Incorporating Misclassification into Capture-Recapture Methodology in the 2012 Census of Agriculture

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

USDA

National Agricultural Statistics Service

Federal Committee on Statistical Methodology (FCSM)

November 6, 2013

Census of Agriculture: history

- Conducted in years ending in 2 and 7.
- Before 1997: U.S. Census Bureau
- 1997-present: USDA/NASS

Census of Agriculture: purpose

- Quantity of interest: number of farms
  - State
  - Farm type
  - Race
  - Gender
- Factor in allocation of funds for Federal agriculture programs
  - e.g. land grant universities

Farm definition

- Target population: operations that meet USDA farm definition.
- An agricultural operation that either
  - Produces at least $1,000 of sales in a year,
  - Normally produced $1,000 in sales,
  - OR 1,000 points of agricultural items

Examples of Point Farms

- 5 horses
- 1 acre of Christmas trees
- 100 acres of pasture land
- $1,000 in government payments

Documented farm misclassification

- Farm Numbers Research Project (2009): tracts in June Area Survey (JAS) were incorrectly identified as non-agricultural when agriculture was present.
- Classification Error Survey following 2007 Census: classification errors made during both Census and JAS.
- June Area Survey (JAS): used as supplemental survey to Census for farm number estimation.

Purpose of talk

- Accounting for farm misclassification in 2012 Census of Agriculture
- Adjusts traditional methodology
  - Dual System Estimation (DSE)
  - Capture-recapture

Dual system estimation or capture-recapture methodology

- Similar methodologies in different contexts (in this presentation, used interchangibly)
- DSE
  - U.S. Census of Population
  - UK Census Coverage Assessment
- Capture-recapture
  - fish and wildlife populations

Main ideas of DSE and capture-recapture

- DSE
  - supplementary (independent) survey to quantify Census undercount
  - i.e. What proportion of units from the supp. survey are counted by the Census?
- Capture-recapture
  - Capture, tag, and return
  - Recapture
  - What proportion of animals captured in second catch are tagged from first catch?

Example: catching trout (taken from UK Census documentation)

- Day One: catch 100 trout. Tag each and release
- Day Two: catch 50 trout. 25 have tags.
- Estimate of total

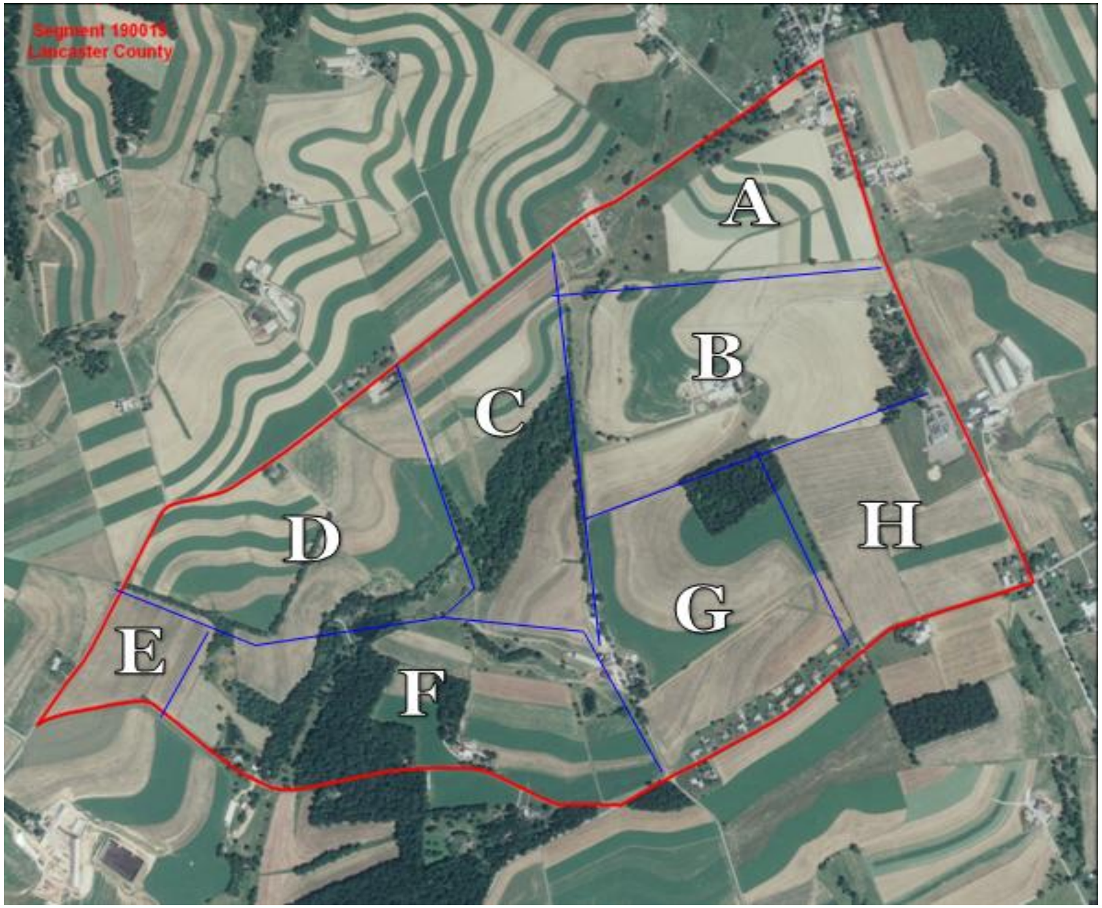
$$\hat{N} = 100\left(\frac{25}{50}\right)^{-1} = 200$$

