Proceedings of Statistics Canada Symposium 2013
Producing reliable estimates from imperfect frames

The detection and treatment of outliers in survey data, in honour of Michel Hidiroglou

Louis-Paul Rivest and Adetona Olabissi Paterne Bada[1]

Abstract

After a brief overview of Michel Hidiroglou's contributions to the detection and treatment of outliers in the processing of survey data, we propose an alternative to the generalized regression estimator that limits the impact of outliers. An estimate of the mean square error of this estimator is proposed. A simulation study based on data from the Canadian Survey of Household Spending shows the benefits of the proposed method.

Key Words: Huber estimator; Generalized regression estimator; GREG; Winsorization.

1. Introduction

1.1 Michel Hidiroglou and the treatment of outliers in survey data

During his career at Statistics Canada, Michel Hidiroglou has made significant contributions in various fields of statistics, especially survey data methodology. We examine his work on the detection and treatment of outliers. An initial key paper is Hidiroglou and Berthelot (1986), which proposes an objective means of detecting outliers in periodic business surveys. This was a landmark paper and is very widely cited. After detection, Michel worked on building a sampling design for populations with large units, a typical issue in economic surveys. For example, Hidiroglou (1986) studies the formation of a take-all stratum, and Lavallée and Hidiroglou (1988) proposes a stratification algorithm that includes a take-all stratum for large units. The Lavallée-Hidiroglou algorithm is now the benchmark stratification method for business surveys. It has been the subject of many studies, and programs to implement it are now widely distributed (*e.g.*, see Baillargeon and Rivest, 2011). In some cases, large values have to be dealt with at the estimation stage; this problem is discussed in Hidiroglou and Srinath (1981), who consider outliers in a simple random sample. A relatively simple treatment method is to Winsorize the largest observation, by replacing it with the second-largest observation before calculating the mean (see Beaumont and Rivest, 2009). The aim of this work is to generalize Winsorization to the general framework of the generalized regression (GREG) estimator.

1.2 Winsorization and GREG estimators

Let $y_1,\ldots,y_n$ be a simple random sample of a highly right-skewed distribution function, such as a log-normal distribution or a Pareto distribution, with mean μ. The sample mean $\bar{y}$ is not the best estimator of μ. The once-Winsorized mean, $\bar{y}_1 = \bar{y} - (y_{(n)} - y_{(n-1)})/n$, where $y_{(n)}$ and $y_{(n-1)}$ represent, respectively, the largest and second-largest values in the sample, has a mean square error smaller than $\bar{y}$. In addition, a nearly unbiased estimator of the mean square error is given by

$$mse(\bar{y}_1) = \frac{s^2}{n} - \frac{(y_{(n)} + y_{(n-1)} - 2\bar{y}_1)(y_{(n)} - 3y_{(n-1)} + 2y_{(n-2)})}{n^2},$$

where s^2 is the sample variance. See Beaumont and Rivest (2009) for a discussion of the statistical properties of $\bar{y}_1$.

[1] Louis-Paul Rivest, Department of Mathematics and Statistics, Université Laval, 1045 rue de la Médecine Québec, Quebec G1V 0A6 Canada (Louis-Paul.Rivest@mat.ulaval.ca); Adetona Olabissi Paterne Bada, Ministry of Development, Economic Analysis and Forecasting, Bénin (badaolabi@yahoo.fr).

The aim of this study is to generalize the once-Winsorized mean to the GREG estimator. This estimator applies to any sampling design with sampling weights w_i and a vector of auxiliary variables **x,** for which $\mathbf{T_x} = \sum_{i \in U} \mathbf{x}_i$ is known where U is the population being studied, which is of size N. The calibrated sampling weights w_i , $i \in s$ where s is the sample, were constructed using a model that expresses the variable under study y as a function of **x**:

$$m: E_m(y_i \mid \mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}_m(y_i \mid \mathbf{X}) \propto v_i = \mathbf{x}_i'\boldsymbol{\lambda} \quad , \quad i \in U,$$

where the index m indicates that the moments are evaluated in relation to the model, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, $\boldsymbol{\lambda}$ is a known vector, and **X** is a matrix with N rows containing $\mathbf{x}_i'$ in its ith row. The GREG estimator of the total of y is written as a function of $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' / v_i\right)^{-1} \sum_{i \in s} w_i \mathbf{x}_i y_i / v_i$, the weighted least squares estimator of $\boldsymbol{\beta}$, $\hat{T}_y^C = \sum_{i \in s} w_i^n y_i = \mathbf{T}_x' \hat{\boldsymbol{\beta}}$, where the w_i^n are the calibrated weights**.** This estimator is sensitive to outliers; a few large values are likely to have a substantial impact on its value.

