Iterative Proportional Fitting in the Census 2000 Dress Rehearsal

Eric Schindler, Bureau of the Census, Washington DC 20233

KEYWORDS: Dual System Estimation, Poststratification, Raking

Iterative proportional fitting, or raking, was employed in addition to the dual system estimation methodology to measure the undercoverage for the Census 2000 Dress Rehearsal conducted during 1998 in three sites. The raking procedure was used to adjust the initial phase estimates for poststrata defined by race/origin/age/sex/tenure to two sets of marginals defined by race/origin/age/sex and tenure estimated by taking the sums of direct dual system estimates for the same poststrata. This procedure was designed specifically to improve reliability and preserve the race/origin/age/sex cells required for congressional and state redistricting and to induce approximately the same coverage differences between owners and renters for each demographic group. This paper discusses the results of the procedure and of several alternative raking matrices with a view towards Census 2000.

I. Introduction

Two years before each decennial census, the Bureau of the Census executes a Dress Rehearsal, a full scale implementation of the census in several small sites. The dress rehearsal is designed as a final test for operations, forms, estimation, and data publication. In theory, after the dress rehearsal only minimal adjustments to correct serious shortcomings should be implemented. The political controversy surrounding Census 2000 has made it difficult for the Census 2000 Dress Rehearsal to follow this prescribed course.

Two of the three Dress Rehearsal sites, Sacramento, CA and Menominee County, WI, mostly an American Indian reservation, used the Integrated Coverage Measurement (ICM) design, which based on a recent Supreme Court ruling will not be used for reapportionment in 2000. The third site, Columbia, SC and several surrounding counties, was collected as a traditional census followed by a Post Enumeration Survey (PES). While the operations and results in Sacramento and Menominee were much as expected, it appears that there was substantial undercoverage in the address listings in the mail delivery areas outside of Columbia, leading to higher than expected estimated undercount rates for owners.

For all sites, the coverage survey was collected in a stratified (by predominant race and home ownership) sample of blocks or block clusters averaging about 30 housing units. The sampled housing units were reinterviewed independently. The persons counted in the initial phase and coverage collection efforts were compared to determine which persons in the initial phase were correctly enumerated and which persons in the coverage survey could be matched to the initial collection.

Estimation of coverage for both designs was by a poststratified dual system estimator. Poststrata were defined by 6 race/Hispanic origin groups, 7 age/sex groups, and tenure, the owner/renter dichotomy. Raking, or iterative proportional fitting (first suggested for the Decennial Census in Schindler and Griffin, 1997), was implemented on a 42 by 2 matrix to help control the standard errors.

Section II discusses the results of the estimation and raking procedures at the site and poststratum level. Section III examines several alternative estimation poststratification and raking options. Section IV concludes the paper by discussing current plans for Census 2000.

II. The Census 2000 Dress Rehearsal

POSTSTRATIFICATION

Poststrata were defined using race, Hispanic origin, age, sex, and tenure. These variables, plus geographic identifiers had been used in the 1990 PES. The six race/Hispanic origin groups are:

1. Non-Hispanic White/Other
2. Non-Hispanic Black
3. Non-Hispanic American Indian/Alaska Native
4. Non-Hispanic Native Hawaiian/Pacific Islander
5. Non-Hispanic Asian
6. Hispanic (of any race)

Persons who marked more than one race box were assigned during estimation to the largest non-white race marked based on 1990 site level census counts. Race/origin groups which were less than 1% of the site total in 1990 were collapsed during estimation into the

---

Eric Schindler is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

largest non-white race based on the 1990 census counts. In Sacramento, American Indians and Native Hawaiians were combined with Hispanics. In Columbia, all minority races were combined with Blacks. In Menominee, Blacks, Native Hawaiians, and Asians were combined with American Indians.

The seven age/sex groups were:

0-17	18-29 Male	18-29 Female
	30-49 Male	30-49 Female
	50+ Male	50+ Female

In the Menominee site some of the age/sex categories were combined for White/Others and for Hispanics.

