BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-87/01

REPORT ON USING REGRESSION MODELS FOR SMALL AREA ADJUSTMENT

by

Linda K. Schultz, Cary T. Isaki, Gregg J. Diffendal
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
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: Kirk M. Wolter

Report completed: January 15, 1987

Report issued: January 15, 1987

REPORT ON USING REGRESSION MODELS FOR SMALL AREA ADJUSTMENT

By

L. Schultz, C. Isaki, G. Diffendal

I. INTRODUCTION

A. Background

This report summarizes the work of the Census Undercount Adjustment for Small Area Group pertaining to the use of regression methodology to implement an adjustment of census counts down to the small area level. While this report concentrates on the investigation and findings based on the use of various regression models it is important to bear in mind that a regression model formed at one level of aggregation does not necessarily apply to another level of aggregation. For example, while a model formed at the state level may perform quite well for states, use of this model at a county level (or other small area) may involve the use of explanatory variables with levels outside the range at which the model was built--this extra extrapolation can result in misleading results. Other reasons why a model formed at the state level may not be applicable to lower levels of aggregation include the possibility that different explanatory variables may be more appropriate at a county or block level as well as the possibility that values for the explanatory variables formed at the state level are not very reliable at the county level. In the work presented here the assumption has been made that after a regression model has been formed at a given level of geography such as state, synthetic estimation would play a role in adjusting down to lower levels of geography such as the block. Synthetic estimation is described in a separate report.

Rather than modelling administrative level estimates of net undercount (states, counties, etc.) via regression, another possible use of regression is to model net undercount estimates that cross administrative boundaries. For example, in statistical synthetic estimation (see Isaki, et.al. 1986), use of a regression model of the adjustment factors is a likely candidate for variance reduction. Regression modelling of such adjustment factors awaits future research.

B. Description of Data Used to Form Regression Models

The data used in the work that follows comes from the 1980 Post Enumeration Program (PEP) and the census.

1. The 1980 PEP was designed to measure the net population undercount for each state and the 16 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 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, containing about 186,000 persons each, 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 an incorrect geographical area, 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 therefore 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

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

the net undercount is defined as

$\hat{Y} = (\hat{N}-N_c)/\hat{N}$. When estimating for a particular geographic area, $\hat{N}$ was first applied to produce separate age-race-sex cell estimates within the area and the results summed over cells. Depending on the size of the area, some categories were collapsed until an adequate amount of sample cases were realized. Dual system estimates were then formed over the collapsed category. Both P- and E-sample estimates include ratio adjustment.

There are deficiencies in the PEP estimates, some of which are mentioned below. The extent to which such deficiencies might affect the regression modelling results that follow is not known. We assume that the deficiencies are not severe enough to affect the results. According to Cowan and Bettin (1982), the proportion of cases in the sample which are missing data after field follow-up is larger than the estimated net undercount. For example, using the PEP 3-8 data, the percent of total persons, Black persons, Non-Black Hispanic persons and Other persons requiring imputation along with their estimated percent net undercounts are provided in Table 1.

Table 1. PEP 3-8 P-Sample Imputation, Follow-up and Undercount Percentages at the U.S. Level

	% Imputation	% Followed-up	% Net Undercount
Total Persons	3.8	15	.8
Black	6.5	20	5.2
Non-Black Hispanic	6.9	18	4.1
Other	3.3	14	-.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 population 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.

2. 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 counts by age-race-sex at small geographic levels, other population characteristics are available for use in adjustment. These include: 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. Such data are available at 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.

C. Measures of Improvement

Several measures of improvement were used in comparing the performance of several regression estimates to the census and the "truth" for each small area of interest. In the work that is to follow PEP 3-8 state estimates are used as the "truth." Regression models were formed at the district office level and then summed to the state level before measures were computed. This was done because the state estimates of population calculated by the PEP are believed to provide a better standard of comparison than PEP 3-8 estimates computed at lower levels of aggregation. In addition, some PEP 3-8 estimates for DO's possessed large sampling errors. We were hesitant in using such estimates as standards with which to compare the regression results, but no other alternatives were available.

The measures of improvement used in the work that follows are described here.

