SUB PROVINCIAL ESTIMATATION OF UNDERCOVERAGE IN THE 1991 CANADIAN CENSUS

PETER DICK AND DON ROYCE

SUB-PROVINCIALESTIMATIONOF UNDERCOVERAGE IN THE 1991 CANADIANCENSUS

Peter Dick and Don Royce, Statistics Canada
Ottawa, Ontario, Canada, K1A 0T6

Abstract

Following each Canadian Census, an evaluation study known as the Reverse Record Check is carried out to estimate the level of undercoverage. The sample size of this study is such that reliable estimates of undercoverage can be produced for each province, and for certain age-sex groups at the national level, but not for age-sex groups at the province level. The use of synthetic estimation techniques has been investigated for this latter purpose, but these have the disadvantage that they are based on inherently unverifiable assumptions. Cressie (1989) proposed a compromise model that combines the direct survey estimate and a synthetic estimate. We generalize Cressie's model to admit the possibility of bias in the survey estimates, and include some results on possible gains to be made by employing this model. These results are illustrated with data from the 1986 Reverse Record Check.

1. Introduction

The Census of Canada is conducted every five years, the most recent having been on June 4th 1991. Census counts serve a variety of uses, such as the allocation of seats in the federal Parliament, the transfer of money between various levels of government, and the planning of essential services such as health, education, and local transportation.

For the period between censuses, Statistics Canada also produces a series of population estimates which are used for many of the same purposes. These population estimates are obtained by adding births and in-migrants to the most recent census counts, and subtracting deaths and out-migrants. When new census results become available, the estimates for the past five years are revised to bring them in line with the new census counts.

Until 1986, this methodology for population estimates was, by and large, acceptable to most users. However the 1986 Census saw a substantial increase in undercoverage. At the national level, estimated undercoverage rose from the 2% level experienced in 1971, 1976 and 1981 to over 3%. There was also considerable variation in undercoverage among provinces, among age and sex groups, and among various other sub-groups of the population.

The unprecedented levels of undercoverage caused considerable disruption in the population estimates program and in several other programs which depend on population estimates. As a result, it was decided in early 1989 to investigate the possibility of changing the methodology of the population estimates program to include an allowance for census undercoverage. It should be noted that the published 1991 Census data themselves will not be adjusted for undercoverage; only the population estimates based on census counts would be affected. From a technical viewpoint, however, the issue is similar to the question of census adjustment which has been the subject of much debate in the United States.

One of the key questions in deciding on adjustment is whether, and if so how, adjustments made at higher levels of aggregation (e.g., provinces) should be "carried down" to lower levels of detail. The coverage studies can only supply reliable estimates of undercoverage at relatively aggregated levels, but adjusting at some levels but not others would cause severe problems for data users. An important part of the research, therefore, was an investigation of estimation techniques for small domains that would maintain the overall consistency of the estimates program.

One such technique is that of synthetic estimation. However, synthetic estimation tends to treat all small areas alike. Areas with reliable estimates of undercoverage are treated the same as areas with unreliable, or no, estimates of undercoverage. In an effort to combine the best features of synthetic estimation and direct survey estimates of undercoverage, Cressie in a series of papers ((1988a), (1988b),(1989)) examined an Empirical Bayes methodology and illustrated it with results from the 1980 U.S. Census Post-Enumeration Program (PEP).

In this paper we examine Cressie's model for undercoverage and analyze the sensitivity of its results to changes in some of the model's key assumptions. Section 2 describes Cressie's model and reviews his basic findings concerning the risk functions for the Census, synthetic, and Bayes estimators of a population total. Section 3 then varies Cressie's model by: (i) investigating the effect on the risks of having to estimate the synthetic adjustment factors, rather than assuming that they are known, (ii) allowing the possibility that the survey estimates of undercoverage may be biased, and (iii) considering the effect of having to estimate the variance components used in the Empirical Bayes estimator. In Section 4 we illustrate the effects of these variations in the model with results from the 1986 Canadian Reverse Record Check. Section 5 summarizes our conclusions and indicates directions for future research work.

2. Cressie's Model of Undercount

Cressie (1989) takes as his basic starting point that the population of interest has been stratified such that undercounting is homogeneous within each stratum. In formulating the model, the following definitions will prove to be useful. Let:

Y_{ji} be the true population count for the j-th stratum (j = 1, 2, ... J) in the i-th area (i = 1, 2, ...I); and

C_{ji} be the corresponding observed Census count for the j-th stratum in the i-th area (assumed to be always non-zero).

The ratio of the true population count to the observed Census count for the j-th stratum and the i-th area is defined as

$$F_{ji} = \frac{Y_{ji}}{C_{ji}} . \qquad (1)$$

The net number of persons missed by the Census in the j-th stratum and the i-th area is defined as

$$M_{ji} = Y_{ji} - C_{ji} . \qquad (2)$$

If all the F_{ji} are known completely then it is easy to see that for any area i the true population can be written as

$$Y_i = \sum_{j=1} F_{ji} C_{ji} \qquad (3)$$

By defining the adjustment factor in this fashion, consistency over any set of areas has been achieved. This can be seen be summing over any two areas and seeing that the higher-level adjustment factor can be written as

$$F_{ji\&i'} = \frac{F_{ji}C_{ji} + F_{ji'}C_{ji'}}{C_{ji} + C_{ji'}} \qquad (4)$$

This shows that the adjustment factor at any higher level of aggregation is merely the weighted average of the lower level adjustment factors.

