Uncertainty Assessment of Finite-population Medians Under Complex Capture-Recapture Sampling Designs

Luca Sartore*† Habtamu Benecha† Valbona Bejleri† Lu Chen*†

Abstract

When the weights associated with data collected from a finite population (i.e., a sample or a census) vary, estimators of the median are generally obtained from the (estimated) cumulative distribution function (CDF). Median estimators based on the CDF are often associated with sorting algorithms that asymptotically require $O(n \log n)$ operations. To improve the computational efficiency, alternative algorithms requiring $O(n)$ operations are investigated and extended under complex sampling designs or, as in the case of a census, after weight adjustments. Furthermore, the uncertainty associated with median estimators is traditionally computed using replicate methods, such as delete-a-group jackknife and bootstrap. Although the bootstrap approach has been shown to be more consistent than the leave-one-out jackknife when estimating the uncertainty of quantiles, it usually requires more iterations than the delete-a-group jackknife. More computationally efficient algorithms that also account for the uncertainty introduced by calibration are desirable. This paper describes and compares several simulation studies that address both accuracy and timeliness of the median standard error.

Key Words: Median estimator, Uncertainty, Complex sampling designs, Accuracy, Precision, Computational efficiency.

1. Introduction

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture every five years to quantify the number of U.S. operations with sales or potential of at least $1,000 in agricultural products. NASS has adopted a Dual-System Estimation (DSE) methodology since 2012 to improve the Census estimates. This technique accounts for those farms that are not captured by the census mailing list (CML). The current practice employs data coming from two independent frames to adjust the estimates for undercoverage, non-response, and misclassification. The June Area Survey (JAS) is an area frame survey conducted each year during the month of June. NASS keeps the JAS sampl independent from the CML, thus one can use any standard capture-recapture technique to estimate the population size of the U.S. farms.

Abreu et al. (2010) studied the impact of JAS procedures on the farm number estimates, and Young et al. (2017) further refined the estimation framework based on standard logistic regression. However, the separate estimation of the probability models for coverage, non-response, response follow-up, and misclassification does not utilize all the available Census and JAS data. Inspired by Alho (1990), a single framework based on a penalized likelihood has been developed to estimate simultaneously all the parameters of three logistic models.

Replicate methods (Kott, 2001, 2005, 2006; Hyman et al., 2021) have been studied and applied for estimating the variances of Census totals, while linearization methods have not been considered due to the estimation complexity of the Census of Agriculture. Furthermore, the analyses of historical data (NASS, 2012, 2017b) and a simple simulation

*National Institute of Statistical Sciences, 1750 K Street NW Suite 1100, Washington, DC 20006, lsartore@niss.org

†National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250

study (see e.g. Appendix A) have identified an upward bias when providing standard errors for the Census medians when using replicate methods (see Figure 1), indicating these methods lead to reporting a greater level of uncertainty associated with an estimated median than is truly present. NASS reports a limited number of medians in its Census publications, and others (with their standard errors) are computed internally and used by the U.S. Congress for policymaking.

A new weighted median approach with linear-time complexity is proposed to accelerate part of the computations required by the replication methods in the following sections. Furthermore, the performances of the proposed delete-a-group jackknife (DAGJ) and nonparametric bootstrap sampling (NPBS) are evaluated to account for the uncertainty associated with both the new estimation methodology and the integer calibration (INCA) algorithm developed by Sartore et al. (2019).

In Section 2, a computationally efficient median estimator is introduced for improving the timeliness of replicate algorithms. Two variance estimation techniques for weighted medians are introduced in Section 3. In particular, a DAGJ estimator is developed in Section 3.1, and a NPBS algorithm is presented in Section 3.2. A simulation study is performed to assess the performances of the proposed variance estimators, and its results are discussed in Section 4. Finally, concluding remarks are given in Section 5.

