Evaluating the Use of Web-Scraped List Frames to Assess Undercoverage in Surveys: Lessons from Local Foods Marketing

Linda J. Young
National Agricultural Statistics Service
March 7, 2018

2015 Local Foods Marketing Survey

- Fall 2015: NASS became aware of the need to conduct a local foods study with results to be published in 2016
- 2015 was chosen as the reference year
- Publication was slated for December, 2016

Key Definition for Local Foods

Operation: A farm having at least $1,000 in sales or potential sales, which in 2015 produced and sold food for humans to eat or drink directly to

- consumers
- retail markets
- institutions
- intermediary businesses marketing the food as being locally produced

Available History on Local Foods

- 2012 Census of Agriculture
  - 144,530 Local Foods Farms
  - $1,309,827,000 in Sales
- 2007 Census of Agriculture
  - 136,817 Local Foods Farms
  - $1,211,270,000 in Sales

Emerging Sectors of Agriculture

- Emerging sectors
  - Urban agriculture
  - Organics
  - Horticulture
  - Local Foods
- These tend to be
  - Smaller
  - More diverse
  - More transient
  - More dispersed

  than the more traditional farms in rural areas
- Hard to Quantify

Challenges with Identifying Small Farms

- NASS list frame
  - List of all known farms and potential farms
  - Known to be incomplete, especially for small farms
  - In 2012 Census of Agriculture, a 12.3% adjustment in the number of farms was due to undercoverage

Challenges with Identifying Small Farms

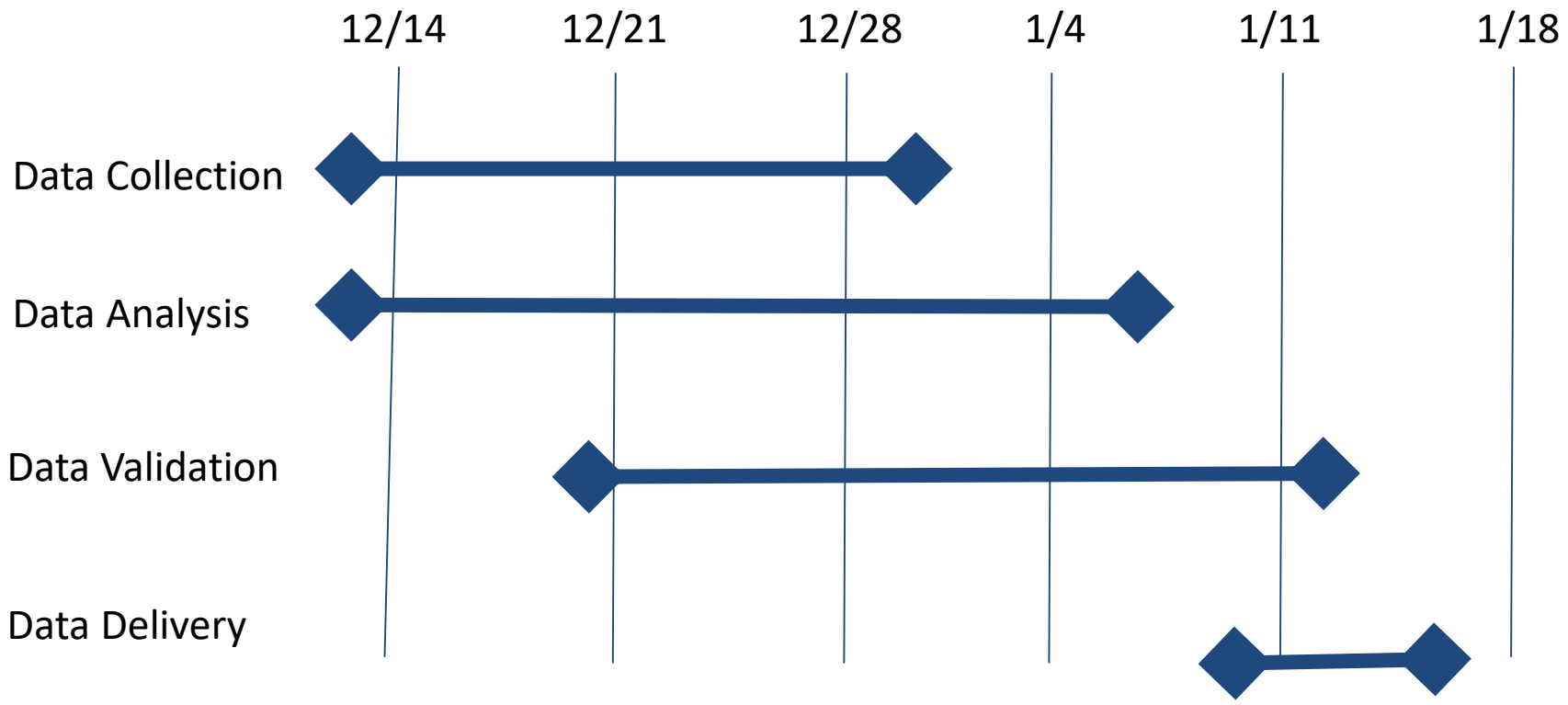
- Need to be able to assess undercoverage on NASS list frame
  - Sampling from NASS area frame not cost-effective when farms are dispersed
  - June Agricultural Survey (JAS) sample from NASS area frame—Insufficient number of small farms
  - Need a new approach

