Journal of Official Statistics, Vol. 23, No. 2, 2007, pp. 223–237

Finite Population Small Area Interval Estimation

Li-Chun Zhang[1]

Small area interval estimation is considered for a finite population, where the small area parameters are treated as fixed constants. Design based direct estimation yields intervals that are too long to be useful. Model based approaches are considered. The design based area-specific coverages are uncontrollable. We propose to use population-specific simultaneous coverage as the basis for evaluating the small area confidence intervals. Wage survey and census household data are used for illustration.

Key words: Design based coverage; random effects mixed model; bootstrap calibration.

1. Introduction

The main current approach to small area estimation is based on prediction models. There is a considerable amount of theory on the estimation (or prediction) of the small area parameters and the associated mean squared error (MSE) with respect to *both* the population and sampling models. We refer to Rao (2003) for a comprehensive overview. For many survey practitioners and users it is appealing to have traditional design based measures of uncertainty, conditional on the given population and with respect to the sampling alone. Rivest and Belmonte (2000) derived such a *conditional* MSE of the composite estimators, which include the empirical best linear unbiased predictor (EBLUP) as a special case. The problem is that the conditional MSE estimator can be very unstable, especially when the shrinkage factor attached to the direct estimator is small. On the other hand, the model based unconditional MSE estimator has been found to track the conditional MSE quite well in small simulation studies (Rao 2003, Section 7.1.6).

In this article we consider the related issue of small area interval estimation. Design based direct estimation is inefficient, and the confidence intervals are too long to be useful in many cases. Since short, area-specific-intervals are impossible to construct, we propose to use design based, population-specific, *simultaneous* coverage as the basis for evaluating the performance of small area interval estimators. It is shown that the simultaneous coverage of interval estimators under the linear mixed models asymptotically achieves the nominal level of confidence, under conditions similar to those for the second-order *unconditional* MSE estimation (Rao 2003). The methodology is set out in Section 2. In Sections 3 and 4 we illustrate the model-based approach using, respectively, the wage survey and census household data. A summary is given in Section 5.

Statistics Norway, Kongens gt. 6, P.B. 8131 Dep., N-0033 Oslo, Norway. Email: lcz@ssb.no
Acknowledgments: I would like to thank an associate editor and a referee for their constructive comments that have greatly improved a previous version of the article.

2. Small Area Interval Estimation

2.1. Simultaneous Coverage

Let $i = 1, \ldots, m$ denote the small areas (or domains). Let θ_i denote the small area parameter of interest. Let $\lambda_i(\alpha)$ denote a confidence interval for θ_i constructed for $100\alpha\%$ nominal level of confidence. Let $I_i = I_i(\alpha) = 1$ if $\theta_i \in \lambda_i(\alpha)$, and 0 otherwise. The design based, *area-specific* coverage of $\lambda_i(\alpha)$ is then

$$\delta_i(\alpha) = P_\pi(I_i = 1) = E_\pi(I_i) \tag{1}$$

where P_π. denotes probability and E_π denotes expectation with respect to sampling from the finite population. We say that $\lambda_i(\alpha)$ achieves the nominal level of confidence if

$$\delta_i(\alpha) = \alpha$$

Regarding the set of $\lambda_i(\alpha)$'s, all of them derived at the same nominal level of confidence, we define their *simultaneous* coverage as

$$\delta(\alpha) = E_\pi\left(m^{-1}\sum_{i=1}^{m} I_i\right) = m^{-1}\sum_{i=1}^{m}\delta_i(\alpha) \tag{2}$$

The simultaneous coverage is the expected proportion of small area parameters covered by the set of confidence intervals on repeated sampling from the population. It is a meaningful design based measure in the context of small area estimation, summarizing all the area-specific coverages in a single number. Since correct area-specific coverages for all the areas imply correct simultaneous coverage, but not the other way around, the simultaneous coverage is a weaker property of the interval estimators. While such population-specific, area averaging performance measures are often used in empirical studies (e.g., Heady and Ralphs 2005), few systematic studies have been reported on how the model based methods *should* behave with respect to such design based measures. Finally, we notice that the simultaneous coverage is not the joint coverage of the $\lambda_i(\alpha)$'s, i.e., $P_\pi(\cap_{i=1}^{m} I_i = 1)$. Nor has it anything to do with the so-called simultaneous intervals in multiple comparison problems.

2.2. Design Based Estimation

Denote by $\hat{\theta}_i$ a design based direct estimator of θ_i. We assume that it is design unbiased with variance ψ_i. That is, let e_i be the sampling error of $\hat{\theta}_i$, and then we have

$$\hat{\theta}_i = \theta_i + e_i \quad \text{where } E_\pi(e_i) = 0 \quad \text{and} \quad V_\pi(e_i) = \psi_i \tag{3}$$

The standard design based approach is then to assume normality of e_i, which yields the $100(2\alpha - 1)\%$ nominal confidence interval

$$\left(\hat{\theta}_i - z_\alpha\sqrt{\psi_i},\quad \hat{\theta}_i + z_\alpha\sqrt{\psi_i}\right)$$

where z_α is the α-quantile of $N(0,1)$, and $\alpha \in (0.5, 1)$. Given the normality assumption, design based intervals achieve area-specific as well as simultaneous coverage.

