2009

The effect of connectedness on the bias and accuracy for prediction of breeding value in swine herds

Norikazu Soga

Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Animal Sciences Commons

Recommended Citation

Soga, Norikazu, "The effect of connectedness on the bias and accuracy for prediction of breeding value in swine herds" (2009).

Graduate Theses and Dissertations. 11116.

http://lib.dr.iastate.edu/etd/11116

This Thesis is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
The effect of connectedness
on the bias and accuracy for prediction of breeding value
in swine herds

by

Norikazu Soga

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Animal Science

Program of Study Committee:

Thomas J. Baas, Major Professor
P.J. Berger
Theodore Bailey

Iowa State University
Ames, Iowa
2009
# TABLE OF CONTENTS

## CHAPTER 1. GENERAL INTRODUCTION

- Introduction 1
- Thesis Organization 3

## CHAPTER 2. LITERATURE REVIEW

- Concepts of Connectedness 4
- Relationship between Connectedness and Bias 6
- Risk of Bias from Low Connectedness 9
- Connectedness Statistics 14
- The Effect of Bias on Genetic Response to Selection 25

## CHAPTER 3. RELATIONSHIP BETWEEN CONNECTEDNESS AND BIAS OF THE U.S. PUREBRED DUROC POPULATION

- Introduction 27
- Connectedness Statistics 28
- Materials and Methods 31
- Results and Discussion 36
- Literature Cited 40

## CHAPTER 4. INVESTIGATION OF CURRENT LEVELS OF CONNECTEDNESS IN THE U.S. PUREBRED DUROC POPULATION

- Introduction 50
- Materials and Methods 51
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Pork is one of the most consumed meats all over the world, representing around 40% of all meat consumption (Ray, 2002). Pork production has been increasing due to increasing world population, which is predicted to increase by 2.9 billion people in the next fifty years. Over the past quarter century, the pork industry has increased lean percentage through genetic selection. This trend toward leaner pork has caused problems such as meat quality issues, susceptibility for specific diseases, and sensitivity for stress. In addition, consumer demands have been expanded to include food safety and animal welfare, as well as meat quality. As a result, meat quality traits and health traits have become more important and have been emphasized in selection programs. Recently, producers have needed to effectively improve these traits, as well as conventional production traits, due to higher livestock feed price and other production costs compared with similar prices a few years ago (Hill, 2009).

However, heritability of most meat quality traits is low to moderate (0.15 to 0.30) and they often have unfavorable correlations with conventional production traits. For example, the genetic correlation between loin muscle area (LMA) and consumer acceptability is negative (Sellier, 1998). Similarly, disease resistance traits also have low heritability (0.08 for serum concentration of immunoglobulin) and have unfavorable correlations with conventional production traits (Edfors-Lilja, 1994).

In breeding programs, precise estimation of breeding value (BV) for each trait is a critical issue. Best Linear Unbiased Prediction (BLUP) models have become the industry standard to estimate BV. Through BLUP, an animal’s phenotypic record can be divided into its genetic and environmental components (Henderson, 1973). Development of accurate
equipment and techniques for measurement of traits decreases environmental variance. Using correlated traits or genetic marker information for indirect selection also helps to increase accuracy of selection. However, even though we attempt to estimate BV for each pig accurately, there is a risk that comparison of individual BV from different herds can be negatively biased (Kuehn et al., 2008). The cause of this bias mainly comes from the assumption that genetic means (means of BVs) of each herd are the same. In many cases this assumption is not valid, which may negatively bias the accuracy of BV comparisons across herds. This bias may also have an unfavorable effect on accuracy of estimation of individual BV from data sets which contain animals from several herds which have different genetic means, as well as on accuracy of estimation of differences in BVs (comparison) (Smith, 1991; Kennedy, 1993).

To indicate the degree of bias, many researchers have studied “connectedness”. Connectedness is defined as genetic similarity across herds. In a population, the more common parents of animals, the greater connectedness they have. If a pair of herds has low connectedness, the genetic mean difference of those herds may tend to be larger, which may make the comparison of animal BVs from the respective herds less accurate (Kuehn et al., 2008). With low connectedness, we cannot precisely estimate the ratio of the genetic component to the environment component for performance differences between different herds (Bunter, 1997). It has been reported that exchanging common sires across herds which would be included in the analysis for estimating BVs is a good method to establish connectedness (Hanocq et al., 1996; Roden, 1996; Lewis and Simm, 2000).

Although many quantitative measures of connectedness have been proposed, each of them has advantages and disadvantages in reduction of bias, computational ease, and ease of
interpretation. Few comparisons of methods are available in the literature. Most researchers have used simulated data, and indications of the current level of connectedness in the swine industry are not available. Moreover, few studies have evaluated the effect of exchanging common sires to reduce the unfavorable effect of bias.

There were three objectives for this study. The first objective was to define several methods of estimating connectedness and to compare advantages and disadvantages of each. The second objective was to evaluate levels of connectedness and prediction error of the difference of BV between animals (PEVD), using purebred swine industry data from 21 herds over 19 years. The third objective was to investigate the relationship between connectedness and bias, and between different methods of estimating connectedness.

These results will help breeders understand the value of establishing connectedness schemes by exchanging common reference sires, which should help to make selection schemes more efficient.

**Thesis Organization**

This thesis is presented as a general introduction, a literature review, 2 individual papers, and a general summary. References cited in the general introduction and literature review follow the general summary section. All reference citations are in compliance with the CBE Style Manual used by the Journal of Animal Science. Each individual paper consists of an abstract and 3 sections: introduction, materials and methods, and results and discussion. Literature cited within the papers is listed after the results and discussion section of each individual paper.
CHAPTER 2. LITERATURE REVIEW

Concepts of Connectedness

In an animal model, if breeding value (BV) differences between pairs of individuals are all estimable, they can be considered to be connected statistically. If not, they can be considered to be disconnected (Searle, 1971). Without some degree of genetic connection, the estimation of differences of BV across herds can be less accurate than with good genetic connection (Kuehn, 2008). It may be possible that this bias in accuracy of comparisons of BV across herds also has an unfavorable effect on accuracy of estimation for each BV itself from data sets which contain animals from several herds which have different genetic means as well as on accuracy of estimation of difference of BVs (comparison) (Smith, 1991; Kennedy, 1993). Connectedness is a quantitative measure of genetic connections across herds. This section will discuss concepts of connectedness to consider the risk for negative bias on accuracy of BV comparison across herds.

The bias in comparison of individual BV from different herds comes mainly from a difference of the genetic mean (average BV) between herds. If connectedness is not considered, and this is usual in conventional schemes for estimation of BV, it is assumed that the average genetic mean of all herds are equal, which is perfect connection. Tosh (1990) defined this assumption in different terms: “that BVs are randomly and perfectly normally distributed across the complete population of base animals.” Of course, this is not always true, rather, not true in most cases (Kennedy, 1993). That is imperfect connection, which may be more common in large populations containing many herds where there is no sharing of breeding animals across herds. If the genetic mean difference across herds differs largely (low connectedness), accuracy for comparison of BV across herds can be negatively biased.
This genetic mean difference across herds is caused by genetic drift or selection after
differentiation into subpopulations and before pedigree and performance data are recorded
for analysis. In this case, in the absence of connectedness, genetic differences are not
considered in the estimation of BV by using best linear unbiased prediction (BLUP)
(Falconer, 1996). Absence of connectedness causes bias in accuracy of comparison for BVs
across herds and accuracy of estimation of BV itself (Kennedy, 1993). It was reported that
the herds of animals can be clustered by the degree of connectedness and those clusters may
be formed not only by geographical discrepancy, but also by different breeding objectives.
They may be desired biological types of those clusters in the sheep industry. It was also
reported that connectedness between animals tends to be lower between those clusters than
within the clusters (Notter, 1998).

Many researchers have discussed the risk of ignoring the bias from low
connectedness. Tong (1980) reported that in order to estimate genetic differences across
herds accurately, connectedness should be established between those herds. He
recommended increasing connectedness by using artificial insemination (AI) for semi-
isolated herds and concluded that establishing connectedness increases the accuracy of
comparison of BV across herds. Kuehn (2008) showed that if a pair of herds has lower
connectedness, the genetic mean difference of that pair might tend to be larger. It might make
the comparison of animal BVs from the respective herds less accurate. He said that
increasing connectedness makes separation of genetic differences from environmental
differences between herds possible. And he proclaimed that the ideal statistic for
connectedness should show a level of risk for selection and also should be able to be used for
breeding programs across herds in order to increase accuracy of estimation of comparison for
BVs across herds. It increases genetic gain response from genetic selection. Lewis (1999) reported that comparisons of BV across herds can be biased when base animals within different herds have different genetic means and when connectedness among those herds is not high enough. Bunter (1997) reported that with low connectedness, we cannot estimate the ratio of the genetic to environment component for performance differences between herds. He suggested that more research about the relationship between connectedness and the degree of bias can be beneficial to develop connectedness as an indication of the risk of comparing BV across herds.

**Relationship between connectedness and bias**

Connectedness and bias in accuracy of comparison for BVs across herds are closely related. In this section, I will discuss the relationship between connectedness and bias and how to decrease that bias in order to increase accuracy comparing BVs across herds.

In this context, the squared bias of prediction was calculated by subtracting unbiased prediction error variances of the animal effect (PEV\textsubscript{unbiased}) from averaged mean square error for the animal effect (MSE) (Kennedy, 1981):

\[
\text{MSE} = \text{PEV}_{\text{unbiased}} + \text{bias}^2.
\]

If a pair of herds is unbiased (connectedness is high or comparison is done within herd), MSE and PEV\textsubscript{unbiased} are equal (bias\textsuperscript{2} is zero) (Kennedy, 1981):

\[
\text{MSE}_{\text{within}} = \text{PEV}_{\text{unbiased}}.
\]

And,

\[
\text{MSE}_{\text{between}} = \text{MSE}_{\text{within}} + \text{bias}^2,
\]

where MSE\textsubscript{between} is MSE of BV comparison between herds and MSE\textsubscript{within} is MSE of BV comparison within herds, respectively. When an animal model does not contain uncorrelated
random effects (usually typical animal models contain the animal residual as a random
effect), unbiased $\text{MSE}_{\text{within}}$ can be approximated with heritability ($h^2$) times the variance of
environmental effects ($V_e$) (Kennedy, 1981):

$$\text{MSE}_{\text{within}} = h^2 \cdot V_e \text{ (no other random effect, unbiased)}.$$  

$\text{MSE}_{\text{between}}$ was defined as variance of contrasts $(L_{ij})$ (Kuehn, 2008):

$$\text{MSE}_{\text{between}} = \text{average}(\text{Var}(\sum L_{ij}))$$

$$= \text{average}(\text{Var}(\sum((u_i - u_j) - \sum(u_i - u_j))))$$

$$= \text{average}(\text{PEV}(\sum(u_i - u_j)))$$

$$= \text{average}(\text{PEV}(\sum(u_i)) + \text{PEV}(\sum(u_j)) - 2*\text{PEC}(\sum(u_i, u_j)))$$

$$= (\text{PEV}(\sum(u_i)) + \text{PEV}(\sum(u_j)) - 2*\text{PEC}(\sum(u_i, u_j)))/2,$$

where $u_i$ and $u_j$ are BV of animals $i$ and $j$, from different herds, respectively, and $u_i$ and $u_j$
are the BLUP estimated BV (EBV) of animals $i$ and $j$, from different herds, respectively.

$\text{PEV}(\sum(u_i - u_j))$ represents the sum of prediction error variance of the difference of
individual BVs from different herds. $\text{PEV}(\sum(u_i))$ is the sum of prediction error variance of
BV from one herd; $\text{PEV}(\sum(u_j))$ is the same but from another herd. $\text{PEC}(\sum(u_i, u_j))$ is the
sum of prediction error covariance of BVs from different herds. These can be biased.

Kuehn (2008) considered analyses of the U.S. sheep industry data for 15 years of
selection with exchanging common sires across herds each year for reducing bias and defined
the average percentage of bias remaining as:

$$\%\text{Bias} = (\text{MSE}_{\text{between}(k)} - \text{MSE}_{\text{within}})/ (\text{MSE}_{\text{between}(0)} - \text{MSE}_{\text{within}}) \times 100,$$

$$\text{MSE}_{\text{between}(k)} = \text{average}(\text{Var}(\sum L_{ij}) \text{ of selection year } k),$$
where $\text{MSE}_{\text{between}(0)}$ is the mean squared error across-flock, $L_{ij}$ in selection year 0 of the current simulation replicate. He illustrated the relationship between this %Bias and flock connectedness correlation ($r_{ij}$) and concluded that the more connectedness, the less bias.