Big Idea: Create an independent list frame using web scraping

A Web-Scraped List Frame for the Local Foods Marketing Practices Survey

- Create a web-scraped list frame of all US local foods farms
  - Farm Name
  - Farm Type (Crops, Livestock, Poultry)
  - Farm Address
  - Farm State
  - Farm Latitude
  - Farm Longitude
  - POC Name
  - POC Address
  - POC State
  - POC Phone
  - POC E-mail

Timeline for Creating the Web-Scraped List



Consequence: Incomplete harvesting of potential open source data

Capture-Recapture: The Big Idea

How many bass are in your pond?

- Catch some bass (say 100)
- Tag each one and return to pond
- Next day catch some more (say 50, 25 are tagged)
- Half in second group have a tag so estimate half in pond have a tag

$$\frac{25}{50} = \frac{100}{N}$$

- Solve to find $N = 200$

List Frames Available for the Survey

- 2,007,110 on NASS List Frame
  - Includes all (not just local foods) operations
  - Consists both of confirmed farms and potential local foods farms
- 33,394 on Web-Scraped List Frame, which only has potential local foods
  - Are not confirmed to be farms
  - In urban ag pilot study about half had agricultural activity

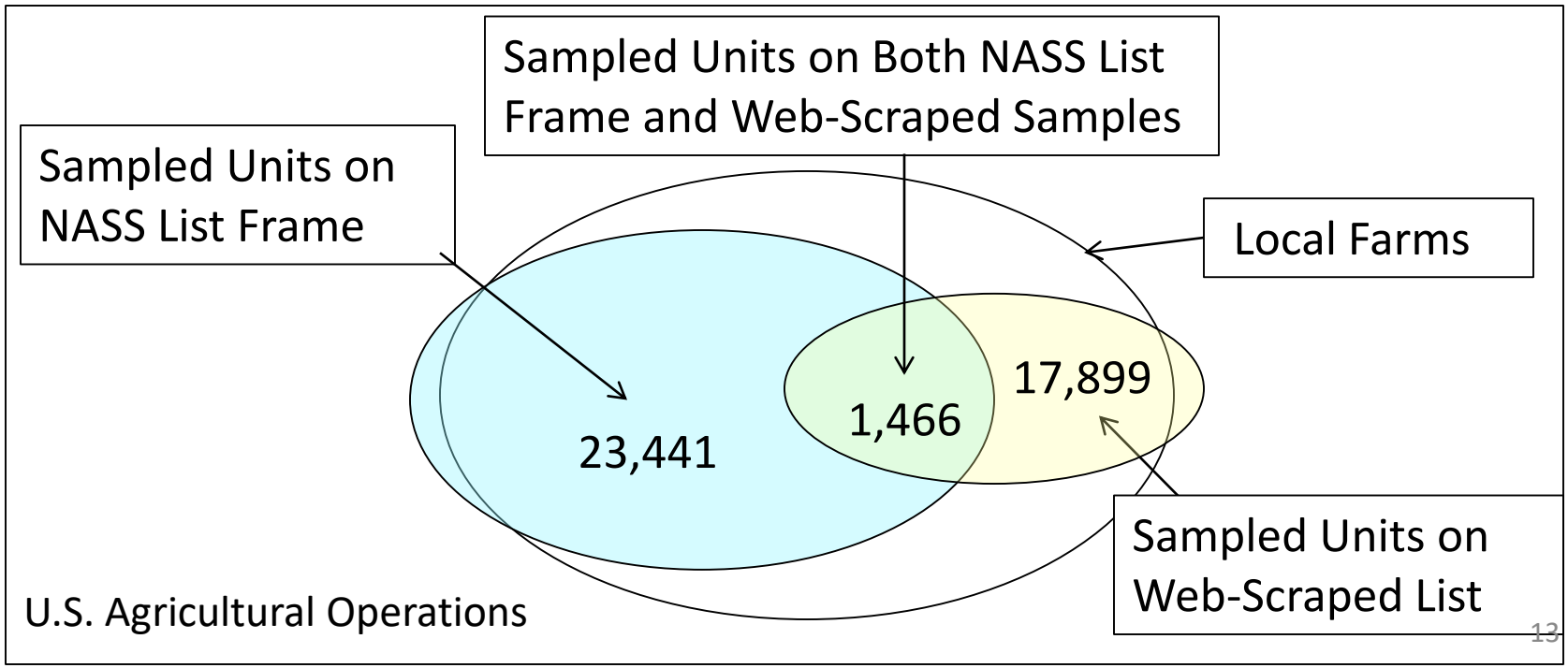
Local Foods Sampling Design

- NASS list – Stratified Sample Design (24,907)
  - Four groups
    - A: Census and Organic respondents + Value of Sales for food
    - B: Local Foods indicator – No Value of Sales
    - C: Potential local foods entities
    - D: All others – stratified by likelihood of local foods
  - Sample Allocation: Target CV's (Value of Sales)
    - US level 2.0 – 3.0
    - Regional 8.0 – 10.0
    - State Level 10.0 – 12.0
- Web-Scraped (WS) list – Systematic Sample (19,365)
  - Ordered by state and web-scraped farm type

1,466 records were in both NASS and Web-Scraped list samples

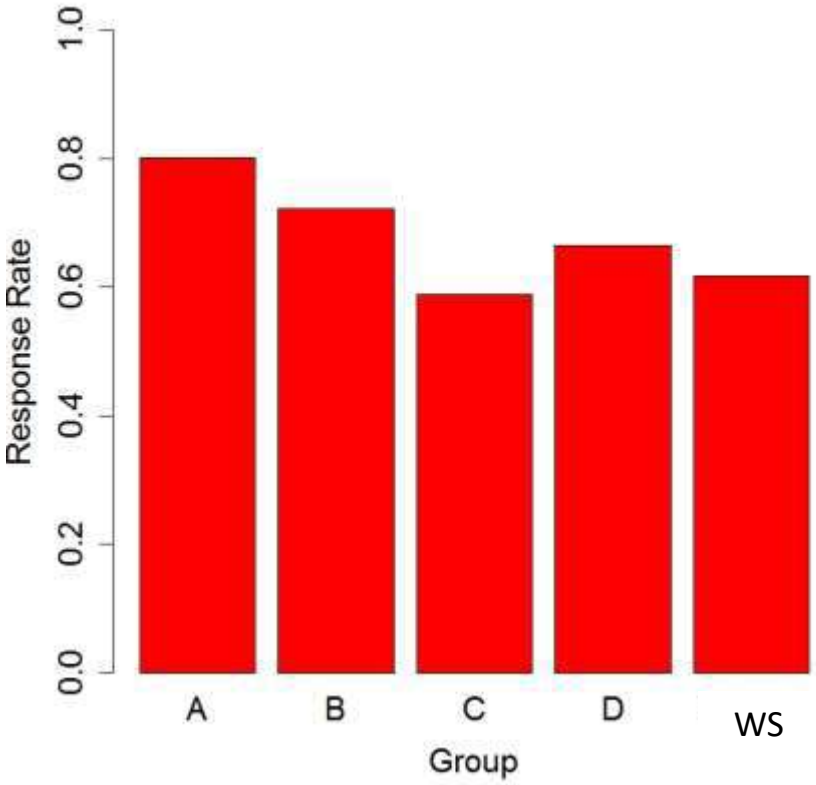
Primary Assumptions for Analysis

- Two Independent Samples:
  - NASS List Frame
  - Web-Scraped List Frame
- Proportion of web-scraped local foods farms captured in the NASS list frame sample is equal to the proportion of the US local foods farms captured by the NASS list frame sample