(i) $$\text{MARE} = L^{-1} \sum_{i=1}^{L} |\text{PEP}_i^{-1} (E_i - \text{PEP}_i)|$$

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

L = number of states.

PEP_i = denotes "truth" for state i.

(ii) weighted squared relative error

$$\alpha = \sum_{i=1}^{N} PEP_i \; [(E_i - PEP_i)/PEP_i]^2$$

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

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

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

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

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

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

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

where

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

(v) $RNAC = C^{-1}E$

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

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

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

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

$$V(PEP_i) = \text{estimated variance of } PEP_i$$

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

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

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

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

(vii) $MP1 = \sum^{N} (P_i^T - P_i^E)^2 /P_i^T$

II. Modelling Results

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, assuming PEP 3-8 state estimates are the "truth." 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. 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 (1984) using a Bayesian hierarchical model.

A. Standard Regression Models

Three models of net undercount using unweighted linear regression will be described below and compared later. The assumed model for the three equations is $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ where $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2\underset{\sim}{I})$. The variables, $\underset{\sim}{X}$, that predict percent net undercount, $\underset{\sim}{Y}$, are variables formed from census tabulations. The 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

and S is the estimated standard error.

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 2.

While it does not seem likely that the percent minority renter variable alone explains the undercount problem fully it does appear to be the only variable one can justify including from a model selection viewpoint. One of the ways this issue was examined was by generating dummy noise variables as suggested by Miller (1984). Then using the regression by the leaps and bounds procedure (Furnival and Wilson (1971)) the 10 best equations of two variables based on the R^2 criterion were found. While model (2) was determined the best of the two variable models with an R^2 of .29, the fifth best two variable model had an R^2 of .27 with one of the explanatory variables being one of the five dummy noise variables. With noise doing almost as well as the percent of the population not attending high school there is further evidence of the large variability in the district office data. This is a major problem. The negative sign of the variable NOHS is of concern. We considered model (2) despite this concern to compare its performance with the other models. Due to the large variability it is difficult to fit models with reasonable explanatory variables. While undercount is most likely a function of many different factors, given the district office data from 1980 there is no 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) one 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.*

$$\text{Net undercount (centralized)} = 19.16+.18\ (\text{MINRENT}) -.26(\text{LISTCOR}) \qquad R^2 = .38$$

$$\text{Net undercount(decentralized)} = -.68+.14\ (\text{CROWD}) +.23\ (\text{BLMALE}) \qquad R^2 = .04$$

$$\text{Net undercount (conventional)} = -2.98+.11\ (\text{URBAN}) -.42(\text{CROWD})+1.59(\text{FOR7580}) \qquad R^2 = .51 \qquad (3)$$

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.

Using indicator variables it was possible to combine the three models listed above into one model. This allows the R^2 and S (standard error)

*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.

values to be compared with those of models 1 and 2. The R^2 for the combined model is .32, the adjusted R^2 is .305 and S is 4.0.

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 variables such as whether a district office was prelisted or not were attempted the results were not as favorable.

To investigate how models (1) - (3) compare when they are used to predict district office population counts, we used the measures described in the previous section. We treat estimated PEP 3-8 state population counts as "truth" comparing them to the district office predicted values summed to the state level. In Table 2 below, data in column (4) are the results of a weighted regression model discussed in section B that follows. The results were combined with that of the simple regression models to conserve space.

Table 2. Comparisons of Adjustment Methods Using Models (1) - (3) and an Examination of the Synthetic Assumption in (4)*
(Based on 51 States)

	Model				
Measure	(1)	(2)	(3)	(4)	Census
MARE (a)	.0121	.0115	.0100	.0100	.0124
α	39467	34631	38298	36837	47950
RSADP (b)	1.105	1.226	1.150	1.143	
PI (b)	.580	.591	.630	.583	
RNAC (b, c)	1.200(24)	1.200(24)	1.300(26)	1.350(27)	
RAC (b)	1.040	1.117	.978	1.001	
MP1x10^{-3}	.17	.15	.17	.16	.21

(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 can be seen from Table 2 all of the measures of improvement indicate that the adjustment models described above improve upon the unadjusted census assuming PEP 3-8 does represent the truth accurately. Overall, there appears to be little difference among the three models in regard to the measures of improvement. In almost all cases an improvement over the census is indicated.

