2000

eremu txikietako zeharkako zenbatespenetarako estatistika-metodologia

metodología estadística para estimaciones indirectas en pequeñas áreas

JON N. K. RAO

39

2000

statistical methodology for indirect estimations in small areas

eremu txikietako zeharkako zenbatespenetarako estatistika-metodologia

metodología estadística para estimaciones indirectas en pequeñas áreas

JON N. K. RAO

39

Eustat

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT-Euskal Estatistika Erakundeak:

- Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
- Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
- Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai horri buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2000ko Otsaila

LOURDES LLORENS ABANDO
EUSTATeko Zuzendari Nagusia

PRESENTATION

In promoting the International Statistical Seminars, EUSTAT-The Basque Statistics Institute wishes to achieve several aims:

- Encourage the collaboration with the universities, especially with their statistical departments.
- Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
- Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

As a complementary activity and in order to reach as many interested people and institutions as possible, it has been decided to publish the papers of these courses, always respecting the original language of the author, to contribute in this way towards the growth of knowledge concerning this subject in our country.

Vitoria-Gasteiz, February 2000

LOURDES LLORENS ABANDO
General Director of EUSTAT

PRESENTACION

Al promover los Seminarios Internacionales de Estadística, el EUSTAT-Instituto Vasco de Estadística pretende cubrir varios objetivos:

- Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
- Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
- Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, Febrero 2000

LOURDES LLORENS ABANDO
Directora General de EUSTAT

BIOGRAFI OHARRAK

Jon N. K. Rao Estatistikako irakasle da Ottawa-ko (Kanada) Carleton Unibertsitatean. Kanadako Estatistika Erakundeko aholkulari eta Metodologiako Aholku Batzordeko kide ere bada. Panel on Estimates of Poverty for Small Geographic Areas elkarteko kide ere bada, Estatu Batuetako Estatistika Batzorde Nazionalean. Royal Society of Canada-ko, American Statistical Association-eko eta Institute of Mathematical Statistics-eko kide da.

Statistical Society of Canada-ren 1994ko Urrezko Domina jaso zuen ikerketa-lorpen aipagarriengatik. Mahaiburu izan zen Nazioarteko Estatistika Erakundearen bi urtez behingo bileran, Helsinkin (Finlandia) 1999an. Annual Morris Hansen-en 8. hitzaldia eman zuen, 1998ko urrian.

Haren uneko ikerguneen artean, honako hauek daude: eremu txikien zenbatespenari buruzko teoria eta metodoak, erantzun-ezeko datuen egozpenaren gaineko bariantza-zenbatespenak, esparru bikoitzeko laginak, inkestetako birlaginketa-metodoak eta laginketa-datuen analisiak.

Eremu txikietako zenbatespenari buruzko ikastaroak eman ditu Washington, DC, Riga eta Erroman. Wiley liburu baten koeditorea da: Small Area Statistics, 1987.

BIOGRAPHICAL SKETCH

Jon N.K. Rao is Professor of Statistics at Carleton University, Ottawa, Canada. He is also a Consultant to Statistics Canada and a Member of the Methodology Advisory Committee. He is also a Member of the Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, USA. He is a Fellow of the Royal Society of Canada, American Statistical Association and Institute of Mathematical Statistics.

He received the 1994 Gold Medal of the Statistical Society of Canada for outstanding research achievements. He served as Program Chair for the International Statistical Institute Biannual meetings, Helsinki, Finland, 1999. He delivered the 8th Annual Morris Hansen Lecture, October 1998.

His current research interests include small area estimation theory and methods, variance estimation under imputation for missing survey data, dual frame surveys, resampling methods in surveys and analysis of survey data.

He has given workshops on small area estimation in Washington, DC, Riga and Rome. He is a Coeditor of a Wiley book: Small Area Statistics, 1987.

NOTAS BIOGRAFICAS

Jon N.K. Rao es Catedrático de Estadística en la Universidad de Carleton, Ottawa, Canadá. También es Consultor de Statistics Canada y Miembro del Comité Consultivo de Metodología. A su vez, es Miembro del Panel on Estimates of Poverty for Small Geographic Areas, Comité Nacional de Estadística, USA. Es Miembro de la Royal Society of Canada, del American Statistical Association y del Institute of Mathematical Statistics.

Recibió la Medalla de Oro 1994 de la Statistical Society de Canadá por sus destacados logros en investigación. Presidió los encuentros bianuales del Instituto Estadístico Internacional, Helsinki, Finlandia, 1999. Pronunció la conferencia del 8° Annual Morris Hansen, Octubre 1998.

Actualmente, su interés en la investigación se centra en la teoría y métodos de la estimación en pequeñas áreas, estimaciones de varianza en supuestos de imputación de no respuesta en encuestas, métodos de remuestreo en encuestas y análisis estadísticos de encuestas.

Ha dado cursos sobre estimación en pequeñas áreas en Washington, DC, Riga y Roma. Es el coeditor de un libro Wiley: Small Area Statistics, 1987.

CONTENTS

SMALL AREA ESTIMATION : METHODS AND APPLICATIONS

J.N.K. Rao
Carleton University, Ottawa, Canada

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area statistics. Traditional area specific direct estimates do not provide adequate precision for small areas because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow data from related areas to increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related areas through supplementary data such as recent census counts and administrative records. The main purpose of this short monograph is to provide an introduction to indirect estimation methods, traditional as well as model-based. Methods for measuring the variability of the estimates are also presented as well as techniques for model validation. A basic area-level linear model is used to illustrate the methods, and then various extensions are presented, including binary response data through generalized linear mixed models and time series data through linear models that combine cross-sectional and time series features. Several recent applications of small area estimation are also given, including those in U.S. Federal Programs. Design issues that have impact on small area estimation are discussed.

The monograph is divided into five parts. Part 1 presents terminology, notation, direct estimation methods and design issues. Demographic methods and tradition indirect estimation methods based on implicit models are studied in Part 2. Basic area level and unit level models are introduced in Part 3 and model-based estimators, methods for measuring their variablility, model diagnostics and several applications are then presented. Various extensions of the basic model are studied in Part 4 as well as a variety of applications. Finally, cautions that need to be exercised in using indirect estimates and some recommendations are presented in Part 5.

1 TRADITIONAL DIRECT ESTIMATORS

1.1 Introduction

A geographical area or more generally any subpopulation (domain) is regarded as a "small area" if the number of domain-specific sample observations is small. Typically, domain sample size tends to increase with the size of the domain, but this is not always true. For example, the U.S. Third National Health and National Examination Survey (NHANES III), was designed to provide reliable estimates for domains classified by race-ethnicity and age. In this example, states may be regarded as small areas because the area-specific sample sizes are small (even zero) for many states.

Small area estimates are needed in formulating policies and programs, in allocation of government funds, in regional programs and so on. Demand for reliable small area statistics from both public and private sectors has grown rapidly in recent years.

Censuses usually provide detailed information on a limited number of items once in five or ten years. Administrative records can provide data more frequently but suffer from coverage problems. On the other hand, sample surveys can provide information on wide-ranging topics at frequent intervals of time and at reduced cost. Data obtained from sample surveys can be used to derive reliable direct estimates for large areas (with large samples), but sample sizes in small areas are rarely large enough for area-specific direct estimators to provide adequate precision for small areas. This makes it necessary to borrow data from related areas to find indirect estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and administrative records. The focus of this monograph is on indirect estimators that include (1) estimators based on implicit models and (2) model-based estimators. Group (1) contains synthetic and composite estimators, while group (2) covers empirical Bayes (EB) and hierarchical Bayes (HB) estimators. Demographers have long been using a variety of methods for small area estimation of population and other characteristics of interest in post-censal years. These methods use implicit models and utilize current data from administrative registers in conjunction with related data from the latest census. In a recent application, model-based estimates of poor school-age children are produced at the county and school district level (National Research Council, 1998). Using these estimates, the U.S. Department of Education allocates over $7 billion of federal funds to counties and school districts.

Ghosh and Rao (1994) and Rao (1999) provided comprehensive overviews of methods for small area estimation. Singh, Gambino and Mantel (1994) discussed design issues that have an impact on small area statistics. Schaible (1996) provided an excellent account of the use of indirect estimators in U.S. Federal Programs.

Prompted by the growing demand for reliable small area statistics, several symposia and workshops have been organized in recent years. Some of the symposia proceedings have also been published, including the following: (1) National Institute on Drug Abuse, Princeton Conference (National Institute on Drug Abuse, 1979); (2) International Symposium on Small Area Statistics, Ottawa (see Platek et al. (1987) for the invited papers and Platek and Singh (1986) for the contributed papers presented at the Symposium); (3) International Scientific Conference on Small Area Statistics and Survey Designs, Warsawa, 1992 (Kalton, Kordos and Platek, 1993); (4) International Association of Survey Statisticians Satellite Conference on Small Area Estimation, Riga, 1999. The published proceedings listed above provide an excellent collection of both theoretical and applied papers.

1.2 Terminology, notation

Terms used to denote a domain with small sample size include small area, small domain, local area, subdomain, small subgroup, subprovince and minor domain. Examples of a small geographical area include: county, school district, municipality, census tract. Even a state can be classified as a small area if the state sample size is small. For example, with a self-weighting sample of $n = 10{,}000$ persons in U.S.A., the expected sample size for Wyoming is only 18. Examples of a small subpopulation include: age-race-sex group within a large geographic area, business firms belongs to a census division by industry group.

A direct estimator is based on data obtained only from the sample units in the area of interest, i.e., it is area-specific. On the other hand, indirect estimators (also called nontraditional, model-based) borrow strength from sample observations of related areas to increase the effective sample size.

We use the following notation:
y = characteristic of interest, $\mathbf{x}$ = vector of auxiliary variables
Y = population total, N = population size
s = overall sample, n = size of s
Y_i = i-th area total, N_i = i-th area size
s_i = i-th area specific sample, n_i = size of s_i
$\bar{Y}_i = Y_i / N_i$ = i-th area mean, P_i = i-th area proportion ($y = 1$ or 0)

Further, w_k denotes the basic design weight attached to the k-th sample unit ($k \in s$), where s is a probability sample. For example, $w_k = N/n$ for simple random sampling. The weight w_k may be interpreted as the number of units in the population represented by the sample unit k.

1.3 Direct estimators

In practice, the weights w_k are adjusted for post-stratification in order to increase the precision of the estimator and to ensure consistency with known totals of auxiliary variables $\mathbf{x} = (x_1,...,x_p)^T$, where the superscript T denotes the transpose of a vector. The adjusted weights are given by $w_k^* = w_k g_k$, where g_k is the adjustment factor (also called g-weight), and the corresponding estimator of the total Y is given by

$$\hat{Y} = \sum_{k \in s} w_k^* y_k ; \tag{1.1}$$

that is, $\hat{Y}$ is the sum of weighted values $w_k^* y_k$ over the units k in the sample s.

Complete post-stratification

Suppose ${}_jN$ denotes the known j-th poststratum (cell) count; for example, projected census age-sex counts obtained from demographic methods. Then letting ${}_js$ be the sample in the j-th poststratum, g_k is given by

$$g_k = {}_jN / {}_j\hat{N}, \quad k \in {}_js \tag{1.2}$$

where ${}_j\hat{N}$ is the sum of weights w_k associated with units k belonging to ${}_js$. The estimator $\hat{Y}$ reduces to ${}_jN$ when y_k is the indicator variable taking the value 1 if the unit k belongs to post-stratum j, and 0 otherwise; that is, $\hat{Y}$ ensures consistency with known totals ${}_jN$.
For the special case of complete post-stratification with g-weights given by (1.2), the estimator (1.1) reduces to

$$\hat{Y} = \sum_j ({}_jN / {}_j\hat{N}) {}_j\hat{Y}, \tag{1.3}$$

where ${}_j\hat{Y}$ is the sum of weighted values $w_k y_k$ of units k belonging to ${}_js$.

Incomplete post-stratification

The numbers of cells increases rapidly if the post-strata are constructed by cross-classifying several variables. To avoid this problem, it is a common practice to use only

marginal counts for each variable. The weights w_k^* are obtained by minimizing a chi-squared distance $\sum_{k\in s}\left(w_k^* - w_k\right)^2 / w_k$ subject to $\sum_{k\in s} w_k^* x_{kl} = X_l, \quad l = 1,..., p,$ where X_l denotes a marginal count. The solution is given by $w_k^* = w_k g_k$ with

$$g_k = 1 + \mathbf{x}_k^T \hat{\mathbf{T}}^{-1}(\mathbf{X} - \sum_s w_k \mathbf{x}_k), \tag{1.4}$$

where $\hat{\mathbf{T}} = \sum_s w_k \mathbf{x}_k \mathbf{x}_k^T$ and $\mathbf{X} = (\mathbf{X}_1,..., \mathbf{X}_p)^T$ (see, for example, Deville and Sarndal, 1992).

Area-specific estimators

A direct estimator of small area total Y_i is given by

$$\hat{Y}_i = \sum_{k\in s_i} w_k^* y_k . \tag{1.5}$$

Similarly, N_i and $\overline{Y}_i$ are estimated by

$$\hat{N}_i = \sum_{k\in s_i} w_k^*, \qquad \hat{\overline{Y}}_i = \hat{Y}_i / \hat{N}_i .$$

The estimator of proportion P_i is a special case of $\hat{\overline{Y}}_i$ with binary y. Note that $\hat{Y}_i$ ensures additivity, i.e., $\sum_i \hat{Y}_i = \hat{Y}$. Note that $\hat{Y}_i$ cannot be used if the sample s_i is empty.

Mean squared error

Mean squared error (MSE) is commonly used to measure the accuracy of the estimator $\hat{Y}_i$. It is given by

$$MSE(\hat{Y}_i) = V(\hat{Y}_i) + \left[B(\hat{Y}_i)\right]^2,$$

where V denotes variance and B denotes the bias of the estimator. The bias of $\hat{Y}_i$ is negligible for large sample sizes n, but the variance of $\hat{Y}_i$ is of the order n_i^{-1} so that the precision of $\hat{Y}_i$ is not adequate if the small area sample size n_i is small.

Relative root mean square error (*RRMSE*) or coefficient of variation (*CV*) of an estimator $\hat{Y}_i$ is defined as the square root of *MSE* of $\hat{Y}_i$ divided by the total Y_i and it is a measure of reliability of the estimator. If the bias of $\hat{Y}_i$ is zero or negligible, then *RRMSE* = *RSE* = relative standard error which is the square root of variance (or standard error (*SE*)) divided by the total Y_i. Advantage of *RRMSE* (or *CV*) is that it does not depend on the unit of measurement. For small areas, *RRMSE* less than 20-25% is often regarded as adequate. In practice, *RRMSE* is estimated from the sample.

