Survey Methodology, December 1989
Vol. 15, No. 2, pp. 237–252
Statistics Canada

Small Area Estimates of Proportions Via Empirical Bayes Techniques

BRENDA MacGIBBON[1] and THOMAS J. TOMBERLIN[2]

ABSTRACT

Empirical Bayes techniques are applied to the problem of "small area" estimation of proportions. Such methods have been previously used to advantage in a variety of situations, as described, for example, by Morris (1983). The basic idea here consists of incorporating random effects and nested random effects into models which reflect the complex structure of a multi-stage sample design, as was originally proposed by Dempster and Tomberlin (1980). Estimates of proportions can be obtained, together with associated estimates of uncertainty. These techniques are applied to simulated data in a Monte Carlo study which compares several available techniques for small area estimation.

KEY WORDS: Logistic regression; Random effects models; Bayes estimation; EM algorithm.

1. INTRODUCTION

1.1 The Problem

Complex multi-stage surveys are used to obtain estimates of proportions in many research disciplines (*e.g.*, epidemiology, economics, criminology *etc.*). Not only are estimates for local areas and other special subgroups required, but there is also a need for reliable measures of the accuracy of these estimates. This suggests to us the need for improved methodologies for this estimation problem and related statistical inference.

In addition, the techniques based on the standard normal theory used by Fay and Herriot (1979) to estimate income, a continuous random variable, in small areas are no longer directly applicable to the problem of estimating proportions for discrete outcome variables. Here, it is the logit transform of the proportion, not the proportion itself, that will be modelled in a linear way. This creates the same problems of estimation as in classical statistical logistic regression theory. (See Haberman 1978.) Unfortunately, fewer attempts have been made to solve this obviously more complex problem in small area estimation.

In order to address the problem of inference from a relatively thinly spread complex, multi-stage survey to small areas or domains not necessarily included in the survey, we have chosen an explicitly model-based approach. This was proposed originally by Dempster and Tomberlin (1980) for the estimation of census undercount from a post-enumeration survey. The methodology uses both a random effects, multiple logistic regression model and empirical Bayes techniques. This directly yields estimates of uncertainty associated with the estimated proportions for small areas via a Bayesian paradigm. This explicitly model-based method differs substantially from the implicitly model-based approach of the synthetic estimation techniques of Gonzalez and Hoza (1976, 1978), Gonzalez and Waksberg (1975), and others.

[1] Brenda MacGibbon, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8 and Département de mathématiques et d'informatique, Université du Québec à Montréal, C.P. 8888, Suc. "A", Montréal, Québec H3C 3P8.
[2] Thomas J. Tomberlin, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8.

As a typical complex survey will often be a nested structure of primary sampling units (PSU's), secondary sampling units (SSU's) within PSU's, tertiary sampling units (TSU's) within SSU's and, finally, households within TSU's; the explicitly model-based approach will allow us to take into account the complexitly of the sample design. The purpose of introducing a random effects model is to allow the data to determine, by empirical Bayes techniques, an appropriate compromise between the classical unbiased estimates which depend only on data in the specific local area, and the fixed effects estimates which pool information across areas.

In Section 1.2, a literature review is given and a solution to the problem of estimating proportions for small areas is proposed. The model and its associated estimates are made explicit in Sections 2 and 3 respectively. The results are applied to simulated data in a Monte Carlo study presented in Section 4.

1.2 The Review and a Proposed Solution to the Problem

Because of the growing need for small area statistics in recent years, and because reliable estimates for small areas or subdomains are not usually directly available by classical sample survey methods, several researchers have focused on the problem of small area estimation. This has necessitated the use of explicitly or implicitly model-based methods which allow for "borrowing strength" across small areas in order to increase the effective sample size for estimation, and hence the accuracy of the resulting estimates. Although much of the research in this area has applied linear model techniques and concentrated on the estimation of means or totals, rather than proportions, a discussion of the literature on these estimators and the criteria used to evaluate them can add valuable insight into our problem.

Classical theory dictates that estimators should be design-consistent and, if possible essentially design-unbiased. However these estimators are not always particularly useful when the sample sizes are small.

Gonzalez (1973) described the method of synthetic estimation as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." It seems its first reported use was by the U.S. National Center for Health Statistics (1968) for the calculation of state estimates of long and short term disability rates. Various authors subsequently tried to formalize this concept of synthetic estimation, in particular, for means of continuous outcome variables, using both *ad hoc* and model-based approaches. Gonzalez (1973), Gonzalez and Waksberg (1975), Gonzalez and Hoza (1976) and Levy and French (1978) used previous census data to form post-strata which are subsequently used to combine information across small areas under the assumption that the mean response is similar across a section of these areas. Levy (1971), Ericksen (1973, 1974) and O'Hare (1976) employed regression methods in order to incorporate auxiliary information in small area estimation.The accuracy of this method has been evaluated in terms of its average sampling mean squared error over all small areas in a region.

Ericksen (1974) warned that there is no systematic methodology for the assessment of the bias or accuracy of synthetic estimators. Despite these shortcomings, synthetic estimation still remains a potentially powerful and attractive tool. There have been many reported empirical evaluations both on actual and simulated data sets of synthetic estimation in recent years, including Levy (1971), Gonzalez (1973), Gonzalez and Hoza (1978), and Schaible (1979). Several of these types of studies are described in a volume edited by Platek and Singh (1986).

Royall (1970, 1973), using a model-based approach, also considered the problem of estimating totals in finite populations, when auxiliary information is available. He established a probability model of the relationship between the variable of interest and the auxiliary variable and then derived optimal subdomain predictors.

