Accounting for Discovery Bias in Genomic Prediction

Jamie T. Parham, Mark Thallman,
Larry Kuehn

June 16, 2016
Beef Improvement Federation Annual Meeting & Symposium
Manhattan, Kansas

South Dakota State University USDA

Genomic Enhanced EPDs (GE-EPDs)

- Enhance response to selection in traits:
  - Difficult to measure
  - Low heritability
  - Measured late in life
  - Sex-limited
- Current methods of genomic selection being used by many breed associations
  - Angus, Hereford, Simmental, Charolais, Red Angus, Limousin, Gelbvieh, Brangus, Santa Gertrudis, etc.

South Dakota State University USDA

Alternative Models Underlying Genomic Selection

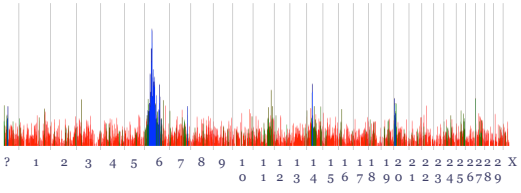
Unweighted	Weighted
• Every marker given equal emphasis ▫ Realistic	• Marker emphasis weighted by [illegible] trait of [illegible]
• Works well [illegible] candidates are closely related to phenotyped and genotyped animals ▫ Accuracy diminishes rapidly with relationships	[illegible]tion candidates are closely related to phenotyped and genotyped animals ▫ Accuracy is less dependent upon relationships

Unweighted analyses not believed to be subject to discovery bias

South Dakota State University USDA

What is Weighting?

- Emphasis given to markers believed to have greater effect on traits of interest
  - Based on training population
  - Different weights for different traits



What is Discovery Bias?

- Occurs because of double counting
  - Same data is used to estimate SNP effects as is used for prediction of breeding values
    - "Winners Curse" (Goddard et al., 2009; Xu et al., 2011)
- Not only a bias of the predictions
  - Also a bias of the accuracy (usually overstated)
    - Difference between model-derived and true accuracies

South Dakota State University USDA

Model Derived Accuracy

- Computed from the inverse of the Mixed Model Equations
  - Prediction Error Variances (PEV)
- Requires that discovery data is analyzed simultaneously with prediction (Single-Step or One-Step)

South Dakota State University USDA

Effect of Discovery Bias

- Accuracy is overstated
  - MBV appear more accurate than they actually are
  - Blending methods of calculating GE-EPD weight genomic portion by its accuracy
    - Too little emphasis placed on phenotypic information
    - Inflation of genomic variance can cause genomic effects to be reported on an inflated scale relative to information derived directly from phenotypes

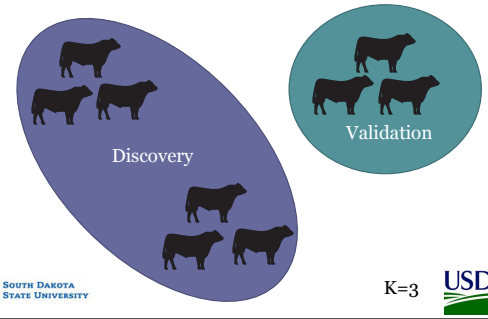
South Dakota State University USDA

Why Use Weighted Analyses?

- Generally obtain greater true accuracy than from unweighted
  - Especially when animals to be predicted are not closely related to those in training
    - At least partly due to the model matching the underlying biology more closely
- Discovery bias is the price we pay for greater true accuracy
  - How do we account for it?