1.4 Design issues

"Optimal" design of samples for use with direct estimators of large area totals received a lot of attention over the past 50 years or so, but survey design issues that have an impact on small area statistics should also be considered. Singh et al. (1994) proposed several methods for use at the design stage to minimize the use of indirect small area estimators. These methods include (i) replacing clusters by using list frames, (ii) use of many strata to provide better sample size control at the small area level and (iii) compromise sample allocations. They presented an excellent illustration of compromise sample size allocations to satisfy reliability requirements at the provincial level as well as sub-provincial level. For the Canadian Labour Force Survey with a monthly sample of 59,000 households, optimizing at the provincial level yields a coefficient of variation (*CV*) for "unemployed" as high as 17.7% for some Unemployment Insurance (UI) regions. On the other hand, a two-step allocation with 42,000 households allocated at the first step to get reliable provincial estimates and the remaining 17,000 households allocated at the second step to produce best possible UI region estimates reduced the worst case of 17.7% *CV* for UI to 9.4% at the expense of a small increase in *CV* at the provincial and national levels: *CV* for Ontario increased from 2.8% to 3.4% and for Canada from 1.36% to 1.51%. Preventive measures, such as compromise sample allocations, should be taken at the design stage, whenever possible, to ensure adequate precision for domains like UI region. Marker (1999) discussed several other methods for use at the design stage, including the use of dual frames, combining data from rolling samples (Kish, 1990) and harmonization (or integration of surveys).

Preventive measures at the design stage may reduce the need for indirect estimators significantly, but for some small areas sample sizes may not be large enough for direct estimators to provide adequate precision even after taking such measures. As noted before, sometimes the survey is deliberately designed to oversample specific domains at the expense of small samples or even no samples in other domains (areas) of interest.

2 TRADITIONAL INDIRECT ESTIMATORS

2.1 Demographic methods

As pointed out earlier, demographers have long been using a variety of methods for local estimation of population and other characteristics of interest in post-censal years. These methods, called Symptomatic Accounting techniques (SAT), utilize current "symptomatic" data from administrative registers (such as the numbers of births and deaths) in conjunction with related data from the latest census. We consider the following SAT methods : Vital Rates (VR), Components and Regression Symptomatic.

VR method

The VR method uses only birth and death data as symptomatic variables. Let $b_t\,(b_0)$ and $d_t\,(d_0)$ denote the number of births and deaths for the local area for current year t (census year 0), where b_t, d_t are obtained from administrative registers. The population p_t for the local area at year t is then estimated by

$$\hat{p}_t = \frac{1}{2}\left(\frac{b_t}{\hat{r}_{1t}} + \frac{d_t}{\hat{r}_{2t}}\right), \tag{2.1}$$

where $\hat{r}_{1t}, \hat{r}_{2t}$ are the estimates of crude birth and death rates, $r_{1t} = b_t/p_t$ and $r_{2t} = d_t/p_t$, for the current year t. The VR method assumes that the updating factors $\phi_1 = r_{1t}/r_{10}$ and $\phi_2 = r_{2t}/r_{20}$ are equal to the corresponding factors for a larger area containing the local area, i.e, $\phi_1 = R_{1t}/R_{10}$ and $\phi_2 = R_{2t}/R_{20}$, where R refers to the larger area for which the population estimate $\hat{P}_t$ is ascertained from official sources. It now follows that

$$\hat{r}_{1t} = \hat{\phi}_1 r_{10}\ , \hat{r}_{2t} = \hat{\phi}_2 r_{20} \tag{2.2}$$

with $\hat{\phi}_1 = \hat{R}_{1t}/R_{10}$, $\hat{\phi}_2 = \hat{R}_{2t}/R_{20}$, $\hat{R}_{1t} = B_t/\hat{P}_t$ and $\hat{R}_{2t} = D_t/\hat{P}_t$, where $B_t(D_t)$ is the number of large area births (deaths) for the current year t. The VR method is simple but the assumption on updating factors is often questionable.

Example : Govindarajulu (1999, ch.17) considered the estimation of population for a small county in Kentucky, USA. We have $b_t = 400$, $d_t = 350$ and from the 1990 census $R_{10} = 2\%$, $R_{20} = 1.8\%$. He assumed $R_{10} = r_{10}$, $R_{20} = r_{20}$. The state rates are $\hat{R}_{1t} = 2.1\%$ and $\hat{R}_{2t} = 1.9\%$ so that $\hat{\phi}_1 = 2.1/2$ and $\hat{\phi}_2 = 1.9/2$. Also, from (2.2)

$$100\hat{r}_{1t} = 2(2.1/2) = 2.1, \qquad 100\hat{r}_{2t} = 1.8(1.9/1.8) = 1.9.$$

It now follows from (2.1) that

$$\hat{p}_t = \frac{1}{2}\left(\frac{400}{0.021} + \frac{350}{0.019}\right) \approx 18{,}735.$$

Components method

The components method takes account of net migration. Denoting $b_{0,t}, d_{0,t}$ and $m_{0,t}$ as the number of births, deaths and net migration in local area during the period [0, t], we have

$$p_t = p_0 + b_{0,t} - d_{0,t} + m_{0,t}, \tag{2.3}$$

where

$$m_{0,t} = i_{0,t} - e_{0,t} + n_{0,t}$$

with $i_{0,t}, e_{0,t}$ and $n_{0,t}$ denoting immigration, emigration and net interstate migration. Administrative records provide $e_{0,t}, i_{0,t}$ and $n_{0,t}$ where $i_{0,t}$ is benchmarked to known immigration figures at state level. The components method is basically designed for population estimation of local areas.

Regression symptomatic procedures

Regression symptomatic procedures use multiple linear regression to estimate local area populations utilizing symptomatic variables as independent variables in the regression equation. Three such methods are the ratio correlation, difference correlation, and sample regression (SR) methods. We focus on the sample regression method (Ericksen, 1974) since it uses the current regression equation unlike the other two methods based on the past two consecutive census years. Let

$Y_i = (p_{it}/P_t)/(p_{i0}/P_0)$ = change in the population proportion for local area i,

$x_{ij} = (s_{ijt}/S_{jt})/(s_{ij0}/S_{j0})$ = change in the j-th symptomatic variable s_j for local area i,

where $S_{jt}(S_{j0})$ are the values for the larger area containing the local area i. The predictor variables x_{ij} are obtained from administrative registers $(j = 1,..., p)$.

The SR method assumes that Y_i is linearly related to $x_{i1},..., x_{ip}$. Survey estimates $\hat{Y}_i$ of Y_i are obtained for k out of m local areas under consideration, and linear regression is

then fitted to the data $(\hat{Y}_i, x_{i1},..., x_{ip}; i = 1,..., k)$ to get the regression coefficients $\hat{\beta}_0, \hat{\beta}_1,..., \hat{\beta}_p$. The sample regression estimators of Y_i are obtained as

$$\tilde{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}, \quad i = 1,..., m \tag{2.4}$$

The current counts p_{it} are then estimated as

$$\tilde{p}_{it} = \tilde{Y}_i (p_{i0}/P_0)\hat{P}_t, \quad i = 1,..., m. \tag{2.5}$$

Note that the sample regression estimator (2.4) does not make use of the direct estimator $\hat{Y}_i$ for the sampled areas, unlike the model-based estimators studied in Part 3. As a result, it is less efficient than the model-based estimator for the sampled areas.

2.2 Synthetic estimators

An estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small area have the same characteristics as the large area (Gonzalez, 1973). U.S. National Center for Health Statistics (1968) pioneered in the use of synthetic estimation for developing state estimates of disability and other health characterisics from the National Health Interview Survey (NHS); sample sizes in many States were too small to provide reliable State estimates.

We give several examples of synthetic estimation to illustrate the underlying implicit models and the methods.

Example 1 : No sampling

Suppose the total Y for a large region covering small areas i is known from administrative sources, as well as sizes N_i and N. Then a synthetic estimator of the small area total Y_i is

$$\hat{Y}_i(S) = (N_i/N)Y \tag{2.6}$$

Implicit model : $\overline{Y}_i = \overline{Y} = Y/N$ for all i.

The synthetic estimator (2.6) has no variance, but can be badly biased if the implicit model is not valid.

Example 2 :

Suppose P_i and P denote proportions in poverty for small area i and a large region covering area i, and suppose a reliable direct estimator $\hat{P}$ of P is available. Then a synthetic estimator of P_i is

$$\hat{P}_i(S) = \hat{P} \tag{2.7}$$

Implicit model : $P_i = P$ for all i.
This model assumes homogeneity of area proportions P_i.
The design bias of $\hat{P}_i(S)$ is

$$B\left[\hat{P}_i(S)\right] = P - P_i.$$

Suppose we want to estimate a total $Y_i = N_i \overline{Y}_i$ and N_i is known. If reliable direct estimators $\hat{Y}$ and $\hat{N}$ are available, then a synthetic estimator of Y_i is

$$\hat{Y}_i(S) = N_i(\hat{Y}/\hat{N}) \tag{2.8}$$

Implicit model : $\overline{Y}_i = \overline{Y}$ for all i.
This model assumes homogeneity of area means $\overline{Y}_i$.
The variances of synthetic estimators $\hat{P}_i(S)$ and $\hat{Y}_i(S)$ are small because they are based on the reliable direct estimators $\hat{P}, \hat{Y}$ and $\hat{N}$. But their biases can be large if the underlying implicit homogeneity model is not satisfied.

Example 3 :

A more realistic model assumes homogeneity only within post-strata j (say age-sex-race groups) for which domain specific post-strata counts ${}_jN_i$ are known :

Implicit model : ${}_j\overline{Y}_i = {}_j\overline{Y}$ for all i and each j.

This model assumes that the post-stratum small area means ${}_j\overline{Y}_i$ are homogeneous across i for each post-stratum j and equal to the regional post-stratum mean ${}_j\overline{Y}$.

If reliable direct estimates ${}_j\hat{Y}$ and ${}_j\hat{N}$ of the post-strata regional totals ${}_jY$ and ${}_jN$ are available, then

$$\hat{Y}_i(S) = \sum_j {}_jN_i\left({}_j\hat{Y}/{}_j\hat{N}\right). \tag{2.9}$$

The synthetic estimators (2.9) add up to the reliable direct estimator $\hat{Y}$ given by (1.3).
If $y = 0$ or 1, then a synthetic estimator of proportion P_i is

$$\hat{P}_i(S) = \left(\sum_j {}_jN_i\,{}_j\hat{P}\right)\Big/\left(\sum_j {}_jN_i\right), \tag{2.10}$$

where ${}_j\hat{P}$ is the direct estimator of the j-th poststratum regional proportion ${}_jP$.

Again the variance of the synthetic estimators $\hat{Y}_i(S)$ and $\hat{P}_i(S)$ is small because it is based on regional estimators ${}_j\hat{Y}\,{}_j\hat{P}$ and ${}_j\hat{N}$, but their bias may be significant if the underlying implicit homogeneity model is not valid.

Application

The synthetic estimator (2.10) was used to produce state estimates of proportions for certain health characteristics from the 1980 U.S. National Natality Survey (NNS) and the 1980 National Fetal Mortality Survey (NFMS). Twenty five post-strata (demographic cells) were formed according to mother's race (white and all others), mother's age group (6 groups) and live birth order (1, 1-2, 1-3, 2, 2+, 3, 3+, 4+).

In this application (Gonzalez et al, 1996), a state in U.S.A. is a small area. For example, suppose i denotes Pennsylvania, $y = 1$ if live birth is jaundiced and $y = 0$ otherwise. The national estimates of percent jaundiced, ${}_j\hat{P}$, were obtained from NSS for each of the 25 demographic cells j. The number of hospital births in each cell ${}_jN_i$ (obtained from the state vital registration data) is then multiplied by ${}_j\hat{P}$ and summed over the cells j to get the numerator of (2.10) as 33,806. Dividing 33,806 by the total hospital births $\sum_j {}_jN_i = 156{,}799$ in Pennsylvania, the synthetic state estimate of percent jaundiced live births in Pennsylvania, 1980 is given by (33,808/156,799)x100=21.6.

Evaluation

Gonzalez et al. (1996) evaluated the accuracy of the synthetic estimator (2.10) for selected health characteristics by comparing the NNS synthetic estimates to "true" state values P_i for five selected states in 1980. These five states covered a wide range in the annual number of births: 15,000 to 160,000. The true state values were obtained from the state vital registration system. The *MSE* of the synthetic estimator was estimated as (est. – true value)^2. Table 1 reports the estimated values of *RRMSE* for the direct and synthetic estimates for the following characteristics: Low birth, prenatal care, Apgar score. Standard errors (*SE*) of the NNS direct estimates were estimated using balanced repeated replication with 20 replicate half-samples (see Morganstein (1998) for an overview of replicating methods for estimating standard errors).

It is clear from Table 1 that the synthetic estimates performed better than the direct estimates, especially for the smaller states (e.g., Montana) with small numbers of sample cases. The values of *RRMSE* ranged from 0.14 (Pennsylvania) to 0.62 (Montana) for the direct estimator while those for the synthetic estimator ranged from 0.00 (Pennsylvania) to 0.24 (Indiana). The NCHS used maximum *RRMSE* of 25% as the standard for reliability of estimates, and most of the synthetic estimates met this criterion for reliability, unlike the direct estimates.

Table 1. RRMSE of Direct and Synthetic Estimates

Characteristic and State	True %	Direct est.		Syn. est.	
		Est.(%)	*RRMSE*(%)	Est.(%)	*RRMSE*(%)
Low birth:					
Pennsylvania	6.5	6.6	15	6.5	0
Indiana	6.3	6.8	22	6.5	3
Tennessee	8.0	8.5	23	7.2	10
Kansas	5.8	6.8	36	6.4	10
Montana	5.6	9.2	71	6.3	13
Prenatal care:					
Pennsylvania	3.9	4.3	21	4.3	10
Indiana	3.8	2.0	21	4.7	24
Tennessee	5.4	4.7	26	5.0	7
Kansas	3.4	2.1	35	4.5	32
Montana	3.7	3.0	62	4.3	16
Apgar score:					
Pennsylvania	7.9	7.7	14	9.4	19
Indiana	10.9	9.5	16	9.4	14
Tennessee	9.6	7.3	18	9.7	1
Kansas	11.1	12.3	25	9.4	15
Montana	11.6	12.9	40	9.4	19

MSE estimation

An estimator of *MSE* of the synthetic estimator $\hat{Y}_i(S)$ is given by

$$mse\left[\hat{Y}_i(S)\right] = \left[\hat{Y}_i(S) - \hat{Y}_i\right]^2 - v(\hat{Y}_i), \tag{2.11}$$

where $v(\hat{Y}_i)$ is a customary variance estimator of the direct estimator $\hat{Y}_i$; for example a variance estimator based on a replication method. The estimator (2.11) is approximately unbiased but can be higly unstable. As a result, it is common practice to take the average of the MSE estimators of the means $\hat{Y}_i(S)/N_i$ over areas i and then multiply this average by N_i^2. This average *MSE* estimator will be stable but it is not an area-specific measure of accuracy. Attempts have been made to provide area-specific measures of accuracy that are more stable (Marker, 1993). One such estimator of *MSE* assumes that the squared bias of the synthetic estimator is approximately equal to the average squared bias of the synthetic estimators over areas i.

