Estimation of Capture Probabilities by Accounting for Sample Designs

Jake Abernethy* Luca Sartore† Habtamu Benecha‡ Clifford Spiegelman§

Abstract

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts the Census of Agriculture every five years to estimate the number of U.S. farms, as well as other agriculturally related population totals. NASS applies a Dual-System Estimation (DSE) methodology on data collected from the Census and the June Area Survey (JAS) to estimate the number of farms in the U.S.. Traditional multinomial-based capture-recapture methodology requires a model to estimate the probability of capture for every captured operation on either survey. Of course, the selection probabilities associated with the JAS area frame design are different from those associated with the Census. Such a difference makes it difficult to compute the exact JAS selection probabilities for farm records captured only by the Census. For this reason, we propose and compare three methods for estimating the overall capture probability. The first two methods involve approximating the JAS selection probabilities and the third conditions them out. We compare these three techniques to investigate their precision through a simulation study.

Key Words: Capture-Recapture, Estimation, Weights

1. Introduction

The National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agricultural every five years to obtain a variety of agricultural totals, the most prominent being the number of U.S. farms. The Census does not capture all farms, so some adjustment is needed to get an estimate of the true population totals. To this end, NASS has adopted a Dual System Estimation (DSE) methodology since 2012. This technique takes into account those farms that are not captured by the Census. Two independent surveys are used to adjust the totals: the Census itself, which is based on a list frame called the Census Mailing List (CML), and the June Area Survey (JAS), which is an area-frame survey conducted each year during the month of June. If it is assumed that the JAS and Census lists are independent given measurable covariates, standard capture-recapture techniques can estimate the total number of farms in the U.S.

Data from the Census is obtained via response to mailed out questionnaires. Thus the probability of response to the Census can be modeled by covariates relating to the propensity to respond, when a second supplementary survey (in our case the JAS) is available. The JAS uses a probability-sampling design to select data. In theory, all the land in the U.S. is divided into segments, each of which has a probability of selection given by the survey design. Segments are then selected at random based on this design and population totals are estimated as a function of the segment totals. In practice, the probabilities are only known for individuals actually selected in the JAS. In addition, the JAS samples consist of segments while traditional capture-recapture methods operate at the individual

*National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250, jake.abernethy@nass.usda.gov

†National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006

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

§Texas A&M University, 3135 TAMU, College Station, TX 77843-3135

(farm) level. These probabilities are only interchangeable if all farms are assumed to fit neatly into a single segment, which is not always the case. In summary, even if the JAS probabilities of selection are almost always available, they cannot be assigned to each individual farm belonging to the same segment, because federal budget constraints do not allow to sustain the high costs of maintenance of an area-frame at the farm-level. To obtain valid estimates, we will deal with the missing JAS probabilities. The first of two proposed solutions approximate the JAS sample design probabilities with logistic models. The last fits a regression model for the CML capture probability by conditioning out the JAS design probabilities.

The paper is organized as follows. In section 2, the capture-recapture framework of (Alho, 1990) is reviewed. Section 3 introduces the notation and reviews the JAS design. In section 4, models for approximating the JAS capture probabilities are discussed. Alternatively, a conditional model to remove the JAS probabilities is developed in section 5. Section 6 presents a simulation study to test the performance of the proposed methods. Finally, section 7 exposes some concluding remarks.

2. Capture-recapture with covariates

The literature on capture-recapture is quite substantial, as the problem setup can vary greatly with the situation. An overview is provided in Amstrup et al. (2010). The field is narrowed in this application as only two lists are available. The simplest choice is the Lincoln-Peterson estimator for two lists. If there are two registries, the estimate of the total population is simply the product of the number of individuals captured by each registry divided by the number captured in both. This method was first implemented by Laplace in 1802 to estimate the population of France (Cochran, 1978). This estimator assumes that the population is closed, the lists are independent, individuals are identified without error, and all individuals have the same probability of capture.

With covariates available from the Census and the JAS, the assumption of equal individual level probabilities of capture may be relaxed. If all variables are discrete, post stratification, where the population is divided into homogeneous groups and the Lincoln-Peterson estimator calculated for each group, may be used (Sekar and Deming, 1949). The disadvantage of this method is that the choice of groups is not always perfect and additional information allows for better partitioning criteria. It also does not allow for the use of continuous variables unless they are discretized. The method of Alho (1990) and Huggins (1989) allows every individual to have a different probability of capture, modeled by observed covariates. This is the method used in this paper.

