Estimating Unemployment for Small Areas in Navarra, Spain

Ugarte, M.D., Militino, A.F., and Goicoa, T.
Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra
Campus de Arrosadía, 31006 Pamplona, Spain
E-mail: lola@unavarra.es

Abstract

In the last few years, European countries have shown a deep interest in applying small area techniques to produce reliable estimates at the county level. The EURAREA project (http://www.statistics.gov.uk/eurarea), founded by the European Union between 2000 and 2003, has investigated the performance of various standard and innovative methods in several European countries. However, the specificity of every European country, the variety of auxiliary information as well as its accessibility, makes the use of the same methodology in the whole of Europe a very difficult task. Navarra is a small autonomous community located at the north of Spain. It has 10.000 km^2 and only 600.000 inhabitants, irregularly distributed in seven subdivisions. Navarra Statistical Institute (NSI) has provided data to the Spanish Statistical Institute (INE) as a member of the EURAREA project. Nowadays, NSI is interested in providing precise estimates of the unemployment population in every of its subdivisions (called "comarcas") in the context of the Spanish Labor Force Survey. In this work we review the current estimation procedure used to provide these estimates. In addition, we discuss the behavior of several design-based, model-assisted, and model-based estimators using different auxiliary information, and provide several methods for estimating the prediction error. We comment on the results and the viability of its implementation. More specifically we comment on the difficulties of estimating in very small areas where the samples are both very scarce and unstable.

1 Introduction

The Spanish Labour Force Survey (SLFS) is a quarterly survey of households living at private addresses in Spain. Its purpose is to provide information on the Spanish labour market that can then be used to develop, manage, evaluate and report on labour market policies. It is conducted by the Spanish Statistical Institute (INE). The target population includes all persons aged 16 or more living in private households. Yet there are multiple aims achieved with this survey, the estimation of unemployment population is one of the most relevant. The survey follows a stratified two-stage cluster design and, for each province, a separate sample is designed. The primary sampling units (PSUs) are Census Sections (areas with a maximum of 500 households) that are grouping into h strata according to the size of municipality ($h = 1, 5, 6, 7, 8, 9$). In Navarra, 91 PSUs are selected in the first stage with probability proportional to the number of households. For each PSU selected, a simple random sampling is applied to draw 18 households, inquirying the overall residents of the household aged 16 or more (about 3000 people). This sampling design produces self-weighting samples at stratum level and then, every household has the same probability of being drawn. In this work we check by simulation the benefits of using different kinds of auxiliary information in alternative estimators according to some measures of precision, later we choose the best the prediction error estimator for the chosen estimator. The scenario is the same as the one used in the SLFS survey, but with samples from the 2001 Census.

The rest of the paper is organized as follows. Section 2 presents the proposed estimators: design-based methods, model-assisted and model-based methods for estimation purposes. Section 3 presents the different indicators measures of the prediction error. Section 4 illustrates the performance of the estimators in a simulation study. In Section 5 we provide different estimators of the mean squared error and finally, we show the conclusions.

2 Alternative estimators of the unemployment

The variable of interest is the number of unemployed by small areas in Navarra (Spain), defined according to the International Labour Organization. Navarra is an autonomous community located at the north

Figure 1: Navarra autonomous community located in the north of Spain

of Spain, (see Figure 1) and it has 7 small areas, called ("comarcas"), (see Figure 2). The proposed estimators are design, model-assisted and model-based estimators, detailed in the following subsections.

2.1 Design-based estimators

In the design-based theory, the variable of interest is a fixed quantity and the probability distribution is induced by the sampling design. It is a distribution-free method mainly focused on obtaining estimates for domains with large samples. Direct estimator only use observations coming from the the domain of interest, but indirect estimators take information outside of the domain. The use of auxiliary information ("borrow strength") is a common tool to improve the precision of design-based estimators, and frequently it comes from other domains. In this paper it consists of (E) age-sex groups, with 6 categories combination of age $(16-24, 25-54, >55)$ and sex. (S) Stratum, that represents the size of the living city and takes 9 values: (1) capital of the province, and the rest of the cities or villages depending of their population, (2) between 20000 and 50000 inhabitants, (3) between 10000 and 20000 inhabitants, (4) between 5000 and 10000 inhabitants, (5) between 2000 and 5000 inhabitants and (6) with less than 2000 inhabitants. Let us note that Navarra strata has only 6 categories (1, 5, 6, 7, 8, 9). (N) educational level has two categories (1) for illiterate, primary or secondary school and (2) for technical workers and professionals. (P) previous unemployment status has three categories: (1) occupied or inactive, (2) unemployed and (3) others. (D) claimant of employ that takes the value 1 if he/she is registered in the employment office of Navarra and 0 otherwise.

There are four design-based estimators considered in this paper: a direct, a post-stratified a synthetic and a composite estimator. In the design-based theory unbiasedness and design-consistency are desirable properties pursued by the majority of estimators. An estimator $\hat{Y}$ of Y is design-unbiased if $E[\hat{Y}] = Y$ and it is design-consistent if it is unbiased and its variance tends to zero as the sample size increases (Rao, 2003). The direct estimator does not make use of any auxiliary information but only of data in the domain. It is design-unbiased but its variability is usually big enough to be considered inappropriate in small-area estimation.

