USE OF HARD-TO-COUNT SCORES AND INCLUSION PROBABILITIES TO IMPROVE DUAL SYSTEM AND CENSUS PLUS ESTIMATES

Rita Petroni, Anne Kearney, and J. Gregory Robinson, Bureau of the Census*
Rita Petroni, Bureau of the Census, Room 3000-4, Washington, DC 20233-9100

Key Words: post-stratification, heterogeneity bias, fourth cell estimates, logistic regression

1. Introduction

Dual-system and Census Plus estimation are alternative techniques the Census Bureau has used to obtain census estimates (Thompson, 1994) and evaluate the completeness of census coverage of population. Both techniques assume the probabilities of enumeration are the same for all persons of the population. Since enumeration probabilities vary by age, sex, race/ethnicity, tenure, and geographic area, the Census Bureau post-stratifies the evaluation samples and the census by these characteristics to define subsets of the population which have more homogeneous enumeration probabilities to reduce heterogeneity bias. However, Alho, et.al. (1993) and Robinson (1996) provide evidence of residual heterogeneity bias after implementation of this post-stratification. Since the Census Bureau will use Dual System Estimation (DSE) for the 2000 Population Census, we are conducting research to identify a way to reduce heterogeneity bias for the 2000 Census.

Additionally, Bell (1991) noted that in the 1990 Census the Census Bureau obtained some negative Dual System Estimates (DSE) of the number of persons missed by both the census and the evaluation sample (i.e. the fourth cell estimates). The Bureau also obtained some negative fourth cell estimates in the 1995 Test Census. Theoretically this can occur because of sampling errors (Bell, 1991). It may also occur if the data reported by census and the evaluation interview differ, hence resulting in differing post-stratification classifications for census and the evaluation survey. If post-strata can be formed to reduce the mean square error, we may reduce the "negative fourth cell problem".

Using the 1995 Test Census data for the Oakland, California site, we researched the potential of alternative post-stratification schemes to reduce heterogeneity bias in DSE and Census Plus estimates and, secondarily, the negative fourth cell phenomenon for DSE. The alternatives build upon the Hard-to-Count (HTC) score and inclusion probability concepts developed respectively by Robinson and Alho et.al.

Section 2 introduces the concepts of DSE, HTC scores, and inclusion probabilities. Sections 3 and 4 respectively discuss our research methodology and results. In section 5, we conclude by summarizing the results, outlining ongoing research, and considering implications for future censuses and surveys.

2. Background

2.1 Dual System Estimation

For DSE in the 1995 Test Census, the Bureau independently obtained a listing of persons from a sample of block clusters (Integrated Coverage Measurement (ICM) blocks) after the regular census enumeration. The Bureau matched persons on the listing (P-sample persons) to the census enumerations to classify each person as being included or not in the census enumeration and as being included or not in the P-sample as follows:

Census Enumeration

P-Sample	In	Out	Total
In	N_{11}	N_{12}	$N_{1.}$
Out	N_{21}	N_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

Except for N_{22} (the fourth cell), all other internal cell values were observable. Hence, under the assumption of independence between inclusion in the census and the P-sample, we modeled the DSE population total as:

$$N_{..} = \frac{N_{.1}\ N_{1.}}{N_{11}} \quad . \quad (Wolter, 1986.)$$

In the DSE model, $N_{.1}$ is the number of distinct and identifiable census persons. The official census count, however, includes erroneous persons and imputations. Hence, the Bureau used the census enumerations in the ICM blocks (E-sample persons) to estimate and adjust for the proportion of erroneous census enumerations and excluded whole-person census imputations.

To reduce heterogeneity bias resulting from the assumption that all persons have the same enumeration probabilities, the population totals were determined separately by post-strata defined by age/sex (0-17, 18-29 male, 18-29 female, 30-49 male, 30-49 female, 50+male, 50+female), race/ethnicity (Black, Non-Black Hispanic, Asian Pacific Islander, Other), tenure (Owner, Renter) and geographic (Oakland) groups.

