VERSITA

Journal of Official Statistics, Vol. 30, No. 2, 2014, pp. 177–189, http://dx.doi.org/10.2478/jos-2014-0012

Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020

Richard A. Griffin[1]

Heterogeneity in capture probabilities is known to produce bias in the dual system estimates that have been used to estimate census coverage in U.S. Censuses since 1980. Triple system estimation using an administrative records list as a third source along with the census and coverage measurement survey has the potential to produce estimates with less bias. This is particularly important for hard-to-reach populations.

The article presents potential statistical methods for the estimation of net census undercount using three systems for obtaining population information: (1) a decennial census; (2) an independent enumeration of the population in a sample of block clusters; and (3) administrative records. The 2010 Census Match Study will create census-like files for the entire nation using federal and commercial sources of administrative records. The 2010 Census Coverage Measurement Survey is an enumeration in a sample of block clusters that is independent of the 2010 decennial Census.

Key words: Heterogeneity; independence; log-linear model.

1. Introduction

Heterogeneity in capture probabilities is known to produce bias in the dual system estimates (DSE) which have been used to estimate census coverage in U.S. Censuses since 1980. Triple system estimation using an administrative records list as a third source along with the census and postenumeration survey (PES) has the potential to produce estimates with less bias. This is particularly important for hard-to-reach populations. Based on theory in Bell (1993), the bias in DSE due to causal dependence or heterogeneity in capture probabilities may be greater for hard-to-reach populations. Some of the many references for the theory and practice of Dual System Estimation are Chandrasekar and Deming (1949), Wolter (1986), Alho (1990), and Mulry and Spencer (1991).

For the 2020 Census postenumeration survey, we are carrying out a preliminary investigation on using Triple System Estimation (TSE). The three systems for obtaining population information for TSE are: (1) a decennial census; (2) an independent enumeration of the population in a sample of block clusters; and (3) administrative records.

[1] U.S. Census Bureau, 4210 Southwinds Place Unit 109, White Plains, Maryland 20695, U.S.A. Email: richard.a.griffin@census.gov

Acknowledgments: This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

© Statistics Sweden

For this article, all data are simulated and there is no sampling. When administrative records are mentioned the reader should bear in mind that any real application would use census-like files for the entire nation, using federal and commercial sources of administrative records similar to those created for the 2010 Census Match Study (Rastogi and O'Hara 2012). Similarly, in practice a PES would be an enumeration in a sample of block clusters independent of the census, like the 2010 Census Coverage Measurement Survey.

For this simulation study, it is assumed that all N individuals in the population are exposed to possible inclusion in all three sources. In practice, sampling is necessary for the postenumeration survey and possibly the administrative list (due to the necessity of follow-up for unresolved match status). Table 1 illustrates the eight cells indicating the possible combinations of captured or not captured by each of the three attempts at enumeration. The count of the population total in each cell is defined as N_{jkl} where the subscripts j, k, and l are 1 or 0 to indicate captured or not captured in the Census list, the postenumeration survey, and the administrative list respectively. For example, N_{110} is the count of persons captured by the Census and PES but not captured by the administrative list. All cells are conceptually observable except N_{000}.

Creation of the simulated populations assumes autonomous independence, which means that the Census list, the postenumeration survey list and the administrative list are created as a result of N mutually independent trials from one person to the next (all persons are captured independently of all other persons, even persons in the same household). The counts in Table 1 and all the estimators studied in this article could be constructed even if autonomous independence did not hold. Autonomous dependence could create additional bias in estimates.

The Census Bureau has used dual system estimation for census net error estimation starting with the 1980 Census. The incomplete 2^3 table of counts for triple system estimation can be divided into one complete 2×2 subtable and one incomplete 2×2 subtable. The additional source from administrative records provides data with which to evaluate the previously untestable assumption of independence between the census and the postenumeration survey. Evidence is available in the triple-system tables for odds ratios in 2×2 subtables formed by restricting consideration to cases observed in the administrative records source. In this case, complete information is available for all four cells defined by capture or noncapture in the census and postenumeration survey. This additional information is used to formulate the triple system estimates using any of an assortment of model assumptions.

