Restricted Multinomial Regression for a Triple-System Estimation with List Dependence

Luca Sartore* Habtamu Benecha† Kelly Toppin† Clifford Spiegelman‡

Abstract

The National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture every five years. In 2012, NASS began using a capture-recapture approach to adjust the Census estimates for under-coverage, non-response, and misclassification. This requires two independent samples. NASS has kept its Census Mailing List (CML) independent from its area frame, which is used for the June Area Survey (JAS) every June. NASS is exploring the use of web-scraping to develop a third list-frame (TL) that would be independent of the CML and the area frame. In this paper, a Triple-System Estimation (TSE) methodology based on regularized multinomial regression is proposed to investigate for possible dependence between the CML and the TF. A simulation study is performed to compare the performance of the estimator based on the proposed methodology, which can take into account the frame dependence with others already presented in the literature.

Key Words: Triple-System, Estimation, Weights, Capture, Probability, Dependence

1. Introduction

Every five years, the National Agricultural Statistics Service (NASS) conducts the United States (US) Census of Agriculture. A capture-recapture approach has been used since 2012 to adjust the Census estimates for under-coverage, non-response, and misclassification. Capture-recapture methodology also known as Dual-System Estimation (DSE) uses two independent surveys to estimate the unknown size of the finite population. The US Census of Agriculture uses the Census Mailing List (CML), a list of all known agricultural operations with the potential of at least $1,000 in sales of agriculture products (O'Donoghue et al., 2009). The June Area survey (JAS), which is based on the NASS area frame, is used as a second survey for the capture-recapture methodology.

NASS is exploring the use of web-scraping technology to develop a third independent list-frame (TL). This concept has the potential to become a standard operating procedure in the future (National Academies of Sciences, Engineering, and Medicine, 2017). Its use requires a more sophisticated methodology than the DSE currently in use.

The advantage of a Triple-System Estimation (TSE) is twofold. This method will be used to provide more accurate estimates of the number of non-captured farms, and it allows for testing the assumption of list-dependence. The resulting "correlation bias" is usually associated with list-dependence and heterogeneity among farms (Chao and Tsay, 1998). The method proposed in this paper can adjust the estimates by modeling both the heterogeneity and possible dependencies.

The problem of census under-count due to under-coverage, non-response and misclassification is not new in the literature. Several solutions are already well-established, and most of them refer to capture-recapture methodology based on two lists. However, only a few articles consider the problem from a different perspective. Darroch et al. (1993) extended

*National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006, lsartore@niss.org

†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

the DSE methodology and introduced a TSE approach that accounts for varying capture probabilities of each sample unit. Multiple-System Estimation was presented by Zaslavsky (1989), and Chao and Tsay (1998) proposed an estimator for the size of a finite population that takes into account both list-dependence and heterogeneity. Griffin (2014) provided a brief investigation of several TSE estimators for potential applications with administrative records.

The proposed new method provides a unified methodology accounting for misclassification, under-coverage, and non-response of the sample units. The estimation of the population totals is performed concurrently with variable selection via regularized multinomial regression.

In section 2, a population partitioning method is introduced. After the population is partitioned, standard multiple-system estimation techniques can be applied to sub-groups of interest. A TSE methodology that takes into account possible dependencies between two surveys is introduced in section 3. A simulation study is performed to assess the performances of the proposed estimator, and its results are discussed in section 4. Final conclusion and remarks are given in section 5.

2. Identifying sub-populations via penalized EM algorithm

NASS collects the data for its Census and other surveys by targeting farms; however, data from other agricultural operations that do not satisfy the definition of farm are also collected. Units belonging to the group of non-farms are removed from the samples before performing any further analysis in order to produce more consistent estimates. The challenge is to correctly separate farms from non-farms.

Several methods have been considered to identify the farms based on the reported data. The most effective consists in classifying the agricultural operations based on the threshold given by the farm definition, however this cannot be done when the data related to the production are reported as revenue intervals/categories instead of a point number. Discriminant analysis is applied in particular to determine the farm status of the small agricultural operations. This involves the estimation of a predictive model that provides an objective framework to select the sample units with the highest probability of being a farm.