If the adjustment factors F_{ji} are unknown, then some assumptions must be made. The simplest assumption that can be made is that the stratification has been carried out perfectly so that within each stratum the adjustment factor for any stratum group j is the same across all the areas i. This would permit the population for any area to be determined by the basic synthetic relationship:

$$Y_i = \sum_j F_j C_{ji} \qquad (5)$$

This reduces the number of parameters in the model from (I x J) to J. However the assumption of perfect stratification seems strong.

Cressie relaxes this assumption by allowing the adjustment factor for any area i and stratum j to come from a distribution with an expected value which depends only on the stratum but with a variance specific to the stratum and the area. This can be written as

$$F_{ji} \sim N (F_j , \tau_{ji}^2) \qquad (6)$$

The distribution has been assumed to be Normal but this is not strictly necessary. As long as the expected value and the variance remain as stated then the results of this section are not compromised. However, this assumption has not reduced the number of parameters in the model since the variance term depends on the area and the stratum. Cressie argues, however, by both a Bayesian and a frequentist argument, that the variance can be written as

$$\tau_{ji}^2 = \frac{\tau_j^2}{C_{ji}} \qquad (7)$$

provided that the Census count C_{ji} is large. This reduces the number of parameters from I x J to 2J.

A further level of randomization occurs in this model because the adjustment factors F_{ji} are not directly observed but have to be estimated using direct estimates X_{ji} from a survey. Cressie states this dependence through the following model:

$$X_{ji} | F_{ji} \sim N (F_{ji} , \sigma_{ji}^2) \qquad (8)$$

where the X_{ji} represents the direct survey estimate of the adjustment factor and σ_{ji}^2 represents the sampling variance. Cressie simplifies this model further by noting that the sample design of the 1980 US Post Enumeration Survey can be assumed to be probability proportional to size within strata. This permits the sampling variance to be modelled as

$$\sigma_{ji}^2 = \frac{\sigma_j^2}{C_{ji}} \qquad (9)$$

In summary then, there are two stages to the Cressie model. The first stage states that the true adjustment factors for each stratum j and area i equals the stratum level adjustment factor plus a random error. The second stage states that the true adjustment factor in the j-th stratum and the i-th area is unbiasedly estimated by the direct survey estimate.

Assuming the above formulation, Cressie uses the results of Lindley and Smith (1972) to show that the posterior distribution of F_{ji}, the true adjustment factor in the j-th stratum and the i-th area, given that the direct survey estimate X_{ji} has been observed, is:

$$F_{ji} \mid X_{ji} \sim N(F_j + \omega_j(X_{ji} - F_j), \; \sigma_j^2 \frac{\omega_j}{C_{ji}}) \qquad (10)$$

where

$$\omega_j = \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2} \qquad (11)$$

If the F_j and the variance components τ_j^2 and σ_j^2 are known, and if a squared error loss function is used, the mean of the above distribution provides the Bayesian estimate of F_{ji}.

Cressie then compares the risks of using different estimators of Y_i using the following loss function:

$$L(\hat{Y}_i) = \frac{(\hat{Y}_i^{(e)} - Y_i)^2}{C_i} \qquad (12)$$

where the population of any area i is estimated by

$$\hat{Y}_i^{(e)} = \sum_j \hat{F}_{ji}^{(e)} C_{ji} \qquad (13)$$

and the superscript (e) represents the estimator that is to be used.

Cressie determines the risks, or expected loss, for three estimators of the true population count Y_i: the actual Census count, a synthetic estimator and the Bayesian estimator. Throughout this development it is assumed that the parameter values are known and the expectation is taken over the model for F_{ji} and X_{ji}.

The three estimators (see footnote 1) for the population of area i can be written as

$$\mathit{Census} \;:\; \hat{Y}_i^{(c)} = \sum_j C_{ji} \;; \qquad (14)$$

$$\mathit{Synthetic} \;:\; \hat{Y}_i^{(s)} = \sum_j F_j C_{ji} \;; \qquad (15)$$

$$\mathit{Bayes} \;:\; \hat{Y}_i^{(b)} = \sum_j (F_j + \omega_j (X_{ji} - F_j)) C_{ji}$$

$$(16)$$

The risks for each estimator can be written as

Census:

$$E \frac{(\hat{Y}_i^{(c)} - Y_i)^2}{C_i} = \sum_j \tau_j^2 \frac{C_{ji}}{C_i} \qquad (17)$$
$$+ \sum_j (1 - F_j)^2 \frac{C_{ji}^2}{C_i} \;;$$

Synthetic:

$$E \frac{(\hat{Y}_i^{(s)} - Y_i)^2}{C_i} = \sum_j \tau_j^2 \frac{C_{ji}}{C_i} \;; \qquad (18)$$

Bayes:

$$E \frac{(\hat{Y}_i^{(b)} - Y_i)^2}{C_i} = \sum_j \frac{C_{ji}}{C_i} \tau_j^2 (1 - \omega_j)$$

$$(19)$$

Clearly the following inequalities concerning the risks are self evident from this development the risks can be ordered as:

$$\mathit{Bayes} \;\le\; \mathit{Synthetic} \;\le\; \mathit{Census} \qquad (20)$$

The implication of these inequalities is that the Bayes estimator will always have a lower risk, or at worst a risk equal to that of the actual Census count.