DUAL SYSTEM ESTIMATION

The site level dual system estimate for the three Census 2000 Dress Rehearsal Sites was defined by:

$$\hat{DSE}_i = C_i \times \frac{\hat{D}_i}{\hat{C}_i} \times \frac{\hat{CE}_i}{\hat{E}_i} \times \frac{\hat{P}_i}{\hat{M}_i}$$

where: C_i is the initial count in poststratum i.

$\hat{C}_i$ is the initial count in poststratum i as estimated from the initial count in the coverage sample areas

$\hat{D}_i$ is the estimated number of initial count persons for whom at least some data was directly collected - only data defined persons are included in the E-Sample

$\hat{E}_i$ is the estimated number of persons in the enumeration sample in poststratum i. This is approximately equal to $\hat{D}_i$

$\hat{CE}_i$ is the estimated number of E-Sample persons in poststratum i who are determined to have been correctly enumerated in the initial collected effort.

$\hat{P}_i$ is the estimated number of persons in the coverage survey in poststratum i.

$\hat{M}_i$ is the estimated number of P-Sample persons in poststratum i who can be matched to a person in the initial collection effort.

Because of the special treatment required for persons who move between Census Day and the time of the coverage interview, the last term of the equation is actually somewhat more complicated:

$$\frac{\hat{P}_i}{\hat{M}_i} = \frac{\hat{P}_{i,nonmover} + \hat{P}_{i,inmover} \times \frac{\hat{P}_{i,outmover}}{\hat{P}'_{i,outmover}}}{\hat{M}_{i,nonmover} + \hat{P}_{i,inmover} \times \frac{\hat{M}_{i,outmover}}{\hat{P}_{i,outmover}} \times \frac{\hat{P}_{i,outmover}}{\hat{P}'_{i,outmover}}}$$

where nonmover, inmover, and outmover are obviously defined and $\hat{P}'_{i,outmover}$ is the estimated number of persons originally enumerated in the P-Sample including some who are later determined to have not been residents. This treatment of movers, known as DSE C, uses the best available data to estimate the number of movers (from the inmovers) and the best available data to estimate the residence and match probabilities (from the outmovers, collected by proxy).

RAKING

A simple raking or iterative proportional fitting procedure was executed to help control the variances. The raking matrix was defined by the collapsed race/Hispanic origin by age/sex categories in the first dimension and tenure in the second dimension. The marginal controls were calculated by adding the dual system estimates for the interior cells. The initial phase estimates were then raked to the marginal controls.

The variance estimates in this paper were estimated by a simple Jackknife procedure which gives similar results to the stratified Jackknife used in the official estimates. Bias^2 at the poststratum level for the raked alternatives is estimated by: (DSE-Rake)^2 - $VAR_{DSE\text{-}Rake}$ and the estimates are often negative; MSE is estimated by VAR_{Rake} + Bias^2 and the estimates are sometimes negative.

DRESS REHEARSAL SUMMARY RESULTS

Table 1 shows some summary results, excluding persons in group quarters, of the initial phase, dual system estimates and raked estimates for the three Dress Rehearsal sites. Past experience has shown that coverage in test censuses is generally worse than in the actual census.

Table 1: Summary Results: Dress Rehearsal Estimation

Sacramento	Initial	DSE	%UC	rakeDSE	%UC
Total	369434	395005	6.47%	395005	6.47%
owner	188202	194398	3.19%	194398	3.19%
renter	181232	200608	9.66%	200608	9.66%
White/Other	160620	168555	4.71%	168555	4.71%
owner	91545	94020	2.63%	93703	2.30%
renter	69075	74535	7.32%	74851	7.72%
Black	59005	64647	8.73%	64647	8.73%
owner	22355	23141	3.40%	23350	4.26%
renter	36650	41506	11.70%	41297	11.25%
Asian	58890	62643	5.99%	62643	5.99%
owner	33057	34153	3.21%	34151	3.20%
renter	25833	28490	9.32%	28492	9.33%
Hispanic +	90919	99161	8.31%	99161	8.31%
owner	41245	43083	4.27%	43193	4.51%
renter	49674	56078	11.42%	55967	11.24%