Holt, Smith and Tomberlin (1979) and Laake (1979) applied the predictive approach of Royall to the problem of small area estimation. Laake (1979) found that in contrast to the synthetic approach, where biased estimators are usually obtained without an explicit method of estimating the bias, the prediction approach yielded estimates of mean squared error (MSE) as a tool for the comparison of estimators. In the problem of estimating small area totals, Holt, Smith and Tomberlin (1979) specified various possibilities of population structure in order to model the assumed relationship across subareas. With a specified model, it becomes possible to determine whether or not it is supported by the data and also to study the effect of model misspecification on the bias of the observed estimators. Under different models, the variance of the estimator, the estimate of the variance and MSE change. They built model-based confidence intervals, which have interpretations in terms of repeated realizations under the superpopulation model.

Purcell and Kish (1979, 1980) reviewed the different existing techniques of small area estimation, subdividing them into the following broad categories, regression-based procedures, the use of empirical Bayes and of Bayesian methods, superpopulation prediction theory, clustering techniques, and categorical data analysis methods. They underlined the fact that small area domain estimation should not be considered as a homogeneous problem, but that there exist many other interacting factors such as domain size which should be taken into account when choosing the type of estimator. Särndal (1984) later confirmed this.

The most serious shortcoming of model-dependent estimators is that useful estimates of mean squared errors are not available using fixed effects models because associated variance estimates do not reflect the bias inherent in estimates based on models having a reduced set of parameters. Two different approaches were then taken to the problem of small area estimation.

Fay and Herriot (1979) used the James-Stein theory of estimation (James and Stein 1961) on sample data to determine estimates of income for small places from the 1970 US Census of Population and Housing. In fact, they used an empirical Bayes approach which originated with Robbins (1955) and has been described by Efron and Morris (1975), thus formalizing the meritorious suggestion of Madow and Hansen (1975) of forming a weighted average of the sample and regression estimates. A similar approach by Schaible, *et. al.* (1977) gives a method for arriving at a composite estimator which is the weighted average of the unbiased and synthetic estimators. For other examples of empirical Bayes methods for small area estimation based on standard normal theory see Stroud (1987) and Cressie (1988).

Battese, Harter and Fuller (1988), using a prediction approach, proposed a nested error regression model in order to estimate means. A more general model, a random coefficients regression model, had been previously proposed for a similar problem by Dempster, Rubin and Tsutakawa (1981). They used Bayesian techniques to estimate fixed and random effects in covariance component models when the covariances and variances are tentatively assumed to be known and the EM algorithm to subsequently estimate these unknown parameters. The introduction of random effects models not only allows for standard maximum likelihood estimation, but also provides measures of the reliability of the final estimates of the parameters in the form of posterior variances.

Ericksen (1980) suggested using the mean squared error (MSE) to evaluate effectiveness of regression in small area estimation. He attempted to answer such questions as: When should more predictor variables be added to the regression equation? Should James-Stein weighting procedures be used when the synthetic and the regression estimate are far apart? He also warned of the effects of outliers on both the resulting estimate and its estimated error. Perhaps the effect on small area estimators of the failure of the linear model assumptions should be more seriously studied.

Although applied to the estimation of counts such as unemployment and mortality statistics, most of these techniques described were designed primarily for continuous outcome variables. Purcell and Kish (1980) introduced a categorical data analysis method for obtaining estimates of counts for small domains. Essentially, their methodology involves fitting log-linear models to the data, omitting some of the higher order interaction terms and obtaining estimates by the iterative proportional fitting algorithm described by Deming and Stephan (1940). We propose to extend these models to the problem of estimation of proportions in small domains as originally conceived by Dempster and Tomberlin (1980) by applying empirical Bayes techniques to logistic regression models with random effects. This would have the added advantage that a measure of uncertainty of the small area estimates would be available through the approximate posterior variances. The estimator proposed here is similar in nature to the composite one used by Schaible *et. al.* (1977) for unemployment rates, the principal difference being in the method for choosing the weights. We feel, however, that the empirical Bayesian paradigm gives a more natural and intuitive method for determining the weights. Empirical Bayes estimation based on simple logistic random effects has already proven useful in studying regional variation in mortality rates by Miao (1977). Somewhat more complex random effects models have been used for proportions on data from the World Fertility Survey (Wong and Mason, 1985) and for Poisson parameters on automobile insurance data (Weisberg, Tomberlin, and Chatterjee 1984 and Tomberlin 1988).

Roberts, Rao and Kumar (1987) fitted logistic regression models to binary outcome data obtained using complex sampling schemes, constructed "pseudo-maximum likelihood" estimators, and compared their estimates to unbiased ones. They also proposed a goodness-of-fit test for their model, which takes the sampling design into account. A fundamental difference between our approach and that of Roberts, *et. al.*, is that by incorporating the characteristics of the sample design into the model, we can estimate parameters, and obtain readily interpretable measures of their reliability by means of standard maximum likelihood techniques.

