SORT 28 (1) January-June 2004, 69-86

Improving both domain and total area estimation by composition*

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

Statistical Institute of Catalonia (Idescat) and Universitat Pompeu Fabra

Abstract

In this article we propose small area estimators for both the small and large area parameters. When the objective is to estimate parameters at both levels, optimality is achieved by a sample design that combines fixed and proportional allocation. In such a design, one fraction of the sample is distributed proportionally among the small areas and the rest is evenly distributed. Simulation is used to assess the performance of the direct estimator and two composite small area estimators, for a range of sample sizes and different sample distributions. Performance is measured in terms of mean squared errors for both small and large area parameters. Small area composite estimators open the possibility of reducing the sample size when the desired precision is given, or improving precision for a given sample size.

MSC: 62J07, 62J10, and 62H12

Keywords: regional statistics, small areas, mean square error, direct and composite estimators

1 Introduction

This study stems from a practical issue. The Institut d'Estadística de Catalunya (IDESCAT) had to develop an Industrial Production Index (IPI) for the Catalan autonomous community. The Instituto Español de Estadística (INE) did not produce any

* **Acknowledgements**: The authors are grateful to Xavier López of the Statistical Institute of Catalonia (IDESCAT) for his help in the elaboration of the figures of the paper, and to Nicholas T. Longford for his detail comments on a previous version of this paper. Eva Ventura acknowledges support from the research grants SEC2001-0769 and SEC2003-04476.

[1] *Address for correspondence:* Statistical Institute of Catalonia (Idescat). Via Laietana, 58 - 08003 Barcelona, Spain. E-mail: acosta@idescat.es.

[2] *Address for correspondence:* Department of Economics and Business. Universitat Pompeu Fabra, 08005 Barcelona, Spain E-mail: albert.satorra@econ.upf.es, eva.ventura@econ.upf.es.

Received: February 2004

Accepted: May 2004

regional IPI for Spain, just a national one. IDESCAT had no budget for conducting a Catalan monthly survey. Instead, IDESCAT estimated the IPI for Catalonia using the Spanish IPI of 150 industrial branches, weighted according to their relative importance in Catalonia. This Catalan IPI is a synthetic estimator and was accepted very well by the analysts of the Catalan economy.

The statisticians of IDESCAT had performed a test prior to publishing the new index. The Instituto Vasco de Estadística (EUSTAT) conducted its own regional survey in the Basque Country and published a Basque IPI. IDESCAT created a synthetic index for the Basque Country, using the methodology applied to the Catalan index. This index was compared to EUSTAT's IPI and the results of such comparison evidenced an acceptable performance for the new index. Both the level value of the synthetic IPI and its rate of variation were very useful in order to follow the Basque economic situation (see Costa and Galter 1994). Based on these results, IDESCAT produced a synthetic IPI for Catalonia. Following this, INE applied the same methodology to obtain a distinct IPI for each of the seventeen Spanish autonomous communities.

The method used by IDESCAT is by no means standard in the Spanish official statistics. The synthetic IPI was criticized within some fields, even when it worked well in Catalonia. Some studies (see Clar, Ramon and Surinach 2000) showed that the synthetic IPI works well in regions that possess a significant and quite diversified industry, such as Catalonia. But it fails in other Spanish regions. This observation encouraged the IDESCAT to investigate the theoretical basis of its synthetic IPI from the context of the small area methods.

There is a varied methodology on small area estimation. The reader can consult Platek, Rao, Särndal and Singh (1987), Isaki (1990), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994) to gain an overview of them. Some of the methods use auxiliary information from related variables in the estimation of area-level quantities. In Spain, recent work by Morales, Molina and Santamaría (2003) deals with small area estimation with auxiliary variables and complex sample designs. We concentrate on methods that use sample information solely from the target variable. These methods include direct and some indirect estimators. Traditional direct estimators use only data from the small area being examined. Usually they are unbiased, but they exhibit a high degree of variation. Indirect, composite and model-based estimators are more precise since they use also observations from related or neighboring areas. Indirect estimators are obtained using unbiased large area estimators. Based on them, estimators can be devised for smaller areas under the assumption that they exhibit the same structure as the large area. Composite estimators are linear combinations of direct and indirect estimators.

The research program on small area estimation carried out jointly by IDESCAT and researchers of the Universitat Pompeu Fabra is characterized by its focus on covariate-free models. An estimator that is based on using auxiliary information from other variables at hand will in general be more efficient, but introduces degrees of subjectivity.

We believe the covariate-free small area estimators are the only ones that are readily usable in the present stage of our official statistics framework.

Costa, Satorra and Ventura (2002) worked with a survey that included direct regional estimators of the Spanish work force. They studied three small area estimators: a synthetic, a direct, and a composite one. The study concluded that the composite estimator and the synthetic estimator were almost identical in Catalonia, because this region's economy is a large component of the whole Spanish economy. The bias of the synthetic estimator was found to be very small for Catalonia.

Costa, Satorra and Ventura (2003) used Monte Carlo methods (with both an empirical[1] and a theoretical population) to compare the performance of several small area estimators: a direct, a synthetic, and three composite estimators. These composite estimators differ in the way the direct and synthetic estimators are combined. One of the composite estimators used theoretical weights (based on known bias and variances). The other two use estimated weights assuming homogeneous or heterogeneous biases and variances across the small areas and concluded that, given the usual sample sizes used in official statistics, the composite estimator based on the assumption of heterogeneity of biases and variances is superior.

