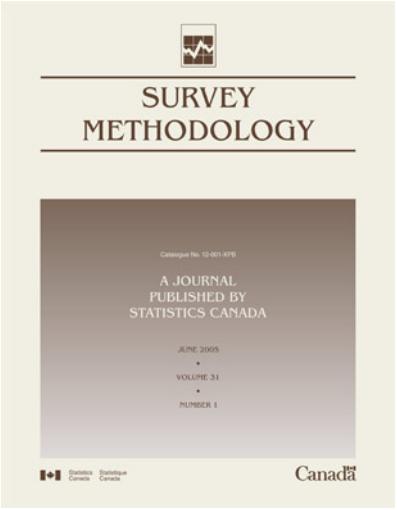
Catalogue no. 12-001-XIE

# Survey Methodology

2005




Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

| | |
|---|---|
| National inquiries line | 1 800 263-1136 |
| National telecommunications device for the hearing impaired | 1 800 363-7629 |
| Depository Services Program inquiries | 1 800 700-1033 |
| Fax line for Depository Services Program | 1 800 889-9734 |
| E-mail inquiries | infostats@statcan.ca |
| Website | www.statcan.ca |

## Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

# Survey Methodology

2005

Published by authority of the Minister responsible for Statistics Canada



November 2005

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 1, pp. 45-52
Statistics Canada, Catalogue No. 12-001

# Adjustment of Unemployment Estimates Based on Small Area Estimation in Korea

**Yeon Soo Chung, Kay-O Lee and Byung Chun Kim** [1]

## Abstract

The Korean Economically Active Population Survey (EAPS) has been conducted in order to produce unemployment statistics for Metropolitan Cities and Provincial levels, which are large areas. Large areas have been designated as planned domains, and local self-government areas (LSGA's) as unplanned domains in the EAPS. In this study, we suggest small area estimation methods to adjust for the unemployment statistics of LSGA's within large areas estimated directly from current EAPS data. We suggest synthetic and composite estimators under the Korean EAPS system, and for the model-based estimator we put forward the Hierarchical Bayes (HB) estimator from the general multi-level model. The HB estimator we use here has been introduced by You and Rao (2000). The mean square errors of the synthetic and composite estimates are derived by the Jackknife method from the EAPS data, and are used as a measure of accuracy for the small area estimates. Gibbs sampling is used to obtain the HB estimates and their posterior variances, and we use these posterior variances as a measure of precision for small area estimates. The total unemployment figures of the 10 LSGA's within the ChoongBuk Province produced by the December 2000 EAPS data have been estimated using the small area estimation methods suggested in this study. The reliability of small area estimates is evaluated by the relative standard errors or the relative root mean square errors of these estimates. We suggest here that under the current Korean EAPS system, the composite estimates are more reliable than other small area estimates.

Key Words: Synthetic estimator; Composite estimator; Hierarchical Bayes; Multi-level model; Jackknife mean square error; Gibbs sampling.

## 1. Introduction

Sample surveys are a more cost-effective way of obtaining information than complete enumerations or censuses for most purposes. The surveys are usually designed to ensure that reliable estimates of totals and means for the population, pre-specified domains of interest, or major subpopulations can be derived from the survey data. There are also many situations in which it is desirable to derive reliable estimates for additional domains of interest, especially geographical areas or subpopulations, from existing survey data.

The Korean National Statistical Office conducts the Economically Active Population Survey (EAPS) in 30,000 sample households every month. The characteristics of the economically active for 16 large areas (7 Metropolitan Cities, 9 Provinces) of the country are based on these monthly EAPS results. The EAPS is a large city or provincial level survey. Many small cities in a large area would prefer to obtain the unemployment figures for individual cities without conducting their own survey, and the most cost-effective way would be to turn to the EAPS data. However, small cities belonging to a large area are unplanned regions in the EAPS and sample sizes for these small cities are typically too small due to the size of small cities. Therefore, if we estimate the unemployment statistics of small areas from the EAPS framework based on large areas, we may be unable to obtain an estimate with adequate precision since the sample size in specific small areas may not be large enough. The direct estimates for specific small areas from the EAPS cannot be sufficiently reliable in this situation. It is hence necessary to "borrow strength" from related areas to obtain more reliable estimates for a given small area. An example of such would be to gather separately published administrative records of related small areas. We define related areas as those areas with similar economic and demographic characteristics as the small area we wish to estimate. Our aim is to adjust the direct estimates derived from the National Statistical Office of Korea through design-based and model-based indirect estimators, and hence secure reliable estimates.

This paper focuses on discussion of the Hierarchical Bayes (HB) estimator using multi-level models, and the composite estimator that takes the weighted average of the direct estimator drawn from the Korean National Statistical Office and the synthetic estimator designed under the Korean EAPS system. The general multi-level model framework for small area estimation has been suggested in Moura and Holt (1999), and the HB estimation method using this multi-level model has been applied in more detail in You and Rao (2000). We use here the HB estimation method as in You and Rao (2000). Detailed accounts of synthetic and composite estimation are given by Ghosh and Rao (1994), Singh, Gambino and Mantel (1994) and Marker (1999).

1. Yeon Soo Chung, Department of Computer Science and Statistics, Korea Air Force Academy, Chungwon, Korea. E-mail: yschung@afa.ac.kr; Kay-O Lee, Gallup Koyed and Chungbuk National University, Seoul, Korea. E-mail: kolee@gallup.co.kr; Byung Chun Kim, Graduate School of Management, KAIST, Seoul, Korea. E-mail: bckim@kaist.ac.kr.

Other references can be found in P.D. Falorsi, S. Falorsi and Russo (1994), and Chattopadhyay, Lahiri, Larsen and Reimnitz (1999). Falorsi *et al.* (1994) produce level estimates for unplanned small area territorial domains from the Italian Labor Force Sample Survey whereas Chattopadhyay *et al.* (1999) give a composite estimation of drug prevalence for sub-state areas to improve on the traditional design-based estimators. It is noted that both studies use supplementary information from the original survey data. For example, Chattopadhyay *et al.* (1999) uses additional information that relates various groups, counties and planning regions to one another.

