BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-85/07

SMALL AREA ESTIMATION RESEARCH FOR CENSUS UNDERCOUNT--PROGRESS REPORT

by

Cary T. Isaki, Linda K. Schultz,
Philip J. Smith and Gregg Diffendal
Statistical Research Division
Bureau of the Census
Room 3524, F.O.B. #3
Washington, D.C. 20233 U.S.A.

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended: Nash Monsour

Report completed: May 9, 1985

Report issued: May 22, 1985

This paper was presented at the International Symposium on Small Area Statistics in Ottawa, Canada on May 22, 1985.

BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-85-07

SMALL AREA ESTIMATION RESEARCH FOR CENSUS UNDERCOUNT--PROGRESS REPORT

by

Cary T. Isaki, Linda K. Schultz,
Philip J. Smith and Gregg Diffendal

Statistical Research Division

Small Area Estimation Research for Census Undercount--Progress Report

by

Cary T. Isaki, Linda K. Schultz, Philip J. Smith and Gregg J. Diffendal
Bureau of the Census

I. Introduction

In 1980, the U.S. Bureau of the Census reported a census count of 226,549,448 persons on Census day, April 1st. No one knows the true number of persons living in the U.S. Beginning in 1950, demographic analysis methods were used to estimate the net census undercount. The most recent estimates of net undercount were 1% for 1980 (assuming 2.0 million illegal aliens included in the 1980 Census enumeration); 2.8% for 1970; 3.3% for 1960 and 4.4% for 1950.

In 1980, a post enumeration program provided a range of net undercount estimates of roughly a .5% overcount to a 2% undercount. Estimates of net undercount for states exhibited a wide range, also. For example, one series of state estimates exhibited a range of a 2% overcount to a 6% undercount. Hence, in addition to variability among net undercount rates for the U.S. we also appear to have differential net undercount among states. There is also evidence of differential undercount among race and sex groups. Such differential undercounting have been the basis for a number of court cases in which various jurisdictions, cities as well as states, have sued the Bureau of the Census to adjust the 1980 census counts. To date, the Bureau is not under any court orders to adjust the census counts.

One factor motivating certain jurisdictions to sue for an adjustment of the census is the use of population counts in determining representation in government as well as in determining the amount of revenues received from the Federal government. Since congressional representation is determined on a relative population basis among states, if every state experienced the same percent net undercount, then the allocation of number of representatives to states would remain unchanged whether the census counts or corrected counts were used. Differential net undercount among states would cause a difference in representation. With respect to revenue sharing allocation to state and sub-state governments it has been reported (Robinson and Siegel, 1979) that undercoverage of the income component in the revenue sharing formula has a greater effect than that due to undercoverage of the population. Nevertheless, undercoverage of the population is still perceived as the cause of incorrect disbursement of revenue sharing funds and hence the continuation of some jurisdictions to request adjustment of census counts. In addition, other methods of income and population adjustments than that used by Robinson and Siegel may alter their results. Finally, some Federal programs base eligibility of jurisdictions on level of total population and allot funds on the basis of other variables. For example, a program for economic development of communities only includes metropolitan cities and urban counties of a sufficient size. The funds alloted to these two types of communities are each based on such variables as population, poverty and housing overcrowding relative to all such communities. A secondary effect is the use of census population counts in determining ratio adjustment factors in on-going surveys such as the monthly labor force

surveys. Estimates of employment status used in disbursing funds would be affected by population undercoverage in this manner.

The research conducted so far has dealt with total population as a characteristic to be adjusted as it is the first characteristic that is to be produced from the census. The first set of population counts is required by state by the end of the calendar year while a second set of population counts is required for legislative re-districting purposes a year after census day. Since the basic unit of census tabulation is the census block consisting of an average 100 persons per block, one possibility is to adjust the census block for undercount. Adjustment at this level will then be consistent at higher levels of aggregation and in cross tabulation. We have not concerned ourselves with the problem of adjusting other characteristics but it is likely that should population be adjusted, housing unit counts, race, age, sex, and other characteristics would require adjustment. Another possibility is to only adjust at higher levels of aggregation. At any rate, the manner in which adjusted counts would be displayed in census publications, if adjustment is implemented, has not been decided upon at this time.

The focus of our small area research so far is in three general directions. The first direction is to look at the results of the 1980 PEP (Post Enumeration Program). The second is to look at demographic analysis results for 1980. The third direction is to use the 1980 census data to simulate and evaluate the performance of potential adjustment methodologies. The remainder of the paper describes the limitations of the data tools previously mentioned, describes the adjustment

methodologies being investigated and provides the results of our work to date.

II. Data Used in Research

The data used in our research comes from the 1980 PEP, demographic analysis and the census. Each data source has favorable and unfavorable features.