Often the statistician is interested on the estimation of both small and large area parameters. In this case, classical estimation methods use sample designs that vary according to the assignment of sample size to the small areas. The following sample designs are considered: a) a proportional design, in which the sample size of each area is proportional to the size of the area in the population; b) uniform design, in which all the areas share the same sample size, regardless of the size of the area, and c) the mixed design, that shares the strategies of a) and b). Clearly, design a) will be optimal when we focus on estimating accurately the large area parameter; while design b) will be chosen when we want to obtain accurate estimates of the small area quantities.

Using Monte Carlo methods on a real population (a labour force census of enterprises) and mixed designs with varying levels in the mixing of uniform and proportional sampling, in the present paper we show how small area estimation improves the estimation of both the small and large are parameters. It will be seen that by using composite small area estimates we can either reduce sample size when precision is given, or improve precision, when sample size is fixed.

The outline of the paper is as follows. Section 2 describes small area estimation. Section 3 describes the fixed, proportional and mixed sampling designs. Section 4 presents the Monte Carlo study. Section 5 describes the results of the Monte Carlo study concerning the direct versus composite estimates. Finally, Section 6 describes how composition improves both large and small area parameters.

1. The empirical population is the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. The small areas are the forty-one Catalan counties.

2 Small area estimation without covariates

Suppose a large country area is divided into small area domains $j = 1, 2, \ldots, J$. Let N be the size of the population, and N_1, N_2,. . ., N_J be the sizes of the J small areas.

Let X be a scalar variable and that we are interested in estimating the mean (or the total) of X for each of the J regions, as well as the overall mean. Let θ_j be the mean of X in the region j, and θ_* be the mean of X in the population. The variance of X in region j is denoted as σ_j^2.

Suppose we have a direct estimator $\hat{\theta}_j$ of the mean of X in each domain, such that $\hat{\theta}_j \sim N\left(\theta_j, \sigma_j^2/n_j\right)$, $j = 1, 2, \ldots, J$, and an estimator $\hat{\theta}_*$ for the large area mean, with $\hat{\theta}_* \sim N\left(\theta_*, \sigma_*^2\right)$. Furthermore, assume a distribution for the mean of area j, $\theta_j \sim N\left(\theta_*, b_j^2\right)$ where b_j^2 is a variance parameter that (possibly) varies with the region.

The design of the survey attends usually to the objective of ensuring precision when estimating the parameters at the country level. For the sake of simplicity, assume that, $\hat{\theta}_*$ is unbiased for θ_* and σ_*^2 is very small. However, some sample surveys have secondary uses of providing information about the small areas. The sample size of most sub-domains is too small, may even be null, to draw accurate inferences about the mean of the small area on the basis of the direct estimate $\hat{\theta}_j$. That is, even though $\hat{\theta}_j$ is an unbiased estimate for θ_j, its variance σ_j^2 is too large to provide an accurate estimation of the small area level parameter.

In this context it is advisable to use composite estimators. They combine linearly the direct estimator and a synthetic (indirect) estimator. The best linear composite estimator of θ_j (in the sense of minimizing the mean squared error, or MSE) is

$$\tilde{\theta}_j = \pi_j \hat{\theta}_* + \left(1 - \pi_j\right) \hat{\theta}_j \tag{1}$$

with

$$\pi_j = \frac{\sigma_j^2/n_j - \gamma_j}{\left(\theta_j - \theta_*\right)^2 + \sigma_j^2/n_j + \sigma_*^2 - 2\gamma_j} \tag{2}$$

where γ_j denotes the covariance between the direct estimator and $\hat{\theta}_*$. For simplicity, assume that the covariance $\gamma_j = 0$ and σ_*^2 is negligible. The value of π_j that minimizes the MSE is

$$\pi_j = \frac{\sigma_j^2/n_j}{\left(b_j^2 + \sigma_j^2/n_j\right)} \tag{3}$$

where $b_j^2 = \left(\theta_j - \theta_*\right)^2$.

The values of the variance σ_j^2 and squared bias b_j^2 are usually unknown, and therefore they must be estimated if we wish to approach the optimal value of $\pi_j$in (3).

There are several procedures for estimating these population parameters. In the present study, we use two estimators, the “classic composite” and the “alternative composite”, as described and investigated in Costa, Satorra and Ventura (2003).

1. Classic composite estimator

The classic composite estimator assumes that the small areas share the same within-area variance (of the baseline data) and a common estimate for the squared bias. Specifically, we assume $\hat{\theta}_j \sim N\left(\theta_j, \sigma^2/n_j\right)$, $j = 1, 2, \ldots, J$, and $\theta_j \sim N\left(\theta_*, b^2\right)$. We obtain the base line variance by a weighted mean of the sample variances from each area as an estimate. Thus, we define the pooled within-area variance

$$\bar{s}^2 = \frac{\sum_{j=1}^{J}\left(n_j - 1\right)s_j^2}{(n - J)}, \tag{4}$$

where n is the size of the entire sample, n_j is the sample size of the small area and s_j^2 is the sample variance of the baseline data of the small area j. If we assume that $\sigma_j^2 = \sigma^2$ for all of j, the estimator of σ_j^2 is $\bar{s}^2$.