2.2 Hard-to-Count Scores

The use of race/ethnicity by tenure by age/sex categories as post-strata for the 1995 Test Census is based on the critical assumption that the overall undercount estimate for a particular post-stratum group is the same for all locations within the site, regardless of the socioeconomic "makeup" of a particular neighborhood. However, Robinson's (1996) research suggests that inclusion of a socioeconomic dimension beyond race and tenure is needed to minimize heterogeneity bias.

Robinson's research made use of composite HTC scores which are based on twelve housing and population/socioeconomic variables related to undercounting (i.e. % of persons living in renter-occupied units, % 25+ not having high school diploma or equivalent, % below poverty threshold; % of occupied buildings with 10+ units; % of occupied units without a telephone, % with 6+ household members, % where householder moved in 1989-90; % of total housing units that are vacant; % of households which are not married couple families, % reporting receipt of public assistance benefits, % where no persons 14+ speak English well; % of civilian population 16+ unemployed.) Composite HTC scores are assigned in several stages. First, each quantified variable is sorted from high to low across all block groups. Then, the variable value for each block group is assigned a HTC score based on a pre-defined scale. The scale assumes that variable values below the 50th percentile do not contribute to a block group being "hard-to-count" and is thus assigned a score of zero. Variable values above the 50th percentile are assigned scores of 1 to 9 in 5 percentile increments. For the top 5 percent of block groups where the distribution is more nonlinear, a value of 10 is assigned to the 95 to 97.5 percentile and a top score of 11 is given to blockgroups above the 97.5 percentile for the particular variable. Once the block groups have been ranked and assigned a HTC score for each individual variable, the scores for all variables are summed within each block group to form a composite HTC Score. (Robinson and Kobilarcik, 1995.)

The HTC scores can be assigned at any geographic level. Our research uses the block group composite HTC score and a similarly defined tract-level composite HTC score to define alternative 1995 Test Census post-strata.

2.3 Inclusion Probabilities

Alho et.al. (1993) used conditional logistic regression to model and estimate probabilities of being enumerated in the 1990 Census and then applied them to measure heterogeneity bias in the Post Enumeration Survey estimates of the 1990 Census. The research showed evidence of residual heterogeneity.

Alho et.al. used logistic regression to model P_{1i} and P_{2i} which are inclusion probabilities corresponding respectively to the i^{th} person being included in the census and the P-sample population as follows (Mulry, et.al., 1991):

$$\log\left(\frac{P_{ji}}{1 - P_{ji}}\right) = X_{ji}^{T}\, a_j$$

where a_j are vectors of parameters estimated iteratively via Newton's method and X_{ji} are vectors of explanatory variables correlated with inclusion.

Our research applies models based on this methodology and 1990 Census data to 1995 Test Census data to obtain inclusion probabilities which we use to define alternative 1995 Test Census post-strata. We used these explanatory variables in nonminority owner, non-minority renter, minority owner, and minority renter models (underlined variables were standardized to provide a level of magnitude equivalent to the indicator variables): intercept, actual age (age), sex, Black [1, 2], Hispanic and not Black [1, 2], household size (HH), % persons in block living in rented units (% renter), % persons in block who are Black or Nonblack Hispanic (% B/NB Hisp), % of persons in multi-unit structures (% multi), % of vacant housing units in block, age*Black [1, 2], % renter*%B/NB Hisp [1, 3], living in central cities in primary metropolitan statistical areas [1], age*age, age*age*age [1, 2], age*sex, not related to person who completed the questionnaire, age*HH [1], living in Northeast Region [1], living in South Region [1], living in West Region [1], % B/NB Hisp*% multi [3], % renter*% multi. 1, 2, and 3 indicate respectively that the variable was not included in the non-minority owner, non-minority renter, or minority owner model.

