Survey Methodology, December 1988
Vol. 14, No. 2, pp. 191–208
Statistics Canada

When Are Census Counts Improved by Adjustment?

NOEL CRESSIE[1]

ABSTRACT

There are persuasive arguments for and against adjustment of the U.S. decennial census counts, although many of them are based on political rather than technical considerations. The decision whether or not to adjust depends crucially on the method of adjustment. Moreover, should adjustment take place using say a synthetic-based or a regression-based method, at which level should this occur and how should aggregation and disaggregation proceed? In order to answer these questions sensibly, a model of undercount errors is needed which is "level-consistent" in the sense that it is preserved for areas at the national, state, county, *etc.* level. Such a model is proposed in this article; like subareas are identified with strata such that within a stratum the subareas' adjustment factors have a common stratum mean and have variances inversely proportional to their census counts. By taking into account sampling of the areas (*e.g.*, by dual-system estimation), empirical Bayes estimators that combine information from the stratum average and the sample value, can be constructed. These estimators are evaluated at the state level (51 states, including Washington, D.C.), and stratified on race/ethnicity (3 strata) using data from the 1980 post-enumeration survey (PEP 3-8, for the noninstitutional population).

KEY WORDS: Emprical Bayes estimation; Loss functions; Measures of improvement; Quantile function; Spatial correlation; Synthetic estimation.

1. INTRODUCTION

This article is of a technical nature, but it is important to present a brief explanation of the political and social ramifications of the "undercount issue" in the United States of America. By December 31 of the year of the decennial census, the U.S. Census Bureau is specified by law to submit state population counts to Congress for the purpose of reapportionment of the House of Representatives, and by March 31, 1991, to submit small-area population counts for the purpose of redistricting. In recent decades, the number of uses to which census data are put have multiplied: revenue-sharing formulas use population and per capita income for each incorporated place, demographic and sociological research at regional, state, and national levels usually rely on census counts, *etc.*

Inaccurate census counts should be cause for concern to the whole nation. That certain groups of people (young black males, illegal aliens, *etc.*) are harder to count than others, is without question; see Ericksen and Kadane (1985), and Freedman and Navidi (1986), and the discussion following these articles. If the hard-to-count groups were distributed in equal proportions throughout the political and administrative regions of the USA there would be far less controversy over what to do about the uncounted people. As it is, many of the large American cities such as Chicago, Detroit, New York, and Los Angeles feel they are losing federal funds because their cities contain more of the types of people that tend to remain uncounted. And certain states such as New York and California feel they are under-represented in Congress, to the benefit of Midwestern states such as Indiana and Iowa.

[1] Noel Cressie, Department of Statistics, Iowa State University, Ames, IA 50011.

Census undercount is defined simply as the difference between the true count and the census count, expressed as a percentage of the true count. My approach to its estimation is model-based, relying on data obtained from the post-enumeration survey (PES). A number of technical aspects of a model-based approach to adjustment will be addressed in this article. Section 2 establishes the model, addresses the question of choice of measures of improvement, and presents results for aggregation and disaggregation based on Bayes and Synthetic estimators. Section 3 gives *empirical* Bayes versions of the results of Section 2. Section 4 summarizes what has been learned from this model-based approach; there is also discussion of the implications of the sufficient conditions that guarantee risks of adjusted counts to be smaller than risks of census counts.

2. THE MIXTURE MODEL AND ITS CONSEQUENCES

At the outset I would like to explain the source of random variation in my model, originally defined in Cressie (1986), and further developed in Cressie (1988). I consider the true population in any well-defined stratum of the USA, to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. In other words, all inference will be performed *conditionally* on the observed census counts.

2.1 The Model

The method of *synthetic estimation* constructs estimators of undercount at a particular level (*e.g.*, the state level) by summing undercounts of various strata (*e.g.*, demographic strata) over the area being considered (*e.g.*, California), where it is assumed that any stratum has a *constant* proportion of true counts to census count regardless of which area is being considered. For example, it would be assumed that the proportion for young black males is the same for California, Delaware and so on. Most often these strata are defined demographically according to the factors of age, race, and sex. However Tukey (1981) suggested that geographic and urban factors should be added. Two such stratifications of the USA are given in Isaki *et al.* (1986).

The mixture model I am proposing assumes a stratification has been defined already, although in Section 4 there is a suggestion how one might determine *post hoc* whether a chosen stratification is satisfactory.

Suppose there are $j = 1, \ldots, J$ strata, and $i = 1, \ldots, I$ areas (*e.g.*, at the enumeration-district level, $I \simeq 300{,}000$, while at the state level, $I = 51$, including the District of Columbia; for demographic stratification, $J = 30$ say, while for the two stratifications in Isaki *et al.*, 1986, $J = 90$ and $J = 96$. Think of stratum j as fixed (for example, stratum j might be the blacks in central cities in those SMSA's whose population's greater than or equal to 250,000, in the New England Census Division). Then as i ranges from $1, \ldots, I$, a sequence of subareas is generated; the subarea indexed by "ji" refers to that part of the i-th area that has stratum j in it. Only subareas with *nonzero census counts* are considered.

Define

$$Y_{ji} \equiv \text{true count in the } j\text{-th stratum of area } i \tag{2.1}$$

$$C_{ji} \equiv \text{census counts in the } j\text{-th stratum of area } i \tag{2.2}$$

$$F_{ji} \equiv Y_{ji}/C_{ji};\ i = 1, \ldots, I;\ j = 1, \ \ldots, J. \tag{2.3}$$

Suppose for the moment that we know the ratios $\{F_{ji}: j = 1, \ldots, J\}$ for the i-th area. Then from the census counts C_{ji}, the true count Y_i can be calculated.

$$Y_i = \sum_{j=1}^{J} F_{ji} C_{ji}. \tag{2.4}$$

The F_{ji} are often called *adjustment factors*. The strata are constructed so that these adjustment factors $\{F_{ji}: i = 1, \ldots, I\}$ are as homogeneous as possible within the j-th stratum; $j = 1, \ldots, J$ (Tukey 1981).

Realistically the adjustment factors are never known; synthetic estimators exploit the homogeneity and replace (2.4) with

$$Y_i^{\text{sya}} = \sum_{j=1}^{J} F_j C_{ji}. \tag{2.5}$$

Now there are only J synthetic adjusment factors $\{F_j: j = 1, \ldots, J\}$ to estimate, which through (2.5) yields an estimate of Y_i. Synthetic estimators have the advantage that the adjustment factors are independent of i and so can be applied to *any* level of aggregation.

The (estimated) adjustment factors could also be modeled by regression on independent variables that may or may not be census variables; for example, percent minority, crime rate, and percent conventionally counted in the census. Consider,

$$Y_i^{\text{reg}} = \sum_{j=1}^{J} \left(\sum_{k=1}^{p} \beta_{k,j} z_{k,ji} \right) C_{ji}. \tag{2.6}$$

To fit the parameters $\beta_{1,j}, \ldots, \beta_{p,j}$ efficiently, various assumptions are made about the error components $\{F_{ji} - \sum_{k=1}^{p} \beta_{k,j} z_{k,ji}\}$, *viz.* independent and identically distributed with mean zero.

