On the performance of small-area estimators: fixed vs. random area parameters *

Alex Costa

Statistical Institute of Catalonia (IDESCAT)

Albert Satorra and Eva Ventura

Department of Economics and Business
Universitat Pompeu Fabra

February 19, 2008

*Detailed and very helpful comments by Nicholas T. Longford on a previous version of this paper are acknowledged.

Abstract

Most methods for small-area estimation are based on composite estimators derived from design- or model-based methods. A composite estimator is a linear combination of a direct and an indirect estimator with weights that usually depend on unknown parameters which need to be estimated. Although model-based small-area estimators are usually based on random-effects models, the assumption of fixed effects is at face value more appropriate. Model-based estimators are justified by the assumption of random (interchangeable) area effects; in practice, however, areas are not interchangeable. In the present paper we empirically assess the quality of several small-area estimators in the setting in which the area effects are treated as fixed. We consider two settings: one that draws samples from a theoretical population, and another that draws samples from an empirical population of a labor force register maintained by the National Institute of Social Security (NISS) of Catalonia. We distinguish two types of composite estimators: a) those that use weights that involve area specific estimates of bias and variance; and, b) those that use weights that involve a common variance and a common squared bias estimate for all the areas. We assess their precision and discuss alternatives to optimizing composite estimation in applications.

KEY WORDS: small area estimation, composite estimator, Monte Carlo study, random effect model, BLUP, empirical BLUP.

AMS CLASSIFICATION (FSC 2000): 62G10,62J02.

1 Introduction

Sample surveys are often used to estimate quantities related not only to the total population but also to a variety of small-area domains. Small-area estimation is concerned with estimating population quantities associated with a partition of the domain (population) into subdomains (small areas or districts) $j = 1, \ldots, J$. Nowadays there is a large body of methodology for small-area estimation; see, e.g., Platek, Rao, Särndal and Singh (1987), Isaki (1990), Ghosh and Rao (1994), Singh, Gambino and Mantel (1994), and Rao (2003).

Large-scale (national) surveys are usually designed to yield estimates of a small number of key national population quantities (means, proportions and the like) that have sufficient precision, without having to adopt any assumptions other than the sampling design. Insisting on a large sample for each district is not realistic, especially when there are many districts, and several of them form a very small fraction of the population.

When estimating a domain quantity, we refer to a *direct estimator* if it is based only on the domain-specific sample. A domain (area) is regarded as small if the direct estimate for the area does not have adequate precision. For a small area one could use *indirect estimators* that borrow strength from values of the variable of interest from related areas and/or time periods. An implicit or explicit model is used to link the different areas and/or time periods, often through the use of auxiliary information such as a census count or some administrative records. An initial classification of small-area estimation divides the methods into *design-based* and *model-based*.

Design-based methods are based solely on the sampling design and do not make use of distributional (model) assumptions about the observed variables. Sampling variation, that is, variation across hypothetical replications of drawing a sample, arises only due

to the variation of the specific units that are selected into the sample, and not due to variation of the population characteristics of interest (such as the small-area means) which are considered fixed because they are constant across replications. In contrast, model-based methods assume stochastic models governing the population values that are the target of the estimation process. Models are used to mediate the process of borrowing strength across the districts (small areas). That is, inference about a district that is represented in the sample by very few observations is supported by the information in the other districts' subsamples. This is most effective when the districts are very similar. Similarity can be enhanced by adjustment for other variables, opening up the potential of regression models.

Borrowing strength, as defined originally by Efron and Morris (1973), is based on the assumption of random effects. In the simplest setting with no covariates, the deviations of the district-level means θ_j from their national mean θ are assumed to be a random sample from a centered distribution with a finite variance, such as $\mathcal{N}(0, \sigma_u^2)$.

Model-based methods for small-area estimation associate the districts with random effects. In applications, however, the districts have their names (labels), and the target quantities θ_j could in principle be established by enumeration. In an hypothetical replication of the survey, the same districts, with the same subpopulations and the same values of θ_j would be involved. Therefore, it is natural to associate the districts with fixed effects. Longford (2007) argues that the assumption of fixed or random effect has a profound effect on standard errors of model-based small-area estimators. In the present paper we consider both design- and model-based estimators, and assess their accuracy in the case of the fixed-effect assumption. Accuracy refers not to average MSE across areas, but to MSE for the particular (fixed) areas. This departs from previous studies in which accuracy was assessed by averaging MSE across areas (see, e.g., Costa, Satorra

and Ventura (2003), and Santamaría, Morales and Molina (2004)).

In the model-based approach, the best linear unbiased predictor (BLUP) of the parameter of interest (the small-area parameter), is a linear combination of a direct and a synthetic estimator with weights that depend on two parameters that are usually unknown: the within- and between-area variances (possibly after controlling for other variables, regressors). Since both parameters are unknown quantities, these two variances have to be estimated, giving rise to the empirical BLUP (EBLUP). This estimation can distort the optimality of the EBLUP. In sections 3 and 4 we assess the consequences on accuracy of the substitution of model parameters by estimated values.

The purpose of the paper is to compare the performance of design- vs. model-based small-area estimators, with a focus on a specific (fixed) set of small areas. Monte Carlo methods are used for this investigation.

Two population frames will be considered in the Monte Carlo study: a) a theoretical population with varying distribution and sample size; b) an empirical population of labor statistics from the affiliation of firms in the NISS (National Institute of Social Security) registers. The choice of the NISS is motivated by current work at IDESCAT (Statistics Bureau of Catalonia).

The plan of the paper is as follows. Section 2 develops the notation and general context of small-area estimation, focusing on the distinction between design-based and model-based methods. Sections 3 and 4 describe the Monte Carlo studies using the theoretical and the empirical population, respectively. Section 5 concludes with a discussion of the results and the avenues for further research.

2 Approaches for small-area estimation

We consider a population stratified into J (small-area) domains (strata), $j = 1, 2, \ldots, J$, and we seek to estimate the stratum parameters θ_j as well as an overall population parameter θ. A direct estimator of θ_j uses sample data only from area j. An indirect or synthetic estimator of θ_j uses data also from outside area j. We suppose that there is a direct estimator $\hat{\theta}_{dj}$ of θ_j and that it is unbiased (but may have large variance), and a synthetic estimator $\hat{\theta}_{sj}$ that has small variance but may be biased for θ_j.

Two perspectives motivate the different small-area estimators. The first assumes that the θ_j are fixed values and that there is sampling variation only within each stratum. In the second, in addition to the random variation within strata, there is also random variation of the θ_j, that are supposed to be realizations from a specific sampling distribution. We now describe these two approaches, design-based (fixed θ_j) and model-based (random θ_j), respectively.

2.1 Fixed-area perspective

Following Rao (2003, Section 4.3), a natural way to balance the potential bias of a synthetic estimator $\hat{\theta}_{sj}$ of θ_j against the instability of a direct estimator $\hat{\theta}_{dj}$ of the same parameter is to take the composite estimator (weighted average)

$$\hat{\theta}_{cj}(\pi_j) = (1 - \pi_j)\hat{\theta}_{dj} + \pi_j \hat{\theta}_{sj}, \tag{1}$$

a function of the weight $0 \leq \pi_j \leq 1$. This estimator has a mean square error (MSE) given by (Rao, 2003, formula (4.3.2)):

$$\begin{aligned} MSE(\hat{\theta}_{cj}, \theta_j) &= (1 - \pi_j)^2 MSE(\hat{\theta}_{dj}, \theta_j) + \pi_j^2 MSE(\hat{\theta}_{sj}, \theta_j) \\ &\quad + 2\pi_j(1 - \pi_j) E\left\{(\hat{\theta}_{dj} - \theta_j)(\hat{\theta}_{sj} - \theta_j)\right\} \end{aligned} \tag{2}$$

where $MSE(\hat{\delta}, \delta)$ denotes the MSE of an estimator $\hat{\delta}$ with respect to the target δ. The expectation in the last term of (2) is taken with respect to the design-based sampling variation. In most applications, $\hat{\theta}_{dj}$ and $\hat{\theta}_{sj}$ are uncorrelated, so this last term vanishes. This is assumed throughout. Denote $\tilde{\theta}_{cj} = \hat{\theta}_{cj}(\tilde{\pi}_j)$.

