ALLOCATION OF THE ICM SAMPLE TO THE STATES FOR CENSUS 2000

Eric Schindler, Bureau of the Census
Bureau of the Census, Washington, DC 20233

KEYWORDS: Dual System Estimation, Reapportionment, Jackknife

ABSTRACT: The introduction of Integrated Coverage Measurement (ICM) for Census 2000 requires 51 state estimates based only on data from each state. The goal is to allocate the available sample of 750,000 housing units so as to achieve coefficients of variation for the Dual System Estimates of 0.5% in all states and standard errors of about 60,000 in the larger states. Data from the 1990 Post Enumeration Survey are restratified and dual system estimates with Jackknife variances are calculated. The need for good data quality in both the initial phase and the ICM phase and the effect on Congressional reapportionment are also discussed.

I. Introduction

Census 2000, as currently planned, will integrate the results of a large coverage survey into the final census estimates at all levels of geography. This paper describes the applied research used to determine an appropriate allocation of the Integrated Coverage Measurement (ICM) sample to the states for Census 2000. For more information on Census 2000 and the design of the ICM program, see Hogan and Waite (1998) or Griffin and Vacca (1998). The following basic facts are considered by the design:

- The total ICM sample size will be about 750,000 housing units. This size was determined by the Census Bureau's ability to implement and control the ICM sample and by statistical requirements. Rough preliminary estimates indicated that this sample size might be enough to produce coefficients of variation of 0.50% in each state. Block clusters averaging about 30 housing units will continue to be the primary sampling unit. The total ICM sample will have about 25,000 block clusters. Data from an independent second enumeration of the ICM block clusters will be compared to the Initial Phase estimate using Dual System Estimation. In comparison, the 1990 Post Enumeration Survey (PES) was only about one-fifth as large.
- A Supreme Court ruling in March 1996 and others have expressed concern about the PES state level total population estimates based on data from several states. The official population of each state and the District of Columbia released on December 31, 2000 will be estimated directly from the data from within the state.
- The primary goal of ICM is to improve the accuracy of the Congressional reapportionment process. In statistical terms, the expected value of most state population estimates should be closer to the true value with ICM than with a traditional census. Without the ICM, the wrong states in terms of their true populations may be competing for the last few seats in the House of Representatives. With ICM, the right states are more likely to be in the competition.
- The primary goal of ICM allocation is to optimize the precision of the apportionment process. In statistical terms, ICM allocation attempts to make the state population estimates close to their expected values[1]. Optimizing the precision of the apportionment process for Census 2000 would require decreasing the standard errors of those four to six states competing for the last three or four seats in the House of Representatives as much as possible and virtually ignoring the other states. However, precensal estimates will not be accurate enough to identify these last few states.
- Since census data are also used for redistricting, for allocation of federal and state funds, for planning purposes, etc., reasonable precision is also required for those states whose apportionment is certain and for synthetic estimates for substate areas and population subgroups.

Section II describes the research leading to the

---

[1] The 1990 reapportionment based on census counts was more precise (closer to its expected value) but less accurate (expected value missed the true value) than the apportionment process will be with ICM in Census 2000. Initial Phase estimates will be close to their expected values which may be far from the true population. ICM state estimates will miss their expected values, which will be closer to the true values, by more than the Initial Phase estimates miss their expected values. However, the ICM estimates will generally miss the "true" values by less than the Initial Phase estimates miss the "true" values.

The author is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. The author wishes to thank John Thompson, Henry Woltman, Richard Griffin, and Alfredo Navarro for their guidance and insight over the last several years.

recommended state ICM sample sizes using data from the 1990 Post Enumeration Survey (PES). Section III discusses the possible effect of changes in data quality from 1990 to 2000. Section IV discusses the effect of ICM sampling errors on congressional reappportionment. Section V provides a brief summary.

