View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Diposit Digital de Documents de la UAB

SORT 30 (1) January-June 2006, 101-122

Improving small area estimation by combining surveys: new perspectives in regional statistics*

Alex Costa[1], Albert Satorra[2] and Eva Ventura[2]

[1] *Statistical Institute of Catalonia (IDESCAT)* [2] *Universitat Pompeu Fabra*

Abstract

A national survey designed for estimating a specific population quantity is sometimes used for estimation of this quantity also for a small area, such as a province. Budget constraints do not allow a greater sample size for the small area, and so other means of improving estimation have to be devised. We investigate such methods and assess them by a Monte Carlo study. We explore how a complementary survey can be exploited in small area estimation. We use the context of the Spanish Labour Force Survey (EPA) and the Barometer in Spain for our study.

MSC: 62J07, C2J10, 62H12

Keywords: composite estimator, complementary survey, mean squared error, official statistics, regional statistics, small areas

1 Introduction

The 1978 Spanish constitution established that the Spanish Government has the exclusive competence over the statistics that are of interest for the whole country (Article 149.1.31a). At the same time, the statutes of the autonomous communities (regions) state that the regional administrations have the exclusive competence over the statistics that

* The authors are grateful to Xavier López and Maribel Garcia, statisticians at IDESCAT, for their help at several stages of this research. The comments by Nicholas T. Longford on a previous version of this paper are very much appreciated.

Address for correspondence: Albert Satorra and Eva Ventura. Department of Economics and Business. Universitat Pompeu Fabra.

Received: February 2006

Accepted: June 2006

are of interest to the region (e.g., Article 33, 1979, Statute of Catalonia).[1] These laws lead to an interesting overlap of competences, since many surveys and administrative registers are of interest to both the country and the regions.

The country's official statistics bureaux have a much longer-standing tradition and greater resources, and are usually in charge of producing survey-based statistics with a country-wide scope. However, the statistics produced at the country level are sometimes not satisfactory for the region. This may arise for three reasons: 1) an issue that is relevant for the region but not for the country is not reported by the survey; 2) data collected at the country level is not reliable at the regional level; 3) statistics collected at the country level may not provide reliable information for small areas of a region.

For an example of the first case, consider the tourism surveys conducted in Catalonia, for which information on cross-border day trips to France or Andorra is highly relevant. The general questionnaire for the Spanish surveys does not include information about these trips since for the country as a whole these trips are of little importance. An example of the second case is that despite being over-represented in the National Labour Force Survey (EPA), less populous regions, such as Murcia and Navarra, still have too small subsample sizes to make any reliable inferences about them. The third problem is that territorially disaggregated information at the county or municipality levels, or for small islands (Isla de Hierro in the Canary Islands, or Formentera in the Balearic Islands, for example), is very important for regions, but at the overall country level they have a much lower priority.

A regional statistics office could conduct a similar survey, duplicated and improved for its purpose, but that would amount to wasting resources and would increase the burden on respondents. Some subjects (companies, households, and the like) would receive virtually identical survey questionnaires, so they are likely to develop an impression that the national and regional statistical offices do not coordinate their activities. In addition to inducing a negative attitude towards official statistics, duplication of the respondents' costs might be unacceptable.

As an alternative, regional statistical offices may ask the National Statistics Institute (INE) to modify its survey design to meet the regions' needs: to expand the questionnaire to cover issues of regional interest, or to increase the sample size to achieve sufficient precision for the inferences of interest. These changes would not cause problems in sporadically conducted surveys. An example is the Survey of Time Usage conducted in Spain on a single occasion in 2004. INE agreed with some regions to increase the subsample size in their areas. This option may not be available in some ongoing (annual, or quarterly) INE surveys that do not meet the regions' needs due to problems related to reliability or territorial disaggregation. Reasons of technical, legal, or professional nature make modifications of the design of the ongoing surveys problematic. The national offices could not cope with the myriad of requests of various kinds from the

[1] Or the Article 135, 2006, of the recently approved Statute of Catalonia.

regional offices. In this paper, we investigate an analytical solution to these problems that is based on supplementing the country survey with auxiliary information available for some of the small areas of interest.

The use of auxiliary information is not a new idea in small area estimation. When the direct estimator for a particular small area is not satisfactory, one may resort to an indirect estimator. The direct estimator uses only information or data from the area and the variable of interest. Direct estimators are usually unbiased, though they may have large variances. An indirect estimator uses information from the small area of interest as well as from other areas and other variables, or even from other data sources. Indirect estimators are based on implicit or explicit models that incorporate information from other sources. For example, information obtained in a survey can be combined with the one collected in a census or an administrative register. Indirect estimators are usually biased, although their variances are smaller than those of the direct (unbiased) estimators, and the trade-off of bias and variance is usually in their favour.

The novelty of our approach is that we use the information of an auxiliary survey instead of census or administrative records. We combine the information of a country-wide survey, called the reference survey (RS), with the information from a complementary survey conducted by the regional statistics office and tailored to the specific needs of the small area.

A complementary survey (CS) is conducted at the regional level and records variables that correlate with the variables in the RS. CS covers one or several regions of the country, or part of a region. We regard CS as a “light survey” since the data will be faster and cheaper to collect than for the RS. For example, in the case of unemployment, a subject in the CS identifies him or herself as unemployed by the response to a single question. In contrast, the RS follow certain guidelines set forth by the International Labour Organization to classify the subject as unemployed (actively searching for work, available to begin working immediately, and so forth), employed and economically inactive. CS can also simplify the process of contacting the subjects (persons, companies, households, etc.) by using telephone contact systems (Computer Assisted Telephone Interviewing, CATI) or other automated survey methods. So, CS provides results similar to those of RS at a much lower cost; however, as CS records the values of a slightly different variable than RS, its results are biased. This is the price for the less elaborate questionnaire, with looser wording.

We differentiate three types of CS:

1) A general complementary survey (GCS) covers all the regions of the country at one or several points in time. With data from many areas we can remove the bias of GCS estimators relative to RS. One example of GCS is the Economically Active Population Survey (EPA) conducted by INE as RS, and the Barometer of Spain conducted by the Centre for Sociological Research (CIS) as GCS. In the Barometer, respondents are asked if they are unemployed. Information from the EPA and CIS is available for all the Spanish regions for several years.

