SORT 27 (1) January-June 2003, 113-136

An empirical evaluation of small area estimators

Àlex Costa[a], Albert Satorra[b] and Eva Ventura[b] *

[a] *Institut d'Estadística de Catalunya (Idescat)*
[b] *Universitat Pompeu Fabra*

Abstract

This paper compares five small area estimators. We use Monte Carlo simulation in the context of both artificial and real populations. In addition to the direct and indirect estimators, we consider the optimal composite estimator with population weights, and two composite estimators with estimated weights: one that assumes homogeneity of within area variance and squared bias and one that uses area-specific estimates of variance and squared bias. In the study with real population, we found that among the feasible estimators, the best choice is the one that uses area-specific estimates of variance and squared bias.

MSC: 62J07, 62J10, 62H12

Keywords: Regional statistics, small areas, root mean square error, direct, indirect and composite estimators

1 Introduction

Official statistics is faced with the need to generate estimates for small administrative units, while working with relatively small samples and within stringent budgetary limit. This conflict has been accentuated in recent years: on the one hand, politics is becoming more and more local, necessitating better local information; on the other hand, the public service nature of official statistics makes it more and more clear that producing quality work in this sphere involves not only optimising some theoretical parameters but also applying appropriate methodological strategies to achieve a positive cost/benefit relationship for society. Within this context, the vital nature and relevance that small

* *Address for correspondence:* A. Costa. Via Laietana, 58. 08003 Barcelona, Spain. acosta@idescat.es (A. Costa), albert.satorra@econ.upf.es (A. Satorra) and eva.ventura@econ.upf.es (E. Ventura)
Received: December 2002
Accepted: January 2003

area statistics have had in the 1990s is understandable, as is the interest generated by official regional statistics.

There is a varied methodology on developing small area estimators. The reader can consult Platek *et al.* (1987), Isaki (1990), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994) to gain an overview of them. An initial classification divides the different existing methods into two categories: traditional and model-based. Traditional models include direct and indirect estimators and their combinations. Traditional direct estimators use only data from the small area being examined. Usually they are unbiased, but their exhibit a high degree of variation. Traditional indirect and model-based estimators are more precise since they also use observations from related or neighbouring areas. Indirect estimators are obtained through unbiased large area estimators. Based on them, it is possible to derive estimators for smaller areas under the assumption that they exhibit the same structure (with regard to the phenomenon being studied) as the initial large area. If this condition is not met, biased estimators could result. Traditional composite estimators are linear combinations of direct and indirect estimators. Model-based estimators can be interpreted as composite estimators, but unlike the traditional estimators, the weighting factors depend on the structure of the estimator's covariances. More information on this topic can be obtained from Cressie (1995), Datta *et al.* (1999), Farrell, MacGibbon and Tomberlin (1997), Ghosh and Rao (1994), Pfeffermann and Barnard (1991), Raghunathan (1993), Singh, Mantel and Thomas (1994), Singh, Stukel and Pfeffermann (1998), and Thomas, Longford and Rolph (1994).

In a previous study (see Costa, Satorra and Ventura, 2002), we began to examine these improved estimators, starting with a scenario in which we have two estimators, neither of which is entirely satisfactory:

- a direct estimator, obtained through the sample data pertaining to the small area; unbiased but in general not very precise.
- an indirect (synthetic) estimator, obtained through auxiliary information from other areas, periods or statistical sources; with smaller variance but generally biased.

Statistical theory of small area estimation proposes a way of combining both estimators in a linear fashion so that the resulting estimator represents a compromise between the absence of bias and minimal variance. The resulting composite estimator is the linear combination of the direct and indirect estimator that minimises mean squared error (MSE).

In our previous study, in which the autonomous regions in Spain are the small areas, the following results were obtained: 1) When the small area is centred and quite large (such as Catalonia), the composite estimator is as efficient as the indirect (or synthetic) one, in that it has a very low bias; 2) The composite estimator works well in general, especially in medium-sized and large areas. In our previous work, we wished to study in more detail the behaviour of composite estimates, but the information with which we were working, the National Statistical Institute (INE)'s Survey of the Active Population

(EPA), was a complex survey that made analysis difficult. For this reason, we decided to learn more about composite estimators in a simpler context in which we could carry out a Monte Carlo experiment. This is the objective of the present article.

Specifically, we decided that we needed to estimate the optimal weighting factors, not an easy task given that the variances and covariances of the estimators must themselves be estimated, as must their bias. For this reason we concentrated on a comparative analysis of the direct and indirect estimators with three composite estimators: one that has optimal weighting factors (theoretical), and two that use estimated weighting factors. One of the estimators based on estimated weighting factors uses the hypothesis of homogeneity of bias and variance for all the areas (this is the so-called *classic* composite estimator). The other estimates the area-specific biases and variances (this is the so-called *alternative* composite estimator). The characteristics of these estimators are studied in relation to the distribution of the mean squared error (MSE) and in a scenario with varied sample sizes.

2 The small area estimators

Consider the random variables $\hat{\theta}_j \sim N\left(\theta_{,}\sigma_j^2\right)$, $\quad j = 1,2,\ldots,J$ and $\hat{\theta}_* \sim N\left(\theta_*,\sigma_*^2\right)$, and γ_j the covariance between $\hat{\theta}_j$ and $\hat{\theta}_*$. Our objective is to estimate θ_j using $\hat{\theta}_j$ and $\hat{\theta}_*$. It is well known that the best linear composite estimator of θ_j (in the sense of minimising the MSE) is

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

with

$$\pi_j = \frac{\sigma_j^2 - \gamma_j}{(\theta_j - \theta_*)^2 + \sigma_j^2 + \sigma_*^2 - 2\gamma_j}$$

For simplicity, assume that the covariance $\gamma_j = 0$ and the σ_*^2 is negligible. We also assume that $\theta_j \sim N\left(\theta_*, b_j^2\right)$. The value of π_j that minimises the MSE is

$$\pi_j = \frac{\sigma_j^2}{\left(b_j^2 + \sigma_j^2\right)} \tag{1}$$

In practice, the values of the variance and bias are unknown (they are population-based parameters), and they must thus be estimated if we wish to approach the optimal value of $\pi_j$in o(1).

The quantity of interest for an area can be estimated "naively", using the sample mean of observations in the small area (direct estimator), or the mean of the observations of the entire population sample (indirect estimator). The direct estimator uses only the information on the area j being examined, while the indirect estimator is based on the sample information gathered in all the areas. It is obvious that the direct estimator is

unbiased for the mean of the area. Nevertheless, it has a high variance (given that if the area is small only few of the observations fall in this area). In contrast, the indirect estimator, based on the sample from the entire population, will have a low variance (given the large sample size), but it will suffer from bias when estimating the characteristics of a certain area, which will almost certainly differ from the common characteristics of the entire population.

An estimator that combines the qualities of the direct and indirect estimators with optimal weighting factors is the composite estimator that uses the value π_j as specified in (1). This estimator constitutes a reference in our study, and it is called the **theoretical composite** estimator denoted by (*theor*). Nevertheless, this estimator is not feasible in practice because it depends on the weighting factor π_j, a value that in turn depends on unknown population parameters.

There are several procedures for estimating these population parameters, all of which lead to different small area estimators. In the present study, we investigate the **classic composite** estimator (*class*) and the **alternative composite** estimator (*altern*) which are described below:

2.1 Classic composite estimator

The classic composite estimator assumes that the areas share the same within-area variance and a common estimate for the squared bias. Specifically, we assume components of variance specification $\hat{\theta}_j \sim N\left(\theta_j, \sigma^2\right)$, $j = 1,2,\ldots,J$ with $\theta_j \sim N\left(\theta_*, b^2\right)$.

Here we use a weighted mean of the sample variances from each area as an estimate of the baseline data variance. Thus we define the pooled within variance

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

in which n is the size of the entire sample, n_j is the sample size of the small area (in our real population example, the county) and s_j^2 is the sample variance of the baseline data of the small area j. Under the assumption that $\sigma_j^2 = \sigma^2$for all of j, the estimator of σ_j^2 is $\bar{s}^2 / n_j$.

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

$$b^2 = \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_j - \hat{\theta}_*\right)^2 \tag{3}$$

which is the mean squared difference of the direct and indirect estimators.

We could also have used a weighted mean of the individual biases; however, the properties of each bias estimator are somewhat different. Specifically, in the case in

which we preferred to use the weighted mean of the individual biases, b^2 would be the estimator of a combination of variances *between* and *within* groups.

