Combining Census and Dual-system Estimates of Population

Alan M. Zaslavsky, Harvard University
Department of Statistics, 1 Oxford Street, Cambridge, MA 02138

Keywords: Undercount, adjustment, Empirical Bayes.

On July 15, 1991 a decision was made to use unadjusted rather than adjusted census populations as the official counts from the 1990 census. The unadjusted counts were more biased (at least at highly aggregated levels) than those adjusted by factors estimated with the Dual System Estimator (DSE), calculated using information from the census and the Post Enumeration Survey (Hogan and Wolter 1988). On the other hand, the variance of the adjusted counts was substantial, since they are based on a relatively small sample of the total population and there are a number of forms of sampling and nonsampling error that may affect the adjusted estimates. The unadjusted census, on the other hand, is regarded as having essentially no variance.

Thus, the choice between the two, as it has been typically formulated, is a choice between a precise but biased estimator of population and a noisy but less biased one. Current analyses of the decision problem (Spencer 1990, Hogan and Mulry 1989) have considered comparing the risk (expected loss) of the two sets of estimates and choosing the set with smaller estimated risk.

This formulation of a bias-versus-variance tradeoff suggests consideration of composite estimators incorporating information from the census and DSE, and possibly from evaluation projects as well. The resulting estimates might have less bias than the unadjusted counts, but less variance than the DSE. Depending on the sizes of the various components of bias and variance, the optimal estimator might be such a composite of the census and DSE, rather than either one alone.

In this paper we consider a number of issues related to the question of how to obtain optimal estimates using the census counts and DSE. In Section 1 we formally specify the objectives of the estimation procedures and models for the underlying processes. In Section 2 we derive a number of estimators of population, and in Section 3 we evaluate them on simulated populations. Finally, in Section 4 we apply these methods to data from the 1991 census and PES, comparing various estimators and examining sensitivity to model assumptions.

1 Specification of objectives and populations

1.1 Notation and loss function

The following vectors will be of interest in forming composite estimates of population: T = true (unknown) population shares of the various domains, C = Census (unadjusted) population shares, D = Dual System Estimator (adjusted) population shares. b_d = bias of the Dual System Estimator in the various domains, $\hat{b}_d$ = estimate of this bias from evaluation projects.

Population shares are of primary importance for allocation of funds and political representation. We will assume that our objective is to minimize some loss function that is an aggregate measure of error in estimating shares, focusing on the loss function $L_X = \sum W_i(X_i - T_i)^2$ with $W_i = 1/T_i$, the size-weighted relative squared error loss function, associated with any estimate of population shares X.

1.2 A model for population shares in the census and DSE

In order to develop procedures for estimation of domain populations, we must specify a model for the relationship among true population shares T, their dual system estimates D and their sampling biases b, and the census counts C. This same model will be used to generate populations in simulations.

The model is specified as follows:

Population model:

$$\begin{pmatrix} b \\ C \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ T \end{pmatrix}, \begin{pmatrix} \sigma_{dd}U & \sigma_{cd}U \\ \sigma_{cd}U & \sigma_{cc}U \end{pmatrix}\right) \quad (1)$$

Sampling (data) model:

$$D|T, b, \hat{b}, C \sim N(T + b, V_d) \quad (2)$$

$$\hat{b}|T, b, C, D \sim N(b, V_b) \quad (3)$$

The interpretations of these distributional specifications are as follows:

Prior distribution of biases. The biases of the dual system estimator ($b = \mathbf{E}\ D - T$) and the census counts ($C - T$) are *a priori* normally distributed around 0. The prior variance-covariance matrix of the components of b is proportional to a matrix U representing the degree of similarity believed to exist between domains with respect to propensity to be under/overestimated by the DSE. The prior variance-covariance matrix of the components of $T-C$ is proportional to the same matrix U. The covariance matrix between the two vectors is also proportional to U, i.e. the propensity to be under/overestimated by the DSE is related to the same propensity with respect to the census, and the relationships between census and DSE biases in different domains are similar to those between biases of the same procedure in the corresponding domains.

Sampling variability of the dual system estimator. The observed DSE D varies around its expectation $T+b$ with variance-covariance matrix V_d, independently of all other measures. (The census is assumed to have no sampling variability.)