2) With a regional complementary survey (RCS) we can assess the bias at the regional level, but not at the small area level because there are no data to compare RS and RCS at the small area level. An example is the Survey on Information and Communication Technologies in Catalonia, conducted by the Statistical Institute of Catalonia (IDESCAT) (RCS) by means of CATI, complementing an equivalent RS conducted by INE. RS has a clustered sampling design in which the 41 counties of Catalonia are not well represented. In contrast, the design of the RCS ensures an even coverage of the counties. In this example, the bias of RCS cannot be disaggregated to counties.

3) A local complementary survey (LCS) is a survey conducted in a specific small area. The bias is unknown since RS does not produce valid results for this small area. One example is a survey similar to the EPA in a single small municipality in Catalonia which has very sparse or no representation in EPA.

The bias of the survey relative to RS can be explicitly modeled in GCS but not in RCS or LCS. We investigate how information from CS can be integrated with RS for making inferences about small areas. We consider the specific context in which EPA is RS and the Barometer is CS, in this case, a GCS. The Barometer contains a few questions regarding the subject's employment status which are at face value highly correlated with the corresponding variable in EPA.

The accuracy in small area estimation can be increased by: *a)* increasing the sample size in the area of interest; *b)* borrowing strength from neighbouring areas (using indirect or composite estimators); *c)* borrowing strength from CS, especially when the variables in RS and CS are highly correlated. We explore all these alternatives, with emphasis on combining the options b) and c). The performance of the estimators and the contribution of the complementary information will be assessed by simulation.

Parallel work on the use of CS has been conducted by Costa *et al.* (2006), who study the Survey on the Uses of Information and Communication Technologies in Catalan households; INE conducts a country-wide survey while IDESCAT is in charge of the RCS.

The present paper is organized as follows. Section 2 reviews established small area estimators, with emphasis on estimating labour statistics. Section 3 describes the specific context of estimating rates of unemployment in Spain. Section 4 assesses the performance of the alternative small area estimators by simulations. Section 5 summarizes the main findings of the paper.

2 Estimators for small areas

In this section we consider a two-stage clustered sampling design, motivated by the sampling design of EPA that is considered later in the paper. We consider a binary variable Y that takes the values $Y_{ij} = 1$ if the characteristic under study is present for subject ij, and $Y_{ij} = 0$ otherwise. Here i ($i = 1, 2, \ldots, n$) and j ($j = 1, 2, \ldots, m_i$) denote *primary sample unit* (PSU) and *secondary sampling unit* (SSU), respectively. We use the convention that capitals (X, Y) denote population values and lowercases (x, y) sample values. Their indexing is implied; that is, in X_{ij} we use population indexing and in x_{ij} we use sample indexing. For every sample, we have a variable W of sampling weights, with w_{ij} representing the sampling weight of subject ij.

The population is divided in K small areas, indexed by $k = 1, 2, \ldots, K$. We use the notation $Y_{k.ij}$ for the values of variable Y on units of area k. For sampling data, the symbol + in the subscript denotes the weighted summation over the sample; for example, $y_{k,i+} = \sum_{j=1}^{m_i} w_{k,ij} y_{k,ij}$. For population data, the symbol + indicates summation without weighting.

Our target is the population ratio $\theta_k = Y_{k,+}/X_{k,+}$ of two totals, for each area k. We consider also the overall population ratio $\theta = Y_+/X_+$. It is assumed that the denominator is positive. Several estimators are considered.

2.1 Direct estimator

A direct estimator of θ_k uses only data from area k. It is defined as

$$\hat{\theta}_k = \frac{\hat{Y}_k}{\hat{X}_k} ,$$

where $\hat{Y}_k = y_{k+}$ and $\hat{X}_k = x_{k+}$. Here the summation extends only over the (say, n_k) PSUs that intersect with area k (we assume $n_k > 2$). Straightforward application of the delta-method yields the following estimator of variance $\mathrm{V}(\hat{\theta}_k)$

$$\hat{V}(\hat{\theta}_k) = \frac{1}{\hat{X}_k^2} \left\{ \hat{V}(\hat{X}_k) - 2\hat{\theta}_k \,\widehat{\mathrm{cov}}(\hat{Y}_k, \hat{X}_k) + \hat{\theta}_k^2 \,\hat{V}(\hat{X}_k) \right\}, \tag{1}$$

where

$$\widehat{\mathrm{cov}}(\hat{Y}_k, \hat{X}_k) = \frac{n_k}{n_k - 1} \sum_{i=1}^{n_k} (z_{k,i}{}^{(y)} - \overline{z}^{(y)})(z_{k,i}{}^{(x)} - \overline{z}^{(x)}), \tag{2}$$

$$z_{k,i}{}^{(y)} = y_{k,i+} \quad \text{and} \quad \overline{z}^{(y)} = n_k^{-1} \sum_{i=1}^{n_h} z_{k,i}{}^{(y)},$$

and similarly for x. We compute $\hat{V}(\hat{X}_k)$ and $\hat{V}(\hat{Y}_k)$ as $\widehat{\mathrm{cov}}(\hat{X}_k, \hat{X}_k)$ and $\widehat{\mathrm{cov}}(\hat{Y}_k, \hat{Y}_k)$, respectively.

In the general case of L strata, the sample values are $y_{h,ij}$, were h denotes strata. The direct estimator of the overall population ratio $\theta = Y_+/X_+$ is $\hat{\theta} = \hat{Y}/\hat{X}$, where $\hat{Y} = y_{+,++}$

and $\hat{X} = x_{+,++}$ (summation over strata, PSUs within the strata, and units within the PSU). The estimator of var $(\hat{\theta})$ is like (1) and (2) with subscript k suppressed and a summation over L strata added to the right hand side of (2).

Information on population totals of some auxiliary variables would allow us to calculate post-stratified or ratio-estimators. This will not be pursued in this study. For more information on those estimators, the reader can consult Rao (2003), Ghosh and Rao (1994) or Mancho (2002). López (2000) considers some of these estimators in the context of small area estimation for EPA in Canary Islands. We consider $\hat{\theta}_k$ and $\hat{\theta}$ as the only direct estimators in this study.