Thus, the estimator of π_j is:

$$\hat{\pi}_j^c = \frac{\bar{s}^2/n_j}{\bar{s}^2/n_j + b^2}, \tag{4}$$

and the composite estimator obtained through the sample data is

$$\tilde{\theta}_j^c = \hat{\pi}_j^c \hat{\theta}_* + \left(1 - \hat{\pi}_j^c\right) \hat{\theta}_j \tag{5}$$

2.2 Alternative composite estimator

Another way of calculating the composite estimator uses direct estimators of each area's variance and bias. In this way the estimator of π_j is:

$$\hat{\pi}_j^a = \frac{s_j^2/n_j}{\left(\hat{\theta}_j - \hat{\theta}_*\right)^2} \tag{6}$$

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

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

The composite estimator obtained through the sample data has the same form as (5), with $\hat{\pi}_j^a$ replacing $\hat{\pi}_j^c$;

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

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

Thus we consider five estimators: the direct $\hat{\theta}_j$, the indirect $\hat{\theta}_*$ (based on the entire sample), the theoretical composite $\tilde{\theta}_j^t$ based on the optimal weights in expression (1), the classic composite $\tilde{\theta}_j^c$ based on the weights of expression (4) and the alternative composite $\tilde{\theta}_j^a$ based on the weights of expression (6).

3 Monte Carlo study in an artificial population

In this section, we investigate the estimators defined in Section 2 using Monte Carlo methods. By generating artificial populations we explore the effect that different population characteristics have on the behaviour of the estimators. Some of the

conclusions drawn in this section are validated through a Monte Carlo simulation in the context of a real population.

The artificial population is defined by the following components of variance model

$$x_{ij} = a + z_j + y_{ij} \qquad i = 1,2,\ldots,n_j \quad j = 1,2,\ldots,J$$

where ij denotes individual i in the area j, n_j denotes the number of individuals in the sample of area j, and J denotes the number of areas considered. In addition, assume that z_j is distributed with a mean 0 and variance b^2, independent of y_{ij}, which has mean 0and variance $\sigma(x)^2_j$. The specific values of the model parameters and the characteristics of the design used in the study are $b^2 = 1$, $a = 10$ and J= 8, with identical sample sizes, $n_j = n^*$. The common sample size n^* varies between 5 and 45. We consider two settings for $\sigma(x)^2_j$: i) common for all the areas, $\sigma(x)^2_j$= 30; and ii) values specific for each area; $\sigma(x)^2_j$ varies from 30 to 240.

We studied the effect of a change in the value of the within-area variance (for $b^2 = 1$). The variation of $\sigma(x)^2_j$ influences the value of the intra-class correlation *cci*, which varies between 0.05 and 0.30. The variation in the distribution, of z_j (variation among groups) and y_{ij} (variation within the area), is also considered. At the same time, we investigated different types of distribution within and between areas, as well as the total number of areas in the population.

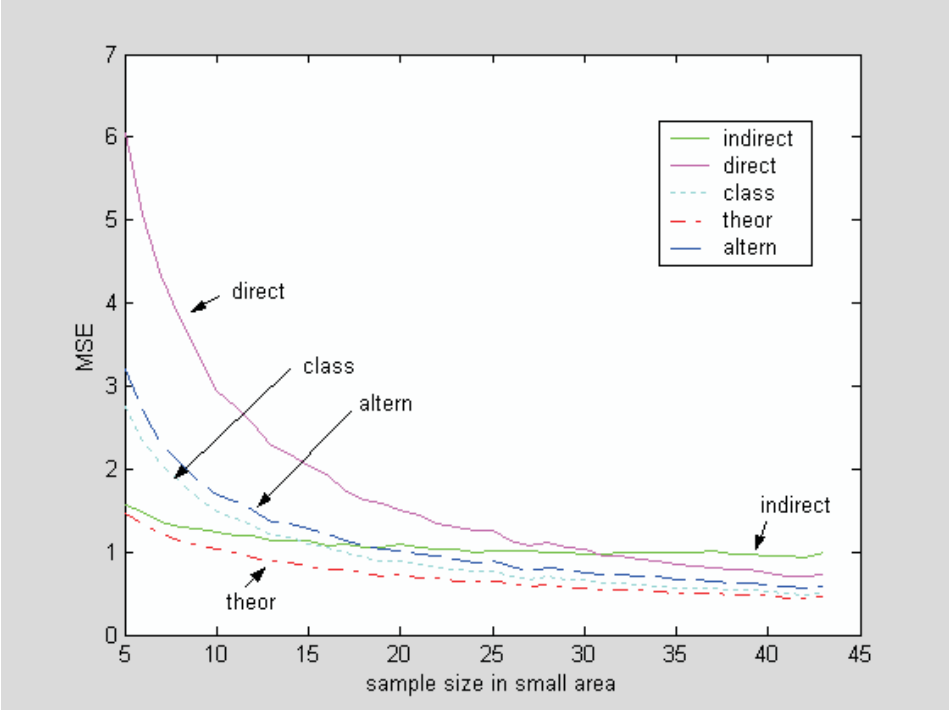

Figure 1: *MSE of the estimators as a function of the sample size n^*: identical within-area variances.*

In summary, we considered the following factors:

i) Sample size n^*(small, medium or large)
ii) Coefficient of intra-class correlation
iii) Homogeneity of the variances within areas
iv) Distribution of the within and between variation

The number of Monte Carlo replications for each combination of factors considered is 6000. In all the cases, we estimated the mean parameter for a specific area. We assess the MSE when estimating this area parameter.

The results of the Monte Carlo study are discussed next. Figures 1-4 show the variation of the MSE of the different estimators analysed (*theor*, *class*, *altern*, *direct* and *indirect*) when we change either the sample size n^* (Figures 1 to 3) or the magnitude of the intra-class correlation coefficient *cci* (Figure 4).

Figure 1 shows the results in the case of normality with identical within-area variances, when we vary the common sample size n^*. The population values used are $b^2 = 1$ and $\sigma(x)_j^2 = 30$, common to all the areas. From Figure 1 we conclude that the MSE is minimal for the *theor* estimator, maximal in the case of the *direct* estimator (except for large sample sizes, in which the MSE of the *direct* estimator can be less than that of the indirect estimator), and that the combined classic and alternative estimators have almost identical MSE, with an intermediate value between the *theor* and direct estimators. MSE are in a wide range for small sample sizes, but it is largely bridged as the sample size grows. The indirect estimator has a MSE greater than that of any of the other estimators for large sample size, but for small sample sizes it behaves similarly to the *theor* estimator. Nevertheless, the wide range of sample sizes for which the *indirect* estimator is better than the *direct* estimator should be noted.

Now consider the case in which there is heterogeneity in the within-area variance $\sigma(x)_j^2$. The results are presented in Figure 2. Here $\sigma(x)_j^2$ varies between 30 and 240 in the eight areas considered. We can consider the two extreme cases in which the area examined has variance 30 and 240, respectively.

This figure shows how the classic estimator improves considerably its performance with respect to the other alternative feasible estimators. Note that the MSE of *class* is, for all sample size, very close to the MSE of *theor*. For the smallest sample sizes, the *class* even improves slightly the performance of the *theor* estimator[1]. It is remarkable that the *altern,* which is mean to account for variation of within-area variance and square bias, performs slightly worst than *class* for all sample sizes considered

The results in Figure 2 correspond to the case in which the area examined has the smallest variances, 30, compared to the largest, 240. In Figure 3, we show the results for the complementary case when the variance of the area examined is 240.

1. This is due to the fact that for small sample sizes and this non-identical within-area case, the variance of the indirect estimator is not negligible. The denominator of *class* of expression (4) has a positive bias that partially accounts for such variance.

