The 2017 Census of Agriculture: Challenges to Be Met

Linda J. Young
Chief Mathematical Statistician
Director, Research and Development Division
USDA National Agricultural Statistics Service

Census of Agriculture

- Conducted since 1840
- Accounts for all U.S. farms and ranches and the people who operate them
- Allows assessment of changes in American agriculture
- Provides foundation for new programs and policies

What is a Farm?

- A farm is any place from which $1,000 or more of agricultural products were produced and sold or normally would have been sold during the year.
  - Examples, some special cases:
    - Christmas trees
    - "government payment" farms
    - "pasture only" farms (at least 100 acres)
    - nurseries and greenhouses
    - exotic livestock
    - large garden

- Point Farm: a farm that does not produce or sell $1,000 or more of agricultural products, but has the potential for that much production.

NASS List Frame

- List of producers and agribusinesses
  - Names, addresses, and telephone numbers
  - Grouped by size and type of unit
- Used as the sampling frame for numerous surveys
- Kept as complete as possible

Census of Agriculture

- Uses a list frame, the Census Mailing List (CML)
- Conducted every 5 years (ending 2 & 7)
- Count of all U.S. agricultural operations ($1,000 or more in sales or potential sales)
- Primarily mailout/mailback data collection

NASS's Area Frame – June Area Survey (JAS)

Frame: All land in U.S. provides a complete frame assuming accurate screening

Sample Unit: A segment, which is typically a 1 square mile area of ~640 acres
Segments divided into tracts, representing unique operations

Design: Stratified Random
Sample of segments, strata based on percent cultivated (>50%, 15%-50%, < 15%).
20% of the sample enters each year and remains for 5 years

USDA

AGRICULTURE NASS COUNTS

June Area Survey (JAS)

- In-person interviewers screen for whether each tract is agricultural or non-agricultural
- Crop and livestock information is collected *only* on the agricultural tracts
- JAS has two primary uses in NASS statistics
  - Provides good direct estimates of large crop commodities
  - Used in multiframe estimates for commodities not fully captured on the list, such as small and medium cattle operations
- Estimating the number of farms based on the JAS is challenging

Agricultural Coverage Evaluation Survey (ACES)

- Supplement to the June Area Survey during Census years, with extended data collection through the end of June
- Allocated to improve the precision of farm demographic estimates

	Number of Segments
June Segments	11,085
ACES	3,291
Total	14,376

USDA

AGRICULTURE NASS COUNTS

Matching Process and Not-on-Mail-List (NML)

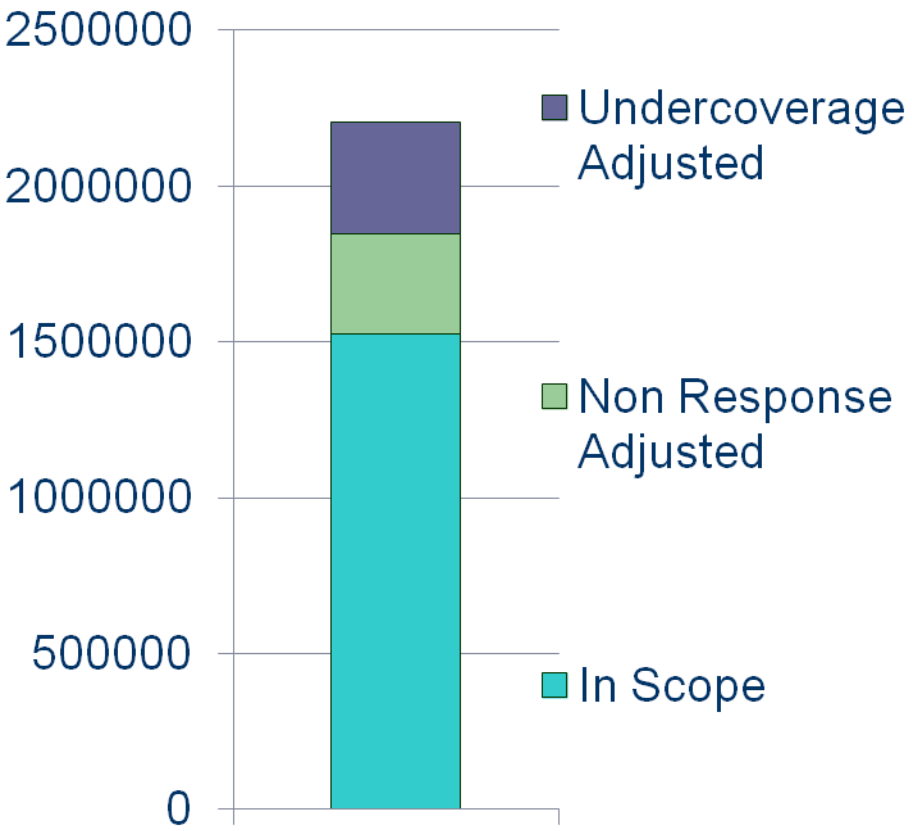
- During screening, JAS tracts are classified as agricultural or non-agricultural with potential, unknown potential, and no potential
- All tracts in the JAS are matched to the Census Mailing List (CML)
- Added non-agricultural tracts with no potential to the 2012 process
- NML records—JAS records that do not match a CML record
- NML records are mailed a census questionnaire

