Statistica Sinica **7**(1997), 1065-1083

EMPIRICAL BAYES ESTIMATORS OF SMALL AREA PROPORTIONS IN MULTISTAGE DESIGNS

Patrick J. Farrell, Brenda MacGibbon and Thomas J. Tomberlin

Acadia University, Université du Québec à Montréal and Concordia University

Abstract: The importance of small area estimation as a facet of survey sampling cannot be over-emphasized. Of late, there has been an increasing demand for small area statistics in both the public and private sectors. It is widely recognized that direct survey estimates for small areas are likely to be unstable due to the small sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to obtain more accurate estimates.

In this study, an empirical Bayes methodology for the estimation of small area proportions is proposed, implemented, and evaluated. The basic idea consists of incorporating into a logistic regression model, random effects that are nested in such a way as to reflect the complex structure of a multistage sample design. This yields both point estimates and associated naive measures of accuracy. The latter do not incorporate the uncertainty that arises from estimating the prior distribution of the random effects. We adjust these naively-estimated measures of uncertainty using the bootstrap techniques developed by Laird and Louis (1987).

The proposed estimation approach is applied to data from a United States Census to predict local labour force participation rates. Results are compared with those obtained using unbiased and synthetic estimation, as well as a domain-adjusted synthetic estimation approach which incorporates predictor variables at both the individual and local area levels.

Key words and phrases: Bootstrap, complex survey design, labour force participation, logistic regression, random effects models.

1. Introduction

The importance of small area estimation as a facet of survey sampling cannot be over-emphasized. Of late, there has been an increasing demand for small area statistics in both the public and private sectors. In small area estimation, estimates of small area parameters and associated measures of uncertainty are required. However, estimates based on classical finite population sampling are typically unstable due to the small sample sizes involved. This deficiency has led to the development of model-based estimates which "borrow strength" from related local areas to obtain estimates which are more accurate. One of the first of these procedures was a synthetic model-based approach proposed by Gonzales (1973), which was subsequently used by Gonzales and Hoza (1978). However, it is now recognized that such a methodology produces estimates which have a tendency to be model-biased and associated measures of uncertainty which are typically misleading.

Such deficiencies have led to the development of other model-based procedures for estimating small area parameters which borrow strength from related areas. Among these are techniques based on empirical and hierarchical Bayes approaches. Ghosh and Rao (1994) reviewed the available techniques for small area estimation and illustrated that small area estimates based on empirical and hierarchical Bayes techniques do not have any of the aforementioned undesirable properties associated with estimates obtained using classical unbiased or synthetic approaches.

One of the earliest uses of empirical Bayes methods based on linear models for small area estimation was that of Fay and Herriot (1979). Using data from the 1970 United States Cesus of Population and Housing, these authors obtained estimates of income for small areas. Hierarchical Bayes estimators based on linear models have also been proposed. Datta and Ghosh (1991) presented a unified Bayesian theory for mixed linear models with particular emphasis on small area estimation.

Several authors have considered the problem of estimating small area rates and binomial parameters using empirical and hierarchical Bayes approaches. Dempster and Tomberlin (1980) proposed an empirical Bayes method for estimating census undercount for local areas based on logistic regression models containing fixed and random effects. This proposal was further developed by MacGibbon and Tomberlin (1989) and Farrell, MacGibbon, and Tomberlin (1994). Similar models have been used by Wong and Mason (1985) to estimate proportions using data from the World Fertility Survey and by Tomberlin (1988) to estimate Poisson rates using automobile accident data. Others have considered hierarchical Bayes approaches. Stroud (1991) studied hierarchical Bayes models for univariate natural exponential families with quadratic variance functions. Malec, Sedransk, and Tompkins (1993) used a fully Bayes approach to estimate proportions using data from the National Health Interview Survey.

In this study, we want to gain the power of a Bayesian approach by borrowing strength from an ensemble, while simultaneously obtaining desirable frequentist operating characteristics. We have chosen an empirical Bayes methodology in this context. More specifically, the work of MacGibbon and Tomberlin (1989), in which the explicitly model-based approach of Dempster and Tomberlin (1980) was employed to estimate small area proportions using a random effects, logistic regression model and empirical Bayes techniques, is extended to accommodate a general multistage sample design. The random effects model allows the data to determine, by empirical Bayes techniques, an appropriate compromise between the classical design-unbiased estimates which depend only on data in the specific local area, and the fixed effects estimates which pool information across local areas.

This approach yields local area point estimates and associated naive measures of accuracy. These measures of accuracy are termed naive since they do not incorporate the uncertainty that arises from estimating the prior distribution of the random effects. This study incorporates the suggestion of Laird and Louis (1987) to use bootstrap techniques for adjusting naive estimates of accuracy.

The proposed empirical Bayes methodology is employed to obtain point and interval estimates of local area female labour force participation rates using United States Census data. Comparisons are drawn with unbiased and synthetic estimation techniques. For such an estimation problem, there are many issues which require attention. They include the selection of predictor variables, model diagnostics, the sample design, the application to be addressed, and the properties of the estimators employed. Here we focus on the properties of the estimators over repeated realizations of the sample design. For many survey practitioners, such properties are of prime importance.

The proposed empirical Bayes procedure for estimating small area proportions is described in Section 2. A simulation study based on data taken from a United States Census is described in Section 3. In this study, the proposed empirical Bayes model-based estimation approach is compared with classical unbiased estimation, as well as with two different model-based synthetic estimation approaches. Finally, conclusions and discussion are presented in Section 4.

2. Estimation Procedures

The objective of this study is the estimation of small area proportions using multistage designs. To achieve this objective, local areas will be treated as the primary sampling units here, although the two need not coincide. In fact, it is often the cased that small areas cut across primary sampling units.

Let p_i represent the proportion of individuals in the ith local area who possess a characteristic of interest. Then

$$p_i = \sum y_C / N_i, \tag{2.1}$$

where N_i is the population size of local area i, and the sum of y_C is over all individuals in local area i, where y_C is an indicator variable associated with the Cth individual for the characteristic of interest.

