# Unresolved Matched Records in Capture-Recapture Methodology

Andrea C. Lamas
Denise A. Abreu
Shu Wang
Daniel Adrian
Linda J. Young

USDA
National Agricultural Statistics Service

2013 FCSM Research Conference

November 6, 2013

# Contents

- Overview of NASS
- Census of Agriculture
- June Area Survey
- Census Bureau
- Dual System Estimation
- Unresolved Records
- Conclusions

# National Agricultural Statistics Service (NASS)

- NASS conducts hundreds of surveys including the Census of Agriculture
- Prepares reports covering every facet of United States agriculture
  - For example:
    - Production and food supplies
    - Prices paid and received by farmers
    - Farm income and finances
    - Number of farms and land in farms

# Farms in the United States

- 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

# Census of Agriculture

- Conducted every 5 years (years ending in 2 and 7)
- Count of all US Agricultural operations ($1000 or more in sales)
- Collects information on agricultural operations' commodities and operator demographics
- Only source of uniform, comprehensive agricultural data for every county or county equivalent in the U.S.
- Starts with list-based frame
  - Census Mail List (CML)
- Primarily mail data collection

# Sources of Error

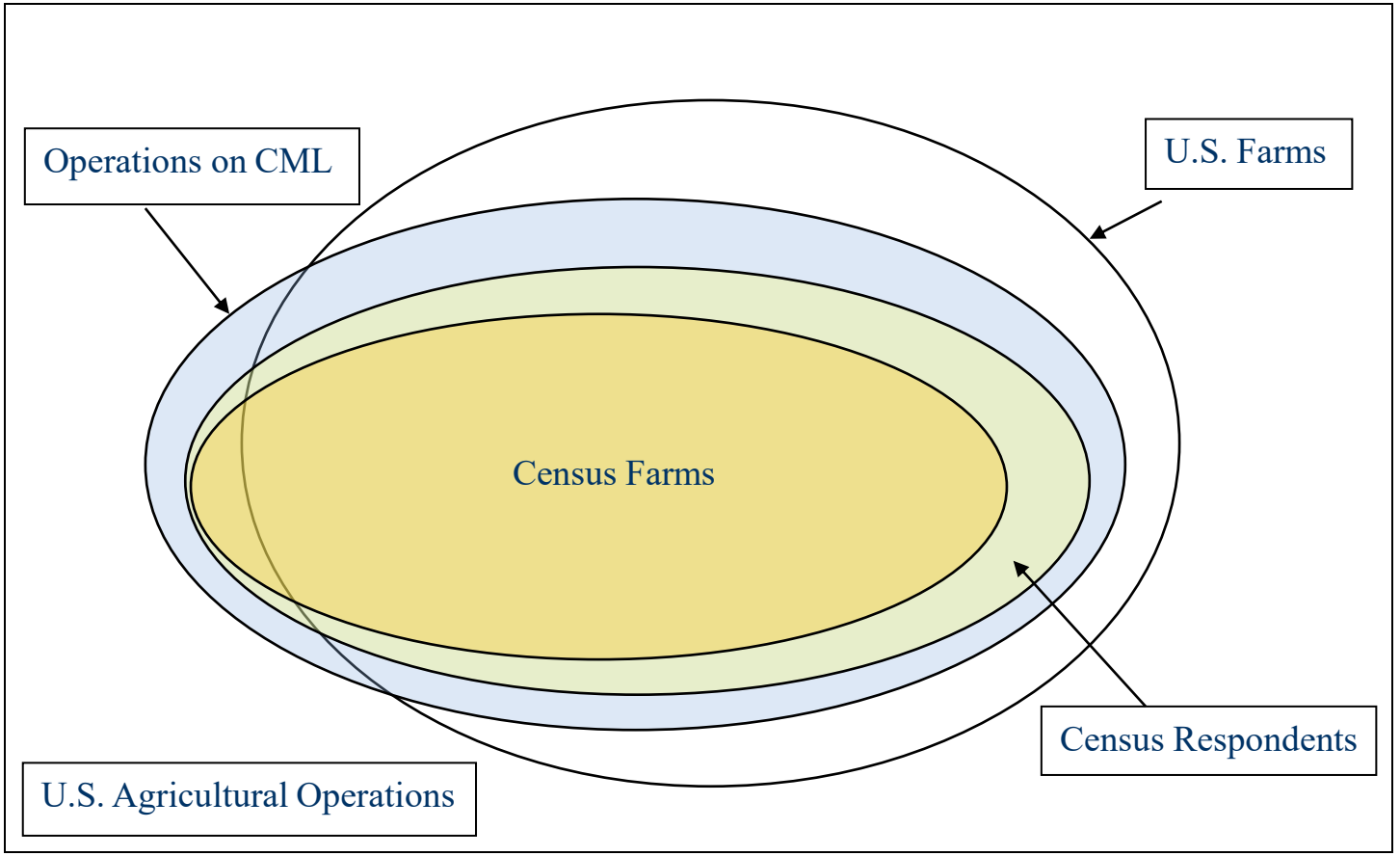
- Under-coverage
  - Incompleteness of the list
    - Not all agricultural operations appear on the Census Mail List
- Non-response
  - Not all agricultural operations on the Census Mail List respond