3. Research Methodology

For this research, we formed post-strata and estimates for the Oakland 1995 Test Census and ICM data based on:

- ranges of tract-based composite HTC scores
- ranges of block group-based composite HTC scores
- ranges of tract-based or block group-based composite HTC scores crossed by demographic characteristics (i.e. tenure, race/ethnicity, age/sex)
- ranges of inclusion probabilities
- ranges of HTC scores crossed by ranges of inclusion probabilities
- ranges of inclusion probabilities crossed by demographic characteristics (i.e tenure, race/ethnicity)

For each post-stratification scheme, we recompute 1995 DSE and CensusPlus estimates. To evaluate the ability of the 1995 post-stratification and each new alternative post-stratification to reduce heterogeneity bias, we use VPLX to compare the coverage ratios of each post-stratification scheme to the classical ICM coverage ratios. The classical ICM coverage ratios are computed assuming all persons belong to the same post-stratum. We also use VPLX to compare the coverage ratios of the new alternative post-stratifications to the coverage ratios of the 1995 post-stratification. Thirdly, for estimates based on the 1995 post-stratification and selected alternative post-stratifications, we compare average tract level undercount rates by enumeration difficulty to assess which alternative works best at lower levels of geography.

Following the precedent of Mulry et al. and Alho et al., we will use 1995 logistic regression-based population estimates as benchmarks to assess which alternative best addresses heterogeneity bias. We will apply the basic modeling approach of Alho et.al. to 1995 Test Census data to obtain inclusion probabilities and then use these probabilities to obtain the LREs as follows:

$$\sum_{n} \frac{1}{\hat{\varphi}_i}$$

where $\hat{\varphi}_i$ is the estimated census inclusion probability based on p_{1i} and p_{2i} and adjusted for cases in the p-sample only and for erroneous enumerations for person i and n includes all non-group quarters persons.

To evaluate the ability of each alternative to reduce the "negative fourth cell problem", we compare the fourth cell estimates and percent of adjustment cells with significantly negative fourth cells.

To accommodate tract level evaluations, we use the census file which was imputed for nonsampled nonresponse cases to obtain all estimates. As a result, original estimates in this paper differ from original estimates which were derived from the census file which contains weights to adjust for non-sampled nonresponse cases.

We divided the HTC scores into three broad ranges and the inclusion probabilities into four ranges for post-stratification purposes.

4. Research Results

4.1 Restrictions and Keys

Since the Census Plus estimates proved to be implausible, we present analysis for DSEs only. Also, since estimates based on block group and tract HTC scores are essentially the same, we only present results for the estimates based on tract HTC scores.

In tables 1-4,

1995 =	1995 post-stratification estimate
Classical =	estimate based on the assumption that all persons belong to the same post-stratum
HTC =	estimate based only on post-stratification by HTC ranges
HTC/R =	estimates based on post-stratification by HTC ranges and race/ethnicity categories
HTC/T =	estimates based on post-stratification by HTC ranges and tenure categories
HTC/R/T =	estimates based on post-stratification by HTC ranges and race/ethnicity and tenure categories

Although the tables do not show them, we also produced estimates based on post-stratification by HTC ranges and age/sex categories, by HTC ranges and race/ethnicity and age/sex categories, and by HTC ranges and tenure and age/sex categories. Results for these correspond very closely to results for HTC, HTC/R, and HTC/T respectively.

4.2 Comparisons of Site-Level Coverage Ratio Estimates to Classical Coverage Ratio Estimates

Tables 1-4 show that total (table 1), Black, Hispanic (table 2), renter (table 3), Hispanic renter, and Other owner (table 4) coverage ratios differ significantly from the classical coverage ratios for all alternative post-stratifications. This observation supports the belief that the 1995 and each new alternative post-stratification may address heterogeneity bias better than the assumption that all persons belong to the same post-strata.

