

A WEIGHT SMOOTHING METHOD FOR DEALING WITH STRATUM JUMPERS IN BUSINESS SURVEYS

Jean-François Beaumont and Louis-Paul Rivest¹

ABSTRACT

The problem of stratum jumpers in stratified business surveys is mainly due to inaccurate size information at the time of sampling, which may result in assigning a large design weight to a large size unit. Such units may be quite influential and substantially inflate the variance of the estimates. They can be viewed as being outliers with respect to the implicit model used at the sampling stage. Therefore, standard robust estimation techniques can be used to handle this problem. Instead, we will consider an alternative weight smoothing method and illustrate its application using data from the Canadian Workplace and Employee Survey.

KEY WORDS: Extreme weight; Model; Robust estimation; Smoothed weight.

RÉSUMÉ

Le problème des unités "sautées" de strate dans les enquêtes stratifiées auprès des entreprises est principalement dû à des informations imparfaites sur la taille de l'entreprise au moment de l'échantillonnage, ce qui peut conduire à assigner un grand poids de sondage à une unité de grande taille. De telles unités peuvent avoir une grande influence et augmenter significativement la variance des estimations. Elles peuvent être vues comme étant des données aberrantes par rapport au modèle implicite utilisé à l'étape d'échantillonnage. Par conséquent, les techniques usuelles d'estimation robuste peuvent être utilisées pour tenir compte de ce problème. Au lieu de ces techniques, nous considérerons une méthode alternative de lissage des poids et illustrerons son application au moyen des données de l'Enquête canadienne sur le milieu du travail et les employés.

MOTS CLÉS : Poids extrême; Modèle; Estimation robuste; Poids lissé.

1. INTRODUCTION

In many business surveys, the population of businesses is stratified by region, industry type and size group. The latter is defined using some measure of business size available on the sampling frame, such as the number of employees or the revenue of the business. Then, the sample is usually selected by stratified simple random sampling without replacement. For efficiency considerations, a large selection probability π_i (and thus a small design weight, $d_i = 1/\pi_i$) is usually assigned to a unit i of large size on the sampling frame while a small selection probability (and thus a large design weight) is assigned to a unit of small size. This strategy is justified on the grounds that business survey variables are usually highly skewed and that there is usually a positive correlation between the size measure and the main variables of interest so that large values of these variables are expected to be assigned to a small design weight.

At the time of collection, we often observe discrepancies between the information available at the design stage and the same information collected from the respondent. These discrepancies can be explained by errors on the sampling frame, which are partly due to outdated information, and the time lag between sampling and collection. They become problematic when a unit that is thought to be of small size at the design stage is actually found to be a large unit. Such units are sometimes called stratum jumpers because they would have been assigned to another stratum had the correct

¹ Jean-François Beaumont, Statistical Research and Innovation Division, Statistics Canada, 16th floor, R.H. Coats Building, Ottawa, Canada, K1A 0T6 (Jean-Francois.Beaumont@statcan.ca) and Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, Cité universitaire, Québec (Québec), Canada G1K 7P4 (lpr@mat.ulaval.ca).

information been available at the time of design. A consequence of this problem is that some units with large values of the variables of interest are unfortunately assigned large design weights, which may result in inefficient design-based estimators. It may actually occur that a few stratum jumpers account for an important percentage, say 20% to 30%, of the estimate of a population total.

At the design stage, the potential impact of stratum jumpers can be reduced to some extent by controlling the maximum design weight to be smaller than a certain threshold. This will usually imply departing from optimal stratification and/or allocation. Rivest (1999) proposed a method for dealing with stratum jumpers, which worked well empirically. It reduces the maximum design weight by a large factor. No matter how carefully the sampling design is chosen, it is likely that the problem will not be completely eliminated as the stratum jumpers occur in a haphazard way.