The direct estimator of the total unemployment in the d-th small area takes the form

$$\hat{y}_d^{direct} = \hat{\bar{y}}_d N_d = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij} I_d(h,i,j)}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} I_d(h,i,j)} N_d = \frac{\sum_{k=1}^{n_d} w_k y_k}{\sum_{k=1}^{n_d} w_k} N_d,$$

where h is the stratum $(h = 1, 5, 6, 7, 8, 9)$, i the cluster in the h stratum, $(i = 1, 2, \ldots, n_h)$, j every unit of cluster i in the h stratum $(j = 1, 2, \ldots, m_{hi})$, y_{hij} takes the value 1 for the jth unemployed person in stratum h, cluster i, and 0 otherwise, N_d is the total population in the d-th small area $(d = 1, \ldots, D)$,

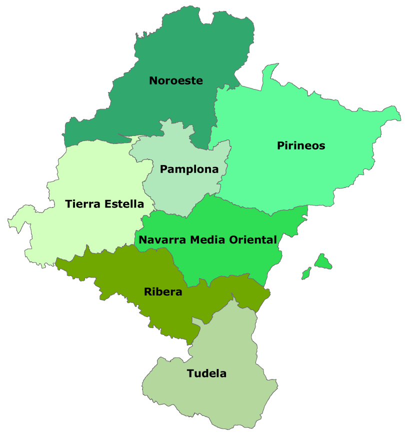

Figure 2: 7 small areas of Navarra

and n_d is the corresponding sample size. The indicator variable $I_d(h,i,j)$ takes the value 1 if person j of cluster i and stratum h is in d and 0 otherwise. The design or sampling weights w_{hij} are the inverse of the inclusion probability but usually they are corrected by non-response effects. They also allow us to incorporate the different sampling plans in the estimation process. The direct estimator does not use any auxiliary variable. However, the rest of design-based estimators one or more auxiliary variables are used to calibrate the final estimation.

The post-stratified estimator of d-th small area incorporates the total of auxiliary variables. It is given by

$$\hat{y}_d^{post} = \sum_g \hat{\bar{y}}_{dg} N_{dg} = \sum_g \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij} I_{dg}(h,i,j)}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} I_{dg}(h,i,j)} N_{dg} = \sum_g \frac{\sum_{k=1}^{n_{dg}} w_k y_k}{\sum_{k=1}^{n_{dg}} w_k} N_{dg} \tag{1}$$

where $g = 1,\dots,G$ define the combination of auxiliary variables and varies from 1 to $6 \times 6 \times 2 \times 3 \times 2$, the indicator variable $I_{dg}(h,i,j)$ takes the value 1 if person j of cluster i and stratum h is in g and and 0 otherwise, and n_{dg} indicates the number of sampled persons in the d-th region belonging to the gth group. The post-stratified estimator may be considered as an assisted-model estimator because it can be derived from a linear model where the predictor variable is the indicator variable of belonging to the gth group.

The synthetic estimator is used for estimating in subareas under the assumption that small areas have the same characteristics as the large area. Synthetic estimators are usually biased. To estimate the total unemployment in the d-th small area the synthetic estimator is written as

$$\hat{y}_d^{synt} = \sum_g \hat{\bar{y}}_g N_{dg} = \sum_g \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij} I_g(h,i,j)}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} I_g(h,i,j)} N_{dg} = \sum_g \frac{\sum_{k=1}^{n_g} w_k y_k}{\sum_{k=1}^{n_g} w_k} N_{dg},$$

where the indicator variable $I_g(h,i,j)$ takes the value 1 if person j of cluster i and stratum h is in g and 0 otherwise.

A natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators. The composite estimators are a linear combination of estimators. In this case composite estimators are defined by a linear combination of a direct estimator and an indirect estimator. It takes the form

$$\hat{y}_d^{comp} = \lambda_d \hat{y}_d^{post} + (1-\lambda_d)\hat{y}_d^{synt}$$

Table 1: Mean of the absolute value of the relative bias (SRAM) and mean of the square root of the relative mean square error (RECMRM) for the post-stratified and synthetic design-based estimators evaluated in the 8 groups of auxiliary variables.

		Design-Based					
		Men		Women		Total	
		SRAM	RECMRM	SRAM	RECMRM	SRAM	RECMRM
Poststratified	E	1.119	47.695	2.022	38.778	1.377	30.964
	ED	6.770	44.306	5.709	36.311	6.152	29.737
	EN	1.521	48.343	3.095	38.445	2.277	30.966
	EP	9.292	44.320	6.178	36.718	7.491	29.855
	ES	2.665	47.862	3.350	39.790	3.039	31.444
	ESD	17.380	45.021	11.367	37.627	13.919	31.656
	ESN	4.562	48.221	5.634	39.322	5.150	31.330
	ESP	17.620	45.703	12.526	38.048	14.688	32.136
Synthetic	E	17.246	22.248	13.585	17.956	13.451	16.811
	ED	12.114	17.867	12.426	16.482	10.746	14.265
	EN	17.869	22.778	13.064	17.679	13.589	16.949
	EP	13.645	18.900	11.526	15.646	10.765	14.171
	ES	6.151	22.312	8.017	18.962	6.022	15.451
	ESD	7.941	22.063	8.471	18.679	6.741	15.283
	ESN	5.419	22.184	8.158	19.150	5.973	15.504
	ESP	7.986	21.791	7.476	18.251	7.534	15.313
Direct		1.043	47.307	1.770	38.789	1.232	30.828

where

$$\lambda_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq \alpha N_d \\ \dfrac{\hat{N}_d}{\alpha N_d} & \text{otherwise} \end{cases}$$

verifying $0 \leq \lambda_d \leq 1$. $\hat{N}_d = \sum_d w_j$ is the estimated total size in region d, and ($\alpha = 2/3, 1, 1.5, 2$). These values provide the names of composite 1, 2, 3 and 4 respectively.