B. Weighted Regression Model Used to Examine Synthetic Assumption

In their work with Bayesian hierarchical models Ericksen and Kadane formed models with three explanatory variables. They chose % minority, % conventional and a crime variable. The models they formed were at the state and central city level. Here we are interested in evaluating their choice of variables at the district office level. Since crime is not available at the DO level, % migration was substituted. The model was fit at the central city and balance of state level of aggregation and

*This table and table 3 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.

then used to predict district office levels of undercount. The main purpose here was to examine the synthetic assumption by fitting the model at the state and central city level and then applying it at the district office level of aggregation. The model formed using weighted regression was estimated (assuming $\underset{\sim}{\varepsilon} \sim N(0, \sigma^2 \underset{\sim}{I} + \underset{\sim}{D}))$

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

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 (i.e., $\underset{\sim}{D}$).

In Table 2 besides examining and comparing models (1)-(3) the measures of improvement were also computed for model (4) when model (4) is applied to the district office level to examine the synthetic assumption. Table 2 indicates that the synthetic assumption has provided comparable results to the other three models.

C. Empirical Bayes Estimation

Our main reference sources concerning Empirical Bayes (EB) estimation with application to undercount adjustment is Ericksen and Kadane (1984) and Freedman and Navidi (1984). The authors of the first paper utilize an hierarchical model (including a regression) of parameter distributions to provide estimates of net undercount at the state level. Their estimator consists of a weighted average of the directly computed net undercount from the PEP state and that obtainable from the regression. The weights consist of estimated model and sampling variances.

The first part of this section briefly lays out the Bayesian hierarchical model. The second part of this section considers the application of Empirical Bayes estimation in connection with 1980 PEP data. Comparisons are made between a model proposed by Ericksen and Kadane and a more parsimonious model with respect to some measures of improvement. Finally, application of EB estimation in conjunction with statistical synthetic estimation is discussed. The Freedman and Navidi paper emphasizes concern with applicability of model assumptions in Bayesian modelling of the PEP data. In the discussion below, we assume that the model assumptions hold.

1. Bayesian hierarchical regression model.

Ericksen and Kadane (1984) advocated the use of EB estimation of total population at the state level using PEP data. The underlying Bayesian hierarchical regression models were developed by Lindley and Smith (1972). Letting $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)^T$ denote the vector of percent net undercount estimates for states, at the first level of the Bayesian hierarchical model it is assumed that

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

$$\underset{\sim}{\phi}^T = (\phi_1, \ldots, \phi_n) \qquad (5)$$

is a vector of mean values for $\underset{\sim}{Y}$, and $\underset{\sim}{D} = \text{diag}(d_{11}, \ldots, d_{nn})$ is a known diagonal matrix of the variances of the net percent undercount estimates. Although the true values of the d_{ii}'s are unknown in the PEP application, they have been taken to be equal to their survey estimates in Ericksen and Kadane's analysis and in the analysis that follows.

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

$$\underset{\sim}{\phi} \sim N\,(\underset{\sim}{X}\,\underset{\sim}{\beta},\ \sigma^2\underset{\sim}{I}) \qquad (6)$$

where $\underset{\sim}{X}$ is an nxp matrix of p explanatory variables, $\underset{\sim}{\beta}$ is a px1 vector of unknown parameters and the value of σ^2 is assumed to be known. In Ericksen and Kadane's analysis, as in ours, the true value of σ^2 is 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

$$\underset{\sim}{\beta} \sim N\,(\underset{\sim}{\gamma},\ \underset{\sim}{\Omega}). \qquad (7)$$

This stage is required to express knowledge about how the explanatory information, $\underset{\sim}{X}$, explains the mean net undercount vector, $\underset{\sim}{\phi}$. The matrix $\underset{\sim}{\Omega}^{-1}$ denotes how precise this knowledge is and in Ericksen and Kadane's analysis $\underset{\sim}{\Omega}^{-1} = \underset{\sim}{0}$ denoting that knowledge is uninformative.

