Outlier Treatment for Disaggregated Estimates

Louis-Paul Rivest
Département de mathématiques et de statistique
Université Laval, Ste-Foy
Québec, Canada, G1K 7P4
lpr@mat.ulaval.ca
and
Mike Hidiroglou
Office of National Statistics
1 Drummond Gate, London
United Kingdom, SW1V 2QQ
mike.hidiroglou@ons.gov.uk

Abstract In many surveys the total sample size is large enough that outliers will have a negligible impact on aggregated estimates. However, their impact can still be substantial at disaggregated levels such as domains. In surveys repeated over time, such as monthly or quarterly, estimates can vary quite a bit between survey occasions, when outliers occur. This article suggests methods to reduce the impact of outliers for disaggregated levels while keeping aggregated estimates unchanged.

The proposed method is akin to using the "Surprise Stratum" procedure proposed by Leslie Kish in his 1965 book, Survey Sampling. In the present context, we illustrate how it can be used for a stratified random sampling plan, where the objective is to reduce the impact of outliers on stratum estimates while keeping the population estimate unchanged. In each stratum a Winsorization cut-off is set; the stratum estimate is the stratum Winsorized estimate plus the stratum share of the data values exceeding their stratum cut-offs.

KEY WORDS: Efficiency, Sample Survey, Stratified Random sampling, Winsorization.

1. Introduction

Outliers in survey sampling are population units that have a very large impact on estimators of population parameters (usually totals). The data associated with these units have large values that have been verified (i.e. they are neither coding nor response errors). Such units with unusually high values would not appear in weighted samples, had this been recognized prior to sampling. Note that if an outlier is considered as a unique value, unrepresentative of the non-sampled part of the population, then it is common practice to reduce its design weight to one (Chambers, 1986). However, outliers are not usually unique; they can appear in both the sampled and the unsampled parts of the population. Their presence in the sample has a large impact on the estimate. A number of procedures have been proposed to curtail their influence, see Lee (1995) Valliant, Dorfman, and Royall (2000, Section 11.1).

Winsorization is a widely used method for curbing the effect of outliers. A threshold is first determined for each variable of interest. The value of survey variables above that threshold are replaced by that threshold. Survey estimates are then computed. This results in an estimator with a downward bias and a variance smaller than that of the corresponding non-Winsorized estimate. Minimizing the mean squared error of a Winsorized estimate has been considered as a technique for selecting the threshold. Kokic and Bell (1994) and Rivest and Hurtubise (1995) suggested methods for calculating the optimal threshold for stratified simple random sampling. The optimal Winsorized mean typically curtails the contribution of few outlier units regardless of the sample size (Rivest and Hurtubise, 1995).

The downward bias of Winsorized estimates is a

major drawback. The bias-variance tradeoff, which is optimal for a single estimate, is not reasonable when estimates are aggregated in time or in space. Aggregated Winsorized estimates have large biases, resulting in less precision than standard aggregated estimates.

This paper investigates new approaches for determining Winsorization thresholds that result in unbiased aggregated estimates. It suggests methods for reducing the impact of outliers on individual survey estimates while keeping aggregated estimates unchanged. This is based on a formalization of the notion of Surprise Stratum defined in Kish (1964, p. 481). Let Y denote the survey variable. Two types of Y-values are considered. The first type assume that the Y-values have a distribution whose mean can safely be estimated for disaggregated levels. The second type consists of large values whose mean is too unstable for disaggregated use. These large values are put in the surprise stratum for which only aggregated estimates are produced. An improved disaggregated estimate is equal to the sample mean for the regular values plus a share of the aggregated estimate for the large values. Techniques for distinguishing the large from the regular values are presented.

Stratified random sampling is briefly reviewed in Section 2. It is shown that one can improve on the stratum sample mean as an estimator of the stratum mean using a one way analysis of variance model for Y. Section 3 discusses Winsorization in stratified random sampling and introduces a method for selecting the threshold that optimizes the estimators of the stratum means. Applications of the new estimator to two real stratified populations are presented in Section 4.

2. A synthetic estimator for stratum means in stratified random sample

Suppose that a stratified random sample has been selected. Let L be the number of strata, W_h, n_h, and N_h respectively denote for $h = 1, \ldots, L$ the stratum weight, the sample size and the population size. Let y_{hi} be the variable of interest associated with the i-th unit $(1, \ldots, n_h)$ within the h-th stratum. The stratum sample mean $\bar{y}_{hs}$ is an unbiased estimator for the stratum mean $\bar{y}_{hU}$. The estimator of the overall population mean $\bar{y}_U$ is given by $\bar{y}_{str} = \sum W_h \bar{y}_{hs}$.

