TRIPLE SYSTEM MODELING OF CENSUS, POST-ENUMERATION SURVEY, AND ADMINISTRATIVE LIST DATA

Alan M. Zaslavsky, Harvard University and Glenn S. Wolfgang, Bureau of the Census
Alan M. Zaslavsky, Department of Statistics, 1 Oxford Street, Cambridge, MA 02138

Dual system measurement of census coverage using a post-enumeration survey (PES) has been criticized for correlation bias, resulting when responses to the census and survey are not independent. Use of a third system (information source) can provide additional information to assess that independence.

The data for this study come from a population subgroup of the 1988 Dress Rehearsal Census and its PES and from rosters from other government sources. This study focuses on Black male adults. Preliminary results using a variety of models confirm that, as previously suspected, their population is underestimated by dual system methods.

Potential problems involving classification and matching errors are also discussed. The results suggest that triple system modeling has great potential for more precise estimation of the hard-to-count population and its census coverage.[1]

1. Introduction

The Post Enumeration Survey (PES) now underway as part of the 1990 census is designed to evaluate the coverage of the census. A sample of census blocks and sub-block areas consisting of approximately 150,000 households will be surveyed to determine whether household members were included in the census. The results of the PES will be used to estimate census coverage rates for poststrata defined by demographic and geographical variables.

The PES, like the census, will not have perfect coverage in the areas surveyed. PES estimates of census coverage rates are based on the assumption that inclusion of a person in the PES is independent of inclusion in the census. This independence assumption underlies the Dual System Estimator (DSE) that is used to estimate coverage rates by poststratum.

One of the major criticisms that has been raised of coverage rate estimates obtained through the DSE is that independence does not in fact hold. If people in the same poststratum have heterogeneous probabilities of coverage, then the same people might be most likely to be omitted from both the census and the PES. In this case, estimates based on the independence assumption would be biased downwards ("correlation bias"). If inclusion in the census has a direct causal effect on a person's propensity to be included in the PES (perhaps because of confusion between the survey and the census, or because of sensitization to the survey process), a correlation bias might be generated in either direction. A general discussion of correlation bias in the PES appears in Ericksen and Kadane (1985) and the following discussion, and is summarized in Zaslavsky (1989). Fay, Passel, and Robinson (1988) discuss the possibility of lack of independence in the 1980 Post Enumeration Program (PEP). They present evidence from demographic estimation suggesting that the most undercovered population group, black male adults, also tended to be underestimated by the PEP (p. 77).

The Administrative List Supplement (ALS) program was part of the research effort around the PES conducted in St. Louis as part of the 1988 test census. In this program, names were collected from a variety of administrative lists to supplement the PES. Issues regarding the scope and processing of the administrative lists are discussed in Section 2.

A dual system estimate incorporating ALS data into the P-source should become less subject to correlation bias: the ALS should improve the independence of the second source because it is collected by a methodology very different from the census. Also, the supplemented PES would have better overall coverage of persons in the PES area. Estimates derived using this approach are presented in Section 3.

Another view of these data is to regard the three sources (census, PES, and administrative lists) as three distinct "systems." (Triple system estimation has been proposed by Marks, Seltzer, and Krotki (1974), among others.) The additional source (the A-source) provides data with which to evaluate the previously untestable assumption of independence between the census and the PES, and to develop models for the interactions between the sources. The statistics for this analysis are the 2 × 2 × 2 tables for counts of cases included or omitted in each of the three sources, for each post-stratum. A number of models applying this approach also are developed and compared in Section 3.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

2. The Triple System Data

2.1. Data Collection

The data are derived from the program of coverage measurement for the 1988 Dress Rehearsal for the Bicentennial Census of the United States. Each source is essentially a list of persons, identified by names, addresses, and characteristics, who lived at specific sample addresses on specific dates.