Ericksen and Kadane (1985) propose the fitting of a regression relation to $\sum_{j=1}^{J} F_{ji} C_{ji} / \sum_{j=1}^{J} C_{ji}$; $i = 1, \ldots, I$. Freedman and Navidi (1986) criticize the approach and point out the consequences of failure of any of the error assumptions. A problem they did not perceive which I emphasize in (2.7) below, is the heteroskedasticity forced onto the problem by working with ratios; Section 2.2 justifies this model choice. Furthermore, in this latter regression approach undercounts across strata are combined, so that variation between strata is shared by both the regression relation and the error variance. More precise estimators can be obtained through (2.6) by allowing each stratum its own regression relation. Homoskedastic errors and a regression model based on the combination of heterogeneous strata, are also assumed by Ericksen and Kadane (1987) and Ericksen, Kadane and Tukey (1987). It seems that the combination of heterogeneous strata was made necessary by the lack of suitable data.

I do not assume F_{ji}'s that depend only on j, nor a regression relation for the F_{ji}'s, but instead reformulate the synthetic assumption $F_{ji} \equiv F_j$, into a (statistical) homogeneity assumption:

$$F_{ji} \sim N(F_j, \tau_j^2 / C_{ji}); \; i = 1, \ldots, I; \; j = 1, \ldots, J, \tag{2.7}$$

where "$\sim$" means "is distributed as," and $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. Using a regression relation for the mean has the potential of explaining more of the variation of the F_{ji}'s at the risk of introducing bias through misspecification. The strata chosen in Section 3 are based on race; it was decided not to cloud this sensitive issue with selection of controversial regression variables. I shall refer to the model (2.7) as a mixing distribution. *The normality assumption is made for convenience and will be relaxed later.* Here F_j is a fixed but unknown mean to be estimated, and $\tau_j^2 = \mathrm{var}(\sqrt{C_{ji}}\, F_{ji})$ is a parameter I shall call the (standardized) *stratum variance*. As a representation of reality, model (2.7) is better at higher levels of aggregation; see Section 3. All distributions in (2.7) are assumed independent.

There are good reasons for weighting the variance by $1/C_{ji}$ (see Cressie 1987a, Appendix and 1988). The most attractive consequence of model (2.7), is that it is *level-consistent*; that is, it is preserved through different levels of aggregation. Specifically,

$$F_{j,i\&i'} \sim N\left(F_j, \frac{\tau_j^2}{C_{j,i\&i'}}\right), \tag{2.8}$$

where

$$F_{j,i\&i'} \equiv \frac{F_{ji}C_{ji} + F_{ji'}C_{ji'}}{C_{j,i\&i'}}, \text{ and } C_{j,i\&i'} \equiv C_{ji} + C_{ji'}. \tag{2.9}$$

This is a very important property that most of the currently proposed statistical models of undercount do *not* possess. It enables the modeler to escape from the geographical and historical accidents that divided up the country into the states, counties, *etc.*, that we now see.

Of course the $\{F_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}$ are not available as data; if they were, $\{Y_i: i = 1, \ldots, I\}$ would be trivial to calculate. In reality, some sampling takes place so that F_{ji} is observed imperfectly. The best way to think of it is that within stratum j of the i-th area, a sample is taken for undercount. Let the outcome be X_{ji} (*e.g.*, X_{ji} is the ratio of dual-system estimator to census count, for the j-th stratum in the i-th area), and model

$$X_{ji} \sim N(F_{ji}, \sigma_j^2/C_{ji});\ i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{2.10}$$

where F_{ji} is an unknown mean parameter to be estimated, and $\sigma_j^2 = \mathrm{var}(\sqrt{C_{ji}}\, X_{ji})$ is a parameter I shall call the (standardized) *sampling variance*. All distributions in (2.10) are assumed independent. When the number of strata is large, a large PES (say, 300,000 households) is needed to obtain data for each area-stratum combination.

Probability-proportional-to-size sampling was used by the U.S. Census Bureau in its 1980 post-enumeration program, which implies a sampling variance of the form given in (2.10). As a consequence of this weighting, (2.10) is also level-consistent.

2.2 Loss Functions (Measures of Improvement) and their Bayes Estimators

The term loss function is used in statistical decision theory (see, for example, Ferguson 1967) to quantify the loss incurred from using $\hat{\theta}$ as a parameter estimator when the true value is θ. For example, a squared-error loss function is $(\hat{\theta} - \theta)^2$. Adopting a more optimistic terminology, the Census Bureau decided in 1986 to use "measure of improvement" instead of "loss function."

Think of (2.10) as a conditional distribution of X_{ji} *given* F_{ji}, and (2.7) as the mixing (or "prior") distribution of F_{ji}. To predict F_{ji} then, the "*posterior*" distribution of F_{ji} given X_{ji} is needed. Notice that a Bayesian terminology is being used since I am thinking of the F_{ji} as random variables whose collection is modeled according to (2.7). But as well as these random parameters, there are fixed but unknown parameters $\{F_j\}$, $\{\tau_j^2\}$, $\{\sigma_j^2\}$ to be estimated. The posterior of $F_{ji} \mid X_{ji}$ is,

$$\frac{(\text{distribution of } X_{ji} \mid F_{ji}) \cdot (\text{``prior'' of } F_{ji})}{\text{marginal of } X_{ji}}. \tag{2.11}$$

For squared-error loss, the usual Bayes estimator of F_{ji} is simply the expectation of F_{ji} with respect to the posterior: $F_{ji}^{\text{uba}} = E(F_{ji} \mid X_{ji})$. Substituting the model (2.7), (2.10) into (2.11), the posterior distribution is easily obtained (see, for example, Lindley and Smith 1972):

$$F_{ji} \mid X_{ji} \sim N\left(F_j + \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2}(X_{ji} - F_j), \frac{\sigma_j^2 \tau_j^2}{\tau_j^2 + \sigma_j^2}/C_{ji}\right), \tag{2.12}$$

for $i = 1, \ldots, I$; $j = 1, \ldots, J$. Hence the posterior expectation is simply

$$F_{ji}^{\text{uba}} = F_j + D_j(X_{ji} - F_j), \tag{2.13}$$

where $D_j \equiv \tau_j^2/(\tau_j^2 + \sigma_j^2)$. To convert (2.13) into an empirical Bayes estimator, estimators have to be found for F_j and D_j; see Section 3.1.

Although the normality assumptions in (2.7) and (2.10) were used to derive (2.13), more generally (2.13) can be shown to be Bayes for squared-error loss, when assuming simply the mean and variance structure of (2.7) and (2.10), and $E(F_{ji} \mid X_{ji}) = a_{ji} + b_{ji}X_{ji}$. Goldstein (1975) has an even more general result of which this is a special case. For ease of exposition I shall continue to assume normality but it should be remembered that there is a nonparametric optimality for all the estimators considered.