As an example, suppose that there are two design strata A and B. Stratum A has 9 selected units considered to be of large size at the design stage, which are assigned a design weight of 1, while stratum B has 41 selected units considered to be of small size at the design stage, which are assigned a design weight of 31. At the collection stage, we observe that one of the 41 units with a large weight is actually a large size unit so that the collection stratum is different from the design stratum for this unit, which is thus called a stratum jumper. Table 1 summarizes the above information. The stratum jumper is found in the middle row of this table.

Table 1. Example showing a stratum jumper

Collection stratum	Design stratum	Number of units	Design weight
A	A	9	1
A	B	1	31
B	B	40	31
Sum over the sample units		50	1280

Assume that the collection strata are homogeneous with respect to the variable of interest y (or at least more homogeneous than the design strata) so that collection stratum A contains large size units associated with large y -values while collection stratum B contains small size units associated with small y -values. On the one hand, the stratum jumper can be viewed as a unit with a large y -value compared to the other 40 units in the same design stratum, which all have the same design weight. Therefore, standard outlier-robust techniques, such as winsorization or M-estimation, could be used to handle this stratum jumper. On the other hand, the stratum jumper can also be viewed as a unit with a large design weight compared to the other 9 units in the same collection stratum although it may have a similar y -value. This is the view we take in Section 3 with our weight smoothing approach.

Before describing our weight smoothing approach, let us first briefly discuss an alternative approach to handling the potential problem of extreme design weights; namely, winsorizing the largest design weights (e.g., Potter, 1990; and Liu, Ferraro, Wilson and Brick, 2004). This actually seems to be the most popular method whenever something is done to deal with the problem of large design weights. Using this approach in the example given in Table 1 would lead to reducing the design weight of the stratum jumper but also of all other units in the same design stratum, which may be less appealing and may reduce efficiency. The main challenge with this method is to determine an appropriate winsorization cut-off. Several methods, sometimes more or less ad hoc, have been considered to address this issue, including the use of outlier detection techniques. One data-driven method is to choose the cut-off so as to minimize an estimate of the design Mean Squared Error (MSE). Unfortunately, this leads to a different cut-off for each variable of interest y , and thus a different winsorized weight for each y -variable. This is not convenient in multipurpose surveys and, thus, a compromise winsorization cut-off is needed. Also, it is worth noting that extreme design weights may not cause any problem if there is no stratum jumper and the design strata are homogeneous.

In Section 2, we briefly describe a weight smoothing approach, proposed by Beaumont (2008), to deal with the general problem of variable design weights, including the presence of extreme weights. This approach is adapted to handle the problem of stratum jumpers in stratified business surveys in Section 3. Unlike winsorization of design weights, a compromise smoothed weight is obtained naturally when there is more than one y -variable. In Section 4, the potential benefits of weight smoothing are illustrated using the data of the Canadian Workplace and Employee Survey (CWES). Finally, in Section 5, we conclude with a brief summary and discussion.

2. A GENERAL WEIGHT SMOOTHING APPROACH

We consider the problem of estimating the population total $T_y = \sum_{i \in U} y_i$ for a population U of size N . A sample s is selected from the population according to a probability sampling design $p(s|\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)'$ and \mathbf{z}_i is a vector of design variables (e.g., the size measure, region or industry) for the i^{th} population unit. Let us also use the notation $\mathbf{I} = (I_1, \dots, I_N)'$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$ and $\mathbf{Y} = (y_1, \dots, y_N)'$, where \mathbf{x}_i is a potential vector of auxiliary variables for the i^{th} population unit that may be used for calibration purposes and I_i is the sample inclusion indicator of unit i ; i.e. $I_i = 1$ if the i^{th} population unit is selected in the sample s and $I_i = 0$, otherwise. It should be kept in mind that the design weight $d_i = 1/\pi_i$ is a function of \mathbf{Z} only and should thus be written $d_i(\mathbf{Z})$. Nevertheless, we still denote it by d_i for convenience.