The weight that minimizes the MSE of $\tilde{\theta}_{cj}$ is

$$\tilde{\pi}_j = \frac{MSE(\hat{\theta}_{dj}, \theta_j)}{MSE(\hat{\theta}_{dj}, \theta_j) + MSE(\hat{\theta}_{sj}, \theta_j)} \tag{3}$$

in which case the (minimum) MSE is

$$MSE(\tilde{\theta}_{cj}, \theta_j) = \tilde{\pi}_j MSE(\hat{\theta}_{sj}, \theta_j); \tag{4}$$

so, the (optimal) composite estimator $\tilde{\theta}_{cj}$ is superior to both the synthetic estimator $\hat{\theta}_{sj}$, since $\tilde{\pi}_j < 1$, and the direct estimator $\hat{\theta}_{dj}$, since $\pi_j > 0$ and $\tilde{\theta}_{cj}(0) = \hat{\theta}_{dj}$. If there was covariation among the synthetic and the direct estimator, $\text{cov}(\hat{\theta}_{dj}, \hat{\theta}_{sj})$ would be subtracted once in the numerator and twice in the denominator.

The expression (2) (with the covariance term ignored) will be used in sections 3 and 4 to compute the exact MSE of various composite estimators arising in a Monte Carlo (see, e.g., Figure 7 below). The exact values of the MSE can be computed since in Monte Carlo studies we know the population values of the parameters. In applications, the MSE will have to be estimated and several estimates are available. Longford (2007) discusses issues arising in the estimation of the MSE in the case of fixed area effects.

When the direct and synthetic estimators are unbiased for θ_j and θ respectively, and the variance of $\hat{\theta}_{sj}$ is small relative to the variance of the direct estimator, we have

$$\tilde{\pi}_j = \frac{\text{var}(\hat{\theta}_{dj})}{\text{var}(\hat{\theta}_{dj}) + (\theta - \theta_j)^2} \tag{5}$$

If $\hat{\theta}_{dj}$ is the sample mean, then $\text{var}\,(\hat{\theta}_{dj}) = \sigma^2_{j\epsilon}/n_j$, where $\sigma^2_{j\epsilon}$ is a within-domain variance

and n_j is the sample size of the jth domain. We could contemplate homoscedasticity and replace $\sigma^2_{j\epsilon}$ by σ^2_ϵ. Then (5) becomes

$$\tilde{\pi}_j = \frac{\sigma^2_{j\epsilon}/n_j}{\sigma^2_{j\epsilon}/n_j + (\theta - \theta_j)^2} \tag{6}$$

For a synthetic estimator (unbiased for θ) whose variance is small compared with the variance of the direct estimator (unbiased for θ_j), we have

$$E(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2 \doteq (\theta - \theta_j)^2 + \text{var}\,(\hat{\theta}_{dj}) \tag{7}$$

and $\tilde{\pi}_j \doteq \text{var}\,(\hat{\theta}_{dj})/(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2$, suggesting the weight

$$\hat{\pi}^\dagger_j = \frac{\widehat{\text{var}}\,(\hat{\theta}_{dj})}{(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2}\,,$$

where $\widehat{\text{var}}\,(\hat{\theta}_{dj})$ is an unbiased estimator of $\text{var}\,(\hat{\theta}_{dj})$. This estimator is very unstable and it could even fall outside the interval $[0, 1]$. In the Monte Carlo study in sections 3 and 4 we use instead the weight

$$\hat{\pi}^*_j = \frac{\widehat{\text{var}}\,(\hat{\theta}_{dj})}{(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2 + \widehat{\text{var}}\,(\widehat{\theta}_{dj})}\,, \tag{8}$$

which satisfies the condition $0 \leq \hat{\pi}^*_j \leq 1$.

The optimal composite estimator that uses the weight in (6) is not feasible in practice because the bias term $(\theta_j - \theta)^2$ and the variance $\sigma^2_{j\epsilon}$ are unknown quantities that need to be estimated. We will see that several alternatives to the estimation of the within-area variance do not induce much difference among estimators; in contrast, alternatives to the estimation of the squared area-bias term will lead to fundamental differences among estimators.

2.2 Random-area perspective

Alternative small-area estimators are based on models. Suppose

$$y_{ji} = X_{ji}\beta + Z_{ji}\gamma_j + \epsilon_{ji} \tag{9}$$

where $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots J$, i and j denoting primary and secondary level units, observations and areas, respectively. X_{ji} and Z_{ji} are vectors of attributes of observation i of area j, β is a vector of regression coefficients and γ_j is a vector of random area effects, independent of ϵ_{ji}, and usually both normally distributed with respective variances σ_u^2 and σ_ϵ^2 (variance matrix Σ_u, instead of σ_u^2, when Z_{ji} is a vector).

To simplify the exposition, consider the simplest version of the model in (9), with Z_{ji} set to the indicator of area j and $X_{ji} = 1$ is empty (there are no covariates); that is,

$$y_{ji} = \mu + u_j + \epsilon_{ji} \tag{10}$$

Variables u_j and ϵ_{ij} are centered random variables with respective variances σ_u^2 (variance "between") and σ_ϵ^2 (variance "within").

Let $y_{j\cdot} = n_j^{-1} \sum_i y_{ij}$ and $y_{\cdot\cdot} = n^{-1} \sum_i \sum_j y_{ij}$ be the respective direct and synthetic estimators of $\theta_j = \mu + u_j$, where $n = \sum_j n_j$ is the overall sample size. Since $\text{var}\,(y_{j\cdot}) = \sigma_u^2 + \sigma_\epsilon^2/n_j$ and $\text{cov}\,(y_{j\cdot}, u_j) = \sigma_u^2$, the best linear predictor (BLUP [1]) of θ_j given $y_{j\cdot}$ is

$$BLUP(\theta_j \mid y_{j\cdot}) = \mu + \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n_j}(y_{j\cdot} - \mu) = (1 - \omega_j)y_{j\cdot} + \omega_j\mu \tag{11}$$

where

$$\omega_j = 1 - \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n_j} = \frac{\sigma_\epsilon^2/n_j}{\sigma_u^2 + \sigma_\epsilon^2/n_j} = \frac{1}{1 + n_j\gamma}$$

[1] A common notation is also BLP, but since BLP is unbiased in the predictive sense, i.e. $E\{BLP(\theta_j) - \theta_j\} = 0$, the terminology of 'best linear unbiased predictor' (BLUP) will be used.