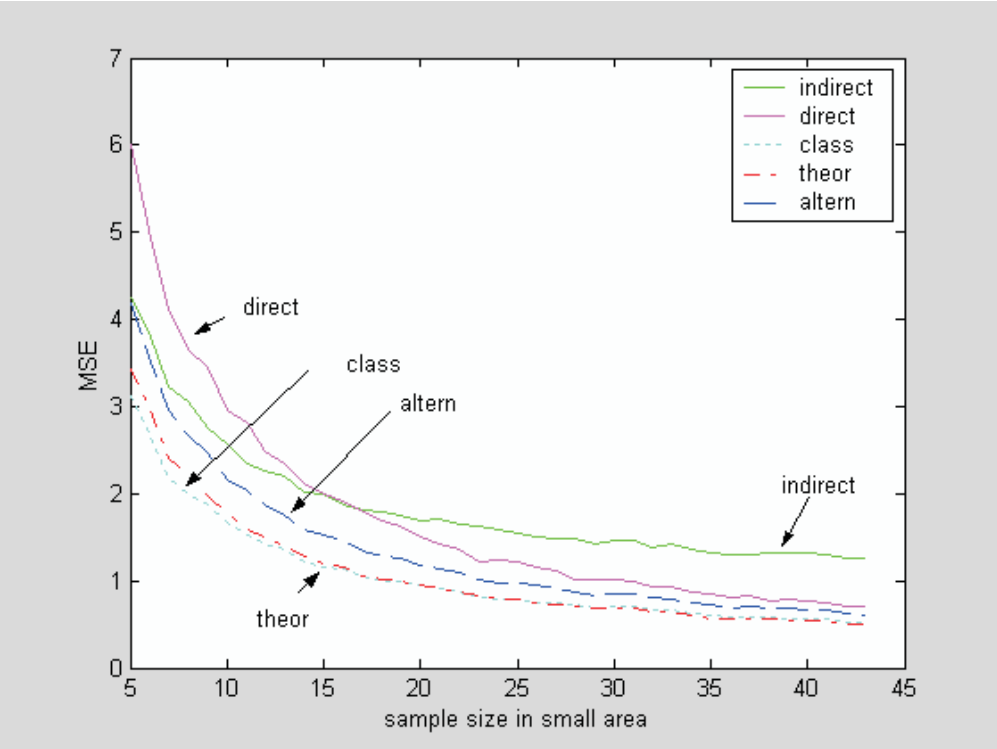

Figure 2: *MSE of the estimators as a function of the sample size n^*: non-identical within-area variances (area examined, the one with the smallest variance).*

In contrast to Figure 2, the indirect estimator exhibits behaviour quite similar to that of the optimal composite estimator (*theor*), while the two feasible composite estimators, *class* and *altern*, are close to each other.

One aspect of the population that could affect the behaviour of the different estimators is the ratio r of variance within the areas with respect to the total variation. The intra-class correlation coefficient *cci*, is equal to 1-r. Note that $cci = 0$ indicates that the entire variation comes from among the areas, while *cci*=1 indicates that the entire variation is within the areas. Figure 4 shows the variation in the MSE when we vary *cci* while maintaining the sample size constant. The population parameters used in the simulation are $b^2 = 1$ and$\sigma(x)_j^2$, constant for all the areas, with values that fluctuate between 2.5 (*cci* close to zero) and 50 (*cci* near 0.3). The sample size is n^* = 10 and the total number of areas is J= 20.

The MSEs of the different estimators now converge as *cci* (the "area effect") increases. The theoretical combined estimator (*theor*) outperforms the other estimators, even though its advantage over the direct estimator decreases toward zero as the *cci* increases. The MSE of the indirect estimator remains constant despite variation of the *cci*, and the classic composite estimator (*class*) outperforms the direct estimator and

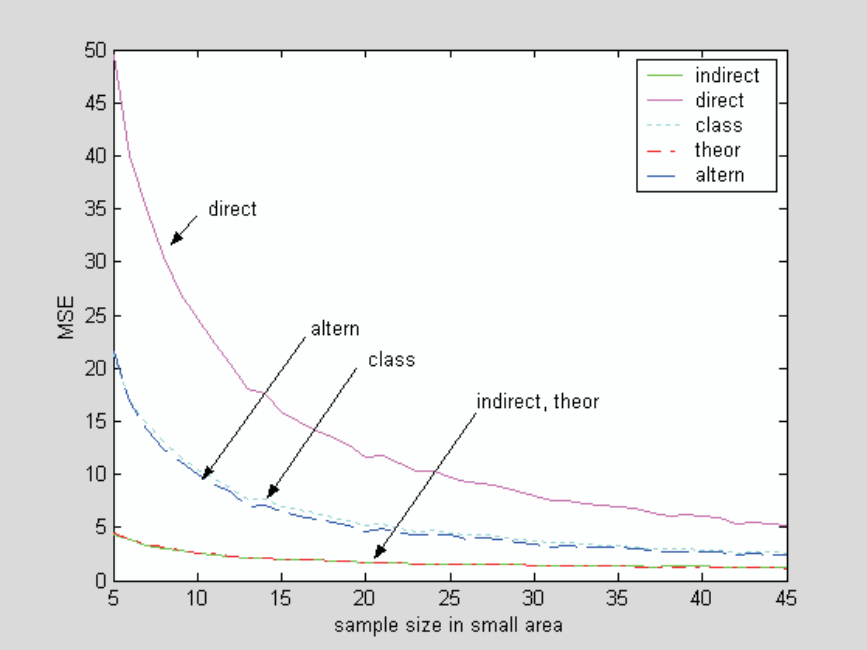

Figure 3: *Variation of the MSE of the estimators as a function of the sample size n^*, with heterogeneity in variances (area examined, the one with the greatest variance).*

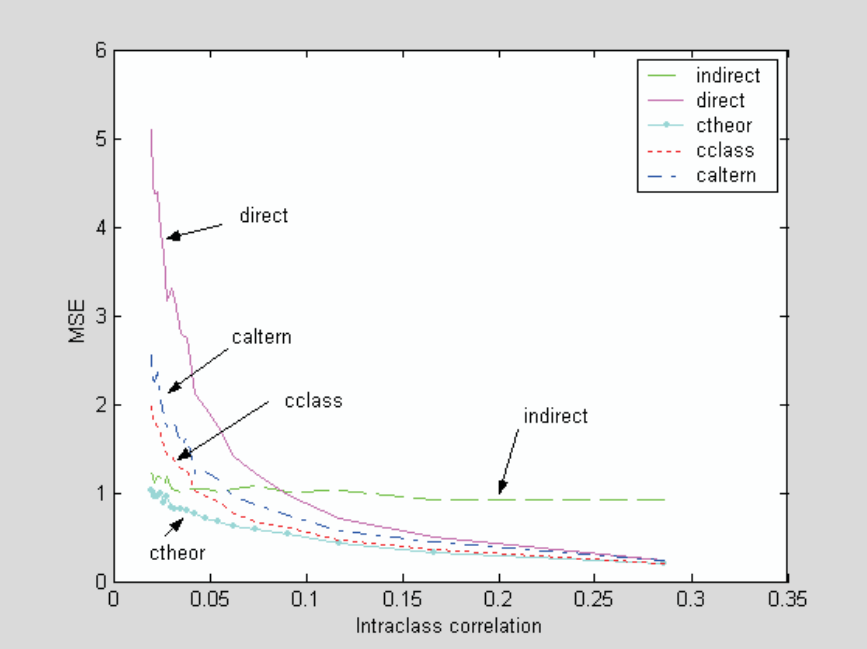

Figure 4: *Variation in the MSE of the different estimators with regard to the intra-class correlation coefficient cci.*

converges with the theoretical composite (*theor*) as the intra-class correlation increases. The alternative and classic composite estimators exhibit similar behaviour, with a slight advantage for the classic. There is a substantial difference among the MSEs of the different estimators, with the indirect clearly outperforming the direct and the classic composite estimators when *cci* is small.

Without documenting further simulations in the same detail, we note that the same conclusions were obtained when: *a*) we violate the normality assumption of the within and between area variation (letting this distribution be a chi-squared distribution with 1 degree of freedom, i.e. a highly right skewed distribution); and, *c*) when we increase the value of J of the number of areas of the population.

4 Simulation study on a real population

In this section we study the behaviour of the composite estimator through a Monte Carlo simulation in which we extract multiple samples from a known population. To do this, 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 enterprise surveyed who are registered with Social 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.

This database contains 243,184 observations from year 2000, divided into 12 groups according to the economic sector, and 41 counties (Catalan "comarques"), the location of a few enterprises was not clarified, they have been excluded from this analysis.

We have eliminated the sector-based classification and have focused solely on the division by counties. Table 1 shows the number of enterprises per county and the mean and variance of individual affiliates per enterprise. The distribution of enterprises is quite uneven, as it is mainly concentrated in densely populated areas.

Next we present the results of the simulation for four sampling designs that differ in size, and compare the behaviour of the five estimators. The sizes of the samples are 10%, 5%, 1.68 % (this is the size used by Idescat in various surveys) and 1% of the population.

4.1 Design of the Monte Carlo simulation

Let x_{j_k} be the number of salaried workers in county j and enterprise k . This is referred to as the *baseline data*. The total number of counties in Catalonia is J.