The E-source is the census itself, that is, those enumerated by ordinary census procedures. The P-source is the Post-Enumeration Survey (PES), a prime vehicle for census coverage evaluation (Childers and Hogan, 1989). The A-source was compiled from pre-census administrative records of state and federal governmental agencies, encompassing Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administration registrants.

The A-source includes persons in the Administrative List Supplement to the PES as well as persons not in that supplement. The Administrative List Supplement (ALS) used the governmental records to test the potential for such data to improve PES coverage and reduce dual system estimation bias (Wolfgang, 1989 and 1990). It was designed to target black male renters, who are believed to be among the most undercounted by the census and the most underestimated by dual system estimates, due to correlation bias (Fay, Passel and Robinson, p. 77). Any persons from those lists but not found in the PES were added to the PES. ALS processing was designed directly to code the supplement and recompute the estimates with the revised data. Other A-source persons come from those not in the supplement because (and only because) they were already in the PES.

The addresses and persons in sample for these data are limited in a variety of ways. Sample addresses are all in St. Louis, Missouri in the seventy blocks of the PES sample design stratum where most residents were expected to be black renters. Administrative list records from addresses outside those blocks were dropped from processing. Persons in scope for this study are selected on the basis of age, race, and sex to fall within four poststrata used in presenting census coverage measurement results: black males aged 20-29 in owned homes, black males aged 30-44 in owned homes, black male renters aged 20-29, and black male renters aged 30-44. For more detail on both sample design stratification and on poststratification for a PES, see Diffendal (1988).

When data were too incomplete or inadequate for coding and eventually classifying persons in all sources, a field followup interview attempted to get clarifications. In particular, residence at a sample address on the dates for census and PES enumerations needed to be verified in the field for A-source persons not found in the census or PES. There was some additional sampling of those followup persons in order to keep interviewing workloads within projections. Supplement persons therefore have a followup selection weight of either 1 or 4.

The data were further classifiied by address register area (ARA), a subdivision of census geography consisting of contiguous blocks. An ARA may contain no or one or several PES sample blocks. There is an average of about two sample block clusters in each ARA that has at least one PES sample block in the A-source stratum.

Three-way cross-classification of the data for each of four poststrata are presented in Table 1.

2.2. Classification Issues.

The validity of triple system estimation depends not only on how well the adopted model fits the unknown reality but perhaps more dramatically on how accurately persons are assigned in three ways: (1) in or out of scope, (2) into poststrata, and (3) into the cells of the cross-classification of sources. Those classifications use many pieces of information. When that information is missing or in error, the classifications will sometimes be wrong. Dual system PES coding and computation procedures for processing incomplete or ambiguous data have been refined over the past decade to minimize and balance classification errors. Triple system efforts have benefitted from those developments but present new issues due to the added interrelationships of sources and, ironically, to the extra information of the third source.

For background on these classification errors, refer to Wolter (1986, p. 339), who defines dual system model assumptions relating to them. Closure and spurious events assumptions specify in-scope issues. Matching and (partial) nonresponse assumptions deal with cell assignment. He also listed a poststratification assumption. Among those who discussed such concerns before Wolter are Marks, Seltzer, and Krotki (1974, pp. 87-124, 408) who also considered the role of errors and biases in three-source data.

Here are a few general classification issues for the present data:

- The A-source explicitly drops those who move away before the PES and, given the pre-census source of the records, misses those who move in since Census Day. Does exclusion of movers affect model assumptions in any way?

- There are many selection forces operating on the A-source. Registration in the administrative lists is not expected to be complete, even for the target population; some people neither drive nor register for the draft nor use employment services nor file tax returns. The A-source was limited geographically to blocks in only one stratum of one test site, even though its data may be generalized to at least two other strata. While most kinds of E-source or P-source missing data are imputed (weighting up for whole-household missing data, hot deck for characteristics used in matching or poststratum assignment, probability imputation based on logistic regression for unresolved E-source in-scope statuses or P-source match statuses), A-source persons with missing age or sex or with unresolved address for the time of the census or PES are declared out of scope. Note that very few cases in E- or P-sources are imputed, but perhaps 25% more A-source persons could be imputed. Does such stringent selection of A-source persons foster or hinder independence; how does it affect error variance?

