SORT 32 (2) July-December 2008, 177-198

Sampling design variance estimation of small area estimators in the Spanish Labour Force Survey*

M. Herrador[a], D. Morales[b], M. D. Esteban[b], A. Sánchez[b],
L. Santamaría[b], Y. Marhuenda[b] and A. Pérez[b]

Abstract

The main goal of this paper is to investigate how to estimate sampling design variances of model-based and model-assisted small area estimators in a complex survey sampling setup. For this purpose the Spanish Labour Force Survey is considered. Sample and aggregated data are taken from the Canary Islands in the second trimester of 2003 in order to obtain some small area estimators of ILO unemployment totals. Several problems arising from the application of standard small area estimation procedures to the survey are described. It is shown that standard variance estimators based on explicit formulas are not applicable in the strict sense, since the assumptions under which they are derived do not hold. In addition two resampling techniques, bootstrap and jackknife, are considered. These methods treat all the considered estimators in the same manner and therefore they can be used as performance measures to compare them. From the analysis of the obtained results, some recommendations are given.

MSC: 62D05, 62J05.

Keywords: Labour Force Survey, small area estimation, linear models, mean squared error, bootstrap, jackknife, unemployment totals, calibrated weights.

1 Introduction

Small area estimation is an increasingly important part of survey sample inference with applications to social and economic statistics. Almost all the methodological developments up to date in this context has been carried out under the assumption that the assumed small area model is true, and that the appropriate measure of accuracy of

* Supported by the grant MTM2006-05693.

[a] Instituto Nacional de Estadística, Spain.

[b] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Spain.

Received: October 2007
Accepted: July 2008

the small area estimator is its repeated sampling variability under random realizations of the population assuming the small area model holds. The fact that the assumed model only approximates reality, and that the measures that capture sampling variability relative to the actual population values are often of primary interest, is often ignored. This paper attempts to redress this imbalance by focusing on the repeated sampling properties of the most commonly used model-based methods of small area estimation.

The paper describes investigations on several issues arising from the application of standard small area estimation techniques, as they have been typically developed to be used under simple random sampling and they do not take into account problems derived from data coming up from surveys with complex sampling design, non response, reliability of population sizes, selection of auxiliary variables, consistency with the officially published data at a higher level of aggregation, estimation of mean squared error in a complex setup and many others. For this sake, some model-assisted and model-based estimators are adapted to the Spanish Labour Force Survey (SLFS) in order to estimate totals of unemployed people by sex and small areas in the Canary Islands. The paper has thus an applied-oriented character, attempting to diminish the gap between theory and practice.

The rest of the paper is organized as follows. Section 2 introduces some standard small area estimators and the corresponding explicit formulas to estimate their variances or mean squared errors. It also describes the auxiliary variables employed to estimate totals of unemployed people in the SLFS. Section 3 discusses two resampling approaches for estimating design-based variances. Section 4 describes technical details of the SLFS, with special emphasis on the sampling design, the separated ratio estimator of totals and the calibration of sampling weights. Section 5 proposes a two-stage bootstrap and a delete-one-cluster jackknife method to estimate sampling variances of small area estimators in the SLFS. These resampling methods produce performance measures to compare estimators of totals. Section 6 gives a discussion on the performance of the small area estimators and on the three methods to estimate their variances. The paper has two appendices. Appendix A presents estimated totals of unemployed people. Appendix B gives figures with estimated coefficients of variation and presents dispersion graphs to illustrate the behaviour of the small area estimators with respect to the basic estimator of the SLFS.

2 Estimators of small areas totals in complex surveys

Let Ω be a population with N units and let $s \subset \Omega$ be a sample of size n selected with a given sampling design. Let $\pi_i = P(i \in s)$ and $w_i = 1/\pi_i$ be the inclusion probability of unit $i \in \Omega$ and its sampling weight. Let y_i and $\mathbf{x}_i$ be the target variable and the vector of auxiliary variables defined for each $i \in \Omega$. Let $\mathbf{y}$ and $\mathbf{X}$ be the vector and the matrix containing the values of y_i and $\mathbf{x}_i$ for all units in the population. The three basic

inferential frameworks in survey sampling are the design-based, the model-based and the model-assisted approaches. In the design-based framework $\mathbf{y}$ and $\mathbf{X}$ are regarded as constants and the only source of randomness is the selection of the sample. In the model-based framework a model is assumed for $\mathbf{y}$ conditioned on $\mathbf{X}$. In the model-assisted framework, both probability sampling design and model have a role (see Särndal, et al. 1992, pp. 227, 238-239). The model is used to propose an estimator with the restriction of being approximately unbiased in the sampling distribution.

We are interested in estimating the total Y_d of a target variable y in a domain d of size N_d. Let $s_d = \Omega_d \cap s$ be the subsample of units in domain d. In this section we introduce some standard small area estimators of Y_d. We also give explicit formulas to estimate the sampling variances of design-based and model-assisted estimators and to estimate the mean squared errors of model-based estimators. As the main goal of this paper is to investigate how to estimate the design-based variance of different types of small area estimators, we consider four of them: a design-based, a model-assisted and two model-based ones. At the end of this section we describe the auxiliary variables employed in the SLFS setup.

2.1 Direct estimator

The direct estimator is the design-based estimator (10.3.6) appearing in Särndal et al. (1992), p. 391, when N_d is known. Its expression is

$$\hat{Y}_d^{dir} = N_d \hat{\bar{Y}}_d^{dir}, \text{ where } \hat{\bar{Y}}_d^{dir} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_j y_j \text{ and } \hat{N}_d = \sum_{j \in s_d} w_j.$$

An explicit-formula estimator of its sampling variance is

$$var(\hat{Y}_d^{dir}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{j \in s_d} w_j(w_j - 1)(y_j - \hat{\bar{Y}}_d^{dir})^2.$$