The subscript C refer to a set of nested sampling characteristics, indicating local area or primary sampling unit (PSU), secondary sampling unit (SSU) within PSU, tertiary sampling unit (TSU) within SSU, and so on. For example, if a three stage sample design is considered, C would contain the components ijk, to indicate the kth individual belonging to the jth SSU within the ith PSU. Generally speaking, let w be a subscript which references the final stage of the sample design where individuals are to be selected, regardless of the number of stages in the design. Then, for the three stage sample design considered above, C would contain the components $ij \ldots w$, with $w = k$.

We wish to estimate the parameters p_i. The predictive model-based approach proposed by Royall (1970) is used to obtain empirical Bayes estimates. Under this approach, the estimator of p_i in (2.1) is

$$\hat{p}_i = (\sum_{C(i)} y_{_C} + \sum_{C(i)'} \hat{\pi}_C)/N_i, \tag{2.2}$$

where the sum over $C(i)$ of $y_{_C}$ is the sum of the values of the outcome variable for sampled individuals from the ith local area, and the sum over $C(i)'$ of $\hat{\pi}_C$ is the sum of the estimated probabilities for nonsampled individuals in the ith local area.

To obtain values for $\hat{\pi}_C$, we employ the explicitly model-based approach proposed by Dempster and Tomberlin (1980). Under this approach, a model which describes the probabilities π_c associated with individuals in the population is as follows:

$$y_{_C}|\pi_C \backsim \text{i.i.d. Bernoulli}(\pi_C), \quad \text{logit}(\pi_C) = \underline{X}_C^T \underline{\beta} + \delta_{[C(w)]}. \tag{2.3}$$

The vector $\underline{X}_C$ represents a vector of predictor variables associated with the fixed effects, while $\underline{\beta}$ is the vector of fixed effects logistic regression parameters. The vector of predictor variables may include covariates at both the individual and aggregate levels. The quantity $\delta_{[C(w)]}$ represents a sum of random effects associated with local areas or PSU's and smaller sampling units within local areas, excluding the last sampling stage referenced by w, where individuals are selected. For example, if the three stage sample design described above is considered, then $\delta_{[C(w)]} = \delta_i + \delta_{ij}$. There would be a random effect for the ith PSU or local area, δ_i, as well as a random effect for the jth SSU within the ith PSU, δ_{ij}. Note, however, that there is no random effect to account for the final stage in the sample design where individuals would be selected from SSU's.

Here we are interested in obtaining point and interval estimates for local area proportions, p_i. We require estimates of π_C, where

$$\pi_C = [1 + \exp\{-(\underline{X}_C^T \underline{\beta} + \delta_{[C(w)]})\}]^{-1}. \tag{2.4}$$

Once empirical Bayes estimates $\hat{\underline{\beta}}$ and $\hat{\delta}_{[C(w)]}$ have been determined, π_C is estimated by

$$\hat{\pi}_C = [1 + \exp\{-(\underline{X}_C^T \hat{\underline{\beta}} + \hat{\delta}_{[C(w)]})\}]^{-1}. \tag{2.5}$$

2.1. Parameter estimates

Following previous empirical Bayes studies (See Laird (1987); Leonard (1988); Tomberlin (1988); Wong and Mason (1985)), a joint multivariate normal prior distribution is assumed for the random effects. For example, under a two stage sample design where individuals are selected from a sample of local areas, the model in (2.3) becomes:

$$\text{logit}(\pi_{ij}) = \underline{X}_{ij}^T \underline{\beta} + \delta_i, \quad \delta_i \backsim \text{i.i.d. Normal}(0, \tau^2), \tag{2.6}$$

where π_{ij} represents the probability that the jth individual within the ith local area possesses the characteristic of interest, $\underline{\beta}$ contains the constant term β_0, and δ_i represents a random effect associated with the ith local area. The random effects are assumed to follow normal distributions, each with a mean of zero, and unknown variance τ^2. It should be noted that, in order to consider the situation of unequal variances and correlated random effects for these prior distributions the theoretical extension is trivial, but the computational implications are not.

To develop empirical Bayes estimates, the EM algorithm described by Dempster, Laird and Rubin (1977) is used. Suppose that the $\delta_{[C(w)]}$ in (2.3) follow a joint multivariate normal prior. In addition, let $\underline{\delta}$ and $\underline{y}$ be vectors containing the random effects and the data y_C. We begin by considering the distribution of the data, $f(\underline{y}|\underline{\beta}, \underline{\delta}, \mathbf{X})\, \alpha\, \Pi_C\, \pi_C^{y_C} (1 - \pi_C)^{1-y_C}$, where $\mathbf{X}$ is a matrix of predictor variables. If a flat prior is placed on the fixed effects parameters, then the distribution of the parameters for the general model in (2.3) is given by $f(\underline{\beta}, \underline{\delta}|\mathbf{\Sigma_p})\, \alpha\, \exp(-\frac{1}{2}\underline{\delta}^T \mathbf{\Sigma_p}^{-1} \underline{\delta})$ where $\mathbf{\Sigma_p}$ is the prior covariance matrix for the random effects parameters. The product of the distribution of the data and the distribution for the parameters is the joint distribution of the data and the parameters, $f(\underline{y}, \underline{\beta}, \underline{\delta}|\mathbf{\Sigma_p}, \mathbf{X})$. This joint distribution can be used to determine the posterior distribution of the parameters:

$$f(\underline{\beta}, \underline{\delta}|\underline{y}, \mathbf{\Sigma_p}, \mathbf{X}) = f(\underline{y}, \underline{\beta}, \underline{\delta}|\mathbf{\Sigma_p}, \mathbf{X})/f(\underline{y}|\mathbf{\Sigma_p}, \mathbf{X}). \tag{2.7}$$

It is not feasible to obtain a closed form expression for the posterior given in (2.7) due to the intractable integration required to determine the marginal distribution of $\underline{y}$. A possible approach could be a stochastic integration method such as Gibbs sampling to replace numerical integration. Here, following MacGibbon and Tomberlin (1989), we prefer to approximate the posterior as a multivariate normal distribution having its mean at the mode of (2.7) and covariance matrix equal to the inverse of the information matrix evaluated at the mode. It should be noted that neither the equations for the mode, nor the covariance matrix involve the intractable denominator in (2.7). When values are specified for the components of $\mathbf{\Sigma_p}$ the resulting mode and covariance matrix associated with (2.7) constitute Bayes estimates for the model parameters in (2.3). Empirical Bayes estimates can be derived for this model by using the EM algorithm to determine a maximum likelihood estimate for $\mathbf{\Sigma_p}$. For details on how the empirical Bayes estimates are obtained for the model based on a two stage sample design in (2.6), see MacGibbon and Tomberlin (1989).