2. A robust alternative to the GREG estimator

2.1 Beaumont-Rivest estimator (2009)

Beaumont and Rivest (2009) discuss the many alternatives to the estimator $\hat{T}_y^C$ that have been proposed in the statistical literature. Most are based on an alternative to the weighted least squares estimator $\hat{\boldsymbol{\beta}}$ constructed using the Huber function $\psi_c(t)$,

$$\psi_c(t) = \begin{cases} t & \text{if} \quad |t| < c \\ \mathrm{sign}(t) \times c & \text{if} \quad |t| \geq c \end{cases} ,$$

where c is a constant to be determined. The Huber estimator limits the contribution of units with a residual greater than c. Lastly, they propose using the robust calibrated estimator equal to $\hat{T}_y^{RC} = \mathbf{T}_\mathbf{x}' \hat{\boldsymbol{\beta}}^{GM}$ where $\hat{\boldsymbol{\beta}}^{GM}$ is a robust estimator of $\boldsymbol{\beta}$ defined as the solution to the equation $\mathbf{U}(\boldsymbol{\beta}) = 0$, where

$$\mathbf{U}_c(\boldsymbol{\beta}) = \sum_{i \in s} \frac{w_i}{h_i} \psi_i^* \left(h_i \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sqrt{v_i}} \right) \frac{\mathbf{x}_i}{\sqrt{v_i}} ,$$

v_i is the residual variance under model m, h_i is a positive measurement of the potential leverage of unit i, and $\psi_i^*(t)$ is a Huber function modified to take the sampling weight into account,

$$\psi_i^*(t) = \frac{t}{w_i} + \frac{(w_i - 1)}{w_i} \psi_c(t) .$$

When c tends toward infinity, $\psi_i^*(t)$ tends toward t for all i, and the solution of $\mathbf{U}_c(\boldsymbol{\beta}) = 0$ is the weighted least squares estimator $\hat{\boldsymbol{\beta}}$ calculated with weights w_i. It is therefore necessary to determine a value for the constant c in the Huber function. Beaumont and Rivest (2009) use an iterative procedure that chooses c as the value that minimizes an estimate of the MSE of $\hat{T}_y^{RC}$. A new way of determining c is proposed in the next section.

2.2 Determining the constant c in the Huber function

After Beaumont and Alevi (2004), we let $h_i = w_i \sqrt{v_i}$. As a result, the estimation function for $\boldsymbol{\beta}$ becomes $\sum_{i \in s} \psi_i^* \left(w_i (y_i - \mathbf{x}_i'\boldsymbol{\beta}) \right) \mathbf{x}_i / v_i = 0$. We propose setting constant c to the smallest value for which the impact of a single observation is reduced, that is, for which $w_i \mid y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{GM} \mid$ is greater than c for only one $i \in s$. In this section, we discuss the calculation of $\hat{\boldsymbol{\beta}}^{GM}$ in this context.

Let k be the unit of s with the largest weighted residual $k = \arg\max\left\{w_i \mid y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\mid\right\}$, and we set $c_0 = w_k \mid y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}\mid$. For small values of $z>0$, the solution $\hat{\boldsymbol{\beta}}_z$ of $\mathbf{U}_{c_0-z}(\boldsymbol{\beta}) = 0$ is a linear function of z, $\hat{\boldsymbol{\beta}}_z = \hat{\boldsymbol{\beta}} + Az$, where $\hat{\boldsymbol{\beta}}$ is the estimator of the weighted least squares associated with the GREG. The slope A is given by

$$A = -\left\{\sum_{i\in s} w_i \frac{\mathbf{x}_i\mathbf{x}_i'}{v_i} - (w_k - 1)\frac{\mathbf{x}_k\mathbf{x}_k'}{v_k}\right\}^{-1} \frac{w_k - 1}{w_k}\operatorname{sgn}(y_k - x_k'\hat{\boldsymbol{\beta}})\frac{\mathbf{x}_k}{v_k} .$$