There are two difficulties with this approach. In the first place, normality of e_i, or even zero expectation of e_i, may not be valid if the within-area sample size is small and/or θ_i is nonlinear. Secondly, in practice ψ_i is seldom known. Replacing ψ_i with the design based direct estimator $\hat{\psi}_i$ is inefficient. Often in practice one chooses to use some smoothed, stable estimator of ψ_i with small bias instead. While both the nonnormality of $\hat{\theta}_i$ and the potential design bias of $\hat{\psi}_i$ may cause problems for the area-specific coverage, the simultaneous coverage often remains quite robust.

2.3. Estimation Based on Best Predictor

Let us start with a special case. Consider the following simple random effects model

$$\hat{\theta}_i = x_i^T\beta + v_i + e_i$$

(Fay and Herriot 1979), where x_i contains the area-level covariates, v_i is the independent random effect with zero mean and variance σ_v^2, and $\hat{\theta}_i$ is the direct design unbiased estimator and e_i is its sampling error as defined in (3). Suppose that the parameters (β, σ_v^2) are known, and so are the ψ_i's. The *best predictor* (*BP*) is then given by

$$\tilde{\theta}_i = x_i^T\beta + \gamma_i\left(\hat{\theta}_i - x_i^T\beta\right) \quad \text{where} \quad \gamma_i = \sigma_v^2/\left(\sigma_v^2 + \psi_i\right)$$

The MSE of the BP is simply the so-called g_1-term (Prasad and Rao 1990), i.e.,

$$g_{1i} \stackrel{\text{def}}{=} E\{(\tilde{\theta}_i - \theta_i)^2\} = \gamma_i\psi_i = (1 - \gamma_i)\sigma_v^2$$

where E denotes expectation with respect to both the population model and the design generated sampling distribution. Given the normality of $\tilde{\theta}_i - \theta_i$, a $100(2\alpha - 1)\%$ nominal level prediction interval for θ_i is given by

$$\lambda_i = \left(\tilde{\theta}_i - z_\alpha\sqrt{g_{1i}}, \quad \tilde{\theta}_i + z_\alpha\sqrt{g_{1i}}\right)$$

The nominal level refers to accuracy of prediction because θ_i is a random variable here. Under the assumed random effects model, λ_i then achieves the unconditional coverage

$$\Delta_i(\alpha) = P(I_i = 1) = \alpha \tag{4}$$

where P is the probability under both the population model and the sampling distribution.

How does this BP based λ_i perform from the design based point of view? Take first the situation where $\psi_i/\sigma_v^2 \approx 0$ and $\sigma_v^2 = O(1)$. We have

$$\gamma_i \approx 1 \quad g_{1i} \approx \psi_i \quad \tilde{\theta}_i \approx \hat{\theta}_i \quad \text{and} \quad \lambda_i \approx \left(\hat{\theta}_i - z_\alpha\sqrt{\psi_i},\ \hat{\theta}_i + z_\alpha\sqrt{\psi_i}\right)$$

Since normality is probably a reasonable assumption in this case, $\lambda_i(\alpha)$ would approximately achieve the design based area-specific coverage, i.e., $\delta_i(\alpha) \approx \alpha$. Next, suppose $\sigma_v^2/\psi_i \approx 0$. We have

$$\gamma_i \approx 0 \quad g_{1i} \approx \sigma_v^2 \quad \tilde{\theta}_i \approx x_i^T\beta \quad \text{and} \quad \lambda_i \approx \left(x_i^T\beta - z_\alpha\sigma_v,\ x_i^T\beta + z_\alpha\sigma_v\right)$$

Since v_i is a constant given the finite population, we basically have $\delta_i = 1$ if $|v_i| \le z_\alpha\sigma_v$, and 0 otherwise. That is, the area-specific coverage degenerates. In summary, the model based interval λ_i approximately attains the design based area-specific coverage as ψ_i/σ_v^2

tends to 0 or, equivalently, $\gamma_i \to 1$. As $\gamma_i \to 0$, the design based area-specific coverage cannot be maintained in general, and tends towards the degenerate case.

Consider now the design based simultaneous coverage of the λ_i's. Let E_1, V_1 and Cov_1 denote, respectively, expectation, variance and covariance with respect to the population model. Given (4) under the assumed model, for all $i = 1, \ldots, m$, we have

$$E(I_i) = E_1\{E_\pi(I_i)\} = E_1\{\delta_i(\alpha)\} = \alpha$$

$$V_1\{\delta_i(\alpha)\} = V(I_i) - E_1\{V_\pi(I_i)\} \leq V(I_i) = \alpha(1-\alpha)$$