The estimator F_{ji}^{uba} given by (2.13) is Bayes for squared-error loss, within the j-th stratum of the i-th area. Define the estimator of Y_i,

$$Y_i^{\text{uba}} \equiv \sum_{j=1}^{J} F_{ji}^{\text{uba}} C_{ji};\ i = 1, \ldots, I, \tag{2.14}$$

and consider the following general loss function:

$$\sum_{i=1}^{I} (Y_i^{\text{est}} - Y_i)^2 f(C_i), \tag{2.15}$$

where $f(C_i)$ is any nonnegative function of the i-th area's census count. Minimizing (2.15) over all $Y_i^{\text{est}} \equiv \sum_{j=1}^{J} F_{ji}^{\text{est}} C_{ji}$ leads to choosing F_{ji}^{est}'s such that $E\ [\ \sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_{ji}^{\text{est}} (F_{ji}^{\text{est}} - F_{ji})^2 \mid \{X_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}\ ]$ is minimized, where the $\lambda_{ji} \geq 0$ only depend on census counts $\{C_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}$. This minimum is achieved by the estimator (2.14), which shows it to possess a certain robustness since it is optimal regardless of which $f(\cdot)$ is chosen.

In accordance with recommendation 7.2 in National Academy of Sciences (1985), choice of $f(C_i) = 1/C_i$ yields an area's contribution to the total loss that reflects the size of its population. Among the loss functions the Census Bureau has been using, the one most like (2.15) with $f(C_i) = 1/C_i$, is

$$\sum_{i=1}^{I} (Y_i^{\text{est}} - Y_i)^2/Y_i; \tag{2.16}$$

it is "most like" in the sense that it is also a weighted sum of squares where each summand yields an area's contribution to the total loss that reflects the size of its population. Here, undercount in more populous areas receive more weight, so that using such loss functions reflects an emphasis on national considerations. The loss function $\sum_{i=1}^{I} (Y_i^{\text{est}} - Y_i)^2/Y_i^2$, which guarantees undercount equity for the I areas, will not be considered in this article.

It is easy to show that the Bayes estimator in the case of loss function (2.16) is given by,

$$Y_i^{\text{est}} = \left[E\left(\left(\sum_{j=1}^{J} F_{ji}C_{ji}\right)^{-1} \mid \{X_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}\right)\right]^{-1}, \qquad (2.17)$$

which is *not* a linear combination of $\{F_{ji}^{\text{uba}}: j = 1, \ldots, j\}$. However to a first approximation, using the δ-method, it can be shown that this $Y_i^{\text{est}} \simeq Y_i^{\text{uba}}$. This is in fact true for a much larger class of loss functions suggested by Cressie (1987b):

$$L^{\lambda} \equiv \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{I} \left\{ Y_i^{\text{est}}\left[\left(\frac{Y_i^{\text{est}}}{Y_i}\right)^{\lambda} - 1\right] + \lambda[Y_i - Y_i^{\text{est}}]\right\}; \lambda \neq 0, -1; \qquad (2.18)$$

the cases $\lambda = 0, -1$ are defined as the respective limits of L^{λ} as $\lambda \to 0, -1$. Read and Cressie (1988, Chapter 8) show that in this case the Bayes estimator is

$$Y_i^{\text{est}(\lambda)} = \left[E\left(\left(\sum_{j=1}^{J} F_{ji}C_{ji}\right)^{-\lambda} \mid \{X_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}\right)\right]^{-1/\lambda}, \qquad (2.19)$$

which reduces to (2.14) when $\lambda = -1$, and to (2.17) when $\lambda = 1$.

The curious fact is that most undercount estimators used are optimal (under various model assumptions) for $\lambda = -1$, but their performance is measured using $\lambda = 1$; *i.e.*, (2.16). The δ-method argument gives $Y_i^{\text{est}(\lambda)} \simeq Y_i^{\text{uba}}$, and recall Y_i^{uba} is optimal for (2.15); therefore squared-error loss estimators of undercount perform well according to a large class of loss functions. This was observed by Kadane (1984) in his heirarchical Bayesian analysis of 1980 census undercount data ($\lambda = -1$ and $\lambda = -2$ were compared), and confirmed on the studies of artificial populations carried out by Cressie and Dajani (1988).

It has just been demonstrated that the estimators (2.13) and (2.14) are Bayes (or approximately so) for a large class of loss functions. However it is not likely that the ensemble properties of $\{F_{ji}^{\text{uba}}: i = 1, \ldots, I; j = 1, \ldots, j\}$, estimate the corresponding ensemble properties of $\{F_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}$, very well. This follows from the inequality $\text{var}(\theta) \geq \text{var}(E(\theta \mid X))$; in other words the posterior mean of the parameter has a smaller variance than the parameter itself. For estimation of state population totals, this does not matter, but for estimation of the *distribution* of say $\{F_{ji}C_{ji}: i = 1, \ldots, 51\}$; $j = 1, \ldots, J$, or $\{Y_i: i = 1, \ldots, 51\}$, (2.13) is ill-suited to the task. Such a distribution is needed in standards research (Mulry-Liggan and Hogan 1986) to determine the proportion of people in a stratum affected by an undercount more severe than u% (Cressie 1988, Section 4).

I shall constrain the estimator of $\{F_{ji}: i = 1, \ldots, I\}$ so that the posterior moments of its (weighted) empirical distribution function match the moments of the estimator's weighted empirical distribution function. This is achieved by modifying the usual Bayes estimator, yielding a constrained Bayes estimator with the right ensemble properties. Louis (1984) presents the details for an equal-variance version of the model (2.7), (2.10), but a straightforward modification of his approach is possible for weighted variances. Cressie (1986) shows that such a *constrained Bayes estimator* is

$$F_{ji}^{\mathrm{cba}} = \zeta_j + G_j(X_{ji} - \zeta_j), \tag{2.20}$$

$$Y_i^{\mathrm{cba}} = \sum_{j=1}^{J} F_{ji}^{\mathrm{cba}} C_{ji}, \tag{2.21}$$

obtained by solving for ζ_j and G_j in:

$$\zeta_j + G_j(X_j. - \zeta_j) = F_j + D_j(X_j. - F_j);$$

$$G_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh}\right)(X_{ji} - X_j.)^2 =$$

$$(I - 1)D_j\sigma_j^2 / \sum_h C_{jh} + D_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh}\right)(X_{ji} - X_j.)^2, \tag{2.22}$$

where

$$X_j. = \sum_{i=1}^{I} X_{ji}C_{ji} / \sum_{h=1}^{I} C_{jh}. \tag{2.23}$$

2.3 Risks of Adjustment; Model Parameters Assumed Known

The model-based approach described in the previous section specifies undercounts in various area-strata combinations, to be random variables. When it comes to comparing the value of one adjustment procedure against another, the expected loss (or the *risk*) is used. Statistical procedures with small risk are preferred.