For populations of size 1,000, this article presents simulations for ten estimators of persons missed on all lists, each of which can be combined with observed counts to

Table 1. Population counts by capture status

	In AL		Out of AL	
	In PES 1	Out of PES 0	In PES 1	Out of PES 0
In Census 1	N_{111}	N_{101}	N_{110}	N_{100}
Out of Census 0	N_{011}	N_{001}	N_{010}	N_{000}

produce estimates of the total population. Each estimate is compared with the corresponding true population value. The ability to estimate the dependence between the census and postenumeration survey for persons *not on* the administrative list (using persons *on* the administrative list) may reduce bias in the estimation of census coverage error. With dual system estimation, we cannot achieve this reduction in correlation bias in the presence of dependence and heterogeneity because we have no data available to estimate the dependence. Log-linear model theory can be useful in formulating and understanding some triple system estimators. These models are supported by empirical evidence that capture in the census or coverage measurement list is only weakly associated with capture on the administrative list. This is plausible since the administrative list is created in a radically different way than the census list and postenumeration survey list, which are independent surveys using similar fieldwork. The ten estimators are described in Section 2.

Seven of these estimators are motivated by hierarchical log-linear models based on Fienberg (1972). Additional references for log-linear models are Bishop et al. (1975), Fienberg (2000), and Agresti (2002). Two of the estimators are based on suggestions from Zaslavsky and Wolfgang (1990 and 1993). For comparison, the traditional dual system estimate (DSE) using only the decennial census and postenumeration survey will be computed.

Other triple system estimators using alternative models, not simulated for this article, are suggested by Darroch et al. (1993). They built an equivalence for the generalized Rasch model and the quasi-symmetric log-linear model. They compared estimates from a partial quasi-symmetry model and a full quasi-symmetry model with the no second order interaction estimator (see Subsection 2.3) as well as with the Zaslavsky and Wolfgang estimators (see Subsections 2.4 and 2.5). Chao and Tsay (1998) developed an estimator that is a function of an expected sample coverage (based on an average over the three lists of the proportions of persons observed as missed on the other two lists) and measures of dependence between lists. They compared their estimator with those of Darroch et al. Fienberg and Manrique-Vallier (2009) looked at a methodology for integrating these multiple system estimation methods with record linkage and missing data issues. Madigan and York (1997) developed a Bayesian methodology that allows for a variety of dependence structures between lists, uses covariates, and explicitly accounts for model uncertainty.

The following assumptions apply to all estimators: (1) Erroneous inclusions have been removed from all lists and (2) Processing and matching procedures have been developed so that there is no matching error as well as no error in the determination that a person is enumerated at the correct address. Section 3 describes the creation of a simulated population of 1,000 persons and Section 4 discusses the replication of this process and the creation of evaluation statistics. Section 5 presents the results and Section 6 provides a discussion.

2. Estimators to Be Simulated

All these estimates are motivated based on an assumption of homogeneity in capture probabilities across individual persons. If the particular log-linear model assumptions hold

and capture probabilities are homogeneous, then these estimators are nearly unbiased. Since individual capture probabilities are heterogeneous in the real world, the simulated populations are created using heterogeneous capture probabilities. Estimates using these models are biased given this heterogeneity and we can compare these estimates with the known population total.

2.1. *Conditionally Independent Models*

In order to use common log-linear model notation, let C denote Census, P denote the postenumeration survey, and A denote the administrative list. For example, consider the log linear model {CP, PA}. This log-linear model notation puts sources together if there is an assumed relationship (dependence) between them. This is a conditional independence model where at each level of P, C and A are independent, a unconditional relationship between C and P and between P and A is allowed but not between C and A. Since there is some empirical evidence (Zaslavsky and Wolfgang 1990, 1993, and Darroch et al. 1993) that the C and P lists are dependent conditional on capture on the A list, this model may be reasonable. The same is true for the model {CP, CA}. The third model with exactly two two-factor terms, {CA, PA}, may not be accurate since it assumes at each level of A that C and P are independent and this is not supported by the empirical evidence. Note that the empirical evidence from Zaslavsky, Wolfgang and Darroch is from the 1988 Census Dress Rehearsal and is based on administrative data limited to a few specific geographic areas and based on sources likely to be very different from any sources that might be used for the 2020 Census postenumeration study.