Since farm status is unknown, its estimation is also affected by uncertainty. If the farm status is considered as a binary latent variable U, a more sophisticated formulation than a logistic regression model is needed. In addition, when the data of a farm are collected by multiple surveys, there are some inconsistencies in classification. Farm indicator variables are essential for the estimation of the probability of observing a farm. An evaluation of the farm status can be achieved by Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

This algorithm estimates the probability of being a farm given the observed realization of a binomial random variable $F|U$. The algorithm starts by assigning random binary values to the latent variable U_i with probabilities f_i/m_i, where $m_i > 0$ represents the total number of surveys capturing the i-th sampled agricultural operation, and f_i is the observed outcome of

$$F_i|U_i \sim \mathsf{Bin}\left[m_i,\ \left\{1+\exp(-\mathbf{x}_i^\top \boldsymbol{\xi}_{U_i})\right\}^{-1}\right],$$

where the vector $\mathbf{x}_i$ is a set of covariates for the i-th sample unit, and $\boldsymbol{\xi}_{U_i}$ denotes the vectors $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_1$ of p parameters to estimate.

The EM algorithm successively iterates the following steps until convergence:

M-step: Fits the parameters of the logistic models for the two sub-populations. The estimation of the parameters is performed via regularized logistic regression (Friedman

et al., 2010), with model selection achieved through a Least Absolute Shrinkage and Selection Operator (LASSO) penalty added to the negative log-likelihood (Tibshirani, 1996). That is:

$$Q(\boldsymbol{\xi}_0, \boldsymbol{\xi}_1 | U_1, \ldots, U_n) = \sum_{i=1}^{n} \left[m_i \log \left\{ 1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\xi}_{U_i} \right) \right\} - f_i \mathbf{x}_i^\top \boldsymbol{\xi}_{U_i} \right] + \sum_{i=1}^{p} \left(\lambda_0 |\xi_{0,i}| + \lambda_1 |\xi_{1,i}| \right),$$

where n corresponds to the number of unique sampled operations, $\xi_{0,i}$ and $\xi_{1,i}$ respectively represent the i-th component of the vectors $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_1$, and the scalars λ_0 and λ_1 control the tradeoff of the LASSO penalties.

E-step: Calculates the binary values of U_i, for all $i = 1, \ldots, n$, such that

$$U_i = \begin{cases} 0, & \text{if } \mathbf{x}_i^\top (\hat{\boldsymbol{\xi}}_1 - \hat{\boldsymbol{\xi}}_0) < 0, \\ u_i, & \text{if } \mathbf{x}_i^\top (\hat{\boldsymbol{\xi}}_1 - \hat{\boldsymbol{\xi}}_0) = 0, \\ 1, & \text{otherwise,} \end{cases}$$

where u_i is drawn from a $\mathsf{Bernoulli}(0.5)$. This result provides an optimal solution that minimizes the function

$$Q(U_1, \ldots, U_n | \hat{\boldsymbol{\xi}}_0, \hat{\boldsymbol{\xi}}_1) = \sum_{i=1}^{n} \left[m_i \log \left\{ 1 + \exp\left(\mathbf{x}_i^\top \hat{\boldsymbol{\xi}}_{U_i} \right) \right\} - f_i \mathbf{x}_i^\top \hat{\boldsymbol{\xi}}_{U_i} \right]$$

with respect to $U_1, \ldots, U_n$.

For simplicity, the sub-population identified via the EM algorithm is momentarily assumed to be homogeneous, i.e. when each observation has the same probability of being collected in a survey. This assumption will be later removed to extend the result. In addition, the probability of belonging to a specific sub-population group is assumed to be independent on the surveys.

NASS is currently using all data in the CML to produce estimate of the total number of farms. The technique adopted adjusts the survey weights by a misclassification factor computed as the probability of having a farm given that the unit i-th is in the CML. This probability can be replaced with $\Pr(U_i | F_i)$, and the weights will be approximated as $w_i^* = \Pr(U_i | F_i) \Pr(C_i, R_i | U_i = 1)^{-1}$, where C_i represents the coverage of the Census, R_i denotes the response to the questionnaire, and U_i the farm status of the observation i.

3. Models for heterogeneous catchability

Another alternative to the standard approach adopted by NASS to deal with the misclassification consists in estimating the total number of units belonging to a sub-population only. To model heterogeneous probabilities of capturing a farm in the CML, the units identified as farms by the discriminant analysis are kept while those identified as non-farms are excluded from the estimation process.