3. Modifications to Cressie's Model

As is evident from the previous section there are many assumptions that have been made in the development of the Bayesian model. This section examines Cressie's basic model when three of his assumptions are relaxed. First, we examine the impact on the risks when estimates of the F_j are used instead of assuming them known (this was described in Cressie (1988b) but is repeated here for the sake of comparison to the risks presented in Section 2). Next, the assumption of unbiased direct survey estimates will be replaced by using survey estimates that are possibly biased. Finally, using results of Prasad and Rao (1990), we examine the effects on the risk of using estimates of the variance components in place of the true, but unknown, variance components.

3.1 Risks Associated with Estimating the F_js

We examine the effect on the risks of the three estimators of population when estimates of the stratum level adjustment factors, the F_js, must be used. It is assumed throughout this section that both the model variance and the sampling variance are known.

Since the Census estimator does not involve any of the model parameters, the risk of using the Census counts will not change and will remain as in (17).

The risk of using the Synthetic model will change because Y_i is now estimated by

$$\hat{Y}_i^{(s)} = \sum_j \hat{F}_j C_{ji} \quad \text{where} \quad \hat{F}_j = \sum_i \frac{C_{ji} X_{ji}}{C_j} \qquad (21)$$

Using the model developed earlier ((6) and (8)), the risk of using an estimate of the Synthetic estimator can be shown to be

$$E\,\frac{(\hat{Y}_i^{(s)} - Y_i)^2}{C_i} = \sum_j \frac{C_{ji}}{C_i}\,\tau_j^2 \qquad (22)$$
$$+ \sum_j \frac{C_{ji}^2}{C_i}\,(\frac{\sigma_j^2}{C_j} - \frac{\tau_j^2}{C_j})\ .$$

This risk is simply the risk of using the original Synthetic model with an added term representing the risk from estimating the F_js. From (22) it can be seen that when the sampling variance and model variance are equal in each stratum the added risk due to estimating the F_js is zero. However, if the sampling variance is larger than the model variance then the risk of using the Synthetic model with estimated F_js will grow. Note also that it is possible to have the risk decline even when estimating the F_j. This occurs when the sampling variance is smaller than the model variance.

Under the model assumptions in (6) and (8), the risk of using the Bayes estimator can be shown to be

$$E\,\frac{(\hat{Y}_i^{(b)} - Y_i)^2}{C_i} = \sum_j \frac{C_{ji}}{C_i}\,\tau_j^2\,(1 - \omega_j)$$
$$+ \sum_j \frac{C_{ji}^2}{C_i\,C_j}\,\sigma_j^2\,(1 - \omega_j)\ . \qquad (23)$$

This is just the risk of using the Bayes estimate when the F_js are assumed known plus an additional term due to the estimation of the F_js. Note that this added term will always add to the risk of using this estimator, unlike the synthetic case where the risk can actually decline with estimation of the F_j. Cressie (1988b) showed that this risk (23) was always less than or equal to the risk of the synthetic estimator (22), and also gave a sufficient condition for the risk of the synthetic estimator (22) to be less than that of of the Census (17).

3.2 Bias in the Direct Survey Estimates

In this sub-section we modify Cressie's model by now permitting the possibility of biased estimates, X_{ji}, from the survey. We can write this modification to Cressie's original model (8) for the survey estimates as

$$X_{ji}|F_{ji} \sim N(F_{ji} + \alpha_{ji}\ ;\ \frac{\sigma_j^2}{C_{ji}}) \qquad (24)$$

where the alpha component represents a bias term that is present in all areas and strata. Another interpretation of this term can be seen by writing the expected value of X_{ji} as

$$E\frac{(\hat{M}_{ji} + C_{ji})}{C_{ji}} = \frac{M_{ji} + C_{ji}}{C_{ji}} + \alpha_{ji} \qquad (25)$$

so that

$$\alpha_{ji} = \frac{E\,(\hat{M}_{ji}) - M_{ji}}{C_{ji}}\ . \qquad (26)$$

That is, alpha represents the bias in the estimate of missed divided by the actual Census count.

Bias in the estimation of undercoverage may arise from many sources. Hogan and Wolter (1988) describe the major sources of error in the U.S. Post Enumeration Survey (PES), while Burgess (1988) describes similar issues for the Canadian Reverse Record Check (RRC). Among these are non-response, matching errors, correlation bias (in the case of the PES) and, in the case of the RRC, the fact that the RRC measures gross, not net, undercoverage. Thus, it would seem prudent, in any assessment of risks, to consider the possibility that the survey estimates of undercoverage are biased.