Suppose now that the variable y can be modelled by

$$y_{hi} = \mu_h + \sigma_h \epsilon_{hi} \quad h = 1, \ldots, L \; i = 1, \ldots, N_h \quad (1)$$

where μ_h and $\sigma_h > 0$ are the location and scale parameters for the hth stratum. Assume that the random variables ϵ_{hi}'s are independent repetitions drawn from a skewed distribution with expectation 0 and variance 1. Note that we have made the unrealistic assumption that the distribution of ϵ_{hi} does not depend on the stratum; this assumption will be relaxed in the next section. The scaled Weibull, log-normal, and contaminated normal distributions are potential candidates to represent the distribution of ϵ_{hi}.

A synthetic estimator for $\bar{y}_{hU}$ is defined using an equivariant estimators of the location and of the scale of each stratum. Let a and b denote arbitrary real numbers. The location estimator $\hat{T}(y_1, \ldots, y_n)$ is equivariant if it satisfies $\hat{T}(a+by_1, \ldots, a+by_n) = a + b\hat{T}(y_1, \ldots, y_n)$. A scale estimator $\hat{S}(y_1, \ldots, y_n)$ must satisfies $\hat{S}(a + by_1, \ldots, a + by_n) = |b|\hat{S}(y_1, \ldots, y_n)$ to be equivariant. The median and the interquartile range are possibly the simplest examples of equivariant robust estimators of location and scale. They will be used for constructing a synthetic estimator.

Let $\tilde{y}_{hs}$ and $r_{hs} = \tilde{y}_{h3s} - \tilde{y}_{h1s}$ respectively denote the sample median and the sample interquartile range of stratum h; here $\tilde{y}_{h1s}$ and $\tilde{y}_{h3s}$ respectively denote the first and the third quartile of the y-values in the sample portion belonging to stratum h. The median and the interquartile range are robust measures of location and scale that are not sensitive to outliers. Thus $\tilde{y}_{hs}$ has a variance that is much smaller than $\bar{y}_{hs}$. It is however biased as an estimator of μ_h. The synthetic estimator for μ_h, $y_{hs}^{(syn)}$, is constructed by adding to $\tilde{y}_{hs}$ a bias correction involving the L strata. It is given by

$$y_{hs}^{(syn)} = \tilde{y}_{hs} + r_{hs} \frac{\sum W_k (\bar{y}_{ks} - \tilde{y}_{ks})}{\sum W_k r_{ks}}. \quad (2)$$

Since $\sum W_h y_{hs}^{(syn)} = \sum W_h \bar{y}_{hs}$, the synthetic and the standard estimates coincide at the aggregated level.

We next proceed to determine both theoretically and via a simulation study the sampling properties of $y_{hs}^{(syn)}$ as an estimator of μ_h. Let $\tilde{\mu}$ and $\tilde{\sigma}$ denote the median and the interquartile range of the distribution of ϵ_{ij}. As the stratum sample sizes tend to ∞ then

$$r_{hs} \to \sigma_h \tilde{\sigma} \quad \tilde{y}_{hs} \to \mu_h + \sigma_h \tilde{\mu} \quad \bar{y}_{hs} \to \mu_h, \; h = 1, \ldots, L$$

where "$\rightarrow$" stands for "converges in probability". These results imply that the synthetic estimator is convergent for μ_h under (1).

Simulations were carried out to compare the sampling properties of the direct and of the synthetic estimator. Two distribution in (1) are included for the errors ϵ_{hi} in the Monte Carlo study which is summarized in Tables 1 and 2. The first one is the contaminated distribution of normal investigated by Hidiroglou and Srinath (1981). The second one is the Weibull distribution considered by Fuller and Rivest and Hurtubise (1995) amongst others.