The parameters of interest are $\theta_j = \left(\sum_{k=1}^{N_j} x_{jk} \right) \Big/ N_j$, $j = 1,2,\ldots J$, the mean number of salaried workers per enterprise in each county, N_j are the numbers of surveyed enterprises in county j. With any sample we have a direct estimator $\hat{\theta}_j \sim N\left(\theta_j, var(\hat{\theta}_j)\right)$

Table 1: *Population values of area means, bias and variances.*

County	Population size	θ_j	$\left(\theta_j - \theta_*\right)^2$	$\sigma(x)_j^2$
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	18.70	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	1356.90
Baix Camp	5530	6.47	6.59	479.54
Baix Ebre	2237	6.31	7.41	534.40
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	12.90	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	246.00
Noguera	1128	5.12	15.30	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	3.70	610.20
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 number of affiliates in the whole of Catalonia, θ_*, is 9.04.
The parameter *cci* is 0.0008.

Table 2: *Results of the simulation. Medium sample size (N = 24,295).*

		Sample distribution means				Weights			Root mean square deviation				
	Sample size	**Direct**	**Theoretical composite**	**Classic composite**	**Alternative composite**	**Theoretical**	**Classic estimate**	**Alternative estimate**	**Direct**	**Indirect**	**Theoretical composite**	**Classic composite**	**Alternative composite**
Alt Camp	128	8.50	9.03	8.93	7.74	1.00	0.66	0.58	21.46	0.26	0.26	3.55	3.39
Alt Empordà	471	5.28	5.44	6.70	5.51	0.04	0.37	0.09	0.55	14.25	0.53	2.61	1.19
Alt Penedès	305	8.89	9.03	8.99	8.61	1.00	0.47	0.77	5.09	0.19	0.19	1.65	1.19
Alt Urgell	74	4.78	5.21	8.04	5.14	0.10	0.77	0.20	2.09	18.83	1.93	11.68	3.70
Alta Ribagorça	14	4.66	6.53	8.78	4.27	0.43	0.94	0.22	13.76	19.86	8.28	17.83	6.42
Anoia	326	7.76	8.58	8.37	8.08	0.64	0.45	0.63	2.13	1.53	0.85	1.03	1.52
Bages	569	8.27	8.87	8.55	8.33	0.79	0.33	0.66	2.30	0.79	0.62	1.24	0.94
Baix Camp	553	6.44	6.74	7.34	6.83	0.12	0.34	0.23	0.72	6.74	0.65	1.30	1.35
Baix Ebre	223	6.32	6.98	7.81	6.77	0.24	0.54	0.36	2.01	7.56	1.63	3.03	2.84
Baix Empordà	463	5.42	5.66	6.80	5.73	0.07	0.38	0.14	0.81	13.06	0.76	2.55	1.75
Baix Llobregat	2054	9.76	9.30	9.68	9.32	0.63	0.13	0.78	0.71	0.65	0.37	0.57	0.53
Baix Penedès	219	5.23	5.43	7.33	5.50	0.05	0.54	0.09	0.70	14.36	0.66	4.89	1.21
Barcelonès	8833	10.63	10.13	10.56	9.99	0.31	0.03	0.45	1.07	2.73	0.94	0.97	1.21
Berguedà	139	5.41	5.76	7.78	5.87	0.10	0.65	0.20	1.23	13.04	1.11	6.03	2.37
Cerdanya	78	3.73	3.90	7.76	3.95	0.03	0.76	0.06	0.88	28.46	0.86	17.04	1.59
Conca de Barberà	61	8.33	9.01	8.96	7.46	0.98	0.80	0.56	21.65	0.72	0.71	2.10	5.28
Garraf	346	6.25	6.82	7.50	6.84	0.21	0.44	0.38	1.66	7.76	1.36	2.34	2.98
Garrigues	51	5.22	5.66	8.36	5.78	0.11	0.82	0.24	1.70	14.55	1.51	10.16	3.21
Garrotxa	190	7.57	8.28	8.44	7.84	0.49	0.58	0.56	2.24	2.49	1.24	1.48	1.68
Gironès	636	9.85	9.16	9.62	9.22	0.84	0.31	0.84	2.87	0.80	0.65	1.51	1.18
Maresme	1171	6.47	6.65	7.00	6.70	0.07	0.20	0.15	0.47	6.79	0.45	0.71	0.84
Montsià	191	5.59	5.93	7.60	5.96	0.10	0.58	0.19	1.12	11.87	1.01	4.56	2.07
Noguera	112	5.09	5.30	7.83	5.40	0.05	0.69	0.09	0.70	15.44	0.66	7.86	1.11
Osona	549	7.09	7.62	7.77	7.47	0.27	0.34	0.42	1.18	3.93	0.93	1.16	1.71
Pallars Jussà	41	4.36	4.96	8.35	4.90	0.13	0.85	0.22	2.77	21.88	2.45	16.29	5.13
Pallars Sobirà	27	4.00	4.38	8.52	4.57	0.08	0.90	0.17	1.86	24.89	1.69	20.22	3.82
Pla d'Urgell	110	6.60	7.31	8.31	7.08	0.29	0.69	0.44	2.10	6.10	1.58	3.42	2.85
Pla de l'Estany	116	6.09	6.45	8.13	6.53	0.12	0.68	0.26	1.01	8.94	0.93	4.67	1.99
Priorat	25	4.16	5.27	8.57	4.00	0.23	0.90	0.12	6.92	24.38	5.48	20.25	3.48
Ribera d'Ebre	62	5.65	6.93	8.36	5.93	0.38	0.80	0.32	5.45	11.21	3.64	7.63	4.42
Ripollès	95	7.79	8.87	8.70	7.71	0.87	0.72	0.64	8.41	1.51	1.28	1.63	3.53
Segarra	59	10.88	9.07	10.04	6.89	0.98	0.80	0.43	129.05	3.54	3.46	17.53	23.75
Segrià	709	7.69	8.19	8.11	8.08	0.37	0.29	0.55	0.91	1.85	0.61	0.69	1.04
Selva	458	7.14	7.64	7.88	7.53	0.26	0.38	0.42	1.20	3.85	0.94	1.22	1.59
Solsonès	50	5.64	6.34	8.44	6.11	0.21	0.82	0.31	2.92	12.07	2.43	8.61	3.64
Tarragonès	744	9.46	9.06	9.38	9.00	0.94	0.28	0.83	2.04	0.32	0.30	1.14	0.93
Terra Alta	29	4.26	4.52	8.49	4.60	0.06	0.89	0.11	1.24	23.00	1.18	18.35	2.25
Urgell	117	6.28	6.99	8.17	6.64	0.26	0.68	0.33	2.35	7.74	1.82	4.18	2.62
Val d'Aran	50	5.30	6.34	8.41	5.51	0.28	0.83	0.27	4.96	14.22	3.71	10.25	4.53
Vallès Occidental	2668	10.31	9.80	10.20	9.76	0.40	0.10	0.63	0.97	1.90	0.72	0.82	1.09
Vallès Oriental	1179	8.43	8.83	57	8.62	0.67	0.2	0.71	0.60	0.51	0.29	0.43	0.43