2. THE MODEL

Following the framework of Dempster and Tomberlin (1980), in its most general form, we specify a model which describes the probabilities associated with individuals in the population as a function of categorical variables, continuous covariates and sampling characteristics. The models we consider in this paper are specific examples of the following,

$$\text{logit}\ (\ \pi_{\mu\nu}) = \theta_\mu + \mathrm{X}_{\mu\nu}\,\underline{\beta} + \phi_\nu \tag{2.1}$$

where $\pi_{\mu\nu}$ represents the probability of a "response" for the ν-th unit in the μ-th cell, the subscript μ refers to a set of categorical variable covariates, and the subscript ν refers to a set of nested sampling characteristics, indicating PSU, SSU within PSU, and so on. The parameter θ_μ represents a sum of fixed classification effects, the parameter ϕ_ν represents a sum of random effects associated with sampling characteristics, the vector $\mathrm{X}_{\mu\nu}$ represents a vector of quantitative covariates, and the parameter $\underline{\beta}$ is a vector of fixed logistic linear regression parameters. The random effects parameters are assumed to have some parametric distribution, usually a multivariate normal distribution. The probabilities $\pi_{\mu\nu}$ are obtained by inverting the logit transformation as follows,

$$\pi_{\mu\nu} = [1 + \exp\{-(\theta_\mu + X_{\mu\nu} + \phi_\nu)\}]^{-1}. \tag{2.2}$$

For purposes of illustration, consider the following simple example. Let the proportion of interest be the labour force participation rate. Suppose we have one classification variable indicating sex and one continuous covariate indicating the age of the individual. Suppose further that the sample design is a simple, two stage cluster sample. In the first stage, a sample of counties is drawn and simple random samples of individuals within selected counties are drawn at the second stage.

For estimation purposes, consider the following model,

$$\text{logit } (\pi_{\mu\nu}) = \theta_\mu + X_{\mu\nu}\beta + \phi_i \tag{2.3}$$

$$\phi_i \sim \text{i.i.d. Normal } (0, \sigma^2). \tag{2.4}$$

Here, the classification subscript, μ, indicates the sex of the individual; the sampling characteristics subscript, $\nu = ij$, indicates the j-th individual within the i-th PSU; $X_{\mu\nu}$ indicates the age of the individual and ϕ_i is a random effect associated with the i-th PSU.

The consequence of assuming that the PSU effects are independent, identically distributed is that PSU departures away from the fixed part of the model are treated as exchangeable; that is, apart from effects of age and sex, no systematic information exists regarding differential employment rates among the counties in the population. Obviously in a realistic situation, such information would exist, for example, dominant industry, distance from principal markets, retail sales, etc. In such cases, this auxiliary information should be incorporated into the model. However, for purposes of illustration, we will continue with this simple model. The choice of a normal distribution of the error terms is a mathematical convenience, and the consequences of this choice must also be evaluated after actual data analysis. Extensions from the simple model described in (2.3-4) to include additional covariates, both categorical and quantitative is straight forward.

In theory, extensions to the model allowing for more complex sample designs is also simple. For example, data drawn using a three stage sample could be modelled using nested random effects as follows.

$$\text{logit } (\pi_{\mu\nu}) = \theta_\mu + X_{\mu\nu}\beta + \phi_i + \phi_{j(i)} \tag{2.5}$$

$$\phi_i \sim \text{Normal } (0, \sigma_1^2)$$

$$\phi_{j(i)} \sim \text{Normal } (0, \sigma_2^2).$$

Here, the sampling characteristics subscript, $\nu = ijk$ refers to the k-th individual within the j-th SSU within the i-th PSU. The parameter ϕ_i is the random effect associated with the i-th PSU, and $\phi_{j(i)}$ is the nested random effect associated with the j-th SSU within the i-th PSU. Stratification variables could also be incorporated within the fixed effects part of the model. While it is simple to write down the models corresponding to sample designs with several stages, without further research, it is not yet clear how difficult it will be to produce estimates based on these more complex models.

In an actual application, it would be necessary to use the data to identify predictor variables. This would require the development of some sort of model selection techniques. While not the primary focus of this paper, one might conceive of such a technique being based on an initial analysis using conventional variable selection techniques for logistic regression models as described by Haberman (1978), for example. Such an analysis could be conducted, ignoring the random effects parameters. Having chosen a set of predictors, the random effects would then be incorporated in the manner dictated by the sample design.

3. ESTIMATES

In this section, we develop empirical Bayes estimates for the simple model described in equations (2.3-4). First, it is assumed that the variance component, σ^2, is known, and Bayesian estimates of the probabilities $\pi_{\mu ij}$ are obtained. Then, the EM algorithm, as described by Dempster, Laird and Rubin (1977), is used to obtain the maximum likelihood estimate of σ^2 allowing for empirical Bayes estimates. Finally, posterior variances of these estimates are obtained. The development of these estimates is similar to that described by Laird (1978) and by Tomberlin (1988).

3.1 Bayes Estimates

As noted by Laird (1978) in her analysis of contingency tables, by Dempster, Rubin and Tsutakawa (1981) in their analysis of variance components for linear models, and by Tomberlin (1988) in his analysis of Poisson data, a Bayesian analysis of a mixed model such as described in (2.3-4) can be obtained by placing a flat prior on the fixed parameters, θ_μ and β and the proper prior given in (2.4) on the random parameters, ϕ_i.