and $\gamma = \sigma_u^2/\sigma_\epsilon^2$. We used $E(y_{j.}) = \mu$. The empirical BLUP (EBLUP) is

$$\hat{\theta}_{cj}(\hat{\omega}) = EBLUP(\theta_j \mid y_{j.}) = (1 - \hat{\omega}_j)y_{j.} + \hat{\omega}_j y_{..}$$

where $y_{..}$ (the overall mean) is used as an estimator of μ and

$$\hat{\omega}_j = \frac{\hat{\sigma}_\epsilon^2/n_j}{\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2/n_j} = \frac{1}{1 + n_j\hat{\gamma}} \tag{12}$$

as the estimator of ω_j, with $\hat{\gamma} = \hat{\sigma}_u^2/\hat{\sigma}_\epsilon^2$. Here

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n - J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - y_{.j})^2 \tag{13}$$

and

$$\hat{\sigma}_u^2 = \frac{1}{J - 1}\sum_{j=1}^{J}(y_{.j} - y_{..})^2 \tag{14}$$

are moment-matching estimators of the variances. The estimator $\hat{\sigma}_\epsilon^2$ could be written as a weighted mean, i.e.

$$\hat{\sigma}_\epsilon^2 = \sum_{j=1}^{J}\left((n_j - 1)/(N - J)\right)\hat{\sigma}_{\epsilon j}^2,$$

of the within-area variance estimates

$$\hat{\sigma}_{j\epsilon}^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - y_{.j})^2. \tag{15}$$

The composite small-area estimators can be based on the assumption of homogeneity of the within-area variances, in which case they will use a common estimate of this variance, such as the estimator of (13), or they may contemplate heteroscedasticity, in which case they may use an area specific estimate such as (15).

As an alternative to these estimators we could use maximum likelihood (ML) estimation of the mixed regression model. For the unbalanced case, this provides alternative EBLUP estimators. It will be evaluated in the Monte Carlo study of sections 3 and 4.

One could also question the quality of these EBLUP estimators when the model (10) deviates from the standard assumptions, such as normality of the within- and between-area distributions, or both, or when there is variation among the within-area variances while equality is assumed.

The Monte Carlo study of Section 3 contemplates normal and highly skewed distributions, both for the first- and second-level distributions. Non-normality of the distribution within each area, and heteroscedasticity of the within-area variances, is present in the Monte Carlo study of Section 4 involving an empirical population.

3 Monte Carlo study: theoretical population

In the simplest set-up, data is generated from a two-level model in which the domain parameters θ_j are realizations of $\theta_j \sim \mathrm{N}\left(\mu = \theta, \sigma_u^2 = 3\right)$ and the observations y_{ji} (subject i in area j) are realizations of $y_{ji} \sim \mathrm{N}\left(\mu = \theta_j, \sigma_\epsilon^2 = 6\right)$. The number of small areas is 40. In one simulation the within-area sample sizes are equal to $n_j = 10$, while in another simulation n_j ranges from 6 to 40.

Next we list the estimators considered in the Monte Carlo study. The direct and synthetic estimators are respectively the sample mean of area j and the overall sample mean $\hat{\theta}$. The composite estimators can be classified according to whether or not the weights are known (*theoretical*) or estimated (*empirical*), and according to whether the estimator of the squared bias term $(\theta_j - \theta)^2$ is *area specific* (weights will be denoted by π_j) or *averaged* across the areas (weights denoted as ω_j). Except for the direct estimator, denoted by D, all estimators considered are composite estimators whose weights are specified as follows:

Design-based estimators

Theoretical composite: TC1

$$\tilde{\pi}_j = \frac{\sigma^2_{j\epsilon}/n_j}{(\theta_j - \theta)^2 + \sigma^2_{j\epsilon}/n_j}$$

Empirical composite: CA

$$\hat{\pi}^*_j = \frac{\hat{\sigma}^2_{j\epsilon}/n_j}{(\hat{\theta}_j - \hat{\theta})^2 + \hat{\sigma}^2_{j\epsilon}/n_j}$$

Note that TC1 and CA use area-specific values for the within-area variance $\sigma^2_{j\epsilon}$ (they allow for heteroscedasticity of this variance across areas).

Model-based estimators

Theoretical composite: TC2

$$\tilde{\omega}_j = \frac{\sigma^2_\epsilon/n_j}{\sigma^2_u + \sigma^2_\epsilon/n_j}\,,$$

with population (true) values for the variances within σ^2_ϵ and between σ^2_u.

Empirical composites: C

C is the composite estimator with $\hat{\omega}_j$ defined in (12), (13) and 14.

ML estimator: CML

Uses the estimator (11) with the population values μ, σ^2_ϵ and σ^2_u substituted by estimates obtained by fitting the model in (9) by ML.

3.1 Monte Carlo study: θ_j random

We first generate the area-level quantities θ_j as random draws from an assumed distribution $\mathrm{N}(\theta, \sigma^2_u)$, independently across replications. The areas cannot be distinguished by any features (they are exchangeable), and so their MSEs are constant for each estimator. As should be expected, the results summarized in Figure 1 indicate that the MSEs for

the different methods are highly correlated. Within a method, the empirical MSE's are not constant because the number of replications is finite (is 3000).

Figure 1 about here

The theoretical design-based estimator TC1 is far more efficient than the others since it uses more information about the true values of the within-area variance and area-bias in each replication. The within-area variance is constant across replications, but this is not the case for the area-bias. The theoretical estimator TC2 that uses variance and bias parameters common across the areas performs similarly as the C and CML estimators (the last two estimators are equivalent, given that the n_j are equal across areas), which are the next in performance. The feasible design-based estimators CA perform poorly. Finally, the direct estimator has the poorest performance.

The gain of TC1 over TC2 can be explained by the fact that TC1 uses information about the squared area-bias $(\theta_j - \theta)^2$, which varies across replications, while TC2 uses only information about the true value of its expectation, the between-area variance (model parameter σ_u^2). Replacement of the parameters by their estimates in the model-based methods does not reduce this efficiency substantially; indeed, the RMSEs of TC2, C and CML are nearly indistinguishable in Figure 1. In contrast, the design-based estimator CA, which is based on substituting an estimate for the true value of the area-bias, incurs a severe loss of efficiency when compared with the theoretical estimator TC1.

3.2 Monte Carlo study: θ_j fixed

Now we assume that the θ_j are fixed across replications, in accordance with the empirical set-up in which the *eccentricity* of an area, i.e. the deviation of the area from the overall mean, is an (unknown) but fixed quantity that remains constant across replications.

Figure 2 reports the empirical root-MSE (RMSE) across replications for each area and for the different estimates. TC1 and TC2 are not feasible in practice since they use true values of population parameters that are not available in a typical application. However, the performance of TC1 and TC2 will shed light on the nature of the accuracy of the alternative estimators.

Figure 2 about here

We see that the theoretical composite estimator TC1 that uses area-specific bias performs better than the theoretical composite TC2 that uses a single parameter (the variance between) to account for the squared bias averaged across the areas. In fact, TC1 performs as the worst estimator for areas with an extreme value of eccentricity (on the far right-hand side of the x-axis), that is, the areas for which the mean deviates highly from the overall area mean. To understand the variation of RMSE across the areas, these have been ordered according to their absolute deviation $\mid \theta_j - \theta \mid$, so that the extreme areas are located at the right-hand side. The lengths of the bars at the bottom of the graph are proportional to these deviations. We see that the largest difference between TC1 and the other statistics arises when $\mid \theta_j - \theta \mid$ is small; on the other hand, TC1 and TC2 nearly coincide when $\mid \theta_j - \theta \mid$ is approximately equal to the between-area variance. The empirical model-based composite estimators also perform poorly for the areas that deviate highly from the overall mean θ. These results can be summarizes as follows:

- TC1 is the most efficient estimator for all the areas. This is a theoretical estimator, not feasible in applications. It provides a benchmark against which other estimators can be compared or related.