Structure preserving estimation (SPREE)

We now describe the method of Structure Preserving Estimation (SPREE) for categorical variables of interest y and categorical post-strata (auxiliary) variables (Purcell and Kish, 1980). For example, the categories, a, of y denote employed/unemployed while the categories, b, of an auxiliary variable denote white/nonwhite. The small area cell counts $\{N_{iab}\}$ from the past census are assumed to be available and $\{M_{iab}\}$ denote the corresponding unknown current counts. Our interest is in estimating the small area counts $M_{ia+} = \sum_b M_{iab}$ utilizing $\{N_{iab}\}$ as well as reliable direct survey estimates $\{\hat{M}_{+ab}\}$ of $\{M_{+ab}\}$ and demographic projections $\{\hat{M}_{i++}\}$ of population counts $\{M_{i++}\}$, where + denotes summation over the subscript. We consider two cases : (1) use only $\{N_{iab}\}$ and $\{\hat{M}_{+ab}\}$; (2) use $\{N_{iab}\}$ and both $\{\hat{M}_{+ab}\}$ and $\{\hat{M}_{i++}\}$.

Case 1 :

SPREE adjusts $\{N_{iab}\}$ to confirm to the "allocation" structure $\{\hat{M}_{+ab}\}$ but preserving the "allocation" structure in $\{N_{iab}\}$ as much as possible. This is accomplished by minimizing a chisquared distance $\sum_{iab} N_{iab}^{-1}(x_{iab} - N_{iab})^2$ with respect to x_{iab} subject to the conditions $\sum_i x_{iab} = \hat{M}_{+ab}$. The resulting solution is given by $\tilde{M}_{iab} = (N_{iab}/N_{+ab})\hat{M}_{+ab}$ and the estimator of M_{ia+} is therefore obtained as

$$\tilde{M}_{ia+} = \sum_b \tilde{M}_{iab} = \sum_b (N_{iab}/N_{+ab})\tilde{M}_{+ab} \tag{2.12}$$

The estimators $\tilde{M}_{iab}$ are also called one-step "raking ratio" estimators because of the ratio form.

Case 2 :

SPREE adjusts $\{N_{iab}\}$ to confirm to the "allocation" structure $\{\hat{M}_{+ab}\}$ and $\{\hat{M}_{i++}\}$ but preserving the association structure in $\{N_{iab}\}$ as much as possible. In this case we do not have a closed-form solution $\{\tilde{M}_{iab}\}$ unlike in case 1. Two-step raking ratio estimators $\{M^*_{iab}\}$ are obtained through iterative cycles, each cycle consisting of two steps. In the first cycle, we use starting values $x^{(0)}_{iab} = N_{iab}$ and do one-step raking to agree with the column margins $\hat{M}_{+ab}$ as in case 1 to get $x^{(1)}_{iab}$. We then do one-step raking using $x^{(1)}_{iab}$ to agree with the row margins $\hat{M}_{i++}$ to get $x^{(2)}_{iab}$ which are used as starting values for the second cycle. Iteration of cycles is continued until some convergence criterion is met. SPREE estimator of M_{ia+} is given by $M^*_{ia+} = \sum_b M^*_{iab}$.

Evaluation

Purcell and Kish (1980) made an evaluation of one-step and two-step SPREE estimators by comparing them to true counts, obtained from Vital Statistics registration system. In this study, SPREE estimates of mortality due to each of four different causes (a) and for each state (i) in U.S.A. were calculated for five individual years ranging over the post-censal period 1960-70. Here the categories b denote 36 age-sex-race groups, $\{N_{iab}\}$ the 1960 census counts and $\{\hat{M}_{+ab} = M_{+ab}\}\{\hat{M}_{i++} = M_{i++}\}$ the known current counts.

Table 2 reports the percent absolute relative differences (*ARD*), where $ARD = |est. - true|/true$. It is clear from Table 2 that the two-step SPREE estimator performs significantly better than the one-step SPREE estimator (2.12). Thus it is important to incorporate through the allocation structure the maximum available current data into SPREE estimation.

Purcell and Kish (1980) also studied SPREE when the full association structure $\{N_{iab}\}$ is not available. For example if only $\{N_{i+b}\}$ are known, SPREE estimators of M_{ia+} is taken as $\sum_b (N_{i+b}/N_{i++})\hat{M}_{+ab}$, assuming proportionality across the categories a.

Table 2. Percent *ARD* of SPREE Estimates

Cause of death	Year	One-step	Two-step
Malignant neoplasms	1961	1.97	1.85
	1964	3.50	2.21
	1967	5.58	3.22
	1970	8.18	2.75
Major CVR diseases	1961	1.47	0.73
	1964	1.98	1.03
	1967	3.47	1.20
	1970	4.72	2.22
Suicides	1961	5.56	6.49
	1964	8.98	8.64
	1967	7.76	6.32
	1970	13.42	8.52
Total other	1961	1.92	1.39
	1964	3.28	2.20
	1967	4.89	3.36
	1970	6.65	3.85

Griffiths (1996) used the one-step SPREE estimator (2.12) to provide estimates for Congressional Districts (CD) in U.S.A. The survey counts $\{\hat{M}_{+ab}\}$ were obtained from the March CPS (Current Population Survey) which collects extensive information such as household income and health insurance coverage in addition to labour force characteristics. The CPS sample was not designed to provide reliable direct estimates at the CD level; the CD sample sizes tend to be too small for direct estimation with desired reliability.

2.3 Composite estimators

Synthetic estimators are simple to implement in practice. But their bias can be quite significant because they make too strong a use of information from other areas and as a result allow too little for local variation. A simple way to balance the potential bias of a synthetic estimator $\hat{Y}_i(S)$ against the instability of a direct estimator $\hat{Y}_i$ is to take a weighted average of the two estimators. This leads to a composite estimator of the form

$$\hat{Y}_i(C) = \phi_i \hat{Y}_i + (1-\phi_i)\hat{Y}_i(S) \tag{2.13}$$

for some suitably chosen weight ϕ_i in the range [0,1]. Optimal weights ϕ_i that minimise the *MSE* of the composite estimator can be obtained (Schaible, 1978), but the weights have to be estimated from the sample data. The estimated weights, however, can be very unstable as they involve the estimated *MSE* of $\hat{Y}_i(S)$, given by (2.11), which is highly unstable. To overcome this difficulty, Purcell and Kish (1979) used a common weight, ϕ, and then minimized the average *MSE* over small areas. This leads to an estimated weight $\hat{\phi}$ of the form

$$\hat{\phi} = 1 - \sum_i v(\hat{Y}_i) \Big/ \sum_i \left[\hat{Y}_i(S) - \hat{Y}_i\right]^2, \tag{2.14}$$

where $v(\hat{Y}_i)$ is the variance estimator of the direct estimator $\hat{Y}_i$. The common weight $\hat{\phi}$ will be stable but the use of a common weight may not be reasonable if the individual variance estimators $v(\hat{Y}_i)$ vary considerably.

Sample-size dependent estimators

Simple weights that depend only on the small area counts N_i and $\hat{N}_i$ have also been proposed (Drew, Singh and Chaudhry, 1982). The weights are given by

$$\phi_i(D) = \begin{cases} 1 & \text{if } \hat{N}_i \geq \delta N_i \\ \hat{N}_i / (\delta N_i) & \text{otherwise,} \end{cases} \tag{2.15}$$

where δ is a specified constant. The resulting composite estimator is called the sample-size dependent (SSD) estimator. In the special case of SRS with design-weights $w_k = N/n$ and no post-stratification, we have $\hat{N}_i = N(n_i/n)$ and $\phi_i(D) = 1$ if n_i is not less than the expected domain sample size $n(N_i/N)$, assuming $\delta = 1$. Note that the SSD estimator reduces to the direct estimator if the realized domain sample size n_i is not less than the expected domain sample size even when the latter is very small. In the latter case, the SSD estimator may not be reliable because n_i could be very small. One could remedy this problem to some extent by choosing a suitable δ larger than 1, but the choice of δ is somewhat subjective. Moreover, the same weight $\phi_i(D)$ is used for all characteristics y regardless of their differences with respect to between area homogeneity.

The SSD estimator is used when the expected domain sample size is sufficiently large for the direct estimator to be reliable and the realized sample size may fall short of the expected sample size. The Canadian Labour Force Survey uses the SSD estimator with $\delta = 2/3$ to produce Census Division (CD) level estimates.

Evaluation

Falorsi, Falorsi and Russo (1994) compared the performance of direct estimator, synthetic estimator, SSD estimator with $\delta = 1$, and optimal composite estimator, through a Monte Carlo study in which the Italian Labour Force Survey design (stratified two-stage design) was simulated using data from the 1981 Italian census. The optimal weight ϕ_i was obtained from the census data. In this study, Health Service Areas (HSAs) are the small areas (unplanned domains) that cut across design strata. The study was confined to the 14 HSAs of the Friuli region and the sample design was based on the selection of 39 primary sampling units (PSUs) and 2,290 second stage units (SSUs); PSU is a municipality and SSU is a household. The variable of interest, y, was the number of unemployed.

The performance of the estimators was evaluated in terms of absolute relative bias (*ARB*) and *RRMSE*. The relative bias (*RB*) and *MSE* of an estimator of the total Y_i are given by

$$RB = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{est_r}{Y_i} - 1\right),$$

and

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(est_r - Y_i)^2,$$

where est_r is the value of the estimator for the r-th simulation run $(r = 1,...,R)$. Note the $RRMSE = (\text{square root of MSE})/Y_i$ and $ARB = |RB|$. They used R=400 simulation runs to calculate *RB* and *MSE* for each HSA and each estimator.

Table 3 reports the average values $\overline{ARB}$ and $\overline{RRMSE}$ of the estimators, where the average is over the 14 HSAs. It is clear from Table 3 that $\overline{ARB}$ values of the direct estimator and SSD estimator are negligible $(< 2.5\%)$ while these of the composite and synthetic estimators are somewhat larger. Synthetic estimator has the largest $\overline{ARB}$ $(\text{about } 9\%)$. In terms of $\overline{RRMSE}$, synthetic and composite estimator have the smallest values (about one half of the value for the direct estimator) followed by the SSD estimator with apprximately 30% higher value.

Falorsi et al. (1994) also examined area level values of *ARB* and *RRMSE*. Synthetic and composite estimators were found to be badly biased in small areas with low values of the ratio $p_1 = (\text{population of HSA})/(\text{population of the set of strata including the HSA})$ but exhibited low *RRMSE* compared to the other alternatives. Considering both bias and efficiency, they concluded that the SSD estimator is preferable over the other estimators. It

may be noted that the sampling rates were relatively high leading to large enough expected domain sample sizes, a case favourable to the SSD estimator.

Table 3. Average Absolute Relative Bias $\left(\overline{ARB}\%\right)$ and Average $RRMSE\left(\overline{RRMSE}\%\right)$ of Estimators

Estimator	$\overline{ARB}$ %	$\overline{RRMSE}$ %
Direct	1.75	42.08
Synthetic	8.97	23.80
Composite	6.00	23.57
SSD	2.39	31.08

James-Stein Estimators

James-Stein estimators belong to the class of composite estimators and the weights ϕ_i are obtained under certain assumptions. Suppose the population area means $\overline{Y}_i$ (or the proportions P_i) are the parameters of interest and $\theta_i = g\left(\overline{Y}_i\right)$ or $g\left(P_i\right)$ for a specified function $g(.)$. For example, $g(a) = a$ or $g(a) = \ln\{a/(1-a)\}$ which is used in the context of proportions P_i. Let $\hat{\theta}_i = g\left(\hat{\overline{Y}}_i\right)$ or $g\left(\hat{P}_i\right)$ be the direct survey estimator of θ_i and assume that the $\hat{\theta}_i$ are independent and normally distributed with means θ_i and a common known variance ψ. Then the James-Stein estimator of θ_i is given by

$$\hat{\theta}_i(JS) = \hat{\phi}_{JS}\hat{\theta}_i + \left(1-\hat{\phi}_{JS}\right)\theta_i^0, i = 1,...,m \tag{2.16}$$

where

$$1-\hat{\phi}_{JS} = \left[(m-2)\psi/S\right], \quad m \geq 3 \tag{2.17}$$

with $S = \sum_i \left(\hat{\theta}_i - \theta_i^0\right)^2$ and θ_i^0 is a guess (or prediction) of θ_i (James and Stein, 1961). The estimator (2.16) is also called a shrinkage estimator because it shrinks $\hat{\theta}_i$ towards the guess θ_i^0. The estimator of $\overline{Y}_i$ (or P_i) is given by $g^{-1}\left(\hat{\theta}_i\right)$.

If $\theta_i = \theta_i^0$ for all areas i, then the total *MSE* of the JS-estimators equals 2ψ whereas the total *MSE* of the direct estimators $\hat{\theta}_i$ equals $m\psi$. This result clearly demonstrates the potential for a drastic reduction of total *MSE*. A practical implication is that we can do much better in terms of *MSE* for the group of small areas as a whole by using J-S estimators instead of the direct estimators. The condition $\theta_i = \theta_i^0$ is of course unrealistic but in practice one might be able to get a prediction θ_i^0 close enough to θ_i by using area-level auxiliary variables $(z_1,...,z_p)$ linearly related to θ_i. For example, one could use the regression predictor $\theta_i^0 = \sum_j \hat{\beta}_j z_j$, where $\hat{\beta}_j's$ the ordinary least squares (OLS) estimator of regression coefficients $\beta_j's$ obtained by regressing $\hat{\theta}_i$ on $z_{i1},...,z_{ip}$, $i = 1,...,m$. In the absence of auxiliary information, θ_i^0 is taken as $\sum_i \hat{\theta}_i / m = \hat{\theta}_\bullet$. A fundamental result of James-Stein is that the J-S estimator uniformly dominates the direct estimator in terms of total *MSE* no matter what the choice of $\theta_i^0 s$ is, provided θ_i^0 is fixed. If θ_i^0 is not fixed, then the JS-estimator needs to be modified to account for the estimation. For example, if $\theta_i^0 = \hat{\theta}_\bullet$, then we change $m-2$ to $m-3$ in (2.17) and assume $m \geq 4$. It should be noted that the dominance property for fixed θ_i^0 is not necessarily true for the retransform of $\hat{\theta}_i(JS)$. That is, the estimators $g^{-1}\left[\hat{\theta}_i(JS)\right]$ of $\bar{Y}_i$ (or P_i) may not dominate the direct estimators $g^{-1}(\hat{\theta}_i) = \hat{\bar{Y}}_i$ (or $\hat{P}_i$) in terms of total *MSE*.