Let the vector of 0-1 outcome variables indicating membership in the labour force be represented by $\underline{y}$ and let $\underline{\pi}$ represent a vector of the individual probabilities $\pi_{\mu ij}$. The data are then distributed as a product binomial given by,

$$p(\underline{y} \mid \underline{\pi}) \propto \prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})}. \tag{3.1}$$

The prior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta \mid \sigma^2) \propto \exp\left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right]. \tag{3.2}$$

Thus, the joint distribution of the data, $\underline{y}$, and the parameters is given by,

$$p(\underline{y}, \underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathbf{X}) = p(\underline{y} \mid \underline{\theta}, \underline{\phi}, \beta, \sigma^2, \mathbf{X})\, p(\underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathbf{X}) \tag{3.3}$$

$$\propto \left[\prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})} \right] \exp\left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right].$$

From this, the posterior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta \mid y, \sigma^2, \mathrm{X}) \frac{p(y, \underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathrm{X})}{p(y \mid \sigma^2, \mathrm{X})}. \tag{3.4}$$

It is not feasible to obtain a closed form expression for the posterior given in (3.4) due to the intractable integration required to obtain the marginal distribution of y. Here we adopt the approximation employed by Laird (1978) and by Tomberlin (1988). The posterior is expressed as a multivariate normal distribution having its mean at the mode of (3.4) and covariance matrix equal to the inverse of the information matrix evaluated at the mode.

Obtaining the mode requires solving the following set of equations. This can be accomplished by using a multivariate Newton-Raphson algorithm.

$$\sum_{\mu ij} y_{\mu ij} \mathrm{X}_{\mu ij} = \sum_{\mu ij} \hat{\pi}_{\mu ij} \mathrm{X}_{\mu ij} \tag{3.5}$$

$$\sum_{ij} y_{\mu ij} = \sum_{ij} \hat{\pi}_{\mu ii} \tag{3.6}$$

$$\sum_{\mu j} (y_{\mu ij} - \hat{\pi}_{\mu ij}) - \frac{\hat{\phi}_i}{\sigma^2} = 0. \tag{3.7}$$

The posterior covariance matrix of the parameters is found by inverting the negative of the second derivative matrix of the log of (3.4) taken with respect to the parameters, and evaluated at the mode. Note that neither the equations for the mode, nor the covariance matrix involve the intractable denominator of (3.4).

Elements of the inverse of the posterior covariance matrix are given by,

$$\frac{-\partial^2}{\partial \beta^2} = \sum_{\mu ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}^2_{\mu ij} \tag{3.8}$$

$$\frac{-\partial^2}{\partial \theta_\mu^2} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \tag{3.9}$$

$$\frac{-\partial^2}{\partial \phi_i^2} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) - \frac{1}{\sigma^2} \tag{3.10}$$

$$\frac{-\partial^2}{\partial \beta \, \partial \theta_\mu} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}_{\mu ij} \tag{3.11}$$

$$\frac{-\partial^2}{\partial \beta \, \partial \phi_i} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}_{\mu ij} \tag{3.12}$$

$$\frac{-\partial^2}{\partial \theta_\mu \, \partial \phi_i} = \sum_{j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}). \tag{3.13}$$

3.2 Empirical Bayes Estimates

To obtain empirical Bayes estimates, the prior variance, σ^2, must be estimated from the data. A reliable estimate requires a reasonable number of PSU's in the sample; otherwise, if the number of PSU's is too small, a purely Bayesian approach is recommended. We propose to estimate the prior variance using an EM algorithm as described by Dempster, Laird and Rubin (1977). The general framework for the estimates is similar to that employed by Laird (1978) for contingency table analysis, and Tomberlin (1988) for Poisson data in a two way classification. The estimates for the simple two-stage sample are obtained in exactly the same way as used by Leonard (1988).

The algorithm is initiated by choosing a starting value, $\sigma^2_{(0)}$, for the variance component. The posterior distribution of the random effects, ϕ_i, is then obtained by carrying out a Bayesian analysis as described in Section 2. This posterior distribution is then used to implement the E-step. The expected value of the sufficient statistic is calculated conditional on the data. The M-step is then completed by merely calculating the maximum likelihood function of the sufficient statistics. For a more complete description of the EM algorithm for regular exponential densities, see Dempster, Laird and Rubin (1977). The process is then repeated with a Bayesian analysis based on the updated estimate of the variance component, $\sigma^2_{(1)}$. The algorithm is continued until it converges.

3.3 Estimates of Small Area Proportions

Estimates together with associated posterior variances and covariances for parameters of the model given in (2.3-4) are presented in Sections 3.1 and 3.2. These estimated parameters are then employed to obtain estimates for small area proportions using a predictive approach. Assuming that the sample sizes within each area are small compared to those of the corresponding populations, this can be accomplished by averaging the individual estimated probabilities:

$$\hat{p}_i = \frac{\sum_{\mu j} \hat{\pi}_{\mu ij}}{N_i} \tag{3.14}$$

where N_i is the number of individuals in the i-th small area, and where the estimated probability associated with the μij-th individual, $\hat{\pi}_{\mu ij}$ is obtained by inverting the logistic function as follows,

$$\hat{\pi}_{\mu ij} = [1 + \exp\{-(\hat{\theta}_\mu + \mathrm{X}_{\mu ij}\hat{\beta} + \hat{\phi}_i)\}]^{-1}. \tag{3.15}$$

To develop posterior variances for the estimates of small area proportions, it is convenient to adopt a more conventional notation for the linear part of the model, using dummy variables to indicate classifications. Let $\boldsymbol{Z}_{\mu ij}$ represent a vector of predictor variables, both quantitative and qualitative, associated with the μij-th individual and let $\underline{\Gamma}$ represent a vector of the parameters of the model. Then,

$$\boldsymbol{Z}^T_{\mu ij}\,\underline{\Gamma} = \theta_\mu + \mathrm{X}_{\mu ij}\,\beta + \phi_i, \tag{3.16}$$

$$\hat{\pi}_{\mu ij} = [1 + \exp(-Z_{\mu ij}^T \hat{\underline{\Gamma}})]^{-1}. \tag{3.17}$$

Then, using a standard Taylor Series method, the posterior variance of the estimated small area proportion can be approximated as,

$$\text{Var}(\hat{p}_i) = \left[\sum_{\mu j} Z_{\mu ij}^T \hat{\pi}_{\mu ij}(1 - \hat{\pi}_{\mu ij})\right] \frac{\hat{\underline{\Sigma}}_{\Gamma}}{N_i^2} \left[\sum_{\mu j} Z_{\mu ij} \hat{\pi}_{\mu ij}(1 - \hat{\pi}_{\mu ij})\right]. \tag{3.18}$$

Here, $\hat{\underline{\Sigma}}_{\Gamma}$ is the posterior covariance matrix of the estimated logistic regression parameters $\hat{\underline{\Gamma}}$.