2.2 Model-Assisted estimators

These estimators use regression models a mean to obtain consistent estimators from the design-based point of view (Särndal, Swensson and Wretman, 1989). The most well known model-assisted estimators are the generalized regression estimators (GREG). Here they have been obtained assisted in a linear and a logit model. Here, the linear model is

$$y_{jd} = \mathbf{x}_{jd}^t \boldsymbol{\beta} + \epsilon_{jd}, \quad j = 1, \ldots, n_d, \tag{2}$$

where for every small area d, y_{jd} takes the value 1 if the jth person is unemployed, $\mathbf{x}_{jd} = (x_{id,1}, x_{id,2}, \ldots, x_{id,p})^t$ is the vector of the p auxiliary variables and $\epsilon_{id} \sim N(0, \sigma^2)$. The GREG estimator of the total number of unemployed is given by

$$\hat{Y}_d^{GREG} = N_d \left(\bar{\mathbf{X}}_d \hat{\boldsymbol{\beta}} + \frac{1}{\hat{N}_d} \sum_{j \epsilon n_d} w_{jd} \left(y_{jd} - \mathbf{x}_{jd}^t \hat{\beta} \right) \right)$$

where

$$\bar{\mathbf{X}}_d = \frac{1}{N_d} \sum_{j=1}^{n_d} w_j \mathbf{x}_j^t = \left(\frac{N_{d1}}{N_d}, \frac{N_{d2}}{N_d}, \cdots, \frac{N_{dp}}{N_d} \right) = \left(\bar{X}_{d1}, \bar{X}_{d2}, \ldots, \bar{X}_{dp} \right)^t$$

is the vector of the p auxiliary variable population means, $N_{d1}, \dots, N_{dp}$ are the population of these p auxiliary variables. The $\boldsymbol{\beta}$ coefficients are estimated with observations coming from the overall areas and then

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{n} w_j \mathbf{x}_j \mathbf{x}_j^t \right)^{-1} \sum_{j=1}^{n} w_j \mathbf{x}_j y_j. \tag{3}$$

Assuming that $y_{jd} \sim B(n_d, p_{jd})$, it is more appropriate to be assisted in a logit model given by

$$\mathrm{logit}(p_{jd}) = \log \left(\frac{p_{jd}}{1 - p_{jd}} \right) = \mathbf{x}_{jd}^t \boldsymbol{\beta}. \tag{4}$$

The GREG estimator of the total number of unemployed in the dth area is given by

$$\hat{Y}_d^{GREG} = \sum_{j=1}^{N_d} \frac{\mathrm{e}^{\mathbf{x}_{jd}^t \hat{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_{jd}^t} \hat{\boldsymbol{\beta}}} + \frac{N_d}{\hat{N}_d} \sum_{j \epsilon n_d} w_{jd} \left(y_{jd} - \frac{\mathrm{e}^{\mathbf{x}_{jd}^t \hat{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_{jd}^t \hat{\beta}}} \right).$$

Usually $\boldsymbol{\beta}$ is estimated by iteratively weighted least squares method.

2.3 Model-based estimators

These estimators use regression models for estimation, prediction and inferential purposes. The model-based theory is called prediction theory and considers $y_1, \dots, y_N$ as realizations of the random variables $Y_1, \dots, Y_N$. Splitting the population in sampling observations (s) and non-sampling observations (r), the total of Y, called T, can be expressed as the sum of sampled and non-sampled observations, $T = \sum_{j \in s} y_j + \sum_{j \in r} y_j$. The prediction theory predicts the non-observed variable, and therefore it provides the estimator

$$\hat{T} = \sum_{j \in s} y_j + \sum_{j \in r} \hat{Y}_j.$$

The common predictors of the non-sampling total are linear combinations of y_j and are based on different models. In this paper, linear models and logit models have been considered.
Assuming a linear model given by

$$y_{jd} = \mathbf{x}_{jd}^T \beta + \epsilon_{jd} \quad j = 1, \dots, n_d \quad d = 1, \dots, 7$$

where $\mathbf{x}_{jd} = (x_{jd,1}, x_{jd,2}, \dots, x_{jd,p})^T$ is a vector of p covariates. The estimator of the total number of unemployed based on a linear model is given by

$$\hat{Y}_d = \mathbf{X_d} \hat{\boldsymbol{\beta}} \tag{5}$$

where $\mathbf{X}_d = (X_{d,1}, X_{d,2}, \dots, X_{d,p})^T$ is the total population vector of the p covariates.
(a) The synthetic estimator, called Linear Synthetic, estimates $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j \epsilon s} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j \epsilon s} \mathbf{x}_j y_j.$$

(b) The synthetic estimator based on a weighted linear model, called Linear Synthetic W, assumes that $\epsilon_{jd} \sim N(0, \sigma^2 / w_i)$ and

Table 2: Mean of the absolute value of the relative bias (SRAM) and mean of the square root of the relative mean square error (RECMRM) for the composite design-based estimators evaluated in the 8 groups of auxiliary variables.

