

Joycelyn Francisco

Small Area Estimation :
An Overview of Existing Methodologies
with Application to the Estimation
of Unemployment Rates in the Philippines

Essai
présenté
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de maître ès sciences (M. Sc.)

Département de mathématiques et de statistiques
FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL

Juillet 2003

Foreword

The completion of this thesis would not have been possible were it not for the support and generosity of several people. First and foremost, I would like to express my deep gratitude to M. Louis-Paul Rivest, my research director. He has been generous both with his time and invaluable advices.

The data for chapter 6 of this thesis came from the National Statistics Office of the Philippines. I am especially grateful to Ms. Carmelita Ericta, the Executive Director, and likewise to her staff. A million thanks go to my friend Ms. Monina Collado of the same office who had been very patient with all my demands.

I would also like to extend my heartfelt thanks to those friends who have been friends indeed. They have given me the moral support to persevere when the going was rough. I would like to thank, too, those who thought it is crazy to go back to school after so many years and who, in so doing, have pushed me to prove them wrong.

Finally, I want to thank my family who has endured a diminished standard of living these past two years. They have been patient, forgiving and understanding. They have given me the moral support and the push to go on and finish what I have started. To them, I dedicate this work.

Résumé

As countries move from away from centralized planning and decision-making, and with the increasing concern with issues of distribution, equity and disparity, the demand for small area statistics has been growing. The response to this has been two-pronged: i) the use of survey designs that can produce statistics at the small area level, and ii) the use of non-conventional (i.e., indirect) estimation methods to produce statistics at the small area level. With limited resources available to survey planners and producers of statistics, the more popular approach has been the latter. This paper presents an overview of the existing small area estimation techniques: direct, indirect and model based. Illustrative examples are given for each of the estimation techniques discussed. In each case, comparisons are made in terms of the efficiency of the resulting small area estimates. A case study is presented at the end of the paper to illustrate the potential use of the empirical Bayes technique in estimating unemployment rates at the region/age group level in the Philippines.

Table of Contents

Résumé	ii
Foreword	iii
List of figures	vii
List of tables	viii
1 Introduction	1
2 Direct estimators	5
2.1 Statistical properties	5
2.2 Applications	7
2.2.1 Estimation of total wages and salaries	7
2.2.2 Direct estimates of the average number of doctor visits, by State (NHIS)	11
3 Synthetic estimators	13
3.1 Statistical properties	13
3.2 Applications	18
3.2.1 Synthetic estimates of the proportion of uninsured adults	18
3.2.2 Synthetic estimates of unemployment	19
3.2.3 Synthetic estimates of total wages and salaries	21
3.2.4 Synthetic estimates of the average number of doctor visits (NHIS)	22

4	Composite estimators	23
4.1	Statistical properties	23
4.2	Applications	26
4.2.1	Composite estimates of total wages and salaries	26
4.2.2	Small area estimates of employment	27
4.2.3	Composite estimates of the average number of doctor visits (NHIS)	29
5	Bayesian estimators	33
5.1	Basic Bayesian concepts	33
5.2	Bayesian modeling for small areas	36
5.3	Empirical Bayes	38
5.3.1	Statistical properties	38
5.3.2	Applications	42
5.3.2.1	Empirical Bayes estimates of income for small places	42
5.3.2.2	Empirical Bayes estimates of net undercoverage in the 1991 Canadian Census	47
5.4	Hierarchical Bayes	50
5.4.1	Statistical properties	50
5.4.2	Application: The 1991 Canadian Census undercoverage	56

6 Case study: Small area estimates of unemployment in the Philippines for April 1994	59
6.1 Definition of terms	59
6.2 Notations	62
6.3 Available data	62
6.4 Model selection	67
6.5 Results and analysis	74
6.6 Conclusion	83
7 Bibliography	84
8 Appendices	89
Appendix 1 The Splus program for the Shapiro-Wilks test	89
Appendix 2 The Splus program for scenario 2	90

List of figures

1	Histogram of the direct estimators	70
2	Histogram of the EB estimators for Scenario 1	70
3	Histogram of the EB estimators for Scenario 2	70
4	Scenario 1: Normalized residuals as a function of predicted values	71
5	Scenario 2: Normalized residuals as a function of predicted values	71

List of tables

2.2.1.1	Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Direct Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia	10
3.1.1	Distribution of the RMSE of Synthetic Estimates by Counties by Size of 1970 Census Unemployment Rate	20
3.2.3.1	Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Synthetic Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia	21
4.2.1.1	Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Composite Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia	27
4.2.2.1	Population Count for the 1999 Labor Force Survey	28
4.2.2.2	Comparison of the Small area Estimators	29
4.2.3.1	Average Number of Doctor Visits, by State	31
4.2.3.2	Variation in State estimates (Range) for average number of doctor visits	32
5.3.1	Comparison of Selected PCI Estimates with 1973 Special Census Values of 1972 PCI	46
5.3.2	Direct and Empirical Bayes Estimates of Adjustment Factors	49
5.4.1	1991 Canadian Census Undercount Estimates and Associated CV's	58

6.1	Percentage Distribution of the Gross Domestic Product at Constant (1985) Prices, 1999 and 2000	61
6.2	Employment Status of the Labour Force Population, Standard Error and Coefficient of Variation of the Unemployment Estimate, by Region and Age Group: April 1994	64
6.3	Age Groups Ranked by Unemployment Rates, By Region	73
6.4	Regions Ranked by Unemployment Rates, By Age Group	74
6.5	Regression Coefficients, Standard Errors, t-statistics and p-values	75
6.6	Unemployment Rates: Direct estimates, Empirical Bayes Estimates and Variance of the Direct Estimates for the Small Areas	77
6.7	Mean Square Errors of the Empirical Bayes Estimates of Unemployment Rates and their Efficiency	81

1 Introduction

1.1 General

The term "small area" usually denotes subsets of the population. These subsets may refer to geographic subdivisions (i.e., small geographical area) such as a county, a municipality or a city or a census division. It may also be used to describe demographic subdivisions (i.e., small subpopulations) such as a specific age-sex-race group of people within a large geographical area. It may also denote a cross classification of a small geographic area and a specific demographic or industrial group. In small area estimation, the interest is usually on the estimation of some attribute of the "small area" or "small domain" such as a mean, a total or a proportion. In this paper, the terms "small area" and "small domain" are used interchangeably.

There has been an increasing demand for small area statistics throughout the world from both the public and private sectors. Since the 1990's, as countries move away from centralized planning and decision-making, and with increasing concern with issues of distribution, equity and disparity, the need for reasonable and accurate estimates of local economic and demographic conditions has been growing. For example, as part of the decentralization process in many countries, national governments have been transferring responsibilities for many social and economic plans and programs to the local governments (e.g., provinces, states, municipalities and cities). Evaluating the success of these local plans and programs requires corresponding small area statistics. The private sector, likewise, needs small area statistics since the policy making of many businesses and industries relies

on local socio-economic conditions. Feasibility studies, for example, require the use of small area statistics.

Small area estimates can be made available from various censuses of population, businesses, housing and agriculture. However, the demand for small area estimates also exists for the intercensal periods when data usually come from sample surveys. In general, sample surveys, whether they are conducted by government statistical offices or by private entities, are designed to produce reasonably accurate direct estimates of characteristics for specific domains of the population in addition to statistics for the whole population; they are not meant to produce reliable estimates for subsets of the population. With the increasing interest on small area statistics, survey organizations are then faced with producing estimates (i.e., of sufficient efficiency) from existing sample surveys for subsets of the population. Unfortunately, sample sizes in small areas tend to be too small, sometimes non-existent, to provide reliable direct estimates for these specific small domains. In other words, for small domains, the direct estimates are too unstable to be used for planning and policy-making purposes as they are likely to produce unacceptably large standard errors due to an unduly small sample size for the small area. Accurate direct estimates for small areas would require a substantial increase in the overall sample size which in turn could overwhelm an already constrained budget and which could further lengthen the data processing time. Consequently, there has been a growing interest in developing a range of estimation techniques to answer this need for small area statistics without further burdening the resources of already constrained survey organizations. The recent book of Rao (2003) reviews many of these developments.

This paper attempts to describe the available small area estimation techniques. Illustrative examples or applications are likewise presented for each of these techniques. These techniques may be classified as direct, indirect

or model-based. Indirect estimators may be further classified as synthetic or composite. Under model-based estimators are the empirical Bayes estimators and the hierarchical Bayes estimators.

1.2 Notations

In the following discussions on these different estimation techniques, a population of size N is considered from which a sample of size n is drawn. We denote by y the attribute or characteristic of interest. We assume that there are m small geographic subdivisions and l socio-demographic or industrial subgroups. Thus, y_{ihk} denotes the value of the characteristic of interest y on the k^{th} unit in the h^{th} socio-demographic or industrial subgroup in the i^{th} small geographic area, where $i = 1, 2, \dots, m$, $h = 1, 2, \dots, l$ and $k=1, 2, \dots, n_{ih}$. The term n_{ih} denotes the sample size in the $(ih)^{th}$ cross classification and

$$\sum_i \sum_h n_{ih} = n \quad .$$

N_{ih} denotes the population size in the $(ih)^{th}$ cross classification. Similarly, we define the total number of units or individuals in the population as

$$\sum_i \sum_h N_{ih} = N \quad .$$

When $n_{ih} = 0$, we define

$$\sum_k y_{ihk} = 0.$$

The usual dot summation is adopted; thus,

$$\sum_h \sum_k y_{ihk} = y_{i\bullet\bullet} \quad .$$

Depending on the purpose of a particular study, a small area could mean a small geographic area i for which $l = 1$; a demographic or industrial subgroup h for which $m = 1$; or a demographic or industrial subgroup h in a small geographic area i . It should be emphasized that these subdivisions, whether they are geographic, socio-demographic or industrial, are mutually exclusive and exhaustive population cells. For the sake of convenience, the following notations are used in this work to denote, respectively, the characteristic of interest for the k^{th} sampling unit in the small area, the sample size and population size for the small area:

1. y_{ik}, n_{ik}, N_{ik} when the small area is a small demographic area,
2. y_{hk}, n_{hk}, N_{hk} when the small area is a socio-demographic subgroup,
3. $y_{ihk}, n_{ihk}, N_{ihk}$ when the small area is a cross classification of a small demographic area and a socio-demographic subgroup.

In addition, an auxiliary variable is denoted by X . Such variable is indexed in the same manner as the variable of interest Y .

In this paper, the discussion centers on two (2) characteristics of interest:

1. the mean for the small area (i.e., \bar{Y}_i, \bar{Y}_h or \bar{Y}_{ih}), with its corresponding estimator (i.e., $\hat{Y}_i = \bar{y}_i, \hat{Y}_h = \bar{y}_h$ or $\hat{Y}_{ih} = \bar{y}_{ih}$) and
2. the total for the small area. (i.e., Y_i, Y_h or Y_{ih}), with its corresponding estimator (i.e., \hat{Y}_i, \hat{Y}_h or \hat{Y}_{ih}).

2 Direct Estimators

2.1 Statistical properties

Direct estimates use data only from sample units in the area or domain of interest. Let us define a small area as a small geographic area i . For the characteristic *total* Y_i , a simple expansion estimator is given by:

$$\begin{aligned}\hat{Y}_i^{\text{exp}} &= \frac{N}{n} \sum_{k=1}^{n_i} y_{ik} && \text{if } n_i \geq 1, \\ &= 0 && \text{otherwise .}\end{aligned}\quad (2.1)$$

When the N_i 's are known, a second but more efficient estimator for Y_i may be used. This is the post-stratified estimator and is given by:

$$\begin{aligned}\hat{Y}_i^{\text{pst}} &= \frac{N_i}{n_i} \sum_{k=1}^{n_i} y_{ik} = N_i \bar{y}_i && \text{if } n_i \geq 1 \\ &= 0 && \text{otherwise .}\end{aligned}\quad (2.2)$$

A direct estimator for the mean for the small area i is given by:

$$\begin{aligned}\hat{\bar{Y}}_i &= \bar{y}_i = \frac{\sum_{k=1}^{n_i} y_{ik}}{n_i} && \text{if } n_i \geq 1 \\ &= 0 && \text{otherwise}\end{aligned}\quad (2.3)$$

The estimators (2.2) and (2.3) are conditionally unbiased for a fixed $n_i \geq 1$. This is shown below:

$$\begin{aligned}E(\bar{y}_i) &= E_{n_i} E_y(\bar{y}_i | n_i) && \text{Note: } E(y) = E_x E_y(y|x) \\ &= E(\bar{Y}_i) \\ &= \bar{Y}_i\end{aligned}$$

$$\begin{aligned}
E(\hat{Y}_i^{pst}) &= N_i E(\bar{y}_i) \\
&= N_i \bar{Y}_i \\
&= Y_i
\end{aligned}$$

The conditional variance for \bar{y}_i , assuming that the n_i 's are fixed is given by:

a) if N_i is known

$$\begin{aligned}
\text{Var}(\bar{y}_i) &= \frac{1-f_i}{n_i} S_{yi}^2 \quad (2.4) \\
&= \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{yi}^2,
\end{aligned}$$

where $f_i = \frac{n_i}{N_i}$, and

$$S_{yi}^2 = \sum_{k=1}^{N_i} \frac{(y_{ik} - \bar{y}_i)^2}{N_i - 1}, \quad N_i \geq 2.$$

An unbiased estimator for S_{yi}^2 is given by:

$$s_{yi}^2 = \sum_{k=1}^{n_i} \frac{(y_{ik} - \bar{y}_i)^2}{n_i - 1}. \quad (2.5)$$

Hence, $\text{Var}(\bar{y}_i)$ may be estimated by:

$$v(\bar{y}_i) = \frac{1-f_i}{n_i} s_{yi}^2 = \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_{yi}^2. \quad (2.6)$$

b) if N_i is unknown

The sampling fraction f_i is replaced by f in formula (2.4) above, and thus we have:

$$\text{Var}(\bar{y}_i) = \frac{1-f}{n} S_{yi}^2,$$

$$\text{where } f = \frac{n}{N} .$$

Similarly, S_{yi}^2 is estimated by s_{yi}^2 defined in formula (2.5) above.

$\text{Var}(\bar{y}_i)$ may then be estimated as:

$$v(\bar{y}_i) = \frac{1-f}{n_i} s_{yi}^2 .$$

The variance of the estimators of the total \hat{Y}_i^{exp} and \hat{Y}_i^{pst} may be estimated in a similar way (Cochran, 1977).

Direct estimators are generally used when the sample size for each small area is sufficiently large to give reasonably accurate estimates. However, as the sources of data are usually sample surveys designed to give national and regional statistics, sample sizes for the small areas (usually subdomains of the original domains of study) are usually unduly small. Consequently, the associated variances are likely to be unacceptably large since the conditional variances (as can be seen above) are of the order n_i^{-1} . Moreover, if information from a national sample is used to make estimates for small areas and there are no sample units in the small area of interest, then obviously direct estimation cannot be used.

2.2. Applications

2.2.1. Estimation of total wages and salaries (direct estimates)

A simulation study is undertaken to compare small area estimates derived using different methodologies (Rao and Choudry, 1995). In this simulation study, a sample of $N=1678$ unincorporated tax filers from Nova Scotia is considered as the overall population. The province of Nova Scotia is divided into $m=18$ census divisions. For each census division, the

unincorporated tax filers are classified into four (4) mutually exclusive industry groups: retail with 553 units, construction with 496 units, accommodation with 114 units and others with 515 units. The small area of interest in this study is defined to be a nonempty census division by industry group combination. There are 18 nonempty census divisions for each of the industry groups of retail, construction and other; there are 16 nonempty census divisions for the industry group accommodation. Hence, there is a total of 70 (i.e., 18+18+18+16) small areas. The variable of interest Y is total wages and salaries. The notation Y_{ih} refers to total wages and salaries for the h^{th} industry group in the i^{th} census division. For this simulation study, two approaches are taken: 1) unconditional comparisons under repeated sampling and 2) conditional comparisons of the estimators by conditioning on the realized sample sizes in the small areas.

Under the first approach (unconditional comparison under repeated sampling), 500 simple random samples, each of size $n = 419$, are drawn from the overall population of $N = 1678$ unincorporated tax filers in the province of Nova Scotia. For each sample of $n = 419$, two (2) kinds of direct estimators of total wages and salaries are estimated: the simple expansion estimator \hat{Y}_{ih}^{exp} as described in formula (1.1) and the poststratified estimator \hat{Y}_{ih}^{pst} given by formula (2.2). For a given estimator \hat{Y}_{ih} are computed the average relative bias (\overline{ARB}), the average relative efficiency (\overline{EFF}) with respect to \hat{Y}_{ih}^{pst} , and the average absolute relative error (\overline{ARE}). These quantities are defined as follows:

$$\overline{ARB}_h = \frac{1}{m} \sum_{i=1}^m \left| \frac{\frac{1}{500} \sum_{t=1}^{500} (\hat{Y}_{iht}^* - Y_{ih})}{Y_{ih}} \right| ,$$

$$\overline{EFF}_h = \left(\frac{\overline{MSE}(\hat{Y}_{ih}^{pst})}{\overline{MSE}(\hat{Y}_{ih}^*)} \right)^{1/2}, \text{ and}$$

$$\overline{ARE}_h = \frac{1}{m} \sum_{i=1}^m \left\{ \frac{\frac{1}{500} \sum_{t=1}^{500} |\hat{Y}_{iht}^* - Y_{ih}|}{Y_{ih}} \right\},$$

where \hat{Y}_{iht}^* is the estimator of total wages and salaries from the t^{th} random sample from the $(ih)^{th}$ small area obtained using a particular method (e.g., simple expansion) and Y_{ih} is the value of total wages and salaries for the whole population of N=1678 unincorporated tax filers from Nova Scotia, $m = 18$ for the industry groups retail, construction and others and $m = 16$ for the industry group accommodation, and

$$\overline{MSE}(\hat{Y}_h^*) = \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{500} \sum_{t=1}^{500} (\hat{Y}_{iht}^* - Y_{ih})^2 \right).$$

The second approach (i.e., conditional comparisons of the estimators by conditioning on the realized sample sizes in the small areas) is considered by the authors to be a more realistic approach because the small area sample sizes n_i are considered random with known distributions. Under this second approach, a simple random sample of size $n = 419$ was drawn and the sample sizes n_{ih} were thereafter determined. The n_{ih} 's were then treated as fixed. The small areas (i.e., a combination of a nonempty census division by industry group) were treated as strata or subgroups and 500 stratified samples were drawn. As with the first approach, the simple expansion estimator \hat{Y}_i^{exp} described by formula (1.1) and the poststratified estimator \hat{Y}_i^{pst} given by formula (2.2) were calculated for each of the 500 samples drawn. As

in the first approach, for a given estimator \hat{Y}_{ih} were computed the average relative bias (\overline{ARB}_h), the average relative efficiency (\overline{EFF}_h) with respect to $\hat{Y}(pst)$, and the average absolute relative error (\overline{ARE}_h), defined in the same way as in the first approach .

The following table presents the computed percentage values of \overline{ARB}_h , \overline{EFF}_h and \overline{ARE}_h for h denoting the construction industry group in the province of Nova Scotia.

Table 2.2.1.1
Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Direct Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia

Quality Measure	Estimator			
	EXP		PST	
	Approach 1	Approach 2	Approach 1	Approach 2
\overline{ARB}	1.9%	25.6%	5.4%	1.6%
\overline{EFF}	74.8%	87.4%	100.0%	100.0%
\overline{ARE}	44.6%	38.4%	32.2%	32.7%

The total sample size of $n=419$ distributed among 70 small areas gives approximately an average of 6 sample units per small area. In the case of the construction industry, there are $N_{\bullet h} = 496$ units distributed among 18 census divisions giving an average of $N_{ih} = 27$ units per small area.

As can be observed from Table 2.2.1.1, except for the \overline{ARB} result for Approach 1, the PST estimator performs better than the simple expansion estimator. Stratification or subgrouping, when done properly, produces large gains in precision by reducing the variance within the strata or group. In this

case, the poststratified estimator was obtained with the use of the expansion factor N_{ih}/n_{ih} , corresponding to the subgroup rather than N_i/n_i , thus giving a more efficient estimate of total wages and salaries.