2.2 Small area estimators without auxiliary information

An indirect estimator of θ_k uses data from outside area k. As an alternative to $\hat{\theta}_k$ we may adopt the overall-country direct estimator $\hat{\theta} = \hat{Y}/\hat{X}$ for every area k. This is an indirect estimator. Being based on much more data than $\hat{\theta}_k$, $\hat{\theta}$ has a much smaller variance than $\hat{\theta}_k$, but is biased for θ_k, unless the θ_k's are all equal. In this case, $\hat{\theta}$ is much more efficient than $\hat{\theta}_k$. But if the θ_k's vary substantially across areas, the bias of $\hat{\theta}$ will be large, and so will be its mean squared error (MSE). An attractive alternative estimator to both $\hat{\theta}_k$ and $\hat{\theta}$ is the composite estimator $\hat{\theta}_k^{(\mathrm{c})}$ defined as the convex combination

$$\hat{\theta}_k^{(\mathrm{c})} = \phi_k \hat{\theta} + (1 - \phi_k)\hat{\theta}_k \tag{3}$$

with $0 \le \phi_k \le 1$. The coefficient ϕ_k is chosen so as to minimize the MSE and is equal to

$$\phi_k = \frac{\operatorname{var}(\hat{\theta}) - \operatorname{cov}(\hat{\theta}_k, \hat{\theta})}{(\theta_k - \theta)^2 + \operatorname{var}(\hat{\theta}_k) + \operatorname{var}(\hat{\theta}) - 2\operatorname{cov}(\hat{\theta}_k, \hat{\theta})} \,. \tag{4}$$

The denominator of (4) is positive; in fact, it is equal to $E\{(\hat{\theta}_k - \hat{\theta})^2\}$. Clearly, ϕ_k depends on some unknown parameters, and itself has to be estimated. Since

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \hat{Y}_k}{\hat{X}} = \frac{\sum_{k=1}^{K} \hat{X}_k \hat{\theta}_k}{\hat{X}} = \sum_{k=1}^{K} q_k \hat{\theta}_k,$$

where $q_k = \hat{X}_k/\hat{X}$, cov $(\hat{\theta}_k, \hat{\theta}) = \sigma_k^2 q_k$, so the optimal weight is

$$\phi_k = \frac{\sigma_k^2(1 - q_k)}{(\theta_k - \theta)^2 + \sigma_k^2(1 - 2q_k) + \sigma^2} \,, \tag{5}$$

where σ_k^2 and σ^2 are the respective sampling variances of $\hat{\theta}_k$ and $\hat{\theta}$ (see Longford, 1999).

When q_k is very small and the survey is large (e.g., the number K of small areas is large and the sample sizes of most of them are small), we can ignore both q_k and the variance σ^2; then, ϕ_k is approximated by

$$\phi_k = \frac{\sigma_k^2}{(\theta_k - \theta)^2 + \sigma_k^2} .$$

We could estimate σ_k^2 from the sample data from area k and use $(\hat{\theta}_k - \hat{\theta})^2$ as an estimator of the denominator in (5). Our experience, shows that this results in a very unstable estimator of ϕ_k (see Costa, Satorra and Ventura 2003, 2004). One way to overcome this difficulty is by averaging the estimators $\hat{\sigma}_k$'s among several areas (or several variables). For example, Purcell and Kish (1979) use a weight common to all areas that minimizes the between-area average of the mean squared errors. By assuming that the within-area variances of Y are equal, the pooled estimator of their common variance is

$$\hat{\sigma}_w^2 = \frac{1}{n-K} \sum_{k=1}^{K} (n-1)\hat{\sigma}_k^2 . \tag{6}$$

By using the following estimator of the square of the bias

$$b^2 = \frac{1}{K} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2, \tag{7}$$

and ignoring both q_k and the var $(\hat{\theta})$, we estimate the weight as

$$\hat{\phi}_k = \frac{\hat{\sigma}_w^2}{\hat{\sigma}_w^2 + b^2} . \tag{8}$$

Expressions (3) and (8) define the *classic composite* estimator (Costa, Satorra and Ventura, 2003).

2.3 Auxiliary information

In the literature on small area estimation, we find many indirect estimators that incorporate auxiliary information. Usually this information consists of data from a census or an administrative register. Rao (2003) describes the regression synthetic estimator, which combines direct estimators with those obtained from census or records, encompassing a large variety of estimators, such as ratio or count estimators.

We use auxiliary information that arises from a CS carried out in each of several small areas. In particular, we assume that for each of a set of small areas we have the direct estimator $\hat{\theta}_k$ as well as an estimator $\hat{\delta}_k$ derived from CS. For these estimators, consider the simple regression equation

$$\hat{\theta}_k = \alpha + \beta\hat{\delta}_k + \epsilon_k, \tag{9}$$

$k = 1, 2, \ldots K$, and the *fitted estimator* of θ_k by the OLS regression,

$$\hat{\theta}_k^F = \hat{\alpha} + \hat{\beta}\hat{\delta}_k \, .$$

When historical data are available for the RS and CS across several areas, the fitted estimator $\hat{\theta}_k^F$ could be based on more advanced regression than just simple OLS. If we have RS and CS at several time points ($t = 1, \ldots, T$) and several areas ($k = 1, 2, \ldots, K$), we could estimate θ_k by an analysis of covariance model. The regression could also involve other covariates. As more variables are incorporated into the regression that links $\hat{\theta}_k$ with $\hat{\delta}_k$, the synthetic estimator $\hat{\theta}_k^F$ will be more efficient for θ_k, although using too many covariates may inflate the sampling variance. For simplicity, we only consider the OLS regression (9). In the Monte Carlo set-up of Section 4, however, we also involve an estimator that is based on a covariate (fixed-effects, FE) regression model that serves as a benchmark for maximum information attainable from CS.