To prove this result, we note that for $c=c_0-z$,

$$\psi_k^*\left(w_k(y_k - \mathbf{x}_k'\boldsymbol{\beta})\right) = y_k - \mathbf{x}_k'\boldsymbol{\beta} + (w_k - 1)\{y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}} - \operatorname{sgn}(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}})z / w_k\} .$$

Thus, we are looking for b, such that $\mathbf{U}_{c_0-z}(\hat{\boldsymbol{\beta}} + b) = 0$. This equation can be written

$$\sum_{i\in s} w_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{x}_i'b)\frac{\mathbf{x}_i}{v_i} - (w_k - 1)\{y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \mathbf{x}_k'b - y_k + \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \frac{\operatorname{sgn}(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}})\mathrm{z}}{w_k})\}\frac{\mathbf{x}_k}{v_k} = 0 .$$

Since $\sum_{i\in s} w_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i / v_i = 0$, we conclude that $b=Az$.

The residuals calculated with $\hat{\boldsymbol{\beta}}_z$ are $r_{iz} = w_i(y_i - x_i'\hat{\boldsymbol{\beta}}_z) = w_i(y_i - x_i'\hat{\boldsymbol{\beta}}) - w_i x_i' Az$. The unit $i\neq j$ has its weight reduced in the estimate if r_{iz} exceeds the upper limit of c_0-z or the lower limit of -c_0+z. This occurs at

$$z_{is} = \frac{c_0 - w_i(y_i - x_i'\hat{\boldsymbol{\beta}})}{1 - w_i x_i' A} \text{ and } z_{i\ell} = \frac{c_0 + w_i(y_i - x_i'\hat{\boldsymbol{\beta}})}{1 + w_i x_i' A}$$

respectively. These two values are always positive, because when $c>c_0$, all the residuals have an absolute value less than c. The optimal value of z is equal to the minimum, over the $i\neq k$, of the z_{is} and the $z_{i\ell}$. The index j, for which the minimum is reached, identifies the unit with the second-largest residual. Hence, the procedure consists in choosing c so that the two largest residuals contribute equally to the equation $\mathbf{U}_c(\boldsymbol{\beta}) = 0$. Lastly, it is worth noting that the residuals for units k and j can have different signs; however, if the distribution of errors in the regression model m is skewed, the largest residuals usually have the same sign. Assuming, for example, that the two largest residuals are positive, we have

$$\hat{T}_y^{RC} = \mathbf{T}_{\mathbf{x}}'(\hat{\boldsymbol{\beta}} + A\frac{w_k(y_k - x_k'\hat{\boldsymbol{\beta}}) - w_j(y_j - x_j'\hat{\boldsymbol{\beta}})}{1 - w_j x_j' A}) = \sum_{i\in s} w_i^n y_i + \mathbf{T}_{\mathbf{x}}' A\frac{w_k(y_k - x_k'\hat{\boldsymbol{\beta}}) - w_j(y_j - x_j'\hat{\boldsymbol{\beta}})}{1 - w_j x_j' A} ,$$

where w_i^n are the calibrated weights from Section 1.2. We call this estimator the Winsorized generalized regression (WGREG) estimator, since it generalizes the Winsorized estimator from Section 1.2. This is demonstrated below.

2.3 A special case: Simple random design without replacement

In a simple random design without replacement, $w_i=N/n$, $\mathbf{x}_i=1$ and $\hat{\boldsymbol{\beta}}$ is the sample mean $\bar{y}$. Suppose that the largest residual is $c_0 = N(y_{(n)} - \bar{y})/n$. Hence, we have $A=-(1-f)/(N-N/n+1)$, where $f=n/N$ is the sampling fraction. Thus,

$$z_{is} = (N - N/n + 1)\frac{y_{(n)} - y_i}{n} \text{ and } z_{i\ell} = \frac{N(N - N/n + 1)}{nN - 2N + 2n}(y_{(n)} + y_i - 2\bar{y}) .$$