In the absence of other considerations (*e.g.*, political, practical, *etc.*), implementing the procedure with the smallest risk is the correct, impartial approach. The statistician knows that adherence to this *modus operandi* will yield better estimates *on the average*, where the average is taken over all problems considered by the statistician. However there is nothing to guarantee that for the particular problem being considered, here estimation of undercount in the 1990 census, a set of area-strata estimates derived from the criterion of minimum risk will actually have smaller loss than another set of estimates. To put it more succinctly, the inequality $E(V^2) < E(W^2)$ does not guarantee that $V^2 < W^2$ for a particular realization. If, in the light of the data collected, a minimum risk prediction did not prove to be the most accurate, the statistical *procedure* should still be seen as optimal.

In the rest of this section, various results about Bayes estimators will be stated (proofs are given in Cressie 1988). Needless to say, these results rely on the correctness of the assumed model. In practice, the more relevant results are for *empirical* Bayes estimators, which are given (with proofs) in Section 3.

The first thing to recall (from Section 2.2) about the usual Bayes estimators (2.13), (2.14) is that they are optimal or near optimal for a large class of loss functions. Moreover the estimators are level-free; *i.e.*, they are not only optimal at the level at which they are constructed, but after aggregation they are also optimal at the higher level. From (2.14),

$$Y_i^{\mathrm{uba}} + Y_{i'}^{\mathrm{uba}} = Y_{i\&i'}^{\mathrm{uba}}, \tag{2.24}$$

where $i\&i'$ denotes the area obtained by combining the two disjoint areas i and i'.

Therefore, one should aim to construct a Bayes estimator at the very lowest level (census blocks) and aggregate up to whatever level is desired, thus ensuring consistency of counts at all levels. In practice this is out of the question, simply because the post-enumeration survey would *never* be large enough to give dual-system estimated undercount data for all the blocks. The same is true at the enumeration-district level and the county level. Moreover, at these lower levels the model (2.7) and (2.10) does not fit as well (Cressie and Dajani 1988); an adequate fit at the state level is shown in Section 3.1.

It is certain that the post-enumeration survey will gather data from each of the 51 states, allowing construction of (empirical) Bayes estimators at the state level. Politically, the state level is the most sensitive; reapportionment of the 50 states' representation (Washington, D.C. is excluded) in the House is the first use made of decennial census counts (mandated to reach Congress by December 31 in the year of the census). Thus at this level, the Bayes estimators (2.13) and (2.14) offer a compromise between a *state's* observed adjustment factors $\{X_{ji}: j = 1, \ldots, J\}$; and the (synthetic) adjustment factors $\{F_j: j = 1, \ldots, J\}$. For example, Mississippi's black undercount is recognized as being potentially different from New York's black undercount, when using the Bayes estimators.

I shall now explore the consequences of *synthetic* estimation at lower levels, after Bayes estimation is carried out at a given level. For consistency of counts at all levels, it is desired to estimate undercount at the block level and aggregate up to whatever level is desired. Suppose an adjustment factor F_{ji}^{est} is estimated for the j-th stratum in the i-th area. Now suppose $i = i_1 \,\&\, i_2$; *i.e.*, the i-th area is split up into two disjoint subareas i_1 and i_2. Then the synthetic method at the lower level posits,

$$F_{ji_1}^{\mathrm{sye}} = F_{ji_2}^{\mathrm{sye}} = F_{ji}^{\mathrm{est}}, \tag{2.25}$$

so that estimators of the true population are given by,

$$Y_{i_1}^{\mathrm{sye}} = \sum_{j=1}^{J} F_{ji_1}^{\mathrm{sye}} C_{ji_1};\; Y_{i_2}^{\mathrm{sye}} = \sum_{j=1}^{J} F_{ji_2}^{\mathrm{sye}} C_{ji_2}. \tag{2.26}$$

Notice that from (2.25) and (2.26).

$$Y_{i_1}^{\mathrm{sye}} + Y_{i_2}^{\mathrm{sye}} = Y_i^{\mathrm{est}} \equiv \sum_{j=1}^{J} F_{ji}^{\mathrm{est}} C_{ji}, \tag{2.27}$$

which is the desired disaggregation-aggregation property.

Compare the risk of using Y_i^{uba}, Y_i^{sya}, and Y_i^{cba} (given by (2.14), (2.5), and (2.21) respectively) to the risk of using C_i, the census count of the i-th area. Using the loss function (2.15), the risks are:

$$\text{uba-risk}_i \equiv E[(Y_i^{\mathrm{uba}} - Y_i)^2 f(C_i)], \tag{2.28}$$

$$\text{cen-risk}_i \equiv E[(C_i - Y_i)^2 f(C_i)], \tag{2.29}$$

$$\text{sya-risk}_i \equiv E[(Y_i^{\mathrm{sya}} - Y_i)^2 f(C_i)], \tag{2.30}$$

$$\text{cba-risk}_i \equiv E[(Y_i^{\mathrm{cba}} - Y_i)^2 f(C_i)]. \tag{2.31}$$

The following sequence of inequalities can be proved (Cressie 1988):

$$\text{uba-risk}_i \leq \text{cba-risk}_i \leq \text{sya-risk}_i \leq \text{cen-risk}_i, \tag{2.32}$$

where the middle inequality requires $\sigma_j^2/\tau_j^2 \leq 3$; $j = 1, \ldots, J$.

Now compare the risk of using $Y_{i_1}^{\text{sye}}$ and $Y_{i_2}^{\text{sye}}$ (estimators of Y_{i_1} and Y_{i_2} respectively) based on F_{ji}^{uba} in (2.25), with the risk of using C_{i_1} and C_{i_2}, where area $i = i_1$ & i_2, the union of disjoint areas i_1 and i_2. It can be shown (Cressie 1988) that the synthetic estimation based on the usual Bayes estimator defined at a particular level but applied at a lower level, always has smaller risk than the census counts.

It is also of interest to determine the behaviour of the census-based risk minus the Bayes-then-synthetic-based risk as a function of the level; the larger this difference, the more advantageous it is to adjust the census counts. Here use $f(C_i) = 1/C_i$ in loss function (2.15). It is possible to show (Cressie 1988) that as disaggregation proceeds to a lower level, the "risk gap" between Bayes-then-synthetic estimation and census counts widens in absolute terms. Although this is proved there for the uba-then-synthetic- based estimator, the same is true for cba-then-synthetic-based and sya-then- synthetic-based estimators, and the ordering of risks (2.32) is preserved at any level of disaggregation. This conclusion depends on the model (2.7) and (2.10) holding at *all* levels. Unfortunately at the lower levels there is some evidence that biases can be substantial. That is, $E(F_{ji}) = F_j + b_{ji}$; $E(X_{ji} \mid F_{ji}) = F_{ji} + d_{ji}$. Realistically b_{ji}'s and d_{ji}'s are *never* zero, but at sufficiently high levels of aggregation they are unimportant. At the block and enumeration-district level they can be substantial (Cressie and Dajani 1988) and could invalidate the risk inequalities proved so far. Moreover, at lower levels, the data $\{X_{ji}\}$ are more variable leading to less precise estimates of $D_j = \tau_j^2/(\tau_j^2 + \sigma_j^2)$ in the *empirical* Bayes version (see Section 3) of the Bayes estimator (2.14). These observations, as well as a recognition of the difference between risk and loss, help to explain the deterioration of the performance of the adjusted counts at lower levels, observed in artificial populations (Schultz *et al.* 1986).