Moreover, given the population, I_i and I_j depend on e_i and e_j, for $i \neq j$, respectively. Given independent sampling within the small areas, we have

$$\delta_i \delta_j = E_\pi(I_i)E_\pi(I_j) = E_\pi(I_i I_j)$$

Thus,

$$Cov_1(\delta_i, \delta_j) = E_1\{E_\pi(I_1 I_j)\} - E_1\{E_\pi(I_i)\}E_1\{E_\pi(I_j)\} = Cov(I_i, I_j) = 0$$

because I_i derives from $\tilde{\theta}_i - \theta_i$ which is a function of (e_i, v_i) and therefore independent of I_j from $\tilde{\theta}_j - \theta_j$. It now follows from the Law of Large Numbers that, as $m \to \infty$,

$$\text{(A)} \quad \delta(\alpha) = m^{-1}\sum_{i=1}^{m} \delta_i(\alpha) \quad \underrightarrow{P} \quad \alpha$$

i.e., there is convergence in probability with respect to the population model. In other words, given a large number of small areas, we may expect the design based simultaneous coverage of the model based intervals to be close to the nominal level of confidence, given (i) correct unconditional area specific coverage, i.e., $\Delta_i(\alpha) = \alpha$ for $i = 1, \ldots, m$, and (ii) independent random effects and independent sampling within the small areas.

2.4. Estimation Based on EBLUP

The model considered above is a member of the class of linear mixed models (LMMs). We refer to Rao (2003) for an account of the various LMMs that have been used in small area estimation. It is clear that the result (A) for the BP based intervals remains valid, provided the random effects are uncorrelated across the areas under the LMM. In practice, however, the parameters are unknown, and the EBLUP is used instead of the BP. Denote by $\hat{\theta}_i^H$ the EBLUP of θ_i. Denote by $\hat{g}_i$ the estimator of MSE ($\hat{\theta}_i^H$). The $100(2\alpha - 1)\%$ nominal level prediction interval for θ_i is then

$$\lambda_i = \left(\hat{\theta}_i^H - z_\alpha\sqrt{\hat{g}_i}, \quad \hat{\theta}_i^H + z_\alpha\sqrt{\hat{g}_i}\right) \tag{5}$$

with respect to both the population and sampling models. The latter may be purely model based, such as in the case of unit-level mixed models. It can also be the design generated sampling distribution, or a model of the actual sampling process.

The design based area-specific coverage of λ_i is uncontrollable as before. When it comes to the design based simultaneous coverage, the condition (4) would still imply $E_1\{\delta_i(\alpha)\} = \alpha$ and $V_1\{\delta_i(\alpha)\} \leq \alpha(1-\alpha)$ as above. However, since both δ_i and δ_j depend on the parameter estimators, they are not uncorrelated as in the case of BP, even if the

sampling is independent within the small areas. Nevertheless, given the EBLUP asymptotically converges to the BP in distribution, the EBLUP based intervals will converge towards the BP based intervals, and we may expect the simultaneous coverage of the intervals (5) to be close to the nominal level of confidence. For this we need consistent parameter estimators, with respect to *both* the population model *and* the sampling distribution. In the usual context where the EBLUP is applied, this is for example the case given the asymptotic settings for the second-order MSE estimation (Rao 2003). In particular, the within-area sample sizes can remain bounded. Since consistent parameter estimation is also necessary for the unconditional coverage (4), what we require is that the conditions (i) and (ii) above hold asymptotically, as $m \longrightarrow \infty$.

2.5. *Normality Assumptions*

Because we are using a model based approach, the intended design based coverage may fail in cases of model misspecifications. Good unconditional coverage of the interval (5) depends on the normality of $\hat{\theta}_i^H - \theta_i$ and the accurate MSE estimator. Thus, if the asymptotic second-order MSE estimators (Prasad and Rao 1990) are being used, then the regularity conditions for the MSE estimation are also needed for the interval estimation. Notice that also the MSE estimation assumes normality of the random effects introduced by the model, in which case the normality of $\hat{\theta}_i^H - \theta_i$ follows from the normality of $\hat{\theta}_i^H$ in addition.

We do not know in general how robust the model based intervals are under nonnormal θ_i and $\hat{\theta}_i^H$ although robustness of the MSE estimation towards nonnormal area-level random effects has been considered by Lahiri and Rao (1995). Taking the simple Fay-Herriot area-level model above, we have

$$\hat{\theta}_i^H = x_i^T\hat{\beta} + \hat{v}_i = x_i^T\hat{\beta} + \hat{\gamma}_i\{v_i + e_i - x_i^T(\hat{\beta} - \beta)\}$$

Of the random variables on the right-hand side, normality of $\hat{\beta}$ is usually a plausible assumption, whereas normality of $\hat{\gamma}_i$ is probably not entirely accurate even if $\hat{\sigma}_v^2$ is normal. The normality of v_i is a model assumption, whereas the normality of e_i is probably not true where the area sample size is small. In short, normality of $\hat{\theta}_i^H - \theta_i$ is unlikely to hold in all the areas even if v_i is normal. However, given normal v_i, nonnormality of e_i may not be crucial to the simultaneous coverage. The reason is that the simultaneous coverage of the direct designed based intervals are often quite robust under nonnormal e_i, which will be illustrated in the simulation studies later. Empirically, diagnostics can be used to check whether there are severe departures from the normality assumptions of the LMM. We refer to Rao (2003) for examples from practices.