2.2 GREG estimator

GREG estimator is a model-assisted estimator. The one presented here is assisted by a linear model. Consider p explanatory variables measured at N population units; i.e. $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,p})$, $j = 1, ., N$. Let

$$\bar{\mathbf{X}}_d = \frac{1}{N_d} \sum_{j \in \Omega_d} \mathbf{x}_j \quad \text{and} \quad \hat{\bar{\mathbf{X}}}_d^{dir} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_j \mathbf{x}_j$$

be the domain means of the auxiliary variables and their direct estimators. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with rows $\mathbf{x}_j$, $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{W}^{-1})$ and $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$. The weighted least square estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y} = (\sum_{j\in s} w_j\mathbf{x}_j^t\mathbf{x}_j)^{-1}(\sum_{j\in s} w_j\mathbf{x}_j^t y_j).$$

Observe that the set of p explanatory variables can include artifical variables. Here the first variable is such that $x_{j,1} = 1$, $j = 1, ..., n$; i.e. we assume a linear model with intercept term. In this way, estimation of $\boldsymbol{\beta}$ does not depend on the type of selected small area in territories with hierarchical structure. In this paper the GREG estimator of a total is a slight modification of the model-assisted estimator (2.4.8) appearing in Särndal et al. (1992, p. 410). Its expression is

$$\hat{Y}_d^{greg} = N_d\hat{\bar{Y}}_d^{dir} + N_d(\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{dir})\hat{\boldsymbol{\beta}}.$$

Observe that

$$\hat{Y}_d^{greg} = \sum_{j\in s} g_{dj}w_j y_j \quad \text{and} \quad N_d\bar{\mathbf{X}}_d = \sum_{j\in s} g_{dj}w_j\mathbf{x}_j$$

where

$$g_{dj} = \frac{N_d}{\hat{N}_d}I_{\Omega_d}(j) + N_d(\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{dir})(\sum_{j\in s} w_j\mathbf{x}_j^t\mathbf{x}_j)^{-1}\mathbf{x}_j^t,$$

and I_{Ω_d} is the indicator function of subset Ω_d. An explicit-formula estimator of its sampling variance is

$$var(\hat{Y}_d^{greg}) = \sum_{j\in s_d} w_j(w_j - 1)g_{dj}^2(y_j - \mathbf{x}_j\hat{\boldsymbol{\beta}})^2.$$

2.3 EBLUPA estimator

The EBLUPA estimator is a composite estimator based on the 2-level linear mixed model (model A)

$$y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + u_d + v_{dj}^{-1/2}e_{dj},$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$ are independent. The model is fitted by calculating maximum likelihood estimators of the regression and variance component parameters with a Fisher-scoring algorithm (see e.g. Rao, 2003, ch. 5-6). The EBLUPA estimator of a total is $\hat{Y}_d^{eblupa} = N_d\hat{\bar{Y}}_d^{eblupa}$, where

$$\hat{\bar{Y}}_d^{eblupa} = \hat{\gamma}_d(\hat{\bar{Y}}_d^{dir} - \hat{\bar{\mathbf{X}}}_d^{dir}\hat{\boldsymbol{\beta}}) + \bar{\mathbf{X}}_d\hat{\boldsymbol{\beta}}, \quad \text{with } \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2/\nu_d)},\ \nu_d = \sum_{j \in s_d} \nu_{dj}.$$

The EBLUPA estimator is in fact a pseudo-eblup estimator studied in work package 4 of the EURAREA project (`http://www.statistics.gov.uk/eurarea/`) and related to the ones proposed by Prasad and Rao (1999) and You and Rao (2002). Mean squared error is estimated by using $g_1 - g_4$ explicit formulas given by Prasad and Rao (1990) and later extended by Das, Jiang and Rao (2001) to more general linear mixed models. Recent results are reviewed by Jiang and Lahiri (2006).

2.4 EBLUPB estimator

The EPLUPB estimator is a composite estimator based on the area-level model (model B)

$$\bar{Y} = \bar{\mathbf{X}}_d\boldsymbol{\beta} + u_d \quad \text{and} \quad \hat{\bar{Y}}_d^{direct} = \bar{Y}_d + \varepsilon_d,$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_d \overset{iid}{\sim} N(0, \sigma_d^2)$ are independent. This model was introduced by Fay and Herriot (1979) to estimate average per capita income for small areas in USA. The model is fitted by the same method as model A. Under model B, EBLUP estimator of total is

$$\hat{\bar{Y}}_d^{eblupb} = \hat{\gamma}_d\hat{\bar{Y}}_d^{dir} + (1 - \hat{\gamma}_d)\bar{\mathbf{X}}_d\hat{\boldsymbol{\beta}}, \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_d^2}.$$

Mean squared error is estimated by using $g_1 - g_3$ explicit formulas given by Prasad and Rao (1990).

2.5 Auxiliary variables to estimate totals of unemployed people in the SLFS

To obtain models with high predictive properties, the selection of adequate explanatory variables is very important. In the case of individual level models, auxiliary variables are needed at both individual and domain level. At the individual level auxiliary variables are obtained from the survey sample and, except for the cases of non-response, their values are available. However it is much more difficult to evaluate auxiliary variables at the domain level, because their values come from external sources which sometimes are not available, have not sufficiently good quality or may even present definition differences with their sample counterparts. Because of these reasons, the number of available auxiliary variables for individual-level models describing unemployment is in general very small. In the real data application of this paper the domains of interest are small areas (provisional geographical divisions for statistical purposes) crossed with

sex. There are $2 \times 27 = 54$ domains in the considered universe (Canary Islands). The following auxiliary variables have been used to estimate totals of unemployed people in the SLFS:

1. Auxiliary variables at aggregated and unit level are:
   - GSAC: groups of sex (1-2), age (1-3) and employment claimant (1-2) with 12 values. Three age groups have been considered: 16-24, 25-54 and ≥ 55.
   - CLUSTER: groups of province and population size of the municipality with 4 values.
2. An auxiliary variable aggregated at the domain level without sample counterpart has been used. This variable, GSAU, has 12 categories representing the groups of sex (1-2), age (1-3) and registered as unemployed in the administrative register of employment claimants (1-2).
3. To estimate totals of unemployed people we use the following auxiliary variables:
   - CLUSTER and GSAC for estimators GREG and EBLUPA.
   - CLUSTER and GSAU for estimators EBLUPB.