Table 3: *Results of the simulation. Small sample size (N = 12,059).*

		Sample distribution means				Weights			Root mean square deviation				
	Sample size	**Direct**	**Theoretical composite**	**Classic composite**	**Alternative composite**	**Theoretical**	**Classic estimate**	**Alternative estimate**	**Direct**	**Indirect**	**Theoretical composite**	**Classic composite**	**Alternative composite**
Alt Camp	64	8.32	9.04	8.99	7.44	1.00	0.73	0.56	39.34	0.45	0.45	6.22	5.24
Alt Empordà	236	5.27	5.58	7.04	5.55	0.08	0.45	0.14	1.14	14.48	1.06	4.21	1.93
Alt Penedès	153	8.88	9.04	9.02	8.30	1.00	0.55	0.69	10.42	0.37	0.37	2.81	2.51
Alt Urgell	37	4.77	5.56	8.26	5.04	0.19	0.81	0.23	4.51	19.06	3.72	13.53	4.60
Alta Ribagorça	7	4.73	7.31	8.86	3.92	0.60	0.95	0.20	30.84	20.10	12.38	18.77	7.15
Anoia	163	7.67	8.74	8.45	7.76	0.78	0.54	0.58	4.30	1.73	1.19	1.67	2.22
Bages	285	8.29	8.95	8.66	8.17	0.88	0.41	0.64	5.19	0.98	0.86	2.26	1.84
Baix Camp	177	6.39	6.94	7.54	6.84	0.21	0.42	0.33	1.46	6.95	1.16	2.14	2.18
Baix Ebre	112	6.40	7.43	8.06	6.74	0.39	0.62	0.42	4.52	7.78	3.02	4.44	3.90
Baix Empordà	232	5.41	5.86	7.12	5.66	0.12	0.46	0.17	1.65	13.29	1.46	4.07	2.28
Baix Llobregat	1027	9.74	9.20	9.66	9.27	0.77	0.18	0.80	1.47	0.83	0.60	1.04	0.81
Baix Penedès	110	5.26	5.63	7.64	5.64	0.10	0.62	0.18	1.49	14.59	1.36	6.75	2.47
Barcelonès	4417	10.67	9.89	10.56	9.66	0.48	0.05	0.69	2.25	2.89	1.65	1.88	1.98
Berguedà	70	5.42	6.06	8.03	5.97	0.18	0.71	0.30	2.47	13.26	2.07	7.65	4.04
Cerdanya	39	3.72	4.05	8.04	4.06	0.06	0.81	0.12	1.76	28.71	1.67	19.70	3.43
Conca de Barberà	31	8.24	9.03	9.04	7.06	0.99	0.84	0.53	41.31	0.92	0.90	3.56	7.64
Garraf	173	6.30	7.23	7.79	6.67	0.34	0.52	0.40	3.55	7.98	2.50	3.74	3.70
Garrigues	26	5.25	6.03	8.52	5.88	0.21	0.86	0.35	3.57	14.78	2.88	11.42	5.25
Garrotxa	95	7.50	8.50	8.53	7.70	0.65	0.65	0.58	4.13	2.69	1.66	2.02	2.30
Gironès	318	9.86	9.11	9.59	9.05	0.91	0.39	0.80	6.60	0.97	0.88	2.85	2.61
Maresme	586	6.48	6.83	7.21	6.85	0.13	0.27	0.27	1.03	7.00	0.93	1.46	1.77
Montsià	96	5.59	6.21	7.87	6.01	0.18	0.65	0.27	2.43	12.10	2.02	6.20	3.31
Noguera	56	5.07	5.46	8.07	5.52	0.10	0.75	0.20	1.59	15.67	1.41	9.59	2.72
Osona	275	7.09	7.92	7.96	7.42	0.43	0.42	0.47	2.62	4.13	1.64	2.06	2.23
Pallars Jussà	21	4.39	5.42	8.49	4.57	0.22	0.88	0.20	6.02	22.13	4.77	17.76	4.78
Pallars Sobirà	14	3.97	4.67	8.61	4.58	0.14	0.91	0.25	3.73	25.14	3.15	21.34	6.46
Pla d'Urgell	55	6.64	7.73	8.48	6.99	0.45	0.75	0.49	4.93	6.32	2.83	4.40	3.84
Pla de l'Estany	58	6.05	6.71	8.31	6.67	0.22	0.74	0.39	2.20	9.16	1.75	5.76	3.42
Priorat	13	4.22	5.97	8.67	3.92	0.36	0.92	0.13	14.44	24.63	9.36	21.52	3.61
Ribera d'Ebre	31	5.64	7.51	8.49	5.73	0.55	0.84	0.34	11.62	11.44	5.77	8.86	5.75
Ripollès	48	7.83	8.95	8.84	7.27	0.93	0.77	0.57	17.61	1.71	1.57	2.67	6.43
Segarra	30	10.65	9.06	10.52	6.59	0.99	0.84	0.43	248.74	3.70	3.67	50.61	31.13
Segrià	355	7.66	8.41	8.22	8.00	0.54	0.37	0.60	1.85	2.05	0.98	1.27	1.58
Selva	229	7.15	7.94	8.08	7.53	0.42	0.46	0.50	2.61	4.06	1.61	2.10	2.25
Solsonès	25	5.62	6.80	8.57	5.93	0.35	0.86	0.34	6.60	12.30	4.39	9.70	5.02
Tarragonès	372	9.40	9.05	9.31	8.90	0.97	0.36	0.81	4.04	0.50	0.48	1.93	1.64
Terra Alta	15	4.30	4.79	8.61	4.92	0.11	0.91	0.23	2.65	23.24	2.41	19.64	5.08
Urgell	59	6.40	7.48	8.39	6.64	0.41	0.74	0.40	5.36	7.95	3.37	5.46	3.63
Val d'Aran	25	5.35	6.95	8.56	5.43	0.43	0.86	0.31	10.37	14.45	6.17	11.58	5.50
Vallès Occidental	1334	10.30	9.58	10.14	9.53	0.57	0.15	0.78	1.89	2.06	1.15	1.50	1.63
Vallès Oriental	590	8.42	8.92	8.64	8.52	0.81	0.27	0.70	1.31	0.70	0.50	0.85	0.73

Table 4: *Results of the simulation. Very small sample size (N = 4,100).*

		Sample distribution means				Weights			Root mean square deviation				
	Sample size	**Direct**	**Theoretical composite**	**Classic composite**	**Alternative composite**	**Theoretical**	**Classic estimate**	**Alternative estimate**	**Direct**	**Indirect**	**Theoretical composite**	**Classic composite**	**Alternative composite**
Alt Camp	22	8.49	9.05	9.40	6.84	1.00	0.78	0.53	132.58	1.20	1.20	37.69	9.07
Alt Empordà	79	5.39	6.16	7.51	5.60	0.21	0.54	0.24	3.85	15.35	3.23	7.64	3.66
Alt Penedès	51	8.98	9.05	9.19	7.92	1.00	0.63	0.66	34.77	1.11	1.11	8.35	6.22
Alt Urgell	13	4.68	6.40	8.39	4.84	0.39	0.84	0.27	13.25	19.95	7.97	15.72	6.17
Alta Ribagorça	2	5.03	8.41	9.08	3.35	0.84	0.97	0.19	134.57	21.00	18.81	23.45	9.89
Anoia	55	7.62	8.93	8.65	7.34	0.91	0.62	0.56	14.45	2.51	2.14	4.24	4.31
Bages	96	8.44	9.03	8.89	7.82	0.96	0.50	0.64	16.30	1.75	1.67	6.29	3.28
Baix Camp	93	6.52	7.63	7.92	6.70	0.44	0.51	0.42	4.67	7.78	3.09	4.53	3.49
Baix Ebre	38	6.33	8.11	8.29	6.36	0.65	0.69	0.44	14.19	8.61	5.54	7.22	5.29
Baix Empordà	78	5.51	6.56	7.55	5.64	0.30	0.54	0.24	5.68	14.15	4.22	7.66	3.14
Baix Llobregat	346	9.74	9.12	9.65	8.97	0.91	0.25	0.81	4.67	1.54	1.38	2.89	1.98
Baix Penedès	37	5.27	6.20	7.97	5.71	0.25	0.69	0.32	4.14	15.46	3.30	9.46	4.73
Barcelonès	1490	10.69	9.49	10.46	9.22	0.73	0.08	0.90	6.77	3.58	3.34	4.84	3.59
Berguedà	24	5.31	6.76	8.21	5.77	0.39	0.76	0.37	7.24	14.13	4.62	9.97	5.83
Cerdanya	13	3.70	4.57	8.25	3.87	0.16	0.84	0.14	5.08	29.63	4.36	22.75	4.20
Conca de Barberà	10	7.75	9.05	9.16	6.17	1.00	0.87	0.47	105.46	1.68	1.67	11.31	14.50
Garraf	58	6.25	7.96	8.08	5.99	0.61	0.61	0.33	10.89	8.82	4.92	6.72	4.34
Garrigues	9	5.23	6.87	8.64	5.47	0.43	0.88	0.38	9.86	15.65	6.03	13.24	7.51
Garrotxa	32	7.44	8.81	8.68	7.24	0.85	0.72	0.58	11.73	3.48	2.74	3.74	4.84
Gironès	107	9.85	9.08	9.63	8.63	0.97	0.48	0.75	21.02	1.68	1.62	7.42	5.34
Maresme	198	6.56	7.35	7.52	6.72	0.32	0.35	0.35	3.39	7.83	2.59	3.64	2.56
Montsià	32	5.66	7.00	8.15	5.86	0.40	0.72	0.35	7.31	12.96	4.84	8.83	4.95
Noguera	19	5.09	6.05	8.29	5.82	0.24	0.80	0.39	4.97	16.55	3.78	11.99	6.13
Osona	93	7.16	8.46	8.20	7.06	0.69	0.51	0.49	7.86	4.94	3.25	4.52	3.45
Pallars Jussà	7	4.20	6.44	8.60	4.52	0.46	0.90	0.28	15.43	23.03	9.00	19.66	7.57
Pallars Sobirà	5	4.06	5.61	8.70	4.20	0.31	0.92	0.26	11.54	26.05	8.02	23.04	8.03
Pla d'Urgell	19	6.65	8.35	8.63	6.34	0.71	0.80	0.46	13.77	7.14	4.84	6.27	6.31
Pla de l'Estany	20	6.11	7.43	8.51	6.34	0.45	0.79	0.45	7.11	10.00	4.24	7.85	5.50
Priorat	4	4.08	7.31	8.77	3.92	0.65	0.94	0.20	41.22	25.54	15.93	23.59	6.64
Ribera d'Ebre	10	5.54	8.32	8.67	5.13	0.79	0.87	0.35	36.97	12.29	9.32	11.60	8.24
Ripollès	16	7.66	9.02	8.94	6.28	0.98	0.82	0.47	54.19	2.49	2.40	6.01	11.37
Segarra	10	11.06	9.06	11.92	5.99	1.00	0.87	0.45	825.17	4.37	4.37	345.24	33.31
Segrià	120	7.72	8.76	8.43	7.72	0.78	0.45	0.61	5.97	2.83	2.06	3.31	3.08
Selva	77	7.18	8.46	8.32	7.20	0.68	0.55	0.52	7.93	4.87	3.13	4.34	3.83
Solsonès	9	5.57	7.64	8.67	5.56	0.59	0.88	0.38	18.29	13.16	7.68	11.34	7.02
Tarragonès	125	9.33	9.06	9.33	8.29	0.99	0.44	0.71	12.31	1.23	1.21	5.01	4.79
Terra Alta	5	4.35	5.57	8.71	4.73	0.26	0.92	0.32	8.88	24.15	6.69	21.40	8.06
Urgell	20	6.37	8.18	8.58	6.43	0.67	0.79	0.47	15.96	8.79	5.87	7.82	5.99
Val d'Aran	8	5.28	7.94	8.71	4.85	0.71	0.89	0.29	29.89	15.32	10.16	14.02	7.18
Vallès Occidental	450	10.27	9.30	10.09	9.26	0.80	0.21	0.80	6.13	2.76	2.25	4.20	3.26
Vallès Oriental	200	8.37	9.00	8.73	8.32	0.92	0.35	0.70	4.06	1.46	1.29	2.45	1.79