For model {CP, PA} the estimate is $\hat{N}^1_{000} = \frac{N_{001}N_{100}}{N_{101}}$. This is the usual dual system estimate for the unobserved cell in the 2×2 table conditional on $P = 0$, using the A list and the C list as sources after removing all individuals captured on the P list and assuming A and C are independent.

Models {CP, CA} leading to the estimate $\hat{N}^2_{000} = \frac{N_{001}N_{010}}{N_{011}}$ and {CA, PA} leading to the estimate $\hat{N}^3_{000} = \frac{N_{010}N_{100}}{N_{110}}$ follow from the appropriate permutations of the capture status indices.

2.2. *Jointly Independent Models*

For example, consider the log-linear model {A, CP} where there is a relationship between C and P, but neither C nor P has a relationship with A. This is a jointly independent model where A is jointly independent of C and P. This is ordinary two-way independence between A and a categorical variable composed of all four combinations of C and P. Given the empirical evidence cited above, this model might be reasonable. The other two jointly independent models, {P, CA} and {C, PA}, assume C and P are independent, but this is not supported by the empirical evidence.

For model {A, CP} the estimate is $\hat{N}^4_{000} = \frac{N_{001}(N_{110}+N_{100}+N_{010})}{N_{111}+N_{101}+N_{011}}$. This is equivalent to a DSE where one list is the administrative list and the other is a list formed by combining the census and PES list (un-duplication required). The combined list is assumed to be independent from the administrative list.

Model {P, CA} leading to the estimate is $\hat{N}^5_{000} = \frac{N_{010}(N_{101}+N_{001}+N_{100})}{N_{111}+N_{011}+N_{110}}$ and model {C, AP} leading to the estimate is $\hat{N}^6_{000} = \frac{N_{100}(N_{011}+N_{001}+N_{010})}{N_{111}+N_{101}+N_{110}}$; both follow from the appropriate permutations of the capture status indices.

2.3. No-Second-Order-Interaction Log-Linear Model

There is only one no-second-order-interaction log-linear model. Model {CP, CA, PA} assumes that the Census and postenumeration survey have dependence but there is no CPA term (three-way interaction). This is the least restrictive log-linear model for which data is available for estimation. All log-linear models from Subsections 2.1 and 2.2 are special cases of the no-second-order-interaction model (i.e., they all assume no second-order interaction along with additional restrictions).

The incomplete 2^3 table of counts in Table 1 is divided into one complete 2×2 subtable and one incomplete subtable. Assume the cross-product ratio is the same in both subtables. Then the estimate of the missing cell in the incomplete 2×2 table can be estimated using the known cross-product ratio from the complete 2×2 table. The assumption is that the dependence in the 2×2 table for C $\times$ P using only those individuals in A is the same as the dependence in the 2×2 table for C $\times$ P using only those individuals not in A. This model is in some sense analogous to the assumption of independence for the 2×2 table used for DSE but is one layer deeper. All pairs of sources can exhibit dependence, but the amount of dependence in each pair is assumed to be unaffected by conditioning on the third source. The estimator for this model is

$$\hat{N}^7_{000} = \frac{(N_{111})(N_{001})(N_{100})(N_{010})}{(N_{011})(N_{101})(N_{110})}.$$

Note that in order to estimate N_{000}, it is necessary to make an assumption about second-order interaction. This assumption does not have to be that the interaction term in the log-linear model is zero; any other fixed value for the interaction coefficient could be used, although some assumptions might be more plausible than others.

2.4. Zaslavsky and Wolfgang 1

This is a DSE, suggested in Zaslavsky and Wolfgang (1990 and 1993), where one source is the administrative list and the other is the combined census and census coverage measurement list. However, persons captured in both the census and postenumeration survey are removed from the administrative list and the combined list.

$$\hat{N}^8_{000} = N_{001}\frac{N_{100}+N_{010}}{N_{101}+N_{011}}$$

The assumption underlying the use of this estimator is that the probability of capture in the administrative list of persons omitted from the census and postenumeration survey is the same as the average probability of capture for those included in either the census or postenumeration survey, but not both. In other words, persons captured by neither C nor P are more like those captured by only the C or P than those captured by both. This estimator was included by Zaslavsky and Wolfgang based on evidence for four poststrata studied taken from the 1988 Census Dress Rehearsal.