3 Design-based variance estimation

The most commonly used methods for design-based variance estimation with complex survey data are linearization and resampling methods. Krewski and Rao (1981) showed the asymptotic consistency of the variance estimates for nonlinear functions of design-unbiased mean estimators based on linearization or on some of the existing resampling methods applied to multistage designs in which the primary sampling units are selected with replacement. The linearization method requires theoretical calculation and subsequent programming of derivatives, which can make it cumbersome to implement. For this reason resampling methods are becoming each time more popular. In this section we review, without being exhaustive, some resampling methods that can be adapted to complex survey sampling designs.

3.1 Bootstrap with replacement

Efron (1979) proposed a bootstrap method that involves generation of independent resamples, each drawn from the original with replacement. For each such resample the statistic of interest is calculated and the obtained values form the basis of inference. The properties of the bootstrap method have been extensively studied for the i.i.d. case. In the framework of survey sampling Efron's original bootstrap requires modifications to handle issues like finiteness of population, without replacement sampling, complexity

of survey designs, weighting schemes and nonlinearity of population parameters and estimators. Under random sampling without replacement the finite population correction factor (f.p.c.), $f = n/N$, plays an important role. If f is not negligible the with bootstrap with replacement (BWR) method tends to overestimate the variance of linear estimators. To overcome this difficulty McCarthy and Snowden (1985) suggested to use a bootstrap sample size $n' = (1 - f)^{-1}(n - 1)$. Rao and Wu (1988) proposed a BWR method which rescales the bootstrap samples so as to recover the f.p.c. factor in the usual simple random sampling without replacement (SRSWOR) variance formula for the design unbiased estimators of the population mean. For interesting related papers dealing with the impact of BRW methods in survey sampling, see Rust and Rao (1996), Sitter (1992), Shao (2003) and Lahiri (2003).

All the mentioned references treat the problem of estimating parameters of the global population. However, in the small area estimation (SAE) setup the application of the BRW method has extra difficulties because the parameters of interest are from non-designed domains with small expected sample sizes. Further in the SAE framework is quite common to use nonlinear model-based estimators (like the EBLUP). For these estimators, there exist a variety of methods to estimate their model-based mean squared error, but not to estimate its design-based variance. For these reasons the bootstrap proposals mentioned above need adaptation to be applied in a SAE setup with complex survey data, so that the naive BWR method becomes a simpler and worthwhile approach to be considered for complex sampling designs like the one of the SLFS. In this paper, the naive BWR method involves the following basic steps:

1. Using a suitable probability sampling scheme, generate resamples from the original sample.
2. From each resample calculate the estimator $\hat{\theta}$. Denote them by $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.
3. Bootstrap estimator of variance is $var_B(\hat{\theta}) = (B - 1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^*_b - \hat{\theta}^*)^2$ with $\hat{\theta}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*_b$.

3.2 Jackknife

Quenouille (1949) introduced the jackknife method to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data. The jackknife has become a more valuable tool since Tukey (1958) found that the jackknife can also be used to construct variance estimators. The first theorem concerning the jackknife variance estimator was given by Miller (1964). Since then jackknife theory has been widely developed (see e.g. Shao and Tu, 1995), although not much work has been done on its adaptation to complex survey designs. One exception is the paper by Rao and Tasui (2004) where jackknife variance estimators are introduced under stratified multistage sampling. Rao and Tasui (2004) consider a population stratified in L strata and from each stratum h, $m_h \geq 2$ clusters are

selected, independently across the strata. They further assume that subsampling within the sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi}, $i = 1, \ldots, m_h$, $h = 1, \ldots, L$. An unbiased estimator of the cluster total Y is given by

$$\hat{Y} = \sum_{s} w_{hik}\, y_{hik},$$

where w_{hik} is the inverse of the first order inclusion probability of unit k in cluster i and stratum h. To obtain the delete-one-cluster jackknife estimator of the variance of $\hat{Y}$ the jackknife weights, when the (g, j)-th cluster is deleted or equivalently in the jackknife sample $s^*_{(g,j)}$, are $w^*_{hik(g,j)} = w_{hik} b_{hi(g,j)}$ with $b_{hi(g,j)} = m_g/(m_g - 1)$ if $h = g$, $i \neq j$, and $b_{hi(g,j)} = 1$ if $h \neq g$. The variance of $\hat{Y}$ can be approximated by

$$var_J(\hat{Y}) = \sum_{g=1}^{L} \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{Y}^*_{(g,j)} - \hat{Y})^2, \quad \text{where} \quad \hat{Y}^*_{(g,j)} = \sum_{s_{(g,j)}} w^*_{hik(g,j)}\, y_{hik}.$$

4 The Spanish Labour Force Survey

The SLFS is a good example of complex survey design where a lot of challenging statistical issues takes place. The research in this paper is motivated by them. This section summarizes the key points of its sampling design as well as some details on how the collected survey data is handled in practice. Additional information can be downloaded from web site of the Spanish Statistical Office (Instituto Nacional de Estadística-INE)

`http://www.ine.es/en/docutrab/epa05_disenc/epa05_disenc_en.pdf.`

SLFS is a quarterly survey following a stratified two-stage random sampling design with separate samples s_p drawn from each province p. The Primary Sampling Units (PSUs) are Census Sections (geographical areas with a maximum of 500 dwellings-approximately 3000 people) and they are grouped in strata according to the size of municipality. Within each stratum, PSUs are selected with probabilities proportional to size according to the number of dwellings. In the second stage sampling, the Secondary Sampling Units (SSUs) are dwellings and a random start systematic sampling is applied to draw a fixed number (18 in most cases) of SSUs from each selected PSU. All people aged 16 years old or more in the selected SSUs are interviewed. The probability that a dwelling v belonging to PSU a of stratum h be selected in s_p is

$$P(Dwe_{hav}) = P(PSU_{ha}) P(Dwe_{hav}) | PSU_{ha}) = m_h \frac{V_{ha}}{V_h} \frac{18}{V_{ha}} = \frac{18 m_h}{V_h},$$

where V_{ha} and V_h are the totals of dwellings in PSU a of stratum h and in stratum h respectively and m_h is the number of sections allocated in stratum h. Because all individuals in a selected dwelling are interviewed, the inclusion probabilities of individuals and dwellings coincide. Therefore, the inclusion probability of individual j in dwelling v and stratum h is

$$\pi_j = \frac{18m_h}{V_h} = \pi_h.$$

This means that all individuals within a given stratum have the same selection probability, i.e. this survey uses what is called a self-weighting design. Afterwards, at stratum level, probabilities π_j are modified to take non-response into account and their inversions produce sampling weights $w_j^{(1)}$ adjusted by non-response. Consequently the survey is still using a self-weighting design inside of each stratum. Up until year 2001 the INE used a ratio estimator, with Demographic Population Projections as auxiliary variable, to estimate the total Y_p of variables y in the province p, i.e.