A. The 1980 PEP was designed to study the net population undercount for each state and the 23 largest metropolitan areas. The PEP consisted of essentially two samples (termed P and E samples in what follows) and a matching process which used dual system estimation to produce net undercount estimates. A detailed description of the PEP can be found in Cowan and Bettin (1982). The first sample consisted of about 186,000 persons in households in an ongoing monthly labor force survey in which a roster of persons in the households was obtained via a supplementary interview. The address was geographically coded to census geography. In fact two separate, non-overlapping monthly samples, April and August, were canvassed in this manner. However, no attempt has been made to combine the results and each sample has been treated separately with respect to dual system estimation. Each of these monthly samples are termed P-samples in the discussion that follows. The other sample consists of a sample of about 231,000 persons selected from the 1980 census from within the same selected primary sampling units associated with the P-sample and is termed the E-sample.

The PEP matched cases in the P-sample to the census files in the general location of the geocoded P-sample address. A status of matched or nonmatched was assigned to each person. Persons with a nonmatched status were sent back into the field for follow-up and then rematched to the census. All cases whose status (matched/not matched) could not be ascertained after the second match had a status imputed. Variations in the treatment of nonresponse cases and the manner of status imputation resulted in several different P-sample estimates.

The underlying concept of dual system estimation is to conduct two independent listings of the population and to measure those that are observed in both listings. In our context, one listing of the population is accomplished by the census and the other is accomplished by the P-sample. However, direct use of the census counts in dual system estimation is not feasible. The census operation includes in its count persons imputed on the basis of vague information and then allocates characteristics to them. Such persons could not be matched and were subtracted from census counts. In addition an estimate of persons coded to incorrect geography, out of scope and persons otherwise erroneously enumerated in the census was obtained via the E-sample and subtracted from the census count. In the E-sample procedure, interviewers returned to the census households. Persons not at the housing unit were followed up or neighbors were asked their whereabouts on census day. As in the P-sample, differing treatment of noninterviews and imputation of enumeration status resulted in several E-sample estimates.

Combinations of P- and E-sample treatments have resulted in 12 dual system estimates of total population by age, race, and sex categories at the U.S. level and with lesser detail at the state and sub-state level. The particular combination of treatments used in our modelling efforts below is termed PEP 3-8 which is based on the April labor force survey sample. Our use of PEP 3-8 estimates (as opposed to any other PEP estimate) was mostly arbitrary. The PEP 3-8 procedure was the designated one prior to implementation of the PEP program. We used PEP 3-8 as an illustration although other PEP estimates are equally viable. In this P-sample all noninterviews are adjusted by a weighting procedure that assumes that the noninterviewed are similar to the interviewed. Also, match status of unresolved cases (those remaining after follow-up) were imputed using as a pool of donors those cases initially sent to follow-up and whose status subsequently were resolved. The E-sample cases lacking enumeration status after follow-up were given to the post office for resolution. Those cases not resolved were imputed using donor pools of like persons whose status were resolved by the post office.

For a particular category, let

N_c ≡ census count of population

N_p ≡ the P-sample based estimate of population

EE ≡ the E-sample based estimate of census population erroneously enumerated

M ≡ the P-sample based estimate of population matched and

II ≡ census count of population imputed.

Then, the dual system estimator of population total used in the PEP is $\hat{N}$ where

$\hat{N} = N_p (N_c\text{-}EE\text{-}II)/M$ and

the net undercount is defined as

$\hat{Y} = (\hat{N}\text{-}N_c)/\hat{N}$. When estimating for a particular geographic area, the categories used were age-race-sex within the area. Depending on the size of the area, the categories were collapsed until an adequate amount of sample cases were realized. Both P- and E-sample estimates include ratio adjustment.

According to Cowan and Bettin (1982), the proportion of cases in the sample which are missing data is larger than the estimated net undercount. For example, for PEP 3-8, the percent of total persons, Black persons, Non-Black Hispanic persons and Other persons requiring imputation were 4.1, 7.2, 7.3 and 3.6 percent, respectively. The estimated net undercount in percent for the same categories were .8, 5.2, 4.1 and -.1. Consequently, the manner of imputation can have a major effect on the final estimates. There is some doubt as to whether independence is actually achieved in the PEP. Without independence the PEP estimates are biased. In addition, the listings are assumed to cover the entire populaton under consideration so as to yield a positive probability of response from every individual. It is questionable whether this was achieved in the PEP because the P-sample suffers from non-coverage. Despite these deficiencies, the PEP provides the only direct estimates of net undercount and gross errors at the sub-U.S. geographic level.

B. The second data source for measuring undercount levels in the 1980 census is the method of demographic analysis. Demographic analysis provides national estimates of the population and of net undercount classified by age, sex, and race. As a tool for census evaluation, demographic analysis involves the combination of different types of demographic data to develop estimates for the population as of the census date; then the estimates are compared with the corresponding census counts. The particular procedure used to estimate the coverage for the various demographic subgroups depends primarily on the nature of the available data. For the population under age 45 in 1980, estimates of the resident population and coverage are based directly on birth, death, immigration and emigration statistics and estimates. For the population aged 65 and over in 1980, estimates are developed from aggregate Medicare statistics adjusted for underenrollment in the Medicare files. For the population aged 45 to 64, the coverage estimates are based on population estimates derived primarily from the analysis of previous censuses. (See Passel, Siegel, and Robinson, 1982, and Passel and Robinson, 1984, for discussion of the demographic method of estimating coverage.)

