

VARIANCE ESTIMATION OF SMALL AREA ESTIMATORS IN THE SPANISH LABOUR FORCE SURVEY

Francisco Hernández¹, Montserrat Herrador¹, Domingo Morales², M^a Dolores Esteban² and Agustín Pérez².

¹ Instituto Nacional de Estadística (INE)

² Centro de Investigación Operativa, Universidad Miguel Hernández de Elche (UMH)

Summary

- The main goal of this paper is to investigate how to estimate sampling design variances of small area estimators in a complex survey sampling setup.
- For this sake, 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 and rates.
- Three variance estimation methods are considered.
- Some recommendations are given.

The Spanish Labour Force Survey (SLFS)

- This quarterly survey follows a stratified two-stage random sampling design and, for each province, a separate sample is extracted.
- The **Primary Sampling Units (PSUs)** are Census Sections. They are selected with a stratified random design and probabilities proportional to the number of contained dwellings.
- 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.
- Since the first trimester of 2002, weights are 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.

The Spanish Labour Force Survey

- *SLFS estimator* of the total Y_p of variable y in province p , is

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

where weights w_j are the obtained calibrated elevation factors

- Consequently SLFS estimator of the total Y_d of a domain is

$$\hat{Y}_d^{lfs} = \sum_{j \in s_d} w_j y_j.$$

- For provinces $P_p = \bigcup_{d=1}^{D_p} P_{pd}$, it holds $\hat{Y}_p^{lfs} = \sum_{d \in P_p} \hat{Y}_d^{lfs}$; i.e. there exist consistency between SLFS estimates at comarca and province level.

Small area estimators

SLFS estimator of a total is

$$\hat{Y}_d^{lfs} = \sum_{j \in s_d} w_j y_j \cdot$$

SLFS estimator of a mean is

$$\hat{\bar{Y}}_d^{lfs} = \frac{\hat{Y}_d^{lfs}}{\hat{N}_d} = \frac{\sum_{j \in s_d} w_j y_j}{\sum_{j \in s_d} w_j} \cdot$$

DIRECT estimator of a total is

$$\hat{Y}_d^{direct} = N_d \hat{\bar{Y}}_d^{lfs}$$

GREG estimator

- Based on a linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, 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 $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$.
- **GREG estimator of a total is**

$$\hat{Y}_d^{greg} = N_d \hat{Y}_d^{lfs} + N_d (\bar{\mathbf{X}}_d - \hat{\mathbf{X}}_d^{epa}) \hat{\boldsymbol{\beta}}, \quad \text{where } \hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W} \mathbf{y}.$$

EBLUPA estimators

- Model A is the regression synthetic estimator based on the 2-level linear model

$$y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj},$$

where $u_d \sim iid N(0, \sigma_u^2)$ and $e_{dj} \sim iid N(0, \sigma_e^2)$ are independent.

- The model is fitted by ML with Fisher-scoring algorithm.

- **EBLUPA estimator** of total is $\hat{Y}_d^{eblupa} = N_d \hat{\bar{Y}}_d^{eblupa}$, with

$$\hat{\bar{Y}}_d^{eblupa} = \hat{\gamma}_d (\hat{\bar{Y}}_d^{lfs} - \hat{\bar{X}}_d^{lfs} \hat{\beta}) + \bar{X}_d^T \hat{\beta}, \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2 / \hat{N}_d)}.$$

EBLUPB estimators

Model B is the area-level model $\bar{Y}_d = \bar{X}_d^T \beta + u_d$ and $\hat{\bar{Y}}_d^{lfs} = \bar{Y}_d + \varepsilon_d$, where

$u_d \sim iid N(0, \sigma_u^2)$ and $\varepsilon_d \sim iid N(0, \sigma_d^2)$ are independent.