Contaminated normal distributions considered by Hidiroglou and Srinath (1981) can be characterized by 5 parameters: the means (μ_1, μ_2) and the variances (σ_1^2, σ_2^2) respectively representing the regular and outlying components in the sample, and the proportion of outliers p in the sample. Their distribution function is

$$F(x) = (1-p)\Phi((x-\mu_1)/\sigma_1) + p\Phi((x-\mu_2)/\sigma_2),$$

where $\Phi(x)$ is the standard $N(0,1)$ distribution. These distributions typically result in populations where the outliers can be clearly differentiated from the non-outliers. Its skewness coefficient κ_3 is given by,

$$\frac{(1-p)(3\mu_1\sigma_1^2+\mu_1^3)+p(3\mu_2\sigma_2^2+\mu_2^3)-3\mu\sigma^2-\mu^3}{(\sigma^2)^{3/2}},$$

where $\mu = (1-p)\mu_1 + p\mu_2$ is the mean and $\sigma^2 = (1-p)\sigma_1^2 + p\sigma_2^2 + p(1-p)(\mu_1-\mu_2)^2$ is the variance of the contaminated normal. The Weibull distribution function is $F(x) = 1-\exp(-x^\alpha)$ for $x > 0$. Its skewness coefficient is

$$\frac{\Gamma(\beta+1) - 3\Gamma(\beta+1)\Gamma(2\beta+1) + 2\Gamma(\beta+1)^3)}{(\Gamma(2\beta+1) - \Gamma(\beta+1)^2)^{3/2}},$$

where $\beta = 1/\alpha$ and $\Gamma(\cdot)$ is the gamma function.

For all the simulations, populations of L strata with $N_h = 500$ units were simulated following (1), using the same location and scale parameters in all strata. One thousand ($T = 1000$) stratified simple random samples were selected without replacement from the population. For each stratum and each replicated sample, the sample mean $\bar{y}_{hs}^{(r)}$ and the synthetic estimator $y_{hs}^{(syn)(r)}$, given by (2), were calculated for $r = 1, \ldots, T$. The relative bias of the synthetic estimator $y_{hs}^{(syn)}$ in stratum h is

$$RB(y_{hs}^{(syn)}) = \frac{|\sum_r (y_{hs}^{(syn)(r)} - \bar{y}_{hU})|}{T\bar{y}_{hU}}$$

			$RRMSE$		
L	n_h	$B(y_{hs}^{(syn)})$	$\bar{y}_{hs}$	$y_{hs}^{(syn)}$	EFF
20	20	1.74	14.18	9.2	2.4
20	50	1.71	8.73	5.94	2.17
20	100	1.71	5.84	4.26	1.86
10	20	1.51	13.83	9.44	2.19
10	50	1.61	8.48	6.04	2
10	100	1.7	5.67	4.28	1.75

Table 1 : Simulation results for a contaminated normal distribution with parameters $\mu_1 = 3$, $\mu_2 = 15$, $\sigma_1 = 1$, $\sigma_2 = 15$, and $p = .02$ and $\kappa_3 = 5.5$. The average relative biases (B) and the average relative root mean squared errors ($RRMSE$) of the synthetic estimator $y_{hs}^{(syn)}$ are reported in percentage.

while the corresponding mean squared error is

$$MSE(y_{hs}^{(syn)}) = \frac{\sum_r (y_{hs}^{(syn)(r)} - \bar{y}_{hU})^2}{T}.$$

Tables 1 and 2 report the average relative biases, $B(y_{hs}^{(syn)}) = \sum_h RB(y_{hs}^{(syn)})/L$, the average relative root mean squared errors ($RRMSE(y_{hs}^{(syn)}) = \sum_h \sqrt{MSE(y_{hs}^{(syn)})}/(\bar{y}_{hU}L)$) of both $y_{hs}^{(syn)}$ and $\bar{y}_{hs}$, and the efficiency of $y_{hs}^{(syn)}$ defined as $EFF = \sum_h MSE(\bar{y}_{hs}) / \sum_h MSE(y_{hs}^{(syn)})$.

Table 1 shows that the synthetic estimator allows important gains in precision when model (1) holds with contaminated normal errors. These gains are most important when the sample sizes, n_h, are small and when the number of strata, L, is large. Simulation results for a Weibull model with a skewness similar to that of the contaminated normals of Table 1 are presented in Table 2. The results reflect that the synthetic estimator does not improve on the stratum mean even if (1) holds. This finding is reasonable because, when L is large, the synthetic estimator is approximately equal to

$$y_{hs}^{(syn)} \approx \tilde{y}_{hs} - r_{hs}\frac{\tilde{\mu}}{\tilde{\sigma}}$$

where $\tilde{\mu}$ and $\tilde{\sigma}$ are the median and the interquartile range of the error distribution ϵ_{hi} in (1). Given the Weibull distribution with $\alpha = .54$, $\tilde{\mu}/\tilde{\sigma} = -0.72$ and, using some standard results on the asymptotic variance-covariance matrix of the quartiles, one can calculate the asymptotic efficiency of the approximate synthetic estimator with respect to the sample mean. The resulting asymptotic efficiency is equal to 83%,