For the squared bias $\left(\theta_* - \theta_j\right)^2$, we define the common estimator

$$b^2 = \frac{1}{J}\sum_{j=1}^{J}\left(\hat{\theta}_j - \hat{\theta}_*\right)^2, \tag{5}$$

i.e., the mean squared difference of the direct and indirect estimators.

Thus, the estimator of π_j is:

$$\hat{\pi}_j^c = \frac{\bar{s}^2/n_j}{\bar{s}^2/n_j + b^2}, \tag{6}$$

and the composite estimator obtained by substituting π_j^c for π_j in (1)

$$\tilde{\theta}_j^c = \hat{\pi}_j^c\hat{\theta}_* + \left(1 - \hat{\pi}_j^c\right)\hat{\theta}_j \tag{7}$$

2. Alternative composite estimator

An alternative for the above classic composite estimator is based on direct estimators of each area's variance and bias. In this way the estimator of π_j is:

$$\hat{\pi}_j^a = \frac{s_j^2/n_j}{\left(\hat{\theta}_j - \hat{\theta}_*\right)^2} \tag{8}$$

Note that $\left(\hat{\theta}_j - \hat{\theta}_*\right)^2$ is biased for $(\theta_j - \theta_*)^2$, but is unbiased for $\sigma_j^2/n_j + b_j^2$, as

$$\begin{aligned} &E\left(\hat{\theta}_j - \hat{\theta}_*\right)^2 = E\left(\hat{\theta}_j - \theta_j + \theta_j - \hat{\theta}_*\right)^2 = \\ &= E\left(\hat{\theta}_j - \theta_j\right)^2 + E\left(\theta_j - \hat{\theta}_*\right)^2 + 2E\left(\hat{\theta}_j - \theta_j\right)\left(\theta_j - \hat{\theta}_*\right) = \\ &= \sigma_j^2/n_j + b_j^2 \end{aligned}$$

which leads to the alternative composite estimator

$$\tilde{\theta}_j^a = \hat{\pi}_j^a \hat{\theta}_* + \left(1 - \hat{\pi}_j^a\right) \hat{\theta}_j \tag{9}$$

If necessary, the weight $\hat{\pi}_j^a$ is truncated to one.

3 Survey design with small areas

$$\sum_{j=1}^{J} n_j = n \tag{10}$$

and

$$\frac{n_j}{n} = \frac{N_j}{N} \quad j = 1, 2. \ldots, J \tag{11}$$

where n_j is the size of the sample belonging to area j.

A **purely fixed** survey design assigns the same sample size to each small area. Therefore

$$n_j = \frac{n}{J} \quad \text{for } j = 1, 2, \ldots, J \text{ and } \sum_{j=1}^{n} n_j = n \tag{12}$$

A **mixed** survey design distributes a fraction of the whole sample in a proportional way among the different areas, with the rest of the sample distributed evenly among the areas (Sing, Mantel and Thomas, 1994). Let k be the fraction of the sample to be assigned to the proportional design.

Then

$$n_j = k\frac{N_j}{N}n + (1 - k)\frac{n}{J} \text{ for } j = 1, 2, \ldots, J \text{ and } \sum_{j=1}^{J} n_j = n \tag{13}$$

A pure proportional sample design minimizes MSE for the estimate of the country-level quantity, while a pure fixed design minimizes the MSE of the estimates at the small area level. In the present paper we use simulation to explore the performance of a mixed design strategy when the interest is in minimizing the MSE of estimates of both the country and region level quantities. For that we consider different sample sizes, different survey design strategies, and different estimators.

4 Monte Carlo study

In this section we conduct a Monte Carlo study in which we extract multiple samples from a known population. We use data from the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. This census contains data on the number of employees from each surveyed enterprise who are registered with the Social

Table 1: *Population characteristics: size, county-mean, square bias, and variance.*

	Population size	θ_j	$\left(\theta_j - \theta_*\right)^2$	$\sigma_j^2(x)$
Alt Camp	1282	8,73[a]	0,09	3250,37
Alt Empordà	4712	5,28	14,11	294,27
Alt Penedès	3052	8,91	0,02	1686,24
Alt Urgell	745	4,71	18,7	158,25
Alta Ribagorça	140	4,59	19,73	205,38
Anoia	3264	7,86	1,37	801,64
Bages	5698	8,24	0,63	1356,9
Baix camp	5530	6,47	6,59	479,54
Baix Ebre	2237	6,31	7,41	534,4
Baix Empordà	4634	5,44	12,92	425,17
Baix Llobregat	20541	9,73	0,48	1642,46
Baix Penedès	2197	5,26	14,23	171,82
Barcelonès	88331	10,63	2,55	10314,88
Berguedà	1397	5,44	12,9	196,15
Cerdanya	788	3,71	28,34	71,93
Conca de Barberà	611	8,29	0,56	1388,95
Garraf	3466	6,28	7,62	685,91
Garrigues	516	5,24	14,42	96,89
Garrotxa	1909	7,51	2,33	419,72
Gironès	6369	9,82	0,62	2037,47
Maresme	11718	6,46	6,64	605,07
Montsià	1918	5,61	11,73	246
Noguera	1128	5,12	15,3	93,29
Osona	5494	7,09	3,77	774,65
Pallars Jussà	410	4,37	21,76	130,37
Pallars Sobirà	272	4,06	24,76	55,46
Pla d'Urgell	1106	6,59	5,95	271,85
Pla de l'Estany	1160	6,07	8,79	143,37
Priorat	254	4,11	24,26	180,17
Ribera d'Ebre	620	5,71	11,07	418,72
Ripollès	959	7,87	1,35	875,92
Segarra	594	10,87	3,35	8171,41
Segrià	7096	7,74	1,69	714,23
Selva	4586	7,11	3,7	610,2
Solsonès	508	5,58	11,93	157,58
Tarragonès	7440	9,42	0,15	1675,66
Terra Alta	297	4,25	22,87	40,28
Urgell	1178	6,28	7,59	312,25
Val d'Aran	503	5,28	14,08	270,11
Vallès Occidental	26683	10,34	1,71	3026,89
Vallès Oriental	11795	8,45	0,34	832,68