2.5. Zaslavsky and Wolfgang 2

For this estimator, also suggested by Zaslavsky and Wolfgang (1990 and 1993), the odds ratio in the 2 × 2 table fixing on capture in the administrative list is calculated. Then, assuming this odds ratio holds, the DSE for the 00+ cell of the marginal C × P table is multiplied by this odds ratio.

$$\hat{N}^{9}_{000} = \left(\frac{N_{001}N_{111}}{N_{011}N_{101}}\right)\left(\frac{N_{01+}N_{10+}}{N_{11+}}\right) - N_{001}$$

The count in the 001 cell is subtracted to obtain an estimate of the 000 cell. The assumption is that the degree of dependence between the C and P sources is similar in the overall population to that in the subpopulation captured by the administrative list. Zaslavsky and Wolfgang note that this assumption may be conservative for the population as a whole, because the administrative list captures are likely to be more homogeneous than the general population and the odds ratio would be closer to 1 (independence would more nearly hold).

2.6. Traditional DSE

For comparison, this is the DSE estimate using only the Census list and postenumeration survey list. The assumption is that C and P are unconditionally independent {C, P}.

$$\hat{N}^{10}_{000} = \frac{N_{10+}N_{01+}}{N_{11+}} - N_{001}$$

2.7. Population Total Estimates

For each of the $t = 1$ to 10 $\hat{N}^{t}_{000}$ estimates calculated, the total population estimate is

$$\hat{N}^{t} = \hat{N}^{t}_{000} + N_{1++} + N_{011} + N_{010} + N_{001}.$$

3. Creating the Simulated Populations

Populations of $N = 1{,}000$ persons will be simulated, allowing for heterogeneous capture probabilities and homogeneous conditional odds ratios. One conditional odds ratio is the odds ratio for the 2 × 2 table of C × P conditional on capture on A and the other is the odds ratio for the 2 × 2 table of C × P conditional on not captured (missed) on A.

3.1. Creating a Specified Conditional Odds Ratio

Omitting any subscript for an individual member of the population, the 2 × 2 table of conditional capture probabilities for census capture and postenumeration survey capture given capture on the administrative list is given in Table 2.

In order to create a simulated population with a given set of conditional odds ratios, the odds ratio formula for a 2 × 2 subtable is written as a function of an unknown proportion in the 11 cell and the known marginal proportions the 1+ and +1 margins.

Table 2. Capture probabilities for Census and PES given capture on administrative list

	In PES 1	Out of PES 0	
In Census 1	P_{11}	P_{10}	P_{1+}
Out of Census 0	P_{01}	P_{00}	
	P_{+1}		

Accordingly, given P_{1+}, P_{+1}, and odds ratio

$$\theta = \frac{P_{11}P_{00}}{P_{10}P_{01}} = \frac{P_{11}(1 - P_{1+} - P_{+1} + P_{11})}{(P_{1+} - P_{11})(P_{+1} - P_{11})},$$

the equation can be rewritten as

$$(1 - \theta)P_{11}^2 + [1 - P_{1+} - P_{+1} + \theta(P_{1+} + P_{+1})]P_{11} - \theta P_{1+}P_{+1} = 0. \quad (1)$$

This equation can be solved for P_{11} using the quadratic formula producing two roots, one of which is between 0 and 1 and is the one we want.

This value of P_{11} and given P_{1+} and P_{+1} provides the desired odds ratio θ.

The process described starting with Table 2 is repeated for Capture Probabilities for census and PES given not captured (missed) on the administrative list, allowing in some simulations for a different conditional odds ratio θ.

3.2. Generating a 1,000 Person Population Allowing for Heterogeneity in Capture Probabilities

We want to generate several populations of size $N = 1{,}000$ persons to have particular capture properties. This is accomplished by specifying two conditional odds ratios.

Let θ_1 be the odds ratio for census and PES given capture on the administrative list and θ_2 the odds ratio for census and PES given *not* captured on the administrative list.