Kennedy (1993) and Kuehn (2008) defined this $\text{MSE}_{\text{between}}$ in different terms such as the “mean prediction error variance of differences” in EBV between animals in different herds (PEVD):

$$\text{PEVD} = \text{MSE}_{\text{between}}.$$  

The change of the bias in the PEVD mainly comes from the change of prediction error covariance. That covariance between animals is zero when no connection exists (zero connectedness). Increasing connections through a sire referencing scheme makes the prediction error covariance between individuals from different herds larger and this positive covariance between different herd animals reflects a correlation of BV comparison across herds, which put errors in the same direction. With this, causes of bias across herds should decrease as prediction error covariance increases (more connected) (Kuehn, 2008).

Kennedy (1993) concluded that genetic connections across herds decrease PEVD to the level of $\text{MSE}_{\text{within}}$ because PEVD for related animals are less than PEVD for unrelated animals and because genetic connections decrease variances of estimates of difference across genetic group effects. He illustrated that establishing connectedness across herds not only increases a positive covariance between herds which reduces the bias of PEVD but also reduces the bias in prediction error of variance of the animal effect because it decreases the sampling error of genetic group effects. So, establishing connectedness increases both accuracy of estimation and comparison BVs with the data sets which contain animals from different herds.
Kuehn (2009) proclaimed that if connectedness is low, MSE_{between} can be twice as large as MSE_{within}. Almost the same ratio can be seen in another paper (Kennedy, 1981). Cundiff (1975) suggested the genetic variance between herds may be as high as that within herds. Spike (1978) also reported the same conclusion. It means ignoring connectedness can have significant risk which can increase prediction error of comparison for BVs across herds two-fold.

**Risk of bias from low connectedness**

To reduce the risk of low connectedness bias, three major reactions can be considered. One way is to compare BV of performance data from central testing directly, which has been historically used to compare BV across herds. However, this way does not establish genetic improvement effectively in the industry because of limited number of comparisons, the risk of pre-test influence on comparisons, and less than constant correlations between central-test and on-herd performance data (Bunter, 1997). Also, the number of test stations has been decreasing and each test station capacity has become smaller and smaller to the extent that it is impossible to compare all possible subunits (Bichard, 1987). In addition to that, cost is another big disadvantage of this method.

The next method to manage the bias in comparing animals across herds is to fit the source of the bias into the genetic evaluation models. Bias of genetic evaluation caused by genetic mean difference between herds (low connectedness) can be eliminated by including genetic groups into the prediction model. By considering genetic groups, genetic mean differences can be absorbed into the remaining effects to be estimated. This can reduce the risk of bias from low connectedness data (Quaas, 1988). This model can be written as:

\[ y = Xb + ZQg + Zu + e, \]
where y is a vector of phenotypes, b is a vector of fixed effects, g is a vector of fixed base animal genetic group effects, u is a vector of random genetic effects expressed as a deviation from the expectation of each animal’s genetic group, and e is a vector of residuals. Incidence matrixes X and Z represent phenotypes to specific combinations of fixed and random genetic effects, respectively, and Q represents the expected proportion of genes in each animal coming from the various genetic groups (Pollak, 1983).

The assumed distribution of random effects in that model is:

\[
\begin{bmatrix}
y \\
u \\
e
\end{bmatrix} \sim N\left(\begin{bmatrix}Zb + ZQg \\
0 \\
0
\end{bmatrix}, \begin{bmatrix}ZA'Va + I*Ve \\
A'Va \\
I*Ve
\end{bmatrix}\right),
\]

where A is the numerator relationship matrix and Va is additive genetic variance.

Estimates of g and predictions of u are obtained as solutions to the resulting

\[
Q'Z'MQ \begin{bmatrix}Q'Z'MZ \\
Z'MQ \\
Z'MQ + \lambda A^{-1}
\end{bmatrix} \begin{bmatrix}g \\
u
\end{bmatrix} = \begin{bmatrix}Q'Z'My \\
Z'My
\end{bmatrix},
\]

with:

\[
M = I - X(X'X)^{-1}X'
\]

\[
\lambda = Ve/Va.
\]

The generalized inverse of the partitioned coefficient matrix in the above equation is:

\[
C^{-1} = \begin{bmatrix}C_{11} & C_{12} \\
C_{12}' & C_{22}
\end{bmatrix}.
\]

The prediction error variance and covariance matrix for fixed effects and random effects can be derived from:

\[
C^{-1}Ve.
\]

From Henderson (1975), the PEV (biased) of estimation BV (MSE) in this model with genetic grouping is:
PEV_{biased} = \text{var} (\hat{a} - a) \\
= (QC_{11}Q' + C'_{12}Q' + QC_{12} + C_{22}) * Ve \geq \text{PEV}_{unbiased}

On the other hand, the PEV of estimation BV in the model without genetic grouping is:

PEV_{biased} = \text{var} (\hat{a} - a) \\
= C_{22} * Ve \geq \text{PEV}_{unbiased}.

If minimization of PEV_{biased} is the goal, compared with PEV_{biased} from the no grouping model, PEV_{biased} from the grouping model is decreased by considering genetic group (decrease $C_{22} * Ve$), whereas it is increased by the error variance of the fixed genetic group effects (increase $(QC_{11}Q' + C'_{12}Q' + QC_{12}) * Ve$) (Tong, 1980). From this, the answer to the question whether genetic group effect should be included into a prediction model or not depends on which effect (decreasing $C_{22} * Ve$ or increasing $(QC_{11}Q' + C'_{12}Q' + QC_{12}) * Ve$) is larger (Kuehn, 2008). So this way to manage the bias from low connectedness is valid only when decreasing the error of random group effects is larger than increasing the error of fixed group effects.

Complicated grouping strategy might cause unexpected confounding with other fixed effects (Quaas, 1998). For example, if data contains only two herds and if the true genetic mean difference is less than the standard error of estimate of estimation for BV between herds, ignoring genetic groups leads to smaller MSE of estimation for BV between those herds (more accurate) (Kennedy, 1981). This study used simulation data so true genetic mean differences can be computable. Thus, which model is better depends on the degree of difference of their genetic differences. That is, the bias elimination ability of each model cannot be relied on without considering the degree of genetic connection, because low connectedness may increase $C_{22}$ (Foulley, 1990). Regardless of which model is used for
estimation of BV, without establishing good connectedness between herds, these systems are still biased and have risk of lower accuracy of estimation (Kennedy, 1981). Thus, connectedness between herds is still required to be established in order to estimate difference of BVs across herds accurately.

Thus, the third way to manage the risk of bias from low connectedness is to reduce bias itself by establishing connectedness. For establishing connectedness between herds, exchanging common reference sires has been considered as the best way. By using common reference sires, producers can decrease genetic mean differences between herds and then increase the effective size of their herd data (Kuehn, 2008), which improves accuracy of comparison of BVs and reduces inbreeding across herds (Lewis, 2000). From this research, programs that promote an exchange of common reference sires among herds effectively have been required to eliminate bias which comes from confounding between genetic and environmental means.

To achieve elimination of bias, it is important to know which of schemes are effective. Hudson (1980) suggested that making each unit (herd) larger can be beneficial, because each reference sire can have more progeny for more accurate comparison of BVs across herds. Schaeffer (1980) suggested that fitting the common reference sire effect as a fixed effect rather than a random effect in the prediction model may not be effective for reducing prediction error variance of those young sires. Tong (1980) reported that more semen exchange and reciprocal exchange rather than less semen exchange and one-directional semen exchange had a larger impact on decreasing bias. Kuehn (2008) compared exchanging schemes for fifteen year simulation among AI sires exchanging schemes (five year, ten year, fifteen year continuous, fifteen year alternating) and natural service (NS)
schemes in terms of connectedness. From this study, they concluded that AI schemes were more efficient than natural service (NS) schemes and that connectedness increased and PEVD decreased by AI. With every year exchanging schemes, PEVD no longer decreased after sire referencing had been going on for five years whereas connectedness continued to increase. This may be due to PEVD levels reaching an asymptotic level near $\text{MSE}_{\text{within}}$ (no bias) already with established connectedness, which meant that the bias and connectedness had a negative correlation but if connectedness of the pair of herds became higher than a specific level, bias can be almost eliminated from the system so PEVD became relatively constant. Although less intense schemes can achieve no-bias-level, as well as more intense schemes can, it may take longer (15 years). It was also concluded that using known recorded sires as reference sires was more effective than using non-recorded sires as reference sires in order to establish connectedness. In terms of effectiveness of connectedness, he reported that it seemed to be difficult to remove bias completely without considering the type of genetic link across herds in only one generation if the herd sizes are small (less than one hundred progeny per year) (Kuehn, 2008). In that study, once bias was removed from the comparisons, it did not again come back, even though the herds stop the program to connect with other herds actively. This stability of established connectedness may be because even after the herd leaves a scheme for establishing connectedness, subsequent genetic changes from selection or random drift can be accounted for by the relationship matrix of the prediction model as long as pedigree and performance data recording continues (Sorensen, 1983).

Smith (1991) reported that establishing connection between herds with sire referencing schemes will reduce PEVD and individual lamb prediction error variance
(PEV_{biased}), especially in small flocks with limited information (low number of records and low heritability) within a single year, due to more accurate estimation of genetic means for base founder animals (no-record). As mentioned in the previous section, the change mainly comes from the change of prediction error covariance or sampling error of genetic group effects.

Thus, to manage bias from low connection, only using central test results or including genetic groups into the prediction model is not enough. Using the breeding program with exchanging common reference sires is the best way for accurate (less or non-bias) estimation and comparisons of BVs across herds.

**Connectedness Statistics**

Until the 1980s, the method of calculating connectedness was relatively simple. Determination of connectedness for fixed effects involved only ascertaining if linear functions of fixed effects were estimable in N-way cross classifications (Peterson, 1978). If the model used is a 2-factor model, it can be displayed by tracing a perpendicular path between non-zero cells in a table which counts the number of data points for each classification (Fernando, 1983). However, these simple ways cannot give a continuous and comparable measure of connectedness (Foulley, 1990). Definitions and features of several connectedness statistics and their advantages and disadvantages are discussed.

PEVD. Because the purpose of connectedness is to indicate the accuracy of comparisons between BVs across herds, a logical statistical measurement of connectedness across herds would be the average prediction error variance (PEV) of all pair-wise estimated BV (EBV) differences between the animals in those herds (PEVD):

\[
\text{PEVD} = \frac{(\text{Var}(\sum((u_i - u_j) - \sum(u_i - u_j))) / 2.}
\]
Prediction error variance of differences (PEVD) is derived from the mean squared error (MSE) of prediction differences between candidates for selection. It can be calculated from the prediction models, both with and without genetic groups. It was reported that if PEVD was used as an indicator of connectedness, a large number of reference sire progeny (20% to 45% of the total number of progeny produced in that herd) were required to accurately compare BVs across herds (Foulley, 1983), and genetic groups should be fitted into the prediction model to manage potential bias from genetic mean difference across herds (Miraei, 1991). Although PEVD is a direct indicator of prediction accuracy of BV difference and the most precise way to show connectedness, computation of PEVD is difficult or impossible for large data sets in most situations and these values are difficult to interpret because no range is decided. Approximations that were used do not usually allow for computation of off-diagonal components of the prediction error variance-covariance matrix, which should affect PEVD, and as a result, alternative measures of genetic connectedness are required (Kennedy, 1993).