- Does imputation of A-source tenure or race on the basis of P-source values for the sake of poststratum assignment constitute a violation of independence?

- Is the probability of resolving or excluding a case the same across the unknown true conditions? In particular, are persons more likely unresolved or declared out of scope if in only one source?

Several examples illustrate more specific quandaries encountered in classification for triple system matching. These are representative of a multitude of scenarios. Most, except age discrepancies between sources, seem to occur quite rarely. Together, they could affect a substantial percentage of data.

- Person A is represented in the P-source with an unresolved PES match status and a 0.67 imputed probability of match to the census for dual system estimation purposes. Should the P-source exclude him on the basis that the A-source droppped all unresolved match status cases? Should he be kept, assigning him as a census match, the result of rounding the match probability?

- Person B, represented by records in all three systems that matched with each other even if age did not perfectly agree, has a missing E-source age, a P-source age of 28, and an A-source age of 30. Given the poststratum cutoff between 29 and 30, in which poststratum should he be assigned?

- Person C, from a racially mixed household, is coded white in the P-source and black in the A-source. Is he in scope for the present study?

- Person D, unresolved in followup at the address listed in the A-source, has been found in the P-source at a different address in a different (or, for another case, the same) PES sample block. Should he be considered in scope? If an A-source person not in the P-source cannot be confirmed to live in a different sample area at the time of the PES, how can such serendipitous finds be balanced?

These examples illustrate that three-way classification of such data still needs procedural investigation and refinement. It will improve with experience and informed feedback. This test run provides an opportunity to raise issues that need evaluation and prescription. At the same time, its data may be considered good enough for exploratory modeling, since they have undergone many months of processing, review, file merges, quality checks, and consistency edits in order to ensure quality and identify issues. The modeling in the sequel shows the value of further development of classification procedures.

3. Triple-System Models

A rich collection of models can be built upon the three-way table for inclusion status with respect to the three sources. In this section, a number of models of this type are developed.

Point estimates were calculated for the combined geographic area, and standard errors were obtained by jackknifing the ARAs in Stratum 11. (Similar results were obtained using the bootstrap but are not displayed.) Each ARA included in this study contains one or more PES sample blocks. Since variances *within* the geographically contiguous ARAs may be smaller than those *between* ARAs, these standard errors are perhaps slightly conservative relative to those that would have been obtained if the individual sampled blocks had been jackknifed.

Weights, except those for ALS subsampling, were ignored in this analysis, but would not be expected to have much effect on the conclusions.

In the following discussion, inclusion in a source is coded as "1" and omission as "0". The coordinates of a cell are given as (e,p,a) so, for example, cell

(0,1,0) refers to persons included in the PES but not in the census or administrative lists. The observed count in the (e,p,a) cell is x_{epa}.

The poststrata are labeled "O2," "R2," "O3" and "R3" in the tables for owners/renters × ages 20-29, 30-44.

3.1. Cross-product ratios

Direct evidence is available in the triple-system tables for various cross-product ratios in 2 × 2 subtables formed by restricting consideration to cases observed in a selected source. Controlling for $e = 1$, the $P \times A$ cross-product ratio is given by $x_{111}x_{100}/x_{101}x_{110}$. Thus, for the subpopulation that is enumerated in the census, complete information is available to calculate the cross-product ratio, while for the full population, the $P \times A$ data are necessarily incomplete for persons included in neither the P- nor the A-source. Similarly, cross-product ratios may be calculated for $E \times A$ given that $p = 1$ and for $E \times P$ given that $a = 1$.

The estimated cross-product ratios are displayed in Table 2. The evidence of the cross-product ratios in the $a = 1$ subtables suggests a substantial lack of independence between the census and the PES, with ratios ranging from 12 to 32 (independence=1).