Given θ_1 and θ_2 (assumed constant over persons) and five beta parameters in the following conditional capture probabilities

$$P_k\langle A\rangle = \frac{\exp(\beta_{10} + \beta_{11}X_k)}{1 + \exp(\beta_{10} + \beta_{11}X_k)}, \quad P_k\langle C|A\rangle = \frac{\exp(\beta_{20} + \beta_{21}X_k)}{1 + \exp(\beta_{20} + \beta_{21}X_k)},$$

$$P_k\langle P|A\rangle = \frac{\exp(\beta_{30} + \beta_{31}X_k)}{1 + \exp(\beta_{30} + \beta_{31}X_k)}, \quad P_k\langle C|notA\rangle = \frac{\exp(\beta_{40} + \beta_{41}X_k)}{1 + \exp(\beta_{40} + \beta_{41}X_k)},$$

$$P_k\langle P|A\rangle = \frac{\exp(\beta_{50} + \beta_{51}X_k)}{1 + \exp(\beta_{50} + \beta_{51}X_k)},$$

for $k = 1$ to 1,000 independently generate $X_k \sim N(0,1)$ and calculate

$$P_k\langle A\rangle, P_k\langle C|A\rangle, P_k\langle P|A\rangle, P_k\langle C|notA\rangle, P_k\langle P|notA\rangle.$$

Note that although the conditional odds ratios are assumed constant over persons, the capture probabilities are heterogeneous since variation in the independent variables is created.

Using θ_1 and $P_k\langle C|A\rangle, P_k\langle P|A\rangle$, we use the methodology from Subsection 3.1 and Equation (1) to solve for the probability of capture in both the census and postenumeration survey given capture on the administrative list. Then complete the 2 × 2 table of capture probabilities given capture on the administrative list. Multiplying each of these conditional probabilities by $P_k\langle A\rangle$ provides $p_{k,111}, p_{k,101}, p_{k,011}, p_{k,001}$.

Then, using θ_2 and $P_k\langle C|notA\rangle, P_k\langle P|notA\rangle$, use the methodology from Subsection 3.1 and Equation (1) to solve for the probability of capture in both the census and postenumeration survey given *not* captured on the administrative list. Then complete the 2 × 2 table of capture probabilities given *not* captured on the administrative list. Multiplying each of these conditional probabilities by $(1 - P_k\langle A\rangle)$ provides $p_{k,110}, p_{k,100}, p_{k,010}, p_{k,000}$.

Next, generate a number u from 0 to 1 from the $U(0,1)$ distribution and use the cumulative distribution of the eight cell probabilities to determine which of the eight cells of Table 1 person k falls into.

After completing the above for each of the 1,000 population persons, tabulate the seven observed counts from Table 1 and using these compute $\hat{R}^t = \frac{\hat{N}^t}{1000}$ for $t = 1$ to 10. This is the ratio of the estimated population count to the true population count and provides a measure of the accuracy of the estimate.

4. Replication

This article presents results for 1,000 independent replications of the population generation as specified in 3.2 for a given θ_1 and θ_2 (assumed constant over persons) and one set of beta parameters (shown in Table 3). This set of beta parameters was selected as they produce average capture probabilities, described in Section 5, that are small (.227), and thus represent what may be considered a hard-to-reach population.

Table 3. Accuracy of alternative estimates of missing count

"Average Capture Probability" = 0.227				
Average R = Estimated Count/True Count (se)				
$\beta_{10} = -0.700$ $\beta_{30} = -1.000$ $\beta_{50} = -1.500$	$\beta_{11} = 0.800$ $\beta_{31} = 0.600$ $\beta_{51} = -0.400$	$\beta_{20} = -1.200$ $\beta_{40} = -2.000$	$\beta_{21} = .500$ $\beta_{41} = -.300$	
Estimator	$\theta_1 = 1.5$ $\theta_2 = 1.2$	$\theta_1 = .75$ $\theta_2 = .85$	$\theta_1 = .75$ $\theta_2 = .75$	$\theta_1 = 1.5$ $\theta_2 = 1.5$
1	.965 (.002)	.958 (.002)	**.972 (.002)**	***.946 (.002)***
2	1.535 (.007)	1.455 (.007)	1.428 (.006)	1.488 (.007)
3	1.021 (.003)	1.178 (.004)	1.240 (.005)	**.947 (.003)**
4	1.100 (.002)	1.093 (.002)	1.091 (.002)	1.086 (.002)
5	1.242 (.003)	1.369 (.004)	1.390 (.004)	1.174 (.003)
6	.968 (.002)	**1.013 (.002)**	***1.032 (.002)***	.937 (.002)
7	1.177 (.008)	1.022 (.007)	1.056 (.007)	1.060 (.006)
8	1.087 (.002)	1.075 (.002)	1.077 (.002)	1.063 (.002)
9	1.197 (.008)	***1.018 (.008)***	1.053 (.008)	1.074 (.007)
10	**.993 (.004)**	1.295 (.004)	1.377 (.007)	.915 (.003)

