The Canadian Journal of Statistics
Vol. 39, No. 1, 2011, Pages 147–164
La revue canadienne de statistique

A domain outlier robust design and smooth estimation approach

Li-Chun ZHANG* and Nina HAGESÆTHER

Statistics Norway, Kongensgate 6, P.O. Box 8131 Dep, Oslo 0033, Norway

Key words and phrases: Domain estimation; outlier robust; threshold sample; Winsorization; prediction

MSC 2010: Primary 62D05; secondary 62G35.

Abstract: Outliers that commonly occur in business sample surveys can have large impacts on domain estimates. The authors consider an outlier-robust design and smooth estimation approach, which can be related to the so-called "Surprise stratum" technique [Kish, (1965)]. The sampling design utilizes a threshold sample consisting of previously observed outliers that are selected with probability one, together with stratified simple random sampling from the rest of the population. The domain predictor is an extension of the Winsorization-based estimator proposed by Rivest and Hidiroglou (2004), and is similar to the estimator for skewed populations suggested by Fuller. It makes use of a domain Winsorized sample mean plus a domain-specific adjustment of the estimated overall mean of the excess values on top of that. The methods are studied in theory from a design-based perspective and by simulations based on the Norwegian Research and Development Survey data. Guidelines for choosing the threshold values are provided. *The Canadian Journal of Statistics* 39: 147–164; 2011

Résumé: Il est fréquent d'observer des valeurs aberrantes dans les enquêtes d'entreprises et celles-ci peuvent avoir des impacts majeurs dans les estimations d'un domaine. Les auteurs considèrent un plan de sondage robuste par rapport à la présence de valeurs aberrantes et une approche d'estimation lisse qui peuvent être reliées à la technique dite de la (voir Kish, 1965). Le plan de sondage utilise un échantillon à seuil qui consiste à combiner toutes les valeurs aberrantes déjà observées à un échantillon aléatoire simple stratifié pour le reste de la population. Le prédicteur du domaine est une généralisation de l'estimateur avec regroupement frontalier proposé par Rivest et Hidiroglou (2004) et il est similaire à l'estimateur pour les populations asymétriques suggéré par Fuller (1991). Il utilise la moyenne échantillonnage avec regroupement frontalier en plus d'un ajustement, spécifique au domaine, de la valeur estimée de la moyenne globale des valeurs excédentaires. Ces méthodes sont étudiées théoriquement d'un point de vue du plan de sondage et par des simulations basées sur les données provenant d'une enquete norvégienne sur la recherche et le développement. Des recommandations pour choisir les valeurs de seuillage sont aussi proposées. *La revue canadienne de statistique* 39: 147–164; 2011

1. INTRODUCTION

In business sample surveys even if outliers do not affect aggregated estimates substantially, their impact can be large for domain estimates. For repeated surveys, the estimate for a given domain can vary greatly over time if outliers occur in the sample in some periods and not others, causing volatility in the change estimate. Explicit model-based small area estimation techniques that are robust towards the presence of outliers have begun to receive interest in the past few years. Chambers and Tzavidis (2006) suggested the use of outlier-robust M-quantile models for

* *Author to whom correspondence may be addressed.*
E-mail: lcz@ssb.no

small area estimation, and Sinha and Rao (2009) proposed robust empirical best linear unbiased predictor (EBLUP) under the linear mixed models.

This study has been motivated by the existing practice in the yearly Norwegian Research and Development Survey (NRDS). The NRDS sample consists of three parts: (1) a self-representing (sub-) sample of the largest enterprises, covering about 80% of the R&D-total in the population; (2) an additional self-representing *threshold* sample containing outliers identified in the previous round of NRDS, that is, units with R&D-value exceeding a chosen threshold, covering just below 10% of the population R&D-total in most cases, and (3) a stratified simple random sample from the rest of the population. Intuitively, the use of such a threshold sample seems sensible since the pool of population outliers from which representative sample outliers might be drawn will be reduced, provided the observed outliers tend to remain large in the following year. On the other hand, one needs to be careful that the threshold sample does not get too large compared to the probability sample; otherwise the estimation precision for the rest of the population may suffer too much.

Moreover, one is concerned with a particular feature associated with the use of a threshold sample. Consider, for example, an outlier that first turns up in the probability sample, which is then placed in the threshold sample in the next round. Even if the unit has the same R&D-value in both years, its contribution to the respective totals will be quite different due to the different weights assigned to it. In other words, the in- and out-flows to the threshold sample may cause instability even though the R&D-value itself is stable. While such "noise" may cancel out on an overall level, the effects can be obvious at a disaggregated level. Thus, justifications on a more theoretical basis are desirable in order to implement the threshold-sample design. In addition, an estimation methodology that is able to control the influence of the probability-sample outliers at the domain level will be useful.

We shall develop a domain estimator that is a prediction extension of the estimator proposed by Rivest and Hidiroglou (2004). Like the threshold-sample design, this estimator can also be related to the "Surprise stratum" technique (Kish, 1965, Section 12.6C). It is constructed as a domain Winsorized sample mean plus a domain-specific adjustment of the estimated overall mean of the excess values on top of that. The sum of these domain estimators is equal to the direct design-unbiased population total estimator, under the assumption that outliers do not cause problems at the aggregated level. Our domain estimators retain this feature. At the same time, our estimator is similar to the estimator for skewed populations suggested by Fuller (1991). The difference is that, whereas the Fuller's estimator uses a robust estimator for the superpopulation mean, we use a Winsorization-based estimator for the finite population mean outside of the observed sample.

The domain outlier-robust design and the smooth estimation approach developed subsequently in this paper have several important features that need to be made clear at once. Firstly, only stratified simple random sampling in combination with a threshold sample will be considered, which is typical in business survey applications. We shall not discuss complex sampling designs beyond that. Secondly, the threshold-sample design relies heavily on a continuing survey environment, in order to identify potential outliers on the basis of historical data. It is not applicable in a one-off survey situation. Thirdly, the domains of interest are known in advance at the design stage, such that direct design-based domain estimators would have been considered acceptable in the absence of outliers. This is the reason that we shall maintain the design-based outlook, instead of resorting to model-based small area estimation techniques. Fourthly, underlying the modified Winsorization-based estimation methodology is the assumption that the outliers are able to destabilize the domain estimators, but not the overall population estimator. (In the case of NRDS, the largest units are already covered by the self-representing cut-in sample. The threshold-sample design only deals with the remaining population, which is about 20% of the total of interest.)

Hence, we shall describe the approach as *smooth* domain estimation, rather than outlier-robust per se. Despite these limitations, however, we believe that this setting is relevant for enough business survey applications to warrant its own treatment.

The rest of the paper is organized as follows. The threshold-sample design is studied in theory in Section 2. In Section 3 we consider smooth domain estimation with and without the presence of a threshold sample, and derive the design mean squared error (MSE). The design and estimation approach is evaluated in Section 4 based on a synthetic population constructed using the NRDS data. Conditions under which the approach can be made outlier resilient are described and examined. Finally, a short summary is given in Section 5.

2. THRESHOLD SAMPLE DESIGN

The idea of surprise stratum is to include potentially large observations (i.e., outliers) in the sample with probabilities that are higher than usual, so as to reduce the weights of these units without introducing bias. In the extreme case where a potential outlier is selected with probability one, it is converted from a representative outlier to a nonrepresentative one (Chambers, 1986), and receives a unit weight in estimation. We define a *threshold sample* to contain all the sample units that exceed a given value (i.e., threshold) in the previous survey. The threshold value itself, though, can be changed from one time point to another. Moreover, we notice that outliers may be cyclical or seasonal in certain populations, in which case one needs to be more careful and alternative ways of defining the threshold sample should be explored. Below we first introduce the necessary notations, and highlight the important factors for design efficiency through a simple motivating example. We then focus on the comparison between simple random sampling (SRS) with and without the use of a threshold sample for level estimation. Finally, we discuss how the results can be applied to stratified SRS design and change estimation.

2.1. A Motivating Example

Consider the estimation of a population total given by $Y = \sum_{i \in U} y_i$, where $U = \{1, 2, ..., N\}$ denotes the population and y_i is the variable of interest for the ith unit. Denote by s a sample of size n, regardless of the sampling design. Given the use of a threshold sample, let s_A denote the threshold sample of size A, and let $s_B = s/s_A$ be a simple random sample of size $n - A$ from the rest of the population, denoted by $U_B = U/s_A$. Let R be the threshold value. Let $U_+ = \{i \in U; y_i \geq R\}$ be of size N_+, and $U_- = \{i \in U; y_i < R\}$ of size N_-. Similarly, let $s_+ = \{i \in s; y_i \geq R\}$ be of size n_+, and $s_- = \{i \in s; y_i < R\}$ of size $n_- = n - n_+$. Given the use of a threshold sample, let $s_{A+} = \{i \in s_A; y_i \geq R\}$ be of size A_+, and $s_{A-} = \{i \in s_A; y_i < R\}$ of size $A_- = A - A_+$, such that $s_{B+} = \{i \in s_B; y_i \geq R\}$ is of the size $n_+ - A_+$ and $s_{B-} = \{i \in s_B; y_i < R\}$ is of the size $n_- - A_-$. An illustration of the setting is given in Figure 1 below.