2.2. Estimates of small area proportions

Once the empirical Bayes model estimates have been obtained, (2.2) is used to obtain empirical Bayes point estimates for small area proportions. In order to obtain an expression for the variance of the estimator defined by (2.2), it is convenient to adopt a more conventional notation for the linear part of the model, using dummy variables to indicate classifications for the random effects parameters. For each C, let $\underline{Z}_C$ be a vector consisting of the fixed effects predictor variables for the Cth individual augmented by a series of binary variables, each indicating whether the individual belongs to a particular sampling unit. Let $\hat{\underline{\Gamma}}$ represent a vector of estimated fixed and random effects parameters such that $\underline{Z}_C^T\hat{\underline{\Gamma}} = \underline{X}_C^T\hat{\underline{\beta}} + \hat{\delta}_{[C(w)]}$.

Since a model-based approach to estimation is employed, the uncertainty in $\hat{p}_i$ arises from repeated realizations of the model in (2.3). Since the approach is also predictive in nature, the $\sum y_C$ term in (2.2) will have zero variance. Thus, the mean square error of $\hat{p}_i$ as a predictor for p_i is estimated as

$$\hat{\mathrm{MSE}}(\hat{p}_i) = \hat{\mathrm{Var}}\,(\frac{\sum_{C(i)'}\hat{\pi}_C}{N_i}) + \frac{\sum_{C(i)'}\hat{\pi}_C(1-\hat{\pi}_C)}{N_i^2}. \tag{2.8}$$

For sampled local areas, where n_i is greater than zero, the first term of (2.8) is of order $1/n_i$, while the second term is of order $1/N_i$. We base our approximation of the mean square error of $\hat{p}_i$ on the first term only, yielding a useful approximation so long as N_i is large compared to n_i. For nonsampled local areas, the first term in (2.8) is of order 1; therefore it always dominates the second term.

To develop an expression to approximate the uncertainty associated with $\hat{p}_i$, a first order multivariate Taylor series expansion of (2.2) is taken with respect to the realized values of the fixed and random effects estimates. The resulting approximation for $\hat{p}_i$ is a linear function of the estimators of the fixed and random effects parameters. By deriving the variance of the linear function, we obtain

$$\hat{\mathrm{Var}}\,(\hat{p}_i) = \sum_{C(i)'} Z_C^T\hat{\pi}_C(1-\hat{\pi}_C)(\frac{\hat{\mathbf{\Sigma}}}{N_i^2})\sum_{C(i)'} Z_C\hat{\pi}_C(1-\hat{\pi}_C), \tag{2.9}$$

where $\hat{\mathbf{\Sigma}}$ represents the covariance matrix of $\hat{\underline{\Gamma}}$.

2.3. Bootstrap adjustment of naive empirical Bayes interval estimates

When the previous naive approach to estimation is employed, empirical Bayes interval estimates are typically too short. Measures of accuracy do not account for the uncertainty which results from estimating the prior distribution of the parameters. A number of authors have suggested methods for dealing with this problem (See Carlin and Gelfand (1990), Deely and Lindley (1981), Laird and Louis (1987), and Morris (1983)). Here, we use the parametric Type III bootstrap proposed by Laird and Louis (1987). For a comparison of the use of Type II and Type III bootstraps in a similar situation, see Farrell (1991).

The method requires the generation of a number of bootstrap samples, N_B, from a given set of data. The procedure for the generation of a single bootstrap sample, say the bth, can described as follows:

(1) For a given set of sample data, the estimation procedures described in Section 2.1 are applied to the model in (2.3) to obtain empirical Bayes estimates of the fixed and random effects, $\underline{\hat{\beta}}$ and $\hat{\delta}_{[C(W)]}$, along with an estimate for the prior distribution.

(2) For each sampled sampling unit excluding those associated with the last stage of the design, a random effect is generated using the estimated prior distribution of the random effects obtained in step (1).

(3) For each sample observation, a probability, π^*_{bC}, is computed by replacing $\underline{\beta}$ and $\delta_{C(w)}$ in (2.4) with $\underline{\hat{\beta}}$ obtained in step (1) and $\delta^*_{b[C(w)]}$, which is determined by simply adding the appropriate random effects generated in step (2).

(4) For each sample observation, a value for y^*_{bC}, is generated from a Bernoulli random variable with parameter π^*_{bC}.

(5) The values obtained for y^*_{bC}, along with the vectors $\underline{Z}_C$ for sampled individuals constitute the data for a bootstrap sample.

For the bth bootstrap sample, the estimation procedures described in Section 2.1 are applied to the model in (2.3) to obtain the estimate $\hat{\mathbf{\Sigma}}^*_{\mathbf{bp}}$. These same procedures are then applied to the original data $\underline{y}$, the vectors $\underline{Z}_C$ for sampled individuals, and $\hat{\mathbf{\Sigma}}^*_{\mathbf{bp}}$ to obtain estimates of the fixed and random effects, $\underline{\hat{\beta}}^*_b$ and $\hat{\delta}^*_{b[C(w)]}$, along with an associated covariance matrix, $\hat{\mathbf{\Sigma}}^*_{\mathbf{b}}$. These estimates are employed to produce an estimate, $\hat{p}^*_{bi}$, for the proportion of local area i by replacing $\hat{\pi}_C$ in (2.2) with $\hat{\pi}^*_{bC}$, where $\hat{\pi}^*_{bC}$ is determined by substituting $\underline{\hat{\beta}}^*_b$ and $\delta^*_{b[C(w)]}$ into (2.5) for $\underline{\hat{\beta}}$ and $\hat{\delta}_{[C(w)]}$, respectively. An estimate of the variability of $\hat{p}^*_{bi}$, $\hat{\mathrm{Var}}\,(\hat{p}^*_{bi})$, is obtained by replacing $\hat{\pi}_C$ and $\hat{\mathbf{\Sigma}}$ in (2.9) by $\hat{\pi}^*_{bC}$ and $\hat{\mathbf{\Sigma}}^*_{\mathbf{b}}$, respectively. The quantities $\hat{p}^*_{bi}$ and $\hat{\mathrm{Var}}\,(\hat{p}^*_{bi})$ are determined for each of N_B bootstrap samples, and used to calculate

$$\hat{p}^*_i = \sum_b \hat{p}^*_{bi}/N_B, \tag{2.10}$$

and the bootstrap-adjusted estimate of variability associated with the empirical Bayes estimator, $\hat{p}_i$:

$$\hat{\mathrm{Var}}^{\,*}(\hat{p}_i) = \frac{\sum_b \hat{\mathrm{Var}}\,(\hat{p}^*_{bi})}{N_B} + \frac{\sum_b (\hat{p}^*_{bi} - \hat{p}^*_i)^2}{N_B - 1}. \tag{2.11}$$