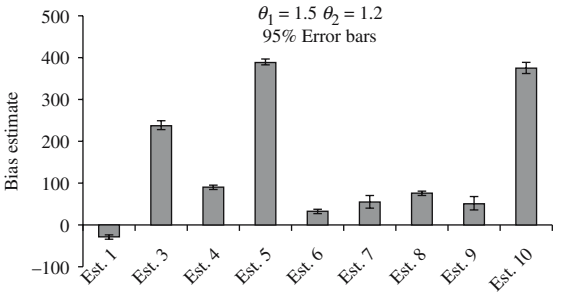

Fig. 1. 95% error bars

For each of the ten estimates, use these 1,000 replicates to compute the empirical mean ratio R^t denoted as $\bar{R}^t$, and its variance, $Var(\bar{R}^t)$.

Note that none of the precise assumptions, particularly homogeneity in capture probability, needed for validity of any of these ten estimators is satisfied by any of these simulated populations. Darroch et al. (1993) provide some arguments that no three-way interaction model may be a fair approximation except for heterogeneity. The kind of person-to-person heterogeneity introduced by these simulations might be expected to be a reasonable representation of the reality of list formation. This heterogeneity produces bias in these estimates even if the model assumptions about the relationship between the capture attempts hold.

5. Results

Table 3 shows results for each of the ten estimator alternatives for one set of β parameters and four sets of odds ratios θ_1 and θ_2 (1.5 and 1.2; .75 and .85; .75 and .75; 1.5 and 1.5). When $\theta_1 = \theta_2$, the odds ratio for census capture or not by postenumeration survey capture status is independent of capture status on the administrative list (no second order interaction). When $\theta_1 \neq \theta_2$ the odds ratio for census capture or not by postenumeration survey capture status is dependent on capture status on the administrative list. The "Average Capture Probability" (ACP) is the average of the five probabilities defined in Section 3 for $X_k = 0$ (the mean of the random variable X). *It is used as a measure of*

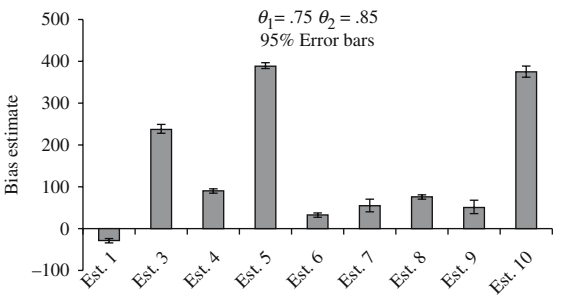

Fig. 2. 95% error bars

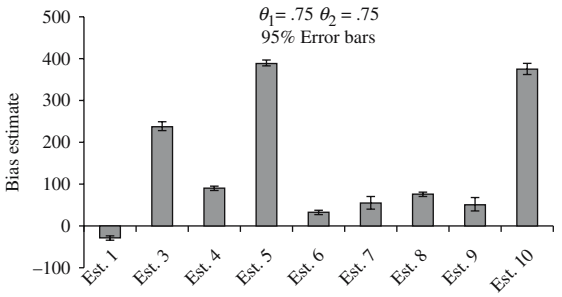

Fig. 3. 95% error bars

"hard to reach" since lower values indicate lower capture probabilities (i.e., harder to reach).

$$ACP = \frac{\sum_{i=1}^{5} \frac{e^{\beta_{i0}}}{1+e^{\beta_{i0}}}}{5}$$

There are ten rows of average ratios R defined in Section 3 with the standard error of the average R in parenthesis, one row for each of the ten estimators of total population. There are four columns, one for each θ_1 and θ_2 combination. For each θ_1 and θ_2 combination, the average R-value that is closest to 1 is in bold. If the second-best average R is not statistically different (single pair comparison) than the best, it is shown in bold italics. The standard errors are small (all coefficients of variation less than 0.01). Thus the results are similar for many of the estimators, except for Estimator 2 which produced a large overestimate (close to 50%) for all four columns. To illustrate this, for each of the four sets of odds ratios, a 95% confidence interval error bar chart for the bias estimate is also provided (as Figures 1 through 4) excluding Estimator 2.

