

# Rewarding High-Quality Data with Higher Accuracies

MARK THALLMAN AND BAILEY ENGLE

---

U.S. MEAT ANIMAL RESEARCH CENTER

BEEF IMPROVEMENT FEDERATION ANNUAL MEETING

CALGARY, ALBERTA – JULY 5, 2023

# Overview

---

- Rewarding High Quality Data
- Better EPDs: A new approach to contemporary groups and other model improvements
- Technical Aspects of Implementation

A photograph of a herd of cows in a grassy field, overlaid with a dark grey semi-transparent layer. The cows are in various colors, including brown, black, and white. The text 'Rewarding High Quality Data' is centered in white. A thin white horizontal line is positioned below the text. A solid green horizontal bar is at the bottom of the image.

# Rewarding High Quality Data

---

# Introduction

---

1. Some breeders report higher quality data for use in genetic evaluation than others.
2. Current genetic evaluations do not reward higher quality data with higher reported accuracies, although the EPDs they produce are, in fact, more accurate.

**Nonetheless, it should be possible to report higher accuracy for EPDs derived from higher quality data.**

# What is Meant by “High Quality Data”?

---

Measurements are made accurately

- Using appropriate instruments that are read correctly

Measurements are made under appropriate and uniform conditions

Pedigrees are recorded accurately and verified (and corrected) by genomics

Contemporary groups are formed correctly

All optional information that is requested is provided

Information is recorded with the intent of obtaining accurate evaluations and not with the intent of obtaining favorable evaluations

# Objective

---

Genetic evaluations should consider data quality in 2 distinct but related ways:

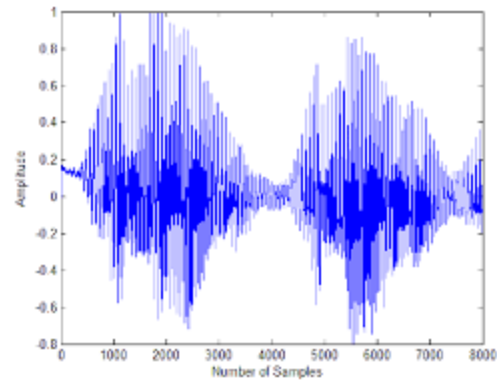
1. They should increase the true accuracy (and effectiveness of selection) of animals whose evaluations are based on high quality data
  - Furthermore, the accuracy that is reported along with EPDs should be more reflective of true accuracy than it currently is
2. They should put more emphasis on high quality data and deemphasize low quality data
  - The effect of this should be to increase the overall accuracy of the genetic evaluation, and consequently, the genetic trend of the population

*Fortunately, the same improvement to the genetic evaluation model can have both effects.*

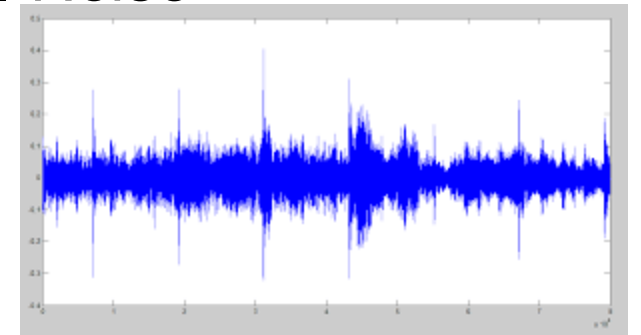
# Signal vs Noise

---

Noise Obscures Signal



Signal is Clear Because of Minimal Noise



# Signal vs Noise

---

In cattle breeding:

- The signal is true breeding value

Noise comes in many forms:

- Measurement error.
- Pedigree error
- Environmental variation
- Pulling bulls from feed intake test to collect semen on them

We can't eliminate noise, but there are things breeders, breed associations, and genetic evaluation providers can do to reduce noise and increase rate of genetic progress

The signal to noise ratio determines the rate of genetic improvement

Less noise results in higher heritability and greater accuracy



# Specific Approach to Rewarding High Quality Data



Matt Spangler has presented an approach to solving this problem previously at BIF

Matt's approach is based on artificial intelligence and is specific to traits and situations

In this paper, the trait is birth weight and the specific data quality problem is that some records are weights and some are not

# General Approach to Rewarding High Quality Data

---


1. Estimate the amount of noise in the data submitted by each ranch
2. Adjust the genetic evaluation model to put more emphasis on high quality data and deemphasize low quality data based on the estimated amounts of noise

*More details on implementation are coming later in the presentation*

This approach is general in that it would be applied automatically to all traits

There would be advantages to superimposing this approach on the specific approach that Matt and colleagues have described

The ranch-specific heritability of each trait could be reported to breeders so they could benchmark themselves against their peers for quality of data submitted

A herd of cows of various colors (black, brown, white, and spotted) is grazing in a lush green field. In the background, there is a dense line of dark evergreen trees. The overall scene is a typical farm landscape.