II. Methodology

Step 1: Redefine Sampling Strata:

For the 1990 PES 112 sampling strata were defined based on the Census division, degree of urbanization, minority population, and tenure. Some of these sampling strata had very small sample sizes. For this work the 1990 PES sampling strata were collapsed to 39 sampling strata. In addition to a national stratum for American Indians living on reservations, each of nine census divisions has zero, one, or two minority redefined strata (total 13, none in New England or the North Central division), and two or three non-minority redefined strata (total 25). Each state has PES block clusters in from two to six of the redefined sampling strata. There are 186 sampling stratum/state substrata for non-American Indian Reservations and 14 for the American Indian Reservations.

Step 2: Remove Outlier Block Clusters:

Thirty-nine block clusters which contribute heavily to the error were identified and removed. These clusters generally have high sampling weights and accounted for a large portion of the undercount or overcount in the sampling stratum/state cell. Outliers of the magnitude found in 1990 could as much as double the standard errors of the affected states. Several proposed design changes will help to control the effect of outliers in Census 2000:

- In 1990 large block clusters (over 80 housing units) were subsampled before PES collection. The subsampling resulted in high weights. Increasing the number of large block clusters in the 2000 ICM will reduce their initial weights. Subsequent subsampling will increase the weights back to a normal level.
- In 1990 only a very small sample of small block clusters (0-2 housing units) was selected. During PES collection some of these block clusters were found to be much larger, giving high weights to a large number of housing units. In 2000 a two stage sample for very small block clusters will control the weights of those block clusters which are found to be much larger than expected.

These approaches involve additional costs and will not succeed completely in eliminating the effects of outlier block clusters.

Step 3: Adjust Weights to Match State Totals:

The PES E-Sample consists roughly of the 1990 census reports of those persons in the PES sample block clusters. For each block cluster, weighted estimates of the E-Sample are calculated. For many states the weighted E-Sample is not a good estimate of the 1990 census count. For each state the sum of the state's weighted E-Sample estimate is ratio adjusted to the 1990 census count. The weighted estimates of erroneous enumerations (persons in the census who should not have been counted), P-Sample (persons enumerated in the second interview in the PES block clusters), and omissions (persons in the P-Sample who could not be matched back to the census) are multiplied by the same ratio.

Step 4: Make State and Stratum Estimates:

For each of the 39 sampling strata, dual system estimates are calculated by:

$$DSE_{Div,k} = C_{Div,k} \frac{E_{Div,k} - EE_{Div,k}}{E_{Div,k}} \frac{P_{Div,k}}{M_{Div,k}}$$

where:

$C_{Div,k}$	is the census count in stratum k or in this case the weighted E-sample,
$E_{Div,k}$	is the E-sample estimate in stratum k for the census division,
$EE_{Div,k}$	is the estimated number of erroneous enumerations in stratum k for the census division,
$P_{Div,k}$	is the P-sample estimate in stratum k for the census division, and
$M_{Div,k}$	is the estimated number of P-sample persons who match to the E-sample in stratum k for the census division.

A Jackknife procedure dropping one block cluster at a time from each census division's PES sample without reweighting[2] is used to estimate standard errors, $SE_{Div,k,n_{k,Div}}$, and variances, $VAR_{Div,k,n_{k,Div}}$, for the E-sample person sample sizes, $n_{k,Div}$, for the DSE in stratum k for division Div.

Since the finite population correction factors are negligible, for the same sample size, $n_{k,Div}$, the CV of state i for stratum k is the same as the CV for the division.

---

[2] The DSE and its population variance can also be estimated by Taylor Series expansion from the erroneous enumeration and omission rates. The results are consistent with those of the approach used here. This more direct approach is preferred because it is simpler and it is consistent with the 1990 and 2000 variance estimation methods. Other options considered included equal allocations to all states and various combinations of the alternatives.

Therefore: $SE_{i,k,n_{k,Div}} = SE_{Div,k,n_{k,Div}} \frac{E_{i,k}}{E_{Div,k}}$.

Step 5: Determine Block Cluster Sample Sizes

We assume n_i^0 = 10,000 E-Sample persons.