In general, with the exception of the result for the second approach for the simple expansion estimator, the \overline{ARB} for the direct estimators, both for the simple expansion and poststratified methods, are relatively small. This gives the impression that these direct estimators may be reasonably accurate. However, the \overline{ARE} are considerably large as they range from 32.2% to 44.6%. These results signify that the sample sizes in the small areas are unduly small; consequently the associated variances are significantly large. One solution to address this situation is to increase the sample sizes at the small area level. One other solution is to use other methodologies.

2.2.2. Direct estimates of the average number of doctor visits by State from the U.S. National Health Interview Survey (NHIS)

The National Health Interview Survey (NHIS) is one of the major surveys undertaken by the National Center for Health Statistics (NCHS) of the United States (NCHS,1999). This survey aims to collect information concerning the health of the U.S. civilian noninstitutionalized population through household interviews throughout the United States.

The NHIS has a three (3)-stage sampling design. A primary sampling unit (PSU) is defined as a group of counties. In the first stage, PSU's are selected using a stratified sampling design to ensure the representation of chosen demographic (e.g., Northeast, Midwest, New England) and economic characteristics. In the second stage, blocks or small groups of blocks are selected as secondary sampling units. In the third stage, within each chosen block, a new cluster of eight (8) housing units is selected each year.

One of the objectives of the NHIS is to provide health statistics for subnational geographic areas – in particular, States. However, an analysis done at the NCHS concluded that State-level statistics from the NHIS have very poor precision for most States. In order to provide reliable State-level health statistics, two studies were done: i) a possible redesign of the NHIS increasing its sample size to improve the precision of State-level statistics, and ii) a study on the potential of various small-area estimation techniques. For the second study, an empirical comparison of small area estimators (i.e., direct estimates, synthetic estimates and composite estimates) that could be used to produce State-level estimates was undertaken using the data collected in the 1988 NHIS.

In this example, the variable of interest y is the number of doctor visits and the small area is defined as a State. The characteristic of interest is the average number of doctor visits \bar{Y} for a particular State i (denoted by \bar{Y}_i). The results are presented in Table 4.2.3.1, with the States arranged in decreasing average number of doctor visits. The direct estimates are simple expansion estimates. It should be noted that for two (2) of the states (North Dakota and Nebraska) no samples were collected and, consequently, no direct estimates exist. The comparison of the different small area estimates for the average number of doctor visits, by State, is made in section 4.2.3.

3 Synthetic estimators

3.1. Statistical properties

The term "synthetic estimates" was first used by the U.S. National Center for Health Statistics (1968) of the United States when it calculated estimates of long and short term physical disabilities based on the National Health Interview Survey. Since then, synthetic estimation has been used to generate small area statistics from a number of surveys. Synthetic estimates of county unemployment rates were produced using regional estimates from the U.S. Current Population Survey (Gonzales and Waksberg, 1973). The Australian Bureau of Census has done empirical studies aimed at generating synthetic estimates of income and work force status for Australian Census Statistical Divisions (Purcell and Linacre, 1976). More recently, small area synthetic estimates of literacy rates, health and morbidity statistics and income have been generated for purposes of local planning.

The method of synthetic estimation has been described by M.E. Gonzales (1973) as follows:

"An unbiased estimate is obtained from a sample for a large area; when this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates."

This is the method of "borrowing information from related subareas in order to increase the effective sample size for estimation and hence the accuracy of the resulting estimates" (Holt, Smith and Tomberlin, 1979).

The method of synthetic estimation assumes the availability of estimates from an inquiry or survey of estimates for a large subset of the population. This large subset may mean a large geographical area (e.g.,

country, province, state) or a demographic group (e.g., age group, sex-group, age-sex group) or an industrial group (e.g., construction industry, retail industry). Appropriate weights are then applied to these large population subset estimates to obtain the desired small area estimates. The censuses are often the sources of these weights.

The synthetic estimation method proposed by Gonzales and Hoza (1978) and elaborated by Holt et al (1979) uses such weights from the census or some other sources of accurate information. In this method, the small area is defined as a small geographic area. It is assumed that reliable direct estimates $\hat{Y}_{\bullet h}$ for each subgroup h are available from a current inquiry. Consider the series of weights p_{ih} such that

$$\sum_h p_{ih} = 1 \quad .$$

From the available estimates $\hat{Y}_{\bullet h}$ the corresponding subgroup mean is computed as:

$$\bar{y}_{\bullet h} = \frac{\hat{Y}_{\bullet h}}{N_{\bullet h}},$$

The weights p_{ih} are then applied to the subgroup means thus obtaining the estimate of the mean for the small area i :

$$\bar{y}_i = \sum_{h=1} p_{ih} \bar{y}_{\bullet h} \quad . \quad (3.1.1)$$

The corresponding estimator of the small area total is:

$$\hat{Y}_i = \sum_h p_{ih} \hat{Y}_{\bullet h} \quad . \quad (3.1.2)$$

If Y_i is the true total, then the average mean squared error of this estimator over all m small areas is:

$$\text{average MSE} = E \left\{ \sum_{i=1}^m (\hat{Y}_i - Y_i)^2 / m \right\}. \quad (3.1.3)$$

Holt et al (1979) suggested the use of population sizes as weights. These population sizes are known and may be obtained from a previous census or some other source of accurate information. The series of weights p_{ih} are obtained as:

$$p_{ih} = \frac{N_{ih}}{N_i} .$$

In general, any suitable auxiliary variable available from a census or some other source of accurate information may be used as weights. Thus, the series of weight p_{ih} may also be expressed as:

$$p_{ih} = \frac{X_{ih}}{X_i} \quad (3.1.4)$$

For example, Gonzales and Hoza used the proportions of the labour force as weights in estimating unemployment rates at the county level.

Purcell and Kish (1979) and Ghosh and Rao (1994) propose a different series of weights p_{ih} which does not sum to 1, *i.e.*, $\left(\sum_h p_{ih} \neq 1 \right)$. The weights p_{ih} are calculated as:

$$p_{ih} = \frac{X_{ih}}{X_{\bullet h}} . \quad (3.1.5)$$

and

$$\sum_h p_{ih} = \sum_h \frac{X_{ih}}{X_{\bullet h}} \neq 1$$

but

$$\sum_i p_{ih} = \sum_i \frac{X_{ih}}{X_{\bullet h}} = \frac{X_{\bullet h}}{X_{\bullet h}} = 1 .$$

Again, the X 's are assumed to be known and available from a census or some other source of accurate information. The resulting estimator \hat{Y}_i has the desired consistency property that $\sum_i \hat{Y}_i$ is equal to \hat{Y}_{\bullet} , the reliable direct estimator for the bigger population subset, i.e.,

$$\hat{Y}_{\bullet h} = \sum_i p_{ih} \hat{Y}_i .$$

The use of these weights imply that the relationship of the small area to the subgroup with respect to the auxiliary variable X , observed during the census year or as shown by the data from some other source of accurate information, holds true for the variable of interest.

Rao and Choudry (1995) suggest the use of a *ratio synthetic estimator*, a modification of the earlier method used by Gonzales and Hoza (1978). Let us define a small area as a small geographic area. The ratio estimator utilises a supplementary *x-data*, or an auxiliary variable, with known or estimated small area totals. Known small area totals for the auxiliary variable, denoted by X_i , may come from a census or some other reliable sources (e.g., government agencies) which gather information as part of their mandate. The data for these covariates may also be estimates, denoted by \hat{X}_i , coming from some

other surveys or inquiries designed to give small area estimates. The assumption is that the population ratios

$$R_i = Y_i / X_i ,$$

Y_i being the value of the characteristic of interest for the i^{th} small area and X_i being the value of the covariate for the same i^{th} small area, are homogeneous. This means that

$$R_i = R = Y/X \quad \text{and} \quad Y_i = R * X_i ,$$

where R , Y and X are the values for the whole population. The value of R is then estimated by

$$\hat{R}_i = \frac{\bar{y}}{\bar{x}} ,$$

where \bar{y} and \bar{x} are the overall sample means. In the case when the data for the auxiliary variable at the small area level is known, the estimator for Y_i is given by

$$\hat{Y}_i = \hat{R}_i * X_i = \frac{\bar{y}}{\bar{x}} * X_i \tag{3.1.6}$$

Otherwise, \hat{Y}_i is obtained by replacing X_i by its estimator \hat{X}_i .

In contrast with the earlier mentioned methods of Gonzales and Hoza (1978) and Purcell and Kish (1979) which assume the availability of a known or estimated total for the variable of interest y at the subgroup level, the method proposed by Rao and Choudry (1995) uses an auxiliary variable x with known or estimated small area totals.

An advantage of the synthetic estimation is its ease of calculation. The variance of the synthetic estimator is of order n^{-1} and, hence, is smaller than that of the post-stratified estimator. However, the synthetic estimates are biased estimates for two (2) reasons. First, the underlying assumption of homogeneity of rates or proportions is often hard to satisfy, i.e., estimated rates for the larger area (for a particular subgroup j) may differ from that of one or more subareas. In other words, the "model assumption" that relations observed in large areas must hold for the small domains may not be always valid. Second, the structure of the population may have changed since the previous census. The synthetic method also fails to account properly for local factors. Unless the grouping variables are highly correlated with the variable of interest, the synthetic estimates will tend to cluster near the mean for the larger area, and fail to reflect the actual effects of local area factors.

3.2 Applications

3.2.1. Synthetic estimates of the proportion of uninsured adults

An example on how small area synthetic estimates may be calculated is given by Suciú et al (2001). The data from the Ohio Family Health Survey (OFHS) are used to provide estimates of "proportion uninsured" for small areas. A small area is defined as a county. In particular, estimates of the proportion of uninsured adults for Marion County are to be calculated. The population of this county is known and is then first divided into eight (8) domains defined by age (<65 , ≥ 65), marital status (married, not married), and annual household income ($<\$30000$, $\geq \$30000$). The proportion of uninsured adults in each of these 8 cells for the whole state of Ohio (the larger area) may be obtained from the OFHS. Suppose that for the whole state of Ohio the proportion of uninsured adults under 65 years of age who are married and with an annual household income of less than \$30000 is 0.21 (as calculated from

the OFHS data). Suppose further that the proportion of the population in the county of Marion that belongs to this group is 0.14. Similar estimates can be obtained for the other seven (7) cells. Using formula (3.1.1) but using the weights as defined by Purcell and Kish, the proportion of uninsured adults in Marion County is then estimated as:

$$\begin{aligned}\bar{y}_i &= \sum_{h=1}^8 p_{ih} * \bar{y}_{\bullet h} \\ &= (0.14)(0.21) + (0.55)(0.04) + \dots + (0)(0) \\ &= 0.108\end{aligned}$$

where \hat{Y}_i is the estimated proportion of uninsured adults in the small area, $P_{\bullet j}$ is the proportion or rate for subgroup j for the larger area, and y_{ij} is the estimated rate or proportion from a survey.

3.2.2. Synthetic estimates of unemployment

Gonzales and Hoza (1978) derived synthetic estimates of unemployment for all counties in the United States using the 1970 Census of Population. In this study, the small area is defined as a county. The subgroup h refers to a particular industry or occupation. The rates p_{ih} are the proportions of the labour force of county i that correspond to cell h (i.e., industry or occupation). For this particular study, estimates are generated for each county for the year 1970 with the objective of comparing the derived synthetic estimates with those of the actual 1970 Census of Population direct estimates. Intercensal estimates are estimated by applying the 1970 census-based p_{ih} on data from the Current Population Survey. The difference between the synthetic estimates and the direct estimates is then used to represent the method error of synthetic estimation. Formula (3.1.1) is used to derive the synthetic estimates of unemployment rates at the county level. The results show that the mean error of the county synthetic estimates for 1970 is

-0.3 per cent. The root mean square error (RMSE) is 1.9 per cent. Table 3.1.1 gives a summary of how the synthetic estimates compare with the direct estimates.

Table 3.1.1
Distribution of the RMSE of Synthetic Estimates
By counties by Size of 1970 Census
Unemployment Rate*

1970 Census unemployment rate (per cent)	Number of counties	RMSE (per cent)	Relative RMSE**
Less than 1.0	21	2.8	5.52
1.0-1.9	171	2.0	1.36
2.0-2.9	493	1.4	0.57
3.0-3.9	679	0.9	0.24
4.0-4.9	580	0.6	0.14
5.0-5.9	363	1.2	0.22
6.0-6.9	232	1.8	0.28
7.0-7.9	137	2.5	0.33
8.0-8.9	88	3.4	0.40
9.0-9.9	51	4.3	0.46
10.0-10.9	30	4.8	0.46
11.0-11.9	22	6.5	0.56
12.0-12.9	23	7.2	0.58
13.0-13.9	10	8.1	0.60
14.0-14.9	2	8.4	0.58
15.0-16.9	6	10.6	0.66
Average=4.2	2908	1.9	0.45

* table taken from Gonzales and Hoza (1978)

** the relative RMSE was calculated by dividing the RMSE by the midpoint of the unemployment rate interval

The average unemployment rate 4.2 % is in the interval 4.0-4.9 and it is in this interval where the minimum RMSE and relative RMSE may be found. The relative RMSE is comparatively high for both the low and high interval of unemployment rates (i.e., the extremes). This would imply that the synthetic

estimation method tends to give values nearer the large area mean than the true small area (county) estimates.

3.2.3. Synthetic estimates of total wages and salaries

We continue with the example of Rao and Choudry (1995) on the estimation of total wages and salaries (Y) by census division by industry group. The small area is defined as a nonempty census division by industry group combination. Synthetic estimates are generated using formula (3.1.6) with gross business income as the auxiliary variable (X). The results for h denoting the construction industry are presented in the following table:

Table 3.2.3.1
Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Synthetic Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia

Quality measure	Approach 1	Approach 2
\overline{ARB}	15.7%	16.8%
\overline{EFF}	232.8%	214.9%
\overline{ARE}	16.5%	17.4%

It is apparent that the results under the two (2) approaches are not very different. These results show that the synthetic estimator performs significantly better than the direct estimators (see Table 2.2.1.1) in terms of the \overline{EFF} and \overline{ARE} . These seemingly better results may be due to the fact that, unlike direct estimators, synthetic estimators take into account the relationship of a particular small area with related small areas by using supplementary information (i.e., the auxiliary data X) which in this case is data on gross business income. As mentioned earlier, synthetic estimators have smaller variances than direct estimators and this fact is reflected in these results. The

reduction in variance, however, is counter-balanced by the bias inherent in synthetic estimates.

3.2.4. Synthetic estimates of the average number of doctor visits, by State, from the U.S. National Health Interview Survey (NHIS)

Continuing with the example in section 2.2.2, State-level synthetic estimates of the average number of doctor visits from the 1988 round of the NHIS have been computed using formula (3.1.1) with $p_{ih} = N_{ih}/N_{i\bullet}$ and where N_{ih} is the 1990 Census total population for subgroup h in State i and $N_{i\bullet}$ is the 1990 Census total population for State i . The resulting estimates are presented in Table 4.2.3.1. The comparison of the different small area estimates for the average number of doctor visits, by State, is made in section 4.2.3.

4 Composite estimators

4.1 Statistical properties

When small area samples are relatively small, the synthetic estimators outperform the simple direct estimators; however, when small area sample sizes are large, the direct estimators outperform the synthetic estimators (Gonzales and Waksberg, 1973; Schaible, Brock and Schnack, 1977). These authors then concluded that a weighted sum of these two (2) estimators would be better than choosing one over the other.

In general, a composite estimator is a weighted average of a direct estimate and an indirect estimate (Ghosh and Rao, 1994; Suciú et al, 2001). Such a combination is designed to give a new, more precise estimator. Ghosh and Rao (1994) suggested the combination of the direct estimator and the synthetic estimator. As such, the composite estimator will hopefully balance the potential bias of the synthetic estimator against the instability of the direct estimator. The composite estimator then takes the general form

$$\hat{Y}_i^C = w_i * \hat{Y}_i + (1-w_i) * \hat{Y}_i^S , \quad (3.2.1)$$

where \hat{Y}_i^C is the composite estimator for small area i , \hat{Y}_i is the direct estimator, \hat{Y}_i^S is the synthetic estimator, w_i is a appropriately determined weight and $0 \leq w_i \leq 1$.

Several methods of weight selection have been proposed. Ghosh and Rao (1994) suggested obtaining optimal weights by minimizing the MSE of \hat{Y}_i^C with respect to w_i assuming that the $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) = 0$. This gives:

$$w_i = \frac{MSE(\hat{Y}_i^S)}{MSE(\hat{Y}_i^S) + V(\hat{Y}_i)} . \quad (3.2.2)$$

Using the assumption that $\text{cov}(\hat{Y}_i, \hat{Y}_i^s) = 0$, an approximately unbiased estimator of $MSE(\hat{Y}_i^s)$, as given by Ghosh and Rao (1994) is:

$$mse(\hat{Y}_i^s) = (\hat{Y}_i^s - \hat{Y}_i)^2 - v(\hat{Y}_i). \quad (3.2.3)$$

An estimator for the optimal weight may then be calculated as:

$$\begin{aligned} \hat{w}_i &= \frac{mse(\hat{Y}_i^s)}{mse(\hat{Y}_i^s) + v(\hat{Y}_i)} = \frac{(\hat{Y}_i^s - \hat{Y}_i)^2 - v(\hat{Y}_i)}{(\hat{Y}_i^s - \hat{Y}_i)^2 - v(\hat{Y}_i) + v(\hat{Y}_i)} \\ &= \frac{(\hat{Y}_i^s - \hat{Y}_i)^2 - v(\hat{Y}_i)}{(\hat{Y}_i^s - \hat{Y}_i)^2}. \end{aligned} \quad (3.2.4)$$

Taking into account the relative instability of the variances associated with direct estimates, the weights resulting from this scheme can be very unstable.

Purcell and Kish (1979) suggested the use of a common weight in the form:

$$w = 1 - \frac{\sum_i v(\hat{Y}_i)}{\sum_i (\hat{Y}_i^s - \hat{Y}_i)^2}, \quad (3.2.5)$$

which is then estimated by

$$\hat{w} = 1 - \frac{\sum_i v(\hat{Y}_i)}{\sum_i (\hat{Y}_i^s - \hat{Y}_i)^2}. \quad (3.2.7)$$

In the case when the variances of the \hat{Y}_i 's are approximately equal, $v(\hat{Y}_i)$ may be replaced by the average

$$\bar{v} = \frac{\sum_i v(\hat{Y}_i)}{m} .$$

The common weight is then estimated as

$$\hat{w} = 1 - \frac{m\bar{v}}{\sum_i (\hat{Y}_i^s - \hat{Y}_i)^2} . \quad (3.2.8)$$

It should be noted, though, that the use of a common weight is not recommended when the individual variances $V(\hat{Y}_i)$ vary considerably.

Simple weights dependent on domain counts have also been proposed by Drew, Singh and Choudry (1982). In this method, the weight is estimated as:

$$w_i = \begin{cases} 1, & \text{if } \hat{N}_i \geq \delta N_i \\ \frac{\hat{N}_i}{\delta N_i}, & \text{otherwise,} \end{cases} \quad (3.2.9)$$

where \hat{N}_i is the direct and unbiased estimator of the known small area population size N_i and δ is chosen subjectively to control the contribution of the synthetic estimator. In the case of simple random sampling of n elements from a population of N elements, \hat{N}_i is estimated as:

$$\hat{N}_i = N \cdot \frac{n_i}{n} ,$$

where n_i is the sample size for the small area i .

Särndal and Hidiroglou (1989) suggested an alternative estimator for the weight w_i :

$$w_i = \begin{cases} 1, & \text{if } \hat{N}_i \geq N_i \\ \left(\frac{\hat{N}_i}{N_i}\right)^{h-1}, & \text{otherwise,} \end{cases} \quad (3.2.10)$$

where h is, again, subjectively chosen. However, h is usually given the value 2.