In order to "borrow strength", we divide the EAPS data into two homogenous sub-regional groups (Cities and Counties), and each sub-regional group is classified into four categories of sex (male, female) and age (15–34, 35 and over). The unemployment characteristics of each category in the given small area are used as supplementary information for small area estimation. We also use the census of 2000 and the Resident Registration Population of 2000 as auxiliary information to calculate the small area estimates.

The contents of this paper are as follows. The Korean EAPS is described briefly in section 2. Section 3 gives the direct estimator drawn from the Korean National Statistical Office. Section 4 introduces design-based and model-based indirect estimators. We suggest synthetic and composite estimators under the current EAPS system, and the mean square errors of these estimates are derived using the Jackknife method. For the model-based indirect estimator we apply the HB multi-level model in estimating small areas. Section 5 illustrates the methodology, studies model selection and presents results employing the EAPS data. Finally, some closing comments are made in section 6.

## 2. Economically Active Population Survey

The Korean National Statistical Office conducts the Economically Active Population Survey (EAPS) on a monthly basis. The characteristics of the economically active (such as employment and unemployment figures) are obtained from the EAPS. The EAPS provides monthly information on the employment trend, which plays an important role in policy making and evaluation for the 7 Metropolitan Cities and 9 Provinces. The interviewees of the EAPS are persons aged 15 and over residing in sample enumeration districts. The survey is conducted during the week just after the reference period, which is the week containing the 15$^{th}$ day of the month. The EAPS is conducted by visiting and interviewing each household.

The sample households for the Korean EAPS are selected from the sampled population using stratified two-stage sampling. The sampled population consists of 22,000 enumeration districts that are ten percent of the 1995 census. According to the classification of major administration regions, the country is divided into 16 large areas; there are 7 Metropolitan Cities and 9 Provinces, and the population is divided into 25 strata; 7 metropolitan strata, and 18 provincial strata consisting of 9 urban strata and 9 rural strata. The number of enumeration districts, which are primary sampling units (PSUs), selected in the 25 strata is computed using a preassigned relative standard error. Then PSUs are systematically selected with a probability proportional to their measure of size within each stratum. Each sampled PSU is divided into the same number of segments as the measure of size of each PSU, each segment containing 8 households on average. Within each PSU, 3 contiguous segments, secondary sampling units (SSUs), are randomly selected, and all households in each selected segment are surveyed. The sample is self-weighting in each stratum while the sampling rates are different from stratum to stratum. The selected sample households are surveyed repeatedly for 5 years without rotating.

The planned domains of the survey design are the 16 large areas (7 Metropolitan Cities and 9 Provinces), and local self-government areas (LSGAs) within those large areas are unplanned sub-regional domains. The sample size for the current EAPS is approximately 1,200 PSUs, and 30,000 households. The purpose of this study is to estimate unemployment statistics of the LSGAs from the EAPS.

## 3. Direct Estimation

The direct estimator $\hat{Y}_{i\cdot}$ representing the total unemployment figure for small area $i$, based on data from the EAPS, is as follows:

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2} {}_s\hat{Y}_{i\cdot} = \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih} = \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \qquad (3.1)$$

for $i = 1, 2, ..., I$, $s = 1, 2$ and $h = 1, 2, ..., n_i$, where $s$ is an index of sex (male or female), $n_i$ denotes the number of sample enumeration districts for small area $i$ from the EAPS, and ${}_sY_{ih}$ is the number of unemployed persons by sex for the $h^{th}$ sample enumeration district within small area $i$ from the EAPS. The multiplier ${}_sM_i = {}_s\hat{X}_{i\cdot}/{}_sX_{i\cdot}$ is calculated under the condition that $\hat{Y}_{i\cdot}$ is an approximately unbiased estimator, where ${}_s\hat{X}_{i\cdot}$ is the estimate of the resident population in small area $i$, and ${}_sX_{i\cdot}$ is the sample survey resident population derived from the EAPS. The variance of $\hat{Y}_{i\cdot}$ in the $i^{th}$ small area is estimated using a linearization – based variance estimator.

## 4. Indirect Small Area Estimation

### 4.1 Synthetic Estimation

For the $i^{th}$ small area belonging to a large area, the direct estimator $\hat{Y}_{i\cdot}$ does not provide adequate precision because

sample sizes in specific small areas are not large enough. The synthetic estimator $\hat{Y}_{i\cdot}^{S}$ is a design-based indirect estimator that borrows strength from related areas through implicit modeling of supplementary data along with the survey data. Suppose that there are $I$ small areas in a large area. We then divide each large area into several homogenous sub-regional groups, in which $I = \sum_{l=1}^{L} I_l$. Each sub-regional group including $I_l$ small areas is classified into $J$ sex-age categories. It is assumed that each small area belongs to one of several sub-regional groups and we obtain auxiliary information from the sub-regional group. The synthetic estimator has a low variance since it is based on a larger sample, but it suffers from bias should the assumption of homogenous sub-regional groups not hold.

The following notations are used: $N_i$, for the number of enumeration districts in small area $i$; $n_i$, for the number of sample enumeration districts allocated to the $i^{\text{th}}$ small area; ${}_jP_{i,2000}^{C}$, for resident population derived from the census of 2000 in cell $(i, j)$; ${}_jP_{i,2000}^{R}$, for Resident Registration Population of 2000 in cell $(i, j)$; ${}_jP_{i,\text{month}}^{R}$, for Resident Registration Population at survey month in cell $(i, j)$; ${}_j\hat{X}_i$, for the direct estimate of resident population in cell $(i, j)$; ${}_jY_{ih}$, for the number of the unemployed in the $h^{\text{th}}$ sample enumeration district in cell $(i, j)$.