- The model is fitted by the same method as model A.
- **EBLUPB estimator** of total is

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

Explicit-formula variance estimation

DIRECT estimator:
$$\text{var}(\hat{Y}_d^{direct}) = \left(\frac{N_d}{\hat{N}_d} \right)^2 \sum_{j \in s_d} w_j (w_j - 1) (y_j - \hat{Y}_d^{direct})^2$$

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

where
$$g_{dj} = \frac{N_d}{\hat{N}_d} I_{P_d}(j) + N_d (\bar{\mathbf{X}}_d - \hat{\mathbf{X}}_d^{lfs}) \left(\sum_{j \in s} w_j \mathbf{x}_j^t \mathbf{x}_j \right)^{-1} \mathbf{x}_j^t.$$

EBLUP estimators: Mean squared errors are estimated by using g_1 - g_4 explicit formulas given by Prasad and Rao (1990)

Two-stage bootstrap variance estimation

Selection of bootstrap samples in stratum $h=1, \dots, H$, 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.

Two-stage bootstrap variance estimation

Variance estimation is done as follows:

- A.** By using the procedure described above, use sample s to draw B bootstrap samples. For every bootstrap sample calculate $\hat{\theta}_b^*$, $b=1, \dots, B$, in the same way as $\hat{\theta}$ was calculated.
- B.** The observed distributions of $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ 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

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

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

$$cv_B(\hat{\theta}) = \frac{\sqrt{\text{var}_B(\hat{\theta})}}{\hat{\theta}} \cdot 100.$$

Jackknife variance estimation (Shao y Tu (1995))

Jackknife samples are obtained by suppressing a census section, so there are as many jackknife samples as census sections there are in the original sample.

- **Jackknife weight** of individual k in the section i of the stratum-province h in the jackknife sample $s_{(g,j)}^*$ is

$$w_{hik(g,j)} = w_{hik} b_{hi(g,j)},$$

where

$$b_{hi(g,j)} = \begin{cases} \frac{m_g}{m_g - 1} & si \quad h = g, i \neq j, \\ 1 & si \quad h \neq g, \end{cases}$$

and m_g is the number of census sections in stratum-province g .

Jackknife variance estimation

- In each jackknife sample, the weights $w_{hik(g,j)}$ are adjusted to obtain calibrated weights $w_{hik(g,j)}^*$ in the same way as it was done with the SLFS sample. These calibrated weights $w_{hik(g,j)}^*$ are used to calculate $\hat{\theta}_{(g,j)}^*$.
- The observed distributions of $\{\hat{\theta}_{(g,j)}^* : g = 1, \dots, L; j = 1, \dots, m_g\}$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.
- The variance of $\hat{\theta}$ can be approximated by

$$\text{var}_J(\hat{\theta}) = \sum_{g=1}^L \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{\theta}_{(g,j)}^* - \hat{\theta})^2$$

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

$$cv_J(\hat{\theta}) = \frac{\sqrt{\text{var}_J(\hat{\theta})}}{\hat{\theta}} \cdot 100$$

Auxiliary variables

- An *auxiliary variable aggregated at the domain level* without sample counterpart has been used. This variable (GSAU) consists in 12 groups of sex – age – registered as unemployment in the administrative register.

The following *auxiliary variables* have been used at **individual level**:

- **GSA**: groups of sex – age with 6 values,
- **GSAC**: groups of sex - age – employment claimant with 12 values,
- **CLUSTER**: groups of province – bistratum with 4 values.

Variable **BISTRATUM**, used in the definition of CLUSTER, is calculated as follows:

- **BISTRATUM** = 1 if SLFS stratum is 1, 2, 3, 4,
- **BISTRATUM** = 2 otherwise.

Auxiliary variables

Estimation procedures for target variables **EMPLOYED** and **UNEMPLOYED** have used the following auxiliary information:

	GREG	MODEL A	MODEL B
GSA	Employed	Employed	Employed
GSAC	Unemployed	Unemployed	
CLUSTER	Employed, Unemployed	Employed, Unemployed	Employed, Unemployed
GSAU			Unemployed

Auxiliary information for target variables: Unemployed and Employed.

Conclusions on the small area estimators

1. The four considered estimators tend to give the same numerical results as LFS estimator when sample size increases.
2. To estimate totals of unemployed people, the four considered estimators are acceptably unbiased with respect to LFS estimator (as shown by dispersion graphs in Appendix). *EBLUPA* estimator is in general the one with the lowest MSE.
3. To estimate rates of unemployed people, *DIRECT* and *GREG* estimators are the most unbiased ones with respect to LFS estimator (see figures in Appendix). *EBLUPA* estimator is in general the one with the lowest MSE.

Conclusions on the estimation of variances

Explicit formulas to estimate variances of direct 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 sample dependent and therefore random.