4.2 Applications

4.2.1. Composite estimates of total wages and salaries

Resuming the example of Rao and Choudry (1995), composite estimates of total wages and salaries by census division by industry group in Nova Scotia are also generated. Recall that in this example, the small area is defined to be a nonempty census division by industry group combination. The number of units N is used to obtain the weights w_i (see formula 3.2.9). The composite estimates are then obtained as the average of the poststratified direct estimate and the synthetic estimate. The following table shows the computed percentage values of \overline{ARB}_h , \overline{EFF}_h and \overline{ARE}_h for subgroup h denoting the construction industry group in the province of Nova Scotia.:

Table 4.2.1.1
Average Absolute Relative Bias \overline{ARB} , Average Relative Efficiency \overline{EFF} and Absolute Relative Error \overline{ARE} of the Composite Estimators of Total Wages and Salaries for the Construction Industry of Nova Scotia

Quality measure	Approach 1	Approach 2
\overline{ARB}	2.9%	4.4%
\overline{EFF}	137.6%	125.4.9%
\overline{ARE}	24.0%	25.5%

Comparing Tables 2.2.1.1 and 3.2.3.1 with Table 4.2.1.1, it may be seen that the quality measures (i.e., \overline{ARB} , \overline{EFF} and \overline{ARE}) for the composite estimates come between those for the direct and synthetic estimates. These results are expected as the composite estimator is a weighted average of these two (2) other estimators. Royall (1978) stipulates that the mean square error of the composite estimator is smaller than the larger of the mean squared errors of the two component estimators. The mean squared error of the composite estimator is smaller than that of either component estimator when an “appropriate weighting system” is used.

4.2.2 Small area estimates of employment

Sostra (2001) made a study on the comparison of the different small area estimation methods as specified in the paper by Rao (1998). In this study, direct estimates (i.e., simple expansion and post-stratified estimators), synthetic estimates and composite estimates of the number of employed, unemployed and inactive people in the labour force were computed using the data from the 1999 Estonian Labour Force Survey. To do so, the sample of the 1999 Estonian Labour Force Survey was used as a fictitious population.

Samples were then drawn from this fictitious population. The numbers of employed, unemployed and inactive people in the labour force for each small area i were generated using the different estimation methods (i.e., direct estimation, synthetic estimation and composite estimation). The direct estimates of the number of people in the labour force who are employed, unemployed and inactive were computed by applying a simple expansion factor to the results obtained from the sample drawn from the fictitious population [see formulas (2.1) and (2.2)]. The synthetic estimates were obtained by applying formula (3.1.2) with $p_{ih} = N_{ih} / N_{.h}$, i denoting a particular small area and h denoting a particular group. The composite estimators were calculated using (3.2.1). It is assumed that the weights p_{ih} have been obtained from a previous statistical inquiry. The following two tables present the results of this study on the comparison of the different small area estimation methods for a particular small area $i = A$. Table 4.2.2.1 shows the total number of people in the fictitious population and in the specified small area A that are employed, unemployed and inactive. Table 4.2.2.2 presents the comparison of the estimates for the specified small area A .

Table 4.2.2.1
Population Count for the 1999 Labor Force Survey

	Employed	Unemployed	Inactive	Total
Population N	6648	893	5031	12572
Small area A N_A	297	32	241	570

Table 4.2.2.2
Comparison of the Small area Estimators

		Employed	Unemployed	Inactive	Total
Descriptive statistics for small area i	Average n_i	9.1	1.0	7.8	17.9
	Minimum	0	0	3	9
	Maximum	19	5	14	33
	Standard deviance	3.1	0.9	2.8	4.6
Estimation method	Direct	302.4	34.1	256.7	593.5
	Poststratified	291.2	32.4	246.2	570.0
	Synthetic	303.4	40.1	226.5	570.0
	Composite	292.0	33.0	244.8	570.0
Bias	Direct	1.8	6.5	6.5	4.1
	Poststratified	-2.0	1.3	2.2	0.0
	Synthetic	2.2	25.4	-6.0	0.0
	Composite	-1.7	3.0	1.6	0.0
MSE	Direct	10472	977	8744	23032
	Poststratified	5911	825	4735	0.0
	Synthetic	277	128	428	0.0
	Composite	4391	657	3646	0.0

Among the four different small area estimators, the composite estimator has the smallest absolute bias (1.7). Its mean squared error, though significantly smaller than that of either the direct and postratified estimators, is more than 15 times that of the synthetic estimator. Again, we see in these results the gain in terms of bias that is often accompanied by an undesirable increase in the mean squared error. This fact is further exemplified by the results for the synthetic estimator which has the biggest absolute bias (2.2) but has the smallest mean squared error.

4.2.3 Composite estimates of the average number of doctor visits by State from the U.S. National Health Interview Survey (NHIS)

Composite estimates of the average number of doctor visits by State (NCHS, 1999) were generated using the direct and synthetic State-level estimates (see 2.2.2 and 3.2.4). The composite estimator is obtained as a linear

combination of direct and synthetic State-level estimates (see formula 3.2.1) where the weights are proportional to the mean square errors of these two estimators, i.e.,

$$\hat{Y}_i^C = w_i \hat{Y}_i + (1 - w_i) \hat{Y}_i^S \quad ,$$

where (see formula 3.2.2)

$$w_i = \frac{MSE(\hat{Y}_i^S)}{Var(\hat{Y}_i) + MSE(\hat{Y}_i^S)} \quad .$$

The resulting composite State-level estimates of the average number of doctor visits, as well as the direct and synthetic estimates, are presented in Table 4.2.3.1. It may be noted that the synthetic estimates and the composite estimates are not consistent with the direct estimators. The state with the highest direct estimate of average number of doctor visits (Vermont) and that with the highest synthetic estimate (Florida) are not the same. The same is true for the state with the lowest direct estimate (North Carolina) and that with the lowest synthetic estimate (Alaska). Both the direct and composite estimators, though, found Vermont with the highest average number of doctor visits and North Carolina with the lowest. Nevertheless, the ranking of the states based on the composite estimates of average number of doctor visits does not coincide with that based on the direct estimates. The shrinkage of the synthetic estimates and, to some extent, the composite estimates toward their respective overall mean may be noted. The computed range for each type of State-level estimates attest to this shrinkage (see Table 4.2.3.2). While the direct estimates span a range of more than 3.0 doctor visits per year (2.685-6.036), the synthetic estimates span a range of less than 0.4 (3.629-4.003) doctor visits per year and the composite estimates less than 2.0 (2.949-4.600). This observation could explain the inconsistencies among the three kinds of small area estimates.

Table 4.2.3.1
Average Number of Doctor Visits, by State

State	Design-unbiased expansion estimator	Synthetic estimator	Composite estimator
Vermont	6.036	3.889	4.600
Delaware	5.648	3.856	4.178
Colorado	4.868	3.831	4.477
District of Columbia	4.706	3.852	4.121
Michigan	4.594	3.852	4.461
Connecticut	4.589	3.926	4.343
Arizona	4.562	3.889	4.350
Nevada	4.530	3.841	4.026
Massachusetts	4.481	3.933	4.362
Rhode Island	4.424	3.959	4.098
Montana	4.394	3.911	4.159
Kentucky	4.333	3.895	4.179
Pennsylvania	4.204	3.967	4.175
Ohio	4.201	3.896	4.162
Wyoming	4.156	3.827	3.912
California	4.111	3.817	4.093
Oklahoma	4.104	3.898	4.041
New Mexico	4.084	3.928	3.898
Maine	4.075	3.811	3.981
Maryland	4.045	4.003	3.970
Florida	3.981	3.876	3.985
Tennessee	3.950	3.871	3.931
Washington	3.869	3.908	3.870
Arkansas	3.840	3.902	3.867
New York	3.828	3.957	3.836
Iowa	3.821	3.913	3.878
New Jersey	3.806	3.975	3.828
West Virginia	3.734	3.975	3.888
Oregon	3.704	3.928	3.800
Utah	3.704	3.740	3.721
Wisconsin	3.692	3.894	3.732
South Carolina	3.677	3.780	3.730
Virginia	3.671	3.819	3.704
Alabama	3.638	3.846	3.698
Hawaii	3.575	3.850	3.796
Kansas	3.573	3.897	3.690
Missouri	3.549	3.913	3.655
Mississippi	3.489	3.763	3.674
Texas	3.450	3.879	3.482
Illinois	3.420	3.862	3.483
Louisiana	3.410	3.754	3.502
Indiana	3.380	3.885	3.516
South Dakota	3.334	3.921	3.702
New Hampshire	3.236	3.879	3.798
Idaho	3.162	3.857	3.626
Alaska	3.093	3.629	3.258
Minnesota	3.034	3.885	3.258
Georgia	3.000	3.755	3.190
North Carolina	2.685	3.842	2.949
North Dakota	...	3.908	3.908
Nebraska	...	3.913	3.913

... no samples taken

Note: Shading indicates the largest and smallest estimates in the column

Table 4.2.3.2
Variation in State estimates (Range)
For average number of doctor visits

Estimates	Range
Direct	2.69-6.04
Synthetic	3.63-4.00
Composite	2.95-4.60

5 Bayesian estimators

5.1 Basic Bayesian concepts

Suppose that a model in the form of a probability distribution $f(\mathbf{y} | \boldsymbol{\theta})$ has been specified for the observed data $\mathbf{y} = (y_1, \dots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. From the perspective of the Bayesian approach, it is supposed that $\boldsymbol{\theta}$ is likewise random and with a prior distribution $\pi(\boldsymbol{\theta} | \boldsymbol{\eta})$, with $\boldsymbol{\eta}$ defined as a vector of known hyperparameters. Inferences concerning $\boldsymbol{\theta}$ are then based on its posterior distribution, defined as:

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{y}, \boldsymbol{\eta}) &= \frac{p(\mathbf{y}, \boldsymbol{\theta} | \boldsymbol{\eta})}{p(\mathbf{y} | \boldsymbol{\eta})} = \frac{p(\mathbf{y}, \boldsymbol{\theta} | \boldsymbol{\eta})}{\int p(\mathbf{y}, \mathbf{u} | \boldsymbol{\eta}) d\mathbf{u}} \\ &= \frac{f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \boldsymbol{\eta})}{\int f(\mathbf{y} | \mathbf{u}) \pi(\mathbf{u} | \boldsymbol{\eta}) d\mathbf{u}} \end{aligned} \quad (5.1.1)$$

The expression (5.1.1) is referred to as *Bayes' Theorem* (Carlin and Louis, 1996). The posterior distribution $p(\boldsymbol{\theta} | \mathbf{y}, \boldsymbol{\eta})$ is expressed as a function of the likelihood $f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\eta})$ and of the prior distribution $\pi(\boldsymbol{\theta} | \boldsymbol{\eta})$. Given a sample of n independent observations, the likelihood $f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\eta})$ may be obtained as $\prod_{i=1}^n f(y_i | \boldsymbol{\theta})$. The denominator of expression (5.1.1) is sometimes rewritten as

$$\int f(\mathbf{y} | \mathbf{u}) \pi(\mathbf{u} | \boldsymbol{\eta}) d\mathbf{u} = m(\mathbf{y} | \boldsymbol{\eta}), \quad (5.1.2)$$

and is interpreted as the marginal distribution of the data \mathbf{y} given the value of the hyperparameter $\boldsymbol{\eta}$. Thus, expression (5.1.1) may be rewritten as

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{m(\mathbf{y}|\boldsymbol{\eta})} \quad (5.1.3)$$

Further, as $\boldsymbol{\eta}$ is known, and hence treated as a constant, the posterior distribution may be written simply as $p(\boldsymbol{\theta}|\mathbf{y})$.

With the Bayes approach, an action is taken such that the *posterior risk* is minimized. This *posterior risk* is defined as:

$$\rho(\pi, \hat{\boldsymbol{\theta}}) = E_{\theta|\mathbf{y}} [l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] = \int l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad , \quad (5.1.4)$$

where $l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is a loss function that gives the loss incurred when $\boldsymbol{\theta}$ is the value of the parameter and $\hat{\boldsymbol{\theta}}$ is taken as its estimator. The loss function $l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ may take any of the following forms (Lee, 1989). The corresponding Bayes estimator $\hat{\boldsymbol{\theta}}^B$ is likewise indicated.

i) the squared error loss (SEL)

$$l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2,$$

in which case

$$\hat{\boldsymbol{\theta}}^B = E(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) \quad (5.1.5)$$

ii) the absolute error loss

$$l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|,$$

in which case

$$\hat{\boldsymbol{\theta}}^B = M_d \quad (5.1.6)$$

where M_d is the median of the posterior distribution of θ

or iii) for the discrete parameter spaces, the 0–1 loss

$$l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}} \\ 1 & \text{if } \boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}} \end{cases} .$$

in which case

$$\hat{\boldsymbol{\theta}}^B = M_o \tag{5.1.7}$$

where M_o is the mode of the posterior distribution of θ .

In the presence of uncertainty as to the proper values for $\boldsymbol{\eta}$ (i.e., $\boldsymbol{\eta}$ is unknown) the proper Bayesian solution is to quantify this uncertainty in a second-stage prior distribution (Carlin and Louis, 1996). This second-stage prior distribution is sometimes referred to as a hyperprior and is, herewith, denoted as $h(\boldsymbol{\eta})$. The desired posterior is then obtained by also marginalizing over $\boldsymbol{\eta}$. Thus,

$$\begin{aligned} p(\theta|\mathbf{y}) &= \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) d\boldsymbol{\eta}}{\iint p(\mathbf{y}, \mathbf{u}, \boldsymbol{\eta}) d\boldsymbol{\eta} d\mathbf{u}} \\ &= \frac{\int f(\mathbf{y}|\boldsymbol{\theta}) \pi(\mathbf{y}|\boldsymbol{\theta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta}}{\iint f(\mathbf{y}|\mathbf{u}) \pi(\mathbf{u}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} d\mathbf{u}} \end{aligned} \tag{5.1.8}$$

The hyperprior for $\boldsymbol{\eta}$ may depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of expression (5.1.7) featuring a second-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda})$ and a third-stage prior $g(\boldsymbol{\lambda})$. This approach of specifying a model over several levels is referred to as hierarchical modeling, with each new distribution forming a new level in the hierarchy. The number of levels may vary with particular problems. In general, though, levels above the second-stage prior are rarely used.

When there is uncertainty as to the values of $\boldsymbol{\eta}$, an alternative approach is to simply replace $\boldsymbol{\eta}$ by an estimate $\hat{\boldsymbol{\eta}}$ obtained as the value which maximizes the marginal distribution $m(\mathbf{y} | \boldsymbol{\eta})$ viewed as a function of $\boldsymbol{\eta}$. Consequent inferences are then based on the estimated posterior distribution $p(\boldsymbol{\theta} | \mathbf{y}, \hat{\boldsymbol{\eta}})$ calculated by replacing $\boldsymbol{\eta}$ by its estimator $\hat{\boldsymbol{\eta}}$. This approach of using the given data to estimate the prior parameter $\boldsymbol{\eta}$ is referred to as *empirical Bayes* analysis.

5.2 Bayesian modeling for small areas

5.2.1 Notations

For consistency, the notations used in the earlier parts of this paper are again adopted. The notation \bar{Y}_i denotes the small area mean for the characteristic of interest y for small area $i, i = 1, \dots, m$; however, from hereon, the index i is used to refer to a small demographic area, a socio-economic subgroup or a cross-classification of these two (2) criteria, as the case may be. For simplicity, focus will be on the small area means. Similar results for area totals Y_i may be obtained using the same approach. Area-specific auxiliary data are denoted by $x_i, i = 1, \dots, m$. New notations will be explained as they appear in the paper.

5.2.2 Bayesian modelling

For the small area level model, it is assumed that area-specific auxiliary data x_i as well as direct estimates \hat{Y}_i of \bar{Y}_i are available whenever $m_i \geq 1$. It

is further assumed that \bar{Y}_i or some suitable function $\theta_i = g(\bar{Y}_i)$ is related to x_i through a linear model with random effects v_i ,

$$\theta_i = \mathbf{x}'_i \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m \quad (5.2.1)$$

where $\boldsymbol{\beta}$ is the vector of regression parameters and the v_i 's are independent and identically distributed (iid) normal random variables with mean 0 and variance σ_v^2 (Rao, 1999). Likewise, the estimator $\hat{\theta}$ of θ may be expressed as

$$\hat{\theta}_i = \theta_i + e_i \quad i = 1, \dots, m \quad (5.2.2)$$

where $\hat{\theta}_i = g(\hat{\bar{Y}}_i)$ and the sampling errors e_i 's are independent $N(0, \psi_i)$ with known ψ_i . It can be seen from expression (5.2.2) that $E(\hat{\theta}_i) = \theta_i$; the direct estimators $\hat{\theta}_i$ are, thus, unbiased. The expressions (5.2.1) and (5.2.2) are combined to yield the area level linear mixed model of Fay and Herriot (1979):

$$\hat{\theta}_i = \mathbf{x}'_i \boldsymbol{\beta} + v_i + e_i, \quad i = 1, \dots, m \quad (5.2.3)$$

and

$$\hat{\theta}_i \sim_{iid} N(\mathbf{x}'_i \boldsymbol{\beta}, \sigma_v^2 + \psi_i), \quad i = 1, \dots, m \quad (5.2.4)$$

The vector of known hyperparameters $\boldsymbol{\eta}$ is defined as

$$\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma_v^2). \quad (5.2.4)$$

Using expression (5.1.3), the posterior distribution of θ_i may be defined as

$$\theta_i \mid \hat{\theta}_i, \boldsymbol{\beta}, \sigma_v^2 \sim N\left(\frac{\sigma_v^2 \hat{\theta}_i + \psi_i \mathbf{x}'_i \boldsymbol{\beta}}{\sigma_v^2 + \psi_i}, \frac{\sigma_v^2 \psi_i}{\sigma_v^2 + \psi_i}\right). \quad (5.2.5)$$

Using (5.1.5), the Bayes estimator of θ_i is derived as

$$\begin{aligned}\hat{\theta}_i^B &= E(\theta_i | \hat{\theta}_i, \beta, \sigma_v^2) \\ &= \frac{\sigma_v^2}{\sigma_v^2 + \psi_i} \hat{\theta}_i + \frac{\psi_i}{\sigma_v^2 + \psi_i} \mathbf{x}_i' \beta\end{aligned}\quad (5.2.6)$$

It may be noted that the expression (5.2.5) gives a composite estimator of θ_i . In fact, $\hat{\theta}_i^B$ is a weighted mean of the direct estimator and the regression estimator with the weights derived from the random effects variance σ_v^2 and the sampling error variance ψ_i .

5.3 Empirical Bayes

5.3.1 Statistical properties

The previous section has been developed under the assumption that $\boldsymbol{\eta}$ is known. In practice, $\boldsymbol{\eta}$ may be unknown and suitable estimates of the parameters of interest may need to be calculated from the given data. Under the empirical Bayes approach, these estimates of $\boldsymbol{\eta}$ are used to derive the posterior distribution $p(\theta | \hat{\theta}, \boldsymbol{\eta})$. In other words, the posterior distribution of the parameters of interest given the data is first calculated. The model parameters are estimated from the marginal distribution of the data, and inferences are then based on the estimated posterior distribution.

Recall from the previous section that the Bayesian estimator of θ_i when $\boldsymbol{\eta}$ is known is given by the expression (5.2.6). The empirical Bayes estimator of θ_i may be then expressed as in formula (5.2.6) with the unknown parameters replaced by estimators. Thus,

$$\begin{aligned}\hat{\theta}_i^{EB} &= \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} \hat{\theta}_i + \frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i} \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \\ &= \tau_i \hat{\theta}_i + (1 - \tau_i) \mathbf{x}_i^T \hat{\boldsymbol{\beta}}\end{aligned}\quad (5.3.1)$$

where

$$\tau_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} \quad (5.3.2)$$

and $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$ are parameter estimators to be defined later.