In particular, this section reviews the capture-recapture model developed in Alho (1990) and Huggins (1989), which assumes that two independent registries are available. For this application, the first is the JAS and the second is the U.S. Census of Agriculture. Furthermore, it is assumed no farms are opened or go out of business between the two surveys (closed population assumption). Finally one more technical assumption is required. Essentially, it is assumed that all operations have a capture probability that is not too small. See Alho (1990) for theoretical details on this assumption. Then an unbiased estimator for the number N of U.S. farms is

$$\hat{N} = \sum_{i \in \mathcal{J} \cup \mathcal{C}} \frac{1}{\phi_i},$$

where $\mathcal{J} \cup \mathcal{C}$ is the event that an individual farm is captured by the JAS or the Census (or both), and ϕ_i is the probability that the farm is captured by at least one of the surveys. If

we further assume that the surveys are independent, then

$$\phi_i = p_{\mathcal{J},i} + p_{\mathcal{C},i} - p_{\mathcal{J},i}p_{\mathcal{C},i},$$

where $p_{\mathcal{C},i}$ is the probability that farm i is captured by the Census, and $p_{\mathcal{J},i}$ is the probability that farm i is captured by the JAS. NASS adjusts this latter probability by the tract-to-farm ratio, which is computed as the acres in the tract covered by the farm i divided by the total acres of the farm i. This adjustment is performed to remove the undercount bias in the final estimates.

For each individual farm i, the capture history follows a multinomial distribution

$$(u_{i\mathcal{J}}, u_{i\mathcal{C}}, u_{i\mathcal{JC}}, u_{i0}) \sim \mathsf{Mult}\left(p_{\mathcal{J},i}(1-p_{\mathcal{C},i}), p_{\mathcal{C},i}(1-p_{\mathcal{J},i}), p_{\mathcal{J},i}p_{\mathcal{C},i}, 1-p_{\mathcal{J},i}-p_{\mathcal{C},i}+p_{\mathcal{J},i}p_{\mathcal{C},i}\right).$$

In practice, the individuals who are not captured (event $u_{i0} = 1$) are unobserved. This can be resolved by working with the conditional likelihood, which is easily shown to be

$$L = \prod_{i=1}^{N_{\text{cap}}} [p_{\mathcal{J},i}(1-p_{\mathcal{C},i})]^{u_{i\mathcal{J}}} [p_{\mathcal{C},i}(1-p_{\mathcal{J},i})]^{u_{i\mathcal{C}}} [p_{\mathcal{J},i}p_{\mathcal{C},i}]^{u_{i\mathcal{JC}}} \phi_i^{-1} \tag{1}$$

where $\phi_i = p_{\mathcal{J},i} + p_{\mathcal{C},i} - p_{\mathcal{J},i}p_{\mathcal{C},i}$, and N_{cap} denotes the total number of captured farms.

Let x_i be a vector of observed covariates for each individual. As in Alho (1990), it is assumed that the probability of capture by the Census is a logistic function of the covariates, i.e.

$$p_{\mathcal{C},i} = [1 + \exp(-\beta^\top x_i)]^{-1},$$

where β is a vector of coefficients to estimate. If the same were possible with the JAS probabilities, then the model would be exactly the same as in Alho (1990), and we could obtain the estimated ϕ_i's via standard likelihood maximization. Doing so would automatically guarantee a loss of accuracy since the JAS probabilities should be known constants obtained from the area frame design. As mentioned previously, obtaining these constants is not feasible due to excessive operational costs. The next two sections provide methods for dealing with the unknown $p_{\mathcal{J},i}$'s.

3. Overview of the JAS Design

The actual methodology of the JAS design is quite complicated and can be found in Abreu et al. (2010). A brief overview of the relevant portions pertaining to obtaining the sampling probabilities $p_{\mathcal{J},i}$ is provided below. In essence, auxiliary information is used to divide the entire landmass of the U.S. into primary sampling units (PSUs). Each PSU is given a known probability of selection. In addition, each PSU will be divided into a pre-specified number of segments N_{PSU}. Note that the segments themselves are not yet drawn, only the number is specified. A stratified random sample of PSUs is drawn. The sampled PSUs are then divided into segments, and one segment is randomly selected from each PSU. Interviewers are sent to the sampled segments to obtain the required data. From this methodology, an exact representation of $p_{\mathcal{J},i}$ is theoretically obtainable. The probability that the i-th farm is selected by the JAS is simply the sum of the selection probabilities of the JAS segments in which it has land.