$$\hat{Y}_p^{lfs,0} = \sum_{h \subset \Omega_p} \frac{N_h}{\hat{N}_h} \sum_{v \subset s_h} \sum_{j \in v} w_j^{(1)} y_j \quad \text{with} \quad \hat{N}_h = \sum_{v \subset s_h} \sum_{j \in v} w_j^{(1)} = w_j^{(1)} n_h,$$

where $\hat{N}_h$ is the projection of the population living in familiar dwellings in stratum h, with reference to the half of the quarter and n_h is the number of people living in the dwellings in the sample, in stratum h, at the time of the interview. Alternatively,

$$\hat{Y}_p^{lfs,0} = \sum_{h \subset \Omega_p} \sum_{j \in s_h} \frac{N_h w_j^{(1)}}{\hat{N}_h} y_j = \sum_{j \in s_p} w_j^{(2)} y_j,$$

with the sample dependent weights

$$w_j^{(2)} = w_j^{(2)}(s_p) = \frac{N_h w_j^{(1)}}{\hat{N}_h} = \frac{N_h}{n_h} \quad \text{if} \quad j \in s_h.$$

Since the first quarter of 2002, reweighting (or calibration) techniques are applied to estimators so as to adjust the survey estimates to some given information from external sources. The reweighting technique (see Deville and Särndal (1992)) requires the availability of K auxiliary variables appearing in the sample s_p and whose populations totals are known, i.e.

$$\sum_{j \in \Omega_p} x_{jk} = X_k, \quad k = 1, \ldots, K.$$

The target is to find a new estimator

$$\hat{Y}_p^{lfs} = \sum_{j \in s_p} w_j y_j$$

with new weights w_j satisfying the balance equations

$$\sum_{j \in s_p} w_j x_{jk} = X_k, \quad k = 1, \ldots, K,$$

and being as similar as possible to $w_j^{(2)}$. The problem aims to find values w_j minimizing

$$\sum_{j \in s_p} w_j^{(2)} G(w_j / w_j^{(2)}) \quad \text{restricted to} \quad \sum_{j \in s_p} w_j x_{jk} = X_k, \quad k = 1, \ldots, K,$$

where G is a function of distance. In the second trimester of 2003 the SLFS weights were calibrated so that their sum coincide with the population projections for individuals aged 16 years and over per groups of sex and age in autonomous communities, and per provinces. In order to obtain the practical solution for this problem, it was employed the CALMAR (CALage sur MARges) software, programmed in SAS code by the INSEE (Institut National de la Statistique et des Études Économiques) in France.

SLFS estimator of the total Y_p of variable y in province p is $\hat{Y}_p^{lfs}$. In this setup direct estimators of the total and the mean (cf. Section 2.1) of domain d are

$$\hat{Y}_d^{lfs} = \sum_{j \in s_p} w_{dj} y_{dj} \quad \text{and} \quad \hat{\bar{Y}}_d^{lfs} = \frac{\hat{Y}_d^{lfs}}{\hat{N}_d}, \quad \text{with} \quad \hat{N}_d = \sum_{j \in s_p} w_{dj}.$$

For provinces, it holds $\hat{Y}_p^{lfs} = \sum_{d \in \Omega_p} \hat{Y}_d^{lfs}$; i.e. there exists consistency between direct estimates at domain and SLFS estimate at province level.

INE publishes estimates of unemployment totals at province level. If in the near future these publications were extended to domain levels it should be necessary to force consistency between both types of data. This is to say that the sum of the estimated totals in all the domains within a province should coincide with the actual estimated total by SLFS in the province. In order to fulfil this consistency criterion, in this paper the following modification of all the considered small area estimators has been implemented.

Let $\hat{Y}_d^{lfs}$ be the SLFS estimator of total Y_p in province p. Assume that province p is partitioned in D_p domains; i.e. $\Omega_p = \cup_{d=1}^{D_p} \Omega_{pd}$ with $\Omega_{d_1} \cap \Omega_{d_2} = \emptyset$ if $d_1 \neq d_2$. Let $\hat{Y}_1, \ldots, \hat{Y}_{D_p}$ be some given estimators of totals $Y_1, \ldots, Y_{D_p}$. In general, the consistency property

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d$$

Table 4.1: *Consistency factors for the totals of unemployed men (left) and women (right) in the SLFS 2003/02 of Canary Islands.*

Province	dir	greg	eblupa	eblupb	dir	greg	eblupa	eblupb
1	0.948	0.944	0.952	0.949	0.937	0.875	0.865	0.932
2	1.005	0.963	0.947	0.981	0.995	0.868	0.879	0.998

is not satisfied. In such cases $\hat{Y}_1, \ldots, \hat{Y}_{D_p}$ can be transformed into consistent estimators by the following calculation

$$\hat{Y}_d^c = \lambda_{yp} \hat{Y}_p \,, \quad \text{where} \quad \lambda = \frac{\hat{Y}_p^{lfs}}{\sum_{d=1}^{D_p} \hat{Y}_d}$$

are *consistency factors*. For consistent estimators, it holds

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d^c.$$

Table 4.1 presents the consistency factors of direct, GREG, EBLUPA and EBLUPB estimators of totals of unemployed men (left) and women (right) in the SLFS 2003/02 of Canary Islands. One can observe that the deviations from the SLFS estimation at province level are at most of 15% for the four small area estimators.

5 Resampling methods for design-based variance estimation in the SLFS

In this section we describe a two-stage bootstrap method as well as a two-stage jackknife method to estimate variances of small area estimators of totals in the SLFS.