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

$$[\underset{\sim}{D}^{-1} + \sigma^{-2}\underset{\sim}{I}]^{-1}\ [\underset{\sim}{D}^{-1}\underset{\sim}{Y} + \sigma^{-2}\underset{\sim}{X}\,\hat{\underset{\sim}{\beta}}]\ . \qquad (8)$$

That is, the Bayesian estimate of percent net undercount is a mixture of the survey estimates, $\underset{\sim}{Y}$, and the modelled predictions, $\underset{\sim}{X}\,\hat{\underset{\sim}{\beta}}$, where $\hat{\underset{\sim}{\beta}}$ is a weighted least squares estimate. The estimator in (8) is termed an EB estimate because the relevant parameters are estimated from the data. The EB estimate is appealing in the following sense: Each component is weighted inversely to its variance so that in (8), if the sampling error of a component of $\underset{\sim}{Y}$ is large, its contribution toward estimating ϕ_i is reduced. Conversely, if σ^2 is large, denoting a poor fit of the regression model, less reliance is made of the modelled prediction.

2. Application of EB estimation to PEP 3-8 DO estimates.

In this section we construct modelled predictors of total population using PEP 3-8 DO estimates. As a standard, we use the survey based PEP 3-8 total population estimates for states. The modelled predictions as well as the EB estimates in (8) are compared using some measures of improvement. The modelled predictions are based on the linear model $\underset{\sim}{Y} = \underset{\sim}{X}\, \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ where $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0},\ \sigma^2 \underset{\sim}{I} + \underset{\sim}{D})$ and $\underset{\sim}{D}$ is the diagonal variance covariance matrix whose elements are the estimated sampling variances of $\underset{\sim}{Y}$ from the PEP 3-8.

Using MINRENT as a single explanatory variable in estimating $\hat{\beta}$ from the PEP 3-8 DO data we have in (9) below,

$$\hat{Y} = .22 + .11\ (\text{MINRENT} \qquad (9)$$

This differs from (1) in that $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}\ \sigma^2\ \underset{\sim}{I}+\underset{\sim}{D})$ rather than $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0},\sigma^2\ \underset{\sim}{I})$. Using the explanatory variables favored by Ericksen and Kadane (%-MIN, %-CONV and %-MIGR) and modelling the PEP 3-8 DO data, we have for $\underset{\sim}{X}\hat{\underset{\sim}{\beta}}$

$$\hat{Y} = -1.90 + .06\ (\%\text{-MIN}) + .003\ (\%\text{-CON}) + .04\ (\%\text{-MIGR}) \qquad (10)$$

This differs from (4) because (4) was fit at the state and central city level while (10) was fit at the DO level. Actually, as previously mentioned Ericksen and Kadane used the crime rate as an explanatory variable rather than the %-MIGR variable. However, crime rate is not measured well at lower geographic levels and is not available at the DO level. Since our intention is to model the DO data and %-MIGR was felt to be somewhat correlated with the crime rate, we used the migration variable instead.

In Table 3 below we present several measures of improvement of each of the four models/estimators of percent net undercount derived at the DO level. We do this by first using the predicted net undercount at the DO level (or using (8)), converting it to the predicted total population of the DO and summing over all DO's separately, in each state. The resulting state figures are compared with the directly computed PEP 3-8 state estimates. Because eight DO estimates (in four states) were omitted from analysis, our standard includes 46 state estimates. Our manner of assessment of the quality of the model predictions and their averaging as presented in (8) is admittedly tangential. However, lacking the actual net undercount for DOs, comparison of their effectiveness at the state level is the best procedure we could devise.

The four predictors/estimators used in Table 3 are denoted 9a, 9b, 10a and 10b. The digit refers to the prediction equation numbers displayed previously and the letter "a" denotes use of the modelled predictions alone. The letter "b" denotes the EB estimate whose form is defined in (8). The first column of data headed by "DO" represents the results of using the sum of the directly computed PEP 3-8 DO estimates within states as an estimate of state total population. The sum of the PEP 3-8 DO estimates within states does not necessarily agree with the

PEP 3-8 directly computed state estimates due to the post-stratification method used to compute the direct estimates. In terms of (8), and at the DO level, each of the directly computed PEP 3-8 DO estimates represents the opposite limit to 9a and 10a in the weighted average . At the DO level and assuming the hierarchical model, 9b and 10b provides smaller sum of expected squared error loss than DO and 9a and 10a, respectively when summed over all DO's. When comparing the estimators' performance at the state level according to our measures of improvement, "DO" outperforms all others. The results are presented in Table 3. Hence, our regression modelling efforts at the DO unit level when applied for comparison at the state level indicates that a trade-off is being made between improving accuracy of estimation for individual DO's and their use in estimating the aggregate at the state level. Ratio estimation (to state PEP totals) can be used with (8). This would likely improve the accuracy of (8) at the state level but its effect at the DO level is not known.