- CML and C are inefficient for areas with the largest deviations from the center

(large eccentricity).

- For the model-based estimators (C, CML and TC2), using estimated or true values of the parameters makes very little difference. This is not the case for the design-based estimators; just compare TC1 with CA.

- CML performs poorly for areas that deviate substantively from the center. The accuracy of CA increases for small or extreme values of eccentricity.

The above difference among estimators would not appear when observing RMSE averaged across areas.

We will see that this results holds in a variety of circumstances, when we vary the sample size, with large or small number of areas, and also with deviation from normality, both in the within-area distributions and in the distribution that generates the fixed realized values of the area means.

3.3 Non-normal data and unequal sample sizes n_j

Now we consider non-normally distributed data and different sample sizes across the areas, with the θ_j fixed. The sample size ranges from 6 to 45, the number of areas is 40. The number of replications is 3000. The results are shown in Figure 3. For clarity of the graph, since across areas the maximum difference of the RMSE for C and CML is .03, only the RMSE for CML is shown.

Figure 3 about here

A similar pattern is observed in Figure 2; that is,

1. The RMSE tends to increase with the eccentricity of the area.

2. TC1 is superior to all the estimators.

3. The feasible estimators C, CML are inefficient for areas that deviate highly from the overall mean (high eccentricity), and so is the theoretical estimator TC2.

4. CA does not do so badly as C and CML for those areas with high or very low values on eccentricity.

5. As expected, the RMSE tends to decrease with the sample size.

We also computed a version of the empirical composite C that estimates the variance-within σ_ϵ^2 as an (unweighted) mean of the within-area estimates $\hat{\sigma}_{\epsilon j}^2$ of (15), but the difference in terms of MSE with the standard version of C was negligible.

We found that the true values of MSE computed according to formulae (2) are indistinguishable from the (estimated) ones computed with 3000 replications and presented in Figure 3. Figure 4 displays the same graph with true RMSE for the three estimators D, TC1 and TC2. In both figures we see the superiority of the design-based estimators (TC1) over the model-based ones (TC2), not only for some areas that deviate highly from the overall mean, but also for those areas that exhibit a small value of eccentricity (the areas on the left of the x-axis).

Figure 4 about here

3.4 Mixture distribution for the area means θ_j

Next we generate the θ_j as realizations of a K-component mixture distribution of the same mean but different variance parameters σ_u^2. Specifically, let the θ_j be realizations of a mixture of $K = 2$ normals, both centered but with different variances, one much larger

than the other. Assume we have information of the component, small or large variance, to which each area belongs. The estimator CN incorporates that information; it is the empirical composite C with an estimator for σ_ϵ^2 specific for each component. Since the RMSE of estimators C and CML differs by less than .007, the graph shows only the values of the RMSE for C. Figure 5 reports the corresponding MSEs of the estimators.

Figure 5 about here

The pattern of variation of RMSE is similar to the one observed in Figure 2, except that now the model-based estimators increase their RMSE considerably and have an efficiency similar to the one of the direct estimator. Except for the different dispersion of the area means, the same population parameters were used. The efficiency of the design-based estimator is not affected by the different dispersion of the area means. Here the estimator that uses information about the group to which the small area belongs, group with small/large within-area variance, is more accurate than CA and improves also the performance of the model-based estimators, with the exception of a few areas (areas that have extreme eccentricity for the group with low within-area variances). Incorporating the information on the two types of areas improves the estimation of the between-area variance and increases the efficiency of the model-based estimators.

4 Simulation study on a real population

In this section we study the behavior of several estimators through a Monte Carlo simulation in which we replicate samples from the Labor Force Census of Enterprises affiliated with the Social Security system in Catalonia. This census contains information on the number of employees who are registered in the Social Security system for each enterprise. The data is available on a quarterly basis from year 1992. We consider only the

population in the first quarter of year 2000. The census contains 243,184 observations for Catalonia in year 2000, divided into 12 groups according to the economic sector to which each firm belongs, and into 41 counties (the 'comarques').

Table 1 about here

We ignore the sector-based classification and focus solely on the division by counties. Table 1 shows the number of enterprises (population size) and the mean and variance of the variable of interest (number of registered employees) in each county. The distribution of the enterprises across Catalonia is very uneven, as they are concentrated mainly in densely populated areas. In our set-up, the small areas are held fixed across resampling over the 1000 replications. In each replication, we extract a proportional stratified sample by county. We used sample sizes representing 10%, 5%, 2% and 1% of the population, which gives sample sizes close to those used by IDESCAT in several surveys). Table 2 summarizes the characteristics of these samples.

Table 2 about here

This population recreates conditions of non-normality, uneven sample sizes, and heterogeneity of within-area variances that are likely to appear in applications. So a Monte Carlo evaluation based on this population will assess the performance of competing estimators in a realistic setting.

We evaluate the performance of the theoretical estimators TC1 and TC2, and the empirical estimators C, CML and CA described in Section 2. The CML estimator was computed using `proc xtmixed` of the software program Stata 9.0. For each estimator, we computed the empirical relative root-MSE (RRMSE) across replications for each

county. Using absolute (instead of relative) root-MSE gave the same pattern of performance as when using RRMSE. The optimization routine to compute CML did not always converge, with the percentage of non-converging replications increasing as sample size decreased. Specifically, the percentage of failed attempts ranges from 8% with the largest sample size to 65.6% with the smallest one. The estimated MSE for CML was based only on the converging runs.

For the 10% and 5% sample sizes, the direct and the composite estimators have similar RRMSE values. For those sampling schemes, D has the smallest RRMSE among the feasible estimators and is more efficient than the theoretical model-based TC2 estimator. We therefore focus on the description of the 2% and 1% sampling designs.

Figure 6 plots the variation of the RRMSE for the estimators and areas for the 2% sampling design. For clarity, the RRMSE of CML is omitted as it is nearly indistinguishable from C. The same pattern of variation is observed for the 1% sample. Areas have been ordered with respect to their eccentricity, i.e. deviation of the area mean from the overall mean (the heights of the bars are proportional to the eccentricity of the area). We see that the RRMSE tends to increase as the areas become more extreme in terms of eccentricity. The area sample size is proportional to the thickness of the bar. We observe that RRMSE tends to decrease as the sample size increases.

The direct estimator D performs poorly. The design-based estimator CA also does poorly while its theoretical counterpart TC1 is the most efficient. The model-based estimators TC2 and C (and CML) do better than the direct estimator D but worse than the theoretical design-based estimator TC1. The poor performance of D and other feasible model-based estimators for some counties, specially for the county SG ('Segarra') stands out. Segarra has both a huge value of the within-county variance and a very small

sample size (see columns 3 and 5 of Table 1). In applications of small-area estimation, it should be of high concern that our area has such extreme features. If we knew the true values of the squared area-bias and the within-area variance (as in TC1 and TC2) then MSE would be reduced dramatically for SG.

The high fluctuation of the performance of the direct estimator is due to the variation of the sample sizes and within-area variances across areas. For extreme areas, the CA estimator performs similarly as the model-based estimator TC1. Estimation of the population parameters has a profound effect on the accuracy for areas that are not extreme. This is the case, for example, for Baix Camp (BC).

For completeness, Figure 7 shows the results for the 10% sampling design. We observe the same pattern of performance across areas as in Figure 6. The distance between the model-based and the design-based estimators is more obvious for areas with greater eccentricity. This graph shows that the direct estimator nearly matches the efficiency of TC1, a clear indication that for such a large sample size, small-area estimation is redundant.