If the distribution of y is skewed, we will obtain the smallest z_i at the second-largest value in the sample, and the estimator $\hat{\boldsymbol{\beta}}^{GM}$ of the mean is

$$\hat{\boldsymbol{\beta}}^{GM} = \bar{y} + A(N - N/n + 1)\frac{y_{(n)} - y_{(n-1)}}{n} = f \times \bar{y} - (1 - f) \times \bar{y}_1 .$$

This is a compromise between the once-Winsorized mean and the mean, as discussed in Section 3 of Beaumont and Rivest (2009). If outliers in both tails of the distribution are possible, the minimum can be obtained with the z_{il} corresponding to the sample minimum. Thus, we have

$$\hat{\boldsymbol{\beta}}^{GM} = \bar{y} - \frac{N-n}{Nn-2N+2n}(y_{(n)} + y_{(1)} - 2\bar{y})$$

$$= f \times \bar{y} - (1-f)\frac{\sum_{i\in s} y_i(1+2/N) - y_{(1)} - y_{(n)}}{n(1+2/N)-2} \approx f \times \bar{y} - (1-f) \times \frac{\sum_{i\in s} y_i - y_{(1)} - y_{(n)}}{n-2}$$

Consequently, if there are extreme values in both tails of the distribution, the proposed estimator is a compromise between the once-trimmed mean and the sample mean.

3. Estimation of the mean square error

3.1 Estimation of the mean square error

Since only one value in the sample is modified, the asymptotic distribution of the WGREG estimator is the same as that of the GREG estimator. Even if the removal of the largest observation has an impact o(1/n) on the variance, that impact may be substantial (for example, see Rivest and Hurtubise (1995)), and it is worthwhile looking for a modification of the standard variance of the GREG that reflects the fact that the contribution of the largest residual has been reduced. We propose such an estimator for with-replacement sampling designs.

For a with-replacement design, the estimator of the variance of the GREG obtained by linearization is written

$$v(\hat{T}_y^C) = \frac{n}{(n-1)}\sum_{i\in s}\{w_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\}^2;$$

it involves the sample variance of the weighted residuals $\left\{nw_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) : i \in s\right\}$. In estimating the variance, we have two cases, depending on whether the two largest residuals identified in the estimation procedure have the same sign or not. If they have the same sign, we use the following generalization of the mean square error for the once-Winsorized estimator:

$$v(\hat{T}_y^{CG}) = \frac{s_r^2}{n} - \frac{(r_{(n)}^* + r_{(n-1)}^*)(r_{(n)}^* - 3r_{(n-1)}^* + 2r_{(n-2)}^*)}{n^2}$$

where s_r^2 is the variance of the residuals $\left\{r_i = nw_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GM}) : i \in s\right\}$ and the $r_{(i)}^*$ are order statistics of the sample of the $|r_i|$. When the two largest residuals have different signs, we propose using s_w^2/n, where s_w^2 is the variance of the Winsorized observations. This is the variance of the residuals $\left\{r_i = nw_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GM}) : i \in s\right\}$ obtained after the largest residual is replaced with minus the second-largest residual. This is a standard estimator of the variance of the trimmed mean.

3.2 Ratio estimator

Our simulation study deals with the case in which only one explanatory variable x is available and, in the model, the variance of the residuals is proportional to x. In this case, $\hat{\boldsymbol{\beta}} = \sum_{i\in s} w_i y_i / \sum_{i\in s} w_i x_i$ and the calibrated weights are $w_i^n = (1/\pi_i)\sum_{j\in U} x_j / \left\{\sum_{j\in s} x_j/\pi_j\right\}$; in addition, the variance estimator is calculated with the residuals $\left\{nw_i(y_i - x_i\hat{\boldsymbol{\beta}}) : i \in s\right\}$. For the WGREG estimator, assuming that the two largest residuals are positive, we have $A = -\{(w_k - 1)/w_k\}/\{\sum_{i\in s} x_i w_i - (w_k - 1)(x_k - w_j x_j / w_k)\}$ and

$$\hat{\boldsymbol{\beta}}_z = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i x_i} - \frac{(w_k - 1)\{y_k - x_k\hat{\boldsymbol{\beta}} - w_j(y_j - x_j\hat{\boldsymbol{\beta}})/w_k\}}{\sum_{i\in s} x_i w_i - (w_k - 1)(x_k - w_j x_j / w_k)} = \frac{\sum_{i\in s} y_i w_i - (w_k - 1)(y_k - w_j y_j / w_k)}{\sum_{i\in s} x_i w_i - (w_k - 1)(x_k - w_j x_j / w_k)}.$$