Menominee	Initial	DSE	%UC	rakeDSE	%UC
Total	4550	4694	3.06%	4694	3.06%
owner	2937	3026	2.93%	3026	2.93%
renter	1613	1668	3.30%	1668	3.30%
Amer Ind +	3859	4024	4.10%	4024	4.10%
owner	2327	2485	6.34%	2450	5.02%
renter	1532	1540	0.49%	1574	2.68%
Columbia	Initial	DSE	%UC	rakeDSE	%UC
Total	628616	693724	9.39%	693724	9.39%
owner	452310	507865	10.94%	507865	10.94%
renter	176306	185858	5.14%	185859	5.14%
White/Other	359854	384073	6.31%	384073	6.31%
owner	286891	308384	6.97%	310670	7.65%
renter	72963	75688	3.60%	73403	0.60%
Black +	268762	309651	13.20%	309651	13.20%
owner	165419	199481	17.08%	197195	16.11%
renter	103343	110170	6.20%	112455	8.10%

+ : In Sacramento Native Hawaiians and American Indians were collapsed with the Hispanic population.
In Menominee, Blacks, Asians, and Native Hawaiians were collapsed with the American Indian population.
In Columbia all groups except non-Hispanic Whites were collapsed with the Black population.
Rounding may cause slight differences between the published results and those presented here.

The results for Sacramento were consistent with those observed in the 1990 PES. The estimated undercount rates were higher for renters than for owners and higher for minorities than for Whites. The total undercount (6.47%) was higher than the 1990 PES, but this is consistent with experience from the 1990 Dress Rehearsal. The raking procedure widened the differential between White owners and renters and narrowed it between Black or Hispanic owners and renters. The standard errors for the raked poststrata average about 80% of the preraking values. The unweighted mean square errors of the 56 postcollapsing poststrata after raking average only 22% of the unraked poststratum variances, indication significant bias reduction.

Although similar 17% and 46% reductions in the average poststratum level standard errors and mean square errors, respectively, occurred, the estimation in the Columbia site did not proceed nearly as smoothly as in Sacramento. The overall undercount rate was about 3% higher. The undercount rate for minorities (almost all Black) was higher than that for Whites, but the undercount rate for owners was higher than that for renters for both race groups. This was especially so for the poststrata for owners under 50 years old. The poststrata for white older persons showed the expected higher undercount rates before raking for renters than for owners. For minorities, the owners over 50 years old had higher undercounts than the over 50 renters but not by as much for the under 50 owners.

Table 2: Undercount Rates by Tenure in Columbia

Race	Age	Owner	Renter
Non-Hispanic White	<50	9.22%	2.57%
	≥50	3.25%	6.43%
All Others	<50	18.78%	5.88%
	≥50	11.74%	5.54%

Investigation showed that the estimates for just the city of Columbia were consistent with those for Sacramento. Also, the estimates for the most rural areas of the site where the Census Bureau enumerators updated the address list while delivering forms were also acceptable. The bulk of the problem occurred for the 60% of the site population who lived in mailout/mailback areas outside of Columbia where the addresses were provided by the Postal Service. It appears that many housing units near the edges of these mailout/mailback areas were not reported. These units are mostly owner occupied. This problem will be addressed in Census 2000 by redesigning the creation of the Master Address File, a major component of which will be a block canvassing operation searching for additional housing units nationwide.

Seven block clusters (out of 665 total block clusters) with varying collection problems, all in the mailout/mailback areas outside of Columbia or the update/leave areas, had a disproportionate impact on the estimates. Table 3 shows the Columbia estimates omitting these seven block clusters. The owner/renter reversal has been greatly reduced, from about 5.8% to 0.7% at the site level and from 8.1% to about 2.8% at the poststratum level.

Table 3: Columbia Results Omitting Outliers

Columbia	Initial	DSE	%UC	RakeDSE	%UC
Total	628616	689593	8.84%	689593	8.84%
owner	452310	497283	9.04%	497283	9.04%
renter	176306	192310	8.32%	192311	8.32%
White	359854	383402	6.14%	383402	6.14%
owner	286891	306135	6.29%	306585	6.42%
renter	72963	77266	5.57%	76817	5.02%
Black	268762	306192	12.22%	306192	12.22%
owner	165419	191147	13.46%	190698	13.26%
renter	103343	115044	10.17%	115494	10.52%

The reversal of undercoverage rates for owners and

renters from the expected direction in Menominee has not been thoroughly investigated.