While simple, this quantity is not obtainable for two reasons. The first is that segments are only drawn once a PSU has been selected in the JAS sample. Thus any Census record that is not in a JAS PSU will not have any segment information. The second is that detailed geographic knowledge of the farms is not available unless a part of a farm is in at least a segment, so it is cumbersome and complex to determine how the farm's area is divided

between different segments and PSUs. If a farm is only captured by the Census, the farm's address and total acreage is the only known geographical information collected. Farms captured by the JAS also have the segment where they were sampled and identified (but not other unsampled segments in which they also have land). In the next section, some strategies to estimate the $p_{\mathcal{J},i}$'s given available information are suggested.

4. Model for the JAS Probabilities

To estimate the probability of capture by at least one survey, we must first obtain a model-based estimate for the JAS probabilities for the approximately 63% of the JAS farms, whose tract-to-farm ratio is less than one. The first possibility is to simply model them as functions of all variables, in the same manner as the Census probabilities, i.e.

$$p_{\mathcal{J},i} = [1 + \exp(-\alpha^\top x_i)]^{-1},$$

where α is a vector of parameters to estimate. Doing this, however, completely ignores the JAS design. Since the JAS strata are created with respect to a relatively small set of aggregate level variables, such as cultivation level, urban/suburban, special crops, etc. (Abreu et al., 2010), it is unclear how accurate a model using farm-level variables would be. Also, the potential for noise is large as many farm-level variables are probably irrelevant to the JAS design.

Another approach is to try using only the variables deemed relevant to the JAS design. To do this, we make the assumption that farms do not cross strata. While this assumption is certainly not perfect, the hope is that it is a good approximation. The assumption is helped by the fact that many of the largest farms are considered "must cases". This means that responses for them are obtained with probability one and hence they do not enter the model (NASS, 2014). We also assume that a farm's address is identified correctly so that it can be associated with a JAS stratum.

Since PSUs (and segments given PSU) are selected at random from strata, a farm's probability of selection given its JAS stratum should only depend on its size. Thus a model for the JAS selection probabilities can be adequately formulated as

$$P_{\mathcal{J},i} = \frac{e^{v_i}}{1 + e^{v_i}}, \text{ where}$$

$$v_i = \alpha_0 + \sum_{j=1}^{S-1} (\alpha_j I_{ij}) + \alpha_S \frac{A(f_i)}{A(str_i)}.$$

Here the index j sums over all strata, I_{ij} is an indicator for farm i being in stratum j, $A(f_i)$ is the area of farm i, and $A(str_i)$ is the area of farm i's stratum. We then plug $P_{\mathcal{J},i}$ into the conditional likelihood and proceed as in (Alho, 1990) to get the point estimates of the coefficients α_j.

The advantage of this model is that it uses all of the available data as in (1), where the model in the next section will not. Confidence intervals can be obtained using the parametric bootstrap detailed in Zwane and Van der Heijden (2003) and Norris and Pollock (1996). The major disadvantage is model misspecification. If the assumptions are not approximately correct, the model will be biased. For example, it is certainly possible that smaller farms happen to sit on or near strata borders. The model would then produce capture probabilities that are too low.

5. Conditional model

Suppose that instead of conditioning on capture by either the Census or JAS as in (1), we condition only on capture by the JAS. Then it is easily shown that the resulting conditional likelihood is of the form

$$L = \prod_{i=1}^{N_{\mathcal{J}}} [p_{\mathcal{C},i}]^{u_{i\mathcal{J}\mathcal{C}}} [1 - p_{\mathcal{C},i}]^{1-u_{i\mathcal{J}\mathcal{C}}},$$

where $N_{\mathcal{J}}$ is the total number of farms collected by the JAS. Here the product is taken over all JAS samples. Notice that the marginalization causes the JAS capture probabilities to "cancel out", which removes the need to approximate them. In addition, the likelihood is equivalent to a simple logistic regression model of the event that the individual is captured by both surveys. Using this methodology leads to the following algorithm. First, estimate the model parameters β by regressing the JAS data against the indicator $u_{i\mathcal{J}\mathcal{C}}$. Then, weight every farm in the Census by its probability of capture, $\widehat{p_{\mathcal{C},i}}$, predicted using the $\widehat{\beta}$ from the first step.