When the non-farms are excluded, the design weights are computed as

$$w_i^* = \Pr(C_i, R_i | U_i = 1)^{-1}.$$

NASS separately estimates for each farm a set of probabilities to compute the design weights as $w_i^* = \Pr(C_i, R_i | U_i = 1)^{-1} = \Pr(R_i | C_i, U_i = 1)^{-1} \Pr(C_i | U_i = 1)^{-1}$. The proposed approach differs by estimating simultaneously all the parameters of a multinomial regression model. This technique associates a capture probability to each farm such that

the heterogeneity among the sample units is taken into account through variations in the covariates. By assuming homogeneous non-response, all the farms selected by either the CML, the JAS, or the TL can respond to the surveys without loss of generality with constant probability. In this paper, it is assumed that all farms are responding to the questionnaires, therefore the probability $\Pr(C_i, R_i|U_i = 1) = \Pr(C_i|R_i, U_i = 1)$ is modeled to produce the estimates of the total number of farms.

Let the variables c (for the CML), j (for the JAS), and t (for the TL) indicate which survey captured the farm i. Any farm can be sampled by the CML only with probability η_{100}, by the JAS only with probability η_{010}, by the TL only with probability η_{001}, or by the CML and JAS with probability η_{110}, by the CML and TL with probability η_{101}, by the JAS and TL with probability η_{011}, or all three lists with probability η_{111}. The sum of these probabilities is used to compute the capture probability of the three lists. It is also possible that a farm is not capture by any of three lists with probability η_{000} (see Table 1 for a summary of the probabilities η_{cjt}).

Table 1: Capture probabilities.

	$\mathbf{t_i = 0}$		$\mathbf{t_i = 1}$	
	$\mathbf{c_i = 0}$	$\mathbf{c_i = 1}$	$\mathbf{c_i = 0}$	$\mathbf{c_i = 1}$
$\mathbf{j_i = 0}$	η_{000}	η_{100}	η_{001}	η_{101}
$\mathbf{j_i = 1}$	η_{010}	η_{110}	η_{011}	η_{111}

By matching the observations from the three surveys, it is possible to estimate the probability that a farm is captured by a survey given that the same farm has been captured. Therefore, the probabilities $\theta_{100,i}$, $\theta_{010,i}$, $\theta_{001,i}$, $\theta_{110,i}$, $\theta_{101,i}$, $\theta_{011,i}$, and $\theta_{111,i}$ can be generically formulated from the capture probabilities exposed in Table 1 as

$$\theta_{\text{cjt},i} = \eta_{\text{cjt},i}(1 - \eta_{000,i})^{-1}, \text{ for any } \mathrm{c}_i,\ \mathrm{j}_i,\ \mathrm{t}_i \in \{0, 1\},$$

such that c_i, j_i, and t_i are not all simultaneously zero.

When these seven probabilities are unknown, a link function is assumed to formulate a multinomial regression model, such that

$$\theta_{\text{cjt},i} \propto \exp(-\mathbf{x}_i^\top \boldsymbol{\zeta}_{\text{cjt}}),$$

where $\boldsymbol{\zeta}_{\text{cjt}}$ is a vector of parameters. Once these are estimated by maximizing the regularized likelihood, they are used to estimate the probabilities exposed in Table 1 under the assumption of dependence between two of the three lists. This means that the covariance between two indicator variables is not zero, i.e. $\mathsf{COV}[\mathrm{c}_i, \mathrm{j}_i] \neq 0$, or $\mathsf{COV}[\mathrm{c}_i, \mathrm{t}_i] \neq 0$, or $\mathsf{COV}[\mathrm{j}_i, \mathrm{t}_i] \neq 0$.

For simplicity, let the model for the CML and the JAS allow for dependence and let the TL be independent on the other two. Any other permutation of the three lists produces an estimates for η_{000} which can be used to study the independence of the three lists. Under these assumptions, the following equality is satisfied:

$$\rho_{\cdot\cdot|0,i} = \rho_{\cdot\cdot|1,i}, \tag{1}$$

where $\rho_{\cdot\cdot|b,i}$ is the conditional correlation between c_i and j_i given that $\mathrm{t}_i = b$, with $b \in \{0, 1\}$. These conditional correlations are computed as

$$\rho_{\cdot\cdot|1,i} = (\theta_{001,i}\theta_{111,i} - \theta_{101,i}\theta_{011,i})[(\theta_{101,i} + \theta_{111,i})(\theta_{001,i} + \theta_{011,i})(\theta_{001,i} + \theta_{101,i})(\theta_{011,i} + \theta_{111,i})]^{-1/2}$$
$$\rho_{\cdot\cdot|0,i} = (\nu_i\theta_{110,i} - \theta_{100,i}\theta_{010,i})[(\theta_{100,i} + \theta_{110,i})(\nu_i + \theta_{010,i})(\nu_i + \theta_{100,i})(\theta_{010,i} + \theta_{110,i})]^{-1/2}$$

where ν_i is an expansion factor such that

$$\begin{aligned} \eta_{000,i} &= (1+\nu_i)^{-1}\nu_i, && \text{if } \mathrm{c}_i, \mathrm{j}_i, \mathrm{t}_i \in \{0\}, \text{ and} \\ \eta_{\mathrm{cjt},i} &= (1+\nu_i)^{-1}\theta_{\mathrm{cjt},i}, && \text{otherwise.} \end{aligned}$$