3. EMPIRICAL BAYES ADJUSTMENT OF CENSUS COUNTS

Obtain from (2.14), (2.21), and (2.5), the estimated (or adjusted) true area counts Y_i^{uba}, Y_i^{cba}, and Y_i^{sya}, respectively. In order to make these functions only of the data, estimators are needed for the unknown parameters F_j, τ_j^2, and σ_j^2; Fay and Herriot (1979) give empirical Bayes estimators in a regression setting, of which the model (2.7), (2.10) is a special case. For reasons of statistical consistency (see Cressie 1986, Section 3.3), choose,

$$\hat{F}_j = X_j. \tag{3.1}$$

$$\hat{\tau}_j^2 = \max\left\{\left[\sum_i C_{ji} I(C_{ji} > 0)(X_{ji} - X_{j\cdot})^2 \Big/ \left(\sum_i I(C_{ji} > 0) - 1\right)\right] - \hat{\sigma}_j^2, 0\right\} \tag{3.2}$$

$\hat{\sigma}_j^2$ is obtained from sampling considerations: it is known for dual-system estimation, and Schultz *et al.* (1986) determine it for their artificial populations by replicating probability-proportional-to-size sampling of 1,440 enumeration districts from the approximately 300,000 total number.

Statistical stability (*i.e.*, small sampling variance) for sample means is easier to achieve than for sample variances. The coefficient of variation of the sample variance is approximately $\sqrt{2}/\sqrt{n}$; therefore to achieve a relative confidence region (0.5, 1.5) for the population variance, a value of $n = 32$ is needed; and to achieve a region (0.95, 1.05) a value of $n = 3{,}200$ is needed. Thus the estimator, $\sum_{i=1}^{I} C_{ji} I(C_{ji} > 0)(X_{ji} - X_{j\cdot})^2/(\sum_{i=1}^{I} I(C_{ji} > 0) - 1)$ of $\tau_j^2 + \sigma_j^2$ is very unstable, particularly when there are a large number of strata and hence $\sum_{i=1}^{I} I(C_{ji} > 0)$ is small (smaller than 30).

One way around this is to introduce a further mixing distribution into the problem, namely, model the $\{\tau_j^2 \colon \mathrm{j} = 1, \ldots, J\}$ as being generated by the reciprocal of a gamma distribution for example. Thus instead of estimating J parameters $\{\tau_j^2 \colon \mathrm{j} = 1, \ldots, J\}$, the problem can be reduced to estimating just two gamma parameters (see *e.g.*, Hui and Berger 1983). Another possibility is to aggregate temporarily some of the strata for the purpose of estimating the stratum variance. In other words, define disjoint groups of strata indices, $A_1, \ldots, A_K$, such that $\cup \{A_k \colon k = 1, \ldots, K\} = \{1, 2, \ldots, J\}$, and $\tau_j^2 = \tau_{j'}^2 = T_k^2$, whenever j and j' belong to the same A_k. In this way, Cressie and Dajani (1988) reduce the number of stratum variance parameters from $J = 96$ down to $K = 4$. For the data analyzed below, since $\sum_{i=1}^{I} I(C_{ji} > 0) = 51$ for each of the three race strata, it was not necessary to "borrow strength" in the ways just described.

3.1 Emprical Bayes Estimators

The usual (see, for example, Morris 1983) and constrained (Louis 1984) empirical Bayes estimators can now be constructed:

$$F_{ji}^{\mathrm{ueb}} = X_{j\cdot} + \{\hat{\tau}_j^2/(\hat{\tau}_j^2 + \hat{\sigma}_j^2)\}(X_{ji} - X_{j\cdot}), \tag{3.3}$$

$$Y_i^{\mathrm{ueb}} = \sum_{j=1}^{J} F_{ji}^{\mathrm{ueb}} C_{ji};\ i = 1, \ldots, I; \tag{3.4}$$

$$F_{ji}^{\mathrm{ceb}} = X_{j\cdot} + \{\hat{\tau}_j^2/(\hat{\tau}_j^2 + \hat{\sigma}_j^2)\}^{1/2}(X_{ji} - X_{j\cdot}), \tag{3.5}$$

$$Y_i^{\mathrm{ceb}} = \sum_{j=1}^{J} F_{ji}^{\mathrm{ceb}} C_{ji};\ i = 1, \ldots, I. \tag{3.6}$$

The usual empirical Bayes estimator (3.3) can also be obtained from standard theory for linear models with random effects (Henderson 1976).

Notice that when $\hat{\tau}_j^2 = 0$, the empirical Bayes estimators of the j-th stratum adjustment factors all reduce to the synthetic estimator $X_{j\cdot}$. The presence of the weight $\{\hat{\tau}_j^2/(\hat{\tau}_j^2 + \hat{\sigma}_j^2)\}^{1/2}$ in the constrained empirical Bayes estimator (3.5) may look a little strange at first, but it is seen in Cressie (1987a) to yield an unbiased estimator of the stratum error $C_{ji}^{1/2}(F_{ji} - F_j)$.

An earlier suggestion for empirical Bayes modeling of undercount came from Dempster and Tomberlin (1980), who proposed that the number of undercounted people in a subarea might be a binomial random variable. They defined a heirarchical Bayes model but did not take into account the heteroskedastic variation. Stroud (1987) introduces a covariate into a two-stage Bayesian model, but his assumptions of homoskedastic variation and equal sample sizes in each subarea, are too restrictive for the problem considered in this article.

Formulas for the bias and mean-squared error of the usual empirical Bayes (ueb) estimators (3.3), (3.4), the constrained empirical Bayes (ceb) estimators (3.5), (3.6), and the synthetic estimators

$$F_{ji}^{\text{syn}} = X_{j}. \tag{3.7}$$

$$Y_i^{\text{syn}} = \sum_{j=1}^{J} F_{ji}^{\text{syn}} C_{ji};\ i = 1, \ldots, I, \tag{3.8}$$

are given in Cressie (1987a, Section 4). Since undercount is a nonlinear function of the true population, its estimators based on $\{F_{ji}^{\text{est}}: i = 1, \ldots, I; j = 1, \ldots, J\}$, viz.

$$u_{ji}^{\text{est}} \equiv 1 - \frac{1}{F_{ji}^{\text{est}}};\ i = 1, \ldots, I; j = 1, \ldots, J, \tag{3.9}$$

$$u_i^{\text{est}} \equiv 1 - \frac{\mathrm{C_i}}{Y_i^{\text{est}}};\ i = 1, \ldots, I, \tag{3.10}$$

are biased; estimated biases and mean-squared errors can be obtained by the δ-method (Cressie 1987a, Section 4). All of these bias and mean-squared error calculations do not take into account variation due to the (nonlinear) estimation of $\tau_j^2/(\tau_j^2 + \sigma_j^2)$.