The basic idea underlying weight smoothing consists of first viewing the design weights as random and then using the model ξ :

$$E_{\xi}(d_i | \mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s), \quad (1)$$

for $i \in s$, where $g_s(\cdot, \cdot)$ is some function that may be sample-dependent and $\boldsymbol{\alpha}_s$ is a vector of unknown model parameters to be estimated from sample data. Specific models for the design weights are given in Pfeffermann and Sverchkov (1999) and Beaumont (2008). Note that nothing precludes y from being a vector so that the problem of finding a smoothed weight in multipurpose surveys boils down to considering more explanatory variables in model ξ . If $\tilde{d}_i = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s)$ were known, we would obtain the smoothed estimator \tilde{T}_y^{SDB} by replacing the design weights d_i by the smoothed weight \tilde{d}_i in a design-based estimator \hat{T}_y^{DB} , such as the Horvitz-Thompson estimator or a calibration estimator (Deville and Särndal, 1992). For instance, if \hat{T}_y^{DB} is the Horvitz-Thompson estimator, i.e., $\hat{T}_y^{DB} = \sum_{i \in s} d_i y_i$, then $\tilde{T}_y^{SDB} = \sum_{i \in s} \tilde{d}_i y_i$. The role of model ξ is to remove the unnecessary variability in the design weights. Beaumont (2008) showed that \tilde{T}_y^{SDB} is unbiased and never less efficient than the corresponding design-based estimator \hat{T}_y^{DB} under the model ξ and the sampling design.

Since $\tilde{d}_i = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s)$ is unknown, we consider a model-consistent estimator $\hat{\boldsymbol{\alpha}}_s$ of $\boldsymbol{\alpha}_s$ and use this estimator to obtain $\hat{d}_i = g_s(\mathbf{x}_i, y_i; \hat{\boldsymbol{\alpha}}_s)$, for $i \in s$. This leads to the smoothed estimator \hat{T}_y^{SDB} , which is obtained by using \hat{d}_i instead of d_i in the design-based estimator \hat{T}_y^{DB} . Note that classical model selection and validation techniques can be used to determine an appropriate model and to estimate $\boldsymbol{\alpha}_s$ since we are interested in estimating the relationship between the design weight variable d and both \mathbf{x} and y conditional on the realized sample and only for sample units. We expect that \hat{T}_y^{SDB} keeps the good properties of \tilde{T}_y^{SDB} in many practical applications provided that the underlying model (1) holds reasonably well. Indeed, Beaumont (2008) showed that if a linear model holds, we still have that the smoothed estimator \hat{T}_y^{SDB} is unbiased and never less efficient than the Horvitz-Thompson estimator under the model ξ and the sampling design.

3. WEIGHT SMOOTHING TO HANDLE STRATUM JUMPERS

Let us now consider the weight smoothing approach in the context of stratum jumpers in business surveys. We have already denoted by \mathbf{Z} , the matrix of design information available at the time of the design. Let us denote by \mathbf{Z}_{col} , the matrix of design information at the time of collection, which is assumed to be measured essentially without errors for sample units. We may hypothesize that, once we know \mathbf{Z}_{col} , \mathbf{X} and \mathbf{I} , the initial design matrix \mathbf{Z} brings no extra

information about \mathbf{Y} . In other words, \mathbf{Y} is independent of \mathbf{Z} after conditioning on \mathbf{Z}_{col} , \mathbf{X} and \mathbf{I} ; i.e., $F(\mathbf{Y} | \mathbf{Z}, \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I}) = F(\mathbf{Y} | \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I})$. This can also be rewritten as $F(\mathbf{Z} | \mathbf{Y}, \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I}) = F(\mathbf{Z} | \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I})$. The latter implies that the design weights are independent of \mathbf{Y} after conditioning on \mathbf{Z}_{col} , \mathbf{X} and \mathbf{I} and that a suitable model for the design weights would be

$$E_{\xi}(d_i | \mathbf{Z}_{col}, \mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{z}_{col,i}, \mathbf{x}_i; \boldsymbol{\alpha}_s), \quad (2)$$

where $\mathbf{z}_{col,i}$ is the vector of design variables at the collection stage for unit i . These design variables are treated here like additional y -variables. The idea here is to keep from the design weights the useful information contained in \mathbf{z}_{col} and \mathbf{x} (since it may have a strong relationship with \mathbf{y}) and remove their extra variability. Model (2) can then be used to construct a smoothed estimator, which should be more efficient than its corresponding design-based estimator.