After some algebra, it is possible to obtain the following equation from (1),

$$\rho_{\cdot\cdot|1,i}\sqrt{a_i\nu_i^2 + b_i\nu_i + c_i} = d_i\nu_i - e_i,$$

where

$$\begin{aligned} a_i &= (\theta_{100,i}+\theta_{110,i})(\theta_{010,i}+\theta_{110,i}), & b_i &= (\theta_{100,i}+\theta_{110,i})(\theta_{010,i}+\theta_{110,i})(\theta_{010,i}+\theta_{100,i}), \\ c_i &= (\theta_{100,i}+\theta_{110,i})(\theta_{010,i}+\theta_{110,i})\theta_{010,i}\theta_{100,i}, & d_i &= \theta_{110,i}, \text{ and} \\ e_i &= \theta_{010,i}\theta_{100,i}. \end{aligned}$$

The solution calculated for ν_i is given as

$$\nu_i = -\frac{\rho^2_{\cdot\cdot|1,i}b_i + 2e_id_i + \text{sign}(\rho_{\cdot\cdot|1,i})\sqrt{(\rho^2_{\cdot\cdot|1,i}b_i + 2e_id_i)^2 - 4(\rho^2_{\cdot\cdot|1,i}c_i - e_i^2)(\rho^2_{\cdot\cdot|1,i}a_i - d_i^2)}}{2(\rho^2_{\cdot\cdot|1,i}a_i - d_i^2)},$$

but this formulation can produce negative values which will be truncated to zero to have only non-negative values as output. Since $\nu_i = \eta_{000,i}(1-\eta_{000,i})^{-1}$, the non-capture probability can be estimated as

$$\hat{\eta}_{000,i} = \hat{\nu}_i\,(1+\hat{\nu}_i)^{-1},$$

therefore the total number of farms can be computed as

$$\hat{N} = \sum_{i\in\mathcal{C}} \frac{1}{1-\hat{\eta}_{000,i}} = \sum_{i\in\mathcal{C}}(1+\hat{\nu}_i),$$

where $\mathcal{C}$ denotes the set of indexes i for the farms captured at least once, i.e. when the inequality $\mathrm{c}_i + \mathrm{j}_i + \mathrm{t}_i > 0$ is satisfied.

4. Simulation study

Simulations are performed under two scenarios to study the performance of the proposed method in the estimation of farm numbers. While farm status of operations is assumed to be known under the first set of simulations, farm status is unknown for the second set of simulations.

A total of 8 simulation studies with population sizes ranging from 2000 to 16000 units are performed under the two scenarios. All the units in the population are assumed to be farms in the first scenario. Under the second scenario, the units of the population are partitioned into farms and non-farms. Each farm in the population has 20 independent and identically distributed covariates x simulated as $\mathsf{N}(0,1)$. The capture history for the CML and JAS is built by simulating four binary values from a multinomial distribution with probabilities

$$\begin{aligned} \Pr\left(\mathrm{c}_i = 1 \cap \mathrm{j}_i = 0\right) &\propto \exp\left(\sum_{k=1}^{20} x_{ik}\beta_{k,1}\right), \\ \Pr(\mathrm{c}_i = 0 \cap \mathrm{j}_i = 1) &\propto \exp\left(\sum_{k=1}^{20} x_{ik}\beta_{k,2}\right), \\ \Pr(\mathrm{c}_i = 1 \cap \mathrm{j}_i = 1) &\propto \exp\left(\sum_{k=1}^{20} x_{ik}\beta_{k,3}\right), \\ \Pr(\mathrm{c}_i = 0 \cap \mathrm{j}_i = 0) &\propto 1, \end{aligned}$$

where $\beta_{k,j} = 0.3 Z_{\mathsf{N}} Z_{\mathsf{B}}$, for any $j = 1, \ldots, 3$ and $k = 1, \ldots, 20$, where the value Z_{N} is drawn from a $\mathsf{N}(0, 1)$ and Z_{B} from a $\mathsf{Bernoulli}(0.7)$. The sample units of the TL are randomly selected with probability

$$\Pr(\mathrm{t}_i = 1) = \left\{ 1 + \exp\left(- \sum_{k=1}^{20} x_{ik} \gamma_k \right) \right\}^{-1},$$

where $\gamma_k = 0.6 Z_{\mathsf{N}} Z_{\mathsf{B}}$, for any $k = 1, \ldots, 20$.