Analogy between DSE and capture-recapture

Capture-recapture	DSE
Day one catch	Records counted by the Census
Day two catch	Records counted by supp. survey
Tagging process	Matching of Census and survey records

$$\hat{N} = N_{Census}\left(\frac{N_{Census\&Supp.S}}{N_{Supp.S}}\right)^{-1}$$

Supp. Survey: June Area Survey



DSE for Census of Agriculture

- Challenge: traditional DSE does not deal with misclassification errors
- Census of Agriculture has 4 types of enumeration errors, including
  - under-coverage
  - non-response
  - 2 types of misclassification.

Errors in Enumeration Process

- List under-coverage: the omission of farms from the Census Mailing List (CML)
- Non-response: The failure of farm operators to return a completed Census questionnaire.
- Both conditioned on farms

Errors in Enumeration Process II

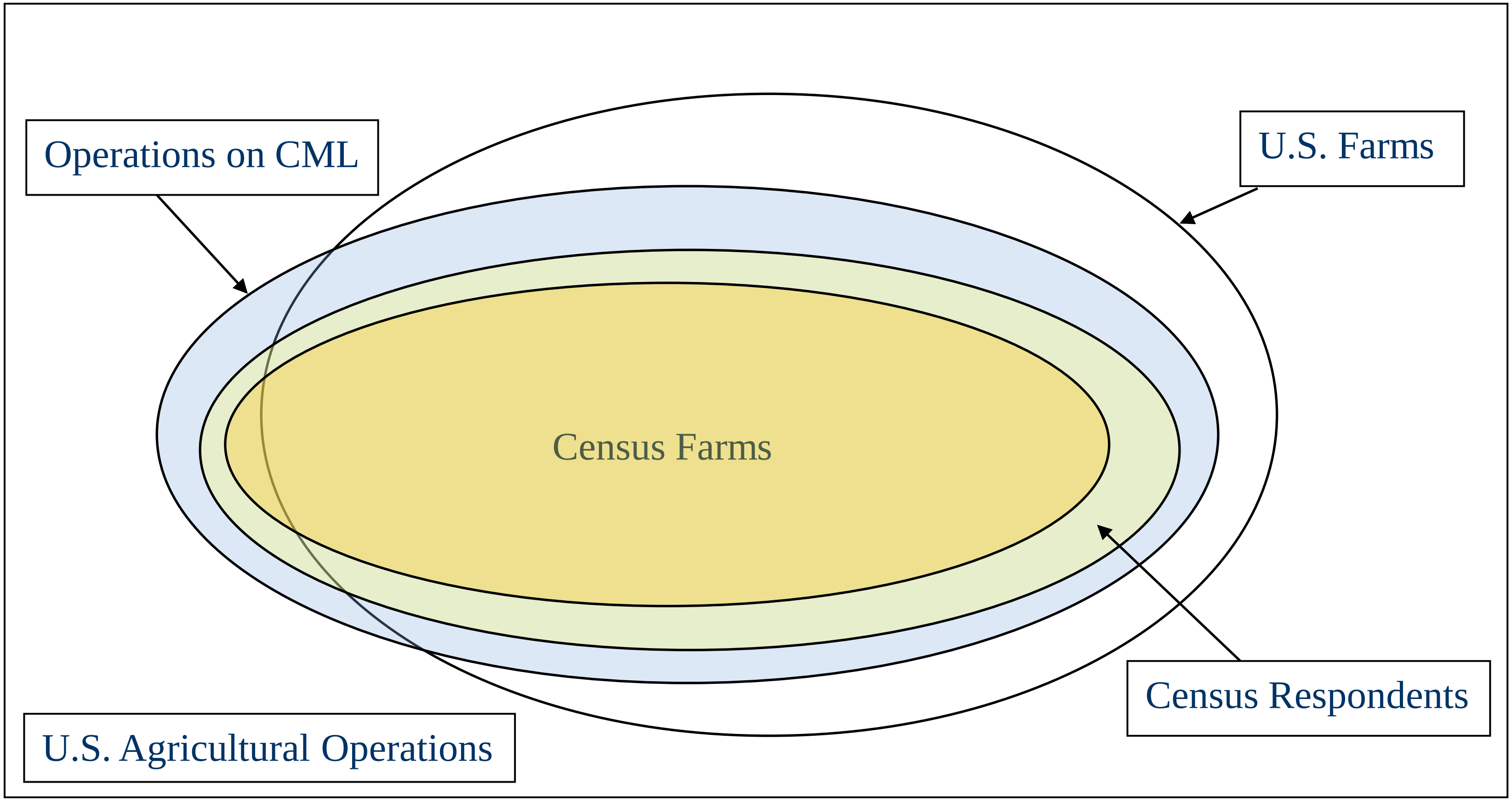
- Misclassification: Errors in Census reporting cause two types of misclassification errors:
  - Farms are classified as non-farms
  - Non-farms are classified as farms