		Design-Based					
		Men		Women		Total	
		SRAM	RECMRM	SRAM	RECMRM	SRAM	RECMRM
Composite 1	E	1.258	47.030	1.776	38.104	1.099	30.424
	ED	6.221	43.683	5.386	35.673	5.762	29.199
	EN	1.024	47.743	2.751	37.880	1.921	30.477
	EP	8.762	43.634	5.814	36.106	7.107	29.310
	ES	2.624	47.299	3.148	39.141	2.853	30.944
	ESD	16.956	44.488	11.072	37.024	13.606	31.165
	ESN	4.250	47.727	5.323	38.779	4.900	30.894
	ESP	17.242	45.121	12.175	37.461	14.384	31.646
Composite 2	E	1.447	43.943	1.233	35.799	1.047	28.641
	ED	5.174	40.808	4.442	33.337	4.830	27.327
	EN	1.119	44.657	2.170	35.619	1.300	28.721
	EP	7.558	40.686	4.836	33.801	6.156	27.484
	ES	2.676	44.604	2.577	37.063	2.470	29.371
	ESD	15.836	42.124	9.987	34.887	12.583	29.467
	ESN	3.891	45.158	4.517	36.799	4.306	29.384
	ESP	16.355	42.687	11.127	35.349	13.473	29.983
Composite 3	E	5.668	33.199	4.089	27.074	4.116	21.754
	ED	3.099	30.150	4.088	24.984	3.277	20.125
	EN	5.470	33.718	3.558	26.895	3.644	21.762
	EP	3.579	29.847	2.925	25.081	2.892	20.035
	ES	3.397	35.194	3.105	29.120	2.858	23.042
	ESD	11.620	33.272	6.644	27.448	8.901	23.004
	ESN	3.717	35.656	3.524	29.021	3.654	23.111
	ESP	12.716	33.759	7.731	27.705	9.990	23.398
Composite 4	E	8.562	27.675	6.429	22.551	6.450	18.567
	ED	4.447	24.432	5.398	20.637	4.636	16.604
	EN	8.570	28.084	5.933	22.301	6.121	18.534
	EP	4.565	24.127	4.812	20.425	4.301	16.425
	ES	4.086	30.124	4.333	24.850	3.586	19.795
	ESD	9.062	28.586	5.686	23.541	6.757	19.704
	ESN	4.101	30.460	4.200	24.800	3.860	19.872
	ESP	10.487	28.967	5.810	23.636	7.954	20.004
Direct		1.043	47.307	1.770	38.789	1.232	30.828

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\epsilon s} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j\epsilon s} w_j \mathbf{x}_j y_j.$$

(c) The estimator based on a linear model with a fixed effect of the area, called Linear F, assumes that one of the explanatory variable is a fixed-effect of the area and $\boldsymbol{\beta}$ is estimated as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j\epsilon s} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j\epsilon s} \mathbf{x}_j y_j.$$

(d) The estimator based on a weighted linear model and fixed-effect of the area, called Linear WF, assumes that $\hat{\boldsymbol{\beta}}$ is estimated with weights w_j and

$$\hat{\boldsymbol{\beta}} = \left(\textstyle\sum_{j\epsilon s} w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \textstyle\sum_{j\epsilon s} w_j \mathbf{x}_j y_j$$

Assuming a logit model

$$\log \left(\frac{p_{jd}}{1 - p_{jd}} \right) = \mathbf{x}_{jd}^T \boldsymbol{\beta}$$

where p_{jd} is the probability of being unemployed, $\mathbf{x}_{jd} = (x_{jd,1}, x_{jd,2}, \ldots, x_{jd,p})^T$ is a vector of p covariates. The estimator of the total number of unemployed based on a logit model is given by

$$\hat{Y}_d = \sum_{j=1}^{N_d} \frac{\mathrm{e}^{\mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}}}}{1 + \mathrm{e}^{\mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}}}} \tag{6}$$

(e) The synthetic estimator based on a logit model, called Logit Synthetic, estimates $\boldsymbol{\beta}$ by iteratively weighted least squares. (f) The synthetic estimator based on a weighted logit model, called Logit Synthetic W, assumes that $\hat{\boldsymbol{\beta}}$ is estimated with weights w_j.
(g) The estimator based on a logit model with fixed-effect of the area, called Logit F, assumes that one of the explanatory variables is a fixed-effect of the area.
(h) The estimator based on a weighted logit model with fixed-effect of the area, called Logit WF, assumes that $\hat{\boldsymbol{\beta}}$ is estimated with weights w_i.

3 Indicators measures of the prediction error

We consider the following indicators measures of prediction error: the absolute value of the relative bias (SRA_d) and its mean (SRA) over the D small areas, given by

$$SRA_d(\hat{y}) = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{\hat{y}_d(k) - Y_d}{Y_d} \right| 100, \qquad SRAM(\hat{y}) = \frac{1}{D} \sum_d SRA_d(\hat{y}),$$

and the square root of the relative mean squared error ($RECMR_d$) and its mean ($RECMR$) over the D small areas, given by

$$RECMR_d(\hat{y}) = \left(\frac{1}{K} \sum_{k=1}^{K} \left(\frac{\hat{y}_d(k) - Y_d}{Y_d} \right)^2 \right)^{\frac{1}{2}} 100, \qquad RECMRM(\hat{y}) = \frac{1}{D} \sum_d EMCR_d(\hat{y}).$$

Table 3: Mean of the absolute value of the relative bias (SRAM) and mean of the square root of the relative mean square error (RECMRM) for model-assisted estimators evaluated in the 8 groups of auxiliary variables.