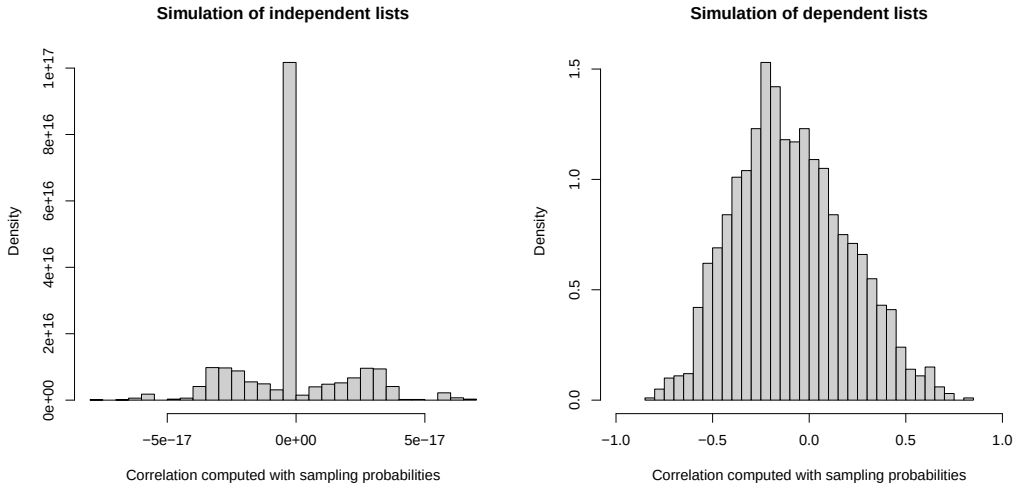

Figure 1: Distribution of correlation between CML and TL (on the left). Distribution of correlation between CML and JAS (on the right).

In each simulation setting, the sample selections of the three surveys are replicated 100 times with a fixed simulated population. The sampling scheme described above allows for list-dependence between the CML and JAS, and maintains TL independent from CML and JAS. In fact, the joint probability $\Pr(\mathrm{c}_i \cap \mathrm{j}_i)$ cannot be obtained as the product of the marginals, while this is possible for both $\Pr(\mathrm{c}_i \cap \mathrm{t}_i)$ and $\Pr(\mathrm{j}_i \cap \mathrm{t}_i)$ since TL is independent by design. Figure 1 shows the empirical distribution of $\mathsf{COR}(\mathrm{c}_i, \mathrm{t}_i)$ on the left and the distribution of $\mathsf{COR}(\mathrm{c}_i, \mathrm{j}_i)$ on the right. These correlations are obtained with the final probabilities used for the simulations of a population with size $N = 2000$. E.g. the correlation between c_i and t_i is computed as

$$\mathsf{COR}(\mathrm{c}_i, \mathrm{t}_i) = \frac{\eta_{00\mathrm{t},i}\eta_{11\mathrm{t},i} - \eta_{10\mathrm{t},i}\eta_{01\mathrm{t},i}}{\sqrt{(\eta_{10\mathrm{t},i} + \eta_{11\mathrm{t},i})(\eta_{00\mathrm{t},i} + \eta_{01\mathrm{t},i})(\eta_{00\mathrm{t},i} + \eta_{10\mathrm{t},i})(\eta_{01\mathrm{t},i} + \eta_{11\mathrm{t},i})}},$$

for any $i = 1, \ldots, 2000$ and $\mathrm{t}_i \in \{0, 1\}$.