There are two drawbacks to this method. The first and most obvious is that conditioning on the capture by the JAS excludes the data captured only by the CML; in fact, the excluded data could improve the estimates $\widehat{\beta}$. Considering that the amount of data obtained from the Census is typically much greater than that obtained by the JAS, a substantial loss in efficiency is possible. The second problem lies in variance estimation. The extra round of conditioning precludes the use of an unconditional, non-asymptotic confidence interval for the population totals. Since we only know the probabilities of being captured by both surveys or the Census, the complete pseudo capture histories required for the parametric bootstrap cannot be generated by this model. Instead, an interval may be obtained following a procedure similar to the original steps of Alho (1990). This is an asymptotic estimator, and symmetric confidence intervals are often not appropriate for capture-recapture data (Yip et al., 1995). Another possibility is to use the non-parametric bootstrap detailed in Zwane and Van der Heijden (2003) and Norris and Pollock (1996) to obtain an interval. A limitation of this method is that it only estimates the variance conditional on capture, which will be smaller than the full variance (Norris and Pollock, 1996; Tilling and Sterne, 1999). We take the latter approach here, as many variables are collected for each survey. This means the assumption for using asymptotics (samples greatly outnumbering variables) is unlikely to be the case, which means the bootstrap is likely more appropriate. In the next section, we evaluate the methods discussed with a simulation study.

6. Simulation Study

To evaluate the methods discussed in this paper, we conducted a simulation study. The goal was to make the setup of the simulated area sample close to the actual JAS sample while only using publicly available data. The USDA has publicly available information on its area frame strata (which are used in the JAS) for some states. In addition, numbers of farms with various attributes are available at the county level from the 2012 Census of Agriculture. In our case, the attribute of interest is the size of the farm. The goal was to find counties that were entirely, or almost entirely, composed of a single stratum. The combination of these two data sets would allow us to roughly estimate the number of farms of a given size per square mile in each stratum. With these data, we then chose a county with multiple strata and "impute" the number of farms in each stratum using the farms per area of the donor counties and the area of the receiver strata. The number of farms in this "imputed" county would be the objective of the study.

The donor counties chosen were West Carroll, Jackson, and East Felicia parishes in Louisiana. West Carroll represented stratum 13 (more than 50 percent cultivated), and Jackson represented stratum 40 (less than 15 percent cultivated). Each of these counties was almost entirely covered by the given strata, so the number of farms of each size per area could be estimated by dividing the area of the county by the Census number of farms. East Felicia had the previous two strata as well as stratum 20 (between 15 and 20 percent cultivated). The unknown number of farms per area of stratum 20 in this county was solved for using the previous two estimates for 13 and 40, along with a rough measurement of the area in the county covered by stratum 20. The receiver county was Morehouse Parish, which had strata in 13, 20, 40, and 31 (urban). For simplicity we took the number of farms in stratum 31 to be zero. A diagram of the process is shown in Figure 1.