3. The Simulation Study

In order to compare local area estimates obtained using the proposed empirical Bayes technique with those based on other commonly used methods, a simulation study was performed. In particular, estimates based on the empirical Bayes technique were contrasted with those obtained using a classical unbiased estimation approach, a synthetic model-based procedure proposed by Gonzales and Hoza (1978), and a domain-adjusted synthetic model-based approach which, unlike the ordinary synthetic estimation procedure, includes local area level predictor variables.

These methods were employed to estimate local labour force characteristics using data from a 1% sample of the 1950 United States Census (United States Bureau of the Census 1984). Here we focus our attention on local area female labour force participation rates, where local areas are more or less confined to states. Each method was used to develop point and interval estimates for 52 local areas. Some of these local areas were sampled, while others were not.

Since the empirical Bayes and synthetic estimation approaches are model-based procedures, predictor variables must be selected. Typically, in a real-life situation, historical data would be available for survey planning purposes. For example, variable selection for purposes of model predictions could be based on previous census data. In order to emulate this situation, we selected a simple random sample of size 2,000 from the 1% sample. Variables for model prediction were obtained using stepwise logistic regression. Random effects were not considered in the model at this stage. The predictor variables selected were age, marital status, and whether the individual had children. The data for estimating local area female labour force participation rates were then obtained from the 1% sample using multistage sample designs, emulating an actual sample survey.

Two different sets of point and interval estimates were obtained using each estimation method; one set was based on data obtained using a two stage sample design, the other set on data drawn using a three stage design. In the two stage design, twenty of the fifty-two local areas were first selected without replacement using probabilities proportional to size (PPS). Then, 50 individuals were randomly selected from each chosen local area, for a total sample size of 1,000. In the three stage design, twenty local areas were once again selected without replacement using PPS. Following this, two state economic areas (SEA) were selected without replacement using PPS from each local area, and then 25 individuals were randomly selected from each chosen SEA, again yielding a total sample size of 1,000.

For each design, two hundred replicates were drawn producing two hundred point and interval estimates for a given local area using each method; thus making it possible to analyze results over repeated realizations of the appropriate sample design. Note that since inference is based conditionally on a particular set of local areas being sampled, resampling was not performed at the local area selection stage when the two hundred replicates for a particular design were drawn. Thus, for a given design, the same twenty local areas were sampled in each of the two hundred replicates.

For both sample designs, the classical design-unbiased estimates are the observed local area sample proportions. Thus, such estimates would not be available for small areas which were not sampled. Variance estimates for the unbiased estimator are based on repeated realizations of the sample design and can readily be obtained (See equations (3.8) and (11.44) in Cochran (1977)).

The two synthetic estimators were based on models similar to that used to illustrate the empirical Bayes estimation approach in Section 2. The only difference is that no random effects are included. For example, in the two stage sample design, both synthetic estimators were based on the model, $\text{logit}(\pi_{ij}) = \underline{X}_{ij}^T \underline{\beta}$. The models used in synthetic estimation included individual level predictor variables for age, marital status, and the presence of children, while the domain-adjusted synthetic estimator is based on models containing not only these individual level covariates, but also local area variables representing average age, the proportions of individuals in various marital status categories, and the proportion of individuals having children.

For the synthetic and fixed effects model-based methods, parameters are estimated using pseudo-maximum likelihood as described by Roberts, Rao, and Kumar (1987). Once estimated, these models are used to produce local area estimates in a fashion similar to that of Section 2.2.

For the two stage sample design, the empirical Bayes estimator was based on the model in (2.6). For the three stage design, the model is given by $\text{logit}(\pi_{ijk}) = \underline{X}_{ijk}^T \underline{\beta} + \delta_i + \delta_{ij}$, $\delta_i \sim$ i.i.d. Nomal$(0, \tau_1^2)$, $\delta_{ij} \sim$ i.i.d. Normal$(0, \tau_2^2)$, where δ_i is the random effect associated with the ith local area and δ_{ij} is the random effect for the jth state economic area within the ith small domain. The models used for empirical Bayes estimation included predictor variables at both the individual and local area levels that were identical to those specified for the models associated with the domain-adjusted synthetic estimator.

Variance estimates for model-based estimators such as synthetic and empirical Bayes are based on repeated realizations of an appropriate model. However, for many survey practitioners, the properties of estimators over repeated realizations of the sample design are considered of prime importance. In order to address this concern, our comparison of the three estimators is based on repeated realizations of the sample design. For a discussion of the differences between design-based and model-based inference, see Hansen, Madow, and Tepping (1978), Royall (1970) and Scott and Smith (1969).