Kennedy (1993) proposed Gene Flow (GF), Genetic Drift Variance (GDV), and Variance of Estimates of Management-Unit Effects (VEM) as three alternatives for computing PEVD. Gene Flow is calculated by:

\[ \text{Gene Flow} = X'ZTQ, \]

where X and Z are incidence matrices for management units and animals, Q represents foundation animals with respect to their management unit of origin, and T is a lower triangular matrix (Henderson, 1976; Thompson, 1979). A simple quantitative measure of connectedness between 2 herds from GF is the absolute difference between the sum of diagonal components of the matrix and the sum of the off-diagonal components of the
A small value means high connectedness. The advantages of this connectedness statistic are ease of computation and that it can be calculated from prediction models both with and without genetic groups, similar to PEVD. Disadvantages are that it requires all foundation animals to be assigned to a management unit of origin and its correlation with PEVD is the lowest (0.671) of the 3 connectedness statistics (Kennedy, 1993). In addition, values are difficult to interpret because no range is decided and it cannot detect which pair of herds that are used for analysis has low genetic connection (only the degree of total genetic connection).

Genetic Drift Variance (GDV) is proposed as an alternative for PEVD (Kennedy, 1993). It can be calculated by:

\[ \text{Genetic Drift Variance} = X'ZAX'X. \]

The average relationship between and within genetic group (including the relationship of an animal to itself) is obtained by dividing diagonal components by the square of the number of records in the genetic group and off-diagonal components by the product of the number of records in each genetic group. The components of this matrix can be interpreted as the genetic components of drift variance and covariance between genetic groups (Sorensen, 1983). A simple quantitative measure of connectedness between 2 herds from GDV is the absolute difference between the sum of diagonal components of the matrix and the sum of the off-diagonal components of the matrix. A small value means high connectedness. The advantages of this connectedness statistic are higher correlation (0.924) with PEVD than GF, it does not require all foundation animals to be assigned to a management unit of origin, and it can be calculated from the prediction models both with and without genetic groups, similar to PEVD (Kennedy, 1993). Disadvantages of GDV are difficulty of computation, values are
difficult to interpret because no range is decided, and it cannot detect which pair of herds used for analysis has low genetic connection. It can detect only the mean of degree of genetic connection across all pairs. This connectedness statistics cannot indicate which comparisons from pairs may be biased.

Variance of Estimates of Management-Unit Effects (VEM) was also proposed as an alternative for PEVD by Kennedy (1993). Variance of estimates of differences of management unit effects can affect PEVD. It can be calculated by:

\[
VEM = [X'X - X'Z(Z'Z + \lambda A^{-1})^{-1}Z'X]^{-1} Ve.
\]

Where \( A^{-1} \) is the inverse of numerator relationship matrix, \( \lambda \) is the ratio of Ve to variance of additive genetic effect (Va). A simple quantitative measure of connectedness between 2 herds from GDV is the absolute difference between the sum of diagonal components of the matrix and the sum of the off-diagonal components of the matrix. In another way, this can be written as:

\[
\text{Var}(h1 - h2) = \text{var}(h1) + \text{var}(h2) - 2\text{cov}(h1, h2),
\]

where \( \text{Var}(h1 - h2) \) is the variance of difference of estimates of management units (\( h1 \) and \( h2 \)). \( \text{Var}(h1) \) and \( \text{Var}(h2) \) represent variances of estimates of management units h1 and h2, respectively. Solutions can be obtained iteratively using procedures suggested by Henderson (1974) and described by Harville (1979). A small value means high connectedness. The advantages of this connectedness statistic are ease of computation, the highest correlation (0.995) with PEVD of the 3 connectedness statistics (Kennedy, 1993), it can detect which pair of herds used for analysis has low genetic connection, and it can be calculated from prediction models both with and without genetic groups, like PEVD. The disadvantages of VEM are that VEM itself and the relationship between VEM and PEVD can be affected by
genetic group population size and data structure, and values are difficult to interpret because no range is decided. Bunter (1997) suggested that if we use VEM as an indicator of connectedness, the impact of data structure on the pattern and levels of connectedness across herds must be carefully considered.

To overcome disadvantages of VEM, Mathur (2002) suggested using Connectedness Rating (CR) to define connectedness, which has been used in the Canadian Swine Improvement Program. Connectedness rating was defined as the correlation between estimates of fixed genetic group effects. Connectedness rating can be calculated as:

\[ CR = \frac{\text{cov}(h_1, h_2)}{\sqrt{\text{var}(h_1) \cdot \text{var}(h_2)}} \]

In terms of variance of the difference of estimates of management units, connectedness rating can be written as:

\[ \text{Var}(h_1 - h_2) = \text{var}(h_1) + \text{var}(h_2) - 2CR \cdot \sqrt{\text{Var}(h_1) \cdot \text{Var}(h_2)}. \]

Connectedness rating values range from 0 to 1 and high values mean more genetic connection (1 is perfect connection). The advantages of CR are that it is less dependent on genetic group size and structure, it can detect which pair of herds used for analysis has low genetic connection, and its computation is relatively easy. In addition, it can be calculated from prediction models both with and without genetic groups like PEVD, it has high correlation with PEVD, and its values are easy to interpret because the range is between 0 and 1. The disadvantage of CR is that its relationship with bias depends on the linking strategy (using common reference sire) that is employed. These features, especially consistency against genetic group size and structure, mean that CR is good for checking the
change of the bias level of specific groups with real industry data (unbalanced data), but is not effective for the comparison of linking strategies. Mathur (2002) reported that CR is usually higher for highly heritable traits such as backfat and age, and lower for low heritability traits such as sow productivity traits. According to Mathur (2002), based on acceptable levels of bias, average CR levels for each pair of groups should be 3% or more for highly heritable traits and 1.5% or more for lowly heritable traits. If substantial importation of genetics into a herd is being considered, the minimum CR should be 5%. Above these thresholds, bias can be almost eliminated.

Another connectedness statistic derived from PEVD is connectedness correlation ($R_{ij}$), proposed by Lewis (1999; 2005). Although $R_{ij}$ is a correlation between the influences of genetic groups like CR, the primary difference between $R_{ij}$ and CR is that $R_{ij}$ uses the means of animal random effects in the genetic groups, instead of using genetic group fixed effects (CR). It can be computed by:

$$R_{ij} = \frac{\text{PEC}(u_i, u_j)}{\sqrt{\text{PEV}(u_i) \cdot \text{PEV}(u_j)}}$$

where $u_{i(j)}$ is the mean EBV of all animals in genetic group $i(j)$, $\text{PEC}(u_i, u_j)$ is the prediction error covariance between these means, and $\text{PEV}(u_i)$ and $\text{PEV}(u_j)$ are the prediction error variances of the mean EBV of genetic groups i and j, respectively. In terms of variance of difference of estimates of EBV means between genetic groups, $R_{ij}$ is:

$$\text{PEV}(u_i - u_j) = \text{PEV}(u_i) + \text{PEV}(u_j) - 2R_{ij} \sqrt{\text{PEV}(u_i) \cdot \text{PEV}(u_j)}.$$  

Values range from 0 to 1 and high values mean more genetic connection (1 is perfect connection). The advantages of $R_{ij}$ are that relatively fewer data are required to obtain
reliable values than for PEVD, CR, Connectedness Index (IC) (Foulley, 1990) and Coefficient of Determination (CD) (Laloë, 1993), and it can be calculated from prediction models both with and without genetic groups. In addition, it can detect which pair of herds used for analysis has low genetic connection, its relationship with bias is less dependent on linking strategy (using common reference sire), it has high correlation with PEVD, and values are easy to interpret. The disadvantages of \( R_{ij} \) are that its computation is more difficult than CR and it may be affected by genetic group size and structure. Based on these features, connectedness correlation \( (R_{ij}) \) works well for numerous situations but is especially suited for simulation data, when compared with CR. Kuehn (2008) suggested that a flock \( R_{ij} \) of 0.05 corresponded to approximately 80% reduction in bias compared with no connection, and a flock \( R_{ij} \) of 0.10 corresponded to approximately 90% reduction in bias compared with no connection in simulated sheep data. This relationship between \( R_{ij} \) and bias was relatively constant and independent of heritability (0.25 and 0.125). Therefore, he concluded that benchmarks of 0.05 for ‘good’ connectedness and 0.10 for ‘superior’ connectedness are warranted.

Foulley (1990) developed Connectedness Index (IC) as a continuous measure of connectedness. This was developed based on the hypothesis that more connected data sets were more nearly orthogonal (perfectly balanced). Connectedness Index (IC) is derived, not from PEVD, but from PEV_{biased} \( (= \text{var}(a^\hat{}-a)) \). Connectedness Index (IC) can be calculated by:

\[
\text{IC}(x) = x'C_R x/x'C_F x,
\]
where \( C_R \) is a portion of the inverse coefficient matrix \( (C_{22}) \) in the model without genetic group effects (reduced model) and \( C_F \) is a portion of the inverse coefficient matrix \( (C_{22}) \) in the model with genetic group effects (full model). High values indicate more genetic connection. If a random effect is perfectly orthogonal to a fixed effect removed in the reduced model, \( \text{PEV}_{\text{biased}} \) is the same for the random effects in both the full and the reduced model. Thus, at that time, IC becomes 1 (perfect connection). The advantage of IC is that values are easy to interpret because its range is between 0 and 1. The disadvantages of IC are that its computation is difficult, especially because it requires calculation of both models with and without genetic grouping, it cannot detect which pair of herds that are used for analysis has low genetic connection, and with only few progeny, IC may be close to 1, even though the accuracy of BV comparison is low because IC does not account for the amount of information (i.e., progeny records) in the analysis.

Foulley (1992) developed the ratio of determinants of inverse coefficient matrices (\( \gamma \)) based on IC to measure connectedness on the whole design. It is calculated by:

\[
\gamma = \frac{|C_R|^n}{|C_F|^n},
\]

where \( n \) is the column rank of the incidence matrix of the parameters to which the subunits \( C_R \) and \( C_F \) relate. This ratio of determinants of inverse coefficient matrices of reduced and full models was proposed using the Kullback-Leibler (Kullback, 1983) distance between the joint density of the maximum likelihood estimators of all parameters in the model with genetic grouping and the product of the marginal densities of the parameters removed from and remaining in the model without genetic grouping. High values indicate more genetic connection. The advantages of \( \gamma \) are that values are easy to interpret because its range is
between 0 and 1 and computation is relatively easier than IC. The one disadvantage of $\gamma$ is that it still requires calculation of both models with and without genetic grouping.

While orthogonality of the data is desirable (IC put emphasis on this), accuracy of estimation of BVs is normally evaluated by the coefficient of determination (CD), which is the squared correlation between the true and estimated BV (Laloë, 1993). This represents the degree of information that contributes to the prediction of breeding value, which had been used in selection indices, so it indicates whether animals could be compared across different environments or, in other words, reliability of estimation of BV. This connectedness statistic is derived not from PEVD but from PEV, unlike connectedness statistics like IC, but only for the model without genetic grouping. Coefficient of determination can be calculated by:

$$CD(x) = 1 - \frac{\lambda^*(x' C_{22} x)}{x' A x}$$

$$= \frac{x'(A - \lambda C_{22})x}{x' A x},$$

where $C_{22}$ is the random effects component of the inverse coefficient matrix for a model without genetic groups. Using the CD to investigate the accuracy of a contrast between animals or sets of animals in different herds would then provide a measure of connectedness. This CD becomes zero if the mean difference is not estimable (no connectedness). The advantages of CD are that its relationship with bias does not depend on the linking strategy (using common reference sire) employed (Kuehn, 2008), and values are easy to interpret because its range is between 0 and 1. The disadvantages of CD are that its computation is difficult, it cannot detect which pair of herds used for analysis has low genetic connection, and it can only be used for the prediction model without genetic grouping unlike IC, PEVD, and other connectedness statistics derived from PEVD (MSE\text{between}). In addition, even if it
becomes large, the proportional amounts of the genetic group mean differences still remain
due to bias in the difference of BVs, which is not considered in the CD calculation (PEV_{biased}
origin), and it can be calculated directly only from simulated data. For calculating
unbalanced industry data, it is necessary to estimate CD using approximations such as
criterion of admission to the group of connected herds (CaCO), which is highly correlated
with CD (0.94) and its computation is easier (Fouilloux, 2008). Another disadvantage of CD
is that designing programs to increase connectedness by increasing CD may risk decreasing
genetic progress from selection due to lower selection intensity because this connectedness
statistic depends not only on the amount of information in the data but also on its structure
(Kuehn, 2008). To decide if a specific system is well connected or not, Laloë (2003)
developed criteria for required CD:

$$\text{required CD} = \frac{n\eta}{n\eta + \lambda},$$

where \( n \) is the number of progeny per sire and \( \eta \) is the proportion of progeny from common
sires. If the CD is larger than this value, it indicates the herds are well connected with each
other. These features mean that CD can be a good indicator only for the comparison of
linking strategy and for simulation data, but is not effective for checking the change of the
bias level of specific genetic groups and for unbalanced industry data. Mainly because of the
difficulty of computation, alternative statistics which are highly correlated with CD like
CaCO are required.

