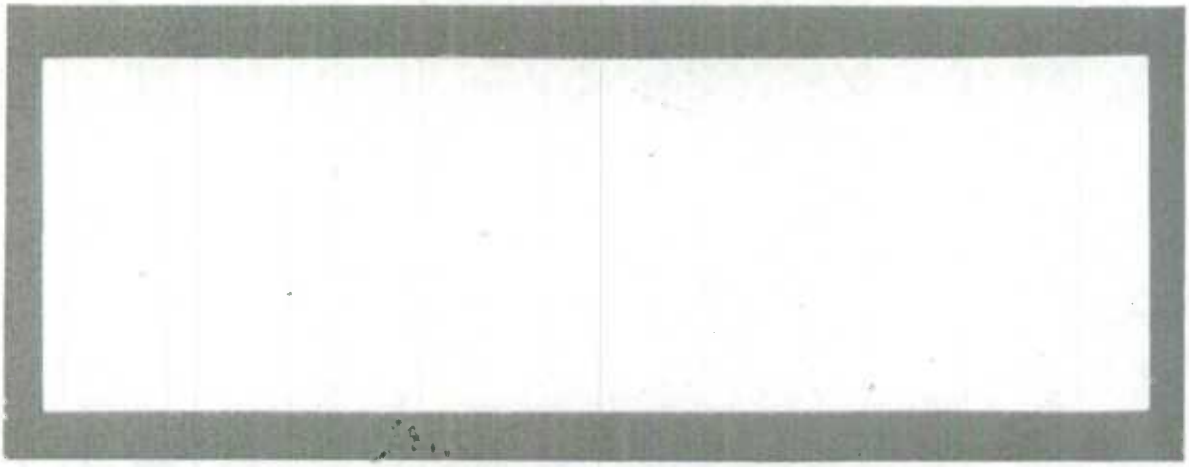

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER

METHODOLOGY BRANCH

ON THE TREATMENT OF INFLUENTIAL OBSERVATIONS IN HOUSEHOLD SURVEYS

HSMD-2001-002E

Jean-François Beaumont and Asma Alavi[1]

Labour Force Survey Methods Section

Household Survey Methods Division

Statistics Canada

January 2001

[1] The work presented here is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada.

ON THE TREATMENT OF INFLUENTIAL OBSERVATIONS IN HOUSEHOLD SURVEYS

Jean-François Beaumont and Asma Alavi[2]

Labour Force Survey Methods Section
Household Survey Methods Division
Methodology Branch, Statistics Canada

ABSTRACT

Household expenditure or income surveys often deal with highly skewed distributions, which potentially lead to samples with some extreme observations. The problem is aggravated by the fact that there usually is a low amount of useful auxiliary information available at the design stage and that the sampling design is complex most of the time, leading to widely dispersed design weights. Therefore, it could happen that a large value be associated with a large design weight and that this combination have a great influence on the estimates produced by the survey. Design consistent estimators, such as the Generalized REGression (GREG) estimator, are usually highly variable in the presence of influential observations but they have a low bias whereas model-based estimators are more stable but they are generally not consistent and more biased. In this paper, a compromise between these two types of estimators is proposed and a simulation study shows that it performs well with respect to the bias and mean squared error (MSE) criteria in comparison with some other robust estimators. Conditions under which the compromise should have a small design bias are also given.

KEYWORDS: GREG estimator; Synthetic estimator; Model-based estimator; M-estimator; Robust estimator; Outliers.

---

[2] Jean-François Beaumont and Asma Alavi, Labour Force Survey Methods Section, Household Survey Methods Division, Methodology Branch, Statistics Canada, Ottawa, Ontario, K1A 0T6.

À PROPOS DU TRAITEMENT DES OBSERVATIONS INFLUENTES DANS LES ENQUÊTES AUPRÈS DES MÉNAGES

Jean-François Beaumont et Asma Alavi[3]

Section des méthodes de l'Enquête sur la population active
Division des méthodes d'enquêtes auprès des ménages
Direction de la méthodologie, Statistique Canada

RÉSUMÉ

Les enquêtes sur les dépenses ou sur le revenu des ménages font souvent face à des distributions très assymétriques, ce qui conduit potentiellement à des échantillons avec des observations extrêmes. Le problème est amplifié par le fait qu'il n'y a généralement qu'une faible quantité d'information auxiliaire utile disponible lors de la conception du plan de sondage et que le plan de sondage est la plupart du temps complexe, ce qui conduit à des poids de sondage très dispersés. Il pourrait donc arriver qu'une grande valeur soit associée à un grand poids de sondage et que cette combinaison ait une grande influence sur les estimations produites par l'enquête. Les estimateurs convergents par rapport au plan de sondage, tel que l'estimateur par la RÉGression Généralisée (REGG), sont généralement très variables en présence d'observations influentes mais ont un faible biais tandis que les estimateurs basés sur un modèle sont plus stables mais ne sont généralement pas convergents et plus biaisés. Dans cet article, un compromis entre ces deux types d'estimateurs est proposé et une étude de simulation montre que ce compromis donne de bons résultats en regard du biais et de l'erreur quadratique moyenne (EQM) en comparaison avec d'autres estimateurs robustes. On donne également les conditions pour lequel ce compromis devrait avoir un faible biais par rapport au plan de sondage.

MOTS-CLÉS: Estimateur REGG; Estimateur synthétique; Estimateur basé sur un modèle; Estimateur-M; Estimateur robuste; Données aberrantes.

---

[3] Jean-François Beaumont et Asma Alavi, Section des méthodes de l'Enquête sur la population active, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6.

1. INTRODUCTION

Household expenditure or income surveys often deal with highly skewed distributions, which potentially lead to samples with some extreme observations. The problem can be aggravated when such extreme observations are associated with large design weights. In business surveys, there are often some useful (well correlated with the variables of interest) auxiliary variables available at the design stage that can be used to reduce the effect of such extreme observations. In fact, the survey methodologist would like to assign a large selection probability (therefore, a small design weight) to units with large values of the variables of interest and vice-versa. This can be achieved with proper stratification or with probability proportional to size sampling when the available auxiliary variables are well correlated with the variables of interest. In household surveys, however, it is generally more difficult to assign a large selection probability to units with large values of the variables of interest because of the low amount of useful auxiliary variables available at the design stage. Moreover, several variables are usually collected in the same survey and appropriate auxiliary variables for one variable of interest may not necessarily be appropriate for another. Also, complex sampling designs are more frequent in household surveys (for example, stratified multi-stage designs), leading to widely dispersed design weights. It may thus well happen that a large value be associated with a large design weight and that this combination have a great influence on the estimates produced by the survey.