4.3 Comparison of New Alternative Site-Level Coverage Ratio Estimates to 1995 Coverage Ratio Estimates

From table 1, we see that none of the new alternative site total coverage ratios differ significantly from the 1995 coverage ratio. Table 2 shows that for all race/ethnicity coverage ratios, the HTC/R/T alternative produces coverage ratios which do not differ significantly from the 1995 alternative. Table 3 shows that all alternatives preserve the correct relationship in coverage ratios between owners and renters and that the owner and renter coverage ratios do not differ significantly from the 1995 coverage ratios for the HTC, HTC/R, and HTC/R/T alternatives.

From table 4, we see:

- For Blacks, all alternatives preserve the correct directional relationship in coverage ratios between owners and renters and the coverage ratios do not differ significantly from the 1995 coverage ratios for the HTC, HTC/R, and HTC/R/T alternatives.
- For Hispanics, all alternatives preserve the correct directional relationship in coverage ratios between owners and renters and the coverage ratios do not differ significantly from the 1995 coverage ratios for the HTC/R and HTC/R/T alternatives.
- For Asian Pacific Islander, the HTC, HTC/T, and HTC/R/T alternatives produce coverage ratios which are not statistically different from the 1995 coverage ratios and preserve the correct directional relationship in coverage ratios between owners and renters.
- For Other, the HTC/R and HTC/R/T alternatives produce coverage ratios which do not differ statistically from the 1995 coverage ratios and preserve the correct directional relationship in coverage ratios between owners and renters.

Thus, none of the new alternative site-level total coverage ratios differ statistically from the 1995 coverage ratio. Various of the new alternatives produce about the same site-level race/ethnicity, tenure, or race/ethnicity by tenure coverage ratios as the 1995 alternative. Only the HTC/R/T alternative produces site-level coverage ratios comparable to the 1995 alternative in all instances.

At least for certain estimates, it is possible that alternatives other than the 1995 or HTC/R/T alternatives may better reduce heterogeneity bias. For example Hispanic, coverage ratios produced by the HTC and HTC/T, alternatives differ from the 1995 coverage ratios, but they may be more plausible. For Blacks, the coverage ratios for owners and renters may have a more reasonable numerical relationship for the HTC/T alternative. Hispanic tenure coverage ratios produced by the HTC and HTC/T alternatives may be better.

To help judge among the alternatives, we analyze tract level results below. Also, comparison of the above results to the LREs should help answer whether the 1995 post-stratification or one of the new alternatives best addresses heterogeneity bias.

4.4 Comparisons of Tract Level Undercover Rates

We compared tract-level undercoverage rates for the 1995 and selected alternative post-stratifications grouped by easy, medium, and difficult-to-enumerate tracts for tenure and race/ethnicity by tenure. As the following table shows for Black owners and Black renters, the 1995 post-stratification does not capture differences in average tract level undercoverage rates by ease of enumeration, while the HTC/T alternative does. Similarly, the HTC, HTC/R, and HTC/R/T alternatives captured differences in average tract level undercoverage rates for Black owners and Black renters (not shown). For other race by tenure groups this did not hold. This may be due to the small sample size for persons in these groups.

These results along with results from section 4.3 suggest that while site level 1995 and HTC/R/T coverage ratios do not differ significantly, tract-level undercoverage rates may be affected differently. They also suggest that the poststrata are not adequately capturing heterogeneity in undercoverage rates at the tract level.

Average Tract-Level Undercoverage Rates for Alternative DSE Estimates by Hard-to-Count Range

Estimate/ Range	Population in Thousands	Alternative DSE Estimates	
		1995	HTC/T
Black Owner	51.2	8.9	5.7
Difficult	12.0	8.8	9.3
Medium	22.5	8.7	6.0
Easy	16.7	9.0	2.4
Black Renter	88.5	9.8	9.7
Difficult	35.4	9.9	12.6
Medium	38.1	9.8	8.2
Easy	15.0	9.7	6.4

4.5 Comparisons of Fourth Cell Estimates

Table 1 shows that all alternative post-stratifications produce improved fourth cell estimates compared to the classical fourth cell estimate. Statistically, none of the new alternatives produce improved fourth cell estimates compared to the 1995 fourth cell estimate. There is also little difference in the percent of significantly negative adjustment cells - 17% or less. However, it is interesting to note that while the 1995 post-stratification has 4.7 and 2.3 times as many cells as the HTC/R and HTC/R/T post-stratifications respectively, it has 8 and 4 times as many significantly negative fourth cells.