Considering the Synthetic estimate first, using the model of the bias developed above (24) (and using (6)) it can be shown that the risk of using the Synthetic estimator when the direct survey estimate X_{ji} is subject to bias can be written as

$$E\,\frac{(\hat{Y}_i^{(s_b)} - Y_i)^2}{C_i} = \sum_j \frac{C_{ji}}{C_i}\,\tau_j^2$$
$$+ \sum_j \frac{C_{ji}^2}{C_i}\,\{(\frac{\sigma_j^2}{C_j} - \frac{\tau_j^2}{C_j}) + (\frac{Bias\,(\hat{M}_j)}{C_j})^2\} \qquad (27)$$

This is same as the Synthetic risk (21) developed earlier with the last term being an added bias term.

The Empirical Bayes model with bias can be developed similarly. It can be shown that the risk of using the Bayes estimate for any area i when the direct survey estimate is subject to bias is

$$E\frac{(\hat{Y}_i^{(b_b)} - Y_i)^2}{C_i} = \sum_j \frac{C_{ji}}{C_i}\,\tau_j^2\,(1-\omega_j)$$
$$+ \sum_j \frac{C_{ji}^2}{C_iC_j}\,\sigma_j^2\,(1-\omega_j) + \sum_j \frac{C_{ji}^2}{C_i}\,(\frac{Bias(\hat{M}_j)}{C_j})^2 \qquad (28)$$

This is the same risk as developed in (22) but with a term added to reflect the additional risk due to using biased estimates from the survey. Note that the added bias term for the risk of the Empirical Bayes estimate is the same as that for the Synthetic estimate. Hence the Empirical Bayes risk will still be less than equal to the Synthetic risk.

3.3 Estimation of the Variance Components in the Bayesian Model

In previous sections, the estimator developed under the Bayesian framework has assumed that the variance components are known. In practice, however, they will not be known and an Empirical Bayes estimator would be used. The Empirical Bayes estimator is actually developed in two stages: first, the Best Linear Unbiased Predictor is obtained assuming the variance components are known, and then the variance components are replaced by estimates of the variance components. However, Prasad and Rao (1990) have noted that ignoring the uncertainty in the variance components and then using the standard Mean Square Error (MSE) calculation of the best linear unbiased predictor of F_{ji} as an approximation to the corresponding MSE of the two-stage estimator can lead to serious understatement of the MSE.

To estimate the MSE of the two-stage estimator, Prasad and Rao quote a result from Kacker and Harville (1984) that states that

$$MSE\ (\ \hat{Y}_i^{(b_2)}\ ) = MSE\ (\ \hat{Y}_i^{(b)}\ ) \qquad (29)$$
$$+\ E\ (\ \hat{Y}_i^{(b_2)} - \hat{Y}_i^{(b)}\ )^2$$

where

$$\hat{Y}_i^{(b_2)} = \sum_j C_{ji}\ (\ \hat{F}_j + \hat{\omega}_j\ (\ X_{ji} - \hat{F}_j\ )\ ) \qquad (30)$$
$$where\ \ \hat{\omega}_j = \frac{\hat{\tau}_j^2}{\hat{\tau}_j^2 + \sigma_j^2}\ .$$

The similarity to (16) is obvious: all parameters in (16) are now replaced with their estimates. Note we assume that the estimate of the sampling variance is not subject to sampling error. The relationship between the MSE(Y_i) and the loss function (12) that Cressie used can be seen to be:

$$\frac{MSE\ (\ \hat{Y}_i\ )}{C_i} = \frac{E\ (\ \hat{Y}_i - Y_i\ )^2}{C_i} \qquad (31)$$

Hence apart from the multiplicative constant ($1/C_i$), Prasad and Rao's results apply directly to the risk as defined by Cressie.

Prasad and Rao show that a second order approximation of (29) can be written (suitably modified to correspond to the original loss function (12)) in two parts. The first component is just the MSE($\hat{Y}_i^{(b)}$) and was shown for the Bayes model in (28) (for the case where the survey estimate is subject to bias). The second part of (29) can be shown to approximately equal to:

$$E\ (\hat{Y}_i^{(b_2)} - \hat{Y}_i^{(b)})^2 =$$
$$\sum_j C_{ji}\{(\frac{\sigma_j^2}{\sigma_j^2 + \tau_j^2})^2\ \frac{Var\ (\ \hat{\tau}_j^2\ )}{\sigma_j^2 + \tau_j^2}\} \qquad (32)$$

where, approximately,

$$Var\ (\ \hat{\tau}_j^2\ ) = \frac{2}{I}(\tau_j^2 + \sigma_j^2)^2 \qquad (33)$$

assuming normality of the model and sampling errors.

Prasad and Rao note that if the model and sample errors have been generated from a normal distribution this second order approximation is satisfactory. They also note that by ignoring the final term in (29) that the MSE calculation of the estimate can be understated by up to 20% depending on the assumed error distribution. The same conclusions can be applied to the risks that were developed for the Bayes model.