Figures 6 and 7 show that the model-based estimators do not perform too badly when the small areas do not show a high value of eccentricity. But for areas that are very extreme these estimators do worse than the the direct estimator, or even than the CA estimator.

This Monte Carlo study with a real population provides a context in which we can recognize different scenarios encountered in an application.

1. For an area with a large sample size, all the small-area estimators are close to each other. This is the case of Barcelona (BN).

2. In areas with a small sample size and a extreme within-area variance, not necessarily extreme in eccentricity, the empirical small-area estimators may perform very poorly. This is the case of Segarra (SG). For such areas, the incorporation of information on the magnitude of the between- and within-area variances may produce dramatic gains on RRMSE. With such additional information, then both model- or design-based estimators are equally efficient.

3. In an area with a small value of eccentricity and small sample size, model-based estimators are less efficient than the design-based estimators. This can be seen in Alt Camp (AC).

4. In an area with a high value of eccentricity and small sample size, design- and model-based estimators are more efficient than the direct estimator. This can be seen in Alta Ribagor ca (AR).

5 Conclusions and agenda

We have seen that in the case of fixed areas, averaging MSE across areas does not provide the complete picture of the performance of alternative small-area estimators. Such averaging of MSEs can be used only to evaluate accuracy in the context of random-area parameters.

We conclude that a composite estimator that uses a common bias estimator for all the areas performs poorly on areas that are extreme. The same is true for the theoretical composite estimator TC2. The problem carries over to the mixed-effects regression, even when the model is not misspecified. Therefore, estimation of the squared bias term for each area becomes crucial.

In the theoretical Monte Carlo exercise the estimation of the variances within the areas seems less important. However, the exercise on a real population shows also the importance of recognizing non-normality and variation across areas of the within-area variance.

These findings indicate that the key to improve small-area estimation is to acknowledge the fixed-effect nature of the data and to improve estimation of the squared area-bias. Differences (heteroscedasticity) of the within-area variances seem to be also critical. Several alternatives arise:

1. Using auxiliary information (such as a census or a previous survey) to estimate the squared bias. Then the same simple and convenient composite estimators could be used.

2. Improving the alternative composite estimator by defining different groups of areas that share a common between-area variance.

3. Estimating the squared bias using small-area methods. This approach has already been used in Longford (2007) for estimating MSEs of model-based estimators.

References

[1] Datta G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S. *Journal of the American Statistical Association, 94*, 1074-82.

[2] 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*, 101-8.

[3] Efron, B., and Morris C.E. (1973). Stein's Estimation Rule and Its Competitors – an Empirical Bayes Approach, *Journal of the American Statistical Association, 68*, 117-130.

[4] Ghosh M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal, *Statistical Science, 9,* 55-93.

[5] Isaki. C. T. (1990). Small-Area Estimation of Economic Statistics, *Journal of Business & Economic Statistics, 8*, 435-41.

[6] Longford, N.T. (2007). On standard errors of model-based small-area estimators, *Survey Methodology, 33*, 69-79.

[7] 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*, 73-84.

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

[9] Raghunathan, T.E. (1993). A Quasi-empirical Bayes Method for Small Area Estimation, *Journal of the American Statistical Association, 88*, 1444-48.

[10] Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley and Sons: New York

[11] Santamaría, L., Morales, D. and Molina, I. (2004). A comparative study of small area estimators, *SORT, 28*, 215-230.

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

[13] Singh, A.C., Mantel, H.J. and Thomas, B.W. (1994). Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology, 20*, Canada Statistics. 33-43.

[14] 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.

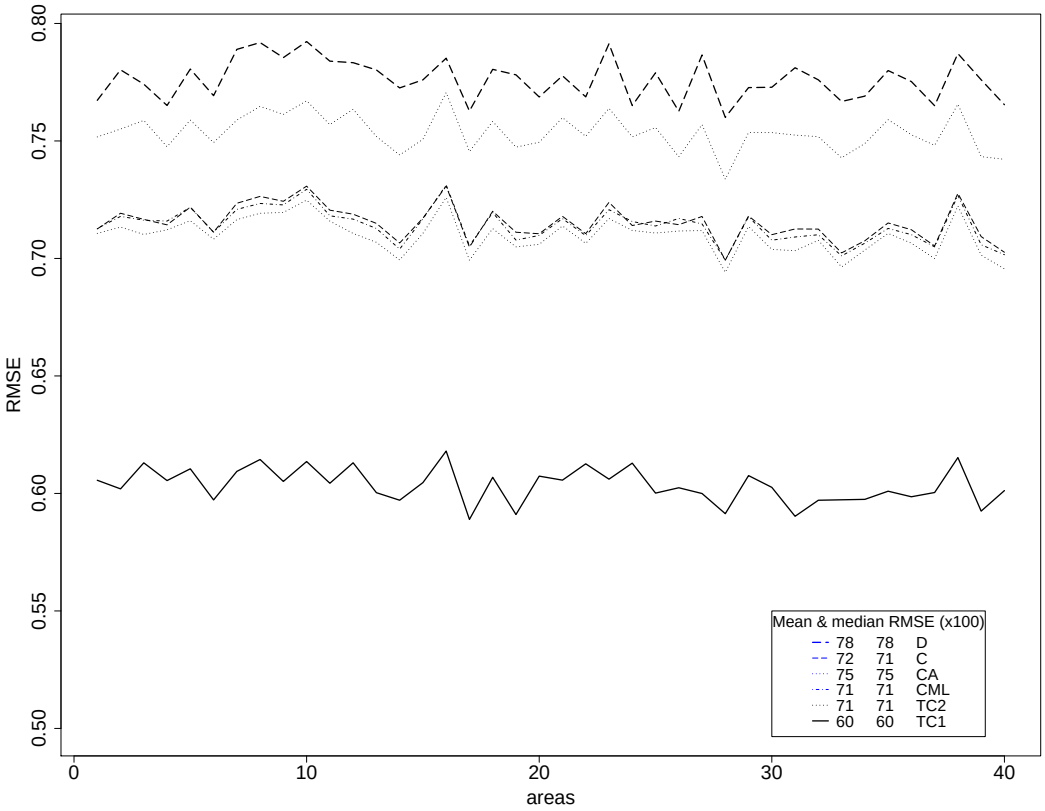

Figure 1: Root MSE (RMSE) for areas $j = 1, \ldots, 40$ when each mean θ_j is random across replications. Within- and between-area distributions are normal (number of replications is 3000, sample size in each area is 10). The mean and the median of the RMSE across areas are shown in the legend.