Since it has been estimated that at least 2 million undocumented aliens were counted in the 1980 census (Warren and Passel, 1983), an allowance for undocumented immigration must be added to the estimated resident population based on demographic analysis to obtain estimates of net undercount of the total population (legal and undocumented residents). The problem of undocumented immigration is a major source of uncertainty in the demographic estimates of coverage,

especially for the nonblack population. For our purposes we assumed a level of 3.5 million illegal aliens assigned to age-race-sex categories on the basis of estimates derived by Warren and Passel (1983). The level of illegal aliens assumed here sets the demographic analysis estimate total population figure to approximately equal the PEP 3-8 total population estimate. From a small area estimation point of view, the lack of sub-U.S. undercount estimates and the illegal immigration problem are important drawbacks of demographic analysis.

C. 1980 Census--The 1980 Census provides much small area data in the way of population, housing and administrative data that are possibly associated with undercount. In addition to age-race-sex counts at small geographic levels, urbanicity, labor force status, education, migration, language, income source, housing unit ownership, housing unit density, address list source, mail returns, substitution and allocation counts of persons are examples of characteristics available for adjustment usage. Such data are tabulated to the district office level at present; the district office (DO) being the smallest level at which PEP 3-8 estimates are available. In the following section, we utilize the data at the DO level to model undercount and evaluate some of the adjustment methods. The DO is the administrative unit that was used to collect census information.

III. Adjustment Methods

The adjustment methods considered to date are either of the synthetic or regression type. Variations of either type arise from the manner in

which data resources are used. For example, net undercount adjustment factors at the total U.S. level could be used in a synthetic adjustment procedure by age-race-sex to provide sub-state level estimates of total population assuming a level of illegal immigration using demographic analysis. Synthetic adjustment could also be used by raking regional PEP 3-8 age-race-sex cell undercounts to state marginals and obtaining individual state age-race-sex cell adjustment factors for application to sub-state census data. Regression models using net undercount as the dependent variable could also be used to obtain sub-state estimates of total population adjusted for undercount.

A. Synthetic Estimation Using Demographic Analysis Estimates

Despite the limitations of demographic analysis data such as the unknown level of illegal immigration and the lack of sub-U.S. detail, synthetic estimation using demographic analysis estimates was investigated and compared with the 1980 census results at the state and DO level under the assumption that the corresponding PEP 3-8 estimates were the "truth." Some of these shortcomings could be reduced or eliminated in future census years. For example, the estimate of the segment of persons 45-64 years old would be reduced to the segment 55-64 years old in the next census and passage of a proposed bill in Congress could provide enough sanctions to enable a count of the number of illegals presently in the country and deter future illegal immigration. Finally, an advantage of synthetic estimation based on demographic analysis is that it could be done in a timely manner.

Basically, construction of demographic analysis based synthetic estimates of total population for an area consists of two steps. In the first step, an adjustment factor for a given age-race-sex cell at the U.S. level is computed as a ratio of the demographic analysis figure to the census figure. In the second step, the corresponding census count of persons in the cell in the small area is multiplied by the relevant factor and such products are summed over all cells in the area. The assumption underlying this process is that undercount for the elements in the cell is uniform over all small areas. This assumption is questionable because it is likely that, for example, minorities in suburban areas are undercounted at a much lower rate than minorities in urban areas. Some comparisons have been made with the census with regard to total population and are presented below. We looked at the performance of three different synthetic estimators with regard to some measures proposed by Schirm and Preston (1984) as well as some other measures.

The three synthetic estimators labelled DA1, DA2 and DA3 differ in the manner of treatment of the Hispanic minority. In DA1, the Hispanics were combined with the Non-Black group. In DA2, the Hispanics were treated separately by assuming they had the same adjustment factors as the Black group. In DA3, the PEP 3-8 undercount rate for Hispanics was used. In all three estimates the same Black adjustment factor was used. In DA2 and DA3, the Non-Black, Non-Hispanic group (termed Other) factor was computed in a straightforward manner by removing the "corrected Hispanic count" from the Non-Black demographic analysis figure. Table 1 below

displays the performance of DA1, DA2, DA3 and the census as estimates of total population for states with respect to the measures listed below and defined in the Appendix.

MARE ≡ Mean absolute relative error.

RSADP ≡ Ratio of the sum of absolute differences of the adjustment proportions to the census proportions.

PI ≡ Proportion improvement after adjustments.

RNAC ≡ Ratio of the number of adjusted estimates within an interval of "truth" to the number of census counts within the same interval.

RAC ≡ Ratio of adjusted population estimates within an interval of "truth" to the census counts within the same interval.