Even though the variance of $\hat{\theta}_k^F$ may be substantially smaller than var $(\hat{\theta}_k)$, $\hat{\theta}_k^F$ may be biased. We improve both $\hat{\theta}_k$ and $\hat{\theta}_k^F$ estimators by considering the composite estimator

$$\hat{\theta}_k^{(c)}(CS) = \phi_k\hat{\theta}^F + (1 - \phi_k)\hat{\theta}_k \tag{10}$$

where

$$\phi_k = \frac{\text{var}\,(\hat{\theta}_k) - \text{cov}\,(\hat{\theta}_k^F, \hat{\theta})}{\Delta_k^2 + \text{var}\,(\hat{\theta}_k) - \ \text{var}\,(\hat{\theta}_k^F)} \tag{11}$$

Here $\Delta_k = \theta_k - E(\hat{\theta}_k^F)$ has to be estimated. If α and β were known, var $(\hat{\theta}_k^F) = \beta^2$ (var $\hat{\delta}_k$). When the regression parameters are estimated, then

$$\text{var}\,\hat{\theta}_k^F = E\left(\text{var}\,\hat{\theta}_k^F \mid \hat{\Theta}\right) + \text{var}\,\left(E(\hat{\theta}_k^F \mid \hat{\Theta})\right),$$

where $\hat{\Theta} = (\hat{\alpha}, \hat{\beta})$ stands for the vector of estimated regression coefficients. Since the expected value of $(\hat{\theta}_k - \hat{\theta}_k^F)^2$ coincides with the denominator in (11), the weight of (11) is estimated as

$$\hat{\phi}_k = \frac{\hat{\sigma}_k^2 - \tilde{\sigma}_k^2}{(\hat{\theta}_k - \hat{\theta}_k^F)^2},$$

where $\hat{\sigma}_k^2$ and $\tilde{\sigma}_k^2$ are the respective estimators of the variances of $\hat{\theta}_k$ and $\hat{\theta}_k^F$. An alternative estimator of ϕ_k is more stable,

$$\hat{\phi}_k = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + b^2}, \tag{12}$$

where $\hat{\sigma}^2$ is given by (6) and

$$b^2 = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_k^F)^2. \tag{13}$$

The estimator $\hat{\theta}_k^c(CS)$ defined by equations (10), (12) and (13) is called the composite complementary survey (CCS) estimator based on OLS regression (or CCS-OLS).

In Section 4, we assess by Monte Carlo the efficiency of the CCS, direct, indirect and classic composite small area estimators. The efficiency of the CCS estimator will be compared against a direct estimator based solely on RS but with sample size increased by r%, with $r = 10, 20, 50, 100$. For the Monte Carlo study we construct an artificial population that resembles Spain in some aspects related to the labour force. The EPA and Barometer Surveys are used for this construction.

3 EPA and Barometer surveys

This section describes general aspects of estimation of unemployment rates in Spain, both at country and area levels. The main source of information about unemployment in Spain is the EPA conducted by the INE, our RS. The Barometer is our CS.

EPA is a quarterly survey that uses a lengthly questionnaire, panel design and face-to-face interview (paper and pencil, PAPI); in contrast, the Barometer is a monthly survey that uses CATI as the interviewing system and an entirely new sample each month. EPA is designed for reliable estimation of several labour market statistics, including the unemployment rate at country level. The Barometer uses the self-perceived labour status of the interviewed individuals as a proxy for unemployment status.

While all the provinces of Spain are represented in the EPA, this is not always the case for the Barometer. Even if a province is represented in the survey, its sample size may be very small. To have a more informative study, we grouped the 50 Spanish provinces (plus the two autonomous cities in the north of Africa) into 25 areas, according to their geographical proximity and similarity of their labour markets. With this grouping, each area is represented by at least 140 observations in CS. The second and the sixth column of Table 1 shows the sample sizes of EPA and CIS surveys across the 25 areas for the fourth quarter of 2003. To give the same time frame to both surveys, the quarterly estimate by Barometer is taken to be the average across three months of the monthly unemployment rates of the Barometer.

While EPA follows the standard International Labour Organization methodology, the Barometer simply asks the subjects how they perceive their employment status. The unemployment rates in the two surveys differ for two reasons: different definitions of the target variables and different samples (sampling error).

Table 1 shows the direct estimates of unemployment rates for the 25 areas for men, women and for all adults. This distinction is important because in Spain the labour market participation of women is lower than of men, so the properties of the estimators for these two categories may differ. The table shows that the estimates of the unemployment rates differ considerably between the two surveys; see, for example, Asturias and Navarra-Rioja. The sample correlation of the sets of estimates are 0.54, 0.66 and 0.26, for the total (T), men (M) and women (W), respectively. Such large correlations supports the use of the Barometer as complementary information for the EPA.

Table 1: *Sample sizes and unemployment rates (in %) for the EPA and CIS surveys (fourth quarter of 2003; Ss = Sample size, T = Total, M = Men, W = Women).*

	EPA survey				CIS survey			
	Ss	T	M	W	Ss	T	M	W
Almería-Granada (AGR)	4 864	15.65	10.22	23.38	324	13.96	12.92	16.30
Málaga(MAL)	3 441	17.33	13.68	23.08	235	18.59	11.90	29.13
Cádiz-Huelva (CHU)	5 287	23.51	18.37	31.97	246	34.24	32.35	38.98
Córdoba-Jaén (CJA)	6 640	18.56	12.83	28.27	237	20.19	16.53	24.37
Sevilla (SEV)	6 411	17.19	12.80	24.01	248	16.74	11.47	25.29
Aragón (ARA)	6 589	6.20	3.71	9.94	232	11.58	7.60	18.26
Asturias (AST)	4 522	10.03	7.00	14.35	218	23.20	12.31	36.59
Baleares (BAL)	3 539	9.38	8.50	10.59	141	13.38	6.41	24.65
Canarias (CAN)	7 748	12.10	9.37	16.10	297	18.37	9.20	33.08
Cantabria (CNT)	3 578	10.32	8.06	13.77	102	22.52	11.50	38.62
Albacete-C. Real (ACR)	4 971	9.28	4.80	16.59	150	28.14	18.89	45.28
Cuenca-Guad.-Tol. (CGT)	6 253	9.89	5.58	17.44	173	12.86	7.44	22.12
Castilla-León (CLE)	15 143	10.91	6.09	18.37	491	11.85	9.50	16.80
Barcelona (BCN)	7 448	9.54	7.37	12.47	919	13.69	11.44	16.44
Gerona-Lérida-Tarr. (GLT)	7 721	7.00	5.09	9.62	259	10.90	5.55	18.04
Alicante-Castellón (ACS)	6 405	10.38	8.16	13.67	345	20.04	18.07	23.28
Valencia (VAL)	5 858	10.08	7.20	14.20	393	20.84	12.85	31.55
Extremadura (EXT)	6 167	17.11	12.51	24.75	201	22.06	13.41	40.97
La Coruña (LCO)	3 472	15.44	10.14	22.17	241	21.11	15.35	29.19
Lugo-Orense-Pont. (LOP)	6 921	11.99	7.73	17.55	293	18.26	15.98	21.83
Madrid (MAD)	7 765	7.00	5.47	9.08	966	15.62	10.51	21.20
Murcia (MUR)	4 043	10.49	7.09	15.87	198	14.03	12.16	18.44
Navarra-Rioja (NRI)	5 362	5.99	4.36	8.45	151	16.81	6.29	30.56
Álava-Guipúzcoa (AGU)	4 489	7.06	5.48	9.28	194	15.21	6.82	26.89
Vizcaya (VIZ)	3 037	11.43	10.06	13.28	212	19.30	17.21	21.47

