Capture-Recapture in the Age of Artificial Intelligence

Robert L. Emmet, Luca Sartore, Habtamu Benecha, and Bruce A. Craig

Federal Committee on Statistical Methodology

October 22nd, 2024

Disclaimer: The findings and conclusions in this report are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Outline

- Motivation
  - US Census of Agriculture
  - Potential uses of administrative data
- Methods
  - Triple-System Estimation
  - AI extension that uses neural networks
- Case Study and Simulation Results
- Conclusions

Motivation – Census of Agriculture

- The Census of Agriculture is a complete count of U.S. farms, ranches, and producers
  - Conducted every 5 years; 2022 Census data recently published
- Based on Census Mailing List (CML)
  - Some undercoverage, mainly for smaller and newer farms
  - USDA definition of a farm is $1,000 in sales or potential sales of agricultural products – can be very small farms

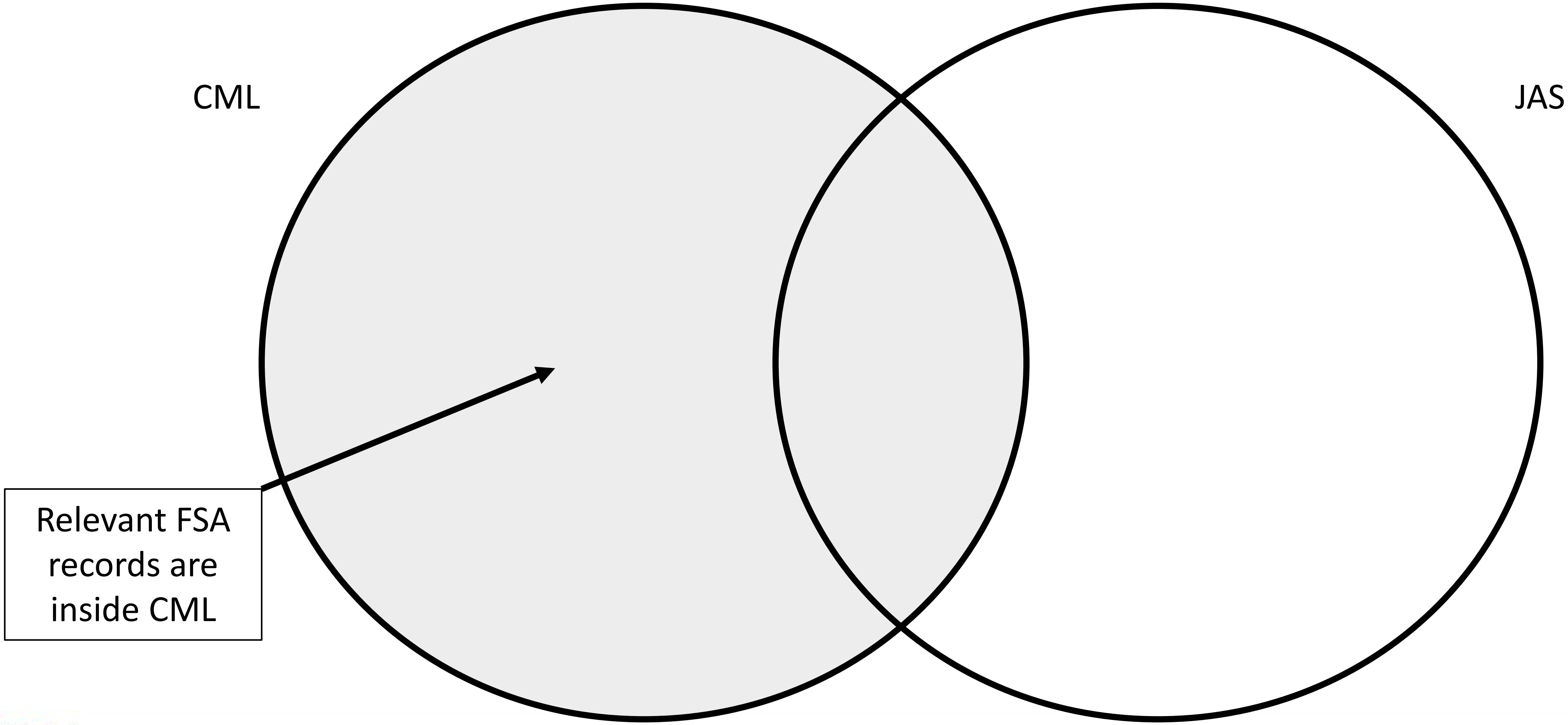
Motivation – Census of Agriculture

- NASS uses the June Area Survey (JAS), an area frame survey, to adjust CML responses for:
  - Undercoverage
  - Nonresponse
  - Misclassification of farms as non-farms, and vice versa
- NASS uses a dual-system estimator based on CML and JAS as two independent lists