As a motivating example, consider the case where there are only *two* distinct y values in the population: one below and one above the threshold value. Let $\hat{Y} = N \sum_{i \in s} y_i / n$ be the expansion estimator given the SRS. Let $\tilde{Y} = \sum_{i \in s_A} y_i + (N - A) \sum_{i \in s_B} y_i / (n - A)$ be the estimator given the

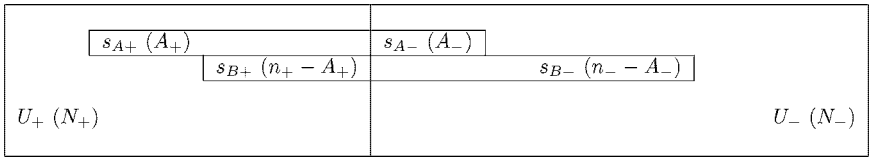

FIGURE 1: An illustration of population and sample division by threshold value. Size in parenthesis.

use of the threshold sample. Denote by V the sampling variance. It is easily verified that

$$\frac{V(\tilde{Y}|A, A_+)}{V(\hat{Y})} = \left(1-\frac{A}{N}\right)^{-1}\left(1-\frac{A}{n}\right)^{-1}\left(1-\frac{A_+}{N_+}\right)\left(1-\frac{A_-}{N_-}\right) \tag{1}$$

It can be seen that the use of the threshold sample can be justified if A_+ is sufficiently large compared to N_+. Thus, as a measure of the effectiveness of the threshold sample, we define the *catch rate* to be

$$\xi = \frac{A_+}{A}$$

A threshold sample is *effective* if the catch rate is high. Expression (1) shows that the ideal is to use as small as possible a threshold sample to catch as many as possible outliers. However, even the most effective threshold sample may not be helpful unless the outliers are "rare" enough. This can be noted in the extreme case of $\xi = 1$, that is, $A = A_+$ and $A_- = 0$, where a necessary condition for $V(\tilde{Y}|A, A_+) < V(\hat{Y})$ is $A/N_+ > A/n$, which amounts to

$$\theta \stackrel{\text{def}}{=} N_+/N < n/N \stackrel{\text{def}}{=} f$$

where θ is the *prevalence* of the outliers in the population, and f is the overall sampling fraction. In practice, the prevalence θ can be reduced by choosing a larger threshold value. However, raising the threshold value also affects the threshold sample size A and, potentially, the catch rate ξ, causing changes to A_+/N_+ and A/n at the same time. To examine the design efficiency more closely we need a general expression for the design effect.

2.2. Design Effect for Level Estimation

Let $\bar{Y} = Y/N$ and $\Delta \stackrel{\text{def}}{=} (N-1)\sigma^2 = \sum_{i\in U}(y_i-\bar{Y})^2$. For $\hat{Y}$ given the SRS, we have

$$V(\hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sigma^2 = \left(\frac{N}{n}\right)\left(\frac{N-n}{N-1}\right)\Delta \tag{2}$$

Let $\bar{Y}_+ = \sum_{i\in U_+} y_i/N_+$ and $\bar{Y}_- = \sum_{i\in U_-} y_i/N_-$. Let $\sigma_+^2 = \sum_{i\in U_+}(y_i-\bar{Y}_+)^2/(N_+-1)$ and $\sigma_-^2 = \sum_{i\in U_-}(y_i-\bar{Y}_-)^2/(N_--1)$. We can rewrite Δ as

$$\begin{aligned}\Delta &= \sum_{i\in U+}(y_i-\bar{Y}_+ + \bar{Y}_+-\bar{Y})^2 + \sum_{i\in U-}(y_i-\bar{Y}_- + \bar{Y}_--\bar{Y})^2 \\ &= (N_+-1)\sigma_+^2 + (N_--1)\sigma_-^2 + \frac{N_+N_-}{N}(\bar{Y}_+-\bar{Y}_-)^2 \end{aligned} \tag{3}$$

Next, let $\bar{Y}_B$ be the target mean in U_B. Given the use of the threshold sample, the conditional sampling variance of $\tilde{Y}$ given $\psi = (A, \xi)$ can be obtained as

$$V(\tilde{Y}|\psi) = E\{V(\tilde{Y}|s_A)|\psi\} + V\{E(\tilde{Y}|s_A)|\psi\}$$

where the outer conditional variance and expectation are with respect to s_A given ψ, and the inner conditional variance and expectation are with respect to s_B given s_A. We have $V\{E(\tilde{Y}|s_A)|\psi\} = 0$ since $E(\tilde{Y}|s_A) = Y$. Moreover, $V(\tilde{Y}|s_A) = (N-A)(N-n)\sigma_B^2/(n-A)$, where

$\sigma_B^2 = \sum_{i\in U_B}(y_i - \bar{Y}_B)^2/(N-A-1)$. It can be shown that

$$V(\tilde{Y}|\psi) = \left(\frac{N-A}{n-A}\right)\left(\frac{N-n}{N-A-1}\right)E(\Delta_B|\psi) \tag{4}$$

where $\Delta_B = (N-A-1)\sigma_B^2$ and, as shown in Appendix,

$$E(\Delta_B|\psi) = (N_+ - A_+ - a_+)\sigma_+^2 + (N_- - A_- - a_-)\sigma_-^2 + \frac{(N_+ - A_+)(N_- - A_-)}{N-A}(\bar{Y}_+ - \bar{Y}_-)^2 \tag{5}$$

for $a_+ = 1-(A_+/N_+)(N_- - A_-)/(N-A)$, and $a_- = 1-(A_-/N_-)(N_+ - A_+)/(N-A)$.

Dividing (4) by (2) we obtain the design effect of the threshold sample as

$$\frac{V(\tilde{Y}|\psi)}{V(\hat{Y})} = \left(\frac{N-1}{N}\right)\left(\frac{N-A}{N-A-1}\right)\left(\frac{n}{n-A}\right)\gamma \quad \text{where} \quad \gamma = \frac{E(\Delta_B|\psi)}{\Delta} \tag{6}$$

In practice we would mostly be concerned with the situations where the first two factors on the right-hand side of (6) are close to unity. The third factor can be interpreted as a penalty term for having only $n-A$ free observations compared to n under SRS. The potential gain from using the threshold sample comes from the last factor γ. We may compare its numerator and denominator, given respectively by (5) and (3), term by term. The ratio of the multipliers is given, respectively, as $(N_+ - A_+ - a_+)/(N_+ - 1) \approx 1 - A_+/N_+$ for σ_+^2, and $(N_- - A_- - a_-)/(N_- - 1) \approx 1$ for σ_-^2, and $(1-A/N)^{-1}(1-A_-/N_-)(1-A_+/N_+) \approx 1-A_+/N_+$ for $(\bar{Y}_+ - \bar{Y}_-)^2$. Let

$$\phi = (N_- - 1)\sigma_-^2/\Delta$$

which represents the contribution to the variance of the non-outliers. Then we have $\gamma \approx \phi + (1-A_+/N_+)(1-\phi)$, and an approximation of $V(\tilde{Y}|\psi)/V(\hat{Y})$ can be given by

$$\alpha = \frac{\phi + \left(1-\frac{A\xi}{N\theta}\right)(1-\phi)}{\left(1-\frac{A}{n}\right)} \tag{7}$$

The conditions that are generally favorable to the use of the threshold sample include a high catch rate ξ, a low prevalence θ, and a small variance contribution ϕ from the non-outliers. The effect of the threshold sample size A is not immediately clear. Given the parameters (ξ, θ, ϕ), the numerator and denominator of γ change in the same direction as A changes. However, since the "penalty" $(1-A/n)^{-1}$ increases with A, one would be more concerned in practice when A gets larger, say, from 1 year to the next. Notice that we have

$$\frac{\partial\alpha}{\partial A} = \frac{\frac{\theta-(1-\phi)\xi f}{n\theta}}{\left(1-\frac{A}{n}\right)^2}$$

It follows that α is a monotone function of A provided $A < n$, and the design efficiency would improve with increasing A if $\partial\alpha/\partial A < 0$, which amounts to

$$\theta < (1-\phi)\xi f \tag{8}$$

Moreover, since $1-\phi \le 1$ as well as $\xi \le 1$, a necessary condition for (8) is $\theta < f$. Notice that this is the same as observed in the special case earlier, where $\xi = 1$ and $\phi = 0$. Provided that

the condition (8) is true then the use of a threshold sample may yield protection against outliers and improve the design efficiency, no matter how large (or small) it is compared to the overall sample size. The necessary condition $\theta < f$ can always be achieved by choosing a sufficiently large threshold value, which in return can be considered as a robust design choice.