Should the samples within small areas be substantial parts of the associated populations within those areas, then some additional gains in precision could be made by predicting only for the non-sampled units, in the spirit of the finite population sampling prediction methods originally described by Royall (1970).

4. THE SIMULATION STUDY

A simulation study was carried out to illustrate the characteristics of three different methodologies for producing local area estimates of proportions. The three methods evaluated were, the classical unbiased estimates, model-based estimates similar to the straightforward "synthetic estimates" of Gonzalez and Hoza (1978), and a modification of the proposed empirical Bayes estimates described in section 3, above. Data were simulated for a two-stage sample design. The 15 primary sampling units (PSU's) were also treated as the local areas for which individual estimates of labour force participation rates were required. Within each of the 15 PSU's, simple random samples of 25 individuals were drawn, for a total sample size of 375. The local area populations were assumed to be infinite so that complications associated with finite population sampling could be avoided.

As evaluations for local area estimates were required, it was decided to simulate resampling at the second stage only. That is, the same 15 PSU's were drawn for each of the simulation studies. Each replicate consisted of a different sample drawn within these PSU's. The study was based on 205 replications.

The data were generated using the model described in equation (2.3). The parameters were defined as follows,

$$\begin{aligned} \theta_1 &= -0.5 \\ \theta_2 &= -1.0 \\ \beta &= \ \ 0.1. \end{aligned} \tag{4.1}$$

The random parameters ϕ_i were generated from a normal distribution having mean zero and standard deviation 0.25. The $\pi_{\mu\nu}$ were obtained by inverting the logistic transformation as given in equation (3.15).

Here, θ_1 and θ_2 are the fixed effects associated with men and women respectively. That is, the odds ratio for labour force participation of men to that of women is $\exp[0.5] = 1.65$. The parameter β is the slope parameter associated with age, and the ϕ_i are the logistic random effects associated with the 15 PSU's, or local areas.

Table 1

Population Labour Force Participation Rates by Local Area

Local Area	1	2	3	4	5	6	7	8
Participation Rate	0.79	0.79	0.96	0.88	0.90	0.95	0.86	0.96
Local Area	9	10	11	12	13	14	15	
Participation Rate	0.61	0.87	0.81	0.91	0.94	0.92	0.83	

The predictor variables, were generated with identical distributions for each of the 15 local areas. Age was distributed uniformly on the interval 20 to 40 years, the sex of each individual was drawn from a Bernoulli distribution with proportion 0.5, and the two predictor variables were assumed to be independently distributed. The population labour force participation rates for the 15 local areas are displayed in Table 1. As each local area was assumed to have the same distribution on the predictor variables, the only source of variation from area to area was the random local area effects, the ϕ_i. The random nature of these effects can produce a substantial variation in local area participation rates as is particularly evidenced by local area 9.

The observed local area sample proportions were used as unbiased estimates. The synthetic estimator was based on the following fixed effects, logit model,

$$\text{logit } (\pi_{\mu\nu}) = \theta_\mu \tag{4.2}$$

where, $\pi_{\mu\nu}$ and θ_μ are defined as for the random effects model in (2.3). Notice, only data from a particular local area are used to form the unbiased estimator while data are pooled from all local areas to obtain the synthetic estimator. However, the synthetic estimators will be biased to a degree which depends on the extent that model (4.2) fails to capture differences between local areas.

The third estimator studied here is a modification of the proposed empirical Bayes estimator described in Section 3. Due to the amount of computer time required to estimate the variance component associated with the local area effects, in fact, the Bayes estimator described in Section 3.1 was employed. The prior variance used for these estimates was the known value of the variance given in (4.1) used to simulate the data. As a result of this compromise, the results for the "empirical Bayes" estimator given below would be expected to be somewhat better than those which would be obtained using a true empirical Bayes estimator. However, sensitivity analyses aimed at determining the effect of changes in the prior variance indicate that the results which would be obtained using the empirical Bayes estimator would not be expected to substantially differ from those reported here for the modified estimator.

To look at bias, (in the classical sense of design-based inference) the estimates were averaged over all 205 replicates. Averages for each of the 15 local areas, for each estimation method are presented in Figure 1. The population rates are plotted as the "True Proportions". These rates are almost exactly the same as the average unbiased estimates, and for the most part, are not visible on the graph. This confirms the unbiased nature of the classical estimates.