Using historical data of the EPA and the Barometer (CIS) surveys of unemployment rates, we fit – for men, women, and all adults, in turn – the following regression with area-fixed effects:

$$\text{EPA}\%_{kt} = \alpha + u_k + \beta\,\text{CIS}\%_{kt} + \epsilon_{kt}, \tag{14}$$

where $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots K$ denote the quarter and the small area respectively. Here u_k is a fixed effect and α and β are the intercept and slope, respectively. The model is estimated using data from the first quarter of 2001 to the fourth quarter of 2003. The monthly data of the Barometer has been averaged over quarters. Table 3 shows parameter estimates with standard errors and t-values, as well as the corresponding R^2 fit measures. The area-effects and the estimated u_k's based on this model are reported in Table 4. These regression estimates and area-effects will be used in the Monte Carlo study to obtain a benchmark estimator (CCS-FE) for unemployment rates that combines RS and CS surveys.

4 Monte Carlo study

In this section, we describe the Monte Carlo study that evaluates the performance of the small area estimators described in Section 2. Based on an EPA sample, we consider an artificial population that resembles that of Spain. We undertake estimation of unemployment rates of total adult population, or just male and female. We expect that a small area estimator that uses auxiliary information will outperform the estimators based only on RS. The gains, however, may diminish when the subsample size of RS in the area of interest is large.

4.1 Design of the simulations

We adopt the sample of the EPA survey in Oct.-Dec. 2003 (see Table 1) as our population. We estimate the unemployment rates of men, women and all adults in this population. The 25 target areas are listed in Table 1. In the simulations, CIS has approximately 2550 monthly observations (7650 observations per quarter). In each area we have a small RS sample and typically a large CS sample.

For simplicity, and to focus on the comparison of the estimators, we apply to the adopted EPA population stratified (by area) simple random sampling (with replacement) proportional to the area size. The sample sizes range from 2500 to 25000, but are fixed within replications. The CS sample is drawn from the realized CIS sample (sample size 7650) treated as the population. The sample size is fixed at 7650. Unemployment rates for men and women are estimated from the total sample by considering just the men and women respectively. Knowing the population values, the MSE's can be estimated with high precision governed by the number of replications.

We use the following RS sample sizes: $5000, 10000, 12500, 25000$, so that the average within-area subsample sizes are $200, 400, 500$ and 1000. Further, the sample for the direct estimator is boosted by $10, 25, 50$ and 100%, but the subsample added is not used in evaluating the other estimators.

For small RS sample sizes, we expect the CCS estimator to be more efficient than the direct estimator, even with a substantially boosted sample size. The specific value of r for which the efficiency of both estimators is similar is likely to vary with the sample size of the RS sample.

The Monte Carlo simulations comprise replications of the following steps.

A) we draw a sample of size m from the population ($N = 147000$). Similarly, we draw a CIS sample. We evaluate the direct, indirect, classic composite, and the composite estimator based on the CS-based composite estimators.

B) We boost the sample size by $r = 10, 25, 50, 100\%$, by drawing an additional subsample from EPA and we evaluate the direct estimator.

Steps A) and B) are replicated 1000 times.

Table 2 summarizes the settings of sample sizes across areas. Our targets are the unemployment rates in the 25 areas, for men, women and all the working force. For each pair of samples, EPA and CIS, we compute the direct, indirect and classic composite, and and two CS-based composite estimators.

Table 2: *Monte Carlo study: average sample sizes across small areas, and sample size increase (in %) of the direct estimator.*

Average Sample size	Sample size increase (%)			
	10	25	50	100
100	110	125	150	200
200	220	250	300	400
400	440	500	600	800
500	550	625	750	1000
1000	1100	1250	1500	2000

For each estimator $\tilde{\theta}$ and area k we evaluate the relative root mean square error as

$$\text{RRMSE}(\tilde{\theta}, k) = \frac{\sqrt{\sum_{j=1}^{1000} \left(\tilde{\theta}_k{}^{(j)} - \theta_k\right)^2 /1000}}{\theta_k},$$

where $\tilde{\theta}_k{}^{(j)}$ is the jth replicate of a specific small area estimator of θ_k. Smallest RRMSE is preferred. The average, median and maximum value of RRMSE across areas is recorded for each estimator.

4.2 Incorporating auxiliary information

Denote the RS and CS direct estimators of the target θ_k by $\hat{\theta}_k$ and $\hat{\delta}_k$, respectively. Because the concepts measured in RS and CS differ slightly, $\hat{\delta}_k$ is likely to be biased. In the fitted estimator $\hat{\theta}_k^F$ we use both $\hat{\theta}_k$ and $\hat{\delta}_k$.

In each replication, we fit the OLS regression (9) with the unemployment rates of EPA and CS as the respective $\hat{\theta}_k$ and $\hat{\delta}_k$ estimators, and compute the OLS fitted $\hat{\theta}_k^F$(OLS). Estimators $\hat{\theta}_k^F$(OLS) and $\hat{\theta}_k$ are then combined to obtain the CS-based CCS-OLS estimator. In parallel, we compute the fitted estimator $\hat{\theta}_k^F$(FE) that is based on the regression coefficients and area effects of the FE regression reported in Section 3. Since this estimator uses more information than the others we can expect it to be more efficient. The FE model is taken as a benchmark model for the information of CS on RS, since among all the regression alternatives that we could think of, many will use more information than the simple OLS model, but less than the FE model.