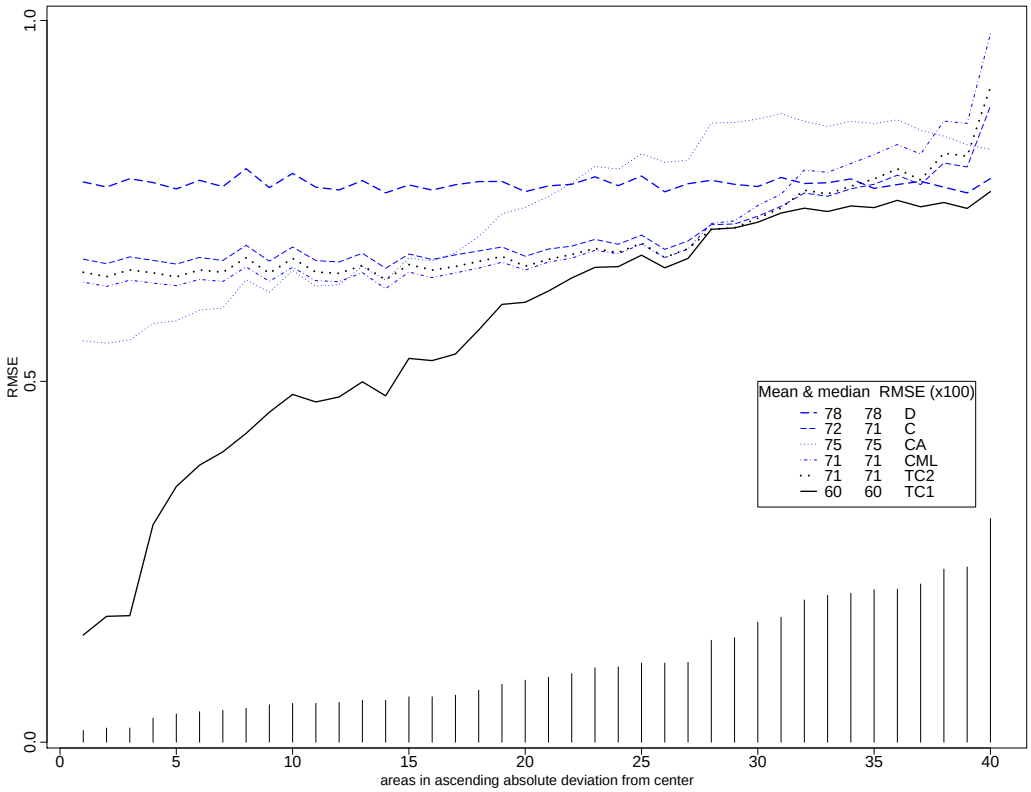

Figure 2: Root-MSE (RMSE) of each area when the θ_j are fixed across replications. The within-area distribution is normal, with homocedastic within-area variances. The area sample size is constant and equal to 10. The number of replications is 3000. The legend shows the mean and the median of the RMSE across areas.

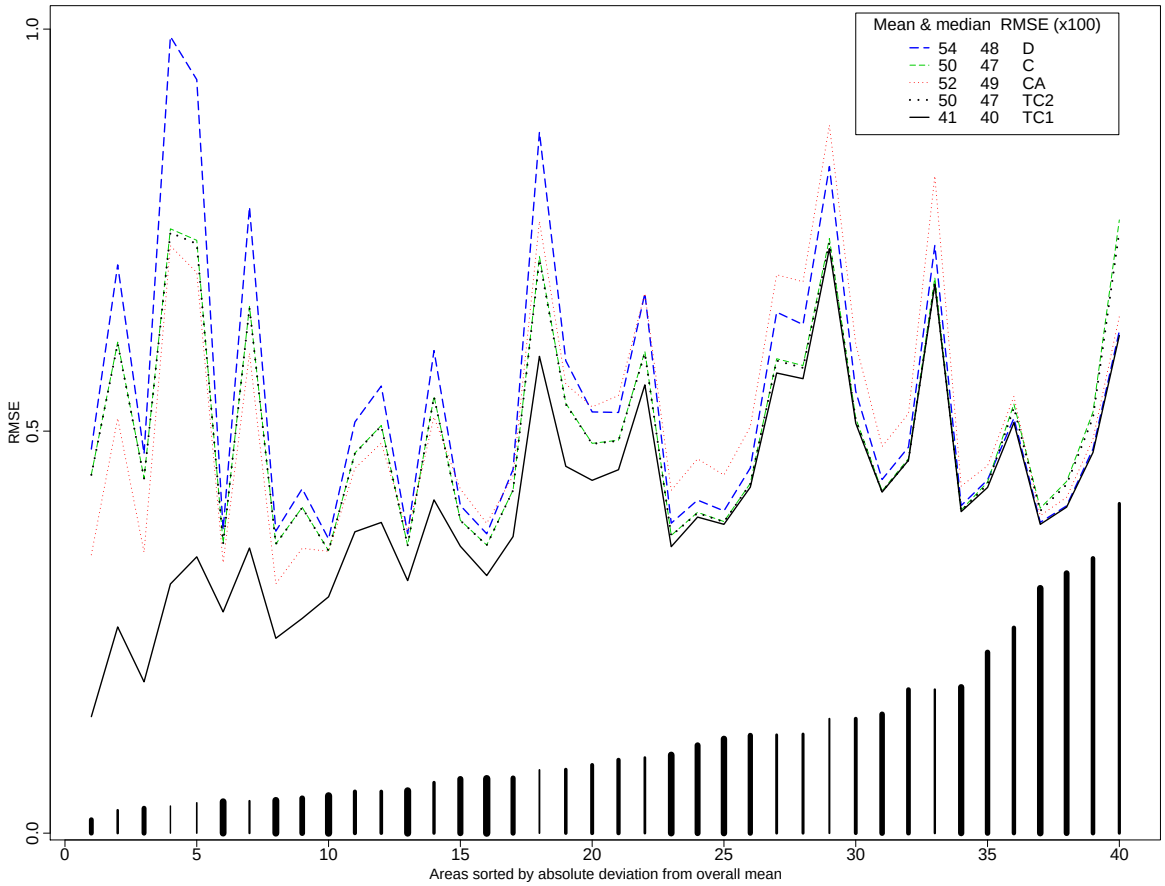

Figure 3: Root-MSE (RMSE) of each area when the θ_j are fixed across replications. The within-area distribution is normal, with homocedastic within-area variances and area sample size ranging from 6 to 45. The number of replications is 3000. The legend shows the mean and median of the RMSE across areas.