We consider the estimation of the total unemployed $Y_{i\cdot}$ for all units belonging to small area $i$. A synthetic estimator for small area $i$ within the sub-regional group including $I_1$ small areas is given by

$$\hat{Y}_{i\cdot}^{S} = \sum_{j=1}^{J} \frac{{}_j\hat{P}_{i\cdot}}{{}_j\hat{X}_{\cdot}} = {}_S\hat{Y}_{\text{dir}},\ i = 1, 2, ..., I_1, \qquad (4.1)$$

where

$${}_j\hat{P}_i = \frac{{}_jP_{i,2000}^{C}\ {}_jP_{i,\text{month}}^{R}}{{}_jP_{i,2000}^{R}},$$

$${}_j\hat{X}_{\cdot} = \sum_{i=1}^{I_1} {}_j\hat{X}_i,$$

$${}_j\hat{Y}_{\text{dir}} = \sum_{i=1}^{I_1}\sum_{h=1}^{n_i} {}_jM_i\ {}_jY_{ih},$$

in which ${}_j\hat{P}_i$ denotes the estimate of resident population obtained from administrative sources for the $j^{\text{th}}$ sex-age category (cell) in small area $i$, ${}_j\hat{X}_{\cdot}$ denotes the estimate of resident population of the $j^{\text{th}}$ sex-age category, ${}_j\hat{Y}_{\text{dir}}$ denotes the direct estimate of the total unemployed of the $j^{\text{th}}$ sex-age category in the EAPS, and the multiplier ${}_jM_i$ is expressed by ${}_jM_i = {}_j\hat{X}_i / {}_jX_{i\cdot}$. Note that ${}_j\hat{Y}_{\text{dir}}$ represent approximately unbiased estimates of ${}_jY_{\cdot\cdot} = \sum_{i=1}^{I_1}\sum_{h=1}^{N_i} {}_jY_{ih}$.

As a measure of accuracy for the synthetic estimator $\hat{Y}_{i\cdot}^{S}$, it is customary to take

$$\text{MSE}(\hat{Y}_{i\cdot}^{S}) = \text{Var}(\hat{Y}_{i\cdot}^{S}) + [\text{Bias}(\hat{Y}_{i\cdot}^{S})]^2. \qquad (4.2)$$

In (4.2), the variance of $\hat{Y}_{i\cdot}^{S}$ is readily estimated, but it is more difficult to estimate the bias of $\hat{Y}_{i\cdot}^{S}$. Under the assumption $\text{Cov}(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S}) = 0$, where $\hat{Y}_{i\cdot}$ is a direct estimator of $Y_{i\cdot}$, the estimator of MSE of $\hat{Y}_{i\cdot}^{S}$ is given by

$$\text{mse}(\hat{Y}_{i\cdot}^{S}) \approx (\hat{Y}_{i\cdot}^{S} - \hat{Y}_{i\cdot})^2 - \hat{\text{V}}\text{ar}(\hat{Y}_{i\cdot}). \qquad (4.3)$$

Note that $\text{mse}(\hat{Y}_{i\cdot}^{S})$ in (4.3) is approximately an unbiased estimator, but is potentially unstable should the number of sample enumeration districts not be large enough. Another measure would be to take the average of these MSE estimators over small areas. This average MSE estimator is expected to be stable, but it is not an area-specific measure of accuracy (Ghosh and Rao 1994).

The Jackknife method is an alternative method that can provide a more accurate area-specific measure. For small area $i$, the estimator for the mean square error of the estimate of the total unemployed is given as follows:

$$\text{mse}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = \hat{\text{V}}\text{ar}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) + [\hat{\text{B}}\text{ias}_{\text{JN}}(\hat{Y}_{i\cdot}^{S})]^2, \qquad (4.4)$$

where

$$\hat{\text{V}}\text{ar}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = \frac{n_i - 1}{n_i}\sum_{h=1}^{n_i}\left[\hat{Y}_{i\cdot}^{S}(h) - \frac{1}{n_i}\sum_{l=1}^{n_i}\hat{Y}_{i\cdot}^{S}(l)\right]^2,$$

$$\hat{\text{B}}\text{ias}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = (n_i - 1)\left[\frac{1}{n_i}\sum_{h=1}^{n_i}\hat{Y}_{i\cdot}^{S}(h) - \hat{Y}_{i\cdot}^{S}\right].$$

Here, $\hat{Y}_{i\cdot}^{S}(h)$ denotes the estimate of $Y_{i\cdot}$ obtained when district $h$ is removed from the sample.

### 4.2 Composite Estimation

For small area $i$, the direct estimator $\hat{Y}_{i\cdot}$ derived from the EAPS does not provide adequate precision because sample sizes in specific small areas are seldom large enough. Also, the synthetic estimator $\hat{Y}_{i\cdot}^{S}$ that borrows strength from related small areas may be biased. A natural way to balance the synthetic estimator $\hat{Y}_{i\cdot}^{S}$ against the instability of the direct estimator $\hat{Y}_{i\cdot}$ is to take a weighted average of the two estimators. The following composite estimator $\hat{Y}_{i\cdot}^{C}$ can be considered to gain adequate precision for small area estimates:

$$\hat{Y}_{i\cdot}^{C} = \omega_i\,\hat{Y}_{i\cdot} + (1 - \omega_i)\hat{Y}_{i\cdot}^{S},\ i = 1, 2, ..., I_1, \qquad (4.5)$$

where $\omega_i$ is the weight having a value between 0 and 1.