Sampling variability of the bias estimate for the dual system estimator. The estimated bias $\hat{b}$ varies around the true bias b with variance-covariance matrix V_b, independently of all other measures.

If there is substantial lack of sampling independence between $\hat{b}$ and D then the covariance V_{bd} could be estimated and included in the model. An argument can be made that this component is likely to be small, and our specification and estimators will assume $V_{bd}=0$.

Note that since the sum of population shares is fixed, the matrices U, V_b and V_d are necessarily singular. As a computational device, and without altering the model specification, we may drop the components corresponding to one arbitrarily chosen domain from each vector, thereby making all of the matrices nonsingular. (However, all components must be included in computation of losses.)

1.3 Specification of U

As noted above, the matrix U describes the extent to which different domains are believed to be similar with respect to the undercount process. We will assume that U is known and elicit it from prior knowledge about factors likely to affect under/overcount.

Elicitation of prior variance-covariance matrix of undercount rates: Suppose that a number of different factors enter, roughly additively, into determining the undercount rate for a domain, so that for a domain with levels $(c_1, c_2, \dots c_F)$ of factors $1, 2, \dots F$ the undercount rate is approximately $\alpha_1(c_1)+\alpha_2(c_2)+\dots+\alpha_F(c_F)$, where $\alpha_f(c_F)$ is *a priori* a random effect with variance Var α_f, assumed to be independent of the effects for distinct f or c_f. Then the covariance between the undercount rates for two domains is the sum of the variances of the random effects corresponding to the factors on which the domains agree, $\sum_{c_f=c'_f}$ Var α_f. Thus, by enumerating a plausible set of factors with relative importances in contributing to undercount, a matrix U^* of *a priori* covariances of undercount rates may be constructed. This procedure requires no prior assumptions about the magnitude of the under/overcount or about which domains are likely to be under/overcounted, but only about the *degrees of relationship* between domains.

The variance-covariance matrices Var b, Var C, and Cov(b, C) are assumed to be proportional to each other. This specification is plausible if the factors that make domains similar to each other with regard to Census undercount are like those which make domains similar to each other with regard to the bias of the DSE.

Converting under/overcount rates into population shares: If X^* is a vector of under/overcount rates and T is the vector of approximate true population shares for the same domains, then the population shares corresponding to X^* are X^{**} where $X_i^{**}=(1+X_i^*)T_i/\sum_j(1+X_j^*)T_j \approx T_i+X_i^*T_i-\sum X_j^*T_j$, as long as $X_i^*\approx 0$. Therefore $X^{**}\approx(I-T\mathbf{1}')X^*$, where $\mathbf{1}$ is a vector of 1's. Finally, the last component of X^{**} may be dropped (to make the covariances nonsingular) by premultiplying by the matrix I_0 (the identity matrix with the last row omitted). Then $X\approx I_0(I-T\mathbf{1}')X^*$; $U=I_0(I-T\mathbf{1}')U^*(I_0(I-T\mathbf{1}'))'$ where $U^*=$ Var X^* and $U=$ Var X.

Sampling variance-covariance matrices for undercount rates, V_b^* and V_d^*, may similarly be converted to the corresponding matrices V_b and V_d for population shares.

2 Estimation of population shares

In this section we develop various composite estimators of population shares.

Pure strategies. First of all, of course, are the unadjusted census counts and the DSE ($X=C$ and $X=D$). Each of the two pure strategies corresponding to these two estimates will have a risk that is a function of the sampling distributions and the (unknown) biases of the census and DSE.

Naive bias correction. If an estimate of the bias of the DSE is available, an unbiased estimate of the true population share may be obtained by subtracting the estimated bias from the DSE estimate $(D - b)$. If the bias estimates and DSE are uncorrelated with each other, the MSE of this estimate X_i will be $v_{bi} + v_{di}$. Since bias estimates have large variance (because evaluation samples are small), the variance of this estimator will be large.

Minimum-risk combinations. Several strategies are based on choosing between or combining the census and DSE estimates so as to minimize the estimated risk of the estimator. The "Better" strategy chooses between the two by comparing estimates of the risk of each set of shares that are functions of unbiased estimates of population parameters. The calculation regards the biases $C-T$ and b of the census and DSE as fixed but unknown quantities. (Thus, for inferential purposes these procedures use only the part of the model that describes the sampling distributions of $\hat{b}$, D.)