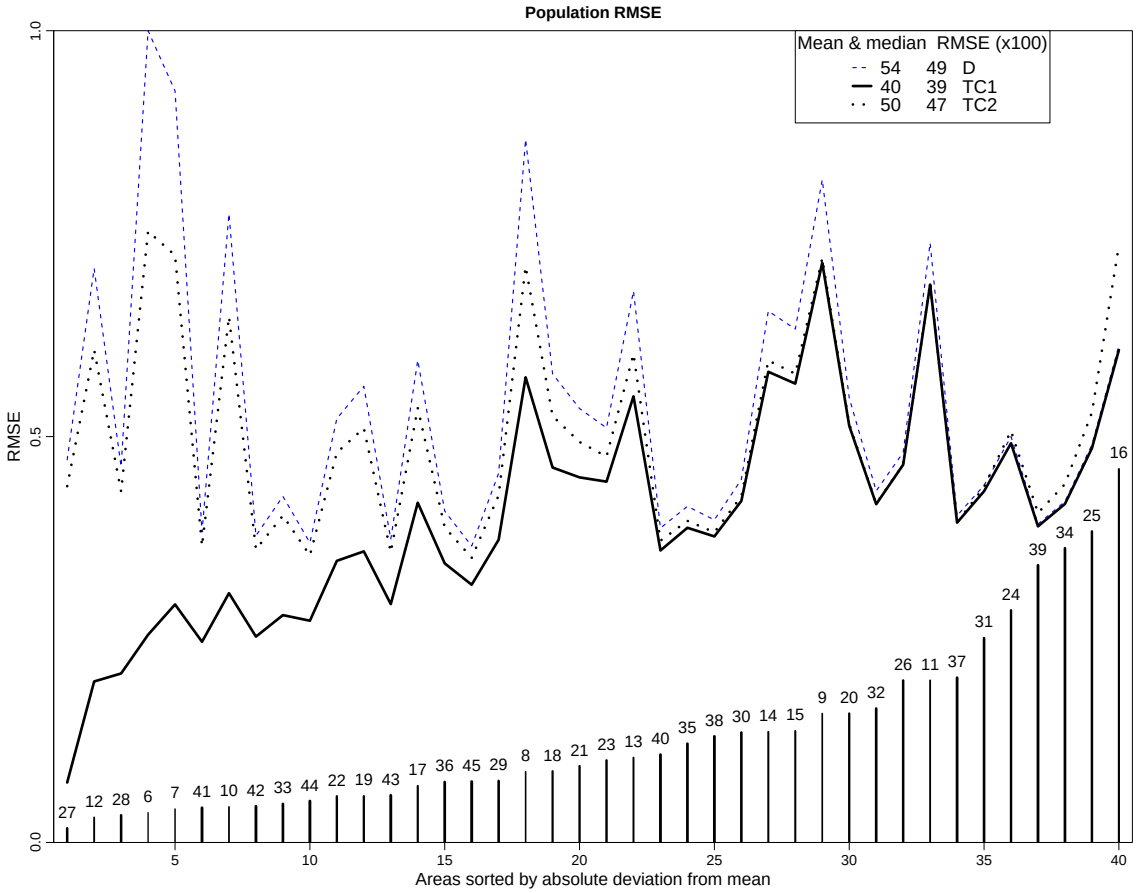

Figure 4: Theoretical values of the RMSE for D, TC1 and TC2 and for each area, for θ_j fixed across replications. The within-area distribution is normal, with homocedastic within-area variances and area sample size ranging from 6 to 45. The legend shows the mean and the median of the RMSE across areas.

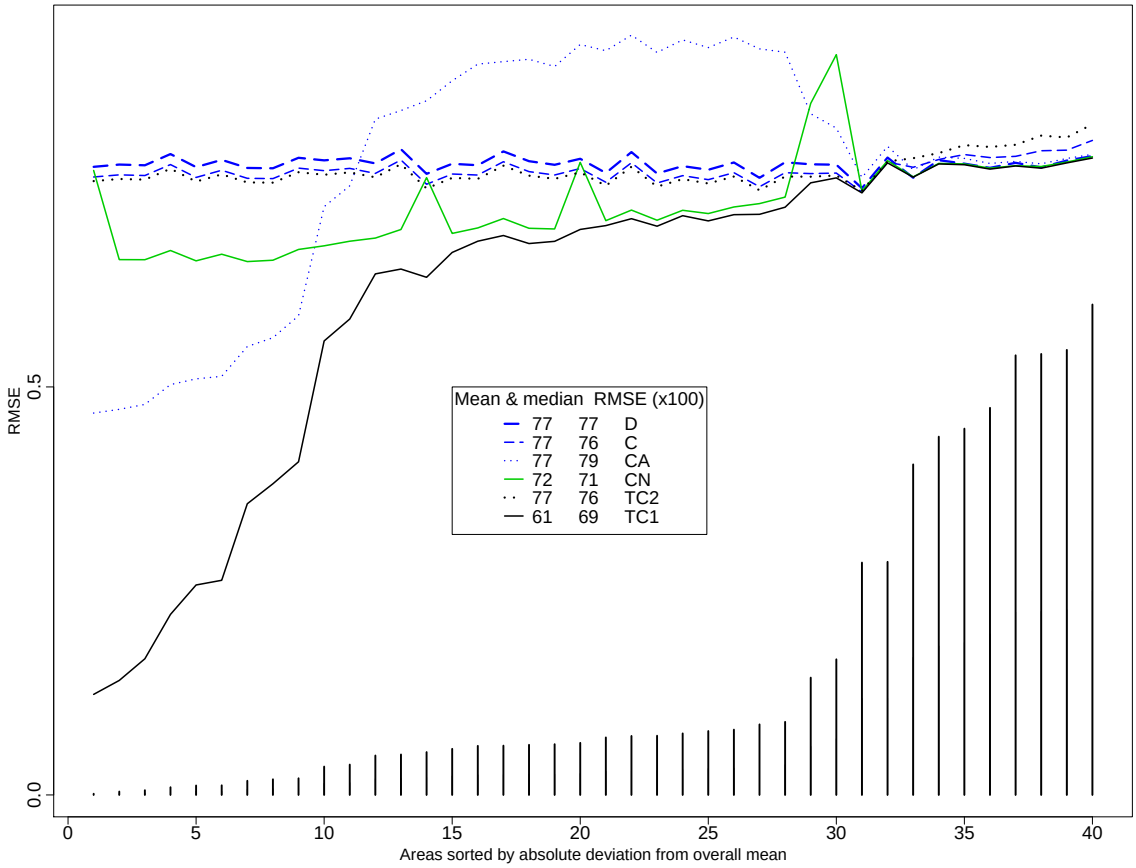

Figure 5: Root-MSE (RMSE) of each area when the θ_j are fixed across replications. The within-area distribution is normal, with homocedastic within-area variances. The area sample size is constant and equal to 10. The number of replications is 3000. The fixed values θ_js are realizations of a mixture of two centered normal distributions with different variances. The estimator CN incorporates information on the within-area variance. The legend shows the mean and the median of the RMSE across areas.

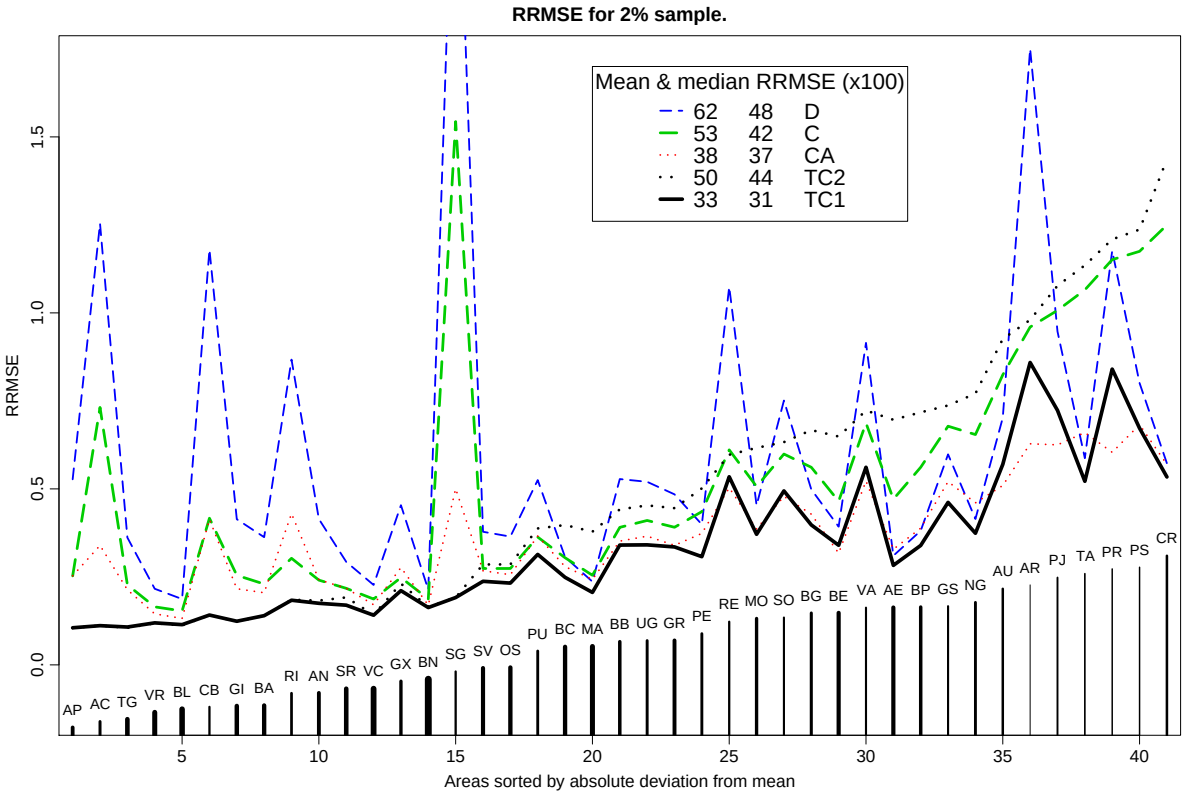

Figure 6: For the Monte Carlo analysis with an empirical population the graph shows the square root of the RRMSE or each county, for the 2% sample. Heights of the bars are proportional to the deviations or the area-level means from the overall mean and their thicknesses are proportional to sample sizes. The legend shows the mean and the median of the RRMSE across areas.

