set at either 70 or 80% resulting in standard errors of EBVs which are close to one-half or one-third of the genetic standard deviation. These two values (70 and 80%) were subsequently used to provide a meaningful threshold to achieve SED which are smaller than one-half or one-third of the genetic standard deviation.

3.3 Results and Discussion

Scenario I

The estimated genetic group solutions at different levels of Xsire (genetic group 3) linkage are shown in Figure 3.2. Since the two flocks considered in this study were treated as two separate genetic groups (genetic groups 1 and 2), there was a complete confounding of effects specific to each group (Figure 3.2) in the absence of linkage (0% linkage). There were no contrasts between the two genetic groups (flocks), which is a prerequisite together with linkage for estimating their separate effects. The lack of contrasts implies that the genetic groups were not connected across different levels of fixed flock effects included in the model. The simulated difference of 2 micron between genetic groups 1 and 2 was not estimable at all levels of linkage (Figure 3.2). This in turn did not allow for the estimation of separate flock effects at different levels of Xsire linkage as shown in Figure 3.3. At all levels of linkage, the simulated flock effects of -1 and +1 micron for flocks 1 and 2, respectively could not be estimated separately whether genetic groups were excluded or included in the model (Figure 3.3). The increase in linkage up to 50% did not have any impact on the estimation of independent flock effects. These results (confounding of genetic group effects) indicate that an accurate partitioning of flock genetic and flock fixed (environmental) effects in across-flock evaluations is also dependent on the data structure (genetic groups need to be cross-classified across different levels of fixed flock effects) besides genetic linkage between flocks and inclusion of genetic groups in the model.

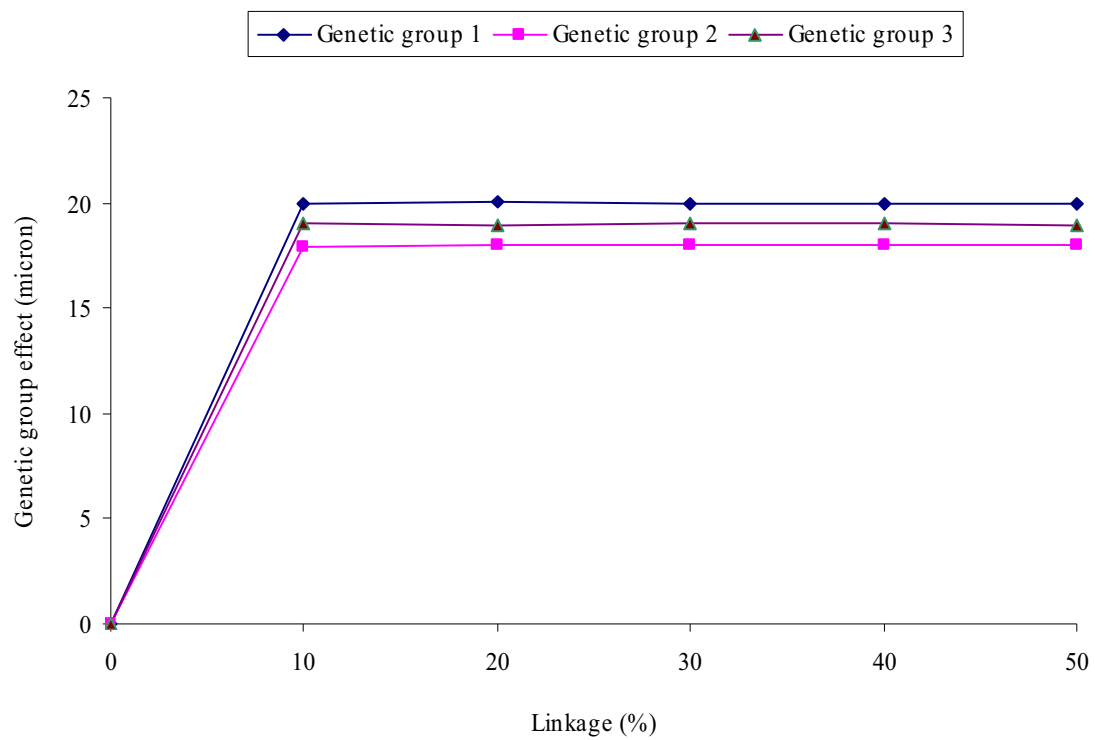


Figure 3.2. Solutions for genetic group effects at different degrees of Xsire linkage (Scenario I)

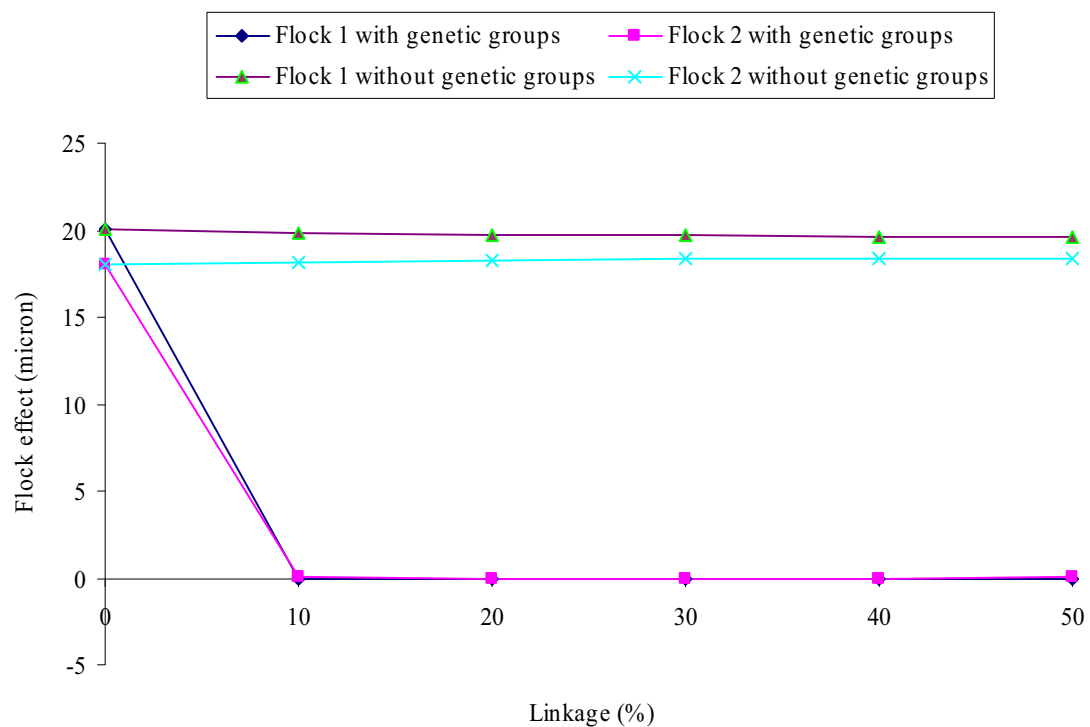


Figure 3.3. Solutions for flock environmental effects for different models and different degrees of Xsire linkage (Scenario I)

The genetic group solutions contain the simulated global mean (19 micron) and flock effects and are thus, biased, as is evident from Figure 3.2. The genetic group effects could not be partitioned from the flock effects in the absence of genetic links. Similar findings were reported by Pollak *et al.* (1977) and Tong *et al.* (1980). Similarly, the estimated flock effects are not independent of the simulated global mean (Figure 3.3). The flock solutions are zero for both the flocks in a groups model. The absence of contrasts between genetic groups results in a design matrix (including fixed flock and genetic group effects) of the MME having linear dependencies between its rows, hence making the matrix not of full rank (Wiggans *et al.*, 1988; Foulley *et al.*, 1990b). This in turn results in the non-estimability of separate genetic group and other systematic effects simultaneously using mixed model methodology as reviewed under section 2.3.2. The estimability of differences between genetic groups is dependent on the groups having animals represented across varying levels of other systematic fixed effects such as flock or herd, year, or season (Tong *et al.*, 1980; Kennedy and Trus, 1993; Kuehn, 2005). In the absence of linkage, as is evident from Figures 3.2 and 3.3, both models (without and with genetic groups) result in similar problems, that is, flock genetic and flock environmental effects cannot be partitioned independent of one another. This signifies the importance of linkage in combination with contrasts or data structure for accurate estimation of genetic group effects in across- flock evaluations.

There was a gradual decrease in correlation between TBVs and EBVs with increasing linkage when genetic groups were accounted for in the model (Figure 3.4). The trend was similar for both external and across-flock sire linkage. This is due to the fact that the EBVs include genetic group solutions that are less accurate. The expectation is that the accuracy of EBVs increases with linkage implying more variance among EBVs. However, when genetic groups were fitted in the model, there was a decline in the variance of EBVs (Figure 3.5) with increasing external sire linkage. This is because an increasing proportion of progeny in each genetic group were derived from common link sires from the third genetic group with an average (zero; -1 and +1 micron for genetic groups 1 and 2, respectively) group effect thereby leading to a reduction in the variance of the TBVs. In the absence of genetic groups in the model, increasing either Xsire or WFsire linkage did not lead to a reduction in the variance of EBVs except when across-flock sire linkage was more than 40%. When genetic groups were

not included in the model, the variance of EBVs was lower for external than for across-flock sire linkage (Figure 3.5) due to the average effect of Xsire group.

At no linkage, the correlation between TBVs and EBVs was 0.5 whether genetic groups were fitted or not in the model. This indicates that there is no impact of including genetic groups in the model on the accuracy of prediction when genetic links do not exist between flocks. However, when linkage is 10%, there is a significant improvement in the correlation (Figure 3.4). The correlation remains constant after that even with an increase in linkage beyond 10%. The inclusion of genetic group effects in the model led to a further improvement in correlations and hence the accuracy of the EBVs (Figure 3.4). This was expected as genetic grouping adds extra information in the BLUP estimation of EBVs by accounting for the differences in the genetic means of base populations (Quaas, 1988). However, these correlations are across-genetic group correlations (the two flocks in Scenario I were treated as two different genetic groups) and must have been affected by the variance between the two genetic groups (Figure 3.4). The variance of the EBVs declines even when genetic groups are included in the model (Figure 3.5) because with increasing linkage more progeny are derived from common link sires, thereby leading to a reduction in the variance of the TBVs and a simultaneous reduction in the accuracy of genetic group solutions that are included in the EBVs. The accuracy of genetic group solutions gets reduced due to less expressions of animals from each genetic group when increasing proportion of progeny is derived from sires coming from another genetic group. Similarly, without any genetic links between groups, the variance of EBVs across all animals remains the same (approximately 0.75 micron²; Figure 3.5) for both models (without and with genetic groups). However, as linkage increases, the variance of EBVs declines for the genetic groups model because a larger proportion of progeny is derived from the same sires (Xsires) with an average (zero) genetic group effect.

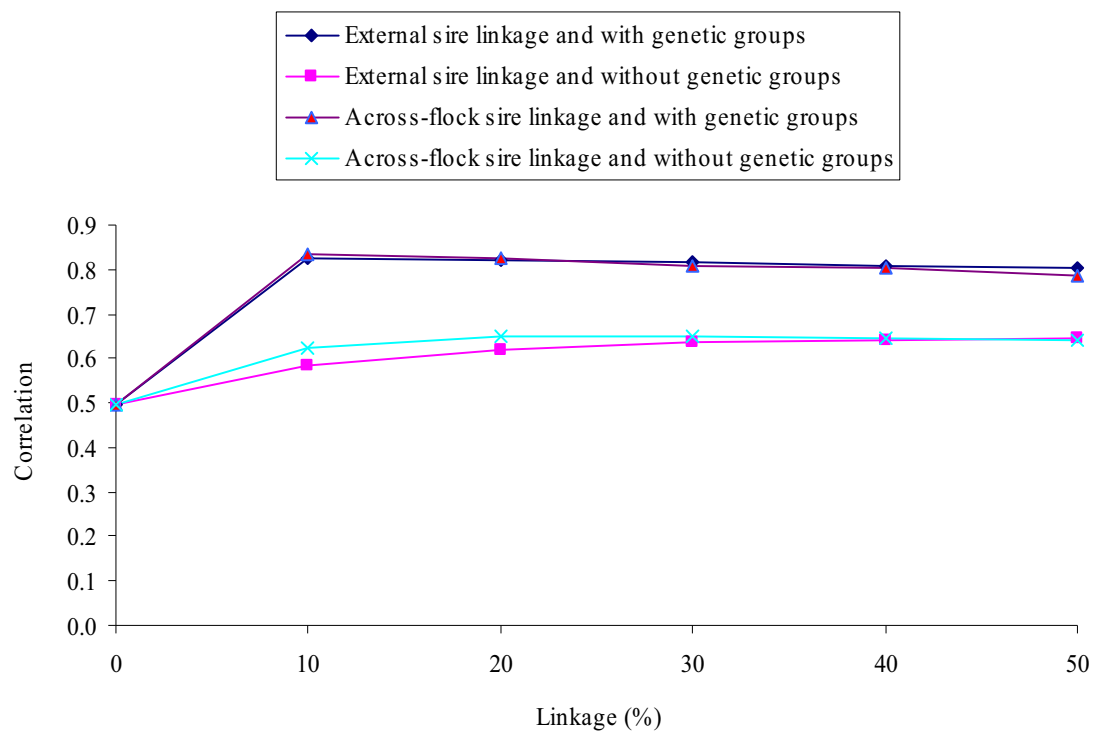


Figure 3.4. Across-flock correlation between the TBVs and the EBVs at varying levels of Xsire and Wfsire linkage (0 to 50%) for different models (Scenario I)

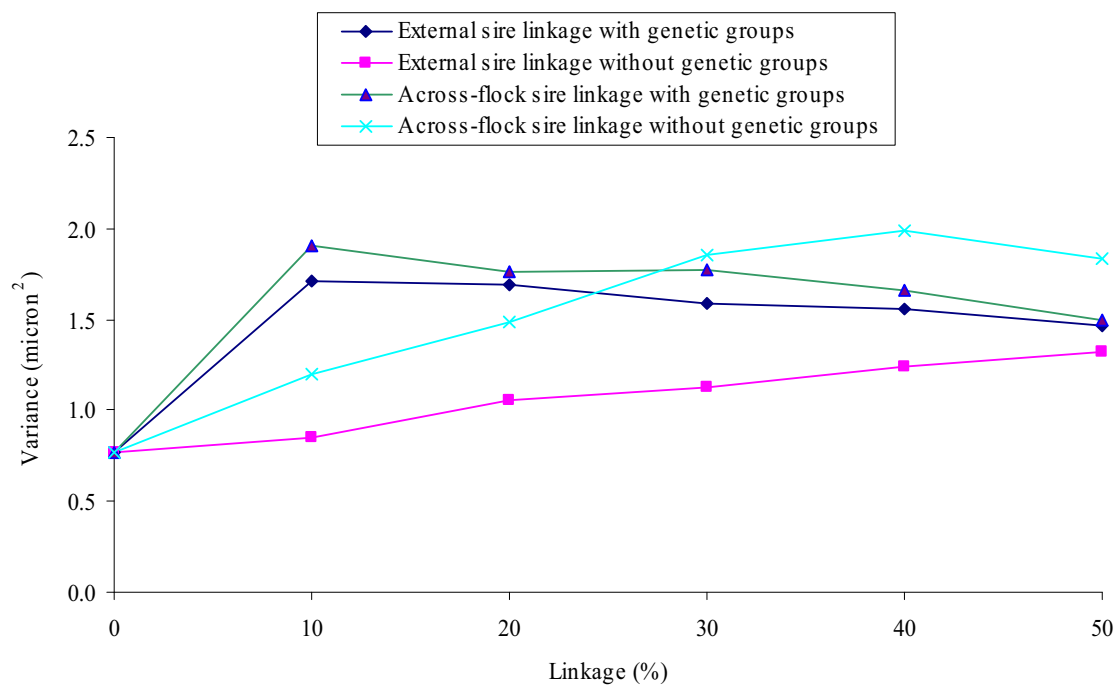


Figure 3.5. The variance of EBVs across all animals at varying levels of Xsire and Wfsire linkage (0 to 50%) for different models (Scenario I)

The difference between across-flock correlations without and with genetic group effects becomes smaller with an increase in both types of linkage (Figure 3.4). A similar conclusion was reached by Kennedy (1981) who stated that with an increase in genetic links across flocks, there is a decrease in the bias due to the differences in their genetic means even if genetic groups are not accounted for in the model. However, there still is a difference in correlations without and with genetic group effects at 50% levels of both forms of linkage (Figure 3.4) due to the fact that the 2 flocks differ in their means and this difference in the means has been accounted for by genetic grouping. However, this accuracy would improve significantly when sufficient linkage is established between the flocks. Although not shown explicitly, the within-flock ranking of sires will not get affected whether genetic groups are fitted or not in the model. Additionally, the level of linkage will not have any impact on the PEV (Prediction error variance) of sires within a flock (or genetic group) because of the large number of direct comparisons between sires within a flock irrespective of the proportion of linkage between flocks as has been demonstrated in previous simulation studies by Kennedy (1981). However, the PEV of across-flock sire EBVs will depend largely on the amount of linkage between flocks (Kennedy, 1981). These results also indicate that accounting for genetic groups in the model in the presence of sufficient proportion of linkage among flocks will increase the accuracy of across-flock EBVs.

In the absence of genetic groups, the within-flock correlations (Figure 3.6) are significantly higher than the across-flock correlations presented in Figure 3.4. This is due to the within-flock correlations being dominated by the within-flock relationships amongst animals. However, the correlation improves when genetic groups are included in the evaluation model for both forms of linkage because genetic grouping accounts for the difference in the means of the two genetic groups. The correlation was slightly higher (at all levels of linkage) when linkage was due to WFsires (known genetic group) than for Xsires (unknown genetic group). The lower correlation for all animals (including Xsires and their progeny) at all levels of Xsire linkage even in the presence of genetic groups could be due to a reduction in the accuracy of estimation for link sires (Xsires) and their progeny as they were evaluated within flock 1 only. Therefore, genetic grouping could not add any extra information as only one genetic mean (for

flock 1) was considered. A similar trend in the correlation between the TBVs and the EBVs was observed for flock 2 (results not shown).

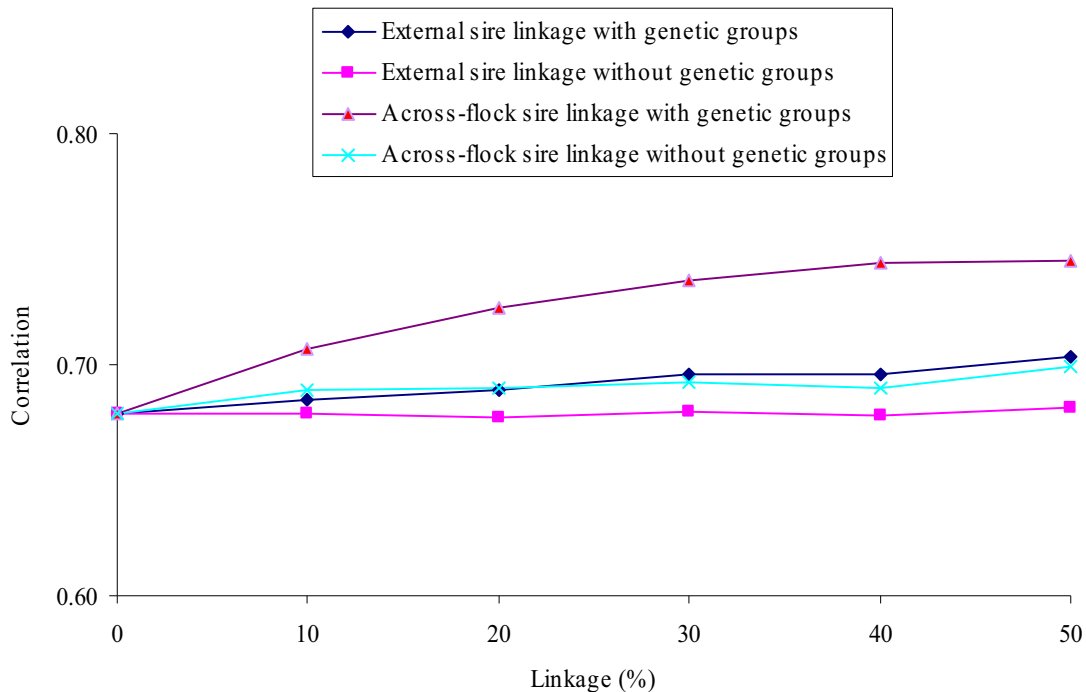


Figure 3.6. Within-flock (genetic group 1) correlation between the TBVs and the EBVs of all animals at varying levels of Xsire and WFsire linkage for different models (Scenario I)

A plot of the TBVs versus the EBVs at 10% Xsire linkage without and with genetic groups for a single replicate is shown in Figures 3.7 and 3.8, respectively. The dark and light colored points correspond to flocks 1 and 2, respectively. It is evident that the two flocks have different mean EBVs when genetic grouping is accounted for in the model (two distinct populations; Figure 3.8). Moreover, it is to be noted that the coefficient of determination (R^2) between the TBVs and the EBVs is the same for both the flocks whether genetic groups are fitted or not in the model because the within-group relationships between animals remains similar in the two flocks when the same proportion (10%) of linkage is created through Xsires (unknown group with missing pedigree) only. Although not represented by Figures 3.7 and 3.8, this relationship between animals within a group will decline with an increase in the level of linkage through sires without any pedigree information (Xsires). Similar findings have been reported by Kuehn *et al.* (2008) who evaluated different measures of connectedness

with regard to their impact on reducing the bias associated with comparing animals across flocks with varying genetic group effects.

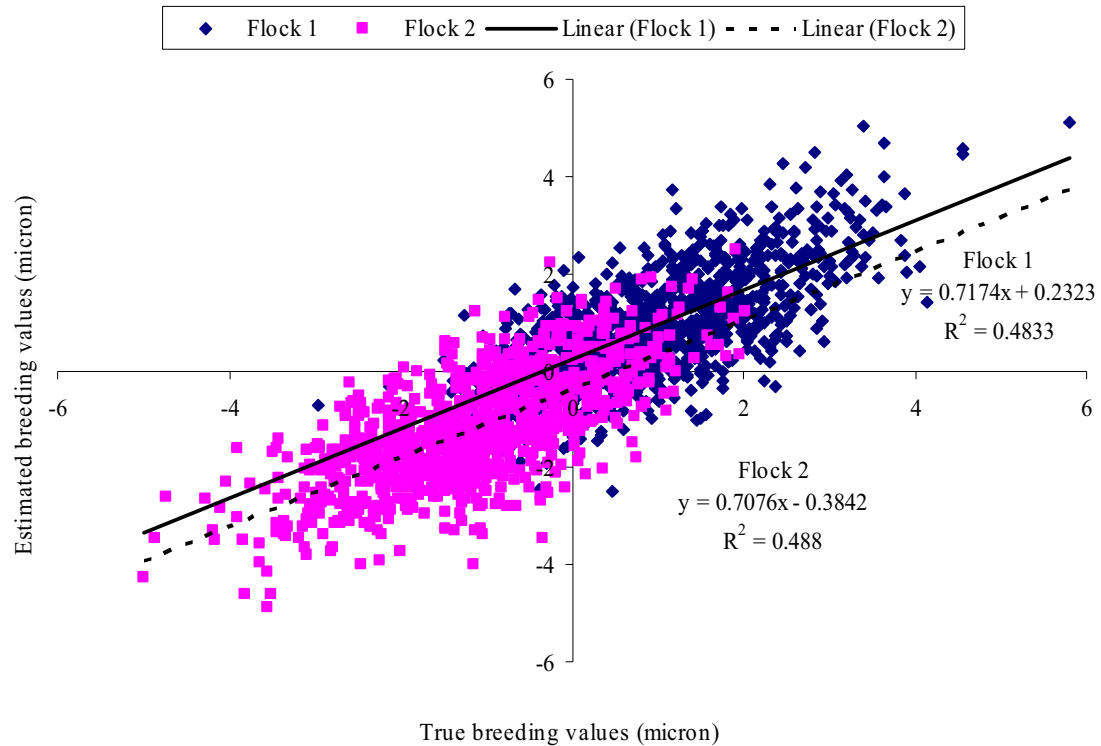


Figure 3.7. True breeding values versus estimated breeding values at 10% Xsire linkage and without genetic groups for a single replicate (Scenario I; dark and light colored points represent flocks 1 and 2, respectively)

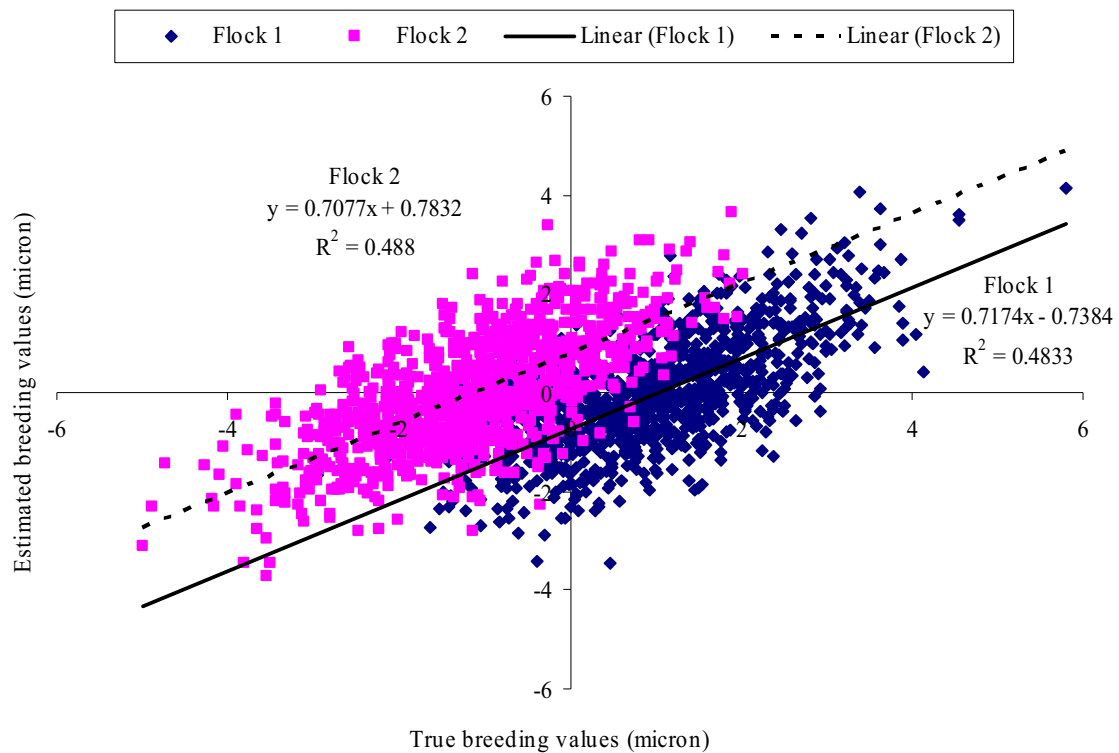


Figure 3.8. True breeding values versus estimated breeding values at 10% Xsire linkage and with genetic groups for a single replicate (Scenario I; dark and light colored points represent flocks 1 and 2, respectively)

The lines in Figure 3.7 are closer than those in Figure 3.8. This is because the difference between the means of the two flocks (genetic groups) is not estimated accurately without accounting for genetic groups in the model. When genetic groups are included in the model, the separated lines in Figure 3.8 represent more accurately (closer to the simulated difference of 2 micron) the estimated difference between the genetic group means of the two flocks. In other words, the difference between the intercepts or estimated means (0.23 and -0.38 micron for flocks 1 and 2, respectively) shown in the regression lines (showing the relationship between TBVs and EBVs) for the two flocks is approximately 0.61 micron when genetic groups are not fitted in the model (Figure 3.7). However, when genetic group effects are included in the model, the difference in the intercepts (-0.73 and 0.78 micron for flocks 1 and 2, respectively) for the two flocks is about 1.5 micron (Figure 3.8). This shows that including genetic groups in the model accounts for the difference in the means of genetic groups in the presence of linkage. However, in the latter case, even though genetic group effects are fitted in the model, still the difference between the two genetic groups is not estimated

accurately (estimated 1.5 micron versus simulated 2 micron) because the genetic group effects of the two flocks (flock 1 and 2) were not estimated accurately under Scenario I.

The actual difference in the flock mean EBVs of WNsires of the two flocks does not remain constant with increasing linkage due to external sires in the absence of genetic groups (Table 3.2). Even at 50% external sire linkage, the difference in the mean EBVs of within-flock sires (WNsires) for each flock cannot be estimated accurately if genetic groups are not included in the model (Table 3.2). These results indicate that even in the presence of higher levels of linkage (50%), the actual simulated difference (2 micron) between the two flocks cannot be estimated accurately without accounting for genetic groups in the model. The accuracy of EBVs in across-flock evaluations can be improved by including genetic group effects in the model even with low amounts (10% linkage) of genetic linkage across flocks (Figures 3.2 to 3.8; Table 3.2). Similar findings have been previously reported by Kuehn (2005).

Table 3.2. Difference between the flock mean EBVs (micron) of WNsires at varying levels of Xsire and WFsire linkage for different models (Scenario I)

Linkage (%)	Xsire linkage with GG*	Xsire linkage without GG	WFsire linkage with GG	WFsire linkage without GG
0	0.00	0.00	0.00	0.00
10	2.02	0.44	2.08	0.85
20	2.01	0.78	2.06	1.28
30	1.99	1.00	1.98	1.45
40	1.94	1.10	2.01	1.54
50	2.02	1.26	2.05	1.57

* Genetic groups

Scenario II

The results ($h^2 = 0.42$) obtained under Scenario II were predictable to a certain extent as has been shown in previous simulation studies by Tong *et al.* (1980) and Kennedy (1981), and reviewed under 2.4.2. The effect of the third genetic group (source of Xsires) is the same as simulated (zero; Table 3.3) when linkage is established through

Xsires. The effect of the first and second genetic groups is the same as simulated when either form of linkage is greater than zero (Table 3.3).

Table 3.3. Genetic group solutions ($\pm$ SE; micron) at varying levels of Xsire and WFsire linkage (Scenario II)

Linkage (%)	Genetic group 1*	Genetic group 2*	Genetic group 3*	Genetic group 1 [#]	Genetic group 2 [#]
0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
1	-0.40 (1.28)	0.40 (1.28)	0.00 (1.44)	-0.40 (0.67)	0.40 (0.67)
2	-0.40 (0.92)	0.40 (0.92)	0.00 (1.02)	-0.40 (0.48)	0.40 (0.48)
3	-0.39 (0.74)	0.40 (0.74)	0.00 (0.81)	-0.40 (0.40)	0.40 (0.40)
4	-0.40 (0.65)	0.39 (0.65)	0.00 (0.74)	-0.41 (0.35)	0.41 (0.35)
5	-0.40 (0.58)	0.40 (0.58)	0.00 (0.65)	-0.40 (0.32)	0.40 (0.32)
10	-0.40 (0.42)	0.40 (0.42)	0.00 (0.47)	-0.41 (0.24)	0.41 (0.24)
15	-0.39 (0.35)	0.40 (0.35)	0.00 (0.39)	-0.40 (0.21)	0.40 (0.21)
20	-0.40 (0.31)	0.40 (0.31)	0.00 (0.34)	-0.40 (0.19)	0.40 (0.19)
25	-0.40 (0.28)	0.40 (0.28)	0.00 (0.32)	-0.41 (0.18)	0.41 (0.18)
30	-0.39 (0.27)	0.40 (0.27)	0.00 (0.29)	-0.40 (0.18)	0.40 (0.18)
40	-0.39 (0.24)	0.40 (0.24)	0.00 (0.27)	-0.41 (0.18)	0.41 (0.18)
50	-0.40 (0.24)	0.40 (0.24)	0.00 (0.26)	-0.40 (0.18)	0.40 (0.18)

* Xsire linkage; [#] WFsire linkage

The genetic group effects are inestimable when there are no genetic links across flocks. However, the genetic group effects are estimable at 1% level of both forms of linkage (Table 3.3). This is in agreement with the study of Hanocq *et al.* (1996) who stated that the differences between subpopulations are estimable even with limited genetic links. However, the standard errors of the effects are large and hence, the solutions are less accurate. In a simulation study, Kennedy (1981) found that the difference between two genetic groups cannot be estimated accurately with limited genetic links. The author argued that at low levels of linkage, the PEV of across-flock comparison of animals is inflated because the variance of prediction of the group difference is a component of that PEV. Cottle and James (1997) stated that the

improvement in linkage will lead to a simultaneous reduction in the PEV of estimated differences between genetic groups. In another simulation study, Tong *et al.* (1980) reported that the difference between two genetic groups can be estimated more accurately with increasing linkage. Hickey *et al.* (2008) stated that the bias in genetic group solutions may bias across-flock EBVs. Therefore, the authors recommended that the accuracy of genetic group solutions must be considered prior to publishing EBVs across genetic groups. With an increase in both types of linkage, there is a gradual reduction in the standard errors of the genetic group effects (Table 3.3). Ashtiani and James (1991, 1992) stated that on average, 33% of progeny should be generated by link sires in different management units (flocks or herds) so as to accurately benchmark the across-unit EBVs for the moderately heritable fleece traits.

The results reported by Ashtiani and James (1991, 1992) were based on the assumption that all sires used for establishing linkage belong to the same genetic group implying only one genetic mean. In the present study, link sires were derived from three different genetic groups with varying genetic means. Therefore, to account for these differences in the means, it is recommended to include genetic groups in the evaluation model. Similar conclusions were made by Hanocq and Boichard (1999). Foulley *et al.* (1983) stated that 10 to 40% of progeny should be derived from link sires for making unbiased across-management group comparisons. However, Ashtiani and James (1991, 1992) and Foulley *et al.* (1983) investigated the level of linkage required to accurately compare animals across management groups, which are more likely to differ in their fixed environmental effects rather than the genetic group effects. Mathur *et al.* (1998) evaluated connectedness between swine herds in Canada and concluded that about 15% of progeny in each herd should be generated from link sires to achieve the minimum level of linkage.

When linkage is established through WFsires, the estimated genetic group effects have lower standard errors than those obtained at any level of linkage through Xsires (Table 3.3). This is because at WFsire linkage only two genetic group effects need to be estimated whereas at Xsire linkage three genetic group effects were estimated. Additionally, there were sufficient progeny records for across-flock sires that allowed for an accurate estimation of their group effect. On the contrary, there were not

enough records on the progeny derived from external sires for the accurate estimation of their group effect. This shows that it would be more advantageous to create linkage through sires from an existing genetic group rather than using sires from an unknown or new genetic group. The standard errors associated with genetic group 1 and 2 solutions are the same at each level and for each type of linkage (Table 3.3) because of the balanced design (same number of animals within each genetic group) of the data simulated. However, from industry's perspective, this would not be the case as genetic groups are likely to vary in their size (number of animals) implying differences in the relative accuracies of genetic group solutions depending on their associated standard errors. Since contrasts between the two genetic groups were available, their effects were estimable. Moreover, the accuracy of comparing animals from different groups is not only dependent on the amount of linkage but also on the size and structure of the units (Cottle and James, 1997; Mathur *et al.*, 1998; Kuehn, 2005). Scenario II allowed enough animals together with linkage and contrasts in each of the four flocks simulated for an accurate across-flock comparison.

At very low levels of both types of sire linkage, the SEDs associated with the difference between the two genetic groups (genetic groups 1 and 2) were relatively large because of small numbers of direct comparisons between animals from different genetic groups (Figure 3.9). Similar findings have been reported by Kennedy (1981). The SEDs decrease gradually with increasing either form of linkage and at 10% and 3% levels for external and across-flock sires, respectively, the SEDs are less than the estimated difference between the two genetic groups (Figure 3.9). This implies that the actual need for accounting for genetic groups in the model is dependent on the level of linkage that exists between the genetic groups. At higher levels of linkage, the need for genetic grouping is imperative provided the difference in the means of genetic groups is relatively large. Kennedy (1981) reported that the need for grouping is greatest at higher levels of linkage and vice versa. However, the author also stated that the proposed criterion for choosing alternative models (without and with group effects) will hold exactly for models involving only two genetic groups, as was the case in the current study.

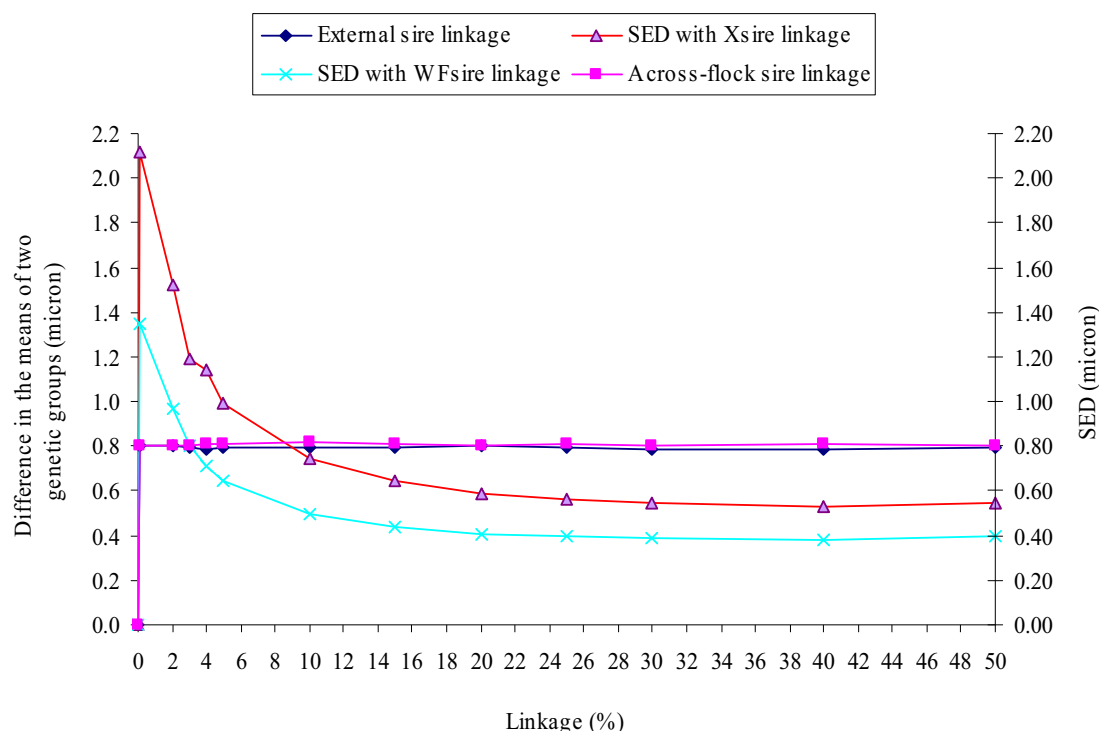


Figure 3.9. Difference in the means of genetic groups 1 and 2 and their associated SEDs at varying levels of Xsire and WFsire linkage (Scenario II)

Although the method proposed by Kennedy (1981) could be treated as an accurate means of determining the need for grouping, it cannot be employed for large scale industry evaluations where large numbers of genetic groups are included in the model such as in the Merinos. Furthermore, the need for accounting for genetic groups in the evaluation scheme on the basis of estimated SED will depend on the magnitude of actual genetic difference between groups. Thus, further research should focus on developing better means and criteria for determining whether genetic groups need to be included or not in the evaluation scheme. In addition, since this was a simulation study, the difference between the two genetic groups was known, therefore, the criterion developed by Kennedy (1981) could be regarded as reasonably accurate. However, as pointed by Tong *et al.* (1980) and Kennedy (1981), in reality this difference is never known. Therefore, Kennedy (1981) suggested making a reasonable guess of the difference or a fairly accurate range of expected minimum and maximum differences. The approximate minimum and maximum values could then be used in a two-tailed statistical test for predicting the difference.

The difference in the means of two genetic groups (genetic groups 1 and 2) was not significant at 5% level ($P < 0.05$) at any level of external sire (Xsire) linkage because the associated SEDs were large (Figure 3.9). However, if external sires were having sufficient progeny records in the two genetic groups (genetic groups 1 and 2) or elsewhere so as to allow an accurate estimation of their genetic group effect, the SEDs would have been smaller and the difference in the means of the two genetic groups significant at 5% level. The difference in the means of genetic groups 1 and 2 was significant at 5% level ($P < 0.05$) at 25% linkage due to across-flock (WFsires) sires. These findings indicate that linkage should be established only through sires (with sufficient progeny records) from those genetic groups whose effects could be estimated with accuracy. The SEDs were lower in case of across-flock sire linkage (Figure 3.9). This was expected due to the difference in the numbers of records available on progeny derived from external and across-flock sires. This shows that establishing genetic links from indigenous or local genetic groups is better than from unknown (external) or new groups. In a simulation study, Kuehn *et al.* (2008) studied the risk of evaluating animals (based on EBVs) across flocks using different measures of connectedness and different types of link sires (homebred versus outside sires). They reported that the bias in across-flock EBVs was lower when only homebred sires were used for establishing connectedness in comparison to when only outside sires were used. However, if both homebred and outside sires are used, then the minimum proportion of linkage required for the accurate partitioning of genetic and environmental differences between flocks would be different to that required in the current investigation.

The SEDs were larger when only external sires were used for creating genetic links between the two genetic groups (Figure 3.9). This is attributed to the fact that there were not enough progeny (generated from external sires) records available for the accurate estimation of their genetic group effect. Keeping in view that this was a simulation study, the actual amount of linkage required in industry applications could vary depending on the magnitude of genetic differences between groups and the structure of the data (number of animals in each genetic group). The choice of using within-flock or external sires for creating links across flocks would depend, to a certain extent, on individual circumstances of producers and their choice of accuracy desired in the across-flock EBVs. The choice of link sires could also depend on the visual

appearance of animals as not all sheep breeders make selection decisions based on EBVs. In the words of Fouilloux *et al.* (2008), “the choice of this level (link sires) is rather arbitrary just like the choice of a level of accuracy for individual genetic values (EBV) to be published.”