			$RRMSE$		
L	n_h	$B(y_{hs}^{(syn)})$	$\bar{y}_{hs}$	$y_{hs}^{(syn)}$	EFF
20	20	5	43.82	47.31	0.87
20	50	5.6	26.71	29.87	0.81
20	100	6.07	17.82	20.56	0.76
10	20	5.07	44.1	47.55	0.86
10	50	6.82	26.81	29.74	0.82
10	100	6.27	17.93	21.07	0.73

Table 2 : Simulation results for a Weibull distribution with parameters $\alpha = .54$, and $\kappa_3 = 5.6$. The average relative biases (B) and the average relative root mean squared errors ($RRMSE$) of the synthetic estimator $y_{hs}^{(syn)}$ are reported in percentage.

which is in the ballpark of the efficiencies reported in Table 2. This highlights the limitations of the synthetic estimator and suggest considering more flexible statistics. Such statistics are considered in the next section.

3. Corrected Winsorized means for disaggregated estimators

The Winsorized mean for stratum h is given by $\bar{y}_h^W = \sum \min(y_{hi}, R_h)/n_h$, where R_h is the chosen Winsorization cut-off point for stratum h. Two methods for selecting the cut-offs R_h are presented in this section. The first method is to use the procedure suggested by Kokic and Bell (1994) and Rivest and Hurtubise (1995). Their selection criterion for R_h was to minimize the mean squared error of $\bar{y}_U^W = \sum W_h \bar{y}_h^W$. The second method is to optimize the Winsorization cut-off points within each stratum.

3.1 Kokic and Bell (1994) and Rivest and Hurtubise (1995) method for selecting the Winsorization cut-offs

The mean squared error of $\bar{y}_U^W$ can be expressed as

$$MSE(\bar{y}_U^W) = \sum_{h=1}^{L} W_h^2 \mathrm{Var}(\bar{y}_h^W) + \{\sum W_h \mathrm{E}(\bar{y}_h^W - \bar{y}_h)\}^2.$$

Minimizing this expression as a function of $R_1, \ldots, R_L$ is equivalent to an L-dimensional optimization problem. Kokic and Bell (1994) and Rivest and Hurtubise (1995) suggested an approximate optimal one-dimensional solution to this problem. When the sampling fraction is negligible, Kokic and Bell (1994) equation (6) and Rivest and Hurtubise equation (3.3) provide the following expression for the approximate solution R_h,

$$R_h = \max\left\{\bar{y}_{hU} + \frac{Rn_h}{nW_h}, 0\right\}, \qquad (3)$$

where R is an unknown quantity determined using numerical algorithms given in the above two papers. The value of R is obtained by minimizing $MSE(\bar{y}_U^W)$.

3.2 Selecting the Winsorization cut-offs to optimize stratum estimators

Known interquartile ranges, r_{hU}, are assumed to be available from historical data for $h = 1, \ldots, L$. This section studies the estimator obtained by replacing the median by the winsorized mean in (2),

$$\bar{y}_h^K = \bar{y}_h^W + r_{hU} \frac{\sum W_k(\bar{y}_{ks} - \bar{y}_k^W)}{\sum W_k r_{kU}}, \qquad (4)$$

where the superscript "K" stands for Kish surprise stratum. Noting that $\bar{y}_{ks} - \bar{y}_k^W = \sum \min(y_{hi} - R_h, 0)/n_h$, estimator (4) is therefore similar to the one proposed by Kish (1964). However the contribution of an outlying unit to the surprise stratum estimate is equal to the difference between its y-value and the stratum cut-off. Noting that $\sum W_h \bar{y}_h^K = \sum W_h \bar{y}_{hs} = \bar{y}_{str}$, the aggregated estimate coincides with the standard estimate for the population mean.

Observe that $\bar{y}_h^K$ depends on the cut-offs R_h for the L strata. When the R_h's are all large, $\bar{y}_h^W = \bar{y}_h^K = \bar{y}_{hs}$, the Winsorized mean is equal to the sample mean in each stratum. The proposed estimator reduces to the standard sample mean. On the other hand, as the R_h's goes to 0, $\bar{y}_h^W = 0$ and $\bar{y}_h^K = r_{hU}\bar{y}_{str}/\sum W_k r_{kU}$ is a predetermined fraction of the stratified estimator for the whole population.