Suppose that the following three U.S. strata (based on race/ethnicity) are chosen: blacks, nonblack hispanics, and others. Data from the post-enumeration survey following the 1980 U.S. Census are given in Cressie (1987a, Table 1). These are from the *noninstitutional* population (Cowan and Bettin 1982) and have been labeled "PEP 3-8" by the U.S. Census Bureau – the "3" refers to census omissions being obtained from an April survey and to imputing missing data, and the "8" refers to erroneous enumerations being obtained from a separate survey that imputed missing data with the help of U.S. Post Office information.

From these data and (3.1), (3.2), Cressie (1987a) estimated the mean of the mixture distribution, and standardized stratum and sampling variances defined in (2.7) and (2.10):

$$\text{blacks:}\quad \hat{F}_1 = 1.06076 \quad \hat{\tau}_1^2 = 673.982 \quad \hat{\sigma}_1^2 = 522.183, \tag{3.11}$$

$$\text{nonblack hispanics:}\quad \hat{F}_2 = 1.04667 \quad \hat{\tau}_2^2 = 308.990 \quad \hat{\sigma}_2^2 = 246.585, \tag{3.12}$$

$$\text{Others:}\quad \hat{F}_3 = 0.99981 \quad \hat{\tau}_3^2 = 242.134 \quad \hat{\sigma}_3^2 = 242.152. \tag{3.13}$$

Based on these parameter estimators and the PEP 3-8 data $\{X_{ji}: j = 1, 2, 3; I = 1, \ldots, 51\}$, Cressie (1987a) gave undercount estimates $\{u_{ji}^{\text{est}}\}$, $\{u_i^{\text{est}}\}$ for ueb-based and syn-based estimators defined by (3.3) and (3.7) respectively.

To check the fit of the model, the residuals $\{C_{ji}^{1/2}(F_{ji}^{\text{ceb}} - F_{j\cdot}^{\text{ceb}}): i = 1, \ldots, I\}$ were computed for each of the three strata. Table 1 shows the results, presented as stem-and-leaf plots for the three race strata; a bell-shaped plot for each is the ideal. The model appears to fit the data, except for the nonblack-hispanic stratum in the state of New York. In light of the lawsuit, Cuomo *vs.* Baldridge, heard by the Southern District Court of New York in 1983, this new way of looking at the data tells an interesting story. The nonblack hispanics in New York State

Table 1

Stem and leaf plots of residuals based on "ceb" estimator

Blacks ($j = 1$)

```
STEM  LEAF                  #
   5  2                     1
   4  58                    2
   3  048                   3
   2  8                     1
   1  259                   3
   0  114566788             9
  -0  876654444322100      15
  -1  83310                 5
  -2  75310                 5
  -3  54                    2
  -4  7                     1
  -5  9872                  4
      ----+----+----+----+
  MULTIPLY STEM.LEAF BY 10**+01

              EXTREMES

LOWEST      ID          HIGHEST       ID
  -59.39(MASSACHU)    34.1312(ILLINOIS  )
-57.7968(GEORGIA  )   38.1621(CALIFORN )
-56.9546(ALABAMA  )     45.401(NEW YORK)
-52.2089(VIRGINIA  )  47.7702(SOUTH CA )
-46.8866(TENNESSE )   52.1999(LOUISIAN  )
```

Nonblack ($j = 2$)

```
STEM  LEAF                       #
   7  1                          1
   6
   5
   4
   3  1                          1
   2  15                         2
   1
   0  111235599                  9
  -0  999877555322211111100     21
  -1  55410000                   8
  -2  4333322                    7
  -3  4                          1
  -4  1                          1
      ----+----+----+----+
  MULTIPLY STEM.LEAF BY 10**+01

              EXTREMES

LOWEST      ID          HIGHEST       ID
-40.7166(ILLINOIS  )   9.31117(MISSOURI  )
  -34.131(CONNECTI)   21.0049(PENNSYLV )
  -24.418(INDIANA   )  24.8855(MARYLAND)
-23.2193(WISCONSI )   30.6594(CALIFORN  )
-23.1509(TEXAS      )  70.7991(NEW YORK)
```

Others ($j = 1$)

```
STEM  LEAF                  #
   4  2                     1
   3  8                     1
   3
   2
   2  012                   3
   1  55568                 5
   1  22                    2
   0  55556889              8
   0  1112344               7
  -0  3200                  4
  -0  98777655              8
  -1  31000                 5
  -1  5                     1
  -2  211                   3
  -2  66                    2
  -3
  -3
  -4  2                     1
      ----+----+----+----+
  MULTIPLY STEM.LEAF BY 10**+01

              EXTREMES

LOWEST      ID          HIGHEST       ID
  -41.719(TENNESSE )   19.6618(WISCONSI )
-26.4394(NEW YORK)    20.5702(ARIZONA  )
  -25.921(PENNSYLV )   22.2817(ILLINOIS  )
-22.0464(LOUISIAN  )       37.52(SOUTH CA)
-20.7451(KENTUCKY)    42.4305(CALIFORN )
```

were grossly undercounted, even in relation to their undercounted fellow nonblack hispanics in other states. Incidentally, the judge decided in favour of the U.S. Department of Commerce (in December 1987) on the grounds that the statistical and demographic professions had not developed adequate methods of adjustment for the whole country by 1980.

When are census counts improved by replacing $\{C_i: i = 1, \ldots, I\}$ with $\{Y_1^{\text{est}}: i = 1, \ldots, I\}$? The next section gives conditions under which an analogous ordering to (2.32) still holds in the *empirical* Bayes setting.

3.2 Adjustment at Different Levels; Model Parameters Estimated

The same comments at the beginning of Section 2.3 apply; in a model-based approach a small risk does not guarantee a small loss in every problem but only on the average. Also the analogous aggregation property to (2.24) holds for ueb-based, ceb-based, and syn-based estimators, namely

$$Y_i^{\text{est}} + Y_{i'}^{\text{est}} = Y_{i\&i'}^{\text{est}} \tag{3.14}$$

for "est" = "ueb," "ceb," and "syn," given by (3.4), (3.6), and (3.8) respectively. Moreover the disaggregation-aggregation property (2.27), namely

$$Y_{i_1}^{\text{sye}} + Y_{i_2}^{\text{sye}} = Y_i^{\text{est}}, \tag{3.15}$$

where $i = i_1 \;\&\; i_2$ and $F_{ji_1}^{\text{sye}} = F_{ji_2}^{\text{sye}} = F_{ji}^{\text{est}}$, holds for any estimator of F_{ji}, including those based on ueb, ceb, and syn.