The synthetic estimates do not vary much from local area to local area. As each local area rate is based on the same pooled, fixed parameter estimates, the only source of variability from

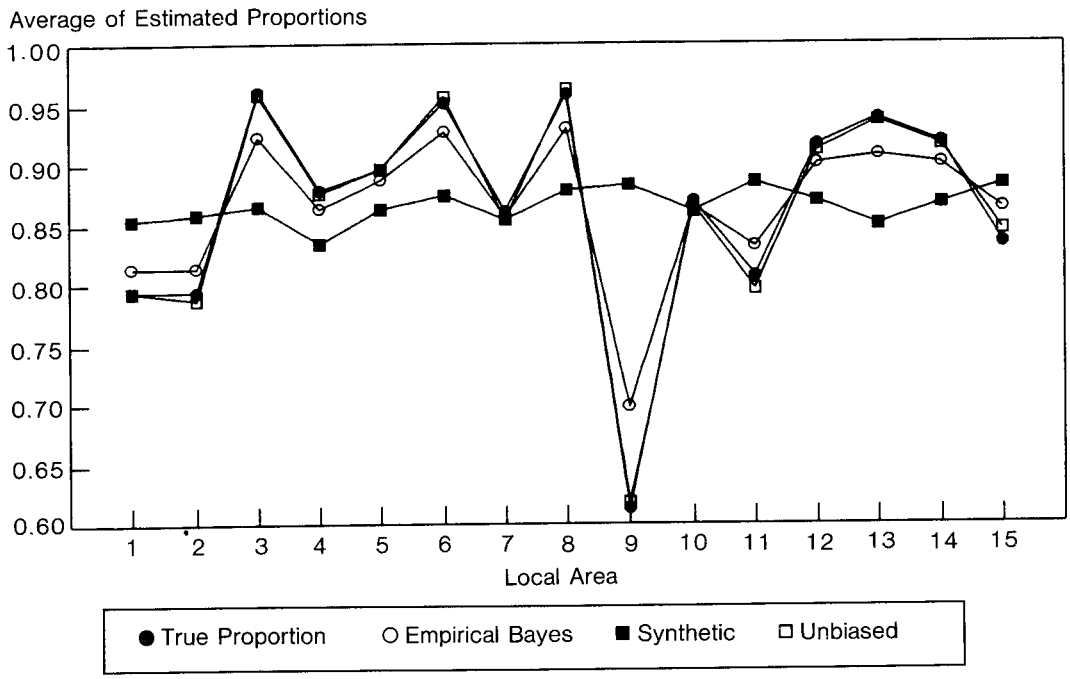

Figure 1. Averages of the estimated labour force participation rates for each of the three estimation methods plotted by local area

local area to local area is the small variability in the realized distributions of the predictor variables. The bias of this estimator can be large, as for example is the case for local area 9, where the synthetic method has a large positive bias. On the other hand, it should be noted that the synthetic method could not be expected to perform very well where there is little variability between the local area distributions of predictor variables.

The averages of the proposed estimates are in between the two extremes of the unbiased and synthetic estimates. They are biased, again in the classical sense, but their biases are smaller than those of the fixed effects model synthetic estimators.

Empirical Root Mean Square Errors (RMSE) were also calculated for each of the three estimators. These are presented in Figure 2. This plot demonstrates graphically where the synthetic estimator performs well and where it performs poorly. For local areas 7 and 10, where the local area effect is close to zero, the expected value of the synthetic estimator is very close to the population proportion. In these areas, the synthetic estimator has by far the smallest RMSE. By pooling data from the whole sample, it obtains a small sampling variance. On the other hand, in local area 9 where the local area effect is quite large, the associated RMSE for the synthetic estimator is also very large, due to its large bias. The modified empirical Bayes estimator obtains most of the reduction in RMSE that results from pooling the data across local areas, without suffering from the large bias associated with the synthetic estimator in those areas with large local area effects. In all but two cases, the modified empirical Bayes estimator achieves a smaller RMSE than the unbiased estimator. For local area 3, the RMSE's for the two estimators are about the same, and for local area 9, with a large local area effect, that of the modified empirical Bayes estimator is somewhat larger than that of the unbiased estimator. In short, the modified empirical Bayes estimator is sometimes the best of the three and never the worst.

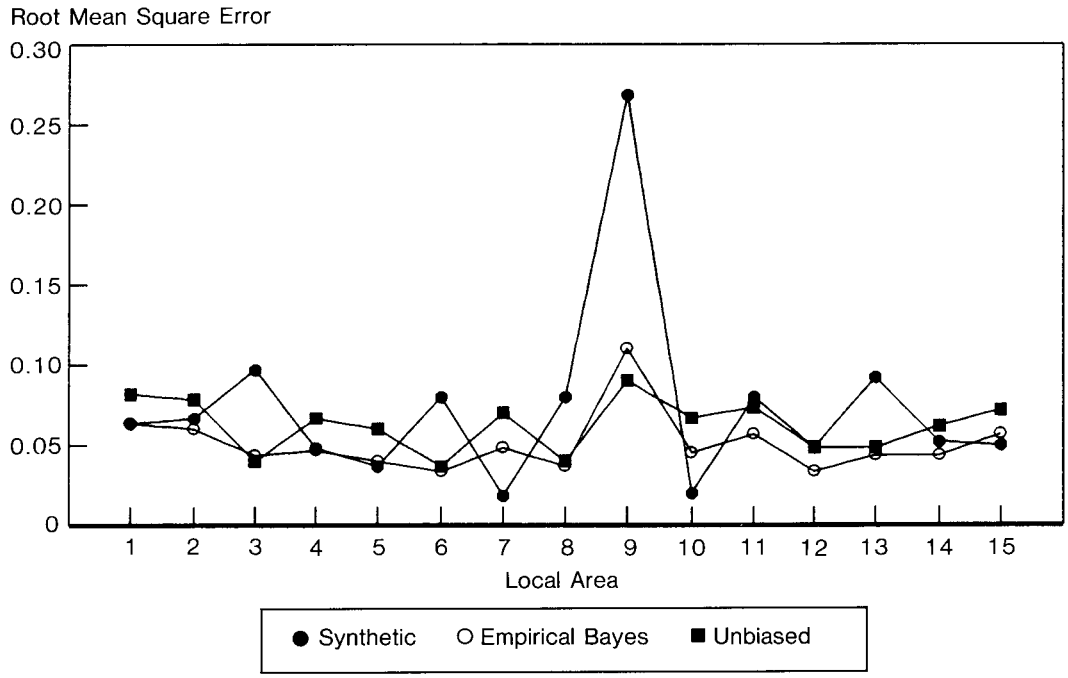

Figure 2. Empirical Root Mean Square Errors associated with each of the three estimation techniques plotted by local area