Laloë (1996) developed 2 overall measures of connectedness from the equation of
CD and related them to the eigenvalues (\( \mu_i \)) and eigenvectors (\( c_i \)) resulting from the
solutions:

$$[(A - \lambda C_{22}) - \mu_i A] c_i = 0.$$
The number of eigenvalues is equal to the number of estimated BV. The smallest eigenvalue is always zero and the other eigenvalues are related with all possible independent contrasts. The two connectedness statistics Laloë (1996) developed from CD are functions of these non-zero eigenvalues. They can be calculated by:

\[ \rho_1 = \sum_{i=1}^{n} \frac{\mu_i}{n-1} \]

and

\[ \rho_2 = \prod_{i=2}^{n} \mu_i \].

These statistics range from 0 to 1, where low values indicate low accuracy in comparison of animal BVs across fixed-effect classes. Both of them provide a sense of the average level of accuracy across all unique contrasts. If more than one eigenvalue is zero, it indicates that at least one contrast is uninformative. In that case, \( \rho_2 \) becomes zero. The advantages of these functions of eigenvalues (\( \rho_1, \rho_2 \)) are that they can handle the information of genetic mean differences unlike CD, values are generalized (0 to 1) and their correlations with CD are high (Hofer, 1994). One disadvantage is that they require extensive computing time for large data sets because their calculations use all elements of the inverse coefficient matrices. Another disadvantage is that designing programs to increase connectedness by increasing CD may decrease genetic progress from selection due to lower selection intensity because these connectedness statistics depend not only on the amount of information in the data, but also on its structure (Kuehn, 2008). So, these overall connectedness measures may be useful for group leaders or scientists overseeing genetic evaluation programs, but not for individual producers who are trying to increase connectedness with other units in the system.

These connectedness statistics are summarized in Table 1.
The effects of bias on genetic response to selection

Without considering connectedness, producers cannot accurately know whether their animals’ performance is better than others because those comparisons of BV across herds can be biased. Several simulation studies have been done to manage this risk. Establishing connectedness could improve genetic gain response by 30 to 35% compared with a selection scheme ignoring connectedness with increasing accuracy of comparison across herds and with lower inbreeding rates (Roden, 1996; Hanocq, 1996; Kuehn, 2008).

Selection intensity can be significantly increased by selecting animals across all members of the scheme (Lewis, 2000), which increases genetic gain because genetic gain response is the product of selection intensity, $h^2$ and $V_p$. However, as mentioned in the previous section, with some connectedness statistics, which are derived from PEVbiased such as CD, $\rho_1$, $\rho_2$, IC and $\gamma$, improving them has the risk of lower selection intensity because they do not consider bias in the difference of BVs.

If the size of herd was relatively small (less than 100), the improvement of genetic gain might be clearer because the likelihood of producing extreme individuals was small in those cases (Wood, 1991). Maraei (1993) showed that the larger the genetic mean difference between herds, the more the genetic gains increased at the first time of their connection. This may be due to lower genetic mean herds could utilize the resources of higher genetic mean herds and then the average of all herd’s performance increased by homogenizing. He also concluded that if herd size was larger than 300, the rate of gain was relatively constant by establishing connectedness. Thus, establishing connectedness across herds should be beneficial for livestock producers in terms of their genetic selection response.
Table 1. Comparison of connectedness statistics.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Computation</th>
<th>Requirement of Genetic Grouping</th>
<th>Correlation with Bias</th>
<th>Value Range</th>
<th>Influence of Size</th>
<th>Influence of Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEVD</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>Unlimited</td>
<td>Influenced</td>
<td>——</td>
</tr>
<tr>
<td>GF</td>
<td>Very easy</td>
<td>Not-required</td>
<td>Low</td>
<td>Unlimited</td>
<td>Influenced</td>
<td>——</td>
</tr>
<tr>
<td>GDV</td>
<td>Hard</td>
<td>Not-required</td>
<td>Normal</td>
<td>Unlimited</td>
<td>Normal</td>
<td>——</td>
</tr>
<tr>
<td>VEM</td>
<td>Easy</td>
<td>Not-required</td>
<td>High</td>
<td>Unlimited</td>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>CR</td>
<td>Easy</td>
<td>Not-required</td>
<td>High</td>
<td>0 to 1</td>
<td>No influence</td>
<td>No influence</td>
</tr>
<tr>
<td>R</td>
<td>Hard</td>
<td>Not-required</td>
<td>High</td>
<td>0 to 1</td>
<td>No influence</td>
<td>Normal</td>
</tr>
<tr>
<td>IC</td>
<td>Easy</td>
<td>Required</td>
<td>Normal</td>
<td>0 to 1</td>
<td>Influenced</td>
<td>——</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Very easy</td>
<td>Required</td>
<td>Normal</td>
<td>0 to 1</td>
<td>Influenced</td>
<td>——</td>
</tr>
<tr>
<td>CD</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>0 to 1</td>
<td>Improved</td>
<td>No influence</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>0 to 1</td>
<td>Improved</td>
<td>No influence</td>
</tr>
</tbody>
</table>

Statistic Detect Weak Pair Effect of Reference Scheme Eliminate Bias Genetic Response
---
PEVD Pair and mean ——— Eliminate all Improved
GF Only mean ——— Eliminate all Improved
GDV Only mean ——— Eliminate all Improved
VEM Pair and mean ——— Eliminate all Improved
CR Pair and mean Varied Eliminate all Improved
R Pair and mean Normal Eliminate all Improved
IC Only mean ——— May remain part of bias May decrease selection intensity
$\gamma$ Only mean ——— May remain part of bias May decrease selection intensity
CD Only mean Constant May remain part of bias May decrease selection intensity
$\rho$ Only mean ——— May remain part of bias May decrease selection intensity

*PEVD: prediction error variance of difference between animals; GF: genetic flow; GDV: genetic drift variance; VEM: variance of estimates of management unit effects; CR: connectedness rating; R: connectedness correlation; IC: connectedness index; $\gamma$: the ratio of determinants of inverse coefficient matrix; CD: coefficient of determination; $\rho$: the functions of eigen value*
CHAPTER 3. RELATIONSHIP BETWEEN CONNECTEDNESS AND BIAS OF THE U.S. PUREBRED DUROC POPULATION

Introduction

Accuracy of estimation of breeding value (BV) and accuracy of comparison of BVs is critical in the livestock industry. When BVs are compared across herds, it is assumed that there is no genetic mean (year-mean BV for that herd) difference between herds. Due to different breeding goals, selection practices, and resulting genetic progress, this assumption is not valid in most cases, which results in bias in prediction error variance and covariance, which then decreases the accuracy of comparison of BVs across herds (Smith, 1991; Kennedy, 1993).

In order to estimate this bias, researchers have studied “connectedness”, which refers to statistics of genetic similarity between groups. Establishing high levels of connectedness between herds can decrease the occurrence of bias because more genetic similarity can decrease genetic mean differences between the herds (Kuehn et al., 2008). It has been reported that exchanging common sires across herds which would be included in the analysis for estimating BVs is a good method to establish connectedness (Hanocq et al., 1996; Roden, 1996; Lewis and Simm, 2000). Participation by those breeders where comparisons between genetic values are made is required for establishing connectedness efficiently through sire exchange programs. To establish the benefits of common sire programs in achieving higher levels of connectedness, more science-based information on connectedness is required.

Although many statistics of connectedness have been proposed and each of them has advantages and disadvantages in reduction of bias, computational ease, and ease of
interpretation, few comparisons of methods are available in the literature. Most researchers have used simulated data, and indications of the current level of connectedness in the swine industry are not available. Useful or accurate connectedness information is valuable to the U.S. purebred swine industry where an across-herd evaluation is conducted and used to facilitate genetic progress at the breed level. Moreover, few studies have evaluated the effect of exchanging common sires to reduce the unfavorable effect of bias.

The main objective of this study was to compare connectedness statistics published in the literature and determine which measure is most practical or suitable for estimating the risk associated with EBV comparisons in an across-herd analysis. Data were obtained from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) on Duroc litters born from 1990 to 2008 (NSR, 2009). Connectedness statistics for 4 common methods (Prediction error variance of all pair-wise estimated BV differences between the animals on those herds (PEVD) (Kennedy, 1993), Connectedness correlation (CR) (Mathur, 2002), Connectedness rating (Lewis, 1999), and Common sire % (CS%)) were estimated. The correlation among connectedness statistics and genetic mean differences was investigated and their effect on bias was evaluated. The effect of bias on selection accuracy and genetic response to selection was also evaluated.

**Connectedness Statistics**

From previous studies, several connectedness statistics are summarized in Table 1. Prediction error variance (PEV) of all pair-wise estimated BV (EBV) differences between the animals on those herds (PEVD) indicates the accuracy of comparisons of BV across herds directly (Kennedy, 1993). Gene Flow (GF), Genetic Drift Variance (GDV), and Variance of
Estimates of Management-Unit Effects (VEM) were proposed as 3 alternatives for computing PEVD (Kennedy, 1993).

To overcome disadvantages of VEM, Mathur (2002) suggested using Connectedness Rating (CR) to define connectedness. Connectedness rating was defined as the correlation between estimates of fixed genetic group effects.

Another connectedness statistic derived from PEVD is connectedness correlation ($R_{ij}$), proposed by Lewis (1999; 2005). The $R_{ij}$ is a correlation between prediction error variance of mean EBV of all animals in genetic group i(j).

Foulley (1990) developed Connectedness Index (IC) as a continuous measure of connectedness. This was developed based on the hypothesis that more connected data sets were more nearly orthogonal (perfectly balanced). Foulley (1992) developed the ratio of determinants of inverse coefficient matrices ($\gamma$) based on IC to measure connectedness on the whole design.

Accuracy of estimation of BVs is normally evaluated by the coefficient of determination (CD), which is the squared correlation between the true and estimated BV (Laloë, 1993). The 2 connectedness statistics Laloë (1996) developed from CD are functions of these non-zero eigenvalues ($\rho_1, \rho_2$).

Common sire % (CS%) between herd $i$ and herd $j$ is the ratio of sum of the number of the progeny of a herd $i$ and herd $j$ which has common sire with herd $j$ and herd $i$ respectively to the sum of the number of total progeny of a herd $i$ and herd $j$.

$$CS\% = \frac{PC_{(i)} + PC_{(j)}}{P_i + P_j} = \frac{k \cdot (S_{c(i)} + S_{c(j)})}{P_i + P_j},$$
where $P_{c(i)}$ and $P_{c(j)}$ are the number of the progeny in herd $i$ and herd $j$ which has common sires with herd $j$ and herd $i$ respectively, $P_i$ and $P_j$ refer to the number of total progeny in herd $i$ and herd $j$ respectively, $S_{c(i)}$ and $S_{c(j)}$ are the number of common sires in the herd $i$ and herd $j$ with herd $j$ and herd $i$ respectively, and $k$ is the average number of progeny/sire. The significant advantage of this statistic is ease of computation when compared to other connectedness statistics. Other advantages are that it can be calculated regardless of the model used, values range from 0 to 1, and it can detect low levels of connectedness. Disadvantages are that the correlation with bias may be less than other connectedness statistics and it may be affected by herd size, which means that common sire % may be less accurate when herd size is small. Numerous researchers have proposed that exchanging common reference sires across herd is an effective method to establish connectedness across herds (Kennedy, 1993; Mathur, 2002; Kuehn, 2008).

From these connectedness statistics PEVD, CR, $R_{ij}$ and CS% were calculated using field data for the Duroc breed from National Swine Registry (NSR, 2009). Even though its computation is difficult, connectedness PEVD was chosen because it is a direct indicator of bias. Of the 5 connectedness statistics derived from PEVD (GF, GDV, VEM, CR and R), CR and R were chosen because they are suited for continuous measurement of connectedness for complicated data sets like swine field data. The computation of CR is relatively easy and it is not affected by data structure. The computation of R is not as easy as CR, but it is not affected by data size. Because they cannot detect which pair of herds has low connectedness, which is one of the objectives of this study, IC and $\gamma$ were not chosen. CD and $\rho$ were not chosen because their computation is difficult and they are more suitable for simulation data, rather than field data. Common sire % (CS%) was chosen because its computation is very
easy. In addition, genetic mean difference (GMD) across herds was calculated because it is an indication of bias and its computation is easy.