2.3. Stratified SRS and Change Estimation

It is straight-forward to apply the threshold-sample design together with a stratified SRS design in the NRDS, that is, among the rest of the population apart from the largest self-representing units. Here, the domains are fixed in advance, and coincide with the design strata. Different threshold values can be specified in different strata. The effect can be evaluated for each domain (i.e., stratum) on its own. Independence of the stratum total estimators makes it easy to evaluate the potential gains for the population total estimation.

Sometimes, however, the domains of interest cut across the design strata. Suppose it is possible to identify all the non-overlapping subsets of domains and strata. We may refer to each of these as a *domain stratum*. Then, each domain total is given as the sum over a number of domain strata, which are sampled independently of each other for a particular domain of concern. The results above can be used to evaluate the threshold-sample design effect within each domain stratum *conditional* on the realized domain-stratum sample sizes. The threshold value can still be set differently in each stratum. The potential gains for the population total estimator can be obtained by considering it as the total of independent stratum population total estimators.

The threshold-sample design effect can easily be applied to the estimation of change in two special situations. In the first case, independent stratified random sampling is administered in two time periods, after the self-representing units and the threshold-sample units have been removed from the frame. This is actually the case in the NRDS. Without the use of the threshold sample, the variance of the change estimator is the sum of the respective variance of each total estimator (Tam, 1984). A similar result holds under the threshold-sample design, conditional on the threshold sample configuration ψ for the two periods. (The derivation is omitted here to avoid the many extra notations needed, but is available from the authors on request.) It follows that the design effect for the change estimator is completely determined by those of the two separate level estimators, and the design efficiency for change estimation is improved provided it is improved for each level estimation. In particular, provided this is the case, the use of threshold sample can be justified despite the noise generated by the in- and out-flows to the threshold sample mentioned earlier.

In the second case, the same sample (i.e., a panel) is used for a number of successive time periods, as for example, in short-term business surveys. It is then natural that the threshold sample units, once chosen, are also held fixed throughout the same periods. The variance formulae (2) and (4) can be applied directly to any change variable, say, $z_i = y_{i,t=2} - y_{i,t=1}$ for period $t = 1$ and $t = 2$, and so on. However, notice that the design threshold value will not be a threshold for the target change variable in this case, that is, a unit above the threshold criterion does not necessarily have a target value z that is greater than a unit below the threshold criterion. The construction of the threshold sample therefore requires more careful consideration.

3. SMOOTH DOMAIN ESTIMATION

An outlier unit in the threshold sample is assigned a unit weight and is hence nonrepresentative. The ability to control the undue influence of the probability-sample outliers, that is, representative outliers, at the domain level is essentially a property of the estimation methodology. Below we describe a Winsorization-based smooth domain estimation approach, both with and without the use of a threshold-sample design. Like the threshold-sample design, the implementation relies heavily on a continuing survey environment to provide pre-fixed smoothing adjustments based

on historic data. Essential to the proposed approach is the assumption that outliers do not cause a problem at the overall population level, so that smoothing between the domain total estimates is able to provide sufficient outlier-resilience at the domain level.

3.1. Under Within-Domain SRS

Let $h = 1, ..., H$ be the domains of interest. Let y_{hi} be the value of interest associated with unit i in domain h. Let R_h be a fixed threshold value for the hth domain. Let

$$z_{hi} = \min(y_{hi}, R_h) \quad \text{and} \quad d_{hi} = \max(0, y_{hi} - R_h)$$

where z_{hi} is the Winsorized value and d_{hi} is the *excess* value, such that $y_{hi} \equiv z_{hi} + d_{hi}$. Let $\bar{Y}_h = \sum_{i \in U_h} y_{hi}/N_h$, where U_h is domain population and N_h is the domain population size. Let $\bar{Z}_h = \sum_{i \in U_h} z_{hi}/N_h$ and $\bar{D}_h = \sum_{i \in U_h} d_{hi}/N_h$. Let $N = \sum_h N_h$ and $W_h = N_h/N$. Let $\bar{Y} = \sum_h \sum_{U_h} y_{hi}/N = \sum_h W_h \bar{Y}_h$, and $\bar{Z} = \sum_h W_h \bar{Z}_h$, and $\bar{D} = \sum_h W_h \bar{D}_h$.

Assume within-domain SRS, that is, stratified SRS from the population where the strata are the domains of interest. An unbiased estimator of $\bar{Y}_h$ is $\hat{\bar{Y}}_h = \bar{y}_h = \sum_{i \in s_h} y_{hi}/n_h$, where s_h is the domain sample and n_h is the domain sample size. However, $\hat{\bar{Y}}_h$ can have a large variance due to the sample outliers. Rivest and Hidiroglou (2004) proposed the following domain estimator

$$\hat{\bar{Y}}_h^R = \bar{z}_h + \lambda_h \hat{\bar{D}} \tag{9}$$

where $\hat{\bar{D}} = \sum_h W_h \bar{d}_h$, and $\bar{z}_h$ and $\bar{d}_h$ are the respective domain sample means, and λ_h is a pre-fixed smoothing adjustment such that $\sum_h W_h \lambda_h = 1$. It follows that $\sum_h W_h \hat{\bar{Y}}_h^R = \sum_h W_h \hat{\bar{Y}}_h$, that is, the sum of $N_h \hat{\bar{Y}}_h^R$ over the domains coincide with that of $N_h \hat{\bar{Y}}_h$. Moreover, $\hat{\bar{Y}}_h^R$ tends to $\hat{\bar{Y}}_h$ as all $R_h \to \infty$, whereas it tends to the synthetic estimator $\lambda_h \sum_g W_g \bar{y}_g = \lambda_h \hat{\bar{Y}}_h$ as all $R_h \to 0$. The idea is to achieve a sensible trade-off between the potential bias against the reduced variance through the choice of R_h and λ_h, so that $\hat{\bar{Y}}_h^R$ may improve on $\hat{\bar{Y}}_h$ in terms of the MSE. Notice that, for fixed λ_h in (9), each $\hat{\bar{Y}}_h^R$ depends on a linear combination of all the outlier excesses. Extreme representative outliers may in fact have unbounded influence on all the domain estimates. We shall therefore refer to the approach as smooth domain estimation, where it is essential that $\hat{\bar{D}}$ is considered acceptable despite the presence of representative outliers.

Rivest and Hidiroglou (2004) set $\lambda_h = r_{hU} / \sum_g W_g r_{gU}$, where r_{hU} is the interquartile range of y_{hi} for $i \in U_h$ based on historic data. In the NRDS, however, the overall proportion of units with positive R&D-value is only around 20% in the population excluding the self-representing units, such that the interquartile range is zero or trivially small in many domains. More generally, we notice that, under the stratified SRS design, the bias of $\hat{\bar{Y}}_h^R$ is given by $E(\hat{\bar{Y}}_h^R) - \bar{Y}_h = \lambda_h \bar{D} - \bar{D}_h$. We therefore propose to aim directly at

$$\lambda_h = \frac{\bar{D}_h}{\bar{D}} \tag{10}$$

In this way the choice of λ_h becomes connected with that of R_h. Rivest and Hidiroglou (2004) set $R_h = \max\{\bar{Y}_h + \beta n_h/(n W_h), 0\}$, for a global choice of β with respect to a MSE-based criterion. But this requires absolute assessment of the current $\bar{Y}_h$, which one may be unwilling to make. In practice, the choice of R_h is likely to be based on historic data, probably involving some form of averaging over several periods. The emphasis will be on smoothness over time. We shall explore the choice of the threshold values in Section 4.

Sometimes, however, a domain may cut across the design strata. Suppose it is possible to identify all the domain strata (as defined earlier) within each domain. One would then replace $\bar{z}_h$ in (9) by a corresponding weighted sum of the Winsorized domain-stratum sample means involved. Both the threshold value R_h and the smoothing adjustment λ_h can still be set directly for each domain of interest, whereas the overall excess mean estimator $\hat{\bar{D}}$ can be obtained as the weighted sum of all the domain-stratum sample excess means. Of course, it is also possible to set the threshold value and the smoothing adjustments for each design stratum, and then build up the domain estimator through the domain strata. We shall not go into the details here.