In this paper, an observation is defined as being influential if its exclusion or inclusion in the sample affects the estimates greatly. Therefore, an observation can be influential because of a large design weight, a large value or the combination of both. An influential observation has to be distinguished from an extreme observation, which is an observation isolated from the bulk of the data. An extreme observation is often associated to a large value of the variable of interest (or a large regression residuals if a regression estimator is used). Note that an extreme observation is not necessarily influential in a large sample. The term *outlier* is also frequently seen in the literature and it usually refers to either an extreme or an influential observation. These definitions are very closely related to those

of Lee (1995) and are suited to survey sampling. They may differ from those seen in the statistical literature not related to survey sampling. It should be noted that the focus of this paper is on influential observations. No matter how the term influential observation is defined, the identification or detection of influential observations in a given sample still remains somewhat arbitrary. Although it is not a main issue in this paper, detection of influential observations is briefly discussed in section 3.

Design consistent estimators, such as the Generalized REGression (GREG) estimator, may be highly variable in the presence of influential observations. Of course, using auxiliary variables that are well correlated with the variables of interest at the design stage as well as at the estimation stage is always recommended and helpful to reduce the variability of any estimator, including the GREG estimator. However, such useful auxiliary variables are often not available and, as a result, more robust (to influential observations) estimators may be needed. Modifying the value (for example, the *Winsorization* technique) or modifying the weight (see, for example, Hidiroglou and Srinath, 1981) of an influential observation are the two traditional approaches that have been used in sample surveys to obtain robust estimators. More recently, the M-estimation technique has been considered to form robust estimators (see, among others, Chambers, 1986; Gwet and Rivest, 1992; Lee, 1991, 1995 and Hulliger, 1995).

In the second section, the notation is introduced along with the usual estimators, such as the GREG estimator. In the third section, some robust estimators, including the Winsorized estimator, are described. These estimators require two steps: a detection step and a treatment step. In the fourth section, other robust estimators, which do not require a detection step, are presented and proposed. Some of them are based on the M-estimation technique. In the fifth section, a simulation study comparing different estimators is described and the results are shown in section 6. Finally, a brief conclusion is found in the last section.

2. BACKGROUND

Let us first assume that we want to estimate the total of a variable of interest y for a population U and denote this unknown population parameter by $t_y = \sum_{k \in U} y_k$. Because it is usually not feasible to observe the variable y for all units (households) in the population, a random sample s is drawn. Let us also assume that we have a vector of auxiliary variables, $\mathbf{x_k}$, available for all units of the sample s and for which the population totals, $\mathbf{t_x} = \sum_{k \in U} \mathbf{x_k}$, are known. The GREG estimator can then be used to estimate the unknown total t_y:

$$\hat{t}_y^G = \hat{t}_y^{HT} + \hat{\mathbf{B}}'\left(\mathbf{t_x} - \hat{\mathbf{t}}_x^{\mathrm{HT}}\right), \tag{2.1}$$

where $\hat{t}_y^{HT}$ and $\hat{\mathbf{t}}_x^{\mathrm{HT}}$ are Horvitz-Thompson estimators given respectively by $\hat{t}_y^{HT} = \sum_{k \in s} w_k y_k$ and $\hat{\mathbf{t}}_x^{\mathrm{HT}} = \sum_{k \in s} w_k \mathbf{x_k}$, w_k is the design weight obtained by the inverse of the selection probability π_k and

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} \frac{a_k}{c_k} \mathbf{x_k} \mathbf{x_k}' \right)^{-1} \sum_{k \in s} \frac{a_k}{c_k} \mathbf{x_k} y_k \,. \tag{2.2}$$

In most practical cases, $a_k = w_k$ and $c_k = \boldsymbol{\lambda}' \mathbf{x_k}$, where $\boldsymbol{\lambda}$ is a vector of known constants. In these cases, it can easily be shown (see Särndal, Swensson and Wretman, 1992, p.231) that (2.1) reduces to the synthetic estimator

$$\hat{t}_y^S = \hat{\mathbf{B}}' \mathbf{t_x} \,. \tag{2.3}$$

In fact, $\hat{t}_y^G$ can be justified through the following model m:

$$y_k = \boldsymbol{\beta}' \mathbf{x_k} + \varepsilon_k \,,$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters, ε_k is a random error with $\mathrm{E}_m(\varepsilon_k)=0$, $\mathrm{E}_m(\varepsilon_k\varepsilon_l)=0$, for $k\neq l$, $\mathrm{E}_m(\varepsilon_k^2)=\sigma^2 c_k$, and σ^2 is an unknown parameter. The determination of c_k in (2.2) can therefore be justified through the assumed model variance. It will be assumed in the following that c_k is a known function of $\mathbf{x}_k$. Also, throughout this paper, all expectations under the model are conditional on the observed values of the auxiliary variables. For instance, we have used $\mathrm{E}_m(\varepsilon_k)=0$ instead of $\mathrm{E}_m(\varepsilon_k \mid \mathbf{x}_k)=0$ to simplify the notation.

Regarding the determination of a_k, the usual choice $a_k=w_k$ makes $\hat{\mathbf{B}}$ a design consistent estimator for the population parameter

$$\mathbf{B}=\left(\sum_{k\in U}\frac{1}{c_k}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_{k\in U}\frac{1}{c_k}\mathbf{x}_k y_k\ ,$$

whether the model holds or not. Note that this population parameter would be the best linear unbiased estimator for $\boldsymbol{\beta}$ under the model, if every household in the population could be observed. On the other hand, the choice $a_k=1$ will make $\hat{\mathbf{B}}$ the best linear unbiased estimator for the vector of parameters $\boldsymbol{\beta}$ under the model, provided that condition (c.2) of section 4 is satisfied (which is basically equivalent to say that the sampling mechanism is ignorable). It is also interesting to note that $\hat{t}_y^G$ is design consistent whether the model holds or not and no matter how a_k is specified.

An important and useful feature of $\hat{t}_y^G$ is that it can be expressed as the following weighted sum:

$$\hat{t}_y^G=\sum_{k\in s}w_k^* y_k\ , \tag{2.4}$$

where

$$w_k^* = w_k + \frac{a_k}{c_k}\mathbf{x}_k'\left(\sum_{k\in s}\frac{a_k}{c_k}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT}\right). \qquad (2.5)$$

Once the estimation weights w_k^* have been calculated, they can be provided to data users who can thus obtain estimates for any variable of interest just by the means of a simple weighted sum. It is also worth mentioning that when these estimation weights are applied to the auxiliary variables, we get the known population totals ($\sum_{k\in s} w_k^*\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$). The last equation is called the calibration equation because the estimation weights are calibrated to get the known population totals.