As a criterion for selecting the L cut-offs R_h, we suggest minimizing the sum of the stratum mean squared errors,

$$\sum_{h=1}^{L} E\{(\bar{y}_h^K - \bar{y}_{hU})^2\}, \qquad (5)$$

where $\bar{y}_{hU} = \sum_1^{N_h} y_{hi}/N_h$. This quantity is estimated by observing that $\bar{y}_h^W = \bar{y}_{hs} - \sum_i \max(y_{hi} - R_h, 0)/n_h$. Thus, $E\{(\bar{y}_h^K - \bar{y}_{hU})^2\}$ is given by

$$E\left\{\left(\bar{y}_{hs} - \bar{y}_{hU} - \frac{\sum \max(y_{hi} - R_h, 0)}{n_h} + d_h \sum W_k \frac{\sum \max(y_{ki} - R_k, 0)}{n_k}\right)^2\right\},$$

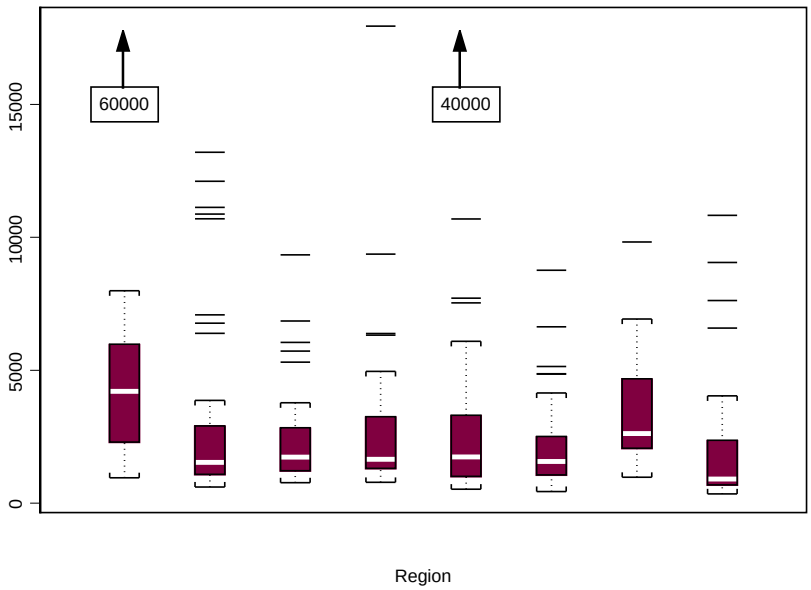

Figure 1: Boxplots of the $REV84$ variable, by regions, for the $MU284$ population

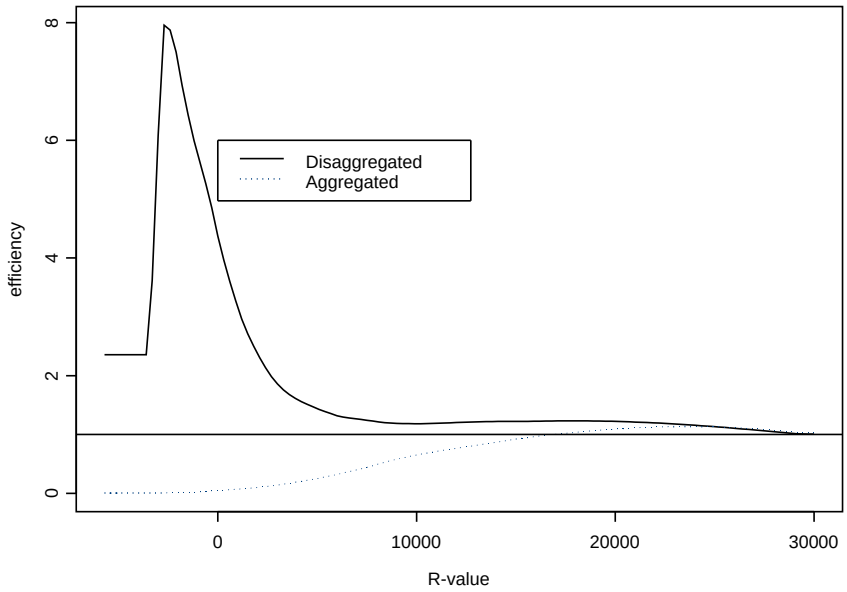

Figure 2: Efficiency plot for determining the optimal value of R for the $MU284$ population

where $d_h = r_{hU}/(\sum W_k r_{kU})$. This mean squared error involves a bias term and some variances.

