HHS Public Access

Author manuscript

Int Stat Rev. Author manuscript; available in PMC 2020 August 19.

Published in final edited form as:

Int Stat Rev. 2018 August ; 86(2): 259–274. doi:10.1111/insr.12248.

Towards a Routine External Evaluation Protocol for Small Area Estimation

Alan H. Dorfman

Bethesda, Maryland 20814, USA

Summary

Statistical criteria are needed by which to evaluate the potential success or failure of applications of small area estimation. A necessary step to achieve this is a protocol-a series of steps-by which to assess whether an instance of small area estimation has given satisfactory results or not. Most customary attempts at evaluation of small area techniques have deficiencies. Often, evaluation is not attempted. Every small area study requires an *external evaluation*. With proper planning, this can be routinely achieved, although at some cost, amounting to some sacrifice of efficiency of global estimates. We propose a Routine External Evaluation Protocol to allow us to judge whether, in a given survey, small area estimation has led to accurate results and sound inference.

Keywords

bias; confidence interval; coverage; mean square error; mean square error estimate

1 Introduction

Small area estimation is employed worldwide in many important applications, for example determining the allocation of funds. It has long history and a rich literature with a variety of ingenious techniques and well-developed theory. Already in 1979, Purcell and Kish (1979) were writing a review article on small area estimation. Rao and Molina (2015) give a current compendium of theory and methods. A recent overview may be found in Pfeffermann (2013).

In 1979, there was a conference on Synthetic Estimation, the prominent small area estimation technique of the day. At its end, Richard Royall, the pivotal figure in current-day model-based sampling theory, issued a warning, which, with a slight modification of terms might still be applied to present day conferences on small area estimation:

‘A workshop of this sort, focused on a specific technique, can spur development, but it can also be dangerous. The danger is that, from hearing many people speak many words about [small area estimation], we become comfortable with the technique. The idea and the jargon become familiar, and it is easy to accept that “Since all these people are studying [small area estimation], it must be okay.” We must remain skeptical and not allow familiarity to dull our

dorfmans@erols.com.

healthy skepticism. (Royall, 1979)'. (Also quoted in 'Indirect Estimators in U.S. Federal Programs' (1996) edited by Wesley L. Schaible, 1996, p. 193.)

Why would someone, this author included, who thinks the proper understanding of survey sample inference lies in the proper use of models, hesitate over small area estimation, a procedure resting as it does on the sophisticated use of models? We will suggest an answer in the succeeding section.

1.1 A Thought Experiment: The 'Small Area Vise'

The enterprise of small area estimation arises because of the collision of two factors:

- **a.** **Demand** by policymakers, often legally mandated demand, for estimates in each of many small areas. These estimates for example may be part of a legal framework by which to allocate resources of one sort or another to the various small areas.
- **b.** **Limited budgets** (of money, time and energy) insufficient to collect enough data to allow straightforward estimation in each small area based on its own proper data ('direct estimation').

There is a general tacit assumption that these two factors are reconcilable; that while we would *rather* have recourse to sufficient data for each area to stand on its own, nonetheless, by adroit modeling of variables across areas, we can, by 'borrowing strength', meet the needs and demands of policymakers. But when is it not 'borrowing strength' but '**borrowing weakness**'?

Here is a thought experiment. Suppose (a) the **demand increases** or at least does not decrease and (b) the **budget decreases**. There is a call for ever increasing refinement in the estimates and there is ever decreasing resources. Policymakers having seen the productiveness of small area estimation, and being reassured by statisticians of its efficacy, cut back the budget ever more from year to year, and at the same time, like the Egyptians requiring bricks from the Hebrews without straw, demand more and more detailed estimates.

Surely there is a limit to how far this cycle could go on. If the resources were to dry up to *zero*, then the production of estimates would clearly be impossible (unless with a *very* strong Bayesian prior). There must be a tipping point, well before resources are non-existent, at which small area estimates become unsatisfactory, where for example their actual relative bias is beyond a bound we would regard as acceptable or associated confidence intervals are misleading. This raises the interesting statistical question: **by what statistical criteria do we judge that resources are too limited to produce satisfactory small area estimates?**

We do not attempt in this paper to answer this question. To answer it, we need to be able in general to evaluate small area projects and to have gained considerable experience in such evaluation. For the most part, our current experience in evaluation of small area estimation is inadequate. 'The main limitation of small area methods … has been the difficulty in validating a particular approach for a given … problem. Standard approaches … are not useful … do not adequately answer the question of how well these methods work compared to … a large sample survey in each locality'. (Srebotnjak *et al.*, 2010)

The problem is exacerbated by the fact that we turn to small area estimation precisely because of the fact that many if not most of the areas of interest are under-sampled or not sampled at all. For example, Table 1 gives the distribution of effective sample sizes (number of in-scope households) per area (county) in a recent year of the National Health Interview Survey carried out by the U.S. Centers for Disease Control, National Center for Health Statistics. The second row gives the number of counties having sample size in the corresponding cell in the first row. There are 3 143 counties and we note that the vast majority (2 307) have no sample at all. Nonetheless, we sometimes seek local estimates in all the counties (e.g. Raghunathan *et al.* 2007).

1.2 Variety of Inadequate Methods of Evaluation

The 'gold standard' of evaluation has been evaluation of results against large external data sets derived from censuses or administrative data (cf. Rao & Molina, 2015, p. xxvi). Such evaluations are large scale projects and can only with difficulty be carried out on a regular basis. Furthermore, the comparisons they offer tend to be surrogates for what we would really like; for example censuses tend to be out of date and some assumptions become necessary to bring their data into line with what the small area estimates are actually meant to target.

A variety of other evaluation procedures have been used over the years, each having some weakness: (1) that the small area estimates are *reliable*, that is do not change much from time to time, or place to place. A uniform estimate of zero, pulled out of a hat, is extremely *reliable*; (2) that the estimates have *smaller estimated mean square error than their direct estimation* counterparts-estimation of mean square error for small areas can be precarious and requires its own validation. Furthermore, if the direct estimates are weak, then being better than weak is not reassuring; (3) that the estimates are *benchmarked*, that is add up to reliable estimates on the large areas that enfold them-this criterion does not distinguish the comparative worth of uniformly equal estimates from disparate estimates adding to the same total; (4) that the *model fits* [for example Pfeffermann (2013, section 8)]-the very nature of estimation on small areas precludes there being enough data on the typical small area to tell whether a particular model fits its data or not; (5) *cross-validation*-again because in too many small areas, there is insufficient data for verification purposes; (6) methods that rely on comparisons of just the *heavily sampled small areas*-these can be outnumbered by the many extremely sparse small areas (including, often enough, those with no data) that might behave quite differently; (7) large scale *simulation studies* from administrative, census or large samples-these can give useful insights but satisfactory extrapolation to the case at hand has to be assumed; (8) evaluation of *previous small area projects* that resemble the current one. This can give important insights but leaves us vulnerable to changing conditions; (9) the fact that the estimates arise from *sophisticated statistical methodology* or heavy computing power; having heavy firepower is desirable but is not self-validating by itself; (10) *plausibility* of point estimates and confidence intervals; verification by subject matter experts, for example can reassure but carries risks of political pressure or dissension.

The key problem is the lack of data precisely where they are needed to verify the validity of assumptions (for example in the 2 307 small areas lacking any sample in Table 1). This is

the reason a model-based sampler might hesitate to simply embrace small area estimation. The means to verify the model are generally too sparse or lacking over groups of areas (for example the smallest or rural areas) that might differ in their behaviour from the (typically larger or urban) areas that are heavily sampled. There can be no built-in robustness to model failure, as, for example that which the model-based sampler seeks to achieve through balanced samples (Valliant *et al.*, 2000).