Because the sampling distribution of the log cross-product ratio is more nearly normal than the untransformed ratio, estimates and jackknife standard errors are presented on the logarithmic scale in the lower part of Table 2. The standard errors of these estimates are large, but in every poststratum the estimated log cross-product ratio is at least three times its standard error. Furthermore, these estimates may be conservative for the population as a whole because the A-source captures are likely to be somewhat more homogeneous than the general population and therefore independence would more nearly hold within the smaller group, i.e. the cross-product ratio would be somewhat depressed.

The cross-products obtained by fixing $e = 1$ or $p = 1$ are closer to 1, suggesting that the administrative lists are more nearly independent of the survey sources. This seems intuitively reasonable since a very different method of data collection might be expected to yield a more nearly independent set of names.

3.2. Models for three-system estimation

With three sources, as with two (the DSE), there is always one cell (in the contingency table for inclusion in the various sources) that represents persons who are not enumerated in any of the sources. Unless the evidence suggests that coverage is good enough to make this cell count negligible, some sort of model must be used to estimate this part of the population in the area subjected to coverage evaluation.

A number of different models are compared, each predicting a count $\hat{x}_{000}$ in the unobserved (0,0,0) cell. The first and third models predict the margin $\hat{x}_{00+}$ and $\hat{x}_{000}$ is then obtained by subtraction; the remaining models explicitly predict $\hat{x}_{000}$. The predicted counts $\hat{x}_{000}$ for each model, with jackknife standard errors, are shown in Table 3. The gross census coverage rates, the ratios of the census count to the total population estimated under each model, is shown in Table 4.

The following models are considered:

"DSE w/o A-source" — This is the ordinary dual system estimator based only on the E- and P-sources, ignoring the A-source: $\hat{x}_{00+} = x_{10+}x_{01+}/x_{11+}$. $\hat{x}_{000}$ is obtained by subtracting x_{001} from the dual system estimate $\hat{x}_{00+}$. The negative estimates $\hat{x}_{000}$ indicate that the DSE estimate $\hat{x}_{00+}$ is less than the actually observed count added by the ALS, x_{001}. This represents a substantial underestimation of population since marginal A-source coverage is low and therefore there are probably additional persons missed by all sources.

"DSE with P+ALS" — This estimator represents the ALS as originally implemented, that is with the ALS combined with the PES to make a single second source. The DSE is applied to estimate directly $\hat{x}_{000} = x_{100}(x_{01+} + x_{001})/(x_{11+} + x_{101})$. Population estimates under this model are smaller than those for any of the following models, although larger than those for the unsupplemented DSE.

The preceding models are entirely dual system estimates, differing in the definition of the second source. The remaining models make use of triple system data.

"DSE with k_2" — An $E \times P$ cross-product ratio k_2 is estimated from the cells with $a = 1$, $k_2 = x_{001}x_{111}/x_{011}x_{101}$. The DSE is recalculated assuming that same cross-product ratio in the marginal $E \times P$ table, $\hat{x}_{00+} = k_2 x_{01+}x_{10+}/x_{11+}$. Thus, the assumption is that the degree of dependence (as measured by the cross-product ratio) between the E- and P-sources is similar in the overall population to that in the subpopulation captured by the administrative lists. This assumption may be conservative for reasons noted in the discussion of cross-product ratios above.

"Ratio r_1" — An estimate of the odds ratio for coverage by the A-source is obtained based on all the cells enumerated in the E- or P-source, $r_1 =$

$(x_{111}+x_{101}+x_{011})/(x_{110}+x_{100}+x_{010})$. This same odds ratio is applied to x_{001} to estimate the count $\hat{x}_{000} = x_{001}/r_1$. The assumption underlying use of this estimator is that the probability of coverage in the A-source of persons omitted from the E- and P-sources is the same as the average probability of coverage for those included in at least one of those sources.