5.1 A naive two-stage bootstrap method

Let θ be a parameter to be estimated with $\hat{\theta}$. Bootstrap (see e.g. Efron and Tibshirani, 1998) is a resampling method which is often used to estimate variances $var(\hat{\theta})$. To implement the proposed two-stage bootstrap method, it is not necessary to construct artificial populations since the procedure generates bootstrap samples directly from the original SLFS sample as it is explained in next lines. Let s be an SLFS sample in a given province. Let $s = \cup_{h=1}^{H} s_h$, where $s_1, \ldots, s_H$ are subsamples by strata. Let $s_h = \cup_{a=1}^{m_h} s_{ha}$, where $s_{h1}, \ldots, s_{hm_h}$ are subsamples in the m_h selected PSUs from the stratum h. Finally, let $s_{ha} = \cup_{v=1}^{m_{ha}} s_{hav}$, where $s_{ha1}, \ldots, s_{ham_{ha}}$ are the subsamples in the m_{ha} visited dwellings in PSU a and stratum h. Selection of bootstrap samples in stratum h, $h = 1, \ldots, H$, is done in the following way:

1. Select a simple random sample with replacement of m_h PSUs from the set of m_h PSUs appearing in the original SLFS sample.
2. Within each selected PSU, draw a simple random sample with replacement of m_{ha} dwellings from the set of m_{ha} dwellings appearing in the given PSU of the original SLFS sample.
3. Select all the individuals aged 16 or more from the dwellings in the bootstrap sample.

Variance estimation is done as follows:

A. By using the procedure described above, use sample s to draw B bootstrap samples ($B = 500$ in this paper). For every bootstrap sample calculate $\hat{\theta}^*_b$, $b = 1, ., B$, in the same way as $\hat{\theta}$ was calculated. So, in each bootstrap sample, the weights $w^{*(2)}_{j,b} = N_h/n^*_{hb}$ (where n^*_{hb} is the number of individuals selected in bootstrap sample b and stratum h) are adjusted by a calibration procedure to obtain calibration weights w^*_j in the same way as in SLFS sample. These calibration weights w^*_j are used to calculate $\hat{\theta}^*_b$.

B. The observed distribution of $\hat{\theta}^*_1, \ldots, \hat{\theta}_B$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.

C. The variance of $\hat{\theta}$ is approximated by

$$var_B(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \hat{\theta}^*)^2, \quad \text{where} \quad \hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b.$$

D. A bootstrap estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$cv_B(\hat{\theta}) = \frac{\sqrt{var_B(\hat{\theta})}}{\hat{\theta}}\,100.$$

An important step when estimating variances through the bootstrap method is to take into account the consistency property of estimators of totals at province level. The consistency property was not required in the bootstrap samples. To estimate variances of consistent estimators, estimated variances of non consistent estimators are multiplied by the square of the consistency factor λ (cf. Section 4). However, for the coefficient of variation this adjustment is not necessary. More concretely, if $\hat{\theta}^c = \lambda\hat{\theta}$ is the consistent version of a total estimator $\hat{\theta}$, where λ is the consistency factor calculated in the original SLFS sample, then bootstrap estimators of the variance and the coefficient of variation of $\hat{\theta}^c$ are

$$var_B(\hat{\theta}^c) = \lambda^2 var_B(\hat{\theta}) \quad \text{and} \quad cv_B(\hat{\theta}^c) = cv_B(\hat{\theta}).$$

5.2 A delete-one-cluster jackknife method

In order to apply the jackknife for variance estimation in SLFS samples, we use the delete-one-cluster jackknife method (see e.g. Rao and Tausi, 2004). To obtain the delete-one-cluster jackknife variance estimator of $\hat{\theta}$, we generate jackknife samples by deleting a PSU each time. So within each province, there are as many jackknife samples as PSUs are in the corresponding SLFS sample.

Consider the jackknife sample, $s^*_{(g,j)}$, obtained by excluding PSU j of stratum g. The jackknife weight of individual k, PSU i and stratum h in the sample $s^*_{(g,j)}$ is $w_{hik(g,j)} = w^{(2)}_{hik} b_{hi(g,j)}$, where $b_{hi(g,j)} = \frac{m_g}{m_g - 1}$ if $h = g$, $i \neq j$, $b_{hi(g,j)} = 1$ if $h = g$, and m_g is the number of PSUs in the stratum g. Note that the case $h = g$ and $i = j$ does not appear in the jackknife sample $s^*_{(g,j)}$. If H is the number of strata in the sample, the variance estimation is done as follows:

A. By using the procedure described above, use sample s to draw jackknife samples $s^*_{(g,j)}$, $g = 1, \ldots, H$, $j = 1, \ldots, m_g$. For every jackknife sample calculate $\hat{\theta}^*_{(g,j)}$ in the same way as $\hat{\theta}$ was calculated. So, in each jackknife sample, the weights $w_{hik(g,j)}$ are adjusted by a calibration procedure to obtain calibrated weights $w^*_{hik(g,j)}$ in the same way as it was done with the SLFS sample. Theses calibrated weights $w^*_{hik(g,j)}$ are used to calculate $\hat{\theta}^*_{(g,j)}$.

B. The observed distribution of $\{\hat{\theta}^*_{(g,j)} : g = 1, \ldots, H, j = 1, \ldots, m_g\}$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.

C. The variance of $\hat{\theta}$ can be approximated by

$$var_J(\hat{\theta}) = \sum_{g=1}^{H} \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{\theta}^*_{(g,j)} - \hat{\theta})^2.$$

D. A jackknife estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$cv_J(\hat{\theta}) = \frac{\sqrt{var_J(\hat{\theta})}}{\hat{\theta}} 100.$$

6 Discussion

6.1 On the small area estimators

In this section a specific analysis of the behaviour of direct, GREG, EBLUPA and EBLUPB estimators of unemployment totals (men and women), in the SLFS of Canary Islands in the second trimester of 2003, is given. Conclusions are mainly based on data

from Table A.1 and in figures presented in Appendix B. In Appendix B explicit-formula, bootstrap and jackknife estimates of the variances or mean squared errors (MSE) of the estimators of totals of unemployed men are plotted. In order to analyze the degree of bias of the estimators of totals, in Figure B.5 they are plotted against the basically unbiased-design SLFS estimator in dispersion graphs. Similar figures have been plotted for the case of women. However, for the sake of brevity, they are not presented here. In relation to the different estimators tested the main conclusions are:

1. The four considered estimators tend to give the same numerical results as LFS estimator when sample size increases. See Table A.1.
2. To estimate totals of unemployed people, the four considered estimators are acceptably unbiased with respect to LFS estimator (see Figure B.5). From the figures in Appendix B we conclude that EBLUPA estimator is in general the one with the lowest MSE.