2. A suitable median estimator for replicate algorithms

The median estimator, $\hat{\theta}$, often requires more computations than the mean. In general, the estimating equation for computing weighted medians is $\hat{F}(\theta) - \frac{1}{2} = 0$, where

$$\hat{F}(\theta) = \sum_{i=1}^{n} w_i \mathbb{1}_{[x_i,\infty)}(\theta) \Big/ \sum_{i=1}^{n} w_i$$

corresponds to the Hájek (1971) estimator.

Standard methods to estimate the median, $\hat{\theta} = \inf\{\theta : \hat{F}(\theta) \geq \frac{1}{2}\}$ as proposed by Kuk (1988), are usually implemented through sorting algorithms that have computational complexity of $O(n \log n)$, such as the Quick-sort algorithm (Hoare, 1962; JaJa, 2000). Instead of sorting the observations, an approach inspired by variations of the Radix-sort algorithm (McIlroy et al., 1993) is implemented by borrowing some ideas from Weiss (2006) and Zhang et al. (2014). This approach results in a weighted median estimator suitable to be used in replication algorithms (such as the DAGJ or the NPBS). The weighted medians are computed with linear complexity, i.e. $O(n)$, by partitioning the observed range iteratively. Thus, the identification of $\hat{\theta}$ is accomplished using histograms with range lengths that converge to zero.

3. Variance estimation of weighted medians

The total variation of the Census estimates begins with the data collection of the Census given that the JAS data are fully available. The variation introduced by the imputation and editing processes is here ignored, although it should be accounted at later development stages. Modeling, which should include variable selection as part of the estimation process, currently provides the most substantial contribution to the uncertainty associated with Census estimates. Furthermore, calibration adjusts the estimated weights computed in the modelling phase producing an integer number for each respondent farm in the CML.

The new Census estimators (for both totals and medians) use only the respondent units in the CML. Hence, the estimation of the standard errors associated with these totals can be

performed using replicate methods (such the DAGJ and the NPBS). Since the estimation of the Census totals is based on a capture-recapture procedure, the combination of Census data with the JAS information should be similarly treated in the replicate method of choice. Because the JAS data are fully available during the Census data collection, the selection scheme should characterize the uncertainty by randomizing only the samples acquired by the Census. Furthermore, some of the Census records are flagged as "Must Cases" and have final integer calibrated weight equal to one; that is, they are selected from the population of interest with probability one. Therefore, "Must Cases" should not be subject to any resampling or deletion procedure and should provide a constant basis to build replicate estimates.

3.1 The DAGJ estimator

Given each observation $i \in \mathcal{C} \cap \mathcal{R}$, where $\mathcal{C}$ and $\mathcal{R}$ respectively represent the sets of indexes for the CML and those that respond to the Census questionnaire. Each observation i is randomly assigned into a group $\mathcal{G}_g$, for $g = 1, \ldots, G$, where G denotes the number of jackknife groups.

Under the considerations provided in the Introduction, the likelihood developed for the Census of Agriculture must use the contribution of the observation i

$$L^{(g)}_{\text{JK},i} = \frac{\left(\pi^{(g)}_{J,i}\right)^{y_{J,i}} \left\{\pi^{(g)}_{C,i}\left(\rho^{(g)}_i\right)^{r_i}\left(1-\rho^{(g)}_i\right)^{1-r_i}\right\}^{y_{C,i}A_i} \left(1-\pi^{(g)}_{C,i}\right)^{y_{J,i}(1-y_{C,i}A_i)}}{K_i^{(y_{C,i}+y_{J,i}-y_{C,i}y_{J,i})(y_{J,i}+A_i-y_{J,i}A_i)}\left(1-\pi^{(g)}_{J,i}\right)^{-y_{C,i}A_i(1-y_{J,i})}},$$

where $y_{C,i}$ and $y_{J,i}$ are respectively indicator variables for Census and JAS coverage, r_i represents the indicator variable for positive response from a farm in the CML. The notation A_i accounts for "Must Cases" and group deletion and is defined as