A drawback of $\hat{\bar{Y}}_h^R$ is that it is generally not equal to $\bar{Y}_h$ even when $f_h = n_h/N_h = 1$. We shall instead adopt a prediction form of the estimator (9) given by

$$\hat{\bar{Y}}_h^P = f_h \bar{y}_h + (1-f_h)\hat{\bar{Y}}_h^{RP} \tag{11}$$

where smooth domain estimation, that is, $\hat{\bar{Y}}_h^{RP}$, is only applied to the population outside of the sample. Let $f = n/N$ and $\hat{\bar{D}} = \sum_h W_h \bar{d}_h$ as before. Let $\bar{d} = \sum_h w_h \bar{d}_h$ and $w_h = n_h/n$. We have

$$\begin{aligned}\hat{\bar{Y}}_h^{RP} &= \bar{z}_h + \lambda_h^P \hat{\bar{D}}_{(s)} \\ \hat{\bar{D}}_{(s)} &= (N\hat{\bar{D}} - d)/N = \hat{\bar{D}} - f\bar{d} \\ \lambda_h^P &= \lambda_h / (\sum_g W_g(1-f_g)\lambda_g)\end{aligned}$$

Notice that we have $\sum_h W_h f_h \bar{d}_h = f \sum_h w_h \bar{d}_h = f\bar{d}$ and $\sum_h W_h(1-f_h)\lambda_h^P = 1$, such that

$$\sum_h W_h \hat{\bar{Y}}_h^P = \sum_h W_h\{\bar{y}_h - (1-f_h)\bar{d}_h + (1-f_h)\lambda_h^P \hat{\bar{D}}_{(s)}\} = \hat{\bar{Y}} - \hat{\bar{D}} + f\bar{d} + \hat{\bar{D}}_{(s)} = \hat{\bar{Y}}$$

that is, the sum of $N_h \hat{\bar{Y}}_h^P$ over the domains coincide with that of $N_h \hat{\bar{Y}}_h$, just as $\hat{\bar{Y}}_h^R$.

We notice that the domain estimator (11) is similar to the Fuller's estimator (Fuller 1991, Equation (7.3)) for skewed populations, which is given as a weighted sum of the observed sample mean and a robust estimator of the superpopulation mean, where the weights are determined by the sampling fraction. The combined estimator is shown to be a minimum MSE estimator of the finite population mean under a model-based framework, provided a minimum MSE estimator of the superpopulation mean is being used. The difference is that $\hat{\bar{Y}}_h^{RP}$ in (11) is a Winsorization-based estimator directed at the population outside of the observed sample, the choice of which is motivated by the design-unbiasedness at the overall population level.

The bias of $\hat{\bar{Y}}_h^P$ is given by $E(\hat{\bar{Y}}_h^P) - \bar{Y}_h = (1-f_h)(\lambda_h^P \bar{D}_{(s)} - \bar{D}_h)$, where

$$\bar{D}_{(s)} = \bar{D} - f\sum_h w_h \bar{D}_h = \bar{D} - \sum_h W_h f_h \bar{D}_h = \sum_h W_h(1-f_h)\bar{D}_h$$

Thus, $\hat{\bar{Y}}_h^P$ is unbiased provided

$$\lambda_h^P = \frac{\bar{D}_h}{\bar{D}_{(s)}}$$

which is true if $\lambda_h = \bar{D}_h/\bar{D}$, so that aiming λ_h at $\bar{D}_h/\bar{D}$ remains plausible for $\hat{\bar{Y}}_h^P$. To obtain the variance of $\hat{\bar{Y}}_h^P$, we write $\hat{D}_{(s)} = \sum_h W'_h \bar{d}_h$ for $W'_h = W_h - w_h f$, and

$$\hat{\bar{Y}}_h^P = \bar{y}_h - (1-f_h)(1-\lambda_h^P W'_h)\bar{d}_h + (1-f_h)\lambda_h^P \sum_{g\neq h} W'_g \bar{d}_g$$

It follows that the MSE of $\hat{\bar{Y}}_h^P$ is given by

$$\mathrm{MSE}(\hat{\bar{Y}}_h^P) = (1-f_h)^2(\lambda_h^P \bar{D}_{(s)} - \bar{D}_h)^2 + (1-f_h)\sigma^2_{x,h}/n_h + (1-f_h)^2 \sum_{g\neq h} (\lambda_h^P W'_g)^2 (1-f_g)\sigma^2_{d,g}/n_g \tag{12}$$

where $x_{hi} = y_{hi} - (1-f_h)(1-\lambda_h^P W'_h)d_{hi}$ and $\sigma^2_{x,h}$ is the domain finite-population variance of x_{hi}. Notice that the MSE expression does not allow for post-sample tuning of λ_h.

3.2. Within-Domain SRS Given Threshold Sample

Given the use of a threshold sample, it is natural to condition on the observed threshold sample total, and to apply domain prediction (11) to the population outside of the threshold sample. The estimator of the domain population mean is then given by

$$\hat{\bar{Y}}_h^{AP} = f_{hA}\bar{y}_{hA} + (1-f_{hA})\hat{\bar{Y}}_{hB}^P \tag{13}$$

where $\bar{y}_{hA}$ is the mean of the within-domain threshold sample s_{hA} and $f_{hA} = n_{hA}/N_h$. The estimator $\hat{\bar{Y}}_{hB}^P$ is obtained by applying the estimator (11) to the domain population outside of s_{hA}

$$\hat{\bar{Y}}_{hB}^P = f_{hB}\bar{y}_{hB} + (1-f_{hB})(\bar{z}_{hB} + \lambda_{hB}^P \hat{\bar{D}}_{(s)B}) \tag{14}$$

where $\bar{y}_{hB}$ is the mean of the within-domain probability sample $s_{h,B}$, and $f_{hB} = n_{hB}/N_{hB}$ and $N_{hB} = N_h - n_{hA}$ is the corresponding population size, and

$$\lambda_{hB}^P = \lambda_h / (\sum_h W_{hB}(1-f_{hB})\lambda_h)$$
$$\hat{\bar{D}}_{(s)B} = \hat{\bar{D}}_B - f_B \bar{d}_B = \sum_h W'_{hB}\bar{d}_{hB}$$

where $W_{hB} = N_{hB}/N_B$ for $N_B = N - \sum_h n_{hA}$, and $\hat{\bar{D}}_B = \sum_h W_{hB}\bar{d}_{hB}$ for domain probability-sample mean $\bar{d}_{hB}$, and $W'_{hB} = W_{hB} - f_B w_{hB}$ for $f_B = n_B/N_B = \sum_h n_{hB}/N_B$ and $w_{hB} = n_{hB}/n_B$. Notice that, since $\hat{\bar{Y}}_{hB}^P$ is derived conditional on the actual threshold sample, the threshold value for estimation can be set independently from the threshold value for the sampling design. For simplicity, however, we have chosen not to make an explicit distinction in the notation. Notice also that the similarity between (11) and Fuller's estimator is retained in (13).

Given $s_A = \cup_{h=1}^H s_{hA}$, we have $E(\bar{y}_{hB}|s_A) = \bar{Y}_{hB}$ and $E(\bar{d}_{hB}|s_A) = \bar{D}_{hB}$, such that

$$E(\hat{\bar{Y}}_h^{AP}|s_A) = \bar{Y}_h + (1-f_{hA})(1-f_{hB})(\lambda_{hB}^P \bar{D}_{(s)B} - \bar{D}_{hB})$$

where $\bar{D}_{(s)B} = \sum_h W'_{hB}\bar{D}_{hB}$. Let $\psi = \{(A_h, \xi_h); h = 1, ..., H\}$ where $A_h = n_{hA}$. Since the total excess value is 0 among the units below the threshold by definition, we have $D_{hB} = D_{hB+} = D_{h+} - d_{hA+} = D_h - d_{hA}$, where d_{hA} is the total excess value in s_A, and d_{hA+} is that of the units in s_A that are above the threshold value, and similarly for D_{hB+} and D_{h+}. Given (A_h, ξ_h), we have $E(\bar{D}_{hB}|\psi) = (1-f_{hA+})D_h/N_{hB}$, where $f_{hA+} = A_{h+}/N_{h+}$, and $A_{h+} = A_h\xi_h$,

and N_{h+} is the within-domain number of units above the threshold, such that $E(\bar{D}_{(s)B}|\psi) = \sum_h W'_{hB}(1-f_{hA+})D_h/N_{hB}$. The bias of $\hat{\bar{Y}}_h^{AP}$ conditional on ψ is then given as

$$E(\hat{\bar{Y}}_h^{AP}|\psi)-\bar{Y}_h = (1-f_{hA})(1-f_{hB})\{\lambda_{hB}^P E(\bar{D}_{(s)B}|\psi)-E(\bar{D}_{hB}|\psi)\} \tag{15}$$

The variance of $\hat{\bar{Y}}_h^{AP}$ conditional on ψ can be given as

$$V(\hat{\bar{Y}}_h^{AP}|\psi) = E\{V(\hat{\bar{Y}}_h^{AP}|s_A)|\psi\} + V\{E(\hat{\bar{Y}}_h^{AP}|s_A)|\psi\} \tag{16}$$