6.2 On the estimation of variances or mean squared errors

In this section advantages and disadvantages of the three considered variance or mean squared error (MSE) estimation procedures (explicit formula, bootstrap and jackknife) are analyzed.

Explicit formulas to estimate the variance or MSE of estimators of totals are easy to implement and require the same sample and auxiliary information than the one needed for the given estimators of totals. These formulas can also be extended to more general types of parameters (e.g. nonlinear) via Taylor linearization.

In the case of design-based or model-assisted estimators the formulas of variances are derived with respect to the sampling distribution with some simplifications to avoid double inclusion probabilities. What is estimated is thus a simplified version of the variance. In addition, elevation factors are treated as if they were inverted inclusion probabilities. Explicit formulas to estimate variances of design-based or model-assisted small area estimators of totals may have the following sources of error:

- They estimate simplified formulas of the variance that do not take into account second order inclusion probabilities.
- They assume that calibrated sampling weights are inverses of inclusion probabilities, when they are in fact sample dependent and therefore random.

In the case of model-based estimators MSE formulas are derived with respect to the model distribution. However, survey sampling statisticians are mainly interested in MSE with respect to the sampling distribution. If the model fits the data well, both types of MSE are usually close enough. In our application to real data, model-based and

jackknife estimators of MSE produce quite close results because the models fit the data acceptably well. Another issue is whether or not to use the sampling weights under the model-based approach, and how to use them.

Explicit formulas to estimate MSEs of model-based small area estimators of totals may have the following sources of error:

- They estimate the MSE with respect to the model distribution when we are interested in the MSE with respect to the sampling distribution.
- They are derived for simple random sampling. Under complex sampling designs, the use of sampling weights is still an unsolved problem.

As a summary we can say that explicit estimators of variances or MSEs are easy to apply, but give unreliable estimates as they are based on assumptions that do not hold in practice. Their use should have an orientate character.

The proposed two-stage bootstrap method generates resamples from the original SLFS sample. The method does not require the generation of bootstrap populations. The idea is that small area estimators in the original sample and in the bootstrap samples have very similar distributions, so that variance of estimators in the original sample could be estimated via Monte Carlo method by using the bootstrap samples. In simple random sampling (nonparametric) the bootstrap method is easy to implement and produce consistent (in an asymptotic sense) variance estimates. However in two-stage sampling this is not at all straightforward and it is quite difficult to check asymptotic properties with respect to PSUs or SSUs.

The bootstrap method needs to generate resamples in the same way that the original sample was generated. Here it is necessary to reproduce all the steps followed with the SLFS sample: extraction of the sample, calibration of weights, consistency of estimators at province level, and so on. However the naïve two-stage bootstrap method produce resamples whose distributions are not close enough to the one of the original sample. The key problem is that resamples are obtained with replacement and the original sample was obtained without replacement. Further research is thus needed to adapt BWR methods to the SLFS. By observing the obtained numerical results we conclude that this method over-estimates the variances of the small area estimators. A positive aspect of the bootstrap method is that variance estimates have a small loss of quality in domains with low sample sizes.

To estimate variances of small area estimators of totals, the naïve two-stage bootstrap method may have the following sources of error:

- Distributions of small area estimators in the original sample and in the resamples are not close enough.
- There exists a tendency to over-estimate variances.
- It is an excessively complex method, which needs a lot of delicate work.

The delete-one-cluster jackknife method generates resamples taking one PSU at a time out of the original sample and by recalibrating the sampling weights. It is a simple and easy method to implement. Main problem of jackknife method is that it works erratically in domains containing very few PSUs in the sample. For those domains this method is unreliable and should not be used.

If we compare the numerical results obtained with the three methods to estimate variances or MSEs, we obtain the following conclusions:

- In domains with large sample sizes, the three methods produce basically the same results.
- The naïve bootstrap method gives higher estimates of the variances than the explicit-formula or jackknife methods, so it seems that our implementation is positively biased.
- Assumptions required to deriving explicit formulas to estimate variances or MSEs do not hold in practice, so their use should have an orientative character.
- The delete-one-cluster jackknife method avoids the theoretical problems of the explicit-formula methods and the difficulty of implementation of the bootstrap method. It works quite well in all the domains except in those with very few sampled PSUs.

Acknowledgements

The authors would like to thank the INE household sampling design unit for their support and helpful comments.

References

Das, K., Jiang, J. and Rao, J. N. K. (2001). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.

Jiang J. and Lahiri P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 18, 199-210.

McCarthy, P. J. and Snowden, C. B.] (1985). The bootstrap and finite population sampling. In Vital and Health Statistics 2-95. Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington.

Miller, R. G. (1964). A trust worthy jackknife. *Annals of Mathematical Statistics*, 35, 1594-1605.

Rao, J. N. K. and Wu, C.-F. J.] (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley.

Rao, J. N. K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistic. Theory and methods*, 33, 2087-2095.

Rust, K. F. and Rao, J. N. K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag.

Shao, J. (2003). Impact of bootstrap on sample surveys. *Statistical Science*, 18, 191-198.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

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.

Quenouille, M. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11, 18-84.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.

You, Y. and Rao J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small-area estimation using survey weights. *Survey Methodology*, 30, 431-439.