$$\begin{aligned} A_i &= z_i + \{1 - \mathbb{1}_{\{g\}}(\tilde{g}_i)\} - z_i\{1 - \mathbb{1}_{\{g\}}(\tilde{g}_i)\}, \\ &= 1 - \mathbb{1}_{\{g\}}(\tilde{g}_i) + z_i\mathbb{1}_{\{g\}}(\tilde{g}_i), \end{aligned}$$

where z_i is one if observation i is a "Must Case", zero otherwise; $\tilde{g}_i$ denotes the Jackknife group randomly assigned to the observation i; and $\mathbb{1}_{\mathcal{A}}(x)$ denotes an indicator function, i.e.

$$\mathbb{1}_{\mathcal{A}}(x) = \begin{cases} 0, & \text{if } x \notin \mathcal{A}, \\ 1, & \text{if } x \in \mathcal{A}, \end{cases}$$

for a generic set $\mathcal{A}$. The normalization constant K_i is defined as

$$\begin{aligned} K_i &= \pi^{(g)}_{C,i}\left(1-\pi^{(g)}_{J,i}\right)\rho^{(g)}_i + \pi^{(g)}_{J,i}\left(1-\pi^{(g)}_{C,i}\right) + \pi^{(g)}_{C,i}\pi^{(g)}_{J,i}\rho^{(g)}_i + \pi^{(g)}_{C,i}\pi^{(g)}_{J,i}\left(1-\rho^{(g)}_i\right), \\ &= \pi^{(g)}_{C,i}\left(1-\pi^{(g)}_{J,i}\right)\rho^{(g)}_i + \pi^{(g)}_{J,i}, \end{aligned}$$

where the modelled probabilities are as follows:

$$\begin{aligned} &\pi^{(g)}_{C,i} = \{1+\exp(-\boldsymbol{X_{c,i}}\boldsymbol{\beta}^{c,g})\}^{-1}, \qquad \pi^{(g)}_{J,i} = \{1+\exp(-\boldsymbol{X_{j,i}}\boldsymbol{\beta}^{j,g})\}^{-1}, \text{ and} \\ &\rho^{(g)}_i = \{1+\exp(-\boldsymbol{X_{r,i}}\boldsymbol{\beta}^{r,g})\}^{-1}, \end{aligned}$$

for any deletion group $g = 1, \ldots, G$.

The G Jackknife replication groups are uniformly assigned at random to each observation $i \in \mathcal{C} \cap \mathcal{R}$. Afterwards, parameter estimates for all model-based probabilities are computed by maximizing the following penalized log-likelihood:

$$\ell_{\text{JK}}^{(g)} = \sum_{i \in \mathcal{J} \cup (\mathcal{C} \cap \mathcal{R})} \log\left(L_{\text{JK},i}^{(g)}\right) + \sum_{i \in \mathcal{C} \cap \mathcal{R}} A_i \log\left(\frac{u_i}{\pi_{C,i}^{(g)} \rho_i^{(g)}}\right) \mathbb{1}_{[u_i, +\infty)}\left(\frac{1}{\pi_{C,i}^{(g)} \rho_i^{(g)}}\right),$$

where u_i corresponds to the upper bound used for to assess the calibration weight of observation i, and $\mathcal{J}$ denotes the set of indexes for the observations covered/captured by the JAS.