# Better EPDs: A new approach to contemporary groups and other model improvements

# Introduction

---

1. Concepts related to contemporary groups and how they are fit in genetic evaluations have remained essentially unchanged for about the last 5 decades
  - The way they were set up originally made sense when the most powerful computers available filled entire rooms and had many orders of magnitude less capacity than a current phone
  - They make much less sense today and genetic evaluations could be improved considerably by making some relatively simple modifications to the model and information collection
2. There are other effects that should be added to models to remove some of the noise from the data
3. Together, these should result in a “quantum leap” in genetic evaluation

# Fitting Contemporary Groups as Random Effects Instead of as Fixed Effects

---

Two different ways of fitting the statistical model:

1. Fixed: Each contemporary group has its own average independent of others
2. Random: Contemporary groups from the same ranch share information such that their averages are more similar than they would be if fit as fixed effects

Changing from fixed to random effects has minimal effect on large contemporary groups but can have substantial effects on small contemporary groups

Fitting contemporary groups as random effects is “theoretically correct” but, for practical reasons, beef contemporary groups have been fit as fixed effects

Random contemporary group effects need an overall average to be centered around and, for beef field data, the overall average should be of the ranch

# Proposed Definition: Ranch ID

---

Distinguishes cattle produced by the same breeder under similar conditions over multiple years

In most cases, this can be the member ID of the owner at the time the performance data is collected

But in cases of comingled ownership within contemporary group designation, as occurs in some family operations and other situations, a Ranch ID should be constructed to represent the group of owners that comingle their cattle.

And some breeders may have ranches in different environments

- They should have separate ranch IDs

# Sex of Calf

---

When managed together as contemporaries, groups should not be split by sex

- e.g., for birth and weaning traits

Splitting contemporary groups on sex is one of the greatest contributors to contemporary groups that are too small to be useful

We can do a better job of estimating differences between sexes by considering far more calves than would ever be included in one contemporary group

# Contemporary groups do not need to become continually sub-divided over time

---

*This is the other primary contributor to contemporary groups that are too small to be useful*

Contemporary groups should be formed based on how cattle were grouped immediately prior to measurement (e.g., grouping from weaning to yearling)

e.g., weaning contemporary group is currently part of the definition of yearling contemporary group

The data should be adjusted to reflect prior (e.g., preweaning) groupings





# Make New Information Requests Optional and Reward Them

---

For example, when a large group of cattle is gathered to be weighed individually, they will shrink while waiting to be weighed

- Ordinarily, this shrink contributes to noise
- If breeders report the time each weight is taken, the genetic evaluation could adjust for the expected shrink of each animal and thereby reduce the noise of those weight records
- But most breeders won't report the times

At first, only those breeders that highly prioritize accuracy and genetic progress will likely participate.

Rewards for participation:

- Improved EPD accuracy
- Beneficial weighting of submitted high quality data
- Those ranches will have higher heritabilities than those that don't report the optional information

# Where Do Changes Need to be Made?

---

## BIF Guidelines

- Define additional terms
- Recommend model improvements

## Breeders

- Correctly group animals and provide optional information

## Breed Associations

- Provide the opportunity for breeders to differentiate sub-classes and provide optional information

## Genetic Evaluation

- Implement improved models

A herd of cows of various colors (brown, black, white, and grey) is grazing in a lush green field. In the background, there is a dense line of evergreen trees under a grey, overcast sky. The text "Technical Aspects of Implementation" is overlaid in white, bold font across the center of the image.

# Technical Aspects of Implementation

# Implementation of Rewarding High Quality Data

---

Fit residual variances that are heterogeneous by Ranch ID

Estimate the heterogeneous residual variances as follows:

- For each trait, condition on all (co)variance parameters except the residual variance parameters for that trait.
- For each Ranch ID,  $i$ , within trait,  $t$ , estimate residual variance as:

$$\hat{\sigma}_{(t)i}^2 = \mathbf{y}_{(t)i}' \hat{\mathbf{u}}_{(t)i} / (n_{(t)i} - r(\mathbf{X}_{(t)i}))$$

- Update residual covariances between traits  $s$  and  $t$  as:

$$\hat{\sigma}_{(st)i} = \rho_{(st)} \sqrt{\hat{\sigma}_{(s)i}^2 \hat{\sigma}_{(t)i}^2}$$

where  $\rho_{(st)}$  is the homogeneous residual correlation between traits  $s$  and  $t$ .