Let $z_{hi} = y_{hi} - \bar{y}_{hU} - (1 - d_h W_h)\max(y_{hi} - R_h, 0)$ and $w_{hi} = \max(y_{hi} - R_h, 0)$. One has

$$\begin{aligned} E\{\bar{y}_h^K - \bar{y}_{hU}\} &= \bar{z}_{hU} + d_h \sum_{k \neq h} W_k \bar{w}_{kU} \quad (6) \\ &= B_h - d_h \sum W_k B_k, \end{aligned}$$

where $B_h = E(\bar{y}_h^W - \bar{y}_{hU}) = -\bar{w}_{hU}$ is the bias of the Winsorized mean in stratum h. Since the sampling is carried out independently between strata, the variance can be expressed as

$$\mathrm{Var}(\bar{y}_h^K) = \frac{1 - f_h}{n_h} S_{hz}^2 + d_h^2 \sum_{k \neq h} W_k^2 \frac{1 - f_k}{n_k} S_{kw}^2,$$

where S_{hz}^2 and S_{hw}^2 are respectively the variances of z and w in stratum h. The mean squared error $E\{(\bar{y}_h^K - \bar{y}_{hU})^2\}$ of (4) is

$$\begin{aligned} MSE(\bar{y}_h^K) &= \frac{1 - f_h}{n_h} S_{hz}^2 + d_h^2 \sum_{k \neq h} W_k^2 \frac{1 - f_k}{n_k} S_{kw}^2 \\ &\quad + (B_h - d_h \sum W_k B_k)^2, \quad (7) \end{aligned}$$

and the sum of the strata mean squared errors, (5), is

$$\begin{aligned} \sum_{h=1}^{L} E\{(\bar{y}_h^K - \bar{y}_{hU})^2\} &= \sum d_h^2 \sum W_h^2 \frac{1 - f_h}{n_h} S_{hw}^2 \\ &+ \sum \frac{1 - f_h}{n_h} S_{hz}^2 - \sum d_h^2 W_h^2 \frac{1 - f_h}{n_h} S_{hw}^2 \\ &+ \sum (B_h - d_h \sum W_k B_k)^2 \end{aligned}$$

This is a relatively complex expression whose partial derivatives with respect to the cut-offs R_h's do not result in simple expressions. Rather than attempting to locate a global minimum, we set R_h equal to (3) and find the optimal R by evaluating (5) repeatedly. This is implemented in the next section.

4. Some examples

This section applies the Winsorization method of section 3.2 to some data sets. The population interquartile ranges of the individual strata are assumed to be known. In each case the loss function (5) is calculated repeatedly for several values of R, with the stratum cut-offs given by (3), to determine the optimal cut-offs. The performance of the resulting individual estimators of the stratum means at the optimal cut-offs are investigated by evaluating (6) and (7) for the known optimal R. Table 3 and Table 4 report on relative biases, $RB(\bar{y}_h^K) = E(\bar{y}_h^K - \bar{y}_{hU})/\bar{y}_{hU}$, and on relative root mean squared errors, $RRMSE(\bar{y}_h^K) = \sqrt{MSE(\bar{y}_h^K)}/\bar{y}_{hU}$, where the bias and the MSE are given by (6) and (7). The coefficient of variation (CV) of the stratum sample mean and the stratum efficiencies of the corrected Winsorized estimator are also presented.

A modified Neyman allocation formula was used to allocate the sample to the strata, for the two examples given in this section. The sample size for stratum h is proportional to $W_h r_{hU}$, where r_{hU} has been defined as the population interquartile range in stratum h. The maximal sampling fraction n_h/N_h per stratum is set to 75%.

4.1 The MU284 populations of Särndal, Swensson and Wretman (1992)

This database contains $N = 284$ Swedish municipalities stratified in $L = 8$ regions. The survey variable is $REV84$, the 1984 revenues, see Särndal et al. (1992). The distribution of $REV84$ for the populations in each of the 8 regions are summarized in Figure 1. The stratum skewness are in the range 1.3 to 5.8. Given an overall sample size of $n = 141$, the coefficient of variation for $\bar{y}_{str}$ is 7.4%. Note that we are using very high sampling fractions in this examples; in such situations outliers are not a problem anymore, at least for aggregated estimators. The efficiency graph for finding the optimal R-value is given in Figure 3. Observe that the graph for the disaggregated estimates has a local maximum around $R = 19000$; this highlights some numerical difficulties in the construction of a general algorithm for finding the optimal value of R.