Under the assumption of $\text{Cov}(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S}) = 0$, the optimal weight $\omega_{i(\text{opt})}$ that minimizes the $\text{MSE}(\hat{Y}_{i\cdot}^{C})$ with respect to $\omega_i$ can be approximated by

$$\omega_{i(\text{opt})} = \frac{\text{MSE}(\hat{Y}_{i\cdot}^{S})}{\text{MSE}(\hat{Y}_{i\cdot}^{S}) + \text{Var}(\hat{Y}_{i\cdot})}. \qquad (4.6)$$

The optimal weight $\omega_{i(\text{opt})}$ in (4.6) may be estimated by substituting the Jackknife estimator $\text{mse}_{\text{JN}}(\hat{Y}_{i\cdot}^{S})$ given

in (4.4) for $\mathrm{MSE}(\hat{Y}_{i\cdot}^{S})$, and replacing $\mathrm{Var}(\hat{Y}_{i\cdot})$ by $\hat{\mathrm{V}}\mathrm{ar}(\hat{Y}_{i\cdot})$, the linearization-based estimator typically used by the National Statistical Office of Korea. The estimated weight $\hat{\omega}_{i(\mathrm{opt})}$ is then given by

$$\hat{\omega}_{i(\mathrm{opt})} = \frac{\mathrm{mse}_{\mathrm{JN}}(\hat{Y}_{i\cdot}^{S})}{\mathrm{mse}_{\mathrm{JN}}(\hat{Y}_{i\cdot}^{S}) + \hat{\mathrm{V}}\mathrm{ar}(\hat{Y}_{i\cdot})}. \qquad (4.7)$$

Using the estimated weight given in (4.7), we can obtain the composite estimator of the total unemployed as follows:

$$\hat{Y}_{i\cdot}^{C} = \hat{\omega}_{i(\mathrm{opt})}\hat{Y}_{i\cdot} + (1-\hat{\omega}_{i(\mathrm{opt})})\hat{Y}_{i\cdot}^{S},\ i=1,\ 2,\ ...,\ I_1. \quad (4.8)$$

The Jackknife method was used to obtain area-specific measures of accuracy.

### 4.3 Hierarchical Bayes Estimation Using Multi-level Models

Suppose that there are $I$ small areas. We consider the following multi-level model that integrates variations within and between the small areas in a single model:

$$Y_{ik} = x_{ik}^{T}\beta_i + e_{ik},\ \beta_i = Z_i\gamma + \nu_i,\ i=1,\ 2,\ ...,\ I;\ k=1,\ 2,\ ...,\ K, \quad (4.9)$$

where $y_{ik}$ are the direct estimates associated with the $k^{\mathrm{th}}$ month in the $i^{\mathrm{th}}$ small area, which may be adjusted through the model (4.9) with the auxiliary variables $x_{ik} = (x_{i1k}, x_{i2k}, ..., x_{ipk})^{T}$ selected from the EAPS, census and administrative records; $\beta_i$ is a $p\times 1$ vector of regression coefficients; $Z_i$ is a $p\times q$ design matrix; $\gamma$ is a $q\times 1$ vector of fixed coefficients; and $\nu_i = (\nu_{i1}, \nu_{i2}, ..., \nu_{ip})^{T}$ is a $p\times 1$ vector of random effects for the $i^{\mathrm{th}}$ small area.

The $\nu_i$'s are assumed to have a joint distribution $\nu_i \overset{\mathrm{ind}}{\sim} N_p(0, \Phi)$ with an unknown variance covariance matrix $\Phi$ and the $e_{ik}$'s are assumed to be independent random error variables with $E(e_{ik}) = 0$ and $\mathrm{Var}(e_{ik}) = \sigma_i^2$. $\nu_i$ and $e_{ik}$ are also assumed to be independent.

To obtain HB estimates for each small area and posterior variances of estimates obtained from (4.9), we apply You and Rao's (2000), HB multi-level model framework as follows:

**Model 1**: HB model with equal error variances.

(i) $[y_{ik}|\beta_i, \sigma_e^2] \overset{\mathrm{ind}}{\sim} N(x_{ik}^{T}\beta_i, \sigma_e^2)$, $i=1,\ 2,\ ...,\ I;\ k=1,\ 2,\ ...,\ K,$ (4.10)

(ii) $[\beta_i|\gamma,\ \Phi] \overset{\mathrm{ind}}{\sim} N_p(Z_i\gamma,\ \Phi),$ (4.11)

(iii) Marginal prior distributions are as follows: $\gamma \sim N_q(0,\ D)$, $\tau_e \sim G(a,\ b)$, and $\Omega \sim W_p(\alpha,\ R)$, where $\tau_e = \sigma_e^{-2}$, $\Omega = \Phi^{-1}$ and $D$, $a$, $b$, $\alpha$ and $R$ are known and $G(a, b)$ denotes a gamma distribution with its density given by $f(x) = [b^a/\Gamma(a)]x^{a-1}e^{-bx}$ $(a>0,\ b>0,\ x\geq 0)$. $W_p(\alpha, R)$ denotes a Wishart distribution.

**Model 2**: HB model with unequal error variances

(i) $[y_{ik}|\beta_i, \sigma_e^2] \overset{\mathrm{ind}}{\sim} N(x_{ik}^{T}\beta_i, \sigma_e^2)$, $i=1,\ 2,\ ...,\ I;\ k=1,\ 2,\ ...,\ K,$ (4.12)

(ii) $[\beta_i|\gamma,\ \Phi] \overset{\mathrm{ind}}{\sim} N_p(Z_i\gamma,\ \Phi),$ (4.13)

(iii) Marginal prior distributions are as follows: $\gamma \sim N_q(0,\ D)$, $\tau_i \overset{\mathrm{ind}}{\sim} G(a_i,\ b_i)$, and $\Omega \sim W_p(\alpha,\ R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$, and $D$, $a_i$, $b_i$, $\alpha$ and $R$ are known.