The mean squared error (MSE) of the unadjusted census counts for domain i is $(C_i - T_i)^2$ and the MSE of the DSE is $b^2 + v_{di}$, where v_{di} is the sampling variance of the DSE for domain i (the i-th diagonal element of V_d).

We may write

$$\mathbf{E}\ \hat{R}_D = \mathbf{E}\ \sum W_i(\hat{b}_i^2 + v_{di}) = R_D + \sum W_i v_{bi}$$

$$\mathbf{E}\ \hat{R}_C = \mathbf{E}\ \sum W_i((C_i - D_i - b_i)^2 - v_{di}) = R_C + \sum W_i v_{bi}$$

so that R_C and R_D may be compared by comparing $\hat{R}_C$ and $\hat{R}_D$. Then the "Better" estimate is $X = C$ if $\hat{R}_C < \hat{R}_D$, $X = D$ otherwise. This is the proposal of Spencer (1990).

Rather than selecting C or D we may choose from a continuum of combinations of the two estimators of the form $X_\lambda = \lambda C + (1-\lambda)D$. The risk $R_\lambda = \sum W_i(\lambda(C_i - T_i) + (1-\lambda)b_i)^2 + (1-\lambda)^2 v_{di})$ of the combined estimator may be estimated by $\hat{R}_\lambda = \sum W_i(\lambda^2(C_i - D_i)^2 + 2\lambda(\hat{b}(C_i - D_i) - v_{di}) + (\hat{b}_i^2 + v_{di}))$. Then $\mathbf{E}\ \hat{R}_\lambda = R_\lambda + \sum W_i v_{bi}$ The estimated risk is minimized for

$$\lambda = \lambda^* = -\frac{\sum W_i(\hat{b}_i(C_i - D_i) - v_{di})}{\sum W_i(C_i - D_i)^2}$$

The "Best" linear combination then is X_{λ^*}. If $v_{di} = 0$ for all i this reduces to the weighted regression of the DSE bias on the difference between the Census and DSE estimates; positive v_{di} increase λ (reduce the weight given to the DSE) relative to the regression coefficient. (This procedure was proposed by Spencer (1980).)

If $\lambda^* < 0$ or $\lambda^* > 1$ then X_{λ^*} may not be as plausible or acceptable as when $0 < \lambda^* < 1$, since it is more extreme than either of the estimates it is based on. (Nonetheless, under some circumstances it may be a sensible estimator, as when the biases of the census and DSE are in the same direction and the census bias always has larger magnitude. The data may provide evidence that this is, in fact, the case.) We therefore also consider a best convex linear combination estimate ("Best01") which differs from the best linear combination only in that λ^* is truncated to lie in the interval $[0,1]$.

(Note that only C, D, $\hat{b}$, and $\{v_{di}\}$, but not the complete covariance matrix V_d nor V_b, are required to calculate the "Better" and "Best" estimates.)

Bayes and Empirical Bayes estimates. From the Bayesian point of view, the estimate which minimizes posterior MSE (and therefore the risk under the loss functions considered here) is the posterior mean vector. We now turn to estimators based on posterior mean calculations.

First, consider posterior distributions conditional on the variance components $(\sigma_{dd}, \sigma_{cd}, \sigma_{cc})$. To complete the model we must specify a prior distribution for the true population shares T. We will assume that any prior information we may possess about the shares is negligible compared to the information derived from the census and DSE, so the prior may be reasonably approximated by a uniform prior on the space of shares summing to unity. (Equivalently, the prior is uniform on the space of shares of all but the last domain.)

We may wish to calculate two posterior distributions. The distribution conditional on C, D, $\hat{b}$ and the variance components makes use of all available data. The distribution conditional on C, D, and the variance components, but not on $\hat{b}$, excludes the information from $\hat{b}$ except as it is used in estimation of variance components. Inference from the latter distribution might be appropriate to a situation in which $\hat{b}$ is regarded as so unreliable that it should not be directly weighted into the estimates, but only used for the essentially evaluative function of estimating variance components.

Inferences for T are by application of Bayes's Theorem to the uniform prior and the data distribution specified by Equations 2 and 3. Calculation of the posterior means given the variance components is straightforward by the usual formulae for the normal distribution.