At very low levels (< 10%) of linkage established through across-flock sires, the SED associated with the estimated difference between the two genetic groups (groups 1 and 2) is large although the difference is still estimable (Figure 3.9). The SED is in the range of approximately 0.5 to 1.34 micron when the level of linkage is less than or equal to 10%. At 25% linkage established from across-flock sires, the SED is 0.40 micron (Figure 3.9) which is approximately one-third of the simulated genetic standard deviation of 1.22 micron. To derive genetic group solutions and hence across-flock EBVs with an accuracy (measured in terms of correlation between TBVs and EBVs) of 70 or 80%, the SED needs to be smaller than one-half of the genetic standard deviation. The minimum threshold for the accuracy of EBVs for body weight that is required before they are published by Sheep Genetics is determined by the product of $\sqrt{h^2} \times 0.8$ where h^2 is the heritability of the trait (0.40 for body weight) and 0.8 is the correlation between the same trait measured at different ages like weaning, yearling, and hogget (Brown, personal communication). The accuracy (square root of h^2) of EBVs is defined as a correlation between the theoretical TBVs and the EBVs. The EBVs whose calculated accuracy is below the threshold of $\sqrt{h^2} \times 0.8$ are not considered to be accurate enough to be reported. However, for reporting across-flock EBVs, a threshold of a minimum of 25% linkage between flocks should also be taken into consideration for precisely estimating the difference between different genetic groups. At 25% linkage between flocks, the SED related with the estimated difference between the means of genetic groups will be approximately one-third of or less than the genetic standard deviation.

The genetic group solutions obtained using Scenario I were confounded with the fixed flock (environmental) effects and were not able to be estimated individually (Figure 3.2). However, the data structure simulated in Scenario II enabled differentiation between the means of the two genetic groups because of sufficient

connectedness of genetic groups across different levels of fixed effects included in the model such as flock. This in turn allowed for the accurate estimation of pairwise differences between varying levels of fixed flock effects as discussed in the following paragraph. These results indicate that the level of linkage, data structure and the magnitude of genetic differences between groups will influence whether genetic groups could be fitted or not in the evaluation model. If the differences between genetic groups are not significantly large and there are limited genetic links between the two, grouping of animals could be counterproductive (Kennedy, 1981; Phocas and Laloe, 2004) thereby affecting the accuracy of prediction. Famula *et al.* (1983) stated that the unnecessary inclusion of group effects in the model will increase the PEV of EBVs.

The flock effects (expressed as deviations from flock 1) at varying levels of external sire linkage and without genetic groups ($h^2 = 0.42$) in the model are shown in Table 3.4. The results show that the flock solutions (Flocks 3 and 4) contain the genetic group effects and are therefore inaccurate. The results presented in Table 3.4 show that without accounting for genetic groups in the model, the accurate flock (environmental) effects cannot be estimated even at higher levels of linkage. Similar results were obtained at different levels of across-flock sire linkage and are therefore not presented. In the presence of either type of linkage and with genetic groups included in the model, the differences (genetic and environmental) between flocks can be estimated as is evident from results shown in Table 3.4 for external sire linkage. However, at low levels of linkage, the standard errors are large and thus, the solutions are less accurate. There is a gradual reduction in the standard errors of flock solutions with increasing linkage. A similar trend was observed (results not shown) when linkage was established through across-flock sires. However, at all levels of across-flock sire linkage, the standard errors of flock solutions were significantly lower than those obtained when linkage was established through external sires (results not shown). This again signifies the superiority of across-flock sires over external sires in establishing linkage. Table 3.4 shows that the estimated flock differences for the model with genetic groups were the same as those simulated for the 4 flocks.

Table 3.4. Estimated flock effects expressed as deviations from Flock 1 ($\pm$ SE; micron) at varying levels of Xsire linkage for different models (Scenario II)

Linkage (%)	<u>Without genetic groups</u>				<u>With genetic groups</u>			
	Flock 1 (-0.5)*	Flock 2 (-0.7)*	Flock 3 (+0.5)*	Flock 4 (+0.3)*	Flock 1 (-0.5)*	Flock 2 (-0.7)*	Flock 3 (+0.5)*	Flock 4 (+0.3)*
0	0.00 (0.07)	0.00 (0.07)	0.00 (0.07)	0.00 (0.07)	0.00 (0.07)	-0.19 (0.07)	1.80 (0.07)	1.59 (0.07)
1	0.00 (0.07)	-0.24 (0.07)	1.77 (0.07)	1.57 (0.07)	0.00 (1.29)	-0.24 (1.29)	1.00 (1.29)	0.76 (1.29)
2	0.00 (0.07)	-0.19 (0.07)	1.78 (0.07)	1.58 (0.07)	0.00 (0.90)	-0.19 (0.90)	0.99 (0.90)	0.78 (0.90)
3	0.00 (0.07)	-0.20 (0.07)	1.75 (0.07)	1.56 (0.07)	0.00 (0.74)	-0.20 (0.74)	0.98 (0.74)	0.78 (0.74)
4	0.00 (0.07)	-0.20 (0.07)	1.73 (0.07)	1.54 (0.07)	0.00 (0.64)	-0.20 (0.64)	0.96 (0.64)	0.76 (0.64)
5	0.00 (0.07)	-0.20 (0.07)	1.72 (0.07)	1.50 (0.07)	0.00 (0.58)	-0.20 (0.58)	1.00 (0.58)	0.79 (0.58)
10	0.00 (0.07)	-0.19 (0.07)	1.65 (0.07)	1.45 (0.07)	0.00 (0.41)	-0.19 (0.41)	0.99 (0.41)	0.79 (0.41)
15	0.00 (0.07)	-0.20 (0.07)	1.59 (0.07)	1.41 (0.07)	0.00 (0.34)	-0.20 (0.34)	1.00 (0.34)	0.81 (0.34)
20	0.00 (0.07)	-0.19 (0.07)	1.56 (0.07)	1.36 (0.07)	0.00 (0.30)	-0.19 (0.30)	1.02 (0.30)	0.82 (0.30)
25	0.00 (0.07)	-0.21 (0.07)	1.50 (0.07)	1.31 (0.07)	0.00 (0.27)	-0.21 (0.27)	0.98 (0.27)	0.79 (0.27)
30	0.00 (0.07)	-0.19 (0.07)	1.48 (0.07)	1.29 (0.07)	0.00 (0.24)	-0.19 (0.24)	0.99 (0.24)	0.80 (0.24)
40	0.00 (0.07)	-0.16 (0.07)	1.45 (0.07)	1.26 (0.07)	0.00 (0.22)	-0.16 (0.22)	1.00 (0.22)	0.82 (0.22)
50	0.00 (0.07)	-0.20 (0.07)	1.39 (0.07)	1.21 (0.07)	0.00 (0.20)	-0.20 (0.20)	1.00 (0.20)	0.83 (0.20)

* Simulated flock effects

In the absence of either linkage or genetic grouping, the flock effects will be confounded with the genetic group effects, thus introducing inaccuracy in the across-flock EBVs. Similar findings were reported by Tong *et al.* (1980), Hanocq *et al.* (1996) and Clement *et al.* (2001) in simulation studies. These results also indicate that including genetic groups in the model will account for the missing pedigree information in the Australian Merino industry datasets by partitioning the environmental (flock) and genetic group effects. A similar argument was put forth by Quaas and Pollak (1981). Although simulation studies by Barnett *et al.* (1998) on Merino sheep have shown that the extra cost involved in obtaining dam pedigrees is generally greater than the value of the extra genetic gain (for production traits only) achieved, it is desirable for producers to record performance data and dam pedigrees (Banks, 1997; Brash, 1999; Casey and Atkins, 1999) keeping in mind the economic importance of dam fertility traits in breeding programs (Brown *et al.*, 2001a). In addition, the relative significance of pedigree information to individual records increases when the trait is lowly heritable as has been shown in studies by Famula (1993), Sullivan (1995), Mark *et al.* (2002) and Cardoso and Tempelman (2003).

The change in across-genetic group correlation between the TBVs and the EBVs without and with genetic group effects is shown in Figure 3.10. The trend in this graph resembles that in Figure 3.4 (Scenario I). However, the correlation shown in Figure 3.10 is across genetic groups and is different from within-genetic group correlation presented in Figure 3.11 for genetic group 1 because of the increased covariance between the TBVs and the EBVs across flocks, which in turn is due to the difference in the means of the two genetic groups. The correlation improved significantly when genetic groups were included in the model with an increase in both forms of linkage (Figure 3.10) up to 15%. A further increase in linkage did not improve correlation any further. Similar trend was observed for Scenario I (Figure 3.4) and is attributed to a reduction in the variance of simulated TBVs in the data with increasing linkage. At higher levels of external sire linkage, a greater proportion of progeny in each flock is derived from the common sires simulated to have a genetic group effect of zero, hence resulting in a decrease in the variance of their TBVs. In addition, with a continual increase in the proportion of linkage, there is a reduction in the accuracy of genetic group solutions thereby resulting

in less accurate EBVs, which in turn leads to a decline in the correlation between the TBVs and the EBVs.

A plot of the difference in the mean EBVs of WNsires of each flock at different levels of external sire linkage without and with genetic groups in the model is presented in Figures 3.12 and 3.13 (the solutions for flock 1 are set to zero in both figures), respectively. In the absence of genetic groups (Figure 3.12), even at very high (50%) levels of linkage, the actual difference in the average genetic merit of WNsires across genetic groups cannot be estimated. The mean EBVs of WNsires for each flock within a genetic group are the same as simulated (Figure 3.13). This difference remains constant with a further increase in linkage up to 50%. This indicates that flocks can be differentiated from each other in their mean performance even with low amounts of linkage when genetic groups are included in the model.

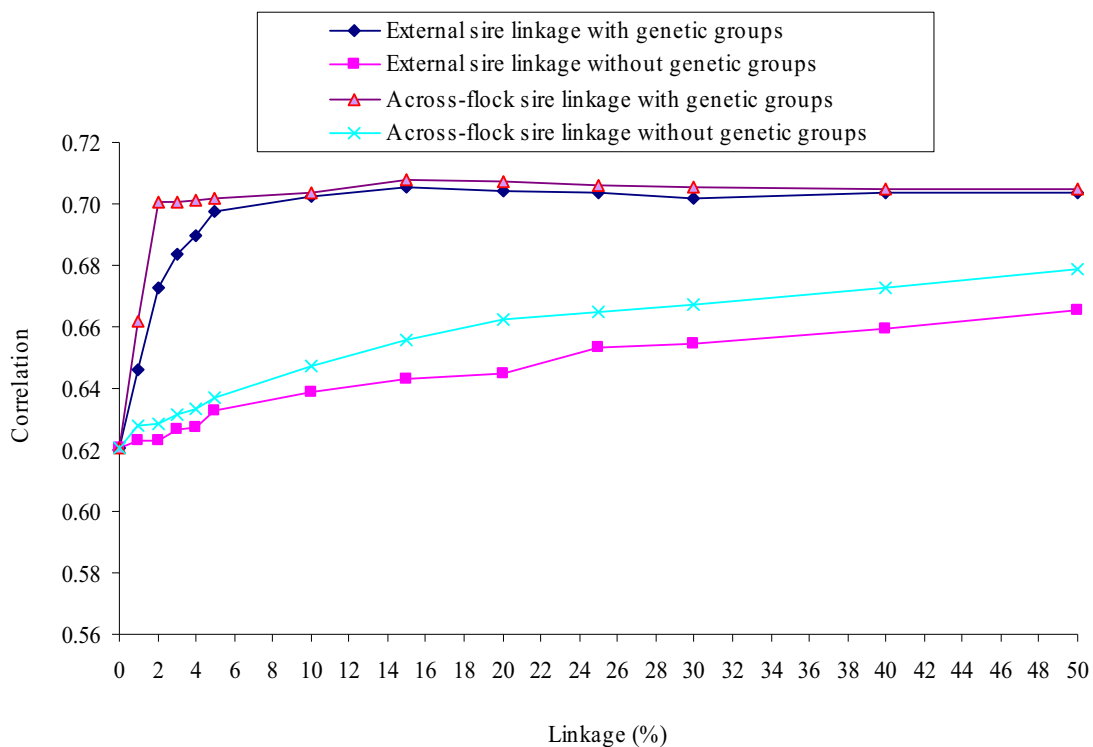


Figure 3.10. Across-genetic group correlation between TBVs and EBVs at varying levels of Xsire and WFsire linkage for different models (Scenario II)

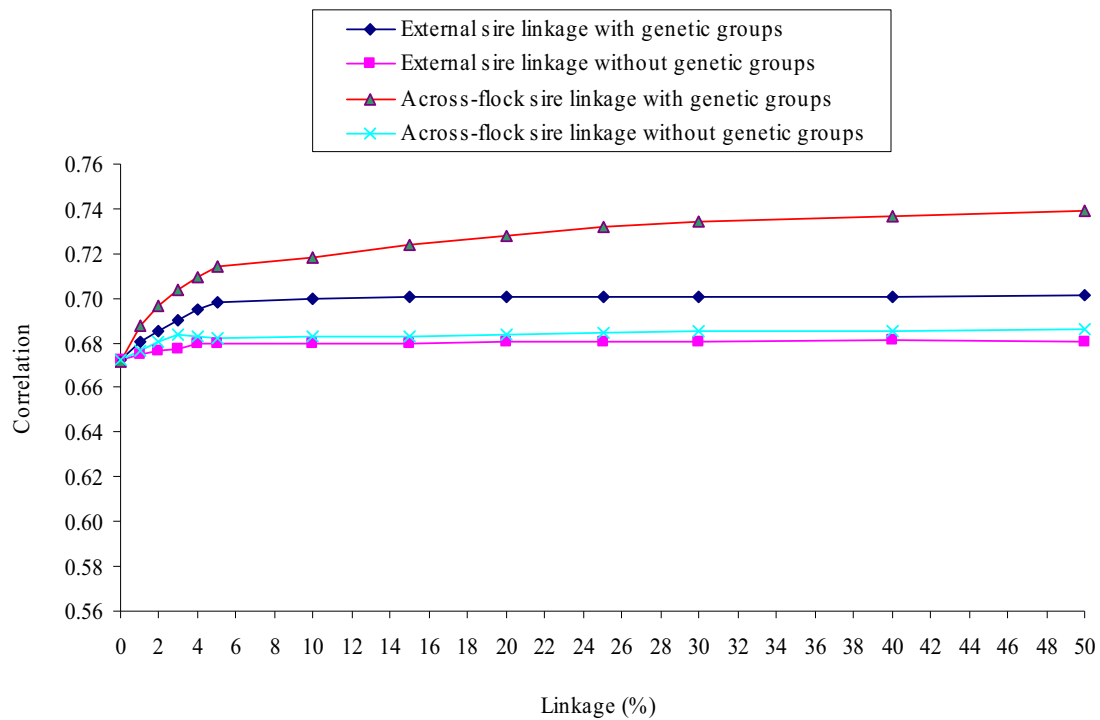


Figure 3.11. Correlation between TBVs and EBVs within genetic group 1 at varying levels of Xsire and WFsire linkage for different models (Scenario II)

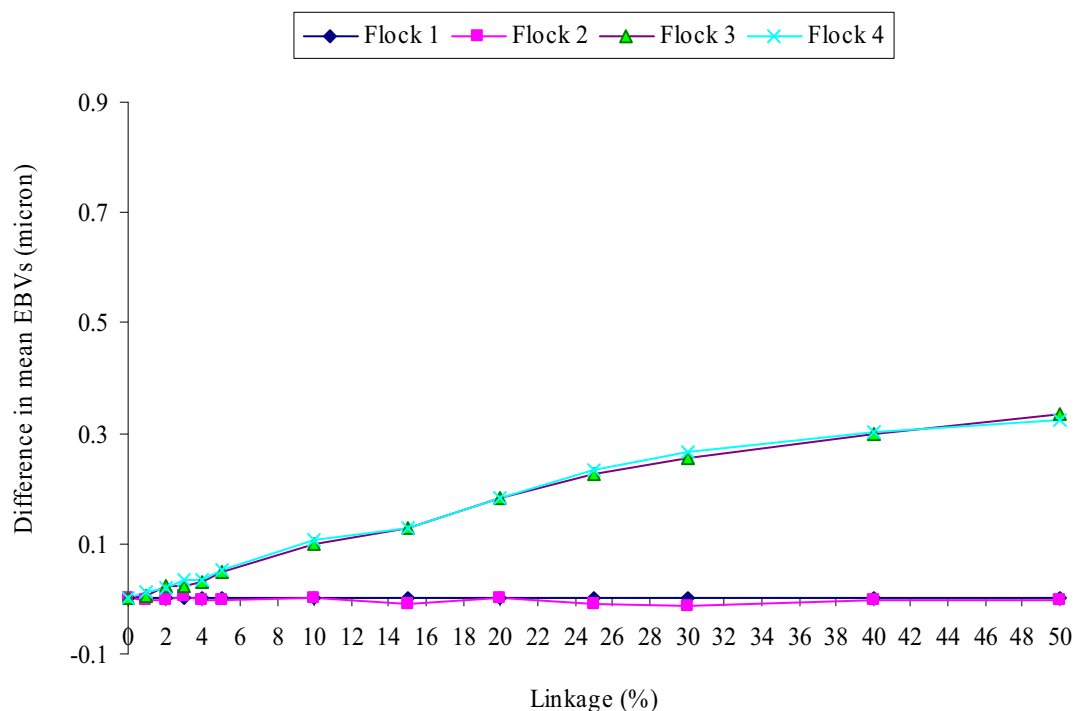


Figure 3.12. Difference in the mean EBVs of WNsires by flock at varying levels of Xsire linkage and without genetic groups (Scenario II)

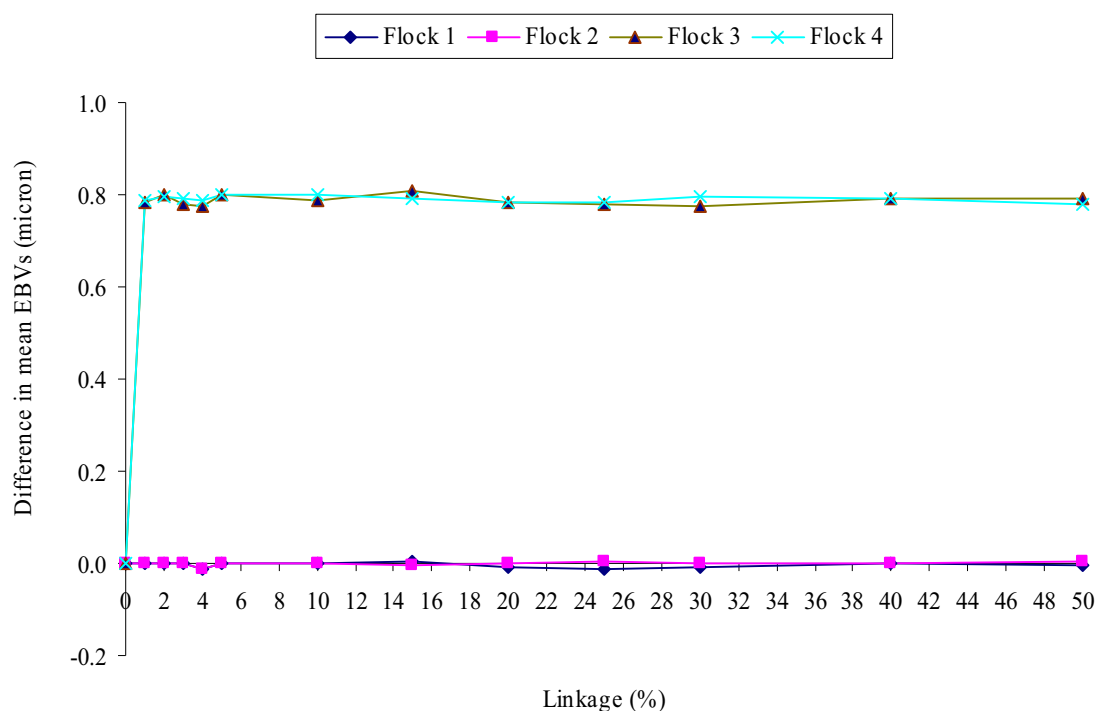


Figure 3.13. Difference in the mean EBVs of WNsires by flock at varying levels of Xsire linkage and with genetic groups (Scenario II)

Similar trends, as in Figures 3.12 and 3.13, were evident when linkage was established through WFsires (results not shown).

Altering heritability

Similar results which are not presented were obtained for h^2 of 0.30 and 0.10. This indicates that for traits with low, moderate or high heritability, genetic group effects are estimable at 5% significance level ($P < 0.05$) when at least 25% linkage from WFsires (indigenous genetic group) exists among flocks with varying genetic means. However, the exact proportion of linkage required could vary with the differences in the means of the genetic groups. The mean standard errors of genetic group solutions were different at comparable levels of genetic links due to external sires across flocks (Figure 3.14). The average standard errors were directly proportional to the level of h^2 . This trend appears to be diminishing gradually with an increase in the level of linkage (Figure 3.14). These results also show that the minimum linkage required across flocks for estimating genetic group effects with acceptable standard errors remains relatively

similar irrespective of the heritability of the trait. Similar conclusions were made by Foulley *et al.* (1983).

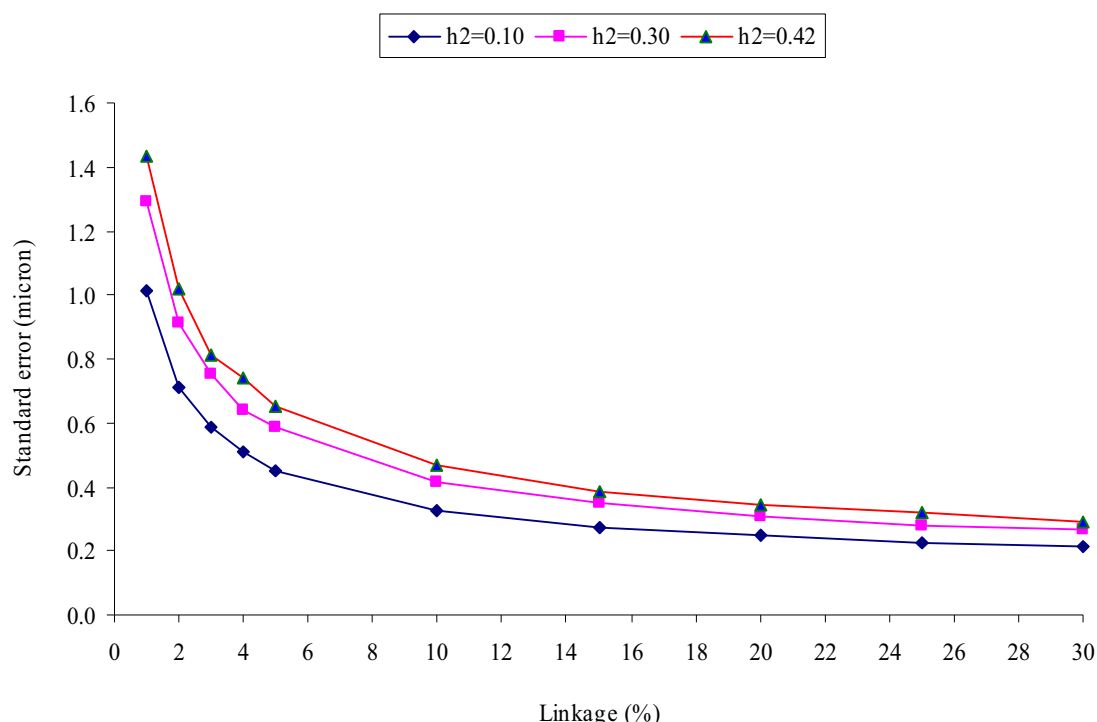


Figure 3.14. Standard errors of estimated genetic group effects (flocks 1 or 2) at varying levels of Xsire linkage and heritability (Scenario II)

Altering number of sires and progeny per sire

The number of progeny per sire (previously was 10) was then varied to 5 and 2 (h^2 of 0.42). The mean genetic group effects were estimated at the same level of both types of sire linkage as reported earlier (results not presented). However, there were larger standard errors (Figure 3.15) compared to 10 progeny per sire. At the same number of link progeny, the average standard errors of estimated genetic group effects were marginally higher when more progeny were generated from less link sires compared to less progeny from more sires (Figure 3.15). In another study, Cottle and Hickson (1996) stated that the average PEV of EBVs is improved when less number of link sires is used. These authors also stated, “if many different sires are used as link sires, the percentage of progeny from common link sires will be less than optimum”. The findings reported by Cottle and Hickson (1996) are in agreement with those of Ashtiani and James (1991, 1992) who advocated that for a given number of link sires,

increasing total progeny (sum of progeny from link and home-bred sires) will reduce the PEVs of link sire EBVs. In addition, an increase in the number of link sires will lead to an equivalent increase in the PEV of sire EBVs due to less comparisons (Ashtiani and James, 1992) in progeny testing programs. However, Ashtiani and James (1991, 1992) investigated the significance of linkage when comparing animals across management units, which generally differ in their environmental effects rather than the genetic effects. On the contrary, the flocks simulated in this study differed both in their environmental and genetic effects.

Although the same number of progeny were generated from link (Xsires and WFsires) and internal (WNsires) sires in the present study, link sires should have enough progeny to accurately evaluate animals across flocks whereas sires born within the flocks must have sufficient progeny to allow them to be progeny tested (Kuehn, 2005) with increased precision. However, the actual number of progeny per link sires required depends on the amount of variation among different populations to be differentiated from one another so that not many progeny per sire are needed when the difference between populations is large (Ashtiani and James, 1993). The effect of h^2 on the number of progeny generated per link sire for achieving optimum levels of linkage across flocks was not investigated in this study. The optimum number of progeny required per link sire decreases with an increase in the level of h^2 (Ashtiani and James, 1992; Huisman *et al.*, 2006) due to relatively small differences in the environmental variance within each flock. In practice, the actual number of progeny required per link sires would also depend on the use of link sires across years. The level of linkage between flocks could vary over a time period depending on the use of same or different link sires (for example, Kuehn *et al.*, 2008). This could influence interpretation of the results obtained in this study. A similar argument was put forth by Kennedy (1981) with respect to linkage across time. Additionally, this could also have an impact on the accuracy of estimation of genetic trend as discussed under 2.3.1.2. If linkage across years is not adequate, animals born (or used) in different years cannot be accurately compared with one another. Although this is rarely the case for ewes, across-year evaluation is highly relevant to selection of sires (Brash, 1999) in the Merino industry.

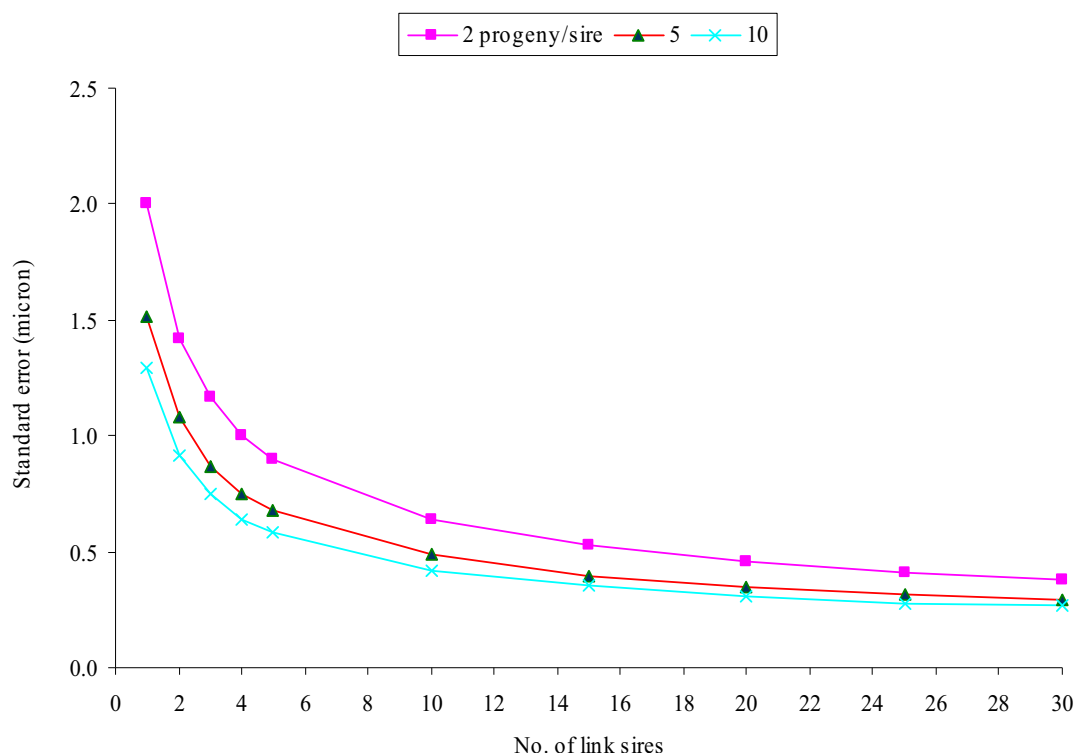


Figure 3.15. Average standard errors of estimated genetic group effects at varying levels of Xsire linkage and progeny per sire (Scenario II)

3.4 Conclusions and Implications

For the population structure, flock size, and the magnitude of genetic difference simulated between the two flocks or genetic groups, an accurate partitioning of genetic and environmental effects on the phenotype of animals in across-flock evaluations could be achieved with 1% linkage from either indigenous or external sires. However, the genetic group solutions are not accurate because of large standard errors. To estimate genetic group effects and hence, across-flock EBVs accurately, at least 25% linkage through sires from a known (with sufficient progeny records to accurately estimate group effects) genetic group must be established across flocks and genetic groups need to be fitted in the model. For obtaining genetic group solutions with an accuracy of 70%, the SED associated with the difference in the means of two genetic groups needs to be approximately one-half of the genetic standard deviation. If the accuracy desired in the estimated genetic group effects is 80% then the SED needs to be approximately one-third of the genetic standard deviation. Under both cases, the standard errors of EBVs are expected to be less than the genetic standard deviation. The results obtained

indicate that establishing linkage from known genetic groups is better than from unknown (without sufficient progeny records to accurately estimate group effects) or external groups. In the present study, the genetic group effect for external sires was not estimable with the same level of accuracy as the effects for the other two genetic groups (source of across-flock sires) due to the difference in the numbers of progeny records that were available for external and across-flock sires. Therefore, the difference between the two genetic groups was not statistically significant at any level of external sire linkage.

The actual level of linkage required for disentangling genetic and environmental differences between flocks could vary depending on the real genetic differences between genetic groups. In other words, the level of significance regarding linkage at which genetic difference between groups is estimable is not only dependent on the standard error associated with the estimated difference, but also on the magnitude of the difference. Although not examined in the present study, the minimum linkage required could be less than 25% if the genetic differences between flocks are significantly large and vice-versa. In addition, in this study, linkage was established reciprocally in the two genetic groups implying that sires from one group were used to create linkage in another group and vice versa, and there were no genetic relationships amongst those sires. Therefore, the interpretation of results obtained in this study could be broader if only one-way linkage is established and genetic relationships exist amongst link sires as has been shown in previous studies by Pollak *et al.* (1977), Tong *et al.* (1980) and Kennedy (1981). Tong *et al.* (1980) found reciprocal linkage to be more desirable than one-way linkage.

The simulated difference between the two genetic groups in Scenario II of this study was sufficiently large to necessitate inclusion of genetic group effects in the model. Although not investigated in this study, if the differences between flocks are not significantly large, the inclusion of group effects in the model could reduce the accuracy of evaluation scheme (Kennedy and Moxley, 1975; Kennedy, 1981; Phocas and Laloe, 2004) by increasing the PEV of across-flock EBVs (Famula *et al.*, 1983; Kuehn *et al.*, 2007). Additionally, in this study, the two types of linkage were tested independently of each other. However, from industry's perspective, it is a usual practice to create genetic

links across flocks simultaneously through both indigenous and external sires. Therefore, the amount of linkage required could vary depending on the proportion of progeny derived from indigenous and outside sires in different flocks. In this study linkage was created through common sires and there were no genetic relationships through dams between the two genetic groups. However, this may not be the case with industry datasets where dam relationships and retained female progeny could also enhance linkage across flocks (Kuehn *et al.*, 2007). Further research should explore the effect of linkage due to both sires and dams on the accuracy of genetic group solutions. Nevertheless, linkage through dams cannot be a substitute for direct sire linkage, which has been found to be most effective in reducing bias in across-flock evaluations (Kuehn *et al.*, 2007).

Depending on the data structure, an appropriate grouping strategy will account for the varying genetic means of base populations and subpopulations from different environments and for traits with varying heritabilities. The environmental effects specific to different flocks cannot be accurately separated from genetic effects when genetic groups are not fitted in the model even in the presence of higher levels of linkage among flocks. The estimability of genetic group effects is an important issue that needs to be investigated prior to estimating genetic group and other fixed effects simultaneously. In the absence of sufficient contrasts or lack of comparison across different levels of fixed effects in different flocks, the genetic group effects would be confounded with other fixed effects in the model such as flock.

The level of heritability and linkage affect the accuracy of estimated genetic group effects. When the heritability is low, genetic group effects are estimable with adequate accuracy at 25% linkage. The effect of heritability on the accuracy of genetic group effects appears to decrease with increasing linkage between flocks. The number of progeny per link sire has relatively less effect on the mean standard errors of genetic group solutions which are more a function of the total number of progeny from link sires. It is recommended that a minimum of 25% lambs should be born to sires from outside flocks (genetic groups with adequate numbers of recorded progeny; such sires also need to have progeny in other flocks) to provide adequate linkage among flocks

with varying phenotypic means and genetic groups be fitted in the evaluation model to accurately evaluate animals across flocks.

This study considered a simple data structure involving only one generation and genetic groups were defined on the basis of flock of origin. However, the impact of genetic grouping may be different if multiple generations are involved. Further research could explore different methods for grouping animals when the data span several generations. Additionally, the data structure simulated had a balanced design (same number of animals in each flock) and linkage across groups was established through the same sires which may not be realistic to adopt in an industry with diverse breeding objectives and where animals are linked to a large number of base populations and subpopulations. Therefore, the results obtained may need certain modifications prior to their actual implementation, that is, linkage required may be less or more than 25% depending on the structure of the data and the actual difference in the genetic means of different base populations.

Previous simulation studies by Laloe and Phocas (2003) have shown that for balanced sire designs (same sires used in each year with same progeny number per sire) where common sires are used for establishing genetic links across herds, the coefficient of determination (CD; higher value for CD implies less bias; used as a measure of accuracy of estimation in this study) is a perfect indicator of the potential bias associated with evaluation of animals across herds. Similar findings were reported by Kuehn *et al.* (2007). Therefore, the results obtained in this study which also focused on evaluating animals across genetically linked flocks are expected to be reasonably accurate. The design of data simulated in this study was such that the criterion proposed by Kennedy (1981) could be used to determine whether to include or exclude genetic groups in the model. However, the approach of Kennedy (1981) cannot be applied to ascertain the need for grouping in large scale industry evaluations such as in the Australian Merinos where large numbers of genetic groups are assigned.

The accuracy of estimation of genetic group effects when treated as fixed depends on the group size, that is, the number of animals with records in each genetic group (Kuehn, 2005; Kuehn *et al.*, 2007; Hickey *et al.*, 2008). If many small groups are

assigned, the error associated with the estimation of breeding values may be increased (Kuehn *et al.*, 2007). There is a large variation in the size of flocks and genetic groups evaluated in the Australian Merino industry. Therefore, if a large number of smaller groups are formed with relatively small effects, the genetic group solutions are likely to be inaccurate when treated as fixed depending on the size of genetic groups that are estimated simultaneously in the mixed model equations. Additionally, it is important to note that besides flocks, genetic groups were treated as additional fixed effects in the model for the two scenarios tested in this study. If a large proportion of animals in a dataset have missing records on multiple-traits, a characteristic of Merinos, treating genetic groups as fixed will affect the accuracy of genetic group solutions. Moreover, considering differences in the size of flocks, the level of linkage between flocks, and the differences in the level of heritability of different traits, treating genetic groups as random effects could be tested as an alternative option.

As genetic group effects are the accumulated effects resulting from past selection in genetically diverse base populations and subpopulations such as in Australian Merinos, it could be assumed that genetic group effects follow a random distribution, thereby implying that they may have an associated variance. However, the choice of the best model (fixed versus random groups) is dependent on the size of the analyses and the knowledge of the pedigree (Fouilloux *et al.*, 2008) which always vary between datasets. Nevertheless, the impact of treating genetic groups as random effects on the accuracy of evaluation procedure merits further research. The next chapter will investigate the effect of flock (or genetic group) size on the accuracy of estimated genetic group effects and genetic group variance.

Chapter 4. The effect of flock size on the accuracy of estimated genetic group effects and genetic group variance

4.1 Introduction

The performance data recorded in Australian Merino sheep industry are characterized by varying flock sizes with significant differences in flock means. Therefore, to account for these variations between flocks, a large number of genetic groups are required in routine Sheep Genetics evaluations (Brown *et al.*, 2007). Genetic groups could be incorporated in the evaluation model as fixed or random effects. When genetic groups are treated as fixed effects, an implicit assumption is that groups are genetically and statistically independent of one another (Sullivan and Schaeffer, 1994). Genetic groups may have animals which are genetically related although they may have differences in performance. For example, groups within flocks or segregated lines developed as a result of selection over a time period. The across-flock EBVs will be expressed as a deviation from the average of all genetic group effects. This could affect the accuracy of across-flock EBVs if significant differences exist between strains/flocks that are not related genetically. Additionally, for smaller flocks, genetic group solutions when fitted as fixed effects will have large standard errors. Thus, considering these factors, the routine Sheep Genetics evaluations are undertaken using a random groups model in OVIS (Brown *et al.*, 2000).

Assuming genetic groups to be random effects requires prior knowledge regarding their distributions and covariances. Depending on the number of flocks being evaluated and their respective means, genetic groups formed could vary in number considerably. Therefore, genetic groups may not be included in the model as fixed effects which would be useful especially when few genetic groups are assigned (Ashtiani and James, 1993). Furthermore, the Australian Merino industry comprises a large number of strains and bloodlines with links to base populations and

subpopulations that presumably differ in their mean genetic merit. Therefore, the assumption of treating their mean breeding values as random effects could be regarded as a reasonable option (Ashtiani and James, 1993), especially taking into consideration the wide variation across flocks. Moreover, treating genetic groups as random has greater advantage over treating as fixed for low heritability traits (Sullivan and Schaeffer, 1994).

Most studies in the past have examined the effect of including fixed genetic group effects on the accuracy of evaluation procedure (Foulley *et al.*, 1990a). Sullivan and Schaeffer (1994) stressed the need for further studies involving varying group sizes and random genetic groups. Cardoso and Tempelman (2004) suggested treating genetic groups as random effects in conjunction with connectedness to avoid confounding with other fixed effects. Quaas (1988) advocated that forming several genetic groups with a few animals within each group should be avoided implying that the size of genetic groups (number of animals in each group) is an important factor that would affect the accuracy of prediction. Foulley *et al.* (1992) suggested assuming random genetic groups when the size of groups is small. In a simulation study, Ugarte *et al.* (1992) examined the use of fixed versus random contemporary groups (CG) in genetic evaluations. They found consistent advantages of random CG effects over fixed in terms of accuracy of EBVs especially when the size of CG is small. Although CG cannot be directly compared with genetic groups, however in statistical sense, the number of animals within each is likely to impact the accuracy of prediction. Huisman and Brown (2007) evaluated Merino industry data on worm egg count (WEC) from Sheep Genetics to identify the best grouping strategy. They reported that the estimated genetic group variance varied across different grouping strategies.

Chapter 3 focused on the level of linkage needed for an accurate partitioning of flock genetic and flock environmental effects when genetic groups were treated as fixed effects in the model. The focus of this study will be to examine the effect of flock (or genetic group) size on the accuracy of estimation of genetic group effects and genetic group variance.

4.2 Materials and Methods

The data structure in this study was simulated different to that in Scenario II of Chapter 3 where only four flocks were considered with two each in a genetic group. The aim was to investigate the accuracy of estimation of a large number of genetic group effects which better reflects that observed in the Australian Merino industry (for example, Brown *et al.*, 2007). Performance data and associated pedigree for a single trait were simulated for 50 flocks with 100 animals each per flock from a single generation. The progeny were obtained from two different types of sires which are described as follows:

1. **WNsires**: internal or within-flock sires of each flock and used only in their own flocks.
2. **Xsires**: external sires used in all flocks but derived from a different flock.

The pedigree of all sires was assumed to be unknown. The aim of this simulation was to create a data structure resembling field situation for examining the variation in the mean EBVs of 50 flocks at varying levels of linkage using models with genetic groups fitted as random effects. Two flocks each were randomly assigned to a genetic group. For each genetic group, a group effect was randomly simulated from a normal distribution with zero mean and variance (σ_{gg}^2) the same as the additive genetic variance (σ_a^2), which was 1.5 micron². The value for genetic group variance was chosen based on previous research in Sheep Genetics data (Brown, Personal communication). The covariance between genetic groups was simulated to be zero. Thus, there were 25 genetic groups with different effects and a 26th genetic group for the Xsires.

Each sire was randomly mated to an unknown dam to generate 10 progeny each. The pedigree of all dams was assumed unknown. The simulated TBVs of all sires and dams included the random genetic group effect plus the additive genetic effect. All the dams were assumed to be indigenous, that is, belonging to the same flock in which progeny were generated. The TBVs of progeny animals in each flock were simulated as:

$$TBV_i = \frac{1}{2} (TBV_{sire} + TBV_{dam}) + \Phi$$

where TBV_i is the TBV of the i th animal, TBV_{sire} and TBV_{dam} are the TBVs of the sire and the dam, respectively of the i th animal, and Φ is a factor for the Mendelian sampling effect which was simulated with mean zero and half of the additive genetic variance. The TBVs of all sires and dams were sampled from a normal distribution with a zero mean and additive genetic variance (σ_a^2) of 1.5 micron². For each progeny, a phenotypic observation (P_i) on a single trait was simulated around a global mean (μ) of 19 micron as:

$$P_i = \mu + TBV_i + F_j + e_{ij}$$

where F_j and e_{ij} are the fixed flock and random residual effects, respectively associated with each phenotypic record. The flock effects (treated as fixed) were sampled from a uniform distribution with effects ranging between 0 and 1 micron. The residual effects were drawn from a normal distribution with zero mean and residual variance (σ_e^2) of 2.25 micron². Therefore, the heritability (h^2) value was 0.40.