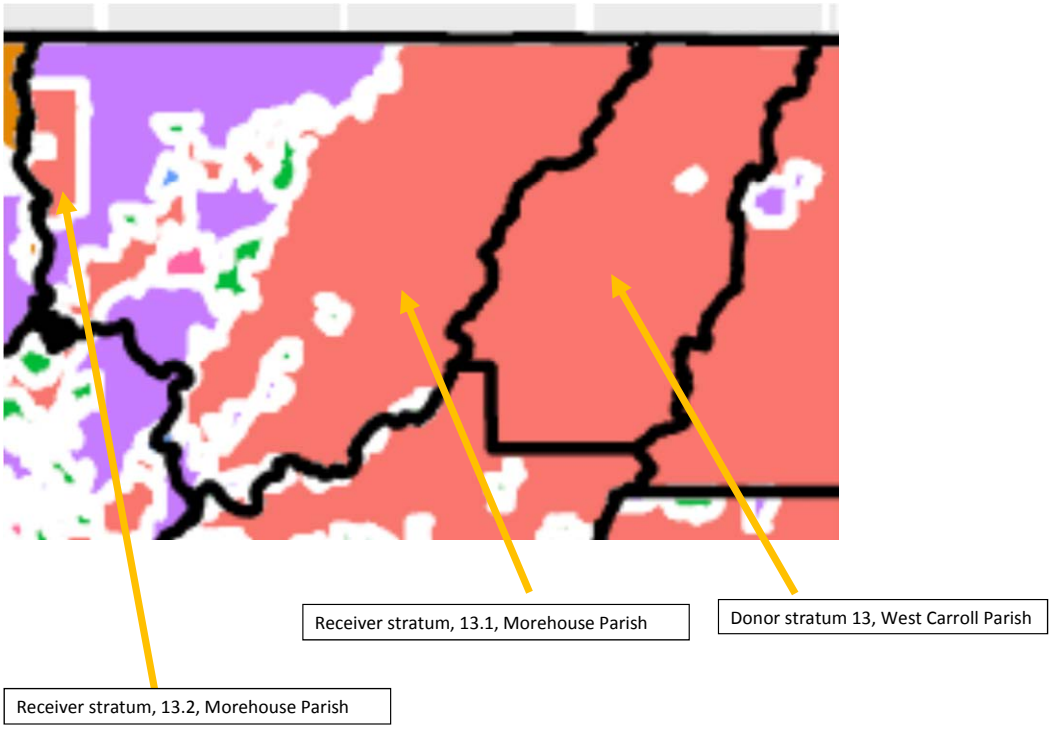

Figure 1: Process for calculation of farm numbers per strata

For example, the area of West Carroll parish (see Figure 1) is 360 square miles. It is almost entirely composed of stratum 13. From the 2012 Census, there are 111 farms with size between 70 and 99 acres in this county. So, we may estimate the number of farms between 70 and 99 acres per square mile in stratum 13 to be $\frac{111}{360}$. In the receiver Morehouse parish, stratum 13 occurs in multiple groups, examples in the figure are 13.1 and 13.2. The number of farms between 70 and 99 acres per square mile in 13.1 is then $\frac{111}{360} \times 392$, where the area of stratum 13.1 is estimated to be 392 square miles.

Using this process and an approximation of Morehouse parish discretized into cells of one square mile each, the input parameters to our simulation are exposed in Table 1.

An image of the cell representation of Morehouse parish is given in Figure 2. Note that there are 907 farms in the population. For each iteration of the simulation we do the following. First, for each farm, generate its size uniformly given its size bounds. For example, if a farm is between one and nine acres, its size is a draw from $\mathsf{Unif}(1, 9)$, converted to square miles. Next, each farm is represented as a circle with area equal to its size. Its center is then randomly placed inside its stratum and group. More specifically, suppose for example a farm is in stratum 13, group 13.1 (see Figure 1 above). Then a cell in group 13.1 is ran-

Table 1: Number of farms of various sizes for each strata

Strata	1-9 Acres	10-49 Acres	50-69 Acres	70-99 Acres
13	23	96	117	140
20	6	4	2	2
40	6	46	10	7

Strata	100-139 Acres	140-179 Acres	180-219 Acres	220-259 Acres
13	112	70	37	39
20	2	0	0	0
40	14	8	8	1

Strata	260-499 Acres	500-999 Acres	1000-1999 Acres	2000-3000 Acres
13	60	39	27	23
20	2	2	0	0
40	4	1	0	0

domly selected, and the center is placed in that cell. Furthermore, the (x, y)-coordinates of the center in the cell are selected at uniform. Once all the farms have been placed, the JAS sample is created by randomly selecting 50, 5, and 20 segments from strata 13, 20, and 40, respectively. Any farm that lies in any of these segments is selected in the sample. Note that under this setup, it is indeed possible for farm boundaries to cross segments (cells) or even strata (groups of cells). Next, five normal random variables x_2, x_3, x_4, x_5, u are generated. In addition x_1, the log of the farm's size is also used. Individuals are then drawn with probability $p_{\mathcal{C},i} = \frac{\exp(5+1.5x_1+x_2+x_3+x_4+x_5+u)}{1+\exp(5+1.5x_1+x_2+x_3+x_4+x_5+u)}$ from the population to be in the Census sample. Each of the 3 models discussed is run on the simulated data to obtain point estimates and confidence intervals (1000 bootstrap replications). The simulation was run 399 times (due to the high computational time). The results are tabulated below.