The mean of the affiliates for the whole of Catalonia is 9.04

Security. The census was carried out in each of the four quarters between the years 1992 and 2000 (inclusive). We limit the analysis to one year, 2000.

The database contains 243,184 observations from year 2000, divided into 12 groups according to the economic sector, and 41 counties (Catalan «comarques»). A few enterprises were excluded from the analysis because their locations were not established.

We eliminated the sector-based classification and focused solely on the division to counties. Table1 shows the number of enterprises per county and the mean and variance of the number of employees per enterprise. The distribution of enterprises is quite uneven, as they are concentrated in a few populous areas.

We consider four sample sizes and five alternative survey designs. The smallest sample size is 2,050 observations. We then repeatedly double the sample and use 4,100, 8,200 and 16,400 observations. For each sample size, we consider a purely proportional sample, a 75%, 50% and 25% mixed sample design (respectively), and a purely fixed sample design, that is, combinations with $k = 1, 0.75, 0.50, 0.25$ and 0.

For the overall sample size of $n = 4100$, Table A1 in the Appendix shows the small area sample sizes in each of those 20 (4 by 5) scenarios. As some of the counties have very small population, we have used sampling with replacement. The number of Monte Carlo replications is 1,000. We evaluate the direct, classic composite and alternative composite estimators for each of the 41 counties as well as for the whole of Catalonia.

5 Direct vs. composite estimators

We computed the MSE of each small area estimator. Table 2 shows a summary of descriptive statistics. The mean, median, variance, minimum and maximum values of the MSE of the small area estimators across the 1,000 replications are presented. Table 3 evaluates the relative performance of the three alternative estimators differently. In that table we calculate the percentage of counties for which the MSE of a particular estimator (in the leftmost column) is lower than the MSE of the other two estimators.

The performance of the small area estimations can be evaluated by several criteria. We have observed in Costa, Satorra and Ventura (2003) that the distribution of the MSE across the Catalan counties is asymmetric. It also exhibits extreme values and it is very dispersed. This is a consequence of the extremely uneven distribution of the population and economic activity in the region. This is a drawback of the simple average of the MSE of the counties. The median is affected less by the presence of extreme values. On the other hand, we may want to put one upper limit on the MSE for each small area. Looking at the maximum MSE of the counties is then appropriate. To keep things simple we present the results graphically using the median evaluation criterion. Tables 2 and 3 and Figure 4 show the results using other evaluation criteria. Those criteria are the average of the MSE of the counties, the maximum and minimum values of the MSE of the counties, and the percentage of counties for which a particular type of estimator

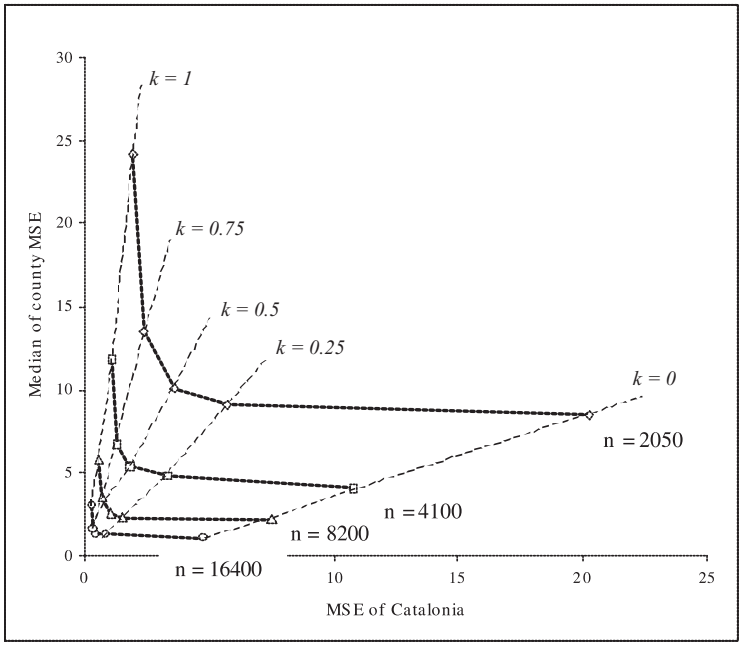

Figure 1: *MSE of direct estimator for various combinations of sample size (n) and sampling design (k).*