We can use the Gibbs sampler to obtain the posterior estimates of $\mu_{ik} = x_{ik}^{T}\beta_i$ for the $k^{\mathrm{th}}$ month in the $i^{\mathrm{th}}$ small area using the posterior distribution of $\beta_i$ given $y = (\{y_{ik}\},\ i=1,\ 2,\ ...,\ I;\ k=1,\ 2,\ ...,\ K)$. Its implementation requires generating samples from full conditional posterior distributions. The necessary full conditional posterior distributions under Model 1 are given by:

For $i=1,\ 2,\ ...,\ I,\ k=1,\ 2,\ ...,\ K,$

(i) $$[\beta_i \mid y,\ \gamma,\ \Omega,\ \tau_e] \overset{\mathrm{ind}}{\sim} N_p\begin{pmatrix}(\tau_e \sum_k x_{ik}\, x_{ik}^{T} + \Omega)^{-1} \\ (\tau_e \sum_k y_{ik}\, x_{ik} + \Omega Z_i\,\gamma), \\ (\tau_e \sum_k x_{ik} x_{ik}^{T} + \Omega)^{-1}\end{pmatrix},$$

(ii) $$[\gamma \mid y,\ \beta,\ \Omega,\ \tau_e] \sim N_q\begin{pmatrix}\left(\sum_i Z_i^{T}\Omega Z_i + D^{-1}\right) \\ \left(\sum_i Z_i^{T}\Omega\beta_i\right), \\ \left(\sum_i Z_i^{T}\Omega Z_i + D^{-1}\right)^{-1}\end{pmatrix},$$

(iii) $$[\Omega \mid y,\ \beta,\ \gamma,\ \tau_e] \sim W_p\begin{pmatrix}\alpha + I,\ R \\ +\frac{1}{2}\sum_i(\beta_i - Z_i\,\gamma)(\beta_i - Z_i\gamma)^{T}\end{pmatrix},$$

(iv) $$[\tau_e \mid y,\ \beta,\ \gamma,\ \Omega] \sim G\begin{pmatrix}a + \frac{IK}{2},\ b \\ +\frac{1}{2}\sum_i\sum_k (y_{ik} - x_{ik}^{T}\beta_i)^2\end{pmatrix}.$$

Using initial values $\gamma^{(0)}$, $\Omega^{(0)}$ and $\tau_e^{(0)}$, we can generate samples iteratively based on (i)–(iv). The $M$ Gibbs samples $\{\beta_i^{(m)}, \gamma^{(m)}, \Omega^{(m)}, \tau_e^{(m)};\ m=1,\ 2,\ ...,\ M\}$ after implementing a "burn-in" period are assumed to be iterative samples from the joint posterior distribution of $\beta_i$, $\gamma$, $\Omega$ and $\tau_e$. The posterior estimates of $\beta_i$ can be calculated using the $M$ iterative samples $\{\beta_i^{(m)};\ m=1,\ 2,\ ...,\ M\}$.

The posterior mean of $\mu_{ik}$ and posterior variance of estimates can be obtained by implementing Markov chain Monte Carlo (MCMC) integration techniques from $M$ Gibbs samples. It should be noted that should the Gibbs samples of the parameters be produced using the WinBUGS program (Spiegelhalter, Thomas and Best 2000), the need to derive the full conditional posterior distributions for the parameters mentioned above ceases to exist. This is due to the fact that the Gibbs samples would be produced by the full conditional posterior distributions of the parameters (inherent in the process of running the program), provided that the applicable model, priors and the initial values of the parameters are given to the WinBUGS program. The full conditional distributions for Gibbs sampling under Model 2 are similar to the above Model 1.

## 5. Data Analysis

### 5.1 Description of the Data and HB Model Fitted

Before we continue, we highlight the point that direct, synthetic, composite and HB estimates were all derived using the EAPS data of December 2000. However, the HB estimates were derived using additional EAPS data of May and July 2000 for model fitting.

The large area ChoongBuk Province in Korea consists of 10 local self-government areas (LSGAs), which are small areas. The number of sample enumeration districts of the ChoongBuk Province allocated in the EAPS is 63, and the number of sample households is 1,512. Under the EAPS, the planned domains are large areas such as the ChoongBuk Province, and hence small areas such as the LSGAs fall under the category of unplanned domains. This leads to the concern that should the estimates of the total unemployed of the LSGAs be derived using only the sample enumeration districts allocated under the LSGAs, the standard errors will become large. To address this problem, we have used data of neighboring small areas with similar economic and demographic characteristics as the areas considered here as complementary information for small area estimation. We have first divided the large area of ChoongBuk Province into two sub-regional groups with similar economic and demographic characteristics. The two sub- regional groups mentioned above are Cities and Counties. We next divided each sub-regional group into four categories of sex (male, female) by age (15–34, 35, and over). The unemployment and economically active population (EAP) estimates for each of the categories of each sub-regional group were derived from the EAPS data.

Using the above estimates and the estimated resident population for each of the four categories of LSGAs produced monthly by the Korean National Statistical Office as supplementary data, we have estimated the synthetic and composite estimates for the unplanned domains (10 LSGAs) within the ChoongBuk Province based on the EAPS data of December 2000.

Let the direct estimate for the $k^{\text{th}}$ month in small area $i$ be $y_{ik}$. The direct estimates derived from the EAPS data of May, July and December 2000 were used as dependent variates in HB multi-level models. The additional auxiliary variates for the $k^{\text{th}}$ month in small area $i$ are as follows:

$$x_{ik} = (x_{i1k},\ x_{i2k},\ x_{i3k},\ x_{i4k})^T,\ i = 1,\ 2,\ \ldots,\ I;\ k = 1,\ 2,\ 3$$

$$= \begin{pmatrix} \left(\frac{{}_1\hat{P}_i}{{}_1\hat{X}_{\cdot}}\,{}_1\hat{Y}_{\text{dir}}\right)_k, \left(\frac{{}_2\hat{P}_i}{{}_2\hat{X}_{\cdot}}\,{}_2\hat{Y}_{\text{dir}}\right)_k, \\ \left(\frac{{}_3\hat{P}_i}{{}_3\hat{X}_{\cdot}}\,{}_3\hat{Y}_{\text{dir}}\right)_k, \left(\frac{{}_4\hat{P}_i}{{}_4\hat{X}_{\cdot}}\,{}_4\hat{Y}_{\text{dir}}\right)_k \end{pmatrix}^T.$$