2.6. *Bootstrap Calibration*

As with any confidence procedure, various departures from the underlying assumptions may cause the true coverage to deviate from the nominal level of confidence. A technique for adjustment is *calibration*. That is, if the $100\alpha\%$ nominal confidence intervals do not have $100\alpha\%$ coverage, intervals with nominal level $100\phi\%$ may do. Bootstrap calibration can be used to explore the mapping between ϕ and α. The idea is straightforward: given

the true population, we can repeatedly draw samples and derive confidence intervals at chosen nominal levels in order to find out their true coverages. In practice, however, we do not know the true population so that the bootstrap calibration must be conducted on the basis of some *plug-in* population. The rest is exactly the same.

Booth, Butler, and Hall (1994) proposed a nonparametric method for constructing the plug-in population based on stratified simple random samples. Let N_h and n_h be, respectively, the population and sample sizes in stratum h, for $h = 1, \ldots, H$. Let m_h be the integer part of N_h/n_h and let $k_h = N_h - m_h n_h$. The plug-in population in stratum h is given by m_h replicates of the within-stratum sample, plus a sample of size k_h selected randomly and without replacement from it. The different stratum populations are formed independently of each other. Thus, resampling from each stratum of the plug-in population does not differ much from direct nonparametric bootstrap resampling (i.e., randomly and with replacement) from the selected within-stratum sample, provided the sampling fraction n_h/N_h is negligible.

The method is explored in Section 4, where it does improve the coverage. However, the extent to which such improvements will hold in general remains an open question at this stage. This is because the bootstrap method is asymptotically justified when the population is considered as a sample from a super-population, where all the N_h's and n_h's diverge to infinity in such a way that each ratio N_h/N_g and n_h/n_g converges to a finite nonzero limit. Whether this holds for the given finite population is a question difficult to answer in general. For instance, if the design strata coincide with the small areas of interest, then many of the area sample sizes are probably not large enough to justify the asymptotic setting of $n_h \to \infty$. If possible, the performance of the bootstrap calibration over repeated sampling needs to be evaluated by simulations before it is endorsed.

2.7. *Generalized Linear Mixed Models*

For categorical data, such as binary or count data, generalized linear mixed models (GLMM) are more appropriate (Rao 2003, Section 5.6). Let μ be the mean of a response vector y. Let $h()$ be a monotonic link function, such that

$$h(\mu) = \eta = X\beta + Zv$$

where X and Z are the design matrices, β is the parameter vector containing the fixed effects, and v is the vector of random effects. In the small area estimation context, $\eta = (\eta_1, \ldots, \eta_m)^T$ is often a vector of m components corresponding to $\theta_1, \ldots, \theta_m$. It is convenient to consider prediction of η_i on the linear scale. An interval for η_i can be turned into an interval for θ_i through the inverse transformation $h^{-1}(\eta)$. Let $\hat{\eta}_i^H$ be the predictor of η_i, and let $\hat{g}_i$ denote its MSE estimator. A $100(2\alpha - 1)\%$ nominal level prediction interval is given as (5) with $\hat{\eta}_i^H$ replacing $\hat{\theta}_i^H$. Asymptotic simultaneous coverage depends on the normality of $\hat{\eta}_i - \eta_i$ as well as accurate MSE estimation. Typically, we need the normality assumption of v, which partly depends on the choice of the link function. Otherwise, the approach is similar to that under the LMM.

There is a harder computational issue under the GLMM. This is an area with rapid on-going research developments. See McCulloch and Searle (2001) for an account of the field. These authors favor the maximum likelihood estimation wherever possible.

In practice, however, a group of the so-called penalized quasi-likelihood (PQL) algorithms are often used because of the easiness of implementation (Schall 1991; Breslow and Clayton 1993; McGilchrist 1994). We will not go into the computational details here.

2.8. *Interval Based on Synthetic Estimator*

Regression-synthetic types of estimator are sometimes used in small area estimation (Rao 2003, Section 4.2). Even when the point estimators are acceptable, the evaluation of the uncertainty in the estimation will be misleading if it is done with respect to the underlying model, because a fixed effects model is unlikely to be able to fully capture the between area variation of the small area parameters of interest. For example, assume the following linear regression model

$$y_{ij} = x_{ij}^T \beta + \varepsilon_{ij}$$

where x_{ij} contains the covariates of the jth unit from the ith area, and ε_{ij} is independent of each other with variance σ_ε^2. The regression synthetic estimator of $\theta_i = \Sigma_{j=1}^{N_i} y_{ij}$ is given by $\hat{\theta}_i^S = X_i^T \hat{\beta}$, where $X_i = \Sigma_{j=1}^{N_i} x_{ij}$ and N_i is the population size, given negligible within area sampling fraction. Denote by τ_i the approximate *design based* sampling variance of $\hat{\theta}_i^S$. The $100(2\alpha - 1)\%$ nominal level interval $\left(\hat{\theta}_i^S - z_\alpha\sqrt{\tau_i},\ \hat{\theta}_i^S + z_\alpha\sqrt{\tau_i}\right)$ would in general have very misleading design based coverage, because it fails to recognize the design based bias in $\hat{\theta}_i^S$ that is always present in practice.