Table 3. Measures of Improvement of Two Modelled Predictors and Two EB Estimators of State Total Population (46 states)

	Model					
Measure	DO	(9a)	(9b)	(10a)	(10b)	Census
MARE	.0032	.0112	.0092	.0104	.0088	.0121
α	6620	29449	24650	27786	24661	32974
RSADP	2.461	1.200	1.289	1.138	1.166	
PI	.579	.488	.496	.444	.466	
RNAC	2.211	1.316 (25)	1.579 (30)	1.421 (27)	1.632 (31)	
RAC	1.823	1.430	1.460	1.396	1.468	
MP1x10^{-3}	.03	.16	.13	.15	.14	.19

3.Application of EB Estimation in the Statistical Synthetic Estimator.

In this section the application of applying Empirical Bayes estimation methodology (or its counterpart from variance components methodology) to statistical synthetic estimation is discussed. Statistical synthetic estimation is defined in Isaki et.al. (1986), briefly speaking, it involves grouping persons and areas not by administrative units (such as state or county) but by persons and areas felt to possess similar undercount rates. An adjustment factor is then formed for each of these groups. For estimating a particular administrative unit the adjustment factors are then applied appropriately to all individuals within each group within a given administrative unit. The adjusted counts are combined to obtain state and county estimates. In Schultz et.al. (1986) it was observed that sampling errors of the adjustment factors affected the measures of performance of the statistical synthetic estimators to the extent that competitors, inferior to it in the absence of sampling error, became comparable in the presence of sampling error. Work examining the possibility of modelling the adjustment factors used to adjust the census counts shows promise in reducing the effect of sampling error on the adjustment factors. Current work lacked a wide choice of explanatory variables tabulated at the factor level so that even better performance may be possible. More research will have to be done to investigate whether the sampling error can indeed be reduced enough so that the statistical synthetic estimator that was found to be superior without sampling error would still be ruled superior after variance reduction methods to reduce sampling error have been applied.

Cressie (1986) is exploring the use of EB estimation within the context of statistical synthetic estimation but targeting the estimation of state totals. His idea is to average the adjustment factors, currently defined for a Census division, with comparable factors for a specified state within the division. In his model, Cressie omits the regression model component although it could, as he states, be considered as well.

D. Other Approaches

Based on suggestions made, other models of undercount were also investigated. Several of the avenues investigated will be discussed here.

1. Modelling of Undercount By Race

The modelling of undercount by race at the state level was attempted. With the explanatory variables available we were not able to form any models that might explain the differences in the three race groups. We were able to show that the three race groups do exhibit differential undercount at the state level. To examine the differential undercount issue at a smaller level of geography we grouped district offices. While we plan to model the district office group this work has not been completed.

2. Modelling of P and E Separately

Using district office data the proportion matched and the proportion correctly enumerated were tabulated separately. This was done so that the issue of modelling the P- and E- sample information separately and combining the two results to use as another predictor of the dual system estimator could be examined.

The model chosen as a predictor of proportion matched was

$$\text{PROP-M} = .98 - .16(\text{MINRENT}) - 1.18(\text{SUBS}) \qquad (11)$$

$R^2 = .60 \qquad S = .028$

where PROP-M = proportion of the population in both the census and the P sample.

MINRENT = proportion of the renter occupied housing rented by Blacks or Hispanics.

SUBS = proportion of total population that are substituted.

Modelling of the proportion enumerated correctly was a much more difficult task. Plots of the data did not indicate any trends. Therefore, the best model to predict the proportion correctly enumerated appears to be the average level which was .9668 for the district office data set.

The dual system estimate has been defined as $\frac{N_p(N_c - EE - II)}{M}$. Using our prediction equation for the proportion matched we have an estimate for $\frac{N_p}{M}$. We estimate N_c-EE-II using the average proportion correctly enumerated multiplied by the Census = .9668* Census.