Table 1. Comparison of Census and Adjustment Methods (States)

		Census	DA1	DA2	DA3
I.	MARE (a)	.0124	.0119	.0110	.0112
II.	RSADP (b)		1.014	1.142	1.088
III.	PI (b)		.505	.707	.688
IV.	RNAC (b,c)		1.100 (22)	1.250 (25)	1.250 (25)
V.	RAC (b)		1.181	1.113	1.113

(a) A smaller number is considered better.
(b) A larger number is considered better.
(c) Numbers in parentheses are counts of states falling in the interval.

The measures in the above table favor DA2 over the other synthetic estimates as well as the census. Table 2 below displays the same estimators and measures when estimation of total population of district offices is of interest. A minor change in the coverage is that the population under consideration is the non-institutional population. The synthetic estimates do better than the census at

this lower level but not nearly as well as at the state level. Note especially that the MARE has at least doubled for all methods.

Table 2. Comparison of Census and Adjustment Methods (DO)

		Census	DA1	DA2	DA3
I.	MARE (a)	.0328	.0308	.0300	.0300
II.	RSADP (b)		1.031	1.051	1.050
III.	PI (b)		.535	.559	.556
IV.	RNAC (b,c)		1.078 (236)	1.123 (246)	1.123 (246)
V.	RAC (b)		1.083	1.137	1.139

(a) A smaller number is considered better.
(b) A larger number is considered better.
(c) Numbers in parentheses are counts of DO's falling in the interval.

The above results assume that PEP 3-8 provides the correct population counts. In the absence of any other direct sub-state estimates of total population such comparisons are the best we can do.

B. Regression Estimation Using PEP Data

Most of the modelling that follows is based on district office PEP 3-8 estimates of population. Several different regression models have been produced and are compared as to how well they predict district office population, assuming PEP 3-8 estimates are the "truth." While regression and synthetic estimation are considered separately here, it is important to point out that should regression be chosen as an adjustment method, synthetic estimation would also be playing a role in adjusting down to lower levels of aggregation. Two types of regression modelling are described below. The first consists of

several unweighted linear regressions and the second involves work by Ericksen and Kadane (1985) using a Bayesian hierarchical model.

Four hundred fourteen of the 422 district offices were used in all of the modelling work based on district offices. It was necessary to eliminate eight of the district offices due to insufficient sample size. Three models of net undercount using unweighted linear regression will be described below and compared later. The assumed model for the three equations is $\underset{\sim}{Y} = \underset{\sim}{X}\beta + \underset{\sim}{\varepsilon}$ where $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2\underset{\sim}{I})$. The carrier variables, $\underset{\sim}{X}$, that predict percent net undercount, $\underset{\sim}{Y}$, are carrier variables formed from census tabulations. The carrier variables selected in the models that follow were chosen based on expert opinions as well as stepwise regression procedures. All variables used are expressed in percent.

In the first two models, described below, all 414 district offices were used to form both equations.

$$Y = -.36 + .17(\text{MINRENT}) \qquad R^2 = .27 \qquad S = 4.1 \qquad (1)$$

$$Y = 1.55 + .20(\text{MINRENT}) - .11(\text{NOHS}) \qquad R^2 = .29 \qquad S = 4.0 \qquad (2)$$

where MINRENT = percent of non-vacant renter occupied housing that is minority

NOHS = percent of total population that has not attended high school

Although model (2) does not appear to be significantly better than model (1), model (2) does seem to do a slightly better job predicting district office populations as can be seen in Table 3.

While we would agree that it does not seem likely that the percent minority renter variable alone explains the undercount problem fully it does appear to be the only carrier variable we feel we can justify including from a model selection viewpoint. One of the ways we examined this issue was to generate dummy noise variables as suggested by Miller (1984). Then using the regression by the leaps and bounds procedure (Furnival and Wilson (1971)) we found the best 10 equations of two carrier variables based on the R^2 criterion. While model (2) was determined the best of the two carrier variable models with an R^2 of .29, the fifth best two carrier variable model had an R^2 of .27 with one of the carrier variables being one of the five dummy noise carrier variables. With noise doing almost as well as the percent of the population not attending high school we have further evidence of the large variability in the district office data. This is a major problem. Due to the large variability it is difficult to fit models with reasonable carrier variables. While undercount is most likely a function of many different factors, given the district office data from 1980 we do not have evidence as to what those factors are except to say that there appears to be a relationship between undercount and minority renters at the district office level. Considering the results from the central city regression (below) we may even conjecture that it is the central cities that are dictating this relationship.