2 Towards a Routine External Evaluation Protocol

We should perhaps stress that we are here addressing the situation where small area estimation is *anticipated*. A survey is being carried out with some primary goals (for example efficient national estimates), but there is also the secondary goal of getting estimates for smaller areas where insufficient local data is anticipated.

2.1 Twin Goals: Accuracy and Sound Inference

We want to keep in mind the twin goals of the survey sample enterprise, which are the same as statistical estimation in general: (a) sharp accuracy (efficiency) and (b) sound inference. Accuracy: how close is the small area estimate to its target? Inference: does a confidence interval or its equivalent, derived in small area estimation typically from an estimate of mean square error, actually cover the target in accord with its stated coverage?

Both accuracy and inference strongly suggest the need for an **external measure of comparison**; data from outside the sample that can validate point estimates *and* interval estimates. We emphasize the need to validate confidence intervals or their equivalents. Validation of intervals is almost never carried out in practice and there is very little to draw on in the literature. Exceptions seem to be Brown *et al.* (2001) and Beresovsky *et al.* (2011).

Having an external basis of comparison does not necessarily mean an external ***census*** or very large alternate survey. Nor does it necessarily require verification for *every* small area. Needed is just a good representative independent sample of the set of small areas for which small area estimates are constructed, particularly those which the overall ‘wide area’ survey will have neglected. What is desirable in general is a procedure that can be done regularly for *any* survey in which the use of small area estimation is anticipated: a built-in Routine External Evaluation Protocol (**REEP**) that will enable us to evaluate the effectiveness of small area estimation in the particular survey at hand. This means we need to *plan* on such evaluation from the very beginning of the survey, at the design stage.

2.2 REEP Design: The Supplementary Sample of Samples

Every survey S where estimation for particular small areas is anticipated should allot a small portion of its resources for a supplementary independent sample S_A of these small areas, with a particular focus on those that will be weakly (or not at all) sampled in S. Then each of the areas a selected into S_A has a sample s_a taken within a sufficiently large to enable construction of a good direct estimate for variables of interest in a, *independently* of estimation using small area estimation from the main sample. The purpose is comparison of the small area estimates to their corresponding direct estimates from the supplementary samples and evaluation of the small area confidence intervals.

Let $A = \{a\}$ be the set of areas for which the global sample S is expected to supply small area estimates. Typically, S will sample the larger of the areas a heavily, with few units (possibly none) sampled in the smaller a. From A, let a not very large appropriate sample S_A of n_A areas be drawn; 'appropriate' may mean, for example simple random sampling (*srs*) or *srs* within the subclass of areas expected to be neglected by the main sample S. From each of the areas a in S_A, a supplementary sample s_a will be taken of size n_a, where n_a is large enough that the direct estimates based on s_a can be regarded as normally distributed with variances well estimated and not large. The direct estimates and variance estimates will then be available for shedding light on the corresponding small area estimates derived from the main sample S.

Note 1. To mitigate confusion, let us emphasize that sampling is here envisaged as taking place in *three* different ways: there is the main sample S, carried out with whatever (usually complex) design is called for and typically primarily aimed at estimates at levels higher than the small areas a; there is the supplementary validation sample S_A, which supplies a collection of areas a; then there are the several samples s_a intended to give accurate estimates for the areas $a \in S_A$, quite independently of any data S might supply. Let us refer to S_A as the *supplementary sample* or the *validation sample* and to the individual s_a's as the *local samples*.

Note 2: There is precedent for designing surveys that intentionally compromise large scale accuracy. The goal has been improved small domain estimates, for example Singh *et al.* (1994), Marker (2001), Longford (2006), Falorsi and Righi (2008), Molefe (2011), Molefe and Clark (2015). The message has been that a minor loss in accuracy in the principal estimates can afford important gains for the small area estimates. Here, the goal is different: evaluation of the small area estimation process itself for the particular survey.

Note 3: Once the supplementary sample has served its primary function of validating the small area estimates and providing diagnostics, there is nothing to prevent combining S with the extra data arising from S_A and getting a revised set of estimates for both small areas and S's primary targets. This point is discussed further in Section 4, but the implication is that the extra data collected can have a dual benefit.

2.3 REEP: Evaluation of Accuracy and of Inference

The data from S_A can be used to produce measures that evaluate small area estimates (including mean square error and interval estimates) and provide diagnostic clues if there are indications of faulty estimation. Some information may be gained by graphing small area estimates against corresponding direct estimates for areas a in S_A. We can get formal measures by getting summary statistics across S_A (or suitable partitions of S_A) on relative biases, relative absolute biases and by comparing small area estimates of mean square error to the squared differences between small area and direct estimates. It is important also to evaluate the confidence level of small area confidence intervals. We give details on possible approaches in the succeeding text.

Our list of techniques is meant to be suggestive, not exhaustive.

Suppose {a} are the targets (truth) for areas a in S_A. Let $\{\hat{\mu}_a\}$ be the corresponding direct estimates based on the supplementary samples, $\{\sigma_a^2\}$ their variances and $\{\hat{\sigma}_a^2\}$ the corresponding variance estimates. As is so frequently done in the small area literature, we shall bypass complications by assuming that $\hat{\sigma}_a^2 = \sigma_a^2$. Let $\{\tilde{\mu}_a\}$ be the small area estimates based on the main sample S, $\{\tau_a^2\}$, $\{b_a\}$ and $\{m_a^2 \equiv \tau_a^2 + b_a^2\}$ their variances, biases and mean square errors, respectively, and $\{\tilde{m}_a^2\}$ the corresponding estimates of mean square error (note that we take it for granted that small area estimates have a potential bias, possibly large relative to the corresponding variance).

If the small area estimation is working as hoped, then the average (mean) across S_A of the relative biases $(\tilde{\mu}_a - \mu_a)/\mu_a$ and the average of the absolute value of the relative biases, $|(\tilde{\mu}_a - \mu_a)/\mu_a|$ will be small. Looking to mean squared error estimates, we anticipate if there is some degree of homogeneity across the areas and if the estimates are on target, that $n_A^{-1}\sum_{a \in S_A}(\tilde{m}_a^2)/n_A^{-1}\sum_{a \in S_A}(m_a^2) \approx 1$ (in principle, we might prefer looking at $n_A^{-1}\sum_{a \in S_A}(\tilde{m}_a^2/m_a^2)$ but this quantity tends to be unstable. (An intermediate statistic would be $G^{-1}\sum_{g=1}^{G}\left[\sum_{a \in S_{Ag}}(\tilde{m}_a^2)/\sum_{a \in S_{Ag}}(m_a^2)\right]$, where S_A has been divided into G subgroups S_{Ag} that we believe to have internal mean squared error homogeneity.) These quantities, and the true confidence level of confidence intervals, depend on unknowns and cannot be calculated from the sample S on which the small area estimates are based. Confidence intervals, assumed here to be of the form $c_a = \left(\tilde{\mu}_a - z_{1-\alpha/2}\sqrt{\tilde{m}_a^2}, \tilde{\mu}_a + z_{1-\alpha/2}\sqrt{\tilde{m}_a^2}\right)$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution, should have $(1 - \alpha)$ 100 percentage or better coverage of the μ_a. Such intervals (based on estimated mean square error rather than variance) tend to be conservative, covering μ_a at least the nominal coverage level, provided the *mse* estimate is on target (cf. for example Cochran, 1977, p. 15). The situation reverses, when the confidence intervals are formed from the root of (estimated) variances (Särndal, *et al.*, 1992, p. 165); for further discussion, see Appendix B.