The total risk for the Empirical Bayes estimate including the risk due to estimation of the variance components can now be summarized by:

$$E\ \frac{(\ \hat{Y}_i^{(\omega_2)} - Y_i\ )^2}{C_i} = \sum_j \frac{C_{ji}}{C_i}\ \tau_j^2\ (\ 1 - \omega_j\ )$$
$$+ \sum_j \frac{C_{ji}^2}{C_i C_j}\ \{\sigma_j^2\ (1-\omega_j)$$
$$+ \frac{Bias\ (\hat{M}_j\ )^2}{C_j}\} + \frac{2}{I}\sum_j \frac{C_{ji}}{C_i}\ \sigma_j^2\ (1-\omega_j)$$
$$(34)$$

4. Empirical Results

4.1. Reverse Record Check Results from 1986

We first apply the methods of Section 2 to the 1986 Reverse Record Check. The objective of the Reverse Record Check is to provide estimates of the number of persons and households missed in the Canadian Census. It is described more fully in Burgess (1988), but briefly the approach is as follows. In 1986, some 36,000 persons were selected for the study from the following four frames:

- persons enumerated in the 1981 Census of Canada;
- persons missed in the 1981 Census of Canada (available in the form of a sample of persons so classified in the 1981 Reverse Record Check);
- a birth frame containing all births in Canada

between the 1981 Census and the 1986 Census;

- an immigrant frame containing a list of immigrants to Canada between the two censuses.

Each person in the sample was then traced to their 1986 Census Day address and the 1986 Census questionnaire for that address was checked to determine if the person had been enumerated or not.

The sample size and the sample design are sufficient to provide reliable estimates at the province level and for some age-sex combinations at the national level. In 1986, for example, for individual provinces the resulting coefficient of variations varied from under 6% for Ontario to over 37% for Prince Edward Island. However estimates at the level of province by age group and sex, which would be required for any adjustment of the population estimates program, are very often not reliable. In 1991, the Reverse Record Check sample has been increased to approximately 50,000 persons but many of the estimates by province-age-sex will still have unacceptably large CVs.

The unknown parameters that have to be estimated are the F_js, the model variance τ_j^2 and the sampling variance σ_j^2. Following the development in Maritz and Lwin (Section 2.8, 1989) we use the method of moments to estimate the F_j with

$$\hat{F}_j = \frac{\sum_i X_{ji} C_{ji}}{\sum_i C_{ji}} \qquad (35)$$

where X_{ji} is the direct survey estimate from the Reverse Record Check of the adjustment factor in the j-th stratum and the i-th area.

Using our notation, the estimate for the model variance can be written as:

$$\hat{\tau}_j^2 = \max\left[\frac{\sum_i C_{ji}(X_{ji} - \hat{F}_j)^2}{I-1} - \hat{\sigma}_j^2, 0\right] \qquad (36)$$

This is the same approach used by Cressie and also described in Prasad and Rao.

The sampling variance must be estimated directly from the survey using sampling considerations. We consider two possible approaches.

The first method is simply to use the direct estimate of the sampling variance that is produced by the Reverse Record Check for each age-sex stratum. The sampling variance estimates from the published table can be determined by taking

$$\hat{\sigma}_{j(d)}^2 = \frac{Var(\hat{M}_j)}{C_j}. \qquad (37)$$

The second approach is to use a Generalized Variance Function as suggested by Wolter (1985). To estimate this we first set

$$\log\left(\frac{Var(\hat{M}_j)}{M_j^2}\right) = \alpha + \beta M_j + \gamma_j \qquad (38)$$

and use least squares to estimate the unknown parameters in (48). We then use the estimating equation

$$\hat{\sigma}_{j(e)}^2 = \frac{\hat{M}_j^2}{C_j} \exp(\hat{\alpha} + \hat{\beta} \hat{M}_j) \qquad (39)$$

to generate estimates of the sampling variance for the number of missed persons (M_j) in each stratum.

Using either of these estimates (37) or (39) and substituting the result directly into (36), it was found that there were a large number of zero estimates for the model variance. Cressie, in this situation, suggests collapsing the strata. Doing so resulted in 4 strata in each case. The strata differed slightly in the details of the collapsing (see Tables 1 and 2).

Table 1 gives the estimated variance components when the sampling variance is estimated directly from the 1986 Reverse Record Check; Table 2 displays the estimated variance components when the sampling variance has been smoothed.

Using the estimates from Table 1 and Table 2, the estimated adjustment factors for the Bayesian model were calculated. The results of the methods of Section 2 are presented in Table 3. The first estimate in each cell gives the direct survey estimate of the adjustment factor calculated from the Reverse Record Check. Note the entries that have an adjustment factor of 1, such as Alberta males 65 and over, mean that the Reverse Record Check estimated no missed persons in this cell. The second row within each cell gives the Synthetic estimate that is determined solely by the total for the age-sex national estimate.

The third row and fourth rows in each cell give the Empirical Bayes estimates. The third row was calculated using variance components based on the direct estimates of the sampling variance from the Reverse Record Check, as displayed in Table 1. The fourth row was based on the smoothed estimates of the sampling variances as given in Table 2.

The extremes of the adjustment factor for each cell entry are always the direct survey estimate and the Synthetic estimate. The two Bayesian estimates represent compromises between these extremes. For cells with very small samples, as is the case in Prince Edward Island for Males 45 - 54, the Bayesian approach smooths the adjustment factor back almost to the

Synthetic estimate. In the larger provinces of Ontario and Quebec the estimates for both the direct survey and the Synthetic estimate are usually very close. Hence the Empirical Bayes estimates does not affect either estimate greatly.