If the objective is to estimate the population mean, $\mu_y = \sum_{k\in U} y_k / N$, where N is the population size, then the following estimator can be used:

$$\hat{\mu}_y^G = \sum_{k\in s} w_k^* y_k \Big/ \sum_{k\in s} w_k^* . \qquad (2.6)$$

If we are interested in estimating a domain total or a domain mean, we may just replace the sums over all units of the sample s in (2.4) and (2.6) by sums over all units of the sample s which belong to the domain of interest.

Although estimators (2.4) and (2.6) are design consistent, they may be highly variable in the presence of influential observations. So, data users may well be willing to use slightly biased estimators in order to significantly reduce the variance. In the next two sections, such robust estimators, which are less affected by influential observations, are studied. Modifying the estimation weight or modifying the value of an influential observation are the two approaches that are considered. These two approaches are very easy to implement in practice and very appealing to users, especially the former.

3. DETECT-AND-TREAT ESTIMATION METHODS

All estimation methods described in this section require two steps: a detection step and a treatment step. Subsection 3.1 briefly discusses detection of influential observations and subsection 3.2 describes methods treating observations that have been identified as being influential.

3.1 Detection of influential observations

There exist several techniques to detect outliers. For example, Lee (1995) discusses a number of them, including the quartile method, which is often used in practice. Although the quartile method may be very efficient for finding extreme observations, it may not necessarily be the most appropriate method for finding influential observations even if a weighted version of the quartile method is used. In a large sample, for instance, an extreme (weighted or not) observation may not have a great impact on estimates and thus may not be considered as being influential. However, extreme observations are likely to have a great impact on estimates for small domains and are of interest even if the emphasis is on influential observations. But, for reasons of simplicity and also because the interest of this paper is more on treatment than detection of influential observations, we rather preferred the following simple rule: a unit l is identified as being influential when

$$\frac{w_l^* y_l}{\sum_{k \in s} w_k^* y_k} \geq p\% \ , \tag{3.1}$$

where p is a predetermined cut-off value and w_k^* is given in (2.5). An unweighted version of (3.1) can also be used if the interest is in identifying observations that are influential uniquely because of their value. Rule (3.1) can be useful to detect influential observations for positive variables of interest. For a variable y that can take negative values, it may be preferable to take the absolute value before applying the rule.

Rule (3.1) is usually applied within a large domain such as province. This may cause a problem for smaller domains since an uninfluential observation in a large domain may well be influential in a smaller domain. Ideally, this rule should thus be applied within each domain of interest known in advance. However, it could result in the identification of too many influential observations for larger domains, and therefore large biases if all these identified observations are treated.

3.2 Treatment of influential observations

In this subsection, methods for treating influential observations identified in the detection step are described. For all these methods, estimates of totals or means are obtained by using (2.4) or (2.6) with a modification of either the value of the variable of interest y or the estimation weight of the identified observations. If the estimation weights are modified then the calibration equation is not satisfied anymore. To avoid this inconvenience, a new set of estimation weights can be obtained using (2.5) and replacing the design weights by the modified weights.

The first method considered is *Winsorization*. This method consists of replacing the y-value of the I largest observations by the y-value of the $(I + 1)^{th}$ largest, where I is the number of influential observations detected (using an unweighted version of rule 3.1, for example). This technique can be useful if influential observations mainly occur because of large values of the variables of interest and not because of large estimation weights.

If the estimation weights have a great impact on the estimates, the *weighted Winsorization* technique, described in Tambay (1988), may be more appropriate. It consists of replacing y_k for the I influential observations (detected using rule 3.1) by C/w_k^*, where C is the value of the $(I + 1)^{th}$ largest weighted y-value. Tambay (1988) shows with an empirical study the slight superiority of the method of Dalén (1987) over the weighted Winsorization technique. The Dalén's method consists of replacing y_k for

the I influential observations by $C/w_k^* + \left(y_k - C/w_k^*\right)/w_k^*$. In fact, when the estimation weights are large, both methods will generally produce similar estimates. In the simulation study described in section 5, these two methods have been tested and they have given very similar results. We therefore report only the results of one of them, namely the weighted Winsorization technique.

The methods that have been discussed so far in this subsection are based on a modification of the values of the variable of interest. They can be applied separately for each variable of interest (as in section 5). They could also be transformed such that we modify the estimation weight instead of the y-values. However, this would require a different weight for each variable of interest, which is not very appealing to data users (especially users of a public use microdata file). To avoid the production of more than one weight, these methods can be applied to only one key variable or to some linear combination of the key variables. The resulting modified weight can then be used for every variable of interest.

A very simple method for reducing the effect of influential observations is the *unit weight reduction* method. It consists of giving a weight of one to all influential observations. To detect influential observations with rule (3.1), one key variable has to be chosen or some linear combination of the key variables. A multivariate outlier detection method could also be used. The important point here is that the detection step must be based on only one variable if a unique estimation weight per household is desired. This feature is extremely desirable in practice, especially if a public use microdata file is produced.

The unit weight reduction method can be too drastic in practice, especially if most of the estimation weights are large. Survey methodologists usually prefer methods modifying the estimation weight of influential observations such that it is between one and its original value. One such interesting method is used by a few household surveys at Statistics Canada and is described in Tremblay (1998). This method requires knowing population totals for a certain number of categories, say n_{cat}, of an auxiliary variable.

Note that this auxiliary variable may also be used to detect influential observations (as in section 5). The modified weights can be obtained through the following procedure:

1. Arrange the categories ($c = 1, 2, ..., n_{cat}$) in decreasing order of the auxiliary variable;
2. Start with $c = 1$ (the highest category of the auxiliary variable);
3. While there is no influential observation in category c and $c \leq n_{cat}$ do $c = c + 1$;
4. If $c > n_{cat}$ then go to 8;
5. Modify the estimation weights of influential observations in category c by an adjustment factor such that the sum of the modified weights over all units of categories higher than or equal to c (1, 2,..., c) equals the sum of the known population totals over categories higher than or equal to c;
6. If the modified weight of an influential observation is less than one then this observation is assigned a weight of one and if it is greater than its original weight then this observation keeps its original weight;
7. Do $c = c + 1$ and return to 3;
8. End of the procedure.

This method is very similar to a poststratification with the constraints that the final weight be between one and the original weight, and that only the weight of influential observations be modified. The only difference is that the adjustment factor within a category c is not necessarily calculated independently from the previous adjustment factors. In the following, this method will be called *constrained poststratification*.

4. OTHER ROBUST ESTIMATION METHODS

In this section, methods that do not need the detection step are considered. These methods avoid the arbitrariness introduced by choosing a cut-off value above which observations are treated. They have the major advantage of dealing naturally with

domains since each observation receives the same treatment. With detect-and-treat methods, an influential observation in a domain is not necessarily treated unless this observation is also influential in the larger domain used at the detection step.