Accounting for Nonresponse and Undercoverage in 2007

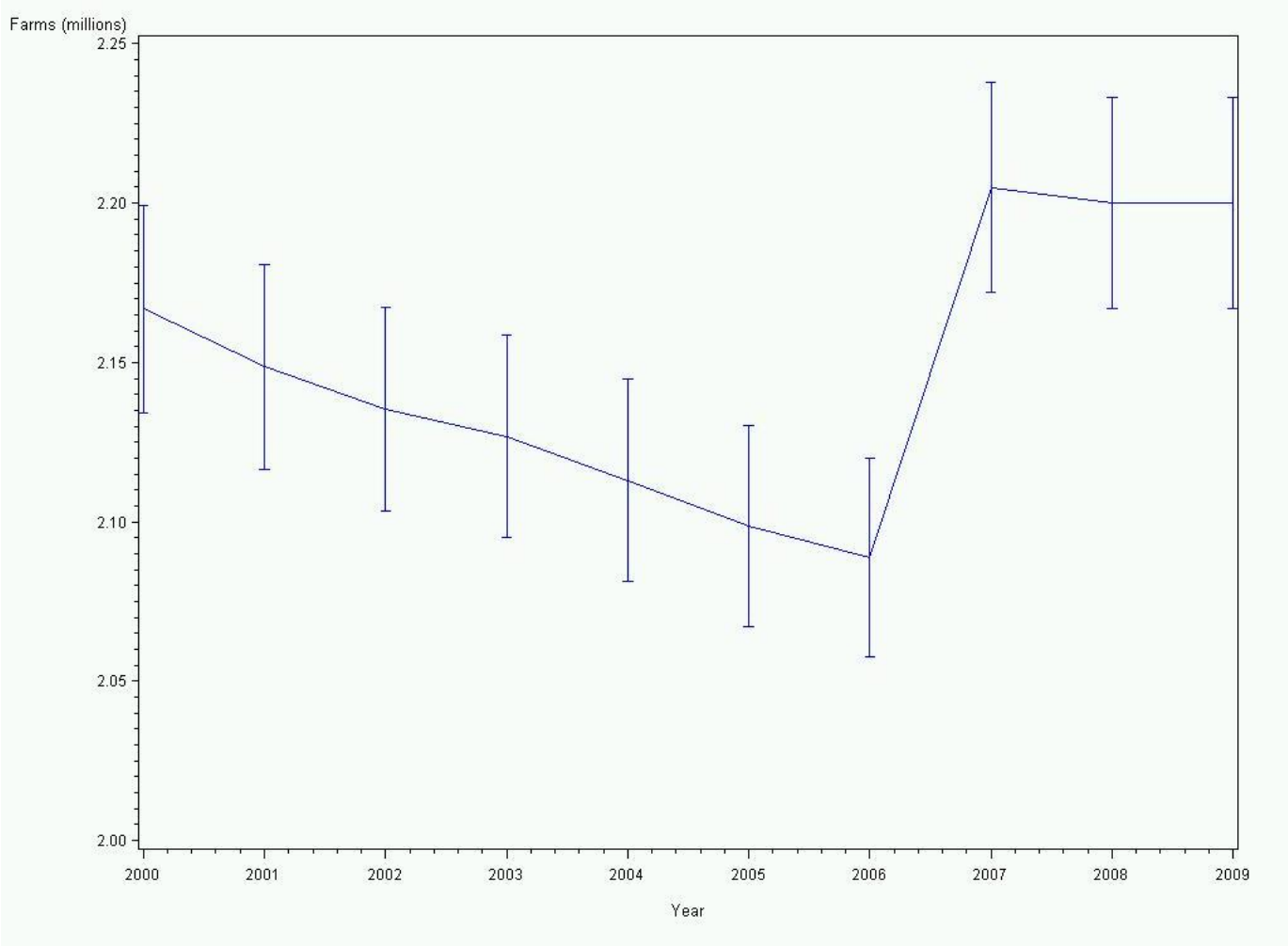
- Under-coverage
  - Census responses from NML records
  - JAS weights provide national estimates
  - Assumptions: Independence of CML and NML, Correct classification by the JAS
- Non-response
  - Probabilities of response estimated using CART
  - Reciprocals provide non-response weights for CML records with responses