The remaining issue for the Bayes/Empirical Bayes approach is estimation of these variance components. It is not difficult to write down unbiased estimators of $(\sigma_{dd}, \sigma_{cd}, \sigma_{cc})$ by equating the expectations of certain quadratic forms to their observed values (see Appendix). Estimates obtained by fixing $(\sigma_{dd}, \sigma_{cd}, \sigma_{cc})$ at these "Method of Moments" estimates are labelled "MM" estimates.

However, the unbiased estimates of the variances may be negative and even if they are not, the estimated covariance matrix may not necessarily be positive definite. This phenomenon is common in variance components problems. The probability of its occurrence is high here since the number of domains used in the evaluation exercise, and therefore the number of degrees of freedom, is small, and the sampling variances may be large. Maximum likelihood estimates may lie on the edge of the parameter space, i.e. MLE variances $\hat{\sigma}_{dd}, \hat{\sigma}_{cc}$ may be 0 or the correlation $r_{cd} = \hat{\sigma}_{cd}/\sqrt{\hat{\sigma}_{dd}\hat{\sigma}_{cc}}$ may be ± 1. Substituting meaningless or implausible variance estimates does not produce plausible results.

Our preferred solution to this problem is to sample from the posterior distribution of the variance components and combine the estimated conditional posterior means $\hat{T}$ by averaging them to estimate an unconditional posterior mean. This fully Bayesian procedure has the following appealing properties: (1) Each draw of the variance components is guaranteed to be valid in the sense of yielding a positive definite covariance matrix for (b, C). (2) Prior beliefs about the components, for example that the ratio σ_{dd}/σ_{cc} is likely to be near some predetermined value, can be incorporated into the distribution. (3) Although closed-form expressions for the distributions are complex, draws from the posterior distributions may be obtained via Gibbs sampling.

Results obtained by this procedure (representing, for each simulated data set, the mean shares over a fixed number of draws from the posterior distribution of the variance components) are reported as "Hierarchical Bayes" ("HB") estimates. Two different prior distributions were considered for the variance components: reparametrizing $\begin{pmatrix} \sigma_{dd} & \sigma_{cd} \\ \sigma_{cd} & \sigma_{cc} \end{pmatrix}$ as two standard deviations and a correlation, $\sigma_c = \sigma_{cc}^{1/2}, \sigma_d = \sigma_{dd}^{1/2}, r_{cd} = \sigma_{cd}/\sigma_c\sigma_d$, then the priors are proportional to $\sigma_c^0\sigma_d^0(1 - r_{cd}^2)^{0.2}$ (Bayes2:HB and Bayes3:HB), or alternatively $\sigma_c^1\sigma_d^1(1 - r_{cd}^2)^{0.2}$ (Bayes2:HB.2 and Bayes3:HB.2).

3 Simulations

The performance of the various estimators of population shares described in Section 1.2 was evaluated by a series of simulations. For each set of simulation trials, values were selected for the following parameters: k (number of domains), T (true population shares), U^* (covariance matrix structure for undercount rates and DSE biases), σ_{cc}, σ_{cd}, and σ_{dd} (scale of variances and covariances of undercount and biases), V_d (sampling variances of the DSE), and V_b (posterior variances of differences between estimated and actual biases of DSE). The simulation conditions are chosen so that on the average over all simulations, the mean squared error of the census and the DSE are equal, allowing us to focus on the effectiveness of the other estimators in improving on both of the pure strategies.

On each trial, values of census and DSE population shares C and D were drawn from the models in Section 1.2. The estimates described in Section 2 were calculated, and for each estimate the loss function described in Section 1.1 was calculated. Ten draws of the simulated populations were performed for each simulation condition.

In some simulations, the structure of U was assumed to be known. In others, a different matrix U_{fit}^* was assumed in fitting the models than the U^* which generated the data, in order to determine sensitivity to misspecification of U.

The Gibbs sampler for the hierarchical Bayes estimators was run for 1500 iterations and the first 100 values were discarded. It was determined that the loss of efficiency at this number of iterations (relative to a very large number of iterations) did not affect the loss comparisons substantially. For the Empirical Bayes estimators using method-of-moments variance estimates, an *ad hoc* correction was used to force the variance estimates to be positive definite when the method-of-moments estimates were not.