The J-S method may perform poorly in estimating those components θ_i with unusually large or small values of $\theta_i - \theta_i^0$. To reduce this undesirable effect, Efrom and Morris (1972) proposed a comprise J-S estimator which offers a compromise between the J-S estimator and the direct estimator. This estimator has both good ensemble and good individual properties, unlike the J-S estimator. It is obtained simply by restricting the amount by which $\hat{\theta}_i(JS)$ differs from $\hat{\theta}_i$ to a multiple of the standard error of $\hat{\theta}_i$:

$$\theta_i^*(JS) = \begin{cases} \hat{\theta}_i(JS) & \text{if } \hat{\theta}_i - c\sqrt{\psi} \leq \hat{\theta}_i(JS) \leq \hat{\theta}_i + c\sqrt{\psi} \\ \hat{\theta}_i - c\sqrt{\psi} & \text{if } \hat{\theta}_i(JS) < \hat{\theta}_i - c\sqrt{\psi} \\ \hat{\theta}_i + c\sqrt{\psi} & \text{if } \hat{\theta}_i(JS) > \hat{\theta}_i + c\sqrt{\psi} \end{cases} \tag{2.18}$$

where $c > 0$ is a suitably chosen constant. The choice $c = 1$ ensures that the *MSE* of individual estimator $\theta_i^*(JS)$ never exceeds twice the variance of the direct estimator, which

retaining more than 80 percent of the reduction in total *MSE* of J-S estimators over the direct estimators. The compromise J-S estimator of $\bar{Y}_i$(or P_i) is taken as $g^{-1}\left[\theta_i^*(JS)\right]$.

Example:

Efron (1975) gave an amusing example of batting averages of major league baseball players to illustrate the superiority of J-S estimators over the direct estimators. Table 4 gives the batting averages of $m = 18$ players after their first 45 times at bat during the 1970 season. These estimates are taken as the direct estimates $\hat{\theta}_i = \hat{P}_i$. The J-S estimates (2.16) were calculated using $\theta_i^0 = \sum_i \hat{P}_i / 18 = 0.265 = \hat{P}_\bullet$ and $\psi = \hat{P}_\bullet(1-\hat{P}_\bullet)/45 = 0.0043$, the binomial variance. The compromise J-S estimates (2.8) were calculated using $c = 1$. To compare the total accuracy of the estimates, the batting average for each player *i* during the reminder of the season (about 370 more at bats on average) was taken as the true $\theta_i = P_i$. Table 4 also reports the values of J-S and compromise J-S estimates.

Since the true values θ_i are assumed to be known, we can compare the relative accuracies using the ratios $R_1 = \sum_i (\hat{P}_i - P_i)^2 \Big/ \sum_i (\hat{P}_i(JS) - P_i)^2$ and $R_2 = \sum_i (\hat{P}_i - P_i)^2 \Big/ \sum_i (P_i^*(JS) - P_i)^2$. We have $R_1 = 3.50$ so that the J-S estimates outperform the direct estimates by a factor of 3.5. Also, $R_2 = 14.67$ so that the compromise J-S estimates perform even better than the J-S estimates in this example. It may be noted that the compromise J-S estimator protects player 1's proportion $\hat{P}_1 = 0.400$ from overshrinking toward the common proportion $\hat{P} = 0.265$.

MSE estimation

It is possible to obtain an estimator of *MSE* of the J-S estimator similar to (2.11) for the synthetic estimator. But it is also highly unstable and one needs to resort to some sort of averaging of *MSE* estimators as in the case of synthetic estimators. Model-based methods in Part 4 do not suffer from these limitations. Morever, they permit validation of models from sample data and extensions to handle complex situations.

Benchmarking

Composite estimators and model-based estimators in Part 4 do not necessarily add up to the reliable direct estimator, $\hat{Y}$, at a large area level. A simple way to ensure consistency with $\hat{Y}$ is to use a ratio adjustment. Suppose $\hat{Y}_i(C)$ is a composite estimator of i-th area total, then the ratio-adjusted estimator is given by

$$\hat{Y}_{ia}(C) = \frac{\hat{Y}_i(C)}{\sum_i \hat{Y}_i(C)} \hat{Y} \tag{2.18}$$

Table 4. Batting Averages for 18 Baseball Players

Player	Direct est.	True Value	J-S est.	Compromise J-S est.
1	0.400	0.346	0.293	0.334
2	0.378	0.298	0.289	0.312
3	0.356	0.276	0.284	0.290
4	0.333	0.221	0.279	0.279
5	0.311	0.273	0.275	0.275
6	0.311	0.270	0.275	0.275
7	0.289	0.263	0.270	0.270
8	0.267	0.210	0.265	0.265
9	0.244	0.269	0.261	0.261
10	0.244	0.230	0.261	0.261
11	0.222	0.264	0.256	0.256
12	0.222	0.256	0.256	0.256
13	0.222	0.304	0.256	0.256
14	0.222	0.264	0.256	0.256
15	0.222	0.226	0.256	0.256
16	0.200	0.285	0.251	0.251
17	0.178	0.319	0.247	0.243
18	0.156	0.200	0.242	0.221

The ratio-adjusted estimators (2.18) add upto the reliable direct estimator $\hat{Y}$ at the large area level.

3 MODEL-BASED ESTIMATORS

3.1 Basic area-level model

Model-based methods of small area estimation have received a lot of attention because of the following advantages: (1) Model-based methods make specific allowance for local variation through complex error structures in the models that link the small areas. Efficient indirect estimators can be obtained under the assumed models. (2) Models can be validated from the sample data. (3) Methods can handle complex cases such as cross-sectional and time series data. (4) Stable area specific measures of variability associated with the estimates can be obtained, unlike the overall measures for synthetic and composite estimators mentioned in Part 2.

We introduce three model-based methods, empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB), using a basic area-level model that uses area-level covariates. This model has two components: (a) The direct estimator $\hat{\theta}_i = g\left(\hat{\bar{Y}}_i\right)$ for a specified function $g(.)$ is equal to the sum of the population value $\theta_i = g\left(\bar{Y}_i\right)$ and the sampling error e_i:

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1,...m \tag{3.1}$$

where the sampling errors e_i are assumed to be independent across areas i with means 0 and known variances ψ_i. (b) A linking model that relates the θ_i's to area-level auxiliary variables $\mathbf{z}_i = \left(z_{i1},...,z_{ip}\right)^T$ through a linear regression model:

$$\begin{aligned} \theta_i &= z_{i1}\beta_1 + ... + z_{ip}\beta_p + v_i, \\ &= \mathbf{z}_i^T\boldsymbol{\beta} + v_i \end{aligned} \tag{3.2}$$

where the model errors v_i are assumed to be independent with means 0 and a common unknown variance σ_v^2 and independent of e_i and $\boldsymbol{\beta}$ is the p-vector of regression parameters $\beta_1,...,\beta_p$. Combining the two components (3.1) and (3.2) we get a combined model

$$\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + v_i + e_i \tag{3.3}$$

which is of the form of a linear mixed effects models with fixed effects β and random small area effects v_i. The parameter σ_v^2 is a measure of homogeneity of the areas after accounting for the covariates $\mathbf{z}_i$. Note that the combined model involve both design-based random variables e_i and model-based random variables v_i.

In practice, sampling variances ψ_i are seldom known, but smoothing of direct estimated variances $\hat{\psi}_i$ is often done to get stable estimates which are then treated as the true ψ_i. Other methods of handling unknown sampling variances are discussed in section 3.3. The assumption that the direct estimator $\hat{\theta}_i$ is design-unbiased for θ_i may not be valid if θ_i is a nonlinear function and the area sample size n_i is small. A more realistic sampling error model assumes that the director estimator $\hat{Y}_i$ of the total Y_i is design-unbiased (or approximately so for large over-all sample size n); that is, $\hat{Y}_i = Y_i + e_i^*$ with zero mean sampling errors e_i^*. But we cannot combine this sampling model with the linking model (3.2) directly to produce a linear mixed combined model. As a result, standard results in linear model theory do not apply, unlike in the case of the linear model (3.3).

The basic area level model (3.3) has been extended to handle correlated sampling errors, spatial dependence of random small area effects, vectors of parameters $\boldsymbol{\theta}_i$ (multivariate case), time series cross-sectional data and others. Part 4 of this monograph presents some of these extensions.

3.2 EBLUP and EB methods

EBLUP, EB and HB methods have played a prominent role for model-based small area estimation. EBLUP method is applicable for linear mixed models whereas EB and HB methods are more generally valid. EBLUP estimators do not require distributional assumptions on the random errors e_i and v_i, but normality is often assumed for *MSE* estimation. Also, EBLUP and EB estimators are identical under normality and nearly equal to the HB estimator, but measures of variability of the estimators may be different.

Estimators

A linear estimator, $\sum l_i \hat{\theta}_i$, with fixed coefficients l_i is called a linear unbiased prediction (LUP) estimator of the realized value of θ_i if the expectation of the deviation $\sum l_i \hat{\theta}_i - \theta_i$ with respect to the combined model (3.3) is zero. BLUP estimator of the realized θ_i is the estimator with minimun *MSE* in the class of LUP estimators.

Appealing to general results for linear mixed models (see Prasad and Rao, 1990), the BLUP estimator of the realized θ_i under the model (3.3) is given by

$$\tilde{\theta}_i\left(\sigma_v^2\right) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\mathbf{z}_i^T \tilde{\boldsymbol{\beta}}\left(\sigma_v^2\right) \tag{3.4}$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$ and $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$ is the weighted least squares estimator of $\boldsymbol{\beta}$ with weights $(\sigma_v^2 + \psi_i)^{-1}$ obtained by regressing $\hat{\theta}_i$ on $\mathbf{z}_i$:

$$\tilde{\boldsymbol{\beta}}(\sigma_v^2) = \left(\sum_i \gamma_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \left(\sum_i \gamma_i \mathbf{z}_i y_i\right) \tag{3.5}$$

It is clear from (3.4) that the BLUP estimator is a weighted combination of the direct estimator $\hat{\theta}_i$ and the "regression synthetic" estimator $\mathbf{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)$ with weights γ_i and $1-\gamma_i$ respectively. The BLUP estimator gives more weight to $\hat{\theta}_i$ when the sampling variance ψ_i is small (or σ_v^2 is large) and moves towards the regression synthetic estimator as ψ_i increases (or σ_v^2 decreases). For the nonsampled areas, the BLUP estimator is given by the regression synthetic estimator, using the covariates $\mathbf{z}_i$ associated with those areas.

A measure of variability associated with the BLUP estimator is given by its $MSE = E(est. - \theta_i)^2$. We have

$$MSE\left[\tilde{\theta}_i(\sigma_v^2)\right] = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2), \tag{3.6}$$

where

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i \tag{3.7}$$

and

$$g_{2i}(\sigma_v^2) = \sigma_v^2 (1-\gamma_i)^2 \mathbf{z}_i^T \left(\sum_i \gamma_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \mathbf{z}_i . \tag{3.8}$$

The results (3.4) and (3..6) do not require distributional assumptions on the random errors v_i and e_i.

The leading term $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ is of order $O(1)$ whereas $g_{2i}(\sigma_v^2)$, due to estimating $\boldsymbol{\beta}$, is of lower order, $O(m^{-1})$, for large number of sampled small areas, m. The leading term shows that MSE of the BLUP estimator can be substantially smaller than the MSE of the direct estimator under the assumed model (3.3) when γ_i is small or the model variance σ_v^2 is small relative to the sampling variance ψ_i. The success of small area estimation, therefore, largely depends on getting good auxiliary data $\{\mathbf{z}_i\}$ that leads to a small model variance

relative to sampling variance. Of course, one should also make a thorough validation of the assumed model.

In practice the model variance σ_v^2 is unknown so we replace it by a suitable estimator $\hat{\sigma}_v^2$ to obtain a two-step or EBLUP estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$. The estimator of the small area mean $\bar{Y}_i$ is then given by $g^{-1}(\tilde{\theta}_i)$. A simple method of moments estimator of σ_v^2 is given by $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$, where

$$(m-p)\tilde{\sigma}_v^2 = \sum_i (\hat{\theta}_i - \mathbf{z}_i^T \boldsymbol{\beta}^*)^2 - \sum_i \psi_i h_{ii} \tag{3.9}$$

with $h_{ii} = \mathbf{z}_i^T \left(\sum_i \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \mathbf{z}_i$ and $\boldsymbol{\beta}^*$ is the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. Alternatively, $\tilde{\sigma}_v^2$ is obtained iteratively as a solution the following nonlinear equation (Fay and Herriot, 1979):

$$a(\sigma_v^2) = \sum_i [\hat{\theta}_i - \mathbf{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)]^2 / (\sigma_v^2 + \psi_i) = m - p, \tag{3.10}$$

where $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$ is given by (3.5). The middle term in (3.10) is the weighted residual sum of squares which is equated its expected value *m-p*. Note that *m-p* is the degrees of freedom associated with the weighted residual sum of squares. We truncate the estimator $\tilde{\sigma}_v^2$ to 0 as in the case of (3.9) to get $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$. Note that if $\hat{\sigma}_v^2 = 0$ then the EBLUP estimator $\tilde{\theta}_i$ reduces to the regression synthetic estimator $\mathbf{z}_i^T \hat{\boldsymbol{\beta}}$ even for the sampled areas, where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$, the estimator obtained by substituting $\hat{\sigma}_v^2$ for σ_v^2 in (3.5). The estimators obtained from (3.9) or (3.10) do not require distributional assumptions on v_i and e_i. If normality of the random errors v_i and e_i is assumed, then the marginal distribution of $\hat{\theta}_i$ is normal with mean $\mathbf{z}_i^T \boldsymbol{\beta}$ and variance $\sigma_v^2 + \psi_i$, and the $\hat{\theta}_i$'s are independent. Using this result, restricted maximum likelihood estimators (REML) of $\boldsymbol{\beta}$ and σ_v^2 can be obtained. The REML estimators for linear mixed models remain asymptotically valid under deviations from normality (Jiang, 1996). Therefore, the EBLUP estimator $\tilde{\theta}_i$ using REML estimator of σ_v^2 is also asymptotically valid (for large *m*) under deviations from nonnormality. We refer the reader to Cressie (1992) for a good introduction to REML estimation in the context of census undercount estimation.

We now turn to empirical Bayes (EB) estimation, assuming normality of the random errors v_i and e_i which implies that the joint distribution of $(\hat{\theta}_i, \theta_i)$ is bivariate normal with means $(\mathbf{z}_i^T\boldsymbol{\beta}, \mathbf{z}_i^T\boldsymbol{\beta})$, variances $(\sigma_v^2 + \psi_i, \sigma_v^2)$ and correlation γ_i. Using the latter result, the minimum *MSE* estimator of the realized θ_i, given by the conditional expectation of θ_i given $\hat{\theta}_i$, $E(\theta_i | \hat{\theta}_i, \boldsymbol{\beta}, \sigma_v^2)$, reduces to

$$\tilde{\theta}_i^B(\boldsymbol{\beta}, \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\mathbf{z}_i^T\boldsymbol{\beta} \quad . \tag{3.11}$$

This estimator, called Bayes estimator, is optimal among all estimators that are functions of the data $\{\hat{\theta}_i, \mathbf{z}_i\}$; linearity or model-unbiasedness of the estimators is not required. The model parameters $\boldsymbol{\beta}$ and σ_v^2 are replaced by REML estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$ to get an EB estimator:

$$\tilde{\theta}_i^{EB} = \tilde{\theta}_i^B(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2). \tag{3.12}$$

The EB estimator is identical to the EBLUP estimator $\tilde{\theta}_i$ under normality, but the EB approach is applicable generally for any joint distribution of $\hat{\theta}_i$ and θ_i. It should be noted that the EB approach is essentially frequentist, because it uses only the sampling model and the linking model which can be validated from the data; no prior distributions on the model parameters unlike in the HB approach considered in section 3.3.