In forming the third model the 414 DO's were split into three groups each represented by its own model. The groups were chosen based on whether the district office was centralized, decentralized or conventional.*

$$\begin{aligned}
\text{Net undercount (centralized)} &= 19.16 + .18\,(\text{MINRENT}) \\
&\quad -.26(\text{LISTCOR}) \qquad R^2 = .38 \\
\text{Net undercount(decentralized)} &= -.68 + .14\,(\text{CROWD}) \\
&\quad +.23\,(\text{BLMALE}) \qquad R^2 = .04 \\
\text{Net undercount (conventional)} &= -2.98 + .11\,(\text{URBAN}) \\
&\quad -.42(\text{CROWD}) + 1.59(\text{FOR7580}) \qquad R^2 = .51
\end{aligned} \tag{3}$$

where

MINRENT = percent nonvacant renter occupied housing that are minority

LISTCOR = percent of occupied housing units that were listed correctly before census day

CROWD = percent of housing units with more than one person per room

BLMALE = percent of population that are Black males 15-39

URBAN = percent of total population that is urban

FOR7580 = percent of total population foreign born and entering U.S. between 1975 and 1980.

*Centralized DO's are located in large cities and canvassed by mail; conventional DO's are located in rural areas and canvassed via enumerators; decentralized DO's were canvassed by mail and constitute the bulk of the DO's.

As can be seen from the three equations above, the minority renter variable, while important in the central city regression, does not appear in the decentralized or the conventional district office equations even though it is the variable most associated with undercount based on the combined set of district offices. While both the centralized and conventional areas can be modelled somewhat adequately, it was not possible to find an adequate model for the decentralized district offices. (We note that roughly two-thirds of their absolute net undercounts were less than two percent.) While other groupings of the district offices based on different variables were attempted the results were not as favorable.

The second method as advocated by Ericksen and Kadane involves the application of Bayesian hierarchical regression models for adjusting the census. These models were developed by Lindley and Smith (1972). Letting $Y = (Y_1,\ldots Y_N)^T$ denote the vector of percent net undercount estimates from the district offices, at the first level of the Bayesian hierarchical model it is assumed that

$$\underset{\sim}{Y} \sim N(\underset{\sim}{\theta},\underset{\sim}{D})$$

$$\underset{\sim}{\theta}^T = (\theta_1,\ldots,\theta_N)$$

is a vector of mean values for $\underset{\sim}{Y}$, and $\underset{\sim}{D} = \text{diag}(d_{11},\ldots,d_{NN})$ is a diagonal matrix of the variances of the net percent undercount estimates which are assumed to be known. Although the true values of the d_{ii}'s are unknown, they have been taken to be equal to their

survey estimates in Ericksen and Kadane's analysis. In addition to this approach, we have experimented with values of the d_{ii}'s obtained from empirical models.

At the second stage in the hierarchical model it is assumed that

$$\underset{\sim}{\theta} \sim N(\underset{\sim}{X}\underset{\sim}{\beta} , \sigma^2\underset{\sim}{I})$$

where $\underset{\sim}{X}$ is a matrix of p carrier variables, $\underset{\sim}{\beta}$ is a vector of unknown parameters, and the value of σ^2 is assumed to be known. In their analysis, Ericksen and Kadane used percent minority, percent conventionally enumerated, and the crime rate as carrier variables to explain percent net undercount for states and cities. In our research we are experimenting with alternative carrier variables in addition to those considered by Ericksen and Kadane. In their analysis as in ours, the true value of σ^2 is actually unknown but taken to be equal to its maximum likelihood estimate.

At the final and third level of the Bayesian hierarchical model it is assumed that

$$\underset{\sim}{\beta} \sim N(\underset{\sim}{\gamma} , \underset{\sim}{\Omega}) \ .$$

This stage is required to express knowledge about how the carrier information, $\underset{\sim}{X}$, explains the mean net undercount vector, $\underset{\sim}{\theta}$. The matrix $\underset{\sim}{\Omega}^{-1}$ denotes how precise this knowledge is

and, as in Ericksen and Kadane's analysis, we let $\underset{\sim}{\Omega}^{-1} = \underset{\sim}{0}$ denoting that our knowledge is uninformative.

Using this Bayesian hierarchical formulation, the estimate of percent net undercount is taken to be the posterior mean of $\underset{\sim}{\theta}$:

$$[\underset{\sim}{D}^{-1} + \sigma^{-2}\underset{\sim}{I}]^{-1} [\underset{\sim}{D}^{-1}\underset{\sim}{Y} + \sigma^{-2} \underset{\sim}{X\hat{\beta}}] .$$

That is, the Bayesian estimate of percent net undercount is a mixture of the survey estimates, $\underset{\sim}{Y}$, and the modelled predictions, $\underset{\sim}{X\hat{\beta}}$, where $\underset{\sim}{\hat{\beta}}$ is a weighted least squares estimate.

Due to our interest in evaluating the carrier variables chosen by Ericksen and Kadane at the district office level and comparing it to our other previously mentioned models, we fit variables very similar to theirs, the only change being that we substituted percent migration for the crime variable. This was necessary because crime rates are not available at the district office level. While the model was to be fit to a state and central city data set which did in fact have the crime variable our intention was to predict district office results. Percent migration was selected as a reasonable proxy for crime. The weighted model was estimated

$$Y = -2.58 + .08\ (\%\text{-MIN}) + .02\ (\%\text{-CONV}) + .04\ (\%\text{-MIGR}) \qquad (4)$$

where

%-MIN = percent of the total population that are Black or Hispanic.