The various estimators were compared by the following measures:

(1) Mean loss across the trials for each estimator, relative to the average of the mean loss for the Census and the DSE for the same simulation conditions (to facilitate a scale-free comparison). ("MEAN")
(2)The number of simulation conditions (paramater values) on which the estimator had the smallest mean loss. ("DOMINATES")

A number of estimators can be eliminated because they make little or no improvement on C or are almost always dominated by another estimator, e.g. Best dominated by Best01.

Estimator	MEAN	DOMINATES
C	0.980	0.013
D	1.020	0.002
Best01	0.525	0.104
D-b	1.146	0.003
Bayes2:HB	0.493	0.275
Bayes3:HB	0.453	0.603

Attention was focused on the Bayes2:HB,

Bayes3:HB and Best01 estimators, which were the only real competitors. Overall, the Bayes3 estimator had smaller loss umder the largest number of simulation conditions. Fitting a logistic regression model to discriminate those conditions under which Bayes3 outperformed its competitors, the following factors were found to be relevant:

1. Bayes3 performed better relative to Bayes2 and Best01 as k (the number of domains) became larger.

2. Bayes3 performed better as r_{cd}, the model correlation between the biases of the DSE and the Census, became positive (i.e. when the DSE tended to underestimate the undercount in domains with large undercounts). Under this condition the correct estimate of the population would lie outside the range between the census and DSE, so the Best01 estimator would be unable to approach the truth while the Bayes3 estimator best makes use of the bias estimates.

3. The Bayes3 estimator improved relative to the Bayes2 estimator as the fraction of DSE variance due to sampling (rather than model variance) increased.

4. The Bayes3 estimator did worse relative to the the other two as the total error of the DSE increased relative to that of the census.

4 Results with 1990 Census and PES data

In this section, the various methods are applied to data from the 1990 Census and PES and the evaluation programs following the PES. In all of the analyses of this section, the domains of interest are the thirteen Evaluation Poststrata (EPS), described in the Undercount Steering Committee report.

The following data were available for analysis from the Census Bureau report on the total error analysis: census estimates of population for each EPS; DSE estimates of population for each EPS; sampling SD of each DSE estimate; estimated bias of DSE by EPS, separated by source of bias, and estimated total bias (obtained by simulation); estimated SD of bias estimates by EPS, separated by source of bias, and combined by simulation.

Covariances of DSEs and biases in different EPS's were not available. DSE's in different EPS's were assumed to be uncorrelated; this is nearly but not exactly true because of the overlap of sample blocks for Black/Hispanic and Other within the same region and urbanicity.

The model variance matrix U was assumed to be of the form $r_U U_0 + (1 - r_U)I$, where $U_0 = \{u_{ij}\}$ and u_{ij} = (number of factors for which EPS i and j have a common value)/3. This corresponds to a model in which r_U is the fraction of the variance across EPS's of the biases of Census and DSE which is explained by a model in which the three factors have *a priori* equal explanatory power, and the remaining variance is due to independent model error in each EPS.

The variance of b was divided into five components: "process" variance v_P associated with bias due to matching, address and E-sample error and PES fabrications, model variance v_M associated with estimation of correlation bias, imputation variance v_I due to imputation of individual unresolved cases, imputation variance v_B due to estimation of the parameters of the imputation model, and imputation variance v_{RA} due to uncertainty among "reasonable alternatives" in the specification of the imputation model. The inter-EPS correlation matrix of each of these error sources was assumed to be of the form $r_c U_0 + (1 - r_c)I$, $c =$P, I, B, RA, where U_0 is the similarity matrix defined above. r_c may be interpreted as the fraction of the corresponding variance component which is explained by the three factors which define the EPS. Then for any selected values $\{r_c\}$, $V_b = \sum_{c=\text{P,I,B,RA}} \text{diag}(v_c)^{1/2}(r_c U_0 + (1 - r_c)I)\text{diag}(v_c)^{1/2}$.

The first part of this analysis assumes fixed values of $\{r_c\}$, henceforth referred to as the "preferred" values, and compares the various estimators of population shares. In principal, all of these values except r_U could be estimated using data very similar to those developed in the PES evaluation programs, but this analysis has not yet been carried out by the Census Bureau, and may not be entirely feasible with existing data, given limitations of sample size and design. In the second part of the analysis, sensitivity to a range of plausible values of these parameters is tested. Since the estimates of correlation bias were regarded as somewhat controversial, sensitivity to omission of this component of error is also tested. An alternative specification of model correlation U is explored as well.