South Dakota State University USDA

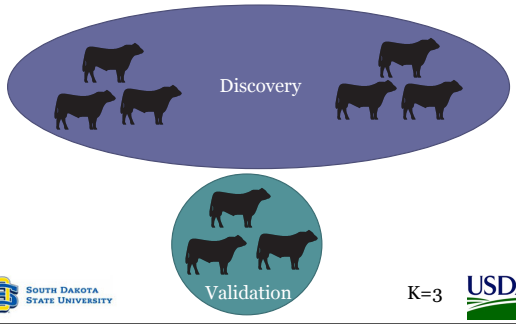
K-Means Validation



K=3

South Dakota State University USDA

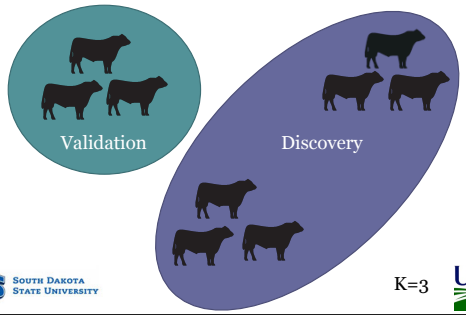
K-Means Validation



K=3

South Dakota State University USDA

K-Means Validation



K=3

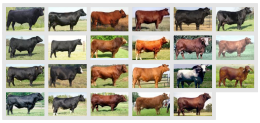
South Dakota State University USDA

Objective

- Investigate the topic of discovery bias and propose a way in which to partially account for it
  - Determine whether removing groups of animals from a pedigree, with their phenotypes, during training would reduce discovery bias resulting from their records being used in training

South Dakota State University USDA

Taking K-Means Validation to the Limit



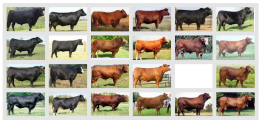
Discovery

- Each animal is its own group.

- Estimate marker effects conditional on Discovery subset of the population
- Apply those effects to the Prediction Animal's marker data.

k = n

Cycle Through Each Animal in the Population



Discovery

- Each animal is its own group.

- This results in a Corrected Molecular Breeding Value (CMBV)
  - own information was not used in estimation of the marker effects.

k = n

Molecular Breeding Value (MBV)

- Summary of the genetic merit of an individual as measured by genomic effects
- Computed as a sum of SNP effects

South Dakota State University USDA

SNP Analysis Model

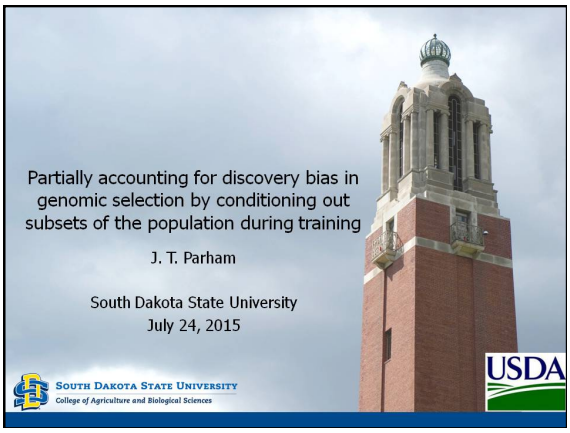
- One random effect per SNP, each with its own variance parameter
  - i.e. Weighting
- BLUP predictions of SNP effects and REML estimates of SNP variances
  - Deterministic algorithm (not MCMC)

South Dakota State University USDA

Correcting for Discovery Bias

- Step 1: Estimate marker effects conditional on the entire population (only once)
  - Results in an MBV for each individual
- Step 2: Adjust the effects for each individual for dropping its information from the data set
  - Results in a Corrected MBV (CMBV)
  - This latter step requires minimal computational effort, even though it is applied to each individual in the population

South Dakota State University USDA



Materials and Methods

- Animals
  - Cattle in the GPE population at USMARC
    - Cycle VII of GPE (Snelling et al., 2010)
    - Represented 18 industry breeds
  - Used 2,600 animals with BovineSNP50 genotypes
    - 107 sire groups (1-107 animals per group)
    - Simulated phenotypes for non parents only
      - Resulted in 107 non overlapping paternal half sib groups
      - Population structure simplified computations for replicated data

South Dakota State University USDA

Simulated Phenotypes

$$\mathbf{y} = \mathbf{Qq} + \mathbf{e}$$

y = vector of simulated phenotypes
Q = matrix of Quantitative Trait Nucleotide (QTN) genotypes
q = vector of simulated QTN effects
e = vector of simulated residuals