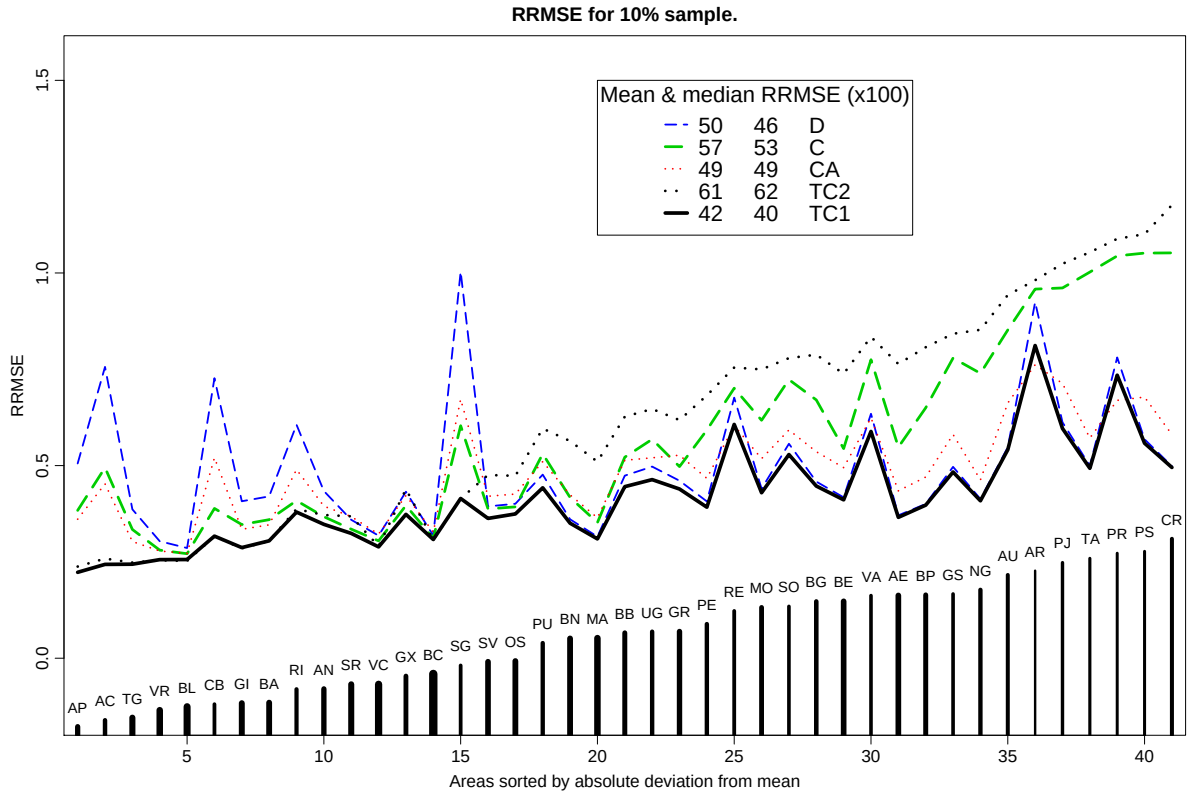

Figure 7: For the Monte Carlo analysis with an empirical population the graph shows the square root of the RRMSE for each county, for the 10% sample. Heights of the bars are proportional to the deviations or the area-level means from the overall mean and their thicknesses are proportional to sample sizes. The legend shows the mean and the median of the RRMSE across areas.

Table 1: Population characteristics

Counties ('Comarques')	Code	Size N_j	Mean θ_j	Squared bias $(\theta_j - \theta)^2$	Variance $\sigma^2_{j\epsilon}$
Alt Camp	AC	1282	8.73	0.09	3250.37
Alt Empordà	AE	4712	5.28	14.11	294.27
Alt Penedès	AP	3052	8.91	0.02	1686.24
Alt Urgell	AU	745	4.71	18.70	158.25
Alta Ribagorça	AR	140	4.59	19.73	205.38
Anoia	AN	3264	7.86	1.37	801.64
Bages	BA	5698	8.24	0.63	1356.90
Baix Camp	BC	5530	6.47	6.59	6479.54
Baix Ebre	BB	2237	6.31	7.41	534.40
Baix Empordà	BE	4634	5.44	12.92	425.17
Baix Llobregat	BL	20541	9.73	0.48	1642.46
Baix Penedès	CP	2197	5.26	14.23	171.82
Barcelonès	BN	88331	10.63	2.55	10314.88
Berguedà	BG	1397	5.44	12.90	196.15
Cerdanya	CR	788	3.71	28.34	71.93
Conca de Barberà	CB	611	8.29	0.56	1388.95
Garraf	GR	3466	6.28	7.62	685.91
Garrigues	GS	516	5.24	14.42	96.89
Garrotxa	GX	1909	7.51	2.33	419.72
Gironès	GI	6369	9.82	0.62	2037.47
Maresme	MA	11718	6.46	6.64	605.07
Montsià	MO	1918	5.61	11.73	246.00
Noguera	NG	1128	5.12	15.30	93.29
Osona	OS	5494	7.09	3.77	774.65
Pallars Jussà	PJ	410	4.37	21.76	130.37
Pallars Sobirà	PS	272	4.06	24.76	55.46
Pla d'Urgell	PU	1106	6.59	5.95	271.85
Pla de l'Estany	PE	1160	6.07	8.79	143.37
Priorat	PR	254	4.11	24.26	180.17
Ribera d'Ebre	RE	620	5.71	11.07	418.72
Ripollès	RI	959	7.87	1.35	875.92
Segarra	SG	594	10.87	3.35	8171.41
Segrià	SR	7096	7.74	1.69	714.23
Selva	SV	4586	7.11	3.70	610.20
Solsonès	SO	508	5.58	11.93	157.58
Tarragonès	TG	7440	9.42	0.15	1675.66
Terra Alta	TA	297	4.25	22.87	40.28
Urgell	UG	1178	6.28	7.59	312.25
Val d'Aran	VA	503	5.28	14.08	270.11
Vallès Occidental	VC	26683	10.34	1.71	3026.89
Vallès Oriental	VR	11795	8.45	0.34	832.68

† The average number of affiliates in Catalonia (overall mean θ) is 9.04.

Table 2: **Sample sizes of the Monte Carlo study (empirical population)**

Overall sample		Sample size in county			
% of pop.	Sample size	Mean	Median	Min.	Max.
1	2431	59.3	19	1	883
2	4863	118.6	38	3	1767
5	12159	296.6	95	7	4417
10	24316	593.1	191	14	8833