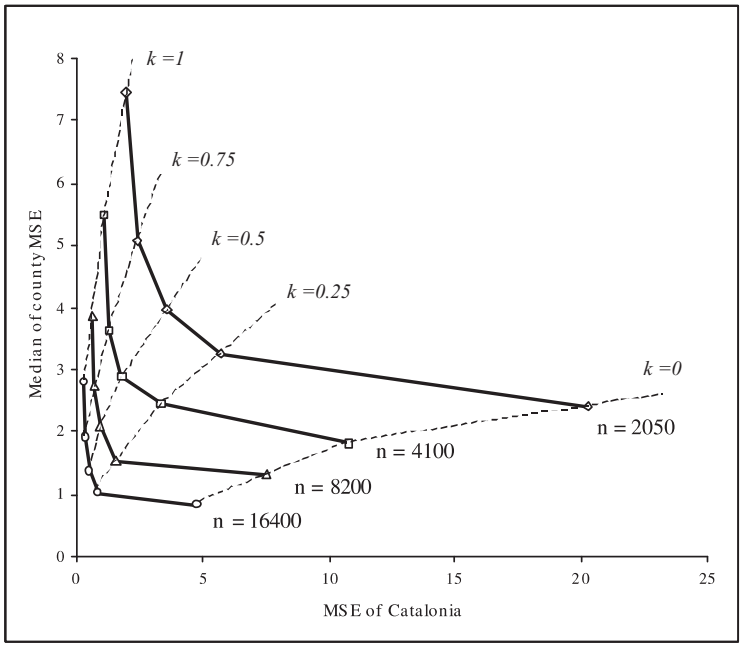

Figure 2: *MSE of composite alternative estimator for various combinations of sample size (n) and sampling design (k).*

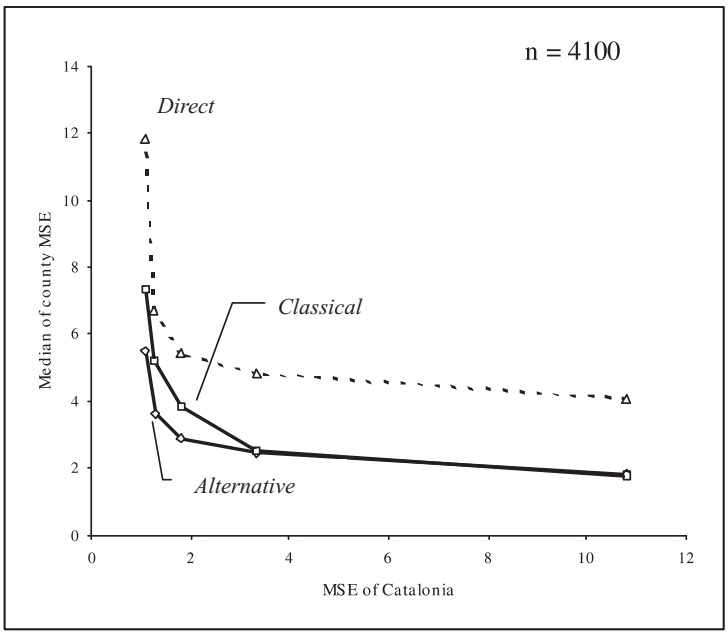

Figure 3: *Comparing three estimators for sample n = 4100.*

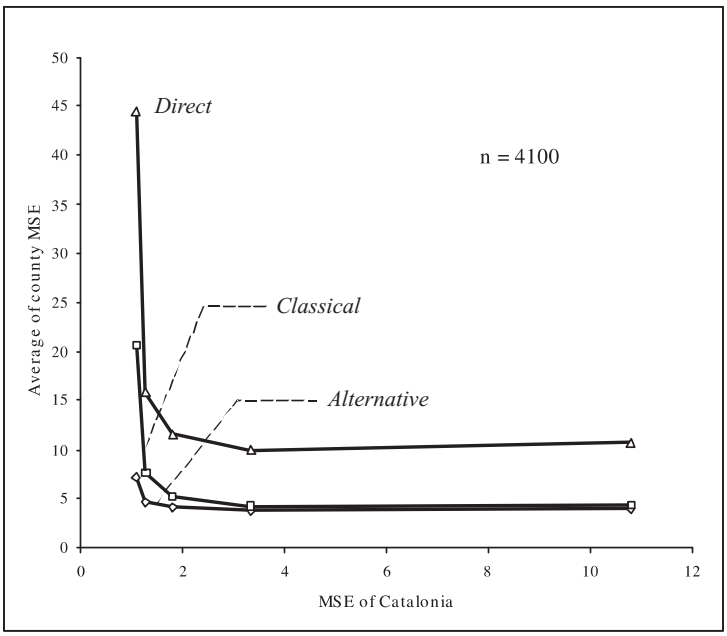

Figure 4: *Comparing three estimators for sample n = 4100 by the criteria of the average county MSE.*

performs better than the other two estimators. Figures 1 to 4 summarize the results visually.