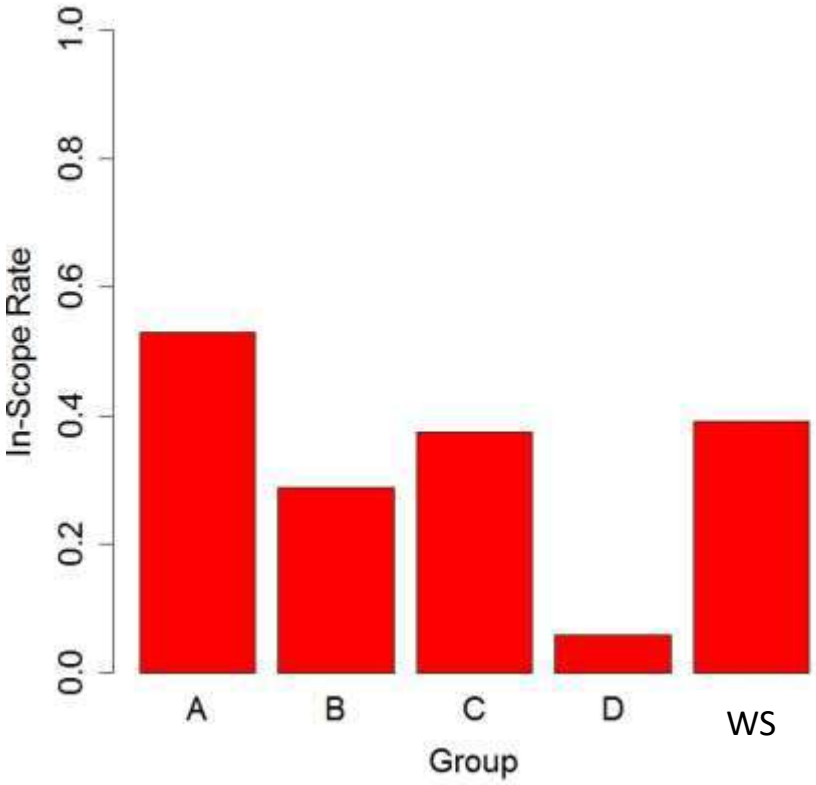


Signal of a Challenge Ahead

Response Rates



In-Scope Rates



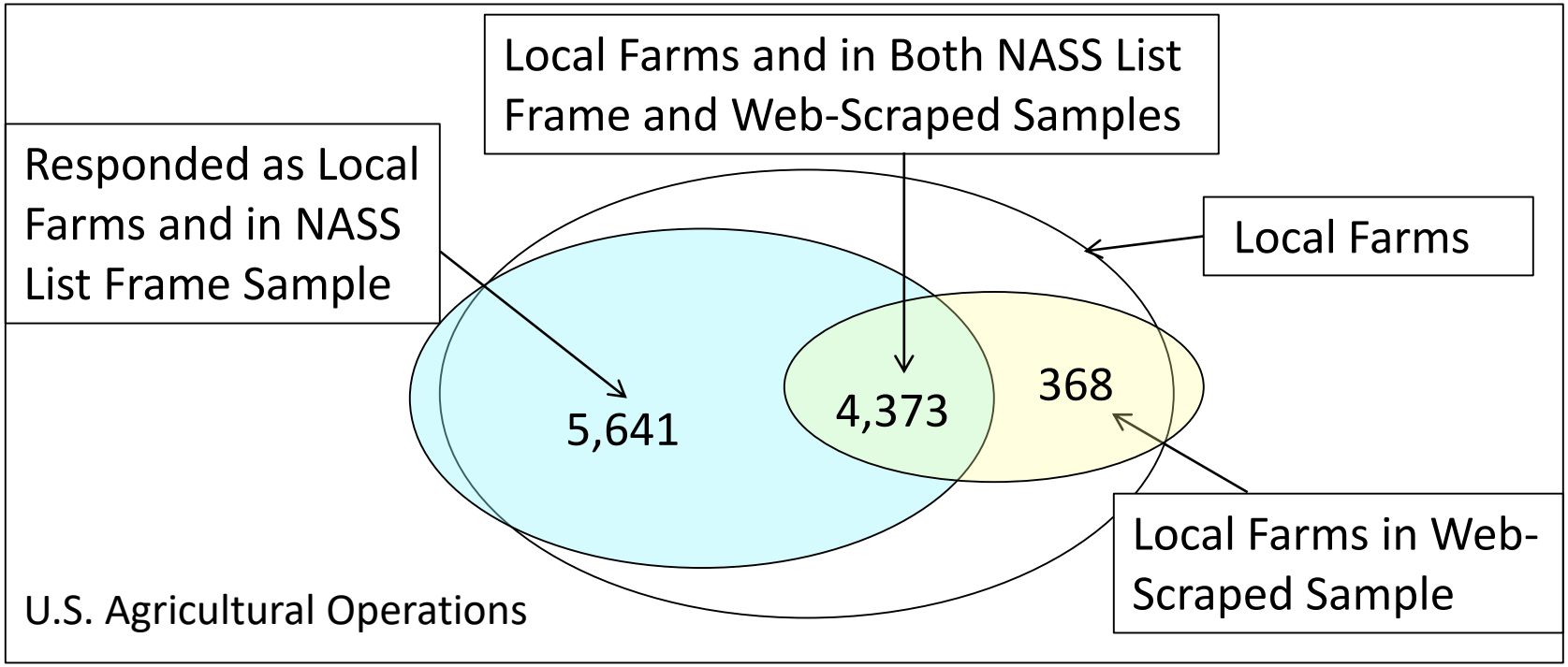
A: Census and Organic respondents with Value of Sales for food

B: Local Foods indicator – No Value of Sales