It is possible to introduce a 'second' model to improve the interval estimation. Put

$$\hat{\theta}_i^S = \theta_i + \xi_i + \varepsilon_i \quad \text{where} \quad \xi_i = E_\pi\left(\hat{\theta}_i^S\right) - \theta_i \quad \text{and} \quad \varepsilon_i = \hat{\theta}_i^S - E_\pi\left(\hat{\theta}_i^S\right)$$

Assume that $\xi_i \sim N(\xi,\ \sigma^2)$, where ξ_i is the design based bias of the synthetic estimator. Notice that this is *not* the 'first' linear regression model under which $\hat{\theta}_i^S$ has been derived. Once introduced, it implies the following $100(2\alpha - 1)\%$ nominal level interval

$$\lambda_i = \left(\hat{\theta}_i^S - \xi - z_\alpha\sqrt{\sigma^2 + \tau_i},\quad \hat{\theta}_i^S - \xi + z_\alpha\sqrt{\sigma^2 + \tau_i}\right)$$

assuming normality of ε_i, and independence between ξ_i and ε_i. In particular, this includes an extra variance component σ^2 not to be found in the naive confidence interval, which can be considered as a uniform adjustment due to the design based bias. An estimator of λ_i is obtained from replacing (ξ, σ^2, τ_i) by their estimates. For this purpose, observe that

$$\hat{\theta}_i^S - \hat{\theta}_i = \xi_i + \varepsilon_i - e_i \approx \xi_i - e_i$$

where $\hat{\theta}_i$ is the design based direct estimator and e_i is its sampling error. This is an LMM with one model based random effect ξ_i and two design generated sampling errors. The approximate two-component model follows from the fact that $\hat{\beta}$ is a global parameter estimator depending on the whole sample such that asymptotically ε_i will have much smaller variance than e_i.

It may be noted that the length of the adjusted interval above varies much less across the areas than the direct design based intervals. Firstly, the term σ^2 is the same everywhere, and is comparable to τ_i in many situations. Secondly, τ_i is the sampling variance of the synthetic estimator $\hat{\theta}_i^S$, and thus does not directly depend on the sample size in the i-th

area. As a result, a smaller area does not necessarily have a longer interval than a larger area. Thus, the efficiency gains over the direct intervals are typically largest for the areas with smallest sample sizes. The two-model approach here is conceptually less appealing than the EBLUP based approach earlier, where the same model is used for point as well as interval estimation. Because the resulting interval does not have an area-specific bias correction, it is on the whole less efficient than the mixed modeling approach. Nevertheless, it provides a more realistic measure of uncertainty for the regression synthetic estimator than that under the original fixed effects model.

3. Simulation Based on Wage Survey Data

The Norwegian wage survey is based on a yearly sample of clusters of wage earners. The clusters are establishments, which are stratified according to the size of the establishments. From each stratum except the largest one (where a census is carried out), a random sample of establishments are selected first, and all the employees from the selected establishment are then included in the sample. The primary variable of interest is the monthly wage of full-time employees in various subgroups of the population. For our simulation study we extracted the data from occupation group 5 (sales and service) in industry group 52 (retailing) in 2001 and 2002. The panel from the three counties in North Norway contains 1,269 persons. We take the 66 municipalities in these three counties as the small areas, and estimate the average monthly wage in each municipality in 2002, using the monthly wage in 2001 as the auxiliary variable.

Let $i = 1, \ldots, m$ (and $m = 66$) denote the small areas. Let y_{ij} be the monthly wage of the jth full-time employee from area i in 2002, and let x_{ij} be the monthly wage of the same person in 2001. Let $\bar{y}_i$ and $\bar{x}_i$ be the area sample means in the retrieved panel. These are fixed to be the area means of the synthetic population for this simulation study, denoted by $\theta_i^* = \bar{Y}_i^*$ and $\bar{X}_i^*$. To generate a sample from this fixed population, we assume negligible sampling fractions in all the areas, and draw randomly and with replacement pairs of residuals from $\{(x_{ij} - \bar{x}_i, y_{ij} - \bar{y}_i); i = 1, \ldots, m \text{ and } j = 1, \ldots, n_i\}$, where n_i is the area sample size in the panel. Adding this to $(\bar{X}_i^*, \bar{Y}_i^*)$, we obtain a pair of simulated observations from the corresponding area, denoted by (x_{ij}^*, y_{ij}^*). Notice that, in this way, the within-area covariance between the simulated pair of variables is the same across all the areas. We have thus a stratified random sampling design with the areas as the strata. The total sample size is 1,269, the area sample size varies from 3 to 215, the median area sample size is 8 and the mean is 19.2.