Allocating these persons proportionally to the states's E-sample population in the redefined strata we have:

$$n_{i,k}^0 = 10000 \frac{E_{i,k}}{\sum_{k'} E_{i,k'}} .$$

The standard errors for these stratum sample sizes are: $SE_{i,k,n_{i,k}^0} = SE_{i,k,n_{k,Div}} \sqrt{\frac{n_{k,Div}}{n_{i,k}^0}}$ for each stratum and

$$SE_{i,n_i^0} = \sqrt{\sum_k SE^2_{i,k,n_{i,k}^0}}$$ for the state total.

The next step is to convert the $n_{i,k}^0$ to $b_{i,k}^0$, the number of block clusters in state i stratum k, using the observed average block cluster E-Sample person size for stratum k within Census division.

If SE_{i,b_i^0} is the standard error for the block cluster sample size b_i^0 corresponding to the 10,000 E-Sample persons, the sample size, in block clusters, required to obtain a desired standard error, SE_i is: $b_i = b_i^0 \frac{SE^2_{i,b_i^0}}{SE_i^2}$.

An allocation of 18,873 block clusters is required to achieve the desired coefficients of variation (CV) of 0.5% in all states. These allocations are shown in column 5 of Table 1.[3]

Step 6: Assure Minimum Sample Size

[3] The allocations for states in the same Census Division are correlated because the same population variance estimates are being used. The differing proportions of the population in each sampling stratum account for the small differences between states. It is possible to repeat this procedure entirely within each state. This eliminates the synthetic estimation from the Census Divisions to the states and the correlation of the allocations. Unfortunately, most stratum/state cells do not have sufficient sample to obtain reliable estimates.

Thirteen states have ICM samples less than 300 block clusters from step 5. These states are concentrated in several divisions with relatively low estimated population variance. Since the estimated population variances, which are subject to high variance, may not be as low in 2000 as in 1990, the samples sizes for these states are increased to 300 block clusters to be more in line with the remaining states. These increases require about 1200 block clusters, increasing the total allocated so far to about 20000. The results are shown in columns 6 and 7 of Table 1.

Step 7: Reduce Expected Standard Errors for States with Populations over 10,000,000

Reserving 350 block clusters for American Indian Reservations, about 4600 block clusters remain to be assigned. These are assigned to the largest states proportionately to the squares of their 1990 census counts[4]. This reduces the estimated standard errors of the largest states from 0.50% of their population to 55672. These decreases are particularly substantial in the largest states: California, New York, and Texas. The sample size for Ohio was increased from 260 to 358, executing steps 6 and 7 simultaneously. The results are shown in columns 8 to 10 of Table 1. Columns 11 and 12 show the number of persons and occupied housing units which can be expected in each state.

AMERICAN INDIAN RESERVATIONS

In 1990 the largest reservations, spread across fourteen states, were covered by a sample of 43 block clusters. American Indians living on reservations or other tribal lands have special legal status. In 1990 variances were high for this hard to count population of about 800,000 people with about a 10% undercount rate. 350 block clusters, about as many as the states with fewer than 10,000,000 residents, 1.4% of the sample, were set aside for this 0.3% of the population.

III. ICM Quality Concerns

The ICM sample sizes calculated above are designed to yield errors of 0.5% or 56,000, whichever is smaller in all states. Table 1 shows that California would require 361 block clusters to achieve a CV of 0.5%. Estimates made for the state of California show a CV of about 0.45% for its 381 1990 PES block clusters. On the other hand, the CVs calculated for the 1995 and 1996 tests were considerably higher than the design estimates. The DSE is roughly the

[4] The use of projected 2000 populations was considered, but the estimated allocations for several states seemed inappropriate.

Initial Phase estimate times the rate of Initial Phase persons who are correctly enumerated times the inverse of the rate of P-Sample persons who could be matched back to census reports:

$DSE = IP \times R_{CE} \times 1 / R_{MATCH}$ where both rates are close to 1. There is comparatively little variance in IP, so (assuming equal design effects and even with some correlation) the variance is proportional to the sum of two PQ type variances:

$$VAR_{DSE} \approx \frac{R_{CE}\times(1-R_{CE})}{n} + \frac{R_{MATCH}\times(1-R_{MATCH})}{n}$$

where n is the ICM sample size[5].