Table 2: *Descriptive statistics of the MSE of the small area estimators, by sample size and sampling choice.*

n = 2050

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	20,45	7,19	5,32	19,20	6,83	5,24	23,22	8,89	5,67	31,05	13,19	6,62	79,69	37,93	8,82
Median	8,43	2,91	2,41	9,12	4,32	3,26	10,04	5,30	3,96	13,49	7,63	5,07	24,20	11,12	7,45
Variance	1141,20	122,53	43,41	1006,35	78,88	27,18	2275,40	177,32	30,53	4737,04	577,39	41,89	48928,68	16967,77	43,10
Min	0,80	1,10	1,34	1,05	1,84	1,76	1,58	2,83	2,66	2,87	3,93	2,97	5,79	4,53	2,99
Max	158,64	55,38	36,97	194,64	56,89	31,89	299,66	87,70	37,24	423,58	154,48	44,76	1382,42	837,40	41,17

n = 4100

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	10,62	4,35	3,90	9,91	4,25	3,78	11,49	5,26	4,06	15,74	7,58	4,75	44,39	20,61	7,23
Median	4,07	1,80	1,82	4,83	2,50	2,46	5,41	3,86	2,90	6,70	5,19	3,63	11,84	7,34	5,51
Variance	298,21	40,74	26,25	260,30	24,03	19,18	484,49	28,65	15,75	1218,45	71,44	17,25	20158,91	4111,94	73,94
Min	0,41	0,57	0,61	0,54	1,21	0,80	0,77	1,50	1,26	1,42	1,94	1,56	2,88	2,27	1,70
Max	78,71	30,84	28,32	94,56	30,84	27,74	133,82	34,52	26,57	212,70	53,80	28,13	904,32	417,57	56,99

n = 8200

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	5,70	2,83	2,83	4,88	2,73	2,64	6,01	3,72	2,94	8,39	5,36	3,53	21,21	10,93	5,24
Median	2,26	1,21	1,32	2,28	1,63	1,54	2,53	2,89	2,10	3,54	3,64	2,74	5,82	5,41	3,88
Variance	112,27	19,88	17,66	66,30	9,59	13,59	145,38	11,38	14,13	408,74	21,91	16,75	3998,63	323,97	34,33
Min	0,19	0,37	0,24	0,28	0,68	0,33	0,40	1,00	0,51	0,74	1,34	0,96	1,60	1,51	1,02
Max	54,53	21,02	23,29	49,34	20,06	23,42	73,93	21,20	25,23	125,17	26,32	27,95	397,33	115,13	38,37

n = 16400

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	2,85	1,75	1,95	2,45	1,78	1,75	2,89	2,45	1,87	4,28	4,03	2,45	10,96	7,38	3,86
Median	1,07	0,72	0,83	1,28	1,16	1,02	1,32	1,87	1,36	1,59	2,50	1,92	2,97	3,76	2,80
Variance	30,51	7,57	9,54	17,80	3,81	7,38	32,95	4,58	7,51	125,96	13,68	13,57	1295,79	70,52	26,49
Min	0,11	0,24	0,12	0,12	0,42	0,13	0,18	0,59	0,20	0,36	0,67	0,43	0,68	0,74	0,62
Max	30,92	13,42	16,78	26,17	12,75	17,36	35,43	12,68	18,26	70,87	17,97	24,85	230,29	42,83	34,26

Table 3: *Comparing estimators under the percentage criterion.*

n = 2050

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		7,32%	7,32%		17,07%	12,20%		19,51%	12,20%		24,39%	7,32%		12,20%	0,00%
Classical	92,68%		34,15%	82,93%		19,51%	80,49%		12,20%	75,61%		7,32%	87,80%		7,32%
Alternative	92,68%	65,85%		87,80%	80,49%		87,80%	87,80%		92,68%	92,68%		100,00%	92,68%	

n = 4100

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		14,63%	21,95%		24,39%	21,95%		34,15%	24,39%		39,02%	17,07%		34,15%	7,32%
Classical	85,37%		56,10%	75,61%		39,02%	65,85%		29,27%	60,98%		21,95%	65,85%		12,20%
Alternative	78,05%	43,90%		78,05%	60,98%		75,61%	70,73%		82,93%	78,05%		92,68%	87,80%	

n = 8200

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		21,95%	31,71%		34,15%	36,59%		41,46%	39,02%		41,46%	39,02%		41,46%	26,83%
Classical	78,05%		68,29%	65,85%		60,98%	58,54%		43,90%	58,54%		41,46%	58,54%		17,07%
Alternative	68,29%	31,71%		63,41%	39,02%		60,98%	56,10%		60,98%	58,54%		73,17%	82,93%	

n = 16400

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		21,95%	43,90%		39,02%	46,34%		43,90%	53,66%		53,66%	60,98%		53,66%	43,90%
Classical	78,05%		60,98%	60,98%		58,54%	56,10%		48,78%	46,34%		43,90%	46,34%		26,83%
Alternative	56,10%	39,02%		53,66%	41,46%		46,34%	51,22%		39,02%	56,10%		56,10%	73,17%	

The horizontal axis corresponds to the MSE of Catalonia $MSE(\hat{\theta}_*)$. The vertical axis corresponds to the median of the 41 counties MSE. The figures show the behaviour of the three estimators considered for different total sample sizes and alternative sample designs. In Figure 3 we overlay the curves of the estimators. We only draw the curves for sample sizes 4,100 and 8,200, to avoid excessive clutter.
The results can be summarized as follows:

- From Table 2 and figures 1 to 3 we see that the MSE for Catalonia is smaller when $k = 1$ and larger when$k = 0$. In contrast, the median and mean county MSE is smaller when $k = 0$ and larger when$k = 1$. This result holds for all three estimators.
- Both the MSE for Catalonia and the median county MSE are reduced as the total sample size is increased, for each estimator. The same result holds for other summaries, such as the average county MSE, or the maximum values of MSE (see Figure 4).
- On average, the alternative composite estimator is the best estimator, as assessed by the MSE of Catalonia and median county MSE, for any sample size and values of k of the mixed design. There are some exceptions, when the total sample sizes are large and we use a fixed sample survey design. This is seen in Figure 3, where, to avoid clutter, we only show the sample size $n = 4100$.
- In Table 3 we see that the two composite estimators are the best in almost all settings. There are some exceptions when the total sample size is large. As expected, the direct estimator improves its behavior as the total sample size increases, independently of the survey design.
- Table 3 shows also that the alternative composite estimator is better than the classical one except for sampling designs with $k = 0$ or $k = 0.25$ (nonproportional survey designs).
- To achieve a particular combination of a small MSE for the large area jointly with small median or average small area MSE, a mixed design strategy is recommended. The desired combination will depend on the preferences about how to use the estimators.

6 Improving survey design by composition

The results in the previous section suggest some clear guidelines for how to improve both the sampling design and the estimation. We examine them now.

Assume that we start with a predetermined sample size and a mixed-design allocation, partly proportional and partly of equal sample sizes (the same sample size in all the counties). More specifically, suppose the budget allows to extract a sample of size $n = 8{,}200$. A fraction of these observations (for example, 35%) is distributed proportionally among the small areas and the remainder is distributed evenly. That

means that each Catalan county would have at least 0.65 * 8200/41 = 130 observations. Some counties with a large population would have up to 500 more observations, while the sample size of others would not surpass 150.

If we decide to use a direct estimator for each small area as well as the county, we will obtain the MSE for large and small areas that corresponds approximately to point A (k = 0.35) in Figure 5.

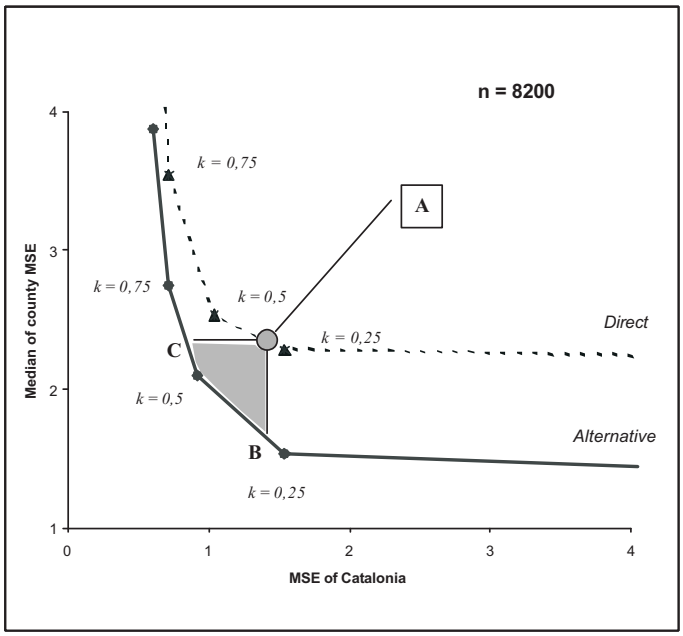

Figure 5: *Opportunities for improvement using composite estimators.*

This figure shows that using the alternative composite estimator (the same would occur if we choose the classical composite estimator) improves estimation over the use of a direct estimation (i.e., non composition). There is a continuum of choices between points B and C (that correspond to varying values of k) that achieves lower or equal MSE for the estimate of both the large and small area parameters. The point B, which has k= 0.35, shows that the use of a composite estimator reduces the median of county MSE, while keeping the MSE of the estimate of Catalonia at the same level than the direct estimate. This point B, with k = 0.35, is the limit that we can move toward egalitarian sampling design, without loosing (by composition) in the estimation of the large area parameter. As we move towards point C, thus adopting a more proportional sampling design, composition improves both the median of county MSE and the MSE of Catalonia over direct estimation. Point C, with k = 0.65, represents the limit we can move on a more proportional design, without loosing (by composition) in the estimation of the small area parameters (i.e., without increasing the median of county MSE over the value obtained with the direct estimate).

That is, the adoption of a small area estimate such as the alternative composite estimator (the classical composite estimator would lead also to the same phenomena) brings room for improvement in the precision of the estimates of both the large and small area parameters, over the use of the simple direct estimation. The main aim of this paper was to illustrate this issue using Monte Carlo data on a real population. We left for further work the development of specific tables to be used for the choice of the required sample sizes and values of *k* required for attaining a priori specified precision.

References

Clar, M., Rarmos and Suriñach, J. (2000). Avantatges i inconvenients de la metodologia del INE per elaborar indicadors de la producció industrial per a les regions espanyoles, *Qüestiió*, 24, 1, 151-186.

Costa, A. and Galter, J. (1994). L'IPPI, un indicador molt valuós per mesurar l'activitat industrial catalana, *Revista d'Indústria*, 3, Generalitat de Catalunya, 6-15.

Costa, A., Satorra, A. and Ventura, E. (2002). Estimadores compuestos en estadística regional: Aplicación para la tasa de variación de la ocupación en la industria, *Qüestiió*, 26, 1-2, 213-243.