For the first term on the right-hand side of (16), we write

$$\begin{aligned}\hat{\bar{Y}}_h^{AP} &= f_{hA}\bar{y}_{hA} + (1-f_{hA})\{\bar{y}_{hB}-(1-f_{hB})(1-\lambda_{hB}^P W'_{hB})\bar{d}_{hB} + \sum_{g\neq h}(1-f_{hB})\lambda_{hB}^P W'_{gB}\bar{d}_{gB}\} \\ &= f_{hA}\bar{y}_{hA} + (1-f_{hA})\bar{x}_{hB} + \sum_{g\neq h}(1-f_{hA})(1-f_{hB})\lambda_{hB}^P W'_{gB}\bar{d}_{gB}\end{aligned}$$

where $\bar{x}_{hB}$ is the probability-sample average of $x_{hi} = y_{hi}-(1-f_{hB})(1-\lambda_{hB}^P W'_{hB})d_{hi}$. Conditional on s_A, $\bar{y}_{hA}$ is a constant, whereas $\bar{x}_{hB}$ is independent of $\bar{d}_{gB}$ for $h \neq g$. We have

$$V(\bar{x}_{hB}|s_A) = (1-f_{hB})\sigma_{x,hB}^2/n_{nB} \quad \text{and} \quad V(\bar{d}_{gB}|s_A) = (1-f_{gB})\sigma_{d,gB}^2/n_{gB}$$

where $\sigma_{x,hB}^2$ and $\sigma_{d,gB}^2$ are, respectively, the within-domain finite-population variances of x_{hi} and d_{gi} outside of s_A. By derivations similar to those that result in (4), we obtain

$$\begin{aligned}E\{V(\hat{\bar{Y}}_h^{AP}|s_A)|\psi)\} &= (1-f_{hA})^2 c_{hB}E(\Delta_{x,hB}|\psi) \\ &\quad + \sum_{g\neq h}((1-f_{hA})(1-f_{hB})\lambda_{hB}^P W'_{gB})^2 c_{gB}E(\Delta_{d,hB}|\psi)\end{aligned} \tag{17}$$

where $c_{hB} = (1-f_{hB})/(n_{hB}(N_{hB}-1)) = (N_h-n_h)/(n_{hB}N_{hB}(N_{hB}-1))$ for $h = 1, ..., H$, and $E(\Delta_{x,hB}|\psi)$ and $E(\Delta_{d,hB}|\psi)$ can be obtained in a similiar way to $E(\Delta_B|\psi)$ in (5), but are now domain indexed and with respect to x_{hi} and d_{hi}. For instance, we have

$$\begin{aligned}E(\Delta_{x,hB}|\psi) &= (N_{h+}-A_{h+}-a_{h+})\sigma_{x,h+}^2 + (N_{h-}-A_{h-}-a_{h-})\sigma_{x,h-}^2 \\ &\quad + \frac{(N_{h+}-A_{h+})(N_{h-}-A_{h-})}{N_h-A_h}(\bar{X}_{h+}-\bar{X}_{h-})^2\end{aligned}$$

for $a_{h+} = 1-(A_{h+}/N_{h+})(N_{h-}-A_{h-})/(N_h-A_h)$ and $a_{h-} = 1-(A_{h-}/N_{h-})(N_{h+}-A_{h+})/(N_h-A_h)$, and $\bar{X}_{h+}$ and $\sigma_{x,h+}^2$ are the finite-population mean and variance of x_{hi} in U_{h+}, and $\bar{X}_{h-}$ and $\sigma_{x,h-}^2$ those in U_{h-}. For the second term on the right-hand side of (16), we write

$$E(\hat{\bar{Y}}_h^{AP}|s_A) = \bar{Y}_h + (1-f_{hA})(1-f_{hB})(\lambda_{hB}^P W'_{hB}-1)\bar{D}_{hB} + \sum_{g\neq h}(1-f_{hA})(1-f_{hB})\lambda_{hB}^P W'_{gB}\bar{D}_{gB}$$

The variance of $\bar{D}_{hB}$ conditional on ψ is independent of $\bar{D}_{gB}$ for $g \neq h$, and is given as

$$V(\bar{D}_{hB}|\psi) = N_{hB}^{-2}V(D_h-d_{hA}|\psi) = N_{hB}^{-2}V(d_{hA}|\psi) = N_{hB}^{-2}V(d_{hA+}|\psi) = N_{hB}^{-2}\kappa_h\sigma_{d,h+}^2$$

where $\kappa_h = A_{h+}(N_{h+}-A_{h+})/N_{h+}$ for $A_{h+} = A_h\xi_h$, and $\sigma^2_{d,h+}$ is the within-domain finite-population variance of d_{hi} in U_{h+}. We obtain, then,

$$V\{E(\hat{\bar{Y}}_h^{AP}|s_A)|\psi\} = ((1-f_{hA})(1-f_{hB})(\lambda_{hB}^P W'_{hB}-1))^2 N_{hB}^{-2}\kappa_h\sigma^2_{d,h+} + \sum_{g\neq h}((1-f_{hA})(1-f_{hB})\lambda_{hB}^P W'_{gB})^2 N_{gB}^{-2}\kappa_g\sigma^2_{d,g+} \quad (18)$$

The MSE of $\hat{\bar{Y}}_h^{AP}$ can now be obtained by (15) and (16), where (16) is obtained from (17) and (18).

4. AN EVALUATION

In this section we investigate the performance of the threshold-sample design and the smooth domain estimation procedure, using a synthetic population constructed based on the NRDS data. It also serves as an illustration of the type of analysis that one would carry out in order to apply the design and estimation approach. The main findings can be summarized as follows:

- The condition (8) can be expected to provide correct indication for the performance of the threshold-sample design. Targeting a sufficiently low global value of the outlier prevalence θ leads naturally to reasonable domain-specific threshold values. Too low a choice of the value of the prevalence, however, may lead to empty threshold sample in many domains and, thereby, loss of potential gains of the threshold-sample design there.
- Similarly, targeting a global value of θ can be used to generate reasonable domain-specific threshold values for smooth domain estimation. The prevalence can be set higher than that for the design of the threshold samples, provided plausible values of λ_h can be obtained based on historic data. In cases where the uncertainty about the values of λ_h is greater, however, a lower value of θ can be used to reduce the sensitivity of the domain estimators.
- The efficiency gains from the threshold-sample design may be small compared to those of the estimation methodology, provided the latter can be tuned appropriately in practice. However, the threshold-sample design can be made robust in a simple manner, such that the potential gains are easily achieved, including that for the population total estimator.

4.1. Data and Set-Up

As mentioned earlier, the NRDS contains a self-representing sub-sample of the largest enterprises, a threshold sub-sample and a probability sub-sample from the rest of population. Table 1 provides a summary of the NRDS sample in 2003. The threshold sub-sample contains 187 units, where 158 of them have an R&D-value above the threshold value. This yields a catch rate of 84.5%, much higher than the other two sub-samples. Moreover, the average R&D-value in the threshold sub-sample (i.e., 5.310×10^6) is rather close to the average among the self-representing units (i.e., 5.576×10^6), and is much higher than that among the randomly selected units (i.e., 0.432×10^6). Both of these seem to support the use of the threshold sample in the NRDS.

For a detailed evaluation we constructed a synthetic population for two successive years, denoted by $t = 1$ and 2, based on the NRDS data in 2003 and 2004 excluding the self-representing units. The population consists of 9,734 units, that is, $N = 9{,}734$, divided into 55 industrial domains, that is, $H = 55$. The domain sampling fractions are set similiar to the NRDS, giving a fixed total sample size of 2,379, that is, $n = 2{,}379$, on each occasion including the threshold sample. Figure 2 contains boxplots of $N_h, f_h = n_h/N_h, \theta_h = N_{h+}/N_h$ for $t = 1$ and $t = 2$, where the prevalence θ_h is calculated with respect to a global *reference* threshold value $R_0 = 1 \times 10^6$ here.

TABLE 1: Self-representing, threshold and probability sub-samples of NRDS 2003.