Write the risk of estimating Y_i by $Y_i^{\text{est}}(= \sum_{j=1}^{J} F_{ji}^{\text{est}} C_{ji})$ as

$$\text{est-risk}_i \equiv E[(Y_i^{\text{est}} - Y_i)^2 f(C_i)]. \tag{3.16}$$

The estimators given by "est" = "ueb," "ceb," and "syn," will be compared to "cen" $(F_{ji}^{\text{cen}} \equiv 1)$ via (3.16). For the rest of this section consider the estimator,

$$F_{ji}^{\text{est}} = r_j X_{ji} + (1 - r_j) X_{j\cdot}\,;\; 0 \le r_j \le 1, \tag{3.17}$$

a convex combination of the data X_{ji} and the synthetic estimator $X_{j\cdot}$. Then

$$\text{est-risk}_i = \sum_{j=1}^{J} \tau_j^2 (1 - r_j)^2 \left\{ C_{ji} - \frac{C_{ji}^2}{\sum_h C_{jh}} \right\} + \sigma_j^2 \left\{ r_j^2 C_{ji} + \frac{(1 - r_j^2) C_{ji}^2}{\sum_h C_{jh}} \right\}. \tag{3.18}$$

It is easy to see that the value of r_j that minimizes (3.18) is $r_j = D_j = \tau_j^2 / (\tau_j^2 + \sigma_j^2)$; *i.e.*, neglecting the effect of estimating τ_j^2 and σ_j^2, I obtain

$$\text{ueb-risk}_i \le \text{est-risk}_i;\; 0 \le r_j \le 1. \tag{3.19}$$

Now compare ueb-risk$_i$ (put $r_j = D_j$ in (3.17)) with cen-risk$_i$; recall from (2.29)

$$\text{cen-risk}_i = \sum_{j=1}^{J} \tau_j^2 C_{ji} f(C_i) + \left[\sum_{j=1}^{J} (F_j - 1) C_{ji}\right]^2 f(C_i). \tag{3.20}$$

Also, by putting $\tau_j^2 = k_j \sigma_j^2$; $j = 1, \ldots, J$,

$$\text{ueb-risk}_i = \sum_{j=1}^{J} \sigma_j^2 \left\{\frac{k_j}{1 + k_j} + \frac{C_{ji}}{\sum_h C_{jh}} \cdot \frac{1}{1 + k_j}\right\} C_{ji} f(C_i). \tag{3.21}$$

A sufficient condition for ueb-risk$_i \leq$ cen-risk$_i$ is,

$$\left\{\frac{k_j}{1 + k_j} + \frac{C_{ji}}{\sum_h C_{jh}} \cdot \frac{1}{1 + k_j}\right\} \leq k_j;$$

that is, if

$$\sigma_j^2/\tau_j^2 \leq \left\{\sum_h C_{jh}/C_{ji}\right\}^{1/2}; j = 1, \ldots, J, \tag{3.22}$$

then

$$\text{ueb-risk}_i \leq \text{cen-risk}_i. \tag{3.23}$$

Similarly, it can be shown that if

$$\sigma_j^2/\tau_j^2 \leq 1; j = 1, \ldots, J, \tag{3.24}$$

then

$$\text{syn-risk}_i \leq \text{cen-risk}_i. \tag{3.25}$$

Finally, if $(\sigma_j^2/\tau_j^2) \leq 1$, and

$$4(\sigma_j^2/\tau_j^2)^2\left(\frac{C_{ji}}{\sum_h C_{jh}}\right)^2 - (\sigma_j^2/\tau_j^2)\left(1 + \frac{2C_{ji}}{\sum_h C_{jh}}\right) + 3 \geq 0; j = 1, \ldots, J, \tag{3.26}$$

then

$$\text{ceb-risk}_i \leq \text{syn-risk}_i. \tag{3.27}$$

once again (from (3.26)), if σ_j^2/τ_j^2 is small, risks can be bounded.

Therefore an analogous sequence of inequalities to (2.32) is possible:

$$\text{ueb-risk}_i \le \text{ceb-risk}_i \le \text{syn-risk}_i \le \text{cen-risk}_i, \tag{3.28}$$

where the middle inequality requires the condition (3.26) and the last inequality requires the condition (3.24). If either of these two inequalities do not hold, at least the ueb-based estimator is an improvement over the census counts if condition (3.22) is satisfied. For the PEP 3-8 data from the 1980 U.S. Census,

$$\hat{\sigma}_1^2/\hat{\tau}_1^2 = 0.77,\ \hat{\sigma}_2^2/\hat{\tau}_2^2 = 0.80,\ \hat{\sigma}_3^2/\hat{\tau}_3^2 = 1.00; \tag{3.29}$$

that is, for the 1980 U.S. decennial census the census risk is larger than the synthetic risk and the usual-empirical-Bayes risk is smallest of all.

Now compare the risk of using $Y_{i_1}^{\text{sye}}$ and $Y_{i_2}^{\text{sye}}$ (estimators of Y_{i_1} and Y_{i_2} respectively, based on F_{ji}^{est} given by (3.17)), with the risk of using C_{i_1} and C_{i_2}, where area $i = i_1$ & i_2 is disaggregated into two disjoint areas i_1 and i_2.

$$\begin{aligned}\sum_{\ell=1}^{2} E\left[(Y_{i_\ell}^{\text{sye}} - Y_{i_\ell})^2 f(C_{i_\ell})\right] \\ = \sum_{\ell=1}^{2}\sum_{j=1}^{J}\Bigg[\tau_j^2\Bigg\{(1-r_j)^2\Bigg(\frac{1}{C_{ji}} - \frac{1}{\sum_h C_{jh}}\Bigg) + \Bigg(\frac{1}{C_{ji_\ell}} - \frac{1}{C_{ji}}\Bigg)\Bigg\} \\ + \sigma_j^2\Bigg\{\frac{1-r_j^2}{\sum_h C_{jh}} + \frac{r_j^2}{C_{ji}}\Bigg\}\Bigg] C_{ji_\ell}^2 f(C_{i_\ell}).\end{aligned} \tag{3.30}$$

It is easy to see that under precisely the same conditions (3.22), (3.24), (3.26), the same sequence of inequalities (3.28) holds; interpret est-risk$_i$ in (3.28) as being equal to (3.30) with $r_j = D_j$ for "est" = "ueb," with $r_j = D_j^{1/2}$ for "est" = "ceb," and with $r_j = 0$ for "est" = "syn". Moreover for the loss function (2.15) with $f(C_i) = 1/C_i$, risk gaps widen as lower levels of aggregation are attained.