Table 2: Simulation results with JAS conditional outliers

Model	Avg. Bias	M.S.E.	Coverage	Mean C.I. Width
All Variables	44	5920	91.48	296
JAS Specific Variables	-13	1852	91.98	167
Conditional	229	9178208	92.73	451631.8

Table 3: Simulation results without JAS conditional outliers

Model	Avg. Bias	M.S.E.	Coverage	Mean C.I. Width
All Variables	44	5920	91.48	296
JAS Specific Variables	-13	1852	91.98	167
Conditional	60	24328	93.2	231020

It is apparent that the conditional model does the worst, especially in terms of mean square error. This is to be expected, as the sample size for this model is much less than that of the other two, as was mentioned before. This likely caused instability leading to the large bias and mean squared error shown in Table 2. It was found that, out of the 399 simulations, there were two outlier point estimates with magnitudes of 8,478 and 60,866. Table 3 gives the summary with the outliers removed. Regardless, even with the outliers removed it performs the worst in terms of bias and MSE, although it does have the closest to desired confidence interval coverage. The mean confidence interval width is still influenced by 32 intervals with upper bound outliers larger than 10,000, caused by the same instability in the parametric bootstrap, which explains there abnormally large width. With these additional

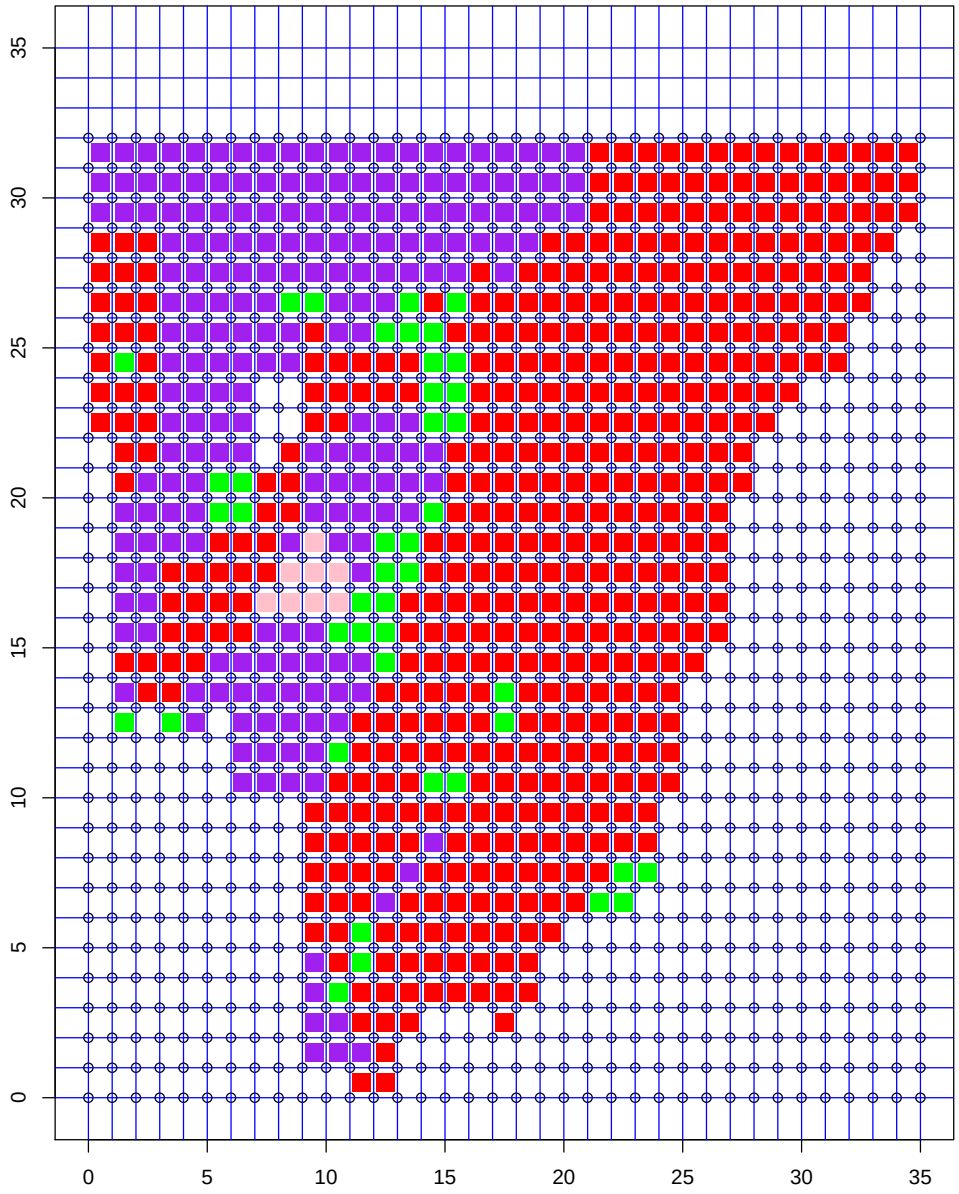

Figure 2: Morehouse parish

outliers removed, the mean confidence interval width becomes 1541, still far larger than the other two.

Also unsurprisingly, the model using only JAS relevant variables to estimate the JAS selection probabilities performed better than the one using the same variables as the Census. It has the lowest mean square error, average confidence interval width, and average bias. It is only outperformed in coverage, where the conditional model is closer to the nominal 95% level. This suggests that it is potentially possible to approximate the JAS design with a model and still obtain meaningful estimates.

While all confidence intervals have below the desired coverage, it should be mentioned

that this is a well known property of the naive bootstrap (Shi, 1992). While the parametric bootstrap was chosen to save computational time, a more advanced (but slower) method, such as the double bootstrap or bootstrap t, may be used in a real application where multiple simulations do not need to be run (Shi, 1992; Hall, 1988).

7. Conclusion and Further Research

This study shows promising signs of employing model-based approximations of survey design probabilities for our capture-recapture application. It demonstrates that the increased sample size allowed by modeling the JAS probabilities allows for a significant increase in model performance when compared to using an asymptotically unbiased model with a smaller sample size (and the potential instability that the smaller sample size yields). This is particularly important as the real capture-recapture procedure for the Census of Agriculture also conditions on capture by the JAS. Considering that the JAS county-level sample sizes are much smaller than in this simulation study and that the JAS sample is less than a tenth of that of the Census sample, the importance of having as much data as possible is even greater. The model-based approximation would allow use of the entire data set.