Table 5: *Results of the simulation. Very small sample size (N = 2,431).*

		Sample distribution means				Weights			Root mean square deviation				
	Sample size	**Direct**	**Theoretical composite**	**Classic composite**	**Alternative composite**	**Theoretical**	**Classic estimate**	**Alternative estimate**	**Direct**	**Indirect**	**Theoretical composite**	**Classic composite**	**Alternative composite**
Alt Camp	13	8.22	9.09	9.47	6.59	1.00	0.79	0.51	184.43	2.16	2.16	65.13	13.40
Alt Empordà	47	5.33	6.49	7.64	5.54	0.31	0.57	0.27	6.80	16.59	4.98	9.83	4.27
Alt Penedès	31	8.98	9.09	9.33	7.50	1.00	0.65	0.60	59.28	2.06	2.06	15.00	8.95
Alt Urgell	7	4.63	7.07	8.55	4.37	0.55	0.87	0.24	27.31	21.24	11.92	18.19	7.16
Alta Ribagorça	1	4.75	8.71	9.31	9.09	0.91	0.97	1.00	235.46	22.29	20.38	34.97	22.29
Anoia	33	7.63	9.02	8.74	7.00	0.95	0.64	0.51	23.69	3.55	3.23	6.39	6.53
Bages	57	8.43	9.08	8.97	7.57	0.97	0.53	0.61	24.59	2.75	2.66	8.35	4.94
Baix Camp	55	6.54	8.00	8.09	6.67	0.57	0.54	0.44	8.08	8.93	4.60	6.77	4.32
Baix Ebre	22	6.52	8.49	8.51	6.08	0.77	0.71	0.41	28.98	9.77	7.68	10.93	7.91
Baix Empordà	46	5.58	7.05	7.76	5.72	0.42	0.57	0.31	9.77	15.38	6.37	10.23	4.48
Baix Llobregat	205	9.76	9.13	9.70	8.93	0.94	0.28	0.77	8.80	2.43	2.27	4.95	3.12
Baix Penedès	22	5.24	6.60	8.10	5.64	0.35	0.71	0.34	6.24	16.71	4.62	11.36	5.37
Barcelonès	883	10.78	9.40	10.44	9.22	0.82	0.10	0.92	12.40	4.39	4.64	7.77	4.44
Berguedà	14	5.42	7.34	8.37	5.46	0.52	0.78	0.36	13.25	15.36	7.17	12.15	6.90
Cerdanya	8	3.75	5.03	8.35	3.87	0.24	0.85	0.16	8.90	31.00	7.04	24.76	4.93
Conca de Barberà	6	7.68	9.09	9.40	5.61	1.00	0.88	0.41	182.38	2.68	2.67	28.73	20.25
Garraf	35	6.15	8.27	8.18	5.83	0.72	0.63	0.33	18.47	9.98	6.54	9.83	5.87
Garrigues	5	5.31	7.48	8.74	5.09	0.57	0.89	0.34	19.47	16.90	9.02	15.11	8.51
Garrotxa	19	7.50	8.94	8.81	7.07	0.90	0.74	0.57	20.45	4.54	3.91	5.91	7.06
Gironès	64	9.94	9.11	9.67	8.37	0.98	0.51	0.70	36.91	2.56	2.51	11.68	8.21
Maresme	117	6.51	7.64	7.66	6.71	0.44	0.38	0.40	5.05	8.98	3.48	5.20	3.46
Montsià	19	5.66	7.46	8.30	5.81	0.52	0.74	0.38	12.11	14.17	6.72	11.10	6.30
Noguera	11	5.11	6.53	8.42	5.57	0.36	0.82	0.38	8.06	17.80	5.55	13.97	7.00
Osona	55	7.12	8.68	8.32	6.97	0.79	0.54	0.50	13.97	6.04	4.48	7.10	5.05
Pallars Jussà	4	4.24	7.15	8.72	4.04	0.60	0.91	0.25	26.85	24.34	12.63	22.06	8.56
Pallars Sobirà	3	4.17	6.28	8.79	3.95	0.43	0.93	0.24	22.26	27.39	12.54	25.07	10.03
Pla d'Urgell	11	6.60	8.61	8.77	6.25	0.81	0.82	0.47	22.90	8.28	6.30	8.58	7.49
Pla de l'Estany	12	6.17	7.85	8.62	6.26	0.58	0.80	0.46	12.61	11.18	6.16	9.67	6.87
Priorat	3	4.14	7.67	8.79	3.80	0.71	0.93	0.21	59.33	26.87	18.82	25.34	7.19
Ribera d'Ebre	6	5.40	8.59	8.74	4.77	0.86	0.88	0.31	56.59	13.50	11.14	14.75	9.97
Ripollès	10	7.91	9.08	9.15	5.87	0.98	0.83	0.42	93.34	3.52	3.45	11.40	15.94
Segarra	6	11.22	9.10	12.60	5.71	1.00	0.88	0.43	1278.82	5.16	5.15	701.02	36.85
Segrià	71	7.77	8.90	8.58	7.55	0.86	0.48	0.60	10.14	3.87	3.14	5.34	4.28
Selva	46	7.19	8.68	8.44	7.00	0.78	0.57	0.51	12.88	5.96	4.38	6.36	5.14
Solsonès	5	5.36	8.07	8.76	5.19	0.73	0.89	0.35	24.51	14.38	9.10	13.53	8.64
Tarragonès	74	9.37	9.10	9.39	8.21	0.99	0.48	0.71	20.14	2.13	2.12	7.74	6.64
Terra Alta	3	4.25	6.04	8.79	4.54	0.37	0.93	0.32	14.08	25.47	9.03	23.09	10.89
Urgell	12	6.40	8.49	8.67	6.11	0.77	0.80	0.45	25.50	9.95	7.45	9.61	6.99
Val d'Aran	5	5.19	8.29	8.83	4.63	0.79	0.89	0.28	49.26	16.56	12.34	17.78	8.46
Vallès Occidental	267	10.28	9.25	10.08	9.07	0.87	0.24	0.80	10.44	3.59	3.20	6.37	4.43
Vallès Oriental	118	8.40	9.06	8.84	8.15	0.95	0.38	0.67	6.91	2.44	2.26	4.23	2.62

for each county j and an indirect estimator $\hat{\theta}_* \sim N\left(\theta_*, var(\hat{\theta}_*)\right)$, which is common to all the counties.

If the variance of x_{jk} is $\sigma(x)_j^2$, then $var(\hat{\theta}_j) = \sigma(x)_j^2 \Big/ n_j$, where n_j is the number of sample observations in county j.

Our simulation exercise allows us to develop an optimal *theoretical composite* estimator, since we can evaluate expression (1).

We also evaluate a *classic composite* estimator and an *alternative composite* estimator as defined in Section 2.