In many business survey applications, model (2) often reduces to $E_{\xi}(d_i | \mathbf{Z}_{col}, \mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{z}_{col,i}; \boldsymbol{\alpha}_s)$. For instance, $\mathbf{X} = \mathbf{Z}_{col}$ in the example of Section 4. A simple approach to approximate the unknown function $g_s(\mathbf{z}_{col,i}; \boldsymbol{\alpha}_s)$, is then to discretize $\mathbf{z}_{col,i}$, for $i \in s$, into homogeneous categories called collection strata; this is considered in Section 4. Assuming that $g_s(\mathbf{z}_{col,i}; \boldsymbol{\alpha}_s)$ is constant within each category, we approximate the unknown model ξ by a simple analysis-of-variance model. The smoothed weight \hat{d}_i is simply obtained as the average of the design weights within the collection stratum containing unit i . The second-to-last column of Table 2 gives the smoothed weight when the above methodology is used in the example provided in Table 1.

Table 2. Smoothed weights in the example given in Table 1

Collection stratum	Design stratum	Number of units	Design weight	Smoothed weight	Smoothed weight (with constraint)
A	A	9	1	4	$1 \times 1.0215 = 1.02$
A	B	1	31	4	$4 \times 1.0215 = 4.09$
B	B	40	31	31	$31 \times 1.0215 = 31.67$
Sum over the sample units		50	1280	1280	$1253 \times 1.0215 = 1280$

For the stratum jumper, the smoothed weight is close to eight times smaller than the design weight. To compensate for this weight reduction, the smoothed weight of other units in collection stratum A became four times larger than the design weight. Since units with a small design weight may be associated to large y -values, it is perhaps preferable not to modify too much the weight of these units as they may become quite influential. One option, tested in Section 4, is to use the constraint that the smallest design weights are kept unchanged so that the 9 units with a design weight of 1 in Table 2 would also be given a smoothed weight of 1. It may then be necessary to adjust all the resulting smoothed weights by a constant factor so that the overall sum of the final smoothed weights is still equal to the overall sum of the design weights. This leads to the last column of Table 2, where the constant factor is $1280/1253 = 1.0215$. This strategy is equivalent to a hybrid approach between winsorization and weight smoothing, where the largest design weights are winsorized within each analysis of variance cell. Under this scheme, the winsorization cut-off is simple to compute as it is the average of the design weights within each collection stratum. Perhaps more sophisticated methods of finding the winsorization cut-off could yield better results. This has yet to be investigated.

A design-based MSE estimator is

$$\text{mse}(\hat{T}_y^{SDB}) = v(\hat{T}_y^{SDB}) + \max\left\{0, \left(\hat{T}_y^{SDB} - \hat{T}_y^{DB}\right)^2 - v(\hat{T}_y^{SDB} - \hat{T}_y^{DB})\right\}. \quad (3)$$

Since \hat{T}_y^{SDB} may have a complicated form, the bootstrap technique (e.g., Rao and Wu, 1988; and Rao, Wu and Yue, 1992) is a natural candidate for obtaining estimators $v(\hat{T}_y^{SDB})$ and $v(\hat{T}_y^{SDB} - \hat{T}_y^{DB})$ of the design variances $\text{Var}(\hat{T}_y^{SDB})$ and $\text{Var}(\hat{T}_y^{SDB} - \hat{T}_y^{DB})$ respectively. Note that the second term in the right-hand side of (3) is an estimator of the squared

design bias, which is restricted to be greater than or equal to zero. Its form is justified in Gwet and Rivest (1992). One can also restrict the estimated MSE not to be greater than $v(\hat{T}_y^{DB})$ since we expect gains in efficiency if model ξ holds. Such an MSE estimator performed well in the empirical study of Beaumont (2008).