%-CONV = percent of the area enumerated conventionally.

%-MIGR = percent of the population over 5 years old who did not live in the same house 5 years ago.

from the state and central city data using estimated, rather than known variances. To investigate how models (1) - (4) compare when they are used to predict district office population counts, we used the same measures used in the synthetic estimation section. Again we treat estimated PEP 3-8 state population counts as "truth" comparing them to the district office predicted values summed to the state level.

Table 3. Comparison of Adjustment Methods Using Unweighted Models.*

Measure	Model (1)	(2)	(3)	(4)
I. MARE (a)	.0121	.0115	.0100	.0100
II. RSADP (b)	1.481	1.626	1.524	1.515
III. PI (b)	.607	.644	.688	.550
IV. RNAC (b)	1.200(24)	1.200(24)	1.300(26)	1.350(27)
V. RAC (b)	1.040	1.117	.978	1.001

(a) A smaller number is considered better.
(b) A larger number is considered better.
(c) Numbers in parentheses are counts of states falling in the interval.

As described previously the Ericksen and Kadane estimates are a mixture of the survey estimates and the modelled predictions. The modelled predictions are based on the linear model $\underset{\sim}{Y} = \underset{\sim}{X}\beta+\underset{\sim}{\varepsilon}$ where $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2\underset{\sim}{I}+\underset{\sim}{D})$, $\underset{\sim}{D}$ being a diagonal variance covariance matrix whose elements are the estimated variances of $\underset{\sim}{Y}$ from the PEP. In the work below we will be comparing the following two models:

*This table and table 4 are designed to make comparisons between possible adjustment models and should not be interpreted as a definitive statement that these models are better than the census.

$$Y = .22 + .11(\text{MINRENT}) \quad (5)$$

$$Y = -1.90 + .06(\%\text{-MIN}) + .003(\%\text{-CONV}) + .04(\%\text{-MIGR}) \quad (6)$$

Table 4. Comparison of Adjustment Methods Using Both Weighted Models and Ericksen and Kadane Models (Based on 46 States)*

		Model (5a)	(5b)	(6a)	(6b)
I.	MARE (a)	.0112	.0092	.0104	.0088
II.	RSADP (b)	35.707	39.420	35.174	36.828
III.	PI (b)	.758	.758	.758	.758
IV.	RNAC (b,c)	1.316 (25)	1.579 (30)	1.421 (27)	1.632 (31)
V.	RAC (b)	1.430	1.460	1.396	1.468

(a) A smaller number is considered better.
(b) A larger number is considered better.
(c) Numbers in parentheses are counts of states falling in the interval.

Models (5a) and (6a) in Table 4 consist of predictions based on equations 5 and 6. Models (5b) and (6b) are mixtures of the direct estimates and their respective modelled predictions. According to Table 4 model (6b) appears to be the best although not by very much.

Since model (5b) using the minority renter variable does almost as well as the three variable model (both using estimated standard errors) it probably could be used without much if any loss in the precision of the adjustment results.

As mentioned in our discussion of the Ericksen and Kadane work, an assumption of known variances is made. In our work, presented here,

*Because eight DO's did not have sufficient sample size to produce estimates of undercount the states containing them had to be removed from this

we have only estimated variances. We have, however, looked into the possibility of using modelled variances instead of the estimated variances but without much success.

C. Synthetic Estimation Using Post Enumeration Survey Data

Examples of synthetic estimates for undercount adjustment using a different PEP estimate than PEP 3-8 have been illustrated in Diffendal, Isaki and Malec (1982) and Diffendal, Isaki and Schultz (1984). In that application, regional PEP age-race-sex distributions were first raked to PEP state estimated marginals. The resulting cell estimates were used to construct state age-race-sex adjustment factors. It is of interest to repeat the process with the PEP 3-8 data and thereby construct DO estimates and compare them as in Table 2 in section A. We intend to do this in the future.

A somewhat related procedure to that described above is due to Tukey (1981) and briefly reported by Ericksen and Kadane (1985). Rather than design a post enumeration survey (PES) to provide jurisdictional estimates of undercount, e.g., regions, states and large cities, the authors suggest designing a PES to provide estimates of undercount that satisfy the assumptions underlying synthetic estimation. This implies grouping together "areas" that are believed to have the same undercount rate. For example, some of the major variables correlated with undercount are minority and rural-suburban-central city. In this setting, the rural areas of South Dakota, North Dakota, Nebraska and Iowa would be presumed to have similar undercount rates. The

central cities of Baltimore, MD and Washington, DC could also be grouped together.

In this procedure, direct estimates of undercount for "areas" are divided by estimated census counts for the same "areas." These resulting adjustment factors are then regressed on related carrier variables and the regression estimate is mixed with the adjustment factor to produce a final adjustment factor. The final adjustment factors are applied to census counts to obtain synthetic estimates. The suggested adjustment scheme contains both a synthetic estimation and a regression component. We use the word area in quotes to distinguish between traditional geographic areas and those likely to arise in the proposed methodology. It should be recognized that there is much similarity in the ideas presented here with those in the paper by Cohen and Kalsbeek (1974).