One of the principal shortcomings of the usual, fixed effects synthetic estimators is the difficulty in obtaining useful measures of associated accuracy. One can only obtain measures of sampling variances. Measures of bias which reflect model inadequacies are not available. For unbiased estimates, on the other hand, the usual estimates of sampling variability are also mean square error estimates as there is no bias. For empirical Bayes estimates, measures of uncertainty are available from the posterior covariance matrix of the parameters. These posterior variances reflect sampling variability as well as the "bias" which comes from simple fixed effects model inadequacies. This latter source of uncertainty is captured via the variability in the local area effects parameters.

The usefulness of these measures of uncertainty are compared graphically in Figure 3. The vertical axis corresponds to the empirical root mean square error (RMSE) which is obtained by comparing the individual replicate estimates with the known population proportions for each local area. The horizontal axis corresponds to the "reported RMSE". For the classical unbiased estimates, these are merely the sampling standard deviations for simple random sampling. For the synthetic estimates, they are also sampling standard deviations, corrected for the cluster sampling. The "reported RMSE" for empirical Bayes estimates are the square roots of the posterior variances of the estimated proportions which were obtained using the methods described in Section 3.2 above.

Note that the points corresponding to the unbiased estimates lie along a line indicating that the reported RMSE's are very close to the empirical RMSE's. This is as expected since there is no bias in these, so the reported RMSE's and the empirical RMSE's are merely sampling standard deviations. As opposed to this, the points corresponding to the synthetic estimates are in a cluster above 0.015 to 0.020 on the horizontal axes. For these estimates,

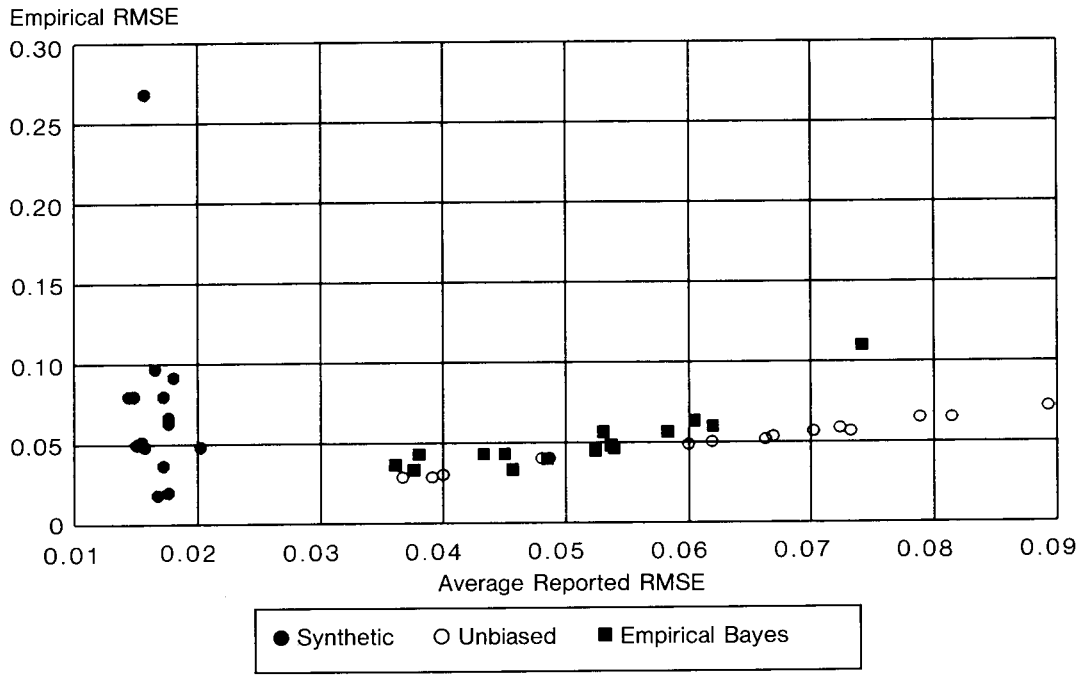

Figure 3. Empirical Mean Square Error vs "Reported Mean Square Errors" for each of the three estimation techniques

the "Reported RMSE's" are estimates of sampling standard deviations, which for these pooled estimates are all quite small. However, the empirical RMSE's for these estimates are quite a different story. They range from 0.015 to 0.100, with one outlier in excess of 0.250 (local area 9). Sampling variances alone are not sufficient to describe the uncertainty associated with the estimates.

The case for the modified empirical Bayes estimators is again in between these two extremes. However, with respect to the relationship between reported RMSE and empirical RMSE it is much closer to the corresponding relationship for the unbiased estimators. With the exception of the point associated with local area 9, the average reported RMSE's are very close to the corresponding empirical RMSE's.

