Using Multiple Imputation to Mitigate a Sample Reduction

Tara Murphy[1]

Darcy Miller[1]

Benjamin Reist[2]

[1] USDA National Agricultural Statistics Service;
[2] National Aeronautics and Space Administration, U.S.A.

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

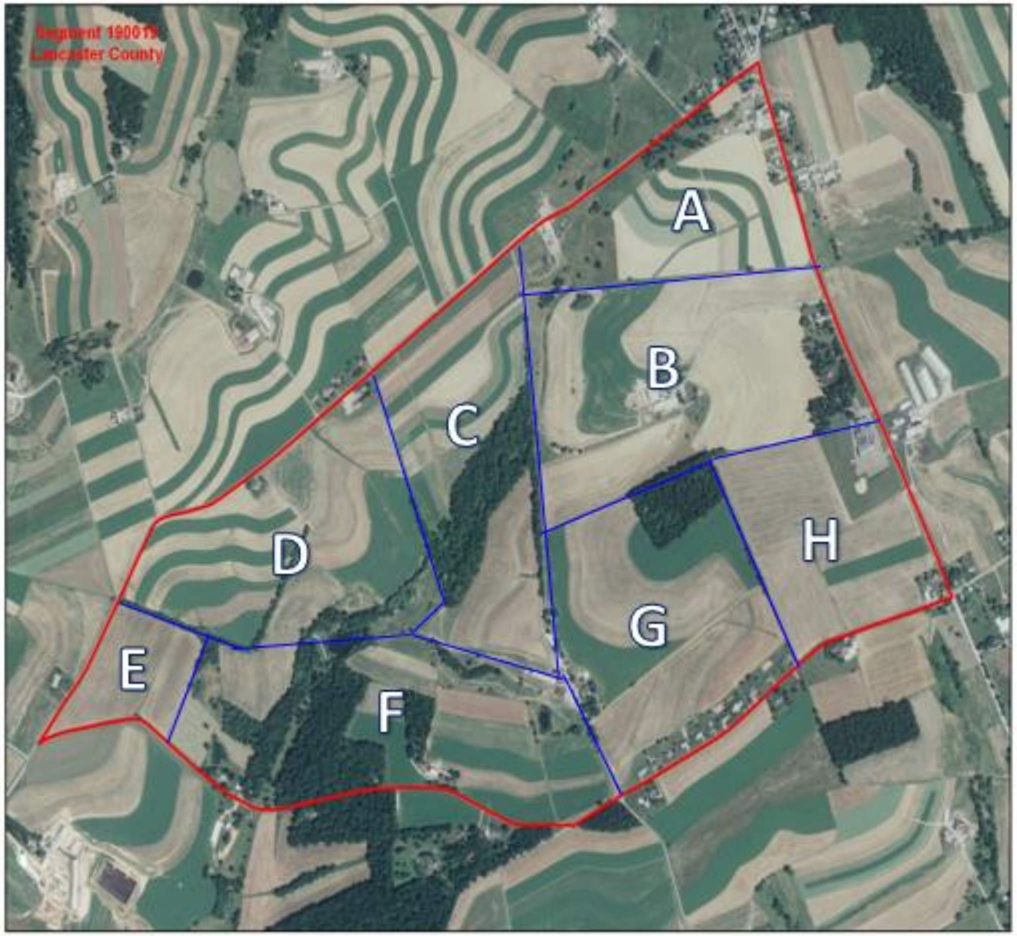
National Agricultural Statistics Service (NASS)

- Statistical arm of the United States Department of Agriculture (USDA)
- Conducts over 100 surveys each year, as well as the Census of Agriculture every five years
- Prepares more than 500 reports annually covering every facet of U.S. agriculture

  For example:

  - Production and food supplies
  - Prices paid and received by farmers
  - Farm income and finances
  - Number of farms and land in farms

June Area Survey (JAS)



- Area-frame based
- Segments of land sampled
- Sampled segments divided into tracts representing unique land operating arrangements
- Conducted annually via in-person interviews

JAS Stratification Design

Stratum	Percent Cultivated	Segment Size	Chance of Selection
10’s	>50% Cultivated	1.00 sq. mi.	Equal
20’s	15-50% Cultivated	1.00 sq. mi.	Equal
31	Ag Urban	0.25 sq. mi.	Equal
32	Commercial	0.10 sq. mi.	Equal
40’s	<15% Cultivated	2.00 sq. mi.	Equal
50	Non-Ag	PPS	Proportional to Seg Size