Fay and Herroit (1979) recommended the use of a compromise EB estimator similar to the compromise James-Stein estimator (2.18); that is, (i) use $\tilde{\theta}_i^{EB}$ if $\tilde{\theta}_i^{EB}$ lies in the interval $\left[\hat{\theta}_i - c\sqrt{\psi_i}, \hat{\theta}_i + c\sqrt{\psi_i}\right]$; (ii) use $\hat{\theta}_i - c\sqrt{\psi_i}$ if $\tilde{\theta}_i^{EB}$ is less than $\hat{\theta}_i - c\sqrt{\psi_i}$, (iii) use $\tilde{\theta}_i^{EB} + c\sqrt{\psi_i}$ if $\hat{\theta}_i$ exceeds $\hat{\theta}_i + c\sqrt{\psi_i}$.

Applications

We present three applications of the EBLUP(EB) estimators: (1) Estimation of per capita income (PCI) for small places (population less than 1000) in U.S.A. (2) Estimation of school-age children in poverty at the county level in U.S.A. (3) Estimation of net undercoverage in the 1991 Canadian Census for 96 small areas defined by sex (2) x age (4) x province (12) combinations.

Application 3.1

The U.S. Census Bureau was required to provide the Treasury Department with the PCI estimates and other statistics for state and local governments receiving funds under the General Revenue Sharing Program. These statistics were then used by the Treasury Department to determine allocations to the local government units (places) within the different states by dividing the corresponding state allocations. Initially, the Census Bureau determined the current estimates of PCI by multiplying the 1970 census estimates of PCI in 1969 (based on a 20 percent sample) by ratios of an administrative estimate of PCI in the current year and a similarly derived estimate for 1969. But the sampling errors of the PCI estimates turned out to be quite large for places having fewer than 500 persons in 1970: RRMSE of about 13 percent for a place of 500 persons and 30 percent for a place with 100 persons. As a result, the Bureau initially decided to set aside the census estimates for these small places and to substitute the corresponding county estimates in their place. But this solution turned out to be unsatisfactory because the census sample estimates for many small places differed significantly from the corresponding county estimates after accounting for sampling errors. Using the EB method, Fay and Herriot (1979) proposed a better solution and presented empirical evidence that the EB estimates have average error smaller than either the census sample estimates or the county averages for small places. The EB estimate used by them is a weighted average of the census sample estimate and a regression synthetic estimate obtained by fitting a linear regression equation to the sample estimates of PCI using as independent variables the associated county averages, tax-return data for 1969 and data on housing from the 1970 census. The Fay-Herriot method was adopted by the Census Bureau in 1974 to form updated estimates of PCI for small places. This was the largest application of EB methods in a U.S. Federal Statistical Program.

We now provide some details of the Fay-Herriot application. First, based on past studies, the CV of the sample estimate, $\hat{\bar{Y}}_i$, of PCI was taken as $3.0/\hat{N}_i$ for the i-th area, where $\hat{N}_i$ is the weighted sample count; $\hat{\bar{Y}}_i$ and $\hat{N}_i$ were available for almost all places. This suggested the use of logarithmic transformation, $\hat{\theta}_i = \ln\left(\hat{\bar{Y}}_i\right)$, with $\text{var}\left(\hat{\theta}_i\right) \approx \left[C.V.\left(\hat{\bar{Y}}_i\right)\right]^2 = 9/\hat{N}_i = \psi_i$. Secondly, four separate regression models were evaluated to determine a suitable combined model, treating the sampling variances ψ_i as known: (1) $z_1 = 1, z_2 =$ logarithm of PCI for the county; $p = 2, (2)$ $z_1, z_2, z_3 =$ ln (value of housing for the place), $z_4 =$ ln (value of housing for the county); $p = 4, (3)$ $z_1, z_2, z_5 =$ ln (average gross income per exemption from the 1969 tax returns for the place), $z_6 =$ ln (average gross income per exemption from the 1969 tax returns for the county); $p = 4, (4)$ $z_1, \ldots, z_6; p = 6$.

Fay and Herriot (1979) calculated the values of $\hat{\sigma}_v^2$ for each of the four models using the iterative method based on (3.10). The values of $\hat{\sigma}_v^2$ provide a measure of the average fit

of the regression models to the sample data, after allowing for sampling errors in $\hat{\theta}_i$'s. A value for $\hat{\sigma}_v^2$ of 0.045 corresponds to $\psi_i = 9.0/200 = 0.045$ for a place of size 200, and equal weighting $(\hat{\gamma}_i = \frac{1}{2})$ of the direct estimate and the regression synthetic estimate. The resulting *MSE*, based on the leading term $g_{1i}(\hat{\sigma}_v^2) = \hat{\sigma}_v^2 \hat{\gamma}_i$, is one-half of the sampling variance; that is, the EB estimate for a place of 200 persons roughly has the same precision as the direct estimate for a place of 400 persons.

Table 3.1 reports the values of $\hat{\sigma}_v^2$ for the states with more than 500 small places (size less than 500). It is clear from Table 3.1 that regressions involving either tax or housing data, but especially those including both, are significantly better than the regression on the county values alone; that is ,model 4 and to a lesser extent models 2 and 3 provide better fits to the data than model 1. Note that the values of $\hat{\sigma}_v^2$ for model 4 are much smaller than 0.045, especially for North and South Dakota, Nebraska, Wisconsin and Iowa.

Table 3.1 Values of $\hat{\sigma}_v^2$ for states with more than 500 small places

State	Model			
	(1)	(2)	(3)	(4)
Illinois	0.036	0.032	0.019	0.017
Iowa	0.029	0.011	0.017	0.000
Kansas	0.064	0.048	0.016	0.020
Minnesota	0.063	0.055	0.014	0.019
Missouri	0.061	0.033	0.034	0.017
Nebraska	0.065	0.041	0.019	0.000
North Dakota	0.072	0.081	0.020	0.004
South Dakota	0.138	0.138	0.014	*
Wisconsin	0.042	0.025	0.025	0.004

* *Regression not fitted because of two few data points.*

The compromise EB estimates were obtained from the EB estimates and transformed back to the origin scale. The latter estimates were then subjected to a two-step raking to ensure consistency with the following aggregate sample estimates: (i) For each of the classes <500, 500-999 and >1000, the total estimated income for all places equals the direct estimate at the state level; (ii) The total estimated income for all places in a county equals the direct county estimate of total income.

Table 3.2 Values of Percent Difference of Estimates from True Values

Special Census area	Direct est.	EB est.	County est.
	Population less than 500		
1	10.2	14.0	12.9
2	4.4	10.3	30.9
3	34.1	26.2	9.1
4	1.3	8.3	24.6
5	34.7	21.8	6.6
6	22.1	19.8	14.6
7	14.1	4.1	18.7
8	18.1	4.7	25.9
9	60.7	78.7	99.7
10	47.7	54.7	95.3
11	89.1	65.8	86.5
12	1.7	9.1	12.7
13	11.4	1.4	6.6
14	8.6	5.7	23.5
15	23.6	25.3	34.3
16	53.6	10.5	11.7
17	51.4	14.4	23.7
Average	28.6	22.0	31.6
	Population between 500 and 999		
1	36.5	28.0	36.0
2	8.5	4.1	9.3
3	7.4	2.7	7.7
4	13.6	16.9	13.6
5	25.3	16.3	25.8
6	33.2	34.1	32.9
7	9.2	7.2	9.9
Average	19.1	15.6	19.3

Evaluation

Fay and Herroit (1979) also made an evaluation study by comparing the estimates for 1972 to "true" values obtained from census of a random sample of places in 1973. Table 3.2 reports the values of percentage difference $\{|\text{est.} - \text{true value}|/\text{true value}\}\times 100$ for the special census areas using direct, county and EB estimates, as well as the values of average percentage difference. It is clear from Table 3.2 that the EB estimates exhibit smaller average errors and a lower incidence of extreme errors than either the direct estimate or the county estimate: 22% compared to 28.6% (for direct estimates) and 31.6% (for county estimates) for

places with less than 500 persons; 15.6% compared to 19.1% (for direct estimates) and 19.3% (for county estimates) for places with population size between 500 and 999. The EB estimates were consistently higher than the special census values, but missing income was not imputed in the special census (unlike in the 1970 sample census) and therefore were subject to a downward bias.

Application 3.2

The basic area-level model (3.3) has been used recently to produce model-based current county estimates of poor school-age children in U.S.A. (National Research Council, 1998). Using these estimates, the U.S. Department of Education allocates over $7 billion of general funds annually to counties, and then states distribute those funds among school districts. In the past, funds were allocated on the basis of estimated counts from the previous census sample counts, but the poverty counts have changed significantly overtime.

In this application, $\theta_i = \ln Y_i$, where Y_i is the true poverty count of the i-th county (small area), and $\hat{Y}_i = 3$–year weighted average of poor school age children (under 18) obtained from the March Income Supplement of the Current Population Survey (CPS). The following area-level predictor variables, obtained from administrative records, were used in fitting the linear regression model (3.3): $z_1 = 1, z_2 = \ln$ (number of child exemptions reported by families in poverty on tax returns), $z_3 = \ln$ (number of people receiving food stamps), $z_4 = \ln$ (estimated population under age 18), $z_5 = \ln$ (total number of child exemptions on tax returns) and $z_6 = \ln$ (number of poor school-age children from the previous census). Counties with CPS sample but no prior school-age children (i.e., $\hat{Y}_i = 0$) were excluded due to the log transformation $(\ln 0 = -\infty)$.

The difficulty with unknown sampling variances ψ_i was handled by (i) using a model of the same form as above for the census year 1990, for which reliable estimates $\hat{\psi}_{ic}$ of sampling variances ψ_{ic} are available and (ii) assuming the census model errors v_{ic} follow the same distribution as the current model errors v_i; that is, normal with mean 0 and variance σ_v^2. Under assumption (ii) an estimate of σ_v^2 was obtained from the census data assuming $\hat{\psi}_{ic} = \psi_{ic}$ and used in the current model, assuming $\psi_i = \sigma_e^2 / n_i$, to get an estimate $\tilde{\sigma}_e^2$ of σ_e^2. The resulting estimate $\tilde{\psi}_i = \tilde{\sigma}_e^2 / n_i$ was treated as the true ψ_i in calculating the EBLUP (EB) estimate $\tilde{\theta}_i$ of θ_i. The county totals Y_i can then be estimated as $\tilde{Y}_i = \exp(\tilde{\theta}_i)$, but a more refined method based on the mean of lognormal distribution was used. The county estimates were then raked to agree with model-based state estimates obtained from a state model; the state estimates were ratio-adjusted to agree with the direct national estimate.

Evaluations

Both internal and external evaluations were conducted for model choice and for checking the validity of the model. In an internal evaluation the validity of the underlying assumptions and features of the model are examined. On the other hand, in an external evaluation the estimates from a model are compared with "true" values that were not used in the development of the model; internal evaluation of regression output precedes external evaluation.

For internal evaluation, the following features were examined: (a) linearity of the regression, (b) constancy of regression coefficients overtime, (c) choice of predictor variables, (d) normality of standardized residuals $r_i/s.e.(r_i)$, where $r_i = \hat{\theta}_i - \mathbf{z}_i'\hat{\beta}$ is the residual and $s.e.(r_i)$ a function of $\hat{\sigma}_v^2$ and ψ_i, (e) homogeneity of the variances of standardized residuals, (f) outliers. No significant departures from the assumed model were observed.

For external evaluations, models were fitted to the 1989 CPS estimates and predictor variables and model estimates of county poverty counts for 1989 were then obtained following the procedures outlined above; the 1989 county estimates obtained from the 1990 census were used to estimate the model variance. Table 3.3 reports the average percent difference $\left(\sum_i |est_i - \text{true value}_i|/\text{true value}_i\right)\times 100$ for three different county estimates by treating the 1990 census estimates as true values: (1) Model estimates; (2) Stable shares estimates: county shares within a state same as those in the 1980 census; (3) Stable rates estimates: county ratios (poor/population) within a state same as those in the 1980 census. Estimates (2) were obtained by benchmarking 1980 census poverty counts to state estimates for 1990, while estimates (3) were obtained by multiplying 1980 census ratios by current population estimates and then benchmarking to state estimates for 1990. Note that (2) and (3) rely heavily on the 1980 census.

Table 3.3 Values of Average Percent Difference for Three Estimates

Estimate :	Model	Stable shares	Stable rates
	16.4	27.1	26.2

The 1990 census estimates that were used in the comparisons were ratio adjusted by a common factor to make the census national estimate of poor school-age children equal the 1989 CPS national estimate. This was done to account for the CPS-census differences in measurement of income and poverty. It is clear from Table 3.3 that the model estimates are much better than the stable shares or stable rates estimates. Apart from the above overall comparison, average percent differences for subgroups were also examined for various types of subgroups. The latter analysis revealed that the use of $z_4 = \ln(\text{estimated population under } 18)$ is better than using $z_4^* = \ln(\text{estimated population under } 21)$ as a predictor variable in the

regression model eventhough the latter led to a slightly smaller overall average percent difference: 15.4. The model using z_4^* (called log number model (under 21)) did not perform well in terms of average percent difference for counties with large numbers of people under age 21 in group quarters.

Application 3.3

Dick (1995) used the basic area-level model (3.3) to estimate net undercoverage rates in the 1991 Canadian Census. The objective here is to estimate 96 adjustment factors $\theta_i = T_i/C_i$ corresponding to sex (2) x age (4) x province (12) combinations, where T_i is the true (unknown) count and C_i is the census count in the i-th domain (area); the net undercoverage rate in the i-th area is given by $U_i = 1 - \theta_i^{-1}$. Direct estimates $\hat{\theta}_i$ were obtained from a post enumeration survey and the associated sampling variances ψ_i were derived through smoothing of the estimated variances. In particular, the variance of estimated number of missing person, $\hat{M}_i$, was assumed to be proportional to a power of the census count and a linear regression equation was fitted to $\left(\log \hat{V}\left(\hat{M}_i\right), \log C_i, i = 1,...,m\right)$, where $\hat{V}\left(\hat{M}_i\right)$ is the estimated variances of $\hat{M}_i$. Using this relationship, the sampling variances were calculated from $\ln \psi_i = -6.13 - 0.28 \ln C_i$ and the resulting ψ_i were treated as true ψ_i.

Explanatory (predictor) variables $\mathbf{z}$ for building the model were selected from a set of 42 variables by backward stepwise regression described in Draper and Smith (1981, chapter6). Internal evaluation of the resulting combined model was then performed, by analysing the standardized residuals $\mathbf{r}_i = \left(\tilde{\theta}_i^{EB} - \mathbf{z}_i^T \hat{\boldsymbol{\beta}}\right) / \left(\hat{\sigma}_v^2 + \psi_i\right)^{\frac{1}{2}}$, where the EB estimator $\tilde{\theta}_i^{EB}$ was calculated from (3.12) using REML estimates of $\boldsymbol{\beta}$ of σ_v^2 and. No significant departures from the assumed model were observed.