A convenient form of the estimator $\hat{\theta}_i^{EB}$ (Rivest and Belmonte, 2000) is given by:

$$\begin{aligned}\hat{\theta}_i^{EB} &= \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} \hat{\theta}_i + \left(1 - \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i}\right) \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \\ &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} (\hat{\theta}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \\ &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \left(1 - \frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i}\right) (\hat{\theta}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \\ &= \hat{\theta}_i - \frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i} \hat{\theta}_i + \frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i} \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \\ &= \hat{\theta}_i - \frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i} (\hat{\theta}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) .\end{aligned}\quad (5.3.3)$$

The estimation of the vector of parameters $\boldsymbol{\eta} = (\beta, \sigma_v^2)$ may be described as follows. A simple moment estimator of σ_v^2 may be obtained as $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$, where

$$\tilde{\sigma}_v^2 = \frac{\sum_i (\hat{\theta}_i - \mathbf{x}_i^T \hat{\beta}^*)^2 - \sum_i \psi_i \left\{ 1 - \mathbf{x}_i^T \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i \right\}}{m - p}, \quad (5.3.4)$$

and

$$\hat{\beta}^* = \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_i \mathbf{x}_i \hat{\theta}_i \right), \quad (5.3.5)$$

is the ordinary least squares estimator of β . On the other hand, $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$ is a weighted least squares estimator of β and is given by:

$$\tilde{\beta}(\sigma_v^2) = \left[\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_v^2 + \psi_i} \right]^{-1} \cdot \left[\sum_i \frac{\mathbf{x}_i \hat{\theta}_i}{\sigma_v^2 + \psi_i} \right]. \quad (5.3.6)$$

The mean square error $MSE(\hat{\theta}^{EB})$ gives an indication of the accuracy of $\hat{\theta}_i^{EB}$ as an estimator of θ_i . This may be defined as

$$MSE(\hat{\theta}_i^{EB}) = E \left[\left(\hat{\theta}_i^{EB} - \theta_i \right)^2 \right] \quad (5.3.7)$$

where the expectation E is taken with respect to the marginal distribution of y_i . An approximation of $MSE(\hat{\theta}^{EB})$ proposed by Rao (2001) is given by :

$$MSE(\hat{\theta}_i^{EB}) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) \quad (5.3.8)$$

where

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i, \quad ,$$

$$g_{2i}(\sigma_v^2) = (1 - \gamma_i)^2 \mathbf{x}_i^T \left[\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_v^2 + \psi_i} \right]^{-1} \mathbf{x}_i, \quad (5.3.9)$$

$$g_{3i}(\sigma_v^2) = \left[\frac{\psi_i^2}{(\sigma_v^2 + \psi_i)^3} \right] h(\sigma_v^2) \quad (5.3.10)$$

and $h(\sigma_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$ for large m . The three (3) components of $MSE(\hat{\theta}^{EB})$, namely $g_{1i}(\sigma_v^2)$, $g_{2i}(\sigma_v^2)$ and $g_{3i}(\sigma_v^2)$, have been interpreted as follows: $g_{1i}(\sigma_v^2)$ is the posterior variance, $g_{2i}(\sigma_v^2)$ is the contribution from estimating the regression parameters β and $g_{3i}(\sigma_v^2)$ is the contribution from estimating the model variance σ_v^2 . Neglected terms in the approximation formula (5.3.8) are of order lower than m^{-1} .

Prasad and Rao (1990) have shown that in the MSE approximation (5.3.8), $g_{2i}(\sigma_v^2)$ and $g_{3i}(\sigma_v^2)$ are both of order $O(m^{-1})$. They have further shown that the estimators of $g_{2i}(\sigma_v^2)$ and $g_{3i}(\sigma_v^2)$ are given simply by $g_{2i}(\hat{\sigma}_v^2)$ and $g_{3i}(\hat{\sigma}_v^2)$. Since $g_{1i}(\hat{\sigma}_v^2)$ is $O(1)$, its estimation is more complex. Prasad and Rao (1990) showed that $E(g_{1i}(\hat{\sigma}_v^2)) = g_{1i}(\sigma_v^2) - g_{3i}(\sigma_v^2) + o(1/m)$, where the third term is negligible. Thus, $g_{1i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)$ estimates $g_{1i}(\sigma_v^2)$. A correct estimator of $g_{1i}(\sigma_v^2)$, then, is obtained by adjusting $g_{1i}(\hat{\sigma}_v^2)$ so that its bias is $O(m^{-1})$. The resulting estimator of the MSE approximation with expectation correct to $O(m^{-1})$ is then given by

$$mse_{PR}(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \quad (5.3.11)$$

For the moment estimator σ_v^2 , Prasad and Rao (1990) have proposed the following estimator for the asymptotic variance $h(\hat{\sigma}_v^2)$:

$$h_{PR}(\hat{\sigma}_v^2) \approx \frac{2}{m^2} \sum_i (\sigma_v^2 + \psi_i)^2 . \quad (5.3.12)$$

Using (5.3.12), the estimator $mse_{PR}(\hat{\theta}_i^{EB})$ simplifies into:

$$\begin{aligned} mse_{PR}(\hat{\theta}_i^{EB}) &\approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \\ &= \frac{\hat{\sigma}_v^2 \psi_i}{\hat{\sigma}_i^2 + \psi_i} + \left(\frac{\psi_i}{\hat{\sigma}_v^2 + \psi_i} \right)^2 \mathbf{x}_i^T \left(\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^T}{\hat{\sigma}_v^2 + \psi_i} \right)^{-1} \mathbf{x}_i \\ &\quad + \frac{4}{m^2} \frac{\psi_i^2}{(\hat{\sigma}_v^2 + \psi_i)^3} \sum_i (\hat{\sigma}_v^2 + \psi_i)^2 \end{aligned} \quad (5.3.13)$$

Rivest and Vandal (2003) have shown that when ψ_i is not known and, thus, has to be estimated along with the parameters β and σ_v^2 , the $mse_{PR}(\hat{\theta}_i^{EB})$ has to be adjusted for the estimation of ψ_i . It is assumed that s_i^2 , the estimator of ψ_i , is a statistic with a $N(\psi_i, \phi_i)$ distribution. The adjustment is done by adding the term $2\phi_i \hat{\sigma}_v^4 / (s_i^2 + \hat{\sigma}_v^2)$, thus giving the generalized Prasad and Rao estimator:

$$mse_{PRg}(\hat{\theta}_i^{EB}) \approx \frac{\hat{\sigma}_v^2 s_i^2}{\hat{\sigma}_v^2 + s_i^2} + \frac{s_i^4 \mathbf{x}_i^T \hat{A}^{-1} \mathbf{x}_i}{(\hat{\sigma}_v^2 + s_i^2)^2} + 2 \frac{s_i^4 \hat{h}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 \hat{\phi}_i}{(\hat{\sigma}_v^2 + s_i^2)^3} \quad (5.3.14)$$

where $\hat{A} = \sum \mathbf{x}_i \mathbf{x}_i^T / (\hat{\sigma}_v^2 + s_i^2)$ and $\hat{h}(\hat{\sigma}_v^2)$ is an estimator of the variance of $\hat{\sigma}_v^2$.

5.3.2 Applications

5.3.2.1 Empirical Bayes estimates of income for small places

In a well documented application of the empirical Bayes approach, Fay and Herriot (1979) developed a procedure for estimating per capita income (PCI) for small places in the U.S. initially for the reference year 1969 .

This procedure became the framework for some later developments in the use of the empirical Bayes approach in small area estimation.

The U.S. State and Local Fiscal Assistance Act of 1972 mandate the allocation of funds to states and units of general-purpose local government for operational or capital expenditures. Statistics on population, per capita income, and adjusted taxes are used to determine the allocation of these funds within the states.

Adopting the notations used earlier, the term \bar{Y}_i is used to denote the PCI for the small area i , here defined as a geographic area with population of less than 1000. The term θ_i , in turn, is defined as $\theta_i = \log(\bar{Y}_i)$.

In the procedure undertaken by Fay and Herriot (1979), census sample estimates \hat{Y}_i were calculated from the results of the 1970 Census of Population and Housing. The auxiliary variables chosen were the county PCI values, tax return data from the IRS for 1969, and data on housing from the 1970 Census of Population and Housing. Logarithmic transformations of the census values and of the auxiliary variables were made and the corresponding $\hat{\beta}$ was calculated using formula (5.3.6). The regression estimates $\hat{\theta}_i^*$ for each small area i were then subsequently computed as follows:

$$\hat{\theta}_i^* = \mathbf{x}_i^T \hat{\beta} \quad , \quad (5.3.15)$$

where \mathbf{x}_i is the matrix of the transformed auxiliary variables. Using formula (5.3.1), the empirical Bayes estimates $\hat{\theta}_i^{EB}$ were generated as a weighted average of the direct estimates $\hat{\theta}_i$ and the regression estimates $\hat{\theta}_i^*$, with the weights reflecting the relative magnitudes of the sampling error variance ψ_i

and the random effects variance σ_v^2 . The resulting estimates were then retransformed back to the original scale.

The methodology adopted to estimate the PCI for the succeeding years after the reference year 1969 (e.g., 1970, 1971, 1972) involved the use of an updating factor f_i . This updating factor was based on changes in administrative data. The methodology allowed the updating of three (3) types of estimates, to wit, the direct estimates $\hat{\theta}_i$, the empirical Bayes estimates $\hat{\theta}_i^{EB}$ and the county based PCI. It should be noted that the large magnitude of sampling errors led to the initial use of county estimates as substitutes to small area estimates.

To verify the validity of the resulting empirical Bayes estimates of PCI for 1972, the Census Bureau conducted complete censuses of a random sample of small areas in 1973, collecting income data for the reference year 1972 on a 100 percent basis. For comparison purposes, the estimates generated from this special census were treated as the true income values. The results of the comparison of the 3 sets of 1972 PCI estimates with the special 1973 census results for some selected small areas are presented in Table 5.3.1.

An analysis of Table 5.3.1 shows that the revised James-Stein estimator (i.e., the empirical Bayes estimator obtained with the use of the Fay-Herriot procedure) shows smaller average errors than either the direct estimates or the county values. In the majority of the small areas shown on the table, the percentage difference of the revised James-Stein/empirical Bayes estimator is lower than either of the two other estimators. It may be noted, though, that the majority of the 1972 PCI estimates are higher than the special 1973 census values. This is true even for the case of the James-Stein/empirical Bayes estimates. One possible reason for this, as suggested by Fay and Herriot, is

that while imputations were done for missing income data in the 1970 census, the same was not done for the 1973 special census. As a consequence, the special census estimates may be subject to a downward bias.

It should be noted that only the 1969 estimates of the PCI were obtained using the empirical Bayes approach. Estimates for the subsequent years (i.e., 1970 to 1972) were generated by applying an updating factor to the 1969 base estimates. In other words, the validity of the 1972 estimates being considered as empirical Bayes estimates may be questioned. Caution should thus be taken when drawing conclusions about the efficiency of the 1972 empirical Bayes estimates through comparison with the results of the 1973 special census.

Table 5.3.1.
Comparison of Selected PCI Estimates with 1973 Special Census Values of 1972 PCI

Special Census Areas	1973 Special Census 1972 PCI	1972 PCI Estimates and Percentage Difference from Special Census PCI					
		Using 1970 Sample Base		Using Revised Base (James-Stein)		Using County Base	
		1972 Estimate	Percentage Difference	1972 Estimate	Percentage Difference	1972 Estimate	Percentage Difference
<u>1970 Census Weighted Sample Population Less than 500</u>							
Newington, Ga.	\$2019	\$2225	10.2	\$2302	14.0	\$2279	12.9
Foosland Village, Ill.	2899	2771	4.4	3199	10.3	3796	30.9
Bonaparte, Iowa	2331	3126	34.1	2942	26.2	2542	9.1
McNary, La.	2333	2303	1.3	2527	8.3	2908	24.6
Freeborn Village, Minn.	2741	3693	34.7	3338	21.8	2922	6.6
Spruce Valley Twp., Minn.	2430	1894	22.1	1949	19.8	2076	14.6
Jacksonville, Mo.	2723	2338	14.1	2611	4.1	3233	18.7
Thayer, Nebr.	2742	2245	18.1	2870	4.7	3452	25.9
Benton Town, N.H.	1788	2874	60.7	3284	78.7	3570	99.7
Nora Twp., N.Dak.	1780	2629	47.7	2754	54.7	3476	95.3
Riga Twp, N.Dak.	1454	2749	89.1	2411	65.8	2711	86.5
Deer Creek, Okla.	2451	2493	1.7	2673	9.1	2762	12.7
Dudley Borough, Pa.	2446	2168	11.4	2411	1.4	2608	6.6
Brookings Twp, S.Dak.	3132	3400	8.6	3309	5.7	2395	23.5
Valley Twp., S.Dak	1574	1946	23.6	1972	25.3	2114	34.3
Bryant Twp., S.Dak.	2412	1120	53.6	2158	10.5	2695	11.7
Parrish Town, Wis.	3567	5399	51.4	4079	14.4	2721	23.7
Ave. % Difference	----	-----	28.6	----	22.0	----	31.6
<u>1970 Census Weighted Sample Population Between 500 and 999</u>							
Caswell Plantation, Maine	\$1946	\$2656	36.5	\$2490	28.0	\$2646	36.0
Sugar Creek Twp., Mo.	2224	2035	8.5	2315	4.1	2018	9.3
Jeromesville, Ohio	3329	3081	7.4	3418	2.7	3072	7.7
Rush Twp., Ohio	2241	2545	13.6	2619	16.9	2546	13.6
Dennison Twp., Pa.	3521	4411	25.3	4095	16.3	4430	25.8
Manor, Tex.	2062	2746	33.2	2765	34.1	2740	32.9
Derby Center, Vt.	2968	2694	9.2	2754	7.2	2675	9.9
Ave. % Difference	-----	----	19.1	----	15.6	----	19.3

5.3.2.2 Empirical Bayes estimates of net undercoverage in the 1991 Canadian Census

The Census of Canada, conducted every five (5) years, is designed to provide accurate population counts by age and sex within each province and territory. Statistics Canada adjusted the Population Estimates Program for the first time in 1991 using the results of two (2) independent coverage surveys, namely the Reverse Record Check (RRC) and the Overcoverage Study (OCS), both of which have been designed to give reliable estimates of net undercoverage for all provinces, some of the larger metropolitan areas and for some large national domains (e.g., males in the age group 20-24). The RRC was used to give estimates for the number of persons missed in the census while the OCS was designed to give estimates on the number of persons erroneously included in the census.

The variable of interest in this example (Dick, 1995) is the population adjustment factor. This adjustment factor is denoted by θ_i and is defined as:

$$\theta_i = \frac{T_i}{C_i} = \frac{M_i + C_i}{C_i} \quad (5.3.16)$$

where C_i is the number of persons in the small area i enumerated in the Census, T_i is the true population count in the same domain i , and $M_i = T_i - C_i$ is the net number of persons missed in this same domain. The term small area is defined here as a province/territory age-sex group. For the empirical Bayes approach, the auxiliary variables x_i considered in the model are the age-sex combination (male 20-29, male 30-34, and female 20-29), sex by age by non-official language (female-language-0 to 19) and tenure by

province (British Columbia renters, Ontario renters, Québec renters, New Brunswick renters, Yukon renters and Northwest Territories renters).

The empirical Bayes estimates $\hat{\theta}_i^{EB}$ were generated using formula (5.3.1). Table 5.3.2 presents and compares two (2) estimates of the adjustment factor for each small area (Dick, 1995).

It may be observed from Table 5.3.2 that the direct estimates and the empirical Bayes estimates are relatively close. This may be due either to the relatively close values of the direct estimates and the regression estimates or to the relatively bigger weights for the direct estimators (i.e., the random effects variance σ_v^2 relatively larger than the sampling error variance ψ_i). The differences between the direct estimates and the empirical Bayes estimates, in absolute values, are less than 0.5% in the larger provinces but are significant in some small areas in the smaller provinces and the territories. As for the accuracy of the empirical Bayes estimates versus the direct estimates, Dick (1995) claims that it is clear that the empirical Bayes estimates have smaller MSE's than the direct estimates. However, the gain in the use of the empirical Bayes estimates are more obvious in the smaller provinces and territories than in the bigger provinces such as Ontario and Québec, a reflection of the fact that the bigger provinces have relatively larger sample sizes which in turn give more reliable direct estimates and more reliable sample variances.

Table 5.3.2 Direct and Empirical Bayes Estimates of Adjustment Factors

Sex	Age	Estimate	B.C.	Alta.	Sask	Man.	Ont.	Qué.	N.B.	N.S.	P.E.I.	Nfld	Yukon	NWT
Male	0-19	Direct	1.017	1.026	1.012	1.029	1.028	1.017	1.022	1.018	1.004	0.999	1.031	1.036
		E.Bayes	1.019	1.013	1.009	1.013	1.029	1.016	1.027	1.010	1.007	1.006	1.026	1.027
	20-29	Direct	1.087	1.036	1.068	1.058	1.113	1.071	1.122	1.063	1.060	1.057	1.098	1.127
		E.Bayes	1.086	1.056	1.065	1.062	1.104	1.074	1.103	1.064	1.063	1.062	1.094	1.122
	30-44	Direct	1.031	1.021	1.028	1.034	1.054	1.047	1.043	1.018	1.025	1.026	1.069	1.080
		E.Bayes	1.039	1.026	1.028	1.030	1.053	1.041	1.046	1.026	1.028	1.028	1.052	1.059
	45 +	Direct	1.019	1.018	1.002	1.014	1.013	1.011	1.014	1.016	1.018	1.016	0.992	1.076
		E.Bayes	1.017	1.011	1.006	1.009	1.019	1.013	1.019	1.010	1.009	1.009	1.021	1.039
Female	0-19	Direct	1.031	1.018	1.017	1.012	1.037	1.029	1.029	1.014	0.995	1.016	1.026	1.054
		E.Bayes	1.030	1.015	1.013	1.015	1.038	1.023	1.030	1.010	1.006	1.010	1.028	1.061
	20-29	Direct	1.068	1.047	1.028	1.020	1.072	1.043	1.070	1.030	1.004	1.041	1.068	1.072
		E.Bayes	1.080	1.036	1.031	1.029	1.070	1.044	1.071	1.031	1.027	1.033	1.069	1.092
	30-44	Direct	1.013	1.009	1.004	1.006	1.027	1.017	1.031	1.019	1.004	1.024	1.031	1.020
		E.Bayes	1.018	1.008	1.007	1.007	1.030	1.017	1.029	1.010	1.007	1.011	1.028	1.026
	45 +	Direct	1.007	1.003	1.018	1.001	1.011	1.011	1.000	1.002	0.993	1.013	1.024	1.007
		E.Bayes	1.014	1.006	1.010	1.006	1.021	1.015	1.020	1.006	1.005	1.009	1.031	1.026

5.4. Hierarchical Bayes

5.4.1 Statistical properties

Recall that the basic area level mixed model of Fay and Herriot (1979) as given in section 5.2 and denoted as formula (5.2.3) is

$$\hat{\theta}_i = \mathbf{x}'_i \beta + v_i + e_i \quad (5.4.1)$$

The hierarchical Bayes (HB) approach is applied to this basic area mixed model assuming a prior distribution on the model parameters $\boldsymbol{\eta} = (\beta, \sigma_v^2)$. Two (2) cases are considered: a) σ_v^2 is known, and b) σ_v^2 is unknown.