The level of linkage across the 50 flocks was varied over 5 levels from 10, 20, 30, 40, to 50%. These values refer to the proportion of progeny produced from Xsires in each flock. Flock sizes of 200, 300 and 400 animals were also examined. Additionally, further testing was performed by generating varying progeny per sire from a constant number of Xsires. This scenario was further examined such that each flock was treated as a separate genetic group, therefore there were 51 genetic groups with the 51st group that of Xsires. Thus, there were less animals allocated to a group than when 26 groups were formed. The aim for simulating 26 versus 51 genetic groups was to examine the effect of the size of genetic groups on the accuracy of their solutions.

Varying the number of animals in each flock

Data for this scenario were simulated as described above except that different numbers of animals were simulated in each flock using a uniform distribution so as to resemble a field scenario where flocks vary in their size. The size of simulated flocks ranged from 150 to 1000 animals. Flocks for this scenario were categorized into three classes; small (150 to 350 animals), medium (350 to 650 animals) and large (650 to 1000 animals). As individual flocks may vary widely in their genetic means, each flock was treated as an individual genetic group. Therefore, the number of genetic groups was

51 with the 51st genetic group that of Xsires. Since varying numbers of animals per flock were simulated, each flock had a different proportion of link progeny derived from Xsires. The number of Xsires used was 2 and the progeny generated from each sire per flock was varied from 1, 2, 5, 10, 25, and 50. Therefore, the results presented for this scenario are at varying levels of link progeny generated from Xsires rather than at different proportions of linkage in terms of percentage.

4.2.1 Genetic analysis

Breeding values, genetic group effects and genetic group variance (σ_{gg}^2) were estimated from each replicate in univariate analyses using an animal model in the ASReml package (Gilmour *et al.*, 2006). The model fitted was:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of the simulated phenotypic observations, $\mathbf{b}$ is a vector of fixed (flock) effects, $\mathbf{u}$ is a vector of random additive genetic effects of the animals, $\mathbf{g}$ is a vector of random genetic group effects, $\mathbf{e}$ is a vector of random residuals; $\mathbf{Q}$ is a design matrix describing the proportion of genes contributed by each genetic group to the animals with records; and $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices associating observations in $\mathbf{y}$ with the fixed flock and random animal (and genetic group effects in $\mathbf{g}$) in $\mathbf{b}$ and $\mathbf{u}$, respectively.

Genetic groups were defined on the basis of flock and were fitted in the model as random effects. The covariance among genetic groups was assumed zero. The starting values for genetic and genetic group variances were assumed to be known. The approach of Robinson (1986), Westell *et al.* (1988) and Quaas (1988) was followed to derive the coefficients (ranging from 0 to 1) representing the proportion of genes derived by animals with records from different genetic groups. The accuracy of prediction was determined using the correlation between the simulated TBVs and the EBVs. The accuracy of estimation was also measured in terms of the coefficient of determination or R^2 and the variance of difference between simulated and estimated genetic group effects. Each analysis was repeated at varying levels of linkage established through Xsires.

Tests of significance for the accuracy of estimated genetic group effects and breeding values

A paired t-test (averaged over all levels of linkage) was used to determine whether the difference in the accuracy (measured in terms of variance of difference) when estimating 26 groups versus 51 groups was statistically significant at 5% ($P < 0.05$) level. The variance of difference for 26 groups was subtracted from that with 51 groups at each level of linkage and flock size. The t-statistic for each pairwise (26 versus 51 groups) comparison was calculated using the following formula:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{Y}_1 - \bar{Y}_2}} \text{ and } S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\bar{Y}_1$, $\bar{Y}_2$, s_1^2 , s_2^2 , n_1 and n_2 denote the means (averaged over different levels of linkage), variances, and number of observations for the variance of difference (between simulated and estimated genetic group effects) when 51 and 26 genetic groups, respectively are fitted in the model, and $S_{\bar{Y}_1 - \bar{Y}_2}$ is the standard deviation of the difference in the means. The computed t-value was required to be greater than the tabulated value of 2.77 at 4 degrees of freedom to be significant at 5% level.

The test (averaged over linkage) described above was repeated to investigate whether the difference in correlation (between TBVs and EBVs) for 26 versus 51 group effects was statistically significant at 5% level. Thus, $\bar{Y}_1$, $\bar{Y}_2$, s_1^2 , s_2^2 , n_1 and n_2 denote the means (averaged over different levels of linkage), variances, and number of observations for the correlation when 26 and 51 group effects, respectively are included in the model. The difference in correlation was obtained by subtracting the correlation with 51 groups from that derived using 26 groups at each level of linkage and flock size. The estimated t-value was required to be greater than the tabulated value of 2.77 at 4 degrees of freedom [(5 – 1) levels of linkage] to be significant at 5% level.

For each number (26 and 51) of genetic groups fitted in the model, the paired t-test was used to examine whether the difference in the variance (of difference between simulated and estimated genetic group effects) averaged over linkage was significant for various flock size combinations. The comparison of calculated and tabulated t-values

was the same as described above. Similarly, for each number (26 and 51) of groups, the test was used to explore if the difference in correlation (between TBVs and EBVs) averaged over flock size was significant for different combinations of linkage. For this test, the calculated t-value was needed to be greater than the tabulated value of 3.182 at 3 degrees of freedom [(4-1) levels of flock size] to be significant at 5% level.

4.3 Results and Discussion

All results unless otherwise stated are presented at varying proportions of linkage (10 to 50%) and averaged over 100 replicates. The R^2 between simulated and estimated genetic group effects improved with increasing flock size and reached close to a maximum value at approximately 25% linkage for different flock sizes (Figure 4.1). A further increase in linkage ($> 25\%$) did not significantly improve R^2 . This shows that genetic group effects can be estimated accurately with a minimum of 25% linkage. These results imply that the accuracy of estimating genetic group effects using a random groups model improves with an increase in the flock size. The difference in R^2 for flock sizes of 100 and 200 is greater than the difference in R^2 for flock sizes of 300 and 400 implying that a further increase in flock size (> 400) will not improve R^2 much. The results also indicate that group effects cannot be estimated precisely for smaller flocks (< 200). The accuracy of prediction is similar for flock sizes of 300 and 400 (Figure 4.1). Hickey *et al.* (2008) evaluated accuracy and bias in genetic evaluations with genetic groups using industry data involving several beef breeds. They concluded that the accuracy of genetic group solutions improved with an increase in the numbers of records within each genetic group. The accuracy of genetic group solutions in a multi-trait evaluation will impact the accuracy of multi-trait EBVs which in turn will affect the selection response. Hickey *et al.* (2008) stated that the response to selection in a breeding program involving multiple genetic groups is affected by both the accuracy of EBVs within and across genetic groups.

In another study, Kuehn *et al.* (2007) concluded that the PEV of across-flock EBVs is increased if many small genetic groups are formed. Similar findings were reported by Cottle and James (1997) who evaluated the impact of linkage on the accuracy of sire evaluation in CTSE (Central Test Sire Evaluation) schemes in Australia. They stated that when small numbers of genetic groups are formed, a larger

proportion of sires are compared within groups and therefore the errors associated with estimating genetic group means do not affect the PEV of EBVs. The authors also argued that when smaller groups are formed, the accuracy of prediction of genetic group means improves due to a larger number of observations per group and also because of superior genetic connections between groups.

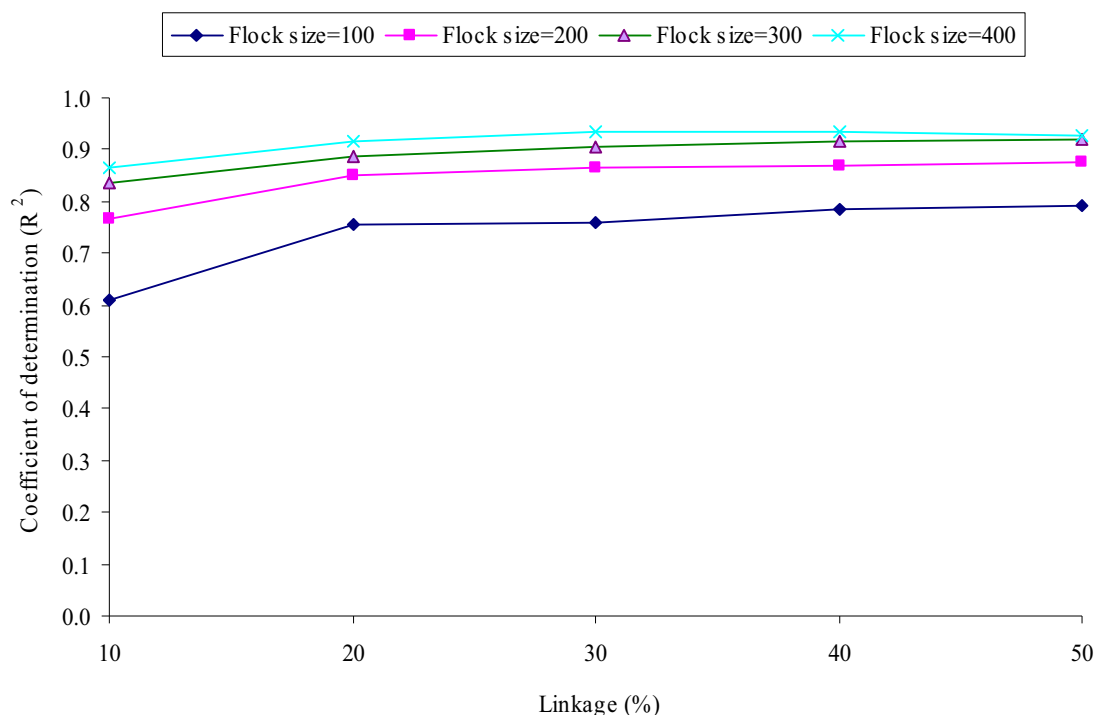


Figure 4.1. Coefficient of determination between simulated and estimated genetic group effects for different flock sizes at varying levels of linkage (26 genetic groups)

At all levels of linkage and for flock sizes investigated in this study, the estimated genetic group variance was not significantly different to that simulated (Table 4.1). However, for large flocks, the standard errors were slightly smaller but not significantly different from those for small flocks. This shows that the accuracy of estimation of genetic group variance is slightly higher for large flocks examined in this study. The variance of difference between simulated and estimated genetic group effects decreased with an increase in flock size and/or linkage (Table 4.2). The results presented in Table 4.2 indicate that at any given linkage, there is a significant effect of flock size on the accuracy of genetic group solutions. The variance of the difference decreases with an increase in flock size at any level of linkage resulting in higher

accuracy of prediction for large flocks. The accuracy of estimation of genetic group variance is expected to improve with a further increase in flock size.

Table 4.1. Estimated genetic group variance ($\pm$ SE; micron²) for different flock sizes at varying levels of linkage (26 genetic groups)

Linkage (%)	<u>Flock size</u>			
	100	200	300	400
10	1.39 (0.06)	1.51 (0.06)	1.46 (0.05)	1.45 (0.04)
20	1.60 (0.06)	1.53 (0.04)	1.47 (0.04)	1.51 (0.04)
30	1.49 (0.05)	1.42 (0.05)	1.44 (0.04)	1.51 (0.05)
40	1.56 (0.06)	1.52 (0.05)	1.56 (0.05)	1.55 (0.05)
50	1.51 (0.05)	1.52 (0.05)	1.56 (0.05)	1.49 (0.04)

Table 4.2. Variance of the difference (micron²) between simulated and estimated genetic group effects for different flock sizes at varying levels of linkage (26 genetic groups)

Linkage (%)	<u>Flock size</u>			
	100	200	300	400
10	0.56	0.33	0.24	0.19
20	0.38	0.23	0.16	0.12
30	0.36	0.19	0.14	0.10
40	0.32	0.20	0.12	0.09
50	0.32	0.18	0.12	0.10

The across-flock correlation between TBVs and EBVs is a mixture of within-flock and between-flock correlations dominated by the within-flock relationships. The across-flock correlation improves with increasing flock size and linkage (Table 4.3). However, the effect is not large. There was a gradual but not significant decline in the correlation with increasing linkage above 20% (Table 4.3). This is due to the fact that at higher levels of linkage, genetic group solutions are less accurate because of less

expressions of animals from the 25 flock-based genetic groups and more from the 26th genetic group. This results in the EBVs which are less accurate thereby leading to a decline in the correlation (Table 4.3). The decline in across-flock correlation is also attributed to a reduction in the variance of TBVs with increasing proportion of linkage between flocks. It is important to note that these correlations (Table 4.3) are across-flock correlations (for all animals across all flocks) and therefore, will be different from the within-flock correlations due to the variance across flocks (genetic group variance) as described in the previous chapter.

Table 4.3. Correlation between TBVs and EBVs for different flock sizes at varying levels of linkage (26 genetic groups)

Linkage (%)	<u>Flock size</u>			
	100	200	300	400
10	0.73	0.77	0.79	0.80
20	0.76	0.79	0.79	0.80
30	0.76	0.77	0.78	0.79
40	0.75	0.77	0.78	0.78
50	0.75	0.76	0.76	0.76

When varying numbers of progeny were generated from a constant number of Xsires, the results obtained were similar to those shown above and are therefore, not presented. Generating same number of progeny from varying number of link sires and vice versa has not affected the accuracy of prediction of genetic group effects and across-flock breeding values. The data simulated in this study had a balanced design with respect to the number of animals per flock and the use of common link sires across flocks for establishing linkage. In a simulation study characterized by a balanced design of the data and the use of link sires, Laloe and Phocas (2003) showed that the accuracy of estimation of a contrast (measured in terms of R^2) between the genetic means of two herds is a function of the number of progeny generated per sire, total proportion of progeny derived from common link sires, and the heritability of the trait. Therefore, generating same number of progeny from varying number of link sires and vice versa is

expected to provide similar results. However, when flocks have varying size and linkage is created through different sires and varying proportions of progeny are generated from different link sires in different flocks, the results are likely to differ from those obtained in this study. Additionally, in this study, the heritability was simulated to be valid for within-flock evaluation only and no account of genetic group variance was taken into consideration, which is unlikely to be true for industry datasets because of the variation in selection history and breeding objectives. Hence, further studies could examine the impact of random groups model on the accuracy of evaluation scheme when heritability estimates vary depending on the genetic group variance across flocks.

When each flock was treated as a separate genetic group and varying numbers of progeny were generated from fixed number (10) of Xsires, the R^2 between simulated and estimated genetic group effects decreased (Figure 4.2 versus Figure 4.1) at comparable levels of linkage. This is due to the difference in the number of genetic group effects estimated (26 versus 51 genetic group effects). The accuracy of prediction was higher for 26 than for 51 genetic group effects because standard error is a function of the number of observations (animals) per genetic group (group size). This difference in R^2 was evident at all levels of linkage. In both cases, R^2 reached a maximum value at 30% linkage. A further increase in linkage did not affect R^2 significantly (Figure 4.2). These results indicate that the precise estimation of genetic group effects is dependent on the group size (number of animals per genetic group). It is also evident from Figure 4.2 that for a flock (or genetic group) size of 400 animals each, the accuracy of estimation was more than for flock sizes of 100, 200 or 300. However, the R^2 will only keep increasing minutely with the size of genetic groups. With small numbers of animals in a group, the genetic group effects were still estimable, however, the accuracy of estimation was slightly reduced (Figure 4.2).

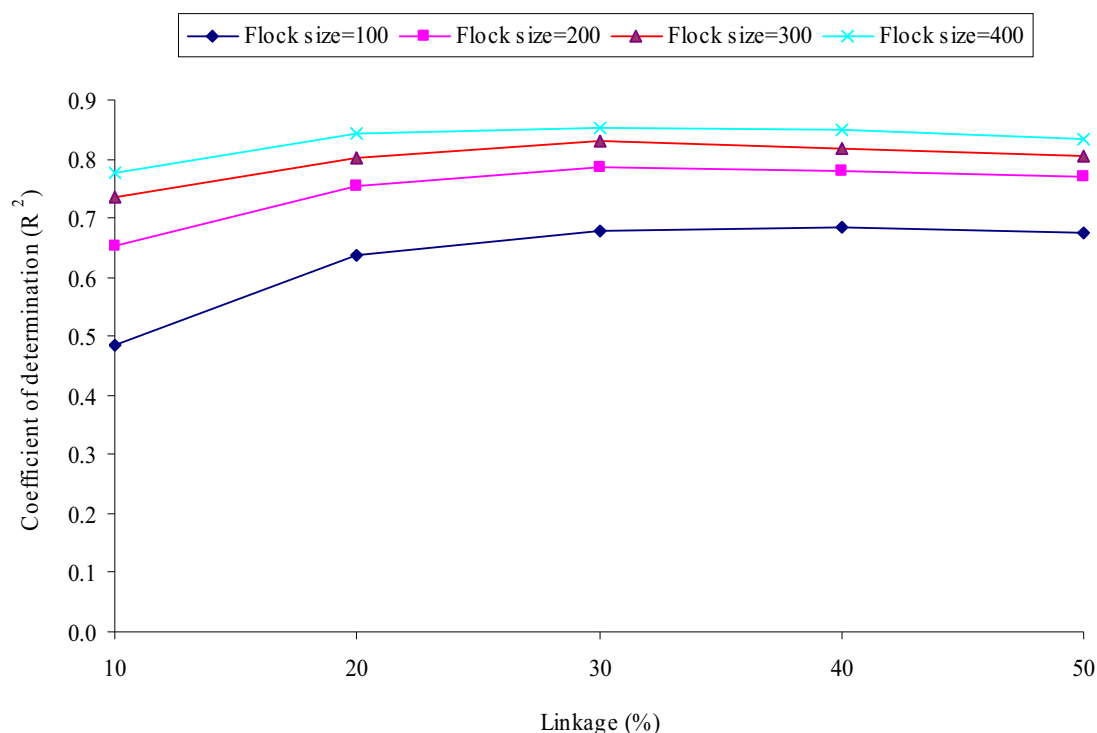


Figure 4.2. Coefficient of determination between simulated and estimated genetic group effects for different flock sizes at varying levels of linkage (51 genetic groups)

When each flock was assumed to be a separate genetic group (51 groups), the estimated genetic group variance was slightly lower than the simulated variance for all levels of linkage and flock sizes (Table 4.4). This is attributed to a reduction in the accuracy of genetic group solutions when smaller groups are estimated. The estimates for genetic group variance are consistently lower for all flock sizes and proportions of linkage because of the reduction in the between-group variance with an increase in the number of genetic groups from 26 to 51. The variance of difference between simulated and estimated genetic group effects decreased with increasing linkage (Table 4.5) as more progeny from common Xsires improved the accuracy of prediction of genetic group effects. However, when 51 groups were fitted, the variance of difference was higher than that shown in Table 4.2 (26 genetic groups) due to the difference in the size of genetic groups. The variance of difference is also dependent on flock size (Table 4.5). As the flock size and the level of linkage increase, the variance of difference between simulated and estimated genetic group effects decreases. This again shows that

the accuracy of prediction of genetic group effects is higher for large than for smaller flocks.

Table 4.4. Estimated genetic group variance ($\pm$ SE; micron²) for different flock sizes at varying levels of linkage (51 genetic groups)

Linkage (%)	<u>Flock size</u>			
	100	200	300	400
10	1.41 (0.07)	1.43 (0.04)	1.43 (0.04)	1.45 (0.04)
20	1.45 (0.04)	1.43 (0.04)	1.41 (0.03)	1.47 (0.03)
30	1.42 (0.05)	1.45 (0.03)	1.46 (0.03)	1.46 (0.04)
40	1.45 (0.04)	1.41 (0.04)	1.36 (0.03)	1.40 (0.03)
50	1.38 (0.04)	1.51 (0.04)	1.40 (0.04)	1.47 (0.03)

Table 4.5. Variance of difference (micron²) between simulated and estimated genetic group effects for different flock sizes at varying levels of linkage (51 genetic groups)

Linkage (%)	<u>Flock size</u>			
	100	200	300	400
10	0.78	0.53	0.39	0.33
20	0.57	0.37	0.29	0.24
30	0.48	0.33	0.25	0.22
40	0.48	0.33	0.26	0.22
50	0.48	0.36	0.28	0.25

When each flock was treated as an individual genetic group (51 genetic groups), the correlation between TBVs and EBVs decreased (Table 4.6) compared to that shown in Table 4.3. This is due to the fact that when 51 genetic group effects were estimated, there was a reduction in the accuracy of prediction of genetic group effects (Table 4.6) as less numbers of animals were allocated to each genetic group. This also implies that the accuracy of prediction of group effects improves with an increase in the group size.

The correlation is highest for flock (or genetic group) size of 400 at all levels of linkage. At low levels of linkage, the correlation improved slightly with increasing flock size. However, with increasing linkage, the effect of flock size on correlation appeared to diminish (Table 4.6). This can be attributed to the fact that when a certain proportion of progeny in small or large flocks is derived from common link sires, there is a reduction in the variance of TBVs and a simultaneous reduction in the variance of EBVs thereby affecting the correlation between the two. However, at low levels of linkage, genetic group solutions for smaller flocks are less accurate which affects the accuracy of EBVs and hence the correlation between TBVs and EBVs is small. For large flocks, the accuracy of genetic group solutions is always higher than that for smaller flocks even at low levels of linkage.

Table 4.6. Correlation between TBVs and EBVs for different flock sizes at varying levels of linkage (51 genetic groups)

Linkage (%)	Flock size			
	100	200	300	400
10	0.59	0.70	0.74	0.76
20	0.69	0.75	0.76	0.78
30	0.71	0.76	0.77	0.78
40	0.72	0.75	0.76	0.77
50	0.71	0.75	0.75	0.76

Test of significance (Difference in the variance of difference between simulated and estimated genetic group effects)

For each flock size and proportion of linkage, the variance of difference with 26 groups was subtracted from that with 51 groups (Table 4.7). Averaged over levels of linkage and flock sizes, the difference in the variance (of difference between simulated and estimated genetic group effects) was not statistically significant ($P > 0.05$) for all levels (Table 4.7). However, when 26 groups were estimated, the difference was significant for a number of flock size combinations (Table 4.8). The difference was significant for flock sizes of 100 and 200 ($P < 0.05$). The difference for combinations of

flock sizes of 100 versus 300 and 100 versus 400 was highly significant (computed t-value > 4.604; $P < 0.01$; Table 4.8). When 51 groups were fitted in the model, the difference was highly significant for 100 versus 400. These results indicate that the accuracy of estimation of genetic group effects is higher for large flocks. For different combinations of linkage (averaged over flock size) and different group numbers (26 and 51), the difference was not significant (Table 4.9) although the trend was that with increasing linkage between flocks, the variance of difference decreased.

Table 4.7. Difference in the variance (of difference between simulated and estimated genetic group effects; micron²) for 26 versus 51 genetic groups for varying flock sizes and levels of linkage*

Linkage (%)	Difference at varying flock sizes			
	100	200	300	400
10	0.22 ^{ns}	0.20 ^{ns}	0.15 ^{ns}	0.14 ^{ns}
20	0.19 ^{ns}	0.14 ^{ns}	0.13 ^{ns}	0.12 ^{ns}
30	0.12 ^{ns}	0.14 ^{ns}	0.11 ^{ns}	0.12 ^{ns}
40	0.16 ^{ns}	0.13 ^{ns}	0.14 ^{ns}	0.13 ^{ns}
50	0.16 ^{ns}	0.18 ^{ns}	0.16 ^{ns}	0.15 ^{ns}

* Obtained by subtracting the variance of difference with 51 groups from that with 26 groups

^{ns} Not significant ($P > 0.05$)

Table 4.8. Difference in the variance (of difference; micron²) averaged over different levels of linkage for 51 genetic groups (above diagonal) and 26 genetic groups (below diagonal) for varying combinations of any two flock sizes^{1, ^}

Flock size	100	200	300	400
100		0.17 ^{ns}	0.26 [*]	0.31 ^{**}
200	0.16 [*]		0.09 ^{ns}	0.13 [*]
300	0.23 ^{**}	0.07 ^{ns}		0.04 ^{ns}
400	0.27 ^{**}	0.11 [*]	0.04 ^{ns}	

¹ Across different levels of linkage; [^] Presented values are obtained by subtracting the mean (averaged over different levels of linkage) of smaller from that of larger flock sizes; ^{ns} Not significant ($P > 0.05$); * $P < 0.05$; ** $P < 0.01$

Table 4.9. Difference in the variance (of difference; micron²) averaged over different flock sizes for 51 genetic groups (above diagonal) and 26 genetic groups (below diagonal) for varying combinations of two levels of linkage^{1, ^}

Linkage (%)	10	20	30	40	50
10		0.14 ^{ns}	0.19 ^{ns}	0.19 ^{ns}	0.17 ^{ns}
20	0.11 ^{ns}		0.05 ^{ns}	0.05 ^{ns}	0.03 ^{ns}
30	0.13 ^{ns}	0.03 ^{ns}		0.00 ^{ns}	0.02 ^{ns}
40	0.15 ^{ns}	0.04 ^{ns}	0.02 ^{ns}		0.02 ^{ns}
50	0.15 ^{ns}	0.04 ^{ns}	0.02 ^{ns}	0.00 ^{ns}	

¹ Across different flock sizes; ^ Presented values are obtained by subtracting the mean (averaged over different flock sizes) of lower from that of higher levels of linkage; ^{ns} Not significant (P > 0.05)

Test of significance (Difference in correlation between TBVs and EBVs)

The correlation between TBVs and EBVs with 51 groups was subtracted from that with 26 groups at each proportion of linkage and flock size (Table 4.10). The difference in correlation was not statistically significant for all flock sizes and proportions of linkage. The trend was that with increasing linkage and for larger flock sizes, the difference between the two correlations decreased (Table 4.10). However, only four degrees of freedom were available to make a comparison because of the way the data were simulated. Therefore, the conclusions drawn are likely to be different if the level of significance is measured at a particular level of linkage rather than computing the means averaged over different levels of linkage as was undertaken in this study. For example, in a field scenario, certain flocks may be adequately linked and others may have only limited genetic links. Moreover, some of the flocks may not be significantly superior to others in terms of average performance. Eady *et al.* (1991) evaluated industry data to compare different strains of Australian Merino sheep for a range of fleece and reproduction traits. They found that the differences between genetic groups were not always statistically significant although there was a trend to improved production for certain strains than for the others. When 26 groups were estimated, the difference in correlation (averaged over linkage) was not significant (P > 0.05) for all

flock size combinations (Table 4.11). However, the trend was that the difference in correlation decreased with an increase in the difference in the size of flocks. Similar results were obtained when 51 genetic groups were fitted (Table 4.11). The difference in correlation (averaged over flock sizes) was not significant for any combination of linkage when 26 and 51 groups were estimated (Table 4.12) although there was a trend for larger difference in correlation at very different proportions of linkage. For example, the difference was larger between 10 and 50% than between 40 and 50% levels of linkage (Figure 4.12).

Table 4.10. Difference in correlation (26 versus 51 genetic groups) between TBVs and EBVs for varying flock sizes and levels of linkage^{*}

Linkage (%)	Difference at varying flock sizes			
	100	200	300	400
10	0.14 ^{ns}	0.07 ^{ns}	0.05 ^{ns}	0.04 ^{ns}
20	0.07 ^{ns}	0.04 ^{ns}	0.03 ^{ns}	0.02 ^{ns}
30	0.05 ^{ns}	0.01 ^{ns}	0.01 ^{ns}	0.01 ^{ns}
40	0.03 ^{ns}	0.02 ^{ns}	0.02 ^{ns}	0.01 ^{ns}
50	0.04 ^{ns}	0.01 ^{ns}	0.01 ^{ns}	0.00 ^{ns}

^{*} Presented values are derived by subtracting correlation with 51 groups from that with 26 groups; ^{ns} Not significant ($P > 0.05$)

Table 4.11. Difference in correlation between TBVs and EBVs averaged over different levels of linkage for 51 genetic groups (above diagonal) and 26 genetic groups (below diagonal) for varying combinations of two flock sizes^{1, ^}

Flock size	100	200	300	400
100		0.06 ^{ns}	0.07 ^{ns}	0.09 ^{ns}
200	0.02 ^{ns}		0.01 ^{ns}	0.03 ^{ns}
300	0.03 ^{ns}	0.01 ^{ns}		0.01 ^{ns}
400	0.04 ^{ns}	0.01 ^{ns}	0.01 ^{ns}	

¹ Across different levels of linkage; [^] Presented values are obtained by subtracting the mean (averaged over different levels of linkage) of smaller from that of larger flock sizes; ^{ns} Not significant ($P > 0.05$)

Table 4.12. Difference in correlation between TBVs and EBVs averaged over different flock sizes for 51 genetic groups (above diagonal) and 26 genetic groups (below diagonal) for varying combinations of two levels of linkage^{1, ^}

Linkage (%)	10	20	30	40	50
10		0.05 ^{ns}	0.06 ^{ns}	0.05 ^{ns}	0.04 ^{ns}
20	0.01 ^{ns}		0.01 ^{ns}	0.00 ^{ns}	0.00 ^{ns}
30	0.00 ^{ns}	0.01 ^{ns}		0.01 ^{ns}	0.01 ^{ns}
40	0.00 ^{ns}	0.01 ^{ns}	0.01 ^{ns}		0.01 ^{ns}
50	0.01 ^{ns}	0.03 ^{ns}	0.02 ^{ns}	0.01 ^{ns}	

¹ Across different flock sizes; [^] Presented values are obtained by subtracting the mean (averaged over different flock sizes) of lower from that of higher levels of linkage; ^{ns} Not significant (P > 0.05)

Varying the number of animals in each flock

There was a significant reduction in the estimated genetic group variance when only one progeny per link sire was generated in each flock (Figure 4.3). However, increasing progeny per link sire led to an accurate estimation of the genetic group variance and the same trend continued up to 50 progeny per sire. This illustrates that genetic group variance can be estimated precisely at different levels of linkage even when flocks have varying sizes (Figure 4.3). The R^2 (results not presented) between simulated and estimated genetic group effects improved with increasing number of progeny per link sire. At a constant number of 100 progeny from two Xsires (50 progeny from each), there was a slight reduction in the standard errors of estimated genetic group effects with increasing flock size indicating that genetic group effects can be estimated with greater accuracy for large flocks (Figure 4.4). The standard errors were large for small flocks because of the high proportion of link progeny and also less progeny in total in those flocks thereby leading to less accurate genetic group solutions. On the other hand, the standard errors were small for large flocks because of less proportion of link progeny (compared to their size) and large size, which led to higher accuracy of genetic group solutions. However, the standard errors of genetic group solutions do not appear to be significantly different for varying flock sizes (Figure 4.4).

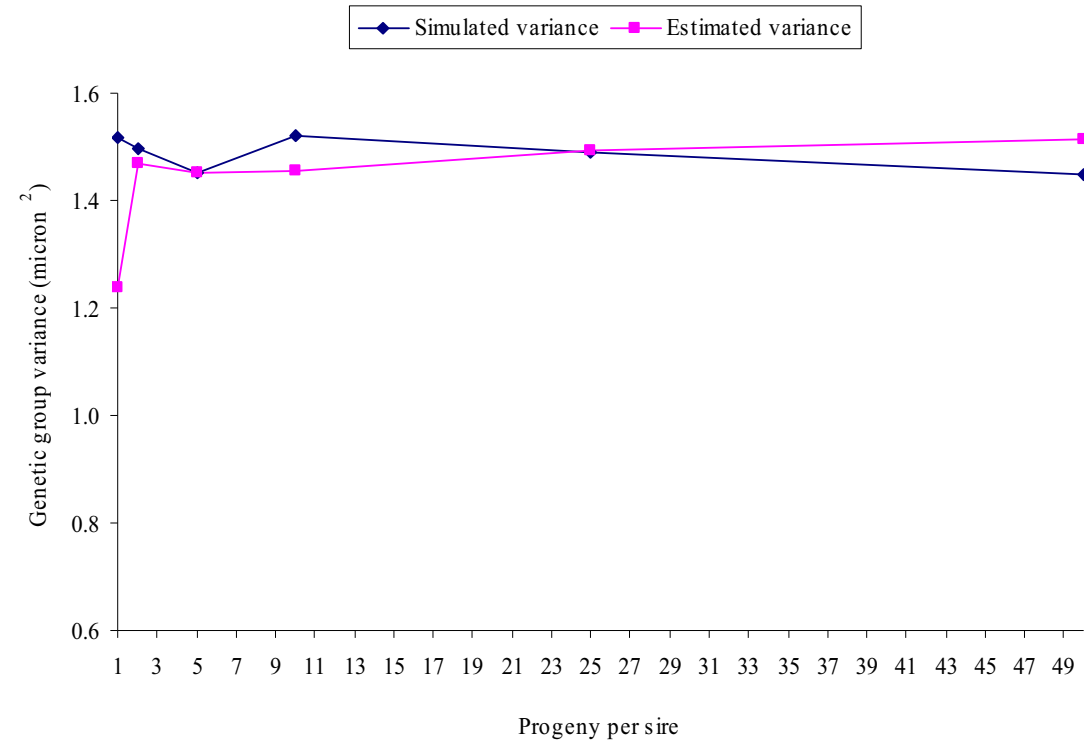


Figure 4.3. Simulated and estimated genetic group variance at different levels of progeny per link sire (51 genetic groups)

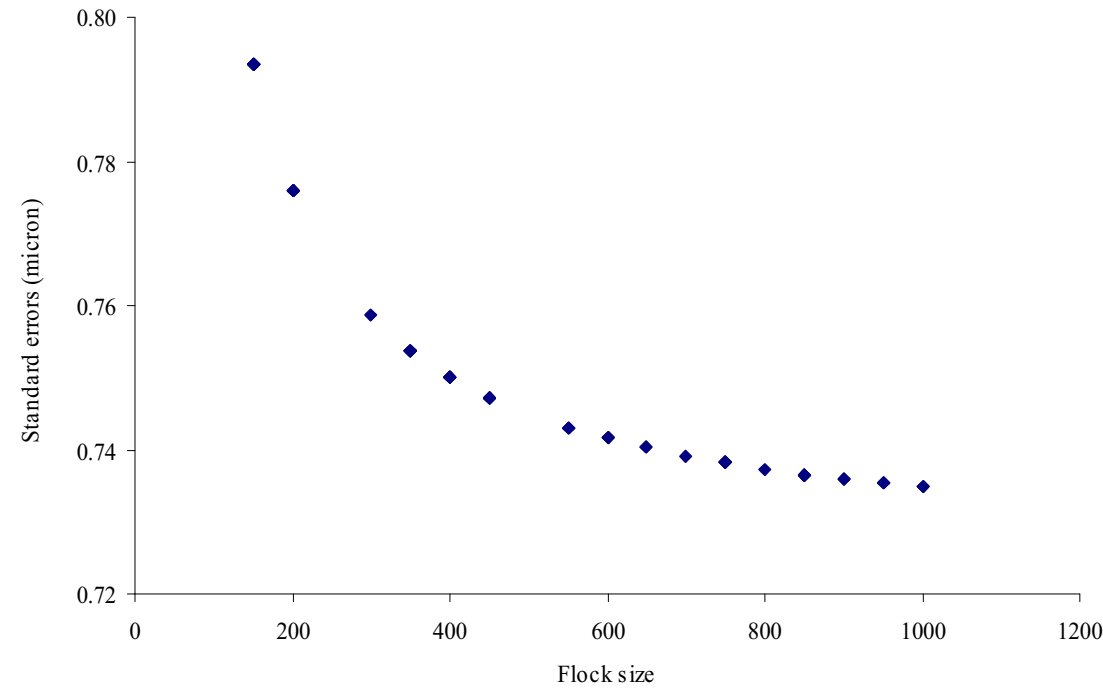


Figure 4.4. Standard errors of estimated genetic group effects for varying flock size (for single replicate at 50 progeny each from two Xsires)

The R^2 in Figure 4.5 started declining when more than 25 progeny per link sire were generated in small flocks. This is because a large proportion of link progeny increases the information about contrasts between different groups and simultaneously reduces the number of within group progeny (progeny of WNsires), therefore reducing the accuracy of estimating the genetic group effect of that flock. At the level of one progeny per sire, the genetic group effects and hence their standard errors were inestimable for some of the replicates. Therefore, the results in Figures 4.5 and 4.6 are not presented at one progeny per sire. The trend in Figure 4.6 shows that the effect of flock size on the accuracy of estimation increases in importance as the number of progeny from link sires increases. This implies that at very low numbers of progeny from link sires, the accuracy of prediction will be similar irrespective of the flock size (Figure 4.6). These results also show that the number of progeny required per link sire will be less for small flocks and vice-versa.

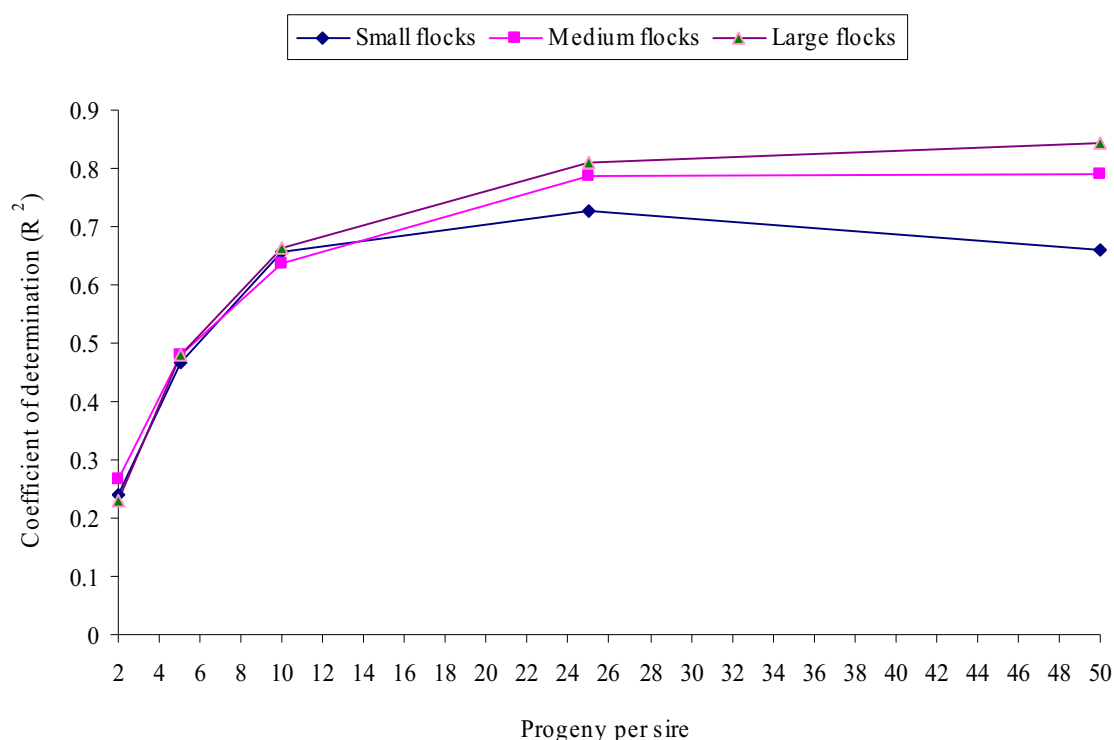


Figure 4.5. Coefficient of determination between simulated and estimated genetic group effects for small, medium and large flocks at varying levels of progeny per link sire

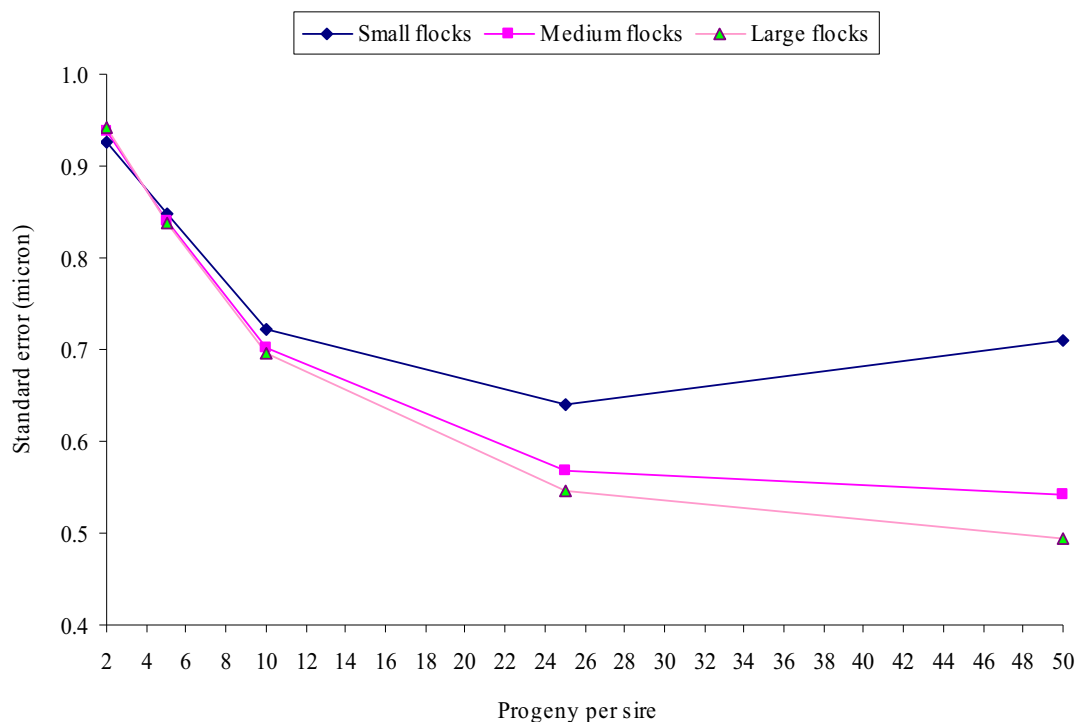


Figure 4.6. Standard errors of genetic group solutions for small, medium and large flocks at varying levels of progeny per link sire

The standard errors of estimated genetic group effects for small flocks (Figure 4.6) started increasing when more than 25 progeny per link sire were generated. This is attributed to the fact that small flocks get a large proportion of their progeny from the same two link sires (less home-bred expressions) thus, leading to less accurate estimates of genetic group solutions. However, the effect is small and non-significant. Kuehn (2005) stated that if genetic groups are included in the model as additional fixed effects, the accuracy of prediction would strongly depend on the number of animals within each group and the level of linkage. In Sheep Genetics Merino evaluations for WEC, each flock having a minimum of 100 animals with records is treated as an individual genetic group (Brown *et al.*, 2006).