# Sources of Error (cont)

Errors in Census reporting

- Misclassification of Census non-farms
  - Non-farms classified as farms
    - Includes a subset of non-farms
- Misclassification of Census farms
  - Farms classified as non-farms
    - Omits a subset of farms

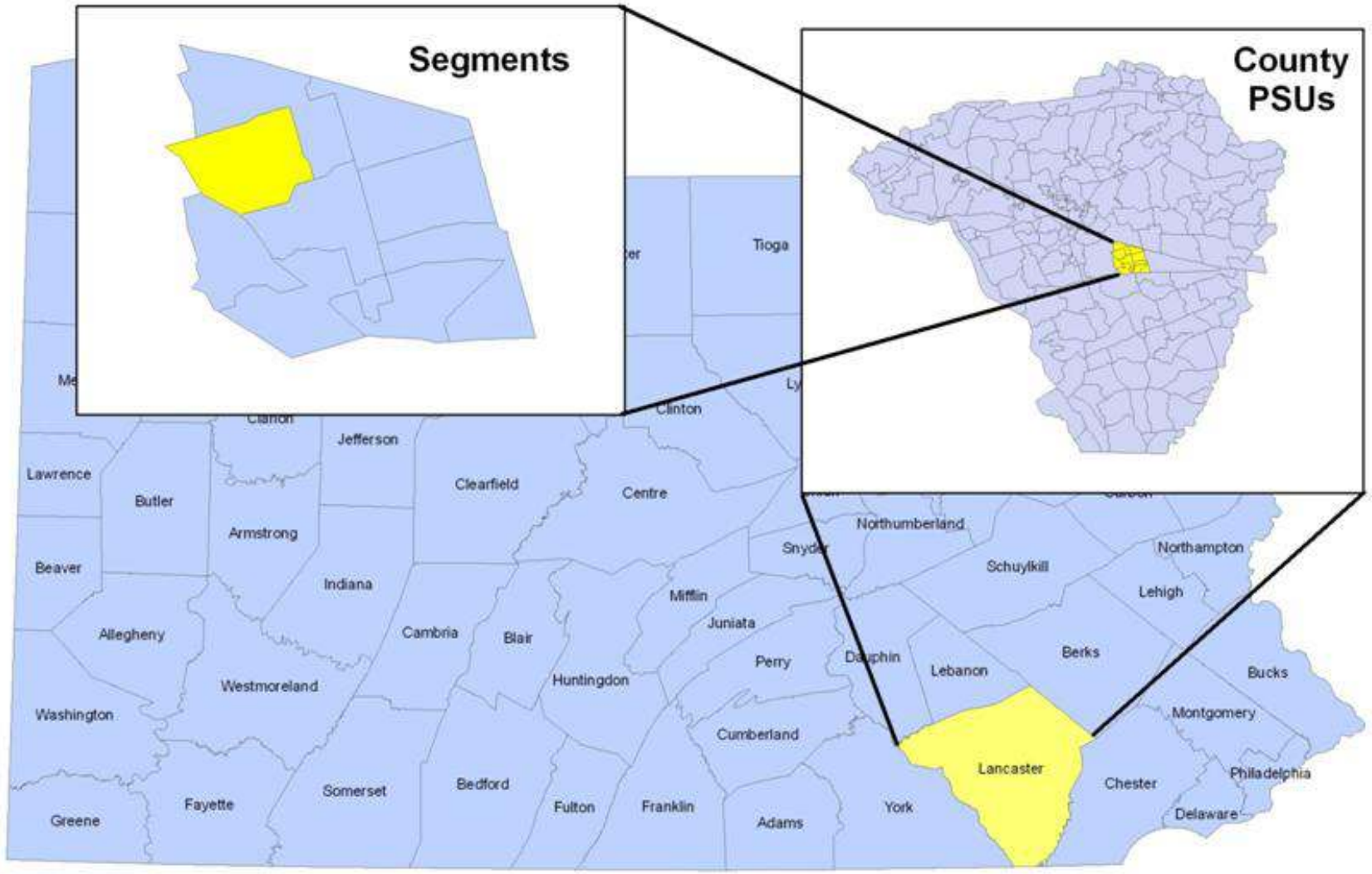
# Sources of Error



# June Area Survey (JAS)

- Area-based survey
- Conducted annually
- Theoretically complete sampling frame
  - No overlap or gaps
- Sample rotation scheme used every year
  - 20% of the sample replace every year
  - Sample rotations remain for 5 years
- Stratified sample based on land-use and percent of cultivation

# NASS Area Sampling Frame



# NASS Area Frame


- Segments of land sampled