A different, but generally similar, approach should be feasible to implement in the MCMC computing approach implemented by BOLT

# Fitting Contemporary Groups (**CG**) as Random Effects

---

Requires Ranch ID fit as fixed effect to avoid bias

- Probably also requires within-ranch regression on year to account for genetic trend

Multiple traits that share CG definitions should have correlated CG effects

- e.g., post-weaning gain, ultrasound traits, and scrotal circumference
- This would allow sharing of CG information across traits

Variance due to CG does not contribute to phenotypic variance for purposes of computing heritability

# Management Code

---

Historically and currently used for 2 different purposes:

1. Indicates systematic differences between groups of contemporaries, e.g.,
  - Range
  - High quality pasture
  - Feedlot
  - Show barn
  - 1<sup>st</sup> Calf heifers if managed separately
2. Designate groups of animals that were managed relatively similarly but not together and thus should not be considered contemporaries

With fixed contemporary groups, the same terminology and data field can be used for both

With random contemporary groups, 2 can be used to designate separate contemporary groups but 1 should be nested within Ranch ID to specify different means

- Therefore, a change in terminology and data capture is needed

# Is Separate But Equal Really Contemporary?

---

The BIF Guidelines currently say YES

- “A contemporary group is a set of same-sex calves that were born within a relatively short window of time and have been managed the same since birth.”
- With fixed CG, there is an argument that small groups of calves that were managed similarly but separately should be combined into one CG
- I have long argued that large groups of calves that are raised in separate groups should be in different CG regardless of how similarly they are managed

With CG fit as random effects nested within Ranch ID, calves raised in separate groups should be reported as different contemporary groups

# What Happens When 2 Groups That Should be Different CG's are Reported as the Same?

**Reported accuracy** is higher than if they were separate CG

Depending on the actual difference between the group means, **true accuracy** may be considerably lower than if they were separate CG.

Consequences are:

- Correlation between EPD and true breeding value is lower
- Response to selection is lower



# How to Combine CG when Groups Get Larger Over Time

---

- Condition on previous contemporary group effects
- Minimal computational cost
- Especially important for cow traits
- Will be even more important when we start collecting various kinds of high-throughput phenotypes on a more-or-less continuous basis

# Regression on Birth Date

---

For many traits (e.g, BWT, WWT, YWT, HP, carcass), it would make sense to fit a random, within-cg regression on age

- After 205-adjustment for WWT
- Fixed linear regression vs random regression

This approach should allow for wider age ranges within CG.

Categorizing by age is perhaps not the issue it once was with current, very short breeding seasons

- But we should eliminate birth weight slices
- Birth contemporary groups should be based on what cows were contemporary

# Effects to Account for ET vs AI vs NS

---

We should probably include a cross-classified effect to account for differences between AI- and cleanup-sired calves

These could be a combination of genetic and temporal effects

- That makes it complicated because the EPDs should be adjusted for the genetic part

But ignoring it may be dangerous as well

- So this may be an area that requires more work

A similar issue may exist for ET vs AI

# Regression on Time of Measurement

---

Many weights are now taken by electronic scales with automatic data transfer

- In many cases, it is possible to automatically record timestamps for the weights
- In other cases, it is feasible to at least collect the order in which animals were weighed
- Within-contemporary group regression on time of measurement would account for shrink

# Carcass Data from Pens that are “Topped Off”

---

Standard practice and BIF guidelines would break each harvest date into a separate and disconnected CG

- This causes loss of variation and big problems

The pen needs to be considered the CG

- Accounting for harvest date determined by selection needs to be accounted for within CG
- How to accomplish this this is not known to me
- Nonetheless, I think it is plausible that this approach could make it feasible to express carcass traits on multiple endpoints
  - If this could work, it would substantially reduce the cost of collecting carcass data

# What Else We Can Add to the Model that Adds Value to the Evaluation?

---

Effects that can be used for management purposes

- Estimation of effects of IVF vs MOET vs AI
- Estimation of alternative IVF protocols on phenotypes

Anything else that could be used to replicate the dairy model of driving data collection for management purposes and scavenging the data for genetic evaluation purposes

A photograph of a herd of cattle in a green field under a blue sky with white clouds. The text "Take Home Points" is overlaid in white, with a thin white horizontal line underneath it. The bottom of the image has a solid green bar.

# Take Home Points

# Fixed Effects to Reduce Noise

---

- Ranch ID
- Management Code nested within Ranch ID
- Sex of Calf (cross-classified)
- Age of Dam Effects (cross-classified)
- ET vs AI vs Natural service nested within Ranch ID (cross-classified)
- Regressions on previous CG effects within Ranch ID
- Regression on birth date within CG
- Regressions on time of weight within CG



# Random Effects to Reduce Noise

---

- Contemporary Group (nested within ranch ID)
- Random effects per Ranch to consider nesting within overall cross-classified fixed effects:
  - Sex of Calf
  - ET vs AI vs Natural service
  - Age of Dam Effects

# Effects on Big Breeders vs Small Breeders

---

Both large and small breeders will benefit

Small breeders will benefit more

# Acknowledgements

- Matt Spangler
- Larry Kuehn
- Warren Snelling



USDA is an equal opportunity provider and employer. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.