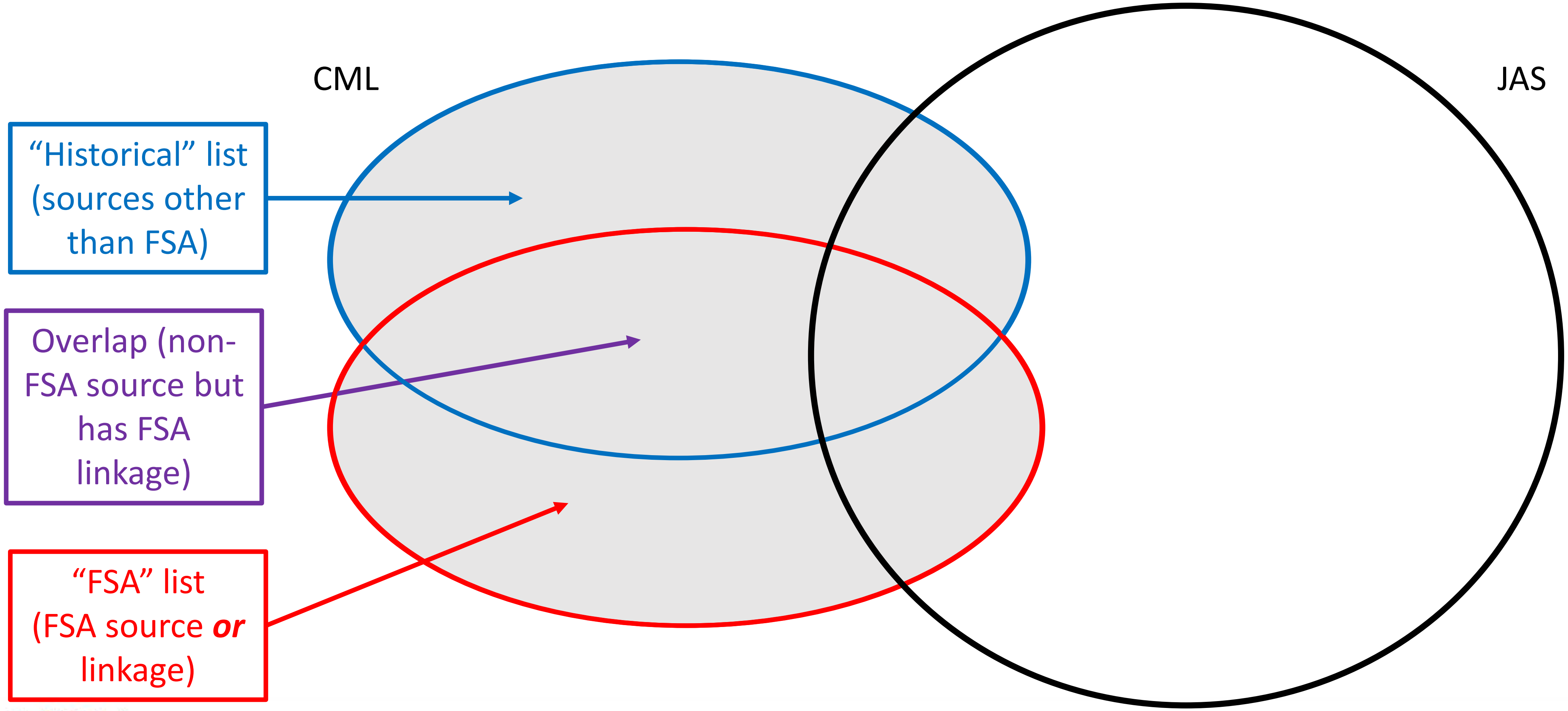
Motivation – Farm Service Agency (FSA) Data as a Third List