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.

Conclusions on the estimation of variances

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

To estimate variances of small area estimators of totals, 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.

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

- It works erratically in domains containing very few PSUs in the sample. For those domains this method is unreliable and should not be used.

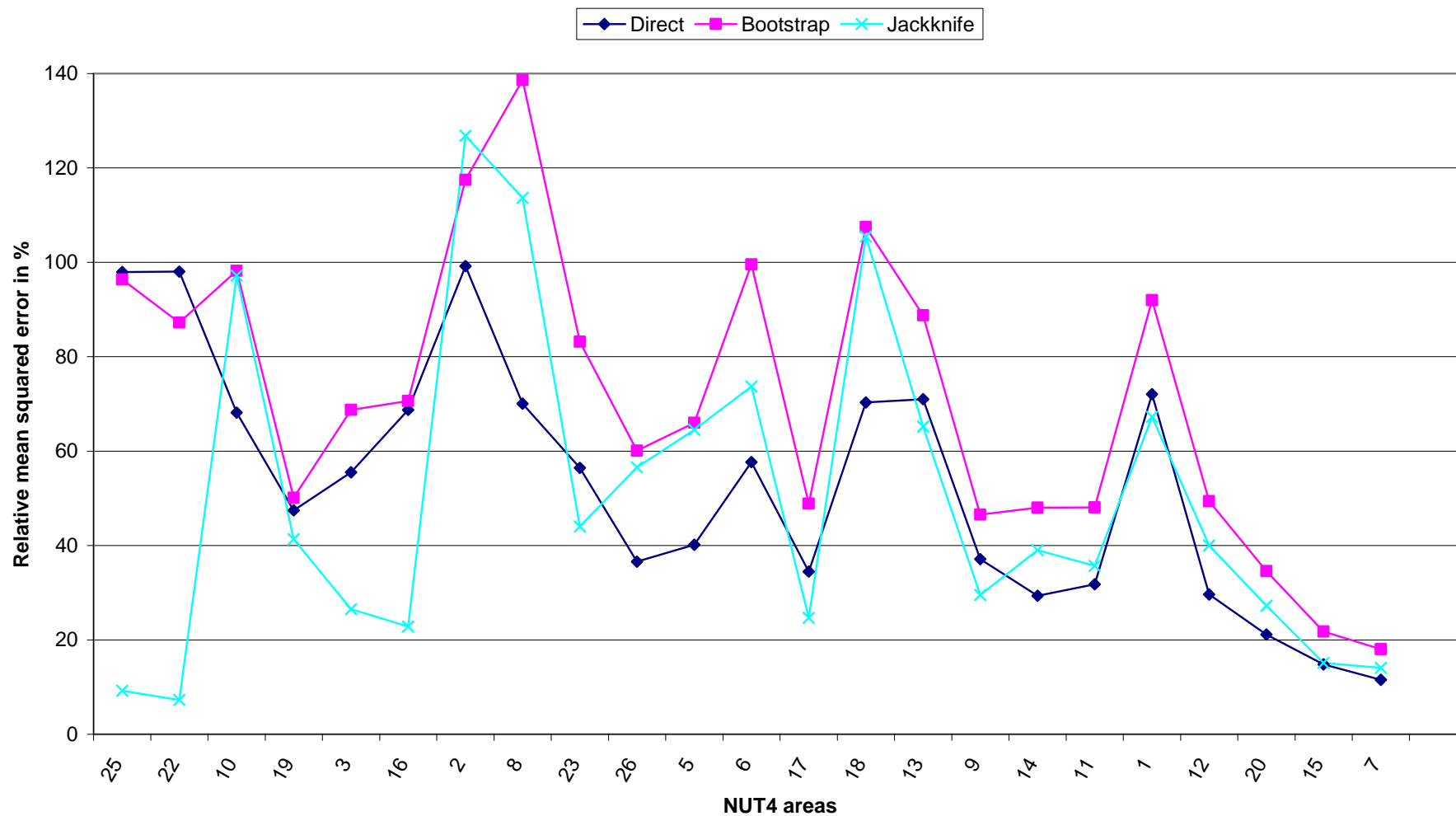
Conclusions on the estimation of variances