**Materials and Methods**

**Data source**

Data were obtained from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) on Duroc litters born from 1990 to 2008. Details of data collection can be found in STAGES (NSR, 2009). Data included pedigree information for each animal, contemporary group, sex of the pig and litter identification, birth date and date weighed, and measurement for weight (WT), backfat (BF), loin eye area (LEA), marbling score, pH, loin color (MINOLTA), and days to 250 lb (Days250).

Numbers of records, animals, litters, and contemporary groups are presented in Table 2. These data were from 21 herds from 11 firms (Table 2). Firms represent breeders with several herds at different locations. Of the firms, 6 firms have a single herd, and 5 firms have multiple herds. Herds represent animals at a single location. Number of herds and firms which have records at each year is in Table 3. Only data from current breeders who employ an active genetic improvement program and follow ‘total herd reporting’ techniques were included. Table 3 indicates that only two of the breeders that currently use STAGES were actively reporting data in 1990. Means for off test weight (WT), backfat (BF), loin eye area (LEA), marbling score, pH, loin color (L*), and days to 250 lb (Days250) are presented in Table 4.

Contemporary groups (Table 2) were defined by individual breeders as a group of pigs that were raised in a common herd under similar environmental conditions. Data on boars, gilts, and barrows were included in the data set.
**Statistical analysis**

Data were analyzed according to the following single-trait model:

\[ Y_{ijkmn} = u + cg_i + sex_j + litter_k + a_{ijkmn} + wt_n + wt_n^2 + e_{ijkmn} \]

where \( Y_{ijkm} \) = the trait measured on pig \( m \) in contemporary group \( i \) of sex \( j \) in litter \( k \); \( u \) = mean; \( cg_i \) = fixed effect of contemporary group \( i \); \( sex_j \) = fixed effect of sex \( j \); \( litter_k \) = effect of litter \( k \), assumed random with \( Var = I\sigma_k^2 \) where \( A^{-1} \) is the inverse of the additive genetic relationship matrix; \( a_{ijklm} \) = effect of animal \( m \) assumed random with \( Var(a_{ijklm}) = A^{-1}\sigma_a^2 \); \( wt_m \) and \( wt_m^2 \) = linear and quadratic effects of the offtest weight of pig \( m \); \( e_{ijklm} \) = random residual error with \( Var(e_{ijklm}) = I\sigma_e^2 \). Variance and covariance components were estimated by the multiple-derivative-free restricted maximum likelihood program (MTDFREML) developed by Boldman (1995).

**Calculation of each connectedness measure**

Progeny are grouped by birth year and herd to define genetic groups. Prediction error variance of differences between the animals (PEVD). PEVD was calculated as:

\[ PEVD = PEV(u_i - u_j) \]

\[ = \frac{\sum PEV(u_i)}{P_i} + \frac{\sum PEV(u_j)}{P_j} - 2 \frac{PEC(\sum u_i, \sum u_j)}{(P_i + P_j)}. \]

Individual EBV and their prediction error variances are calculated and averaged by genetic group:

\[ \frac{\sum PEV(u_i)}{P_i}, \]

where \( P_i \) is the number of total progeny in genetic group \( i \).
Prediction error covariance for genetic group $i$ and $j$ was calculated by calculating contrasts of the sum of the individual BVs for genetic group $i$ and $j$ and averaged by the number of progeny:

$$\text{PEC}(\sum u_i, \sum u_j)/(P_i + P_j),$$

where $P_i$ and $P_j$ are the number of total progeny in herd $i$ and herd $j$, respectively.

Kennedy (1993) stated that as connectedness increases, PEVD approaches a specific level, and the unbiased level of PEVD should be equal to heritability ($h^2$) multiplied by the variance of environment ($Ve$) if there was no bias and the model contained only one random effect. This unbiased level of PEVD is defined as Base Line (BL$_{PEVD}$) where

$$\text{BL}_{PEVD} = h^2 \times Ve \text{ (only one random effect)}.$$  

The model used in this study contains a litter effect as a second random effect, so a modification of this equation is required. The ratio of the variance of the litter effect to the variance of the phenotypic effect is $L^2$. The animal effect is nested in the litter effect, so the mean squared error of estimation of the animal effect can be written as $(h^2-L^2) \times Ve$. Thus, the modified equation is:

$$\text{BL}_{PEVD} = (h^2-L^2) \times Ve.$$

The difference between PEVD and this modified BL$_{PEVD}$ can be considered as bias:

$$\text{PEVD} = \text{BL}_{PEVD} + \text{bias}.$$

Connectedness correlation ($R$). Prediction error variance (PEV($\sum u_i$)) and covariance (PEC($\sum u_i, \sum u_j$)) for genetic group $i$ and $j$ were estimated by calculating contrasts of the sum of individual BVs for each genetic group. Then,

$$R = \frac{\text{PEC}(\sum u_i, \sum u_j)}{\sqrt{\text{PEV}(\sum u_i) \times \text{PEV}(\sum u_j)}}.$$
Connectedness rating (CR). Variance of estimation of each genetic group effect \( \text{Var}(\sum g_i) \) and covariance \( \text{Cov}(\sum g_i, \sum g_j) \) for genetic group \( i \) and \( j \) were estimated by calculating contrasts of the sum of GCG effects in each genetic group. Then,

\[
CR = \frac{\text{Cov}(\sum g_i, \sum g_j)}{\sqrt{\text{Var}(\sum g_i) \ast \text{Var}(\sum g_j)}}.
\]

Common sire % (CS%). Progeny were grouped by birth year and their herd of origin to define genetic groups. Then,

\[
CS\% = \frac{(P_{c(i)} \ast + P_{c(j)})}{(P_i + P_j)}
\]

for each pair of genetic groups.

**Effect of Bias on Accuracy and Genetic Gain**

The MSE of estimation of BV can be written as the fraction of additive genetic variance \( \text{Va} \) not considered by prediction (Mrode, 2005):

\[
\text{MSE} = (1-r^2) \ast \text{Va} \leq \text{BL}_{\text{PEVD}},
\]

where \( r^2 \) is reliability of estimation of BV. For comparison of BV,

\[
\text{MSE} = \text{PEVD}
\]

and MSE becomes equal to \( \text{BL}_{\text{PEVD}} \) for both cases when the system is unbiased (perfect connection):

\[
\text{PEVD} = \text{BL}_{\text{pevd}} + \text{bias}.
\]

Thus, accuracy is:

\[
r = \sqrt{1 - \text{PEVD} / \text{Va}} \leq \sqrt{1 - \text{BL}_{\text{pevd}} / \text{Va}}.
\]

From this equation, it can be assumed that if low connectedness increases PEVD due to bias, then bias decreases accuracy for comparison of BVs. Here,

\[
\text{Va} = h^2 \ast \text{Vp},
\]
where \( V_p \) is phenotypic variance. From this, we can also assume that when \( h^2 \) becomes lower, the effect of bias on accuracy may become larger.

Here, PEVD can be also written as:

\[
PEVD = BL_{pevd} \times (1 + \text{bias} \%) \\
= (h^2 - L^2) \times Ve \times (1 + \text{bias} \%),
\]

where bias\% refers to the ratio of the difference between PEVD and BL\( _{PEVD} \) (bias) to BL\( _{PEVD} \). From the above equation, \( Ve' \) is identified as biased \( Ve \), where

\[
Ve' = Ve \times (1 + \text{bias} \%).
\]

So,

\[
PEVD = (h^2 - L^2) \times Ve'.
\]

With this equation and the definition of heritability \( (h^2) \), biased heritability \( (h^2') \) is:

\[
h^2' = \frac{Va}{Va + Vl + Ve'}
\]

\[
= \frac{Va}{Va + Vl + Ve \times (1 + \text{bias} \%)},
\]

where \( Va \) is variance of animal effect, and \( Vl \) is variance of litter effect. Here, the equation of genetic gain of selection is:

\[
\Delta G = i \times h^2 \times Vp,
\]

where \( \Delta G \) is genetic gain and \( i \) is selection intensity. With biased heritability, this can be written as:

\[
\Delta G = i \times h^2' \times Vp
\]

\[
= i \times \frac{Va}{Va + Vl + Ve \times (1 + \text{bias} \%)} \times Vp
\]
\[ i = \frac{h^2}{h^2 + L^2 + \text{PEVD}/(h^2 - L^2)} \times \text{Vp}. \]

From this, we can assume that bias may decrease genetic gain of selection.

**Results and Discussion**

**Correlation between connectedness and bias**

Correlations among 3 connectedness statistics (CR, R, CS%), PEVD, and the absolute value of genetic mean difference between herds (GMD) are presented in Table 5. High correlation estimates were obtained among the different measures of connectedness in the current data. The relationship between PEVD (\(= \text{bias} + BL_{\text{pevd}}\)) and the 3 connectedness statistics were moderately negative (-0.34 to -0.24). R had the highest correlation with PEVD whereas CS% had the lowest correlation with PEVD. In terms of computational requirement, CS% is significantly less demanding when compared to R, whereas the computation difficulty of R is similar to that of PEVD, and CR is intermediate to R and CS%. These correlations were lower than previous studies (Kuehn, 2009), likely because of the unbalanced nature of the dataset used in this study. In the current study, each genetic group has largely different progeny numbers, whereas previous studies used simulation data with equal progeny numbers in each genetic group and their genetic standard deviations across genetic groups were equal.

The relationship between GMD and the 3 connectedness statistics also had moderate negative correlations with each other. The correlations between GMD and CR, R, and CS% were relatively similar (-0.30 to -0.29) (Table 5). However, correlations between PEVD and the 3 connectedness measures involved more variation in estimates.
(-0.34 to -0.24) (Table 5). Finally, the relationship between GMD and PEVD was 0.11 (Table 5). This value was also much smaller than in a previous study (Kennedy, 1993). The reason for this difference may be that the previous research compared only 2 groups, so decreasing bias directly associated with decreasing GMD, in comparison to 21 herds in the current study.

The relationship between the 3 connectedness statistics and PEVD are presented in Figures 1, 2 and 3. Each plot represents the pair of genetic groups (herd and birth year). These graphs indicate that establishing connectedness (CR, R, and CS%) decreases PEVD level until it reaches the BL\textsubscript{pevd} level (derived by \(h^2 L^2 \ast Ve\)). Red points are pairs of herds from same firms. Black points are pairs of herds from different firms. Based on these, pairs of herds from same firms tend to have higher connectedness and lower PEVD than pairs of herds from different firms. It may be because of two reasons: a) sires are more commonly used across herds within a firm when compared to across firms, and b) herds within a firm generally follow the same leadership, and thus selection schemes are similar—which should decrease the GMD. Of the 3 connectedness statistics, CS% are more scattered than other 2 connectedness statistics (Fig. 3). Most of those outliers were pairs of genetic groups which contained less than 100 pigs, so if average herd size is more than 100, CS% can be used as an accurate indicator of connectedness.

For all the 3 connectedness statistics (CR, R, and CS %), when levels of connectedness become higher than approximately 0.1, PEVD seems to be suppressed to the level of BL\textsubscript{pevd} (unbiased). Therefore, 0.1 can be considered as a criterion for non-biased levels of connectedness. These criteria are approximately equal to those reported in previous studies (Kuehn, 2008; Mathur, 2002).
Effects of establishing connectedness

Variance components of these data with a single trait model for backfat are shown in Table 6. When the data are unbiased (PEVD = BL_{pevd}), the accuracy associated with comparison of BV is higher than when the data are biased (high PEVD; low connectedness; Table 7). Within the current study, accuracy decreases from 77.6% (unbiased) to 54.6% (biased) (Table 7).