There are several operational changes from 1990 in the design for Census 2000 which may decrease either the correct enumeration rate and/or the match rate. A 3% decrease in both the correct enumeration rate and the match rate from 97% to 94% would not change the estimate much, but it could double the estimated variance, multiplying the estimated CV by about 1.4. These changes include:

- The easy availability of Be Counted Forms could increase the number of erroneous enumerations, decreasing the correct enumeration rate.
- The use of a five person form instead of a seven person form could increase the number of persons, especially children and nonrelatives, missed in the initial phase, decreasing the match rate.
- The tight schedule and decreased public cooperation could increase the number of whole household imputations in the initial phase, decreasing the match rate. The rate was about 1% in 1990 but about 8% in the 1996 test in Chicago.
- Not performing a surrounding block search for additional matches or performing a limited surrounding block search could decrease the match rate.

There are few counterbalancing changes to improve the data quality[6]. Therefore, it should be expected that the calculated standard errors for Census 2000 may be somewhat higher than those predicted by the design.

[5]There is a third ratio in the DSE formula which adjusts for whole person imputations in the Initial Phase. This term adds little to the variance, but it corrects for census persons who cannot be matched to by P-Sample persons because their census data is imputed.

[6]The Be Counted Forms should decrease the number of nonmatches and decreased weight variation should make the sample more efficient.

IV. Effect on Reapportionment

The 435 seats in the House of Representatives are reapportioned to the states using the Hill Algorithm which works as follows:

- Assign each state one seat.
- For each state calculate: $R_i = POP_i / \sqrt{N_i \times (N_i + 1)}$

  where POP_i is the population of state i being used in the apportionment, and N_i is the number of seats already assigned to state i.
- Assign the next seat to the state with the largest value of R_i.
- Calculate new R_is and repeat the process until all 435 seats are assigned.

Using the 1990 PES instead of the 1990 census counts would have given one more seat to California at the expense of Wisconsin. For the two apportionments the 435th seat was assigned as follows:

- 1990 census count: Washington
  Massachusetts was next and would have needed 12605 more inhabitants to take the last seat instead of Washington.
- 1990 PES estimate: Pennsylvania
  Wisconsin was next and would have required 12573 more inhabitants to take the last seat. Montana was fourth in line but would have required only 3919 more inhabitants to take the last seat.

The 1990 PES estimate for state i can be viewed as a random draw from the normal distribution about the true value. That is: PES_i is selected from $N(T_i, SE_{PES,i})$. Since we know PES_i, we can reverse the situation and obtain 100 possible target values of the truth for each state, i, by selecting $T_{i,j}$, j=1,100 from $N(PES_i, SE_{PES,i})$. For each target estimate $T_{i,j}$, we can obtain 100 estimates of possible values that the ICM would produce, $ICM_{i,j,k}$, by sampling from $N(T_{i,j}, SE_{ICM,i})$. Thus, it is possible to compare the apportionment from the 1990 Census to 100 1990 targets and the apportionments from 10000 ICM estimates to the same 100 1990 targets. The results are shown in table 2.

- Using either census counts or ICM, there is only a small probability that the apportionment process will assign all 435 seats to the correct states.
- Over the 10000 simulations of ICM estimates, the 1990 census and 2000 ICM apportionments had the same number of errors compared to the target "true" apportionments 3738 times. The 1990 census apportionment had fewer errors 1982 times. The 2000 ICM apportionments had fewer errors 4280 times.
- Over the 10000 simulations there were 42032 instances where a state had a difference between its 1990 census apportionment and its 2000 ICM apportionment. In these instances, the 1990 census apportionment matched the target apportionment

18270 (43.47%) times. The 2000 ICM apportionment matched the target apportionment 23762 (56.53%) times.

- Using the 1990 PES estimates or the ICM estimates, the states with the most variation in the target apportionments; that is, the states which may deserve one more or one less seat, are bunched around the 435th selection for both the target and the ICM apportionments. On the other hand, even though there is only one difference between the 1990 census and the 1990 PES apportionments, the states with variation in their target apportionments are not the states clustered around the 435th selection in the 1990 census data apportionment.