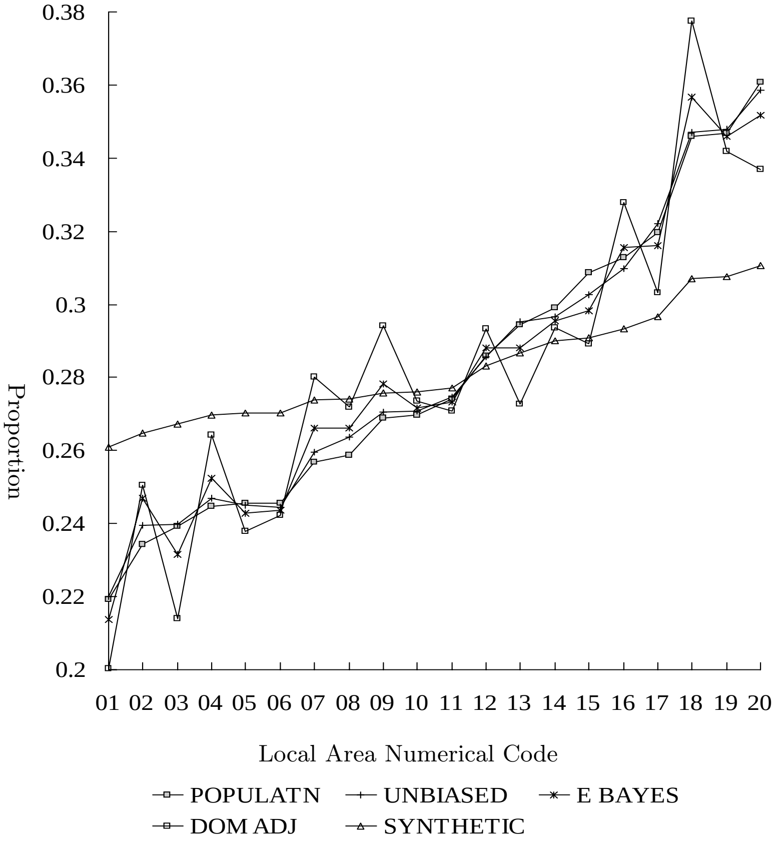

Figure 1. Average point estimates for the proportions of sampled local areas

As indicated earlier, two hundred samples were drawn from the Census data set using each of the two and three stage sample designs previously described. The results for the two stage design will now be presented. For this design, unbiased, empirical Bayes, synthetic, and domain-adjusted synthetic point estimates of female participation rates for each of the twenty sampled local areas along with associated measures of uncertainty were determined for each replicate. For the remaining thirty-two local areas which were not sampled, analogous estimates for only the empirical Bayes and the two synthetic estimators can be obtained, since the unbiased estimates are based on observed local area sample proportions.

For each estimation method, average estimated rates (over all 200 replicates) for each of the twenty sampled local areas are presented in Figure 1, arranged in ascending order according to the population rates. The population rates are also plotted. Note that there is little difference between the average of the unbiased estimates and the population rates, empirically confirming their unbiasedness. When the population proportions are relatively small or large, the synthetic estimator does not perform well with respect to design bias (See the local areas near the two extremes of the horizontal axis). As the expected value of the synthetic estimator increases, the bias associated with the estimator increases from large negative to large positive values. This correlation can occur if the outcome variable is correlated with covariates at the local area level. This is demonstrated by the point estimates associated with the domain-adjusted synthetic approach, where this correlation has been removed by including covariates at the local area level in the model. To evaluate the design bias of the estimators across all sampled local areas, the mean absolute difference between the small area proportions and the average estimated rates was determined for each estimator. The bias associated with the domain-adjusted synthetic estimator over all sampled local areas (0.0153) is somewhat smaller than that of the synthetic estimator (0.0203). Across all sampled local areas, the design bias of the empirical Bayes estimator is quite small (0.0058), matching that of the unbiased estimator (0.0021), while being dramatically better than that of the two synthetic estimators. The graph reflects the fact that the empirical Bayes estimation methodology is a compromise between the unbiased and domain-adjusted synthetic procedures.

The empirical root mean square errors (RMSE) over the two hundred replicates for the twenty sampled local areas are presented in Figure 2. This figure demonstrates graphically where the synthetic estimator performs well and where it performs poorly. When the expected value of the synthetic estimator is very close to the population proportion, the synthetic estimator has by far the smallest empirical RMSE. By pooling data from the whole sample, it obtains a small sampling variance. On the other hand, for the local areas near the two extremes of the horizontal axis, the empirical RMSE for the synthetic estimator is large. This can be explained by the large model bias of the synthetic estimator in these local areas. Results obtained for the domain-adjusted synthetic estimator can be explained similarly.

The empirical Bayes estimator obtains some of the reduction in empirical RMSE that results from pooling the data across local areas, without suffering from the sometimes large model bias associated with the two synthetic estimators. Figure 2 indicates that the empirical Bayes estimator has a smaller empirical RMSE than the synthetic estimators for most local areas. In addition, the empirical Bayes estimator always has a smaller empirical RMSE than the unbiased estimator. In this case, the empirical Bayes estimator gives more weight to the unbiased estimator than to the domain adjusted synthetic estimator. Averaging over the twenty sampled local areas, the empirical Bayes estimator obtains the smallest average empirical RMSE (0.0472), the unbiased estimator obtains the largest (0.0622), while the synthetic and domain-adjusted synthetic estimators obtain averages between these two (0.0521 and 0.0481). Thus, the average empir-

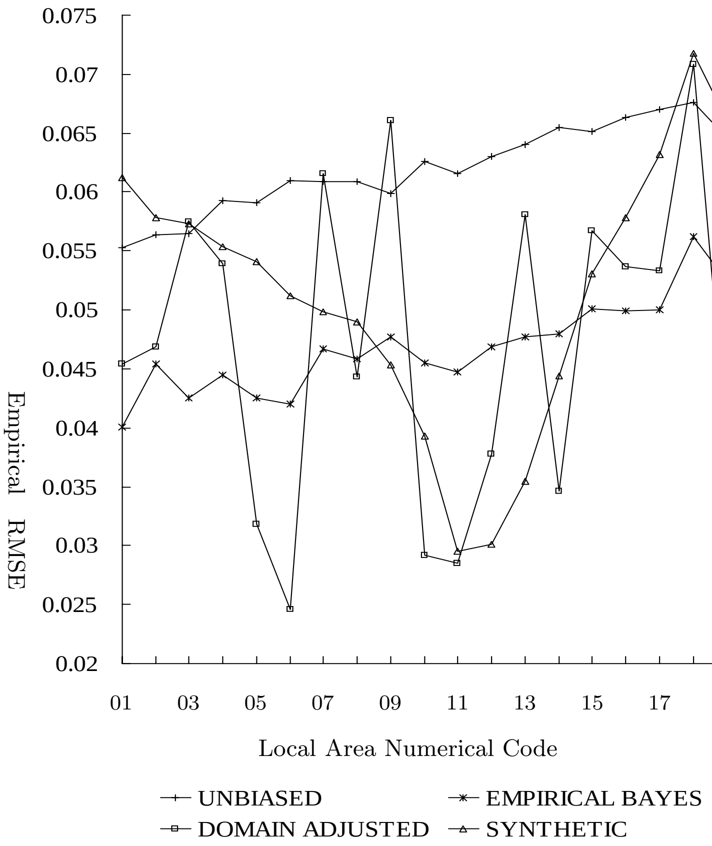

Figure 2. Empirical root mean square errors for sampled local areas

ical RMSE for the empirical Bayes estimator is very close to that of the domain-adjusted synthetic estimator. On the other hand, it should be noted that the RMSE's associated with individual local areas are relatively constant for the empirical Bayes estimator, but highly variable for the domain-adjusted synthetic estimator. This characteristic makes the empirical Bayes estimator easier to justify to users who are concerned to a greater extent about estimation for individual local areas than about the average performance over the ensemble.