We look to the validation sample to provide ‘mirrors’ (indirect information) on the aforementioned quantities and on confidence levels.

The relative bias is assayed by $n_A^{-1}\sum_{a \in S_a}\{(\tilde{\mu}_a - \hat{\mu}_a)/\hat{\mu}_a\}$, the relative absolute bias by $n_A^{-1}\sum_{a \in S_a}|(\tilde{\mu}_a - \hat{\mu}_a)/\hat{\mu}_a|$ and the ratio of estimated mean square error to mean square error $\sum_{a \in S_a}\{\tilde{m}_a^2 + \hat{\sigma}_a^2\}/\sum_{a \in S_a}\{(\tilde{\mu}_a - \hat{\mu}_a)^2\}$. The $\hat{\sigma}_a^2$ intrudes in the numerator to account for the sample variation in $\hat{\mu}_a$. We shall refer to these three quantities as ‘diag rel bias’, ‘diag rel abs bias’ and ‘diag mse est’, diagnostics for the relative bias, relative absolute bias and ratio of estimated mean square error to mean square error, respectively. We can expect that there will be some distortion in our ‘mirrors’ due to the sampling variability of the validation estimates. Nevertheless, these diagnostics can provide valuable information as to how the small area estimation is working, much like residuals in regression can provide information about the true error structure.

The confidence interval c_a contains μ_a if and only if $t_a = \frac{\tilde{\mu}_a - \mu_a}{\sqrt{\tilde{m}_a^2}}$ lies in $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$, so if we could calculate t_a, then we could appraise the coverage by looking at the distribution of the t_a's across areas. But t_a is inaccessible because μ_a is unknown. Instead, we can, for a in S_A, calculate $t_{diff,a} = \frac{\tilde{\mu}_a - \hat{\mu}_a}{\sqrt{\tilde{m}_a^2 + \hat{\sigma}_a^2}}$. If σ_a^2 is reasonably small, the behaviour of $t_{diff,a}$ should be a good indicator of the behaviour of t_a (for more discussion, see Appendix B). We can appraise the behaviour of $t_{diff,a}$ by looking at its values across the a in S_A and this can provide a window into the behaviour of $t_a = \frac{\tilde{\mu}_a - \mu_a}{\sqrt{\tilde{m}_a^2}}$.

Note 4: A summary measure of $t_{diff,a}$ is the coverage $p_{cov} = P(|t_{diff,a}| \leq z_{1-\alpha/2})$, which can be taken as the average (mean) of $I(|t_{diff,a}| \leq z_{1-\alpha/2})$ over all areas of concern (e.g. Group 1 in the example later). If this is seriously less than the nominal, it will arouse concerns about the actual behaviour of t_a. However, p_{cov} itself is inaccessible, because we only have a sample S_A from the areas of concern. We must rely on an estimate of coverage $\hat{p}_{cov} = n_A^{-1} \sum_{a \in S_A} I(|t_{diff.a}| \leq z_{1-\alpha/2})$. This is a random variable whose variation allows for the possibility of misleading evidence. The frequency of misleading estimates of coverage will depend on n_A and on p_{cov}. For example suppose $n_A = 60$ and $p_{cov} = 95\%$ then, assuming that $\hat{p}_{cov}$ is binomial, the probability that $\hat{p}_{cov} < 90\%$ is about 3%. For the same n_A and $p_{cov} = 99\%$, the probability of $\hat{p}_{cov} < 95\%$ is about 0.3 %. This suggests it is worthwhile including a look at nominal coverage higher than 95%. Also, in planning, it suggests a consideration in deciding how large to take n_A.

3 Illustration and a Simulation Study-Modified Lahiri-Rao Populations

3.1 Fay-Herriot Model

We will consider variants of the Lahiri and Rao (1995) population, which has served as an illustrative basis in a great many small area papers since its inception and is based on the Fay-Herriot area level model (Fay & Herriot, 1979): The small area targets are $\mu_a = v_a + \eta_a$, $a = 1; : : : ; A$

Here, $\eta_a \sim N(0; \psi_a)$ a stochastic component, ψ_a, typically assumed unknown and v_a is fixed unknown. In the case of the Lahiri-Rao population, these components are assumed constant across areas: $\psi_a = \psi$ and $v_a = v$. The data available from the sample S are $Y_a = v + \eta_a + \varepsilon_a \equiv \mu_a + \varepsilon_a$ with $\varepsilon_a \sim N(0, D_a)$ the sampling error and η_a, ε_a independent of each other and across areas.

The sampling variances D_a are typically assumed *known*. There has been important recent work dealing with the fact that they are unknown and their estimates often volatile, for example Bell (2008), Hawala and Lahiri (2010), Maiti *et al.* (2014); see also, Rao and Molina (2015, section 6.4.1). However, to avoid complications, we shall treat the D_a as known in this paper.

Then we have the estimates:

$\tilde{\mu}_a = \gamma_a Y_a + (1 - \gamma_a)\tilde{v}$, where

$$\gamma_a = \tilde{\psi}/(\tilde{\psi} + D_a)$$

$$\tilde{v} = \sum_a \frac{Y_a}{\tilde{\psi} + D_a} / \sum_a \frac{1}{\tilde{\psi} + D_a},$$

where $\tilde{\psi}$ satisfies $\sum_a \frac{(Y_a - \tilde{v})^2}{\tilde{\psi} + D_a} = A - 1$, A the number of areas in S.

This is the original Fay-Herriot estimator of ψ It has many competitors but we limit ourselves to it here for simplicity.

We will use estimates of the mean square errors $\tilde{m}_a^2$ derived in (Datta *et al.*, 2005); these, for convenience, are given in Appendix A.

We can form confidence intervals $c_a = \left(\tilde{\mu}_a - z_{1-\alpha/2}\sqrt{\tilde{m}_a^2}, \tilde{\mu}_a + z_{1-\alpha/2}\sqrt{\tilde{m}_a^2}\right)$ that should contain μ_a in at least $(1 - \alpha)$ 100% instances.

3.2 A Lahiri-Rao Population and Variants

In the Lahiri and Rao (1995) population, the areas divide into five groups, where, within each group, samples of the same size are taken. They consider groups of small equal size, but we, mimicking the data in Table 1, will allow the groups to be quite large and we will focus on estimation in just one of them.

In our case, the sample variances within the five groups are taken to be D = (10000, 25, 4, 0.6, 0.1), respectively. (We will follow the Lahiri-Rao notation.) The first group in Table 1 had sample sizes = 0, which corresponds to infinite variance; to avoid programming exceptions, we instead simply assume a very large variance for direct estimates in the first group. Our focus will be estimation, inference and validation for this first group, where data are 'missing'. In all cases, our working model in constructing estimates will be this Fay-Herriott-Lahiri-Rao structure and we will use the estimates given above and in Appendix A.

We will consider four populations. In all cases, the number of areas in the five groups will be $\mathbf{N}_g$ = (1 200, 800, 500, 400, 100). From each population, we take a single sample S that comprises samples from all areas a; each having variance D_a depending on which group g the area belongs to.