5. Conclusions

Currently available research results based on the Oakland, California test site show that the new alternative post-stratifications based on HTC scores may potentially reduce heterogeneity bias at lower geographic levels. The results also suggest that the post-stratifications have little impact on the "negative fourth cell" problem.

Admittedly, the results are preliminary. We are currently producing estimates based on ranges of inclusion probabilities and comparing them to the other alternatives. We also are comparing the site-level estimates to LREs to help assess which alternatives produce the least biased estimates. Additionally, since current results suggest the potential to reduce heterogeneity bias at lower geographic levels, we may research this in one of two ways. The first is to repeat this research using the 1990 Post Enumeration Survey. This will increase the number of cases in the various race/ethnicity by tenure groups. The second is to research the use of the HTC ranges or inclusion probability ranges as a third dimension or an alternative classification variable of the raking approach proposed by Schindler and Griffin (1997). This approach offers a potential way to reduce the small sample size concern.

If research on the potential to reduce heterogeneity bias at lower geographic levels supports our initial findings, we will consider using the HTC scores or inclusion probabilities in forming post-strata for evaluating the 2000 population census. If our findings are positive, other censuses and surveys which produce small geographic level estimates should consider our approach.

References

Alho, J.M., Mulry, M.H., Wurdeman, K. and Kim, J. (1993). Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. Journal of the American Statistical Association, Volume 88, pp. 1130-1136, September 1993.

Bell, W. (1991). Use of Alternative Dual System Estimators to Measure Correlation Bias. 1990 Coverage Studies and Evaluation Memorandum Series #0-3, Project P13, July 10, 1991, Bureau of the Census.

Mulry, M., Wurdeman, K., Kim, J. and Alho, J.A. (1991). Estimation of Correlation Bias Using Conditional Logistic Regression. Proceedings of the Social Statistics Section, American Statistical Association, pp. 260-265.

Robinson, J.G. (1996). Demographic Review of the Housing and Population Results of the 1995 Test Census. Population Division's Review of 1995 Census Test Results, Memorandum No. 2, March 12, 1996, Bureau of the Census.

Robinson, J.G. and Kobilarcik, E.L. (1995). Identifying Differential Undercounts at Local Geographic Levels: A Targeting Database Approach. Presented at the 1995 annual meeting of the Population Association of America, San Francisco, California.

Schindler, E. and Griffin, R. (1997). Census 2000: Stratification and Poststratification. Presented at the 1997 annual meeting of the American Statistical Association, Anaheim, California.

Thompson, J. (1994). 1995 Test Census Design. Presented at the 1994 annual meeting of the American Statistical Association, Toronto, Canada.

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

Acknowledgments

The authors wish to thank George McLaughlin and Diana Smith for programming support, Tina Arbogast for typing our paper, and Rajendra Singh and Yves Thibaudeau for reviewing our paper.

* 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.

Table 1. 1995 DSE Total Population Estimates [1] and Coverage Ratios Based on Alternative Post-Stratification, Oakland, California

Alternatives	Estimate	C.R.[2]	Fourth Cell Estimates[3]	Number of Adjustment Cells	Adjustments with 4th Cells Significantly <0	
					%	#
1995	369,554	1.087^{c}	14,302^{c}	56	14	8
Classical	363,462	1.069^{x}	6,270^{x}	1	0	0
HTC	366,632	1.078^{c}	10,319^{b}	3	0	0
HTC/R	367,358	1.080^{c}	11,501^{c}	12	8	1
HTC/T	368,464	1.084^{c}	13,216^{c}	6	17	1
HTC/R/T	369,482	1.086^{c}	14,390^{c}	24	8	2