JAS Panel Design

No. of Segments selected in each Substratum	Rotating Replication Numbers by Survey Year				
	Rep Group 1	Rep Group 2	Rep Group 3	Rep Group 4	Rep Group 5
2	1	2			
3	1	2	3		
4	-	2	3	4	1
5	1	2	3	4	5
6	1	2	3	4	5,6
7	1	2	3	4,7	5,6
8	1,6	2,7	3,8	4	5
9	1,6	2,7	3,8	4,9	5
10	1,6	2,7	3,8	4,9	5,10
11	1,6,11	2,7	3,8	4,9	5,10
12	1,6,11	2,7,12	3,8	4,9	5,10
13	1,6,11	2,7,12	3,8,13	4,9	5,10
14	1,6,11	2,7,12	3,8,13	4,9,14	5,10
15	1,6,11	2,7,12	3,8,13	4,9,14	5,10,15
16	1,11,16	2,7,12	3,8,13	4,9,14	5,6,10,15
17	1,11,16	2,12,17	3,8,13	4,7,9,14	5,6,10,15

JAS Purpose

- Provides key indications for many agricultural aspects, including:
  - Planted acreage for most row crops and small grains
  - On-farm grain stocks
  - Land values
  - Technology use
  - Farm number estimates
- Measures the incompleteness of the NASS List Frame
- Serves as the sampling frame for not-on-list follow-on surveys and row crop objective yield surveys
- Used in the Dual System Estimator for the Census of Agriculture

Problem

- Budget cuts
  - JAS incurs the largest data collection costs to NASS, outside of the Census of Agriculture and reimbursable surveys
  - As a result, a reduction of the JAS sample was determined by the NASS Senior Executive Team

Past Remedies

- "Freeze" sample in 2017
  - No new segments rotated into the sample and no segments rotated out
    - This provided a reduction in cost since newly sampled segments are more expensive to enumerate
    - Required a panel to remain in the survey for six years instead of five through 2021
      - Increases respondent burden, which may lead to increased nonresponse or increased measurement error due to fatigue

Past Remedies

- Cut sample in 2018
  - Two panels were rotated out (those samples drawn in 2012 and 2013) and one panel rotated in, leaving four panels for data collection and estimation
    - Decreased sample size
    - In a rotation scheme / longitudinal study, this led to issues in sample design

Potential Past Remedy

- Impute some segments in non-speculative states in lieu of in-person interviews
  - Helps respondent burden to maintain response rates
  - Bonus that it helps with budget

Case Study – 2018

- Conducted in 2019, with 2018 JAS data, under urgent constraints
- Consisted of 99 simulated JAS response data sets
  - Approximately 50% of the non-speculative state segments randomly set to missing
    - First stratified by segment year, state, and sampling stratum
  - New segments not eligible to be set to missing
  - Led to approximately 9% of segments being imputed, yielding an estimated cost savings of about $232,000[+]

[+] Based on cost estimates provided for the simulation study

Case Study Methods

- Predictive mean matching implemented using SAS PROC MI with multiple imputation
  - Utilized current year collected data and previous year collected data as well as any other appropriate sample design information