We replicate 1,000 proportional samples from the enterprise census and apply the five estimators. The results are summarised in Tables 2 to 5.

4.2 Results of the simulation

Tables 2 through 5 summarise the results of the simulations for four scenarios. These scenarios differ in the sample size. In Table 2, the sample size is large, ***large*** sample size: precisely 24,295 observations in each total sample, which corresponds to 10% of the population. In Table 3, 5% of the population is sampled, resulting in 12,059 sample observations (***medium***-sized). The third sample represents slightly more than 1.68% of the population, yielding an average of 100 county observations (***small*** sized). However, the sample was extracted proportionally and the observations per county are distributed between a minimum of two in the county of *Alta Ribagorça* and a maximum of 1,490 in the county of *el Barcelonès*. The total number of observations from Catalonia is 4,100. Table 5 shows the fourth sample, which represents 1% of the population (***very small*** sized). The total number of observations is only 2,431.

To illustrate the form of the distribution of the MSE across counties, in Figures 5 we show the distributions of the MSEs for the four feasible estimators and two contrasting sample sizes: n =24295, a large sample; and n = 4100, a small sample. Overall we can say that these distributions of MSEs have the following common characteristics:

1. They are asymmetrical
2. They have extreme values (very noticeable in the case of the direct estimator)
3. They reveal a high degree of variation

These characteristics make it difficult to evaluate the different estimators based solely on their mean MSEs, especially given the presence of skew distributions and extreme values. For this reason we have decided to mix three comparison criteria, allowing us to make a more refined evaluation than just comparing simple means. These criteria are:

1. Comparison of the mean MSEs .
2. Comparison of the median of MSEs.
3. Comparison of the percentage of counties with lower MSEs (this criterion will be used for each pair of estimators)

Table 6: *Statistics on the distribution of the MSEs for each estimator, by sample size.*

ESTIMATORS (n = 24,295)	**direct**	**indirect**	**com teor**	com clas	**com alt**
mean	6.44	9.14	1.48	5.98	2.89
variance	399.01	64.14	2.33	39.22	12.88
average	1.86	7.56	0.94	3.03	1.99
Minimum value	0.47	0.19	0.19	0.43	0.43
Maximum value	129.05	28.46	8.28	20.25	23.75

ESTIMATORS (n = 12,059)	**direct**	**indirect**	**com teor**	com clas	**com alt**
mean	12.82	9.35	2.48	7.98	4.16
variance	1,478.30	64.50	5.65	83.73	21.12
average	3.73	7.78	1.65	4.40	3.42
Minimum value	1.03	0.37	0.37	0.85	0.73
Maximum value	248.74	28.71	12.38	50.61	31.13

ESTIMATORS (n = 4,100)	**direct**	**indirect**	**com teor**	com clas	**com alt**
mean	41.45	10.17	4.78	18.57	6.35
variance	16,316.69	65.45	13.88	2,723.66	24.62
average	11.54	8.61	3.78	7.82	5.34
Minimum value	3.39	1.11	1.11	2.45	1.79
Maximum value	825.17	29.63	18.81	345.24	33.31

ESTIMATORS (n = 2,431)	**direct**	**indirect**	**com teor**	com clas	**com alt**
mean	66.38	11.29	6.48	30.91	8.33
variance	39,254.52	67.72	18.23	11,342.14	36.74
average	20.14	9.77	5.15	11.10	6.99
Minimum value	5.05	2.06	2.06	4.23	2.62
Maximum value	1,278.82	31.00	20.38	701.02	36.85

In Tables 6 and 7 the results of the synthesis can be seen, along with other complementary data, allowing the estimators to be evaluated. Based on the tables, we conclude:

1. For all sample sizes and for any of the three criteria used, the best estimator is the **theoretical composite** estimator. This result is as expected. Although not so important in practice, since this estimator is not accessible in real life applications. It is useful as a benchmark.
2. The best estimator among the four feasible ones is the **alternative composite**. For the four sample sizes and the three evaluation criteria (twelve combinations), the alternative composite estimator is better. The only exception to this is when we have a large sample size and we use the criterion of the counties with lowest MSE. In that case, the direct estimator is better than the alternative composite estimator.

Table 7: *Comparison of the estimators according to the criterion based on the percentage of counties with best MSE*^{2}.

n=24,295	direct	indirect	com teor	com clas	com alt
direct		73.17	0.00	60.98	60.98
indirect	26.83		0.00	19.51	19.51
com teor	100.00	100.00		100.00	95.12
com clas	39.02	80.49	0.00		34.15
com alt	39.02	80.49	4.88	65.85	

n=12,059	direct	indirect	com teor	com clas	com alt
direct		65.85	0.00	48.78	36.59
indirect	34.15		0.00	24.39	24.39
com teor	100.00	100.00		100.00	90.24
com clas	51.22	75.61	0.00		26.83
com alt	63.41	75.61	9.76	73.17	

n=4,100	direct	indirect	com teor	com clas	com alt
direct		36.59	0.00	34.15	4.88
indirect	63.41		0.00	39.02	36.59
com teor	100.00	100.00		100.00	70.73
com clas	65.85	60.98	0.00		12.20
com alt	95.12	63.41	29.27	87.80	

n=2,431	direct	indirect	com teor	com clas	com alt
direct		26.19	0.00	19.05	0.00
indirect	73.81		7.14	52.38	38.10
com teor	100.00	92.86		97.62	54.76
com clas	80.95	47.62	2.38		7.14
com alt	100.00	61.90	45.24	92.86	

Only if we grant this last criterion as much importance as the other two criteria, or more, can we say that for larger sample sizes the direct estimator is better. This specific advantage in one criterion disappears with the medium sample size (N = 12,059), so that in general the conclusion that the alternative composite estimator is better is warranted.

3. The **direct estimator** exhibits acceptable behaviour for the largest sample size, but its performance declines as sample size is reduced. In effect, for the large sample size (N =24,295), the direct estimator is the best according to the criterion of percentage, the second best according to the criteria of the average, and the third best according to the criteria of the mean MSE (it is surpassed by the two composite estimators, both classic and alternative). Its performance declines considerably in small samples since it is the second best according to the criterion of the average for medium-sized samples (N = 12,059), the third best according to the criteria of percentage and the worst according to the criteria of the mean of the MSEs. For small and very small samples, the direct estimator performs worse than any other estimator for all three criteria.
4. The **classic composite** is the one usually used in small areas. It is an estimator that always performs worse than the alternative composite, but it exhibits certain

interesting results in relation to the other estimators. In brief, for large sample sizes it is better than the indirect, while for small sample sizes it is better than the direct. For medium-sized samples it most likely obtains the best-combined results, since (if we keep aside the alternative estimator) it performs the best in both average and percentage. For small sample sizes it competes with the indirect, since for the small sample the indirect performs better on the average and percentage criteria, but worse on the MSE mean criterion. For the smallest sample size, it is clearly outperformed by the indirect estimator.

5. The last estimator to be examined is the **indirect estimator**, or the synthetic estimator. This "naive" estimator shows its qualities in small samples. Although it performs the worst of all the estimators for samples larger than 10,000, in the sample containing 4,100 it outperforms the direct estimator according to all three criteria used, and in the smallest sample it is the best estimator after the alternative composite.

5 Conclusions and research programme

The following general conclusions can be drawn from the Monte Carlo studies on artificial and real populations.

a) When the within-area variances are identical, the differences among the MSEs of the estimators examined are great when the (area) sample size is small, and tend to disappear in large sample sizes, although the indirect estimator shows a lesser degree of variation as a result of varying the sample size. Thus, there is a direct relationship between the sample size and convergence of the MSEs of the estimators.

b) When the within-area variances are identical, the differences among the MSEs of the direct, theoretical composite, classic composite and alternative composite estimators is large in the case of small intra-class correlation, but it disappears as the intra-class correlation increases. Thus, there is a direct relationship between the intra-class correlation and convergence of the MSEs of the estimators. An increase in the intra-class correlation does not lead to a reduction of the MSE in the indirect estimator.

c) As the sample size increases, the behaviour of the MSEs of the indirect estimator reflects a rate much lower than that of the other estimators, both in the improvements in its estimates and in its convergence with the rest. The greatest improvement when faced with increases in sample sizes is that of the direct estimator. The composite estimators have intermediate rates. Thus, each estimator has a different degree of sensitivity to increases in sample sizes.

d) There is a sample size below which the indirect estimator (or synthetic estimator), which uses information from all the areas, is the best alternative for