2007 Census Results

- Number of records on CML and respond as a farm
  - 1,523,826
- Estimated farms totals after nonresponse adjustment
  - 1,846,814
- Estimated farm totals after nonresponse and undercoverage adjustment
  - 2,204,972
- Total effect of adjustment:
  - 681,146 – 45% Increase
  - From nonresponse: 322,988
  - From undercoverage: 358,158



Numbers of U.S. Farms



USDA

AGRICULTURE COUNTS NASS

Why Change?

- Under-coverage
  - Assumption that there was no misclassification when tracts were pre-screened was found invalid during the Farm Numbers Research Project (FNRP) in 2009.
- Non-response weights
  - Probabilities of response modeled using CART
  - CART designed for classification—provides biased estimates of probabilities
  - Assumption that the probability of a farm operation responding is equal to the probability of a non-farm responding is suspect
- Standard errors
  - Based on random effects model—statistical flaws could not be adequately addressed
  - In 2007, the total adjustment for non-response and under-coverage was 681,146, representing about 30% of the published number of 2.2 million farms.
  - The reported standard error was 4,775.

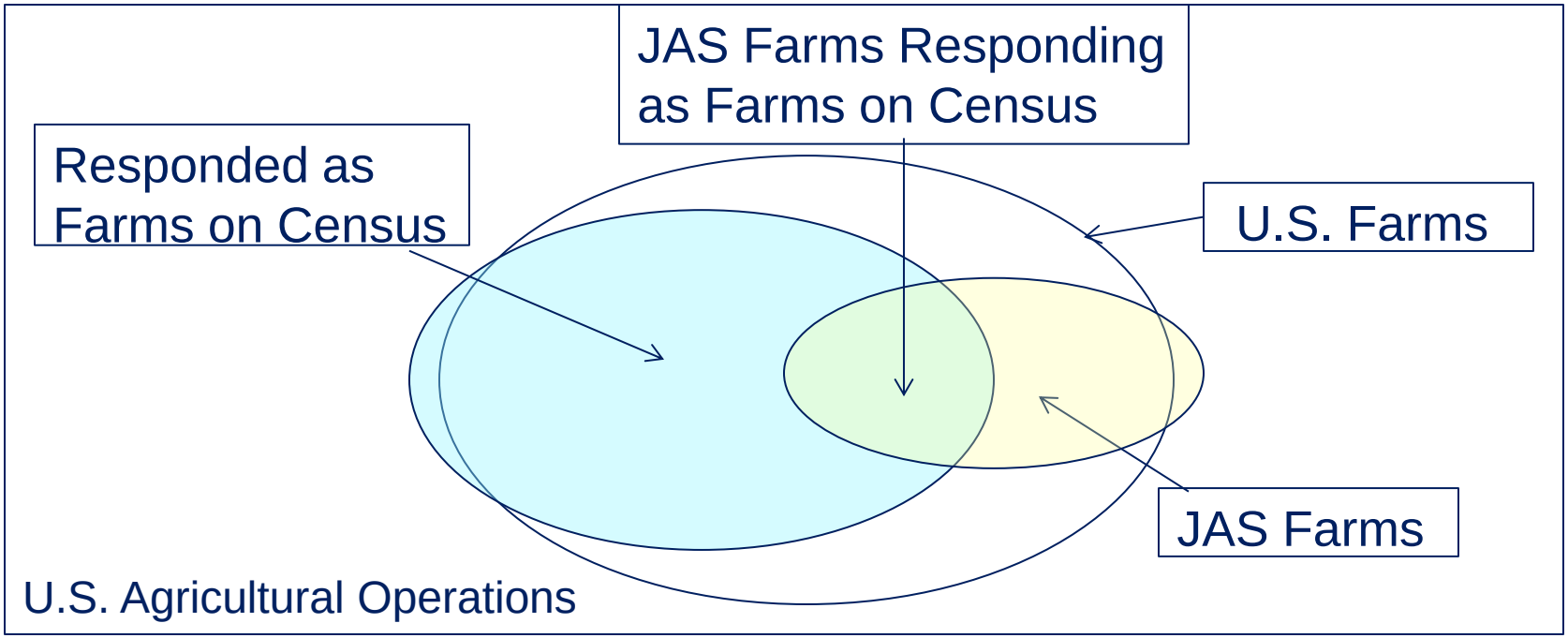
Goal

A unified framework

- for adjusting Census-based estimates for non-response, under-coverage, and misclassification
- for providing appropriate measures of uncertainty for estimates at all geographical levels

The Primary Assumptions