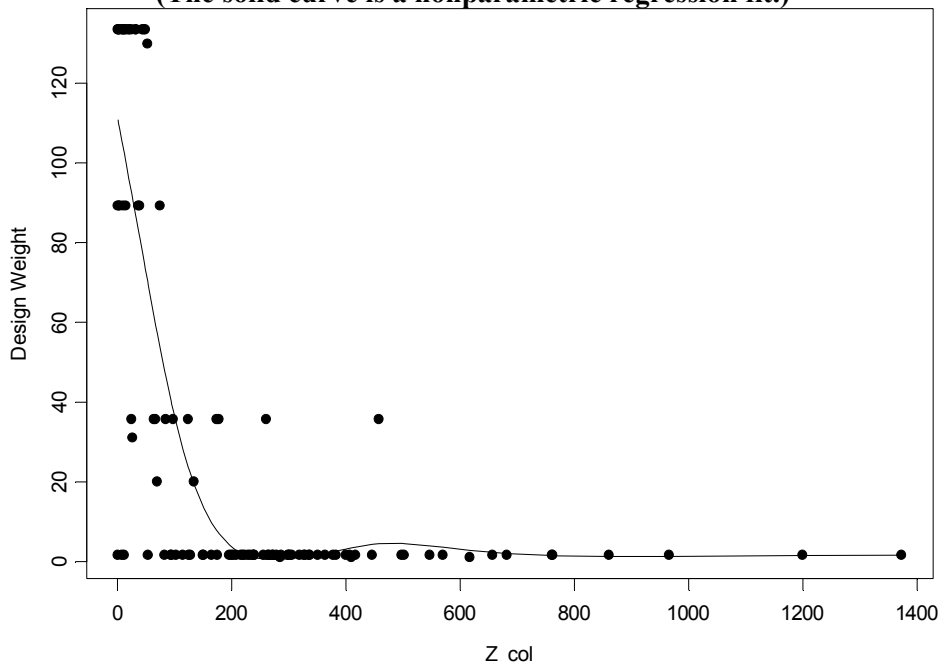
4. AN ILLUSTRATION USING THE CANADIAN WORKPLACE AND EMPLOYEE SURVEY

To illustrate the benefits of the weight smoothing method when handling stratum jumpers, we have used the 2003 CWES data. The CWES is a longitudinal survey that started in 1999 and that collects information on employers and their employees. In this application, we focus on the employer portion of the CWES and consider only a subset of its variables. They are five widely-used financial variables, which will be denoted by Y_1, Y_2, Y_3, Y_4 and Y_5 . Every other year, a sample from the population of births is selected from the Business Register. Therefore, the 2003 sample contains units selected in 1999, 2001 and 2003. In each of these years, employers are selected by stratified simple random sampling without replacement and the strata are formed by crossing six regions, fourteen industry groups and three size groups. The size variable corresponds to the number of employees available on the Business Register. To simplify the example, we restrict to a single region-industry group leading to a sample size of 112.

In this survey, there is a single auxiliary variable $x_i = z_{col,i}$, with $z_{col,i}$ being the number of employees obtained at the collection stage for employer i . The design-based estimator \hat{T}_y^{DB} is the ratio estimator $\hat{T}_y^{DB} = \sum_{i \in s} w_i y_i$, where $w_i = d_i T_x / \sum_{i \in s} d_i x_i$ and $T_x = \sum_{i \in U} x_i$. The population total T_x is obtained from a reliable external survey (the Survey of Employment Payroll and Hours) and will be assumed without error to simplify the example.