If each value is substituted into the genetic response equation, the equation becomes:

$$\Delta G = \frac{i \times 0.43 + 0.0133}{0.4358 + \text{PEVD}/0.4242}$$

$$= \frac{i \times 0.00243}{0.1849 + \text{PEVD}}.$$

From this, when the pair of genetic groups are highly connected (no bias; PEVD = BL_{pevd}), $\Delta G$ is higher than when PEVD is large (biased) (Table 7). Within the current study, the genetic response decreases from 0.013*i (unbiased) to 0.012*i by bias (Table 7). Therefore, established connectedness helps to increase genetic gain across herds. This example includes variance components associated with BF ($h^2 = 0.43$), thus, the level of decrease in genetic gain due to low levels of connectedness is not as substantial as potential situations with characters associated with lower heritability. From the equation, if lower heritability traits such as maternal traits are used, this effect becomes larger, which means the risk of bias becomes higher compared with high heritability traits.

Thus, the relationship between PEVD ($=$ bias + BL_{pevd}) and the 3 connectedness statistics were moderately negative (-0.34 to -0.24). R had the highest correlation with PEVD, whereas CS% had the lowest correlation with PEVD. Establishing connectedness
(CR, R, and CS%) decreases PEVD level until it reaches the BL<sub>pevd</sub> level (derived by \( (h^2 - L^2) \ast Ve \)). Pairs of herds from same firms tend to have higher connectedness and lower PEVD than pairs of herds from different firms. If average herd size is more than 100, CS% can be used as easy indicator of connectedness. For CR, R, and CS %, when levels of connectedness become higher than approximately 0.1, PEVD seems to be suppressed to the level of BL<sub>pevd</sub> (unbiased). Therefore, 0.1 can be considered as a criterion for non-biased levels of connectedness.

For further study, in order to investigate effects of heritability and record density in multiple trait models, analysis using more traits of interest such as maternal traits which have lower heritability and which are sparsely recorded will be required. Different heritability levels can be evaluated with the current dataset (i.e., pH vs. backfat). An attempt to investigate effects of progeny number with another data set may be also useful. Because this paper used a representative sample of herds for the Duroc breeds, other breed should be studied. Using these results, a more detailed simulation study of the effect of exchanging common sires is warranted. Further studies can revolve around the needed number of common sires for maternal traits where records are obtained only in one sex and later in life, as well as simulation studies that examine the effectiveness of sire sampling programs.
Literature Cited


Table 1. Comparison of connectedness statistics.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Computation</th>
<th>Requirement of Genetic Grouping</th>
<th>Correlation with Bias</th>
<th>Value Range</th>
<th>Influence of Size</th>
<th>Influence of Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEVD</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>Unlimited</td>
<td>Influenced</td>
<td>—----------</td>
</tr>
<tr>
<td>GF</td>
<td>Very easy</td>
<td>Not- required</td>
<td>Low</td>
<td>Unlimited</td>
<td>Influenced</td>
<td>—----------</td>
</tr>
<tr>
<td>GDV</td>
<td>Hard</td>
<td>Not- required</td>
<td>Normal</td>
<td>Unlimited</td>
<td>Normal</td>
<td>—----------</td>
</tr>
<tr>
<td>VEM</td>
<td>Easy</td>
<td>Not- required</td>
<td>High</td>
<td>Unlimited</td>
<td>—</td>
<td>Influenced</td>
</tr>
<tr>
<td>CR</td>
<td>Easy</td>
<td>Not- required</td>
<td>High</td>
<td>0 to 1</td>
<td>—</td>
<td>No influence</td>
</tr>
<tr>
<td>R</td>
<td>Hard</td>
<td>Not- required</td>
<td>High</td>
<td>0 to 1</td>
<td>No influence</td>
<td>Normal</td>
</tr>
<tr>
<td>IC</td>
<td>Easy</td>
<td>Required</td>
<td>Normal</td>
<td>0 to 1</td>
<td>Influenced</td>
<td>—----------</td>
</tr>
<tr>
<td>γ</td>
<td>Very easy</td>
<td>Required</td>
<td>Normal</td>
<td>0 to 1</td>
<td>Influenced</td>
<td>—----------</td>
</tr>
<tr>
<td>CD</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>0 to 1</td>
<td>—</td>
<td>—----------</td>
</tr>
<tr>
<td>ρ</td>
<td>Very hard</td>
<td>Required</td>
<td>High</td>
<td>0 to 1</td>
<td>—</td>
<td>—----------</td>
</tr>
<tr>
<td>CS%</td>
<td>Very easy</td>
<td>Not- required</td>
<td>Normal</td>
<td>0 to 1</td>
<td>Influenced</td>
<td>—----------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Detect Weak Pair</th>
<th>Effect of Reference Scheme</th>
<th>Eliminate Bias</th>
<th>Genetic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEVD</td>
<td>Pair and mean</td>
<td>————</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>GF</td>
<td>Only mean</td>
<td>————</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>GDV</td>
<td>Only mean</td>
<td>————</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>VEM</td>
<td>Pair and mean</td>
<td>————</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>CR</td>
<td>Pair and mean</td>
<td>Varied</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>R</td>
<td>Pair and mean</td>
<td>Normal</td>
<td>Eliminate all</td>
<td>Improved</td>
</tr>
<tr>
<td>IC</td>
<td>Only mean</td>
<td>————</td>
<td>May remain part of bias</td>
<td>May decrease selection intensity</td>
</tr>
<tr>
<td>γ</td>
<td>Only mean</td>
<td>————</td>
<td>May remain part of bias</td>
<td>May decrease selection intensity</td>
</tr>
<tr>
<td>CD</td>
<td>Only mean</td>
<td>Constant</td>
<td>May remain part of bias</td>
<td>May decrease selection intensity</td>
</tr>
<tr>
<td>ρ</td>
<td>Only mean</td>
<td>————</td>
<td>May remain part of bias</td>
<td>May decrease selection intensity</td>
</tr>
<tr>
<td>CS%</td>
<td>Pair and mean</td>
<td>————</td>
<td>Eliminate all</td>
<td>————</td>
</tr>
</tbody>
</table>

PEVD: prediction error variance of difference between animals; GF: genetic flow; GDV: genetic drift variance; VEM: variance of estimates of management unit effects; CR: connectedness rating; R: connectedness correlation; IC: connectedness index; γ: the ratio of determinants of inverse coefficient matrix; CD: coefficient of determination; ρ: the functions of eigen value; CS%: common sire
Table 2. Structure of data set of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>156,101</td>
</tr>
<tr>
<td>Animals</td>
<td>160,325</td>
</tr>
<tr>
<td>Litters</td>
<td>33,210</td>
</tr>
<tr>
<td>Contemporary Groups</td>
<td>3,065</td>
</tr>
<tr>
<td>Herds</td>
<td>21</td>
</tr>
<tr>
<td>Firms</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3. Number of herds and firms of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008 in each year.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Herds</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Firms</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Herds</td>
<td>15</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td>16</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Firms</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 4. Number of observation, mean, and standard deviation for measured traits of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th>Traits</th>
<th>Number of observation</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight (kg)</td>
<td>156,101</td>
<td>108.86</td>
<td>12.68</td>
</tr>
<tr>
<td>Backfat (cm)</td>
<td>156,091</td>
<td>1.45</td>
<td>0.43</td>
</tr>
<tr>
<td>Loin eye area (cm2)</td>
<td>142,133</td>
<td>17.50</td>
<td>2.51</td>
</tr>
<tr>
<td>Marbling score</td>
<td>1,410</td>
<td>2.48</td>
<td>0.96</td>
</tr>
<tr>
<td>pH</td>
<td>1,372</td>
<td>5.76</td>
<td>0.22</td>
</tr>
<tr>
<td>Meat color (L*)</td>
<td>1,383</td>
<td>49.77</td>
<td>3.75</td>
</tr>
<tr>
<td>Days to 250 lb (Days250) (day)</td>
<td>156,095</td>
<td>161.9</td>
<td>12.8</td>
</tr>
</tbody>
</table>

Table 5. Correlations between genetic mean difference (GMD), prediction error variance prediction error variance of all pair-wise estimated BV differences between the animals on those herds (PEVD), common sire% (CS%), connectedness correlation (R), and connectedness rating (CR) of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th></th>
<th>GMD</th>
<th>PEVD</th>
<th>CS%</th>
<th>R</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMD</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>PEVD</td>
<td>0.115275</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>CS%</td>
<td>-0.30066</td>
<td>-0.24416</td>
<td>--------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>R</td>
<td>-0.29389</td>
<td>-0.34138</td>
<td>0.968131</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>CR</td>
<td>-0.29604</td>
<td>-0.31302</td>
<td>0.963768</td>
<td>0.989214</td>
<td>--------</td>
</tr>
</tbody>
</table>
Table 6. Parameters of data of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008 with buckfat single trait model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h^2$</td>
<td>0.43</td>
</tr>
<tr>
<td>$L^2$</td>
<td>0.063</td>
</tr>
<tr>
<td>Ve</td>
<td>0.00635</td>
</tr>
<tr>
<td>Va</td>
<td>0.0655</td>
</tr>
<tr>
<td>BL\textsubscript{pevd}</td>
<td>0.00233</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>PEVD</th>
<th>Accuracy of comparison BV</th>
<th>$\Delta G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0023 (=BL\textsubscript{pevd})</td>
<td>76.6%</td>
<td>$i*0.01300$</td>
</tr>
<tr>
<td>0.0040 (biased)</td>
<td>54.6%</td>
<td>$i*0.01286$</td>
</tr>
</tbody>
</table>
Figure 1. Relationship between connectedness rating (CR) and prediction error variance of all pair-wise estimated BV differences between the animals on those herds (PEVD) of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

\(^a\) Horizontal line is BL\(_{pevd}\).

\(^b\) Red and black points are pair of herds from same and different firms, respectively.

\(^c\) No-bias zone is the range in which CR is more than 0.1.
Figure 2. Relationship between connectedness correlation ($R$) and prediction error variance of all pair-wise estimated BV differences between the animals on those herds (PEVD) of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

a Horizontal line is $BL_{pevd}$.

b Red and black points are pair of herds from same and different firms, respectively.

c No-bias zone is the range in which CR is more than 0.1
Figure 3. Relationship between common sire% (CS%) and prediction error variance of all pair-wise estimated BV differences between the animals on those herds (PEVD) of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

a Horizontal line is BL_{pevd}.

b Red and black points are pair of herds from same and different firms, respectively.

c When CS% is more than 0.1, bias becomes small.
CHAPTER 4.

INVESTIGATION OF CURRENT LEVELS OF CONNECTEDNESS IN THE U.S. PUREBRED DUROC POPULATION

Introduction

Accuracy of estimation of breeding value (BV) and accuracy of comparison of BVs is critical to the U.S. purebred industry where an across-herd genetic evaluation provides the selection tools utilized by independent purebred breeders. When BVs are compared across herds, it is assumed that there is no genetic mean (average of BVs for that herd) difference between herds. This assumption is not valid in most cases, which results in bias in prediction error variance and covariance, which then decreases the accuracy of comparison of BVs across herds (Smith, 1991; Kennedy, 1993).

In reference to this bias, researchers have studied “connectedness”, which refers to statistics of genetic similarity between herds. Establishing high levels of connectedness between herds can decrease the risk of bias, as more genetic similarity can decrease genetic mean differences between individual herds (Kuehn et al., 2008). It has been reported that exchanging common sires across herds which would be included in the analysis for estimating BVs is a good method to establish connectedness (Hanocq et al., 1996; Roden, 1996; Lewis and Simm, 2000). Participation by a large number of breeders is required for establishing connectedness efficiently through sire exchange programs. To establish the benefits of common sire programs in achieving higher levels of connectedness, more science-based information on connectedness is required.
Although many researchers have studied connectedness, few have investigated how the level of connectedness in the swine industry has changed, what the current level of connectedness is, and if improvements in levels of connectedness are beneficial.

The main objective of this study was to investigate trends in connectedness levels over 19 years utilizing field data from the U.S. Duroc breed (NSR, 2009). A second objective was to investigate changes in bias levels and genetic mean differences (GMD) across herds within the data. The importance of the improving connectedness for the swine industry was evaluated, and desirable number of sires to be exchanged across herds to eliminate bias was investigated. Within the current study, three connectedness measures were evaluated: 1) Connectedness Rating (CR); 2) Connectedness Correlation (R); and 3) Common Sire % (CS%).