Model based weights $\tilde{w}_i^{(g)} = \left(\hat{\pi}_{C,i}^{(g)} \hat{\rho}_i^{(g)}\right)^{-1}$, for any $i \in (\neg \mathcal{G}_g) \cap \mathcal{C} \cap \mathcal{R}$ are successively calibrated through INCA using group-specific benchmarks. Thus, the integer calibrated weights $\hat{w}_i^{(g)}$, for any $i \in (\neg \mathcal{G}_g) \cap \mathcal{C} \cap \mathcal{R}$, are then used for computing the group-specific medians (or totals) $\hat{\theta}^{(g)}$ appearing in the following Jackknife estimator:

$$\mathsf{VAR}[\,\hat{\theta}\,] = \frac{G-1}{G} \sum_{g=1}^{G} \left(\hat{\theta}^{(g)} - \hat{\theta}^{*}\right)^2,$$

where $\hat{\theta}^{*}$ corresponds to the final Census estimate. For implementation purposes, the jackknife calibrated weights $\hat{w}_i^{(g)}$ must be set to zero, for all $i \in \mathcal{G}_g \cap \mathcal{C} \cap \mathcal{R}$.

3.2 The NPBS estimator

As for the DAGJ approach, the assumptions considered above in Section 3 still hold when developing a NPBS estimator for the standard errors of Census medians (and totals).

All observations acquired through the JAS are going to be available throughout each bootstrap replicate, as well as for the "Must Cases" that will be selected with probability one in our replicate subsamples since they have weights equal to one. Therefore, only Census records with final integer calibrated weights in the set $\{2, 3, \ldots, 6\}$ will be resampled without replacement. No record in the Census is allowed to have weight larger than 6.

To avoid the construction of a pseudo-population, each record $i \in \mathcal{C} \cap \mathcal{R}$ is associated with a randomly generated integer number $S_i^{(b)} \sim \mathsf{Bernoulli}(\hat{w}_i^{-1})$, where $\hat{w}_i^{-1}$ is the final integer calibrated weight associated with record i for any bootstrap replicate $b = 1, \ldots, B$, where B is a large integer number. Thus, the random variable S_i represents the selection indicator of the record $i \in \mathcal{C} \cap \mathcal{R}$.

The contribution of a generic record i to the likelihood at the bootstrap iteration b is

$$L_{\text{BS},i}^{(b)} = \frac{\left(\pi_{J,i}^{(b)}\right)^{y_{J,i}} \left\{\pi_{C,i}^{(b)} \left(\rho_i^{(b)}\right)^{r_i} \left(1 - \rho_i^{(g)}\right)^{1-r_i}\right\}^{y_{C,i} V_i^{(b)}} \left(1 - \pi_{C,i}^{(b)}\right)^{y_{J,i}(1 - y_{C,i} V_i^{(b)})}}{K_i^{(y_{C,i} + y_{J,i} - y_{C,i} y_{J,i})(y_{J,i} + V_i^{(b)} - y_{J,i} V_i^{(b)})} \left(1 - \pi_{J,i}^{(b)}\right)^{-y_{C,i} V_i^{(b)} (1 - y_{J,i})}},$$

where $V_i^{(b)}$ accounts for "Must Cases" and bootstrap selection, and it is defined as

$$\begin{aligned} V_i^{(b)} &= z_i + S_i^{(b)} - z_i S_i^{(b)}, \\ &= S_i^{(b)} + z_i \left(1 - S_i^{(b)}\right). \end{aligned}$$

The parameter estimates for all probabilities are computed by maximizing the following penalized log-likelihood:

$$\ell_{\text{BS}}^{(b)} = \sum_{i \in \mathcal{J} \cup (\mathcal{C} \cap \mathcal{R})} \log\left(L_{\text{BS},i}^{(b)}\right) + \sum_{i \in \mathcal{C} \cap \mathcal{R}} V_i^{(b)} \log\left(\frac{u_i}{\pi_{C,i}^{(b)} \rho_i^{(b)}}\right) \mathbb{1}_{[u_i, +\infty)}\left(\frac{1}{\pi_{C,i}^{(b)} \rho_i^{(b)}}\right).$$