C: Potential local foods entities

D: All others – stratified by likelihood of local foods

Responding Local Foods Farms for Capture-Recapture

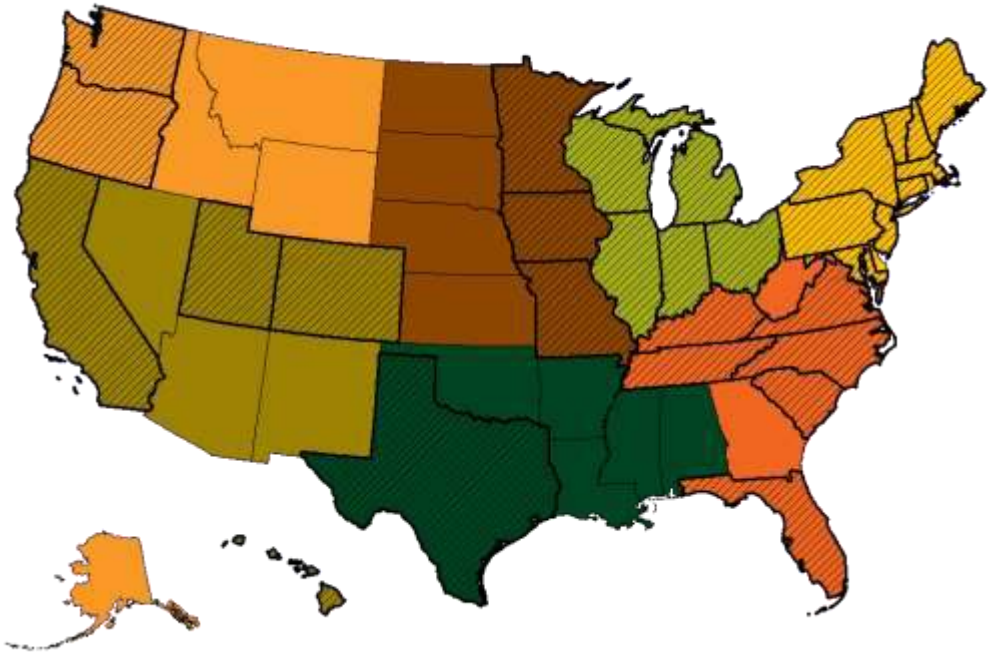


Operations Selling Directly: Count and Sales Through all Marketing Channels, 2015