Case Study

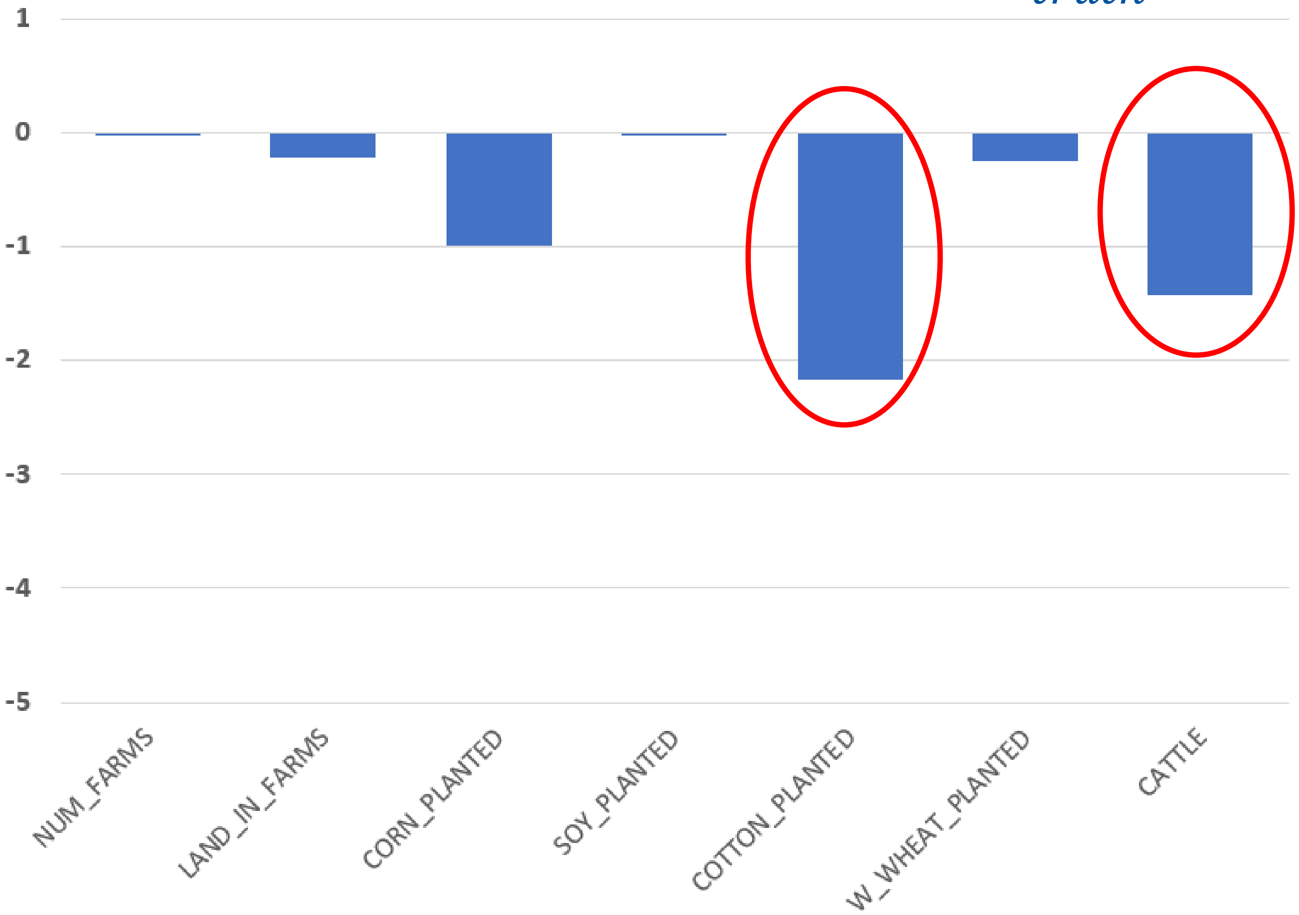
- List of key variables considered:

Number of Farms
Total Land in Farms
Corn Acres Planted
Soybean Acres Planted
Cotton Acres Planted
Winter Wheat Acres Planted
Total Cattle