Ghangurde (1989) and Pelletier and Rancourt (1998) have studied the determination of the variance structure of the random error ε_k. Although these approaches are different, the idea in both papers is to give a reduced weight to outliers (or in our case, influential observations) when estimating $\boldsymbol{\beta}$. In fact, this is also equivalent to giving a smaller value of a_k to influential observations. As mentioned in section 2, $\hat{t}_y^G$ is design consistent no matter how a_k is specified. However, it can be highly variable even when a_k is appropriately chosen. An alternative to $\hat{t}_y^G$ is the following composite estimator:

$$\hat{t}_y^C = \hat{\mathbf{B}}'\mathbf{t}_x + \theta\left(\hat{t}_y^{HT} - \hat{\mathbf{B}}'\hat{\mathbf{t}}_x^{HT}\right), \tag{4.1}$$

where θ is an unknown parameter to be estimated from the data or to be chosen subjectively and $\hat{\mathbf{B}}$ is given in (2.2). In fact, estimator (4.1) is a weighted average of the design consistent estimator $\hat{t}_y^G$ and the usually more stable (but not necessarily design consistent) synthetic estimator $\hat{t}_y^S$ ($\hat{t}_y^C = \theta\hat{t}_y^G + (1-\theta)\hat{t}_y^S$). So, $\hat{t}_y^C$ reduces to $\hat{t}_y^S$ when $\theta = 0$ and it reduces to $\hat{t}_y^G$ when $\theta = 1$. Therefore, a choice of θ between 0 and 1 will usually yield a compromise between the desire to have a low bias and the opposite desire to have a low variance. However, it should be noted that there is no compromise possible when $a_k = w_k$ and $c_k = \boldsymbol{\lambda}'\mathbf{x}_k$ since in that case we have seen in section 2 that $\hat{t}_y^C = \hat{t}_y^G = \hat{t}_y^S$. Lee (1991, 1995) also proposed an estimator with the same form as $\hat{t}_y^C$.

A nice property of $\hat{t}_y^C$ is that it can still be written as the weighted sum (2.4), but with the following estimation weights:

$$w_k^* = \theta w_k + \frac{a_k}{c_k}\mathbf{x}_k'\left(\sum_{k\in s}\frac{a_k}{c_k}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\mathbf{t}_x - \theta\hat{\mathbf{t}}_x^{\mathrm{HT}}\right). \qquad (4.2)$$

Estimator (2.6) can again be used if the interest is in the estimation of a population mean rather than in the estimation of a population total.

An important issue with the composite estimator (4.1) is the determination of a_k. The idea is to give a smaller value of a_k to influential observations, that is, those having a large design weight, or those having a large regression residual. Several options are possible to achieve this. For example, Pelletier and Rancourt (1998) used the Cook's distance to reduce the effect of extreme observations on estimates. In this paper, we study the following form for a_k:

$$a_k = w_k^{\alpha}\exp(-\delta z_k)\ , \qquad (4.3)$$

where α and δ are unknown non-negative parameters to be estimated from the data or to be chosen subjectively and z is a variable to be appropriately chosen. The exponential function is used to avoid having some $a_k \le 0$.

The variable z in (4.3) should be positive with large values associated to extreme observations since large z-values have a smaller value of a_k. The Cook's distance could be used. In this paper, we rather considered, as the variable z, the absolute value of the standardized regression residuals. To calculate the standardized regression residuals, an initial estimate $\hat{\mathbf{B}}^{(0)}$ of $\boldsymbol{\beta}$ is required, which can be obtained in replacing a_k by $a_k^{(0)} = w_k^{\alpha}$ in (2.2). The absolute value of the standardized regression residuals can then be given by

$$z_k^{(0)} = \frac{\left|e_k^{(0)}\right|}{\hat{\sigma}^{(0)}\sqrt{c_k}}\ ,$$

where $e_k^{(0)} = y_k - \hat{\mathbf{B}}'^{(0)}\mathbf{x}_k$ is the regression residual for unit k and

$$\hat{\sigma}^{(0)\,2} = \frac{\sum_{k\in s} a_k^{(0)}\, e_k^{(0)\,2}/c_k}{\sum_{k\in s} a_k^{(0)} - q}\ , \tag{4.4}$$

where q is the dimension of $\mathbf{x}_k$. Of course, it would be possible to do an iterative procedure calculating alternately a_k, $\hat{\mathbf{B}}$ and z_k until some convergence criterion is reached. The estimator $\hat{\mathbf{B}}$ of $\boldsymbol{\beta}$ obtained after convergence could be viewed as an M-estimator since it can be shown that it is the solution of the following system of equations:

$$\sum_{k\in s} w_k^{\alpha}\, \psi\left(\frac{y_k - \boldsymbol{\beta}'\mathbf{x}_k}{\sigma\sqrt{c_k}}\right)\frac{\mathbf{x}_k}{\sigma\sqrt{c_k}} = 0\ , \tag{4.5}$$

where $\psi(t) = t \times \exp(-\delta|t|)$. This function is known as a redescending ψ function. For positive values of t (and $\delta > 0$), the function $\psi(t)$ is approximately equal to t for small values of t, is increasing for $t < 1/\delta$ and is decreasing toward 0 after that point. The situation is reversed for negatives values of t.

The iterative procedure that has just been described to solve (4.5) is known as the iteratively reweighted least squares (IRLS) algorithm (Beaton and Tukey, 1974). Note that an estimating equation for the unknown parameter σ^2 is also required to solve (4.5). This estimating equation (which is not explicitly defined here) is solved simultaneously with (4.5). This can be seen from equation (4.4), which is recalculated at each iteration of the IRLS algorithm.

In fact, if z_k is any function of the regression residuals and if a sufficient number of iterations are completed then $\hat{\mathbf{B}}$ is an M-estimator for $\boldsymbol{\beta}$. M-estimators are usually obtained through an iterative procedure and this may be very time-consuming, especially if a replication technique, such as the jackknife or the bootstrap, is used for variance estimation. However, it has been empirically shown that a one-iteration procedure, starting with reasonable initial values, is often as good as the fully iterated procedure (Lee, 1991). For practical reasons, only one-iteration procedures have been considered in this paper. Therefore, estimation weights of the composite estimator are obtained in replacing a_k by $a_k^{(1)} = w_k^{\alpha} \exp(-\delta z_k^{(0)})$ in (4.2).

Gwet and Rivest (1992) considered a generalization of M-estimators called GM-estimators. These estimators can be used to reduce the influence of extreme values in the residuals as well as in the auxiliary variables. However, most auxiliary variables are categorical in household surveys and extreme values in the auxiliary variables are usually not a problem. This is the reason why GM-estimators have not been considered in this paper.