- Sampled segments divided into tracts representing unique land operating arrangements
- In-person interviewers
- All tracts pre-screened and classified as:
  - Agricultural
  - Non-agricultural
    - With potential
    - With unknown potential
    - With no potential
- Crop and livestock information is collected *only* on the operations classified as agricultural

# Census Bureau Methods

- U.S. Census Bureau studied a capture-recapture adjustment method for 2000 Decennial U.S. Census
  - Used:
    - U.S. Decennial Census
    - Accuracy and Coverage Evaluation (A.C.E.) survey
- Methodology also used by Office of National Statistics (ONS) in the U.K.

# Capture Recapture Methodology

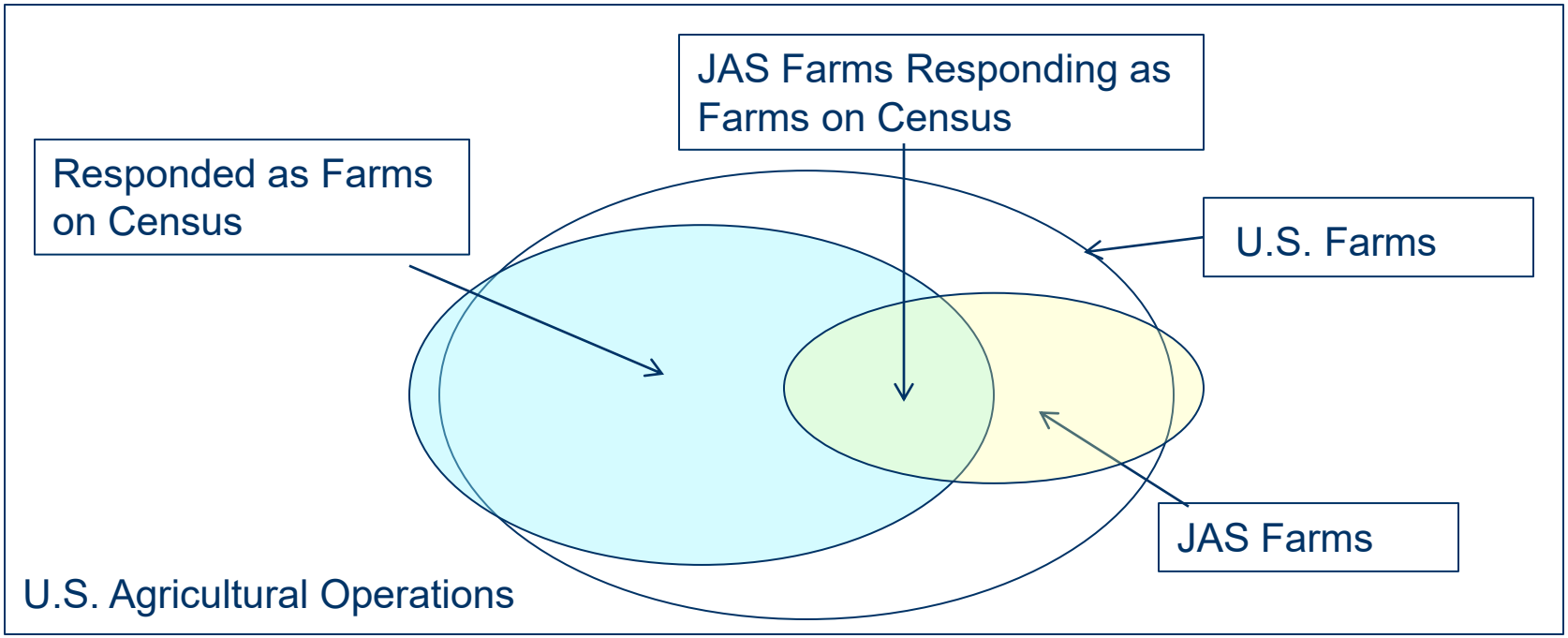
- Dual system estimation
- Requires two independent surveys