The EB adjustment factors $\tilde{\theta}_i^{EB}$ were converted to estimates of missing persons, $\tilde{M}_i^{EB}$, and these estimates were then subjected to two-step raking to ensure consistency with reliable direct estimates of marginal totals $\hat{M}_{p+}$ and $\hat{M}_{+a}$, where "p" denotes a province, "a" denotes an age-sex group and $M_i = M_{pa}$. The raked EB estimates $\tilde{M}_{pa}^R$ were used as the final estimates. These estimates were further divided into single year of age estimates by using simple synthetic estimation:

$$\tilde{M}_{pa}(q) = \tilde{M}_{pa}^R \left[C_{pa}(q)/C_{pa}\right],$$

where q denotes a sub-age group and $C_{pa}(q)$ is the associated census count.

MSE Estimation

An advantage of the model-based approach is that it permits stable, area-apecific estimators of *MSE*, unlike the average measures used for synthetic and composite estimators. *MSE* estimation under the model-based approach is highly technical, and we simply present some key results here.

First, a "naïve" estimator of *MSE* of EBLUP estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$ can be obtained from the formula (3.6) for the *MSE* of BLUP estimator $\tilde{\theta}_i(\hat{\sigma}_v^2)$ by substituting $\hat{\sigma}_v^2$ for σ_v^2. But this leads to significant underestimation of true *MSE* because the effect of estimating σ_v^2 is ignored. Prasad and Rao (1990) obtained a second-order correct (or approximately unbiased) estimator of *MSE* of EBLUP estimator $\tilde{\theta}_i$, assuming normality of $\{v_i\}$ and $\{e_i\}$. For the simple moment estimator (3.9) of σ_v^2, it is given by

$$mse(\tilde{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \tag{3.13}$$

where $g_{1i}(\hat{\sigma}_v^2)$ and $g_{2i}(\sigma_v^2)$ are as given in (3.7) and (3.8) and $g_{3i}(\sigma_v^2) = \left[\psi_i^2/(\sigma_v^2+\psi_i)^3\right] h(\sigma_v^2)$ with $h(\sigma_v^2) = 2m^{-2}\sum_i (\sigma_v^2+\psi_i)^2$.

Lahiri and Rao (1995) showed that (3.13) is robust in the sense that it remains valid under moderate nonnormality of $\{v_i\}$. The estimator (3.13) depends on z_i but not on the area-specific direct estimator $\hat{\theta}_i$. An alternative estimator that depends on $(\hat{\theta}_i, z_i)$ through the least squares residuals $r_i = \hat{\theta}_i - \mathbf{z}_i^T\hat{\boldsymbol{\beta}}$ is given by

$$mse_a(\tilde{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{4i}(r_i^2, \hat{\sigma}_v^2), \tag{3.14}$$

where

$$g_{4i}(r_i^2, \sigma_v^2) = \left[\psi_i^2/(\sigma_v^2+\psi_i)^4\right] r_i^2 h(\sigma_v^2).$$

Both (3.13) and (3.14) are approximately unbiased in the sense that their bias is of lower order than m^{-1}, for large m.

Recently, Jiang, Lahiri and Wan (1999) proposed a jackknife estimator of *MSE* which is also approximately unbiased. An advantage of the jackknife method is that it permits extension to more complex models such as logistic regression with random small area effects.

We write the EB estimator (3.12) of θ_i as

$$\tilde{\theta}_i^{EB} = k\left(\hat{\theta}_i, \hat{\boldsymbol{\varphi}}\right),$$

where $\boldsymbol{\varphi} = \left(\boldsymbol{\beta}, \sigma_v^2\right)$ denotes the model parameters $\boldsymbol{\beta}$ and σ_v^2. The jackknife steps are then as follows:

(i) Calculate $\hat{\boldsymbol{\varphi}}(l)$, the estimator of $\boldsymbol{\varphi}$ when l-th area data $\left(\hat{\theta}_l, \mathbf{z}_l\right)$ is deleted. Let

$$\tilde{\theta}_i^{EB}(l) = k\left[\hat{\theta}_i, \hat{\boldsymbol{\varphi}}(l)\right]$$

(ii) Calculate

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}\left[\tilde{\theta}_i^{EB}(l) - \tilde{\theta}_i^{EB}\right]^2$$

(iii) Calculate

$$\hat{M}_{1i} = g_{1i}\left(\hat{\sigma}_v^2\right) - \frac{m-1}{m}\sum_{l=1}^{m}\left[g_{1i}\left(\hat{\sigma}_v^2(l)\right) - g_{1i}\left(\hat{\sigma}_v^2\right)\right]^2$$

(iv) Jackknife estimator of *MSE* is calculated as

$$mse_J\left(\tilde{\theta}_i^{EB}\right) = \hat{M}_{1i} + \hat{M}_{2i}. \tag{3.15}$$

Note that $\hat{M}_{1i}$ estimates *MSE* when $\boldsymbol{\varphi}$ is known and $\hat{M}_{2i}$ estimates the extra variability in *MSE* due to estimating the model parameters $\boldsymbol{\varphi}$.

3.3 HB method

The hierarchical Bayes (HB) approach is straightforward, inferences are exact and complex problems can be handled using recently developed Monte Carlo Markov Chain (MCMC) methods, such as the Gibbs sampler. In the HB approach, the population values θ_i as well as the model parameters $\boldsymbol{\varphi} = \left(\boldsymbol{\beta}, \sigma_v^2\right)$ are regarded as random, and a prior distribution on the model parameters (also called hyperparameters) is specified. The inferences on θ_is are based on the marginal posterior distribution (conditional distribution of θ_i given the data $\left\{\left(\hat{\theta}_i, z_i\right), i = 1,\ldots,m\right\}$: $f\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right)$, where $\hat{\boldsymbol{\theta}}$ is the vector of direct estimates $\hat{\theta}_i$. In particular, θ_i is estimated by the posterior mean $E\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right)$ and the variability of the estimate is measured by

the posterior variance $V\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right)$, assuming squared error loss which corresponds to *MSE* in the EBLUP or EB approach.

σ_v^2 known

We first consider the case of known σ_v^2 and specify a prior distribution on $\boldsymbol{\beta}$. If the prior is proportional to a constant (called improper prior) and normality of v_i and e_i is assumed, then the posterior mean $E\left(\theta_i \middle| \hat{\boldsymbol{\theta}}, \sigma_v^2\right)$ is identical to the BLUP estimator $\tilde{\theta}_i\left(\sigma_v^2\right)$ given by (3.4). Further, the posterior variance $V\left(\theta_i \middle| \hat{\boldsymbol{\theta}}, \sigma_v^2\right)$ is identical to *MSE* of the BLUP estimator given by (3.6). Therefore, if σ_v^2 is known, HB and EBLUP approaches lead to identical inferences. The improper prior on $\boldsymbol{\beta}$ reflects absence of prior information on $\boldsymbol{\beta}$.

σ_v^2 unknown

In practice, σ_v^2 is seldom known. We therefore assume a prior distribution on σ_v^2 and prior independence of $\boldsymbol{\beta}$ and σ_v^2. This leads to the marginal posterior distribution, $f\left(\sigma_v^2 \middle| \hat{\boldsymbol{\theta}}\right)$ However, an improper prior on σ_v^2 could lead to an improper posterior of θ_i s (Hobert and Casella, 1996). To avoid this difficulty, the prior on $\tau_v = \sigma_v^{-2}$ is assumed to be a gamma distribution with parameters $a > 0$ and $b > 0$, denoted by $G(a,b): f(\tau_v) \propto \exp(-a\tau_v)\tau_v^{b-1}$. Small values of a and b are chosen to reflect the absence of prior information on σ_v^2.

Using the marginal posterior $f\left(\sigma_v^2 \middle| \hat{\boldsymbol{\theta}}\right)$, the HB estimator $E\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right)$ is obtained as

$$\tilde{\theta}_i^{HB} = E\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right) = \int \tilde{\theta}_i\left(\sigma_v^2\right) f\left(\sigma_v^2 \middle| \hat{\boldsymbol{\theta}}\right) d\sigma_v^2 . \tag{3.16}$$

We may write (3.16) as $E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}\left[\tilde{\theta}_i\left(\sigma_v^2\right)\right]$, where $E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}$ denotes the expectation with respect to the marginal posterior distribution $f\left(\sigma_v^2 \middle| \hat{\boldsymbol{\theta}}\right)$. Similarly, the posterior variance $V\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right)$ is obtained as

$$V\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right) = E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}\left[g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right)\right] + V_{\sigma_v^2|\hat{\boldsymbol{\theta}}}\left[\tilde{\theta}_i\left(\sigma_v^2\right)\right], \tag{3.17}$$

where $V_{\sigma_v^2|\hat{\theta}}$ denotes the variance with respect to $f\left(\sigma_v^2\middle|\hat{\boldsymbol{\theta}}\right)$.

No closed form expressions for the integrals (3.16) and (3.17) exist, but in this simple case the integrals can be evaluated numerically using only one-dimensional numerical integration. For complex models, high-dimensional integration is often involved and it becomes necessary to use MCMC-type methods to overcome the computational difficulties.

It follows from (3.16) that $\tilde{\theta}_i^{HB}$ is approximately equal to the EBLUP (EB) estimator $\tilde{\theta}_i\left(\hat{\sigma}_v^2\right)$, but (3.17) shows that ignoring the uncertainty about σ_v^2 and then using $g_{1i}\left(\hat{\sigma}_v^2\right)+g_{2i}\left(\hat{\sigma}_v^2\right)$ as a measure of variability can lead to significant overestimation because the last term in (3.17) is positive.

Gibbs sampling

Gibbs sampling is a MCMC method that can be used to evaluate (3.16) and (3.17). To implement Gibbs sampling we need the following Gibbs-conditional distributions:

(i) $\boldsymbol{\beta}\left|\boldsymbol{\theta},\sigma_v^2,\hat{\boldsymbol{\theta}}\right.$ which is a p-variate normal with mean $\left(\sum \mathbf{z}_i\mathbf{z}_i^T\right)^{-1}\left(\sum \mathbf{z}_i\theta_i\right)$ and covariance matrix $\sigma_v^2\left(\sum \mathbf{z}_i\mathbf{z}_i^T\right)^{-1}$.

(ii) $\theta_i\left|\boldsymbol{\beta},\sigma_v^2,\hat{\boldsymbol{\theta}}\right.$ which is a univariate normal with mean $\tilde{\theta}_i^B\left(\boldsymbol{\beta},\sigma_v^2\right)$ given by (3.11) and variance $g_{1i}\left(\sigma_v^2\right)=\gamma_i\psi_i$.

(iii) $\tau_v=\sigma_v^{-2}\left|\boldsymbol{\beta},\boldsymbol{\theta},\hat{\boldsymbol{\theta}}\right.$ which is a gamma with parameters $\tilde{a}=\frac{1}{2}\sum\left(\theta_i-\mathbf{z}_i^T\ \right)^2+a$ and $\tilde{b}=\frac{m}{2}+b$.

The Gibbs algorithm is as follows: (a) Using starting values $\theta_i=\theta_i^{(0)}$ and $\sigma_v^2=\sigma_v^{2(0)}$, draw $\boldsymbol{\beta}^{(1)}$ from (i). (b) Draw $\theta_i^{(1)}, i=1,...m$ from (ii) using $\boldsymbol{\beta}=\boldsymbol{\beta}^{(1)}$ and $\sigma_v^2=\sigma_v^{2(0)}$. (c) Draw $\sigma_v^{2(1)}$ from (iii) using $\theta_i=\theta_i^{(1)}$ and $\boldsymbol{\beta}=\boldsymbol{\beta}^{(1)}$. Steps (a), (b), and (c) complete one cycle. Perform a large number of cycles, say t, called "burn-in" period until convergence and then treat $\left\{\beta^{(t+j)},\sigma_v^{2(t+j)},\theta_1^{(t+j)},...,\theta_m^{(t+j)};j=1,...,J\right\}$ as J simulated samples from the joint posterior distribution of $\boldsymbol{\beta},\sigma_v^2,\theta_1,...,\theta_m$. Alternative methods use multiple parallel runs instead of a single long run as above. But parallel runs can be wasteful because initial "burn-in" periods are discarded from each run. On the other hand, a single long run may leave a significant portion of the space generated by the joint posterior distribution unexplored.

For the single long run, we may use $\theta_i^{(0)} = \tilde{\theta}_i^{EB}$ and $\sigma_v^{2(0)}$ = REML estimator of σ_v^2 as starting values. For multiple parallel runs, we need multiple starting values.

Using the J simulated samples, the posterior mean and the posterior variance of θ_i are estimated as

$$\tilde{\theta}_i^{HB} \approx \frac{1}{J}\sum_j \tilde{\theta}_i\left[\sigma_v^{2(t+j)}\right] = \frac{1}{J}\sum_j \tilde{\theta}_i(j) = \tilde{\theta}_i(\cdot) \tag{3.18}$$

and

$$V\left(\theta_i \middle| \hat{\boldsymbol{\theta}}\right) \approx \frac{1}{J}\sum_j \left[g_{1i}\left(\sigma_v^{2(t+j)}\right) + g_{2i}\left(\sigma_v^{2(t+j)}\right)\right] + \frac{1}{J}\sum_j \left[\tilde{\theta}_i(j) - \tilde{\theta}_i(\cdot)\right]^2 . \tag{3.19}$$

For the basic area level model (3.3), all the Gibbs-conditional distributions (i), (ii) and (iii) are in a closed-form and, therefore, samples can be generated directly. But, for more complex models, some of the Gibbs-conditional distributions may not have closed form in which case alternative algorithms, such as Metropolis-Hastings within Gibbs and adaptive rejective sampling, are needed to draw samples from the joint posterior distribution. We refer the reader to Brooks (1998) for an excellent overview of the MCMC methods. Software, called BUGS (Bayesian Inference using Gibbs Sampling) and CODA (Convergence Diagnostics) are readily available for implementing MCMC and convergence diagnostics. (website: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.html). BUGS can handle a wide variety of models, using adaptive rejective sampling and Metropolis-Hastings within Gibbs.

Caution should be exercised in using MCMC methods and associated software. For example, Hobert and Casella (1996) demonstrated that the Gibbs sampler could lead to seemingly reasonable inferences about a nonexistant posterior distribution. This happens when the posterior distribution is improper and yet all the Gibbs-conditional distributions are proper. Another difficulty with MCMC is that the convergence diagnostics tools can fail to detect the sort of convergence failure that they were designed to be identify (Cowles and Carlin, 1996). Further difficulties include the choices of t for the burn-in period, number of simulated samples, J, and the starting values especially for multiple parallel runs.