In this study, linkage was established across flocks using the same link sires (Xsires) which is not realistic under field conditions due to differences in management practices (for example, lack of use of AI in some flocks) and breeding objectives which vary across studs. Therefore, the results presented and the conclusions drawn from these results may need certain modifications in terms of industry scenarios that generally

involve multiple generations of performance data, in which case, the influence of genetic grouping on the accuracy of prediction might be different. Therefore, the minimum amount of linkage required to estimate genetic group solutions and genetic group variance accurately may vary depending on these factors.

4.4 Conclusions and Implications

The accuracy of estimation of genetic group effects and their variance is dependent on the size of genetic groups and the level of linkage. However, treating genetic groups as random effects requires prior assumptions regarding (co)variances. In this study, genetic group variance was assumed to be the same as the additive genetic variance, which in reality may not be true. For example, breed differences are also genetic and genetic groups could be assigned on the basis of breed of origin as was undertaken by Sullivan *et al.* (1999) and Hickey *et al.* (2008). However, that does not imply that the within-breed variance would be the same as the between-breed variance. Therefore, it would be erroneous to make assumptions regarding prior distributions and co(variances) of genetic groups.

The genetic group variance is expected to depend on the selection history, which for most datasets cannot be traced back to the unselected base population. The genetic group variance could be estimated without making prior assumptions. Alternatively, the assumed variance due to genetic groups could be varied (different from genetic variance) in simulation studies and then estimated to examine whether it is reasonable to make such an assumption. However, this would depend on the design of data simulated and the structure of genetic groups, that is, the criteria used for defining groups. Huisman and Brown (2007) analyzed Merino data on WEC from Sheep Genetics and reported that the estimates of genetic group variance are dependent on the grouping strategy adopted. They found that the estimated additive genetic and residual variance were very similar across different grouping strategies tested. They also stated that different grouping strategies are required for different traits depending on the amount of data and selection history. If the assumption of same additive and genetic group variance is not correct then the EBVs obtained in a genetic groups model are likely to be inaccurate thereby affecting the ranking of animals across flocks.

The accuracy of estimated genetic group effects is dependent on the size of genetic groups. In other words, the accuracy of estimation of genetic group effects is dependent on the number of animals with effective records allocated to each group. On the basis of classification of flocks into small, medium and large, as considered in the present study, and the relative accuracy of estimation of their respective genetic group effects, it can be concluded that the size of genetic groups formed should be large for obtaining across-flock EBVs with accuracy. In this study, the difference in the accuracy (variance of difference) of estimation of genetic group effects and the difference in the correlation between TBVs and EBVs for 26 versus 51 groups was not statistically significant ($P > 0.05$). This could be due to the same value simulated for additive and genetic group variances. However, the accuracy of prediction was always higher for larger flock and genetic group sizes. Although not examined in this study in the strict sense, besides group size, the statistical power of estimation of genetic group effects will also be dependent on the magnitude of real genetic differences between groups (Woolaston, 1997). This implies that the power of estimation will also depend on the magnitude of real genetic group variance.

In the present study, linkage was established across genetic groups through the same link sires, which is not practical from an industry's perspective due to differences in management practices and breeding objectives. For example, breeders in a low rainfall area might not be interested in selecting for worm or fleece rot resistance. Therefore, the amount of linkage required for estimating genetic group effects and genetic group variance may vary if different link sires from different genetic groups are used. This study considered only a simple design with one generation of progeny with records. However, the impact of genetic grouping may be smaller or larger if multiple generations are involved, a likely feature of some Merino industry datasets with extended pedigree information. In addition, the within-flock or within-genetic group variance was simulated to be the same for all flocks, which may not be the case with industry data. However, modeling of these effects could be tedious due to the number of parameters to be estimated and the type of data structure required for such estimations (Hickey *et al.*, 2008).

Flocks with varying sizes participate in across-flock evaluations and therefore, establishing sufficient genetic links among flocks is not always possible due to the above mentioned reasons. Additionally, not all sheep breeders use sires from outside flocks to establish genetic links. Hence, in such situations, there is a potential problem of reduced accuracy of prediction of genetic group effects for a large number of small flocks with likely significant differences in their genetic means. Therefore, the need arises for pooling genetically similar and smaller flocks (with respect to their genetic group means) that have comparable group effects such that these problems are overcome. The next chapter will examine the feasibility of clustering flocks on the basis of their estimated genetic group effects so as to overcome the above-mentioned problem and simultaneously maximize the accuracy of evaluation procedure.

Chapter 5. The effect of cluster size on the accuracy of clustering mechanism in single-trait genetic evaluation

5.1 Introduction

Field data on fleece traits in the Australian Merino industry are collected across flocks exhibiting large phenotypic variation. Previous chapters have demonstrated that genetic group effects are estimable in across-flock evaluations when sufficient genetic links exist between flocks. However, generating a certain proportion of progeny in all flocks from common sires is not always practical due to limited use of outside sires by some breeders. Hence, flocks that use a large proportion of outside sires will be linked either directly or indirectly to other flocks whereas flocks that do not use outside sires will be completely unlinked. Additionally, genetic group effects are estimated with reduced accuracy for groups with small numbers of animals (Chapter 4). In order to account for the differences in the means, varying numbers of genetic groups need to be fitted in the evaluation model. However, for smaller flocks or those poorly linked, estimated genetic group effects are likely to be inaccurate due to large standard errors.

Flocks without adequate genetic links with other flocks could be pooled into one single group. Additionally, there is a possibility of several small flocks with similar genetic group means. Thus, smaller flocks could be pooled into the same group on the basis of their estimated genetic group solutions. This will reduce the number of genetic group effects that need to be estimated thereby improving the accuracy of prediction. Hence, clustering flocks on the basis of their estimated genetic group solutions could be explored to combine similar flocks into one large cluster or genetic group. The number of clusters required will vary with the data structure. Furthermore, pooling of several flocks into one large cluster will impact the within-cluster (variance within a cluster) and the between-cluster variance (variance between clusters) depending on the magnitude of estimated genetic group effects.

Several studies (Amar *et al.*, 1998; Weigel and Rekaya, 2000; Zwald *et al.*, 2003; Fouilloux *et al.*, 2008; Vasconcelos *et al.*, 2008) have explored the application of cluster analysis in genetic evaluation of livestock. The purpose for using cluster analysis ranged from investigating differences among genetically diverse groups of wool sheep for crossbreeding to condensing large numbers of connected genetic groups into compact clusters in order to estimate connectedness among herds of beef cattle. Vasconcelos *et al.* (2008) examined the impact of clustering herds in dairy cattle genetic evaluations by varying the minimal number of observations required in each contemporary group. They reported that clustering based on management similarities was a practical approach when inadequate numbers of animals exist in certain contemporary groups per level of fixed effects included in the model. This study will evaluate the effect of different cluster sizes on the accuracy of prediction when flocks are clustered based on their estimated genetic group effects. The effect of varying proportions of linkage and cluster sizes on the estimated between-cluster variance will be examined. In addition, the impact of varying cluster sizes on the accuracy of prediction of breeding values for small, medium and large flocks in an across-flock evaluation will also be investigated.

5.2 Materials and Methods

5.2.1 Data simulation

The Sheep Genetics Merino database contains performance and pedigree information from a large number of participating flocks which vary widely in their size (number of animals). The proportion of genetic links between flocks vary and therefore for certain flocks genetic group effects cannot be estimated precisely which is likely to impact the pattern and the numbers of clusters formed. Therefore, to resemble industry scenario, data were simulated in the same manner as in Chapter 4 (Varying the number of animals in each flock), briefly that is, 50 flocks with varying numbers of animals were simulated and the 51st flock was that of Xsires. Each flock was simulated to represent a single genetic group drawn from a normal distribution. The number of Xsires used was 2 and the progeny generated from each sire per flock was varied from 1, 2, 5, 10, 25, and 50. Since each flock had a different number of animals and varying

numbers of progeny were generated from the two Xsires, linkage varied across flocks. Therefore, the results shown are at varying numbers of progeny generated per link sire rather than in terms of linkage expressed as percentage. As in Chapter 4, the number of progeny generated per Xsire or home-bred sire of each flock was the same. Additionally, since the aim of this study was to evaluate the impact of linkage and cluster size on the estimated between-cluster variance, genetic group effects were simulated ($\sigma_{gg}^2 = 1.5 \text{ micron}^2$) and subsequent cluster effects (estimated genetic group effects from clustered flocks) were treated as random. The aim of generating varying numbers of progeny from Xsires was to investigate the impact of different proportions of linkage on the effectiveness of clustering. The size of simulated flocks ranged from 150 to 1000 animals which better reflects that observed in the Merino industry as described by Brown *et al.* (2007). Depending on the range from 150 to 1000 animals, flocks were categorized into small, medium and large as described in Chapter 4.

5.2.2 Genetic analysis I (Non-clustering)

Data were analyzed exactly in the same way as in Chapter 4. Thus, genetic group effects (treated as random) were estimated from simulated data in the ASReml package (Gilmour *et al.*, 2006). The animal model fitted was as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of the simulated records, $\mathbf{b}$ is a vector of fixed (flock) effects, $\mathbf{u}$ is a vector of random genetic effects of the animals, $\mathbf{g}$ is a vector of random genetic group effects, $\mathbf{e}$ is a vector of random residuals; $\mathbf{Q}$ is a design matrix describing the proportion of genes contributed by each genetic group to the animals with records; and $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices associating records in $\mathbf{y}$ with the fixed flock and random animal (and genetic group effects in $\mathbf{g}$) in $\mathbf{b}$ and $\mathbf{u}$, respectively.

5.2.3 Clustering flocks

Flocks were clustered on the basis of their estimated genetic group effects using the FASTCLUS procedure of SAS (1990). Briefly, clustering was based on genetic group solutions obtained from a mixed model analysis and only genetic group solutions

without their associated standard errors were used as an input for the FASTCLUS procedure. The clustering mechanism (*K*-means clustering) is based on Euclidean distances (straight line distance between any two points in a plane) estimated from either one or more quantitative variables. The FASTCLUS procedure finds initial clusters using a standard iterative algorithm which minimizes the sum of squared distances between different cluster means. Initially, a set of points known as cluster seeds is selected as the means of the clusters. Temporary clusters are formed by assigning each observation to the nearest seed. Subsequently, cluster seeds are replaced by the means of temporary clusters. This process is repeated until no further changes take place in the means of clusters formed. In a given dataset, observations that are close to one another (depending on the Euclidean distance from cluster mean) are generally assigned to the same cluster and those that are different from each other are allocated to different clusters. The above principle will be valid when flocks are clustered for a single trait only. However, when flocks are clustered jointly for multiple traits, the clustering mechanism is likely to be dominated by traits with a larger variance. To overcome this problem, the estimated genetic group effects for each trait need to be standardized (by expressing them in genetic standard deviations) to get an even influence of each trait on the clustering mechanism (refer to **Appendix** for more information).

In the computer program using the FASTCLUS procedure, either the Euclidean distance or the maximum number of clusters required for a given dataset could be specified. Therefore, if the Euclidean distance between any two data points is pre-specified, then the number of clusters required cannot be manipulated and the algorithm would form any number of clusters. In fact, the number of clusters formed may vary with each run as the mechanism is based on the initial random assignment of data points to cluster seeds that are later replaced by cluster means. However, when the number of clusters required is specified, then the Euclidean distance cannot be manipulated and is determined by the FASTCLUS procedure depending on the number of clusters. The objective of this study was to compare the effect of different numbers of clusters (cluster sizes) for 50 flocks on the accuracy of prediction. Therefore, the Euclidean distance between the genetic group means of different flocks was different for varying

cluster sizes. Additionally, it was examined whether clustering small, medium and large flocks leads to similar or different levels of accuracy of prediction. The underlying principles of K -clustering have been reviewed in detail under section 2.7 of Chapter 2.

5.2.4 Genetic analysis II (Clustering)

Cluster effects (genetic group effects from clustered flocks), between-cluster variance (variance between clusters; σ_{cl}^2) and breeding values for the simulated trait were estimated from each replicate in univariate analyses using the ASReml package (Gilmour *et al.*, 2006). In vector notation, the animal model fitted was:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{Q}\mathbf{c} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypic records, $\mathbf{b}$ is a vector of fixed flock effects, $\mathbf{u}$ is a vector of direct additive genetic effects of the animals, $\mathbf{c}$ is a vector of random cluster (genetic group) effects from clustered flocks, $\mathbf{e}$ is a vector of random residuals; $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices associating records in $\mathbf{y}$ with the fixed flock and random animal effects in $\mathbf{b}$ and $\mathbf{u}$, respectively; and $\mathbf{Q}$ is a matrix which describes the proportion of genes derived by an animal from different clusters (genetic groups). The diagrammatic illustration of the methodology used for the estimation of genetic group and subsequent cluster effects is depicted in Figure 5.1. The accuracy of prediction was estimated using the correlation between the simulated flock mean TBVs and the EBVs. Unless otherwise stated, all results presented are averaged over 100 replicates.

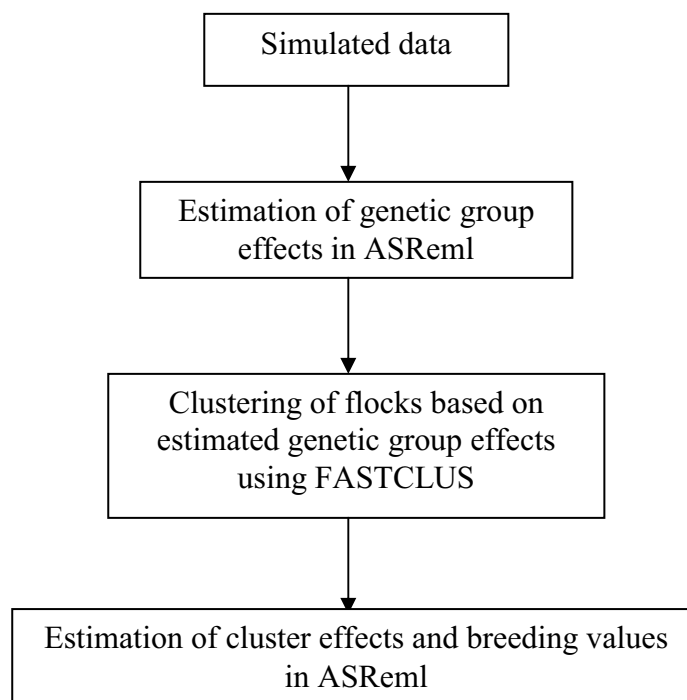


Figure 5.1. Flow diagram of the methodology used for the estimation of genetic group and cluster effects

5.3 Results and Discussion

At the level of one and two progeny per sire, the genetic group effects and hence their standard errors were inestimable for some of the replicates. Therefore, the results presented are at the level of 5, 10, 25, and 50 progeny per link sire. The between-cluster variance estimated using each cluster size varies depending on the number of clusters formed and the proportion of linkage (Figure 5.2). Except for 2 clusters, the fewer the number of clusters formed, the greater is the difference between simulated genetic group (σ_{gg}^2) and estimated cluster variance. When 2 clusters are formed with 25 flocks in each, the within-cluster variance (not shown) is larger than the between-cluster variance (Figure 5.2). This is due to the condensing of a large number of genetic groups into 2 clusters which results in an increased variance within clusters. However, with 5 clusters, fewer flocks are assigned to each cluster which results in smaller within-cluster than the between-cluster variance. The difference in estimated between-cluster variance with 5 and 25 clusters is due to the fact that when 5 clusters are formed, each cluster has more numbers or expressions of animals whereas when 25 clusters are formed for the same number of flocks, each cluster has less numbers of animals. Therefore, with

increasing numbers of clusters, the estimated between-cluster variance decreases (Figure 5.2).

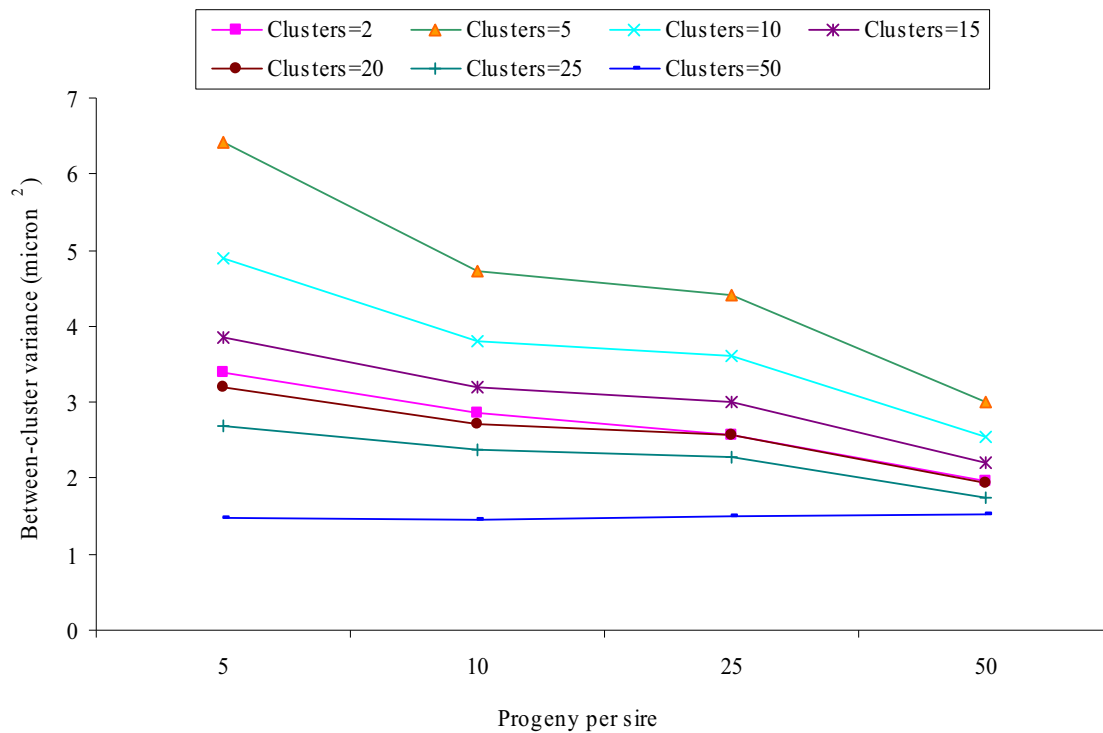


Figure 5.2. Estimated between-cluster variance for different cluster sizes at varying levels of progeny per link sire ($\sigma_{gg}^2 = 1.5 \text{ micron}^2$)

When the number of clusters formed is increased from 2 to 10, 15, 20, and 25, the trend observed for 2 clusters (within-cluster variance is higher than between-cluster variance) is reversed leading to a gradual reduction in the between-cluster variance (Figure 5.2). Therefore, the between-cluster variance is significantly smaller with 5, 10, 15, and 20 clusters. At low levels of progeny per link sire, the between-cluster variance is considerably higher than at higher levels. This is attributed to the fact that at higher levels of progeny per link sire or linkage, an increasing proportion of progeny in each flock is derived from the same common link sires and less proportion of progeny from home-bred sires thereby leading to a reduction in the variance of their TBVs and hence their EBVs. At all levels of progeny per link sire, the estimated between-cluster variance is the same as the simulated genetic group variance (1.5 micron^2) when 50 clusters are formed (Figure 5.2).

In the current study, genetic groups and clusters were treated as random effects which may not be suitable for industry evaluations due to the following reasons. Firstly, limited genetic links between flocks and small flock or group sizes may affect the accuracy of genetic group solutions thereby influencing the accuracy of cluster allocation. Secondly, as seen from the trend in Figure 5.2, clustering usually increases the between-cluster variance in comparison to the between-flock (or between-genetic group) variance. This would impact the choice of the covariance structure that needs to be fitted in the model for between-flock effects. The covariance structure required would depend on the pattern (number and size) of clusters formed which in turn would depend on the clustering procedure. The choice of the covariance structure required in the model for between-flock effects may also impact the variance of across-flock EBVs. Thus, it appears that treating clusters as random effects may not be the best option for industry evaluations. Hence it appears that fitting a model with fixed cluster effects may be a more appropriate alternative for industry evaluations.

There is a decline in the standard errors of estimated between-cluster variance with increasing progeny per sire and the total clusters formed (Figure 5.3). This is because with increasing proportion of link progeny in different flocks, the genetic group effects are estimated with greater accuracy. The decline in standard errors is steeper for 2 and 5 clusters when compared to other cluster sizes. The minimum amount of linkage required for accurately estimating the between-cluster variance would vary with the size of the flocks assigned to each cluster. The actual proportion of linkage required for the accurate estimation of between-cluster variance will also be dependent on the genetic group variance. Since each flock had a different number of animals and varying numbers of progeny were generated from Xsires, linkage varied across flocks. The linkage shown in Figure 5.4 is averaged over all flock sizes in each category, that is, small, medium and large. In other words, each flock had a different number of animals and the number of progeny generated per link sire was constant across flocks. Therefore, depending on the size, each flock had a different proportion of progeny derived from link sires. Besides flock size, the proportion of linkage required would also depend on the magnitude of genetic differences between various flocks or genetic groups.

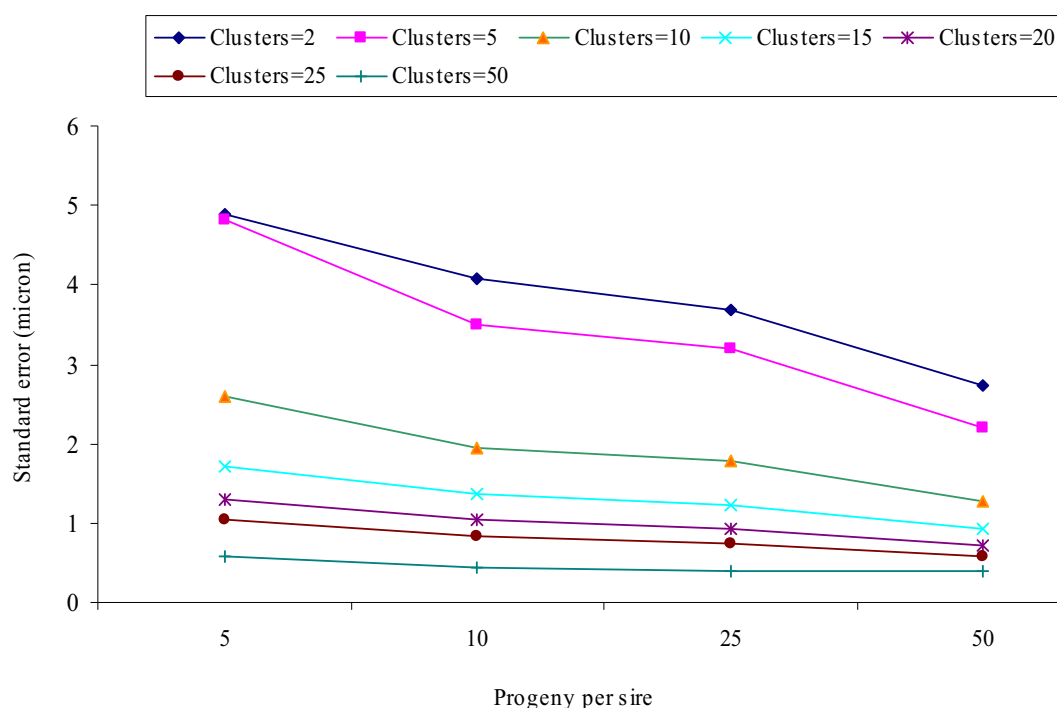


Figure 5.3. Standard errors of estimated between-cluster variance for different cluster sizes at varying levels of progeny per link sire

Although forming more clusters implies reduction in the number of animals per cluster, standard errors of estimated between-cluster variance become smaller with increasing numbers of clusters (Figure 5.3). This is due to the larger differences between genetic group and cluster means when fewer clusters are formed. The accuracies of estimated between-cluster variance for 2 and 5 clusters are significantly reduced at all levels of progeny per sire (Figure 5.3). This is due to the fact that condensing 51 genetic groups into 2 clusters leads to a significant decrease in the between-cluster variance and an equivalent increase in the within-cluster variance.

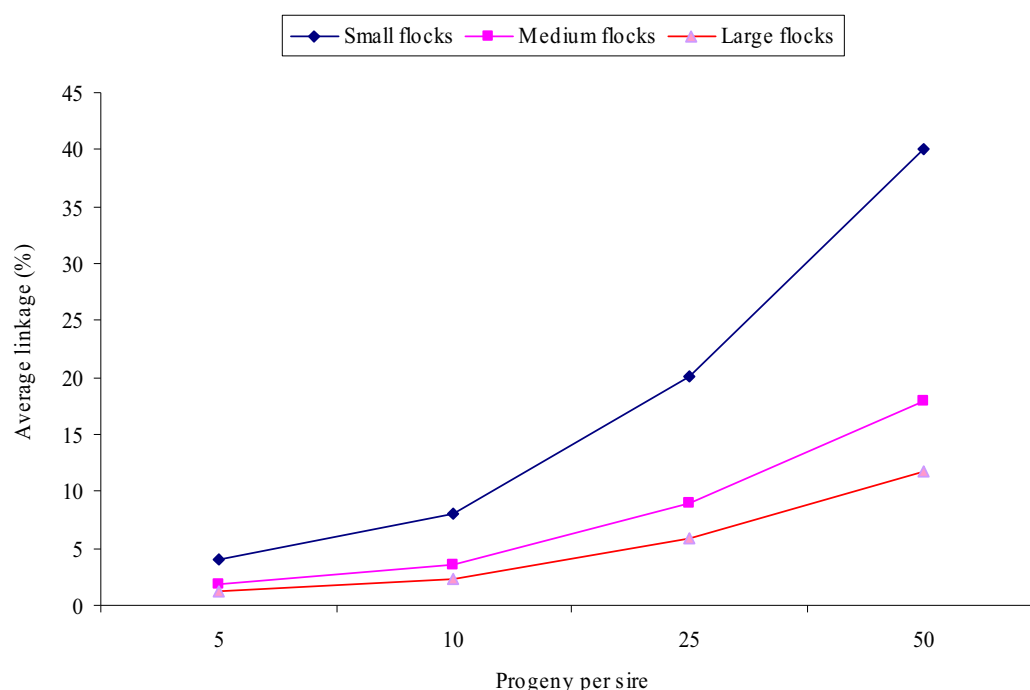


Figure 5.4. Average linkage (averaged over numbers of flocks in each category) in each category of small, medium and large flocks at varying levels of progeny per link sire

The correlation between simulated flock mean TBVs and flock mean EBVs for small flocks is low for 2 and 5 clusters as categorizing 50 flocks with varying genetic group effects into either 2 or 5 clusters is not an accurate approach of reducing the number of genetic group effects to be estimated (Figure 5.5). This is because of large distances between the genetic group and cluster means. The correlation for 10, 15, 20, 25 and 50 clusters is the same at all levels of progeny per sire. However, the level of accuracy for each cluster size is different as shown for the estimated between-cluster variance (Figure 5.3). The level of accuracy is higher (low standard errors) for 20 and 25 clusters (Figure 5.3). Since the correlation starts declining when more than 25 progeny per sire is used for all cluster sizes (Figure 5.5), it can be concluded that 25 progeny per sire (20% average linkage across small flocks; Figure 5.4) is adequate for small flocks. The maximum correlation reached at this level of linkage is approximately 0.88 for small flocks. The decline in correlation with increasing linkage is steeper if 2 and 5 clusters are formed (Figure 5.5) due to larger variance associated with these cluster sizes as discussed above for Figure 5.2.

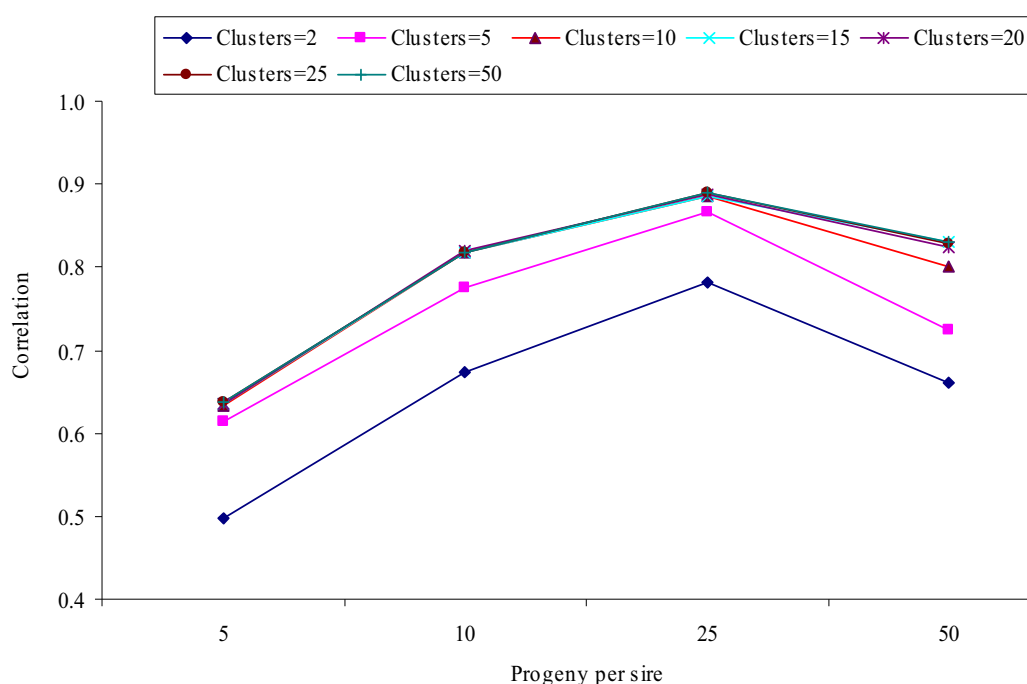


Figure 5.5. Correlation between simulated flock mean TBVs and EBVs for varying cluster sizes at different levels of progeny per link sire for small flocks

Similarly, for medium and large flocks, a similar trend is observed (results not presented). For medium and large flocks, a maximum correlation of 0.90 is obtained at the same level of progeny (25) per link sire. These results indicate that clustering flocks on the basis of estimated genetic group effects is an effective method for all flocks irrespective of their size and could be used for reducing the number of genetic group effects fitted in the model only if there is sufficient amount of linkage. However, in real scenarios, certain flocks are not adequately (at least 20% linkage required) linked. The formation of 10 to 50 clusters provides almost the same correlation between simulated flock mean TBVs and EBVs for small flocks (Figure 5.5). Similar results were obtained for medium and large flocks and are therefore not presented. There is virtually no difference in the accuracies of each cluster size relative to one another depending on the flock size (Figure 5.6).

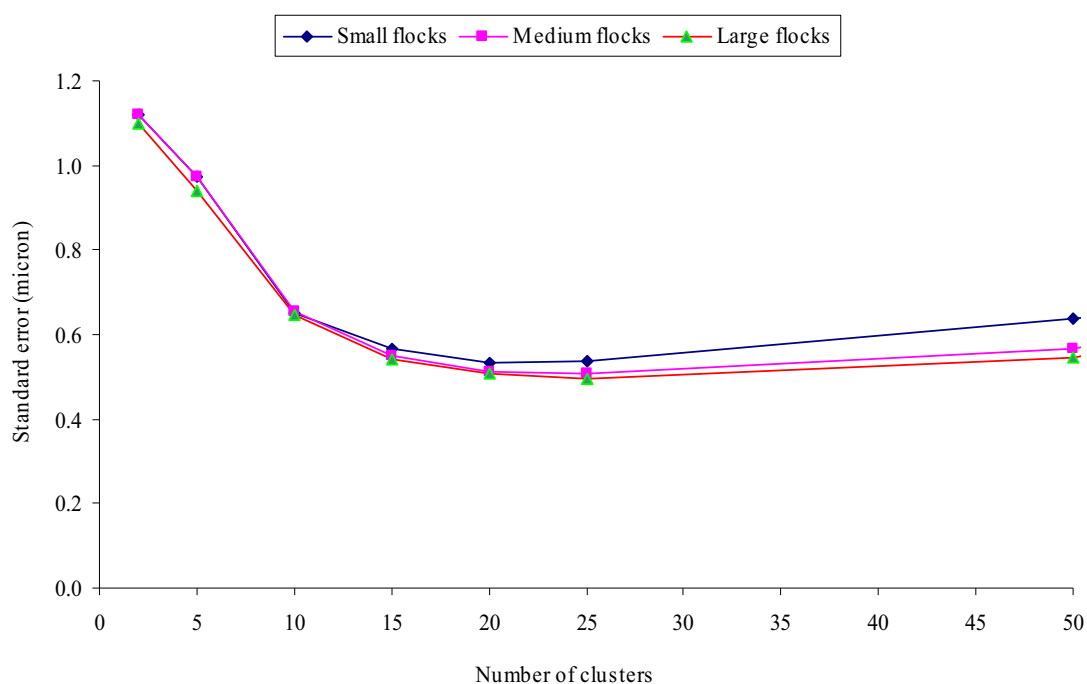


Figure 5.6. Standard errors of estimated cluster effects for different cluster sizes for small, medium and large flocks (25 progeny per link sire)

There is a gradual reduction in the standard errors of estimated cluster effects with increasing clusters from 2 to 25 (Figure 5.6). This is expected because an increase in the number of clusters formed for a given number of genetic groups led to an improvement in the precision of estimated cluster effects as a result of lesser number of flocks per cluster. When a large number of flocks are allocated to each cluster, the difference in the flock and cluster means is also large and more variable which leads to the reduced accuracy of cluster solutions. If less numbers of flocks are assigned to each cluster, the difference between flock and cluster means is not large thereby improving the accuracy of cluster solutions. If each cluster comprises only a single flock, the difference between flock and cluster means will be reduced significantly and cluster solutions will be similar to genetic group solutions derived without clustering. Alternatively, for the magnitude of genetic group effects simulated in this study, the single flock clusters lead to an over-fitted model. However, when more than 25 clusters are formed (fewer animals per cluster), there is an increase in the standard errors of all the flocks irrespective of their size (Figure 5.6). The difference in the standard errors of estimated cluster effects for small, medium and large flocks is not significant for all cluster sizes (Figure 5.6). The trends in Figures 5.6 and 5.7 indicate that cluster effects

are estimable with slightly higher accuracy for large than for small and medium flocks at all levels of linkage (progeny per link sire). However, this difference in the accuracy is not large enough to be significant. At all levels of progeny per sire, the accuracy of estimated cluster effects is comparable for small, medium and large flocks when 25 clusters are formed (Figure 5.7). A similar trend is observed with 10 clusters, however the standard errors are larger at all levels of progeny per sire (Figure 5.7). For small, medium and large flocks, with increasing progeny per sire, there is a decline in the standard errors of estimated cluster effects (Figure 5.7). This is due to the fact that at higher levels of linkage, the accuracy of genetic group solutions is increased which in turn improves the accuracy of cluster solutions. This also shows that the accuracy of cluster solutions depends on the accuracy of estimated genetic group effects. The decline in the standard errors of estimated cluster effects is more rapid in large than in small and medium flocks. The level of accuracy (based on standard errors of estimated cluster effects) of EBVs from 10 and 25 clusters is significantly different (Figure 5.7). The difference in accuracy is higher when more progeny per link sire is generated. The accuracy of prediction is higher for 25 than 10 clusters.

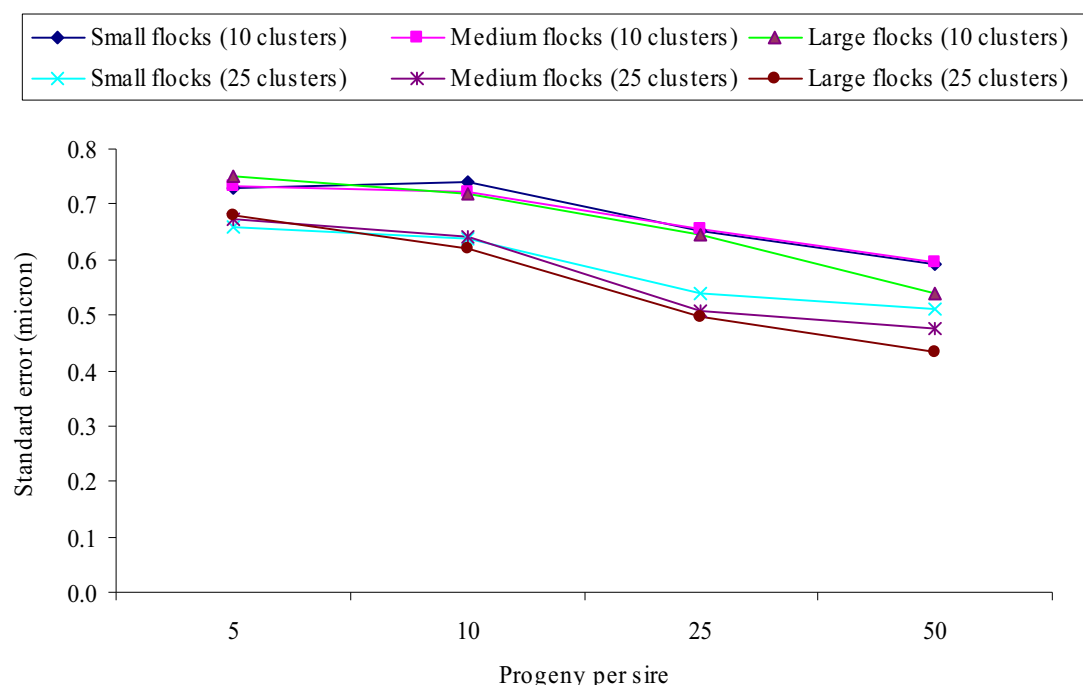


Figure 5.7. Standard errors of estimated cluster effects for small, medium and large flocks for 10 and 25 clusters at varying levels of progeny per link sire

Figure 5.8 shows the correlation between flock mean TBVs and EBVs obtained in a genetic groups model (estimated genetic group effects from flocks without clustering) along with correlation between flock mean TBVs and EBVs obtained in a cluster model (estimated genetic group effects from clustered flocks) with different cluster sizes and at different levels of progeny per sire. The trend in this graph shows that the correlation improves with increasing linkage and is significantly low for cluster size of 2. For all other cluster sizes, the correlation is similar at each level of progeny per sire. The correlation decreases slightly when less than 50 clusters are formed and is the least when 2 clusters are formed (Figure 5.8). The correlation between flock mean EBVs derived from a genetic groups model and the EBVs obtained from varying cluster models was also high for all cluster sizes except for 2 and 5 clusters when the level of linkage reaches 25 progeny per sire (Figure 5.8). A further increase in linkage leads to a reduction in correlation for 10, 15, 20 and 25 clusters due to reduction in the variance of TBVs when increasing proportion of progeny is derived from common link sires (Xsires). These results indicate the similarity between genetic grouping and clustering on the basis of estimated genetic group effects only when greater than 5 clusters are formed. However, the actual number of clusters required in a particular scenario would depend on the magnitude of genetic group effects and their variance. If the variance of genetic group effects is large, the number of clusters required would be more and vice-versa.

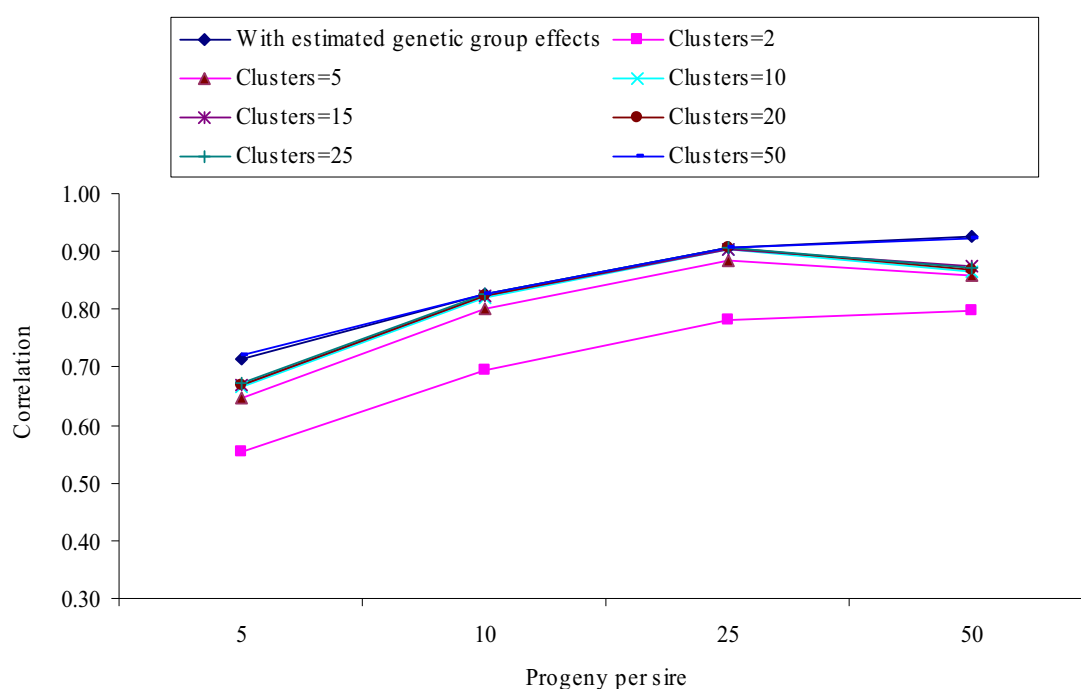


Figure 5.8. Across-flock correlation between simulated flock mean TBVs and EBVs for varying cluster sizes at different levels of progeny per link sire

The correlation between simulated TBVs and EBVs for all animals is shown in Figure 5.9. Except for 2 clusters, all cluster sizes provide similar correlation (Figure 5.9). When the number of clusters is increased from 2 to 50, there is only a very slight increase in the accuracy of prediction. This is due to a reduction in the distance between the genetic group and cluster means with increasing numbers of clusters. The correlation starts declining gradually when more than 25 progeny is generated per link sire (Figure 5.9). This is due to a reduction in the variance of the TBVs when increasing proportion of progeny is derived in each flock from the common link sires (Xsires). This emphasizes the relative importance of both link sire progeny and the progeny derived from local or home-bred sires (WNsires) in estimating cluster effects and hence the EBVs with sufficient accuracy. However, the correlation shown in Figure 5.9 is across-flock correlation which has been inflated due to the between-cluster variance. For all cluster sizes and levels of linkage except 5 progeny per sire, the correlation between simulated TBVs and EBVs within flock (Figure 5.10) is less than that shown in Figure 5.9.