Each of the estimators have associated measures of uncertainty. For the unbiased estimates, these are simply the classical sampling standard errors. For the two synthetic estimates, these are the usual estimates of standard errors associated with maximum likelihood estimates. For the empirical Bayes estimates, these are the square roots of the naive estimates given in (2.9). The usefulness

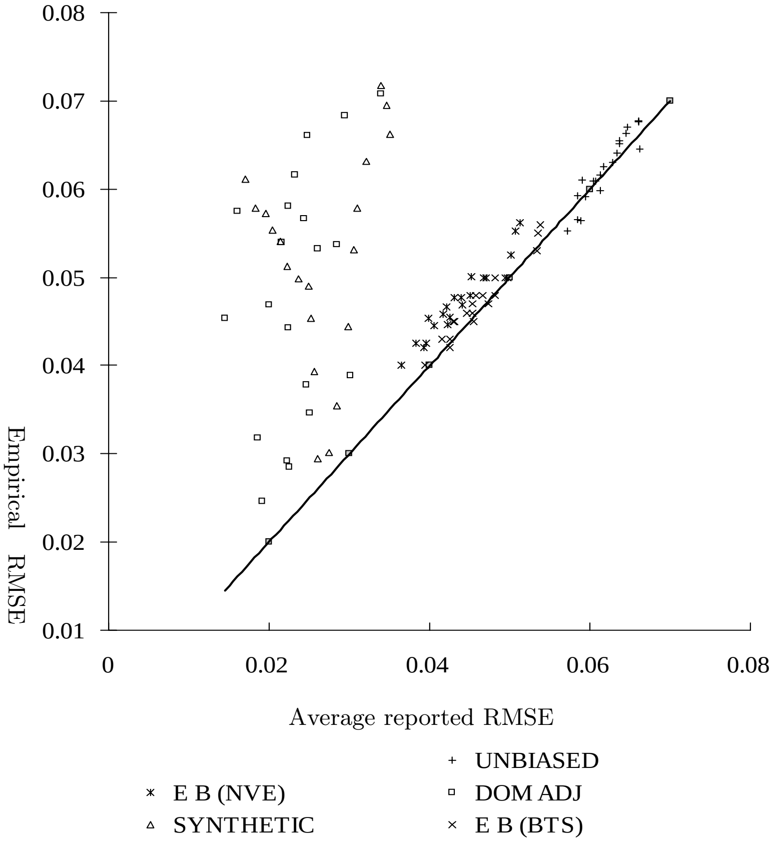

Figure 3. Empirical RMSE versus average reported RMSE for sampled local areas

of these computed measures of uncertainty, called "reported RMSE" here, are compared graphically for all three estimators in Figure 3. The vertical axis corresponds to the empirical RMSE while the horizontal axis corresponds to the average reported RMSE over the two hundred replicates.

For the unbiased estimates, the average reported and empirical RMSE's are almost equal. In contrast, the points corresponding to the two sets of synthetic estimates are in a cluster above 0.015 to 0.035 on the horizontal axis of Figure 3. For these estimates, the average reported RMSE's are quite small, while the empirical RMSE's are much larger, ranging from 0.03 to 0.07. This indicates that the reported RMSE's are not useful for describing the uncertainty associated with the synthetic estimators.

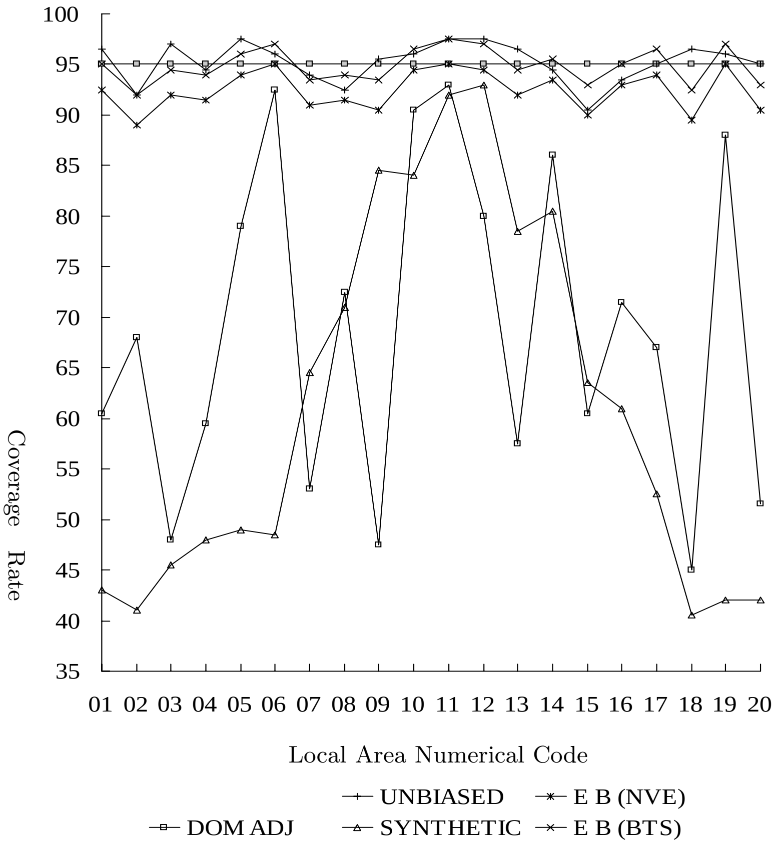

Figure 4. Coverage rates for sampled local areas

For the empirical Bayes estimator, the reported RMSE's slightly underestimate the true variability since they do not incorporate the uncertainty that arises from having to estimate the prior distribution of the random effects. Bootstrap-corrected measures of reported RMSE have also been determined using (2.10) and (2.11). From each of the 200 samples, 100 bootstrap samples were generated using the procedure described in Section 2.3. The average of the bootstrap-corrected measures of reported RMSE (this average being taken over the 200 replicates) is also plotted in Figure 3. When compared to the naive estimates, these adjusted measures of uncertainty appear to be more representative of the accuracy of the empirical Bayes local area estimates.