Summary of Enumeration Errors

1. CML under-coverage of farms
2. Farm non-response
3. Farms are misclassified as Census non-farms
4. Non-farms are misclassified as Census farms

1-3: undercount, 4: overcount

Venn Diagram



Operations on CML

U.S. Farms

Census Farms

Census Respondents

U.S. Agricultural Operations

Adjustments to traditional DSE

- Probability of farm imputed for unresolved records: where Census and JAS disagree on farm status
- Three undercount errors are combined into “capture”
- Account for “differential catchability”
- Correct for misclassification overcount

Definition of Capture

- An operation is “captured” by the Census if it is
  - on CML | Farm
  - Responds | CML, Farm
  - Classified as Census Farm | CML, Responds, Farm

Product of probabilities

P(capture) = P(on CML | Farm)

x P(Responds | CML, Farm)

x P(Classified as Census Farm | CML, Responds, Farm)

Or

$$p_{Cap} = p_{Cov} p_{\mathrm{Re}sp} p_{CCF}$$

Dual System Estimation

- Dual System Estimator using capture:

$$\hat{N} = \left(\frac{F_{JC}}{F_J} \right)^{-1} F_C$$

  - F_{JC}: number of farms captured by JAS and Census
  - F_J: number of farms captured by JAS
  - F_C: number of farms captured by Census

Two problems with first-adapted DSE

- Doesn’t account for heterogeneity in capture probabilities
- Doesn’t account for misclassification of non-farms as Census farms (overcount)

Example: Heterogeneity in capture probabilities

- Catching trout, roach, and catfish.

	First day catch, #Tagged	Second day catch, Fraction tagged	Estimate of Total
Trout	100	25/50	200
Roach	50	5/20	200
Catfish	10	1/10	100
All fish	160	31/80	413

- Account for differential capture rates = 500
- Don’t account = 413

Heterogeneity in capture probabilities: estimator

- Partition farms into groups so that the probability of capture is about the same within each group.
- Example: by state, farm sales, farm type, race, gender
- Sum DSE's for each group

$$\hat{N} = \sum_{i=1}^{n\_groups} \left(\frac{F_{JC,i}}{F_{J,i}} \right)^{-1} F_{C,i}$$

Logistic regression

- Logistic regression extends this approach: allows
  - continuous variables
  - more complex models.
- Each Census record has its own capture probability.

$$\hat{N} = \sum_{j=1}^{n\_records} p_{Cap,j}^{-1}$$

Logistic regression

- The 0/1 capture indicators Y_i follow a Bernoulli(π_{Ci}) distribution, where

$$\pi_{Ci} = \mathrm{logit}^{-1}(x_{Ci}'\beta_C)$$

- β_C is estimated using the matched dataset (of JAS and Census)
- Then Census record probabilities of capture are

$$p_{Cj} = \mathrm{logit}^{-1}(x_{Cj}'\hat{\beta}_C)$$

Two problems with traditional DSE

- Doesn’t account for heterogeneity in capture probabilities
- Doesn’t account for misclassification of non-farms as Census farms (overcount)

Adjustment for Misclassification

- The probability of correct Census farm classification is

  p_{CCFC} = P(Farm | Census Farm)

- Multiplied by capture weights to correct for overcount.

$$\hat{N} = \sum_{j=1}^{n\_records} \frac{p_{CCFC,j}}{p_{Cap,j}}$$

Final estimator

- The final estimator is obtained after expanding the capture probability into its components.

$$\hat{N} = \sum_{j=1}^{n\_records} \frac{p_{CCFC,j}}{p_{Cov,j}\, p_{\mathrm{Re}sp,j}\, p_{CCF,j}}$$

Computing probabilities

- The 4 probabilities
  - p_{Cov} = P(on CML | Farm)
  - p_{Resp} = P(Responds | CML, Farm)
  - p_{CCF} = P(Classified as Census farm | CML, Responds, Farm)
  - p_{CCFC} = P(Farm | Census farm)
- use different subsets of the matched dataset depending on the conditions.

Wrap-up

- Census of Agriculture
- Adjusts traditional DSE/capture-recapture methods for misclassification.

Thank you!