On the basis of each bootstrap sample, we derive the direct design based interval, as well as the model based interval (5), under the following nested-error regression model

$$y_{ij}^* = \beta_0 + x_{ij}^* \beta_1 + \upsilon_i + e_{ij}$$

(Battese, Harter, and Fuller 1988), where υ_i and e_{ij} are independent normal errors. The use of actual wage data implies that the normality of the random errors remains a model assumption. We record the proportion of θ_i^*'s that are covered by each set of confidence intervals. The average of this proportion over independent repeated bootstrap simulations yields then a Monte Carlo approximation to the simultaneous coverage of the corresponding nominal level of confidence.

In Table 1 the confidence intervals are compared to each other with respect to the simultaneous conditional coverage and the average relative length (ARL), given by

$$m^{-1}\sum_i 2z_\alpha\sqrt{\hat{g}_i}/\theta_i$$

for the $100(2\alpha - 1)\%$ model based intervals. The simultaneous coverage of the direct design based intervals approximately achieves the nominal level of confidence. The EBLUP based intervals (5) have simultaneous coverage about 2–4 percent below the nominal levels that we have looked at. Diagnostics for normality of v_i suggest that the distribution of v_i has heavier tails than the normal distribution on both ends, which seems to be the reason that the under-coverage increases slightly with the nominal level of confidence. However, considerable gains in efficiency have been achieved: the model based interval on average reduces the length of the direct confidence interval by more than 50%.

In Figure 1 the design based area specific coverages of the direct intervals are compared to those of the model based intervals. The nominal level of confidence is 95% in this case. It can be seen that the coverage of the model based intervals can be quite erratic as the area sample size decreases, for reasons explained earlier. The coverages of the design based intervals are closer to the nominal level of confidence. In particular, the average coverage across the small areas appears robust even when the sample size is as low as three.

4. Simulation Based on Census Household Data

4.1. Simulation Design

Small area household compositions are area population counts of households with respect to some classification of interest (e.g., the size of the household). Complete enumeration of the population is only available in the census. Post-censal updates need to be based on household surveys conducted at the national statistical offices. For our simulation study we retrieved the Norwegian census household compositions in 1990 and 2001. The 89 economic regions (analogous to the NUTS4 regional classification of EUROSTAT) are fixed as the domains of interest, i.e., $m = 89$. We set the area proportion of single-person

Table 1. Design based simultaneous coverage and average relative length of confidence intervals based on 250 simulated wage survey samples: A, Direct estimation; B, EBLUP under nested-error regression model. Monte Carlo standard error in parentheses

	Simultaneous coverage				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.806 (0.003)	0.896 (0.002)	0.942 (0.002)	0.964 (0.001)	0.979 (0.001)
B	0.780 (0.003)	0.863 (0.002)	0.910 (0.002)	0.936 (0.002)	0.958 (0.001)
	Average relative length				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.141 (0.000)	0.180 (0.001)	0.215 (0.001)	0.246 (0.001)	0.283 (0.001)
B	0.062 (0.000)	0.080 (0.000)	0.095 (0.001)	0.109 (0.001)	0.125 (0.001)

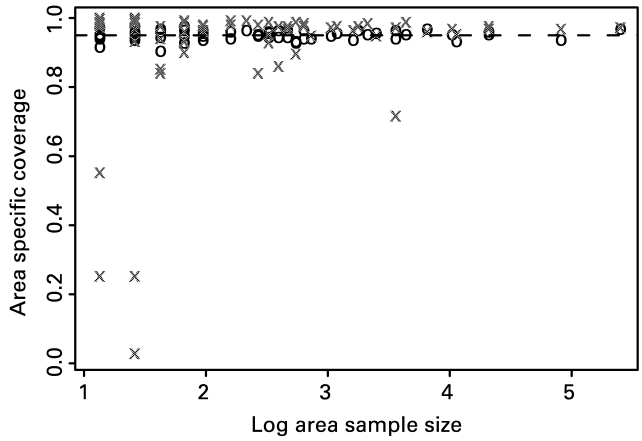

Fig. 1. Design based area-specific coverage of 95% nominal level confidence intervals (marked by the dotted horizontal line). Direct estimation (o) and EBLUP based estimation (x)

households in 2001 as the interest of estimation, using the corresponding area proportion in 1990 as the auxiliary variable.

Let θ_i be the area proportion of interest in 2001. Let p_i be the corresponding proportion in 1990. Let $\eta_i = \log(\theta_i) - \log(1 - \theta_i)$ and let $x_i = \log(p_i) - \log(1 - p_i)$. Simple regression of η_i on x_i yields approximate normal residuals with appreciable variance. We shall therefore adopt the following GLMM for this population

$$\text{logit}(\theta_i) = \eta_i = \beta_0 + x_i\beta_1 + v_i \quad \text{and} \quad v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$$

For sampling from the population we use stratified simple random sampling with the 89 areas as the design strata. The sampling design varies with respect to the sampling fraction, denoted by f. Simulations are carried out at $f = 1/50$, 1/150 and 1/500. On the basis of each simulated sample, we derive direct design based intervals as well as the GLMM based intervals, using the PQL algorithm outlined by McGilchrist (1994) to estimate the parameters $(\beta_0, \beta_1, \sigma_v^2)$ and the random effects v_i. In particular, σ_v^2 is estimated by the REML procedure. The MSE of $\hat{\eta}_i^H - \eta_i$ is estimated on the basis of a normal approximation to the quasi log-likelihood for $(\beta_0, \beta_1, v_1, \ldots, v_m)$.