		Model-Assisted					
		Men		Women		Total	
		SRAM	RECMRM	SRAM	RECMRM	SRAM	RECMRM
Linear GREG	E	1.011	46.713	1.733	38.126	1.149	30.236
	ED	0.445	43.395	1.724	36.271	0.865	28.998
	EN	0.960	46.686	1.710	38.151	1.129	30.265
	EP	0.870	42.793	1.347	35.975	0.929	28.608
	ES	0.993	46.192	0.983	37.543	0.582	29.893
	ESD	2.010	42.514	1.346	35.480	0.761	28.387
	ESN	1.050	46.162	0.989	37.565	0.564	29.909
	ESP	1.040	42.094	1.150	35.319	1.319	28.209
Logit GREG	E	1.011	46.713	1.733	38.126	1.149	30.236
	ED	0.442	43.616	1.772	36.356	0.948	29.169
	EN	1.006	46.657	1.714	38.138	1.130	30.261
	EP	0.749	42.962	1.345	36.076	0.983	28.819
	ES	0.940	46.531	1.708	37.946	1.034	30.087
	ESD	1.206	43.253	2.007	35.920	1.499	28.859
	ESN	0.963	46.478	1.702	37.963	1.045	30.103
	ESP	3.459	42.909	2.443	35.847	2.604	28.754
Mixed Logit GREG	E	1.005	46.732	1.706	38.158	1.116	30.244
Direct		1.043	47.307	1.770	38.789	1.232	30.828

4 Simulation Results

We have conducted 500 simulations from the 2001 Census based on the same scenario as the one used in the (Spanish Labour force Survey) SLFS. The aim of these simulations is to choose the best estimator of the unemployed people by small areas between those with less $SRAM$ and $RECMRM$. There are a total of 7 small regions in Navarra for which we have evaluated 4 composite estimators, 3 GREG estimators and 3 model-based estimators. In all of them 8 combinations of auxiliary variables have been used: (E) age-sex, (ED) age-sex-claimant, (EN) age-sex-educational level, (EP) age-sex-previous employment status, (ES) age-sex-stratum, (ESD), age-sex-stratum-claimant, (ESN) age-sex-stratum-educational level and (ESP) age-sex-stratum-previous employment status. There is no optimal estimator with the smallest $SRAM$ and $RECMRM$ simultaneously, but a good trade off between both criteria has bean reached by composite 4 (with $\alpha = 2$) estimator.

Table 1 and 2 show the mean of the absolute value of the relative bias (SRAM) and the mean of the square root of the relative mean square error (RECMRM) of the design-based estimators for the 8 combinations of auxiliary variables. Composite 4 ED ($\alpha = 2$) presents a good balance of bias and mean squared error for men and total. Synthetic ES and ESN are also competitive, but Composite 4 EP ($\alpha = 2$) presents the best balance between bias and mean squared error for men, women and total, because a SRAM less than 4 is considered negligible. Table 3 shows the same indicators for assisted-model estimators. The bias ($SRAM$) of all the estimators, is roughly the same as the direct one, but with the Linear GREG ESP a less RECMRM with regard to the direct one is attained. Table **??** shows again the indicators measures of the precision error for the model-based estimators in the 8 combinations of auxiliary variables. The bias of all the estimators are higher than the direct one but the Logit Synthetic ES estimator presents the smallest REMRM. Summing up, the composite 4 EP estimator is the simplest and more precise small area estimator for the total number of unemployed in Navarra. Table 6 provides the estimate of the number of unemployed with Composite 4 EP (Age-Sex-unemployment population register in the SNE) by small areas. It is evident the good approximation provided by the estimator in all the small areas,

Table 4: Mean of the absolute value of the relative bias (SRAM) and mean of the square root of the relative mean square error (RECMRM) for the linear model-based estimators evaluated in the 8 groups of auxiliary variables.

		Model-Based					
		Men		Women		Total	
		SRAM	RECMRM	SRAM	RECMRM	SRAM	RECMRM
Linear Synthetic	E	18.752	23.589	14.035	18.603	14.604	17.884
	ED	12.897	18.428	12.606	16.765	11.382	14.771
	EN	19.429	24.346	13.234	17.931	14.349	17.736
	EP	14.384	19.521	11.862	15.981	11.286	14.709
	ES	6.182	21.700	8.735	18.506	6.218	15.096
	ESD	7.726	19.985	9.338	17.788	7.367	14.315
	ESN	5.539	21.494	8.615	18.489	6.250	15.117
	ESP	7.961	20.056	7.876	17.090	7.479	14.441
Linear Synthetic W	E	17.246	22.248	13.585	17.956	13.451	16.811
	ED	12.202	17.904	12.458	16.441	10.784	14.308
	EN	17.829	22.841	12.893	17.394	13.206	16.702
	EP	13.466	18.755	11.558	15.648	10.704	14.164
	ES	6.155	21.787	8.653	18.554	6.176	15.146
	ESD	7.708	20.096	9.248	17.831	7.303	14.368
	ESN	5.510	21.580	8.529	18.534	6.215	15.169
	ESP	7.937	20.210	7.794	17.159	7.465	14.517
Linear F	E	3.613	43.794	3.815	33.671	2.215	27.742
	ED	6.817	38.015	5.339	30.544	4.062	24.984
	EN	3.736	43.743	3.857	33.716	2.256	27.757
	EP	5.580	38.522	4.792	30.988	3.552	25.310
	ES	3.645	43.192	4.267	33.555	2.346	27.410
	ESD	7.555	37.607	6.019	30.428	4.462	24.806
	ESN	3.757	43.099	4.292	33.602	2.382	27.424
	ESP	6.034	38.043	5.334	30.944	3.827	25.098
Linear WF	E	2.654	43.702	3.298	33.663	1.801	27.683
	ED	6.041	37.972	4.981	30.546	3.652	24.941
	EN	2.730	43.638	3.362	33.700	1.827	27.693
	EP	4.673	38.475	4.312	30.999	3.018	25.286
	ES	3.390	43.279	4.031	33.645	2.169	27.481
	ESD	7.218	37.702	5.797	30.534	4.206	24.874
	ESN	3.500	43.184	4.062	33.687	2.208	27.493
	ESP	5.686	38.153	5.078	31.029	3.536	25.178
Direct		1.043	47.307	1.770	38.789	1.232	30.828

Table 5: Mean of the absolute value of the relative bias (SRAM) and mean of the square root of the relative mean square error (RECMRM) for the logit model-based estimators evaluated in the 8 groups of auxiliary variables.