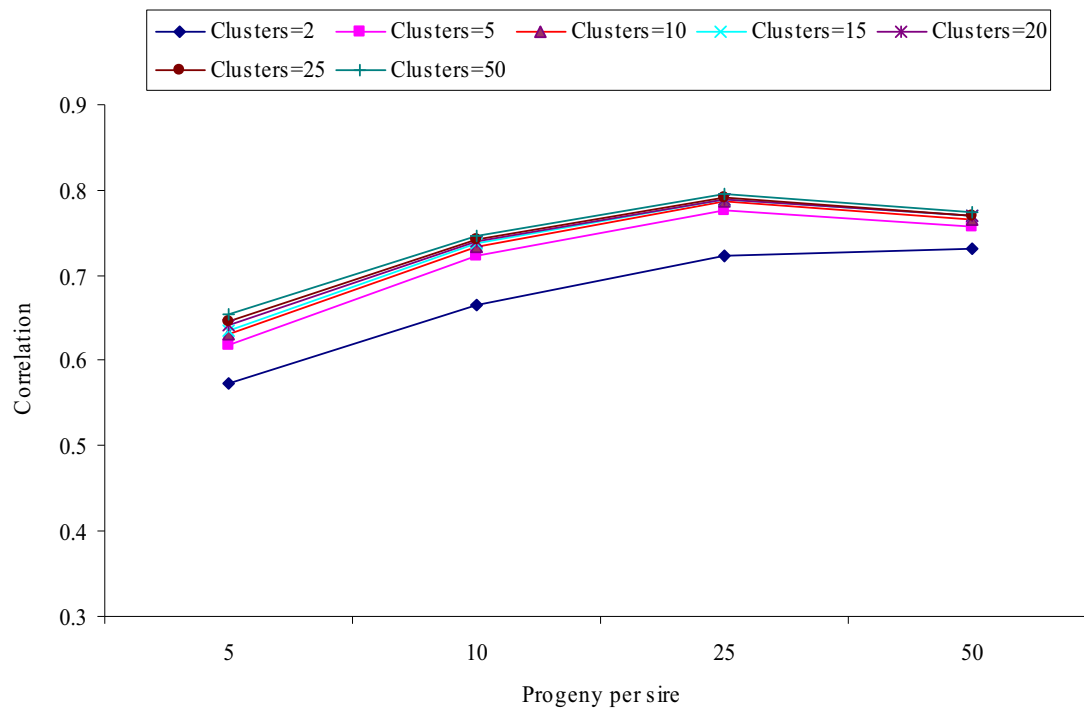


Figure 5.9. Across-flock correlation between simulated TBVs and EBVs with estimated cluster effects for varying cluster sizes at different levels of progeny per link sire

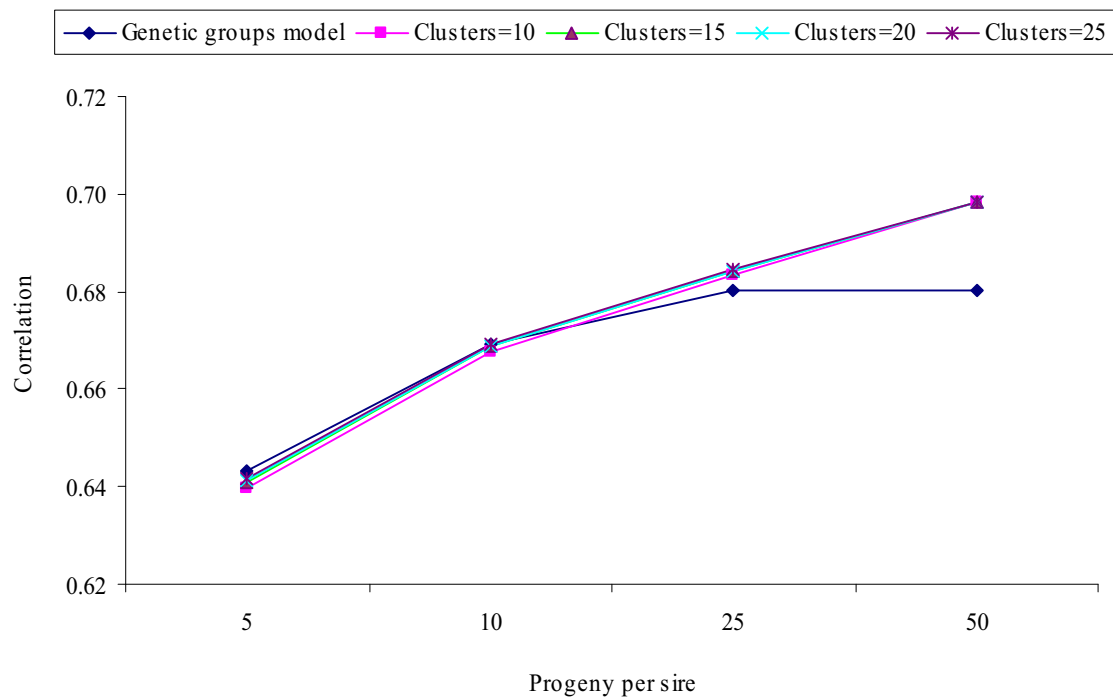


Figure 5.10. Within-flock correlation between simulated TBVs and EBVs from genetic groups model and different clustering models at varying levels of progeny per sire

5.4 Conclusions and Implications

Different numbers of clusters could be formed on the basis of estimated genetic group effects of flocks. The accuracy of clustering is dependent on the number of cluster effects that needs to be estimated in a model, the number of genetic groups and their associated effects, and the level of linkage across genetic groups. For a given number of genetic groups, the number of clusters required for maximizing the accuracy of prediction is a function of the variance of genetic group effects which are the basis for clustering. The accuracy of estimated cluster variance is dependent on the number of clusters formed and the proportion of linkage across genetic groups. An increase in the number of clusters leads to an improvement in the accuracy of estimated cluster variance and effects, and vice versa. In the absence of adequate genetic links across flocks, the accuracy of estimated genetic group effects would get reduced due to confounding with environmental differences between flocks. Additionally, if significant differences exist between genetic groups regarding the amount of data represented by each group, their solutions will be less accurate. This could affect the accuracy of clusters formed. Therefore, the minimum linkage required should be established between different flocks and each flock should have adequate data to accurately estimate their genetic group effects. Although not examined in this study, the actual proportion of linkage required will vary with the magnitude of genetic differences between flocks. Clustering is an effective method that could be used for reducing the number of genetic groups that need to be fitted in the model used for Sheep Genetics data without compromising the accuracy of evaluation procedure.

The design of data simulated in this study allowed for clustering flocks based on their genetic group solutions. However, in Merino industry evaluations, not all flocks are adequately linked which could lead to estimability problems (regarding their genetic group effects) for those flocks and therefore clustering may not be an effective method. Therefore, flocks should be adequately linked prior to estimating their genetic group effects and subsequently clustering flocks on the basis of their estimated genetic group effects. Alternatively, the clustering procedure needs to take this issue into consideration. Another issue relevant to the conclusions drawn from this study is the estimability of cluster effects. The FASTCLUS procedure (K -clustering) assigns flocks

to unique clusters based on the Euclidean distances between the genetic group means of different flocks. In addition, the level of linkage between flocks was not considered by the FASTCLUS procedure prior to forming clusters implying that flocks assigned to the same cluster might not necessarily be adequately linked. However, this could be a limitation of the clustering algorithm. Therefore, the clustering procedure needs to consider the issue of linkage between flocks prior to clustering. As reviewed under 2.7, each clustering mechanism is unique and is suitable for specific datasets. Furthermore, K -clustering does not provide the same pattern of clusters with each run because the clusters formed are dependent on the initial random assignment of data points. This implies that the results obtained are dependent on the number of clusters pre-specified. In reality, this may not necessarily be the ideal configuration. Although this type of effect could be reduced by multiple runs on the data and subsequent examination of the frequency of assigning of a flock to the same cluster, the allocation of flocks to different clusters could be inaccurate. Therefore, prior to the actual clustering process, not only distances between the data points but also linkage and estimability of cluster effects should be taken into consideration. Thus, future research should focus on developing algorithms and procedures that would take these issues into consideration.

For the 50 flocks simulated in this study, except for 2 and 5 clusters, all other cluster sizes (10, 15, 20, and 25 clusters) provided a similar level of accuracy of prediction. Nevertheless, the actual number of clusters required in a particular scenario or field situation would also depend on the size and the total number of genetic groups identified prior to evaluation. Additionally, the actual number of clusters needed in a particular situation will be determined by the magnitude of differences between genetic group means of different flocks. If the differences among genetic group means are not accurately estimated, the accuracy of clustering will be reduced. In this study, the genetic group solutions obtained from a mixed model analysis were considered for clustering flocks without taking into account their associated standard errors. Therefore, clustering of flocks could have been slightly inaccurate if the estimates of genetic group effects had large standard errors. The fact that FASTCLUS procedure could not account for the standard errors of estimated genetic group solutions appears to be a fairly serious limitation. This would be especially true for some of the flocks being clustered in

industry data having either small size (number of animals) or without adequate genetic links with other flocks. Hence further studies should focus on developing algorithms and evaluation procedures that would allow for the option of using genetic group solutions along with their related standard errors as an input for clustering flocks. This may improve the accuracy of clustering algorithm.

The results obtained and the conclusions drawn in this study may have been influenced by the small size of the simulated dataset used for the estimation of genetic group effects. Additionally, in the current analyses, the Euclidean distance was different for different cluster sizes. If the Euclidean distance is varied, the pattern and the size of clusters formed will get affected. However, the choice of Euclidean distance would depend on the level of similarity desired between estimated genetic group effects of different flocks. Future studies can investigate the impact of the size of dataset and different values for Euclidean distance on the accuracy of clustering mechanism. These issues need to be evaluated prior to the application of cluster analysis in Sheep Genetics Merino evaluations.

Depending on the differences in selection history and data recording, the usual convention of allocating animals to the same genetic group for different traits may not be the correct technique. When multiple traits are evaluated jointly, the animals could be assigned to different genetic groups for various traits based on their selection history and the amount of data available on each trait. The next chapter will explore different options for assigning genetic groups for different traits in a multiple-trait animal model evaluation.

Chapter 6. Multiple-trait animal model with differential genetic grouping

6.1 Introduction

Sheep farmers and breeders record multiple traits for improving the economic efficiency of their flocks. Therefore, it is more appropriate to use multi-trait genetic evaluation for predicting the aggregate genotype. There are several benefits of a multi-trait evaluation in comparison to single-trait evaluation. The main advantage of multi-trait estimation of breeding values is that the genetic correlation between traits is taken into consideration in the variance-covariance matrix of the MME which increases the accuracy of multi-trait EBVs. As reviewed in Chapter 2, different genetic grouping strategies have been proposed in the literature to improve the accuracy of evaluation procedures. However, grouping strategies for multi-trait evaluations may be different from those of single trait evaluations. For example, key Merino traits like fibre diameter (FD) and greasy fleece weight (GFW) are characterized by a long history of selection and are generally well recorded. However, traits like worm egg count (WEC) have a recent selection history and usually have fewer records. Hence, significant differences could exist in the numbers of missing records on different traits. Moreover, animals with a large body weight may not be in the same higher genetic group for a fertility trait like litter size. Therefore, alternative strategies for grouping animals differently for different traits in a multi-trait evaluation need to be explored.

Golden *et al.* (1994) argued that in multiple-trait evaluations, missing parents assigned to a genetic group for a trait may not be in the same group for another trait(s). The authors emphasized that further studies are required to investigate additional genetic grouping strategies for such situations. Hickey *et al.* (2008) evaluated Irish beef cattle industry data and reported large differences in the accuracy of genetic group solutions for different traits and breeds. Furthermore, depending on the mean, genetic variance, and the number of records available for each trait, the clustering mechanism would be different for different traits (Chapter 5) when evaluated jointly. The accuracy

of clustering flocks for different traits will depend on the accuracy of estimated genetic group solutions which in turn will depend on the number of records available on each trait. Therefore, a simultaneous evaluation of FD, GFW and WEC may affect the accuracy of estimated genetic group solutions for WEC with more recent selection history and fewer numbers of records.

Currently, animals are allocated to the same genetic group for all traits in the MERINOSELECT evaluation by Sheep Genetics for fleece, body weight and carcass traits. This traditional system will be termed as conventional genetic grouping. The multi-trait EBVs obtained using conventional genetic grouping may be biased due to differences in data recording and selection history between traits. To account for these differences, differential genetic grouping could be adopted. The term “differential” implies that an animal may be allocated to different genetic groups for different traits. As reviewed in section 2.2.3 of Chapter 2, genetic groups can be incorporated in the evaluation model through **Q**-matrix and **W**-matrix approaches. The **Q**-matrix is a design matrix which describes the proportion of genes derived by an animal from different genetic groups. The **W**-matrix is similar to the numerator relationship matrix (**A**) except that it includes genetic groups as extra animals. Therefore, the dimensions of **W**-matrix are equal to the number of animals plus the number of genetic groups.

The aim of this study is to explore and develop new and better genetic grouping strategies for Sheep Genetics based around differential genetic grouping. Additionally, computationally feasible MME for a multi-trait animal model with differential genetic grouping will be derived. A numerical example will be presented to investigate the use of **Q**-matrix and **W**-matrix approaches in predicting solutions for genetic groups and breeding values for individual animals using a hypothetical dataset.

6.2 The animal model and mixed model equations

Genetic group effects can be included in the evaluation model using two different approaches. The first approach fits genetic groups as additional fixed effects in the matrix **X** (incidence matrix for systematic fixed effects) and through matrix **Q** (describes the proportion of genes contributed by each genetic group to real animals) in the MME as described earlier in section 2.2.3 of Chapter 2. The second approach

involves inclusion of genetic groups through an expansion of the inverse of the numerator relationship matrix $\mathbf{A}$ which results in a new matrix $\mathbf{W}$, described previously in section 2.2.3 of Chapter 2. However, the solutions obtained are identical under both approaches as the models fitted are equivalent. This will be illustrated using a small numerical example. The two approaches can be extended to accommodate differential genetic grouping in the MME as shown in the following section.

6.2.1 Differential genetic grouping via Q-matrix

Scenario 1

For simplicity, only two traits and two genetic groups (GG1 and GG2) will be considered. In Scenario 1, genetic groups are assigned for Trait 1 only and for Trait 2, all animals are assumed to be in the same genetic group 2 (GG2). Thus, genetic grouping is undertaken only for the first trait and ignored for the second. The multiple-trait animal model with genetic grouping as proposed by Westell (1984) and Westell and Van Vleck (1987) can be written in vector notation as below:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations, $\mathbf{b}$ is a vector of fixed effects other than genetic groups, $\mathbf{g}$ is a vector of fixed genetic group effects, $\mathbf{u}$ is a vector of random animal effects; $\mathbf{X}$ and $\mathbf{Z}$ are the known incidence matrices associating fixed effects in $\mathbf{b}$, and animal and genetic group effects in $\mathbf{u}$ and $\mathbf{g}$, respectively with observations in $\mathbf{y}$; $\mathbf{Q}$ is a matrix which describes the proportion of genes contributed by different genetic groups to the animals of interest; and $\mathbf{e}$ is a vector of random residuals. The corresponding MME for the above model (Westell and Van Vleck, 1987) can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{ZQ} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1} \otimes \mathbf{G}_0^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZQ} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZQ} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where $\mathbf{R} (= \mathbf{I} \otimes \mathbf{R}_0)$ and $\mathbf{G} (= \mathbf{A} \otimes \mathbf{G}_0)$ are the Kronecker products of $\mathbf{I}$ and $\mathbf{R}_0$, and $\mathbf{A}$ and $\mathbf{G}_0$, respectively; $\mathbf{R}_0$ and $\mathbf{G}_0$ are the residual and genetic co(variance) matrices, respectively; $\mathbf{I}$ and $\mathbf{A}$ are the identity and additive genetic relationship matrices,

respectively; and $\otimes$ is the Kronecker product ($EBV = \hat{g} + \hat{u}$). These equations can be modified so as to group animals for one trait and ignore for the other. This will be illustrated using both approaches.

Scenario 2

For Scenario 2, it is assumed that some animals belong to the same genetic group (GG1 or GG2) for both traits and some of the animals belong to a genetic group (for example, GG1) for a given trait and to a different genetic group (for example, GG2) for the second trait. Let us consider the following two linear mixed models for the two different traits measured on the same animal:

$$y_1 = X_1 b_1 + Z_1 Q_1 g_1 + Z_1 u_1 + e_1$$

$$y_2 = X_2 b_2 + Z_2 Q_2 g_2 + Z_2 u_2 + e_2$$

where subscripts **1** and **2** correspond to the first and second traits, respectively and all the matrices are as specified above. These models can be combined and the modified MME accounting for differential genetic grouping can be expressed as:

$$\begin{bmatrix} X_1' R^{11} X_1 & X_1' R^{12} X_2 & X_1' R^{11} Z_1 & X_1' R^{12} Z_2 & X_1' R^{11} Z_1 Q_1 & X_1' R^{12} Z_2 Q_2 \\ X_2' R^{21} X_1 & X_2' R^{22} X_2 & X_2' R^{21} Z_1 & X_2' R^{22} Z_2 & X_2' R^{21} Z_1 Q_1 & X_2' R^{22} Z_2 Q_2 \\ Z_1' R^{11} X_1 & Z_1' R^{12} X_2 & Z_1' R^{11} Z_1 + M^{11} & Z_1' R^{12} Z_2 + M^{12} & Z_1' R^{11} Z_1 Q_1 & Z_1' R^{12} Z_2 Q_2 \\ Z_2' R^{21} X_1 & Z_2' R^{22} X_2 & Z_2' R^{21} Z_1 + M^{21} & Z_2' R^{22} Z_2 + M^{22} & Z_2' R^{21} Z_1 Q_1 & Z_2' R^{22} Z_2 Q_2 \\ Q_1' Z_1' R^{11} X_1 & Q_1' Z_1' R^{12} X_2 & Q_1' Z_1' R^{11} Z_1 & Q_1' Z_1' R^{12} Z_2 & Q_1' Z_1' R^{11} Z_1 Q_1 & Q_1' Z_1' R^{12} Z_2 Q_2 \\ Q_2' Z_2' R^{21} X_1 & Q_2' Z_2' R^{22} X_2 & Q_2' Z_2' R^{21} Z_1 & Q_2' Z_2' R^{22} Z_2 & Q_2' Z_2' R^{21} Z_1 Q_1 & Q_2' Z_2' R^{22} Z_2 Q_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X_1' R^{11} y_1 + X_1' R^{12} y_2 \\ X_2' R^{21} y_1 + X_2' R^{22} y_2 \\ Z_1' R^{11} y_1 + Z_1' R^{12} y_2 \\ Z_2' R^{21} y_1 + Z_2' R^{22} y_2 \\ Q_1' Z_1' R^{11} y_1 + Q_1' Z_1' R^{12} y_2 \\ Q_2' Z_2' R^{21} y_1 + Q_2' Z_2' R^{22} y_2 \end{bmatrix}$$

where M^{11} , M^{12} , M^{21} , and M^{22} are the direct products of A^{-1} with g^{11} , g^{12} , g^{21} , and g^{22} , respectively; g^{11} , g^{12} , g^{21} , and g^{22} are the corresponding elements of the inverse of the genetic co(variance) matrix G_0 ; and all other matrices are as specified earlier. It is important to note that Q_1 and Q_2 are different for differential genetic grouping.

6.2.2 Differential genetic grouping via W-matrix

Scenario 1

The MME for a multiple-trait animal model described above with genetic grouping through matrix W (Westell *et al.*, 1988) can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & 0 \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{g}^{11} & \mathbf{g}^{12} \\ 0 & \mathbf{g}^{21} & \mathbf{g}^{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ 0 \end{bmatrix}$$

where matrices $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{Q}$, and $\mathbf{R}$ and vectors $\mathbf{y}$, $\hat{\mathbf{g}}$ and $\hat{\mathbf{b}}$ are as above; vector $\hat{\mathbf{a}}$ contains animal solutions plus genetic group effects; $\mathbf{g}^{11}$ (Animal-Animal), $\mathbf{g}^{12}$ (Animal-Group), $\mathbf{g}^{21}$ (Group-Animal), and $\mathbf{g}^{22}$ (Group-Group) are the sub-matrices from the direct product of $\mathbf{W}^{-1}$ and $\mathbf{G}_0^{-1}$ ($\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$) where superscripts 1 and 2 correspond to animal and group, respectively and $\mathbf{G}_0$ is the genetic co(variance) matrix; matrix $\mathbf{W}$ results from the expansion of the inverse of matrix $\mathbf{A}$ including genetic groups. The $\mathbf{W}$ -matrix is similar to the numerator relationship matrix $\mathbf{A}$ except that it includes genetic groups as extra animals. The dimensions of $\mathbf{W}$ -matrix are equal to the number of animals including phantom parents plus the number of genetic groups. These MME can be modified by manipulating $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ to assign genetic groups for a given trait and ignore for the other. The following rule can be followed to construct $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ accounting for differential genetic groups:

1. Account for the covariance between the genetic group effects of Trait 1 and the additive effect of Trait 2 in the Group-Animal ($\mathbf{g}^{21}$) and Animal-Group ($\mathbf{g}^{12}$) portions of $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ where $\mathbf{g}^{21}$ and $\mathbf{g}^{12}$ are transpose of one another, and are the sub-matrices of modified $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$.

Scenario 2

The modified MME for this scenario are presented below:

$$\begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{11}\mathbf{X}_1 & \mathbf{X}'_1\mathbf{R}^{12}\mathbf{X}_2 & \mathbf{X}'_1\mathbf{R}^{11}\mathbf{Z}_1 & \mathbf{X}'_1\mathbf{R}^{12}\mathbf{Z}_2 & 0 & 0 \\ \mathbf{X}'_2\mathbf{R}^{21}\mathbf{X}_1 & \mathbf{X}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{X}'_2\mathbf{R}^{21}\mathbf{Z}_1 & \mathbf{X}'_2\mathbf{R}^{22}\mathbf{Z}_2 & 0 & 0 \\ \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{X}_1 & \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{X}_2 & \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{Z}_1 + \mathbf{D}_{aa}^{11} & \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{Z}_2 + \mathbf{D}_{aa}^{12} & \mathbf{D}_{ag}^{11} & \mathbf{D}_{ag}^{12} \\ \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{X}_1 & \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{Z}_1 + \mathbf{D}_{aa}^{21} & \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{Z}_2 + \mathbf{D}_{aa}^{22} & \mathbf{D}_{ag}^{21} & \mathbf{D}_{ag}^{22} \\ 0 & 0 & \mathbf{D}_{ga}^{11} & \mathbf{D}_{ga}^{12} & \mathbf{D}_{gg}^{11} & \mathbf{D}_{gg}^{12} \\ 0 & 0 & \mathbf{D}_{ga}^{21} & \mathbf{D}_{ga}^{22} & \mathbf{D}_{gg}^{21} & \mathbf{D}_{gg}^{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \\ \hat{\mathbf{g}}_1 \\ \hat{\mathbf{g}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{11}\mathbf{y}_1 + \mathbf{X}'_1\mathbf{R}^{12}\mathbf{y}_2 \\ \mathbf{X}'_2\mathbf{R}^{21}\mathbf{y}_1 + \mathbf{X}'_2\mathbf{R}^{22}\mathbf{y}_2 \\ \mathbf{Z}'_1\mathbf{R}^{11}\mathbf{y}_1 + \mathbf{Z}'_1\mathbf{R}^{12}\mathbf{y}_2 \\ \mathbf{Z}'_2\mathbf{R}^{21}\mathbf{y}_1 + \mathbf{Z}'_2\mathbf{R}^{22}\mathbf{y}_2 \\ 0 \\ 0 \end{bmatrix}$$

where matrices $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{R}$ and vectors $\mathbf{b}$, $\mathbf{a}$ and $\mathbf{g}$ are as above; $\mathbf{D}^{11}$, $\mathbf{D}^{12}$, $\mathbf{D}^{21}$, and $\mathbf{D}^{22}$ are the corresponding sub-matrices from modified $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ ($\mathbf{D}^{-1}$) and are the direct products of $\mathbf{W}^{-1}$ with $\mathbf{g}^{11}$, $\mathbf{g}^{12}$, $\mathbf{g}^{21}$, and $\mathbf{g}^{22}$, respectively; $\mathbf{g}^{11}$, $\mathbf{g}^{12}$, $\mathbf{g}^{21}$, and $\mathbf{g}^{22}$ are the corresponding elements of the inverse of the genetic co(variance) matrix $\mathbf{G}_0$. The superscripts 1 and 2 correspond to Trait 1 and Trait 2 and the subscripts a and g correspond to animal and group, respectively. Additionally, the sub-matrices $\mathbf{D}^{12}$ and $\mathbf{D}^{21}$ are transpose of one another. The rules for constructing $\mathbf{D}^{-1}$ (modified $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$) accounting for differential genetic grouping when considering two traits and two genetic groups for each trait are as follows:

Animal-Animal portion ($\mathbf{D}_{aa}$)

1. For every animal assigned to a different genetic group for each trait, $\mathbf{D}_{aa}$ is the same as conventional $\mathbf{A}$ (numerator relationship matrix) matrix and is the same for both traits. Although phantom parents might be different for the two traits, the inverse of the matrix $\mathbf{A}$ is the same. However, the matrix $\mathbf{A}$ is not the same for animals with missing pedigree information.

Group-Group portion ($\mathbf{D}_{gg}$)

1. For every animal allocated to a genetic group for a given trait, add the inverse of variance (from $\mathbf{G}_0^{-1}$) of that trait to the diagonal corresponding to that trait.
2. The covariance (from $\mathbf{G}_0^{-1}$) between the two traits is added to the off-diagonals if the same animal belongs to a different genetic group for the second trait.
3. The remaining elements are zeros as there is no relationship between the two genetic groups of the same trait.

Group-Animal portion ($\mathbf{D}_{ga}$)

1. For animals belonging to the same genetic group for both traits and for animals belonging to different genetic groups for different traits, variances and covariances (from $\mathbf{G}_0^{-1}$) with changed signs (positive to negative and vice versa) need to be added to

the corresponding diagonals and off-diagonals, respectively. The same rule can be used to construct Animal-Group portion ($\mathbf{D}_{ag}$) as Group-Animal ($\mathbf{D}_{ga}$) and Animal-Group ($\mathbf{D}_{ag}$) portions are transpose of one another.

6.3 Numerical example

The aim of this example is to show that differential genetic grouping is feasible using the modified MME and to examine whether same solutions are obtained using $\mathbf{Q}$ -matrix and $\mathbf{W}$ -matrix approaches. Let us consider the following pedigree and performance data information (Table 6.1) on animals from 1 to 5 on FD and GFW. Animals 1 to 4 have missing sire and dam pedigrees and have been assigned to genetic group 1 (GG1) and genetic group 2 (GG2). Animal 6 is the dam assigned to animal 5. Animal 3 is assumed to have a missing observation on GFW. Modified MME will be solved using the two approaches described above for differential genetic grouping in the following section.

Table 6.1. Hypothetical dataset for illustrating differential genetic grouping

Animal	Sire	Dam	FD (micron)	GFW (Kg)
1	GG1	GG1	18.5	2.8
2	GG1	GG1	17.5	3.4
3	GG2	GG2	21.5	Missing record
4	GG2	GG2	22.0	3.0
5	3	6	22.5	4.2
6	GG2	GG2	No record	No record

6.3.1 Q-matrix approach

Scenario 1

Throughout the calculations, traits and genetic groups are ordered within animals. From Table 6.1, the vector $\mathbf{y}$ of observations can be written as:

$$\mathbf{y}' = [18.5 \ 2.8 \ 17.5 \ 3.4 \ 21.5 \ 22 \ 3 \ 22.5 \ 4.2]$$

When constructing matrix $\mathbf{A}$, animal 6 will be placed first according to the usual rule of ordering animals (ancestor animals come first) in the pedigree (6, 1, 2, 3, 4, and 5). Using the above pedigree, the numerator relationship matrix $\mathbf{A}$ and its inverse ($\mathbf{A}^{-1}$) can be written as:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 1 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} 1.5 & 0 & 0 & 0.5 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 1.5 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 2 \end{bmatrix}$$

The genetic and residual co(variance) matrices and their inverse are:

$$\mathbf{G}_o = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \quad \mathbf{G}_o^{-1} = \begin{bmatrix} 0.6 & -0.2 \\ -0.2 & 0.4 \end{bmatrix} \quad \mathbf{R}_o = \begin{bmatrix} 5 & 3 \\ 3 & 10 \end{bmatrix} \quad \mathbf{R}_o^{-1} = \begin{bmatrix} 0.24390 & -0.07317 \\ -0.07317 & 0.12195 \end{bmatrix}$$

The incidence matrices for the fixed (two trait means) and genetic group effects in $\hat{\mathbf{b}}$ and $\hat{\mathbf{g}}$, respectively and animal effects in $\hat{\mathbf{u}}$, can be expressed as:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where column numbers 1, 2, 3, 4, and 5 in $\mathbf{X}$ correspond to Trait 1 mean, Trait 2 mean, GG1 effect for Trait 1, GG2 effect for Trait 1 and GG2 effect for Trait 2, respectively. The first two columns in $\mathbf{Z}$ are null as animal 6 is a phantom animal and does not have a record on either trait. Similarly, the remaining columns in $\mathbf{Z}$ correspond to observations on the remaining animals for both traits. Although animal 3 has a missing observation for Trait 2, a null column corresponding to this has been added to $\mathbf{Z}$. The inverse of $\mathbf{R}$ ($\mathbf{R} = \mathbf{I} \otimes \mathbf{R}_0$, where $\mathbf{I}$ is an identity matrix of order 5) is as below:

$$\mathbf{R}^{-1} = \begin{bmatrix} 0.243 & -0.073 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.073 & 0.122 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.244 & -0.073 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.073 & 0.122 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.244 & -0.073 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.073 & 0.122 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.244 & -0.073 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.073 & 0.122 & 0 \end{bmatrix}$$

Since animal 3 has a record for Trait 1 only, the row (or column) corresponding to animal 3 in $\mathbf{R}^{-1}$ above contains 0.2 which is the inverse of 5 only from $\mathbf{R}_0$. As there is a dependency in the resulting coefficient matrix, two restrictions are imposed by setting the rows and columns corresponding to GG2 effect for Trait 1 and GG2 effect for Trait 2 to zero to make the coefficient matrix of full rank. The solution vectors (values are rounded to three decimal places) obtained after solving the modified MME can now be expressed as:

$$\hat{\mathbf{b}} = [21.961 \quad 3.282]'$$

$$\hat{\mathbf{g}} = [-3.905]$$

$$\hat{\mathbf{u}} = [0.084 \quad 0.113 \quad 0.137 \quad -0.103 \quad -0.165 \quad 0.019 \quad -0.071 \quad 0.035 \quad 0.015 \quad -0.063 \quad 0.090 \quad 0.186]'$$

The estimates in $\hat{\mathbf{g}}$ are not the absolute solutions and are the differences between the two genetic group means for Trait 1. Thus, genetic group solutions in $\hat{\mathbf{g}}$ need to be added to the solutions in $\hat{\mathbf{u}}$ to obtain the final EBVs in $\hat{\mathbf{a}}$ ($\hat{\mathbf{a}} = \hat{\mathbf{g}} + \hat{\mathbf{u}}$) which are:

$$[0.084 \quad 0.113 \quad -3.768 \quad -0.103 \quad -4.070 \quad 0.019 \quad -0.071 \quad 0.035 \quad 0.015 \quad -0.063 \quad 0.090 \quad 0.186]$$

Scenario 2

Genetic grouping of animals for Scenario 2 is presented in Table 6.2. Animals 2 (GG1 and GG2 for Traits 1 and 2, respectively) and 3 (GG2 and GG1 for Traits 1 and 2, respectively) have been assigned genetic groups differentially for the two traits. Animal 5 has both parents identified (Table 6.1), therefore, has not been assigned to any genetic group. However, the incidence matrix for fixed effects (trait means plus GG1 and GG2 effects for each trait) including genetic groups, contains coefficients for GG1 and GG2 effects for Trait 2 recorded on animal 5. This is justified by the fact that the sire (animal 3) and phantom dam (animal 6) of animal 5 belong to GG1 and GG2, respectively for Trait 2.

Table 6.2. Pedigree illustrating Scenario 2 of differential genetic grouping

Animal	Trait 1	Trait 2
1	GG1	GG1
2	GG1	GG2
3	GG2	GG1
4	GG2	GG2
5	3	6
6	GG2	GG2

The incidence matrix $\mathbf{X}$ for the fixed (including genetic groups) and genetic group effects in $\hat{\mathbf{b}}$ and $\hat{\mathbf{g}}$, respectively can be written as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

The incidence matrix for the animal effects in $\hat{\mathbf{u}}$ is the same as in Scenario 1. The order of animals in $\mathbf{X}$ is the same as that in $\mathbf{A}$, that is, animals 6, 1, 2, 3, 4, and 5. The column numbers 1, 2, 3, 4, 5, and 6 in $\mathbf{X}$ correspond to Trait 1 mean, Trait 2 mean, Trait 1(GG1), Trait 1(GG2), Trait 2(GG1), and Trait 2(GG2), respectively. For Trait 2, animal 5 belongs to GG1 (0.5) and GG2 (0.5). The order of columns in $\mathbf{Z}$ is the same as for Scenario 1 described above. There is a dependency in the coefficient matrix which is removed by deleting the rows and columns corresponding to GG2 effects for Trait 1 and Trait 2. The solutions obtained are as below:

$$\hat{\mathbf{b}} = [21.979 \quad 3.520]'$$

$$\hat{\mathbf{g}} = [-3.948 \quad -0.640]'$$

$$\hat{\mathbf{u}} = [0.081 \quad 0.122 \quad 0.139 \quad -0.012 \quad -0.154 \quad -0.034 \quad -0.079 \quad 0.042 \quad 0.013 \quad -0.118 \quad 0.082 \quad 0.204]'$$

The vector $\hat{\mathbf{a}}$ of EBVs ($\hat{\mathbf{a}} = \hat{\mathbf{g}} + \hat{\mathbf{u}}$) can be written as:

$$[0.081 \quad 0.122 \quad -3.810 \quad -0.652 \quad -4.102 \quad -0.034 \quad -0.079 \quad -0.598 \quad 0.013 \quad -0.118 \quad 0.082 \quad -0.116]$$

6.3.2 W-matrix approach

Scenario 1

The incidence matrices for the fixed (two trait means) and animal effects, and the inverse of the additive genetic relationship matrix $\mathbf{W}$ (containing phantom animal and genetic groups) are as follows:

$$\mathbf{X}' = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{W}^{-1} = \begin{bmatrix} 2 & 0 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 3 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & -1 & 1.5 & 0 & 0 & 0.5 & 0 & -1 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0.5 & 0 & 0 & 1.5 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 0 & 2 \end{bmatrix}$$

The two rows in $\mathbf{X}'$ correspond to the two trait means. An extra null column (first column) has been added to $\mathbf{Z}$ which corresponds to GG1 effect for Trait 1 and the remaining columns in $\mathbf{Z}$ are the same as described earlier. The direct product of $\mathbf{W}^{-1}$ and $\mathbf{G}_0^{-1}$ ($\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$) is now modified by deleting rows and columns corresponding to GG1 effect for Trait 2 and GG2 effects for Traits 1 and 2 because genetic group effects are

only applicable for Trait 1. The modified $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ accounts for the covariance between GG1 effect of Trait 1 and the additive genetic effect of Trait 2 itself.

$$\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1} =$$

$$\begin{bmatrix} 1.2 & 0 & 0 & -0.6 & \mathbf{0.2} & -0.6 & \mathbf{0.2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9 & -0.3 & 0 & 0 & 0 & 0 & 0.3 & -0.1 & 0 & 0 & -0.6 & 0.2 \\ 0 & -0.3 & 0.6 & 0 & 0 & 0 & 0 & -0.1 & 0.2 & 0 & 0 & 0.2 & -0.4 \\ -0.6 & 0 & 0 & 0.6 & -0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0.2} & 0 & 0 & -0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.6 & 0 & 0 & 0 & 0 & 0.6 & -0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0.2} & 0 & 0 & 0 & 0 & -0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & -0.1 & 0 & 0 & 0 & 0 & 0.9 & -0.3 & 0 & 0 & -0.6 & 0.2 \\ 0 & -0.1 & 0.2 & 0 & 0 & 0 & 0 & -0.3 & 0.6 & 0 & 0 & 0.2 & -0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & -0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.2 & 0.4 & 0 & 0 \\ 0 & -0.6 & 0.2 & 0 & 0 & 0 & 0 & -0.6 & 0.2 & 0 & 0 & 1.2 & -0.4 \\ 0 & 0.2 & -0.4 & 0 & 0 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & -0.4 & 0.8 \end{bmatrix}$$

The solutions obtained after solving the MME are:

$$\hat{\mathbf{b}} = [21.961 \ 3.282]'$$

$$\hat{\mathbf{g}} = [-3.904]$$

$$\hat{\mathbf{a}} = [0.084 \ 0.113 \ -3.768 \ -0.103 \ -4.069 \ 0.019 \ -0.071 \ 0.035 \ 0.015 \ -0.063 \ 0.090 \ 0.186]'$$

These solutions are the same as those derived using the $\mathbf{Q}$ -matrix approach.

Scenario 2

The incidence matrix $\mathbf{X}$ for the fixed effects (two trait means) in $\hat{\mathbf{b}}$ is the same as shown above for Scenario 1 and the incidence matrix $\mathbf{Z}$ for the animal effects in $\hat{\mathbf{u}}$ can be written as:

$$\mathbf{Z}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The modified $\mathbf{W}^{-1} \otimes \mathbf{G}_0^{-1}$ using the rules presented above can be written as:

$$\begin{bmatrix} 1.2 & -0.2 & 0 & -0.2 & 0 & 0 & -0.6 & 0.2 & -0.6 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.2 & 0.8 & -0.2 & 0 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & 0 & 0 \\ 0 & -0.2 & 1.8 & -0.4 & -0.6 & 0.2 & 0 & 0 & 0 & 0 & -0.6 & 0.2 & -0.6 & 0.2 & 0 & 0 \\ -0.2 & 0 & -0.4 & 1.2 & 0.2 & -0.4 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & 0.2 & -0.4 & 0 & 0 \\ 0 & 0 & -0.6 & 0.2 & 0.9 & -0.3 & 0 & 0 & 0 & 0 & 0.3 & -0.1 & 0 & 0 & -0.6 & 0.2 \\ 0 & 0 & 0.2 & -0.4 & -0.3 & 0.6 & 0 & 0 & 0 & 0 & -0.1 & 0.2 & 0 & 0 & 0.2 & -0.4 \\ -0.6 & 0.2 & 0 & 0 & 0 & 0 & 0.6 & -0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & -0.4 & 0 & 0 & 0 & 0 & -0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.6 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0.6 & -0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & -0.4 & 0 & 0 & 0 & 0 & -0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & -0.6 & 0 & 0.3 & -0.1 & 0 & 0 & 0 & 0 & 0.9 & -0.3 & 0 & 0 & -0.6 & 0.2 \\ 0 & -0.4 & 0.2 & 0 & -0.1 & 0.2 & 0 & 0 & 0 & 0 & -0.3 & 0.6 & 0 & 0 & 0.2 & -0.4 \\ 0 & 0 & -0.6 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & -0.2 & 0 & 0 \\ 0 & 0 & 0.2 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.2 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.6 & 0.2 & 0 & 0 & 0 & 0 & -0.6 & 0.2 & 0 & 0 & 1.2 & -0.4 \\ 0 & 0 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & 0 & 0 & 0.2 & -0.4 & 0 & 0 & -0.4 & 0.8 \end{bmatrix}$$

There is a dependency in the resulting coefficient matrix. Therefore, the rows and columns corresponding to GG2 effect for Trait 1 (5th) and GG2 effect for Trait 2 (6th) are set to zero and the solutions obtained are as follows:

$$\hat{\mathbf{b}} = [21.979 \ 3.519]'$$

$$\hat{\mathbf{g}} = [-3.948 \ -0.639]'$$

$$\hat{\mathbf{a}} = [0.081 \ 0.122 \ -3.810 \ -0.652 \ -4.102 \ -0.034 \ -0.079 \ -0.598 \ 0.013 \ -0.118 \ 0.082 \ -0.116]'$$

The solutions in $\hat{\mathbf{a}}$ are exactly the same as those derived using the $\mathbf{Q}$ -matrix approach.

6.4 Discussion

As seen in the previous sections, differential genetic grouping can be implemented using the modified MME for the two scenarios described. Differential genetic groups can be included in the evaluation scheme either as additional fixed effects in the model ($\mathbf{Q}$ -matrix approach) or through a modified $\mathbf{A}$ inverse ($\mathbf{W}$ -matrix approach). Both approaches provide identical solutions. However, if the animals to be evaluated belong to a large number of distinct and overlapping generations, deriving coefficients for the matrix $\mathbf{Q}$ can be tedious. On the other hand, estimating elements for the $\mathbf{W}$ -matrix will be relatively simpler. This is due to the advantage of $\mathbf{W}$ -matrix approach in that it generates matrix $\mathbf{Q}$ automatically and therefore the final animal solutions include estimated genetic group effects. Additionally, solving the MME using the $\mathbf{Q}$ -matrix approach may involve more computational time than the $\mathbf{W}$ -matrix approach. However, the coefficient matrix for an animal model with genetic grouping is always singular (Da *et al.*, 1989). Therefore, certain restrictions need to be imposed to obtain the inverse of the coefficient matrix when genetic groups are included in the evaluation procedure.