Population 1. *Our primary population is generated according to the Lahiri-Rao formulation. Specifically, we take* v = (16, 16, 16, 16, 16) *(common mean for all areas in all groups) and* ψ = (1, 1, 1, 1, 1) *(common variance of the area deviations* η_a*). It may be worth noting that the coefficients of variation within each of the five groups are respectively*

$$cv \equiv \sqrt{D}/v = (6.250, 0.312, 0.125, 0.048, 0.020)\,.$$

In addition to Population 1, where the postulated model coincides with how the data actually arises, we will consider three deviant populations.

Population 2. *deviates from Population 1 only in having* $\psi = (4, 1, 1, 1, 1)$*; that is the variance of the area deviations is larger for Group 1 than for the groups where data are available.*

Population 3. *deviates from Population 1 only in having* $v = (18, 16, 16, 16, 16)$*; that is the fixed mean for each area in Group 1 differs from the corresponding means in the other groups.*

Population 4. *differs in structure from the others. It assumes the presence of a highly skewed (standard lognormal) size variable x, ordered so that the the smallest x are in Group 1 and the largest in Group 5, and further assumes that the area means satisfy* $v_a = \beta x_a$*; we took* $\beta = 1/2$*. The quartiles of x are given in* Table 2.

In all cases, we took the Lahiri-Rao model as the working model and employed the estimates for area means and for mean square error given earlier. Thus, we expect things to work well in Population 1 and possibly to misbehave in the other three populations. The question is how well our proposed diagnostics, employing data from the validation sample, reflect the underlying actual behaviour of the small area point estimates, their corresponding estimates of mean square error and the confidence intervals constructed from these.

3.3 Results

3.3.1 Behaviour of small area estimates across Group 1—

Table 3 gives the values of the percent relative bias, averaged over the 1 200 areas in Group 1 for each of the four populations. For Population 1, everything is well behaved, as anticipated: bias is small, on average, the mean square estimator approximates the average of the mean square error and coverage is on target. Each of the other populations goes awry. Population 2's bias is not too large, but the estimated mean square error seriously underestimates the actual mean square, so that nominal coverage of intervals seriously overstates actual coverage. Population 3 has serious biases and underestimates mean square error, with consequent poor coverage. Population 4 is a bit of an anomaly: the coverage is actually conservative, despite there being the most serious bias. The estimates of mean square error are somehow taking the bias into account and tracking the mean square error.

We emphasize that none of the earlier results would be known to the analyst, because they all require knowledge of the unknown μ_a's.

3.3.2 Diagnosis using a sample of 60 areas from each population—

We take a single simple random sample S_A of 60 areas from Group 1 from each of the populations, respectively. For each of the areas a selected into S_A, we take a sample having variance $D_a = 0.4$ (so intermediate to the sampling intensity in Groups 4 and 5). For each of the selected areas, we calculate the diagnostics described in Section 2.3 earlier and we average over the 60 areas. Table 4 gives the results for each population. We emphasize that these results *would* be available to the analyst.

The results reflect the hidden reality of Table 3. The reflection is not perfect. In Population 1, the 95% coverage is a bit low; in Population 4, the bias estimates are exaggerated. On the whole though, the underlying situation seems to be mirrored pretty well through these diagnostics. Table 4 gives averages across the areas in the supplementary sample, but one can also learn by looking at results for individual areas. Figure 1 plots the values of $t_{diff,a}$ for each of the 60 sampled areas. Ideally, most of the values will be spread between −2 and 2, getting sparser away from 0. This holds for Population 1. Population 2 sees a greater spread and the indication of problems is very clear for Populations 3 and 4.

3.3.3 Multi-runs—The results in Section 3.3.2 are for a single random chosen sample of 60 areas from the 1 200 areas composing Group 1, for each of the populations, and illustrate how one might go about making use of the data arising from a supplementary sample. In this section, we repeatedly take such samples to see how much variation there might be in our ability to assess the underlying situation.

For each population, we take 500 validation samples of size $n_A = 60$ in Group 1 using simple random sampling. Local samples are taken with variance equal to $D_a = 0.4$. For each run, summary statistics are calculated as in Table 4. Figures 2–5 show the distribution via histograms of each of the summary statistics for each of the populations, respectively.

In the main, the sort of indications that our single sample gave hold up across the runs.

In Population 1, none of the samples suggest anything seriously amiss with respect to bias. There is one isolated sample with coverage around 85% that might make us question our small area inferences. The mean square ratio seems the least stable of our indicators with a fair portion of samples suggesting that the mean square estimator is too small. The 95% coverage of *t.diff* actually leans to being greater than 95%, which is in keeping with idea that confidence intervals based on mean square error tend to be conservative.

In Population 2, there are one or two samples that might suggest inference is okay, but by and large, the coverages reflect well that our small area inferences are doing poorly. The mean square diagnostic points in the same direction, but there is considerable overlap with what was seen for Population 1. For the bias diagnostics also, a large number of samples would not clearly delineate between a Population 1 and Population 2 situation.

Thus, there is a suggestion that the *t-diff* statistic may be the most sensitive of the indicators.

Population 3 is unambiguous on all four diagnostics: relative bias is consistently negative, the estimated mean square error is consistently low, and the coverage gives a clear warning signal in all runs.

In Population 4, the diagnostics across runs mirror the mixed picture we saw in the population (Table 3), with often an indication of sharp bias, but satisfactory or conservative coverage.

4 Discussion

Current practice in small area estimation makes us vulnerable to our using very elegant and persuasive techniques that leave us in the dark as to whether they are actually working in the particular survey to which they are applied. This is a serious matter, especially because small area estimates are often used to make judgments on funding and other matters important to the body politic.

Although sporadic attempts at validation are made, they are often flawed, relying themselves on judgments that embody assumptions and speculations, as described in Section 1.2.

In this paper, we have suggested that every small area estimation project should carry with it means for checking validity in the form of an independent sample of areas that ordinarily go sparsely sampled or unsampled and so institute REEP, a Routine External Evaluation Protocol.

The data gathered from appropriately selected small areas can in the end be incorporated into overall estimates, having served their main purpose of validating the small area estimates (Note 3 earlier).

But what if the diagnostics indicated that the model was not adequate? Should we give up on doing SAE for the problem? Not necessarily. The first step would be to try an alternative model suggested by the results of the validation study. For example if the x variable in Population 4 were available, we might try incorporating it into the model (although, the coverage being satisfactory, we might rest with the original model, despite recognizing some bias in the estimates, so long as the estimated mean square errors were palatable.) Where opportunity presents, we would make use of internal diagnostics as well. We would try an alternative model and do a revalidation, in the same manner as the original validation process. This would be iterated until we had verified a model or exhausted possibilities. If the former, then we would take a final step of incorporating S_A into the model to get final estimates. If the latter, then we might have to acknowledge that in the present instance, small area estimation is failing.

The illustrative examples in Section 3 gave results that are cleaner than what we are likely to encounter in practice. To keep the examples clear, we assumed that the only set of areas of concern was Group 1, where the areas were essentially unsampled. If we were to include say Groups 2 and 3, we would expect any model failure in them to be less severe, because their data contribute more to the estimation of parameters, and we would therefore anticipate that the diagnostics will show up less sharply as well. Lesser problems might still be of concern but will be harder to detect.

The major questions facing us in putting REEP into practice are (1) how many small areas need to be sampled in our validation sample? (2) how heavily must each area in the sample be sampled? (3) what diagnostics based on the supplementary data will be illuminating?