- 167,009 ± 5845 operations used direct marketing practices to sell food in the US.
- $8,747 million ± $892 million of food was sold through direct marketing practices, including value-added products at the first point of sale.
  - $4.8 billion were direct food sales of raw commodities.
  - $3.9 billion were food sales of value-added commodities.

Local Food Marketing Practices Publication Levels

Levels of Publication: US, Regional, and 30 States



Count of Published items by level	
US	393
7 Regions	33
30 States	15

=States with published data.

A Closer Look at the Assumptions

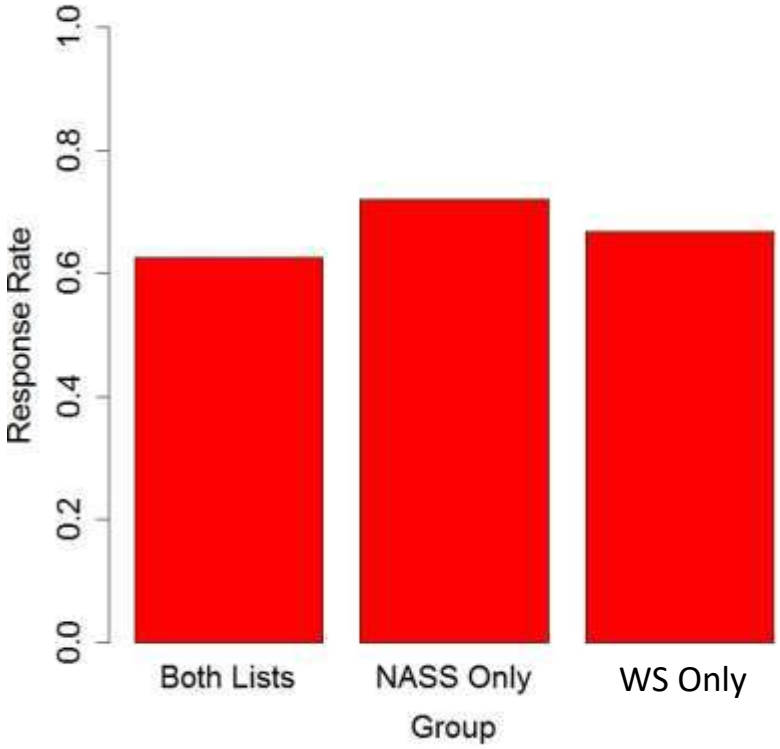
- The population is closed (no “births” or “deaths” during the time between the two samples)
  - Samples collected during the same timeframe
- The two lists are independent
  - Web sources used in developing the NASS list frame
  - Lack of independence introduces bias
- All farms are equally likely to be captured in each sample
  - Tried to control for this using logistic regression or by forming categories
  - Heterogeneity tends to cause downward bias