In the numerical example presented above, the covariance between genetic groups of different traits was the same as the genetic covariance. However, in reality, this would depend on the structure of the data and selection pressure each trait had undergone in the past (selection history) which in most cases could be unknown. When genetic groups are fitted as fixed effects in the evaluation model, they do not have a covariance structure and the common term that arises from the inverse of matrix $\mathbf{R}_0$ (residual covariance matrix) simply accounts for the correlated errors, as with systematic fixed effects. However, when genetic groups are treated as random effects in

the differential groups model, the covariance between genetic groups of different traits would be a function of matrix $\mathbf{G}_0$ (genetic covariance matrix). Alternatively, the covariance among genetic groups for different traits could be estimated using field data. Nevertheless, further research should investigate this issue in more detail. The two approaches used for differential genetic grouping for the numerical example considered in this study could be extended to routine Sheep Genetics evaluations using OVIS. This would provide alternative procedures of accounting for genetic groups for a large number of traits evaluated by Sheep Genetics that are characterized by differences in selection history and data recording. For such traits, the strategy of differential genetic grouping will provide more accurate multi-trait EBVs.

A large number of reproduction traits like the number of lambs born alive, number of lambs weaned per ewe, and other fertility traits are also being recorded regularly and can be improved to maximize the profitability of a sheep enterprise. Therefore, when reproduction traits are evaluated jointly with other production traits like greasy fleece weight and fibre diameter, differential grouping strategy appears to be an alternative option which would account for the differences among traits regarding their selection history, recording or data structure.

6.5 Conclusions and Implications

Differential genetic grouping is feasible using the modified mixed model equations presented for a multiple-trait animal model. The modified mixed model equations can be constructed and can accommodate varying numbers of missing observations on different trait(s). Accounting for differential genetic groups as additional fixed effects in the model ($\mathbf{Q}$ -matrix approach) and as extra equations in the $\mathbf{A}$ inverse ($\mathbf{W}$ -matrix approach) provide the same solutions in this small example. For large datasets with complex pedigrees and involving a large number of traits, solving mixed model equations with differential genetic grouping may increase computational time especially if maternal genetic effects are also included in the model because building the coefficient matrix is a direct function of the square of the number of factors. This could also over-parameterize the model, thus leading to either estimability problems or inaccurate solutions.

Future studies should focus on the development of formulations and algorithms for the REML estimation of variance components with differential genetic grouping. Depending on the data structure and selection history, differential genetic grouping can be incorporated in OVIS for routine Sheep Genetics evaluations. Further studies could also focus on incorporating maternal genetic effects in the mixed model equations accounting for differential genetic grouping for multiple traits. In this approach, genetic groups would be defined for maternal genetic effects on the basis of different criteria depending on the data structure. The rules defined for differential grouping for additive effects can be used for allocating groups differentially for maternal genetic effects as well. However, the (co)variance structure required for maternal genetic groups of different traits will need further investigation.

Depending on data recording and selection history, the strategy of differential genetic grouping will allow additional options for allocating animals to genetic groups for a range of traits evaluated by Sheep Genetics. The next chapter (Chapter 7) will examine the application of conventional and differential genetic grouping to industry data from Sheep Genetics database. Chapter 7 will also investigate whether there are any possible benefits of differential over conventional grouping when flocks are clustered differently based on their estimated genetic group solutions for different traits.

Chapter 7. Conventional versus differential genetic grouping in across-flock multiple-trait evaluation

7.1 Introduction

Field data in the Australian Merino sheep industry are characterized by varying amounts of missing records on different traits. For some traits like worm egg count (WEC), significant differences also exist in the selection history and the amount of data recorded compared to other traits in the analysis. The differences in selection history imply that animals with records may not be in the same genetic group for different traits. The previous chapter demonstrated that in a multiple-trait evaluation animals can be allocated to differential genetic groups for various traits. In this chapter, this concept is applied to real data.

Currently, Sheep Genetics evaluations utilize the same grouping strategy based on flock of origin and time period for all traits. In this approach, genetic groups are formed based on different combinations of flock and time period in order to estimate breeding values and genetic trends across flocks with increased precision. The inclusion of genetic groups accounts for the change in the genetic merit of base animals over a time period (Brown *et al.*, 2007). This system of grouping will be termed as conventional genetic grouping. When traits are evaluated jointly, the conventionally used strategy of assigning the same genetic group to an animal for all traits may affect the accuracy of across-flock EBVs by inaccurately estimating the genetic group effects for some traits. This is due to the fact that certain groups may have adequate data for some trait(s) but may not have sufficient data for the other trait(s). This may also affect the accuracy of clustering flocks (if groups are formed based on flock) which is dependent on the accuracy of estimated genetic group solutions.

When groups are allocated differentially, the EBVs for traits with adequate data are less likely to be affected by the missing records on other traits. In such cases, differential grouping could remove the bias in the multi-trait EBVs by more accurately predicting the genetic group effects for all traits. The objective of this study is to compare the ability to predict progeny performance using multi-trait across-flock EBVs obtained from industry data using conventional and differential genetic grouping. Additionally, the effect of missing data for a trait on EBVs obtained from the two strategies of genetic grouping will also be investigated.

7.2 Materials and Methods

7.2.1 Data

A subset of data from Australian Merino sheep was extracted from the MERINOSELECT database of Sheep Genetics. The subset of data was chosen by selecting the 15 flocks which Sheep Genetics submitted the most sires to CTSE (Central Test Sire Evaluation). The data and pedigree files contained 110,380 and 118,986 animals, respectively. The number of animals with both sire and dam pedigree missing were 54,273; those with both sire and dam pedigree known were 26,643; those with sire known and dam unknown were 27,906; and those with dam known and sire unknown were 1,558. The animals in the data file belonged to 15 flocks and had pedigree links to 140 flocks without performance data. These additional flocks tracing back in the pedigree file were pooled into one genetic group. The traits considered for analysis were body weight (BWT) and fibre diameter (FD) recorded at hogget age (average of 18 months). The data were pre-adjusted for fixed effects of birth type, rearing type, age of animal at measurement, and linear and quadratic effects of dam age using OVIS (Brown *et al.*, 2000). The observations and contemporary group (CG) codes for BWT and FD were extracted for further analysis. The CG was defined on the basis of flock, year of birth, sex, date of measurement and management group.

The level of linkage due to common sires varied between flocks (Table 7.1). The number of animals in each flock (flock size) varied with different proportions of animals derived from link sires. The number of link sires was 45 and each link sire had varying numbers of progeny in different flocks. In some flocks, linkage was due to only

one link sire while in others linkage was due to many link sires. For some flock combinations, there were no direct genetic links (Table 7.1). The basic summary statistics on the data analyzed and the number of records available for each trait evaluated is presented in Table 7.2. More than twice as many animals had FD recorded than BWT.

Table 7.1. The numbers of progeny derived from common link sires for different combinations of flocks (values on diagonal are the total number of animals in each flock)

Flock	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	5,825	1,544	548	0	0	2,037	101	117	790	360	360	117	360	365	1,638
2	1,544	9,283	548	0	0	2,207	243	117	665	360	360	241	360	553	140
3	548	548	12,544	0	0	360	0	35	548	381	360	0	360	0	246
4	0	0	0	3,903	93	90	0	0	0	0	0	0	0	0	0
5	0	0	0	93	544	0	0	0	0	0	0	0	0	0	0
6	2,037	2,207	360	90	0	13,558	586	0	360	360	360	197	360	106	1,339
7	101	243	0	0	0	586	4,488	0	0	0	0	0	0	0	0
8	117	117	35	0	0	0	0	900	117	0	0	117	0	117	20
9	790	665	548	0	0	360	0	117	10,115	360	360	117	360	207	90
10	360	360	381	0	0	360	0	0	360	1,785	437	0	360	0	0
11	360	360	360	0	0	360	0	0	360	437	2,027	0	360	0	0
12	117	241	0	0	0	197	0	117	117	0	0	9,869	0	117	0
13	360	360	360	0	0	360	0	0	360	360	360	0	5,798	0	0
14	365	553	0	0	0	106	0	117	207	0	0	117	0	6,883	337
15	1,638	140	246	0	0	1,339	0	20	90	0	0	0	0	337	22,858

Table 7.2. Summary statistics of Sheep Genetics data used for estimating breeding values for BWT and FD using conventional and differential genetic grouping

Trait	Number of records	Number of sires	Number of CGs ¹	Mean	Maximum	Minimum
BWT (Kg)	47,731	1,402	267	48.09	82.12	23.90
FD (micron)	106,292	1,402	790	18.57	24.36	13.37

¹ Contemporary groups

7.2.2 Data analysis

Estimation of genetic group effects

Genetic groups were assigned to all animals with data on the basis of flock of origin, thus, 15 initial genetic groups were formed. A 16th genetic group comprised all the additional flocks to which animals in the extended pedigree belonged. Animals belonging to the 16th genetic group had no performance data. An animal model was fitted in a bivariate analysis to estimate variance components and genetic group effects treated as fixed for BWT and FD. The ASReml package (Gilmour *et al.*, 2006) was used for all statistical analyses. Maternal genetic effects were fitted for BWT, this improved the log likelihood of the model. The animal model fitted can be written in vector notation as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{Z_m m} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations on BWT and FD, $\mathbf{b}$ is a vector of fixed effects including mean and CG, $\mathbf{g}$ is a vector of fixed genetic group effects, $\mathbf{u}$ is a vector of direct additive genetic effects of the animals, $\mathbf{m}$ is a vector of maternal genetic effects fitted for BWT only, $\mathbf{e}$ is a vector of random residuals; $\mathbf{X}$ is the incidence matrix associating records in $\mathbf{y}$ with the fixed flock effects in $\mathbf{b}$; $\mathbf{Z}$ and $\mathbf{Z_m}$ are the incidence matrices associating records in $\mathbf{y}$ with the random animal and dam genetic effects in $\mathbf{u}$ and $\mathbf{m}$, respectively; and $\mathbf{Q}$ is a design matrix which describes the proportion of genes derived by an animal from each genetic group. The covariance between additive and maternal genetic effects of BWT was assumed to be zero. Maternal genetic effects for

BWT were included in the genetic covariance matrix ($\mathbf{G}_o$). The $\mathbf{G}_o$ and $\mathbf{R}_o$ (residual covariance matrix) used were:

$$\mathbf{G}_o = \begin{pmatrix} 11.20 & 0.00 & 0.42 \\ 0.00 & 1.36 & 0.01 \\ 0.42 & 0.01 & 0.64 \end{pmatrix} \text{ and } \mathbf{R}_o = \begin{pmatrix} 14.20 & 0.43 \\ 0.43 & 0.89 \end{pmatrix}$$

The same $\mathbf{G}_o$ and $\mathbf{R}_o$ matrices were used for estimating genetic group and subsequently, cluster effects (genetic group effects from clustered flocks) for conventional and differential genetic grouping strategies. The methodology used for the comparison of EBVs obtained from conventional and differential grouping is illustrated diagrammatically in Figure 7.1.

Conventional clustering

The estimated genetic group effects for each trait were standardized to get an even influence of each trait on the clustering mechanism (refer to **Appendix** for more information). Additionally, to eliminate the effect of difference in the means of two traits, the data were transformed as per routine Sheep Genetics evaluations (Brown *et al.*, 2006; 2007). Flocks were clustered simultaneously (multivariate) on the basis of their estimated genetic group effects (from the previous analysis) for BWT and FD using the FASTCLUS procedure of SAS (1990). The principle of clustering using this procedure has been described in Chapter 5. Only genetic group solutions were the basis for clustering flocks without taking into account their associated standard errors due to the limitation of the FASTCLUS procedure. A $\mathbf{Q}$ -matrix common for both traits was derived. The Euclidean distance between the estimated genetic group effects of flocks was not specified in order to manipulate the number of clusters.

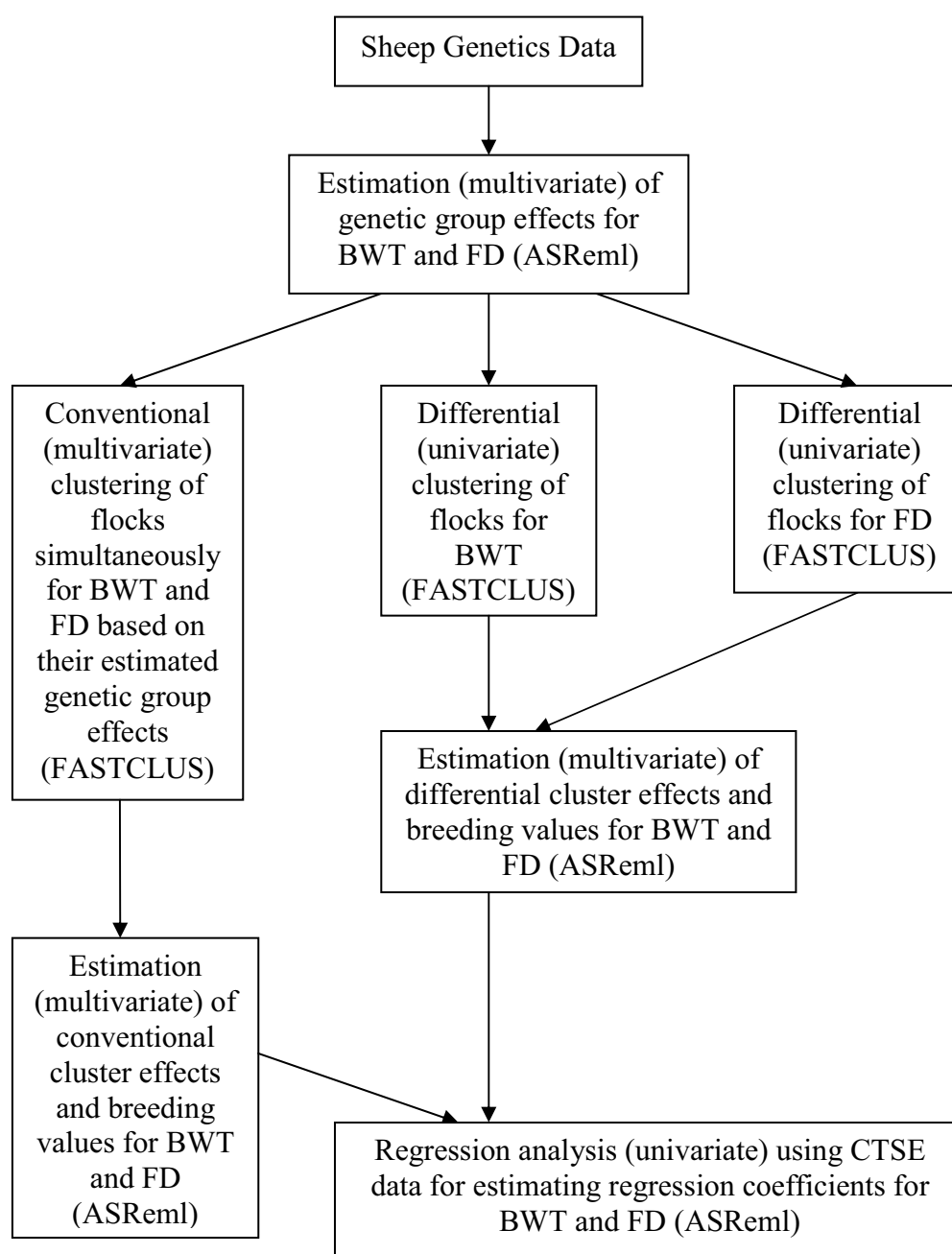


Figure 7.1. Flow diagram depicting the methodology used for the comparison of EBVs from conventional and differential genetic grouping strategies for BWT and FD

Clustering grouped the 16 flocks or genetic groups into 5 groups with similar genetic merit for both traits. The purpose of clustering was to reduce the number of genetic group effects that need to be estimated and therefore, improve their accuracy, as

demonstrated in Chapter 4. For this analysis, the genetic group effects estimated after clustering flocks into groups based on cluster allocation will be termed cluster effects. The cluster effects and breeding values for the two traits were estimated in an animal model using bivariate analysis. The animal model fitted was equivalent to the one described above except for the fixed cluster effects replacing fixed genetic group effects. CG was also fitted as fixed effect in the model.

Differential clustering

Flocks were clustered separately for each trait univariately based on their respective estimated genetic group effects (same as those for conventional clustering) using the FASTCLUS procedure of SAS (1990). A different **Q**-matrix was derived for each trait. The number of clusters formed was 5 for each trait to match conventional clustering. Breeding values and cluster effects for each trait were obtained in an animal model using bivariate analysis similar to the previous one except for differential cluster effects. In vector notation, the animal models fitted for each trait can be expressed as:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{Q}_1\mathbf{c}_1 + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_m\mathbf{m} + \mathbf{e}_1 \\ \mathbf{y}_2 &= \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{Q}_2\mathbf{c}_2 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}_2 \end{aligned}$$

where **y**, **X**, **b**, **u**, **m**, and **e** are as described above and subscripts 1 and 2 correspond to traits 1 (BWT) and 2 (FD), respectively; **Q**₁ and **Q**₂ are the **Q**-matrices for traits 1 (BWT) and 2 (FD), respectively; **Z**₁ and **Z**_m are the incidence matrices associating records on BWT in **y**₁ with random animal and maternal genetic effects in **u**₁ and **m**, respectively; **Z**₂ is the incidence matrix associating records on FD in **y**₂ with random animal effects in **u**₂; and **c**₁ and **c**₂ are the vectors of fixed cluster effects for traits 1 (BWT) and 2 (FD), respectively.

Validation of estimated breeding values (Regression analysis)

The EBVs (calculated as random animal plus fixed cluster effects) derived using conventional and differential clustering were then validated using the CTSE data which comprised progeny test performance records. The data were transformed to a percentage by dividing progeny records by their respective contemporary group means for each

trait as per routine Sheep Genetics evaluations (Brown *et al.*, 2006). The data were pre-adjusted for fixed effects of birth type, rearing type, age of animal at measurement, and linear and quadratic effects of dam age using OVIS (Brown *et al.*, 2000). The validation of sire EBVs (estimated from Sheep Genetics data) was achieved by regressing progeny performance (recorded in CTSE database) on the sire EBVs. Regression coefficients (averaged across contemporary groups) were derived in univariate analyses fitting a fixed effects model in ASReml (Gilmour *et al.*, 2006) which was as below:

$$y = \text{mean} + \text{CG} + b\text{EBV}_{\text{sire}}$$

where y is the observation on progeny for either trait, CG is the contemporary group effect for a trait, and b is a regression coefficient with an expectation of 0.5 as half of the difference in sire EBVs is transmitted to progeny. The summary statistics of the CTSE data used for regression analysis is shown in Table 7.3. The data were only a subset of those sires from the Sheep Genetics data which also had progeny performance recorded in the CTSE dataset. The progeny data in the CTSE dataset were recorded over years 1984 to 2003. For BWT, out of 159 sires whose EBVs were used for regression analysis, 47 sires had their own records in the Sheep Genetics dataset. For FD, all 193 sires whose EBVs were used for estimating regression coefficients had their own records in the Sheep Genetics dataset.

Table 7.3. Summary statistics of CTSE data used for estimating regression coefficients (transformed as per Sheep Genetics evaluations)

Trait	Number of records	Number of sires	Number of CGs ¹	Mean	Maximum	Minimum
BWT (Kg)	6,065	159	170	48.90	76.45	26.64
FD (micron)	6,165	193	227	18.43	24.22	14.33

¹ Contemporary groups

Effect of missing observations on genetic grouping and regression coefficients

The likely benefit of differential grouping would be most obvious if differences exist regarding selection history and the numbers of records available on different traits. Since industry data are characterized by varying amounts of missing records on a number of traits, the effect of missing observations on the correlation between EBVs obtained from conventional and differential genetic grouping models was examined. The aim of deleting data was to study its effect on the clustering mechanism for different traits. To achieve this, data on BWT were deleted from 8 flocks with the least numbers of BWT records. Flocks having less than or equal to 3,000 records on BWT were assumed to have missing records on BWT. Therefore, out of 15 flocks, only 7 flocks were having observations on BWT (39,709 records). The data on FD remained unchanged. The number of genetic groups formed remained the same for both traits, that is, 16. Flocks were clustered on the basis of their estimated genetic group effects (re-estimated using reduced data) using conventional and differential strategies as described previously. The (co)variance estimates were slightly different from those obtained using the complete dataset. Therefore, the $\mathbf{G}_o$ and $\mathbf{R}_o$ matrices used for reduced data were:

$$\mathbf{G}_o = \begin{pmatrix} 10.93 & 0.00 & 0.41 \\ 0.00 & 1.72 & 0.00 \\ 0.41 & 0.00 & 0.64 \end{pmatrix} \text{ and } \mathbf{R}_o = \begin{pmatrix} 14.59 & 0.45 \\ 0.45 & 0.89 \end{pmatrix}$$

7.3 Results and Discussion

7.3.1 Complete dataset

The estimated genetic group effects for the two traits without clustering flocks are shown in Figure 7.2. The estimated genetic group effects for the two traits after standardization are also shown in Table 1 of **Appendix**. Cluster allocation using conventional and differential strategies is shown in Table 7.4. The correlation between EBVs from conventional and differential strategies was 0.89 and 0.92 for BWT and FD, respectively. These correlations indicate that the cluster effects estimated using conventional and differential genetic grouping have a small impact on the EBVs for

BWT and FD. However, this finding will apply specifically to the dataset analyzed in this study. Data structure is an important issue for making such comparisons between the two strategies of genetic grouping. In certain cases, correlation is not always a very reliable measure for predicting the differences between evaluation models or procedures. Additionally, it is important to note that the conventional strategy of clustering leads to 5 clusters only, however with differential clustering, 5 clusters are obtained for each trait. Thus effectively more clusters are formed when groups are assigned differentially. Therefore, the statistical comparison of the two techniques of clustering based on the numbers of clusters becomes difficult. Hence, the comparison of conventional versus differential genetic grouping should only be undertaken on the basis of EBVs.

The across-flock correlation (for all animals including sires) between the EBVs obtained from two different strategies of clustering was 0.92 and 0.96 for BWT and FD, respectively. The scatter plot of sire EBVs for BWT derived from either strategy of genetic grouping is shown in Figure 7.3. The overlapping of the two types of EBVs in this graph and a high coefficient of determination ($R^2 = 0.92$) indicate the similarity between conventional and differential grouping strategies. A similar pattern was observed for FD ($R^2 = 0.93$; Figure 7.4). However, for sires in flock 7, the EBVs derived using the two techniques of genetic grouping were different for FD. This is evident from the parallel lines in Figure 7.4 which are due to the differences between estimated cluster effects using the two techniques. Those sires from flock 7 were allocated to cluster 1 (cluster effect of -1.55 micron) and cluster 3 (cluster effect of +1.61 micron) for FD using conventional and differential grouping, respectively (Table 7.4). The parallel lines indicate that some flocks or genetic groups are affected more than the others. Although the two strategies of genetic grouping are not significantly different for BWT (Figure 7.3), there is still a difference in the allocation of flocks to different clusters using either strategy (Table 7.4) which results in the difference between estimated cluster effects for both strategies of grouping.

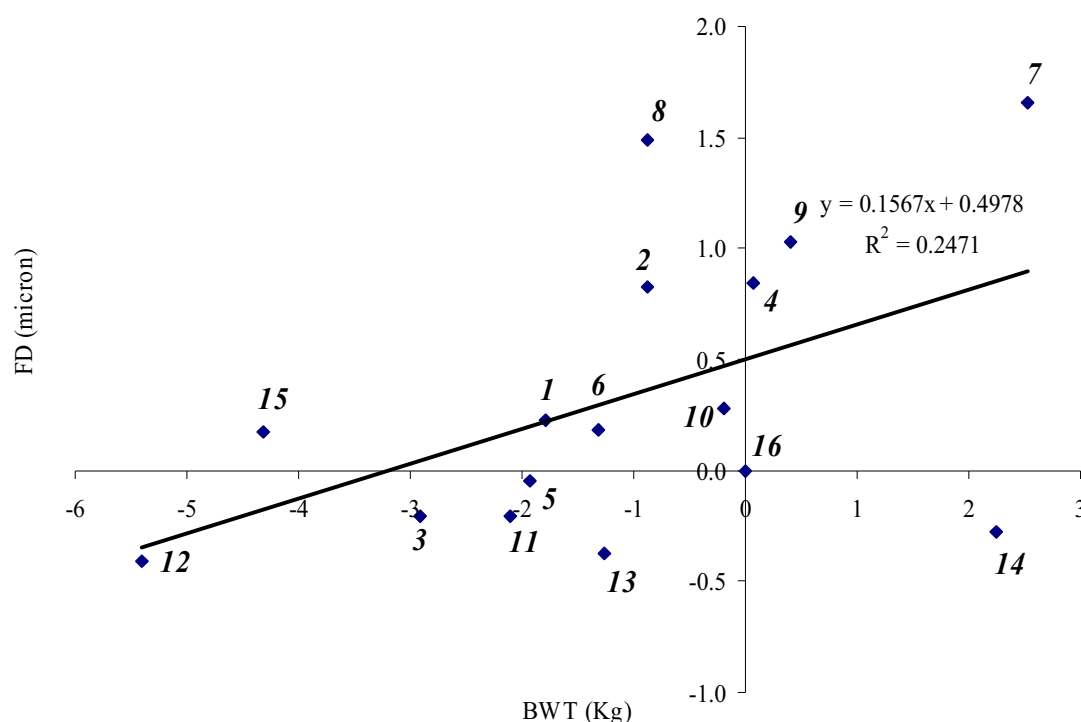


Figure 7.2. Scatter plot of estimated genetic group solutions (without standardization) for BWT and FD (numbers in bold and italics represent different genetic groups; complete dataset)

Table 7.4. Cluster allocation using conventional and differential genetic grouping (complete dataset)

Cluster ID	<u>Flock ID</u>		
	Conventional	Differential (BWT)	Differential (FD)
1	7	7, 14	2, 4, 9
2	2, 4, 8, 9	2, 4, 8, 9, 10, 16	1, 6, 10, 15
3	14	1, 3, 5, 6, 11, 13	7, 8
4	12, 15	12	12, 13, 14
5	1, 3, 5, 6, 10, 11, 13, 16	15	3, 5, 11, 16

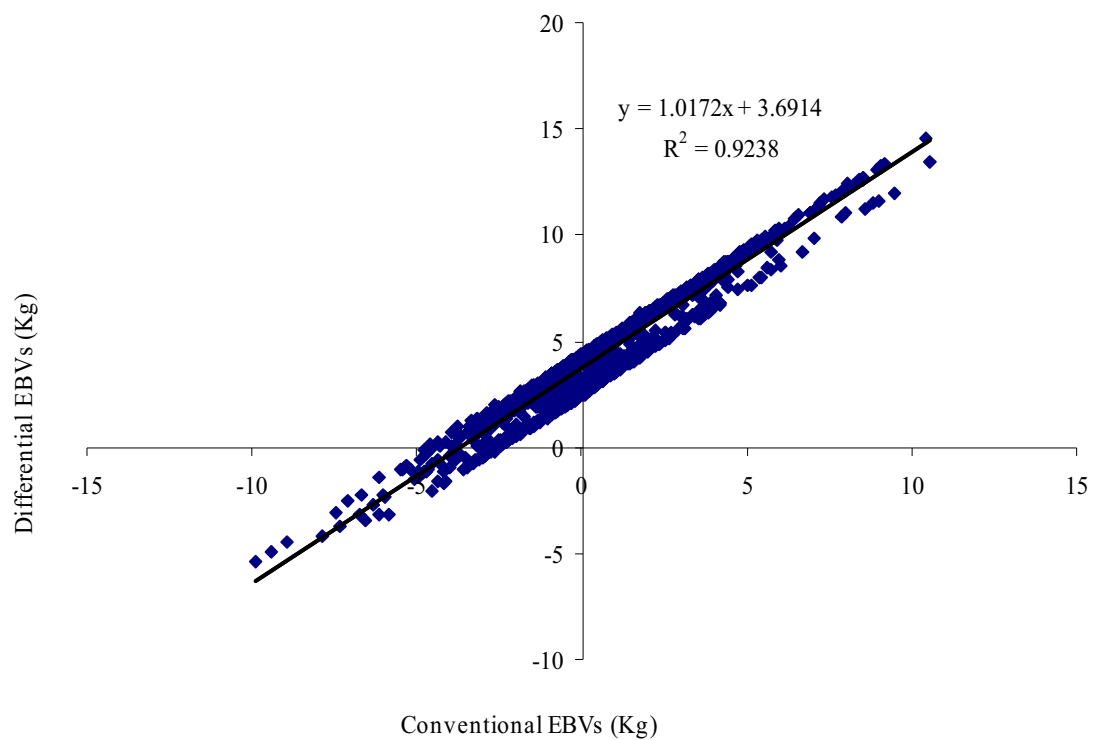


Figure 7.3. Scatter plot of sire EBVs for BWT from conventional and differential genetic grouping (complete dataset)

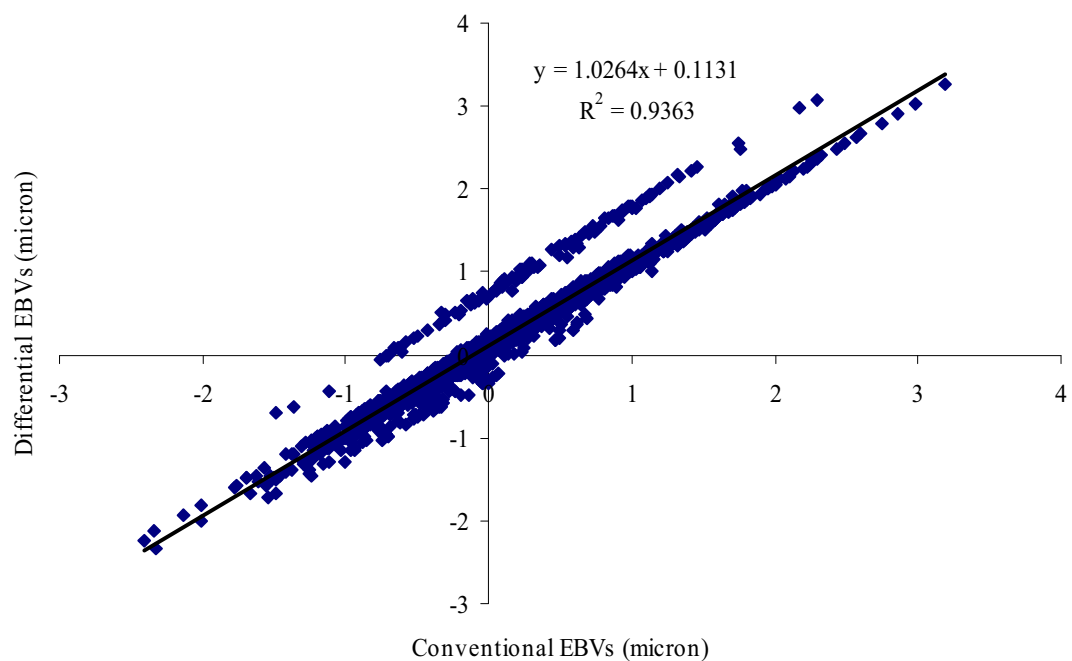


Figure 7.4. Scatter plot of sire EBVs for FD from conventional and differential genetic grouping (complete dataset)

7.3.1.1 Regression analysis using complete dataset

The estimated regression coefficient (averaged across contemporary groups) has an expectation of 0.5 as half of the difference in sire EBVs is transmitted to progeny. The estimated regression coefficients (derived by regressing progeny performance on sire EBVs) were similar for both strategies of genetic grouping (Table 7.5). The regression coefficients for BWT were much lower than those for FD (Table 7.5) for both strategies of genetic grouping. This difference in regressions could be due to the following two reasons. Firstly, the low levels of linkage between some flocks may have led to inaccurate genetic group solutions for those flocks which in turn may have reduced the accuracy of clustering. Secondly, the grouping strategy adopted and the model fitted may have biased the sire EBVs derived from Sheep Genetics data. The small differences in regressions for BWT between the two strategies of genetic grouping are due to the fact that out of 159 sires in the CTSE dataset whose EBVs for BWT were used in regression analysis, 54 sires were bred in the 16th genetic group. The 16th genetic group comprised all extra animals in the pedigree without any records on BWT implying that those 54 sires had no records of their own to accurately estimate their genetic group effects and hence their EBVs for BWT using the two strategies of genetic grouping. Additionally, out of 6,065 progeny records on BWT in the CTSE dataset (Table 7.3) that were used for estimating regressions, almost half the records (2,715) were on the progeny of those 54 sires originating in the 16th genetic group. The remaining records on BWT (3,350) were on the progeny of sires allocated to genetic groups 1 to 15.

Table 7.5. Regression coefficients ($\pm$ SE) derived using all sires for BWT and FD using conventional and differential clustering and without clustering (complete dataset)

Trait	<u>With clustering</u>		<u>Without clustering</u>
	Conventional	Differential	
BWT (Kg)	0.24 $\pm$ 0.03	0.22 $\pm$ 0.03	0.23 $\pm$ 0.02
FD (micron)	0.55 $\pm$ 0.02	0.57 $\pm$ 0.02	0.52 $\pm$ 0.02

Further analysis revealed that out of 159 sires whose EBVs were used for estimating regressions for BWT, 48 sires had less than 10 progeny records on BWT in the CTSE dataset. Additionally, only 114 sires had progeny records on BWT in the Sheep Genetics data and the remaining 45 sires had no records on their progeny for BWT. Further investigation also revealed that out of 114 sires with progeny records in the Sheep Genetics data, 47 sires had their own records on BWT. The accuracy of EBVs for BWT and FD was determined by the magnitude of associated standard errors. The standard errors of EBVs for BWT were in the range of 0.68 to 3.50 Kg (one-fifth to one standard deviation unit) for conventional and differential grouping. The standard errors of EBVs for FD were in the range of 0.08 to 0.79 micron (one-tenth to one standard deviation unit) for both conventional and differential grouping. Although the standard errors of EBVs for BWT were large due to larger variance than that for FD, the differences in the magnitude of standard errors between the two traits may indicate that the EBVs for FD were more accurate than those for BWT. This is attributed to the fact that out of 193 sires used for estimating regressions for FD, 180 sires had sufficient numbers of progeny records in the Sheep Genetics dataset. The EBVs of those 45 sires without progeny records on BWT derived from Sheep Genetics data had large standard errors (> 1.0 Kg) which were approximately one-half of the genetic standard deviation of 3.41 Kg and therefore were inaccurate. For FD, the number of sires with high accuracy of EBVs (conventional and differential) was 180 because all these sires had sufficient progeny records in Sheep Genetics data and therefore their EBVs (conventional and differential) for FD were more accurate than those for BWT. The manner in which clustering was undertaken did not significantly impact the regressions obtained for BWT and FD (Table 7.5). This indicates that either method of clustering could be used as an effective and alternative strategy for reducing the number of genetic group effects that need to be estimated in industry evaluations without compromising the accuracy of prediction.

The regression coefficients for FD were close to 0.5 for both strategies of genetic grouping. The results for both traits may suggest that conventional grouping is slightly better than differential. However, this difference in the two strategies is not significant based on the standard errors of regressions (Table 7.5). Using only flock

genetic groups was as accurate in this data as the clustering approach. The accuracy of predicting progeny performance through regression analysis using sire EBVs obtained from the two strategies of clustering depends on the number of progeny records in Sheep Genetics and CTSE datasets (Table 7.6). To further investigate the impact of numbers of progeny records on the accuracy of sire EBVs and hence the regressions obtained, different thresholds were set by using all sires, sires without any progeny records, and sires having at least or more than 20, 40, 60, 80, 100, and 120 progeny records each in the Sheep Genetics data. The threshold for CTSE data was set by using sires with at least 10 or more progeny records each. Therefore, only one threshold was used for the CTSE data. When a large number of sires without any progeny records in the Sheep Genetics dataset were used, the regressions obtained for BWT were significantly lower than expected and were different for conventional and differential clustering (Table 7.6). These sires were the animals whose EBVs were most affected by inaccurate genetic group solutions. Therefore, without and with clustering (conventional and differential), the regressions were significantly lower than expected for BWT (Tables 7.5 and 7.6).

Table 7.6. Regression coefficients ($\pm$ SE) for BWT and FD estimated using different threshold of Sheep Genetics data for conventional (Con) and differential (Diff) clustering (complete dataset)

Threshold ¹ (Records per sire)	BWT				FD			
	Sires	Records ²	Con	Diff	Sires	Records ²	Con	Diff
All sires	112	6,007	0.24 (0.03)	0.22 (0.03)	110	6,048	0.55 (0.02)	0.57 (0.02)
0 ³	37	2,069	0.21 (0.35)	0.13 (0.10)	11	1,328	0.55 (0.05)	0.57 (0.05)
20	55	3,041	0.32 (0.05)	0.29 (0.04)	90	4,958	0.54 (0.02)	0.57 (0.02)
40	40	2,307	0.28 (0.06)	0.23 (0.06)	70	4,194	0.53 (0.02)	0.55 (0.02)
60	28	1,593	0.23 (0.10)	0.20 (0.08)	52	3,458	0.50 (0.03)	0.53 (0.03)
80	18	1,015	0.20 (0.25)	0.15 (0.26)	49	3,342	0.48 (0.06)	0.54 (0.05)
100	10	695	0.79 (0.50)	0.62 (0.56)	41	3,071	0.46 (0.04)	0.48 (0.05)
120	8	637	1.32 (0.60)	1.26 (0.70)	31	2,518	0.48 (0.06)	0.47 (0.07)

¹ Sires without any progeny records³ and those with at least X or more progeny records in Sheep Genetics dataset; ² Number of progeny records corresponding to selected sires in CTSE dataset

When sires having more than 20 progeny records were used, there was an improvement in the regressions for BWT which were almost the same for the two strategies of clustering (Table 7.6). As there was an increase in the number of progeny records per sire in the Sheep Genetics dataset, there was an improvement in the regression coefficients for the two strategies of clustering. However, the standard errors associated with regressions were large because with an increase in the number of progeny records per sire in the Sheep Genetics dataset, there was a simultaneous decrease in the number of sires and hence the total records used for estimating regressions (Table 7.6). When sires with more than 100 progeny records each were used, the estimated regressions were significantly higher than the expected value of 0.5 due to a decline in the number of sires with those many records and hence the total records used. For all the thresholds tested for BWT, the two strategies of clustering were not significantly different (Table 7.6).

For all the thresholds tested for FD, conventional and differential grouping were not significantly different (Table 7.6). At 40 progeny records per sire in Sheep Genetics data, the regressions for FD were 0.53 and 0.55 for conventional and differential grouping, respectively (Table 7.6). However, with increasing progeny records per sire, there was a decrease in the number of sires and their progeny records which led to the deviation of regressions from the expected value of 0.5 with large associated standard errors.

7.3.2 Reduced dataset

The objective of deleting data on BWT was to examine its effect on conventional versus differential clustering for the two traits. The scatter plots of estimated genetic group effects for BWT and FD using complete and reduced datasets are shown in Figures 7.5 and 7.6, respectively. The bigger dots in Figure 7.5 represent those 8 flocks from which BWT data were removed. The estimated genetic group effects for FD were similar to those obtained from the complete dataset. The similarity in genetic group solutions for FD using complete and reduced datasets was due to the same data used for FD (Figure 7.6). The estimated genetic group effects for BWT for those 8 flocks from which data were deleted were either zeros (4 flocks) or non-zeros

with very large standard errors (4 flocks) and hence, were not accurate (Figure 7.5). The non-zero genetic group solutions were due to the fact that the animals whose data were deleted had relatives in other flocks with BWT recorded. These results show that in a multiple-trait evaluation involving a large number of flocks with varying genetic means where genetic group effects need to be fitted in the model, the type of data structure will affect the accurate estimation of genetic group effects for different traits using conventional and differential strategies. The difference between the two strategies of genetic grouping may be more marked (reflected in across-flock EBVs) when there are varying amounts of data on different traits.