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

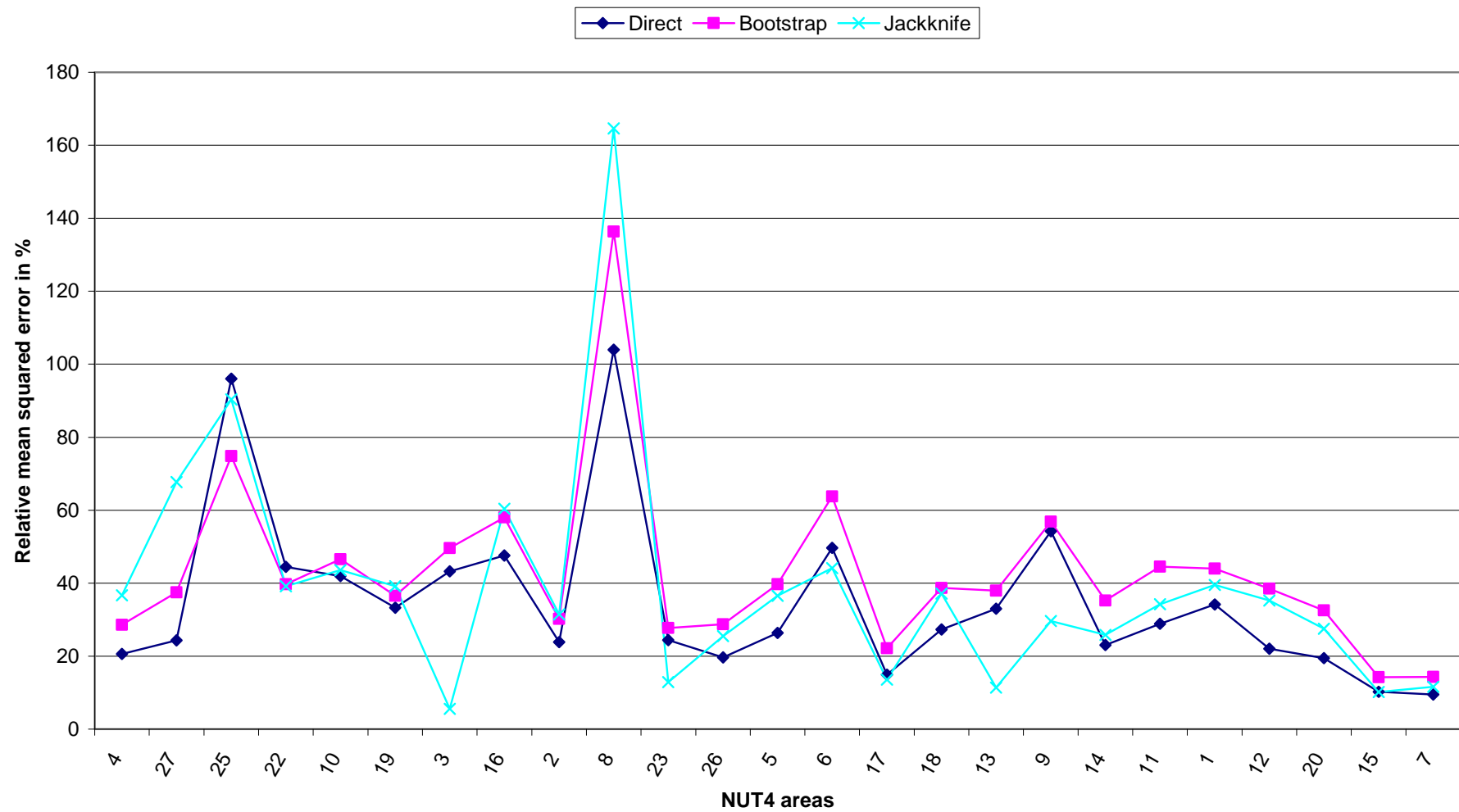
- In domains with large sample size, the three methods produce basically the same results.
- Bootstrap method gives higher estimations of the variances than the explicit-formula or jackknife methods, so it seems that our implementation is positively biased.
- Assumptions required to deriving explicit formulas to estimate variances or MSEs do not hold in practice, so their use should have an orientate character.
- Jackknife method avoids the theoretical problem 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 one with very few sampled PSUs.