Weights $\tilde{w}_i^{(b)} = \left(\hat{\pi}_{C,i}^{(b)} \hat{\rho}_i^{(b)}\right)^{-1}$, for any $i \in \mathcal{C} \cap \mathcal{R}$: $V_i^{(b)} = 1$ are successively calibrated through INCA using specific benchmarks at every bootstrap iteration. The integer calibrated weights $\hat{w}_i^{(b)}$, for any $i \in \cap\mathcal{C} \cap \mathcal{R}$: $V_i^{(b)}$, are then used for computing the group-specific medians (or totals) $\hat{\theta}^{(b)}$ appearing in the following NPBS estimator:

$$\mathsf{VAR}[\,\hat{\theta}\,] = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \hat{\theta}^{*}\right)^2,$$

where $\hat{\theta}^{*}$ corresponds to the final Census estimate. For implementation purposes, the bootstrap calibrated weights $\hat{w}_i^{(b)}$ must be set to zero, for all $i \in \mathcal{C} \cap \mathcal{R}$: $V_i^{(b)} = 0$.

4. Practical application

The data from the 2017 US Census of Agriculture have been considered for two states: Connecticut and Illinois. These states have very different agricultural practices. In fact, farms in Connecticut predominantly produce hay and alfalfa, while corn and soybeans are the major crops cultivated in Illinois. In 2021, Connecticut had about 5,500 farms and about 380,000 acres of land devoted to agriculture. In Illinois, the number of farms was approximately 70,900 while the total farm land in 2021 was about 27,000,000 acres. The two 2017 datasets are, therefore, much different in the number of records to process and the type of information provided by the quantitative variables.

For this application, state-level medians and totals for land in farms are estimated using a capture-recapture methodology that simultaneously estimates the parameters for coverage and nonresponse using all the data available from both the CML and JAS. The estimated weights are then calibrated using the algorithm proposed by Sartore et al. (2019). The state-level totals for the number of farms are also provided to better assess the magnitude of the uncertainty for different quantities of interest. In particular, the coefficients of variation (CV) for both medians and totals are computed using the variances provided by the two replicate methods presented in Section 3.1 and 3.2. For the DAGJ methodology, the number of deletion groups has been set to $G = 10$, and the number of replicates $B = 100$ has been considered for the NPBS. The results computed for Connecticut are shown in Table 1, and those for Illinois are presented in Table 2.

Table 1: Estimated totals and medians in Connecticut, with respective CVs, based on DAGJ (with $G = 10$) and NPBS (with $B = 100$)

	N. Farms	**Land in Farm**
Estimated totals	5,565	405,807
CV based on DAGJ	5.05%	102.44%
CV based on NPBS	1.51%	31.62%
Estimated medians	–	20
CV based on DAGJ	–	15.73%
CV based on NPBS	–	4.26%

The calibrated total number of farms in Connecticut during 2017 is quite close to the published number of farms during 2021. This suggests that the number of agricultural operations in Connecticut has not changed much, and the total land operated by farms (based on the proposed methodology) has decreased when compared to the published number for 2017. These considerations, however, do not account for the variability of the estimates obtained with the proposed method. The CVs computed for the number of farms and the land devoted to farming using the DAGJ are larger than those produced with the NPBS. This aspect might be due to the violation of homogeneity assumptions among the groups. Furthermore, the variations observed between the estimates produced for 2017 and the number published in 2021 are not high enough to suggest any substantial change between 2017 and 2021. On the other hand, although a single Connecticut farm has an average of approximately 73 acres of land, the median value is 20 acres. This highlights a positive skewness of the distribution of land associated with individual farms.