For Table 3, the average capture probability was .227. For $\theta_1 = 1.5$ and $\theta_2 = 1.2$, Estimator 10 was the best with an average R of .993 (se = .004). For $\theta_1 = .75$ and $\theta_2 = .85$, Estimator 6 was the best with an average R of 1.013 (se = .002). For $\theta_1 = .75$ and $\theta_2 = .75$, Estimator 1 was the best with an average R of .972 (se = .002). For $\theta_1 = 1.5$ and $\theta_2 = 1.5$, Estimator 1 was the best with an average R of .947 (se = .003).

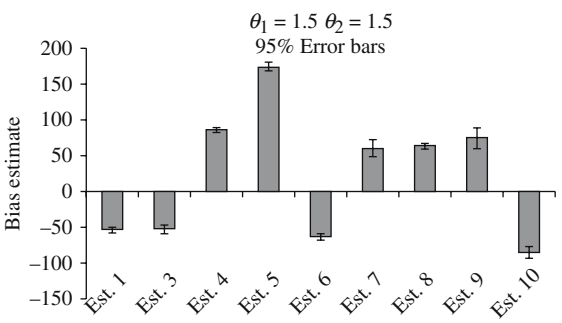

Fig. 4. 95% error bars

6. Discussion

Different conditional odds ratios and beta parameters, as well as a new generation of the independent random X variables for each iteration, would produce different results; thus the simulations shown here serve as an example and only as an indication of what may be expected with varying parameters. Even with the same odds ratios and beta parameters, generating a new 1,000 person population, as described in Section 3, produces different results.

Although it is clear from these results that three sets of capture attempts can produce more accurate estimates than two capture attempts, there are additional things worth considering. First, the cost of three enumeration attempts is considerably greater than for two enumeration attempts. Second, there is likely to be greatly increased matching error going from two attempts to three attempts. For two attempts at capture, there are only four cells in a 2 × 2 table, and given the marginal counts of the total count for each of the attempts, matching is only necessary to obtain the 11 cell (captured in both attempts). For three attempts, there are eight cells. For Estimate 7, no second-order interaction, counts are required for all the other seven cells in order to estimate the 000 cell. Estimate 7 makes a less restrictive assumption (no second-order interaction) than the estimators from Subsection 1.1 (conditionally independent models) and Subsection 1.2 (jointly independent models). In theory, Estimate 7 should be the better than the estimates in Subsections 1.1 and 1.2 as well 1.6 (traditional DSE), if in reality there is a second-order interaction and if there are no errors in obtaining the counts. Second-order interaction and heterogeneity in capture probabilities are likely in the real world for most populations. For example, both the 111 cell and the 110 cell are required so that both the count of captured in the first two attempts and in the third attempt *and* captured in the first two attempts but missed in the third are necessary. Obtaining all these counts from a complex matching operation may be error prone. Further research using some reasonable matching-error models is planned to investigate whether it may be more effective to use less optimal estimators that require less matching but may be more robust to matching error.

For the simulations in Table 3, Table 4 shows the average ratios R for the total population estimate for each of the four sets of θ_1 and θ_2 for the DSE and the best (lowest ABS(R-1)) of all ten estimators of total population. Note that although the standard errors of average R values are small, for some simulations the second-best estimator was not significantly different than the best estimator. The absolute value (ABS) of R-1 is shown for DSE and the best of the ten estimators. This is the absolute relative error.

Table 4. Accuracy of total population estimate: R = average estimated total population/1,000

θ_1	θ_2	Average R for DSE	ABS(R-1) for DSE	Estimator with best average R	Best average R	ABS(S-1) for Best	Difference in absolute error DSE – Best
1.5	1.2	0.993	0.007	10 (DSE)	0.993	0.007	0.000
.75	.85	1.295	0.295	6 {C,AP}	1.013	0.013	0.282
.75	.75	1.377	0.377	1{CP,PA}	0.972	0.028	0.349
1.5	1.5	0.915	0.085	1{CP,PA}	0.946	0.054	0.031

For example, the maximum difference is found in Table 3 for $\theta_1 = .75$ and $\theta_2 = .75$ where the absolute relative error for the best estimator, Estimator 1, is 2.8% and the absolute relative error for DSE is 37.7%, a 34.9 percentage point difference.