- NASS Application:
  - Goal: Want to get estimates for Census of Agriculture
    - Sample 1: Census of Agriculture records overlapping JAS tracts (not all census records)
    - Sample 2: June Area Survey tracts

# Primary Assumptions

- Independence of the Census and JAS
- Proportion of JAS farms captured by Census = Proportion of U.S. farms captured by Census
  - Adjusts for farms not captured by either the Census or JAS

# Census Matched to JAS



# What Data for DSE?

- Match Census records and JAS tracts
  - JAS tracts and CML records were brought together into link groups
    - Should represent the same operation
  - Link groups classified into 3 types
    - matches, possible matches and non-matches
  - Thorough staff review of possible matches
    - Matches / non-matches
- Records kept:
  - CML respondent/non-respondent matching JAS tract
  - Farms on JAS that are not on Census Mail List

# Unresolved Records

- In the matched dataset, farm status based on the Census and the JAS agree in most cases
  - Resolved farm status
- Some records are identified as farms (non-farms) on the JAS and non-farms (farms) on the Census
  - Unresolved farm status

# 2012 Census of Agriculture and JAS Records with Unresolved Farm Status

| | 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.)

# Resolutions to Unresolved Farm Status

| | Census Farm | Census Non-Farm | Totals |
|---|---|---|---|
| Both JAS and Census Correct | 240 | 235 | 475 |
| JAS Farm Status Correct | 235 | 1268 | 1503 |
| Census Farm Status Correct | 2075 | 378 | 2453 |
| Resolved By Correcting Matches | 131 | 233 | 364 |
| Cannot Resolve | 120 | 162 | 282 |
| Totals | 2801 | 2276 | 5077 |

# Unresolved Records

- Develop logistic model of the probability of an operation being a farm based on records with resolved farm status
  - Resolved farms are given a farm status value of 1
  - Resolved non-farms are given a farm status of 0
  - Normalized JAS weights are used in the model
- Unresolved operations are given an imputed farm status value
  - Estimated probability of being a farm obtained from missing data model

# Missing Data Model

- Farm-level data collected on either the Census or JAS can be used to develop the models
- Less information is available for the non-farms than for the farms
  - Limits the covariates available for model development

# Unresolved Records

- Probability that an unresolved record is a farm will be reflected in a reduction of the associated JAS weights

$$w_{Fi} = w_i \hat{p}_{Fi}$$

where $w_i$ = initial JAS weight

$\hat{p}_{Fi}$ = predicted probability that a record is a farm

$w_{Fi}$ = adjusted weight that will be used in the regression models

# Conclusions

- DSE adjusts for farms not captured by either the Census or JAS
- Unresolved records that cannot be resolved through staff review can be given an imputed farm status value
- All records will be used in DSE

# Thanks!

Andrea Lamas

Andrea.Lamas@nass.usda.gov