The parameter α in (4.3) and (4.5) controls the impact of the design weights on $\hat{\mathbf{B}}$. This parameter should normally be between 0 and 1. The closer to zero is α, the smaller is the impact of the design weights on $\hat{\mathbf{B}}$ (and inversely). If influential observations are not due to the presence of large design weights, then α should be close to 1. The extreme case of $\alpha = 1$ is usually preferred by design-based survey statisticians. In the other extreme case ($\alpha = 0$), the design weights are not involved in the estimation of $\boldsymbol{\beta}$. This is the case that model-based survey statisticians usually prefer. A value of α between these two extreme cases can be viewed as an interesting compromise for both types of statisticians. The parameter δ controls the impact of the variable z on $\hat{\mathbf{B}}$ and must be greater than or equal to 0. If there are no or few extreme observations in the sample, then δ should be close to 0 and it should be larger when there are many extreme observations in the sample.

The parameters θ, α and δ can be estimated from the data for each variable and domain of interest. However, if these parameters depend on the variable and the domain of interest, then having a unique estimation weight per household becomes impossible. We therefore suggest to determine a compromise value for each of these parameters based on several variables and domains, and then use these compromise values for all variables and domains of interest. This will yield a unique estimation weight per household. Similarly, it is also suggested to define only one variable z to be used for every variable and domain of interest.

Let us now find the conditions under which the design bias of $\hat{t}_y^C$ should be small. To achieve this, it is interesting to evaluate the anticipated bias of $\hat{t}_y^C$, $\mathrm{E}_m\mathrm{E}_p\left(\hat{t}_y^C - t_y\right)$, where the subscript m refers to the model and the subscript p refers to the sampling mechanism. The order of the expectations can first be changed. Then, if the sampling mechanism does not depend on the error term of the model (which is basically equivalent to say that the sampling mechanism is ignorable), the anticipated bias can be written as: $\mathrm{E}_p\mathrm{E}_m\left(\hat{t}_y^C - t_y\right)$. It is also easily shown that $\mathrm{E}_m\left(\hat{t}_y^C - t_y\right) = 0$ if $\mathrm{E}_m\left(\hat{\mathbf{B}}\right) = \boldsymbol{\beta}$. The last condition will be true when a_k is not correlated with ε_k. Therefore, the anticipated bias $\mathrm{E}_m\mathrm{E}_p\left(\hat{t}_y^C - t_y\right)$ will be equal to 0 if the following three conditions are satisfied:

(c.1) The relation between the variable of interest y and the auxiliary variables $\boldsymbol{x}$ is well specified by model m;

(c.2) The sampling mechanism does not depend on the error term of model m;

(c.3) a_k is not correlated with ε_k.

Condition (c.1) does not mean that the model is perfect, that every possible explanatory variable has been included into it or that it has a good predictive power but only that $\mathrm{E}_m\left(y_k\right)$ is well specified. In other words, if the true (superpopulation) relationship between y and $\boldsymbol{x}$ is linear, then the linear model m is well specified. For instance, if the auxiliary variables can be expressed as binary variables each associated to a different

group of units (or poststratum) then a linear model is appropriate and condition (c.1) is respected even if the predictive power of the model is poor. For auxiliary variables that are treated as continuous, the relation between y and $\boldsymbol{x}$ should be examined before choosing a model. To verify if condition (c.1) is respected, the plot of the regression residuals versus the predicted values may be useful. Any trend in this plot may indicate that $\mathrm{E}_m(y_k)$ is not well specified.

The sampling design generally involves design variables, such as stratum indicators, used to improve the efficiency of estimates produced by the survey. Condition (c.2) will be respected if these design variables are not correlated with the error term of the chosen model. In other words, condition (c.2) will be respected if the design variables add nothing more to the chosen model, which often seems to be the case in household surveys due to the low amount of useful auxiliary information available at the design stage. If this condition is not satisfied then the design variables could be incorporated into the model in order to make the condition true. A simple way to evaluate the validity of condition (c.2) is to plot the regression residuals versus the design weights or the selection probabilities. When condition (c.2) is valid, the design weights are not correlated with the errors and this plot should not show any particular trend.

Condition (c.3) is not respected if a_k is defined as in (4.3), $\delta \neq 0$ and z_k is a function of the regression residuals. Therefore, unless θ is very close to 1, it is important to choose a δ close to 0 if a low bias is desired. However, if $\delta = 0$ and condition (c.2) is respected then condition (c.3) is automatically respected. To verify the validity of this condition, a plot of the regression residuals versus a_k may be useful.

It should be noted that the design bias is approximately 0 for large populations if the anticipated bias is 0 (or, at least, approximately 0). Of course, the conditions for which the anticipated bias is 0 (conditions c.1, c.2 and c.3) have to be satisfied. This is so because, under the model, the variance of the design bias, $\mathrm{V}_m\mathrm{E}_p\left(\hat{t}_y^C - t_y\right)$, should be

small for large populations and, therefore, the design bias should be close to its model expectation, which is 0.

For large samples, it is also easily shown that the design bias of $\hat{t}_y^C$ can be approximated by:

$$\mathrm{E}_p\left(\hat{t}_y^C - t_y\right) \approx (\theta - 1)\left(t_y - \mathrm{E}_p\left(\hat{\mathbf{B}}'\right)\mathbf{t}_\mathbf{x}\right),$$

where

$$\mathrm{E}_p\left(\hat{\mathbf{B}}\right) \approx \left(\sum_{k\in U} \frac{a_k}{w_k c_k}\mathbf{x}_\mathbf{k}\mathbf{x}_\mathbf{k}'\right)^{-1} \sum_{k\in U} \frac{a_k}{w_k c_k}\mathbf{x}_\mathbf{k}\, y_k \ .$$

Therefore, the design bias of $\hat{t}_y^C$ is approximately 0 for large samples when $\theta = 1$ or when $a_k = w_k$ and $c_k = \boldsymbol{\lambda}'\mathbf{x}_\mathbf{k}$. In other cases, it is straightforward to show that the anticipated bias of $\hat{t}_y^C$ is approximately 0 if condition (c.1) is respected as well as the following condition:

(c.4) a_k / w_k is not correlated with ε_k.

Therefore, only conditions (c.1) and (c.4) are necessary to show that the anticipated bias is approximately 0 for large samples. When $a_k = w_k$, it is clear that condition (c.4) is respected. In fact, if $\delta = 0$ and the design weights are not correlated with the errors of the model then condition (c.4) is respected. Again, it is not satisfied when $\delta \neq 0$. To verify this condition, a plot of the regression residuals versus a_k / w_k may be useful.