Two CCS estimators are considered, the CCS-OLS which uses the $\hat{\theta}_k^F$(OLS) obtained from the OLS regression, and the CCS-FE which uses the $\hat{\theta}_k^F$(FE) of the FE regression. Only CCS-OLS is feasible in applications, since CCS-FE uses information that will not generally be available. While the CCS-OLS does fit the OLS regression in each replication, the CCS-FE is based on just one regression fit common to the whole Monte Carlo study. Table 3 reports the regression coefficients used in obtaining the CCS-FE. The area fix-effects that arise from the FE regression are reported in Table 4. The alternative of a random-effect regression model was considered. No substantial change of the performance of the CCS estimator was observed.

***Table 3**: Parameter estimates, se and t-values of the fixed effect regression, for total* (T), *men* (M) *and women* (W) *unemployment rates. Various R^2s are reported: overall (R^2), within (R^2_w) and between (R^2_b).*

	T			M			W		
Model par.	$\hat{\beta}$	se ($\hat{\beta}$)	t-test	$\hat{\beta}$	se ($\hat{\beta}$)	t-test	$\hat{\beta}$	se ($\hat{\beta}$).	t-test
β	0.06	0.02	3.22	0.02	0.02	1.34	0.05	0.02	2.82
α	10.42	0.34	31.08	7.81	0.20	39.81	15.70	0.43	36.25
R^2_w	0.04			0.01			0.03		
R^2_b	0.69			0.70			0.66		
R^2	0.41			0.32			0.35		

Table 4: *Estimated fixed area effects. (*T = *Total,* M = *Men,* W = *Women).*

	Unemployment rate		
Small area	T	M	W
Almería-Granada	3.68	2.65	5.47
Málaga	3.69	4.22	3.56
Cádiz-Huelva	11.08	8.47	16.40
Córdoba-Jaén	7.77	5.65	12.77
Sevilla	7.98	6.80	10.28
Aragón	–5.63	–4.43	–7.83
Asturias	–2.14	–1.44	–3.07
Baleares	–3.68	–2.10	–6.65
Canarias	–0.75	0.24	–1.99
Cantabria	–1.49	–1.29	–2.07
Albacete-Ciudad Real	–2.68	–2.53	–2.16
Cuenca-Guadalajara-Toledo	–1.23	–2.48	1.50
Castilla-León	–0.79	–1.76	0.65
Barcelona	–1.44	–0.47	–3.53
Gerona-Lérida-Tarragona	–4.13	–3.21	–6.36
Alicante-Castellón	–1.95	–1.26	–3.40
Valencia	–0.72	–0.17	–1.81
Extremadura	5.45	4.09	8.28
La Coruña	0.99	0.78	0.74
Lugo-Orense-Pontevedra	0.03	–0.55	0.04
Madrid	–4.05	–2.93	–6.42
Murcia	–0.35	–0.46	–0.25
Navarra-Rioja	–5.74	–4.39	–8.24
Álava-Guipúzcoa	–4.20	–3.35	–5.93
Vizcaya	0.28	–0.09	0.01

5 Results

Tables 5-7 show results of the simulations for the whole labour force, men and women respectively. The tables have identical layouts giving summaries of RRMSEs for the different estimators (columns) and sample sizes (blocks of rows). For each estimator and sample size, the average, mean and maximum RRMSE across the 25 areas is given. The column at the extreme right contains the RRMSE summaries for the benchmark CCS-FE.

***Table 5**: Estimation of total unemployment rates. For different sample sizes, the table shows the RRMSE average (across areas) of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	OLS$^{(2)}$	FE$^{(+)}$
	Average sample size **100**								
Average	419	403	376	345	295	326	323	307	251
Median	415	401	373	347	294	237	312	307	254
Max	578	571	549	501	411	1332	562	550	353
	Average sample size **200**								
Average	298	285	269	245	210	316	251	240	188
Median	298	285	272	239	208	226	236	238	190
Max	421	410	377	350	292	1322	428	454	255
	Average sample size **400**								
Average	210	203	191	172	151	311	191	184	142
Median	211	191	186	164	156	219	179	182	143
Max	299	286	286	242	218	1313	303	335	193
	Average sample size **500**								
Average	190	181	171	158	137	310	174	170	132
Median	188	178	171	157	137	221	167	169	134
Max	264	243	239	229	212	1318	268	305	179
	Average sample size **1000**								
Average	138	132	124	115	100	308	131	129	105
Median	137	132	120	113	097	220	129	127	108
Max	211	209	209	205	193	1319	188	210	140

(i) These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

Of course, the RRMSEs are reduced for the direct estimator when the sample size is boosted. Boosting with $r = 0\%$ is the direct estimator $\hat{\theta}_k$. The efficiency of the composite estimator $\hat{\theta}_k^{(c)}$ is comparable with boosting of the sample by about 50% for small sample size (100), but much less (about 10%) for large sample (1000). That is, the composite estimator $\hat{\theta}_k^{(c)}$ is much more effective for small sample sizes than for large ones; its effectiveness, over the direct estimator, decreases with sample size.

The composite estimator CCS-OLS that makes use of the CS is only slightly more efficient than $\hat{\theta}_k^{(c)}$ for all sample sizes, but the maximum over the areas is slightly higher in most cases. The benchmark CCS-FE estimator (last column of the table), is more efficient than all the other estimators. In fact, its performance is comparable to boosting of the sample by 100%.

***Table 6**: Estimation of unemployment rates for men. For different sample sizes, the table shows the RRMSE average (across areas) of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	OLS$^{(2)}$	FE$^{(+)}$
Average sample size **100**									
Average	669	638	601	546	469	404	478	487	388
Median	645	603	571	525	469	313	447	462	374
Max	915	892	852	776	685	1526	795	803	529
Average sample size **200**									
Average	475	450	423	385	330	385	371	374	281
Median	470	429	409	372	328	300	347	339	272
Max	663	631	590	558	473	1503	643	658	384
Average sample size **400**									
Average	335	316	300	270	238	373	285	287	205
Median	325	299	292	263	231	300	266	264	198
Max	492	454	438	390	344	1484	462	495	294
Average sample size **500**									
Average	298	283	268	247	213	372	259	263	185
Median	294	267	266	233	206	297	250	244	186
Max	411	392	380	347	299	1493	428	454	245
Average sample size **1000**									
Average	212	204	193	178	153	368	196	199	141
Median	206	195	185	170	150	296	194	190	144
Max	292	286	275	243	214	1488	311	331	184

$^{(i)}$ These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

Similar conclusions are arrived at by inspecting the results for men and women. The RRMSEs for women tend to be larger than for men for a fixed sample size, because their rates or unemployment are higher than for men and RRMSEs are approximately proportional to $\sqrt{p/(1-p)}$, where p is the unemployment rate.