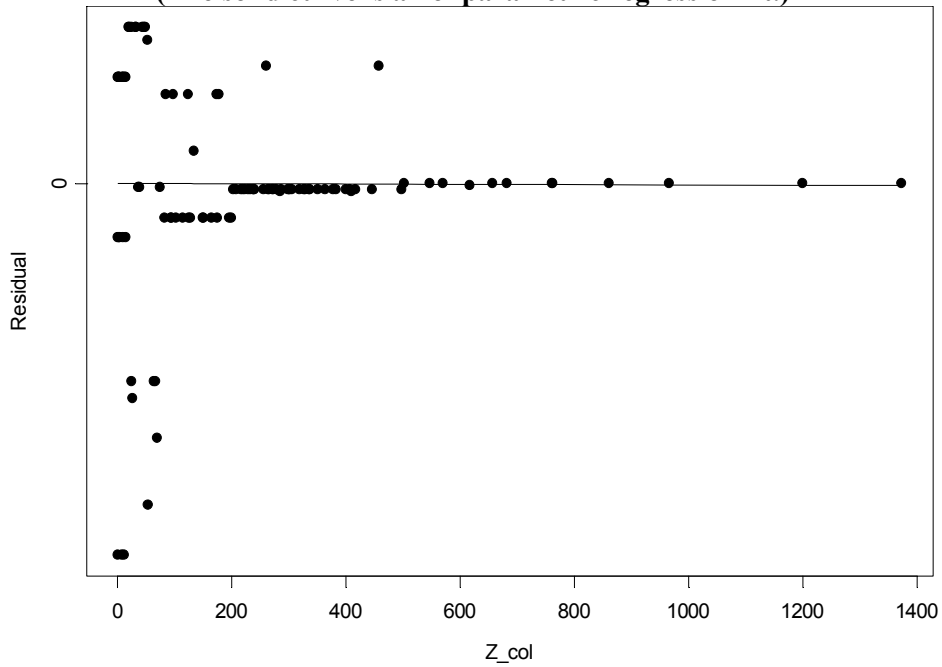
Figure 1 shows the relationship between the design weights d_i and $z_{col,i}$. The solid curve has been obtained using the procedure TPSPLINE of SAS. It is a nonparametric smoothing spline method based on penalized least squares estimation. We can see that there is a unit with a relatively large design weight of about 35 and a large value of z_{col} . Standard winsorization of the design weights may not reduce at all the weight of this unit, depending on the cut-off point. Weight smoothing will be more efficient by reducing the weight of this unit so that it has less influence on the estimates. Note also that the points are aligned along horizontal lines that represent the different strata.