The maximum efficiency of 795% is obtained at $R = -2700$; 282 out of the 284 data points are winsorized. Table 3 gives the characteristics of the optimal corrected Winsorized estimators. The correction stabilizes the stratum relative root mean squared errors. For the stratum sample means, the median CV is 21% with a range of 24%. These statistics are reduced to respectively 6.5% and 4.3% with the proposed estimation method. The improvement is most noticeable in Region 8 where there are important outliers despite a small mean and a small interquartile range. Note that in Table 3, Region 8 had the most variable stratum estimates.

h	$\bar{y}_{hU}$	$100n_h/N_h$	$CV(\bar{y}_{hs})$	$RB(\bar{y}_h^K)$	$RRMSE(\bar{y}_h^K)$	EFF
1	6413	72	22	-0.6	10.1	1123.7
2	2971	48	16.9	0.1	7	583.9
3	2475	42	16.5	0.1	7.5	485.8
4	2916	50	17.2	0.1	8.7	526.5
5	3046	55	20.8	0.1	10.7	609
6	2262	38	15.9	0.1	7.1	504.3
7	3648	53	16	0.1	6.4	629.8
8	2269	45	25.3	0.7	8.5	880.2

Table 3: Evaluation of the performance of the corrected Winsorized estimators, with $R = -2700$, and of the stratum sample means for the $MU284$ population. The sampling fraction, the bias, the CV, and the relative root mean squared error are given in percentage.

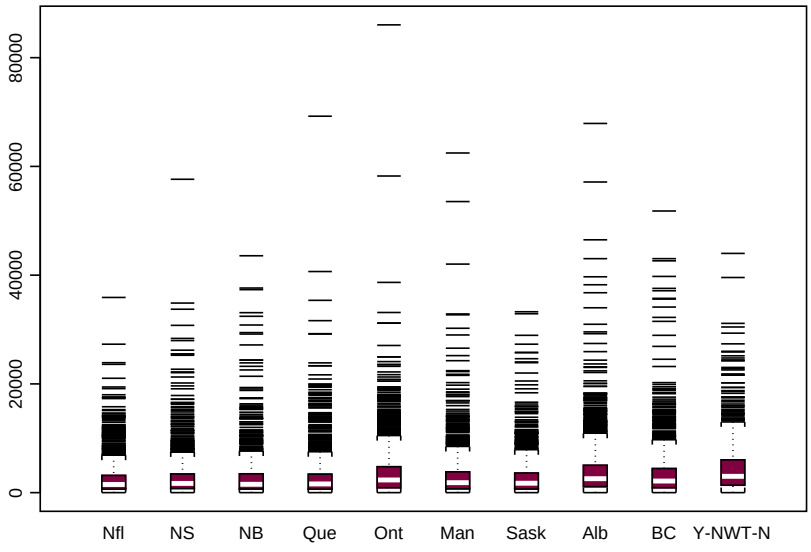

Figure 3: Boxplots for recreation spending by province

With $L = 8$ strata, the maximum efficiency of 8 in Figure 2 means that the optimal corrected Winsorized estimators have a mean squared error comparable to the one associated with $\bar{y}_{str}$.

The optimal aggregated Winsorized estimator corresponds to $R = 24000$; only the two outliers of Figure 1 are Winsorized. The efficiency of the corresponding estimator of $\bar{y}_U$ is 114%. The corresponding stratum Winsorized estimates have efficiencies ranging between 100% and 125%. Given the large sampling fractions, the outliers have a limited impact on the aggregated estimator.

4.2 Statistics Canada Survey of Household Spending

This section uses the data on the variable "Recreation Spending" of the 2001 Survey of Household Spending. The $N = 16012$ households interviewed in this survey is the population under study. It is stratified in $L = 10$ provinces and territories, see Figure 3. The mean yearly spending goes from 2.6K (Newfoundland) to 4.7K (Yukon-North West Territory-Nunavut) while the skewness is in the range 2.6 (Yukon-North West Territory-Nunavut) to 5.3 (Quebec-Ontario). Taking $n = 1000$ yields a 4% CV for $\bar{y}_{str}$. Figure 4 gives the plot of the efficiency as a function of the cut-off R. For disaggregated estimates, the optimal value of R is $R = -2970$; this Winsorizes 91% of the data values . The biases and the mean squared errors of the corrected Winsorized means are given in Table 4. The gains in precision are substantial.

Winsorizing for à la Kokic, Bell, Rivest and Hurtubise, the optimal value is $R = 29334$, and 44 data values are Winsorized; the stratified Winsorized estimator has an efficiency of 108%. The stratum Winsorized estimators all have a negative bias. Their efficiencies with respect to the sample stratum means range between 100% and 121%. Once again, the outliers have a small impact on aggregated estimators.