(1) Taking samples of 60 areas worked pretty well in the artificial populations of this paper. Taking more will give greater precision in summary diagnostics. It is desirable this question

be explored further in a variety of practical settings. A particular concern will be to limit false negatives, for example low coverage using $t_{diff,a}$ when actually true coverage matches the nominal. See *Note* 4 in Section 2.3 earlier.

Our criterion for (2) is that the areas entering into the supplementary sample should be sampled heavily enough that estimates based on the data within an area will be precise and reasonably assumed to follow a normal distribution. In the present paper, we took samples that were intermediate between those most heavily sampled in the main survey and those sampled more moderately. Again, it will be worthwhile to explore how various choices in this regard play out in practical settings.

Criterion (2) has somewhat greater importance than (1). We might still be able to learn a good deal if the number of areas sampled is lessened, but if the the samples from the areas within the validation sample are too small, our measures cannot be expected to be satisfactory.

(3) We explored various diagnostics dependent on the small area estimates and the estimates from the supplementary sample. Perhaps the most useful of these, as verifying (or not) our inferences is $t_{diff,a}$. We anticipate that additional measures will be developed down the road; Brown *et al.* (2001) suggest diagnostics that might prove useful in the REEP context.

We have not discussed many small area methods, for example Bayesian methods and quantile approaches, where doubtless some modification to the diagnostics we have suggested will be in order. But the basic REEP idea should apply to them.

Routine External Evaluation Protocol is analogous to quality control in industrial production. It carries a cost of course, one to which survey administrators may be reluctant to agree. At bottom, the cost is some sacrifice in efficiency in upper level estimates and in areas that are typically heavily sampled. Precedent for such sacrifice is testified to by the several papers cited in Section 2.2 that aim at increased overall efficiency including for the small area estimates. There will, however, generally be a trade-off between achieving overall efficiency and being able to set up an adequate validity protocol such as REEP. Just as there is often a trade-off between bias and variance, so too there is an intrinsic tension between efficiency and validity. In the small area estimation literature, the focus has been almost exclusively on efficiency. Some balance is overdue.

APPENDIX A.: Estimation of Mean Square Error under the Lahiri-Rao Model

$$\tilde{m}_a^2 \equiv \tilde{E}\left[(\tilde{\mu}_a - \mu_a)^2\right] = g1 + g2 + 2g3 - (1 - \gamma_a)^2 b$$

$$g1 = \tilde{\psi} D_a / (\tilde{\psi} + D_a)$$

$$g2 = \{D_a/(\tilde{\psi} + D_a)\}^2 / \sum_a 1/(\tilde{\psi} + D_a)$$

$$g3 = 2A\left\{D_a^2/(\tilde{\psi} + D_a)^3\right\} / \sum_a 1/(\tilde{\psi} + D_a)^2$$

$$b = 2\left(At_2 - t_1^2\right)/t_1^3$$

$$t_1 = \sum_a 1/(\tilde{\psi} + D_a)$$

$$t_2 = \sum_a 1/(\tilde{\psi} + D_a)^2$$

APPENDIX B.: Anticipated Coverage of Confidence Intervals for Differences

Let $X \equiv \hat{\mu} \sim N(\mu, \sigma^2)$, where μ is a desired target and $Y \equiv \tilde{\mu} \sim N(\mu + \delta, \tau^2)$ where δ is Y's bias and $m^2 \equiv \delta^2 + \tau^2$ represents the mean square error of Y. Both X and Y are taken as estimating μ, so that their difference $Y - X$ estimates 0, albeit with a bias. We inquire in this appendix about the coverage properties of confidence intervals based (a) on the variance of $Y - X$ and (b) on the mean square error of $Y - X$. As noted in Section 2.3, for the case of a single variate, it is well known that (a) will tend to have lower than nominal coverage and that (b) tends to be conservative. It is convenient to spell this out for the situation we address here, namely, the difference of two variables, each aiming at the same target, one unbiased, the other (possibly) biased.

For (a), we seek $p_{1-\alpha,\,conv} = P\left(z_{\alpha/2} \le \frac{Y - X}{\sqrt{\tau^2 + \sigma^2}} \le z_{1-\alpha/2}\right)$, the coverage probability arising from a conventional confidence interval that ignores bias.

For (b), we want $p_{1-\alpha,\,mse} = P\left(z_{\alpha/2} \le \frac{Y - X}{\sqrt{m^2 + \sigma^2}} \le z_{1-\alpha/2}\right)$, the coverage probability arising from a confidence interval that implicitly incorporates bias into the component representing degree of accuracy. It is straightforward to show that

$p_{1-\alpha,\,conv} = F\left(z_{1-\alpha/2} - \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}\right) - F\left(z_{\alpha/2} - \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}\right)$ and

$p_{1-\alpha,\,mse} = F\left(z_{1-\alpha/2}\sqrt{1 + \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}} - \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}\right) - F\left(z_{\alpha/2}\sqrt{1 + \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}} - \frac{\delta}{\sqrt{\tau^2 + \sigma^2}}\right)$, where F is the cumulative distribution function of the standard normal distribution. We note that both expressions are scale invariant, that is unchanged if σ, τ, δ are replaced by σ^*, τ^*, δ^* respectively with $\sigma^* = |k|\sigma$, $\tau^* = |k|\tau$, and $\delta^* = k\delta$, for $k \neq 0$. Thus in calculating values, it is enough to hold τ fixed at some convenient value, say $\tau = 1$, and consider the effect of

different ratios σ/τ and δ/τ. It is worth noting also that the larger σ is (the larger the variability of X), the smaller the adjustment terms in either expression, and the less we are able to gain information about the bias and variance of Y from the distribution of $t_{diff} = \frac{Y - X}{\sqrt{m^2 + \sigma^2}}$. This fact is illustrated in Tables B1 and B2, based on the earlier expressions for $p_{1-\alpha,conv}$ and $p_{1-\alpha,mse}$

The basic message is: if we properly take into account mean square error, we get more and more conservative as the bias increases, and as the variance of the unbiased estimator shrinks. If we improperly aim only at getting variance, coverage gets weaker and weaker with larger bias and smaller sigma.

In the small area estimation context of this paper, we use neither variance nor mean square error, but rather an estimate of the mean square error. If the estimate is on target, we would be as in Table B2,

If sigma is not large, we should get a very good picture of how well the combination of our estimate and its accompanying mean square estimate are doing. The two tables are not extremes-the estimate of mean square error can be larger than the mean square error or lower than the variance; nonetheless, these tables serve as guideposts and give us an idea of what to expect.

Table B1.

Coverage probability arising from a conventional confidence interval, for difference of variables.

δ/τ σ/τ	0.1	0.2	0.5	1	1.5	2	5	10
0.1	94.89	94.55	92.12	83.11	67.96	48.8	0.13	0
0.2	94.89	94.56	92.2	83.47	68.73	49.95	0.16	0
0.5	94.91	94.63	92.68	85.45	73.13	56.78	0.6	0
1	94.94	94.77	93.56	89.1	81.45	70.7	5.76	0
1.5	94.96	94.86	94.11	91.41	86.77	80.14	20.8	0.02
2	94.98	94.91	94.43	92.68	89.71	85.45	39.12	0.6
5	95	94.98	94.89	94.56	94	93.22	83.47	49.95
10	95	95	94.97	94.89	94.74	94.55	92.12	83.11

Table B2.

Coverage probability arising from a confidence interval based on mean square error, for difference of variables.