		Model-Based					
		Men		Women		Total	
		SRAM	RECMRM	SRAM	RECMRM	SRAM	RECMRM
Logit Synthetic	E	18.752	23.589	14.035	18.603	14.604	17.884
	ED	13.797	19.060	12.448	16.948	12.382	15.339
	EN	19.181	24.073	13.336	18.038	14.422	17.772
	EP	15.536	20.364	12.043	16.316	12.322	15.382
	ES	6.046	22.259	7.971	18.838	5.973	15.340
	ESD	7.340	21.463	8.351	18.315	6.555	14.909
	ESN	5.526	22.067	8.110	18.861	5.997	15.363
	ESP	7.823	21.481	7.319	17.834	7.297	15.068
Logit Synthetic W	E	17.247	22.248	13.585	17.957	13.451	16.811
	ED	12.885	18.471	12.254	16.535	11.570	14.743
	EN	17.598	22.595	12.966	17.493	13.279	16.736
	EP	14.402	19.518	11.712	15.899	11.437	14.697
	ES	6.056	22.253	7.980	18.839	5.982	15.342
	ESD	7.350	21.460	8.351	18.313	6.555	14.909
	ESN	5.527	22.061	8.129	18.864	6.011	15.369
	ESP	7.833	21.485	7.328	17.839	7.299	15.069
Logit F	E	1.410	46.985	1.755	38.242	1.174	30.481
	ED	1.385	44.263	2.003	36.749	1.120	29.434
	EN	1.411	46.948	1.711	38.282	1.152	30.516
	EP	1.917	44.010	1.664	36.414	1.182	29.161
	ES	0.815	46.880	1.717	38.351	1.036	30.377
	ESD	1.168	44.305	1.789	36.843	1.029	29.413
	ESN	0.815	46.820	1.690	38.389	1.022	30.404
	ESP	1.350	44.034	1.477	36.565	0.981	29.153
Logit WF	E	1.059	46.879	1.842	38.271	1.179	30.410
	ED	1.130	44.315	1.829	36.783	0.922	29.432
	EN	1.042	46.835	1.816	38.310	1.162	30.445
	EP	1.636	43.978	1.567	36.472	1.046	29.150
	ES	0.942	47.038	1.738	38.504	1.077	30.469
	ESD	1.893	44.177	2.131	36.824	1.488	29.364
	ESN	0.954	46.979	1.721	38.538	1.066	30.497
	ESP	4.222	43.623	2.583	36.462	2.644	29.034
EB mixed logit	E	28.909	33.919	17.840	23.217	16.279	21.109
Direct		1.043	47.307	1.770	38.789	1.232	30.828

	Number of unemployed								
	Men			Women			Total		
	Census	Composite 4 EP	Direct	Census	Composite 4 EP	Direct	Census	Composite 4 EP	Direct
Pirineo	199	215	196	247	262	235	446	476	434
Navarra Media Oriental	517	511	530	722	692	719	1.239	1.202	1.250
Tierra Estella	570	559	556	884	834	878	1.454	1.394	1.432
Ribera	989	962	989	1.046	1.106	1.036	2.035	2.067	2.027
Noroeste	775	845	772	1.014	1.094	985	1.789	1.938	1.758
Tudela	1.573	1.469	1.576	1.877	1.864	1.855	3.450	3.333	3.432
Pamplona	5.975	5.842	5.962	8.720	8.402	8.604	14.695	14.244	14.569
Navarra	**10.598**	**10.401**	**10.581**	**14.510**	**14.254**	**14.313**	**25.108**	**24.653**	**24.901**

Table 6: Estimate of the number of unemployed with Composite 4 EP (Age-Sex-unemployment population register in the SNE)

even in those with a scarce populations such those of Pirineos and Noroeste.

5 Estimators of the mean squared error

Three methods are presented to calculate the MSE of the composite 4 estimator: two re-sampling methods (jakknife and bootstrap) and the variance linearization method. Jackknife and bootstrap use sub-samples from the original sample. In jackknife method we take as many sub-samples as clusters we have in the sample, because they are obtained leaving out the clusters from the original sample. For every sub-sample new weights are defined and with them the composite 4 estimator is calculated. To obtain the variance and bias of these estimators and therefore its MSE we proceed as is indicated in subsection (5.2), the jackknife is applied to the overall expression of the MSE. In the bootstrap, the sub-samples are obtained by random sampling, but we need to determine how many we need. Analogously for every sub-sample new weights are defined and then the estimator is calculated. The MSE is estimated as detailed in subsection (5.3). The variance linearized method consists of applying the Taylor series as detailed in subsection (5.1).

5.1 Variance Linearization Method

The linearization method or delta method consist of applying a Taylor series to the (function of the total estimators (Woodruff, 1971). Let us define the following indicator variables $I_k(h,i,j) = 1$ if person j of cluster i and stratum h is in group k, $z_{hij} = y_{hij} I_k(h,i,j)$ and $v_{hij} = w_{hij} I_k(h,i,j)$.
Post-stratified and synthetic estimators of the mean θ_d^k can be written as

$$\widehat{\widehat{\theta}}_d^k = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} z_{hij} \right) / v... \text{ , where } \qquad v... = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} \tag{7}$$