As a preliminary step in studying the above synthetic/regression procedure we assume that the PES will again be a dual system but that the sampling units at the last stage will be blocks. The issues of forming sampling strata as well as adjustment "areas" require study. We briefly describe a simulation procedure that we intend to pursue using 1980 Census data that are felt to be associated with undercount. Some of the variables to be considered are allocations by race; urban, rural; race; female headed households; renter occupied; all levels of geography, etc. For our study, some of the variables will be used as a pseudo undercount variable while others will be used to form sampling strata, adjustment "areas" and in the

regression modelling of adjustment factors. It is necessary to create a pseudo undercount variable for the study because direct estimates of census undercount at the "area" level are not likely to be available. Using the allocation variable plus the census count as a pseudo variable for the true number in the population, we can measure the error of the procedure in estimating total population at the enumeration district (ED) level, place level, county level and so on. In this case we are assuming that the allocation variable has a distribution similar to that of the actual undercount and we are assessing the model error due to failure of the assumptions underlying the synthetic procedure. Since the adjustment factors will be estimated in practice (using a dual system estimator in the numerator), a way must be found to simulate the sampling and model errors in the estimation of the adjustment factor. We are currently trying to solve this issue.

While the basic sampling unit is the block, without a special tabulation of the census variables to the block level, the data variables mentioned in the above are available at the ED level (combination of from 1 to 20 blocks) only. We plan to conduct our preliminary study at the ED level initially, then proceed to a block level analysis.

IV. Summary

We have attempted to present a brief account of our current efforts in developing methods for census undercount adjustment in small areas. In the time allotted, it was not possible to present other issues that

affect adjustment. For example, the adjustment method must allow for the timeliness of census operations and its operating schedule. The adjustment method must be able to produce output consistent with census publication output and be internally consistent as well. Methodology needs to be developed to add or delete persons from census files in accordance with undercount adjustment estimates. All of these are currently being addressed with regards to a test of adjustment related operations to be conducted next year.

Lastly, we are currently researching how to assess the effectiveness of each of our adjustment methodologies. Determining which methodology is the best will pose special difficulties since the standard by which we would like to rank our methodologies, the actual population sizes in specific geographical areas, is unknown. All of our efforts have essentially used variables with some deficiencies. It has been suggested that since the Black population is affected very little by illegal immigration that demographic analysis U.S. figures be used. Use of such a standard (if indeed it is true) may be satisfactory for the total Black population but the problem of a standard for other races remains.

Appendix

Definition of measures I. through V.

I. $$\text{MARE} = L^{-1} \sum_{i=1}^{L} |PEP_i^{-1} (E_i - PEP_i)|$$

where E_i = denotes the estimated total of area i using method E

L = number of areas.

II. $RSADP = (PSAE^E)^{-1} (PSAE^C)$

where $PSAE^C = \sum_{i=1}^{L} |P_i^C - P_i^T|$

$PSAE^E = \sum_{i=1}^{L} |P_i^E - P_i^T|$

$P_i^C = (\sum_{i=1}^{L} census_i)^{-1} census_i$

$P_i^T = (\sum_{i=1}^{L} PEP_i)^{-1} PEP_i$

$P_i^E = (\sum_{i=1}^{L} E_i)^{-1} E_i$

III. $PI = (\sum_{i=1}^{L} PEP_i)^{-1} \sum_{i=1}^{L} IMPV_i$

where

$$IMPV_i = \begin{cases} PEP_i & \text{if } |P_i^E - P_i^T| < |P_i^C - P_i^T| \\ 0 & \text{otherwise .} \end{cases}$$

IV. $RNAC = C^{-1}E$

where $E = \sum_{i=1}^{L} R_i$, $C = \sum_{i=1}^{L} S_i$

$$R_i = \begin{cases} 1 & \text{if } E_i \in D_i \\ 0 & \text{otherwise} \end{cases}$$

$$S_i = \begin{cases} 1 & \text{if census}_i \in D_i \\ 0 & \text{otherwise} \end{cases}$$

$$D_i = PEP_i \pm V(PEP_i)^{1/2}$$

$V(PEP_i)$ = estimated variance of PEP_i

V. $RAC = (C')^{-1}E'$

where $E' = \sum_{i=1}^{L} R_i'$, $C' = \sum_{i=1}^{L} S_i'$

$$R_i' = \begin{cases} PEP_i & \text{if } E_i \in D_i \\ 0 & \text{otherwise .} \end{cases}$$

$$S_i' = \begin{cases} PEP_i & \text{if census}_i \in D_i \\ 0 & \text{otherwise .} \end{cases}$$

References

1. Bailar, Barbara A. (1983). Affidavit submitted to U.S. District Court, Southern District of New York, in Cuomo vs. Baldridge, 80 Civ., 4550 (JES).