δ/τ σ/τ	0.1	0.2	0.5	1	1.5	2	5	10
0.1	95	95	95.1	96.15	97.88	99.12	100	100
0.2	95	95	95.1	96.11	97.81	99.07	100	100
0.5	95	95	95.07	95.84	97.37	98.71	100	100

δ/τ σ/τ	0.1	0.2	0.5	1	1.5	2	5	10
1	95	95	95.03	95.39	96.37	97.62	99.99	100
1.5	95	95	95.01	95.16	95.67	96.54	99.87	100
2	95	95	95	95.07	95.32	95.84	99.48	100
5	95	95	95	95	95.01	95.04	96.11	99.07
10	95	95	95	95	95	95	95.1	96.15

References

Bell WR (2008). Examining sensitivity of small area inferences to uncertainty about sample error variances In Proceedings of Section on Survey Research Methods, pp. 327–334. Alexandria: American Statistical Association.

Brown G, Chambers R, Heady P & Heasman D (2001). Evaluation of small area estimation methods-an application to unemployment estimates from the UK LFS In Proceedings of Statistics Canada Symposium.

Beresovsky V, Burt CW, Parsons V, Schenker N & Mutter R (2011). Application of hierarchical Bayesian nodels with poststratification for small-area estimation from complex survey data In Proceedings of Section on Survey Research Methods, pp. 4745–4756. Alexandria: American Statistical Association.

Cochran WG (1977). Sampling Techniques, 3rd ed New York: John Wiley and Sons.

Datta G, Rao JNK & Smith DD (2005). On measuring the variability of small area estimators under a basic area level model. Biometrika, 92, 183–196.

Falorsi PD & Righi P (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. Surv. Method, 34(2), 223–234.

Fay RE & Herriot RA (1979). Estimation of income from small places: an application of James-Stein procedures to census data. J. Amer. Statist. Assoc, 74, 269–277.

Hawala S & Lahiri P (2010). Variance modeling in the U.S. small area income and poverty estimates program for the American community survey In Proceedings of Section on Survey Research Methods. pp. 4655–4663. Alexandria: American Statistical Association.

Lahiri P & Rao JNK (1995). Robust estimation of mean squared error of small area estimators. J. Am. Stat. Assoc, 82, 758–766.

Longford NT (2006). Sample size calculation for small-area estimation. Surv. Method, 32(1), 87.

Maiti T, Ren H & Sinha S (2014). Prediction error of small area predictors shrinking both means and variances. Scand. J. Stat, 41, 775–790.

Marker DA (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. Surv. Method, 27, 183–188.

Molefe WB (2011). Sample Design for Small Area Estimation, Ph.D Thesis, University of Wollongong (available online).

Molefe WB & Clark RG (2015). Model-assisted optimal allocation for planned domains using composite estimation. Surv. Method, 27, 183–188. 41, 377–387.

Pfeffermann D (2013). New important developments in small area estimation. Stat. Sci, 28(1), 40–68.

Purcell NJ & Kish L (1979). Estimates for small domains. Biometrics, 35, 365–384.

Raghunathan TE, Xie D, Schenker N, Parsons V, Davis WW, Dodd K & Feuer EJ (2007). Combining information from multiple surveys for small area estimation: A Bayesian approach. J. Am. Stat. Assoc, 102, 474–486.

Rao JNK & Molina I (2015). Small Area Estimation, 2nd ed Hoboken: John Wiley and Son Inc.

Royall RM (1979). Prediction models in small area estimation In Synthetic Estimates for Small Areas, National Institute on Drug Abuse, Research Monograph, Vol. 24 Washington: U.S. Government Printing Office.

Särndal E-K, Swensson B & Wretman J (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Schaible WL (1996). Indirect Estimation in Federal Programs. New York: Springer-Verlag.

Singh MP, Gambino J & Mantel HJ (1994). Issues and strategies for small area data. Surv. Method, 20, 3–22.

Srebotnjak T, Mokdad AH & Murray CJ (2010). A novel framework for validating and applying standardized small area measurement strategies. Public Health Metric, 8, 8–26.

Valliant R, Dorfman AH & Royall RM (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley and Son Inc.

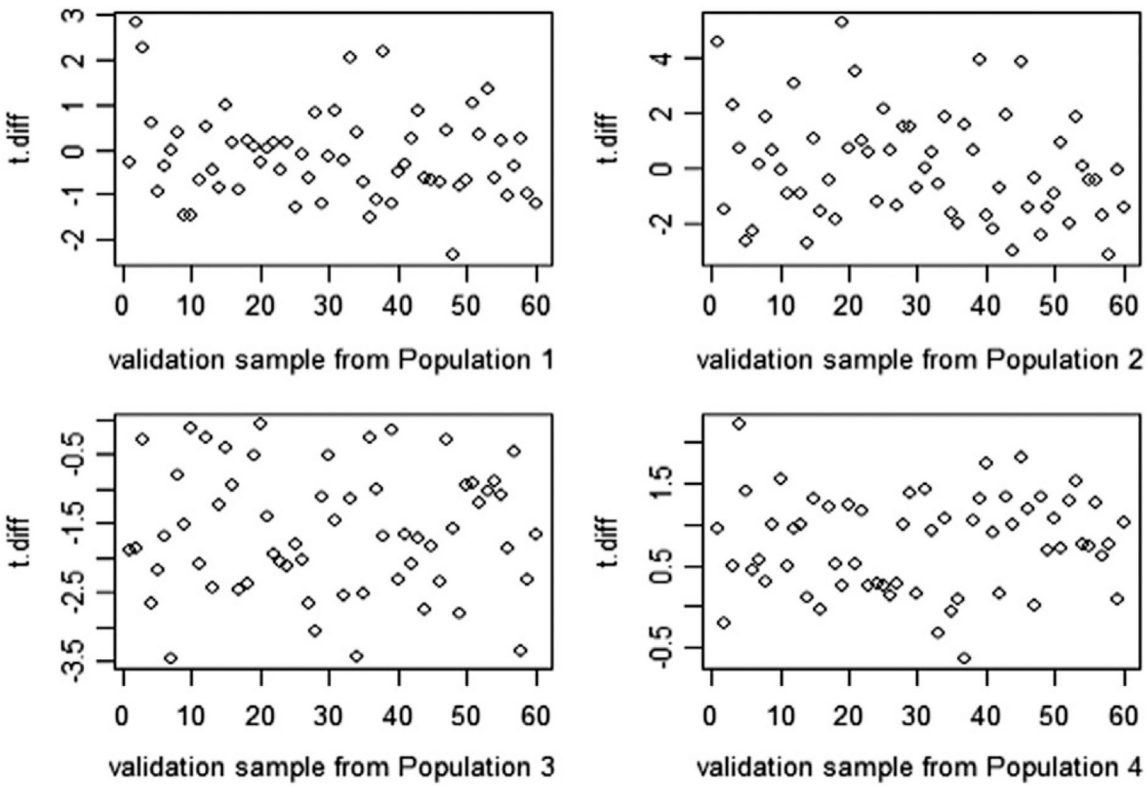

Figure 1.
t-Values differences across a validation sample of 60 areas from each of 4 populations.