4. DISCUSSION

Various assumptions are made in deriving the risk inequalities (3.28), all of which deserve further investigation. The model (2.7) and (2.10) is assumed to fit, and in particular the independence of distributions between subareas is assumed. Moreover, the effect of estimating D_j in the empirical Bayes estimators of F_{ji} is assumed negligible. Notice however that synthetically estimated F_{ji}'s do not use an estimate of D_j and so those risk inequalities only rely on the appropriateness of the model (2.7), (2.10).

The conditions which order the various risks and bound them below the census risk in (3.28), all depend on σ_j^2/τ_j^2 being "small." The practical implication is that a large number of households need to be chosen in the post-enumeration survey (PES) or there can be no guarantee that census counts can be improved by adjustment. With prior knowledge of stratum variation (*e.g.*, from a previous census), the PES could be *designed* so that the conditions are satisfied.

After the survey has been conducted and the data $\{X_{ji}: i = 1, \ldots, I; j = 1, \ldots, J\}$ are available, the various conditions (3.22), (3.24), and (3.26) can all be checked by using the estimators $\hat{\tau}_j^2$ and $\hat{\sigma}_j^2$ given by (3.2).

Concentrate on the best convex combination of X_{ji} and $X_{j\cdot}$, namely F_{ji}^{ueb} given by (3.3). Then, ueb-risk$_i$ ≤ cen-risk$_i$, if (3.22) holds; *i.e.*, if

$$\sigma_j^2/\tau_j^2 \leq \left\{\sum_h C_{jh}/C_{ji}\right\}^{1/2}; j = 1, \ldots, J. \tag{4.1}$$

Notice that the condition is less stringent when the i-th area has a small census population; conversely, areas of large census population may have a ueb-based estimated population further from the truth than census. A sufficient condition for (4.1) to hold is, $\sigma_j^2/\tau_j^2 \leq 1$; $j = 1, \ldots, J$, which is also the condition that guarantees the syn-based estimated population improves over census. This condition was satisfied for the 1980 PEP 3-8 data (see Section 3.2).

Finally, the condition (4.1) becomes less stringent at lower levels, and indeed the results of Section 3.2 show that the risk gap between the adjusted population and the census population widens. This deserves comment. The results are true provided the model holds at lower levels, but this is probably not the case at the block and the enumeration-district level. Presence of bias in (2.7) and (2.10); namely

$$E(F_{ji}) = F_j + b_{ji}; \; E(X_{ji} \mid F_{ji}) = F_{ji} + d_{ji}, \tag{4.2}$$

could cause a reversal in some of the risk inequalities. At the state level however, Table 1 and Cressie (1988) show through an examination of residuals, that (2.7) and (2.10) does fit for the 1980 PEP 3-8 data. And since (3.29) implies that condition (4.1) is satisfied, one can be confident that ueb-based adjusted state totals are closer to the truth than census state totals. That may not be true at the block level; clearly a decision regarding the level at which it is most important to have accurate census counts, needs to be made. The first use of U.S. Census data is the reporting of *state* totals to Congress for the purpose of redistricting House seats. One might include a number of large cities in with the states, and create *e.g.*, the "states" New York City, and New York State Except New York City. It seems to me that this "state" level is the most sensitive politically and that accurate totals at this level should receive the highest priority.

ACKNOWLEDGEMENT

This research has benefited from the input of members of the Undercount Research Staff at the U.S. Bureau of the Census, and from the perceptive comments of J.B. Kadane and J.W. Tukey. Support from Joint Statistical Agreements JSA 86-5, JSA 87-4, JSA 87-10, and JSA 88-13, between Iowa State University and the Census Bureau is gratefully acknowledged. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

COWAN, C.D., and BETTIN, P.J. (1982). Estimates and missing data problems in the post enumeration survey. Internal Report. Statistical Methods Division, Bureau of the Census, Washington, D.C.

CRESSIE, N. (1986). Empirical Bayes estimation of undercount in the decennial census. *Statistical Laboratory Preprint 86-58*, Iowa State University, Ames, IA.

CRESSIE, N. (1987a). Empirical Bayes estimation of undercount in the decennial census. Manuscript submitted to *Journal of the American Statistical Association.*

CRESSIE, N. (1987b). Comment on "Census undercount adjustment and the quality of geographic population distributions," by A.L. Schirm and S.H. Preston. *Journal of the American Statistical Association, 82*, 980-983.

CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. *Proceedings of Bureau of the Census Fourth Annual Resarch Conference.* Bureau of the Census, Washington, D.C., 123-150.

CRESSIE, N., and DAJANI, A. (1988). Empirical Bayes estimation of U.S. undercount based on artificial populations. *Statistical Laboratory Preprint 88-17*, Iowa State University, Ames, IA.

DEMPSTER, A.P., and TOMBERLIN, T.J. (1980). The analysis of census undercount from a post-enumeration survey, in *Proceedings of the 1980 Conference on Census Undercount.* Bureau of the Census, Washington, D.C. 88-94.

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*, 98-131.

ERICKSEN, E.P., and KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census, in *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 23-45.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1987). Adjusting the 1980 census of housing and population. *Technical Report No. 401*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.

FAY, R.E. III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein to census data. *Journal of the American Statistical Association, 74*, 269-277.

FERGUSON, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census. *Statistical Science, 1*, 3-39.

GOLDSTEIN, M. (1975). Approximate Bayes solutions to some nonparametric problems. *Annals of Statistics, 3*, 512-517.

HENDERSON, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32, 69-83.

HUI, S.L., and BERGER, J.O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association, 78*, 753-760.

ISAKI, C.T., DIFFENDAL, G.J., and SCHULTZ, L.K. (1986). Statistical synthetic estimates of undercount for small areas. *Proceedings of Bureau of the Census Second Annual Research Conference.* Bureau of the Census, Washington, D.C., 557-569.

KADANE, J.B. (1984). Allocating Congressional seats among the states when state populations are uncertain. *Technical Report No. 309*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.

LINDLEY, D.V., and SMITH, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34*, 1-41.

LOUIS, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association, 79*, 393-398.

MORRIS, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association, 78*, 47-55.

MULRY-LIGGAN, M., and HOGAN, H. (1986). Research plan on census adjustment standards. *Proceedings of Bureau of the Census Second Annual Research Conference*. Bureau of the Census, Washington, D.C., 381-392.

NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, (Eds. C.F. Citro and M.L. Cohen.) Washington: National Academy Press.

READ, T.R.C., and CRESSIE, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

SCHULTZ, L.K., HUANG, E.T., DIFFENDAL, G.J., and ISAKI, C.T. (1986). Some effects of statistical synthetic estimation on census undercount of small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 321-325.

STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation, in *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 124-137.

TUKEY, J.W. (1981). Discussion of "Issues in adjusting for the 1980 census undercount," by Barbara Bailar and Nathan Keyfitz, presented at the Annual Meeting of the American Statistical Association, Detroit, MI.