- The Census and June Area Survey are independent.
- The proportion of JAS farms captured by the Census is equal to the proportion of U.S. farms captured by the Census.



Capture-Recapture: The Big Idea

Suppose each U.S. farm, which includes all true JAS farms, has the same probability of responding as a farm on the Census. For the moment, assume that probability is 0.5.

Then, through the Census, we have "captured" half of all US farms.

Capture-Recapture: The Big Idea

Estimate of the number of U.S. farms: double the number of farms responding to the Census, which is equivalent to dividing by 0.5, the probability that a farm responds to the Census.

Note: It does not matter why a farm is not recorded as a farm on the Census. It could be that it did not. respond when mailed a Census form. It could be it did not receive a form. All that matters is whether or not it was identified as a Census farm.

What Does It Take to Capture a Farm?

- For a true farm to be captured by the Census, it must
  - Be on the CML
  - Send in a Census response
  - Be classified as a farm on the Census

Probability to be estimated

$$\pi(\text{CML}, \text{Responded}, \text{Census Farm} \mid \text{Farm})$$
$$= \pi(\text{CML} \mid \text{Farm})\pi(\text{Responded} \mid \text{CML}, \text{Farm})$$
$$\pi(\text{Census Farm} \mid \text{CML}, \text{Responded}, \text{Farm})$$

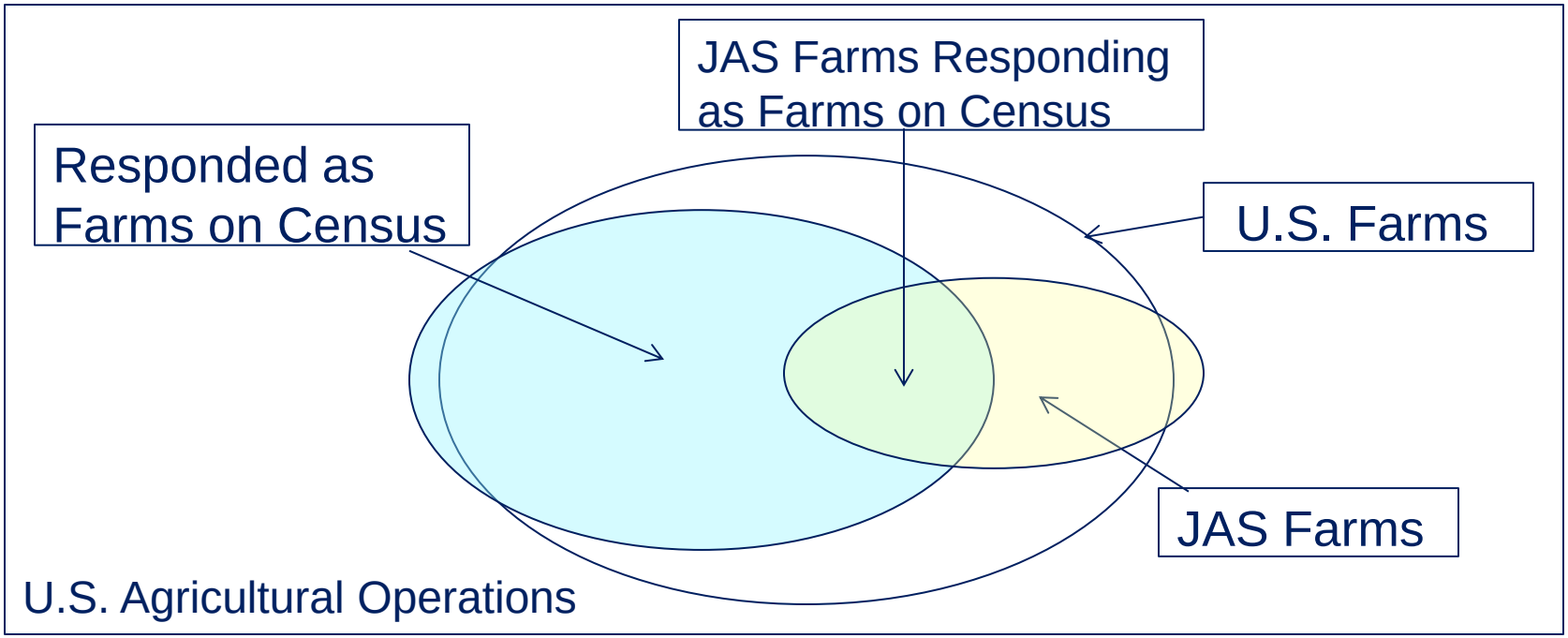
Reality Check: Probability for Capture is NOT the Same for All Farms