3.3 Monte Carlo experiment

This Monte Carlo study compares the GREG and WGREG estimators. Does reducing the largest residual's contribution to the GREG estimator make it biased? Does this operation lead to a substantial reduction of the mean square error of the estimator of the total? The study also assesses the mean square error estimator from Section 3.1. Is $v(\hat{T}_y^{CG})$ a better estimator of the mean square error of $\hat{T}_y^{CG}$ than the standard variance estimator obtained with the linearization method?

To answer these questions, we used the N=15,222 records from the 2005 edition of Statistics Canada's Survey of Household Spending (SHS). The explanatory variable x is household income; our study concerns four dependent variables: y_1, leisure expenses; y_2, clothing expenses; y_3, recreational vehicle expenses; and y_4, life insurance expenses. The last two variables are 0 more than 60% of the time. The skewness of the residuals in the regression model for the four variables is 4, 0.3, 9 and 38. Samples of n=100, 300 and 500 units were selected from the N units in the population using a with-replacement design, with probabilities proportional to the SHS selection probabilities. The simulation is based on 1,000 repetitions. For the four y variables, family income is auxiliary data, and the ratio estimator is more precise than the Horvitz-Thompson estimator.

Figure 3.3.1
Percentage relative bias of the WGREG estimator

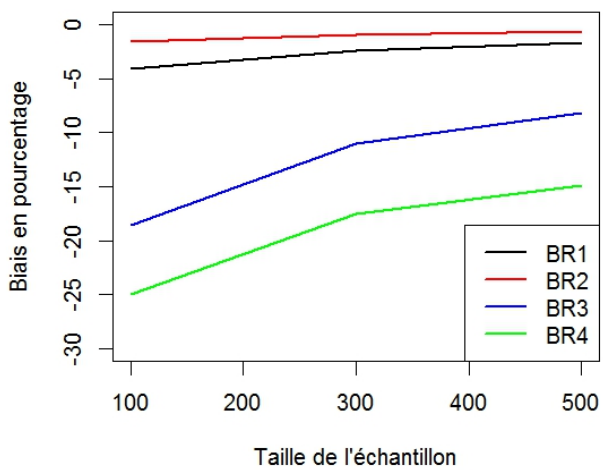
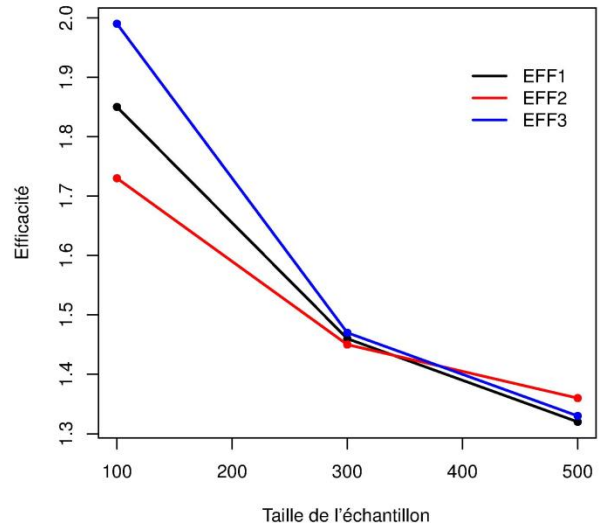


Figure 3.3.1 shows the relative bias of the WGREG estimator for the four dependent variables. In general, the outliers have positive residuals and, therefore, limiting their contribution results in a negative bias. For the first two variables, that bias is relatively small. It is larger for the last two variables, which are, as we have seen, highly skewed.