2. Bryce, Herrington J (1980). "The Impact of the Undercount on State and Local Government Transfers," Conference on Census Undercount, U.S. Government Printing Office, 112-124, Washington, DC.

3. Chandrasekar, C. and Deming, W. Edwards (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration," Journal of the American Statistical Association, 44, 101-115.

4. Coale, Ansley J. and Rives, Norfleet W., Jr. (1973). "A Statistical Reconstruction of the Black Population of the United States, 1880-1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility," Population Index, 39(1), pp. 3-36.

5. Coale, Ansley J. and Zelnik, Melvin (1963). "New Estimates of Fertility and Population in the United States," Princeton University Press, Princeton, NJ.

6. Cohen, S. B. and Kalsbeek, William (1977). "An Alternative Strategy for Estimating the Parameters of Local Areas," Proceedings of the Social Statistics Section of the American Statistical Association, Washington, DC.

7. Cowan, Charles D. and Bettin, Paul J. (1982). "Estimates and Missing Data Problems in the Postenumeration Program," technical report, U.S. Bureau of the Census, Washington, DC.

8. Diffendal, Gregg J., Isaki, Cary T., and Malec, Donald J. (1982). "Examples of Some Adjustment Methodologies Applied to the 1980 Census," technical report, U.S. Bureau of the Census, Washington, DC.

9. Diffendal, Gregg, Isaki, Cary and Schultz, Linda (1984). "Small Area Adjustment Methods for Census Undercount," invited papers to the Data Users Conference on Small Area Statistics, U.S. Department of Human Services, Washington, DC, 52-56.

10. Ericksen, Eugene P. (1974). "A Regression Method for Estimating Population Changes of Local Areas," Journal of the American Statistical Association, 69, 867-875.

11. Ericksen, Eugene P. and Kadane, Joseph B. (1985). "Estimating the Population in a Census Year - 1980 and Beyond," Journal of the American Statistical Association, 80, 98-109.

12. Fay, Robert E., III and Herriot, Roger A (1979). "Estimates of Income for Small Places - An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association, 79, 269-277.

13. Furnival, George M. and Wilson, Robert W. Jr. (1974). "Regression by Leaps and Bounds," Technometrics, 16(4), 499-511.

14. Hill, Robert (1980). "The Synthetic Method: Its Feasibility for Deriving the Census Undercount for States and Local Areas," Conference on Census Undercount, U.S. Government Printing Office, Washington, DC, 129-141.

15. Lindley, D.V. and Smith, A.F.M. (1972). "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, Ser. B, 34, 1-19.

16. Miller, Alan J. (1984). "Selection of Subsets of Regression Variables," Journal of the Royal Statistical Society, Ser. A, 147, 389-425.

17. Passel, Jeffrey S. and Robinson, J. Gregory (1984). "Revised Estimates of the Coverage of the Population in the 1980 Census Based on Demographic Analysis: A Report on Work in Progress," paper presented at the Meetings of the American Statistical Association.

18. Purcell, Noel, and Kish, Leslie (1979). "Estimation for Small Domains," Biometrics, 35, 365-384.

19. Robinson, J. Gregory and Siegel, Jacob S. (1979). "Illustrative Assessment of the Impact of Census Underenumeration and Income Underreporting on Revenue Sharing Allocations at the Local Level," paper presented at the Meetings of the American Statistical Association.

20. Schirm, Allen L., and Preston, Samuel H (1984). "Census Undercount Adjustment and the Quality of Geographic Population Distributions, technical report, University of Pennsylvania.

21. Siegel, Jacob S. (1968). "Completeness of Coverage of the Nonwhite Population in the 1960 Census and Current Estimates, and Some Implications," Social Statistics and the City, Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University.

22. Siegel, Jacob S., and Jones, Charles D. (1980). "The Census Bureau Experience and Plans," Conference on Census Undercount, U.S. Government Printing Office, Washington, DC, 15-24.

23. Slater, Courtenay M. (1980). "The Impact of Census Undercoverage on Federal Programs," Conference on Census Undercount, U.S. Government Printing Office, Washington, DC, 107-111.

24. Tukey, John W. (1981). Discussion of "Issues in Adjusting the 1980 Census Undercount," by Barbara Bailar and Nathan Keyfitz, paper presented at the Annual Meeting of the American Statistical Association, Detroit, MI.

25. U.S. Bureau of the Census (1982). "Coverage of the National Population in the 1980 Census by Age, Race, Sex," Current Population Reports, Ser. P-23, No. 115, U.S. Government Printing Office, Washington, DC.

26. Warren, Robert (1981). "Estimation of the Size of the Illegal Alien Population in the United States," Agenda Item B of the meeting of the American Statistical Association - Census Bureau Advisory Committee, November.

27. Warren, Robert and Passel, Jeffrey S. (1983). "Estimates of Illegal Aliens from Mexico Counted in the 1980 United States Census," paper presented at the Meetings of the Population Association of America.