A Closer Look at the Assumptions

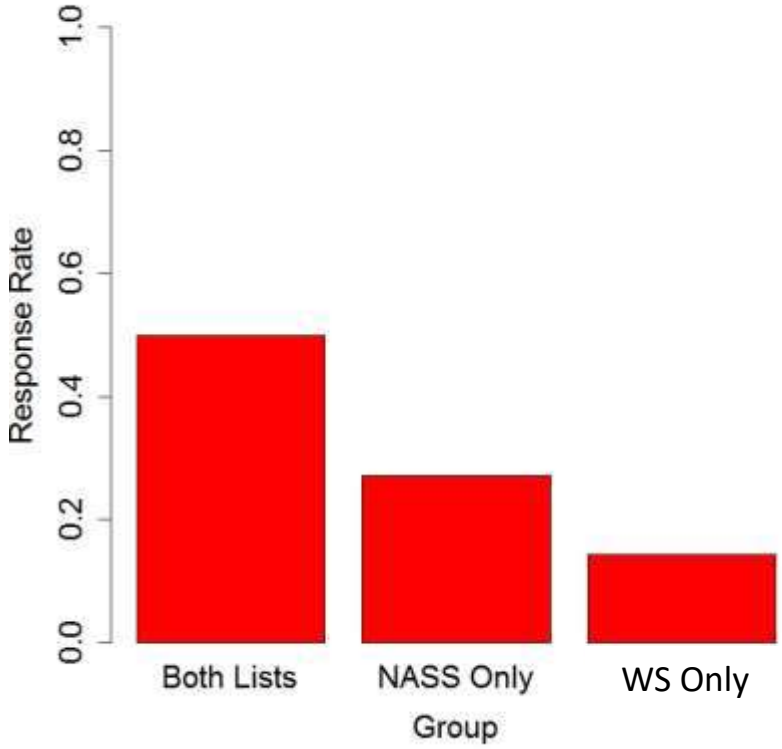
- Capturing a farm in one sample does not affect its catchability in the other sample
  - Operations in both samples only receive one questionnaire
- Farms caught in the first sample can be identified if they are caught in the second sample
  - Assumes perfect record matching

Discussion: List Comparison

Response Rates



In-Scope Rates



A: Census and Organic respondents with Value of Sales for food

B: Local Foods indicator – No Value of Sales

C: Potential local foods entities

D: All others – stratified by likelihood of local foods

Discussion

- Web scraping for list building
  - More thorough web scraping
  - Prescreening to determine farm status
  - Coverage
- Capture-recapture modeling
  - Same population for both lists?
  - Should sample design emphasize records not on NASS list frame?
  - Probability of capture

Local Foods Team Members

Mark Apodaca
Adam Cline
Jeff Bailey
James Barham—RD
Jeremy Beach
Jeff Beranek
Doug Boline
Kara Daniel
Saira Farooq
Ginger Harris
Andrew Brosier
Donald Buysse
Vincent Davis
Sarah Goodale
Pat Gregory

Brandon Hopkins
Mike Hyman
Troy Joshua
Doug Kilburg
Tom Kruchten
Megan Lipke
Christy Meyer
Peter Quan
Scott Shimmin
Dominique Sims
Holly Smith
Elanor Starmer – AMS
Danielle Tarpley
Nate Vandermeer
Stephen Vogel – ERS
Krissy Young

Selected References

Abreu, D.A., J.S. McCarthy, L.A. Colburn (2010). Impact of the Screening Procedures of the June Area Survey on the Number of Farms Estimates. Research and Development Division. RDD Research Report Number RDD-1003. Washington, DC: USDA, National Agricultural Statistics Service.

Alho, J.M. (1990) Logistic regression in capture-recapture models. *Biometrics,* 46,623-635.

Alho, J.M. (1994) Analysis of sample based capture-recapture experiments. *Journal of Official Statistics,* 10, 245-256.

Alho, J.M., M.H. Mulry, K. Wurdeman, and J. Kim (1993) Estimating heterogeneity in the probabilities of enumeration for dual-systems estimation. *Journal of the American Statistical Association* 88: 1130-1136.

Armstrup, Steven C., Trent L. McDonald, and Bryan F.J. Manly (eds) (2005) *Handbook of Capture- Recapture Analysis.* Princeton University Press: Princeton, NJ.

Braeye, Toon, Jan Verheagen, Annick Mignon, Wim Flipse, Denis Pierard, Kris Huygen, Carole Schirvel, and Niel Hens (2016) Capture-recapture estimators in epidemiology with applications to pertussis and pneumococcal invasive disease surveillance. *POS One*. http://dx.doi.org/10.1371/journal.pone.0159832.

Chao, Anne (2001) An overview of closed capture-recapture methods. *Journal of Agricultural, Biological, and Environmental Statistics,* 6, 158-175.

Gemmell, I., T. Millar, and G. Hay (2004) Capture-recapture estimates of problem drug use and the use of simulation based confidence intervals in a stratified analysis. *Journal of Epidemiology and Community Health* 58: 758-765. Doi: 10.1136/203.008755.