Under strictly model-based considerations and when conditions (c.1), (c.2) and (c.3) are satisfied, the choice $\theta = 0$ (synthetic estimator) and $a_k = 1$ ($\alpha = \delta = 0$) minimizes the model variance of $\hat{t}_y^C$, $\mathrm{V}_m\left(\hat{t}_y^C\right)$. Also, if these three conditions are satisfied, the anticipated bias is 0 with this choice and the design bias should be negligible for large

populations. Therefore, this choice is very attractive since it yields a reduction in variance without increasing the bias significantly. However, for domain estimation, the anticipated bias is not necessarily 0 (even if conditions c.1, c.2 and c.3 are respected) unless population totals associated with the auxiliary variables are known at the domain level. Because it is usually impossible to have auxiliary information at the domain level for all domains of interest, it is thus suggested to choose $\theta = 0$ with $a_k = w_k^{\alpha}$, where $0 < \alpha < 1$. When $c_k = \boldsymbol{\lambda}' \mathbf{x}_k$, this choice is a compromise between the model-based estimator obtained with $\alpha = 0$ and the design consistent estimator obtained with $\alpha = 1$. It also has the advantage of requiring only one parameter (α) to determine. The following simulation study shows that this compromise gives good results even for domain estimation and that the design bias remains relatively low.

5. SIMULATION STUDY

In order to compare different estimators in the presence of influential observations, we performed a simulation study. We used data from the Statistics Canada's 1998 Survey of Household Spendings (SHS) to serve as the population. The survey had a stratified multi-stage design and contains information about 15,457 households on several variables. We also looked at a domain of interest from the population, namely: the domain of households with size equal to 1. We looked at four key variables from the survey, namely: *Income*, *Total Expenditure*, *Food* and *Renovation/Repair*. The last one was considered for its potential of having extreme values. Table 1 gives the summary statistics (in dollars) for the chosen variables.

Table 1: Summary Statistics for Key Variables

Variable	Mean	Standard Deviation	Skewness	Domain Mean
Income	48102.70	39362.62	4.06	25263.73
Total Expenditure	47341.23	35296.93	3.25	26207.00
Food	5607.69	3285.32	1.47	3018.29
Renovation/Repair	367.24	1124.28	12.60	188.14

From the population of households, 1000 samples of expected sample size 300 were selected using Poisson sampling. In Poisson sampling the sample size is a random variable with expected value equal to the sum of inclusion probabilities over the population and each unit is selected independently. We wanted to give households quite dispersed probabilities of selection, that would result into diverse weights. We assigned probabilities of inclusion for households such that they were proportional to the inverse of the SHS design weights (which include a nonresponse adjustment factor). The inclusion probabilities were calculated as

$$\pi_k = \pi_k^* \left(\frac{300}{\sum_{k \in U} \pi_k^*} \right),$$

where π_k^*, $k = 1, 2, \ldots, 15{,}457$, is the reciprocal of the design weight (including a nonresponse adjustment factor) from the SHS data.

As discussed earlier, we had different estimators to compare. We assumed that only one population total was known, which was the total number of households in the population ($x_k = c_k = 1$, for all households k). Table 2 gives the names and notations for the detect-and-treat estimators considered.

Table 2: Detect-and-treat Estimators and Respective Notations

Estimator	Notation
1. Winsorization	WIN
2. Weighted Winsorization	WWIN
3. Unit Weight Reduction	UWR
4. Constrained Poststratification	CP

We used two versions of rule (3.1) to detect the influential observations. The weighted version was used in WWIN, UWR and CP, while the unweighted version was used for WIN. The estimator UWR used Total Expenditure as the key variable for detection of influential observations because it is one of the most important variables of the SHS. For

CP, Income was used as auxiliary variable (with 12 categories) and for the detection step, in order to have a strategy similar to what is used in practice for the SHS. For WIN and WWIN, the detection step was carried out separately for each variable of interest. We specified a cut-off value at 3% level. The choice of the cut-off point was purely subjective. We looked at various values and picked the one that resulted in neither too few nor too many influential observations.

For the composite estimator, we used various combinations of θ, α and δ values. Table 3 gives the combinations we used in our simulation study as well as the corresponding notation. Note that, when $\delta = 0.5$, the variable z is chosen as the absolute value of the standardized regression residuals obtained in using the key variable Total Expenditure. This variable z is used for every of the four variables of interest. Also, only one iteration is performed to obtain an M-estimate for $\boldsymbol{\beta}$ (when $\delta = 0.5$).

Table 3: θ, α and δ Values and Notation for the Composite Estimator

Notation	θ	α	δ
GREG-M	1	0	0
GREGM-M	1	0	0.5
GREG-C	1	0.5	0
GREGM-C	1	0.5	0.5
GREG-D	1	1	0
GREGM-D	1	1	0.5
SYN-M	0	0	0
SYNM-M	0	0	0.5
SYN-C	0	0.5	0
SYNM-C	0	0.5	0.5
SYN-D	0	1	0
SYNM-D	0	1	0.5

In this simulation study, we subjectively chose predetermined combinations for θ, α and δ rather than using any optimality criterion. Other combinations might produce better results. We also tried the case $\theta = 0.5$ although the results are not shown in section 6. The results for that case were generally in between the cases $\theta = 0$ and $\theta = 1$.

For each of the 1000 samples of expected size 300, the population mean estimates and population domain mean estimates for each key variable were calculated. The estimated relative bias (RB), expressed as a percentage of population mean, was calculated using the formula

$$\text{est(RB)} = \left[\left(\frac{1}{1000}\right)\sum_{i=1}^{1000}(\hat{\mu}_i - \mu)\right]\frac{1}{\mu}\times 100\%,$$

where $\hat{\mu}_i$ is the population (or domain) mean estimate for the i^{th} sample, and μ is the population (or domain) mean. An estimate of the relative root mean squared error (RRMSE) expressed as a percentage can be calculated as:

$$\text{est(RRMSE)} = \sqrt{\frac{\sum_{i=1}^{1000}(\hat{\mu}_i - \mu)^2}{1000}}\times\left(\frac{1}{\mu}\right)\times 100\%.$$

An estimate of the relative standard error can be calculated by the relationship: $\text{est(RSE)} = \sqrt{\text{RRMSE}^2 - \text{RB}^2}$.

We also looked at the distribution of the number of detected influential observations in the 1000 samples for all four variables. On average, for the Income and Total Expenditure variables, respectively 4.1 and 4.2 influential observations per sample were detected using the weighted version of rule (3.1). The Renovation/Repair variable was on one extreme with 7.0 influential observations on average detected per sample. The Food variable had the lowest number of influential observations detected on average; only 3.7 per sample. For the unweighted version of rule (3.1), almost the same type of pattern with smaller values was observed, except for the Renovation/Repair variable, for which the unweighted version had about the same average number of influential observations detected than its weighted version. The smallest average was for Food (at zero) and the largest average was for Renovation/Repair (at 7.0).