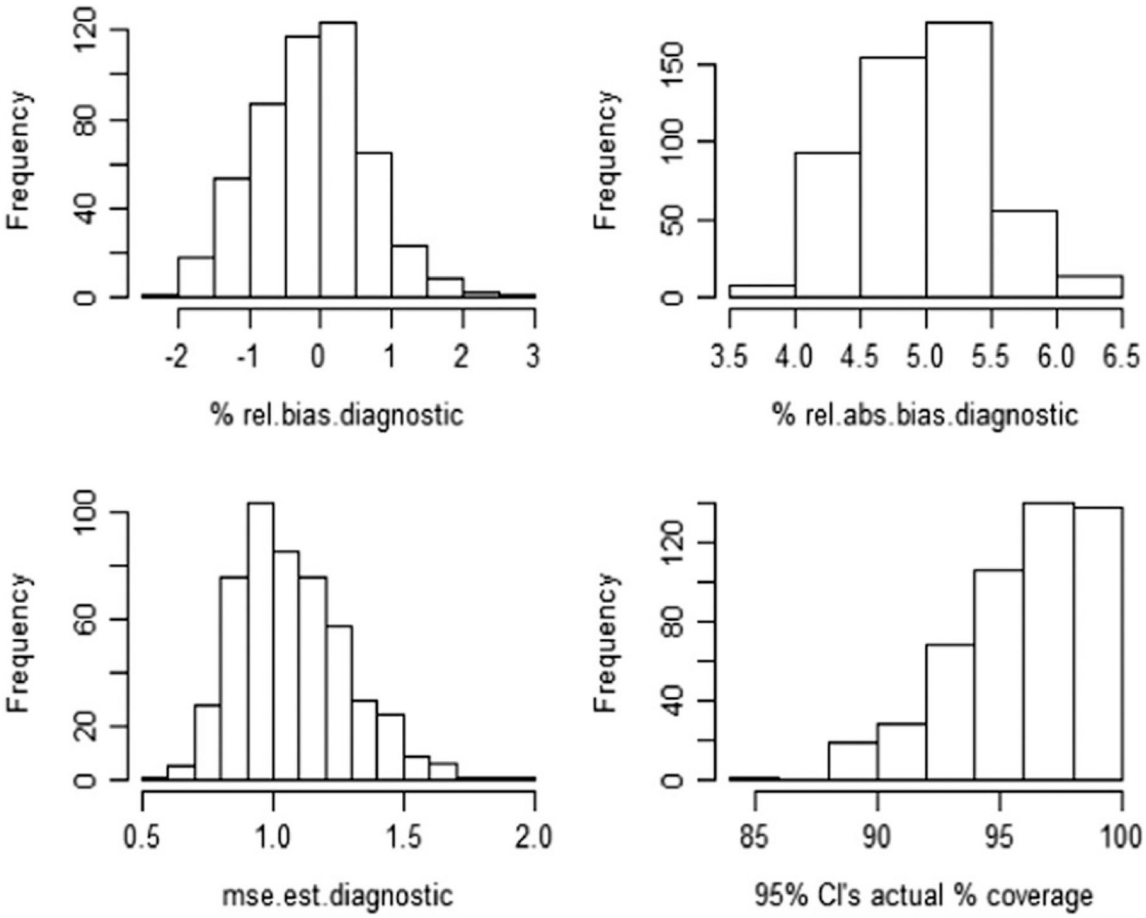

Figure 2.

Populations 1. Distributions of four diagnostics over 500 runs each a sample of size n_A = 60.

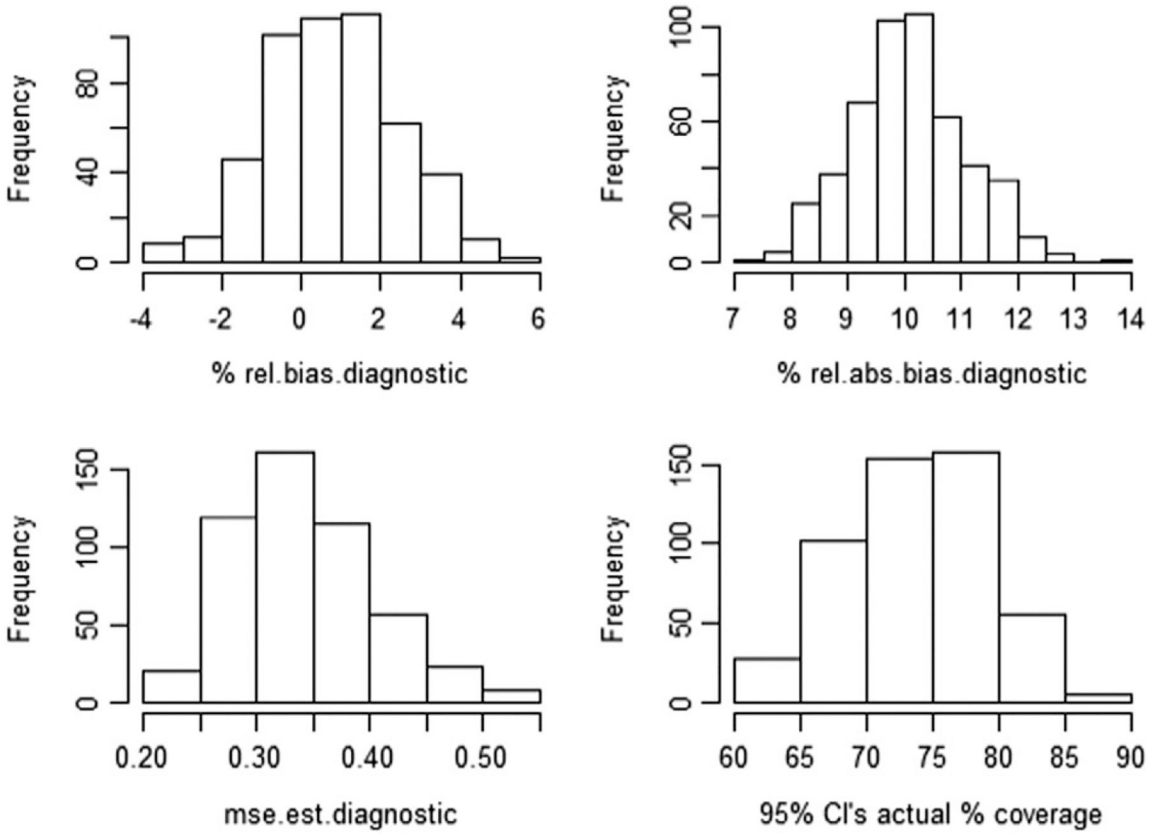

Figure 3.

Populations 2. Distributions of four diagnostics over 500 runs each a sample of size n_A = 60.