4.2. Coverages

Denote by $\lambda_i^{(k)}$ the interval calculated from the kth simulated sample, for $i = 1, \ldots, m$. Let $I_i^{(k)} = 1$ if θ_i is covered by $\lambda_i^{(k)}$ and $I_i^{(k)} = 0$ if not. The Monte Carlo approximation of the true simultaneous coverage based on B simulated samples is then given by

$$\hat{\delta} = m^{-1}\sum_{i=1}^{m}\hat{\delta}_i = m^{-1}\sum_{i=1}^{m}\left(B^{-1}\sum_{k=1}^{B} I_i^{(k)}\right) = B^{-1}\sum_{k=1}^{B}\delta^{(k)} \quad \text{where}$$

$$\delta^{(k)} = m^{-1}\sum_{i=1}^{m} I_i^{(k)}$$

The $\delta^{(k)}$ is indeed the proportion of the θ_i's covered by the set of intervals based on the kth sample. It is the *realized* simultaneous coverage. The simultaneous coverage δ is its expectation on repeated sampling. But one would also be interested in the whole distribution of $\delta^{(k)}$. In particular, an estimate of the median of this distribution is given by the median of $\delta^{(1)}, \ldots, \delta^{(B)}$, whereas the Monte Carlo standard error of $\hat{\delta}$ is an estimate of the standard deviation of $\delta^{(k)}$ divided by $\sqrt{B}$.

In Table 2 the design based simultaneous coverage and the ARL of the confidence intervals are given. Also shown are the median values of $\delta^{(k)}$. It can be seen that the coverages of the design based direct intervals are on the whole quite good. Negative bias arises as the sampling fraction decreases, but is not serious even at $f = 1/500$. The main problem is inefficiency. The direct intervals are simply too long to be useful even at $f = 1/50$.

The GLMM based intervals are much shorter. The relative efficiency compared to the direct estimation increases as the sampling fraction decreases. The coverages of the model based intervals, however, are negatively biased. The problem has little practical consequences at $f = 1/50$, but increases as the sampling fraction decreases. The loss of coverage occurs mainly between $f = 1/50$ and $f = 1/150$. For example, the simultaneous coverage of the 95% nominal intervals drops from 94.1% at $f = 1/50$ to 85.2% at $f = 1/150$, whereas the coverages are about the same at $f = 1/150$ and $f = 1/500$, i.e., within the margins of Monte Carlo error. The median values of the realized simultaneous coverages $\delta^{(k)}$ are very robust as the sampling fraction decreases, and are in close agreement with the corresponding nominal confidence levels. The discrepancy between the mean and median values of $\delta^{(k)}$, however, suggests skewed distributions of the realized simultaneous coverage over repeated sampling at $f = 1/150$ and 1/500. The Monte Carlo standard errors increase quite a lot from $f = 1/50$ to $f = 1/150$ and 1/500. Together they indicate that the realized coverage may be particularly low on certain occasions.

4.3. Calibration

We explore the bootstrap calibration to see if it can improve the simultaneous coverage of the GLMM based intervals. For each simulated sample from the population, we carry out bootstrap calibration by stratified resampling using the method of Booth, Butler, and Hall (1994). Let B be the number of simulated samples from the true population, and let K be the number of resamples for each bootstrap calibration. The total number of simulated samples is then $B \times K$.

Table 3 shows the simulation results at $f = 1/150$ with $B = 500$ and $K = 40$. The Monte Carlo estimates of simultaneous coverages are consistent with the corresponding estimates in Table 2 within the margins of Monte Carlo error. Each pair of a nominal level of confidence and the corresponding calibrated simultaneous coverage is an estimate of the mapping between ϕ and α by the bootstrap calibration. Thus, according to the results in Table 3, the 'true' coverage is 0.863 at the nominal level 0.95, and it is 0.934 at the nominal level 0.99, and so on. In particular, a nominal level of 0.995 should yield a simultaneous coverage of 0.949, or approximately 95%. On account of the Monte Carlo error and the fact that the calibrated coverage can only be checked at chosen grid values,

Table 2. Design based simultaneous coverage and average relative length of confidence intervals based on simulations of household data: A, Direct estimation; B, GLMM based estimation. Median value of the realized simultaneous coverages (in italics). *Monte Carlo standard error in parentheses*