Case Study Results

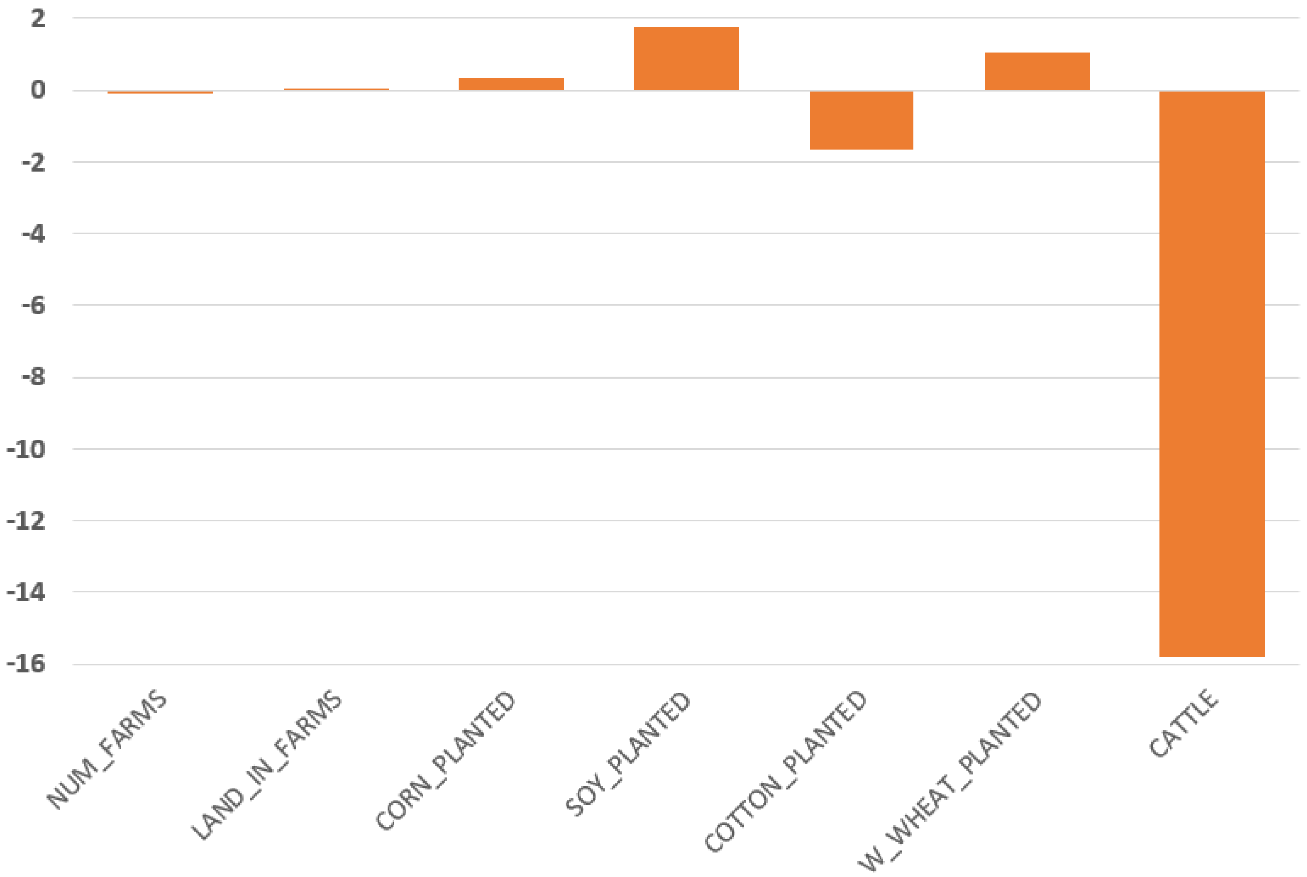
- Average national level percent differences in estimates

$$\frac{simulated - truth}{truth} \times 100$$



Case Study Results

- Average national level percent differences in standard errors



Case Study Results

- Preliminary results using initial models showed promise that imputation could be a viable substitution for data collection on some segments