Solution: Use available information on demographics of farmers and farm-level information to estimate the probability that a farm with a given set of characteristics is captured by the Census.

If only categorical variables are used, then groups are formed. If continuous variables are also used, then each farm could have a separate probability.

Another Challenge: Misclassification

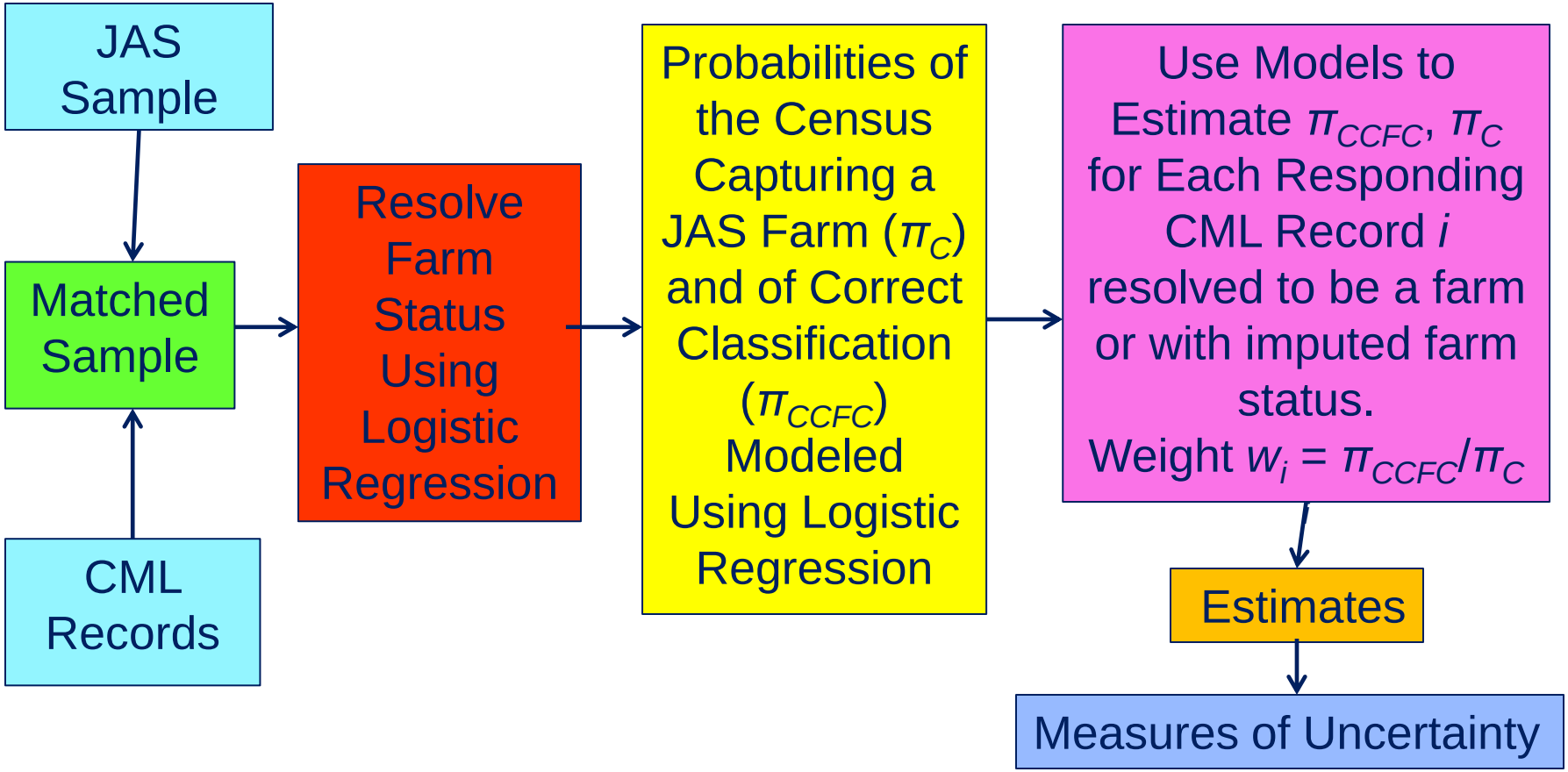
Some misclassification occurs on both the JAS and Census (2009 Farms Number Research Project and Classification Error Survey (CES)).