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.

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

Isaki, C.T. (1990). Small-Area Estimation of Economic Statistics, *Journal of Business & Economic Statistics*, 8, 4, 435-441.

Morales, D., Molina, I. and Santamaría, L. (2003). Modelos estándar para estimaciones en áreas pequeñas. Extensión a diseños complejos. *Proceedings of the XXVII Congreso Nacional de Estadística e Investigación Operativa* (Lleida, Abril de 2003).

Platek, R., Rao, J.N.K., Särndal, C.E. and Singh, M.P. (eds.) (1987). *Small Area Statistics: An International Symposium*. New York. John Wiley and Sons.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and Strategies for Small Area Data, *Survey Methodology*, 20, Statistics Canada, 3-22.

Resum

Els estimadors de petita àrea poden ser utilitzats no només per aproximar els paràmetres d'una petita àrea, sinó també per estimar els paràmetres de l'àrea gran. Quan l'objectiu és estimar un paràmetre en totes dues àrees, l'estratègia òptima s'aconsegueix mitjançant un disseny mostral en dues parts: una part que es distribueix proporcionalment entre les petites àrees i una altra part que es distribueix fixa. S'utilitza un mètode de simulació per avaluar el comportament tant de l'estimador directe com de dos estimadors compostos de petita àrea. La bondat de les estimacions es valora en termes de l'error quadràtic mitjà dels estimadors dels paràmetres de les dues àrees, la gran i la petita. Els estimadors compostos de petita àrea obren la possibilitat bé de reduir la mida mostral quan el nivell de precisió està donat, bé de millorar la precisió quan la mida mostral està donada.

MSC: 62J07, 62J10, and 62H12

Paraules clau: estadística regional, petites àrees, error quadràtic mitjà, estimadors directe i compost

APPENDIX

***Table** A.1: Sample sizes of the small areas, by sampling choice, for n = 4100.*

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Proportional	Fixed	Total	Proportional	Fixed	Total	Proportional	Fixed	Total	Proportional	Fixed	Total	Proportional	Fixed	Total
Alt Camp	0	100	100	6	74	80	10	50	60	16	26	42	22	0	22
Alt Empordà	0	100	100	20	74	94	40	50	90	58	26	84	80	0	80
Alt Penedès	0	100	100	14	74	88	26	50	76	38	26	64	52	0	52
Alt Urgell	0	100	100	4	74	78	6	50	56	10	26	36	12	0	12
Alta Ribagorça	0	100	100	2	74	76	2	50	52	2	26	28	2	0	2
Anoia	0	100	100	14	74	88	28	50	78	40	26	66	56	0	56
Bages	0	100	100	26	74	100	48	50	98	72	26	98	96	0	96
Baix Camp	0	100	100	24	74	98	46	50	96	70	26	96	94	0	94
Baix Ebre	0	100	100	10	74	84	18	50	68	28	26	54	38	0	38
Baix Empordà	0	100	100	20	74	94	40	50	90	58	26	84	78	0	78
Baix Llobregat	0	100	100	88	74	162	174	50	224	254	26	280	346	0	346
Baix Penedès	0	100	100	10	74	84	18	50	68	28	26	54	38	0	38
Barcelonès	0	100	100	386	74	460	742	50	792	1100	26	1126	1488	0	1488
Berguedà	0	100	100	6	74	80	12	50	62	18	26	44	24	0	24
Cerdanya	0	100	100	4	74	78	6	50	56	10	26	36	14	0	14
Conca de Barberà	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Garraf	0	100	100	16	74	90	30	50	80	44	26	70	58	0	58
Garrigues	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Garrotxa	0	100	100	8	74	82	16	50	66	24	26	50	32	0	32
Gironès	0	100	100	28	74	102	54	50	104	80	26	106	108	0	108
Maresme	0	100	100	52	74	126	98	50	148	146	26	172	198	0	198
Montsià	0	100	100	8	74	82	16	50	66	24	26	50	32	0	32
Noguera	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Osona	0	100	100	24	74	98	46	50	96	68	26	94	92	0	92
Pallars Jussà	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Pallars Sobirà	0	100	100	2	74	76	2	50	52	4	26	30	4	0	4
Pla d'Urgell	0	100	100	4	74	78	10	50	60	14	26	40	18	0	18
Pla de l'Estany	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Priorat	0	100	100	2	74	76	2	50	52	4	26	30	4	0	4
Ribera d'Ebre	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Ripollès	0	100	100	4	74	78	8	50	58	12	26	38	16	0	16
Segarra	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Segrià	0	100	100	32	74	106	60	50	110	88	26	114	120	0	120
Selva	0	100	100	20	74	94	38	50	88	58	26	84	78	0	78
Solsonès	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Tarragonès	0	100	100	32	74	106	62	50	112	92	26	118	126	0	126
Terra Alta	0	100	100	2	74	76	2	50	52	4	26	30	6	0	6
Urgell	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Val d'Aran	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Vallès Occidental	0	100	100	116	74	190	226	50	276	332	26	358	448	0	448
Vallès Oriental	0	100	100	50	74	124	100	50	150	148	26	174	198	0	198
TOTAL	0	4100	4100	1066	3034	4100	2050	2050	4100	3034	1066	4100	4100	0	4100