3.4 Basic unit-level model

A basic unit level population model assumes that the unit y-value y_{ij}, associated with the unit j in the area i is related to unit-level auxiliary variables $\mathbf{x}_{ij}$ for which the area population mean vector $\overline{\mathbf{X}}_i$ is assumed to be known. If y is a continous variable, we assume a one-way nested error linear regression model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij},\ j = 1,...,N_i;\ i = 1,...m \tag{3.20}$$

where the random small area effects v_i have mean 0 and common variance σ_v^2, and independently distributed. Further, the v_i are independent of the residual errors e_{ij} which are independently distributed with mean 0 and variance $a_{ij}\sigma_e^2$ for specified constants a_{ij}. The parameters of interest are the totals Y_i or the means $\overline{Y}_i$. To handle count or categorical y-variables (for example, binary y-variable), generalized linear nested error regression models are often used (see Part 4).

The sample data $\{y_{ij}, \mathbf{x}_{ij},\ j = 1,...,n_i;\ i = 1,...m\}$ are assumed to obey the population model. This means that the sample design is "ignorable" or selection bias is absent which is satisfied, for example, for simple random sampling within areas. In general, the sample indicator variables should be unrelated to y_{ij}, conditional on $\mathbf{x}_{ij}$. Model-based estimators for unit level models do not depend on the survey weights so that design-consistency as n_i increases is forsaken except when the weights are equal, as in the case of simple random sampling within areas. The area-level model (3.3) is free of this limitation but assumes that the sampling variances ψ_i are known. The unit-level model is free of the latter assumption and it is possible to incorporate survey weights using model-assisted estimators (Kott, 1990; Prasad and Rao, 1999).

We assume, for simplicity, that the area sampling fractions n_i/N_i are negligible and the error variances are equal, i.e., $a_{ij} = 1$. Then the EBLUP (EB) estimator, $\overline{y}_i^{EB}$, of the mean $\overline{\mathbf{Y}}_i \approx \overline{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i$ is a weighted average of the "survey regression" estimator $\overline{y}_i + \left(\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i\right)^T\tilde{\boldsymbol{\beta}}$ and the regression synthetic estimator $\overline{\mathbf{X}}_i^T\tilde{\boldsymbol{\beta}}$ with weights $\hat{\gamma}_i$ and $1-\hat{\gamma}_i$ where $\left(\overline{y}_i, \overline{\mathbf{x}}_i\right)$ are the sample means for area i, $\hat{\gamma}_i = \hat{\sigma}_v^2/\left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i\right)$ and $\tilde{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$. The estimators of variance components σ_v^2 and σ_e^2 are obtained either by REML or the method of fitting of constants. Note that as the small area sample size, n_i, increases $\hat{\gamma}_i$ tends to 1 and the EB estimator approaches the survey regression estimator. For small n_i and small $\hat{\sigma}_v^2/\hat{\sigma}_e^2$, the EB estimator gives more weight to the regression synthetic estimator.

The leading term of the *MSE* of $\bar{y}_i^{HB}$ is given by $g_{1i}(\sigma_v^2.\sigma_e^2) = \gamma_i \sigma_e^2/n_i$, where σ_e^2/n_i is the *MSE* of the sample regression estimator and $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_i)$. This shows that considerable gain in efficiency over the sample regression estimator is achieved if γ_i is small. Therefore, models with smaller $\hat{\gamma}_i$ should be prefered, provided they provide adequate fit in terms of residual analysis and other model diagnostics. This is similar to the model choice in Application 3.1.

Prasad and Rao (1990) obtained a second-order correct (or approximately unbiased) estimator of *MSE* of the EB estimator, assuming normality of $\{v_i\}$ and $\{e_{ij}\}$. The jackknife method outlined in Section 3.2 can also be used to estimate the *MSE*.

Application 3.4

Battese, Harter and Fuller (1988) applied the nested error regression model (3.20) to estimate areas under corn and soybeans for each of 12 counties in North-Central Iowa, using farm-interview data in conjunction with LANDSAT satellite data. Each county was divided into area segments, and the areas under corn and soybeans were ascertained for a sample of segments by interviewing farm operators; the number of sample segments in a county ranged from 1 to 5. Auxiliary data in the form of number of pixels (a term used for "picture elements" of about 0.45 hectares) classified as corn and soybeans were also obtained for all the area segments, including the sample segments, in each county using the LANDSAT satellite readings. In this application, $\mathbf{x}_{ij} = (1, x_{1ij}, x_{2ij})'$ where x_{1ij} and x_{2ij} respectively denote the number of pixels classified as corn and the number of pixels classified as soybeans in the j-th area segment of the i-th county, and y_{ij} denotes the number of hectares of corn (or soybeans) in the j-th area segment of the i-th county. Further, the errors e_{ij} have a common variance σ_e^2, i.e., $a_{ij} = 1$.

Battese etal (1988) used the EBLUP estimates $\bar{y}_i^{HB}$ of the means $\bar{Y}_i$ and associated second-order correct estimates of *MSE*, where $\bar{Y}_i$ is taken as $\beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i}$ and the population means $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are known from the satellite readings for all the segments in each county i. Table 3.3 reports the EB estimates of corn and standard errors (square root of estimated *MSE*s) of the EB and the survey regression estimates.

Table 3.3 EB Estimates with Standard Errors

County	n_i	$\bar{y}_i^{EB}$	Standard errors	
			EB	Survey regression
1	1	122.2	9.6	13.7
2	1	126.3	9.5	12.9
3	1	106.2	9.3	12.4
4	2	108.0	8.1	9.7
5	3	145.0	6.5	7.1
6	3	112.6	6.6	7.2
7	3	112.4	6.6	7.2
8	3	122.1	6.7	7.3
9	4	115.8	5.8	6.1
10	5	124.3	5.3	5.7
11	5	106.3	5.2	5.5
12	5	143.6	5.7	6.1

The ratio of the standard error of the EB estimator to that of the survey regression estimator decreases from about 0.97 to 0.77 as the number of sample area segments, n_i, decreases from 5 to 1. The reduction in standard error is considerable when $n_i \leq 3$.

The EB estimates were adjusted to agree with the survey regression estimate for the entire area covering the 12 counties; the latter estimate has relatively small standard error. This adjustment produced a very small increase in the standard error of the small area estimates.

Battese etal. (1988) also reported some methods for validating the assumed model. First, they introduced quadratic terms x_{1ij}^2 and x_{2ij}^2 into the model and tested the null hypothesis that the coefficients of the quadratic terms are zero. The null hypothesis was not rejected at the 5% level. Secondly, they tested the hypothesis that the error term v_i and e_{ij} in the nested error linear regression model are normally distributed, by using the fact that the transformed residuals $(y_{ij} - \hat{\alpha}_i \bar{y}_i) - (\mathbf{x}_{ij} - \hat{\alpha}_i \bar{\mathbf{x}}_i)' \tilde{\boldsymbol{\beta}}$ with $\hat{\alpha}_i = 1 - [\hat{\sigma}_e^2 / (\hat{\sigma}_e^2 + n_i \hat{\sigma}_v^2)]^{\frac{1}{2}}$ are independent normal with mean 0 and variance σ_e^2 under the null hypothesis. Th well-known Shapiro-Wilk W statistic, applied to the transformed residuals, gave values of 0.985 and 0.957 for corn and soybeans respectively. Under the null hypothesis, the P-values (probabilities of getting values less than those observed) equal 0.921 and 0.299, respectively. Therefore, there is no reason to reject the hypothesis that the errors v_i and e_{ij} in the nested error regression model are normally distributed. A limitation of this test is that the transformation of residuals may mask the effects of individual errors. To study the effect of individual units (ij), we can

examine the standardized EBLUP residuals $\left(y_{ij} - \mathbf{x}_{ij}^T\tilde{\boldsymbol{\beta}} - \tilde{v}_i\right)/\hat{\sigma}_e$ where $\tilde{v}_i$ is the EBLUP estimator of v_i. If the model is valid, the standardized residuals are approximately independent normal with mean 0 and variance 1. Residual plots of the standardized residuals can reveal the effects of individual units. To check for the normality of the v_i's and to detect outlier v_i's, a normal probability plot of the EBLUP estimates may be examined.

3.5 Simulation Study

Rao and Choudhry (1995) studied the relative performances of some direct and indirect estimators using real and synthetic populations. For the real population, a sample of 1678 unincorporated tax filers (units) from the province of Nova Scotia, divided into 18 census divisions, was treated as the overall population. In each census division, units were further classified into four mutually exclusive industry groups. The objective was to estimate total wages and salaries (Y_i) for each nonempty census division by industry group (small areas of interest). We focus on the industry group "construction" with 496 units and average small area size = 27.5. Gross business income, available for all the units, was used as an auxiliary variable (x); overall correlation coefficient between y and x for construction equals 0.64

To make comparisons between estimators under customary repeated sampling, R = 500 samples, each of size n = 149, from the overall population of N = 1678 units were selected by simple random sampling. From each sample, the following estimators were calculated: (1) Post-stratified estimator $(PST): N_i\bar{y}_i$ if $n_i \geq 1; = 0$ if $n_i = 0$, where N_i and n_i are the population and sample sizes in the i-th area (n_i is a random variable). (ii) Ratio synthetic estimator $(SYN): (\bar{y}/\bar{x})X_i$, where $\bar{y}$ and $\bar{x}$ are the overall sample means in the industry group and X_i is the x – total for the i-th area. (iii) Sample size dependent estimator (SSD): (2.15) with $\delta = 1$. (iv) EBLUP estimator using the nested error regression model (3.15) with $\mathbf{x}_{ij}^T\boldsymbol{\beta} = \beta x_{ij}$ and $a_{ij} = x_{ij}^{\frac{1}{2}}$. To examine the aptness of this model, the model was fitted to the 496 population pairs (y_{ij}, x_{ij}) from the construction group and the standardized EBLUP residuals $\left(y_{ij} - \tilde{\beta}x_{ij} - \mathbf{v}_i\right)/\left(\hat{\sigma}_e x_{ij}^{\frac{1}{2}}\right)$ were examined. A plot of these residuals against x_{ij} indicated a reasonable but not good fit in the sense that the plot revealed an upper shift with several values larger than 1.0 but none below -1.0. Several variations of the model, including a model with an intercept item, did not lead to better fits.

For each estimator and industrry group, the values of average absolute relative bias $\left(\overline{ARB}\right)$, the average ralative efficiency $\left(\overline{ARE}\right)$ and the average absolute relative error $\left(\overline{ARE}\right)$ were calculated:

$$\overline{ARB} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{500}\sum_{r=1}^{500}(est_r/Y_i - 1)\right|$$

$$\overline{EFF} = \left\{\overline{MSE}(PST)/\overline{MSE}(est)\right\}^{\frac{1}{2}}$$

$$\overline{ARE} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{500}\sum_{r=1}^{500}\left|\frac{est_r}{Y_i} - 1\right|,$$

where the average is taken over m=18 census divisions in the industry group (m=16 for accommodation). Here est_r denotes the values of the estimator for the r-th simulated sample (r = 1,2,...,500), Y_i is the true small area total and

$$\overline{MSE}(est) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{500}\sum_{r=1}^{500}(est_r - Y_i)^2 ;$$

$\overline{MSE}(PST)$ is obtained by changing est_r to $POST_r$, the value of the post-stratified estimator for the r-th simulated sample. Note that $\overline{ARB}$ measures the bias of an estimator while $\overline{ARE}$ and $\overline{EFF}$ both measure the accuracy of the estimator.

Table 3.4 reports the percent values of $\overline{ARB}$, $\overline{EFF}$ and $\overline{ARE}$ for the construction group. It is clear from Table 3.4 that SYN and EBLUP perform significantly better than PST and SSD in

Table 3.4 Comparison of Estimators: Construction

	Estimator			
Measure	PST	SYN	SSD	EBLUP
$\overline{ARB}$%	5.4	15.7	2.9	11.3
$\overline{EFF}$%	100.0	232.8	137.6	261.1
$\overline{ARE}$%	32.2	16.5	24.0	13.5

terms of $\overline{EFF}$ and $\overline{ARE}$, leading to larger $\overline{EFF}$ values and smaller $\overline{ARE}$ values; for example, $\overline{EFF}$ for EBLUP is 266.1% compared to 137.6% for SSD. In terms of $\overline{ARB}$, SYN has the largest value (15.7%) as expected followed by the EBLUP estimator with $\overline{ARB}$ = 11.3; PST has smaller $\overline{ARB}$ (5.4%). Overall, EBLUP is somewhat better than SYN ($\overline{EFF}$ value of 261.1% versus 232.8% and $\overline{ARE}$ value of 13.5 % versus 16.5%). It is gratifying that EBLUP under the assumed model performed well despite the not so good fit.

Rao and Choudhry (1995) also compared the estimators using a synthetic population generated from the assumed model with the real population x-values; the parameter values used were the estimates obtained by fitting the model to the real population pairs (y_{ij}, x_{ij}): $\beta = 0.21, \sigma_v^2 = 1.58$ and $\sigma_e^2 = 1.34$. As expected, EBLUP and SYN performed even better because the synthetic population was generated from the assumed model; a plot of the standardized EBLUP residuals obtained by fitting the model to the synthetic population showed an excellent fit as expected. Rao and Choudhry (1995) also made conditional comparisons of the estimators by conditioning on the realized sample sizes in the small areas. This is a more realistic approach because the domain sample sizes, n_i, are random with known distributions. To make conditional comparisons under repeated sampling, they first selected a simple random sample of size $n = 419$ to determine the sample sizes, n_i, in the small areas. Treating n_i as fixed, 500 stratified random samples were then selected treating the small areas as strata, and the conditional values of $\overline{ARB}$, $\overline{EFF}$ and $\overline{ARE}$ were computed. The conditional performances were similar to unconditional performances, but different when two separate values for each measure were computed by averaging first over areas with $n_i < 6$ only and then over areas with $n_i \geq 6$; $\overline{EFF}\left(\overline{ARE}\right)$ for EBLUP much larger (smaller) than the value for SSD when the domain sample sizes is small (<6). Rao and Choudhry (1995) also demonstrated that the SSD does not take advantage of the between area homogeneity, unlike EBLUP. They generated a series of synthetic populations using the previous parameter values $\beta = 0.21, \sigma_v^2 = 1.58$ and $\sigma_e^2 = 1.34$ and the model $y_{ij} = \beta x_{ij} + v_i \theta^{\frac{1}{2}} + e_{ij} x_{ij}^{\frac{1}{2}}$ by varying the parameter θ from 0.1 to 10 ($\theta = 1$ corresponds to the previous synthetic population). Note that for a given ratio σ_v^2/σ^2, the between area homogeneity increases as θ decreases. Table 3.5 reports the unconditional values of $\overline{EFF}$ and $\overline{ARE}$ for the estimators SSD and EBLUP, as θ varies from 0.1 to 10. It is clear from Table 3.5 that $\overline{EFF}$ and $\overline{ARE}$ for SSD remain essentially unchanged as θ increases from 0.1 to 10.

On the other hand, $\overline{EFF}$ for EBLUP is largest when $\theta = 0.1$ (i.e., when the between area is very small relative to the within-area variation) and decreases as θ increase to 10 (i.e., when the between area variation is large relative to the within-area variation). Similary, $\overline{ARE}$ for EBLUP is the smallest when $\theta = 0.1$ and increases as θ increase to 10.