If θ_d^k is the post-stratified estimator, k is the group g in the domain d but, when θ_d is the synthetic estimator $k = g$. The linearized estimator of the variance is the following

$$\widehat{\mathrm{Var}}_L(\widehat{\bar{\theta}}_d^{\,k}) = \sum_{h=1}^{H} \widehat{\mathrm{Var}}_h(\widehat{\bar{\theta}}_d^{\,k}) \quad \text{where} \quad \widehat{\mathrm{Var}}_h(\widehat{\bar{\theta}}_d^{\,k}) = \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \left(U_{hi.} - \bar{U}_{h..}\right)^2,$$
$$U_{hi.} = \frac{1}{v_{...}} \sum_{j=1}^{m_{hi}} v_{hij} \left(z_{hij} - \widehat{\bar{\theta}}_d^{\,k}\right) \quad \text{and} \quad \bar{U}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} U_{hi.} \tag{8}$$

The variance of the total $\widehat{\theta}_d^k$ is calculated as

$$\widehat{\mathrm{Var}}_L(\hat{\theta}_d^k) = \sum_k \widehat{\mathrm{Var}}_L(\hat{\bar{\theta}}_d^{\,k}) N_k^2 \tag{9}$$

The synthetic estimator is biased, then we proceed to calculate the bias (Ghosh and Särndal, 2001) given by $\mathrm{Bias}(\hat{y}_d^{sint}) = -\sum_{j=1}^{N_d} \epsilon_j$ and its estimator

$$\widehat{\mathrm{Bias}}(\hat{y}_d^{sint}) = -N_d \frac{1}{n_d} \sum_{j=1}^{n_d} \epsilon_j \quad \text{where} \quad \hat{\epsilon}_j = y_j - \widehat{\bar{y}}_g \tag{10}$$

Therefore

$$\widehat{MSE}_L(\hat{y}_d^{sint}) = \widehat{\mathrm{Var}}_L(\hat{y}_d^{sint}) + \widehat{Bias}^2(\hat{y}_d^{sint})$$

and finally

$$\widehat{MSE}_L(\hat{y}_d^{comp}) = \lambda_d^2 \widehat{MSE}_L\left(\hat{y}_d^{post}\right) + (1-\lambda_d)^2 \widehat{MSE}_L\left(\hat{y}_d^{sint}\right). \tag{11}$$

5.2 Jackknife Estimator

Jackknife method was introduced by Quenouille (1949, 1956) as a method to reduce the bias, and later Tukey (1958) proposed to use it for estimating the variance and confidence intervals. To apply jackknife we need to drop a cluster (post-code section) each time. Let $\hat{\theta}_{d(hi)}^k$ be the estimator $\hat{\theta}_d^k$ obtained from dropping a cluster i from the h stratum. To calculate $\hat{\theta}_{d(hi)}^k$ we define the new weights

$$w_{j(hi)} = \begin{cases} w_j & \text{if } j \text{ is not in the } h \text{ stratum} \\ 0 & \text{if } j \text{ is in cluster } i \text{ of the } h \text{ stratum} \\ \frac{n_h}{n_h-1} w_j & \text{if unit } j \text{ is in the } h \text{ stratum but not in cluster } i \end{cases} \tag{12}$$

The jackknife estimator of the MSE of $\hat{\theta}_d$ estimator can be obtained as

$$\widehat{MSE}_{JK}(\hat{\theta}_d^k) = \sum_{h=1}^{H} \frac{n_h-1}{n_h} \sum_{i=1}^{n_h} [\hat{\theta}_{d(hi)}^k - \hat{\theta}_{d(h.)}^k]^2 \tag{13}$$

where $\hat{\theta}_{d(h.)}^k = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{\theta}_{d(hi)}^k$. The jackknife estimator of the post-stratified estimator is given by

$$\widehat{MSE}_{JK}(\hat{y}_d^{post}) = \sum_{h=1}^{H} \left[\frac{n_h-1}{n_h} \sum_{i=1}^{n_h} [\hat{y}_{d(hi)}^{post} - \hat{y}_{d(h.)}^{post}]^2 \right] \tag{14}$$

where $\hat{y}_{d(h.)}^{post} = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{y}_{d(hi)}^{post}$, and $\hat{y}_{d(hi)}^{post}$ is similar to $\hat{y}_d^{post}$ but substituting v_{hij} by $w_{j(hi)}$ (see expression (12)). The jackknife estimator of the MSE of a synthetic estimator is given by

$$\widehat{MSE}_{JK}(\hat{y}_d^{sint}) = \sum_{h=1}^{H} \left[\frac{n_h-1}{n_h} \sum_{i=1}^{n_h} [\hat{y}_{d(hi)}^{sint} - \hat{y}_{d(h.)}^{sint}]^2 + \left((n_h-1)(\hat{y}_{d(h.)}^{sint} - \hat{y}_d^{sint}) \right)^2 \right], \tag{15}$$

where $\hat{y}^{sint}_{d(h.)} = \frac{1}{n_h}\sum_{i=1}^{n_h} \hat{y}^{sint}_{d(hi)}$, and $\hat{y}^{sint}_{d(hi)}$ is similar to $\hat{y}^{sint}_d$ but substituting v_{hij} by $w_{j(hi)}$.Finally

$$\widehat{MSE}_F(\hat{y}^{comp}_d) = \lambda_d^2 \widehat{MSE}_F\left(\hat{y}^{post}_d\right) + (1-\lambda_d)^2 \widehat{MSE}_F\left(\hat{y}^{sint}_d\right). \tag{16}$$

5.3 Bootstrap Estimator

The re-scaled bootstrap estimator in a stratified random sampling has been provided by Rao and Wu (1988). It assumes the following steps

1) Given the h stratum we have a sample of n_h clusters. From the sample of the h stratum, we draw a sub-sample of $n_h - 1$ clusters by random sampling with replacement

2) For every sub-sample r ($r = 1, 2, \ldots, R$) we redefine the new weight

$$w_{hij}(r) = w_j \frac{n_h}{n_h - 1} m_i(r) \tag{17}$$

where $m_i(r)$ is the number of times that cluster i is chosen in the sub-sample and we calculate $\hat{\theta}^*_r$ using the new weight $w_{hij}(r)$.