When σ_v^2 is known and a “flat” prior on β is given by $f(\beta) \propto 1$ (Rao, 2003, p.237), the HB model based on formula (5.4.1) may be rewritten as :

$$\begin{aligned} \text{(i)} \quad & \hat{\theta}_i | \theta_i, \beta, \sigma_v^2 \stackrel{ind}{\sim} N(\theta_i, \psi_i) \quad , \quad i = 1, \dots, m \\ \text{(ii)} \quad & \theta_i | \beta, \sigma_v^2 \stackrel{ind}{\sim} N(\mathbf{x}'_i \beta, \sigma_v^2) \quad , \quad i = 1, \dots, m \\ \text{(iii)} \quad & f(\beta) \propto 1 \end{aligned} \quad (5.4.2)$$

The HB estimator of θ_i is then given by:

$$\begin{aligned} \tilde{\theta}_i^{HB} &= E(\theta_i | \hat{\boldsymbol{\theta}}, \sigma_v^2) \\ &= \mathbf{x}'_i \hat{\beta} + \gamma_i (\hat{\theta}_i - \mathbf{x}'_i \hat{\beta}) \\ &= \gamma_i \hat{\theta}_i + (1 - \gamma_i) \mathbf{x}'_i \hat{\beta} \end{aligned} \quad (5.4.3)$$

where

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_m),$$

$$\hat{\beta} = \tilde{\beta}(\sigma_v^2) = \left[\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_v^2 + \psi_i} \right]^{-1} \cdot \left[\sum_i \frac{\mathbf{x}_i \hat{\theta}_i}{\sigma_v^2 + \psi_i} \right], \quad (5.4.4)$$

$$\text{and} \quad \gamma_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i} \quad (5.4.5)$$

The posterior variance of θ_i is given by:

$$V\left(\hat{\theta}_i \mid \hat{\boldsymbol{\theta}}, \sigma_v^2\right) = M_{1i}(\sigma_v^2) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \quad (5.4.6)$$

where

$$\begin{aligned} g_{1i}(\sigma_v^2) &= \gamma_i \psi_i, \quad \text{and} \\ g_{2i}(\sigma_v^2) &= (1 - \gamma_i)^2 \mathbf{x}_i^T \left[\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_v^2 + \psi_i} \right]^{-1} \mathbf{x}_i \end{aligned} \quad (5.4.7)$$

Formulas (5.4.3) to (5.4.7), as may be observed, are the same as those for the EB estimator. Thus, it may be concluded that the HB and EB approaches give identical point estimates and variances for θ_i when σ_v^2 is assumed to be known and the prior on β is given by $f(\beta) \propto 1$.

When σ_v^2 is unknown, as is often the case, the HB model is given by :

$$\begin{aligned} \text{(i)} \quad & \hat{\theta}_i \mid \theta_i, \beta, \sigma_v^2 \stackrel{ind}{\sim} N(\theta_i, \psi_i), \quad i = 1, \dots, m \\ \text{(ii)} \quad & \theta_i \mid \beta, \sigma_v^2 \stackrel{ind}{\sim} N(\mathbf{x}_i' \beta, \sigma_v^2), \quad i = 1, \dots, m \\ \text{(iii)} \quad & f(\beta, \sigma_v^2) = f(\beta) f(\sigma_v^2) \propto f(\sigma_v^2), \end{aligned} \quad (5.4.8)$$

where $f(\sigma_v^2)$ is a prior on σ_v^2 . It is assumed that $\sigma_v^{-2} \sim G(a, b)$, i.e., a gamma distribution with shape parameter a and scale parameter b , with $a > 0$ and $b > 0$ (Rao, 2003, p.240). The parameter σ_v^2 is then distributed

as an inverted gamma $IG(a, b)$, with $f(\sigma_v^2)$ proportional to $\exp(-b/\sigma_v^2)(1/\sigma_v^2)^{a+1}$. The shape parameter a and scale parameter b are usually specified by the user.

The HB estimator of θ_i is obtained as :

$$\hat{\theta}_i^{HB} = E(\theta_i | \hat{\boldsymbol{\theta}}) = E_{\sigma_v^2} [\tilde{\theta}_i^{HB}] \quad , \quad (5.4.9)$$

where $\tilde{\theta}_i^{HB}$ is given by formula (5.4.3) and $E_{\sigma_v^2}$ denotes the expectation with respect to $f(\sigma_v^2 | \hat{\boldsymbol{\theta}})$, the posterior distribution of σ_v^2 . The posterior variance of θ_i is:

$$V(\hat{\theta}_i | \hat{\boldsymbol{\theta}}) = E_{\sigma_v^2} [M_{1i}(\sigma_v^2)] + V_{\sigma_v^2} [\tilde{\theta}_i^{HB}] \quad (5.4.10)$$

where $V_{\sigma_v^2}$ denotes the variance with respect to $f(\sigma_v^2 | \hat{\boldsymbol{\theta}})$. The evaluation of $\hat{\theta}_i^{HB}$ and $V(\theta_i | \hat{\boldsymbol{\theta}})$ involves single dimensional integrations.

In the case when σ_v^2 is unknown, Monte Carlo Markov Chain (MCMC) methods are used to generate simulated samples $\{\theta_i^{(j)}, \dots, \theta_m^{(j)}; j = 1, \dots, J\}$ (Rao, 2001 and Rao, 2003). Simulated samples may be generated from the following conditional distributions:

$$(i) \quad \beta | \boldsymbol{\theta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N_p [\beta^*, \sigma_v^2 (\mathbf{x}_i \mathbf{x}'_i)^{-1}] \quad (5.4.11)$$

$$(ii) \quad \theta_i | \beta, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N_p [\hat{\theta}_i^{EB}, \gamma_i \psi_i] \quad , \quad i = 1, \dots, m \quad (5.4.12)$$

$$(iii) \quad \sigma_v^{-2} | \beta, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim G \left[\frac{m}{2} + a, \frac{1}{2} \sum_i (\theta_i - \mathbf{x}'_i \beta)^2 + b \right] \quad (5.4.13)$$

where

$\sigma_v^{-2} \sim G(a, b)$, $a > 0$, $b > 0$ and are set to be very small ,
 $G(a, b)$ denotes a gamma function with a as the shape
parameter and b as the scale parameter

(Note: The software BUGS uses $a=b=0.001$
as the default setting.)

(Note: BUGS= Bayesian Analysis Using the Gibbs Sampler)

$$\beta^* = \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_i \mathbf{x}_i \hat{\theta}_i \right), \text{ as defined by formula (5.3.6)}$$

$N_p(\cdot)$ denotes a p -variate normal , and

$\hat{\theta}_i^{EB}$ is the EB estimator as defined by formula (5.3.1) .

Denote the MCMC samples as $\{(\beta^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma_v^{2(k)}), k = d + 1, \dots, d + D\}$,
where d is the length of burn-in and D is chosen to be sufficiently large.
Using the MCMC simulated samples, the HB estimator for θ_i (Rao, 2003) is
given by:

$$\hat{\theta}_i^{HB} = \frac{1}{D} \sum_{k=d+1}^{d+D} \hat{\theta}_i^{EB}(\beta^{(k)}, \sigma_v^{2(k)}) \quad . \quad (5.4.14)$$

The posterior variance is estimated by:

$$\begin{aligned} \hat{V}(\theta_i | \hat{\boldsymbol{\theta}}) &= \frac{1}{D} \sum_{k=d+1}^{d+D} g_{1i}(\sigma_v^{2(k)}) \\ &+ \frac{1}{D-1} \sum_{k=d+1}^{d+D} \left[\hat{\theta}_i^{EB}(\beta^{(k)}, \sigma_v^{2(k)}) - \hat{\theta}_i^{HB} \right]^2 \end{aligned} \quad (5.4.15)$$

In the case when L independent runs are generated instead of a single long run, the estimator of the posterior mean θ_i^{HB} may be calculated as:

$$\hat{\theta}_i^{HB} = \frac{1}{Ld} \sum_{l=1}^L \sum_{k=d+1}^{2d} \theta_i^{(lk)} = \frac{1}{L} \sum_{l=1}^L \theta_i^{(l\cdot)} = \theta_i^{(\cdot\cdot)} \quad , \quad (5.4.16)$$

where $\theta_i^{(lk)}$ is the k^{th} retained value in the l^{th} run of length $2d$ with the first d burn-in iterations deleted. The corresponding posterior variance is estimated as:

$$\hat{V}(\theta_i | \hat{\boldsymbol{\theta}}) = \frac{d-1}{d} W_i + \frac{1}{d} B_i \quad , \quad (5.4.17)$$

where

$$B_i = \frac{d \sum_{l=1}^L (\theta_i^{(l\cdot)} - \theta_i^{(\cdot\cdot)})^2}{(L-1)}$$

is the between-run variance and

$$W_i = \sum_{l=1}^L \sum_{k=d+1}^{2d} \frac{(\theta_i^{(lk)} - \theta_i^{(l\cdot)})^2}{[L(d-1)]} \quad (5.4.18)$$

is the within-run variance.

An alternative HB estimator (Belmonte, 1998) that uses solely the sample data, i.e., without simulation, and under the assumptions that $\hat{\theta}_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ and that $m \geq 4$, is :

$$\hat{\theta}_i^{HB} = \hat{\theta}_i - \frac{m-3}{s^2} \left[1 - H_{n-3} \left(\frac{s^2}{2} \right) \right] (\hat{\theta}_i - \bar{\hat{\theta}}) \quad (5.4.19)$$

where

$$H_r(v) = \begin{cases} \frac{v^{r/2}}{\left[\frac{r}{2} \right]! \left\{ e^v - \sum_{i=0}^{(r/2)-1} \frac{v^i}{i!} \right\}} & \text{if } r \text{ is even} \\ \frac{v^{r/2}}{\Gamma\left(\frac{r}{2} + 1\right) \left\{ e^v [2\Phi(\sqrt{2v} - 1)] \right\} - S(r, v)} & \text{if } r \text{ is odd} \end{cases} \quad (5.4.20)$$

$$S(r, v) = \sum_{i=0}^{(r-3)/2} \frac{v^{i+1/2}}{\Gamma(i + 3/2)} \quad ,$$

Φ is the standard normal cumulative distribution function ,

$$s^2 = \sum_{i=1}^m (\hat{\theta}_i - \bar{\hat{\theta}})^2$$

and $\hat{\theta}_i$ is the usual direct estimator of θ_i .

Formula (5.4.15) may be more simply expressed as:

$$\hat{\theta}_i^{HB} = \hat{\theta}_i + g_i(\hat{\boldsymbol{\theta}}) \quad (5.4.21)$$

where

$$g_i(\hat{\boldsymbol{\theta}}) = -\frac{m-3}{s^2} \left[1 - H_{n-3} \left(\frac{s^2}{2} \right) \right] (\hat{\theta}_i - \bar{\hat{\theta}}) \quad (5.4.22)$$

The corresponding estimator of the posterior variance is given by:

$$\begin{aligned}
mse(\hat{\theta}_i^{HB}) = & \\
1 - 2(m-3) & \left[\frac{2(\hat{\theta}_i - \bar{\theta})^2}{(s^2)^2} \left(-1 + H_{m-3}\left(\frac{s^2}{2}\right) - \frac{(m-3)}{2} H_{m-3}\left(\frac{s^2}{2}\right) \left(1 - \frac{s^2}{(m-3)} - H_{m-3}\left(\frac{s^2}{2}\right) \right) \right) \right. \\
& \left. + \left(\frac{1 - H_{m-3}\left(\frac{s^2}{2}\right)}{s^2} \right) \left(1 - \frac{1}{m} \right) \right] + \frac{(m-3)^2}{s^4} \left(1 - H_{m-3}\left(\frac{s^2}{2}\right) \right)^2 (\hat{\theta}_i - \bar{\theta})^2 \quad (5.4.23)
\end{aligned}$$

5.4.2 Application: The 1991 Canadian Census Undercoverage

The following notations are used to illustrate this example on the use of HB in small area estimation: i to denote the small area (i.e., in this case, a province), C the census count, M the number of persons missed in the census and U the undercoverage rate.

In 1991, Statistics Canada decided to adjust the 1991 Census-based population estimates by adding to it the estimated net undercoverage (Dick and You, 1997). The term “net undercoverage” is defined here as the estimated difference between gross undercoverage and gross overcoverage.

You and Rao (2002) estimated the number of missing persons M_i and the undercoverage rate U_i , defined as $U_i = M_i / (M_i + C_i)$, for each province i ($i = 1, \dots, 10$). The log transformation of the census count C_i was

used as the auxiliary variable, i.e., $x_{li} = \log(C_i)$ and the linking model for U_i parallel to formula (5.2.1) is:

$$\theta_i = \log \left\{ \frac{U_i}{(U_i + C_i)} \right\} = \beta_0 + x_{li} \beta_1 + v_i \quad (5.4.24)$$

The prior on the model parameters $(\beta_0, \beta_1, \sigma_v^2)$ is given by formula (5.4.8) with $\sigma_v^{-2} \sim G(0.001, 0.001)$ to reflect the lack of prior information on σ_v^2 . Gibbs samples were simulated with $L=8$ independent runs and $d=500$ burn-in iterations for each run.

Using formulae (5.4.16) and (5.4.17), the estimated posterior mean M_i^{HB} and variance $\hat{V}(M_i | \hat{\mathbf{M}})$ were calculated. Estimates for the posterior mean of the undercoverage rates, U_i^{HB} , and the corresponding estimate of the posterior variance, $\hat{V}(U_i | \hat{\mathbf{U}})$, were also generated using the same formulae. Table 5.4.1 (You and Rao, 2002) presents the HB estimates of the posterior means and their associated CV's. For purposes of comparison, the direct estimates, together with their corresponding CV's, are likewise shown in this table.

Table 5.4.1. 1991 Canadian Census Undercount Estimates
And Associated CVs

Prov	\hat{M}_i	$CV(\hat{M}_i)$	\hat{M}_i^{HB}	$CV(\hat{M}_i^{HB})$	$\hat{U}_i(\%)$	$CV(\hat{U}_i)$	$\hat{U}_i^{HB}(\%)$	$CV(\hat{U}_i^{HB})$
Nfld	11566	0.16	10782	0.14	1.99	0.16	1.86	0.13
PEI	1220	0.30	1486	0.19	0.93	0.30	1.13	0.19
NS	17329	0.20	17410	0.14	1.89	0.20	1.90	0.14
NB	24280	0.14	18948	0.17	3.25	0.13	2.55	0.17
Qué	184473	0.08	189599	0.08	2.58	0.08	2.65	0.08
Ont	381104	0.08	368424	0.08	3.64	0.08	3.52	0.08
Man	20691	0.21	21504	0.14	1.86	0.20	1.93	0.14
Sask	18106	0.19	18822	0.14	1.80	0.18	1.87	0.13
Alta	51825	0.15	55392	0.12	2.01	0.14	2.13	0.12
BC	92236	0.10	89929	0.09	2.73	0.10	2.67	0.09

It may be observed from Table 5.4.1 that, on the basis of their CV's, the HB estimates \hat{M}_i^{HB} and \hat{U}_i^{HB} for the provinces of Newfoundland, PEI, Nova Scotia, Manitoba, Saskatchewan and Alberta perform better than their direct estimates counterpart. For the province of New Brunswick, the direct estimates outperformed the HB estimates. In the case of the two biggest provinces (Ontario and Québec), the CV's of the direct estimates and the HB estimates are equal and this could be attributed to their relatively big sample sizes; those for British Columbia, the third biggest province in terms of population, are nearly equal.

6 Case Study: Small area estimates of unemployment in the Philippines for April 1994

One of the key indicators of economic and social health is unemployment. Statistics on unemployment are important tools for making, implementing and evaluating government programs and policies, both at the national and subnational (e.g., regional, provincial) levels. The need for them is more acute at the subnational level where limited resources have to be targeted to areas which need them most. In other words, local governments use unemployment statistics for planning and budgetary purposes and to determine the need for local employment and training services.

6.1 Definition of terms

The National Statistics Office (NSO) of the Philippines follows the standard definitions and guidelines on labour force statistics recommended by the United Nations International Labour Organisation (ILO). The term labour force or economically active population refers to the group of individuals in the population who are 15 years old or older who are either employed or unemployed. Employed persons are those who are 15 years old or over during the reference period (i.e., in the labour force) and are either at work or with a job but not at work. A person is “at work” if he had done any work even for one hour during the reference for pay or profit, or work without pay on the farm or business enterprise operated by the a member of the same household. A person is “with a job but not at work” if he has a job or business but is not

at work due to some temporary illness/injury, vacation or other reasons. A person who expects to report for work or to start operation of a farm or business enterprise within two weeks from the date of the enumerator's visit is considered employed.

Unemployed persons include those who are in the labour force and who have no job or business and are actively looking for work. They may also be not looking for work because of their belief that no work is available or because of temporary illness/disability, bad weather, pending job application or waiting for a job interview. The unemployment rate is then the proportion of the labour force that is unemployed (NSO, 2003).

In this case study, a small area is defined as a cross classification of region and age group. There are fifteen (15) regions in the Philippines, to wit:

1. National Capital Region (NCR)
2. Cordillera Administrative Region (C.A.R.)
3. Region I – Ilocos Region
4. Region II – Cagayan Valley
5. Region III – Central Luzon
6. Region IV – Southern Tagalog
7. Region V – Bicol Region
8. Region VI – Western Visayas
9. Region VII – Central Visayas
10. Region VIII – Eastern Visayas
11. Region IX – Western Mindanao
12. Region X – Northern Mindanao
13. Region XI – Southern Mindanao
14. Region XII – Central Mindanao
15. Autonomous Region in Muslim Mindanao (A.R.M.M.)

The relative sizes of these regions, in terms of contribution to the gross domestic product (GDP) for the years 1999 and 200, may be discerned from Table 6.1. It may be observed that the region NCR, where the capital city

Manila is located, is by far the biggest region in terms of percent contribution to the GDP. The second biggest region, Region IV, is contiguous to NCR and is about half the size of the latter. The medium-sized regions account for about 6% to 9 % of the GDP (i.e., Regions III, VI, VII AND XI) with the rest of the regions contributing significantly less and are classified as small.

Table 6.1 Percentage Distribution
of the Gross Domestic Product
at Constant (1985) Prices,
1999 and 2000

REGION	1999	2000
PHILIPPINES	100.00	100.00
NCR	30.59	31.09
C.A.R.	2.43	2.33
I	3.10	3.18
II	2.34	2.26
III	9.03	8.90
IV	15.37	15.18
V	2.82	2.71
VI	7.12	7.02
VII	6.87	6.81
VIII	2.37	2.40
IX	2.79	2.83
X	3.80	3.82
XI	6.19	6.31
XII	2.69	2.69
A.R.M.M.	1.00	0.96
CARAGA	1.48	1.50

Note: Caraga is a new region created in 1995 and carved out of Regions X and XI.

Source: National Statistical Coordination Board, Philippines

Six (6) age groups are considered, namely:

1. 15 – 19 years old
2. 20 – 24 years old
3. 25 – 34 years old
4. 35 – 44 years old
5. 45 – 54 years old
6. 55 – OVER

The resulting cross classification of 15 regions and 6 age groups yields 90 small areas.

6.2 Notations

In addition to the previously adopted notations, the following will be used:

1. \hat{E}_{ij} : estimated employment level for region i and age group j
2. \hat{U}_{ij} : estimated unemployment level for region i and age group j
3. r_{ij} : direct estimate of unemployment rate for region i and age group j
4. r_{ij}^{EB} : empirical Bayes estimate of unemployment rate for region i and age group j
5. SE_{ij} : standard error of \hat{U}_{ij}
6. CV_{ij} : coefficient of variation of \hat{U}_{ij} estimated as SE_{ij}/\hat{U}_{ij}

6.3 Available data

Labour force information is collected primarily through the Labour Force Survey (LFS). The LFS collects information on demographic characteristics and labour force status (i.e., employed, unemployed, or not in the labour force). In the Philippines, this survey is one of the modules of the Integrated Survey of Households (ISH) of the National Statistics Office (NSO) and, as such, adopts the sampling design of the latter. The ISH has a stratified two-stage sampling design (Barrios, 1998). The *barangays* (i.e., villages)

which comprise a province, and, likewise, a city or municipality, are classified into urban or rural. The sample design involves the selection of *barangays* as the primary sampling unit (PSU) and then of the households within each sample *barangay* as the secondary sampling unit (SSU).

This case study focuses on the generation and analysis of unemployment rates in the small areas of the Philippines for the period April 1994. From the April 1994 round of the LFS were obtained regional estimates, by age group, of employment and unemployment as well as their respective standard errors and CV's. Table 6.2 contains the data used for this case study.