Another Challenge: Misclassification

- Misclassification can occur on both the JAS and the Census.
- Misclassification in the JAS and Census affects the estimate of the probability of capture.
- Census misclassification also affects the number of reporting farms
- **Unresolved Farm Status:** Agricultural operations classified as farm (non-farm) on the Census and non-farm (farm) on the JAS
- The probability of an operation with unresolved farm status being a farm may be modeled using logistic regression.

Overview of Estimation Process



Results of Matching the Census and JAS

	CML Non-Farm	CML Farm	CML Non-Resp.	NML	NML Domain Non-Farm	Total
Jas Non-Farm	2,938	2,942	1,463	1,090	19,675	28,108
JAS Farm	2,354	34,415	7,835	2,013	1,439	48,056
Total	5,292	37,357	9,298	3,103	21,114	76,164

Numbers of records with unresolved farm status are highlighted.

Subset of records with unresolved farm status were sent to the Regional Field Offices for manual review in an effort to resolve conflicts. (CML non-respondents were not reviewed.)

Imputation to Resolve Farm Status

	Census Non-Farm	Census Farm	CML Non-respondent	Totals
JAS Non-Farm	22,613	4,032	1,463	28,108
JAS Farm	3,793	36,428	7,835	48,056
Totals	26,406	40,460	9,298	76,164

The highlighted records are used to model the probability of an operation being truly a farm based on the region, state and stratum, tract size, and U.S. Census demographics at the county level using logistic regression.

The probabilities are then used to adjust the JAS weights, which are used in the models for coverage, response, and correct classification. But, is this the proper way to account for the survey design?

Resolving Farm Status

- The estimated logistic function is applied to operations with unresolved farm status to obtain the estimated probability of the operation being a farm.
- The probabilities are then used to adjust the weights in the models for capture and misclassification.

Implementation of Capture-Recapture

Once farm status is resolved or imputed, the probability of capture is modeled:

$$\pi_C(\text{CML, Respond, Census Farm} \mid \text{Farm})$$

Then the probability of correct Census farm classification (CCFC) is modeled:

$$\pi_{CCFC} = P(\text{Farm} \mid \text{CML, Responded, Census Farm})$$

These estimated probabilities are used to estimate the probability of capture for each responding farm on the CML based on a Dual System Estimator (DSE):

$$DSE = \sum_{i=1}^{n} \left(\frac{\hat{\pi}_{CCFCi}(\text{Farm} \mid \text{Farm on Census, Responded, CML})}{\hat{\pi}_{Ci}(\text{CML, Responded, Farm on Census} \mid \text{Farm})} \right)$$

Logistic Regression

- The probabilities of capture and correct classification are modeled using logistic regression on the matched dataset with JAS weights.
- For the records with imputed probabilities of being a farm, the JAS weights are multiplied by the predicted probability of being a farm.
- Covariates must be observed in both the Census and the JAS
- Categorical variables must be grouped so that estimated probabilities will be stable.
- The 2007 Census results will be used to establish the groupings used for the 2012 Census.