Figures

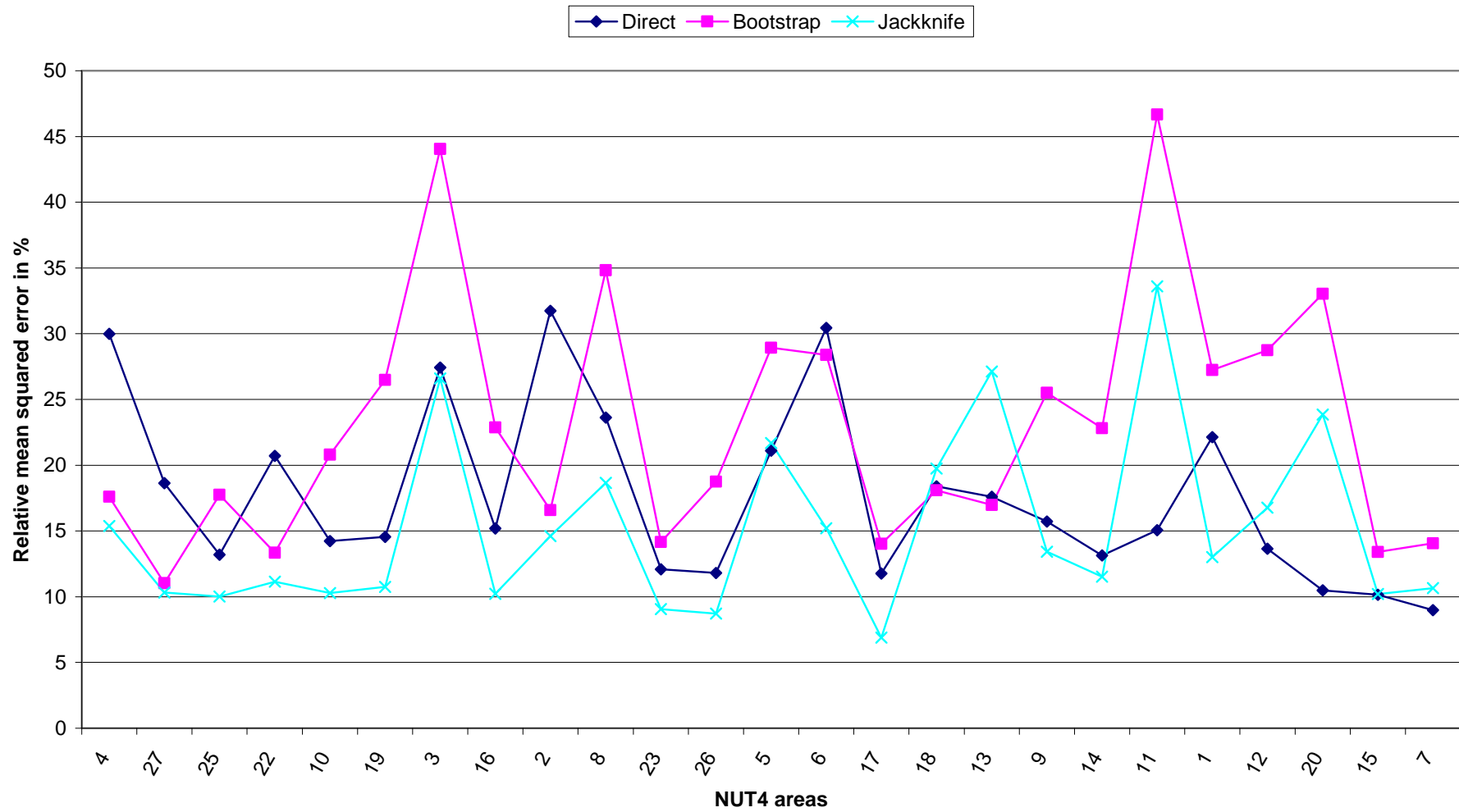
Canary Islands - SLFS 2003/02 - Totals of unemployed men - DIRECT