4.2 Evaluation of Risks

4.2.1 Estimated Risks of each Procedure in the 1986 Reverse Record Check

The risk of using each procedure to estimate the true adjustment factor for each area I can be evaluated for the 1986 Census. Recall the final risks for each of Census (17), Synthetic (27) and Bayes (34).

Substituting the estimates of the variance components and the stratum level adjustment factors F_j into the above, we can estimate the components and the total risk for the three estimators (see footnote 2). These are displayed in Table 4. The sequence of the terms in Table 4 for the risk of using the Synthetic estimate and the Bayes estimate are ordered to correspond to the terms in (27) and (34). The bias component was estimated by assuming an overall relative bias of approximately 5%.

The first point to notice about Table 4 is that for every area the risk of the Census count is always considerably higher than the risk of either the synthetic estimate or the Empirical Bayes estimate. Since the "model" component of the synthetic estimator is equal to the first component of the Census risk (see equation (17)), it can be seen that almost all of the Census risk arises from the second term in (17). Second, unlike the results of Sections 2, 3.1 and 3.2, the risk of the Empirical Bayes estimator is actually **higher** than that of the synthetic estimator. The reason is because of the additional term in the risk representing the effect of estimating the variance components. Without this latter component, the risk of the Empirical Bayes estimate would have been lower than that of the synthetic estimate. The effect of the estimation of the variance components on the total risk of the Empirical Bayes estimate is substantial.

The figures in Table 4 assumed a relatively small amount of relative bias (5%) in the estimation of the number of persons missed. To examine the potential impact of higher levels of bias on the total risk, the bias component of Prince Edward Island was re-written as

$$Bias\ (\hat{M}_j) = \gamma\ M_j \qquad (54)$$

and then gamma was allowed to vary from 0, representing an unbiased estimate of missed persons from the survey, to 1. However, only when gamma approaches 1 does the risk from either the Synthetic or the Bayes estimate approach the risk for the Census. Thus, the impact of biased estimates of missed will only impact on the ordering of the relative risks in extreme situations.

5. Conclusions and Future Research

Cressie in his papers demonstrated that the risk of using the Bayes estimate was less than the risk of using the Synthetic estimate which in turn was always less than using the actual Census count, for both the usual Bayes and Empirical Bayes methods. When the effect of the estimation of variance components are considered, however, it appears that these relationships may no longer hold. Although the risks for both the synthetic and Empirical Bayes were always found to be less than the Census counts, taking the effect of the estimation of the variance components into account can result in a situation where the synthetic estimator has a lower risk than the Empirical Bayes estimator.

In the future, we hope to identify specific algebraic conditions under which the risk for the Empirical Bayes estimator, including the component due to the estimation of the variance components, is less than or equal to the risk for the synthetic estimator. This would represent an extension of conditions given in Cressie(1988b). In the numerical example given, the estimated sampling variances were much larger than the model variances. Since the additional term in the risk is a function of the sampling variance, it could be that lower sampling variances would lead to a situation where the risk of the Empirical Bayes estimator, even allowing for the effect of estimating the variance components, would still be lower than the synthetic estimator.

We will also investigate other models for the adjustment factors. The model that Cressie proposed (6) results in consistency with the national age - sex adjustment factors but not with the provincial level estimates. To create consistency on both margins, other models will be investigated. A recent paper by Barry (1990) describes an Empirical Bayes approach, using a logit model, to the estimation of binomial probabilities (e.g., undercoverage rates) in two-way tables that preserves both row and column margins. Empirical Bayes methods that combine direct survey estimates with Iterative Proportional Fitting estimates will also be investigated.

6. References

Barry, D. (1990) Empirical Bayes estimation of binomial probabilities in one-way and two-way layouts. **The Statistician 39, pp. 437-453.**

Burgess, R.D. (1988) Evaluation of Reverse Record Check estimates of undercoverage in the Canadian Census of Population. **Survey Methodology 14: 137-156**

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

Cressie, N. (1988b) When are census counts improved by adjustment? **Survey Methodology 14: 191-208.**

Cressie, N. (1989) Empirical Bayes estimation of the undercount in the decennial census. **Journal of the American Statistical Association 84: 1033-1044.**

Hogan, H.R. and Wolter, K.M. (1988) Measuring accuracy in a post-enumeration survey. **Survey Methodology 14: 99-116.**

Kackar, R.N. and Harville, D.A. (1984) Approximation for standard errors of estimators of fixed and random effects in mixed linear models. **Journal of the American Statistical Association 79: 853-862.**

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

Maritz, J.S. and Lwin, T. (1989) Empirical Bayes Methods, 2nd edition. **London: Chapman and Hall.**

Prasad, N.G.N. and Rao, J.N.K. (1990) The estimation of the mean squared error of small-area estimators. **Journal of the American Statistical Association 85: 163-171.**

Wolter, K.M. (1985) Introduction to Variance Estimation. **New York: Springer-Verlag**

Footnotes

[1] Cressie also considered a constrained Bayes estimator, however we do not deal with it in this paper.

[2] In fact, Prasad and Rao show that an unbiased estimator, to $o(I^{-1})$, of the first component of the Bayes risk in (34) is equal to the sum of the first and last terms in (34) with estimates of the variance components substituted thus the figures in Table 4 for the Bayes risk are underestimates.