III. Design Alternatives

Raking, as applied in the Census 2000 Dress Rehearsal, approximately imposed a uniform difference in estimated undercount rates on all pairs of owner/renter poststrata. This difference was about 5.5% in Sacramento and -8.75% in Columbia. It is possible to lessen this consistent effect by defining more marginal cells in the second dimension of the raking matrix. Recall that the first dimension was defined by the six race/Hispanic origin categories crossed by the seven age/sex categories which were then collapsed to eliminate small cells. The second dimension was defined by the two tenure cells only. Two types of additional variables can be defined in the second dimension. The first is designed to capture interactions between race/origin or age/sex and tenure by using reduced versions of the variables in the first dimension. The second type of variable, "new" variables, would allow the formation of more homogeneous poststrata.

A. Interactions

A race/age/tenure interaction was observed in Columbia which was masked by the raking. Therefore, two additional dichotomous variables were defined: over/under 50 and nonminority/minority. These additional variables can be included in the second dimension of the raking matrix, yielding eight marginal cells instead of two. Since these variables are already included in the first dimension of the raking matrix in an expanded form there is no change to the before raking estimates at summary levels at the race/origin by age/sex level. The addition of these variables to the second dimension of the raking matrix retains most of the gains of raking for standard errors and reduces the changes made from the before raking poststratum estimates.

The estimates for Black children in Sacramento are typical, with lower MSEs and almost 100 fewer persons moved from renter to owner when the interactions are included. With the interactions in the raking procedure, the estimates for Black children are less biased (in fact, the estimate of $BIAS^2$ is negative), have lower mean square errors, and preserve almost all of the reduction in standard error of the simple raking procedure.

Table 4A: Estimates for Black Children in Sacramento

Estimate	Owner	Renter
Initial Estimate	6477	14855
DSE before Raking (se) Undercount Rate	6624 (265) 2.22%	17203 (747) 13.65%
Rake by Tenure (se) Undercount Rate RMSE / BIAS Difference from DSE	6942 (204) 6.70% 259 / 161 +318	16885 (605) 12.02% 626 / 161 -318
Rake by Tenure, Minority Status, and +/- 50 (se) Undercount Rate RMSE / BIAS Difference from DSE	6857 (209) 5.54% 190 / n/a +233	16971(628) 12.47% 621 / n/a -233

The real advantage of including the interactions occurs in Columbia where the results were not as expected. White owners had worse coverage than White renters, but not White owners over 50. Raking by just tenure forced the older population to follow the overall pattern. For White males over 50, 781 persons, 1.6% of the before raking DSE, were moved from the unraked renter poststratum to the raked owner poststratum and error estimates increased substantially with large biases. Raking with race and age as well as tenure in the second marginal includes the interactions , stops the imposition of average coverage factors, significantly reduces the bias, reduces the MSEs especially for renters, and reduces the number of White males over 50 moved by raking to 159, only 0.3% of the population.

Table 4B:Estimates for White Males over 50 in Columbia

Estimate	Owner	Renter
Initial Estimate BIAS	46432 1697	5073 448
DSE before Raking (se) Undercount Rate	48317 (875) 3.90%	5670 (395) 10.53%
Rake by Tenure (se) Undercount Rate MSE / BIAS Difference from DSE	49097 (888) 5.43% 1118 / 678 +781	4889 (189) -3.76% 704 / 678 -781
Rake by Tenure, Minority Status, & +/- 50 (se) Undercount Rate MSE / BIAS Difference from DSE	48476 (878) 4.22% 868 / n/a +159	5511 (295) 7.95% 266 / n/a -159
$MSE = (Est-DSE)^2 - Var(Est-DSE) + Var(Est)$ $Bias^2 = (Est-DSE)^2 - Var(Est-DSE)$		