Table 2: Number of Seats Shifted Compared to Target Apportionments over 100 Simulations for 1990 Census or 10000 Simulations for 2000 ICM

	1990 Census	2000 ICM
Census or ICM apportionment equals target apportionment	3	1285
One seat shifted by census or ICM apportionment compared to target apportionment	41	5015
Two seats shifted	51	3104
Three seats shifted	5	554
Four seats shifted	0	41
Five seats shifted	0	1
Average number shifted	1.58	1.31

V. Summary

- For the allocation proposed based on the underlying population variances, it is estimated that a total ICM sample size of 24,650 block clusters (if allocated appropriately and assuming data quality equivalent to 1990) is sufficient to (1) achieve coefficients of variation of 0.50% in states with populations less than 10,000,000, (2) allocate each state at least 300 block clusters, and (3) achieve standard errors of 55672 for states with a population over 10,000,000.
- The expected CVs or SEs calculated above are just that: EXPECTED. The increase in population since 1990 will increase the standard errors of the largest states from 55672 to about 60000. Estimates show that the CVs of the estimated CVs or SEs are about 20%. That means that, even if the average state CV is 0.50%, about 16% of the states (8 states) will likely have CVs or SEs at least 20% larger than the expected values or above 0.60% or 72,000; and about 2% (1 state) will likely have a CV or SE 40% larger than the expected value or above 0.70% or 84,000. Any decrease in data quality compared to 1990 may further increase the CVs or SEs in 2000.
- The proposed ICM sample sizes in 2000 will be sufficient to assure that the correct states are in the competition for the last few seats in the House of Representatives, but they will not be sufficient to assure that all 435 seats are apportioned perfectly. The apportionment process is very sensitive to minor population variations and no affordable ICM sample size can reduce the standard errors enough to assure perfect apportionment. However, a traditional census count would virtually insure that the apportionment would be incorrect.
- It is necessary to explain this technical decision to non-technical audiences. This option is relatively simple. Since it is similar to the variance estimation methods for 1990 and 2000, it should already be familiar to many of the involved parties.

Issues which have not been investigated are:

- Only the variance from the ICM sample has been considered. Variance from sampling for nonresponse follow-up and housing units returned as vacant by the post office, will be small at the state level and will have no significant effect on the estimates. A third source, variance due to imputation of missing data, could be more substantial.
- For the allocation of the ICM sample within each state, sampling strata and estimation poststrata will be developed to permit adequate estimates for race, Hispanic origin, age, sex, tenure, and geographic subpopulations. Oversampling small but visible subpopulations could increase the state level errors.

References:

Hogan, H. and Waite, J. (1998) "Statistical Methodologies for Census 2000 - Decisions Issues, and Preliminary Results," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Griffin, R. and Vacca, E.A. (1998) "Estimation in the Census 2000 Dress Rehearsal," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