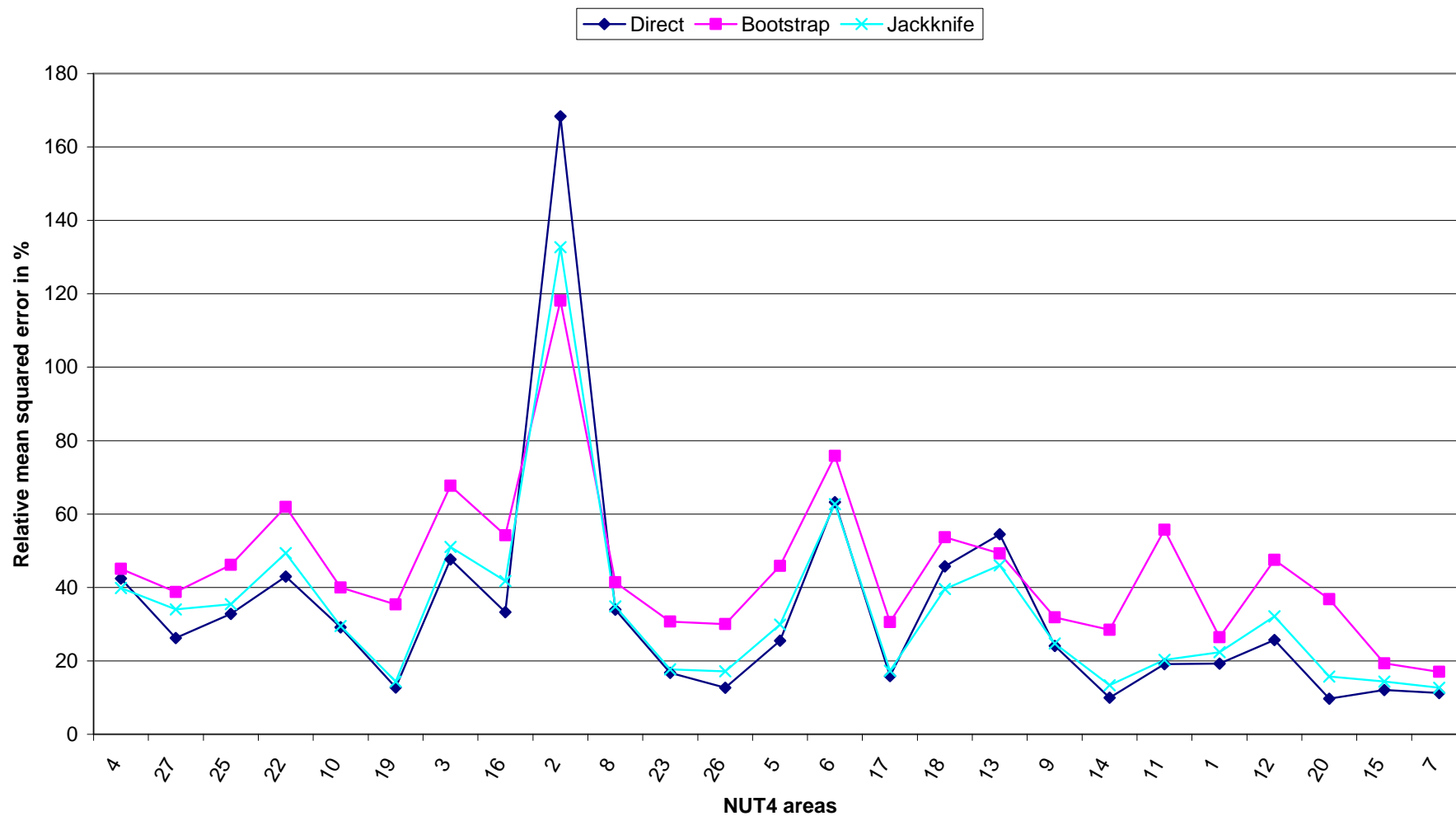
Canary Islands - SLFS 2003/02 - Totals of unemployed men - GREG