Table 6.2 Employment Status of the Labour Force Population,
Standard Error and Coefficient of Variation of the Unemployment Estimate,
by Region and Age Group: April 1994

Area/ age-group	Employment	U n e m p l o y m e n t		
	Estimates	Estimates	SE	CV
PHILIPPINES	25,561,864	3,175,852	67,270	2.12
15-19	3,037,601	1,079,632	36,432	3.37
20-24	3,190,936	865,971	26,437	3.05
25-34	5,961,736	589,499	20,795	3.53
35-44	5,742,527	262,039	12,732	4.86
45-54	3,950,384	179,099	10,343	5.77
55- OVER	3,678,680	199,613	11,607	5.81
NATIONAL CAPITAL REGION	3,044,675	609,739	20,404	3.35
15-19	187,329	120,001	9,604	8.00
20-24	469,977	207,358	10,936	5.27
25-34	896,915	158,758	8,911	5.61
35-44	773,617	67,245	6,054	9.00
45-54	447,084	33,569	3,883	11.57
55- OVER	269,753	22,809	3,308	14.50
CORDILLERA ADM. REGION	535,649	44,055	7,761	17.62
15-19	59,657	17,500	4,615	26.37
20-24	61,277	11,950	2,566	21.47
25-34	140,561	7,530	2,004	26.61
35-44	113,927	3,509	1,341	38.23
45-54	74,881	2,217	1,035	46.69
55- OVER	85,346	1,350	855	63.35
REGION I - ILOCOS REGION	1,327,630	172,186	14,516	8.43
15-19	120,162	62,054	9,367	15.10
20-24	163,877	50,858	6,427	12.64
25-34	309,344	26,509	4,655	17.56
35-44	279,261	12,598	2,748	21.82
45-54	233,838	6,885	2,043	29.67
55- OVER	221,148	13,283	2,859	21.52

Table 6.2

Area/ age-group	Employment Estimates	U n e m p l o y m e n t		
		Estimates	SE	CV
REGION II -CAGAYAN VALLEY	1,220,068	97,222	13,526	13.91
15-19	181,933	43,817	8,723	19.91
20-24	140,559	16,804	2,928	17.42
25-34	262,145	17,339	3,888	22.42
35-44	283,574	4,470	1,894	42.37
45-54	180,772	5,322	1,725	32.40
55- OVER	171,085	9,471	3,368	35.56
REGION III - CENTRAL LUZON	2,269,373	373,907	21,614	5.78
15-19	245,250	123,930	9,195	7.42
20-24	368,185	108,174	8,816	8.15
25-34	551,601	60,695	6,251	10.30
35-44	503,588	30,520	4,605	15.09
45-54	335,965	26,953	3,707	13.75
55- OVER	264,784	23,636	4,077	17.25
REGION IV- SOUTHERN TAGALOG	3,392,789	342,465	23,099	6.74
15-19	371,232	92,322	9,498	10.29
20-24	434,384	89,187	8,507	9.54
25-34	852,754	77,254	8,399	10.87
35-44	793,052	38,466	4,832	12.56
45-54	509,362	26,819	4,201	15.66
55- OVER	432,005	18,417	3,198	17.37
REGION V- BICOL REGION	1,902,115	182,633	19,898	10.89
15-19	277,630	69,976	9,379	13.40
20-24	192,285	46,047	8,472	18.40
25-34	392,064	28,707	5,613	19.55
35-44	402,685	11,783	2,779	23.59
45-54	300,920	10,439	2,954	28.30
55- OVER	336,531	15,682	3,445	21.97
REGION VI- WESTERN VISAYAS	2,288,206	331,727	24,772	7.47
15-19	270,266	131,009	15,109	11.53
20-24	279,943	77,264	7,933	10.27
25-34	478,588	51,153	6,332	12.38
35-44	526,454	24,159	4,176	17.28
45-54	386,296	21,004	3,546	16.88
55- OVER	346,659	27,138	4,526	16.68

Table 6.2

Area/ age-group	Employment Estimates	U n e m p l o y m e n t		
		Estimates	SE	CV
REGION VII- CENTRAL VISAYAS	1,989,155	229,545	17,857	7.78
15-19	215,254	83,260	10,607	12.74
20-24	221,047	65,332	7,166	10.97
25-34	419,443	41,682	5,700	13.68
35-44	410,688	16,714	3,466	20.74
45-54	310,159	10,078	2,783	27.62
55- OVER	412,564	12,478	2,703	21.66
REGION VIII- EASTERN VISAYAS	1,433,351	148,382	18,679	12.59
15-19	201,203	64,750	9,801	15.14
20-24	144,456	23,409	5,139	21.95
25-34	266,765	24,702	5,578	22.58
35-44	291,489	11,167	3,400	30.45
45-54	229,351	8,621	2,472	28.68
55- OVER	300,087	15,734	3,617	22.99
REGION IX - WESTERN MINDANAO	988,074	64,602	9,121	14.12
15-19	136,257	25,216	6,023	23.88
20-24	115,077	18,502	4,134	22.34
25-34	231,542	9,443	2,402	25.43
35-44	211,338	5,671	1,736	30.61
45-54	160,802	2,468	1,189	48.18
55- OVER	133,058	3,302	1,372	41.54
REGION X - NORTHERN MINDANAO	1,613,748	191,960	16,722	8.71
15-19	242,589	61,630	9,406	15.26
20-24	178,239	54,936	6,811	12.40
25-34	349,068	33,619	4,955	14.74
35-44	343,790	15,224	2,804	18.42
45-54	262,623	9,755	2,441	25.02
55- OVER	237,439	16,796	3,215	19.14

6.4 Model selection

The characteristic of interest r in this case study is the unemployment rate. The selected auxiliary variables are regional identification and age group, both of which are categorical in nature. Note that these 2 variables are also used to define a small area.

Following its definition in section 6.1, the estimator of unemployment rate for a given small area is computed as:

$$r_{ij} = \frac{\hat{U}_{ij}}{\hat{U}_{ij} + \hat{E}_{ij}} \quad , \quad i = 1, \dots, 15 \quad , \quad j = 1, \dots, 6 \quad (6.1)$$

We define $\hat{\theta}_{ij} = g(r_{ij})$ and assume that :

$$\hat{\theta}_{ij} \sim_{iid} N(\theta_{ij}, \psi_{ij})$$

$$\theta_{ij} \sim_{iid} N(\theta_{ij}, \sigma_v^2)$$

Two (2) possible ways of defining θ_{ij} are explored:

$$\text{Scenario 1: } \hat{\theta}_{ij} = r_{ij} \quad \text{and} \quad (6.2)$$

$$\text{Scenario 2: } \hat{\theta}_{ij} = \log(r_{ij}) \quad . \quad (6.3)$$

In the first scenario (i.e., $\hat{\theta}_{ij} = r_{ij}$), we calculate ψ_{ij} as

$$\hat{\psi}_{ij} = \left(\frac{SE_{ij}}{\hat{E}_{ij} + \hat{U}_{ij}} \right)^2 \quad ; \quad (6.4)$$

in the second one (i.e., $\hat{\theta}_{ij} = \log(r_{ij})$) as

$$\hat{\psi}_{ij} = (CV_{ij})^2 . \quad (6.5)$$

Note that $\hat{\psi}_{ij}$ is calculated with E_{ij} treated as fixed. Formulas (6.4) and (6.5) are used to approximate ψ_{ij} since the published estimates of variances are available only at the regional and national levels without the age group breakdowns. The resulting estimates of the variances aggregated at the regional and national levels reveal some overestimations when compared with the published figures. These overestimations are, nevertheless, deemed negligible.

In both scenarios, formulas (5.3.4), (5.3.5) and (5.3.6) are used to compute for the estimates of the vector of parameters $\boldsymbol{\eta} = (\beta, \sigma_v^2)$. To do this, the model shown below is fitted:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (6.6)$$

where

θ_{ij} = function of the unemployment rate

μ = overall mean effect

α_i = effect of region i

β_j = effect of age group j

ε_{ij} = random error component

This model is fitted with the constraint that

$$\alpha_{15} = \beta_6 = 0 . \quad (6.7)$$

To determine which of the two (2) scenarios [i.e., $\hat{\theta}_{ij} = r_{ij}$ or $\hat{\theta}_{ij} = \log(r_{ij})$] is more appropriate for the estimation of unemployment rates using the empirical Bayes approach, two (2) kinds of analysis are done: first, the distribution of r_{ij}^{EB} under each scenario is compared with that of r_{ij} and, second, the validity of the regression model under the 2 scenarios is evaluated by analyzing the normalized residuals.

The comparison of the distribution of the empirical Bayes estimators under each of the 2 scenarios with that of the direct estimators may be done using figures 1, 2 and 3. Figure 1 presents the histogram of the direct estimators r_{ij} ; figure 2 shows that of the empirical Bayes estimators under scenario 1; and figure 3 contains that of the empirical Bayes estimators under scenario 2. It is obvious that, between the 2 scenarios, the distribution of the empirical Bayes estimators under scenario 2 (i.e., log-linear model) better approximates that of the direct estimators.

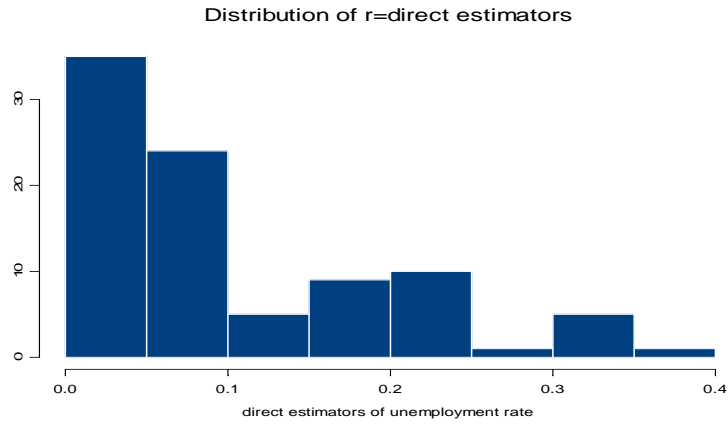


Figure 1. Histogram of the direct estimators

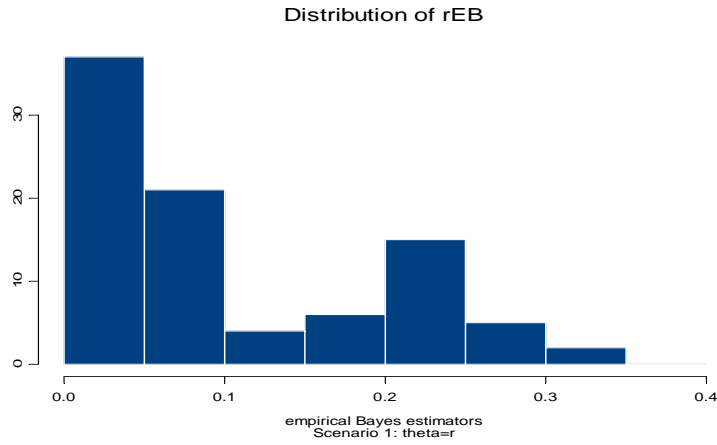


Figure 2. Histogram of the EB estimators for Scenario 1

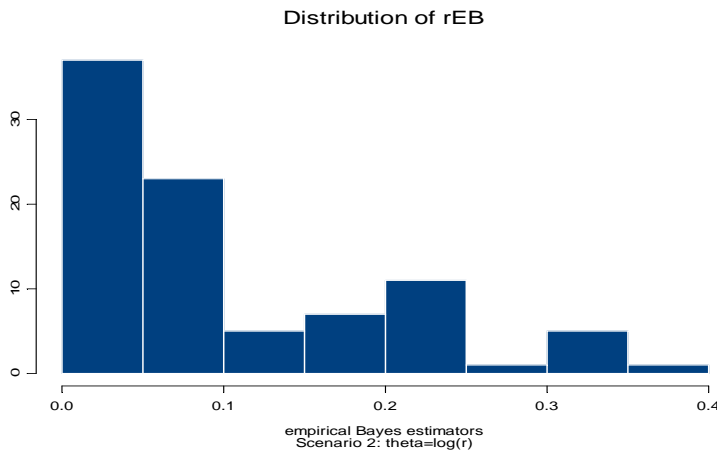


Figure 3. Histogram of the EB estimators for Scenario 2

Distribution of normalized residuals

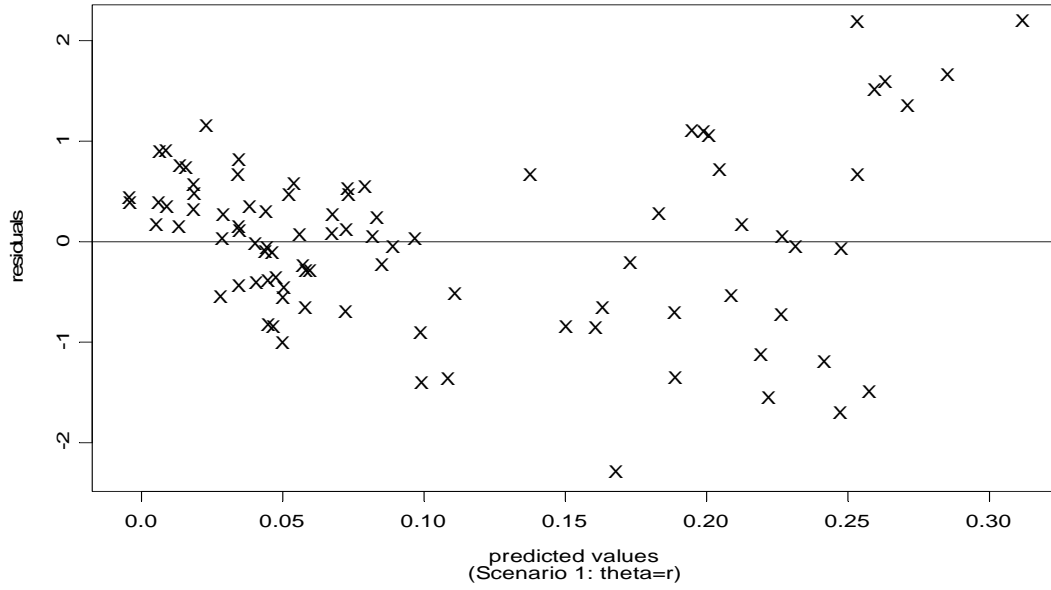


Figure 4. Scenario 1: Normalized residuals as a function of predicted values

Distribution of normalized residuals

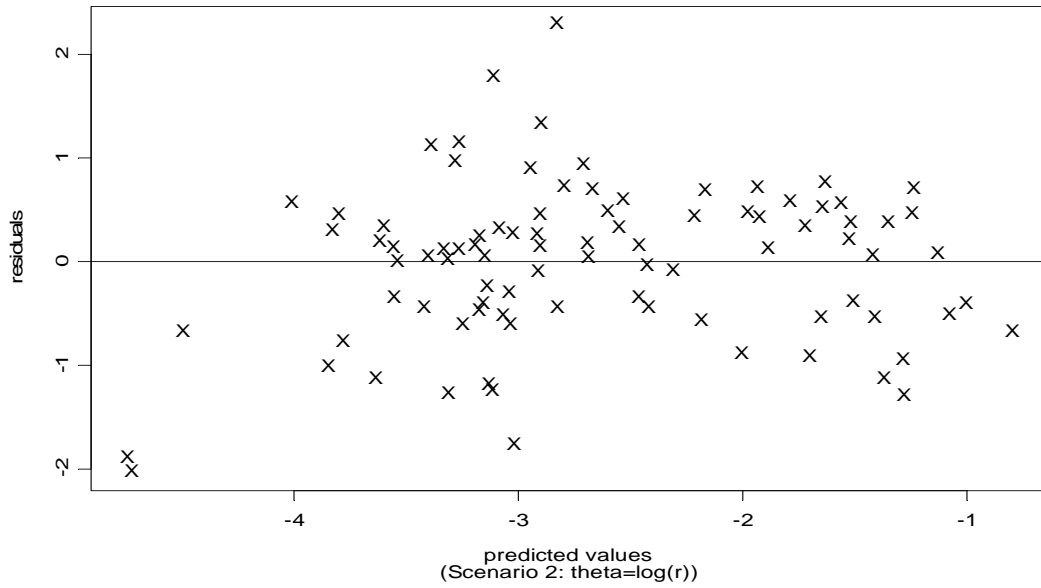


Figure 5. Scenario 2: Normalized residuals as a function of predicted values

The validation of the regression models for scenario 1 and scenario 2 may be done with the aid of the graphs of normalized residuals versus the predicted values, which in this case are the $\hat{\theta}_{ij}$. Figures 4 and 5 (see preceding page) contain these graphs. It may be observed that the form of the graph shown by Figure 4 (that for Scenario 1 where $\theta=r$) resembles that of a funnel. This signifies that the residuals have a tendency to increase with a corresponding increase in predicted values and, hence, the assumption of independence of observations is not satisfied. The funnel shape of the graph also means that the assumptions of the homogeneity of the variance of the residuals and the linearity of relations between the dependent and the independent variables are violated. These observations imply that some kind of transformation or redefinition of variable(s) may be in order before the regression model (6.5) could be applied.

Figure 5 (that for Scenario 2 where $\theta=\log(r)$) shows the graph of the normalized residuals versus the predicted values $\hat{\theta}_{ij}$ under scenario 2. It is apparent that the normalized residuals are more or less randomly distributed around zero and that they are more or less contained within the band $[-2,2]$. This form of the graph signify that the assumptions of independence of observations, homogeneity of the variance of the residuals, linearity of the relations between the independent and the dependent variables, the absence of outliers and normality of residuals are all satisfied. The statistic W from the Shapiro-Wilks test ($W=0.97996$) and its corresponding p-value (0.47869) indicate the normality of the residuals.

The analysis of figures 1 to 5 implies that, in this particular case study, a logarithmic transformation of r_{ij} should be adopted when fitting the

regression model (6.5), i.e., $\hat{\theta}_{ij} = \log(r_{ij})$. The exclusion of the interaction effect $(\alpha\beta)_{ij}$ in the regression model (6.5) is justified by the fact that scenario 2 already satisfies the basic assumptions of a linear regression model. It may also be interesting to note that when the empirical Bayes estimates of unemployment rates, obtained in scenario 2, are used to rank the age groups in each region in terms of unemployment rates the results are more or less homogeneous among the 15 regions. As shown in Table 6.3, the rates of unemployment among the youth (i.e., age groups 15-19, 20-24 and 25-34) in all the regions seem to be very pronounced.

Table 6.3 Age Groups Ranked by Unemployment Rates,
By Region

	15-19	20-24	25-34	35-44	45-54	55-OVER
NCR	6	5	4	2	1	3
C.A.R.	6	5	4	1	2	3
REG. I	6	5	4	2	1	3
REG. II	6	5	4	1	2	3
REG. III	6	5	4	1	2	3
REG. IV	6	5	4	1	3	2
REG. V	6	5	4	1	2	3
REG. VI	6	5	4	1	2	3
REG. VII	6	5	4	2	1	3
REG. VIII	6	5	4	1	2	3
REG. IX	6	5	4	2	1	3
REG. X	6	5	4	2	1	3
REG. XI	6	5	4	1	2	3
REG. XII	6	5	4	1	2	3
A.R.M.M.	6	5	4	1	2	3

Note: 1 denotes lowest unemployment rate within region

Similarly, Table 6.4 shows the ranking of regions by unemployment rates but this time within each age group. It may be observed that, in all the age groups, the rate of unemployment is highest in NCR, the biggest region

of the country. It is apparent that the big regions (e.g., REG.III, REG. VI, REG.XI) have the higher unemployment rates compared with the small regions (e.g., A.R.M.M., REG.XII, REG.II). The observations made using these 2 tables lead to the conclusion that unemployment rates are highest among the youth in the big regions.

Table 6.4 Regions Ranked by Unemployment Rates,
By Age Group

	15-19	20-24	25-34	35-44	45-54	55-OVER
NCR	15	15	15	15	15	15
C.A.R.	5	5	4	5	5	4
REG. I	12	13	9	11	9	11
REG. II	4	3	5	4	4	5
REG. III	14	14	14	14	14	14
REG. IV	7	7	8	12	12	8
REG. V	6	8	6	6	6	7
REG. VI	13	10	13	13	13	13
REG. VII	10	9	10	8	8	6
REG. VIII	9	6	7	7	7	9
REG. IX	2	4	3	3	2	2
REG. X	8	12	11	10	10	12
REG. XI	11	11	12	9	11	10
REG. XII	3	2	2	2	3	3
A.R.M.M.	1	1	1	1	1	1

Note: 1 denotes smallest unemployment rate within age group

6.5 Results and Analysis

In this section, the results and the analysis for scenario 2 [i.e., $\hat{\theta}_{ij} = \log(r_{ij})$] are presented. Table 6.5 shows the resulting regression coefficients, the corresponding standard errors, t-statistics and p-values.