The tables contain a lot of detail that is difficult to digest and do not indicate the performance of the estimators for the individual areas. Figures 1 and 2 display RRMSEs of four small area estimators: direct, marked as 0; indirect, I; composite, 1; and CCS-OLS, 2. It shows also the benchmark estimator, marked as +; and the direct estimator with sample size boosted by 50%, marked as b. We regard estimators 0,1,2 and I as feasible because they use information that would normally be available. Estimators + and b are a benchmark and comparator, respectively. They use information that would not be available in practice. At the outset, I is discarded as competitor of 0, 1 and 2.

Table 7: *Estimation of unemployment rates for women. For different sample sizes, the table shows the average (across areas) of RRMSE of the various estimators evaluated in the Monte Carlo study. The estimators are: RS-based estimators (direct with sample size boosted by r%; indirect, $\hat{\theta}$; composite, $\hat{\theta}_k^{(c)}$), and the CCS estimators based on fixed effects (FE) and simple (OLS) regression. All the values of RRMSE have been multiplied by* 1000.

	r%							CS-based	
Summary	$0^{(0)}$	10	25	$50^{(b)}$	100	$\hat{\theta}^{(I)}$	$\hat{\theta}_k^{(c)(1)}$	$\text{OLS}^{(2)}$	$\text{FE}^{(+)}$
	Average sample size **100**								
Average	539	514	476	437	378	315	386	365	327
Median	537	512	475	445	376	215	383	345	335
Max	757	736	716	648	540	1164	618	621	443
	Average sample size **200**								
Average	376	361	338	309	267	304	292	273	235
Median	383	368	327	305	266	200	279	265	240
Max	556	521	493	444	377	1155	485	517	331
	Average sample size **400**								
Average	269	259	241	218	190	298	228	210	180
Median	262	261	237	211	192	192	217	210	182
Max	419	374	374	314	269	1147	393	410	267
	Average sample size **500**								
Average	238	230	215	197	172	296	206	194	163
Median	234	228	211	194	170	191	198	195	161
Max	352	324	314	283	250	1151	333	381	232
	Average sample size **1000**								
Average	172	166	155	143	126	293	157	151	129
Median	168	164	152	136	123	190	153	155	130
Max	252	241	238	229	218	1153	239	284	192

[i] These superscripts show the symbols used in Figures 1 and 2 to represent the RRMSEs.

The areas are ordered according to their unemployment rates. The diagrams show, for instance, that estimator I has serious weaknesses although it is the most efficient for a few areas in the middle of the range. In general, the composite estimators 1 and 2 are the most efficient among the feasible estimators.

In Figure 1 we have a graphical representation of the RRMSE for the alternative estimators in the case of large sample size (1000). Small areas are on the vertical axis and different symbols represent the different estimators. RRMSE is on the horizontal axis, so that efficiency corresponds to being on the left. For this large sample case, the indirect estimator (I) is generally very inefficient: each RRMSE summary has been truncated at 0.25, except for areas whose unemployment rate is close to the national rate. The composite estimators, the RS-based (1) and CS-based (2) are the most efficient estimators, after the benchmark estimator CCS-FE (+). The composite estimators 1 and

2 (RS and CS based, respectively) are generally as efficient as b, direct estimator with sample size boosted by 50%.

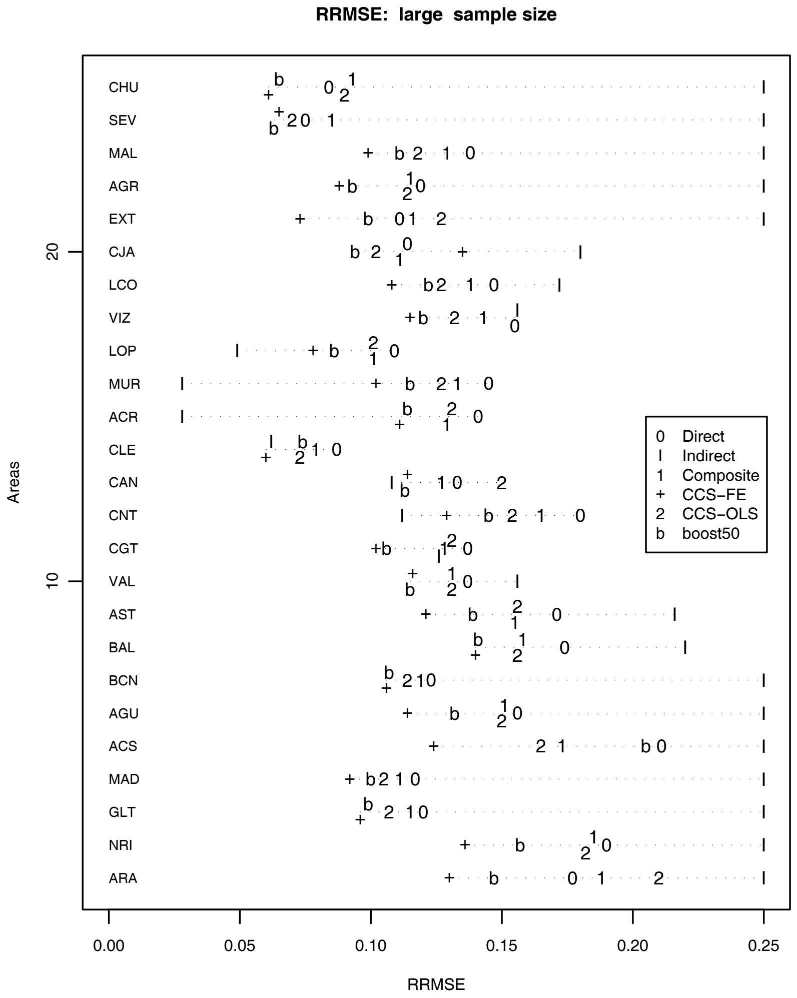

Figure 1: *RRMSEs for the areas and for estimators in the case of large sample (average small area sample size is* 1000*). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at* 0.25*.*

Figure 2 shows the same information, using the same layout and symbols, for small sample size (100). Again, the indirect estimator (I) is the best for the areas which unemployment value is at the middle range (around the national rate), although they are very inefficient for the areas with extreme rates of unemployment. In those areas, the composite RS-based and the feasible new CS-based estimators (1 and 2, respectively) are more efficient than the direct estimator b for most of the areas. For most of the areas, 2 is the most efficient, after the benchmark estimator +.