Table 3.5 Comparison of Estimators: Synthetic Population

Estimator	θ-value					
	0.1	0.5	1.0	2.0	5.0	10.0
			$\overline{EFF}\%$			
SSD	136.0	136.0	135.8	135.6	134.7	133.1
EBLUP	324.3	324.6	319.1	305.0	270.8	239.9
			$\overline{ARE}\%$			
SSD	25.6	25.7	25.9	26.3	27.2	28.2
EBLUP	11.5	11.6	11.8	12.5	14.5	16.7

4 EXTENSIONS

Various extensions of the basic area-level and unit-level models have been studied in the literature. We provide brief account of some recent extensions.

4.1 Area-level models

Extensions of the basic area level model (3.3) include multivariate and time series models and models for disease mapping.

Multivariate models

Datta et al (1996) studied multivariate area level models with a vector of parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \theta_{i3})^T$, where θ_{i1}, θ_{i2} and θ_{i3} denote the true median incomes of four-, three-, and five-person families in U.S. state (small area) i, and θ_{i1}'s are the parameters of interest; estimates of θ_{i1} were used for administering an energy assistance program to low-income families. Adjusted census median income and base-year census median income for the three groups were used as explanatory variables. Direct survey estimates $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ and associated sampling covariance matrix were obtained from the Current Population Survey. HB estimates of θ_{i1}'s, denoted by HB^3, were obtained from the trivariate model and compared to the direct estimates and univariate and bivariate HB estimates, HB^1 and HB^2, treating the 1979 estimates, available from the 1980 census data, as true values. In terms of absolute relative error averaged over the states ($\overline{ARE}$) the three HB estimates performed similarly, outperforming the direct estimates. In this application, the univariate estimates HB^1 worked well in terms of $\overline{ARE}$ and, therefore, it is not necessary to use more complicated estimates based on multivariate models; see Table 4.1 where HB^{2a} and HB^{2b} refer to bivariate models with $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T$ and $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i3})^T$ respectively.

Table 4.1 Average Absolute Relative Error (%)

HB^1	HB^{2a}	HB^{2b}	HB^3
2.07	2.04	2.06	2.02

In terms of standard errors (square root of posterior variances), HB^{2b} obtained from the bivariate model with $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i3})^T$ performed better than the other HB estimates.

Time series models

Suppose θ_{it} denotes a parameter of interest for small area i at time t and $\hat{\theta}_{it}$ is a direct estimator of θ_{it}. A sampling model assumes that the vector $\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_{i1},...,\hat{\theta}_{iT})^T$, given θ_{it}'s, has a multivariate normal distribution with mean $\boldsymbol{\theta}_i = (\theta_{i1},...,\theta_{iT})^T$ and known covariance matrix $\boldsymbol{\psi}_i$, where T is the current time period. A linking model assumes

$$\theta_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta} + v_i + u_{it}, \tag{4.1}$$

where v_i's are independent normal with common mean 0 and common variance σ_v^2, and u_{it} follows either a first order autoregressive model, $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, $|\rho| < 1$ (Rao and Yu, 1994) or a random walk model $u_{it} = u_{i,t-1} + \varepsilon_{it}$ (Datta, Lahiri and Maiti, 1999) where ε_{it}'s are independent of v_i's and normal with mean 0 and common variance σ^2. Linking models of the form (4.1) have been iextensively studied in the econometric literature.

Datta, Lahiri and Maiti (1999) used EB estimators to estimate median income of four-person families by U.S. states, using time series and cross-sectional data $\{\hat{\theta}_{it}, \mathbf{x}_{it}\}$. They employed the linking model (4.1) with a random walk model on the u_{it}'s. Using the 1979 estimates available from the 1980 census data as the true values, they compared the EB (EBLUP) estimates with the HB estimates and CPS direct estimates. The EB estimates were obtained using REML estimates of model parameters. In terms of absolute relative error, avarged over the states, EB performed better than HB and both EB and HB performed nuch better than the direct estimates. Table 4.2 gives the distribution of coefficient of variation (CV) over the states for the three estimates. It is clear from Table 4.2, that in terms of coefficient of variation, EB again performs better than HB and direct estimates.

Table 4.2 Distribution of coefficient of variation (%)

	C.V.		
Est.	2 – 4 %	4 – 6 %	≥ 6%
CPS	6	7	38
HB	10	37	4
EB	49	2	0

Disease mapping

Area local models have also been used in the context of disease mapping or estimating regional mortality and disease rates. A simple model assumes that (*i*) the observed small area counts y_i, given the true incidence rate λ_i, are independent Poisson variables with parameters $n_i\lambda_i$, where n_i is the number exposed in area *i*; (*ii*) The true rates λ_i are independent and identically distributed as gamma variables $G(a,b)$. Maiti (1998) used $\beta_i = \ln\theta_i$ and assumed that β_i's independent and identically distributed as normal with mean μ and common variance σ^2. He also considered a spatial dependence model for β_i's, using conditional autoregression (CAR) that relates each β_i to a set of neighborhood areas of area *i*. He used the HB method to estimate the lip cancer incidence in Scottland for each of 56 counties. The HB estimates of λ_i's were very similar for the two models, but the standard errors (square root of posterior variances) were smaller under the spatial dependence model.

4.2 Unit-level models

Extensions of the basic uni-level model (3.20) include two-level, multivariate and logistic linear mixed models.

Two-level models

Moura and Holt (1999) generalized the basic unit-level model by allowing some or all of the regression coefficients to be random and to depend on area level auxiliary variables, thus effectively integrating the use of unit level and area level covariates into a single model. They obtained EBLUP estimators and associated second-order correct estimators of MSE. They applied the results to data from a sample of 951 retail stores in southern Brazil classified into 73 small areas. Comparison of standard errors with those under the nested error regression model demonstrated improved efficiency from two-level models. You and Rao (1999) applied HB methods to the Brazilian data under three different two-level models: (1) equal error variances as in Moura and Holt (1999), (2) random error variances, (3) unequal error variances. Bayesian diagnostics revealed that model (3) fits the data better than models (1) and (2).

Multivariate models

Datta, Day and Basawa (1999) extended the basic unit-level model to the multivariate case with vector responses $(y_{1ij},...,y_{qij})$. This extension leads to a multivariate nested error regression model. They conducted a simulation study using the sample sizes and auxiliary variable values given in Application 3.4 (Battese, Harter and Fuller, 1988). Further, they estimated the model parameters for the multivariate model using the actual data $\{y_{1ij}, y_{2ij}, x_{1ij}, x_{2ij}\}$, where x_{1ij} and x_{2ij} are as before and $y_{1ij}(y_{2ij})$= number of hectares of corn

(soybeans) in the j-th area segment of the i-th county. Treating the estimated parameters as true values, they simulated samples from the model and showed that the multivariate approach can achieve substantial improvement over the univariate approach in terms of efficiency.

Logistic linear mixed models

Logistic linear mixed models have been extensively used for the case of binary response y_{ij} (0 or 1). The sampling model assumes that y_{ij}'s, for given θ_{ij}'s, are independent Bernoulli variables with parameters θ_{ij}. The linking model is a logistic regression model $\{\theta_{ij}/(1-\theta_{ij})\} = \mathbf{x}_{ij}^T + v_i$ with v_i's independent and identically distributed as normal with mean 0 and common variance σ_v^2. Jiang, Lahiri and Wan (1999) obtained EB estimators and associated jackknife standard errors by the method outlined in Section 3.2. Malec et al (1997) used the HB approach to estimate proportions for demographic groups within U.S. states, using data from the National Health Interviews Survey. For one of the binary variables observed for respondents to the 1990 census long form, they compared the estimates from alternative methods and models with the very accurate census estimates of true values.

5 CONCLUSIONS

We briefly discussed, in section 1.4, survey design issues that have an impact on small area statistics. Preventive measures, such as those outlined in section 1.4, may reduce the need for indirect estimates significantly. But, for many applications sample sizes in some domains of interest may not be large enough to provide adequate precision even after taking such measures. As noted in section 1, sometimes the survey is deliberately designed to oversample specific domains at the expense of small samples or even no samples in other domains (or areas) of interest.

We have provided a brief account of model-based small area estimation. The methodological developments and applications are both impressive, but it is necessary to exercise caution in using model-based methods because of the underlying assumptions. Good auxiliary information related to the variables of interest plays a vital role in model-based inference. As noted by Schaible (1996), expanded access to auxiliary information through coordination and cooperation among federal agencies is needed.

Model validation also plays an important role in model-based estimation. We have presented some methods for model validation in Part 3 and illustrated their application, but the available methods for handling models with random effects are not as extensive as those used for the standard regression models with only fixed effects. More work on model diagnostics for random effects models is needed.

Area-level models have wider scope than the unit-level models because area-level auxiliary information is more readily available than unit-level auxiliary data. But the assumption of known sampling variances, ψ_i, is quite restrictive, although the methods used in the applications (section 3.2) seem to be promising. It should be noted that errors in estimating ψ_i do not affect the model-unbiasedness of EBLUP (EB) estimators provided the mean of θ_i in the linking model (3.2) is correctly specified. But the efficiency of the estimators is affected as well as the validity of the *MSE* estimators. More work on obtaining good approximations to the sampling variances is needed. This task becomes more difficult when using multivariate and time series area-level models because sampling covariances are also needed.

The hierarchical Bayes (HB) approach is a powerful method for small area estimation because it can handle complex problems and the inferences are "exact". But, as noted in section 3.3, caution should be exercised in the choice of improper priors on the model parameters.

We focussed on indirect estimation of small area totals or means, but such estimators may not be suitable if the objectives is to identify domains (or areas) with extreme population values or to rank domaine or to identify domains that fall below or above some prespecified level. Ghosh and Rao (1994) reviewed some methods for handling the latter cases.

Finally, we should emphasise the need for developing an overall program that covers issues relating to sample design and data development, organization and dissemination, in addition to those pertaining to methods of estimation for small areas.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **92**, 1555-1562.

Brooks, S.P. (1998) Markov Chain Monte Carlo method and its application. *The Statistician*, **47**, 69-100.

Cowles, M.K. and Carlin, B.P. (1996) Markov Chain Monte Carlo convergence diagonstics: a comparative review. *Journal of the American Statistical Association*, **91**, 883-904.

Cressie, N. (1992) REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, **18**, 75-94.

Datta, G.S., Ghosh, M., Nangia, N. and Natarajan, K. (1996) Estimation of median income of four-person families: a Bayesian approach In: *Bayesian Analysis in Statistics and Econometrics* (Berry, D.A., Chaloner, K.M. and Geweke, J.K. Eds.). Wiley, New York, pp.129-140.

Datta, G.S., Lahiri, P. and Maiti, T. (1999) Empirical Bayes estimation of median income of four-person by states using time series and cross-sectional data. *Techical Report*, Department of Statistics, University of Georgia-Athens.

Datta, G.S., Day, B. and Basawa, I. (1999) Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, **75**, 269-279.

Deville, J.-C. and Sarndal, C.E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Dick, P. (1995) Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, **21**, 45-54.

Draper, N.R and Smith, H. (1981) *Applied Regression Analysis*. Wiley, New York.

Drew, D., Singh, M.P. and Choudhry, G.H. (1982) Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, **8**, 17-47.

Efron, B. (1975) Biased versus unbiased estimation. *Advances in Mathematics*, **16**, 259-277.

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

Falorsi, P.D., Falorsi, S. and Russo, A. (1994) Empirical comparison of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology*, **20**, 171-176.

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 American Statistical Association*, **74**, 269-277.

Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: an appraisal. *Statistical Science*, **9**, 55-93.

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

Gonzalez, J.F, Placek, P.J. and Scott, C. (1996) Synthetic estimation in followback surveys at the National Center for Health Statistics. In *Indirect Estimators in U.S. Federal Programs* (Schaible, W.L. Ed.), Springer, NewYork, pp. 16-27.

Govindarajulu, Z. (1999) *Elements of Sampling Theory and Methods.* Prentice Hall, Upper Saddle River, NJ.

Griffiths, R. (1996) Current Population Survey small area estimation for congressional districts. *Proceedings of Section on Survey Research Methods*, American Statistical Association, pp. 314-319.

Hobert, J.P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461-1479.

Jiang, J. (1996) REML estimation: asymptotic behaviour and related topics. *Annals of Statistics*, **24**, 255-286.

Jiang, J., Lahiri, P.A. and Wan, S. (1998) Jackknifing the mean squared error of empirical best predictor. *Techical Report*, Department of Statistics, Case Western Reserve University.

James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, Berkeley, pp. 361-379.

Kalton, G., Kordos, J. and Platek, R. (Eds.) (1993) *Small Area Statistics and Survey Designs Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion.* Central Statistical Office, Warsaw.

Kish, L. (1990) Rolling samples and censuses. *Survey Methodology*, **16**, 63-71.

Kott, P. (1989) Robust small area estimation using random effects modelling. *Survey Methodology*, **15**, 3-12.

Lahiri, P.A. and Rao, J.N.K. (1995) Robust estimation of mean squares error of small area estimators. *Journal of the American Statistical Association*, **82**, 758-766.

Maiti, T. (1998) Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, **69**, 339-348.

Malec, D., Sedransk, J., Moriarity, C.L. and LeClere, F. (1997) Small area inference for binary variables with National Health Interview Survey. *Journal of the American Statistical Association*, **92**, 815-826.

Marker, D.A. (1999) Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. Paper presented at the Internal Conference on Small Area Estimation, Riga, Latvia.

Morganstein, D.A. (1998) The Replication Method for Estimating Sampling Errors. Book no. 37, EUSTAT – The Basque Statistical Institute.

Moura, F. and Holt, D. (1999) Small area estimation using multilevel models. *Survey Methodology*, **25**, 73-80.

National Center for Health Statistics (1968) Synthetic State Estimates of Disability. P.H.S. Publication 1759. U.S. Government Printing Office, Washington, D.C.

National Research Council (1998) *Small Area Estimation of School-Age Children in Poverty.* Interim Report 2, National Research Council, Washington, D.C.

Platek, R. and Singh, M.P. (1986) *Small Area Statistics: Contributed Papers.* Laboratory For Research in Statistics and Probablility, Carleton University.

Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P. (1987) *Smalll Area Statistics.* Wiley, New York.

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

Prasad, N.G.N. and Rao, J.N.K. (1999) On robust estimation using a simple random effects model. *Survey Methodology*, **25**, 67-72.

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). *International Statistical Review*, **48**, 3-18.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **26** (in press).

Rao, J.N.K. and Choudhry, G.H. (1995) Small area estimation: overview and empirical study. In: *Business Survey Methods* (Cox, B.G. etal. Eds.), Wiley, New York, pp. 527-542.

Rao, J.N.K. and Yu, M. (1994) Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, **22**, 511-528.

Schaible, W.L. (Ed.) (1996) *Indirect Estimators in U.S. Federal Programs*. Lecture Notes in Statistics no. 108, Springer, New York.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994) Issues and strategies for small area data. *Survey Methodology*, **20**, 3-22.

You, Y. and Rao, J.N.K. (1999) Hierarchical Bayes estimation of small area means using multi-level models. Proceedings of IASS Satellite Conference on Small Area Estimation, 171-185, Riga, Latvia.