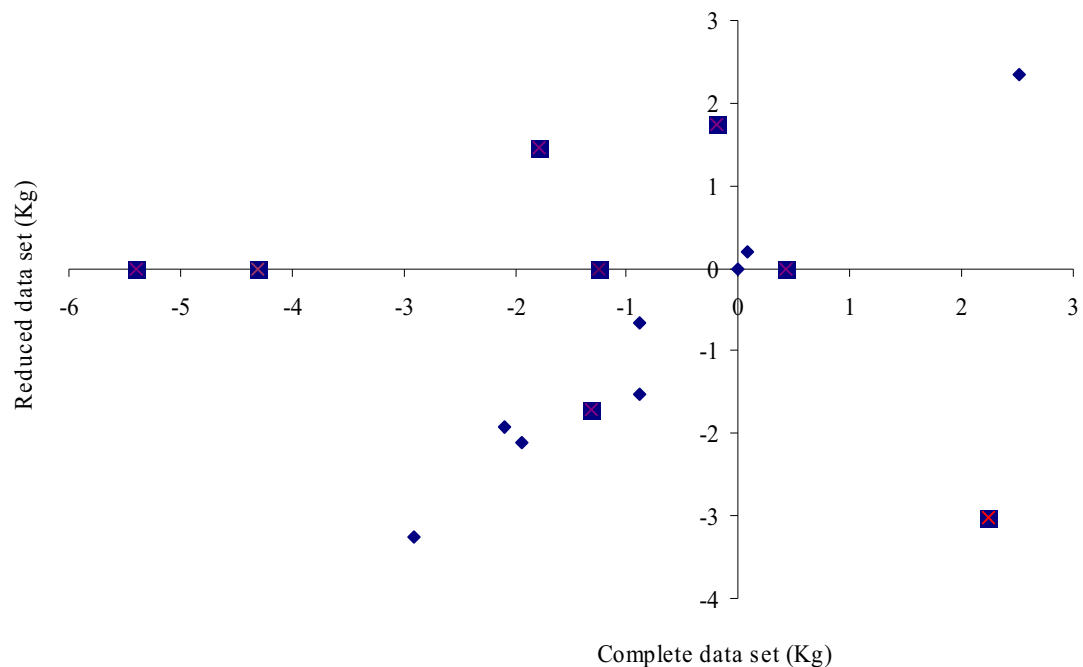


Figure 7.5. Scatter plot of estimated genetic group effects for BWT using complete and reduced datasets (bigger dots represent flocks from which BWT data were deleted)

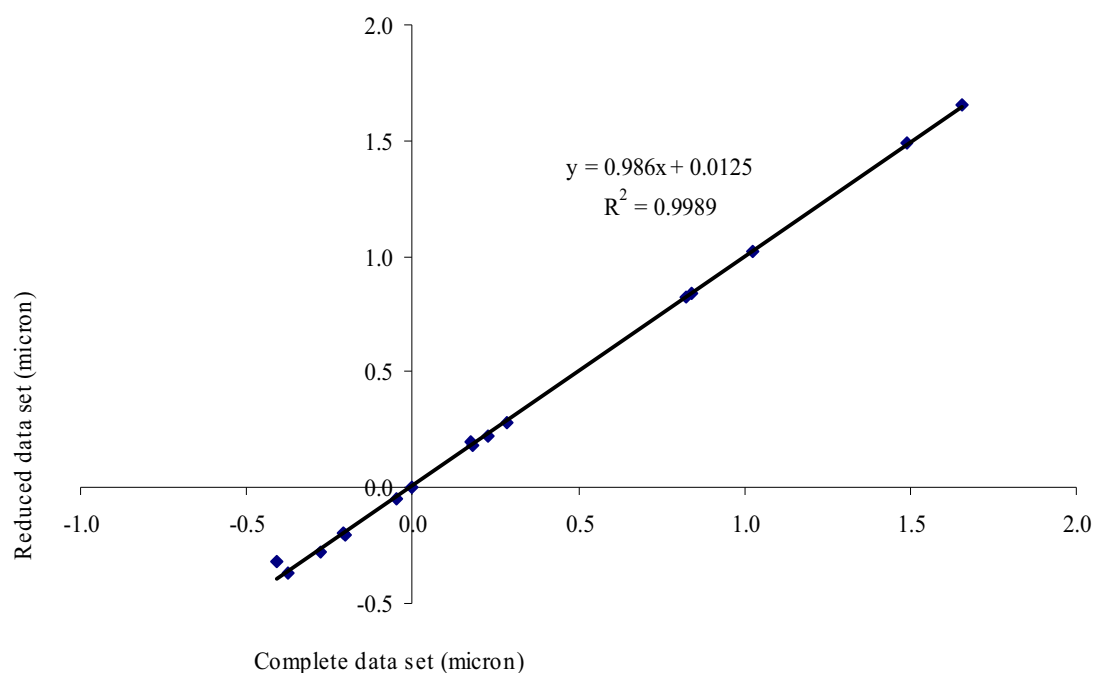


Figure 7.6. Scatter plot of estimated genetic group effects for FD using complete and reduced datasets

The variances of estimated genetic group effects were 2.66 Kg^2 and 0.41 micron^2 for BWT and FD, respectively. The corresponding estimates of variance were 4.29 Kg^2 and 0.42 micron^2 when complete data were used. There was a reduction in the variance of estimated genetic group effects of the two traits when data on BWT were deleted from certain flocks. This can be attributed to sampling effects as flocks from which data on BWT were deleted included two flocks with the lowest genetic group means. The across-flock correlations between EBVs derived from conventional and differential genetic grouping models for all animals were 0.80 and 0.82 for BWT and FD, respectively. The corresponding correlation for sires only was 0.75 and 0.78 for BWT and FD, respectively. The scatter plots of sire EBVs obtained from conventional and differential strategies for BWT and FD are presented in Figures 7.7 and 7.8, respectively. The values of R^2 are much lower than those obtained using a complete dataset (Figures 7.3 and 7.4). This reduction in correlation for BWT is due to reduced accuracy of estimated genetic group effects for BWT (due to deletion of data) which impacted cluster allocation. Although data on FD remained the same, still the correlation decreased because conventional clustering was based on the estimated

genetic group effects of both BWT and FD. Thus there is some evidence that the alternative hypothesis is correct implying that differential genetic grouping has more effect on allocation of flocks to clusters when traits have different amounts of data.

The wider spread in the EBVs shown in Figures 7.7 and 7.8 indicate that varying cluster effects are being added to the animal solutions using conventional and differential clustering. There is a clear difference in the EBVs obtained using the two strategies of genetic grouping for BWT and FD. This is manifested in the differences in the patterns observed in Figures 7.3 and 7.7 for BWT and in Figures 7.4 and 7.8 for FD. Hence, when some flocks have no observations for a trait, there is a difference between the two strategies of genetic grouping.

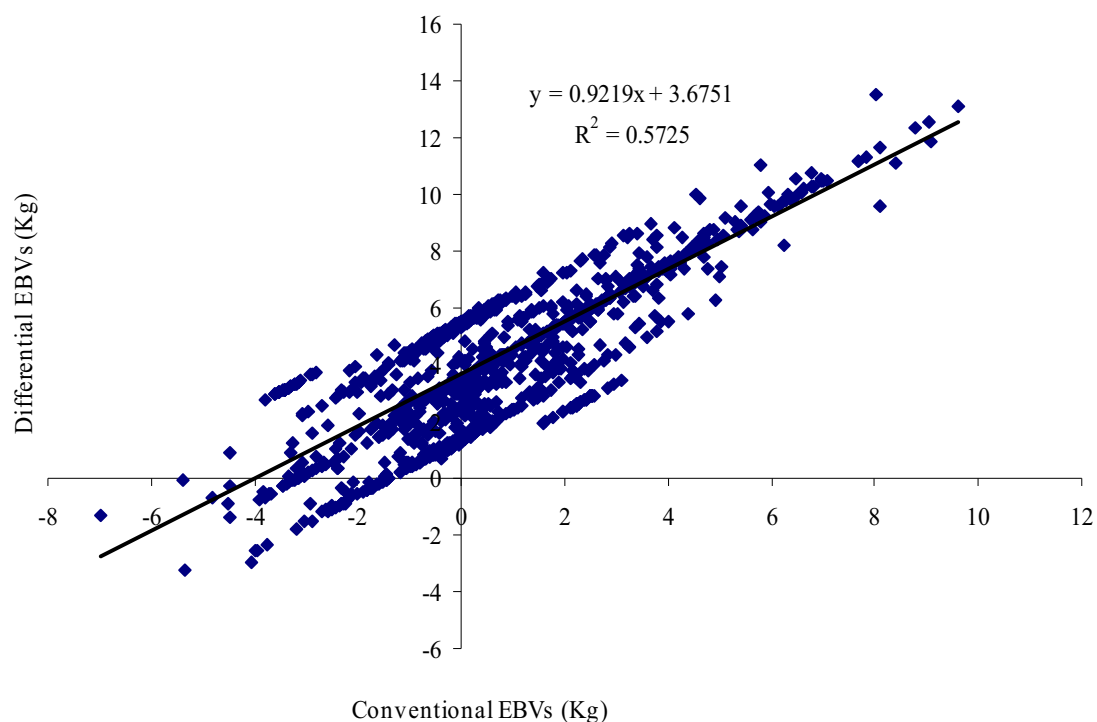


Figure 7.7. Scatter plot of sire EBVs for BWT from conventional and differential genetic grouping (reduced dataset)

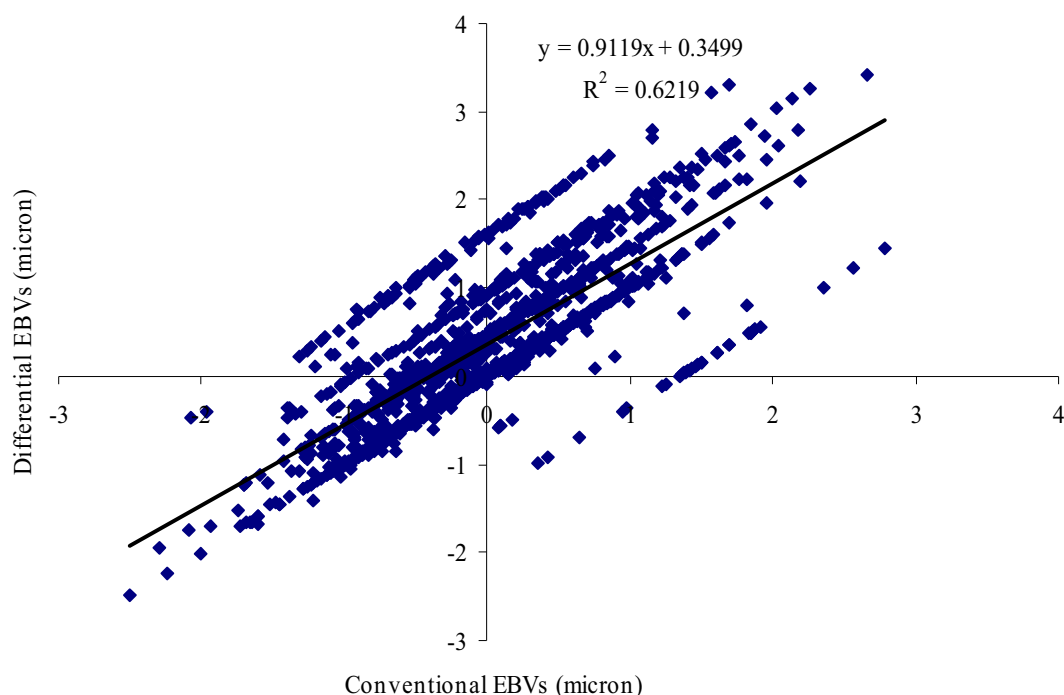


Figure 7.8. Scatter plot of sire EBVs for FD from conventional and differential genetic grouping (reduced dataset)

7.3.2.1 Regression analysis using reduced dataset

The regressions for BWT were significantly different for both strategies of genetic grouping (Table 7.7). Differential genetic grouping is less accurate than conventional for BWT. The low regression coefficient for BWT using differential strategy is due to assigning of flocks to varying clusters based on their estimated genetic group effects for BWT only including effects of those flocks from which data were deleted. The deletion of data on BWT reduced the accuracy of sire EBVs which in turn affected the regression obtained for differential grouping. These results emphasize the importance of the amount of data on each trait that is used for the estimation of genetic group and subsequent cluster effects (estimated genetic group effects from clustered flocks) in a multi-trait evaluation.

Table 7.7. Regression coefficients ($\pm$ SE) derived using all sires for BWT and FD for conventional and differential clustering and without clustering (reduced dataset)

Trait	<u>With clustering</u>		<u>Without clustering</u>
	Conventional	Differential	
BWT (Kg)	0.29 ± 0.04	0.16 ± 0.03	0.25 ± 0.03
FD (micron)	0.50 ± 0.02	0.55 ± 0.02	0.52 ± 0.02

For FD, the regression was lower and closer to 0.5 for conventional than for differential genetic grouping when all sires were used (Table 7.7). This is because flocks were allocated to different clusters (Table 7.8) for FD under conventional and differential clustering which results in different estimated cluster effects being added to the EBVs for FD. Since data for FD remained the same, differential genetic grouping for FD is almost the same whether complete or reduced datasets were used except for clusters 4 and 5 (Tables 7.4 and 7.8). Different flocks have been assigned to clusters 4 and 5 under differential genetic grouping for FD using complete and reduced data. However, the difference between the estimated cluster effects of clusters 4 and 5 is the same for complete and reduced datasets.

Table 7.8. Allocation of flocks to different clusters using conventional and differential genetic grouping (reduced dataset)

Cluster ID	<u>Flock ID</u>		
	Conventional	Differential (BWT)	Differential (FD)
1	7	2	2, 4, 9
2	1, 10	1, 7, 10	1, 6, 10, 15
3	2, 4, 8, 9	5, 6, 8, 11	7, 8
4	3, 5, 6, 11, 14	4, 9, 12, 13, 15, 16	5, 16
5	12, 13, 15, 16	3, 14	3, 11, 12, 13, 14

The conclusions drawn in this study regarding the two strategies of clustering are dependent largely on regressions and correlations from within-cluster relationships amongst animals which are very strong, as expected. However, the between-cluster

relationships are different for each cluster size. For example, the differential strategy for FD using complete dataset (Table 7.4) is very similar to that for the reduced dataset (Table 7.8). Using either dataset, three clusters (Cluster ID 1, 2 and 3) include the same flocks and the other two (Cluster ID 4 and 5) are very similar in their estimated cluster effects. The same data were used for the estimation of genetic group effects for FD. However, the plots depicted in Figures 7.4 (complete data) and 7.8 (reduced data) for FD are significantly different. When complete dataset (Figure 7.4) is used, the between-cluster deviations from the zero mean for conventional versus differential clustering are less than 1 micron whereas for reduced dataset (Figure 7.8), the between-cluster deviations are in the range of -1 to 1.5 micron. This again shows that when conventional clustering was undertaken using reduced data (Table 7.8), flocks were allocated to clusters different from those using complete data for both traits (Table 7.4). This is attributed to the fact that deletion of data affected the magnitude of estimated genetic group effects for both traits which in turn affected the between-cluster variance thereby impacting cluster allocation for both traits using conventional grouping.

The correlation between EBVs derived using the two strategies of genetic grouping for the 159 sires used for regression analysis for BWT was 0.98 when complete dataset was used. However, when reduced dataset was used, the corresponding correlation for BWT was 0.85. This decrease in correlation is due to the fact that some of the sires (out of 159) belonged to those flocks from which data on BWT were deleted. Similarly, the corresponding correlations obtained using complete and reduced datasets for the 193 sires used for estimating regressions for FD were 0.99 and 0.94, respectively. These correlations are high as data remained the same for FD in both the cases. These results show that the 193 sires used for regression analysis for FD are not affected by the change in genetic grouping strategy (conventional versus differential) nor the removal of data on BWT. The results also indicate that the need for genetic grouping decreases with an increase in the number of progeny records available. The latter conclusion will be true only when sire EBVs are considered. Pollak and Quaas (1983) showed that the differences between genetic groups decrease with an increase in the information included in the numerator relationship matrix. In other words, the genetic group solutions become less relevant if an animal has many progeny records.

For sires having a reasonable number of progeny with performance records, the impact of estimated genetic group effects on the EBVs is small. For animals with only own performance records, the estimated genetic group solutions have a much higher weight on the resulting EBVs. However, the EBVs of 159 sires for BWT are influenced more by the two strategies of genetic grouping.

The same progeny thresholds as described under 7.3.1.1 for the complete dataset for investigating the effect of the numbers of progeny records each sire had in the Sheep Genetics data on the accuracy of regression coefficients was also tested for the reduced dataset (Table 7.9). When 37 sires without any progeny records on BWT were used, the conventional strategy of clustering was better than the differential based on the estimated regressions (Table 7.9). This is attributed to the fact that even though there were no progeny observations on BWT to accurately estimate their genetic group solutions still those sires were allocated to clusters based on the correlated information on FD, own records, and those on their relatives when conventional clustering was undertaken. However, when differential clustering was used, the accuracy of cluster allocation and therefore that of cluster solutions was significantly reduced due to the low accuracy of genetic group solutions for BWT.

Table 7.9. Regression coefficients ($\pm$ SE) for BWT and FD estimated using different threshold of Sheep Genetics data for conventional (Con) and differential (Diff) clustering (reduced dataset)

Threshold ¹ (Records per sire)	<u>BWT</u>				<u>FD</u>			
	Sires	Records ²	Con	Diff	Sires	Records ²	Con	Diff
All sires	112	5,619	0.30 (0.04)	0.17 (0.03)	110	6,048	0.52 (0.02)	0.56 (0.02)
0 ³	37	2,069	0.25 (0.05)	0.10 (0.03)	11	1,328	0.55 (0.05)	0.57 (0.05)
20	48	2,941	0.33 (0.06)	0.22 (0.05)	90	4,958	0.54 (0.02)	0.58 (0.02)
40	36	1,879	0.27 (0.08)	0.21 (0.07)	70	4,194	0.49 (0.03)	0.51 (0.03)
60	23	1,256	0.20 (0.09)	0.24 (0.08)	52	3,458	0.50 (0.03)	0.53 (0.03)
80	13	893	0.19 (0.22)	0.14 (0.24)	49	3,342	0.47 (0.05)	0.45 (0.06)
100	10	695	1.32 (0.60)	1.26 (0.71)	41	3,071	0.44 (0.09)	0.45 (0.10)
120	6	379	1.18 (0.70)	1.15 (0.72)	31	2,518	0.47 (0.12)	0.46 (0.11)

¹ Sires without any progeny records³ and those with at least X or more progeny records in Sheep Genetics dataset; ² Number of progeny records corresponding to selected sires in CTSE dataset

At the threshold of 60 progeny records per sire, differential clustering was slightly better than conventional for BWT however the standard errors of regression coefficients were large. For FD, although the data remained the same, still the regressions for conventional clustering were slightly different in comparison to those derived using the complete dataset (Table 7.6). This is due to the fact that under conventional strategy, flocks were clustered on the basis of estimated genetic group effects of both BWT and FD. For the different thresholds examined for FD, conventional clustering was not significantly different from differential based on the regressions shown in Table 7.9.

At different progeny thresholds examined, the number of sires whose progeny records were used for estimating regressions for BWT were almost similar for complete and reduced datasets (Tables 7.6 and 7.9). This shows that deletion of BWT data from some flocks did not significantly reduce the number of sires used for estimating regressions and therefore the numbers of progeny records and regressions were very

similar for BWT (except at the threshold of 100 progeny records per sire) for complete and reduced datasets (Tables 7.6 and 7.9). At the threshold of 100 progeny records per sire, the regressions for BWT were significantly different for both conventional and differential strategies although the numbers of sires (10 sires) and the numbers of their progeny records (695 records) were the same for complete and reduced datasets (Tables 7.6 and 7.9). This is attributed to the differences in the EBVs (conventional and differential) of those 10 sires derived from complete and reduced data. The differences in both EBVs (10 sires) could be due to the differences in the information used on relatives as a result of deletion of data. The differences in the EBVs could also be due to the differences in cluster allocation using either grouping strategy. However, the standard errors were large and therefore the regressions were not accurate for both strategies of genetic grouping using complete and reduced data (Tables 7.6 and 7.9). These results show that the flocks from which data on BWT were removed were generally not the flocks with sires whose progeny performance for BWT were recorded in the CTSE dataset. For FD, the regressions and their associated standard errors were similar for both conventional and differential grouping using complete and reduced datasets due to exactly the same sires and the same numbers of their progeny records (Tables 7.6 and 7.9).

The graphs in Figures 7.9 to 7.11 show the scatter plots of EBVs for BWT for different types of (159 sires used for estimating regressions for BWT) sires from conventional clustering model and those from differential clustering model as described below (complete dataset):

1. Sires which have progeny in their own flock only (13 sires)
2. Sires having progeny in one flock only (122 sires)
3. Sires having progeny in more than one flock (24 sires)

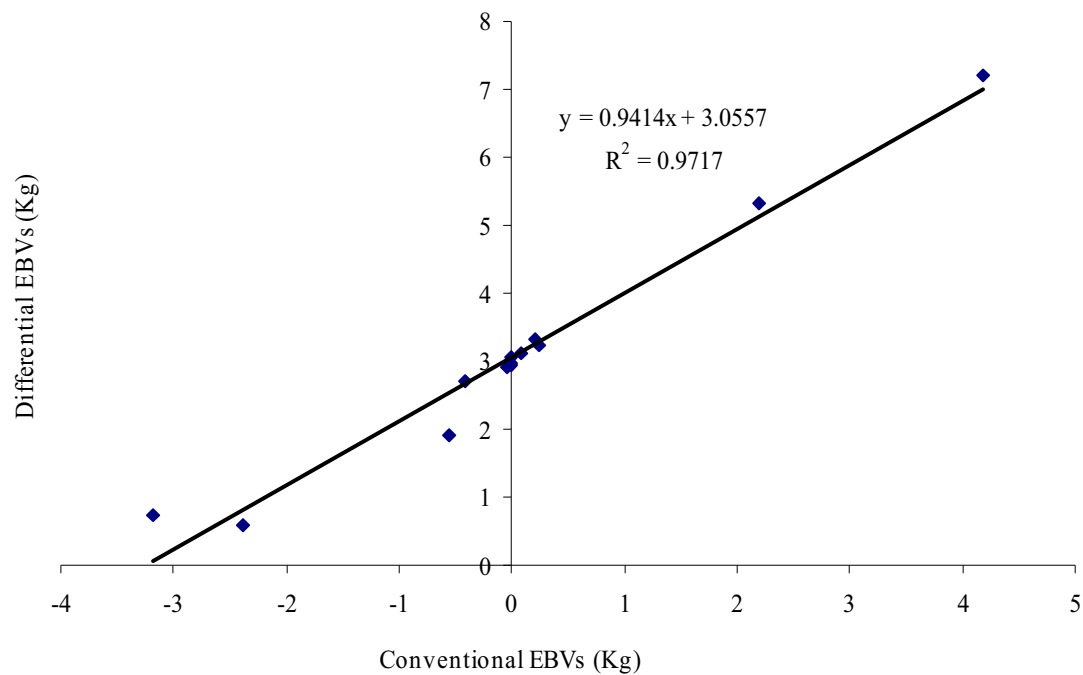


Figure 7.9. Scatter plot of sire EBVs for BWT from conventional and differential clustering models (sires having progeny in their own flocks only; complete dataset)

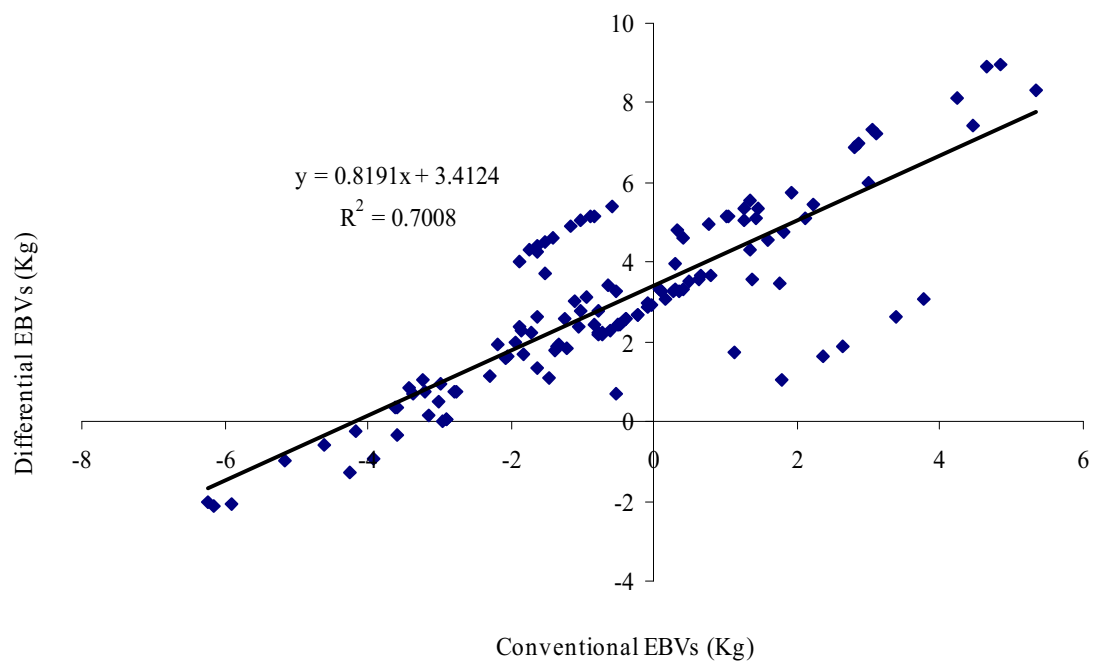


Figure 7.10. Scatter plot of sire EBVs for BWT from conventional and differential clustering models (sires having progeny in one flock only; complete dataset)

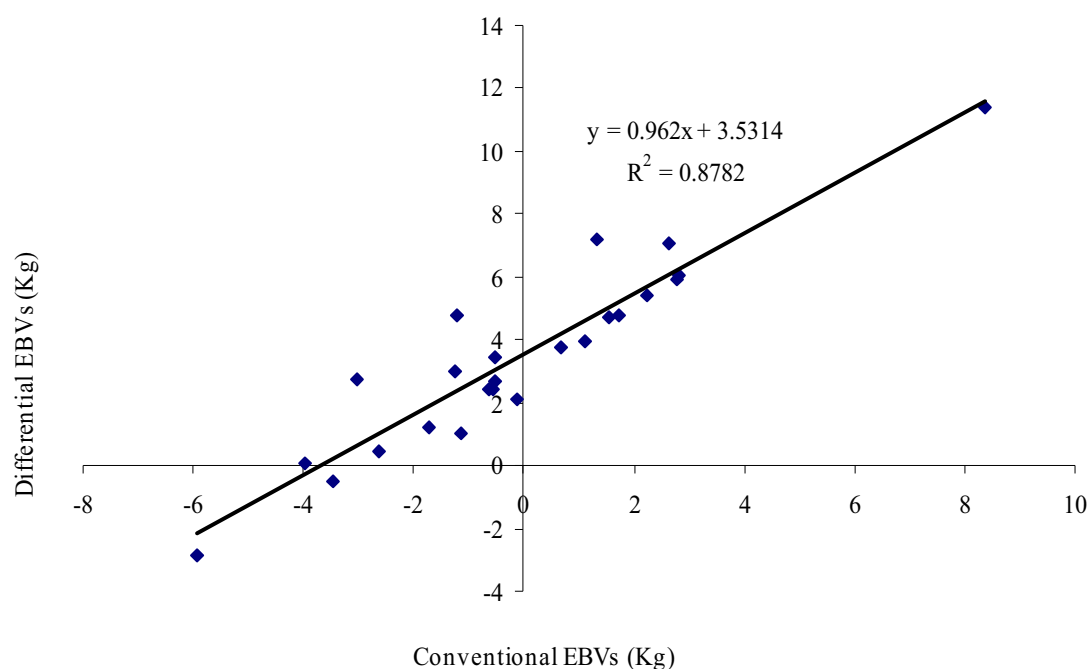


Figure 7.11. Scatter plot of sire EBVs for BWT from conventional and differential clustering models (sires having progeny in more than one flock; complete dataset)

The variance in the difference of EBVs for BWT (obtained from conventional and differential genetic grouping) of different types of sires described above varies with the type of sires examined with minimum variance for animals which have progeny only in their own flocks (Figure 7.9). This is due to the fact that only within-flock variation (more influence of their genetic group effect) has been accounted for in their EBVs. This affects the variance of their within-flock EBVs. For sires which have progeny only in their own flocks, the coefficient of determination (R^2) is 0.97 (Figure 7.9) implying that for these animals the EBVs derived using either strategy of genetic grouping are similar. For sires with progeny in one flock only (Figure 7.10) and for those having progeny in more than one flock (Figure 7.11), the variation in EBVs and the ranking of animals are also different. The variation in EBVs is more for sires having progeny in only one flock. This is the reason for a low R^2 (0.70) for such sires. For these sires, the EBVs obtained using different strategies of genetic grouping are different. For sires having progeny in more than one flock, the spread in EBVs is slightly more (Figure 7.11).

7.4 Conclusions and Implications

The results obtained in this study indicate that conventional and differential genetic grouping strategies for the dataset analyzed in this study are not significantly different from each other. When reduced data were used, the two strategies of genetic grouping were still not different although cluster allocations using either strategy were different. The regressions obtained were very similar for both conventional and differential grouping. The type and amount of data available on different traits will affect genetic group and breeding value estimation using the two strategies of genetic grouping. Depending on the amount of data available for a trait(s), the number of genetic groups allocated for each trait, and the genetic variance of a trait(s), conventional or differential genetic grouping could be adopted. However, the statistical comparison of the two techniques of genetic grouping will be slightly complex due to the fact that conventional generated five clusters combined for the two traits whereas differential produced five clusters each for a trait. The criterion of similarity (Euclidean distance) desired between flocks with respect to their estimated genetic group effects was not specified in the FASTCLUS procedure. Depending on the Euclidean distance specified prior to clustering, different patterns of clusters will be observed for a given data structure. Therefore, the level of similarity desired between flocks will affect the clustering outcome which may lead to significant differences between the two strategies of genetic grouping.

The manner in which data were deleted in this study was not the best way of examining the effect of missing observations on the EBVs obtained through conventional and differential strategies because a large number of sires used for the regressions were generally not in the flocks from which data on BWT were deleted. Additionally, missing data on traits from some flocks may not necessarily affect evaluation for those traits in other flocks. However, as shown in the current study, this will affect the accuracy of genetic group solutions for flocks with missing data which in turn will impact the clustering outcome. Alternatively, to overcome the problem of missing data, genetic groups and clusters could be treated as random effects especially if adequate genetic links exist between flocks. Treating genetic groups and clusters as random effects would also improve the accuracy of genetic group and cluster solutions

for smaller flocks. Pieramati and Van Vleck (1993) reported that the impact of genetic grouping (when treated as fixed effects) on estimated additive genetic variance fluctuates depending on the amount and type of missing information (for example, percentage of data deleted and the pattern of deletion). However, the authors found that when genetic groups were treated as random effects, the estimates of additive genetic variance were unaffected.

The way clustering was undertaken in this study could have impacted the results obtained as only estimated genetic group effects without their associated standard errors were used as an input for the FASTCLUS procedure. There were large standard errors associated with estimated genetic group solutions for some flocks (or genetic groups) depending on the amount of data for each trait, the amount of pedigree information, and the proportion of genetic links with other flocks. The lack of genetic links between some flocks contributed to the large standard errors associated with their estimated genetic group effects which could have impacted the accuracy of clustering. Moreover, one of the limitations of the FASTCLUS procedure is that the pattern and the number of clusters formed for a given dataset depends on the initial random assignment of each data point to a cluster. Although different numbers of clusters were not tested in the current study, the expectation is that similar results would be observed at each level of clusters formed for a given number of genetic groups. Furthermore, the number of clusters tested for each grouping strategy could have impacted the results observed in this study. However, the actual number of clusters that should be formed for a given data structure remains a subjective issue. Since the objective of this study was to compare the ability to predict progeny performance using sire EBVs from conventional and differential genetic grouping strategies, genetic groups were not included in the model fitted to the CTSE data for estimating regression coefficients.

In the data analyzed in this study, the level of linkage between flocks and the size of flocks varied. Additionally, genetic groups and clusters were treated as fixed effects. If genetic group and cluster effects are treated as random, differences in the genetic group variance of traits could impact the clustering outcome if flocks are clustered based on their estimated genetic group effects. Further research should examine the impact of conventional and differential grouping on the accuracy of multi-

trait EBVs when genetic group and cluster effects are treated as random especially if significant differences exist in the genetic group variance of different traits. For those flocks from which data were deleted, the estimated genetic group solutions were either zeros or were different from zero due to information on relatives in other flocks. Hence, the estimated genetic group effects of traits which are not correlated will be different due to differences in the amount of data available on each trait. Further research could also examine the effect of these two strategies of genetic grouping on multi-trait evaluation involving traits that are not genetically correlated. Additionally, since industry data were used in this study, it was difficult to distinguish between the two strategies of genetic grouping because TBVs (True breeding values) were unknown. In this study, data were deleted only on one trait (BWT). Further studies can use simulation to compare these two strategies of genetic grouping and to investigate the impact of amount and pattern of deletion of data on different traits on the accuracy of multi-trait EBVs derived from conventional and differential grouping.

This study failed to show any potential benefit of differential over conventional grouping in terms of regression analysis. This could be due to the pattern of removal of BWT data from some flocks and the set of sires (with progeny performance recorded in CTSE data) that was used to compare the two methods of genetic grouping. Additionally, the model used and the grouping strategy (based on flock) adopted could have affected the results obtained and the conclusions drawn. Moreover, estimability of cluster effects was not considered which would depend on the proportion of genetic links between flocks having similar estimated genetic group effects. However, if adequate linkage does not exist between some flocks or if the size of flocks is small, their estimated genetic group solutions may not be accurate although clustering would allocate those flocks to the same large group or cluster. Therefore, in that case, it would be better not to cluster flocks based on their estimated genetic group effects. It is important to note that without clustering flocks, each animal's EBV will contain weighted contributions from different genetic groups whereas when flocks are clustered, each animal's EBV will contain weighted contributions from different clusters. The former will depend on the accuracy of genetic group solutions which in turn will depend on the level of linkage between flocks and the size of flocks. However, the latter will

Chapter 7. Conventional versus differential genetic grouping in multiple-trait evaluation

depend on the accuracy of estimated cluster solutions, the numbers of clusters formed and the level of similarity desired between flocks in their estimated genetic group effects.

Chapter 8. General Discussion

The aim of this research was to examine the significance of linkage and genetic grouping, and to develop alternative and better strategies of genetic grouping in relation to across-flock evaluation for the Australian Merino sheep industry. Different types of analyses were undertaken using simulated and industry data to address this objective and mixed model equations accounting for differential genetic grouping in an animal model multi-trait evaluation were presented. In Chapter 2, different approaches (**Q**-matrix and **W**-matrix) for including genetic groups in the model were reviewed along with various aspects associated with genetic grouping like linkage, fixed versus random groups, impact of genetic grouping on estimated variance components, effect of genetic grouping on the accuracy of across-flock Merino sire evaluation in Australia, data structure required for accurately estimating genetic group effects, etc.

Chapter 3 examined the significance of linkage and genetic grouping for an accurate partitioning of genetic and environmental differences between flocks. The accuracy of prediction of genetic group effects when treated as fixed was evaluated. The results obtained have shown that genetic linkage in conjunction with including genetic groups in the model are a prerequisite for an accurate partitioning of genetic and environmental effects in an across-flock evaluation. Genetic group effects can already be partitioned from flock environmental effects in the presence of 1% (proportion of progeny generated from common link sires in each flock) linkage, however with low accuracy of genetic group solutions. For the population structure simulated in this study, at least 25% linkage from known sires (with enough data to estimate their genetic group effects) is required for obtaining genetic group solutions and hence, across-flock EBVs with acceptable accuracy. However, the proportion of linkage required also depends on the magnitude of real genetic differences between flocks. To derive genetic group solutions and hence across-flock EBVs with an accuracy of 80%, the standard error associated with the estimated difference (SED) between genetic groups should be less than one-third or 33% of the genetic standard deviation. If the level of accuracy desired in the genetic group solutions is 70%, the SED needs to be less than one-half of the genetic standard deviation. However, the choice regarding the accuracy of EBVs may

vary across breeders. In the absence of adequate linkage between flocks, genetic group and fixed effects included in the model such as flock will be confounded.

The level of heritability and linkage impact the accuracy of genetic group solutions. The effect of heritability on the accuracy of estimated genetic group effects appears to diminish with an increase in the proportion of genetic links between flocks. The number of progeny generated from each link sire has less effect on the standard errors of estimated genetic group effects. The standard errors of estimated genetic group effects are more a function of the total number of progeny derived from link sires. Establishing linkage through sires from a known genetic group is more beneficial than from sires without any information regarding their genetic group effect. In other words, for creating genetic links between flocks, sires from those groups with adequate data are better than those from groups without sufficient data for an accurate estimation of their genetic group effects.

A prior test should be used to examine whether genetic groups need to be fitted in the model. The criterion developed by Kennedy (1981) could be used for this purpose. At a given level of linkage, if the estimated difference between two genetic groups is smaller than or equal to the associated standard error then genetic groups could be ignored in the model. On the contrary, if the estimated difference between two genetic groups is greater than the associated standard error then genetic groups need to be included in the model. However, this criterion cannot be implemented to large scale Sheep Genetics Merino evaluations where a large number of genetic groups are assigned due to differences in the numbers of records available on each trait and the level of genetic links between flocks. For trait(s) with adequate data, the standard error of the difference between two genetic groups will be less than the actual difference. On the contrary, for traits with less effective records, the standard error of the difference between two genetic groups will be more than the actual difference. Therefore, the need for grouping based on Kennedy's (1981) approach could be erroneous for Sheep Genetics evaluations. Furthermore, for some flocks with suboptimal linkage, the accuracy of estimated genetic group solutions will be reduced and therefore the standard error associated with the difference between two genetic groups would be larger than the real difference.

The results obtained in Chapter 3 also indicate that the need for genetic grouping increases with an improvement in the level of linkage. This is due to the fact that when more progeny are generated in each group from common link sires, the differences in the group means still exist (due to unlinked progeny) which can only be accounted for by including genetic groups in the model. If the differences between genetic groups are not significantly large, then accounting for group effects in the model could be counterproductive (Kennedy and Moxley, 1975; Phocas and Laloe, 2004). Another important issue that requires attention is that the performance of link rams could vary depending on the environment in which their progeny are raised implying genotype by environment interaction (James and Cottle, 1997). The genotype by environment interaction could be a significant source of variation for some traits and insignificant for the others. Further research should investigate this issue in more detail using industry data from Sheep Genetics. However, modeling of such effects would be a tedious task due to the number of parameters that need to be estimated. This study treated genetic groups as fixed effects because only three groups were involved. However, considering large numbers of genetic groups defined in routine Sheep Genetics evaluations in OVIS (Brown *et al.*, 2000), the option of treating group effects as random is possible. Therefore, additional studies should determine the amount of linkage required for the accurate partitioning of genetic and environmental differences between flocks in across-flock evaluations when genetic groups are treated as random effects.

The effect of flock size on the accuracy of estimated genetic group effects and genetic group variance was explored in Chapter 4. The results obtained indicate that the accuracy of estimation of genetic group effects depends on the size of groups. The higher the number of animals with data in a genetic group, the greater is the accuracy of estimation, and vice-versa. Genetic group effects can be estimated with higher statistical power by increasing the numbers of animals measured in each group (Woolaston, 1997). The prior assumptions regarding (co)variances of genetic groups may not necessarily be true for industry evaluations where animals with records are linked to large numbers of base populations and subpopulations that are genetically diverse and unrelated. Additionally, Huisman and Brown (2007) have shown significant differences between various strategies of genetic grouping on the estimates of genetic group

variance. Therefore, the grouping strategy required will depend on the data structure. It can also be concluded that the accurate estimation of genetic group variance in across-flock evaluation will also depend on the manner in which genetic groups are defined. Alternatively, genetic group variance could be estimated using simulated and industry data using different strategies or structures of genetic groups so that the optimal structure could be identified.

The results obtained in Chapter 4 have shown that averaged over linkage, the difference in the accuracy of estimation (measured in terms of variance of difference between simulated and estimated genetic group effects) between small and large flocks was statistically significant ($P < 0.05$) whether 26 or 51 group effects were estimated. However, averaged over flock size, the difference in the accuracy of estimation was not significant between any two levels of linkage although the trend was that with increasing linkage the accuracy of prediction of genetic group effects improved. This could be due to the magnitude of differences across genetic groups and the structure of the data simulated. For example, the simulated genetic variance within each flock or genetic group and the genetic group variance between all flocks or genetic groups was the same which may not be true for field data. The difference in correlation between TBVs and EBVs for different flock sizes when fitting 26 versus 51 group effects was not statistically significant. However, the trend was that with increasing linkage and flock size, the difference in correlation (26 versus 51 groups) decreased. Within each category of groups (26 or 51), the difference in correlation between TBVs and EBVs (averaged over linkage) was not significantly different. Similarly, within each category of groups (26 or 51), the difference in correlation between TBVs and EBVs (averaged over flock sizes) was not significant. Therefore, the proportion of linkage and the size of flocks required for accurately estimating genetic group effects would vary with the data structure.

In Chapter 5, different cluster sizes (number of clusters) were compared for clustering flocks on their estimated genetic group effects with the objective of reducing the numbers of group effects that need to be estimated in a model. Clustering flocks on their estimated genetic group effects is an effective method that could be used in Sheep Genetics evaluations. However, the results obtained and the conclusions drawn in this

study are likely to be dependent on the structure of the data simulated. Additionally, the Euclidean distance between genetic group means of flocks varied for all cluster sizes tested. This could have impacted the pattern of clusters formed because only those flocks whose estimated genetic group effects satisfied the criterion of similarity (Euclidean distance) were allocated to the same unique cluster. However, if the Euclidean distance is varied, flocks assigned to a cluster may be allocated to another cluster and therefore the between-cluster variance may get altered.

In the current study, the issues regarding linkage and estimability of genetic group effects were not considered. However, in industry evaluations, lack of genetic links between flocks and/or inestimability for a large number of small flocks need to be considered prior to clustering. If the proportion of linkage is sub-optimal and/or genetic group solutions are less reliable, the accuracy of clustering will be affected. Thus, the accuracy of estimated genetic group effects should be evaluated prior to clustering flocks. The actual number of clusters required for a specific scenario will depend on several factors like the number of flocks, differences in pedigree recording across flocks, genetic differences between flocks, accuracy of genetic group solutions and the proportion of linkage between flocks. In the present study, the proportion of linkage between clusters and the estimability of cluster effects were not taken into consideration. This could have led to erroneous results because certain flocks that were not adequately linked may have been allocated to the same cluster based on their respective estimated genetic group effects which in turn could have affected the accuracy of estimated cluster effects. Therefore, further research should explore the impact of linkage between clusters and hence the estimability of cluster effects on the accuracy of clustering flocks based on their estimated genetic group solutions.

The model used for analyzing simulated data in Chapter 5 treated genetic groups and clusters as random effects. Clustering flocks based on their estimated genetic group effects implies that flocks with comparable genetic level are assigned to the same cluster. Although not examined in this study, treating clusters as fixed effects could be a better option especially if the proportion of genetic links between flocks is low. However, as seen in the current study, clustering impacts the between-cluster variance depending on the numbers of clusters formed. Therefore, the between-cluster variance is

generally more than the between-flock variance. Hence, when clusters are treated as random effects, the assumptions regarding the (co)variance structure for between-flock effects will need to be altered according to the number of clusters formed. Thus, it appears that treating clusters as random effects in industry evaluations may lead to erroneous results if limited genetic links exist across flocks. Future studies should compare fixed versus random cluster models using simulated and industry data to evaluate the impact of linkage on the accuracy of clustering mechanism.

Another important issue that warrants attention is the fact that only genetic group solutions were used as an input for clustering flocks without taking into consideration their standard errors. This implies that for those flocks (depending on their size) whose genetic group effects were estimable without sufficient accuracy may have impacted the accuracy and therefore the pattern of clusters formed. However, the FASTCLUS procedure does not allow for the option of including standard errors of genetic group solutions. Hence, further research could focus on the development of algorithms and procedures that would allow for this option. Additionally, for the data simulated in Chapter 5, the optimum cluster size could not be identified. Therefore, the actual number of clusters required for a given data structure to achieve maximum level of accuracy needs further investigation.