	Number of Units			R&D-Value ($\times 10^6$)	
Sub-Sample	Total	Above Threshold	Catch Rate (%)	Total	Average
Self-representing	1,737	558	32.1	9,685	5.576
Threshold	187	158	84.5	993	5.310
Probability	2,510	228	9.1	1,085	0.432

For the subsequent evaluation in this paper we would like to vary the threshold value as well as the catch rate. The domain threshold sample sizes at $t = 1$ are set in two steps as follows. First, a *reference* value $A_{h+,t=1}(R_0)$ is chosen for $t = 1$ that corresponds to the reference threshold value $R_h = R_0$, that is, the number of units in the domain threshold sample that are above the reference threshold value R_0 at $t = 1$. For a given *theoretical* catch rate ξ, the corresponding domain threshold sample size at $t = 1$ is then given by $A_{h,t=1}(R_0, \xi) = A_{h+,t=1}(R_0)/\xi$. Next, for any other choice of R_h, we calculate the corresponding $A_{h+,t=1}(R_h)$ by

$$A_{h+,t=1}(R_h)/N_{h+,t=1}(R_h) = A_{h+,t=1}(R_0)/N_{h+,t=1}(R_0)$$

where $N_{h+,t=1}(R_h)$ is the number of units above the threshold value R_h in the domain population. For a given theoretical catch rate ξ, we obtain $A_{h,t=1}(R_h, \xi) = A_{h+,t=1}(R_h)/\xi$ as before. The

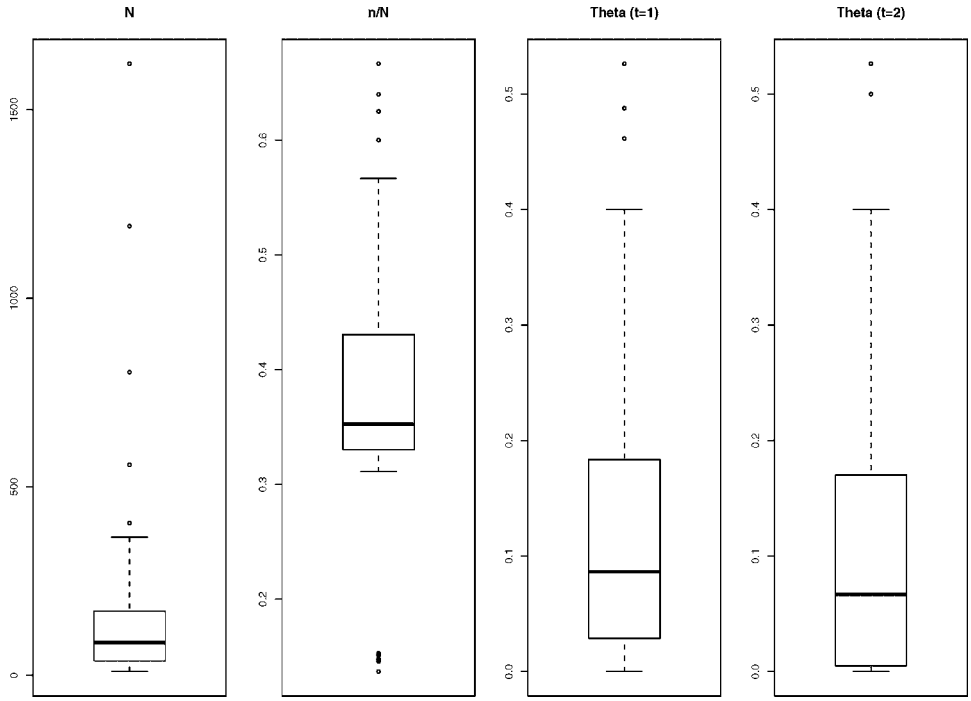

FIGURE 2: Boxplot of the synthetic population. *N*, Domain population size. *n*/*N*, Domain sampling fraction. Theta ($t = 1$), domain outlier prevalence at $t = 1$ corresponding to global threshold value $R_0 = 1 \times 10^6$. Theta ($t = 2$), domain outlier prevalence at $t = 2$ corresponding to global threshold value $R_0 = 1 \times 10^6$.

threshold sample size for $t=2$ is set at its expectation, which is given by

$$A_{h,t=2}(R_h,\xi) = A_{h+,t=1}(R_h) + (n_h - A_{h,t=1}(R_h,\xi))\frac{N_{h+,t=1}(R_h) - A_{h+,t=1}(R_h)}{N_h - A_{h,t=1}(R_h,\xi)}$$

Also the number of units above the threshold value at $t=2$ is set at its expectation, given by

$$A_{h+,t=2}(R_h,\xi) = \min\{N_{h+,t=2}(R_h),\, A_{h,t=2}(R_h,\xi)N_{h++}(R_h)/N_{h+,t=1}(R_h)\}$$

that is, subject to the constraint of $A_{h+} \leq N_{h+}$, where $N_{h++}(R_h)$ is the number of units above the threshold value on both occasions. Notice that, after rounding, the actual catch rates are given by $A_{h+,t}/A_{h,t}$. These are used in the calculations instead of the theoretical value ξ.

4.2. Design Effect

Table 2 contains selected results concerning the relative efficiency (RE) of the threshold-sample design, for alternative settings of the theoretical catch rate ξ and the domain threshold values R_h. The following observations seem worth noting.

Firstly, the condition (8) can be expected to provide correct guideline in a given domain h if $\mathrm{RE}_h < 1$ and $\theta_h < (1-\phi_h)\xi_h f_h$, or if $\mathrm{RE}_h \geq 1$ and $\theta_h \geq (1-\phi_h)\xi_h f_h$. Let $I_h = 1$ if this is the case and $I_h = 0$ if it is not. The results in Table 2 show that the condition works well. Take, for example, the results obtained at $R_h = 5\times 10^6$ and $\xi = 0.8$. The condition (8) is expected to provide correct guidance in all the domains since $\sum_h I_h = 55 = H$ for both $t=1$ and $t=2$. By comparison the condition fails in two domains for $t=1$ when $R_h = 1\times 10^6$, where $\sum_h I_h = 53 = H-2$, although this undoubtedly would still be very valuable in practice.

Secondly, there is the choice of the threshold value. A convenient approach is to set R_h at some global choice, that is, $R_h = R$. The results obtained at $R_h = 1\times 10^6$ and $R_h = 5\times 10^6$, where ξ is set at 0.8 in both cases, suggest that the choice of $R_h = 5\times 10^6$ is more efficient for the population total estimation, at both time points as well as for change. It is also more robust. Across the domains, $R_h = 1\times 10^6$ results in a wider range of RE_h, where, for example, the maximum value of RE_h for $t=1$ is 1.35 with $R_h = 1\times 10^6$ and 1 with $R_h = 5\times 10^6$. Indeed, given a high threshold value, a threshold sample is worse off than pure SRS only if it is apparently ineffective. For instance, there are three domains with $\mathrm{RE}_h > 1$ at $t=2$ when $R_h = 5\times 10^6$, and all of them have $A_{h+} = \xi_h = 0$ and $A_h = 2$. Such cases are easily detected in light of the condition (8).

While raising the threshold value improves outlier resilience, too high a threshold value will eventually result in an empty threshold sample, thus keeping one away from the possible gains of the threshold sample. For instance, there are 21 domains with $\mathrm{RE}_h < 1$ at $R_h = 1\times 10^6$ and $\xi = 0.8$, whereas there are only 12 domains with $\mathrm{RE}_h < 1$ when $R_h = 5\times 10^6$, because many domains have in fact no threshold sample at all in the latter case. As an alternative to a global choice of $R_h = R$, we consider the strategy of setting R_h such that the resulting domain prevalence θ_h (averaged over the 2 years here) is as close as possible to a global choice of θ, denoted by $R_h(\theta)$. Using these domain-specific threshold values $R_h(\theta)$, one seems to be able to make more out of the threshold-sample design, while maintaining its robustness. For instance, given the same effectiveness at $\xi = 0.8$, the overall efficiency is quite similar using $R_h(\theta)$ for $\theta = 0.05$ or $R_h = 5\times 10^6$. However, the threshold-sample design performs better for domain estimation using $R_h(\theta)$ for $\theta = 0.05$: the range of RE_h is shifted towards 0, and the number of domains with $\mathrm{RE}_h < 1$ is increased. To be sure, the domains with $\mathrm{RE}_h > 1$ at $t=2$ when using $R_h(\theta)$ for $\theta = 0.05$ are all apparently ineffective, all of them having $A_{h+} = \xi_h = 0$ and $1 \leq A_h \leq 3$. Thus, the use of $R_h(\theta)$ for a sufficiently low value of θ and keeping control of the condition (8) can provide a practical approach to the threshold-sample design. The strategy is both easier to understand and to implement than completely free individual choice of each R_h.

TABLE 2: Effect of threshold sample for alternative settings of theoretical catch rate ξ and threshold value.