These results are consistent with summary statistics presented in table 1. Renovation/Repair was the most skewed and was the most receptive to both versions of the detection method. The Food variable was the least skewed and hence no extreme observation was detected (for all samples) when weights were not taken into account in rule (3.1). Income and Total Expenditure were moderately skewed and had moderate number of extreme observations detected on average. The unweighted version (with the cut-off value of 3%) resulted in far fewer influential observations than the weighted version. This phenomenon was consistent with the intuition that the weights play a major role in making an observation influential in household surveys.

6. RESULTS

In this section, we present the results of the simulation study. We discuss the results for the estimation of the population mean first, which are shown in table 4. A special attention should be given on the Total Expenditure variable. This variable has been used to detect influential observations for the UWR estimator and has also been used to form the variable z for the GREG and synthetic estimators.

For the Renovation/Repair variable, all estimators have larger RRMSE and RSE values than the other variables. This is not surprising since it is the most skewed variable. It is interesting to note that the different estimators did not perform consistently across variables. Also, if a particular estimator resulted in a low relative bias, it often had a higher relative variance. Overall, among the detect-and-treat estimators, CP performed well for all variables. The synthetic estimator with certain parametric combinations for α and δ represents a potential alternative.

It is evident from table 4 that all GREG estimators perform similarly. They all have low RBs and relatively large RRMSEs and RSEs. These estimators provide a basis for comparison since they are asymptotically unbiased and design consistent, but volatile in the presence of influential observations. Considering the detect-and-treat estimators,

UWR results in large bias, especially for the Income and Total Expenditure variables, as it gives unit weight to influential observations. The GREG estimators and WIN are performing similarly for the Income, Total Expenditure, and Food variables since very few influential observations were detected using the unweighted version of rule (3.1). For the Renovation/Repair variable, WIN has a much larger RB but relatively low RRMSE and RSE as compared to the GREG estimators. For all four variables, WWIN has a large RB and RRMSE, but a low RSE. The distinction between WIN and WWIN is obvious from table 4. The WWIN estimator uses the weighted version of rule (3.1) and resulted in far more detected influential observations than for WIN (which uses the unweighted version), hence the RB and RRMSE of WWIN are higher than those of WIN for all four variables. Among the detect-and-treat estimators and for the Income, Total Expenditure and Food variables, CP stands out as the most reasonable with respect to the RB, RRMSE and RSE criteria, between two extremes of GREG and WIN, and, WWIN and UWR. For Renovation/Repair, the results are not that straightforward. Overall, CP always performs well with respect to RB and RRMSE. This is not surprising since CP uses more auxiliary information (population totals are known for 12 categories of Income).

Table 4: Results from the Simulation Study: Estimation of the Population Mean

	Income			Total Expenditure			Food			Renovation/Repair		
Estimator	RB	RRMSE	RSE	RB	RRMSE	RSE	RB	RRMSE	RSE	RB	RRMSE	RSE
Detect-and-Treat Estimators												
WIN	-0.02	9.06	9.06	0.13	8.21	8.21	0.22	6.16	6.16	-17.10	27.31	21.30
WWIN	-9.69	11.55	6.29	-8.64	10.58	6.11	-6.29	8.11	5.12	-38.47	40.81	13.62
UWR	-10.73	12.44	6.28	-10.55	12.04	5.81	-5.39	7.46	5.17	-11.19	27.36	24.96
CP	-3.87	7.28	6.16	-2.91	6.98	6.34	-1.03	5.83	5.74	-1.06	32.91	32.89
GREG Estimator ($\theta = 1$)												
GREG-M	0.36	9.30	9.30	0.32	8.28	8.27	0.34	6.31	6.31	2.80	39.04	38.94
GREGM-M	0.31	9.38	9.38	0.28	8.31	8.30	0.37	6.42	6.41	2.81	39.00	38.90
GREG-C	0.29	9.30	9.30	0.26	8.29	8.29	0.27	6.17	6.17	2.79	38.89	38.79
GREGM-C	0.27	9.40	9.40	0.24	8.33	8.32	0.28	6.18	6.17	2.79	38.87	38.77
GREG-D	0.23	9.29	9.29	0.23	8.31	8.31	0.22	6.16	6.16	2.77	38.91	38.81
GREGM-D	0.24	9.35	9.34	0.22	8.32	8.31	0.22	6.17	6.16	2.77	38.91	38.82
Synthetic Estimator ($\theta = 0$)												
SYN-M	5.75	7.30	4.50	4.74	6.33	4.19	12.84	13.39	3.79	-10.20	17.87	14.67
SYNM-M	-4.11	5.73	4.00	-2.86	4.81	3.87	17.05	17.53	4.11	-6.62	18.04	16.78
SYN-C	0.21	5.56	5.56	-0.11	5.14	5.14	3.36	5.20	3.97	-3.26	22.55	22.31
SYNM-C	-5.34	7.21	4.85	-4.37	6.44	4.74	4.72	6.21	4.05	-1.42	24.28	24.24
SYN-D	0.23	9.29	9.29	0.23	8.31	8.31	0.22	6.16	6.16	2.77	38.91	38.81
SYNM-D	-2.20	8.67	8.39	-1.50	8.08	7.94	0.63	6.24	6.21	3.55	40.14	39.98

For the synthetic estimators, we have very interesting results. As mentioned in section 4, synthetic estimation has been proposed as a way to reduce variance (while keeping the bias reasonably low). However, the strictly model-based estimator, SYN-M, can be too much biased even if it performed very well with respect to RSE. On the other hand, the design consistent estimator SYN-D has a low RB, but relatively high RSE. As mentioned in section 2, SYN-D is in fact identical to GREG-D. It seems that SYN-C, which is a compromise between the model-based estimator (SYN-M) and the design consistent estimator (SYN-D), is an interesting alternative with respect to all criteria considered. It resulted in RB values for all variables that were less than 5%, even for domain estimation (see table 5). Although other estimators (for example, CP) had this property too, the RRMSEs and RSEs were higher than those obtained with SYN-C. In general, the RB of the synthetic estimators has a lower bound equal to that of the GREG estimators, and the RRMSE and RSE of the synthetic estimators have upper bounds equal to those of the GREG estimators. When $\delta = 0.5$, the RB was still in general reasonably close (but

slightly larger) to the case $\delta = 0$, which indicates that the value of this parameter was not set too high. For example, SYNM-C performed comparably to SYN-C for all four variables.