Using the above equations the dual system estimate of population was predicted for each district office and then all of the district offices within a state were summed to arrive at an estimate of the state population. Then using the same measures of improvement defined earlier we compared these numbers to PEP 3-8 results. Observe the table below.

Table 4. Comparison of Census and Adjustment Methods Based on Modelling the P and E Separately
(Based on 51 States)

	Census	SEP-P-E
MARE(a)	.0124	.0166
α	47950	98051
RSADP(b)		.879
PI(b)		.568
RNAC(b,c)		1.000(20)
RAC(b)		.897
$MP1x10^{-3}$	.21	.26

(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.

Comparing these results with other results based on 51 states we find that the method termed SEP-P-E consisting of modelling the P and E components of the dual system estimator separately appears to be inferior to other adjustment models described previously. While we seem to be able to do a fairly good job of modelling the proportion matched we have a very difficult time with the proportion correctly enumerated which may be causing the poor result when the two models are combined.

III. CONCLUSIONS AND RECOMMENDATIONS

Using PEP 3-8 as a standard the regression estimators performed better than the census for total population of states according to the measures of improvement used. However, using PEP 3-8 as a standard is somewhat arbitrary since no one has any way of knowing whether PEP 3-8 state estimates of population are closer to the truth than the adjusted results or the census.

As stated previously, building regression models at higher level of geography and applying the same models at lower levels can result in adjusted numbers that are inaccurate for small areas. Likewise, forming models at lower levels of geography also may not be desirable. The errors in the explanatory variables as well as the directly computed undercount estimates could be substantial at low levels of aggregation. Even if one could achieve reasonable levels of accuracy in the explanatory variables at low levels of aggregation it would still be necessary to use some sort of synthetic estimation to adjust down to the block level. Therefore, to consider regression as a viable method of adjustment in 1990, more research would have to be done to determine the level of aggregation at which the model should be formed. This is because it is not clear that regression based DO estimates are satisfactory. After building the model synthetic estimation would need to be used to adjust census counts down to the block level.

In the work involving statistical synthetic estimation it has been documented that sampling variability in forming the adjustment factors becomes a factor in the selection of the superior statistical synthetic method. However, using a regression approach to model the factors it is possible to reduce the variability. While preliminary results indicate improvements in our ability to reduce the effect of sampling error, the modelling of adjustment factors needs more attention and research if statistical synthetic estimation is to be considered a viable alternative in a possible census adjustment.

V. References

1. 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, D.C.

2. Cressie, N.H. (1986), Note on Empirical Bayes Modelling, private communication.

3. Ericksen, E.P. and Kadane, J.B. (1984), "Estimating the Population in a Census Year: 1980 and Beyond," Technical Report No. 260, Carnegie-Mellon University, Pittsburgh, Dept. of Statistics.

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

5. Freedman, D.A. and Navidi, W.C. (1984), "Regression Models for Adjusting the 1980 Census," Technical Report No. 35, University of California, Berkeley, Dept. of Statistics.

6. Furnival, G.M. and Wilson, R.W., Jr. (1974), "Regression by Leaps and Bounds," Technometrics, 16 (4), 499-511.

7. Isaki, C.T., Schultz, L.K., Smith, P.J. and Diffendal, G.J. (1985), "Small Area Estimation Research for Census Undercount--Progress Report," SRD Report Series RR-85-07, Bureau of the Census, Washington, D.C.

8. Isaki, C.T. (1986a), "Report on Statistical Synthetic Estimation," draft of internal report, Statistical Research Division, Bureau of the Census.

9. Isaki, C.T., Diffendal, G.J. and Schultz, L.K. (1986), "Statistical Synthetic Estimates of Undercount for Small Areas", Proceedings of the Bureau of the Census' Second Annual Research Conference, pg. 557-569, Reston, Virginia

10. 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.

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

12. Schultz, L.K., Huang, E.H., Diffendal, G.J. and Isaki, C.T. (1986), "Some Effects of Statistical Synthetic Estimation on Census Undercount of Small Areas," paper to be presented at the Annual Meeting of the American Statistical Association, Survey Methodology Section, Chicago, IL.

13. Tukey, J.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.