The element of $x_{ik}$, $({}_j\hat{P}_i / {}_j\hat{X}_{\cdot})\,{}_j\hat{Y}_{\text{dir}}$ $(j = 1,\ 2,\ 3,\ 4)$, denotes the estimate of the total unemployed of the $j^{\text{th}}$ sex-age category in small area $i$, which is given in (4.1). We tried to adjust the direct estimates, $y_{ik}$, through the HB multi-level model with auxiliary variates, $x_{ik}$. The random regression coefficient vector $\beta_i = (\beta_{i1},\ \beta_{i2},\ \beta_{i3},\ \beta_{i4})^T$ of the $i^{\text{th}}$ small area in (4.9) was assumed to have the following structure:

$$\beta_{i1} = \gamma_{10} + \nu_{i1};\ \beta_{i2} = \gamma_{20} + \nu_{i2};\ \beta_{i3} = \gamma_{30} + \nu_{i3};\ \beta_{i4} = \gamma_{40} + \nu_{i4},$$

where the fixed regression parameter vector $\gamma = (\gamma_{10},\ \gamma_{20},\ \gamma_{30},\ \gamma_{40})^T$ is an unknown value, and the random effect vector $\nu_i = (\nu_{i1},\ \nu_{i2},\ \nu_{i3},\ \nu_{i4})^T$ of the $i^{\text{th}}$ small area follows $N_4(0,\ \Phi)$.

Using the vague proper priors for $\gamma, \tau$ and $\Omega$ determined by setting $D = \text{diag}(10^4,\ 10^4,\ 10^4,\ 10^4)$, $\alpha = 4$, $a = b = a_i = b_i = 0.001$ and $R$ with diagonal elements of 1 and off-diagonal elements of 0.001, we generated 6,000 Gibbs samples iteratively. Using the 3,000 samples after the "burn-in" period (3,001–6,000), the posterior means of unemployed persons of the $i^{\text{th}}$ small area and the posterior variances of the estimates were calculated. The data analysis was conducted using the WinBUGS program.

### 5.2 Model Selection

We considered model checking and comparison using MCMC methods under the two assumed HB multi-level model frameworks. First, we examined the posterior means of standardized residuals,

$$\text{resid}_{ik} = \frac{y_{ik} - E(y_{ik})}{\sqrt{\text{Var}(y_{ik})}},\ i = 1,\ 2,\ \ldots,\ 10;\ k = 1,\ 2,\ 3,$$

which are directly computable in WinBUGS. Here $y_{ik}$ are the direct estimates obtained from the data of the EAPS, and $E(y_{ik})$ and $\text{Var}(y_{ik})$ are obtained from the predictive distribution of $y_{ik}$. Figure 1 and Figure 2 give their normal Q-Q plots, both revealing a high degree of agreement with normality.

To make a comparison between the assumed HB multi-level models, we calculated a negative cross-validatory log-likelihood, $-\sum_{i,k} \log f(y_{ik} \mid y_{(ik)})$, and a posterior mean deviance, $-2\sum_{i,k} \log f(y_{ik} \mid \theta)$, for each model. The two measures are also computable using the WinBUGS program. $y_{(ik)}$ denotes all data except $y_{ik}$ and $\theta$ represents the parameters of the predictive distribution of $y_{ik}$. Table 1 gives the results for the HB multi-level model checks based on a 3,000 iteration BUGS run. Model 2 yielded a negative cross-validatory log-likelihood of 121.52 and a posterior mean deviance of 243.05, both of which are smaller than the corresponding Model 1 values. For our data, Model 2 provides a better fit than Model 1.

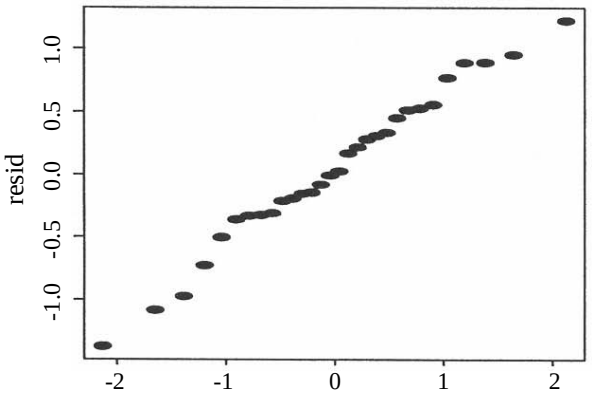



**Figure 1**. Normal Q-Q Plot (Model 1).

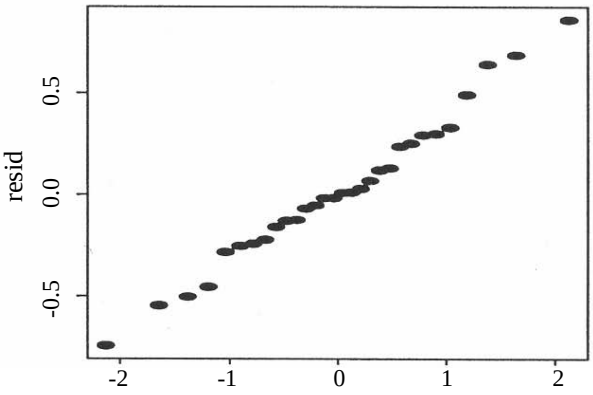



**Figure 2**. Normal Q-Q Plot (Model 2).

**Table 1**
Relative Comparison Between HB Multi-level Models

| HB Model | Negative Cross-Validatory Log-likelihood | Deviance |
|---|---|---|
| Model 1 | 188.67 | 377.30 |
| Model 2 | 121.52 | 243.05 |

In order to study how the direct estimates $y_{ik}$ support the HB multi-level models, we employed conditional predictive ordinate (CPO) values (You and Rao 2000, page 178). The CPO values under Model 1 are calculated by

$$\hat{\text{CPO}}_{ik}^{\text{HB}} = \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{f(y_{ik}\,|\,\beta_i^{(m)},\sigma_e^{2(m)})}}$$

for $i = 1, 2, ..., 10$, $k = 1, 2, 3$, where $f(y_{ik}\,|\,\beta_i, \sigma_e^2)$ are the conditional normal densities given by (4.10). For model 2, the CPO values are calculated with $\sigma_i^{2(m)}$. Using the Gibbs sampler, we can calculate the CPO values for all points (see Gelfand (1995) for a more detailed discussion). Figure 3 gives a CPO comparison plot for the two assumed HB multi-level models.