Grau, Eric, Frank Potter, Steve Williams, and Nuria Diaz-Tena. 2006. Nonresponse adjustment using logistic regression: to weight or not to weight. *Proceedings of the 2006 Joint Statistical Meetings, ASA Section on Survey Research Methods* 3073-3080.

Hickman, Matthew, Stuart Cox, Julie Harvey, Samantha, Howes, Michael Farrell, Martin Frischer, Gerry Stimson, Colin Taylor, and Kate Tilling (1999) Estimating the prevalence of problem drug use in inner London: a discussion of three capture-recapture studies. *Addiction* 11: 1653-1662.

Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association* 88: 1047-1066.

Hopper, Kim, Marybeth Shinn, Eugene Laska, Morris Meisner, and Joseph Wanderling (2008) Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *American Journal of Public Health* 98: 1438-1442. doi: 10.2105/AJPH.2005.083600.

Lamas, Andrea C., Denise A. Abreu, Pam Arroway, Kenneth K. Lopiano, and Linda J. Young (2010) Modeling misclassification in the June Area Survey. *Proceedings of the 2010 Joint Statistical Meetings, ASA Section on Survey Research Methods* 2480-2488.

Selected References

Laplace, P.S. (1786) Sur les naissances, les mariages et les morts. In *Histoire de l'académie royale des sciences*. Année. 1783, Paris.

LaRuche, G., D. Dejour-Salamanca, P. Bernillon, I. Leparc-Goffart, M. Ledrans, A. Armengaud, M. Debruyne, G.A. Denoyel, S. Bichler, L. Ninove, P. Despres, and M. Gastellu-Etchegorry (2013) Capture–recapture method for estimating annual incidence of imported fengue, France, 2007–2010. *Emerging Infectious Diseases* 19: 1740-1748. https://dx.doi.org/10.3201/eid1911.120624.

Lincoln, F.C. (1930) Calculating waterfowl abundance on the basis of banding returns. *Circular of the U.S. Department of Agriculture* No. 118: 1-4.

Mule, Thomas (2012) *2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States*. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01. Washington, DC; U.S. Census Bureau.

Office of National Statistics (2005) *Census 2001 review and evaluation: one number census evaluation report.* ONS, London.

Pollock, K.H., S.C. Turner, and C.A. Brown (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology* 20: 117-124.

Seber, G.A.F. (2002). *The Estimation of Animal Abundance and Related Parameters*, 2nd edition. The Blackburn Press: Caldwell, New Jersey.

Sekar, C. Chandra and W. Edward Deming (1949). On a method of estimating birth and death reates and the extent of registration. *Journal of the American Statistical Association* 44: 101-115.

Sudman, S., M.G. Sirken, and C.D. Cowan (1922) Sampling rare and elusive populations. *Science* 240: 991-996. doi: 10.1128/science.240.4855.991.

U.S. Census Bureau (2004) *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*. September, 2004. Online: http://www.Census.gov/prod/2004pubs/dssd03-dm.pdf.

U.S. Census Bureau (2008). *2010 Census Coverage Measurement Estimation Methodology*. October, 2008. Online: http://www.Census.gov/coverage_measurement/pdfs/2010-E-18.pdf.

U.S.D.A. Economic Research Service (ERS) (2013). *Rural-Urban Continuum Codes: Documentation*. U.S.D.A: Washington DC. Online: http://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation.aspx

U.S.D.A. National Agricultural Statistics Service (NASS) (2014). *U.S. Census of Agriculture: United States Summary and State Data*, Vol. 1: Geographic Series, Part 51. Report AC-12-A-51., U.S.D.A.: Washington DC. Online: http://www.agcensus.usda.gov/Publications/2012/Full_Report/Volume_1,_Chapter_1_US/usappxa.pdf

Xu, Yuan, Murray Fyfe, Liz Walker, and Laura L.E. Cowen (2014) Estimating the number of injection drug users in greater Victoria, Canada using capture-recapture methods. *Harm Reduction Journal*11:9. doi: 10.1186/1477-7517-11-9.

Young, Linda J. Partitioning the capture-recapture estimate of the Census of Agriculture. Research and Development Division. Washington, DC: USDA, National Agricultural Statistics Service.

Thank you!

Linda.Young@nass.usda.gov