Table 2. 1995 DSE Race/Ethnicity Estimates [1] and Coverage Ratios [2] Based on Alternative Post-Stratifications, Oakland, California

Alternative	Black		Hispanic		Asian Pacific Islander		Other	
	Estimate	C.R.	Estimate	C.R.	Estimate	C.R.	Estimate	C.R.
1995	154,025	1.105^{c}	58,272	1.203^{c}	60,136	1.075	97,121	1.007^{c}
Classical	149,138	1.069^{x}	51,816	1.069^{x}	59,808	1.069	102,880	1.069^{x}
HTC	151,143	1.084^{c}	53,207	1.098^{b}	61,041	1.091^{c}	101,241	1.051^{b}
HTC/R	153,251	1.099^{c}	58,485	1.208^{c}	59,163	1.058^{x}	96,459	1.000^{c}
HTC/T	151,969	1.090^{c}	53,436	1.103^{b}	61,275	1.096^{c}	101,784	1.057^{x}
HTC/R/T	153,378	1.100^{c}	59,203	1.223^{c}	60,043	1.073	96,858	1.005^{c}

Table 3. 1995 DSE Tenure Estimates [1] and Coverage Ratios [2] Based on Alternative Post-Stratifications, Oakland, California

Alternative	Owner		Renter	
	Estimate	C.R.	Estimate	C.R.
1995	155,191	1.060	214,363	1.108^{c}
Classical	156,520	1.069	207,122	1.069^{x}
HTC	155,839	1.064	210,793	1.089^{c}
HTC/R	156,197	1.067	211,161	1.091^{c}
HTC/T	154,026	1.052^{x}	214,438	1.108^{c}
HTC/R/T	154,661	1.056	214,821	1.110^{c}

Table 4. 1995 DSE Race/Ethnicity x Tenure Estimates[1] and Coverage Ratios[2] Based on Alternative Post-Stratifications, Oakland, California

Race/Ethnicity	Alternative	Owner		Renter	
		Estimate	C.R.	Estimate	C.R.
Black	1995	56,172	1.097	97,853	1.109
	Classical	54,735	1.069	94,403	1.069
	HTC	54,994	1.074	96,149	1.089[c]
	HTC/R	55,870	1.091	97,381	1.104[c]
	HTC/T	54,280	1.060[x]	97,689	1.107[c]
	HTC/R/T	55,806	1.090	97,572	1.106
Hispanic	1995	20,644	1.176[c]	37,628	1.219[c]
	Classical	18,773	1.069[x]	33,043	1.069[x]
	HTC	19,130	1.089[b]	34,077	1.103[b]
	HTC/R	20,957	1.193[c]	37,528	1.216[c]
	HTC/T	18,835	1.072[x]	34,601	1.120[c]
	HTC/R/T	20,757	1.182[c]	38,446	1.246[c]
API	1995	24,055	1.034	36,081	1.105
	Classical	24,880	1.069	34,928	1.069
	HTC	25,102	1.075	36,029	1.103[c]
	HTC/R	24,946	1.072	34,217	1.047
	HTC/T	24,676	1.060	36,599	1.121[c]
	HTC/R/T	23,975	1.030	36,068	1.104
Other	1995	54,320	.999[c]	42,801	1.019
	Classical	58,132	1.069[x]	44,748	1.069
	HTC	56,703	1.043[b]	44,538	1.064
	HTC/R	54,424	1.001[c]	42,035	1.000[c]
	HTC/T	56,235	1.034[b]	45,549	1.090[x]
	HTC/R/T	54,123	.995[c]	42,735	1.018[c]

[1] Estimates include persons in group quarters.
[2] Coverage ratio = estimate/census estimate.
[3] Estimated number missed by Census and ICM.
[c] Statistically different from the classical, but not from the original estimate ($\propto$ = .10).
[x] Statistically different from the original, but not from the classical estimate ($\propto$ = .10).
[b] Statistically different from both the original and classical estimates ($\propto$ = .10).