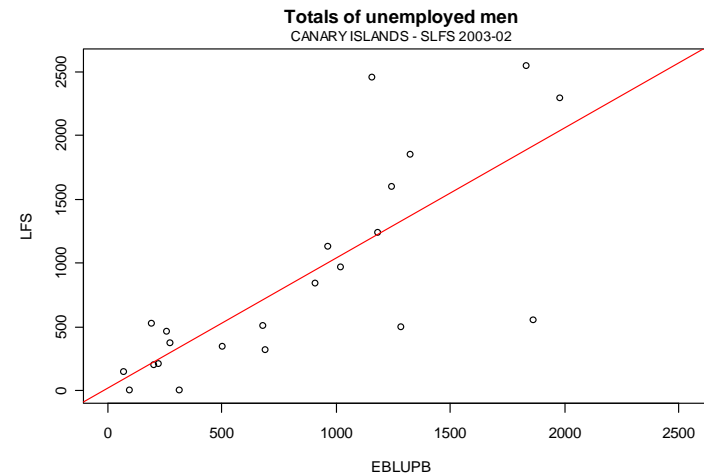
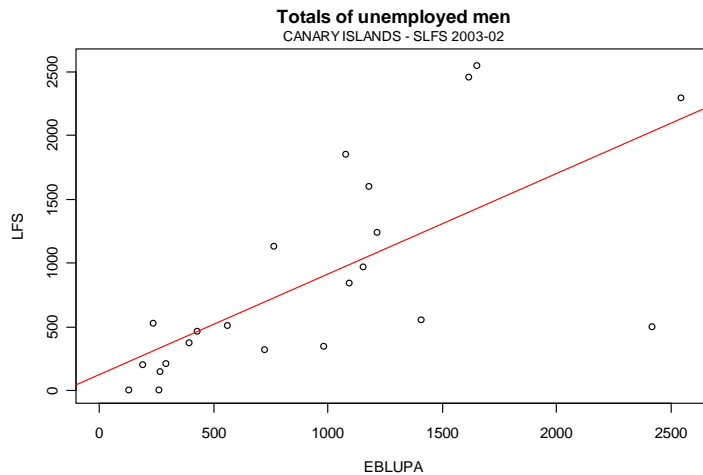
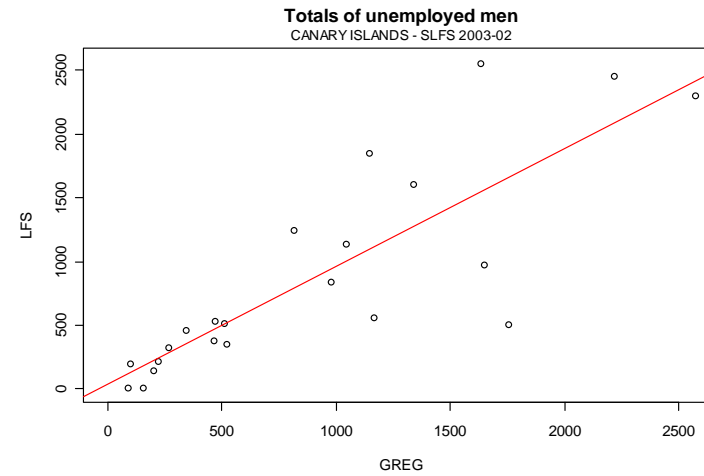
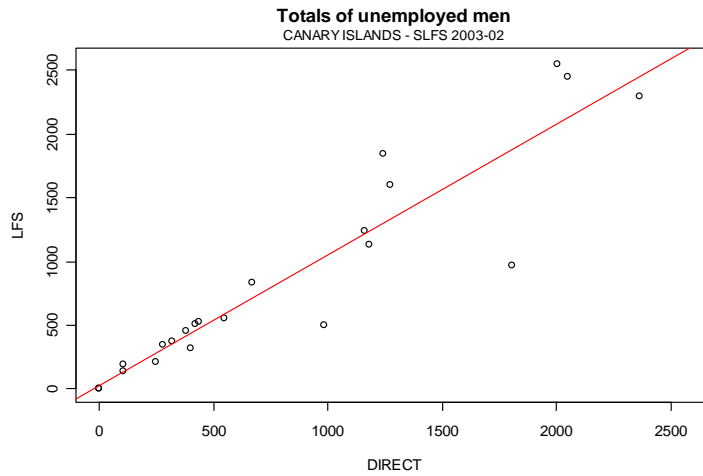


Canary Islands - SLFS 2003/02 - Totals of unemployed men - EBLUPA



Canary Islands - SLFS 2003/02 - Totals of unemployed men - EBLUPB





Dispersion graphs of LFS versus DIRECT, GREG, EBLUPA and EBLUPB estimates of totals of unemployed men.