Symmetric 95% confidence intervals were obtained using the point estimates and reported RMSE's for each estimator. To investigate the coverage properties of the intervals, confidence intervals for each of the twenty sampled local areas were determined for each of the 200 replicates. The coverage rates for each local area are presented in Figure 4. For the unbiased estimates, the coverage rates range from 90.5% to 97.5% with an average of 95.20%, extremely close to the nominal rate of 95%. On the other hand, the coverage rates for the synthetic estimator range from 40.5% to 93.0% with an average of 61.23%, a value that is far below the 95% nominal rate. Coverage rates for the domain-adjusted synthetic estimator are only slightly better, ranging from 45.0% to 93.5%, with an average of 68.55%. The coverage rates for the naive empirical Bayes confidence intervals are close to but consistently below the corresponding unbiased estimate coverage rates. They range from 89.0% to 95.0% with an average of 92.43%, slightly below the 95% nominal rate. The bootstrap-adjusted coverage rates are very close to the 95% nominal rate with an average of 94.88%. Thus, the bootstrap-corrected measures of accuracy seem capable of incorporating most of the uncertainty that arises from having to estimate the prior distribution for sampled local areas.

Thirty-two of the fifty-two local areas were not sampled. It is possible to estimate female labour force participation rates in these nonsampled local areas using the empirical Bayes and synthetic estimation approaches; however, unbiased estimates are not available. Empirical Bayes point estimates and associated measures of uncertainty can be developed for nonsampled local areas using (2.2) and (2.9) by basing the estimate of the local area effect and its associated measure of uncertainty on the estimated prior distribution, and assuming that the effect is independent of all other local area effects and the fixed effects parameters. Synthetic and domain-adjusted synthetic estimates are obtained in exactly the same fashion as they were for sampled local areas, since no local area effects are included in the model upon which the estimates are based.

Average point estimates for each of the 32 nonsampled local areas, for each available method, are presented in Figure 5 along with the population proportions, while the associated coverage rates appear in Figure 6. Over all nonsampled local areas, the mean absolute difference between the small area proportions and the average estimated rates associated with the empirical Bayes estimator (0.0180) was extremely close to that of the domain-adjusted synthetic estimator (0.0184), and was substantially better than that of the synthetic estimator (0.0324). The coverage rates for the synthetic estimates range from 35.0% to 71.5% with an average of 51.36%, which is far below the 95% nominal rate. Results for the domain-adjusted synthetic estimator are only slightly better, ranging from 44.5% to 74.0%, with an average of 60.78%. On the other hand, the coverage rates for the naive empirical Bayes confidence intervals are much better. They range from 78.5% to 95.0% with an average of 89.31%. Bootstrap-corrected coverage rates are even closer to the 95% nominal rate with an average of 94.73%.

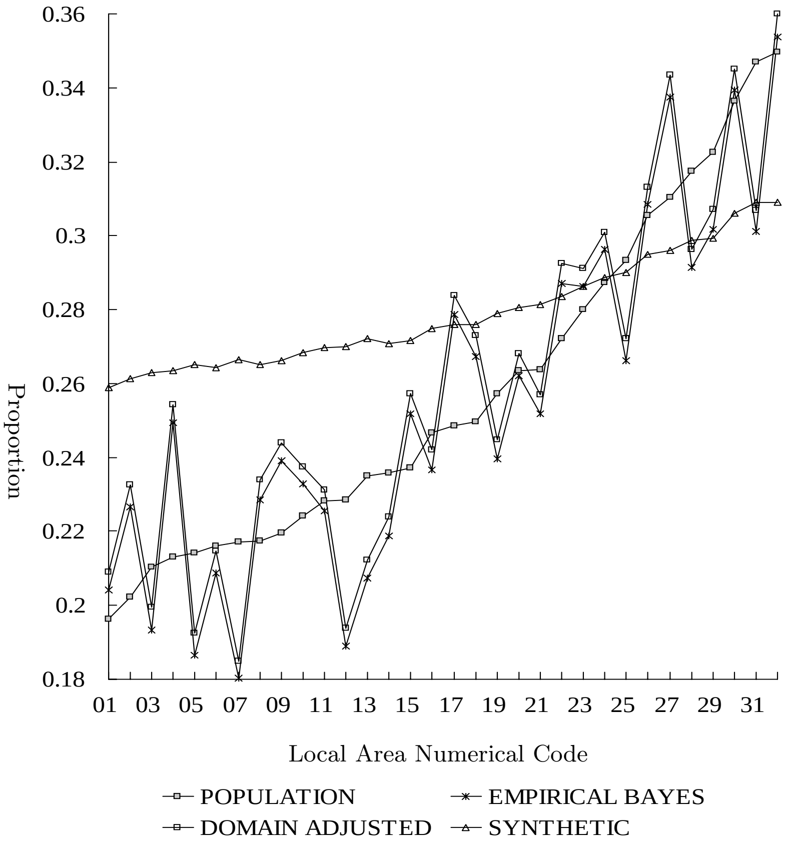

Figure 5. Average point estimates for the proportions of nonsampled local areas

The simulation study for the three stage design gave analogous results (See Farrell (1991)). However, the performance of the three various estimators deteriorated somewhat. This can be attributed to the additional variability associated with the more complex design.