4.1 Comparison of estimators with preferred parameter values

We assumed $r_P = .1$ because the "process" error estimates are nearly independent across EPS's; $r_{RA} = .6$ because the choice of alternatives would have similar effects on adjustment factors in every EPS; $r_B = .2$ because some but not all of the parameters in the imputation model are common between EPS; $r_I = 0$ since each person is imputed independently of all others, given the parameters and model specification; and $r_M = .5$ because various assumptions about correlation bias might be expected to have generally similar effects on correlation bias estimates in every EPS. We somewhat arbitrarily assume $r_U = .5$, i.e. that half the model variance between EPS's is explained by the three-factor model. (With a more complicated model, not explored here, r_U could also be estimated from the data.)

Five estimators of undercount (not including the unadjusted census, adjustment factor = 1) were compared (see table).

The most striking feature of the estimates is the close agreement between the different estimators for each EPS. In every EPS (with the possible exception of EPS 6, 7 and 12 where the DSE relative adjustment factor is very close to 1), every estimate is closer to the DSE than to the Census. By definition, the Best01 estimates are shrunk toward the census by a constant factor; the estimated value of λ is 0.123 so the shrinkage factor is $1 - \lambda = 0.877$. The amount of shrinkage under the Empirical Bayes models varies from EPS to EPS.

The Empirical Bayes estimates are generally closer to D than to $D - b$. This suggests that the model fits did not find much evidence for the existence of substantial bias in the DSE. This conclusion is supported from classical arguments, since the test statistic for non-zero biases (under independence) $\sum(\text{bias/SE})^2$, nominally χ^2 with 13 degrees of freedom, evaluates to 20.53 ($p = 8.3\%$). The Bayes3 estimators generally gave more weight to the bias estimates than did the Bayes2 estimators. The choice of prior (first versus second set of Empirical Bayes estimators) appeared to have very little effect on the estimates.

4.2 Comparison of alternative parameter assumptions

There are two reasons for examining the sensitivity of estimates to the choice of parameter values. Several parameters (r_P, r_{RA}, r_B, the numerical value of R_M) describe quantities that are potentially estimable from existing or obtainable data. While it is interesting to know how much uncertainty our ignorance of these quantities introduces into the current estimates, this does not reflect fundamentally on the proposed methods, and the contribution of the individual parameters is not of great importance. Another set of parameters (r_U, the specification of U_0, and the choice of inclusion or exclusion of correlation bias in the bias estimates) represent modeling assumptions that may not be directly verifiable from the type of studies that have been carried out.

Potentially estimable parameters. Varying assumptions about the estimable but unknown parameters have effects that are small compared to the effects of the modeling parameters, the differences between the EPS's, or the standard errors of the estimates.

Modeling assumptions. The point estimates are more sensitive to the modeling assumptions than to the potentially estimable covariances. However, the range of estimates for each EPS is still small compared to the standard errors or differences between EPS's.

The weight r_U appears to have a systematic effect on the estimates, especially when race is given high weight in the model. Taking account of correlation bias appears to have an effect in a few EPS's, particularly EPS 1,2, 5 and 11, where ignoring correlation bias pulls the adjustment factors slightly toward 1.

Overall, however, the estimates are remarkably insensitive to the choice of assumptions as well as to the choice of estimation procedure.

	Estimator				
	D	D-b	Bayes2	Bayes3	Best01
1	1.046	1.054	1.045	1.047	1.041
2	0.972	0.971	0.974	0.974	0.975
3	1.033	1.028	1.022	1.022	1.028
4	0.979	0.974	0.985	0.985	0.982
5	1.035	1.021	1.035	1.035	1.031
6	0.998	0.995	0.998	0.998	0.999
7	0.997	1.002	0.997	0.997	0.998
8	1.018	1.034	1.019	1.027	1.016
9	0.992	0.991	0.991	0.992	0.993
10	0.983	0.983	0.986	0.986	0.985
11	1.039	1.047	1.039	1.041	1.035
12	1.000	1.024	1.000	1.001	1.000
13	0.997	0.991	0.998	0.994	0.998