"Ratio r_2" — An estimate of the odds ratio for coverage by the A-source is obtained based on the cells enumerated in E- or P-source but *not* both, $r_2 = (x_{101} + x_{011})/(x_{100} + x_{010})$. This same odds ratio is applied to x_{001} to estimate the count $\hat{x}_{000} = x_{001}/r_2$.

The assumption underlying use of this estimator is that the probability of coverage in the A-source of persons omitted from the E- and P-sources is the same as the average probability of coverage for those included in either the E-source or the P-source but not both, but that the persons enumerated in both the E- and P-sources are not necessarily comparable in this respect. In other word, persons captured by neither source are more like those captured by one than those captured by both. This assumption may be more plausible than that of "Ratio r_1" in light of the evidence of Table 4, which shows sample coverage rates for the A-source for each cell of the $E \times P$ table, x_{ep1}/x_{ep+}. Sample coverage rates for cases missed by the E- or P-source but not both are usually lower than those for cases captured in both surveys.

"Estimate, $k_3 = 1$" — The assumption of this estimator is that the 3-way crossproduct ratio in the $E \times P \times A$ table is 1, so $\hat{x}_{000} = x_{100}x_{010}x_{001}x_{111}/x_{110}x_{101}x_{011}$. This is called the "multilist" assumption by Ericksen and Kadane (1985). This assumption yields the largest estimates $\hat{x}_{000}$ of all the models considered here. Standard errors for this estimator are also large, since every observed cell count appears in the ratio.

3.3. Comparison of estimates

It is striking that even without consideration of estimates of people who are not captured by any source, the A-source adds more people than were estimated by the conventional DSE to be present but unobserved in the sample blocks. The dual system estimator with the ALS added to the PES as a supplement, as originally envisioned, yields slightly larger estimates than would be obtained simply by adding the ALS cases to the persons actually observed.

Perhaps the most striking observation is that every triple-system estimator considered yields lower estimates of the census coverage rate than those derived from the DSE. As shown in Table 4, standard DSE gross coverage estimates by poststratum range from 0.738 to 0.864, while triple-system E-source gross coverage estimates range (over various models) from 0.277 to 0.696.

The "DSE with k_2" and "$k_3 = 1$" estimators are both based on projecting the cross-product ratio in the $a = 1$ subtable to the rest of the population; since these cross-product ratios are so large, the estimates are also large although fairly close to each other. The assumptions underlying these estimates attempt to use the cross-product ratio as an assumedly invariant measure of association between two sources. Since this invariance has not been empirically tested, the assumptions may be questionable.

The two ratio estimators are based on readily interpretable assumptions about coverage rates for the A-source in different cells of the $E \times P$ table. They yield estimates that are more modest than those based on cross-product ratios. The "r_2" estimates are slightly the larger of the two and are based on slightly more plausible assumptions, namely that persons missed by both the E- and P-sources are more like those missed by one source than like those included in both sources.

While, taken at face value, these estimates appear to imply a much larger undercoverage than the PES estimates alone have indicated, they must be interpreted with great caution. Although every possible effort was taken in processing, the unduplication and matching procedures used in triple system estimation have not be subjected to the same scrutiny as those used in the DSE. The extremely low estimated coverage rates from the triple-system methods must also be considered in light of the fact that the target population has exceptionally poor coverage, as shown by the DSE. The results do suggest, however, that it would be worthwhile to pursue some form of triple system estimation as a second-order evaluation of the PES program which evaluates the census.

References

Childers, D. and H. Hogan (1989), "The 1988 Post Enumeration Survey: Methods and Preliminary Results," internal memorandum, Statistical Research Division, Bureau of the Census.

Diffendal, G. (1988), "The 1986 Test of Adjustment Related Operations in Central Los Angeles County," *Survey Methodology*, 14, 71-86.

Ericksen, E. and J. Kadane (1985), "Estimating the Population in a Census Year: 1980 and Beyond," *Journal of the American Statistical Asso-*

ciation 80:98-109.