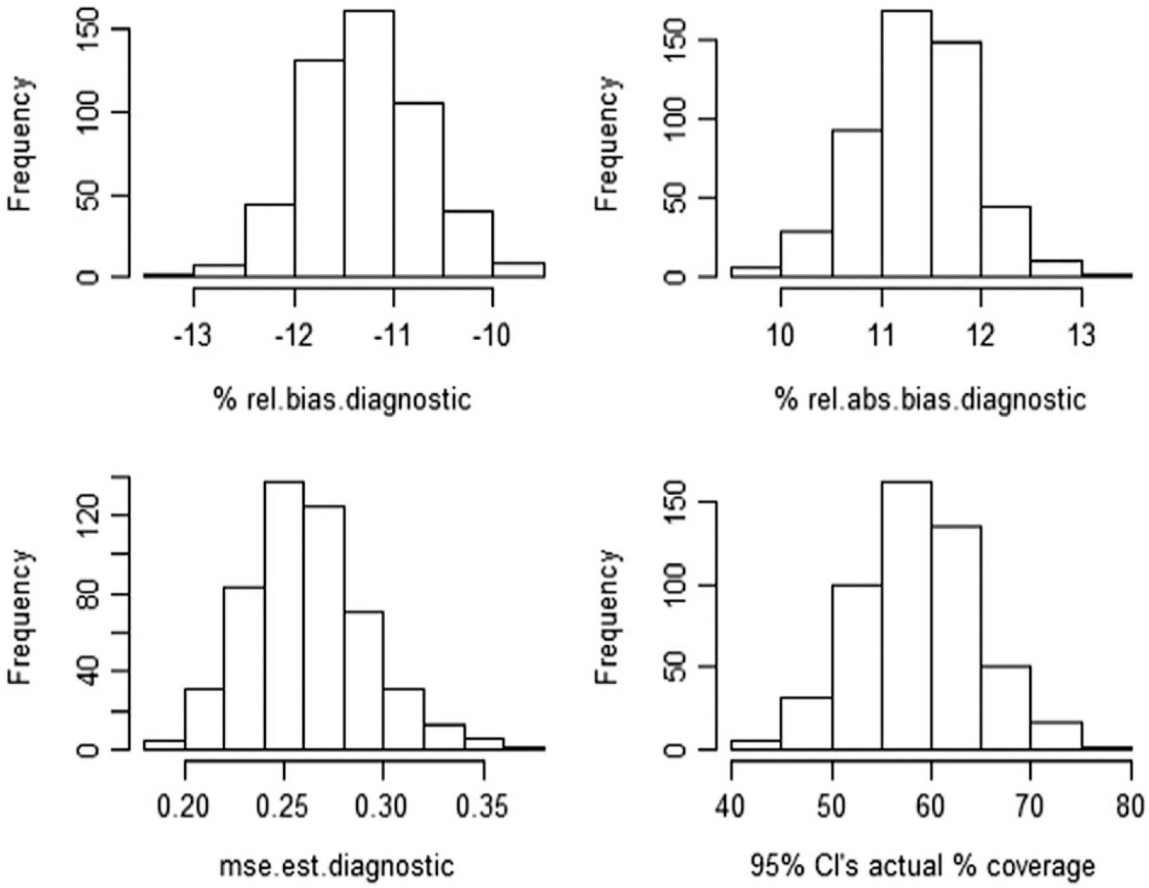

Figure 4.

Populations 3. Distributions of four diagnostics over 500 runs each a sample of size n_A = 60.

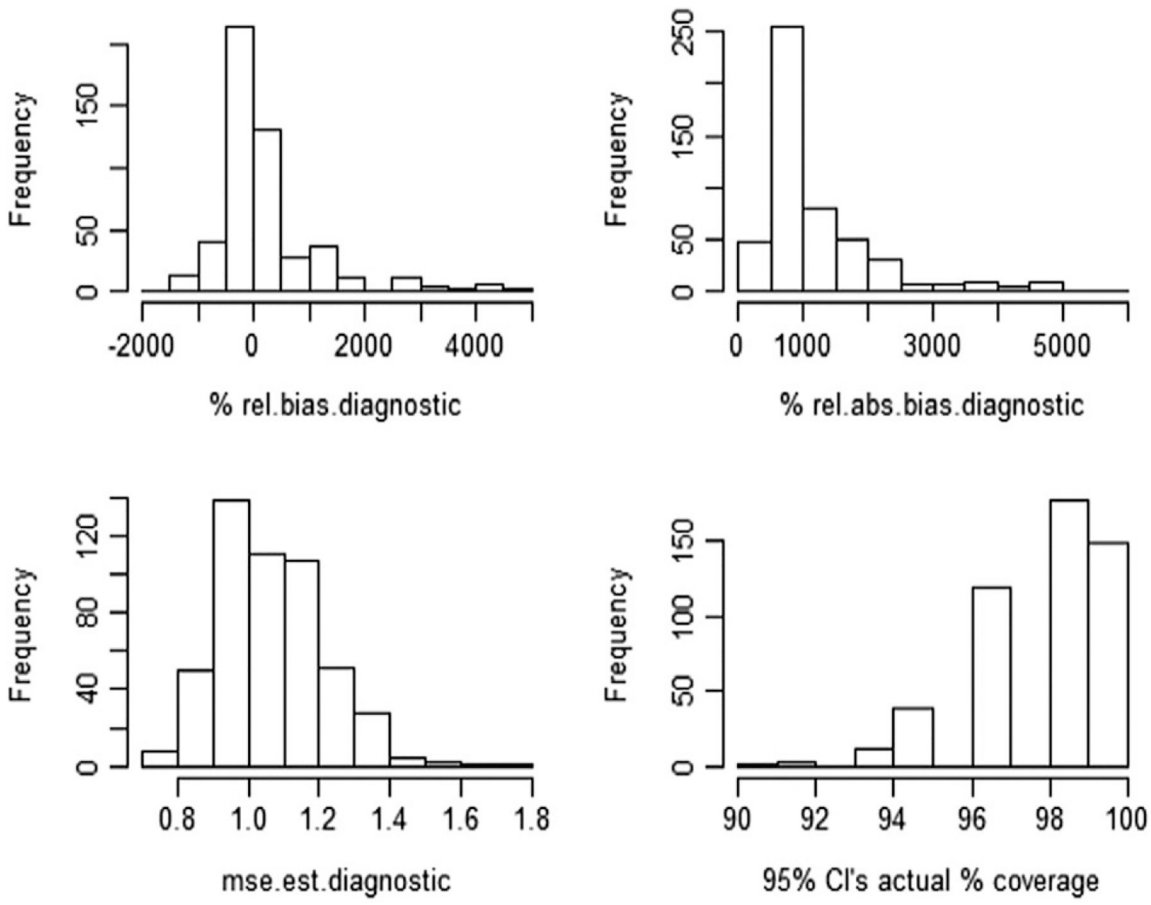

Figure 5.
Populations 4. Distributions of four diagnostics over 500 runs each a sample of size n_A = 60.

Table 1.

Frequency of counties having effective sample size in recent U.S. National Health Interview Survey.

Effective number of sampled units in area	0	(0,100)	[100,300)	[300,600]	(600,900]	>900
Frequency	2 307	497	251	68	11	9

Table 2.

Quantiles of size variable x for population 4.

Minimum	25.00%	50.00%	75.00%	Maximum
0.04	0.52	1	2.01	39.11

Table 3.

Summary statistics of small area estimates for 1 200 areas lacking sample in group 1 of 4 populations.

	% Relative Bias	% Relative absolute Bias	Mean Estimated mse/Mean mse	Nominal 95% coverage	Nominal 99% coverage
Pop1	−0.1	4.99	1.08	95.84	99.09
Pop2	0.96	10.17	0.27	69.92	82.5
Pop3	−11.25	11.34	0.21	48.92	74.25
Pop4	234.34	236.42	1.06	98.83	100

Table 4.

Summary statistics for small area statistics relative to validation values for a sample of 60 areas in group 1 in each of 4 populations.

	Diag % Rel Bias	**Diag % Rel Abs Bias**	**Diag Mean estimated mse**	***"95% Cov" %*** $\|t_{diff,a}\| \leq z_{.975}$	***"99% Cov" %*** $\|t_{diff,a}\| \leq z_{.995}$
Pop1	−0.51	5.41	1.02	91.67	98.33
Pop2	2.76	11.99	0.26	70	83.33
Pop3	−10.84	10.84	0.29	63.33	86.67
Pop4	2 909.26	3 631.94	1.08	98.33	100