5. Conclusions

The disaggregated estimation method proposed in this work has done surprisingly well in the two examples of Section 4. More than 80% of the data points are Winsorized, going beyond the control of outliers. Model (1), with $\mu_h = 0$ for all h, might hold relatively well in these two populations. For this model $R = -\infty$ is optimal and stratum estimates are fixed fractions of the population mean estimate. In Figures 2 and 4 the range of R-values for which the disaggregated efficiencies are really large is relatively small; it would be interesting to evaluate the loss of efficiency associated with using sample interquartile ranges in (4). The stratum sample means are approximately normal; they could be improved upon by using standard small area techniques as reviewed in Rao (2003). This will be compared with the techniques proposed in this paper in future work.

REFERENCES

Chambers, R. L. (1986). Design- Adjusted Parameter. *Journal of the Royal Statistical Society. Series A*, **149**, pp. 161-173.

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

Hidiroglou, M. and Srinath, K. P. (1981). Some Estimators for a Population Total from Simple Random Samples Containing Large Units. *Journal of the American Statistical Association*,**76**, pp. 690-695.

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

Kokic, P. N. and Bell, P. A. (1994). Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. *Journal of Official Statistics*, **10**, pp. 419-435.

Lee, H. (1995). Outliers in Business Surveys. In Cox, B., Binder, D., Chinappa, B., Christianson, A., Colledge, M. and Kott, P. (eds.) *Business Survey Methods* . New York, Wiley, pp. 503-526.

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

Rivest, L.-P. and Hurtubise, D. (1995). On Searls' Winsorized Means for Skewed Populations. *Survey Methodology*, **21**, pp. 119-129..

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Verlag: New York.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000) *Finite Population Sampling and Inference. A Prediction Approach.* Wiley: New York.

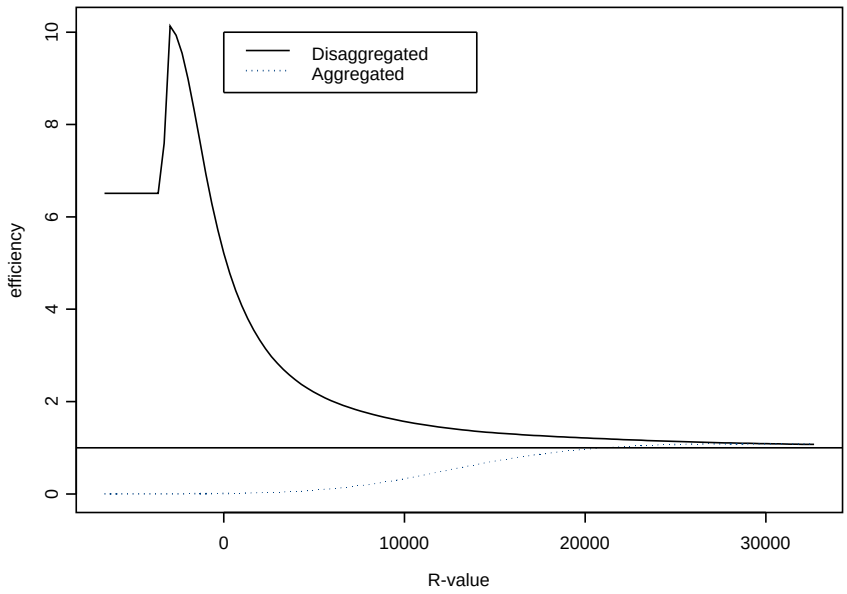

Figure 4: The efficiency of the corrected Winsorized mean for recreation spending

h	$\bar{y}_{hU}$	n_h/N_h	$CV(\bar{y}_{hs})$	$RB(\bar{y}_h^K)$	$RRMSE(\bar{y}_h^K)$	EFF
Nfl	2614	5	14.9	-0.3	3.8	1530.2
NS	2860	5	14.8	-0.6	3.8	1511.5
NB	2822	5	15.4	-0.1	3.9	1556.9
Que	2752	5	13	-0.1	4	1073.5
Ont	3699	7	9.5	0.4	4.1	526.5
Man	3127	6	14.3	-0.1	4	1294.9
Sask	2816	6	13.4	0.1	4.1	1074.9
Alb	4014	8	10.5	-0.1	3.9	728.9
BC	3423	7	10.5	0.3	4.1	648.1
Y-NWT-N	4672	9	12.8	0.1	3.9	1062.6

Table 4: Result for the optimal Kish estimator ($R = -2970$) for recreation spending. All the results are reported in percentage.