In this simulation study, the simple mean model, $y_k = \beta + \varepsilon_k$, has been used. Because all units have the same value of the auxiliary variable ($x_k = 1$, for all k), they have necessarily the same expectation under the model. Therefore, $E_m(y_k)$ is well specified by this simple mean model (given the available auxiliary information) and condition (c.1) is respected. As mentioned in section 4, when $\delta = 0$ and condition (c.2) is respected, then condition (c.3) is also respected. If all three conditions are satisfied then the design bias of the model-based estimator, SYN-M, should be small. However, the relatively large biases observed in SYN-M are an indication that condition (c.2) is not completely satisfied and that the design variables are correlated (to a certain extent) with the error term of the chosen model. Adding useful auxiliary variables to the model would certainly reduce the bias of SYN-M. However, the validity of condition (c.1) should then be verified, especially if continuous auxiliary variables are added. It is also interesting to note that the difference between SYN-M and the design consistent estimator, SYN-D, will diminish as the fit of the chosen model increases. If the fit of the model is perfect ($y_k = \hat{\mathbf{B}}'\mathbf{x}_k$, for all k), then SYN-M will exactly be equal to SYN-D.

The results shown in table 5 for the population domain mean were somewhat similar to those for the population mean. The GREG estimators and WIN are producing identical results for all variables except Renovation/Repair, where the GREG estimators have lower RB, while WIN has lower RRMSE and RSE values. An important difference between table 4 and table 5 was for the detect-and-treat estimators. The RRMSE and RSE values were consistently slightly lower than those obtained by the GREG estimators, and RB values were also close to the GREG estimators RBs. The reason for this phenomenon is that the influential observations were detected at the population level and carried through for the domain mean estimation. It is possible that very few or even no influential observation at all were detected in the domain of interest for a given variable,

but some or many were detected in the whole population. For the synthetic estimators, the RBs are slightly higher than in the preceding table. Again, SYN-C and SYNM-C are interesting alternatives with respect to all criteria considered.

Table 5: Results from the Simulation Study: Estimation of the Population Domain Mean

	Income			Total Expenditure			Food			Renovation/Repair		
Estimator	RB	RRMSE	RSE	RB	RRMSE	RSE	RB	RRMSE	RSE	RB	RRMSE	RSE
Detect-and-Treat Estimators												
WIN	1.17	19.53	19.49	0.73	17.90	17.88	0.47	12.76	12.75	-13.90	59.36	57.71
WWIN	-2.40	14.93	14.73	-2.36	14.95	14.76	-0.73	11.24	11.22	-36.64	47.59	30.37
UWR	-4.55	15.34	14.65	-5.05	14.70	13.80	-1.26	11.67	11.60	-5.74	80.89	80.69
CP	-0.12	17.15	17.15	-0.10	16.88	16.88	0.28	12.54	12.53	0.52	94.00	94.00
GREG Estimator ($\theta = 1$)												
GREG-M	1.30	19.51	19.47	0.88	17.94	17.92	0.59	12.67	12.65	0.34	93.55	93.55
GREGM-M	1.33	19.53	19.48	0.89	17.88	17.85	0.61	12.69	12.67	0.35	93.56	93.56
GREG-C	1.20	19.46	19.42	0.80	17.87	17.86	0.51	12.68	12.67	0.44	93.67	93.66
GREGM-C	1.20	19.43	19.39	0.78	17.81	17.79	0.52	12.68	12.67	0.45	93.69	93.68
GREG-D	1.17	19.53	19.49	0.77	17.91	17.90	0.47	12.76	12.75	0.61	94.13	94.13
GREGM-D	1.16	19.50	19.46	0.76	17.88	17.87	0.47	12.77	12.76	0.62	94.15	94.15
Synthetic Estimator ($\theta = 0$)												
SYN-M	14.30	18.71	12.07	13.03	19.34	14.29	11.46	14.20	8.38	-12.44	39.02	36.98
SYNM-M	20.28	22.92	10.67	15.96	18.90	10.12	13.45	16.12	8.89	-9.84	40.24	39.02
SYN-C	3.31	13.46	13.04	2.93	13.77	13.46	2.89	9.37	8.91	-4.84	55.64	55.43
SYNM-C	5.04	12.68	11.64	3.28	11.87	11.40	3.60	9.81	9.13	-3.49	57.66	57.56
SYN-D	1.17	19.53	19.49	0.77	17.91	17.90	0.47	12.76	12.75	0.61	94.13	94.13
SYNM-D	1.66	18.54	18.47	0.97	17.17	17.15	0.69	12.88	12.86	1.16	95.65	95.65

7. CONCLUSION

In this paper, a number of estimators have been proposed and discussed to deal with the problem of influential observations, which occur because of extreme values, large design weights or the combination of both. All these estimators can take advantage of useful auxiliary information at the design stage and at the estimation stage. In fact, alternatives to the GREG estimator may not be needed when such useful auxiliary variables are available. However, in household surveys, useful auxiliary information is often not available, especially at the design stage. Therefore, large design weights are not

necessarily associated to small values of the variables of interest and more robust estimators are needed.

If the presence of influential observations can be justified, at least in part, by the presence of large design weights, then a model-based estimator may be useful under some conditions. In this paper, a compromise (SYN-C) between a strictly model-based estimator (SYN-M) and the design consistent estimator (SYN-D, which is identical to GREG-D), has been shown to be very attractive through a simulation study using real life survey data. Finally, whatever estimator is chosen, it is always a good idea to verify empirically the relationships between the variables of interest and the auxiliary variables used at the estimation stage, even if a design consistent estimator is preferred.

ACKNOWLEDGEMENTS

The authors would like to thank Zdenek Patak, Eric Rancourt and Johanne Tremblay of Statistics Canada for their useful comments.

REFERENCES

Beaton, A.E., and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-185.

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

Dalén, J. (1987). Practical estimators of a population total wich reduces the impact of large observations. R & D report, Stockholm, Statistics Sweden.

Ganghurde, P.D. (1989). Outliers in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 736-739.

Gwet, J.-P., and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

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

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

Lee, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the Annual Research Conference*, Washington, DC, U.S. Bureau of the Census, 178-202.

Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., and Kott, P.S. (editors), Chapter 26, New-York, John Wiley & Sons, Inc.

Pelletier, E., and Rancourt, E. (1998). Spécification du paramètre de la structure de variance du modèle dans l'estimateur de régression généralisé. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 159-164.

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

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 229-234.

Tremblay, J. (1998). Détection des observations influentes pour l'Enquête sur les finances des consommateurs (EFC) et l'Enquête sur la dynamique du travail et du revenu (EDTR). Internal report, Statistics Canada, Household Survey Method Division.