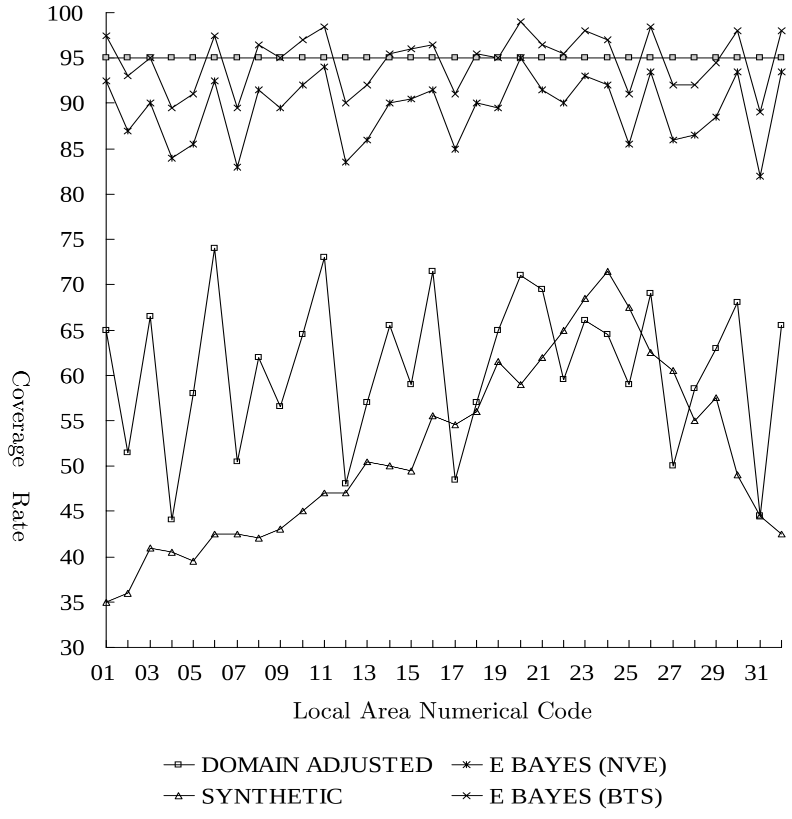

Figure 6. Coverage rates for nonsampled local areas

4. Conclusion

Unbiased, synthetic and empirical Bayes methods have been used to estimate local area female labour force participation rates using data taken from the 1950 United States Census. The models used for classical synthetic estimation contained only individual level predictor variables, while the models specified for the empirical Bayes and domain-adjusted synthetic estimators contained covariates at both the individual and local area levels. The models upon which the synthetic estimator is based were found to be mis-specified, as the bias of the estimator is correlated with its expected value. By including local area level covariates, the domain-adjusted synthetic estimator does not suffer from this mis-specification.

As a compromise between the unbiased and domain-adjusted synthetic estimators, the empirical Bayes estimator performed well in terms of design bias, empirical RMSE, and coverage rates. Finally, bootstrap techniques were shown to be useful for improving naive empirical Bayes estimates of uncertainty. Bootstrapped interval estimates based on an empirical Bayes approach were found on average to attain the desired level of coverage. These interval estimates performed much better than counterparts based on the domain-adjusted synthetic estimator for both sampled and nonsampled local areas, even though the point estimates for the empirical Bayes and domain-adjusted synthetic estimation approaches were very similar in the case of nonsampled local areas.

Acknowledgements

This research was supported by SSHRC and NSERC of Canada, and FCAR of Quebec. The authors are grateful to the associate editor and the referee for their valuable comments and suggestions, particularly those leading to the domain-adjusted synthetic estimator.

References

Carlin, B. P. and Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.* **85**, 105-114.

Cochran, W. G. (1997). *Sampling Techniques*, 3rd Edition. Wiley, New York.

Datta G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Ann. Statist.* **19**, 1748-1770.

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76**, 833-841.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Dempster, A. P. and Tomberlin, T. J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.

Farrell, P. J. (1991). Empirical Bayes estimation of small area proportions. PhD Dissertation, Department of Management Science, McGill University, Montreal, Quebec, Canada.

Farrell, P. J., MacGibbon, B. and Tomberlin, T. J. (1994). Protection against outliers in empirical Bayes estimation. *Canad. J. Statist.* **22**, 365-376.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269-277.

Gonzales, M. E. (1973). Use and evaluation of synthetic estimation. *Proc. Amer. Statist. Assoc. Social Statistics Section*, 33-36.

Gonzales, M. E. and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. *J. Amer. Statist. Assoc.* **73**, 7-15.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9**, 55-93.

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1978). On inference and estimation from sample surveys. *Proc. Amer. Statist. Assoc. Survey and Research Methods Section,* 82-107.

Laird, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581-591.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82**, 739-750.

Leonard, K. J. (1988). Credit scoring via linear logistic models with random effects parameters. PhD Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montreal, Quebec, Canada.

MacGibbon, B. and Tomberlin, T. J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology* **15**, 237-252.

Malec, D., Sedransk, J. and Tompkins, L. (1993). Bayesian predictive inference for small areas for binary variables in the national health interview survey. In *Case Studies in Bayesian Statistic* (Edited by Constantine Gatsonis, James S. Hodges, Robert E. Kass and Nozer D. Singpurwalla), 377-389. Springer Verlag, New York.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78**, 47-65.

Roberts, G., Rao, J. N. K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1-12.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377-387.

Scott, A. and Smith, T. M. F. (1969). Estimation in Multi-Stage surveys. *J. Amer. Statist. Assoc.* **64**, 830-840.

Stroud, T. W. F. (1991). Hierarchical Bayes predicative means and variances with application to sample survey inference. *Comm. Statist. Theory Methods* **20**, 13-36.

Tomberlin, T. J. (1988). Predicting accident frequencies for drivers classified by two factors. *J. Amer. Statist. Assoc.* **83**, 309-321.

United States Bureau of the Census (1984). Census of the population, 1950: Public use microdata sample technical documentation. Edited by J. G. Keane, Washingto, D.C.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist Assoc.* **80**, 513-524.

Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, B0P 1X0, Canada.

E-mail: pat.farrell@acadiau.ca

Département de Mathématiques et d'Informatique, Université du Québec à Montréal, C.P. 8888, Succ. "A", Montréal, Québec, H3C 3P8, Canada.

E-mail: macgibbon.brenda@uqam.ca

Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec, H3G 1M8.

E-mail: jtomb@vax2.concordia.ca

(Received September 1994; accepted November 1995)