- NASS uses FSA administrative data for a variety of purposes
  - Including list-building for the CML
- Using FSA as a third list may reduce variance, but
  - Most FSA records are referred onto the CML
  - Dependence between FSA list and NASS' pre-existing list frame
- A Triple-System Estimator (TSE) can account for list dependence, undercoverage, and nonresponse, using FSA records on the CML as a third list
- **We propose an Artificial-Intelligence TSE (AITSE) to better model nonlinear effects in the data**

Methods – Splitting the CML to Bypass FSA Referral Problem



Methods – Splitting the CML to Bypass FSA Referral Problem



Methods – Triple-System Estimator

- The model uses a multivariate Bernoulli distribution
  - Different link functions to model different conditional list coverage probabilities
  - Allows calculation of conditional and marginal coverage
- Joint coverage and response probabilities are summed for CML respondents to estimate total number of farms

$$\hat{N}^* = \sum_{i \in C \cap R} \frac{1}{\hat{\pi}^*_{i,1111} + \hat{\pi}^*_{i,1101} + \hat{\pi}^*_{i,1011} + \hat{\pi}^*_{i,1001} + \boxed{\hat{\pi}^*_{i,0111}} + \hat{\pi}^*_{i,0011}}$$

Not on historical, on JAS, on FSA, responded

United States Department of Agriculture
National Agricultural Statistics Service

Methods – Traditional Probability Models

- For given predictors $X_1, X_2, \dots, X_p$, the model is specified for a generic probability
$$\pi^*_{i,y_1y_2y_3y_4} = \Pr\left(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4 \middle| X_1, X_2, \dots, X_p\right),$$
where $y_1, y_2, y_3, y_4 \in \{0, 1\}$ are binary observed responses

- Generalized linear model relates the mean of binary responses to the predictors via a link function $h(\cdot)$ to perform regression as
$$h\left(\pi^*_{i,y_1y_2y_3y_4}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Methods – AI Probability Models

- Based on the theory of additive logistic regression
- AI model relates the mean of binary responses to the predictors via logistic regression as
  $h\left(\pi^*_{i,y_1y_2y_3y_4}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + f_1\left(X_1, \dots, X_p\right) + \cdots + f_m\left(X_1, \dots, X_p\right)$
  where the function
  $$f_j\left(X_1, \dots, X_p\right) = g\left(\gamma_{j0} + \gamma_{j1} X_1 + \cdots + \gamma_{jp} X_p\right)$$
  for all $j = 1, \dots, m$, and a nonlinear activation function $g(\cdot)$
- The model uses TensorFlow for production reliability

Methods – Regularization

- Lasso regularization avoids overfitting and improves model stability
- The difference between the conditional log-likelihood, $\ell(\boldsymbol{\theta})$, and the lasso penalty, $\lambda\, ||\boldsymbol{\theta}||_1$, provides penalized conditional log-likelihood

$$\ell_\lambda(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) \; - \lambda\, ||\boldsymbol{\theta}||_1$$

- The objective function $\ell_\lambda(\boldsymbol{\theta})$ is maximized during training for a given hyperparameter λ (controlling the shrinkage on parameter vector $\boldsymbol{\theta}$)
- This hyperparameter is typically tuned via cross-validation

Case Study

- Using the 2022 US Census of Agriculture data from Michigan
  - Used a subset of predictors – farm size and type, demographics
  - Compared a linear-logistic triple-system estimator (TSE) and new AITSE in terms of total farms and land in farms

Bootstrap Simulation

- Used parametric bootstrap to simulate subsampled observations from Michigan TSE model (assuming historical-FSA list dependence)
- Fitted linear-logistic TSE and AITSE to bootstrapped data for testing (potential) nonlinear effects
- Calculated bias and variance for **total farms** and **land in farms**