5. CONCLUSIONS

In the simple simulation of a two-stage sample where PSU's correspond to local areas, the modified empirical Bayes estimators have been shown to be superior, overall to two standard alternatives. These have been evaluated in three ways, design-bias, root mean square error, and validity of estimable measures of uncertainty. The classical estimator is shown to be superior in terms of design-bias, as expected since it is design unbiased. In addition, valid estimates of RMSE's are available using standard techniques. However, these estimators suffer from large RMSE's due to the fact that they are formed from limited amounts of data. Indeed, unlike the other two alternatives, no estimates can be formed at all for local areas not in the sample.

At the other extreme, the synthetic estimator is far more stable than either of its competitors. Since all estimates are based on data from the whole sample, associated sampling variances are much smaller than those of the other two estimators. On the other hand, this estimator is unable to adjust for local areas which are quite different from the rest. This is the case, even when data are available in the sample that would indicate such a difference. As important, estimates of uncertainty in the form of sampling standard deviations for this estimator are particularly misleading since they are unable to account for departures from the fixed effects model.

As a compromise between these two estimators, the modified empirical Bayes estimator performs well on all three assessments. By using the data from the specific local areas to the extent it is reliable, this estimator avoids the large biases associated with the synthetic estimator. On the other hand, by pooling information from the whole sample, it has smaller sampling variances than the unbiased estimator, and generally smaller RMSE's. Finally, posterior variances are available as useful measures of uncertainty.

Several tasks remain in the investigation of the proposed estimators. First, the effect of using true empirical Bayes estimators instead of modified ones must be assessed. Some guidelines for minimum number of sampling units for valid empirical Bayes inference are required. True empirical Bayes estimates employ estimated prior variances and methods which account for this additional uncertainty are required. For example, the bootstrap techniques investigated by Laird and Louis (1987) could be used. Second, the estimation techniques need to be generalized to handle three and more stages of sampling. While the theoretical extension is trivial, the computational implications are not. Finally, these techniques must be applied to real data before recommending their adoption as a standard alternative for local area estimation.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful suggestions of an associate editor and a referee, and the financial support of NSERC of Canada, and Concordia University.

REFERENCES

BATTESE, G. E., HARTER, R. M., and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36.

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

DEMING, W. E., and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of The Royal Statistical Society*, Ser. B, 39, 1-38.

DEMPSTER, A. P., RUBIN, D. B., and TSUTAKAWA, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.

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

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography* 10, 137-159.

ERICKSEN, E. P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

ERICKSEN, E. P. (1980). Can regression be used to estimate local undercount adjustments? *Proceedings of the Conference on Census Undercount*, 55-61.

EFRON, B., and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.

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

GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 33-36.

GONZALEZ, M. E., and HOZA, C. (1976). Small area estimation of unemployment. *Proceedings of the Section on Social Statistics, American Statistical Association*, 437-443.

GONZALEZ, M. E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

GONZALEZ, M. E., and WAKSBERG, J. L. (1975). Estimation of the error of synthetic estimates. Unpublished paper presented at the first meeting of the International Association of Survey Statisticians, Vienna.

HABERMAN, S. J. (1978). *Analysis of Qualitative Data Volume 1: Introductory Topics.* New York: Academic Press.

HOLT, D., SMITH, T. M. F., and TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

JAMES, W., and STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 361-379.

LAAKE, P. (1979). A predictive approach to subdomain estimation in finite populations. *Journal of the American Statistical Association*, 74, 355-358.

LAIRD, N. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.

LAIRD, N., and LOUIS, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association*, 82, 739-757.

LEONARD, K. J. (1988). Credit scoring via linear logistic models with random parameters. Ph. D. Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montréal.

LEVY, P. S., (1971). The use of mortality data in evaluating synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 323-331.

LEVY, P. S., and FRENCH, D. K. (1978). Estimation of health characteristics. *Vital and Health Statistics*, Ser. 2, No. 75, NCHS, Washington, DC.

MADOW, W. G., and HANSEN, M. H. (1975) On statistical models and estimation in sample surveys. Contributed Papers, 40th Session of the International Statistical Institute, Warsaw, Poland, 554-557.

MIAO, L. L. (1977). An empirical Bayes approach to analysis of inter-area variation, Ph. D. Dissertation, Department of Statistics, Harvard University.

MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-54.

O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.

PLATEK, R., and SINGH, M. P. (1986). *Small Area Statistics--An International Symposium '85* (Contributed Papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University-University of Ottawa, Canada.

PURCELL, N. J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

PURCELL, N. J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.

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

ROBBINS, H. I. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium*. Berkeley: University of California Press, 157-164.

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

ROYALL, R. M. (1973). Discussion of papers by Gonzalez and Ericksen. *Proceedings of the Section on Social Statistics, American Statistical Association*, 42-43.

SÄRNDAL, C. E. (1984). Design consistent versus model dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W. L. (1979). A composite estimator for small area statistics. In *Synthetic Estimates for Small Areas* (NIDA Research Monograph 24), edited by J. Steinberg. Rockville, MD: National Institute on Drug Abuse, 36-53.

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.

TOMBERLIN, T. J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.

U.S. National Center for Public Health Statistics (1968). *Synthetic State Estimates of Disability*, PHS Publication No. 1759.

WEISBERG, H. I., TOMBERLIN, T. J., and CHATTERJEE, S. (1984). Predicting insurance losses under a cross-classification: a comparison of alternative approaches. *Journal of Business and Economic Statistics*, 2, 170-178.

WONG, G. Y., and MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.