Chapter 6 investigated the option of allocating genetic groups differentially for different traits using the mixed model methodology. The modified mixed model equations presented can be used to assign genetic groups differentially for various traits in Sheep Genetics evaluations. This would provide an alternative strategy of grouping animals in a multi-trait evaluation especially with varying amounts of missing data across traits. Groups can be assigned differentially through **Q**-matrix approach and **W**-matrix approach. Both methods provide identical solutions, however deriving coefficients for the **Q**-matrix could be tedious for large scale evaluations especially if the pedigree involves large numbers of distinct and overlapping generations. The **W**-matrix approach appears to be relatively simple to implement in industry evaluations. This study examined the feasibility of accounting for differential genetic groups for additive genetic effects only in a multi-trait evaluation. Further research could explore

the options for assigning differential genetic groups for both direct and maternal genetic effects in a multi-trait evaluation.

For some traits, maternal genetic effects are significant, therefore the choice of the correct genetic groups model for evaluation is a critical one. The choice of the best model could be determined from the data structure. If a large proportion of data on dams from past generations is missing, then groups should be assigned differentially for maternal effects. However, this type of choice will also be dependent on the fact whether the traits to be evaluated have a high or low genetic correlation. Additionally, the issue of (co)variances between genetic groups of different traits merits further attention when differential grouping is adopted. However, the actual test of application of various strategies of genetic grouping using industry data will be a challenging task.

Chapter 7 compared conventional versus differential genetic grouping in multiple-trait evaluation using data from Sheep Genetics. When complete dataset was used, no differences were observed between conventional and differential genetic grouping. However, when reduced dataset was used, the differences between the two strategies of genetic grouping became apparent. Although cluster allocations were different using conventional and differential grouping, the regressions obtained were similar. In multi-trait clustering, allocations are more influenced by the traits exhibiting larger variance. When clustering flocks for different traits on the basis of their genetic group solutions, the underlying mechanism of clustering is dependent both on the accuracy and the magnitude of genetic group solutions. Therefore, the estimated genetic group effects for different traits need to be standardized prior to clustering. In this study, the default criterion or distance measure for clustering was not manipulated for either trait so that the number of clusters could be varied. This could have affected the pattern of clusters observed. However, if the criterion or measure of distance is varied, the pattern and the number of clusters formed will get affected.

In Chapter 7, the size and structure of Sheep Genetics data used for estimating genetic group effects and sire EBVs, and that of CTSE data used for regression analysis could have impacted the results obtained and the conclusions drawn. Future studies should test the application of the two strategies (conventional and differential) of

genetic grouping on real data in large scale industry evaluations. Simulated data could also be used to investigate whether the two techniques of genetic grouping are significantly different. The results obtained in this study indicate that clustering did not impact the regression coefficients (regression of sire EBVs on progeny records) estimated for the two traits. The low regressions obtained for body weight using either type of grouping can be attributed to insufficient levels of linkage between some flocks thereby leading to inaccurate genetic group solutions. The grouping strategy adopted and the model used for deriving sire EBVs from Sheep Genetics data could have impacted the regressions for body weight. Furthermore, significant numbers (54 out of 159) of sires whose progeny records were used for estimating regressions originated in the genetic group which comprised animals without any data. Therefore, the genetic group solutions for those sires and hence their EBVs were biased thereby leading to low regressions for body weight. The relevance of genetic grouping will decrease if a large number of records are available on the progeny of animals of interest.

Depending on factors like the phenotypic variance, heritability, and numbers of records available, estimated genetic group effects of different traits could have varying levels of accuracy in industry evaluations. Hickey *et al.* (2008) evaluated Irish beef cattle data and found large differences in the accuracy of genetic group solutions for different traits and breeds. For some traits clustering would be more accurate than for the others. The traits analyzed in the current study (BWT and FD) are generally well recorded and are characterized by long selection history. Further research should investigate evaluation of more than two traits which are characterized by differences in selection history using industry data by assigning genetic groups differentially. Additionally, future research could also extend differential genetic grouping to maternal effects and its subsequent application in multi-trait evaluation of industry data. The differences between the two strategies of genetic grouping tested in Chapter 7 were not significant when complete data were used. However, the differences became more apparent when reduced dataset was analyzed implying that the numbers of records available on different traits relative to each other will determine the choice of grouping strategy required.

If the number of records available on a trait with a recent selection history like WEC are sufficient to represent genetic groups, conventional grouping can be used. Alternatively, if a large number of observations on some traits are missing, separate analyses should be conducted, one for important production traits and the other for WEC, as is currently undertaken by Sheep Genetics in Australia (Huisman and Brown, 2007; Brown *et al.*, 2007). Nevertheless, the information on performance data will always be varying for traits of interest because of differences in breeding objectives and market requirements. Therefore, when significant differences exist in the numbers of records on different traits, differential grouping strategy could be used as an alternative. Hence, a threshold regarding the number of observations required for each trait should be determined in addition to other factors like the structure of the data and the variation exhibited by each trait below which differential genetic grouping needs to be adopted. Further research can focus on investigating these issues in more detail using Sheep Genetics data.

As in Chapter 5, in Chapter 7 also the standard errors of estimated genetic group solutions were not considered for clustering flocks which could have affected the accuracy of the number and the pattern of clusters obtained. In the analysis undertaken in Chapter 7, a total of five clusters were formed for the two traits when animals were grouped with the conventional strategy. When differential grouping was adopted, five clusters were formed for each trait. Therefore, accurate statistical comparison of the two strategies of genetic grouping becomes difficult due to the difference in the numbers of clusters formed.

The results obtained and the literature reviewed in this thesis indicate that the choice of an evaluation model (without and with groups) should be made depending on certain criteria like the proportion of linkage between flocks, the amount and structure of the data available, and the magnitude of genetic differences between flocks. The ongoing advancement in the field of molecular genetics techniques like DNA fingerprinting and identification of genetic markers in various livestock species for traits of interest may assist in solving the problem of missing pedigree information. This would reduce the need for accounting for genetic groups in the model. The reduction in the cost associated with DNA fingerprinting over the next few years could lead to

significant changes in pedigree recording systems in the Australian Merino industry (Barnett *et al.*, 1997). The estimation of breeding values using accurate pedigree information across years will also be advantageous in the estimation of genetic trend with improved accuracy. In addition, the uptake of DNA fingerprinting by commercial breeders will also facilitate the implementation of marker assisted selection for a range of production traits to maximize genetic progress (Barnett *et al.*, 1997). Although the advancements in the development of better models and algorithms have significantly contributed to the improvement in the accuracy of genetic evaluation systems, the need for grouping exists because livestock data in most cases has missing pedigree and there is no substitute for completely and accurately recording the data and the pedigree information.

Chapter 9. General Conclusions

In routine genetic evaluation of animals, genetic groups are generally included in the model to account for unknown genetic merit of base animals. In Merino sheep evaluations in Australia, genetic groups are used to account for differences in the genetic means of base populations of different flocks. Chapter 3 focused on the disentangling of genetic and flock environmental effects in across-flock evaluations. Linkage and genetic grouping are required in conjunction for an accurate partitioning of genetic and environmental differences between flocks. At least 25% linkage is needed for accurately estimating genetic group effects and across-flock breeding values. To derive estimated genetic group effects and hence across-flock estimated breeding values with a minimum of 80% accuracy (measured in terms of correlation between true breeding values and estimated breeding values), the standard error of the difference associated with the estimated difference between the two genetic groups should be approximately one-third of the genetic standard deviation. If the accuracy desired in genetic group solutions is 70% then the standard error of difference needs to be at least one-half of the genetic standard deviation. In the absence of linkage and genetic groups in the model, genetic group effects will be confounded with flock environmental effects.

Data structure is an important issue that needs to be considered prior to estimating genetic group effects in across-flock evaluations in the Merino sheep industry. However, the limitations imposed by Merino data, genetic differences between flocks, and different strains and bloodlines will always remain a challenge to an unbiased across-flock evaluation for Australian Merinos. The level of heritability and linkage affect the accuracy of estimation of genetic group effects. The impact of heritability on the accuracy of estimated genetic group effects decreases with an increase in the level of linkage between flocks. The number of progeny generated per link sire has relatively less impact on the standard errors of estimated genetic group solutions. The standard errors of genetic group solutions are more a function of the total number of progeny derived from link sires.

The impact of flock size on the accuracy of estimated genetic group effects and genetic group variance in across-flock evaluations was investigated (Chapter 4). The results show that the accuracy of estimation of genetic group effects and genetic group variance is dependent on the size of genetic groups and the proportion of linkage between genetic groups. The size of genetic groups should be large for precisely estimating their effects. Although in the current study, the difference in the accuracy of prediction for small and large flocks was not significant, larger flocks and genetic groups exhibited higher accuracy of prediction. Making prior assumptions regarding covariances of genetic group effects would be erroneous as it would depend on the selection history which varies with each dataset. Alternatively, the covariances between genetic groups could be estimated from field data.

Clustering flocks on the basis of their estimated genetic group effects could be effectively adopted in Sheep Genetics evaluations without compromising the accuracy of prediction (Chapter 5). The accuracy of clustering strategy will depend on the accuracy of estimated genetic group effects. Further studies are required incorporating the standard errors of genetic group solutions in the clustering algorithms so that flocks having large standard errors could be regressed more towards the average of group solutions. The estimability of genetic group and subsequent cluster effects (estimated genetic group effects from clustered flocks) based on the level of linkage between flocks should also be considered prior to the actual implementation of the technique of clustering to industry evaluations of Merino data by Sheep Genetics.

Mixed model equations for an animal model multi-trait evaluation accounting for differential genetic grouping for traits characterized by differences in data recording and selection history were presented in Chapter 6. Differential genetic grouping is feasible and can be incorporated in the evaluation procedure through modified mixed model equations. The **Q**-matrix and **W**-matrix approaches used for differential genetic grouping provide identical solutions. The strategy of differential genetic grouping could be extended to routine Sheep Genetics evaluations in OVIS.

A comparison of conventional and differential strategies of genetic grouping was performed in a multi-trait evaluation using industry data from Sheep Genetics (Chapter

7). When complete data were used, both strategies produced similar results in terms of prediction of progeny performance. However, when a reduced dataset was used, some differences were observed between the two techniques of genetic grouping although it could not be determined which technique was better in terms of regression analysis. Therefore, depending on the numbers of records available on different traits, conventional or differential genetic grouping could be adopted. When large numbers of records on some traits are missing, differential grouping should be used. The relevance of genetic groups in the evaluation scheme will decrease with an increase in the number of progeny records available.

Some aspects of industry data were not considered in the experimental analyses undertaken in this research. For example, multiple and overlapping generations and differential grouping criteria for direct and maternal genetic effects. Further research should investigate the impact of genetic grouping on the accuracy of evaluation procedure taking these factors into consideration. Simulated data were used to examine the effect of linkage (established through the same link sires) on the accuracy of estimated genetic group effects. Additionally, genetic groups were assigned based on the flock of origin. The minimum linkage required for the accurate estimation of genetic group effects could vary if many different link sires are used for creating genetic links between flocks. The amount of linkage required may also vary when genetic groups are allocated based on combinations of criteria like flock, year of birth, etc. The actual proportion of linkage needed for an unbiased across-flock evaluation will also depend on the magnitude of real genetic differences between flocks.

The improvement in the models of genetic evaluation would lead to further development of current techniques used for accounting for missing data and pedigree information. The advancements in molecular genetics techniques like DNA finger printing and gene mapping may in future reduce the need for grouping animals. However, animals will need to be grouped as base populations and subpopulations are expected to have differences in their mean genetic values. Furthermore, data and pedigree information are missing in most livestock datasets used for genetic evaluation.

BIBLIOGRAPHY

- Alfonso, L. and Estany, J. (1999). An expression of mixed animal model equations to account for different means and variances in the base population. *Genetics Selection Evolution* **31**, 105-113.
- Amar, P.S., Singh, B., Poonia, J.S. (1998). Genetic divergence among different grades of sheep. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production* **24**, 67-70.
- Amores, B.G., Hinch, G.N., Mortimer, S.I., Sivarajasingam, S. (1999). Genotype x environment interactions for lamb traits of merino sheep. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **13**, 110-113.
- Ashtiani, S.R.M. and James, J.W. (1991). Efficient use of link rams in merino sire reference schemes. *Proceedings of Australian Association of Animal Breeding and Genetics* **9**, 388-391.
- Ashtiani, S.R.M. and James, J.W. (1992). Optimum distribution of progeny in sire referencing schemes. *Proceedings of Australian Association for the Advancement of Animal Breeding and Genetics* **10**, 476-479.
- Ashtiani, S.R.M. and James, J.W. (1993). Optimum family size for progeny testing in populations with different strains. *Genetics Selection Evolution* **25**, 339-352.
- Atkins, K., Gilmour, A.R., Thompson, R., Coelli, K.A., Casey, A.E. (1998). Prospects for across-flock genetic evaluation in Australian merino sheep. *Wool Technology and Sheep Breeding* **46**, 1-11.
- Atkins, K., Casey, A.E., Coelli, K.A. (1999a). Sire evaluation for merino ram breeders. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **13**, 138-143.

- Atkins, K., Mortimer, S.I., Casey, A.E. (1999b). Do we need to include genetic groups to improve the precision of merino sire evaluation? *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **13**, 58-61.
- Banks, R.G. (1997). Implications of across-flock/herd evaluations for breeders and service providers. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **12**, 784-787.
- Barnett, N.L., Purvis, I.W., Stewart, V.A.M., Franklin, I.R. (1997). At what price is DNA pedigreeing cost effective for Merino breeders? *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **12**, 425-428.
- Barnett, N.L., Swan, A.A., Purvis, I.W. (1998). Benefit cost analysis of recording dam pedigrees in merino sheep breeding programs. *Wool Technology and Sheep Breeding* **46**, 119-123.
- Brash, L.D. (1999). Diminished discrimination in merino genetic evaluations. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **13**, 102-105.
- Brown, D.J., Tier, B., Reverter, A., Banks, R., Graser, H.-U. (2000). OVIS: A multiple trait breeding value estimation program for genetic evaluation of sheep. *Wool Technology and Sheep Breeding* **48**, 285-297.
- Brown, D.J., Reverter, A., Tier, B. (2001a). Influence of environmental factors and trait representation on the genetic evaluation of reproductive traits in sheep. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **14**, 131-134.
- Brown, D.J., Tier, B., Banks, R. (2001b). Implications of data quality on the accuracy of estimated breeding values for weaning weight in sheep. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **14**, 441-444.
- Brown, D.J., Ball, A.J., Huisman, A.E., Swan, A.A., Atkins, K.D., Graser, H.-U., Banks, R., Swan, P., Woolaston, R.R. (2006). Sheep Genetics Australia: a national

- genetic evaluation system for Australian sheep. *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, Communication No. 05-03.
- Brown, D.J., Huisman, A.E., Swan, A.A., Graser, H.-U., Woolaston, R.R., Ball, A.J., Atkins, K.D., Banks, R.G. (2007). Genetic evaluation for the Australian sheep industry. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **17**, 187-194.
- Cantet, R.J.C., Schaeffer, L.R., Smith, C. (1992a). Reduced animal model with differential genetic grouping for direct and maternal effects. *Journal of Animal Science* **70**, 1730-1741.
- Cantet, R.J.C., Fernando, R.L., Gianola, D., Misztal, I. (1992b). Genetic grouping for direct and maternal effects with differential assignment of groups. *Genetics Selection Evolution* **24**, 211-223.
- Cantet, R.J.C., Gianola, D., Misztal, I., Fernando, R.L. (1993). Estimates of dispersion parameters and of genetic and environmental trends for weaning weight in Angus cattle using a maternal animal model with genetic grouping. *Livestock Production Science* **34**, 203-212.
- Cardoso, F.F. and Tempelman, R.J. (2003). Bayesian inference on genetic merit under uncertain paternity. *Genetics Selection Evolution* **35**, 469-487.
- Cardoso, F.F. and Tempelman, R.J. (2004). Genetic evaluation of beef cattle accounting for uncertain paternity. *Livestock Production Science* **89**, 109-120.
- Casellas, J., Piedrafita, J., Varona, L. (2007). Bayes factor for testing between different structures of random genetic groups: A case study using weaning weight in *Bruna dels Pirineus* beef cattle. *Genetics Selection Evolution* **39**, 39-53.
- Casey, A.E. and Hygate, L.C. (1992). Selection practices and management systems used by NSW Merino ram breeders. *NSW Agriculture Animal Industries Report No. 2*, p.22.

- Casey, A.E. and Atkins, K.D. (1999). 'Merino-benchmark'- developing on-stud across-flock performance evaluation. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **13**, 165-168.
- Clement, V., Bibe, B., Verrier, E., Elsen, J., Manfredi, E., Bouix, J., Hanocq, E. (2001). Simulation analysis to test the influence of model adequacy and data structure on the estimation of genetic parameters for traits with direct and maternal effects. *Genetics Selection Evolution* **33**, 369-395.
- Cottle, D.J. and Hickson, J. (1996). Using industry-recommended merino link sires to benchmark on-farm performance. *Wool Technology and Sheep Breeding* **44**, 10-16.
- Cottle, D.J. and James, J.W. (1997). Linkage in the merino central test sire evaluation schemes. *Wool Technology and Sheep Breeding* **45**, 108-119.
- Culbertson, M.S., Mabry, J.W., Misztal, I., Bertrand, J.K., Nugent, R. (1998). Influence of genetic groups on estimates of genetic parameters and inbreeding depression in swine. *Journal of Animal Science* **76** (Supplement 1), 67 (Abstract).
- Curi, R.A., Oliveira, H.N., Gimenes, M.A., Silveira, A.C., Lopes, C.R. (2005a). Effects of CSN3 and LGB gene polymorphisms on production traits in beef cattle. *Genetics and Molecular Biology* **28**, 262-266.
- Curi, R.A., Oliveira, H.N., Silveira, A.C., Lopes, C.R. (2005b). Effects of polymorphic microsatellites in the regulatory region of IGF1 and GHR on growth and carcass traits in beef cattle. *Animal Genetics* **36**, 58-62.
- Da, Y., Grossman, M., Misztal, I. (1989). Prediction error variance and restricted maximum likelihood estimation for animal model with relationship grouping. *Journal of Dairy Science* **72**, 2125-2135.
- Da, Y. and Grossman, M. (1991). Multitrait animal model with genetic groups. *Journal of Dairy Science* **74**, 3183-3195.

- Dempfle, L. (1990). Problems in the use of the relationship matrix in animal breeding. In: *Advances in Statistical Methods for Genetic Improvement of Livestock* (Eds. Gianola, D., Hammond, K.), Springer-Verlag, pp. 454-473.
- Diaz, C., Moreno, A., Carabano, M.J. (2002). Effects of genetic groups on estimates of variance components for a maternally influenced trait in the Spanish Limousine. *Journal of Animal Breeding and Genetics* **119**, 325-334.
- Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* **3**, 32-57.
- Eady, S.J., O'Rourke, P.K., Connelly, P.T. (1991). Comparison of wool production and reproductive rate of South Australian strain Merinos and locally bred Peppin Merinos in the semi-arid tropics of Queensland. *Australian Journal of Experimental Agriculture* **31**, 737-743.
- Famula, T.R., Pollak, E.J., Van Vleck, L.D. (1983). Genetic groups in dairy sire evaluation under a selection model. *Journal of Dairy Science* **66**, 927-934.
- Famula, T.R. (1985). An equivalence between models of restricted selection and genetic groups. *Theoretical and Applied Genetics* **71**, 413-416.
- Famula, T.R. (1992). Simple and rapid inversion of additive relationship matrices incorporating parental uncertainty. *Journal of Animal Science* **70**, 1045-1048.
- Famula, T.R. (1993). The contribution of progeny of uncertain paternity to the accuracy of sire evaluation. *Journal of Animal Science* **71**, 1136-1141.
- Fernando, R.L., Gianola, D., Grossman, M. (1983). Identifying all connected subsets in a two-way classification without interaction. *Journal of Dairy Science* **66**, 1399-1402.
- Fikse, W.F. (2003). Fuzzy classification of phantom parent groups in an animal model. *Proceedings of the 54th Annual Conference of the European Association of Animal Production*, Rome.

- Fouilloux, M., Clement, V., Laloe, D. (2008). Measuring connectedness among herds in mixed linear models: From theory to practice in large-sized genetic evaluations. *Genetics Selection Evolution* **40**, 145-159.
- Foulley, J.L., Schaeffer, L.R., Song, H., Wilton, J.W. (1983). Progeny group size in an organized progeny test program of AI beef bulls using reference sires. *Canadian Journal of Animal Science* **63**, 17-26.
- Foulley, J.L., Gianola, D., Planchenault, D. (1987). Sire evaluation with uncertain paternity. *Genetics Selection Evolution* **19**, 83-102.
- Foulley, J.L., Thompson, R., Gianola, D. (1990a). On sire evaluation with uncertain paternity. *Genetics Selection Evolution* **22**, 373-376.
- Foulley, J.L., Bouix, J., Goffinet, B., Elsen, J.M. (1990b). Connectedness in genetic evaluation. In: *Advances in Statistical Methods for Genetic Improvement of Livestock* (Eds. Gianola, D., Hammond, K.), Springer-Verlag, pp. 277-308.
- Foulley, J.L., Hanocq, E., Boichard, D. (1992). A criterion for measuring the degree of connectedness in linear models of genetic evaluation. *Genetics Selection Evolution* **24**, 315-330.
- Fries, L.A. (1998). Connectability in beef cattle genetic evaluation: the heuristic approach used in MILC.FOR. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production* **27**, 449-450.
- Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J., Thompson, R. (2006). ASReml User Guide Release 1.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.
- Golden, B.L., Bourdon, R.M., Snelling, W.M. (1994). Additive genetic groups for animals evaluated in more than one breed association national cattle evaluation. *Journal of Animal Science* **72**, 2559-2567.

- Hanford, K.J., Thallman, R.M., Kachman, S.D., Van Vleck, L.D. (2004). Including genetic groups for QTL effects in marker assisted selection. *Journal of Animal Science* **82** (Supplement 1), 414 (Abstract).
- Hanocq, E., Boichard, D., Foulley, J.L. (1996). A simulation study of the effect of connectedness on genetic trend. *Genetics Selection Evolution* **28**, 67-82.
- Hanocq, E. and Boichard, D. (1999). Connectedness in the French Holstein cattle population. *Genetics Selection Evolution* **31**, 163-176.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-447.
- Henderson, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69-83.
- Henderson, C.R. and Quaas, R.L. (1976). Multiple trait evaluation using relatives' records. *Journal of Animal Science* **43**, 1188-1197.
- Henderson, C.R. (1988a). A simple method to account for selected base populations. *Journal of Dairy Science* **71**, 3399-3404.
- Henderson, C.R. (1988b). Use of an average numerator relationship matrix for multiple-sire joining. *Journal of Animal Science* **66**, 1614-1621.
- Heyer, L.J., Kruglyak, S., Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Research* **9**, 1106-1115.
- Hickey, J.M., Keane, M.G., Kenny, D.A., Cromie, A.R., Mulder, H.A., Veerkamp, R.F. (2008). Estimation of accuracy and bias in genetic evaluations with genetic groups using sampling. *Journal of Animal Science* **86**, 1047-1056.
- Hickson, J.D. (1997). Using national sire evaluation results in merino breeding programs. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **12**, 705-708.

- Hu, X. and Yoo, I. (2004). Cluster ensemble and its applications in gene expression analysis. Paper presented at: *The 2nd Asia-Pacific Bioinformatics Conference*, Dunedin, New Zealand.
- Huisman, A.E., Tier, B., Brown, D.J. (2005). Multiple trait linkage across flocks. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **16**, 250-253.
- Huisman, A.E., Tier, B., Brown, D.J. (2006). On assessing contrasts between groups of animals. *Livestock Production Science* **104**, 254-267.
- Huisman, A.E. and Brown, D.J. (2007). What genetic group structure to fit? A Bayesian approach applied to yearling worm egg count data in merino sheep. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **17**, 399-402.
- James, J.W. and Cottle, D.J. (1997). Detection of genotype by environment interaction using merino sire evaluation data. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **12**, 461-465.
- Kennedy, B.W. and Moxley, J.E. (1975). Comparison of genetic group and relationship methods for mixed model sire evaluation. *Journal of Dairy Science* **58**, 1507-1514.
- Kennedy, B.W. (1981). Bias and mean square error from ignoring genetic groups in mixed model sire evaluation. *Journal of Dairy Science* **64**, 689-697.
- Kennedy, B.W. and Trus, D. (1993). Considerations on genetic connectedness between management units under an animal model. *Journal of Animal Science* **71**, 2341-2352.
- Kinghorn, B.P. and Shepherd, R.K. (1990). The impact of across-flock genetic evaluation on sheep breeding structures. *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production* **15**, 7-16.
- Kinghorn, B.P., Kennedy, B.W., Smith, C. (1993). A method of screening for genes of major effect. *Genetics* **134**, 351-360.

- Kuehn, L.A., Lewis, R.M., Nieuwhof, G.J. (1999). Evaluating connectedness over time in a group breeding scheme using a sheep paradigm. *Journal of Animal Science* **83** (Supplement 1), p.102. (Abstract).
- Kuehn, L.A. (2005). Implications of connectedness in the genetic evaluation of livestock. PhD Thesis, Virginia Polytechnic Institute and State University, Blacksburg, USA.
- Kuehn, L.A., Lewis, R.M., Notter, D.R. (2007). Managing the risk of comparing estimated breeding values across flocks or herds through connectedness: a review and application. *Genetics Selection Evolution* **39**, 225-247.
- Kuehn, L.A., Notter, D.R., Nieuwhof, G.J., Lewis, R.M. (2008). Changes in connectedness over time in alternative sheep sire referencing schemes. *Journal of Animal Science* **86**, 536-544.
- Laloe, D. (1993). Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution* **25**, 557-576.
- Laloe, D., Phocas, F., Menissier, F. (1996). Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics Selection Evolution* **28**, 359-378.
- Laloe, D. and Phocas, F. (2003). A proposal of criteria of robustness analysis in genetic evaluation. *Livestock Production Science* **80**, 241-256.
- Lewis, R.M., Crump, R.E., Kuehn, L.A., Simm, G., Thompson, R. (1999). Assessing connectedness in across-flock genetic evaluations. *Proceedings of British Society of Animal Science*, Scarborough, p.121. (Abstract).
- Lopes de Oliveira, C.A., Nunes Martins, E., Ribeiro de Freitas, A. (2001). Heterogeneity of variances in the genetic groups formed by the Canchim breed. *Revista Brasileira de Zootecnia* **30**, 1212-1219.

- Machado, M.B.B., Alencar, M.M., Pereira, A.P., Oliveira, H.N., Casas, E., Coutinho, L.L., Regitano, L.C.A. (2003). QTL affecting body weight in a candidate region of cattle chromosome 5. *Genetics and Molecular Biology* **26**, 259-265.
- Mackowski, M., Szymoniak, K., Szydlowski, M., Kamyczek, M., Eckert, R., Rozycki, M., Switonski, M. (2005). Missense mutations in exon 4 of the porcine LEPR gene encoding extracellular domain and their association with fatness traits. *Animal Genetics* **36**, 135-137.
- MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press **1**, 281-297.
- Mark, T., Fikse, W.F., Emanuelson, U., Philipsson, J. (2002). Effect of phantom parent grouping and properties of deregression for a low heritability trait. *Journal of Dairy Science* **85**, 2393-2395.
- Mathur, P.K., Sullivan, B., Chesnais, J.P. (1998). A new method for assessing connectedness between herds. *Record of Proceedings of National Swine Improvement Federation Conference Annual Meeting*, No. **23**.
- Mathur, P.K., Sullivan, B.P., Chesnais, J.P. (2002). Measuring connectedness: concept and application to a large industry breeding program. *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production*, Communication No. 20-13.
- McClintock, A.E. and Taylor, J.F. (1982). Developments in the use of BLUP for estimating genetic merit. In: *Future Developments in the Genetic Improvement of Animals* (Eds. Barker, J.S.F., McClintock, A.E.), Academic Press, Australia, pp. 157-178.
- Miglior, F., Sullivan, P.G., Van Doormaal, B.J. (2003). Analysis of genetic groups in MACE evaluations with different time-edit on data. *Interbull Bulletin* **31**, 180-184.

- Mortimer, S.I. and Atkins, K.D. (1989). Genetic evaluation of production traits between and within flocks of merino sheep. I Hogget fleece weights, body weight and wool quality. *Australian Journal of Agricultural Research* **40**, 433-443.
- Mortimer, S.I. and Atkins, K.D. (2001). Genetic differences among merino bloodlines in fleece rot resistance from wether comparisons. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **14**, 183-186.
- Mortimer, S.I., Hopkins, D.L., Stanley, D.F., McMillan, D.C., Anderson, S.L. (2006). Strain differences in merinos for carcase and meat quality. *Proceedings of the 26th Biennial Conference of Australian Society of Animal Production*. Short Communication No. 2.
- Pereira, A.P., Alencar, M.M., Oliveira, H.N., Regitano, L.C.A. (2005). Association of GH and IGF-1 polymorphisms with growth traits in a synthetic beef cattle breed. *Genetics and Molecular Biology* **28**, 230-236.
- Perez-Enciso, M. and Fernando, R.L. (1992). Genetic evaluation with uncertain parentage: a comparison of methods. *Theoretical and Applied Genetics* **84**, 173-179.
- Peskovicova, D., Groeneveld, E., Wolf, J. (2004). Effect of genetic groups on the efficiency of selection in pigs. *Livestock Production Science* **88**, 213-222.
- Phocas, F. and Laloe, D. (2004). Should genetic groups be fitted in BLUP evaluation? Practical answer for the French AI beef sire evaluation. *Genetics Selection Evolution* **36**, 325-345.
- Pieramati, C. and Van Vleck, L.D. (1993). Effect of genetic groups on estimates of additive genetic variance. *Journal of Animal Science* **71**, 66-70.
- Pietsch, S.G., Mortimer, S.I., Crook, B.J. (1997). Bloodline differences and changes in hogget merino wool quality traits during bloodline substitution. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **12**, 167-171.

- Pollak, E.J., Ufford, G.R., Gross, S.J. (1977). Comparison of alternative models for within-herd genetic evaluation of beef cattle. *Journal of Animal Science* **45**, 1010-1014.
- Pollak, E.J. and Quaas, R.L. (1983). Definition of group effects in sire evaluation models. *Journal of Dairy Science* **66**, 1503-1509.
- Quaas, R.L. and Pollak, E.J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science* **51**, 1277-1287.
- Quaas, R.L. and Pollak, E.J. (1981). Modified equations for sire models with groups. *Journal of Dairy Science* **64**, 1868-1872.
- Quaas, R.L. (1988). Additive genetic model with groups and relationships. *Journal of Dairy Science* **71**, 1338-1345.
- Quaas, R.L., Pollak, E.J., Cunningham, B.E., Klei, L. (1998). Bayesian procedure with an autoregressive prior for smoothing estimated time trends. *Journal of Animal Science* **76 (Supplement 1)**, 57 (Abstract).
- Redman, K.J., Golden, B.L., Bourdon, R.M. (1993). Multi-breed animal model BLUP with additive direct and maternal breed groups and regressions on outcrossing. *Journal of Animal Science* **71 (Supplement 1)**, 103 (Abstract).
- Richards, J.S., Atkins, K.D., Mortimer, M., Semple, S.J. (2006). Pedigree assignment by electronic matching of lambs and dams. *Proceedings of the 26th Biennial Conference of Australian Society of Animal Production*. Short Communication No. 31.
- Robinson, G.K. (1986). Group effects and computing strategies for models for estimating breeding values. *Journal of Dairy Science* **69**, 3106-3111.
- Roso, V.M., Schenkel, F.S., Miller, S.P. (2002). Estimation of the degree of connectedness between test groups of station-tested beef bulls. *Proceedings of the*

- 7th World Congress on Genetics Applied to Livestock Production*, Communication No. 02-28.
- Roso, V.M., Schenkel, F.S., Miller, S.P. (2004). Degree of connectedness among groups of centrally tested beef bulls. *Canadian Journal of Animal Science* **84**, 37-47.
- Rozzi, P., Schaeffer, L.R., Burnside, E.B., Schlote, W. (1990). International evaluation of Holstein-Friesian dairy sires from three countries. *Livestock Production Science* **24**, 15-28.
- Russell, B.C. and Cottle, D.J. (1992). The effect of genetic parameter estimates on breeding value estimation in merino sire evaluation programs. *Proceedings of Australian Association of Animal Breeding and Genetics* **10**, 460-463.
- SAS (1990). SAS language: Reference, Version 6, First Edition. SAS Institute Inc., Cary, NC.
- Schaeffer, L.R. (1975). Disconnectedness and variance component estimation. *Biometrics* **31**, 969-977.
- Schaeffer, L.R. (1985). Model for international evaluation of dairy sires. *Livestock Production Science* **12**, 105-115.
- Schaeffer, L.R. (1994). Multiple-country comparison of dairy sires. *Journal of Dairy Science* **77**, 2671-2678.
- Simm, G., Lewis, R.M., Collins, J.E., Nieuwhof, G.J. (2001). Use of sire referencing schemes to select for improved carcass composition in sheep. *Journal of Animal Science* **79** (E. Supplement), E255-E259.
- Smith, C. and Banos, G. (1991). Selection within and across populations in livestock improvement. *Journal of Animal Science* **69**, 2387-2394.
- Sorensen, D.A. and Kennedy, B.W. (1984). Estimation of genetic variances from unselected and selected populations. *Journal of Animal Science* **59**, 1213-1223.

- Sullivan, P.G. and Schaeffer, L.R. (1994). Fixed versus random genetic groups. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production* **18**, 483-486.
- Sullivan, P.G. (1995). Alternatives for genetic evaluation with uncertain parentage. *Canadian Journal of Animal Science* **75**, 31-36.
- Sullivan, P.G., Wilton, J.W., Miller, S.P., Banks, L.R. (1999). Genetic trends and breed overlap derived from multiple-breed genetic evaluations of beef cattle for growth traits. *Journal of Animal Science* **77**, 2019-2027.
- Theron, H.E., Kanfer, F.H.J., Rautenbach, L. (2002). The effect of phantom parent groups on genetic trend estimation. *South African Journal of Animal Science* **32**, 130-135.
- Thompson, R. (1979). Sire evaluation. *Biometrics* **35**, 339-353.
- Togashi, K., Lin, C.Y., Yokouchi, K. (2004). Overview of genetic evaluation in dairy cattle. *Animal Science Journal* **75**, 275-284.
- Tong, A.K.W., Kennedy, B.W., Moxley, J.E. (1980). Potential errors in sire evaluation from regional genetic differences. *Journal of Dairy Science* **63**, 627-633.
- Tosh, J.J. and Wilton, J.W. (1990). Degree of connectedness in mixed models. *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production* **13**, 480-483.
- Tosh, J.J. and Wilton, J.W. (1994). Effects of data structure on variance of prediction error and accuracy of genetic evaluation. *Journal of Animal Science* **72**, 2568-2577.
- Ugarte, E., Alenda, R., Carabano, M.J. (1992). Fixed or random contemporary groups in genetic evaluations. *Journal of Dairy Science* **75**, 269-278.
- Van der Werf, J.H.J. and de Boer, I.J.M. (1990). Estimation of additive genetic variance when base populations are selected. *Journal of Animal Science* **68**, 3124-3132.

- Van der Werf, J.H.J. (1990). Models to estimate genetic parameters in crossbred dairy cattle populations under selection. PhD Thesis, Wageningen Agricultural University, Wageningen, The Netherlands.
- Van Vleck, L.D. (1970). Misidentification in estimating the paternal sib correlation. *Journal of Dairy Science* **53**, 1469-1474.
- Van Vleck, L.D. (1990a). Breeding value prediction with maternal genetic groups. *Journal of Animal Science* **68**, 3998-4013.
- Van Vleck, L.D. (1990b). Absorption of equations for non-parents for an animal model with maternal effects and genetic groups. *Journal of Animal Science* **68**, 4014-4025.
- Van Vleck, L.D., Hanford, K.J., Macneil, M. (2003). Effects of genetic groups to account for selection on estimates of genetic parameters for a line of Hereford cattle. *Journal of Animal Science* **81 (Supplement 1)**, 86 (Abstract).
- Vasconcelos, J., Santos, F., Bagnato, A., Carvalheira, J. (2008). Effects of clustering herds with small-sized contemporary groups in dairy cattle genetic evaluations. *Journal of Dairy Science* **91**, 377-384.
- Weigel, K.A. and Rekaya, R. (2000). A multiple-trait herd cluster model for international dairy sire evaluation. *Journal of Dairy Science* **83**, 815-821.
- Westell, R.A. (1984). Simultaneous genetic evaluation of sires and cows for a large population of dairy cattle. PhD Thesis, Cornell University, New York, USA.
- Westell, R.A. and Van Vleck, L.D. (1987). Simultaneous genetic evaluation of sires and cows for a large population of dairy cattle. *Journal of Dairy Science* **70**, 1006-1017.
- Westell, R.A., Quaas, R.L., Van Vleck, L.D. (1988). Genetic groups in an animal model. *Journal of Dairy Science* **71**, 1310-1318.

- Wiggans, G.R., Misztal, I., Van Vleck, L.D. (1988). Animal model evaluation of Ayrshire milk yield with all lactations, herd-sire interaction, and groups based on unknown parents. *Journal of Dairy Science* **71**, 1319-1329.
- Wilmink, J.B.M., Meijering, A., Engel, B. (1986). Conversion of breeding values for milk from foreign populations. *Livestock Production Science* **14**, 223-229.
- Wolf, J., Peskovicova, D., Wolfova, M., Groeneveld, E. (2002). Impact of genetic groups and crossbred information on the prediction of breeding values in pig sire breeds. *Czech Journal of Animal Science* **47**, 219-229.
- Woolaston, R.R. (1997). Detecting genetic differences between groups of sheep during parasite infection. *International Journal of Parasitology* **27**, 839-841.
- www.en.wikipedia.org/wiki/Data_cluster. The free online encyclopedia.
- www.home.dei.polimi.it/matteucc/Clustering. Clustering online tutorial.
- www.sheepgenetics.org.au. Official website of Sheep Genetics Database.
- www.ucl.ac.uk/oncology/MicroCore. Hierarchical Clustering.
- Zwald, N.R., Weigel, K.A., Fikse, W.F., Rekaya, R. (2003). Application of a multiple-trait herd cluster model for genetic evaluation of dairy sires from seventeen countries. *Journal of Dairy Science* **86**, 376-382.

Appendix. Standardization of estimated genetic group effects

The across-flock correlation between EBVs derived using the two strategies of genetic grouping was 0.96 and 0.83 for BWT and FD, respectively. This difference in correlation (higher for BWT) is due to the fact that when flocks are clustered for the two traits together using conventional strategy, the underlying clustering mechanism is dominated by the trait with a larger variance (BWT) when compared to the second trait (FD) (Brown, Personal communication). The estimated genetic group effects along with their standard errors (SE) for the two traits are presented in Table 1. There is a significant difference in the variance of the two genetic group effects (4.75 Kg^2 and 0.43 micron^2 for BWT and FD, respectively). This difference in the variance has led to a high correlation (0.96) for BWT. As expected, the estimated genetic group effects for the 16th group were zeros for both traits (Table 1) due to missing data.

Since there was a large difference in the estimated genetic group variance of BWT and FD, the estimated genetic group effects of the two traits were standardized prior to clustering by subtracting the mean from each estimate and then dividing by its standard deviation so that the two estimated genetic group effects had similar variance. Additionally, there was a large difference in the across-contemporary group means of the two traits (48.28 Kg for BWT and 18.74 micron for FD). Therefore, to eliminate the effect of difference in the means, records on each trait were divided by their respective contemporary group means and then multiplied with the overall mean (standardized and scaled data) and genetic group effects were re-estimated. The adjusting of data in this manner prior to analysis is a routine practice in Sheep Genetics (Brown *et al.*, 2006) evaluations.

Table 1. Estimated genetic group (GG) effects ($\pm$ SE) for BWT and FD using raw and standardized data (complete dataset)

GG	<u>BWT (Kg)</u>		<u>FD (micron)</u>	
	Raw	Standardized	Raw	Standardized
1	-1.38 ± 1.25	-1.78 ± 1.30	0.23 ± 0.12	0.23 ± 0.12
2	-0.92 ± 0.62	-0.88 ± 0.64	0.83 ± 0.10	0.83 ± 0.10
3	-2.69 ± 1.53	-2.90 ± 1.58	-0.23 ± 0.22	-0.21 ± 0.22
4	-0.07 ± 1.12	0.08 ± 1.14	0.84 ± 0.22	0.84 ± 0.22
5	-2.05 ± 1.03	-1.94 ± 1.07	-0.04 ± 0.21	-0.05 ± 0.20
6	-1.33 ± 1.00	-1.32 ± 1.02	0.18 ± 0.08	0.18 ± 0.08
7	2.74 ± 0.88	2.52 ± 0.91	1.68 ± 0.19	1.66 ± 0.19
8	-0.95 ± 1.41	-0.88 ± 1.42	1.51 ± 0.29	1.49 ± 0.28
9	0.36 ± 0.93	0.41 ± 0.95	1.04 ± 0.13	1.03 ± 0.13
10	-0.16 ± 1.58	-0.20 ± 1.63	0.29 ± 0.18	0.28 ± 0.18
11	-2.28 ± 1.03	-2.10 ± 1.07	-0.21 ± 0.26	-0.20 ± 0.25
12	-5.39 ± 1.22	-5.40 ± 1.26	-0.42 ± 0.46	-0.41 ± 0.46
13	-1.26 ± 1.36	-1.26 ± 1.39	-0.38 ± 0.21	-0.37 ± 0.21
14	2.29 ± 1.90	2.24 ± 1.96	-0.29 ± 0.37	-0.28 ± 0.36
15	-4.85 ± 1.60	-4.31 ± 1.64	0.17 ± 0.41	0.18 ± 0.41
16	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00