Appendix A: Estimated totals of unemployed people in Canary Islands

Table A.1: *Estimated totals of unemployed men (left) and women (right) in Canary Islands with the SLFS 2003/02.*

Area	lfs	dir	greg	ebluba	eblupb	lfs	dir	greg	ebluba	eblupb
1	550	546	1172	1411	1865	3168	3022	3107	1993	2823
2	141	105	206	270	72	187	128	255	297	0
3	526	435	473	236	196	119	85	110	177	141
4	0	0	94	131	100	229	240	180	133	207
5	1131	1183	1047	765	967	467	490	866	892	638
6	458	379	347	432	260	635	464	511	499	527
7	13544	14081	13332	13275	14264	15637	16382	14044	13912	16118
8	319	401	274	724	695	289	350	470	829	107
9	1239	1161	818	1217	1185	1263	968	1260	1630	920
10	369	319	467	395	275	328	276	415	378	345
11	2451	2049	2219	1620	1162	959	760	855	1213	1237
12	2295	2364	2575	2546	1980	1889	2005	3097	3218	2107
13	343	277	527	987	506	787	798	1053	1311	863
14	2548	2006	1638	1656	1833	1791	1414	1381	1589	1515
15	9261	9928	9489	9505	10389	11802	12120	11071	11140	11422
16	507	420	514	564	681	681	614	592	665	746
17	1848	1241	1147	1079	1328	2530	1650	1409	1342	1302
18	496	985	1760	2419	1289	1253	2171	2960	2933	2321
19	966	1809	1650	1158	1022	426	717	818	1194	998
20	5502	5339	4303	3458	3848	5054	4890	4576	3788	4575
21			162	155	184			166	164	334
22	210	251	223	295	226	472	569	321	335	163
23	837	670	981	1095	911	1528	1388	1284	1190	1350
24			311	300	327			310	308	187
25	194	108	103	191	206	203	108	138	176	310
26	1599	1276	1344	1183	1244	446	353	957	1116	1056
27	0	0	159	263	316	545	726	483	269	377

Appendix B: Figures

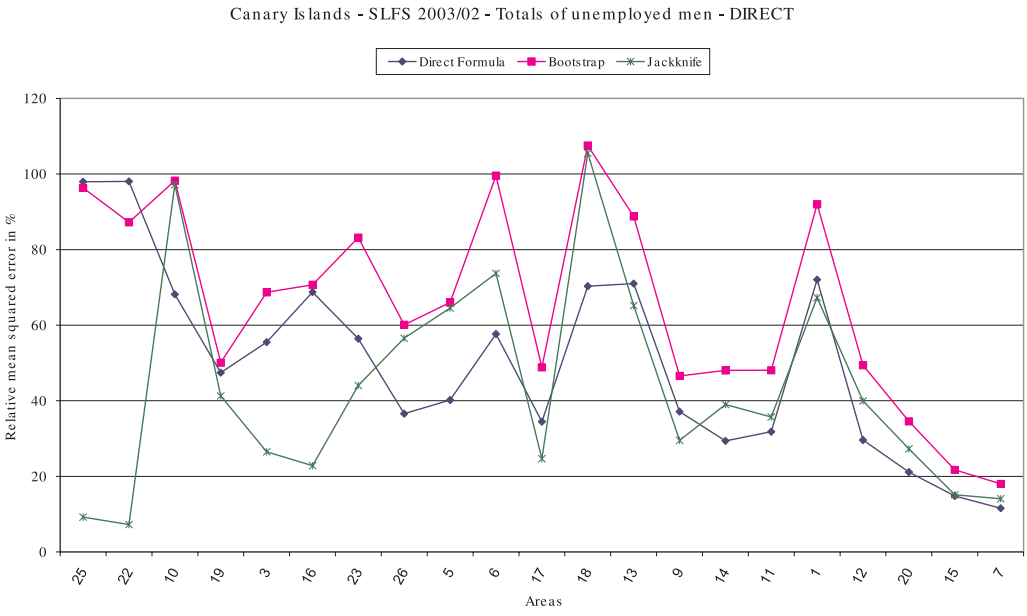

Figure B.1: *Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of direct estimates of totals of unemployed men.*

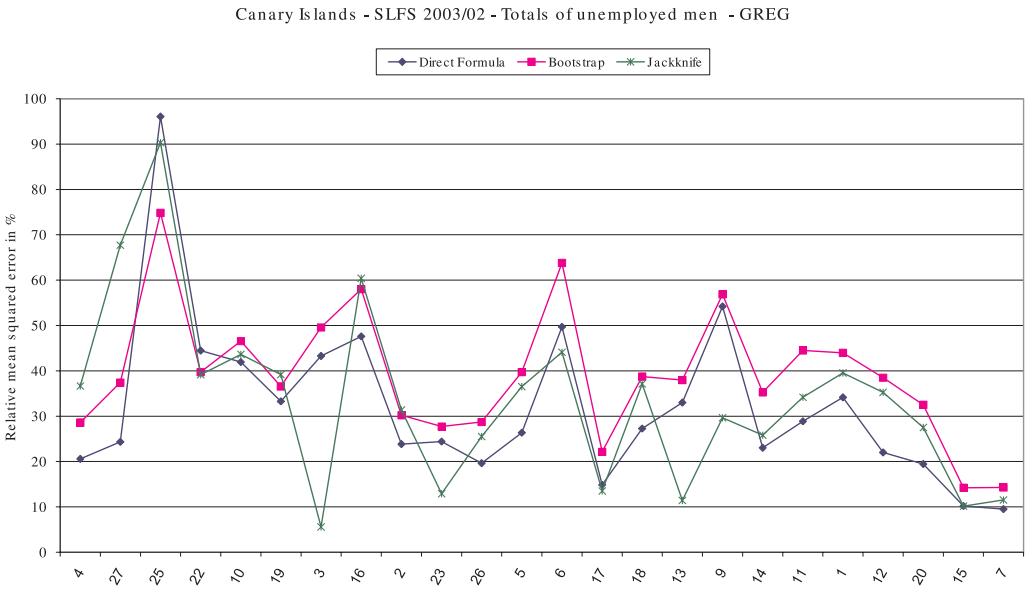

Figure B.2: *Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of GREG estimates of totals of unemployed men.*