While we tried to make the simulation study as realistic as possible, using only publicly available data (to allow for the release of this paper) limited how closely we could approximate actual stratum level farm numbers. The detail-oriented reader may have noticed that the actual number of farms in Morehouse parish published in the 2012 Census is only half the imputed number used in this study. While having twice as many farms is actually conservative in the sense of our objective (more farms means it is more likely segment and strata boundaries are crossed), more accurate inputs to the study can be obtained using (non-publicly available) individual level data. Seeing if this changes the results of the study will be a focus of future research. In addition, the effect of the tract-to-farm ratio on the estimates needs to be further investigated.

It should be noted that the existing capture-recapture methodology that NASS uses to estimate the number of farms takes into account misclassification of an operation's farm status, as well as the fact that Census capture is a two-stage process (NASS, 2014). In addition, other models to estimate the number of farms are also in development. Since we were concerned only about the effect of estimating the JAS probabilities of selection on the bias of a capture-recapture estimator, we used a simpler model assuming no misclassification and treating the Census as a registry, e.g. the model developed in (Alho, 1990). Incorporating these ideas into models that more fully capture the characteristics of the Census of Agriculture is an area of future research.

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.

Amstrup, S. C., McDonald, T. L., and Manly, B. F. (2010). *Handbook of capture-recapture analysis*. Princeton University Press.

Cochran, W. G. (1978). Laplace's ratio estimator. *Contributions to survey sampling and applied statistics (New York, 1978)*, pages 3–10.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pages 927–953.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

NASS (2014). Appendix A – Census of Agriculture Methodology. In *2012 Census Full Report*. United States Department of Agriculture, Washington D.C.

Norris, J. L. and Pollock, K. H. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, 3(3):235–244.

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115.

Shi, S. G. (1992). Accurate and efficient double-bootstrap confidence limit method. *Computational statistics & data analysis*, 13(1):21–32.

Tilling, K. and Sterne, J. A. (1999). Capture-recapture models including covariate effects. *American journal of epidemiology*, 149(4):392–400.

Yip, P., Bruno, G., Tajima, N., Seber, G., Buckland, S., Cormack, R., Unwin, N., Chang, Y., Fienberg, S., Junker, B., et al. (1995). Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology*.

Zwane, E. and Van der Heijden, P. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & probability letters*, 65(2):121–125.