TABLE 1: ICM Sample Sizes for CV/SE Combinations For Three Variance Alternatives

CEN DIV	State		1990 Census	Estimated DSE	0.5% CV	300 ClusterMin		SE for States>10000000			Persons	Occ HUs
					18873	20042	CV	24650	CV	SE	1835172	690062
(1)	(2)		(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	CT	9	3287116	3343185	377	377	0.500%	377	0.500%	16716	27942	10406
New	ME	23	1227928	1246851	309	309	0.500%	309	0.500%	6234	21676	8365
England	MA	25	6016425	6118807	375	375	0.500%	375	0.500%	30594	27790	10359
	NH	33	1109252	1126308	307	307	0.500%	307	0.500%	5632	21544	8320
	RI	44	1003464	1020492	373	373	0.500%	373	0.500%	5102	27607	10302
	VT	50	562758	571129	285	300	0.487%	300	0.487%	2782	20687	8074
2	NJ	34	7730188	7863038	461	461	0.500%	461	0.500%	39315	32003	12067
Middle	NY	36	17990454	18228030	470	470	0.500%	1261	0.305%	55672	96137	35936
Atlantic	PA	42	11881642	12089483	496	496	0.500%	585	0.461%	55672	41216	15514
3	DE	10	666168	687353	413	413	0.500%	413	0.500%	3437	29380	11062
South	DC	11	606900	640519	384	384	0.500%	384	0.500%	3203	28914	10993
Atlantic	FL	12	12937926	13330671	363	363	0.500%	520	0.418%	55672	38813	14775
	GA	13	6478216	6670979	399	399	0.500%	399	0.500%	33355	28240	10852
	MD	24	4781468	4933145	368	368	0.500%	368	0.500%	24666	27528	10384
	NC	37	6628637	6816377	400	400	0.500%	400	0.500%	34082	28330	10827
	SC	45	3486703	3601446	422	422	0.500%	422	0.500%	18007	29587	11230
	VA	51	6187358	6355420	371	371	0.500%	371	0.500%	31777	27033	10368
	WV	54	1793477	1834188	425	425	0.500%	425	0.500%	9171	28548	11113
4	AL	1	4040587	4132465	417	417	0.500%	417	0.500%	20662	26712	10207
East	KY	21	3685296	3725204	447	447	0.500%	447	0.500%	18626	28588	10939
South	MS	28	2573216	2635900	402	402	0.500%	402	0.500%	13179	26432	9698
Central	TN	47	4877185	4979805	433	433	0.500%	433	0.500%	24899	27716	10786
5	AR	5	2350725	2396511	494	494	0.500%	494	0.500%	11983	35395	13204
West	LA	22	4219973	4309683	595	595	0.500%	595	0.500%	21548	43282	15929
South	OK	40	3145585	3208831	426	426	0.500%	426	0.500%	15891	30590	11598
Central	TX	48	16986510	17418396	795	795	0.500%	1945	0.320%	55672	143215	52277
6	IL	17	11430602	11589356	351	351	0.500%	380	0.480%	55672	29531	11040
East	IN	18	5544159	5568146	276	300	0.479%	300	0.479%	26687	21839	8362
North	MI	26	9295297	9365360	317	317	0.500%	317	0.500%	46514	23746	8997
Central	OH	39	10847114	10917940	260	358	0.427%	358	0.427%	46574	26135	10020
	WI	55	4891769	4915103	288	300	0.490%	300	0.490%	23477	21772	8329
7	IA	19	2776755	2806367	151	300	0.355%	300	0.355%	9959	21029	8206
West	KS	20	2477574	2510872	157	300	0.361%	300	0.361%	9077	21569	8301
North	MN	27	4375099	4434536	145	300	0.348%	300	0.348%	15405	21759	8330
Central	MO	29	5117073	5185737	158	300	0.363%	300	0.363%	18841	21543	8296
	NE	31	1578385	1600796	175	300	0.382%	300	0.382%	6013	21216	8244
	ND	38	638800	647888	149	300	0.352%	300	0.352%	2175	21007	8202
	SD	46	696004	705880	162	300	0.368%	300	0.368%	2440	20612	8137
8	AZ	4	3665228	3783790	492	492	0.500%	492	0.500%	18384	36175	13874
Mountain	CO	8	3294394	3390812	479	479	0.500%	479	0.500%	16954	35754	13589
	ID	16	1006749	1050778	412	412	0.500%	412	0.500%	4763	33979	12612
	MT	30	799065	833975	420	420	0.500%	420	0.500%	3762	34342	12841
	NV	32	1201833	1239255	468	468	0.500%	468	0.500%	6196	35765	13650
	NM	35	1515069	1570331	481	481	0.500%	481	0.500%	7155	35291	13267
	UT	49	1722850	1773525	478	478	0.500%	478	0.500%	8617	35836	13631
	WY	56	453588	470907	418	418	0.500%	418	0.500%	2355	34251	12783
9	AK	2	550043	567494	334	334	0.500%	334	0.500%	2837	27291	10407
West	CA	6	29760021	30769103	361	361	0.500%	2753	0.181%	55672	232642	84161
	HI	15	1108229	1142269	283	300	0.485%	300	0.485%	5545	24272	9308
	OR	41	2842321	2927930	320	320	0.500%	320	0.500%	14640	25993	9830
	WA	53	4866692	5007784	332	332	0.500%	332	0.500%	24883	26918	10060