Simulation Results

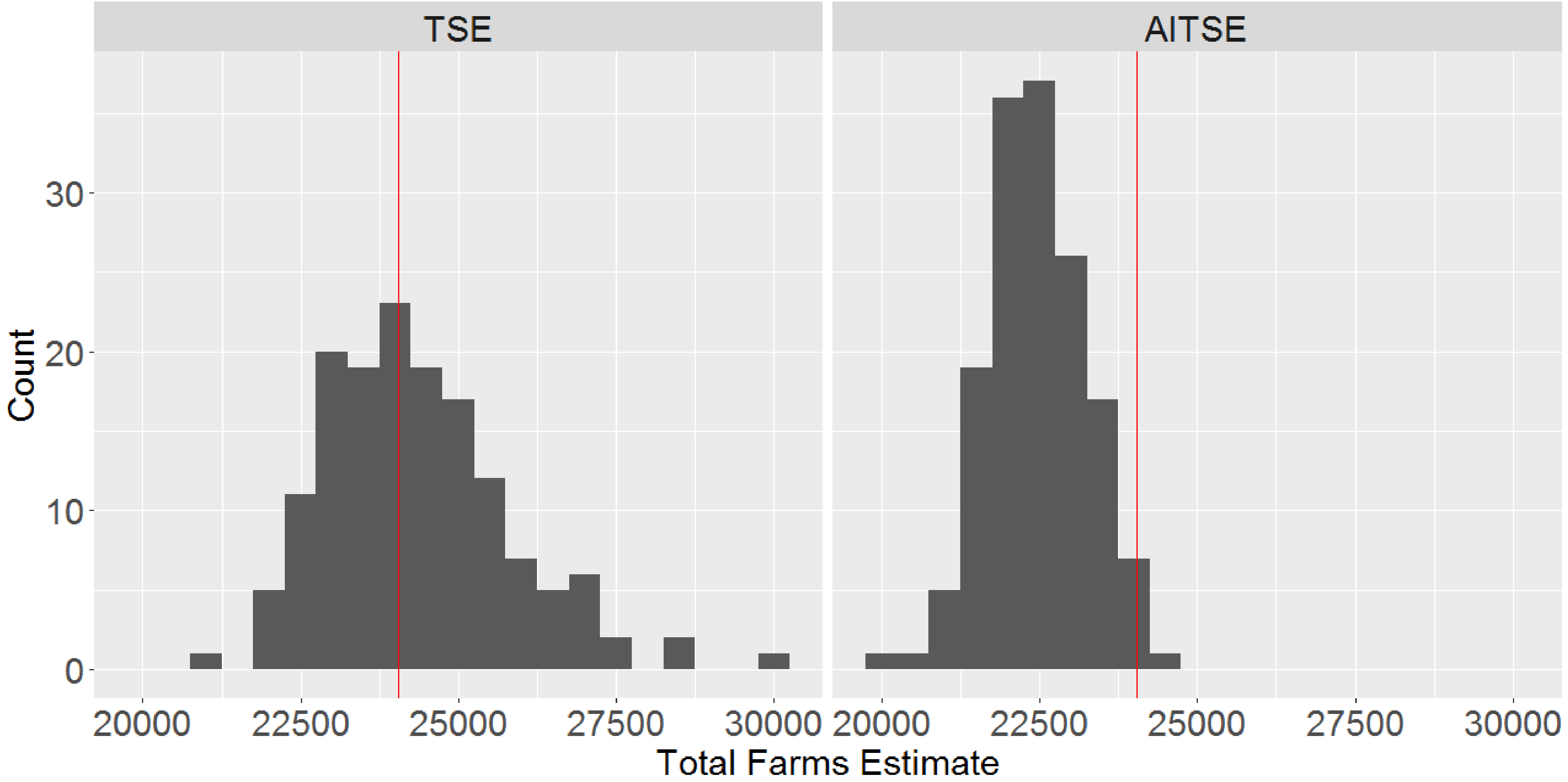
- Total farms simulation – true value is 24,048

	TSE	AITSE
Bias in total farms	1.3%	-6.7%
Simulation 2.5% quantile	22,173	20,952
Simulation 97.5% quantile	27,413	23,807
CV	6.0%	3.4%