Table 6.5 Regression Coefficients, Standard Errors, t-statistics and p-values

	coef	std.err	t.stat	p.value
Intercept	-4.50099	0.176706	-25.4717	0.000000
a ₁	2.03220	0.181743	11.1818	0.000000
a ₂	1.18478	0.211859	5.5923	0.000000
a ₃	1.59228	0.190856	8.3429	0.000000
a ₄	1.10678	0.200896	5.5092	0.000001
a ₅	1.82534	0.183716	9.9356	0.000000
a ₆	1.46013	0.184288	7.9231	0.000000
a ₇	1.32210	0.192490	6.8684	0.000000
a ₈	1.69972	0.185531	9.1614	0.000000
a ₉	1.47653	0.188741	7.8231	0.000000
a ₁₀	1.40969	0.196018	7.1916	0.000000
a ₁₁	0.94239	0.206835	4.5562	0.000021
a ₁₂	1.54888	0.188690	8.2086	0.000000
a ₁₃	1.58468	0.186212	8.5101	0.000000
a ₁₄	0.89491	0.212593	4.2095	0.000075
b ₁	1.66639	0.085132	19.5743	0.000000
b ₂	1.38530	0.085289	16.2424	0.000000
b ₃	0.48535	0.087319	5.5584	0.000000
b ₄	-0.24646	0.092277	-2.6709	0.009401
b ₅	-0.22875	0.096101	-2.3803	0.020027

Note: The a_i and the b_j are the estimators of α_i and β_j , respectively.

The computed F-statistic is 63.764 with a p-value < 0.0001 and $R^2 = 0.9454$. All these imply that the fitted regression equation may be retained. The value of $\hat{\sigma}_v^2$ is 0.0338, a value smaller than most of the $\hat{\psi}_{ij}$ (see

Table 6.6). Its estimated standard error, calculated as the square root of the estimated variance $\hat{h}(\hat{\sigma}_v^2)$, is 0.018328 .

Table 6.6 presents the resulting estimates of the unemployment rates for the small areas. It may be observed that, in general, the empirical Bayes estimates of unemployment rates do not differ significantly from the direct estimates. For the large regions (i.e., NCR, REG. IV and REG. VII), the difference between the direct estimates and the empirical Bayes estimates is almost negligible. The difference between these 2 types of estimates is more noticeable among the smallest regions (i.e., Reg. II, REG. XII and A.R.M.M.). The biggest difference between the direct and the empirical Bayes estimates is observed for the small area [A.R.M.M., 15-19] where the direct estimate of unemployment rate is 0.167533 while that of the empirical Bayes is 0.069528; the direct estimate is more than twice the empirical Bayes estimate.

Table 6.6 Unemployment Rates: Direct Estimates,
Empirical Bayes Estimates
and Variance of the Direct Estimates for the Small Areas

Area	AgeGroup	r	rEB	rsigmai2
NCR	15-19	0.390463	0.399142	0.000977
NCR	20-24	0.306138	0.308478	0.000261
NCR	25-34	0.150386	0.149251	0.000071
NCR	35-44	0.079971	0.077099	0.000052
NCR	45-54	0.069840	0.069131	0.000065
NCR	55-OVER	0.077963	0.080477	0.000128
C.A.R.	15-19	0.226810	0.202810	0.003578
C.A.R.	20-24	0.163191	0.152442	0.001228
C.A.R.	25-34	0.050847	0.056209	0.000183
C.A.R.	35-44	0.029880	0.028642	0.000130
C.A.R.	45-54	0.028756	0.028855	0.000180
C.A.R.	55-OVER	0.015572	0.033983	0.000097
REG. I	15-19	0.340552	0.318629	0.002643
REG. I	20-24	0.236841	0.230610	0.000896
REG. I	25-34	0.078930	0.083416	0.000192
REG. I	35-44	0.043165	0.042852	0.000089
REG. I	45-54	0.028601	0.038656	0.000072
REG. I	55-OVER	0.056661	0.055428	0.000149
REG. II	15-19	0.194095	0.185048	0.001493
REG. II	20-24	0.106785	0.118953	0.000346
REG. II	25-34	0.062039	0.057437	0.000194
REG. II	35-44	0.015518	0.024141	0.000043
REG. II	45-54	0.028598	0.027153	0.000086
REG. II	55-OVER	0.052455	0.036880	0.000348
REG. III	15-19	0.335690	0.339585	0.000620
REG. III	20-24	0.227085	0.234366	0.000343
REG. III	25-34	0.099127	0.102037	0.000104
REG. III	35-44	0.057142	0.055780	0.000074
REG. III	45-54	0.074267	0.066585	0.000104
REG. III	55-OVER	0.081950	0.075537	0.000200
REG. IV	15-19	0.199161	0.210859	0.000420
REG. IV	20-24	0.170344	0.174529	0.000264
REG. IV	25-34	0.083068	0.081630	0.000082
REG. IV	35-44	0.046260	0.043216	0.000034
REG. IV	45-54	0.050019	0.044569	0.000061
REG. IV	55-OVER	0.040888	0.044012	0.000050

Table 6.6

Area	AgeGroup	r	rEB	rsigmai2
REG. V	15-19	0.201308	0.207726	0.000728
REG. V	20-24	0.193205	0.179267	0.001264
REG. V	25-34	0.068225	0.067914	0.000178
REG. V	35-44	0.028429	0.030920	0.000045
REG. V	45-54	0.033527	0.033240	0.000090
REG. V	55-OVER	0.044524	0.042799	0.000096
REG. VI	15-19	0.326482	0.325056	0.001418
REG. VI	20-24	0.216300	0.222307	0.000493
REG. VI	25-34	0.096562	0.097217	0.000143
REG. VI	35-44	0.043877	0.045526	0.000058
REG. VI	45-54	0.051569	0.050054	0.000076
REG. VI	55-OVER	0.072601	0.066978	0.000147
REG. VII	15-19	0.278915	0.271659	0.001263
REG. VII	20-24	0.228131	0.218667	0.000626
REG. VII	25-34	0.090392	0.086129	0.000153
REG. VII	35-44	0.039106	0.038466	0.000066
REG. VII	45-54	0.031470	0.036288	0.000076
REG. VII	55-OVER	0.029357	0.039347	0.000040
REG. VIII	15-19	0.243464	0.242274	0.001358
REG. VIII	20-24	0.139451	0.162866	0.000937
REG. VIII	25-34	0.084751	0.078004	0.000366
REG. VIII	35-44	0.036897	0.035880	0.000126
REG. VIII	45-54	0.036227	0.036173	0.000108
REG. VIII	55-OVER	0.049819	0.047102	0.000131
REG. IX	15-19	0.156162	0.152734	0.001391
REG. IX	20-24	0.138510	0.123195	0.000958
REG. IX	25-34	0.039185	0.043706	0.000099
REG. IX	35-44	0.026133	0.023225	0.000064
REG. IX	45-54	0.015116	0.021520	0.000053
REG. IX	55-OVER	0.024215	0.027732	0.000101
REG. X	15-19	0.202584	0.229984	0.000956
REG. X	20-24	0.235600	0.226836	0.000853
REG. X	25-34	0.087850	0.086667	0.000168
REG. X	35-44	0.042405	0.041603	0.000061
REG. X	45-54	0.035814	0.039442	0.000080
REG. X	55-OVER	0.066065	0.058462	0.000160
REG. XI	15-19	0.313999	0.308916	0.000723
REG. XI	20-24	0.225914	0.223288	0.000635
REG. XI	25-34	0.087007	0.087315	0.000124
REG. XI	35-44	0.038399	0.039899	0.000033
REG. XI	45-54	0.039981	0.041885	0.000091
REG. XI	55-OVER	0.052600	0.053418	0.000109

Table 6.6

Area	AgeGroup	r	rEB	rsigmai2
REG. XII	15-19	0.174014	0.157003	0.001194
REG. XII	20-24	0.125172	0.113259	0.001240
REG. XII	25-34	0.026984	0.039651	0.000088
REG. XII	35-44	0.012787	0.019794	0.000035
REG. XII	45-54	0.024776	0.022064	0.000114
REG. XII	55-OVER	0.031426	0.027842	0.000162
A.R.M.M	15-19	0.167533	0.069528	0.004946
A.R.M.M	20-24	0.105383	0.050150	0.002264
A.R.M.M	25-34	0.024069	0.018698	0.000136
A.R.M.M	35-44	0.003393	0.007621	0.000035
A.R.M.M	45-54	0.003393	0.007625	0.000114
A.R.M.M	55-OVER	0.008188	0.010537	0.000162

Note: r=direct estimate; rEB=empirical Bayes estimates;
rsigmai2=variance of r

Table 6.7 contains the 2 measures of the mean squared error of the empirical Bayes estimates of the small area unemployment rates [i.e., $mse_{PR}(r_{ij})$ and $mse_{PRg}(r_{ij})$] as well as some measures of their efficiency. To obtain $mse_{PR}(r_{ij})$ and $mse_{PRg}(r_{ij})$, the Prasad-Rao estimator $mse_{PR}(\hat{\theta}_{ij}^{EB})$ and the generalized Prasad-Rao estimator $mse_{PRg}(\hat{\theta}_{ij}^{EB})$ are first computed; the first using formulas (5.3.11) and (5.3.13) and the latter with formula (5.3.14). Note that, in this case, the sampling variances are estimated. Moreover, ϕ_{ij} , the sampling variance of the estimated variance, is not available and hence has to be estimated. To do so, it is assumed that the CV of the sampling variance is equal to the CV of the direct estimators. Thus, the generalized Prasad-Rao estimator $mse_{PRg}(\hat{\theta}_{ij}^{EB})$ is obtained using the estimator $\hat{\phi}_{ij} = (s_{ij}^2 * CV(\hat{\theta}_{ij}))^2$. The resulting $mse_{PRg}(\hat{\theta}_{ij}^{EB})$ obtained in this manner is consistent with and not significantly different from the usual Prasad-Rao estimator $mse_{PR}(\hat{\theta}_{ij}^{EB})$.

The generated $mse_{PR}(\hat{\theta}_{ij})$ and $mse_{PRg}(\hat{\theta}_{ij})$ are then transformed into $mse_{PR}(r_{ij})$ and $mse_{PRg}(r_{ij})$ with the use of the relationship (Rao, 2003, p. 133):

$$mse(r_{ij}) = \exp(2\hat{\theta}_{ij}^{EB}) * mse(\hat{\theta}_{ij}^{EB}) \quad (6.8)$$

The efficiency of an empirical Bayes estimate of unemployment rate is calculated as:

$$eff_{ij} = \frac{\hat{\psi}_{ij}}{mse(r_{ij})} \quad (6.9)$$

The measures of efficiency of the empirical Bayes estimates, except for 3 small areas (REG. III, 20-24; REG. IV, 15-19; REG. VII, 55-over), are all higher than 1. The big gainers are the smaller regions (C.A.R., REG. I, REG. VIII, REG. IX, REG. XII and A.R.M.M.) with the highest efficiency observed in the region A.R.M.M. The empirical Bayes estimators for the big and more developed regions (e.g., NCR, REG. IV, REG. VII) show the most moderate gains in efficiency. The 2 measures of efficiency are consistent, as expected, and give the same results.

In those instances when the efficiency of the EB estimates are inferior to 1, the sampling variances corresponding to the small areas are relatively small [i.e., 0.000343 for REG. III, 20-24; 0.000420 for REG. IV, 15-19; 0.000040 for REG. VII, 55-over]. In such a case when the sampling variance ψ_{ij} is small, the addition of $g_{2ij}(\hat{\sigma}_v^2)$ and $2g_{3ij}(\hat{\sigma}_v^2)$ to $g_{1ij}(\hat{\sigma}_v^2)$ to obtain $mse_{PR}(\hat{\theta}_{ij}^{EB})$ may result in the latter being bigger than the sampling variance ψ_{ij} . This could explain the observed efficiencies inferior to 1.

Table 6.7 Mean Squared Errors of the Empirical Bayes Estimates of Unemployment Rates and their Efficiency

Area	AgeGroup	msePR	msePRg	effPR	effPRg
NCR	15-19	0.000966	0.000968	1.01139	1.00930
NCR	20-24	0.000260	0.000260	1.00385	1.00385
NCR	25-34	0.000069	0.000069	1.02899	1.02899
NCR	35-44	0.000045	0.000045	1.15556	1.15556
NCR	45-54	0.000056	0.000056	1.16071	1.16071
NCR	55-OVER	0.000107	0.000108	1.19626	1.18519
C.A.R.	15-19	0.001540	0.001568	2.32338	2.28189
C.A.R.	20-24	0.000715	0.000725	1.71748	1.69379
C.A.R.	25-34	0.000120	0.000122	1.52500	1.50000
C.A.R.	35-44	0.000040	0.000041	3.25000	3.17073
C.A.R.	45-54	0.000044	0.000045	4.09091	4.00000
C.A.R.	55-OVER	0.000066	0.000068	1.46970	1.42647
REG. I	15-19	0.001819	0.001835	1.45300	1.44033
REG. I	20-24	0.000730	0.000734	1.22740	1.22071
REG. I	25-34	0.000153	0.000154	1.25490	1.24675
REG. I	35-44	0.000052	0.000053	1.71154	1.67925
REG. I	45-54	0.000055	0.000056	1.30909	1.28571
REG. I	55-OVER	0.000086	0.000087	1.73256	1.71264
REG. II	15-19	0.000913	0.000926	1.63527	1.61231
REG. II	20-24	0.000321	0.000325	1.07788	1.06462
REG. II	25-34	0.000100	0.000102	1.94000	1.90196
REG. II	35-44	0.000027	0.000028	1.59259	1.53571
REG. II	45-54	0.000030	0.000031	2.86667	2.77419
REG. II	55-OVER	0.000059	0.000060	5.89831	5.80000
REG. III	15-19	0.000610	0.000610	1.01639	1.01639
REG. III	20-24	0.000346	0.000347	0.99133	0.98847
REG. III	25-34	0.000100	0.000100	1.04000	1.04000
REG. III	35-44	0.000055	0.000055	1.34545	1.34545
REG. III	45-54	0.000069	0.000069	1.50725	1.50725
REG. III	55-OVER	0.000120	0.000122	1.66667	1.63934
REG. IV	15-19	0.000426	0.000427	0.98592	0.98361
REG. IV	20-24	0.000256	0.000256	1.03125	1.03125
REG. IV	25-34	0.000070	0.000070	1.17143	1.17143
REG. IV	35-44	0.000025	0.000025	1.36000	1.36000
REG. IV	45-54	0.000037	0.000037	1.64865	1.64865
REG. IV	55-OVER	0.000041	0.000042	1.21951	1.19048
REG. V	15-19	0.000653	0.000657	1.11485	1.10807
REG. V	20-24	0.000750	0.000759	1.68533	1.66535
REG. V	25-34	0.000116	0.000117	1.53448	1.52137
REG. V	35-44	0.000029	0.000030	1.55172	1.50000
REG. V	45-54	0.000040	0.000040	2.25000	2.25000
REG. V	55-OVER	0.000053	0.000054	1.81132	1.77778
REG. VI	15-19	0.001230	0.001236	1.15285	1.14725
REG. VI	20-24	0.000473	0.000475	1.04228	1.03789
REG. VI	25-34	0.000124	0.000124	1.15323	1.15323
REG. VI	35-44	0.000044	0.000044	1.31818	1.31818
REG. VI	45-54	0.000052	0.000052	1.46154	1.46154
REG. VI	55-OVER	0.000091	0.000092	1.61538	1.59783

Table 6.7

Area	AgeGroup	msePR	msePRg	effPR	effPRg
REG. VII	15-19	0.001019	0.001024	1.23945	1.23340
REG. VII	20-24	0.000517	0.000519	1.21083	1.20617
REG. VII	25-34	0.000114	0.000115	1.34211	1.33043
REG. VII	35-44	0.000039	0.000040	1.69231	1.65000
REG. VII	45-54	0.000045	0.000046	1.68889	1.65217
REG. VII	55-OVER	0.000043	0.000044	0.93023	0.90909
REG. VIII	15-19	0.001079	0.001088	1.25857	1.24816
REG. VIII	20-24	0.000764	0.000776	1.22644	1.20747
REG. VIII	25-34	0.000181	0.000184	2.02210	1.98913
REG. VIII	35-44	0.000049	0.000050	2.57143	2.52000
REG. VIII	45-54	0.000048	0.000049	2.25000	2.20408
REG. VIII	55-OVER	0.000068	0.000069	1.92647	1.89855
REG. IX	15-19	0.000776	0.000789	1.79253	1.76299
REG. IX	20-24	0.000473	0.000480	2.02537	1.99583
REG. IX	25-34	0.000067	0.000069	1.47761	1.43478
REG. IX	35-44	0.000022	0.000023	2.90909	2.78261
REG. IX	45-54	0.000024	0.000025	2.20833	2.12000
REG. IX	55-OVER	0.000037	0.000038	2.72973	2.65789
REG. X	15-19	0.000954	0.000962	1.00210	0.99376
REG. X	20-24	0.000680	0.000684	1.25441	1.24708
REG. X	25-34	0.000129	0.000130	1.30233	1.29231
REG. X	35-44	0.000040	0.000041	1.52500	1.48780
REG. X	45-54	0.000049	0.000050	1.63265	1.60000
REG. X	55-OVER	0.000083	0.000084	1.92771	1.90476
REG. XI	15-19	0.000661	0.000662	1.09380	1.09215
REG. XI	20-24	0.000551	0.000553	1.15245	1.14828
REG. XI	25-34	0.000106	0.000106	1.16981	1.16981
REG. XI	35-44	0.000028	0.000028	1.17857	1.17857
REG. XI	45-54	0.000053	0.000054	1.71698	1.68519
REG. XI	55-OVER	0.000071	0.000072	1.53521	1.51389
REG. XII	15-19	0.000698	0.000706	1.71060	1.69122
REG. XII	20-24	0.000509	0.000519	2.43615	2.38921
REG. XII	25-34	0.000072	0.000074	1.22222	1.18919
REG. XII	35-44	0.000021	0.000021	1.66667	1.66667
REG. XII	45-54	0.000025	0.000026	4.56000	4.38462
REG. XII	55-OVER	0.000039	0.000040	4.15385	4.05000
A.R.M.M	15-19	0.000288	0.000295	17.17361	16.76610
A.R.M.M	20-24	0.000155	0.000158	14.60645	14.32911
A.R.M.M	25-34	0.000022	0.000023	6.18182	5.91304
A.R.M.M	35-44	0.000004	0.000004	8.75000	8.75000
A.R.M.M	45-54	0.000004	0.000004	28.50000	28.50000
A.R.M.M	55-OVER	0.000007	0.000008	23.14286	20.25000

Note: msePR=Prasad-Rao mean squared error; msePRg=generalized msePR;
 effPR=efficiency of msePR; effPRg=efficiency of msePRg

6.6 Conclusion

The importance of statistics on unemployment in small areas cannot be overemphasized. However, statistics for these small areas generated using the conventional way (i.e., direct estimates) are often unreliable. This is especially true in this case study where, as may be seen in Table 6.2, the CV estimates can get significantly large, notably for the smaller regions (e.g., 42.37 % for Reg. II, age group 35-44; 45.97 for Reg. XII, age group 35-44). A way out of this dilemma is to use indirect model-dependent estimates. The empirical Bayes approach has been used in this case study. Such use has proved to be advantageous. As shown in table 6.6, the EB estimates of the unemployment rates are more efficient than the direct estimates. Nevertheless, the distribution of the EB estimates has been shown to be not significantly different from that of the direct estimators (see figures 1 and 3).

An important finding from the analysis of the resulting EB estimates is that the unemployment profile from region to region does not differ significantly. In other words, in every region, unemployment is highest among the youth. However, the unemployment rates are highest in the bigger regions like NCR. These two observations may lead to the conclusion that the unemployment problem in the Philippines is at its worst among the youth in the big regions.