Table 5 summarizes the results by averaging over the 84 poststrata. In Sacramento, raking reduces the variances and the mean square errors substantially with or without the interactions, with no measurable bias in either case. In Columbia, raking reduces the estimated variances but the added bias increases the mean square error 18% overall and 30% for persons over 50. Including the interactions adds back about half of the variance reduction, but the reduced bias decreases the mean square error to a 13% increase overall and an 8% decrease for persons over 50. Including the interactions also decreases the number of persons shifted by raking, especially for those over 50. In Sacramento raking changes the 56 collapsed coverage factors by an average of 2.3%. Including the interactions decreases this change to 2.0%. In Columbia, the corresponding decrease for the 28 collapsed poststrata is from a 3.7% change to a 2.7% change; for the 8 collapsed poststrata for persons age 50 and over the decrease is from a 4.5% change to a 1.6% change. As expected, including the interactions is preserving the difference in coverage by tenure for those over or under age 50 which is lost if only tenure is included in the second dimension of the raking matrix.

Table 5: Average Statistics for Raking for two Options over 84 Poststrata (24 for over age 50)

	Sacra-mento	Columbia	
		Total	Over50
No Raking: $\overline{VAR}$	44929	356755	168830
Rake by tenure			
$\overline{VAR}$	30078	259479	156408
$\overline{moved\ by\ raking}$	39	112	165
$\overline{MSE}$	16640	421955	218656
$\overline{BIAS^2}$	-13437	162477	62248
Rake by tenure, minority status, and over/under 50			
$\overline{VAR}$	32821	307986	161927
$\overline{moved\ by\ raking}$	30	83	57
$\overline{MSE}$	20474	402549	155709
$\overline{BIAS^2}$	-12348	94563	-6218

The results comparing the raked estimates with or without the interactions to the unraked DSEs for the individual poststrata are displayed in the graphs at the end of the paper. Including the interactions had little effect in Sacramento where the corresponding squares and diamonds are fairly close to one another. However, in Columbia, including the interactions produces 10% differences for the White renters over age 50.

B. Additional Variables

Additional variables have been proposed for poststratification in the second dimension of the raking matrix for Census 2000. These include:

- Geographic variables. These could be for major areas such as Census regions, Census Divisions, or states, or for subareas such as urban versus nonurban areas or mailout/mailback versus update/leave areas. These variables were not applicable in Sacramento, but could have been important in the Columbia site. Investigation of these variables at the national level with the 1990 PES data can be found in Farooque (1999).
- Neighborhood characteristics such as mail return rate, percentage minority, or poverty rate. These can be used separately or combined into a short form or long form neighborhood "hard-to-count" score. Variables based on data from the Census 2000 "long form" would have to use 1990 data, while a "short form" score could use Census 2000 data. Farooque (1999) is finding that mail return and minority rates are statistically significant.
- Household composition variables which attempt to identify households and residents which are more likely to have good coverage. A simple variable with some effectiveness is whether the first two persons in a household are married. A more complex variable which is very significant in the logistic regression work in Farooque (1999) places single persons over 50 and married couples over 30 and their minor children (and then only if the only other persons in the household are older children and at most one elderly parent) in Class 1 and everyone else in Class 2. This particular variable is about as important as race/Hispanic origin or age/sex as an indicator of coverage. Unfortunately, these variables are influenced by coverage which tends to be better in the P-Sample. This results in more people being in Class 1 in the E-Sample and in Class 2 in the P-Sample. This imbalance leads to biases which cannot be eliminated. Unless a workable definition of this significant variable can be found, it probably should not be used.

IV. Conclusions

- Raking, or iterative proportional fitting, can be a valuable tool in reducing both the variance and the total error in the dual system estimation for Census 2000.
- Inclusion of interactions for race and age can control

the bias due to raking when the simpler raking model is inappropriate for a particular subpopulation as occurred in Columbia.

V. References

Farooque, Golam, (1999) " Variables Selection and Determination of Raking Dimensions," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, to appear.

Schindler, Eric and Griffin, Richard (1997), "2000 Census ICM: Stratification and Poststratification," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, pp689-694.