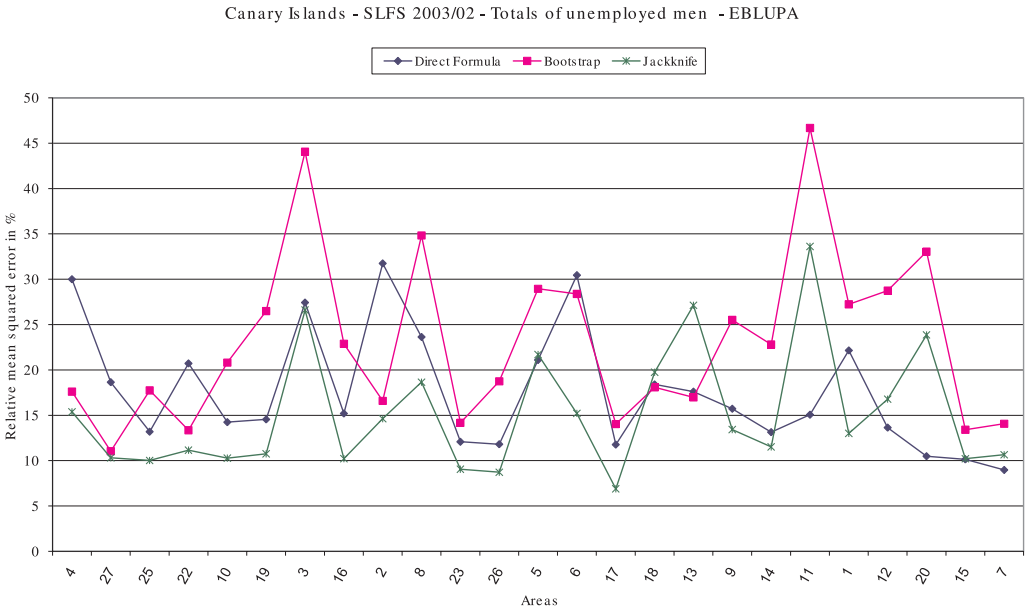

Figure B.3: *Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of EBLUPA estimates of totals of unemployed men.*

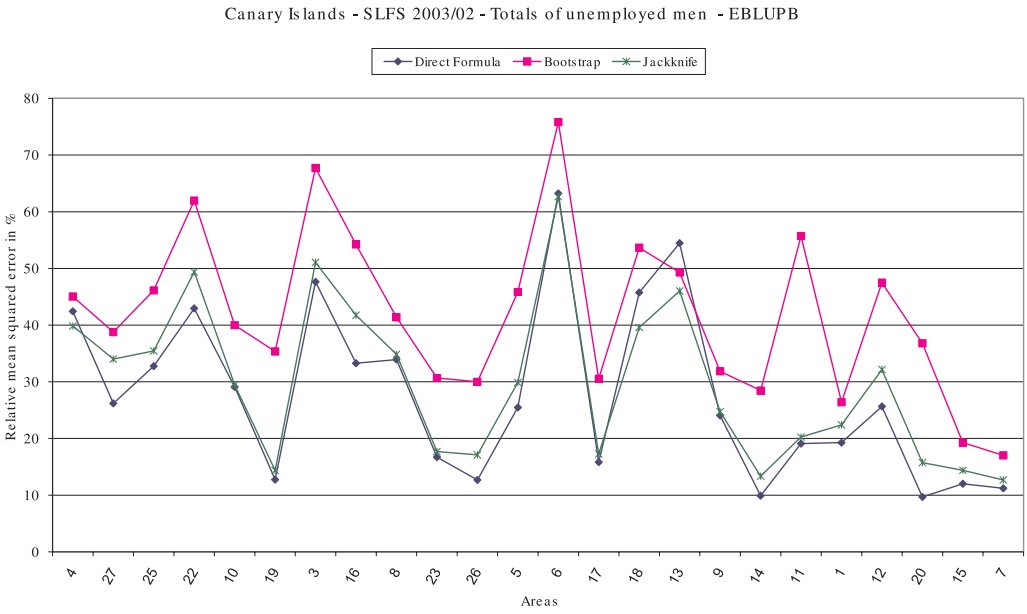

Figure B.4: *Direct-formula, bootstrap and jackknife estimates of coefficients of variation in % of EBLUPB estimates of totals of unemployed men.*

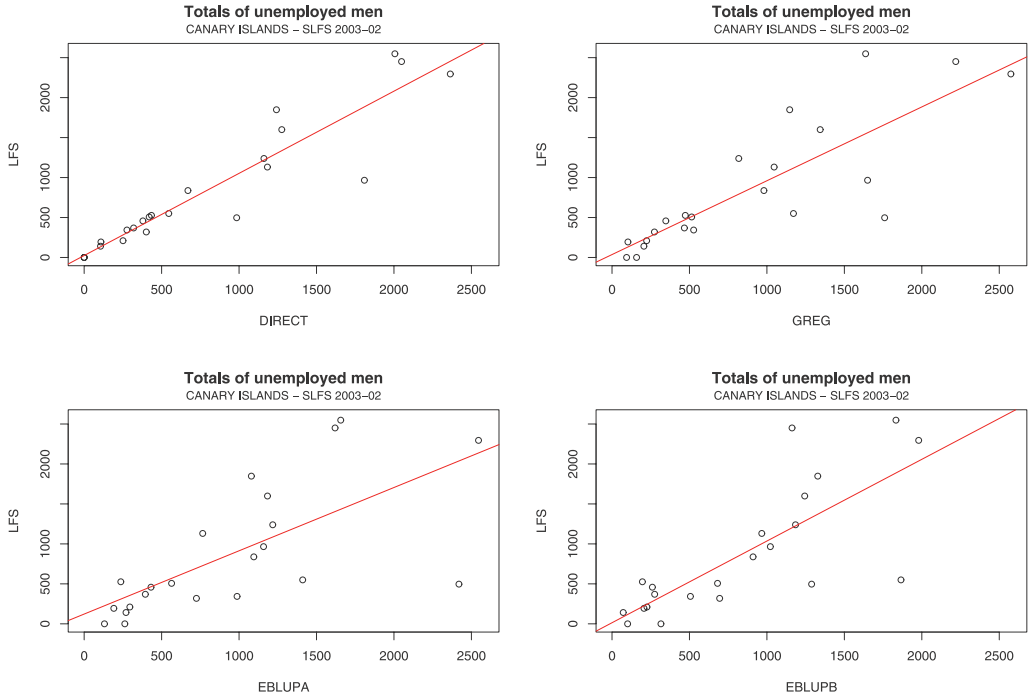

Figure B.5: *Dispersion graphs of LFS versus direct, GREG, EBLUPA and EBLUPB estimates of totals of unemployed men.*