7 Bibliography

Arora, V. and Lahiri, P. (1995). On the Superiority of the Bayesian Method Over the BLUP in Small Area Estimation Problems. *SSC Annual Meeting, July 1995: Proceedings of the Survey Methods Section*, 39-45.

Barrios, E. B. (1998). Small Area Estimation of Selected Socio-Economic Indicators. *MIMAP (Micro Impacts of Macroeconomic Adjustment Policies Project) research Paper No. 36*, Metro Manila (Philippines).

Belmonte, E. (1998). L'estimation dans les petites regions: survol des methods de Bayes et presentation d'un estimateur conditionnel de l'EQM. *Mémoire présenté à la faculté des études supérieures, Département de mathématiques et de statistique, Université Laval, Québec*.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Carlin, B. P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Cox, B. et al, editors (1995). *Business Survey Methods*. John Wiley & Sons, New York.

Datta, G.S.; Lahiri, P. and Lu, K.L. (1999). Rates for the States of the U.S. *Journal of the American Statistical Association*, **94**, 1074-1082.

Datta, G.S. and Ghosh, M. (1991). Bayesian Prediction in Linear Models: Applications to Small Area Estimation. *The Annals of Statistics*, **19**, 1748-1770.

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

- Dick, P. and You, Y. (1997). Bayes and Census Undercoverage. *SSC Annual Meeting, June 1997: Proceedings of the Survey Methods Section*, 57-65.
- Drew, D., Singh, M. P., and Choudry, G. H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. *Survey Methodology*, **8** 17-47.
- Ericksen , E.P.; Kadane , J.B.; Tukey, J.W. (1989). Adjusting the 1980 Census of Population and Housing (in Applications and case Studies). *Journal of the American Statistical Association*, **84**, 927-944.
- Farrel, P. J. (2000). Bayesian Inference for Small Area Proportions. *Sankhya: The Indian Journal of Statistics*, 62, 402-416.
- Fay, R.E. III and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 269-277.
- Gelman, A. et al. (2000). *Bayesian Data Analysis*. Chapman and Hall, Boca Raton.
- Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, **9**, 55-93.
- Gonzales, M.E. and Hoza, C. (1978). Small-Area Estimation with Application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, **73**, 7-15.
- Holt, D.; Smith, T.M.F. and Tomberlin, T.J. (1979). A Model-Based Approach to Estimation for Small Subgroups of a Population. *Journal of the American Statistical Association*, **74**, 405-410.
- Hulting, F.L. and Harville, D.A. (1991). Some Bayesian and Non-Bayesian Procedures for the Analysis of Comparative Experiments and for Small-Area Estimation: Computational Aspects, Frequentist Properties and Relationships. *Journal of the American Statistical Association*, **86**, 557-568.
- International Labour Organisation (1988). *Current International Recommendations on Labour statistics*, 1988 Edition, Geneva, 49-50.

- Isaki, C.T. (1990). Small-Area Estimation of Economic Statistics, *Journal of Business and Economic Statistics*, **8**, 435-441.
- Laake, P. (1979). A Predictive Approach to Subdomain Estimation in Finite Populations. *Journal of the American Statistical Association*, **74**, 355-358.
- Lee, P.M. (1989). Bayesian Statistics: An Introduction. John Wiley & Sons, New York.
- Marker, D.A. (2001). Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators. *Survey Methodology*, **27**, 183-188.
- Maritz, J.S. and Lwin, T. (1989). Empirical Bayes Methods. Chapman and Hall, London.
- Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47-65.
- Moura, F.A.S. and Holt, D. (1999). Small Area Estimation Using Multilevel Models. *Survey Methodology*, **25**, 73-80.
- National Statistics Office of the Philippines (2003). Technical Notes on the Labour Force Survey, NSO, Manila (Philippines).
- National Center for Health Statistics (1977). Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. DHEW Publication 78-1349, Maryland.
- National Center for Health Statistics (1979). Small Area Estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States. DHEW Publication 80-1356, Maryland.
- National Center for Health Statistics (1999). National Health Interview Survey: Research for the 1995-2004 Redesign. Vital Health Stat 2(126), Washington DC.
- Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, **85**, 163-171.

- Prasad, N.G.N. and Rao, J.N.K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, **25**, 67-72.
- Platek, R.; Rao, J.N.K.; Sarndal, C.E and Singh, M., editors (1987). Small Area Statistics: An International Symposium. John Wiley & Sons, New York.
- Purcell, N. J. and Kish, L. (1979). Estimation for Small Domain. *Biometrics* **35**, 365-384.
- Rao, J.N.K. (1998). Small Area Estimation. *Encyclopedia of Statistical Sciences*, **2**, 621-628.
- Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology*. **25**, 175-186.
- Rao, J.N.K. (2001). On Measuring the Quality of Indirect Small Area Estimates. *Proceedings of Statistics Canada Symposium 2001*.
- Rao, J.N.K. (2003). Small Area Estimation. John Wiley & Sons, New Jersey.
- Rao, J.N.K. and Choudry, G. H. (1995). Small Area Estimation: Overview and Empirical Study, in B.G. Cox et al (editors), *Business Survey Methods*, John Wiley & Sons, New York, 527-542.
- Rivest, L.-P. and Belmonte, E. (2000). A Conditional Mean Squared Error of Small Area Estimators, *Survey Methodology*, **26**, 67-78.
- Rivest, L.-P. and Vandal, N. (2003). Mean Squared Error Estimation for Small Areas When the Small Area Variances are Estimated. Proceedings of the International Conference on Recent Advances in Survey Sampling, J.N.K. Rao, ed., to appear.
- Sarndal, C.-E. (1984). Design-Consistent versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association*, **79**, 624-631.
- Sarndal, C.-E. and Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association*, **84**, 266-275.

Schaible, W.L. (1978). Choosing Weights for Composite Estimators for Small Area Statistics, *Proceedings of the Survey Research Methods Section*, 741-746.

Singh, M.P.; Gambino, J. and Mantel, H.J. (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, **20**, 3-22.

Sostra, K. (2001). Small Area Estimation Methods, Statistical Office of Estonia.

Suciu, G. et al (2001). Uninsured Estimates by County: A Review of Options and Issues, Ohio Department of Health.

You, Y. and Rao, J.N.K. (2000). Hierarchical Bayes Estimation of Small Area Means Using Multi-Level Models. *Survey Methodology*, **26**, 173-181.

You, Y. and Rao, J.N.K. (2002). Small Area Estimation Using Unmatched Sampling and linking Models. *The Canadian Journal of Statistics*, **30**, 3-15.

8 Appendices

8.1 The Splus program for the Shapiro-Wilks Test

```
# *****
#      shapiro.wilk.test for normality of a given set of data
# *****
# obtained from http://www.biostat.wustl.edu/archives/html/s-news/1999-
08/msg00153.html
SW_ function(x)
{
#      "shapiro.wilk.test"
#
# This function is an S version of the procedure described by
# J. P. Royston (1982) in "An Extension of Shapiro and Wilk's
# W Test for Normality to Large Samples", Applied Statistics,
# Vol. 31, No. 2, pp 115-124.
#
  n <- length(x)
  index <- 1:n
  m <- qnorm((index - 0.375)/(n + 0.25))
  y <- sort(x)
  mu <- mean(y)
  SSq <- sum((y - mu)^2)
  astar <- 2 * m
  ends <- c(1, n)
  astar.p <- astar[ - ends]
  if(n <= 20)
    m <- n - 1
  else m <- n
  if(m < 20)
    aa <- gamma(0.5 * (m + 1))/(sqrt(2) *
      gamma(0.5 * m + 1))
  else {
    f1 <- (6 * m + 7)/(6 * m + 13)
    f2 <- exp(1)/(m + 2)
    f3 <- (m + 1)/(m + 2)
    f3 <- f3^(m - 2)
    aa <- f1 * sqrt(f2 * f3)
  }
  astar.1 <- (aa * sum(astar.p^2))/(1 - 2 * aa)
  astar.1 <- sqrt(astar.1)
  astar[1] <- - astar.1
  astar[n] <- astar.1
  A <- astar/sqrt(sum(astar^2))
  W <- (sum(A * y)^2)/SSq
  if(n <= 20) {
    u <- log(n) - 3
    lambda <- 0.118898 + 0.133414 * u +
      0.327907 * u^2
  }
}
```

```

logmu <- -0.37542 - 0.492145 * u -
          1.124332 * u^2 - 0.199422 * u^3
logsigma <- -3.155805 + 0.729399 * u +
            3.01855 * u^2 + 1.5558776 * u^3
}
if(n > 20) {
  u <- log(n) - 5
  lambda <- 0.480385 + 0.318828 * u +
            0.0241665 * u^3 + 0.00879701 *
            u^4 + 0.002989646 * u^5
  logmu <- -1.91487 - 1.37888 * u -
            0.04189209 * u^2 + 0.1066339 *
            u^3 - 0.03513666 * u^4 -
            0.01504614 * u^5
  logsigma <- -3.83538 - 1.015807 * u - 0.331885 *
              u^2 + 0.1773538 * u^3 - 0.01638782 * u^4 -
              0.03215018 * u^5 + 0.003852646 * u^6
}
mu <- exp(logmu)
sigma <- exp(logsigma)
y <- (1 - W)^lambda
z <- (y - mu)/sigma
p <- 1 - pnorm(z)
if(n < 7) {
  warning("n is too small for this program
to correctly estimate p"
)
  p <- NA
}
if(n > 2000) {
  warning("n is too large for this program
to correctly estimate p"
)
  p <- NA
}
out <- list(W = W, n = n, p = p)
out
}

```

8.2 The Splus program for scenario 2

```

*****
# **APPLICATION OF THE EMPIRICAL BAYES METHOD TO SMALL AREA ESTIMATION**
#   *** OF UNEMPLOYMENT RATES IN THE PHILIPPINES ***
#*****
#
#   *** 1. AVAILABLE DATA   ***
#
essai.data_read.table("e:/essai/write_up/Splus_program/data.txt",
  header=T,sep=";",row.names=NULL)
#print(essai.data)
Area_essai.data[,1]
AgeGroup_essai.data[,2]
Employed_essai.data[,3]
Unemployed_essai.data[,4]
SE_essai.data[,5]
CV_essai.data[,6]
# rsigmaii=standard error of the unemployment rate
# sigmaii=standard error of log(unemployment rate)
rsigmaii_SE/(Employed+Unemployed)
sigmaii_(CV/100)
#
# imputations for ARMM, Age Groups 35-44, 45-54 & 55-Over
#   note: the CV and sigmaii values of the corresponding age groups
#         for Region XII are adopted as values for ARMM
essai.data[88,6]_essai.data[82,6]
essai.data[89,6]_essai.data[83,6]
essai.data[90,6]_essai.data[84,6]
sigmaii[88]_sigmaii[82]
sigmaii[89]_sigmaii[83]
sigmaii[90]_sigmaii[84]
sigmaii2_sigmaii^2
sigmaii2_round(sigmaii2,digits=6)
#   note: sigmaii2 = psi for small area i (see section 5.3)
#         psi assumed known
rsigmaii[88]_rsigmaii[82]
rsigmaii[89]_rsigmaii[83]
rsigmaii[90]_rsigmaii[84]
rsigmaii2_rsigmaii^2
rsigmaii2_round(rsigmaii2,digits=6)
Input.data_cbind(essai.data,sigmaii,sigmaii2)
Input.data
m_15*6
p_1+14+5
#   note: number of regions=15, number of age groups=6
#   model: log[y(i,,j)] = mu + z(i) + b(j) + e(i,,j)
#
#-----
#
#
#   *** 2. GENERATION OF THE X-MATRIX   ***
#
X_matrix(0,90,20)
X[,1]_rep(1,90)
X[,2]_c(rep(1,6),rep(0,84))
X[,3]_c(rep(0,6),rep(1,6),rep(0,78))
X[,4]_c(rep(0,12),rep(1,6),rep(0,72))
X[,5]_c(rep(0,18),rep(1,6),rep(0,66))

```

```

X[,6]_c(rep(0,24),rep(1,6),rep(0,60))
X[,7]_c(rep(0,30),rep(1,6),rep(0,54))
X[,8]_c(rep(0,36),rep(1,6),rep(0,48))
X[,9]_c(rep(0,42),rep(1,6),rep(0,42))
X[,10]_c(rep(0,48),rep(1,6),rep(0,36))
X[,11]_c(rep(0,54),rep(1,6),rep(0,30))
X[,12]_c(rep(0,60),rep(1,6),rep(0,24))
X[,13]_c(rep(0,66),rep(1,6),rep(0,18))
X[,14]_c(rep(0,72),rep(1,6),rep(0,12))
X[,15]_c(rep(0,78),rep(1,6),rep(0,6))
X[,16]_rep(c(1,rep(0,5)),15)
X[,17]_rep(c(0,1,rep(0,4)),15)
X[,18]_rep(c(0,0,1,rep(0,3)),15)
X[,19]_rep(c(0,0,0,1,rep(0,2)),15)
X[,20]_rep(c(0,0,0,0,1,0),15)
#X
#-----
X1_matrix(0,90,19)
X1[,1]_c(rep(1,6),rep(0,84))
X1[,2]_c(rep(0,6),rep(1,6),rep(0,78))
X1[,3]_c(rep(0,12),rep(1,6),rep(0,72))
X1[,4]_c(rep(0,18),rep(1,6),rep(0,66))
X1[,5]_c(rep(0,24),rep(1,6),rep(0,60))
X1[,6]_c(rep(0,30),rep(1,6),rep(0,54))
X1[,7]_c(rep(0,36),rep(1,6),rep(0,48))
X1[,8]_c(rep(0,42),rep(1,6),rep(0,42))
X1[,9]_c(rep(0,48),rep(1,6),rep(0,36))
X1[,10]_c(rep(0,54),rep(1,6),rep(0,30))
X1[,11]_c(rep(0,60),rep(1,6),rep(0,24))
X1[,12]_c(rep(0,66),rep(1,6),rep(0,18))
X1[,13]_c(rep(0,72),rep(1,6),rep(0,12))
X1[,14]_c(rep(0,78),rep(1,6),rep(0,6))
X1[,15]_rep(c(1,rep(0,5)),15)
X1[,16]_rep(c(0,1,rep(0,4)),15)
X1[,17]_rep(c(0,0,1,rep(0,3)),15)
X1[,18]_rep(c(0,0,0,1,rep(0,2)),15)
X1[,19]_rep(c(0,0,0,0,1,0),15)
#
# -----
#
#
# *** 3. COMPUTE r=DIRECT ESTIMATORS ***
#
r_Unemployed/(Employed+Unemployed)
theta_log(r)
theta_matrix(theta,nrow=90)
theta[89]_theta[88]
# note: theta[89]computed as 0; assigned value of next small area
theta_round(theta,digits=6)
r_exp(theta)
r_round(r,digits=6)
#
# -----
#
#
# *** 4. COMPUTE EMPIRICAL BAYES ESTIMATORS rEB ***
#
reg.data_cbind(X,theta)
# a. compute hii
Xt_t(X)
XXt_Xt%%X
XXinv_solve(XXt)

```

```

hii_X%*%XXinv%*%Xt
# b. compute betastar
reg_lsfit(X1,theta[,1],intercept=T)
ls.print(reg,digits=6)
betastar_reg$coef
#betastar
Xbetastar_X%*%betastar
Xbetastar
# c. compute sigma2v
tli_(theta-Xbetastar)^2
t1_sum(tli)
t1
t2i_sigmai2*(1-diag(hii))
t2_sum(t2i)
t2
sigma2v_max((t1-t2)/(m-p),0)
sigma2v
# d. compute betahat
weight_1/(sigmai2+sigma2v)
reghat_lsfit(X1,theta[,1],wt=weight,intercept=T)
ls.print(reghat,digits=6)
betahat_reghat$coef
betahat
# e. compute thetaEB, rEB
Xbetahat_X%*%betahat
thetaEB_theta-( (sigmai2/(sigmai2+sigma2v) ) * (theta-Xbetahat) )
# note: see formula 5.3.3
rEB_exp(thetaEB)
rEB_round(rEB,digits=6)
#
#-----
#
#
# *** 5. ANALYSIS OF RESIDUALS ***
#
# a. graphique analysis
residuals_reghat$residuals
# note: Xbetahat = predicted values for theta
graphdata_cbind(residuals,Xbetahat)
x_Xbetahat
y_residuals/(sqrt(sigmai2+sigma2v))
Y
plot(x,y,pch="X",main="Distribution of normalized residuals",
sub="(Scenario 2: theta=log(r))",xlab="predicted values",ylab="residuals")
abline(h=0)
# b. test for goodness of fit of model : using the Shapiro-Wilks test
# note: SW is the Shapiro-Wilks test
SW(y)
#
#-----
#
# *** 6. DISTRIBUTION OF r, rEB ***
#
hist(r,main="Distribution of r=direct estimators",
xlab="direct estimators of unemployment rate")
hist(rEB,main="Distribution of rEB",sub="Scenario 2: theta=log(r)",
xlab="empirical Bayes estimators")

```

```

#-----
#
#      *** 7. RANK AGE GROUPS ACCORDING TO rEB, BY REGION ***
#
regions_c("NCR", "C.A.R.", "REG.I", "REG.II", "REG.III", "REG.IV", "REG.V", "REG.VI",
"REG.VII", "REG.VIII", "REG.IX", "REG.X", "REG.XI", "REG.XII", "A.R.M.M.")
agegroup_c("15-19", "20-24", "25-34", "35-44", "45-54", "55-OVER")
reb_matrix(rEB, ncol=6, byrow=T)
for(i in 1:15){
print(regions[i])
x_c(reb[i,1:6])
xrank_rank(x)
if(i==1) yrank_xrank
else
yrank_cbind(yrank, xrank)
}
dimnames(yrank)_list(agegroup, regions)
yrank
# note: yrank = matrix containing the ranks
#
#-----
#
#
#      *** 8. COMPUTE MSE ***
#
# a. compute mselPR=mse(thetaEB) of Prasad-Rao
# note: weight = 1/(sigma11+sigma2v)
g1_sigma2v*sigma11*weight
Xw_diag(weight)%*%X
A_t(Xw)%*%X
Ainv_solve(A)
g2_((sigma11*weight)^2) * diag(X%*%Ainv%*%t(X))
g3_(2/(m^2)) * (sigma11^2 * weight^3) * sum((1/weight)^2)
mselPR_g1 + g2 + 2*g3
# b. compute msePR=mse(rEB)
msePR_(exp(2*thetaEB))*mselPR
msePR_round(msePR, digits=6)
#
#-----
#
#
#      *** 9. COMPUTE GENERALIZED MSE PRASAD-RAO ***
# Note: the term g4=[2*gamma1*sigma4v/(sigma11+sigma2v)^3]
# is added to mse of Prasad-Rao
gamma1_(sigma11*CV/100)^2
g4_2*gamma1*sigma2v^2/((sigma11+sigma2v)^3)
mselPRg_mselPR + g4
msePRg_(exp(2*thetaEB))*mselPRg
msePRg_round(msePRg, digits=6)
#
#-----
#
#
#      *** 10. COMPUTE EFFICIENCY OF rEB ***
#
effPR_rsigma11/msePR
effPRg_rsigma11/msePRg

```

```

# -----
#
#
#     *** 11. THE RESULTS ***
#
essai.frame_data.frame(essai.data)
Area_essai.frame[,1]
AgeGroup_essai.frame[,2]
Employed_essai.frame[,3]
Unemployed_essai.frame[,4]
SE_essai.frame[,5]
CV_essai.frame[,6]
Area_data.frame(Area)
Employed_data.frame(Employed)
Unemployed_data.frame(Unemployed)
SE_data.frame(SE)
CV_data.frame(CV)
sigmai2_data.frame(sigmai2)
cbind(Area, AgeGroup, Employed, Unemployed, SE, CV)
sigma2v
cbind(Area, AgeGroup, r, rEB, rsigmai2)
cbind(Area, AgeGroup, msePR, msePRg, effPr, effPrg)
#
# -----

```