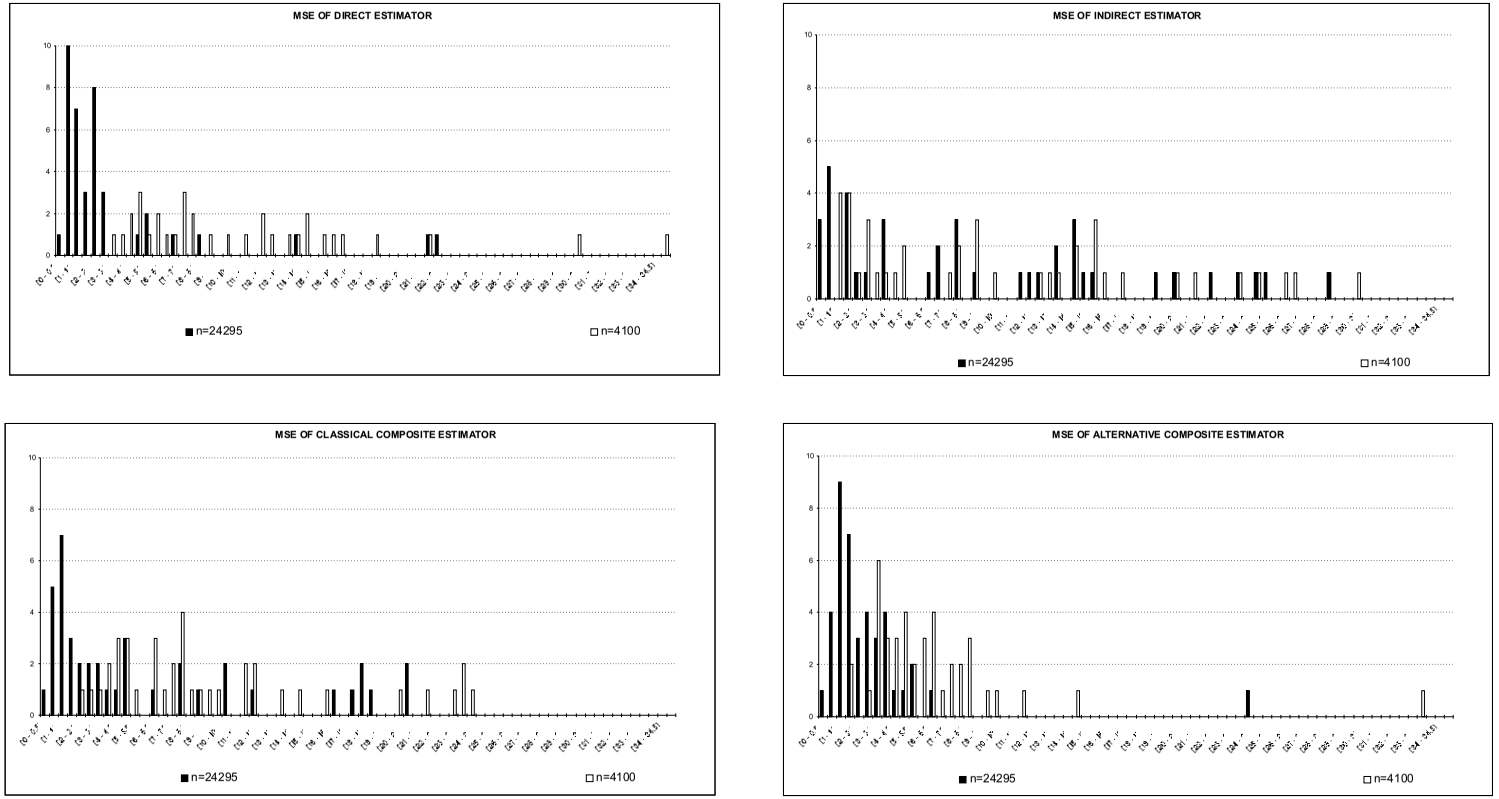

Figure 5: *MSE of the feasible estimators.*

estimating a parameter in a small area. In the real population examined in this study, below a certain sample size (specifically, the very small sample size) the best alternative to estimate the mean of a specific area, or the means of all the areas, is the indirect or synthetic estimator.

e) In the real population examined, the alternative composite estimator achieves the same degree of precision as a direct estimator with a sample size that is four times larger. In general, this estimator presents the best performance with regard to the MSEs for almost all the sample dimensions examined and for the different criteria applied.

f) For small or very small samples, in the empirical population studied, the direct estimator exhibits the worst performance with regard to the MSE. Thus, each of composite estimators considered performs better than the direct estimate.

As extensions to the present study, which constitute a research programme for the immediate future, we shall examine a series of points grouped in two different sections: theoretical developments and simulations, and applications:

Theoretical developments and simulations

1. Estimates of inter-annual variation rates: we wish to replicate the evaluation of the five estimators studied when we examine the most important type of statistics for economic analysis, inter-annual variation rates. This extension could present surprising conclusions given the complexity of the variances of these indicators.
2. Analysis of the estimated weighting factors of the composite estimators: to better understand the comparative performance of the various composite estimator, it would be interesting to analyse in what way the weighting factor estimates are distributed compared to the theoretical weighting factors.
3. Sampling design for small area estimators: how should the sample size n be allocated to the areas when small area estimation is considered? Answering this question through some theoretical development or through simulations is highly relevant to the practical work carried out by statistical organisations that need to provide information both at the area and country-level. In the initial phase it could be enlightening to compare proportional allocation (used here) with fixed and optimal classic allocation (depending on the variances in each stratum).

Practical applications

1. County-level estimates of unemployment rates: in addition to their intrinsic interest for territorial economic analysis, these rates have at least three additional features: we can use sources we have already worked with and with which we are familiar, such as the INE's EPA; this is one of the surveys that has drawn the attention of

recent international literature on small data estimation (Datta *et al.*, 1999), and finally, we will soon have census data on county-wide unemployment when the 2001 census information is disseminated.

2. County-level estimates of the use of ICT (Information and Communication Technology): Idescat is currently researching this topic through a biannual survey undertaken since 2000, with samples slightly under 4,000 families. Currently, the Secretariat of the Information Society of the Generalitat de Catalunya (the sponsor of these surveys) has requested Idescat to generate a series of county-level estimates; this is therefore a natural point to begin applying small area estimators in official statistics.
3. Estimates of the IPI (Industrial Production Index) for Catalonia and its counties: IPI is a fundamental anchor in short-term economic analyses, and constitutes the first experience Idescat has had with small area estimation (the IPI for Catalonia), using a methodology that was later temporarily adopted by INE for all the autonomous regions within Spain. It is a case in which inter-annual variation rates are applied. In the future we will attempt to apply small area estimators to disaggregate the general IPI index provided by INE as of 2003 for Catalonia in two directions: by industrial sectors and by counties.

Acknowledgements

The authors are grateful to the following statisticians from Institut d'Estadística de Catalunya: Xavier López, Maribel García, Cristina Rovira and Antoni Contel for their help at several stages of this paper. We are also indebted to Nick T. Longford for detail comments on a previous version of this paper.

6 References

Clar, M., Ramos, R. 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.

Cressie, N. (1995). Bayesian smoothing of rates in small geographic areas. *Journal of Regional Science*, 35 (4), 659-73.

Datta, G. S., *et al.* (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94 (448), 1.074-82.

Farrell, P. J., Macgibbon, B. and Tomberlin, T. J. (1997). Empirical Bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business & Economic Statistics*, 15 (1), 101-8.

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), 435-41.

Longford, N. T. (2001). Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited, *Statistics in Medicine*, 20, 3.189-3.203.

Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9 (1), 73-84.

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.

Raghunathan, T. E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, 88 (424), 1444-48.

Singh, M. P., Gambino, J. and Mantel, H. J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 1, Statistics Canada, 3-22.

Singh, A. C., Mantel, H. J. and Thomas, B. W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 1, Statistics Canada, 33-43.

Singh, A. C., Stukel, D. M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, b, 60, 377-396.

Thomas, N., Longford, N. T. and Rolph, J. E. (1994). Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in Medicine*.

Resum

Aquest paper compara cinc estimadors d'àrea petita. S'utilitzen mètodes de Monte Carlo primer en un context de població artificial i després en un context de població real. Juntament amb els estimadors directe i indirecte, es considera un estimador compost òptim amb pesos que són funció de valors poblacionals, i dos estimadors compostos amb pesos estimats: un que assumeix homogeneïtat de variàncies dintre àrees i biaix al quadrat, i un altre que considera estimadors específics de variància i biaix. En l'estudi basat en població real, s'observa que entre els estimadors factibles, el millor és aquell que empra estimadors específics de variància i biaix.

MSC: 62J07, 62J10, 62H12

Paraules clau: Estadística regional, àrea petita, error quadràtic mitjà, estimadors directe, indirecte i compost