Materials and Methods

Data source

Data were obtained from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) on Duroc litters born from 1990 to 2008. Details of data collection can be found in STAGES (NSR, 2009). Data included pedigree information for each animal, contemporary group, sex of the pig and litter identification, birth date and date weighed, and measurement for weight (WT), backfat (BF), loin eye area (LEA), marbling score, pH, loin color (MINOLTA), and days to 250 lb (Days250).

Numbers of records, animals, litters, and contemporary groups are presented in Table 1. These data were from 21 herds from 11 firms (Table 1). Firms represent breeders with several herds at different locations. Of the firms, 6 firms have a single herd, and 5 firms have multiple herds. Herds represent animals at a single location. Number of herds and firms
which have records at each year is in Table 2. Only data from current breeders who employ an active genetic improvement program and follow ‘total herd reporting’ techniques were included. Table 2 indicates that only two of the breeders that currently use STAGES were actively reporting data in 1990. Means for off test weight (WT), backfat (BF), loin eye area (LEA), marbling score, pH, loin color (L*), and days to 250 lb (Days250) are presented in Table 3.

Contemporary groups (Table 1) were defined by individual breeders as a group of pigs that were raised in a common herd under similar environmental conditions. Data on boars, gilts, and barrows were included in the data set.

**Statistical analysis**

Data were analyzed according to the following single-trait model:

\[
Y_{ijkmn} = u + c g_i + s e x_j + l i t t e r_k + a_{ijkmn} + w t_n + w t_n^2 + e_{ijklm}
\]

where \(Y_{ijkm} = \) the trait measured on pig \(m\) in contemporary group \(i\) of sex \(j\) in litter \(k\); \(u = \) mean; \(c g_i = \) fixed effect of contemporary group \(i\); \(s e x_j = \) fixed effect of sex \(j\); \(l i t t e r_k = \) effect of litter \(k\), assumed random with \(\text{Var} = I \sigma_k^2\) where \(A^{-1}\) is the inverse of the additive genetic relationship matrix; \(a_{ijkm} = \) effect of animal \(m\) assumed random with \(\text{Var}(a_{ijkm}) = A^{-1} \sigma_a^2, wt_m\) and \(wt_m^2 = \) linear and quadratic effects of the offtest weight of pig \(m\); \(e_{ijklm} = \) random residual error with \(\text{Var}(e_{ijklm}) = I \sigma_e^2\). Variance and covariance components were estimated by the multiple-derivative-free restricted maximum likelihood program (MTDFREML) developed by Boldman (1995).
**Definition of connectedness statistics**

PEVD. Because the purpose of connectedness is to indicate the accuracy of comparisons between BVs across herds, a logical statistical measurement of connectedness across herds would be the average prediction error variance (PEV) of all pair-wise estimated BV (EBV) differences between the animals in those herds (PEVD):

\[
\text{PEVD} = \frac{(\text{Var}\left(\sum_i((u_i - \bar{u})) - \sum_i(u_i - u_j))\right)}{2}.
\]

Prediction error variance of differences (PEVD) is derived from the mean squared error (MSE) of prediction differences between candidates for selection. It can be calculated from the prediction models, both with and without genetic groups. It was reported that if PEVD was used as an indicator of connectedness, a large number of reference sire progeny (20% to 45% of the total number of progeny produced in that herd) were required to accurately compare BVs across herds (Foulley, 1983), and genetic groups should be fitted into the prediction model to manage potential bias from genetic mean difference across herds (Miraei, 1991). Although PEVD is a direct indicator of prediction accuracy of BV difference and the most precise way to show connectedness, computation of PEVD is difficult or impossible for large data sets in most situations and these values are difficult to interpret because no range is decided. Approximations that were used do not usually allow for computation of off-diagonal components of the prediction error variance-covariance matrix, which should affect PEVD, and as a result, alternative measures of genetic connectedness are required (Kennedy, 1993).

Mathur (2002) suggested using Connectedness Rating (CR) to define connectedness, which has been used in the Canadian Swine Improvement Program. Connectedness rating
was defined as the correlation between estimates of fixed genetic group effects.

Connectedness rating can be calculated as:

\[
CR = \frac{\text{cov}(\vec{h}_1, \vec{h}_2)}{\sqrt{\text{var}(\vec{h}_1) \cdot \text{var}(\vec{h}_2)}}
\]

In terms of variance of the difference of estimates of management units, connectedness rating can be written as:

\[
\text{Var}(\vec{h}_1 - \vec{h}_2) = \text{var}(\vec{h}_1) + \text{var}(\vec{h}_2) - 2CR \cdot \sqrt{\text{Var}(\vec{h}_1) \cdot \text{Var}(\vec{h}_2)}.
\]

Connectedness rating values range from 0 to 1 and high values mean more genetic connection (1 is perfect connection). The advantages of CR are that it is less dependent on genetic group size and structure, it can detect which pair of herds used for analysis has low genetic connection, and its computation is relatively easy. In addition, it can be calculated from prediction models both with and without genetic groups like PEVD, it has a high correlation with PEVD, and its values are easy to interpret because the range is between 0 and 1. The disadvantage of CR is that its relationship with bias depends on the linking strategy (using common reference sire) that is employed. These features, especially consistency against genetic group size and structure, mean that CR is good for checking the change of the bias level of specific groups with real industry data (unbalanced data), but is not effective for the comparison of linking strategies. Mathur (2002) reported that CR is usually higher for highly heritable traits such as backfat and age, and lower for low heritability traits such as sow productivity traits. According to Mathur (2002), based on an acceptable level of bias, average CR levels for each pair of groups should be 3% or more for highly heritable traits and 1.5% or more for lowly heritable traits. If substantial importation
of genetics into a herd is being considered, the minimum CR should be 5%. Above these thresholds, bias can be nearly eliminated.

Another connectedness statistic derived from PEVD is connectedness correlation ($R_{ij}$), proposed by Lewis (1999; 2005). Although $R_{ij}$ is a correlation between the influences of genetic groups like CR, the primary difference between $R_{ij}$ and CR is that $R_{ij}$ uses the means of animal random effects in the genetic groups, instead of using genetic group fixed effects (CR). It can be computed by:

$$r_{ij} = \frac{\text{PEC}(u_i, u_j)}{\sqrt{\text{PEV}(u_i) \cdot \text{PEV}(u_j)}},$$

where $u_{i(j)}$ is the mean EBV of all animals in genetic group $i(j)$, $\text{PEC}(u_i, u_j)$ is the prediction error covariance between these means, and $\text{PEV}(u_i)$ and $\text{PEV}(u_j)$ are the prediction error variances of the mean EBV of genetic groups $i$ and $j$, respectively. In terms of variance of difference of estimates of EBV means between genetic groups, $R_{ij}$ is:

$$\text{PEV}(u_i - u_j) = \text{PEV}(u_i) + \text{PEV}(u_j) - 2R_{ij}^* \sqrt{\text{PEV}(u_i) \cdot \text{PEV}(u_j)}.$$

Values range from 0 to 1 and high values mean more genetic connection (1 is perfect connection). The advantages of $R_{ij}$ are that relatively fewer data are required to obtain reliable values than for PEVD, CR, Connectedness Index (IC) (Foulley, 1990) and Coefficient of Determination (CD) (Laloë, 1993), and it can be calculated from prediction models both with and without genetic groups. In addition, it can detect which pair of herds used for analysis has low genetic connection, its relationship with bias is less dependent on linking strategy (using common reference sire), it has a high correlation with PEVD, and values are easy to interpret. The disadvantages of $R_{ij}$ are that its computation is more
difficult than CR and it may be affected by genetic group size and structure. Based on these features, connectedness correlation ($R_{ij}$) works well for numerous situations but is especially suited for simulation data, when compared with CR. Kuehn (2008) suggested that a flock $R_{ij}$ of 0.05 corresponded to approximately 80% reduction in bias compared with no connection, and a flock $R_{ij}$ of 0.10 corresponded to approximately 90% reduction in bias compared with no connection in simulated sheep data. This relationship between $R_{ij}$ and bias was relatively constant and independent of heritability (0.25 and 0.125). Therefore, he concluded that benchmarks of 0.05 for ‘good’ connectedness and 0.10 for ‘superior’ connectedness are warranted.

Common sire % (CS%) between herd $i$ and herd $j$ is the ratio of sum of the number of the progeny of a herd $i$ and herd $j$ which has common sire with herd $j$ and herd $i$ respectively to sum of the number of total progeny of a herd $i$ and herd $j$. 

$$CS\% = \frac{(PC_{(i)} + PC_{(j)})}{(P_i + P_j)}$$

$$= \frac{k \cdot (Sc_{(i)} + Sc_{(j)})}{P_i + P_j}$$

where $PC_{(i)}$ and $PC_{(j)}$ are the number of the progeny in herd $i$ and herd $j$ which has common sire with herd $j$ and herd $i$ respectively, $P_i$ and $P_j$ refer to the number of total progeny in herd $i$ and herd $j$ respectively, $Sc_{(i)}$ and $Sc_{(j)}$ are the number of common sires in the herd $i$ and herd $j$ with herd $j$ and herd $i$ respectively, and $k$ is the average number of progeny / sire. The significant advantage of this statistic is ease of computation when compared to other connectedness statistics. Other advantages are that it can be calculated regardless of model used, values range from 0 to 1, and it can detect low levels of connectedness. Disadvantages are that the correlation with bias may be less than other connectedness statistics and it may be
affected by herd size, which means that common sire % may be less accurate when herd size is small. Numerous researchers have proposed that exchanging common reference sires across herds is an effective method to establish connectedness across herds (Kennedy, 1993; Mathur, 2002; Kuehn, 2008).

**Calculation of each connectedness measure**

Progeny are grouped by birth year and herd to define genetic groups.

Prediction error variance of differences between the animals (PEVD). PEVD was calculated:

\[
PEVD = \frac{\sum PEV(u_i)}{P_i} + \frac{\sum PEV(u_j)}{P_j} - 2 \cdot \frac{PEC(\sum u_i, \sum u_j)}{(P_i + P_j)}.
\]

Individual EBV and their prediction error variances are calculated and averaged by genetic group:

\[
\frac{\sum PEV(u_i)}{P_i},
\]

where \(P_i\) is the number of total progeny in genetic group \(i\).

Prediction error covariance for genetic group \(i\) and \(j\) was calculated by calculating contrasts of the sum of the individual BVs for genetic group \(i\) and \(j\) and averaged by the number of progeny:

\[
PEC(\sum u_i, \sum u_j)/(P_i + P_j),
\]

where \(P_i\) and \(P_j\) are the number of total progeny in herd \(i\) and \(j\), respectively.

Kennedy (1993) stated that as connectedness increases, PEVD approaches a specific level, and the unbiased level of PEVD should be equal to heritability \(h^2\) multiplied by the
variance of environment (Ve) if there was no bias and the model contained only one random effect. This unbiased level of PEVD is defined as Base Line (BL_{PEVD}) where

$$BL_{PEVD} = h^2 \cdot Ve$$

(only one random effect).

The model used in this study contains a litter effect as a second random effect, so a modification of this equation is required. The ratio of the variance of the litter effect to the variance of the phenotypic effect is \(L^2\). The animal effect is nested in the litter effect, so the mean squared error of estimation of the animal effect can be written as \((h^2 - L^2) \cdot Ve\). Thus, the modified equation is:

$$BL_{PEVD} = (h^2 - L^2) \cdot Ve.$$ 

The difference between PEVD and this modified BL_{PEVD} can be considered as bias:

$$PEVD = BL_{PEVD} + \text{bias}.$$ 

Connectedness correlation (R). Prediction error variance (PEV(\sum u_i)) and covariance (PEC(\sum u_i, \sum u_j)) for genetic group \(i\) and \(j\) were estimated by calculating contrasts of the sum of individual BVs for each genetic group. Then,

$$R = \frac{PEC(\sum u_i, \sum u_j)}{\sqrt{PEV(\sum u_i) \cdot PEV(\sum u_j)}}.$$ 

Connectedness rating (CR). Variance of estimation of each genetic group effect (\(Var(\sum g_i)\)) and covariance (\(Cov(\sum g_i, \sum g_j)\)) for genetic group \(i\) and \(j\) were estimated by calculating contrasts of the sum of GCG effects in each genetic group. Then,

$$CR = \frac{Cov(\sum g_i, \sum g_j)}{\sqrt{Var(\sum g_i) \cdot Var(\sum g_j)}}.$$ 

Common sire % (CS%). Progeny were grouped by birth year and their herd of origin to define genetic groups. Then,

$$CS\% = \frac{(Pc_{(i)} + Pc_{(j)})}{(P_i + P_j)}.$$
for each pair of genetic groups.