Logistic Regression: Model Assessment

- 5-fold cross validation
  - The records used to develop each model will be split into 5 groups.
  - In turn, a group will be deleted and the logistic model fit on the remaining four groups
  - A logarithmic penalty function will be used to assess the fit of the model to the data in the omitted group.

- Avoids overstating the accuracy of the model due to fitting and evaluating the model on the same set of data.

Calibration

The calibration process is independent of model adjustments.

The estimates obtained from adjusting for undercoverage, nonresponse, and misclassification provide calibration targets with some allowance for uncertainty.

Administrative data available for commodities are also used as calibration targets again with some allowance for uncertainty.

Restrictions, such as having only a limited amount of land within a county, are incorporated into calibration.

Given these constraints, the calibration process minimizes the changes in the records' weights. So, for example, if the number of cattle needed to be increased to meet a target, the weights of the records with the largest number of cattle tended to be increased first. That could result in 2 producers with more than $10 million in value of sales in the same county, an obvious error.

Measures of Uncertainty

A group-jackknife approach was used to quantify uncertainty

The uncertainty was quantified after adjusting estimates for undercoverage, nonresponse, and misclassification.

The uncertainty was quantified again after calibration and integerization.

For most items, the uncertainty was greater after calibration and integerization. However, for some primary national targets, such as the number of farms, the uncertainty was substantially less after calibration and integerization.

The conservative approach of taking the maximum of the two quantities was adopted.

Reality Check

The analysis is conducted during a short timeframe. We had one week to develop all models initially. We were given one or two days for revising the models after data errors were corrected.

More than 3 million census forms were mailed out. Over 1.3 million records were classified as farms based on the census responses. All statistical methods must be implemented on a tight production schedule.

Often the ideal solution cannot be implemented in the allotted time. One must be ready to provide something good that can be completed on time.

The Rest of the Story …
Editing and Imputation

Some data were missing; reporting errors were also present. Consequence: Data were edited and missing values imputed. The imputation rate was particularly high for the demographic variables.

Imputation Approach:

1. Responses to past surveys
2. If operator had not changed, demographic information taken from the last census.
3. Measure of similarity of report form to that from potential donors calculated using Euclidean distance with each similarity characteristic scaled. The most similar record provided data for imputation.
4. Each imputation conducted independently so different donors could be used on different items

How can/should this be improved for 2017?

The Rest of the Story …
Editing and Imputation

Once records are complete through editing/imputation the data are treated as observed.

Consequences/Concerns:

When are the data so sparse so that the record should be treated as a nonresponse?

How much of the observed misclassification is a consequence of the editing process?

The standard errors do not consider the impact of editing/imputation and thus are biased downwards. How important is it to account for this source of variation? How can we incorporate that source, perhaps expanding the timeline for this 1 to 3 days.

The Rest of the Story … Misclassification

In resolving farm status, the potential misclassification of operations that were farms on both the census and JAS and those that were non-farms on both were assumed to be correctly classified. There are probably errors in both of these groups, but especially in the non-farm group.

For the non-farm group, most of these were screened to be non-farms on the JAS, and they were not on the CML. Thus, very little information is available for these records. How can we begin to quantify the amount of misclassification given such little information?

The Rest of the Story … Models

The JAS is a stratified random sample. Each operation has an associated sampling weight. In the logistic regression, the sampling weights were normalized and used as the weights in the logistic regression. Is there a better approach?

When using the five-fold cross validation for model selection, what is the best way to select the model, given that the best fitting model differs with group?

Model uncertainty is not accounted for in the standard errors. How can/should we account for this source of variation?

The Rest of the Story … Calibration

Remember: NASS reports results at the county level. The county-level values may receive the greatest scrutiny because people know about agriculture at the local level.

We want to have an integrated process for adjusting for undercoverage, nonresponse, and misclassification and for calibrating to commodity targets. Ideas?

Recall that we had two measures of uncertainty and decided to take the larger one to be conservative. What should be done in this circumstance?

Conclusion

The Census of Agriculture will be released in May based on the best methods we have at this time.

Numerous improvements can/should be made for 2017.

What should receive the top priorities?

How can we best balance the desire to provide the very best solution with the need to provide results on a tight timeframe?

Thank You!!