3) Repeat steps 1 and 2 R times.

4) To derive the bootstrap estimatorwe calculate

$$\widehat{MSE}_B(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}^*_r - \hat{\theta}\right)^2 \tag{18}$$

The bootstrap estimator of the post-stratified estimator is given by

$$\widehat{MSE}_B(\hat{y}^{post}_d) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{y}^{post(*)}_{d(r)} - \hat{y}^{post}_d\right)^2 \tag{19}$$

where $\hat{y}^{post(*)}_{d(r)}$ is similar to $\hat{y}^{post}_d$ but substituting v_{hij} by $w_{hij}(r)$ detailed in expression (17).
The bootstrap estimator of the synthetic estimator is given by

$$\widehat{MSE}_B(\hat{y}^{sint}_d) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{y}^{sint(*)}_{d(r)} - \hat{y}^{sint}_d\right)^2, \tag{20}$$

where $\hat{y}^{sint(*)}_{d(r)}$ is similar to $\hat{y}^{sint}_d$ but substituting v_{hij} by $w_{hij}(r)$ detailed (17). Finally

$$\widehat{MSE}_B(\hat{y}^{comp}_d) = \lambda_d^2 \widehat{MSE}_B\left(\hat{y}^{post}_d\right) + (1-\lambda_d)^2 \widehat{MSE}_B\left(\hat{y}^{sint}_d\right). \tag{21}$$

We obtain 500 simulations with post-stratified, synthetic and composite 4 estimator using the auxiliary variables: age-sex (E) and unemployed according to the SNE (P). We consider $R = 200, 500, 1000$ and 4000. From small values of R we find different performance of the estimator, but from $R = 1000$ and higher, the performance of the estimators are similar. In figures 3 and 4 we see the all the coefficients of variations obtained for men and women respectively, where

$$\widehat{CV}(\hat{\theta}) = \frac{\sqrt{R\widehat{MSE}(\hat{\theta})}}{\hat{\theta}}$$

We also provide the real coefficient of variation obtained from the Census data. All the methods proposed here tend to overestimate the MSE, particularly when the sample size is small, but in this case the best performance is attained by the jackknife estimator.

Acknowledgments: We thank to the Navarra Statistical Institute for providing the data. Research supported by the Ministry of Science and Education (project MTM2005-00511) and OTRI project (2005 07 019).

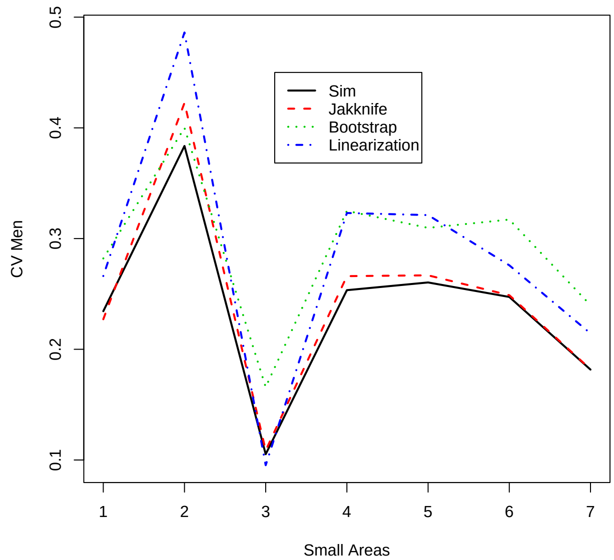

Figure 3: Coefficient of Variation of Composite 4 EP in Men

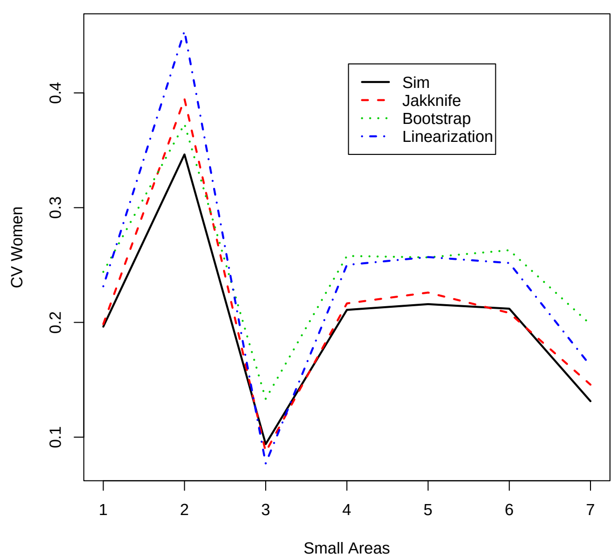

Figure 4: Coefficient of Variation of Composite 4 EP in Women

References

Ghosh, M. and Rao, J.N.K. , (1994), *Statistical Science*, **9**, 55-93.

González, M.E. , (1973). Use and Evaluation of Synthetic Estimates. *Proceedings of the Social Statistics Section* 33-36. American Statistical Association. Washington, D.C.

Militino A. F.,Ugarte, M. D., and Goicoa,T. , (2007), A BLUP Synthetic Versus an EBLUP Estimator: An Empirical Study of a Small Area Estimation Problem . *Journal of Applied Statistics*, **34**, 153-165.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology.

Särndal, C. E., Swensson B. and Wretman, J. H. , (1992). Model assisted survey sampling. Springer.