Future Strategy – 2016 Study

- 2016 JAS dataset used
  - All states were included

- Removed one panel entirely and imputed

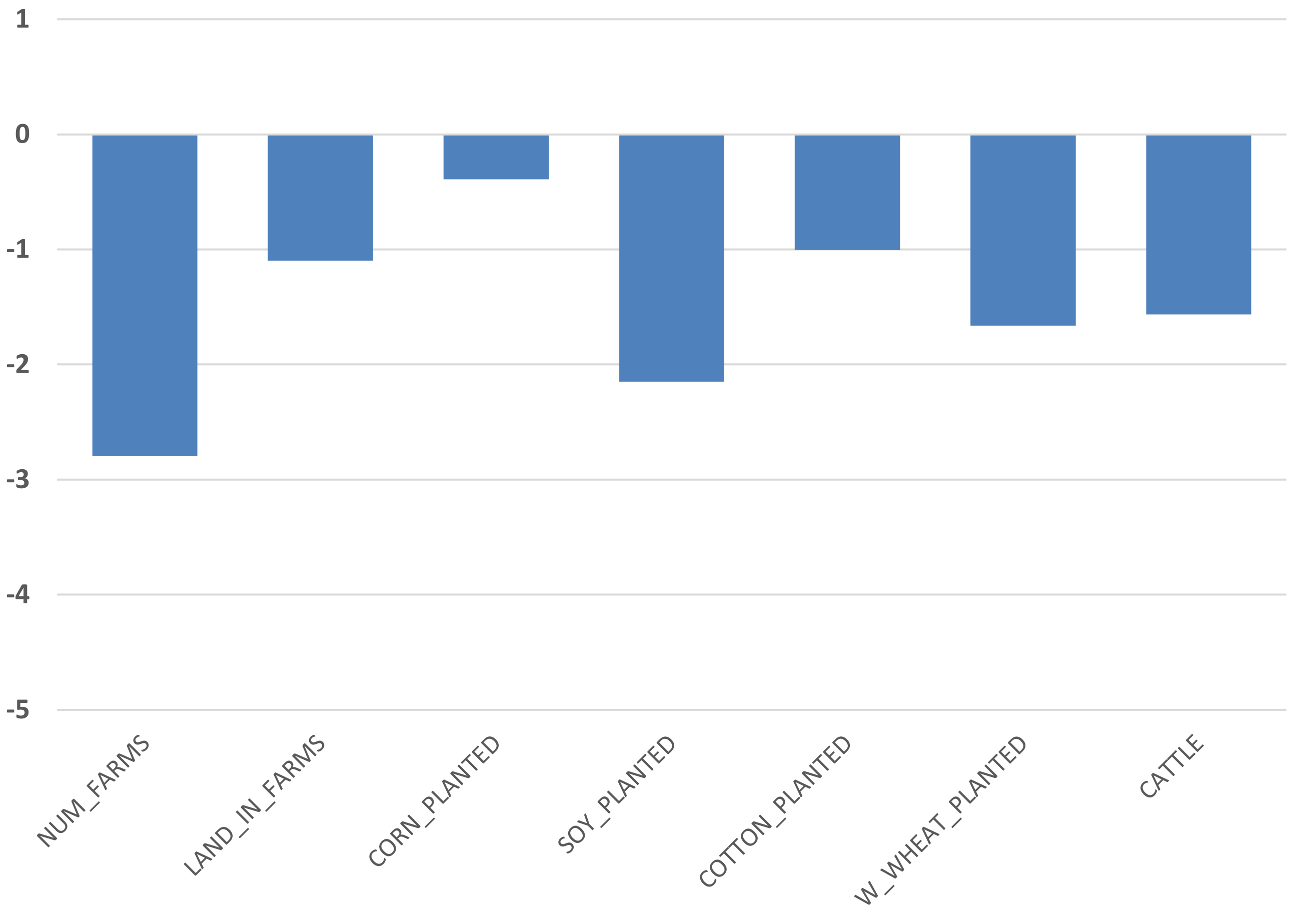
2016 Study

- List of key variables considered:

Number of Farms
Total Land in Farms
Corn Acres Planted
Soybean Acres Planted
Cotton Acres Planted
Winter Wheat Acres Planted
Total Cattle

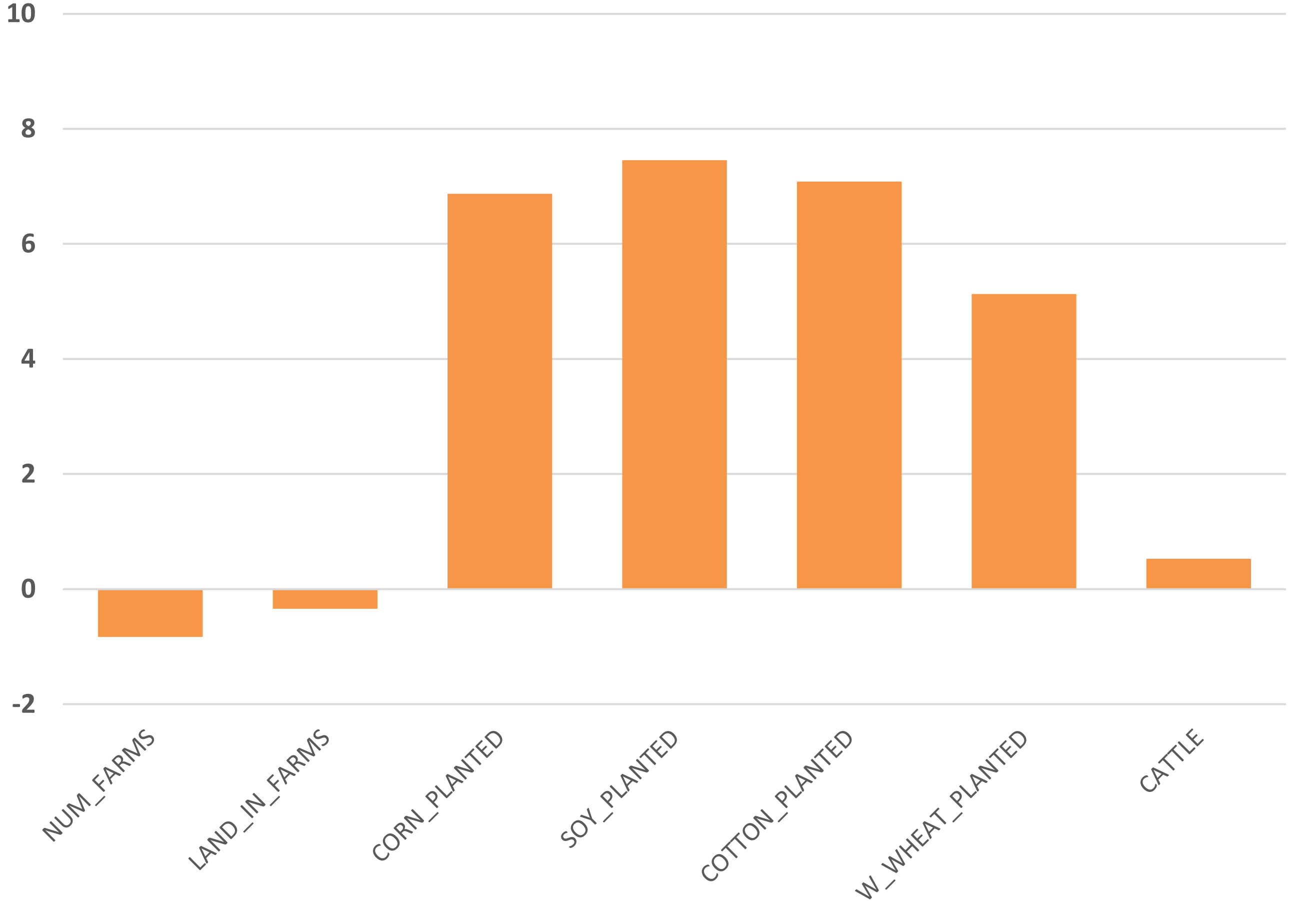
2016 Study

- National level percent differences in estimates



2016 Study

- National level percent differences in standard errors



Future Work

- Review state and regional level key estimates for 2016 study
- Remove panel and reweight 2016 JAS
- Add imputed panel to 2016 JAS, resulting in six total sample panels, where two are imputed
  - Selected segment would follow:
    - Years 1 & 2: data collection
    - Year 3: imputed, no data collection
    - Years 4 & 5: data collection
    - Year 6: imputed, no data collection
  - Does not change way segments are sampled
  - Process is consistent each year

References

- Andridge R. and Little R.J.A. (2010), "A Review of Hot Deck Imputation for survey Non-Response." International Staitstical Review 78 (1): 40-64
- Barboza, W. and Young, L. J. (2018), "Reduction of Segments in the June Area Survey." NASS Decision Memorandum, DM-03-18.
- Cotter, J. and Nealon, J. (1987), "Area Frame Design for Agricultural Surveys." U.S. Department of Agriculture, National Agricultural Statistics Service. Washington, D.C.
- Davies, C. (2009), "Area Frame Design for Agricultural Surveys." U.S. Department of Agriculture, National Agricultural Statistics Service. Washington, D.C. RDD Research Report Number RDD- 09-xx. https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf
- Little, R.J. A. (1988), "Missing –Data Adjustments in Large Surveys (with Discussion)." Journal of Business Economics and Statistics 6(3): 287-301.
- Rubin, D.B. (1976), "Inference and Missing Data." Biometrika, 63, 581-592.
- Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Rubin, D.B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations". Journal of Business Economics and Statistics 4 (1):87-94.
- Schenker, N. and J.M.G. Taylor (1996), "Partially Parametric Techniques for Multiple Imputation." Computation Statistics and Data Analysis 22 (4): 425-46.
- van Buuren , S., Brand , J. P.L., Groothuis-Oudshoorn, C. G.M., and Rubin, D.B. (2006), "Fully conditional specification in multivariate imputation." Journal of Statistical Computation and Simulation, 76:12, 1049-1064, DOI: 10.1080/10629360600810434

Thank you!

Tara Murphy

Tara.Murphy@usda.gov