Table 1

Variance Components Estimated Directly from Survey

Stratum	$\hat{\sigma}_j^2$	$\hat{\tau}_j^2$	$\hat{\omega}_j$
Male 20 - 24	61.43	7.76	0.112
Male 15 - 19, 25 - 44	54.35	11.32	0.172
Male 0 -14, 45 plus	30.86	8.59	0.172
Female	37.88	13.99	0.270

Table 2

Variance Components Estimated from Smoothed Variances

Stratum	$\hat{\sigma}_j^2$	$\hat{\tau}_j^2$	$\hat{\omega}_j$
Male 15 - 24	73.58	9.04	0.109
Male 25 - 44	42.23	21.80	.340
Male 0 -14, 45 plus	29.48	9.97	0.253
Female	39.52	12.35	0.238

Table 3.
Estimated Adjustment Factors

MALE											FEMALE									
BC	ALTA	SASK	MAN	ONT	QUE	NB	NS	PEI	NFLD		BC	ALTA	SASK	MAN	ONT	QUE	NB	NS	PEI	NFLD
1.030	1.028	1.011	1.008	1.032	1.022	1.008	1.021	1.008	1.015	00-14	1.028	1.025	1.028	1.047	1.034	1.020	1.004	1.015	1.000	1.023
1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025		1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027
1.026	1.028	1.022	1.022	1.027	1.025	1.022	1.024	1.021	1.023		1.027	1.027	1.027	1.032	1.029	1.025	1.021	1.024	1.020	1.026
1.027	1.026	1.022	1.021	1.027	1.024	1.021	1.024	1.021	1.023		1.027	1.027	1.027	1.032	1.029	1.025	1.022	1.024	1.021	1.026
1.074	1.023	1.060	1.077	1.066	1.040	1.061	1.031	1.020	1.019	15-19	1.058	1.056	1.038	1.012	1.048	1.042	1.031	1.090	1.029	1.000
1.054	1.054	1.054	1.054	1.054	1.054	1.054	1.054	1.054	1.054		1.046	1.046	1.046	1.046	1.046	1.046	1.046	1.046	1.046	1.046
1.058	1.049	1.059	1.058	1.056	1.052	1.055	1.050	1.048	1.048		1.049	1.048	1.044	1.037	1.046	1.045	1.042	1.058	1.041	1.033
1.056	1.051	1.057	1.057	1.055	1.053	1.055	1.052	1.050	1.050		1.049	1.048	1.044	1.038	1.046	1.045	1.042	1.056	1.042	1.035
1.164	1.148	1.121	1.147	1.143	1.118	1.074	1.068	1.068	1.054	20-24	1.151	1.084	1.109	1.082	1.084	1.073	1.112	1.057	1.061	1.087
1.132	1.132	1.132	1.132	1.132	1.132	1.132	1.132	1.132	1.132		1.090	1.090	1.090	1.090	1.090	1.090	1.090	1.090	1.090	1.090
1.136	1.134	1.131	1.134	1.133	1.130	1.126	1.127	1.125	1.123		1.106	1.091	1.095	1.088	1.088	1.085	1.096	1.081	1.082	1.089
1.135	1.134	1.131	1.134	1.133	1.131	1.126	1.127	1.125	1.123		1.104	1.091	1.094	1.088	1.088	1.086	1.095	1.082	1.083	1.089
1.095	1.062	1.040	1.067	1.064	1.079	1.080	1.079	1.036	1.041	25-34	1.065	1.043	1.022	1.022	1.045	1.048	1.038	1.023	1.075	1.037
1.071	1.071	1.071	1.071	1.071	1.071	1.071	1.071	1.071	1.071		1.045	1.045	1.045	1.045	1.045	1.045	1.045	1.045	1.045	1.045
1.075	1.069	1.066	1.070	1.070	1.072	1.072	1.072	1.065	1.066		1.050	1.044	1.038	1.039	1.045	1.045	1.043	1.039	1.053	1.042
1.079	1.068	1.060	1.070	1.069	1.074	1.074	1.074	1.059	1.061		1.049	1.044	1.038	1.039	1.045	1.045	1.043	1.040	1.052	1.043
1.057	1.017	1.055	1.031	1.033	1.052	1.076	1.026	1.011	1.028	35-44	1.026	1.019	1.027	1.005	1.014	1.024	1.019	1.000	1.000	1.006
1.041	1.041	1.041	1.041	1.041	1.041	1.041	1.041	1.041	1.041		1.018	1.018	1.018	1.018	1.018	1.018	1.018	1.018	1.018	1.018
1.044	1.037	1.043	1.039	1.039	1.043	1.047	1.038	1.036	1.039		1.020	1.019	1.021	1.015	1.017	1.020	1.019	1.013	1.013	1.015
1.046	1.033	1.046	1.039	1.038	1.045	1.053	1.036	1.031	1.036		1.020	1.018	1.020	1.015	1.017	1.019	1.016	1.014	1.014	1.015
1.025	1.019	1.011	1.031	1.036	1.018	1.010	1.013	1.148	1.026	45-54	1.030	1.040	1.023	1.038	1.017	1.016	1.013	1.015	1.003	1.010
1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026		1.021	1.021	1.021	1.021	1.021	1.021	1.021	1.021	1.021	1.021
1.026	1.025	1.023	1.027	1.028	1.025	1.023	1.023	1.062	1.026		1.023	1.026	1.021	1.025	1.020	1.019	1.018	1.019	1.016	1.016
1.026	1.024	1.023	1.028	1.029	1.024	1.022	1.023	1.057	1.026		1.023	1.025	1.021	1.025	1.020	1.020	1.019	1.019	1.016	1.018
1.055	1.020	1.023	1.018	1.026	1.017	1.013	1.032	1.018	1.011	55-64	1.057	1.022	1.030	1.006	1.027	1.029	1.057	1.000	1.000	1.011
1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026	1.026		1.029	1.029	1.029	1.029	1.029	1.029	1.029	1.029	1.029	1.029
1.032	1.024	1.025	1.024	1.028	1.024	1.023	1.027	1.024	1.023		1.037	1.025	1.030	1.023	1.029	1.029	1.037	1.021	1.021	1.024
1.033	1.024	1.025	1.024	1.026	1.023	1.022	1.027	1.024	1.022		1.036	1.026	1.030	1.024	1.029	1.029	1.036	1.022	1.022	1.025
1.051	1.000	1.003	1.022	1.024	1.022	1.008	1.058	1.000	1.021	65+	1.042	1.028	1.025	1.027	1.024	1.039	1.031	1.029	1.047	1.038
1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025	1.025		1.031	1.031	1.031	1.031	1.031	1.031	1.031	1.031	1.031	1.031
1.031	1.019	1.020	1.024	1.025	1.024	1.021	1.032	1.019	1.024		1.034	1.030	1.029	1.030	1.029	1.033	1.031	1.030	1.035	1.033
1.031	1.019	1.019	1.024	1.025	1.024	1.021	1.033	1.019	1.024		1.034	1.030	1.030	1.030	1.029	1.033	1.031	1.031	1.035	1.033
										PROVINCIAL RATE	1.057	1.040	1.035	1.037	1.042	1.040	1.037	1.034	1.029	1.026
											1.041	1.043	1.041	1.042	1.042	1.042	1.042	1.042	1.041	1.042
											1.044	1.042	1.040	1.040	1.042	1.041	1.041	1.040	1.039	1.039
											1.045	1.042	1.040	1.040	1.041	1.042	1.041	1.040	1.039	1.039