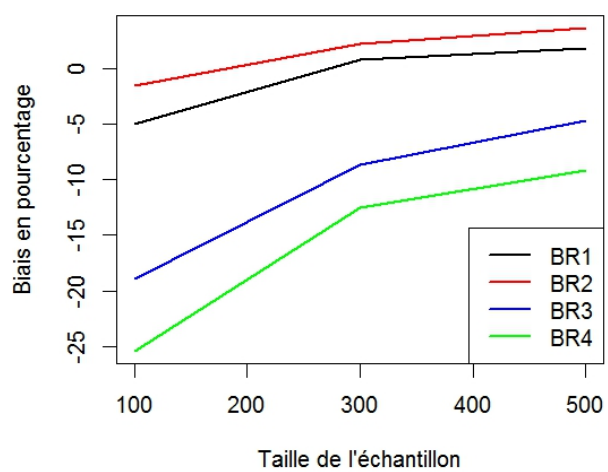
Figure 3.3.2 shows the efficiency of the WGREG estimator compared with the GREG estimator, the efficiency being defined as the ratio $EQM(\hat{T}_y^{C}) / EQM(\hat{T}_y^{CG})$. The efficiency for the 'life insurance expenses' variable is greater than 10 for n=100 and is, therefore, not included in Figure 3.3.2 to simplify its interpretation. Even with a sample size of 500, limiting the largest residual's impact produces a substantial 30% efficiency gain for the three variables shown in Figure 3.3.2. For the first two variables, the CV of the standard GREG estimator decreases from about 20% when n=100, to 10% when n=500. The corresponding values for the WGREG estimator are about 15.5% and 8.5% respectively. For these two variables, giving one outlier special treatment helps to make the estimate more acceptable for publication. The estimates of the totals for the last two variable have CVs greater than 25% for all the sample sizes and all the estimators. These estimates are, in all cases, too imprecise to be published.

Figure 3.3.2
Efficiency of the WGREG estimator compared with the GREG estimator



Lastly, Figure 3.3.3 shows the relative bias of the MSE estimator proposed in Section 3.1. For the first two variables, that bias is relatively small for all sample sizes considered. In most cases, the two largest residuals have the same sign, and the generalization of the MSE estimator for the Winsorized mean is used. The standard with-replacement variance of the GREG estimator, $v(\hat{T}_y^C)$, overestimates the actual mean square error of $\hat{T}_y^{CR}$. This bias is corrected by the estimation in Section 3.1, at least for the first two *y* variables.

Figure 3.3.3
Percentage relative bias of the MSE estimator for the WGREG



4. Discussion

Limiting the contribution of a small number of influential observations appears to be the key to stabilizing the GREG estimator of the total of the variable of interest. The rule used here for choosing the Huber constant *c* is similar to the rule proposed in Section 3.4 of Beaumont et al. (2012), which minimizes the maximum of the conditional biases. The approach proposed here generalizes the methods already available for Winsorizing a sample of an infinite population to the GREG estimator. The MSE estimator proposed in Section 3.1 produced good results in the simulation studies. It is simple, and merits further study.

References

Baillargeon, S. and L.-P. Rivest (2011), Élaboration de plans stratifiés en R à l'aide du programme stratification, *Techniques d'enquête*, 37, 59-72.

Beaumont, J.-F. and A. Alavi (2004), Estimation robuste par la régression généralisée, *Techniques d'enquête*, 30, 217–231.

Beaumont, J.-F and L.-P. Rivest (2009), Dealing with Outliers in Survey Data. D. Pfeffermann and C. R Rao (Eds.) *Handbook of Statistics – Sample Surveys: Design, Methods and Applications*, Volume 29A, The Netherlands, North-Holland, 247–280.

Beaumont, J.-F., Haziza, D., and A. Ruyz-Gazen (2013), A unified approach to robust estimation in finite population sampling, *Biometrika*, 100, 555-569.

Hidiroglou, M. and K.P. Srinath, (1981), Some estimators of a population total from simple random samples containing large units, *Journal of the American Statistical Association*, 76, 690-695.

Hidiroglou, M. and J.-M. Bertholot (1986), Contrôle statistique et imputation dans les enquêtes-entreprises périodiques, *Techniques d'enquête*, 12, 79-89.

Hidiroglou, M. (1986), The construction of a self-representing stratum of large units in survey design, *The American Statistician*, 40, 27-31.

Lavallée, P. and M. Hidiroglou (1988), Sur la stratification de populations asymétriques, *Techniques d'enquête*, 14, 35-45.

Rivest, L.-P. and D. Hurtubise (1995), Moyennes winsorisées de Searls pour populations asymétriques, *Techniques d'enquête*, 21, 119-129.