Table 2 and 3 show the true number of farms, the average of the model estimated number of farms, the standard errors, the relative bias for the entire estimation process. The relative bias is chosen as a criterion to evaluate the accuracy of the proposed estimator. Through the analysis of the estimates obtained via simulations, it is possible to compute the relative bias by computing an optimal adjusting factor φ, which is obtained by minimizing the mean square error with respect to the true value, i.e.

$$\hat{\varphi} = \arg\min_{\varphi} \sum_{k=1}^{100} \left(\hat{N}_k \varphi - N \right)^2 = N \sum_{k=1}^{100} \hat{N}_k \left(\sum_{k=1}^{100} \hat{N}_k^2 \right)^{-1}.$$

Table 2: Results from simulation with known farm status

	Proposed estimator			Estimator for independent lists		
Farms	Average	Std. Err.	Rel. Bias	Average	Std. Err.	Rel. Bias
2000	2102	102	0.051	2100	30	0.048
4000	4153	203	0.039	4196	48	0.047
6000	6179	113	0.029	6323	53	0.051
8000	8179	109	0.022	8419	64	0.050
10000	10204	170	0.020	10508	76	0.048
12000	12204	137	0.017	12609	92	0.048
14000	14254	131	0.018	14744	89	0.050
16000	16243	142	0.015	16815	93	0.048

Table 3: Results from simulation with unknown farm status

	Proposed estimator			Estimator for independent lists		
Farms	Average	Std. Err.	Rel. Bias	Average	Std. Err.	Rel. Bias
1240	1363	110	0.096	1326	28	0.066
2454	2627	103	0.067	2648	38	0.073
3738	3917	248	0.050	3968	54	0.058
4870	5076	110	0.041	5193	53	0.062
6224	6527	192	0.047	6680	64	0.068
7357	7612	188	0.034	7846	63	0.062
8550	8932	117	0.043	9205	71	0.071
9876	10182	121	0.030	10531	82	0.062

The relative biases reported in Table 2 and 3 are computed as $\tilde{R} = 1 - \hat{\varphi}$.

As can be seen from Table 2, all the estimates are biased upwards. The averages of the estimated farm numbers that are obtained with the proposed estimator are close to the true values. This happens for all the considered population sizes when the status of farms is assumed to be known. The averages are within 2 standard deviations from the true value, and the relative bias decreases as the population size increases.

The estimator

$$\hat{\nu}_i = \frac{1}{3}\left(\frac{\hat{\theta}_{001,i}\hat{\theta}_{010,i}}{\hat{\theta}_{011,i}} + \frac{\hat{\theta}_{100,i}\hat{\theta}_{010,i}}{\hat{\theta}_{110,i}} + \frac{\hat{\theta}_{001,i}\hat{\theta}_{100,i}}{\hat{\theta}_{101,i}}\right)$$

proposed by Chao and Tsay (1998) can be used for a comparison. It provides consistent results when the three lists are considered to be independent. In this case, this estimator does not perform as well due to list-dependence. In fact, the average of the estimates is not within 2 standard deviations from the true value, and the relative bias does not improve as the population size increases.

When the population has two groups and the farm status is assumed to be unknown, the two estimators mostly behave as in the previous scenario (see Table 3). Even if similar features are evident, the effect of the farm status uncertainty is consistent with the increments on the relative biases. The standard errors are also relatively larger than the first scenario.

5. Conclusion

Estimating the total number of farms plays a central role in NASS. Several challenges in producing more precise estimates lead to innovative solutions to integrate several source

of information. This information can be used to develop more accurate DSE weights. The study of any spurious dependence between two surveys is allowed by the introduction of a third survey. Under certain assumptions, it is possible to estimate the degree of dependence between two surveys. The developed model allows for heterogeneity so that each farm has its own sampling distribution (which allows for many forms of dependence). Two simulation scenarios are performed to study the bias of the proposed estimator. A general tendency to over-estimate the population totals was observed, and the standard error of the estimates are more stable when the true farm status is known. The proposed estimator is also able to better handle dependent lists. The current proposal can be further improved in the future by exploring the performances with other forms of dependence, introducing non-response adjustments, developing bias reduction techniques, and studying the robustness to model misspecification.

References

Chao, A. and Tsay, P. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association*, 93(441):283–293.

Darroch, J. N., Fienberg, S. E., Glonek, G. F., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88(423):1137–1148.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Griffin, R. A. (2014). Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020. *Journal of Official Statistics*, 30(2):177–189.

National Academies of Sciences, Engineering, and Medicine (2017). *Principles and practices for a federal statistical agency, Sixth Edition*. The National Academies Press, Washington, DC.

O'Donoghue, E., Hoppe, R., Banker, D., Korb, P., et al. (2009). Exploring alternative farm definitions: implications for agricultural statistics and program eligibility. *Economic Information Bulletin-USDA Economic Research Service*, 49.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Zaslavsky, A. M. (1989). Multiple-system method for census coverage evaluation. In *Proceedings: Fifth Annual Research Conference*, pages 657–672. U.S. Department of Commerce, Bureau of the Census.