True Breeding Value (TBV) = **u** = **Qq**

South Dakota State University USDA

SNP and QTN Selection

- Used real data from the BovineSNP50 chip
- Total of 2,500 SNP used as markers for MBVs
  - Selected regions of 250 nearly contiguous SNP on each of 10 chromosomes (2,500 total SNP)
  - Monomorphic SNP were removed
- Total of 35 SNP selected as QTN
  - Moderate frequency (q = 0.29-0.30)
  - Located within regions of the 2,500 marker SNP

South Dakota State University USDA

Analysis of Simulated Data

- Step 1: Estimate marker effects conditional on the entire population (only once)
- Step 2: Adjust the effects for each paternal half sib group by dropping their information from the data set
  - Removed closest relatives for each individual

South Dakota State University USDA

Realized vs. Model Derived Accuracy

Comparison of Accuracies of Simulation (n_r = 105)

	Realized Accuracy		Model Derived Accuracy		Discovery Bias
	Mean	SE	Mean	SE	Mean
MBV	0.687	0.006	0.960	7.11e-5	0.273
CMBV	0.620	0.007	0.954	7.59e-5	0.334

MBV = Uncorrected Molecular Breeding Value; CMBV = Corrected MBV

- Realized Accuracy is the correlation between MBV and TBV
- Model derived accuracy computed from PEV

South Dakota State University USDA

Problem…

- Dropping phenotypes of each individual and their closest relatives had a negative effect on the realized accuracy
- Challenge to incorporate phenotypes without influencing the estimation of SNP effects

South Dakota State University USDA

A Solution: Two Trait Post-Analysis

- Used to reintroduce phenotype into the analysis
- Trait 1: simulated phenotype
  - GE-EPD are predictions of trait 1
- Trait 2: MBV/CMBV from analysis excluding paternal half sib group

Realized vs. Model Derived Accuracy

Comparison of Accuracies of Simulation (n_s = 105)

	Realized Accuracy		Model Derived Accuracy		Discovery Bias
	Mean	SE	Mean	SE	Mean
MBV	0.687	0.006	0.960	7.11e-5	0.273
CMBV	0.620	0.007	0.954	7.59e-5	0.334
GE-EPD	0.716	0.006	0.942	6.85e-5	0.226
★CGE-EPD	0.721	0.007	0.865	5.84e-5	0.144

MBV = Uncorrected Molecular Breeding Value; CMBV = Corrected MBV;
GE-EPD = Uncorrected Genomic Enhanced Expected Progeny
CGE-EPD = Corrected GE-EPD

Conclusions from Thesis

- Method of accounting for bias decreased gap between model derived and realized accuracy
  - Believed to partially account for bias
- True accuracies of genomic prediction were higher when accounting for bias than when not accounting for it
- Accounting separately for polygenic effects improved accuracy for uncorrected predictions
  - Using the two trait model
- Overall, promising "first step" method

Moving Forward

- Removing paternal half sib groups was an expedient proof of principle
  - Not a final solution
- Optimal correction for each individual in a general pedigree is yet to be determined

The Traditional Genetic Prediction Paradigm

- A single analysis conducted for an entire population
  - Optimized to predict differences among all members of the population simultaneously
  - Results in an unnecessary constraint on optimization

A New Genetic Prediction Paradigm

- Individual analyses conducted to optimize prediction for each animal within a population
  - Excluded data customized for each individual
  - Predicts the individual relative to population
    - Rather than directly to specific individuals
  - May require small base adjustment for each individual
- Requires thinking outside of the traditional "box"

Conclusion

- Weighted and Unweighted analyses differ
  - Weighted analyses subject to discovery bias, but can result in more accurate predictions
    - Improvement in accuracy especially important for beef cattle populations
- Although we may not be able to eliminate the effects of discovery bias, we can mitigate them
  - Results of thesis research promising and supportive

South Dakota State University USDA

Questions?

South Dakota State University USDA