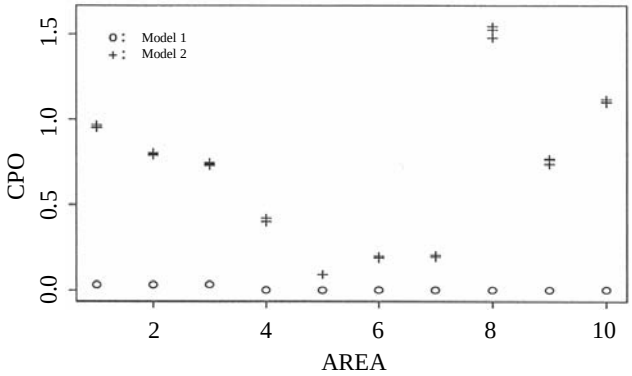



**Figure 3.** CPO comparison plot.

Model 2 proves to be the better of the two HB multi-level models, since its CPO values are significantly larger in every small area than those for Model 1. Therefore, we conclude that Model 2 with unequal error variances is a good model for our data.

### 5.3 Estimation Results

Table 2 shows the estimates of the total unemployed of the 10 LSGAs within the ChoongBuk Province under the EAPS data of December 2000. The estimated standard errors of the direct and HB estimates are provided together with the Jackknife root mean square errors of the synthetic and composite estimates.

In general the direct estimates prove to be highly unstable. Studying the Jackknife root mean square errors of the estimates of the total unemployed in the LSGAs, we find that in comparison to the direct estimates, synthetic and composite estimates are more stable. Although the estimated standard errors of the HB estimates are clearly smaller than those of the direct estimates over all the LSGAs, they turn out to be highly variable in certain LSGAs such as areas 3, 4, and 5. Overall, the composite estimates are more stable than other estimates for our data.

In order to evaluate the reliability of the direct and HB estimates of each LSGA, the relative standard errors of these estimates were obtained. Similarly, the reliability of synthetic and composite estimates was evaluated by the relative bias values and the relative root mean square errors of these estimates. Denoting $\hat{Y}_{i\cdot}^{*}$ as the estimator of the total unemployed in the $i^{\text{th}}$ small area, its relative bias (RB), relative standard error (RSE) and relative root mean square error (RRMSE) are given by the following respectively:

$$\text{RB}(\hat{Y}_{i\cdot}^{*}) = \frac{\hat{\text{Bias}}(\hat{Y}_{i\cdot}^{*})}{\hat{Y}_{i\cdot}^{*}} \times 100,$$

$$\text{RSE}(\hat{Y}_{i\cdot}^{*}) = \frac{\sqrt{\hat{\text{Var}}(\hat{Y}_{i\cdot}^{*})}}{\hat{Y}_{i\cdot}^{*}} \times 100,$$

$$\text{RRMSE}(\hat{Y}_{i\cdot}^{*}) = \frac{\sqrt{\text{mse}(\hat{Y}_{i\cdot}^{*})}}{\hat{Y}_{i\cdot}^{*}} \times 100.$$

Under the condition that $\hat{Y}_{i\cdot}^{*}$ is an unbiased estimator, the RSE and the RRMSE of $\hat{Y}_{i\cdot}^{*}$ are identical.

Table 3 shows the RB, RSE and RRMSE values of the estimates of the total unemployed of the 10 LSGAs within the ChoongBuk Province.

When comparing the bias values of synthetic and composite estimates, the average relative bias value of the composite estimates ($Av.\,\mathrm{RB} = 10.26\%$) is somewhat smaller than that of the synthetic estimates ($Av.\,\mathrm{RB} = 12.24\%$). However, both the synthetic and composite estimators show large values of bias in most small areas with the exception of two areas (areas 3 and 10).

We evaluate the reliability of these estimates based on the RSE (or RRMSE) values of small area estimates. It should be noted that since the direct estimates shown in Table 3 are unbiased, the RSE and RRMSE values of these direct estimates are identical. The National Statistical Office of Korea expects an approximate maximum RSE (or RRMSE) limit of 25% as the standard for reliability of small area estimates. With the exception of area 1, the RSE values of direct estimates do not satisfy this criterion for reliability. It follows that under the current EAPS system, direct estimates are unreliable. In contrast, both the RRMSE values of synthetic and composite estimates and the RSE values of the HB estimates were much smaller than the RSE(= RRMSE) values of the direct estimates in all LSGAs considered.

**Table 2**
Estimates of the Total Unemployed for Ten Local Self-Government Areas (LSGA) in ChoongBuk (December, 2000)

| Area No. | Direct | | Synthetic | | Composite | | Hierarchical Bayes (Model 2) | | $n_i$ |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{Y}_{i\cdot}$ | Est. se | $\hat{Y}_{i\cdot}^{S}$ | $\sqrt{\mathrm{mse}_{\mathrm{JN}}}$ | $\hat{Y}_{i\cdot}^{C}$ | $\sqrt{\mathrm{mse}_{\mathrm{JN}}}$ | $\hat{\mu}_{ik}^{\mathrm{HB}}$ | Est. se | |
| 1 | 8,517 | 1,733 | 7,969 | 580 | 8,023 | 493 | 8,514 | 358 | 22 |
| 2 | 3,949 | 1,445 | 2,823 | 725 | 3,050 | 607 | 3,773 | 474 | 11 |
| 3 | 365 | 390 | 1,830 | 110 | 1,723 | 101 | 399 | 152 | 4 |
| 4 | 503 | 373 | 612 | 234 | 581 | 196 | 440 | 106 | 2 |
| 5 | 781 | 676 | 1,164 | 169 | 1,140 | 158 | 567 | 261 | 3 |
| 6 | 1,275 | 577 | 1,230 | 282 | 1,238 | 233 | 1,138 | 270 | 3 |
| 7 | 1,032 | 646 | 1,459 | 295 | 1,384 | 252 | 1,035 | 117 | 5 |
| 8 | 1,795 | 893 | 1,825 | 346 | 1,821 | 306 | 1,790 | 69 | 6 |
| 9 | 1,023 | 602 | 2,888 | 574 | 2,000 | 270 | 970 | 200 | 5 |
| 10 | 512 | 384 | 872 | 94 | 851 | 92 | 511 | 63 | 2 |