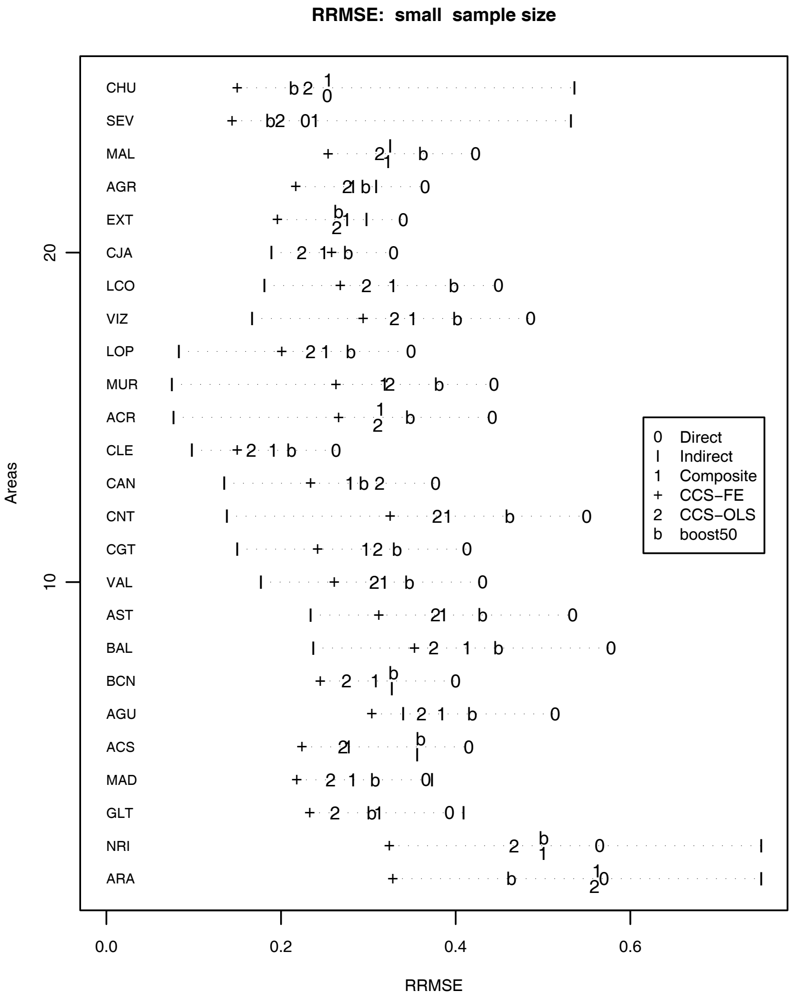

Figure 2: *RRMSEs for the areas and for estimators in the case of small sample (average small area sample size is* 100*). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at* 0.75.

For an area and setting of the simulations (sample size), we define the pattern of RRMSEs by their order for the estimates 0, 1 and 2. For example, for the setting in Figure 1 (large sample size), the pattern for Aragón is 012, which means that the RRMSE for 0 is the smallest and the RRMSE for 2 is the largest of the three; see bottom of the diagram. We say that estimator 2 is the winner for an area if the pattern of RRMSEs is 201 or 210. In Figure 2, we see that 2 is the winner in 22 areas and 1 is the winner in the remaining three areas. This insight can not be gained from Tables 5-7. With large sample size, estimator 2 wins only 17 areas, so it is still preferable, but less decisively so. We also see that 2 is more efficient than b in 22 areas for small sample, but in only two areas for large sample.

Tables 5-7 and Figures 1 and 2 corroborate the prior expectation that composite estimators outperform direct estimators in almost all settings and for almost all areas, and that the indirect estimator is efficient only in areas with small sample size.

We summarize our findings from the simulations as follows:

1) CCS-OLS (with sample data at one time point) is less efficient than the benchmark estimator CCS-FE.
2) Only for very large samples (1000), CCS-OLS has no gains over the direct estimator. For smaller samples, CCS-OLS is comparable with the benchmark estimator with sample size boosted by up to 50%.
3) A substantial part of the gains attained by the benchmark estimator is attained also by the CCS-OLS estimator.
4) The CCS-OLS estimator is slightly more efficient than the estimators that use information solely from RS.
5) The behaviour of the small area estimators does not change much whether we consider total, male or female unemployment rates.
6) In the context of the estimation of Spanish unemployment rates and for moderate area sample sizes (say, 200 subjects in the area), the simplest CCS-OLS estimator is comparable with an increase of sample size by up to 50%.

As a concluding remark, our results show that regional statistics are not in conflict with the statistics produced at the country-wide level. Rather, a regional survey can be combined with a country survey to improve the precision of estimators for small areas, avoiding the costly solution of increasing the region's subsample size.

6 References

Costa, A, Garcia, M., Lopez, X., and Pardal, M. (2006). Estimació de les taxes de desocupació comarcal a Catalunya. Aplicació d'estimadors de petita àrea amb combinació d'enquestes, Working Document, IDESCAT, Barcelona.

Costa, A., Satorra, A. and Ventura, E. (2003). An empirical evaluation of small area estimators, *SORT (Statistics and Operations Research Transactions)*, 27 (1), 113-135.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation, *SORT (Statistics and Operations Research Transactions)*, 28 (1), 69-86.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9 (1), 55-93.

Longford, N. T. (2004). Missing data and small-area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society A*, 167, 341-373.

López, R. (2000). Estimaciones para áreas pequeñas. *Estadística Española*, 42, 146, 291-338.

Mancho, J. (2002). Técnicas de estimación en áreas pequeñas. *Cuaderno Técnico del Eustat*.

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

StataCorp. (2003) *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation.