Using a Web-scraped List-frame for an Agricultural Survey

Habtamu Benecha, Bruce A. Craig, Grace Yoon, Zachary Terner*
Denise A. Abreu, Linda J. Young

National Agricultural Statistics Service (NASS)
United States Department of Agriculture

*National Institute of Statistical Sciences (NISS)

FCSM 2021 Research & Policy Conference

November 3, 2021

Disclaimer

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

Background

The June Area Survey (JAS)

- Conducted annually
- Based on an area-frame that covers all land in the continental U.S.
- In-person enumeration
- NASS uses the JAS to produce comprehensive estimates of land use and agricultural activities
  - Number of farms and land in farms
    - $1000 in sales or potential sales
  - Undercoverage adjustment for the Census of Agriculture and several NASS surveys
- The JAS is an expensive survey; in-person enumeration, maintenance of the area-frame, rotation of segments

Background

- NASS is exploring ways to lower costs and improve data collection by leveraging new statistical methods and technologies
- The June Area Research Project (JARP): 2019 Pilot study
  - Assess the viability of using a web-scraped list-frame to replace the area-frame
  - Evaluate the effectiveness of collecting data by mail, telephone and the web
- Parallel study design: the 2018 JAS and the 2017 Census are used to evaluate the estimates and the quality of information from the pilot study

Goals: Discuss the estimation approaches and explore alternatives

2019 JARP: Frames and Surveys

- Four states: KS, NE, NY, PA
- A web-scraped list-frame developed for each state by scraping for potential farms
- State specific and national open-data sources used
  - National open-data sources for two of the states
  - National and state-specific sources for the other two states
- Record linkage to NASS' main list-frame
- Data collection in two phases

2019 JARP: Phase 1 Survey

- Sample from records on the web-scraped frame not linked to NASS' list-frame
- Use the data to estimate coverage weights for NASS list frame records
- Screening questionnaire
  - Determine farm status
  - Collect information for farm records: mail, web, and telephone

2019 JARP: Phase 2 Survey

- Sample from records on NASS' list-frame
- Some of the Phase 2 sample records linked to both frames
- Obtain information consistent with JAS by using mail, web, and telephone
- Produce JAS estimates using capture-recapture approach
- Farms and non-farms

Estimation

- Number of farms, land in farms, and coverage adjustments for six surveys
- Capture-recapture approach (Young et.al, 2017† ; Hyman et.al, 2021)
- Coverage and response probabilities

Number of farms and land in farms: Capture-recapture weight

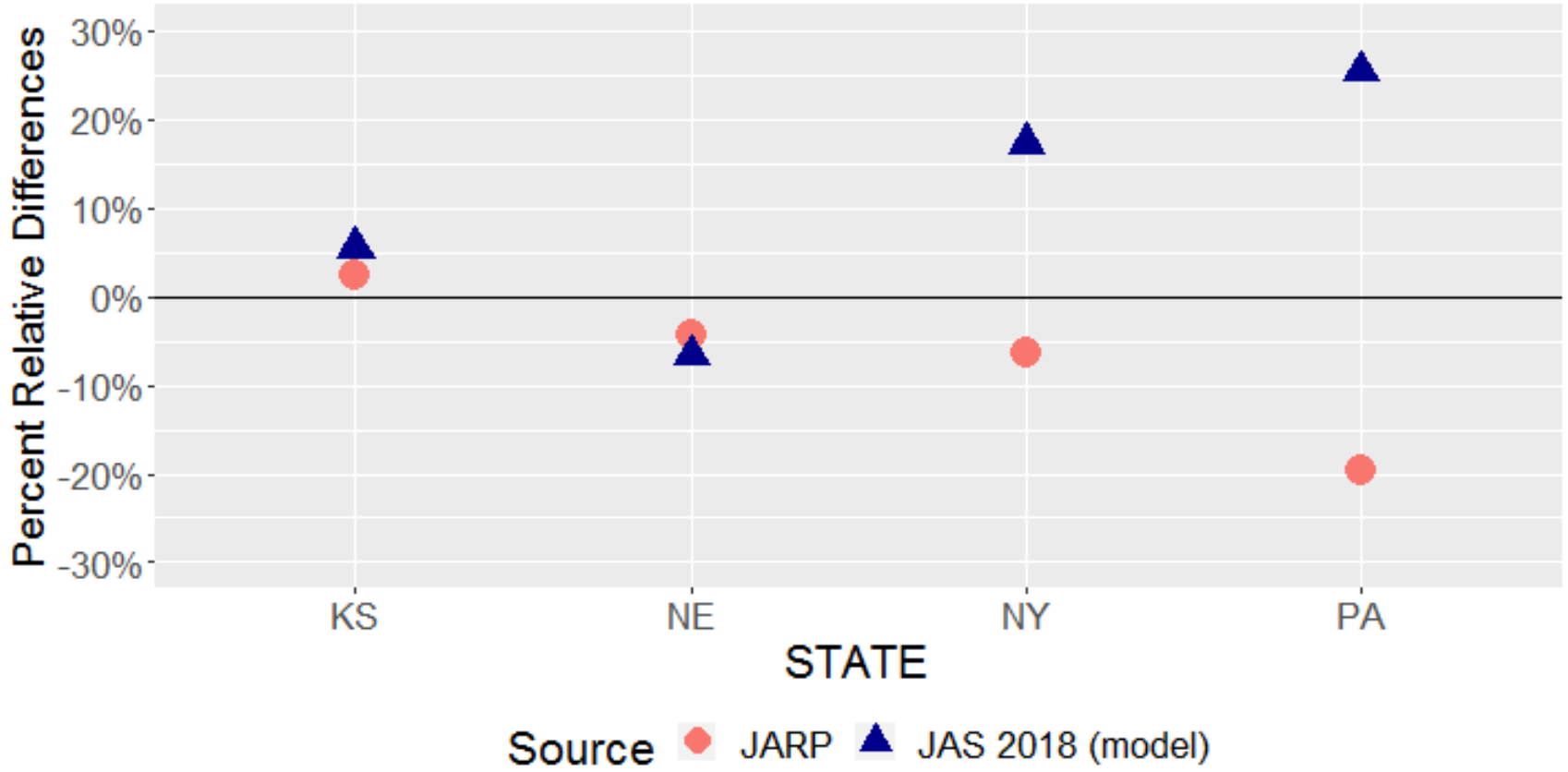
$$\pi_{si} = \frac{w_{si}}{\pi_{si}^{R} \times \pi_{si}^{C}} \qquad (1)$$

- Coverage probabilities (π_{si}^{C}) and response probabilities (π_{si}^{R}) – Logistic regression models
- Categorical covariates

† Young et al. (2017) The 2012 Census of Agriculture: A capture–recapture analysis. Journal of Agricultural, Biological and Environmental Statistics. 22, 523-539.

Results

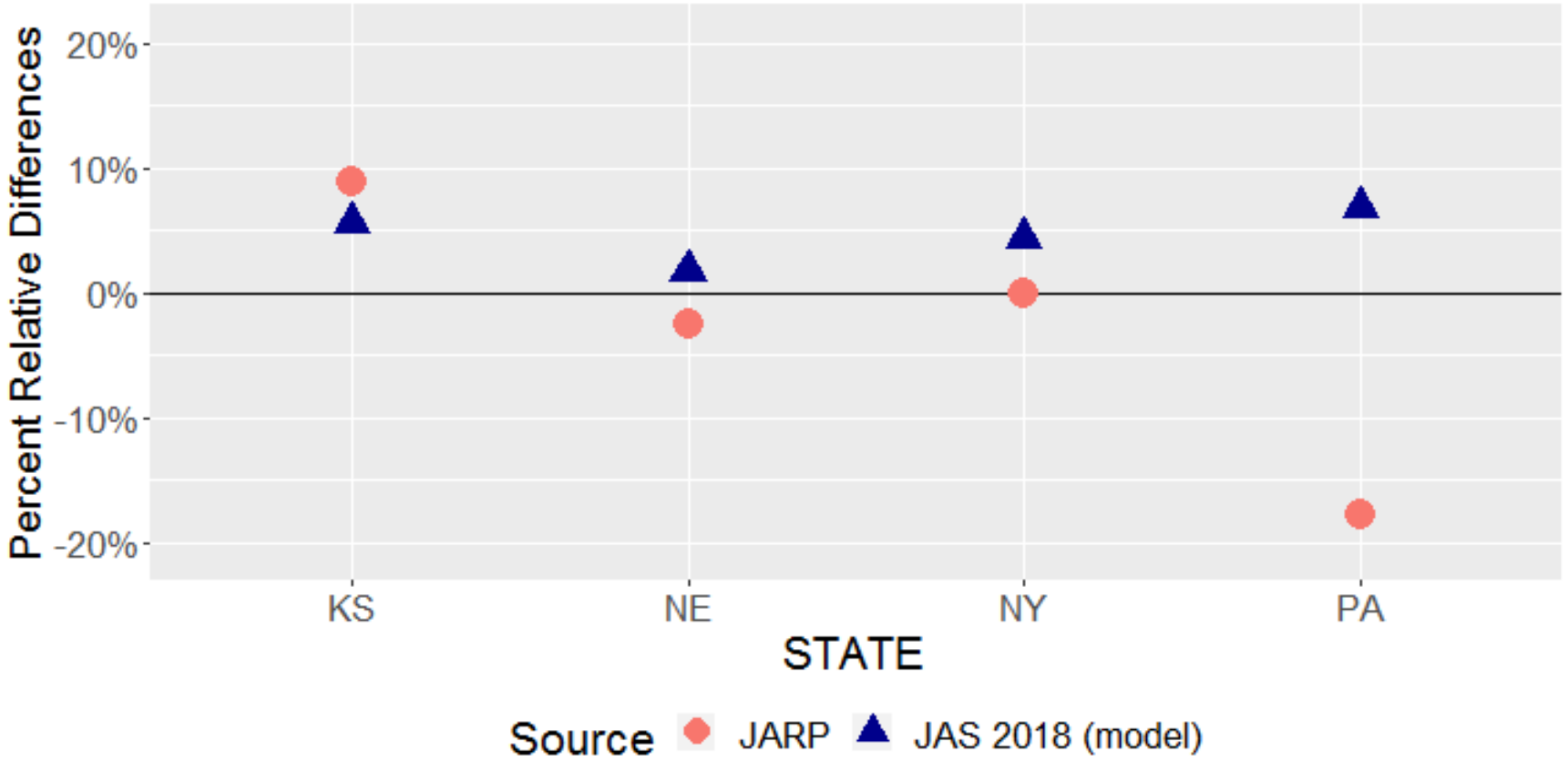
Percent relative differences of the **number of farms** estimates from 2019 JARP and 2018 JAS compared to the 2017 Census estimates



- **Reference to 2017 Census**

Results

Percent relative differences of the **land in farms** estimates from JARP and the JAS compared to the 2017 Census estimates



➢ **Reference to 2017 Census**

An Alternative Capture-recapture Model: Sampling from the Entire Web-scraped Frame

- Neither frame covers the population of farms
- Extension to Alho (1990)*
  - Each frame represents a large sample from the population of potential farms
  - The two surveys (S_1 & S_2) provide sub-samples from the population
- Probability of inclusion in a survey is defined as a products of
  - The probability of capture by the corresponding frame
  - The conditional probability of inclusion in the sample given the record was captured by the frame

* Alho (1990) Logistic Regression in Capture-Recapture Models. *Biometrics*, 46(3): 623-635

An Alternative Capture-recapture Model

- Capture-status: Multinomial distribution
- Four sets of logistic regression models
- The probability of inclusion in at least one survey, ϕ_i, estimated

Total number of potential farms

$$\widehat{N} = \sum_{i \epsilon \mathcal{C}} \frac{1}{\widehat{\phi_i}} \quad (\mathcal{C}\text{: the set of all captured potential farms}) \qquad (2)$$

The number of farms

$$\widehat{F} = \sum_{i \epsilon \mathcal{F}} \frac{1}{\widehat{\phi_i}} \quad (\mathcal{F}\text{: the set of all captured farms}) \qquad (3)$$

Simulation Study

- Two sets of different covariates to generate records for the frames and the surveys; four Bernoulli distributions
- Farm status: a function of one variable
- 500 replicates

True Number of Potential Farms (N)	Model Estimated Mean of $\widehat{N}/N$
25,000	1.064
40,000	1.015
60,000	1.010
100,000	1.009
250,000	1.006

True Number of Farms (F)	Model Estimated Mean of $\widehat{F}/F$
21,031	1.071
33,618	1.015
50,576	1.010
84,296	1.009
210,543	1.006

Summary and future work

- Promising results
  - Many of the estimates are close to results from the JAS, the Census and multiple-frame analysis
- Main challenge: under-estimation of some commodity values
- Low response-rate
- Quality and quantity of covariates
- Scraping at state and national levels may improve estimates
- Sampling from the entire web-scraped frame (several options for estimation)

Thank You!

Questions?

habtamu.benecha@usda.gov

Frames and samples