When considering the accuracy of DSE, which requires only two sets of enumeration attempts and is less subject to matching error, it is important to compare two enumeration attempts with one enumeration attempt. For example for $\theta_1 = .75$ and $\theta_2 = .75$ the difference in absolute error between the best triple system estimator and the DSE was 34.9 percentage points. The average capture probability is .227. If the capture probability was a constant .227, one capture attempt for the population of 1,000 would have an expected capture of 227 persons and absolute error of 77.3%. The DSE absolute error of 37.7% is much less. Thus two capture attempts followed by DSE (with a 37.7% absolute error) may produce a substantial gain over one capture attempt (with a 77.3% absolute error) even if the absolute relative error of DSE is still rather high. In practice, while likely not sufficient for a Decennial Census, the two independent capture attempts, (1) an attempted 100% enumeration of a hard-to-reach population and (2) the creation of a list using administrative records, followed by dual system estimation may produce a much more accurate population estimate than relying on only one capture attempt.

7. References

Agresti, A. (2002). Categorical Data Analysis (2nd edition). New York: John Wiley and Sons.

Alho, J.M. (1990). Logistic Regression in Capture-Recapture Models. Biometrics, 46, 623–635.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Bell, W.R. (1993). Using Information from Demographic Analysis in Post-Enumeration Survey Estimation. Journal of the American Statistical Association, 88, 1106–1118. DOI: http://www.dx.doi.org/10.1080/01621459.1993.10476381

Chandrasekar, C. and Deming, W.E. (1949). On a Method of Estimating Birth & Death Rates and the Extent of Registration. Journal of the American Statistical Association, 44, 101–115.

Chao, A. and Tsay, P.K. (1998). A Sample Coverage Approach to Multiple-System Estimation with Aoolication to Census Undercount. Journal of the American Statistical Association, 93, 283–293.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catch-ability. Journal of the American Statistical Association, 88, 1137–1148. DOI: http://www.dx.doi.org/10.1080/01621459.1993.10476387

Fienberg, S. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. Biometrika, 59, 591–603. DOI: http://www.dx.doi.org/10.1093/biomet/59.3.591

Fienberg, S.E. (2000). Contingency Tables and Log-Linear Models: Basic Results and New Developments. Journal of the American Statistical Association, 95, 643–647. DOI: http://www.dx.doi.org/10.1080/01621459.2000.10474242

Fienberg, S.E. and Manrique-Vallier, D. (2009). Integrated Methodology for Multiple Systems Estimation and Record Linkage using a Missing Data Formulation. AStA Advances in Statistical Analysis, 93, 49–60. DOI: http://www.dx.doi.org/10.1007/s10182-008-0084-z

Madigan, D. and York, J.C. (1997). Bayesian Methods for Estimation of the Size of a Closed Population. Biometrika, 84, 19–31. DOI: http://www.dx.doi.org/10.1093/biomet/84.1.19

Mulry, M.H. and Spencer, B.D. (1991). Total Error in PES Estimates. Journal of the American Statistical Association, 86, 839–855. DOI: http://www.dx.doi.org/10.1080/01621459.1991.10475122

Rastogi, S. and O'Hara, A. (2012). 2010 Census Match Study. 2010 Census Program for Evaluations and Experiments, Center for Administrative Records Research and Applications. November 19, 2012.

Wolter, K.M. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, 338–346. DOI: http://www.dx.doi.org/10.1080/01621459.1986.10478277

Zaslavsky, A.M. and Wolfgang, G.S. (1990). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data. In Proceedings of the Survey Research Methods Section of the American Statistical Association (Anaheim, CA, August 1990).

Zaslavsky, A.M. and Wolfgang, G.S. (1993). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data. Journal of Business & Economic Statistics, 11, 279–288. DOI: http://www.dx.doi.org/10.1080/07350015.1993.10509955

Received February 2013
Revised January 2014
Accepted February 2014