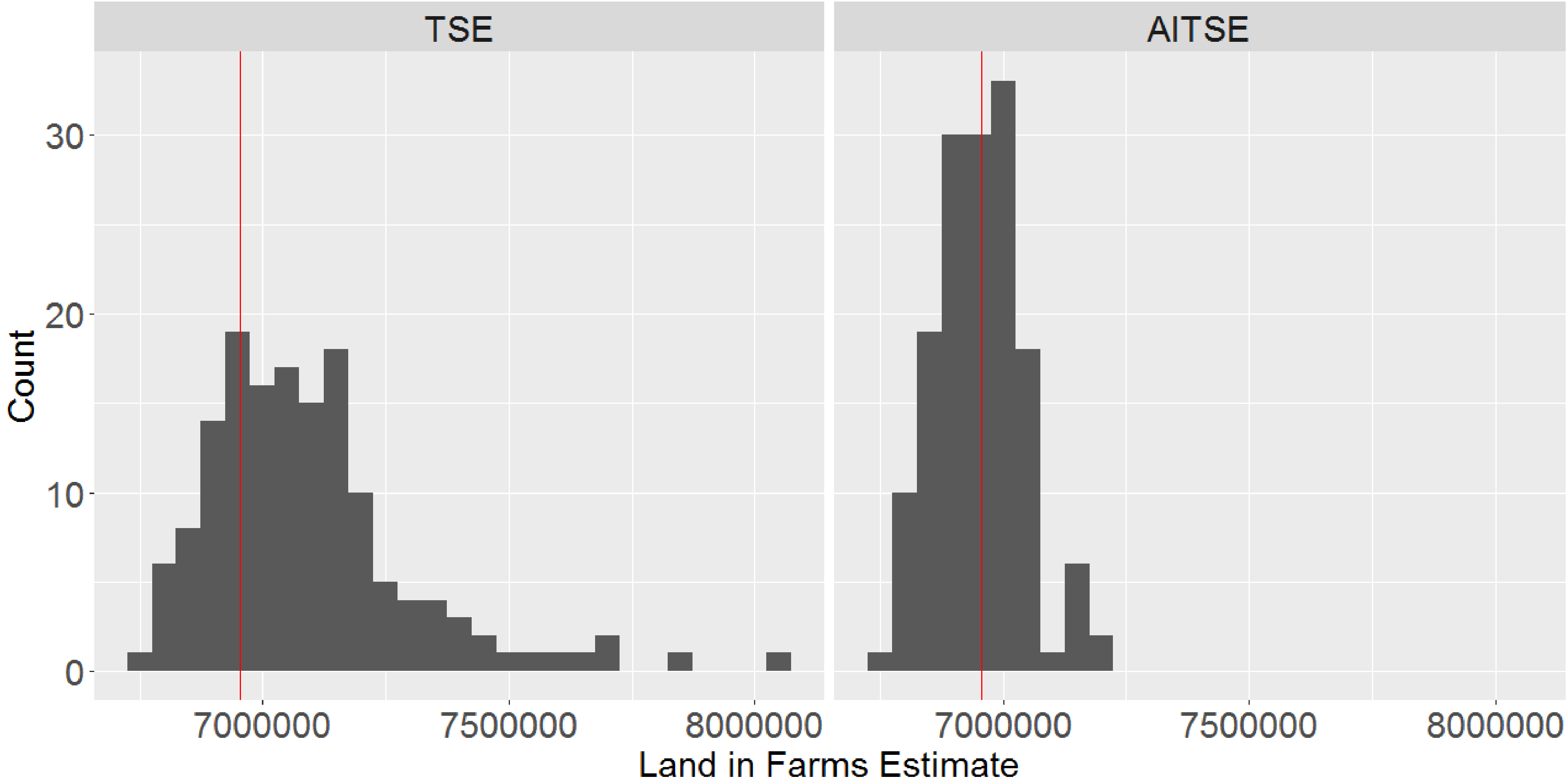
- Land in farms simulation – true value is 6,954,461

	TSE	AITSE
Bias in total farms	1.9%	-0.1%
Simulation 2.5% quantile	6,801,371	6,789,885
Simulation 97.5% quantile	7,676,946	7,156,600
CV	3.0%	1.3%

Visualizing Results for Total Farms



Visualizing Results for Land in Farms



Conclusions

- It is not clear if AITSE outperforms TSE in terms of bias
- AITSE outperforms TSE in terms variance
- Referred records can be counted as a separate list if **record source** and **record linkage** data are retained
- Future research will assess calibration adjustments

Questions?

Robbie.Emmet@usda.gov

Luca.Sartore@usda.gov

Thank you!