	Level at $t=1$	Level at $t=2$	Change
Threshold value $R_h = 1 \times 10^6$; theoretical catch rate $\xi = 0.5$			
RE	1.04	0.95	1.01
(25%, 50%, 75%) Quantiles of $\theta_{h,t}$	(0.03, 0.09, 0.18)	(0.01, 0.07, 0.17)	—
Summary condition $\sum_h I_h$	53	55	—
(Min, median, max) of RE_h	(0.80, 1, 1.48)	(0.02, 0.90, 1.39)	(0.68, 0.96, 1.45)
# ($\text{RE}_h < 1$, $\text{RE}_h = 1$, $\text{RE}_h > 1$)	(11, 31, 13)	(34, 9, 12)	(34, 9, 12)
(Median, max) of ξ_h given $\text{RE}_h > 1$	(0.5, 0.5)	(0.33, 1)	—
Threshold value $R_h = 1 \times 10^6$; theoretical catch rate $\xi = 0.8$			
RE	1.00	0.95	0.98
(25%, 50%, 75%) Quantiles of $\theta_{h,t}$	(0.03, 0.09, 0.18)	(0.01, 0.07, 0.17)	—
Summary condition $\sum_h I_h$	53	55	—
(Min, median, max) of RE_h	(0.02, 1, 1.35)	(0.02, 0.89, 1.57)	(0.02, 0.93, 1.44)
# ($\text{RE}_h < 1$, $\text{RE}_h = 1$, $\text{RE}_h > 1$)	(21, 26, 8)	(34, 7, 14)	(38, 7, 10)
(Median, max) of ξ_h given $\text{RE}_h > 1$	(0.71, 0.8)	(0, 1)	—
Threshold value $R_h = 5 \times 10^6$; theoretical catch rate $\xi = 0.8$			
RE	0.95	0.78	0.88
(25%, 50%, 75%) Quantiles of $\theta_{h,t}$	(0, 0.02, 0.04)	(0, 0.01, 0.04)	—
Summary condition $\sum_h I_{h,t}$	55	55	—
(Min, median, max) of RE_h	(0.54, 1, 1)	(0.01, 1, 1.07)	(0.49, 1, 1.01)
# ($\text{RE}_h < 1$, $\text{RE}_h = 1$, $\text{RE}_h > 1$)	(12, 43, 0)	(26, 26, 3)	(26, 26, 3)
(Median, max) of ξ_h given $\text{RE}_h > 1$	—	(0, 0)	—
Threshold value R_h (θ) for $\theta = 0.05$; theoretical catch rate $\xi = 0.8$			
RE	0.93	0.79	0.87
(25%, 50%, 75%) Quantiles of $\theta_{h,t}$	(0.03, 0.05, 0.07)	(0.00, 0.03, 0.05)	—
Summary condition $\sum_h I_{h,t}$	55	55	—
(Min, median, max) of RE_h	(0.02, 1, 1)	(0.01, 0.80, 1.08)	(0.02, 0.88, 1.02)
# ($\text{RE}_h < 1$, $\text{RE}_h = 1$, $\text{RE}_h > 1$)	(23, 32, 0)	(36, 9, 10)	(40, 9, 6)
(Median, max) of ξ_h given $\text{RE}_h > 1$	—	(0, 0)	—

Threshold: global value $R_h = R$ or domain specifc value $R_h(\theta)$ for global choice of x.

Finally, for the efficiency of the threshold-sample design we would like the catch rate $\xi_h = A_{h+}/A_h$ to be as high as possible. However, given R_h, ξ_h is determined solely by the dynamics in the population, and is a characteristic that is beyond the control of the survey statistician. The experience from the NRDS showed that the overall catch rate could be maintained at around $\xi = 0.8$ over time. However, it is important to explore the possible damages if the catch rate drops. How the domain threshold sample design varies with different values of ξ, in order to compute the results in Table 2, has been described previously in Section 4.1. Compare the results obtained at $\xi = 0.5$ and $\xi = 0.8$, both at $R_h = 1 \times 10^6$. There is a noticeable loss of efficiency at $t = 1$ from $\xi = 0.8$ to $\xi = 0.5$, both in terms of the overall RE and the domain-wise RE_h. Of course, a reduction of the catch rate from 0.8 to 0.5 is rather dramatic, and one should have been alerted in advance provided the catch rates are closely monitored over time.

4.3. Estimation Effect

To examine the effect of the estimation methodology on its own, we assume within-domain SRS, and calculate the RE as the MSE ratio between $\hat{\bar{Y}}_h^P$ given by (11) and the direct estimator $\hat{\bar{Y}}_h = \bar{y}_h$. There will be no overall effect here now that the two agree with each other at the population total level. As explained earlier, setting λ_h according to (10) leads to unbiased domain estimation. We shall refer to the corresponding value of λ_h as the *unbiased* choice. In reality, deviation from the unbiased choice is necessarily the case. Sensitivity of the smooth estimator (11) depends on how well it performs as λ_h deviates from the unbiased choice.

We carry out the following simulations. For each setting of $\{R_h; h = 1, ..., H\}$, we calculate the corresponding unbiased choices, denoted by $\lambda_h^{(0)}$ for $h = 1, ..., H$. At each simulation b for $b = 1, ..., B$, we generate independent $\lambda_h^{(b)} \sim N(\lambda_h^{(0)}, \tau\lambda_h^{(0)})$ where τ is the relative standard deviation (RSD) of $\lambda_h^{(b)}$ over the simulations, that can be controlled by the simulation design. We then calculate the RE of the corresponding domain estimators, denoted by $\mathrm{RE}_h^{(b)}$, and evaluate the minimum, median, and maximum value of $\mathrm{RE}_h^{(b)}$ among $h = 1, ..., H$, and the number of domains where $RE_h^{(b)} < 1$, $=1$ or >1, respectively. Table 3 gives the minimum, median, and maximum values of these summary statistics over the simulations for $t = 1$. The results for $t = 2$ are similar and those for the change can be derived from the two time points; the details are omitted here. For instance, with $R_h = 1 \times 10^6$ and $\tau = \mathrm{RSD}(\lambda_h) = 20\%$ and $B = 500$, the minimum value of $\min_h \mathrm{RE}_h^{(b)}$ over some 500 simulations is seen to be 0.079, whereas the maximum value of $\min_h \mathrm{RE}_h^{(b)}$ over these 500 simulations is seen to be 0.148. That is,

$$\min_{b=1,...,B}\{\min_{h=1,...,H} \mathrm{RE}_h^{(b)}\} = 0.079 \text{ and } \max_{b=1,...,B}\{\min_{h=1,...,H} \mathrm{RE}_h^{(b)}\} = 0.148$$

While these two values represent, respectively, the best and worst case of $\min_h \mathrm{RE}_h$ under the given setting of R_h and τ, the median value roughly corresponds to the expected case of $\min_h \mathrm{RE}_h$ where λ_h is set at the unbiased choice. The other summary statistics can be viewed similiarly.

The following observations are worth noting from Table 3. First of all, the use of smooth domain estimation apparently yields much greater efficiency gains than the use of threshold-sample design. Compare, for example, the results under $R_h = 1 \times 10^6$ in Table 3 with those under $R_h = 1 \times 10^6$ in Table 2 for $t = 1$ and at $\xi = 0.8$. The median RE_h by smooth estimation lies between 0.3 and 0.4 while that by threshold-sample design is 1; the number of domains with $\mathrm{RE}_h < 1$ varies between 45 and 48 by estimation, whereas that by design is only 21. Similar improvements can be observed under the other settings. Next, the results under $R_h = 1 \times 10^6$ and $R_h = 5 \times 10^6$ suggest apparently that a lower threshold value is more efficient than a higher one, because it leads to smaller domain variances of the Winsorized values $\bar{z}_{hi}$. However, notice that this holds only if plausible values of λ_h are obtainable, in which case variance reduction can be achieved without much increase in the bias and, hence, gains in the MSE. Notice also that domain estimation appears less sensitive at a higher global threshold value. For instance, in the worst case the maximum RE_h can be as high as 3.362 under $R_h = 1 \times 10^6$, compared to 1.046 under $R_h = 5 \times 10^6$. Similar comparisons can be made between $R_h(\theta)$ for $\theta = 0.2$ and $\theta = 0.05$. Thus, the threshold values should be set higher as the uncertainty surrounding the unbiased choices of λ_h increases. Finally, using domain-specific R_h seems preferable to setting a global threshold value. For instance, the median of RE_h under $R_h(\theta = 0.2)$ is close to that under $R_h = 1 \times 10^6$, while the worst case of $\max_h \mathrm{RE}_h$ is 1.433 compared to 3.362 under $R_h = 1 \times 10^6$. Moreover, when the threshold values $R_h(\theta)$ are raised from $\theta = 0.2$ to $\theta = 0.05$, the sensitivity is reduced, while the efficiency losses are not as big as when $R_h = 1 \times 10^6$ is raised to $R_h = 5 \times 10^6$.

TABLE 3: Robust domain estimation at $t = 1$ over 500 simulations.