**Table 3**
Relative Standard Errors (RSE) of Direct and HB Estimates for Ten Local Self-Government Areas (LSGA). Relative Bias (RB) Values and Relative Root Mean Square Errors (RRMSE) of Synthetic and Composite Estimates for Ten LSGAs (December, 2000)

Unit %

| Area No. | Direct | Synthetic | | Composite | | Hierarchical Bayes (Model 2) |
|---|---|---|---|---|---|---|
| | $\mathrm{RSE}_i$ | $\mathrm{RB}_i$ | $\mathrm{RRMSE}_i$ | $\mathrm{RB}_i$ | $\mathrm{RRMSE}_i$ | $\mathrm{RSE}_i$ |
| 1 | 20.35 | 6.92 | 7.27 | 5.99 | 6.15 | 4.20 |
| 2 | 36.59 | 23.77 | 25.69 | 18.39 | 19.91 | 12.56 |
| 3 | 106.91 | −2.95 | 5.99 | −2.87 | 5.89 | 37.97 |
| 4 | 74.15 | 16.26 | 38.30 | 14.37 | 33.73 | 24.00 |
| 5 | 86.58 | −7.04 | 14.51 | −6.67 | 13.84 | 45.94 |
| 6 | 45.23 | 17.56 | 22.90 | 14.43 | 18.80 | 23.69 |
| 7 | 62.56 | 14.86 | 20.25 | 13.29 | 18.21 | 11.28 |
| 8 | 49.77 | 15.25 | 18.97 | 13.49 | 16.78 | 3.87 |
| 9 | 58.83 | 15.01 | 19.88 | 10.20 | 13.50 | 20.65 |
| 10 | 74.93 | −2.75 | 10.79 | −2.82 | 10.79 | 12.29 |
| *Av.* RB | | 12.24 | | 10.26 | | |
| *Av.* RSE | 61.59 | | | | | 19.65 |
| *Av.* RRMSE | | | 18.46 | | 15.73 | |

*Av.* RB = average absolute relative bias over all LSGAs.
*Av.* RSE = average relative standard error over all LSGAs.
*Av.* RRMSE = average relative root mean square error over all LSGAs.

It has been noted that both composite and synthetic estimators produced reliable estimates for all the LSGAs, and also that the estimates were similar to each other. However, we stress that the composite estimator showed higher gains in efficiency against the synthetic estimator in all the LSGAs. Despite being efficient and reliable in eight of the LSGAs (areas 1, 2, 4, 6, 8, 9, and 10), the HB estimates fall below the criterion of reliability in the other two LSGAs (areas 3 and 5).

The RRMSE values of the composite estimates are on average 70.66% smaller than the RSE(=RRMSE) values of the direct estimates, with this figure ranging from 45.59% (area 2) to 94.49% (area 3). In comparing RSE values of the direct and HB estimates, HB estimates are on average 69.44% smaller than the direct estimates, with this figure ranging from 46.94% (area 5) to 92.22% (area 8). It is notable that $RSE_3 = 37.97\%$ and $RSE_5 = 45.94\%$ in HB estimation, which reflects not only that there are large variations within areas 3 and 5, but also possible variations of the estimates within each area for different months. For such areas as 3 and 5, it is suggested that additional sample enumeration districts should be allocated to reduce the standard errors of the estimates. Thus we come to the conclusion that under the current EAPS system, the composite estimator were more stable and reliable than the other estimators, and while the model-based HB estimator can be efficient in most areas, it has a major shortcoming in that it is highly variable in some areas.

## 6. Conclusion

The Korean EAPS is a nation-wide sample survey, and the only official source producing monthly employment and unemployment figures. The monthly-published data includes the unemployment rate, employment rate, the economically active rate and also the demographic characteristics of the productive population. However, the EAPS design focuses on figures for large areas such as Metropolitan Cities and Provincial levels, and hence is a less than suitable source on its own for obtaining unemployment figures of unplanned sub-regional domains such as the LSGAs, especially since these areas are increasingly attracting interest. We have suggested the design-based indirect estimators (synthetic and composite estimators) and HB multi-level model estimators for deriving unemployment figures for the LSGAs within large areas, using only the EAPS data and the official figures of the Korean National Statistical Office (supplementary administrative information). The Jackknife mean square errors of the synthetic and composite estimates were introduced as measures of accuracy for the small area estimates. The posterior variances of the HB estimates were also used as measures of precision for the small area estimates.

The results using the EAPS data show that the small area estimators (synthetic, composite and HB multi-level model estimators) were much more effective in comparison to results obtained using the direct estimator, and moreover most of these estimates had significantly lower standard errors (or root mean square errors) than that of the direct estimates. In terms of gains in efficiency, the composite estimator performed much better than other estimators.

The Korean EAPS is conducted every month, in addition to which an overall review and redesign of the survey is carried out every five years. In constructing a new survey, a general review of population stratification, sample allocation and clustering is being considered so that the reliability of small area level estimates can be strengthened. Studies to estimate other relevant domains such as sex, age and education in addition to the existing sub-regional domains within large areas are under consideration, based on the new survey design.

## References

Chattopadhyay, M., Lahiri, P., Larsen, M. and Reimnitz, J. (1999). Composite estimation of drug prevalences for sub-state areas. *Survey Methodology*, 25, 81-86.

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

Gelfand, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman and Hall, 145-161.

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

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

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

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

Spiegelhalter, D., Thomas, A. and Best, N. (2000). *WinBUGS Version 1.3 User Manual*. MRC Biostatistics.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.