Results and Discussion

Trends in PEVD of all pairwise BV differences, R, and CR from 1990 to 2008 for the largest 3 Duroc firms in the STAGEs program are presented in Figures 1, 3, and 5 respectively. Trends in the mean and standard deviation of PEVD of all pairwise BV differences, R, and CR for all Duroc herds in the STAGES program are presented in Figures 2, 4, and 6, respectively. Trends in CS% and GMD from 1990 to 2008 for the largest 3 Duroc firms in the STAGES program are presented in Figures 7 and 9 respectively. Trends in the mean and standard deviation of CS% and GMD for all Duroc herds in the STAGES program are presented in Figures 8 and 10, respectively. The PEVD decreased until 1997 and then slightly increased or was stable for the largest 3 firms. The same tendency can be seen for the mean of all herds. The R, CR, and CS% increased until 2002, and then decreased. The same tendency can be seen for the mean of all herds. The GMD decreased until 1997 and then slightly increased or was stable for the largest 3 firms and the same tendency can be seen for the mean of all herds.

The current levels of connectedness within the U.S. purebred Duroc breed are lower than each criteria associated with increased risk of bias (CR: 3-5%, R: 5-10%, CS%: 5-10%) (Soga, 2009; Mathur, 2002; Kuehn, 2008). This means with current connectedness levels, there is significant risk of bias, which decreases the accuracy of comparison of BVs across those herds (Soga, 2009).

Thus, there is a need to establish increased levels of connectedness within U.S. purebred nucleus herds for more accurate comparison of BVs through effective exchange of
common reference sires across nucleus herds. The required number of common reference
sires to eliminate bias was investigated using the CS% statistic. The definition of CS% is:

\[
CS\% = \left( \frac{P_{c(i)} + P_{c(j)}}{P_i + P_j} \right) \frac{k(S_{c(i)} + S_{c(j)})}{P_i + P_j},
\]

where \( P_{c(i)} \) and \( P_{c(j)} \) are the number of the progeny in herd \( i \) and herd \( j \) which have a
common sire with herd \( j \) and \( i \) respectively, \( P_i \) and \( P_j \) refer to the number of total progeny in
the herd \( i \) and herd \( j \) respectively, \( S_{c(i)} \) and \( S_{c(j)} \) are the number of common sires in herd \( i \)
and herd \( j \) with herd \( j \) and \( i \) respectively, \( k \) is the average number of progeny / sire. If
herd \( i \) and herd \( j \) have same number of progeny and common sires, the equation becomes:

\[
CS \% = \frac{kS_{c(i)}}{P_i}.
\]

To eliminate bias, CS% of more than 10% is required (Soga, 2009). So, the criteria can be
written as:

\[
S_{c(i)} > \frac{P_i}{10k}.
\]

If \( S_{c(x)} \) is the number of common sires in year \( x \), REP is the replacement rate of each sire,
and \( Ex \) is the number of common sires that are used, then,

\[
S_{c(x+1)} = S_{c(x)} \times (1-REP) + Ex.
\]

Thus, the \( Ex \) which is required to achieve CS% of 10% in \( Y \) years is:

\[
Ex = \frac{10+P}{10k + \sum_{1}^{year}(1-REP)^{Y-1}}.
\]

In this study, average genetic group size (\( P \)) was approximately 600 and \( k \) was approximately
30. So, if REP is 40%,

\[
Ex = \frac{2}{\sum_{1}^{year}(1-REP)^{Y-1}}.
\]
Thus, if CS% is increased to 10%, the required Ex is 2, 1.25 and 1.02 for years 1, 2 and 3 respectively. So, in order to establish enough connectedness level, the minimum and effective exchanging sire number may be 2 progenies per year. From previous studies (reference), it has been recommended that the number of sires exchanged between herds is two per year (Kuehn, 2008). However, if total progeny number is less than 100 in the genetic group, the CS% values contain a lot of outliers, which means low accurate (Soga, 2009). As a result, recommendations of previous studies only hold true when genetic group size is larger than 100 and a 10% CS% level is used as a threshold.

For further study, in order to investigate the effect of heritability and record density in multiple trait models, analysis using additional traits of interest such as maternal traits which have lower heritability and are sparsely recorded will be required. Different heritability levels can be evaluated with the current dataset (i.e., pH vs. Backfat). An attempt to investigate the effect of progeny number with additional data may be also useful and other breeds should also be studied. Using these results, a more detailed simulation study of the effect of exchanging common sires is warranted. Further studies could evaluate the needed number of common sires for maternal traits where records are obtained only in one sex and later in life, as well as simulation studies that examine the effectiveness of sire sampling programs.


**Literature Cited**


Soga, N.. 2009. M.S. Thesis. Iowa State University, Ames, IA.
Table 1. Structure of data of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th>Number of Records</th>
<th>Number of Animals</th>
<th>Number of Litters</th>
<th>Number of Contemporary Groups</th>
<th>Number of Herds</th>
<th>Number of Firms</th>
</tr>
</thead>
<tbody>
<tr>
<td>156,101</td>
<td>160,325</td>
<td>33,210</td>
<td>3,065</td>
<td>21</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2. Number of herds and firms in each year of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Herds</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Firms</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herds</td>
<td>15</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td>16</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Firms</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 3. Number of observation, mean, and standard deviation for measured traits of U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

<table>
<thead>
<tr>
<th>Traits</th>
<th>Number of observation</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight (lb)</td>
<td>156,101</td>
<td>108.86</td>
<td>12.68</td>
</tr>
<tr>
<td>Backfat (in.)</td>
<td>156,091</td>
<td>1.45</td>
<td>0.43</td>
</tr>
<tr>
<td>Loin eye area (in2.)</td>
<td>142,133</td>
<td>17.50</td>
<td>2.51</td>
</tr>
<tr>
<td>Marbling score</td>
<td>1,410</td>
<td>2.48</td>
<td>0.96</td>
</tr>
<tr>
<td>pH</td>
<td>1,372</td>
<td>5.76</td>
<td>0.22</td>
</tr>
<tr>
<td>Meat color (L*)</td>
<td>1,383</td>
<td>49.77</td>
<td>3.75</td>
</tr>
<tr>
<td>Days to 250 lb (Days250) (day)</td>
<td>156,095</td>
<td>161.9</td>
<td>12.8</td>
</tr>
</tbody>
</table>
Figure 1. Trend in prediction error variance of all pair-wise estimated BV differences between animals in those herds (PEVD) for the largest 3 firms in the U.S. purebred Duroc population from the National Swine Registry. Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

Figure 2. Trend in the mean and standard deviation of change of prediction error variance of all pair-wise estimated BV differences (PEVD) for all herds in the U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.
Figure 3. Trend in connectedness correlation (R) for the largest 3 firms in the U.S. purebred Duroc population from the National Swine Registry. Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

Figure 4. Trend in the mean and standard deviation of change of connectedness correlation (R) for all herds in the U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.
Figure 5. Trend in connectedness rating (CR) for the largest 3 firms in the U.S. purebred Duroc population from the National Swine Registry. Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

Figure 6. Trend in the mean and standard deviation of change of connectedness rating (CR) for all herds in the U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.
Figure 7. Trend in common sire % (CS%) for the largest 3 firms in the U.S. purebred Duroc population from the National Swine Registry, Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

Figure 8. Trend in the mean and standard deviation of common sire % (CS%) for all herds in the U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.
Figure 9. Trend in genetic mean difference (GMD) for the largest 3 firms in the U.S. purebred Duroc population from the National Swine Registry. Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.

Figure 10. Trend in the mean and standard deviation of change of genetic mean difference (GMD) for all herds in the U.S. purebred Duroc population from the National Swine Registry Swine Testing and Genetic Evaluation System (STAGES) from 1990 to 2008.
CHAPTER 5. SUMMARY

Even though breeding values (BV) for each pig are estimated accurately, there is a risk that comparison of individual BV from different herds can be negatively biased. The cause of this bias mainly comes from the assumption that genetic means (means of BVs) of each herd are the same. In many cases this assumption is not valid, which may negatively bias the accuracy of BV comparisons across herds. To indicate the degree of bias, many researchers have studied “connectedness”, genetic similarity between herds. Few comparisons between many connectedness statistics are available. Most researchers used simulation data to calculate connectedness rather than field data due to computational requirements in large populations.

In addition, no research indicates current connectedness levels of the swine industry in the United States. There were 3 objectives for this study. The first objective was to define several methods of estimating connectedness and compare advantages and disadvantages of each. The second objective was to evaluate levels of connectedness and prediction error of difference of BV between animals (PEVD), using purebred swine industry data of the Duroc breed from 21 herds over 19 years. The third objective was to investigate the relationship between connectedness and bias, and between different methods of estimating connectedness.

In this study, connectedness rating (CR), connectedness correlation (R), common sire % (CS%) were chosen as indicators of connectedness. The results of this study indicate that increasing values for connectedness decreases bias (low PEVD) of comparisons of BV across herds. These 3 connectedness statistics were highly and positively correlated. The correlations between these 3 connectedness statistics and bias were largest for R, and
smallest for CS%. Even though R was the most accurate indicator of connectedness, this
statistic was computationally demanding to obtain. Therefore, CR was most suitable as an
indication of connectedness, due to its ease of computation. The CS% can also be an
indication of connectedness, but only when pairs of herds are relatively large (>100). When
these connectedness statistics become lower than approximately 10%, the risk of bias
significantly increases. These biases decrease accuracy of comparison of BV, which in turn
decrease genetic response to selection. Level of connectedness in the current Duroc breed
over 19 years (1990 to 2008) was estimated. As a general tendency, connectedness (CR, R
and CS %) increased until 2003 but after that, they started to decrease. Because these
connectedness levels were less than 10% from 1990 to 2008, there is risk of bias in reporting
accuracy of BV prediction in the Droc breeds. Thus, exchanging common reference sires is
required to establish well connected herds for accurate comparison of BV across herds.

For further study, in order to investigate the effect of heritability and record density
in multiple trait models, analysis using additional traits of interest such as maternal traits
which have lower heritability and are sparsely recorded will be required. Different
heritability levels can be evaluated with the current dataset (i.e., pH vs. Backfat). An attempt
to investigate the effect of progeny number with additional data may be also useful and other
breeds should also be studied. Using these results, a more detailed simulation study of the
effect of exchanging common sires is warranted. Further studies could evaluate the needed
number of common sires for maternal traits where records are obtained only in one sex and
later in life, as well as simulation studies that examine the effectiveness of sire sampling
programs.
LITERATURE CITED


Soga, N. 2009. M.S. Thesis. Iowa State University, Ames, IA.


ACKNOWLEDGEMENT

I would like to express my sincere appreciation to all the people who supported and challenged me over my years as a graduate student.

I wish to extend my deep appreciation to Dr. Thomas J. Baas for his sincere guidance and friendship during my years at Iowa State University as my major professor. His advice encouraged me many times. I appreciate his tremendous patience.

My thanks also go out to my committee members. Dr. Clint Schwab helped me a lot to solve many problems I faced during this research and gave me a lot of advice. Dr. Theodore Bailey advised me to improve this project. Dr. P.J. Berger was dedicated to my education. Without his great help, I could not accomplish my work. Thanks also to Dr. Matt Spangler. Discussion with him gave me many ideas. I also want to thank Hiroshi Takahasi, who helped me a lot to prepare to study abroad and to choose this great environment to study.

I am truly grateful to my parents, Masashi Sato, Keiko Sato and all my families. They encouraged me many times, and supported me in many ways. You always supplied me with the confidence to accomplish all of my goals. I’m proud of all of you, from my deep heart. And, to my grandfather, Tatsuo Soga, thanks to you, I could find my dreams. As you said to me, I will continue to do my best to contribute to the world. Again, thank you.

Finally, Yoko, my wife, I really appreciate you for being with me for good times and tough times. You always trust me, help me, make me smile, and love me. I’m so glad that I can walk with you. Thank you, with all my love.