	Summary of RE_h			Number of Domains		
Summary Over Simulations	Minimum	Median	Maximum	$RE_h < 1$	$RE_h = 1$	$RE_h > 1$
Threshold value $R_h = 1 \times 10^6$; $RSD(\lambda_h) = 20\%$						
Minimum	0.079	0.330	1	45	7	0
Median	0.087	0.344	1	48	7	0
Maximum	0.148	0.403	3.362	48	7	3
Threshold value $R_h = 5 \times 10^6$; $RSD(\lambda_h) = 20\%$						
Minimum	0.234	0.666	1	37	17	0
Median	0.235	0.670	1	38	17	0
Maximum	0.253	0.716	1.046	38	17	1
Threshold value $R_h(\theta)$ for $\theta = 0.05$; $RSD(\lambda_h) = 20\%$						
Minimum	0.109	0.436	1	47	8	0
Median	0.119	0.447	1	47	8	0
Maximum	0.172	0.503	1	47	8	0
Threshold value $R_h(\theta)$ for $\theta = 0.2$; $RSD(\lambda_h) = 20\%$						
Minimum	0.086	0.323	1	46	7	0
Median	0.094	0.356	1	48	7	0
Maximum	0.141	0.413	1.433	48	7	2

Threshold value: global $R_h = R$ or domain specific $R_h(\theta)$. $RSD(\lambda_h)$, relative standard deviation around the unbiased choice.

In practice, the values of λ_h will most likely be adjusted based on historic data and monitored closely over time. Small area estimation techniques targeting unbiased choices of λ_h may also be employed in this respect, using historic or current data. For instance, a Fay–Herriot type area-level model (e.g., Rao 2003, Chapter 5) may be fitted to obtain estimates of λ_h. A particular advantage of doing this is that a measure of uncertainty of the estimated λ_h can be obtained explicitly. Such a measure would provide a valuable indication of the potential deviation from the unbiased choice. As noted above, when there is greater uncertainty about the unbiased values of λ_h, the threshold values should be set relatively higher.

4.4. Combined Design and Estimation Approch

Combining the threshold-sample design and smooth domain estimation does not pose any extra challenges in theory. According to (13), domain estimation is applied to the rest of the population outside of the threshold sample. Conditional on the threshold sample, the design effect and the estimation effect will enhance each other, and are in this sense additive. Table 4 illustrates the RE of the combined approach against direct estimation $\hat{\bar{Y}}_h = \bar{y}_h$ based on within-domain SRS alone, at $t = 1$ for alternative settings of the threshold values and for theoretical catch rate $\xi = 0.8$ and λ_h given by (10). The combined gains of efficiency are as expected when compared to Tables 2 and 3. The situation is similar for level estimation at $t = 2$ and change estimation. Notice that, since smooth domain estimation has no effect for the population total estimator, the performance of the combined approach at that level is entirely determined by the corresponding design effect.

There are, nevertheless, some practical issues that need to be decided on. First of all, there is the question of whether or not to use the combined approach. On the one hand, the design effects

TABLE 4: Combined use of threshold-sample design and smooth domain estimation at $t = 1$.

	Summary of RE_h			Number of Domains		
($\xi = 0.8$) Threshold Value	Minimum	Median	Maximum	$RE_h < 1$	$RE_h = 1$	$RE_h > 1$
$R_h = 1 \times 10^6$	0.080	0.287	1	50	5	0
$R_h = 5 \times 10^6$	0.221	0.640	1	42	10	3
$R_h(\theta)$ for $\theta = 0.05$	0.107	0.421	1	49	6	0
$R_h(\theta)$ for $\theta = 0.2$	0.086	0.317	1	50	5	0

may be small compared to the estimation effects. On the other hand, the threshold-sample design can be made robust in a simple manner, such that the potential gains are easily achieved. Next, there is the freedom to choose different threshold values at the design and estimation stages. On the one hand, extra gains of efficiency can be expected if the two sets of threshold values are chosen separately. On the other hand, to use a single set of threshold values is easier to manage and, as the results in Table 4 suggest, may be able to achieve much of the efficiency gains if these are chosen reasonably. Optimization according to a particular MSE-criterion is unlikely to be feasible, because these must depend on assumptions of the catch rates ξ_h and the choices of λ_h that are difficult to control tightly in reality. In the end, the answers to these questions will have to depend on the actual data under consideration and the experiences obtained over time.

5. SUMMARY

We have studied an outlier-robust threshold-sample design, which utilizes a threshold sample that is selected with probability one together with stratified simple random sampling from the rest of the population. Condition (8) can be expected to provide a useful condition for the design efficiency of the threshold sample. To use domain-specific threshold values that aim at a sufficiently low prevalence of outliers has been shown to be a practical design approach.

In addition, we have considered a smooth domain estimation approach for reducing the impact of representative outliers on disaggregated estimates. The domain estimators (13) are a prediction extension of the estimator proposed by Rivest and Hidiroglou (2004). The MSE is derived with respect to the sampling design. In practice, a plug-in approach would be used for the MSE estimation. We have not investigated its performance in this paper. Evaluations based on the NRDS data suggest that considerable gains of efficiency are achievable. Again, allowing for domain-specific threshold values, which are regulated through a target level of the outlier prevalence, provides a sensible approach for improving the efficiency of estimation.

ACKNOWLEDGEMENTS

We would like to thank the referees very much for insightful comments and helpful suggestions for reshaping the paper.

BIBLIOGRAPHY

R. L. Chambers (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.

R. L. Chambers & N. Tzavidis (2006). M-Quantile models for small area estimation. *Biometrika*, 93, 255–268.

W. Fuller (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137–158.

L. Kish (1965). "*Survey Sampling,*" Wiley, New York.

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

L.-P. Rivest & M. Hidiroglou (2004). Outlier Treatment for Disaggregated Estimates. In "*Proceedings of the Section on Survey Research Methods,*" American Statistical Association, pp. 4248–4256.

S. K. Sinha & J. N. K. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37, 381–399.

S. M. Tam (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288–289.

6. APPENDIX

Let $(N_+-A_+-1)\sigma^2_{B+} = \sum_{i\in U_B\cap U_+}(y_i-\bar{Y}_{B+})^2$, and $(N_--A_--1)\sigma^2_{B-} = \sum_{i\in U_B\cap U_-}(y_i-\bar{Y}_{B-})^2$. Let $\bar{Y}_{B+}$ be the mean of interest in $U_B\cap U_+$, and $\bar{Y}_{B-}$ that in $U_B\cap U_-$. We have

$$\Delta_B = (N_+-A_+-1)\sigma^2_{B+} + (N_--A_--1)\sigma^2_{B-} + \frac{(N_+-A_+)(N_--A_-)}{N-A}(\bar{Y}_{B+}-\bar{Y}_{B-})^2$$

Let $\bar{Y}_{A+}$ be the mean of $s_A\cap U_+$, and let $(A_+-1)\sigma^2_{A+} = \sum_{i\in s_A\cap U_+}(y_i-\bar{Y}_{A+})^2$. Notice that

$$\begin{aligned}(N_+-1)\sigma^2_+ &= \sum_{i\in s_A\cap U_+}(y_i-\bar{Y}_{A+}+\bar{Y}_{A+}-\bar{Y}_+)^2 + \sum_{i\in U_B\cap U_+}(y_i-\bar{Y}_{B+}+\bar{Y}_{B+}-\bar{Y}_+)^2\\ &= (A_+-1)\sigma^2_{A+} + A_+(\bar{Y}_{A+}-\bar{Y}_+)^2 + (N_+-A_+-1)\sigma^2_{B+} + (N_+-A_+)(\bar{Y}_{B+}-\bar{Y}_+)^2\end{aligned}$$

where $E_{s_A}\{(\bar{Y}_{A+}-\bar{Y}_+)^2|\psi\} = \left(1-\frac{A_+}{N_+}\right)\frac{\sigma^2_+}{A_+}$ and $E_{s_A}\{(\bar{Y}_{B+}-\bar{Y}_+)^2|\psi\} = \left(1-\frac{N_+-A_+}{N_+}\right)\frac{\sigma^2_+}{N_+-A_+}$. Thus, $E_{s_A}(\sigma^2_{B+}|\psi) = \sigma^2_+$, taking expectation on both sides above and noting that $E_{s_A}(\sigma^2_{A+}|\psi) = E_{s_A}(\sigma^2_{B+}|\psi)$. Similarly, we obtain $E_{s_A}(\sigma^2_{B-}|\psi) = \sigma^2_-$. Finally,

$$E_{s_A}\{(\bar{Y}_{B+}-\bar{Y}_{B-})^2|\psi\} = (\bar{Y}_+-\bar{Y}_-)^2 + \frac{A_+\sigma^2_+}{N_+(N_+-A_+)} + \frac{A_-\sigma^2_-}{N_-(N_--A_-)}$$

The expression (5) follows then from that of $E_{s_A}(\sigma^2_{B+}|\psi)$, $E_{s_A}(\sigma^2_{B+}|\psi)$ and $E_{s_A}\{(\bar{Y}_{B+}-\bar{Y}_{B-})^2|\psi\}$.

Received 14 July 2010
Accepted 6 November 2010