X ji
F j
F ji using direct variance
F ji using regression variance estimates

Table 4
Estimated Risk of Using Various Estimates Averaged over Age - Sex Groups

Risk using estimates from Table 1

	BC	ALTA	SASK	MAN	ONT	QUE	NB	NS	PEI	NFLD
Census	1526.31	1295.36	530.37	569.97	4859.50	3541.80	389.63	478.39	77.90	313.19
Synthetic										
Total	19.10	18.00	14.27	14.49	35.02	28.72	13.64	14.06	78.04	13.26
Model	11.83	11.82	11.76	11.81	11.84	11.86	11.83	11.83	11.79	11.81
Estimation	3.24	2.77	1.12	1.19	10.29	7.46	0.81	0.99	0.14	0.65
Bias	4.03	3.42	1.38	1.49	12.90	9.39	1.01	1.24	66.11	0.80
Bayesian										
Total	22.86	21.84	17.80	18.05	39.51	32.94	17.20	17.65	81.50	16.83
Model	9.02	9.03	8.97	9.01	9.03	9.05	9.02	9.02	8.99	9.01
Estimation	3.56	3.02	1.23	1.31	11.30	8.19	0.88	1.09	0.16	0.71
Bias	4.03	3.42	1.38	1.49	12.90	9.39	1.01	1.24	66.11	0.80
Var. Comp	6.26	6.36	6.22	6.25	6.28	6.31	6.29	6.29	6.25	6.30

Risk using estimates from Table 2

	BC	ALTA	SASK	MAN	ONT	QUE	NB	NS	PEI	NFLD
Census	1485.29	1255.13	521.17	558.74	4739.26	3441.37	382.73	469.55	77.80	310.51
Synthetic										
Total	19.99	19.04	15.25	15.49	35.41	29.25	14.70	15.10	77.94	14.29
Model	13.02	13.15	12.80	12.89	12.97	13.07	12.93	12.92	12.81	12.86
Estimation	3.05	2.58	1.09	1.15	9.86	7.06	0.78	0.97	0.14	0.64
Bias	3.92	3.30	1.35	1.45	12.58	9.12	0.98	1.22	64.99	0.79
Bayesian										
Total	23.05	22.08	18.18	18.42	39.38	32.86	17.61	18.05	80.82	17.28
Model	9.61	9.70	9.49	9.54	9.59	9.65	9.57	9.57	9.50	9.53
Estimation	3.44	2.91	1.22	1.29	11.07	7.95	0.88	1.08	0.16	0.72
Bias	3.92	3.30	1.35	1.45	12.58	9.12	0.98	1.22	64.99	0.79
Var. Comp	6.08	6.17	6.12	6.13	6.14	6.14	6.18	6.19	6.17	6.23

NOTE: Bias 0.05 = $\frac{45,600}{883,898}$