Fay, R. E., J. S. Passel and J. G. Robinson (1988), "The Coverage of the Population in the 1980 Census," Evaluation and Research Report PHC80-E4, Bureau of the Census, Washington, D.C.

Marks, E., W. Seltzer and K. Krotki (1974), *Population Growth Estimation*, New York: The Population Council.

Wolfgang, G. S. (1989), "Using Administrative Lists to Supplement Coverage in Hard-to-Count Areas of the Post-Enumeration Survey for the 1988 Census of St. Louis." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Wolfgang, G. S. (1990), "Final Report on The 1988 Administrative List Supplement to the Post-Enumeration Survey in St. Louis," internal memorandum, Statistical Support Division, Bureau of the Census.

Wolter, K. M. (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association* 81:338-346.

Zaslavsky, A. M. (1989), "Multiple-System Methods for Census Coverage Evaluation," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Table 1: Three-source data

	Enumeration status for E, P, and A sources							
	$e = 1$				$e = 0$			
	$a = 1$		$a = 0$		$a = 1$		$a = 0$	
$p =$	1	0	1	0	1	0	1	0
O2	79	19	13	31	19	59	8	0
R2	58	12	69	41	11	43	34	0
O3	91	13	36	62	10	35	10	0
R3	72	7	69	32	13	43	24	0

Table 2: Cross-product ratios α with standard errors, and logarithms with jackknife standard errors.

	Poststratum			
	O2	R2	O3	R3
α for $e = 1$	9.92	2.87	12.06	4.77
α for $p = 1$	2.56	2.60	2.53	1.93
α for $a = 1$	12.91	18.89	24.50	34.02
$\log \alpha$ $(P \times A)$	2.29	1.06	2.49	1.56
S.E.	0.67	0.41	0.70	0.94
$\log \alpha$ $(E \times A)$	0.94	0.96	0.93	0.66
S.E.	0.56	0.45	0.72	0.67
$\log \alpha$ $(E \times P)$	2.56	2.94	3.20	3.53
S.E.	0.40	0.50	0.72	1.11

Table 3: Estimates for $\hat{x}_{000}$ under various dual and triple system models, with jackknife standard errors, and totals of observed cells.

	Poststratum			
	O2	R2	O3	R3
DSE w/o ALS	-44.3	-24.2	-23.2	-32.8
S.E.	17.2	14.8	10.4	15.1
DSE, P+ALS	24.0	26.0	24.4	17.3
S.E.	10.4	8.9	10.0	8.4
DSE with k_2	130.5	311.8	254.4	305.2
S.E.	63.5	171.0	201.9	431.5
ratio r_1	26.2	76.4	33.2	58.4
S.E.	10.5	26.6	10.4	28.8
ratio r_2	60.6	140.2	109.6	120.4
S.E.	26.2	53.3	54.6	83.0
est, $k_3 = 1$	246.3	381.7	421.9	378.7
S.E.	182.2	240.0	488.8	565.13
Sum w/o x_{000}	228.	268.	257.	260.
Sum w/o ALS	169.	225.	222.	217.
DSE w/o ALS	254.2	344.4	290.2	318.4

Table 4: Estimates of gross E-source (census) coverage rates under various dual and triple system models, with jackknife standard errors.

	Poststratum			
	O2	R2	O3	R3
DSE w/o ALS	0.773	0.738	0.864	0.792
S.E.	0.039	0.040	0.032	0.039
DSE, P+ALS	0.563	0.612	0.718	0.649
S.E.	0.061	0.058	0.041	0.065
DSE with k_2	0.396	0.310	0.395	0.318
S.E.	0.081	0.090	0.146	0.206
ratio r_1	0.559	0.523	0.696	0.565
S.E.	0.068	0.076	0.045	0.088
ratio r_2	0.492	0.441	0.551	0.473
S.E.	0.069	0.075	0.085	0.126
est, $k_3 = 1$	0.299	0.277	0.298	0.282
S.E.	0.118	0.097	0.178	0.213