	Sampling fraction $f = 1/50$, number of simulations $B = 500$ Simultaneous coverage				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.797 (0.002) *0.798*	0.896 (0.001) *0.899*	0.947 (0.001) *0.944*	0.972 (0.001) *0.978*	0.988 (0.000) *0.989*
B	0.795 (0.005) *0.831*	0.891 (0.004) *0.921*	0.941 (0.003) *0.966*	0.966 (0.003) *0.989*	0.982 (0.002) *1.000*
	Average relative length				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.238 (0.000)	0.306 (0.000)	0.365 (0.000)	0.417 (0.000)	0.479 (0.000)
B	0.103 (0.001)	0.133 (0.001)	0.158 (0.001)	0.180 (0.001)	0.207 (0.002)
	Sampling fraction $f = 1/150$, number of simulations $B = 500$ Simultaneous coverage				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.795 (0.002) *0.798*	0.892 (0.002) *0.899*	0.941 (0.001) *0.944*	0.968 (0.001) *0.966*	0.985 (0.001) *0.989*
B	0.709 (0.010) *0.798*	0.800 (0.010) *0.899*	0.852 (0.009) *0.955*	0.883 (0.008) *0.978*	0.909 (0.007) *1.000*
	Average relative length				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.409 (0.000)	0.525 (0.000)	0.626 (0.000)	0.716 (0.000)	0.823 (0.000)
B	0.110 (0.002)	0.141 (0.002)	0.168 (0.003)	0.192 (0.003)	0.220 (0.004)

	Sampling fraction $f = 1/500$, number of simulations $B = 500$				
	Simultaneous coverage				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.780 (0.002) *0.775*	0.867 (0.002) *0.865*	0.917 (0.001) *0.921*	0.950 (0.001) *0.955*	0.969 (0.001) *0.966*
B	0.715 (0.010) *0.775*	0.800 (0.009) *0.899*	0.851 (0.008) *0.955*	0.885 (0.007) *0.978*	0.917 (0.006) *1.000*
	Average relative length				
Nominal	0.80	0.90	0.95	0.975	0.99
A	0.720 (0.001)	0.925 (0.001)	1.102 (0.001)	1.260 (0.001)	1.448 (0.001)
B	0.137 (0.003)	0.175 (0.004)	0.209 (0.005)	0.238 (0.005)	0.274 (0.006)

Table 3. Bootstrap calibration with 500 × 40 simulations, $f = 1/150$

Nominal confidence level	0.95	0.99	0.995	0.999	0.9995
Simultaneous coverage	0.857	0.911	0.923	0.944	0.950
Calibrated simultaneous coverage	0.863	0.934	0.949	0.970	0.975
Bias of bootstrap calibration	−0.006	−0.023	−0.026	−0.026	−0.025

the calibrated coverage cannot always be brought to the exact nominal level of confidence, even after repeated trials. For example, in this case we got 94.9% instead of exactly 95%. It now follows that the difference between the simultaneous coverage and the calibrated simultaneous coverage is an estimate of the coverage bias of the calibrated intervals over repeated sampling. For instance, the coverage bias is −0.026 for calibrated intervals aiming at the true coverage 0.95, which is a great improvement compared to the bias of the uncalibrated intervals.

5. Summary

The design based properties of the model based small area interval estimation have been considered. The area-specific coverages are uncontrollable in general. We propose to use the simultaneous coverage to evaluate the small area confidence intervals. Being the mean of the area-specific coverages, the simultaneous coverage is the expected proportion of areas covered by the set of confidence intervals, all of which are derived at the same nominal level of confidence. It is a meaningful concept in small area estimation, summarizing all the area specific coverages in a single number.

It has been shown that the coverage of the EBLUP based intervals achieves the nominal level of confidence asymptotically, as the number of small areas tends to infinity, given correct unconditional area specific coverage with respect to both the population model and the sampling distribution, and asymptotically independent sampling within the small areas. The asymptotic settings are similar to those required for the second-order MSE estimation under the LMMs. The adaption to the GLMMs is straightforward. Bootstrap calibration may improve the coverage over repeated sampling in situations where the normal approximation is poor. We have also discussed interval estimation based on regression synthetic estimators. While this is less efficient than estimation based on the random effects models, it provides more realistic measures of uncertainty than the naive intervals derived under the fixed effects model underlying the synthetic estimator.

6. References

Battese, G.A., Harter, R.M., and Fuller, W.A. (1988). An Error-components Model for Prediction of County Crop-Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83, 28–36.

Booth, J.G., Butler, R.W., and Hall, P. (1994). Bootstrap Methods for Finite Populations. Journal of the American Statistical Association, 89, 1282–1289.

Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88, 9–25.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 74, 269–277.

Heady, P. and Ralphs, M. (2005). EURAREA: An Overview of the Project and Its Findings. Statistics in Transition, 7, 557–570.

Lahiri, P. and Rao, J.N.K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association, 82, 758–766.

McCulloch, C.E. and Searle, S.R. (2001). Generalized, Linear, and Mixed Models. John Wiley and Sons.

McGilchrist, C.A. (1994). Estimation in Generalized Mixed Models. Journal of the Royal Statistical Society, Series B, 56, 61–69.

Prasad, N.C. and Rao, J.N.K. (1990). The Estimation of Mean Square Errors of Small Area Estimators. Journal of the American Statistical Association, 85, 163–171.

Rao, J.N.K. (2003). Small Area Estimation. New York: Wiley.

Rivest, L.-P. and Belmonte, E. (2000). A Conditional Mean Squared Error of Small Area Estimators. Survey Methodology, 26, 67–78.

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. Biometrika, 78, 719–727.

Received June 2004
Revised January 2007