Table 2: Estimated totals and medians in IL, with respective CVs, based on DAGJ (with $G = 10$) and NPBS (with $B = 100$)

	N. Farms	**Land in Farm**
Estimated totals	72,698	26,791,391
CV based on DAGJ	0.56%	39.19%
CV based on NPBS	0.27%	7.70%
Estimated medians	–	98
CV based on DAGJ	–	6.42%
CV based on NPBS	–	1.79%

The calibrated total number of farms in Illinois during 2017, on the other hand, is larger than the published number of farms during 2021. This suggests a decrease in the number of the agricultural operations in Illinois through time; however, the total land operated by farms has increased slightly. Furthermore, when accounting for the variability of the estimates obtained with the proposed method, no substantial changes for the land devoted to agriculture is observed. This, however, cannot be said for the total number of Illinois farms. In fact, the decrease noted between the estimates produced with the proposed methodology for 2017 is quite substantial if one considers the CV computed with the NPBS. However, one can draw opposite conclusions when looking at the CV produced with the DAGJ. In any case, the CVs produced with the DAGJ tend to be consistently higher than those produced with the NPBS, showing potential for improvement in producing unbiased variance estimates. Furthermore, the average land operated by a single farm in Illinois is estimated to be 369 acres while the median value is 98 acres. Thus, the distribution of an Illinois farm devoted to agriculture is highly skewed, and the average is affected by very large operations.

Similarly, NPBS CVs have been attained by looking at subsets of 10 replicates from the $B = 100$ replicates originally computed to produce the results in both Table 1 and 2. This aspect shows the NPBS produces satisfactory results even with a low number of replicates at the expense of statistical efficiency that is usually gained when $B \to \infty$. Nonetheless, considering $B = 10$ could be a convenient solution for implementing time-expensive computational tasks within a limited amount of time.

5. Conclusion and final remarks

The existing methodology to compute the variances is designed to obtained unbiased estimates of Census totals; however, it is not optimal for the variance estimation of Census

medians. Because, in some cases, medians are more informative than averages, especially for skewed distributions (e.g. land-in-farm and economic variables), these are often estimated by leveraging sorting algorithms. To improve the computational performances of replication methods, a new algorithm is proposed to quickly estimate the medians and avoid the sorting of the data.

While improving the computational algorithm to estimate the medians from a finite population by providing minor accelerations, a much larger set of operations is required for computing the variances when using a replication method. In fact, model fitting and calibration must be performed at each replicate requiring several minutes (on average 20) to obtain a set of calibrated weights. Furthermore, the DAGJ usually requires less replications than the NPBS. However, the NPBS has produced more reliable results (as shown in Section 4).

Extensions of the proposed methodology for all states are currently under investigation. However, this process will benefit greatly from the inclusion of an automatic covariate selection. Because each state requires a different set of covariates to produce sound results (mostly due to differences in agricultural practices), high-performance computing capabilities could allow processing several models simultaneously. These modern technologies have a potential to reduce the computational time required to improve model selection and obtain replicate weights through parallel computing devices (such GPUs).

Acknowledgements

The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy. This research was supported in part by the intramural research program of the US Department of Agriculture, National Agricultural Statistics Service (NASS).

A. Simulation results using 2017 variance methodology

To inspect the performance of the 2017 variance methodology (NASS, 2017a), a simple simulation study has been performed by considering the following steps:

1. Simulate a population of size N.
2. Generate uniformly distributed sampling weights, $w_i \in \{1, \ldots, 6\}$.
3. Perform Poisson sampling with selection probabilities $\pi_i = w_i^{-1}$.
4. Estimate the CVs of weighted medians using a direct approximation, the DAGJ, and the current 2017 Census methodology.
5. Iterate steps 3. and 4. to study the empirical distribution of the CVs assuming w_i are known for any $i = 1, \ldots, N$.

The box-plots shown in Figure 1 have been obtained with $1,000$ samples drawn from a population of size $N = 10,000$. The direct estimates of the CVs have been computed according to the following formula:

$$\mathsf{CV}[\,\hat{\theta}\,] = \frac{1}{2\sqrt{n}\, f(\hat{\theta}_y)\,\hat{\theta}},$$

where $f(\hat{\theta})$ is evaluated using the first derivative of a smooth approximation of the empirical cumulative distribution functions (Sedransk and Sedransk, 1979), $\hat{F}(\hat{\theta})$.