**Figure 1. Plot of the design weights versus $z_{col,i}$.
(The solid curve is a nonparametric regression fit.)**

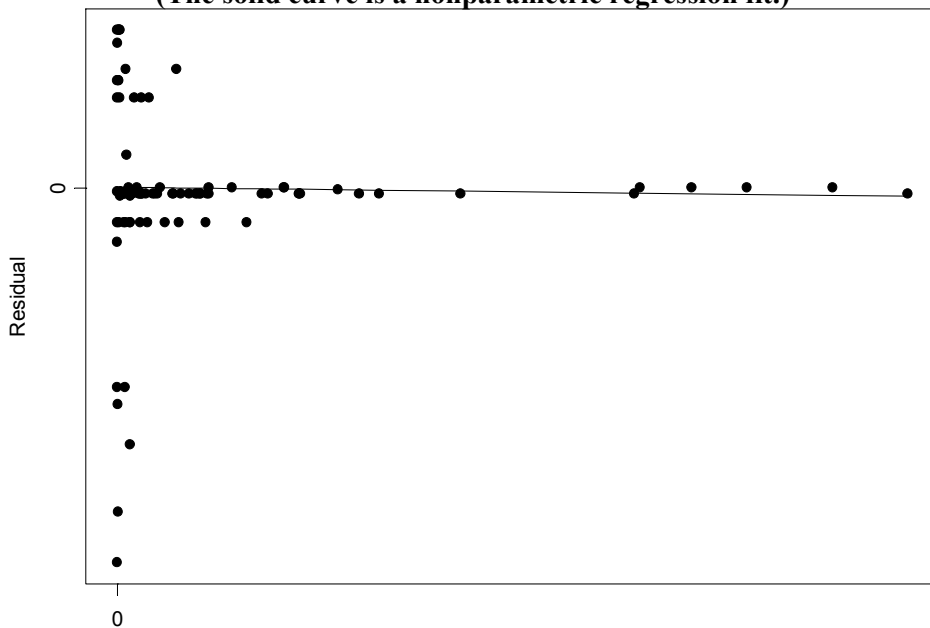


To smooth the weights we have used a one-way analysis of variance model with five categories (SR-5) obtained by discretizing z_{col} as proposed in Section 3. There are three categories for $z_{col} \leq 200$, as the slope of the smoothed spline curve is quite abrupt for small z_{col} and two categories for $z_{col} > 200$. The analysis of variance residuals are plotted against z_{col} and Y_4 in Figures 2 and 3. The smoothing splines in these figures do not show obvious trends in the residuals; thus this model for the design weights is satisfactory. Although they are not provided here, plots of the residuals against the other y -variables were similar. Thus, the assumption $F(\mathbf{Z} | \mathbf{Y}, \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I}) = F(\mathbf{Z} | \mathbf{Z}_{col}, \mathbf{X}, \mathbf{I})$ made in Section 3 is reasonable.

**Figure 2. Plot of the analysis-of-variance residuals versus $z_{col,i}$.
(The solid curve is a nonparametric regression fit.)**



**Figure 3. Plot of the analysis-of-variance residuals versus $Y_{4,i}$.
(The solid curve is a nonparametric regression fit.)**



In this empirical study, we have also considered a common mean model (SR-1) for the weights. Under SR-1, the smoothed weight of all the sample units is simply the average weight. The simulations also feature two estimators, WIN10 and WIN100, obtained by winsorizing the largest design weights with cut-off points of 10 and 100 respectively. Versions of SR-1 and SR-5 that left the design weights less than 2 ($d_i < 2$) unchanged, as described in the example shown in the last column of Table 2, were also calculated. All the smoothed weights were winsorized to insure that the final adjusted weight, say d_i^F , lies in the range $0.1d_i \leq d_i^F \leq 10d_i$. This step did not change the results in a significant way; it controlled the bias by preventing large weight adjustments, especially when the constraint on the smallest design weights was not used.

For each smoothed or winsorized estimator, \hat{T}_y^{*DB} say, associated to the design-based ratio estimator \hat{T}_y^{DB} , three quantities have been computed: i) the relative difference as a percentage, $RD = 100(\hat{T}_y^{*DB} - \hat{T}_y^{DB})/\hat{T}_y^{DB}$; ii) the Wald statistic, $Wald = (\hat{T}_y^{*DB} - \hat{T}_y^{DB})^2 / v_B(\hat{T}_y^{*DB} - \hat{T}_y^{DB})$, which can be used to give an indication of the design bias of the smoothed or winsorized estimators; and iii) the bootstrap relative efficiency as a percentage, $RE_Bootstrap = 100v_B(\hat{T}_y^{DB})/mse_B(\hat{T}_y^{*DB})$. The notation $v_B(\cdot)$ is used to indicate that the Rao-Wu bootstrap variance estimator has been chosen. The estimator $mse_B(\hat{T}_y^{RC})$ is obtained by using the Rao-Wu bootstrap method for the estimation of both variance terms involved in (3). One thousand bootstrap replicates have been used in this empirical study. A summary of the results is reported in Table 3.