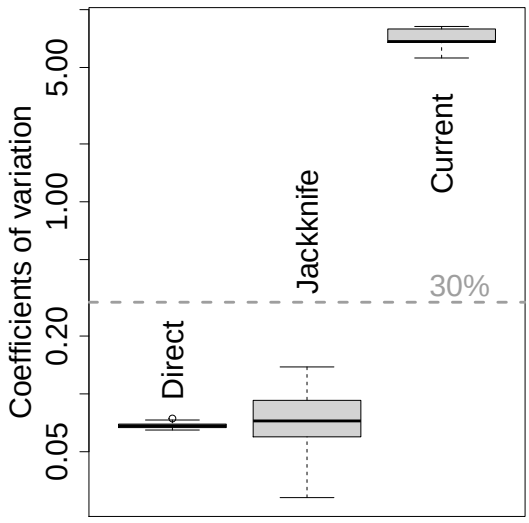

Figure 1: Results from a simulation study on the distribution of the variance estimates obtained through three different methods (direct approximation, DAGJ, and the current 2017 Census methodology).

Although the results computed with both the direct approximation and the DAGJ method are unbiased (i.e. with distributions centered and aligned according to a common median value), the direct approximation is a more statistically efficient estimator than the DAGJ method. On the other hand, the estimator based on the current 2017 methodology has been producing unbiased variance estimate for weighted totals but not for weighted medians (as seen in Figure 1).

References

Abreu, D. A., McCarthy, J. S., Colburn, L. A., et al. (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-10-03. Washington, DC: USDA, National Agricultural Statistics Service*.

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

Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one". *The foundations of survey sampling*, 236.

Hoare, C. A. (1962). Quicksort. *The computer journal*, 5(1):10–16.

Hyman, M., Sartore, L., and Young, L. J. (2021). Capture–recapture estimation of characteristics of US Local Food Farms using a web-scraped list frame. *Journal of Survey Statistics and Methodology*.

JaJa, J. (2000). A perspective on quicksort. *Computing in Science & Engineering*, 2(1):43–49.

Kott, P. S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17(4):521.

Kott, P. S. (2005). Using the delete-a-group jackknife in an economic survey of US farms. *55th Session of the International Statistical Institute, Sydney, Australia (CD-Rom)*.

Kott, P. S. (2006). Delete-a-group variance estimation for the general regression estimator under Poisson sampling. *Journal of Official Statistics*, 22(4):759.

Kuk, A. Y. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75(1):97–103.

McIlroy, P. M., Bostic, K., and McIlroy, M. D. (1993). Engineering radix sort. *Computing systems*, 6(1):5–27.

NASS, U. (2012). Census of Agriculture. Washington, DC: USDA National Agricultural Statistics Service.

NASS, U. (2017a). Appendix A: Census of Agriculture methodology. *United States Department of Agriculture, National Agricultural Statistics, available at:* `https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/usappxa.pdf`.

NASS, U. (2017b). Census of Agriculture. Washington, DC: USDA National Agricultural Statistics Service.

Sartore, L., Toppin, K., Young, L., and Spiegelman, C. (2019). Developing integer calibration weights for Census of Agriculture. *Journal of Agricultural, Biological and Environmental Statistics*, 24(1):26–48.

Sedransk, N. and Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *Journal of the American Statistical Association*, 74(368):754–760.

Weiss, B. (2006). Fast median and bilateral filtering. In *ACM SIGGRAPH 2006 Papers*, pages 519–526. Association for Computing Machinery, New York, NY, United States.

Young, L. J., Lamas, A. C., and Abreu, D. A. (2017). The 2012 Census of Agriculture: a capture–recapture analysis. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):523–539.

Zhang, Q., Xu, L., and Jia, J. (2014). 100+ times faster weighted median filter (wmf). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2830–2837.