Table 3: Comparison of smoothed and winsorized estimators using CWES data

Variable	Method	RD	Wald	RE_Bootstrap	RD	Wald	RE_Bootstrap
		Without constraint on the smallest design weights			With constraint on the smallest design weights		
Y_1	SR-1	8.88	0.59	434.9	11.96	3.68	56.6
	SR-5	-3.77	0.10	323.2	-10.31	0.81	238.4
	WIN10	10.08	1.90	143.8	10.08	1.90	143.8
	WIN100	2.34	1.99	92.3	2.34	1.99	92.3
Y_2	SR-1	10.10	0.62	448.9	14.19	4.05	53.6
	SR-5	-5.49	0.16	418.8	-12.98	0.98	318.7
	WIN10	11.70	2.14	135.2	11.70	2.14	135.2
	WIN100	2.53	1.78	92.8	2.53	1.78	92.8
Y_3	SR-1	18.45	10.82	11.3	10.12	20.02	26.5
	SR-5	2.22	0.32	155.4	-1.66	0.13	136.4
	WIN10	15.20	16.07	15.6	15.20	16.07	15.6
	WIN100	1.72	7.03	88.6	1.72	7.03	88.6
Y_4	SR-1	137.32	73.97	0.5	29.06	26.68	10.8
	SR-5	39.66	6.91	6.3	12.36	3.56	40.5
	WIN10	95.05	61.67	1.1	95.05	61.67	1.1
	WIN100	7.30	34.28	64.2	7.30	34.28	64.2
Y_5	SR-1	47.46	8.65	13.7	27.00	15.33	28.4
	SR-5	-13.01	1.26	211.0	-12.22	0.92	215.3
	WIN10	39.47	13.17	18.2	39.47	13.17	18.2
	WIN100	5.32	8.60	88.9	5.32	8.60	88.9

From Table 3, we can make the following remarks:

- For all variables but Y_4 , the SR-5 estimator was not significantly biased, according to the Wald statistic, and it was more efficient than the ratio estimator. Also, the constraint on the smallest design weights led generally to a small loss of efficiency.

- For variable Y_4 , the SR-5 estimator was significantly biased although less biased than its competitors according to the Wald statistic. This resulted in an inefficient estimator. The use of the constraint on the smallest design weights substantially reduced the bias and improved the efficiency.
- For variables Y_1 and Y_2 , the SR-1 estimator was the most efficient when no constraint on the smallest design weights were used, although only marginally more efficient than estimator SR-5. However, imposing this constraint resulted in a significant loss of efficiency and an increase in bias for these two variables. For the other three variables, the SR-1 estimator had a very large bias which made the estimator inefficient.
- Both winsorized estimators did not perform well. The WIN10 estimator was sometimes significantly biased while the WIN100 never led to gains in efficiency. We tried several other winsorization cut-offs but were not able to find any satisfactory compromise. Note also that the constraint on the smallest design weights had no effect on the winsorized estimators.
- Overall, the SR-5 estimator is the best. Also, the constraint on the smallest design weights seems to offer protection against bias at the expense of a slight loss of efficiency when the bias of the smoothed estimator is not significant.

Figures 2 and 3 suggest that the analysis-of-variance model used in this example is adequate. Using an argument similar to Beaumont (2008), the resulting smoothed estimator SR-5 should be asymptotically unbiased and more efficient than the ratio estimator, under the model and the sampling design, provided that this linear model holds. This is in agreement with results of Table 3, except for variable Y_4 . The bias for variable Y_4 may thus be explained either by a slight model misspecification that is difficult to detect by a graphical analysis or by an error of the Wald test since the Wald statistic is not that extreme. From a single sample, it is difficult to determine the exact cause of this bias. It is worth mentioning again that the constraint on the smallest design weights seems to bring some robustness against model misspecification and potential bias. Another alternative could have been to include Y_4 in the model, in addition to z_{col} , in order to reduce the impact of the possible model misspecification.

As a final comment on this example, note that the SR-1 estimator is equivalent to a model-based ratio estimator when no constraint on the design weights is used. Such a model-based estimator ignores completely the design weights and should work well if its underlying model explains satisfactorily the relationship between the y -variables and x . Apparently, this might have been the case for variables Y_1 and Y_2 . However, results in Table 3 also indicate that it may be risky in general to blindly use this estimator unless one is confident that its underlying model holds reasonably well.

5. SUMMARY AND DISCUSSION

We have adapted a weight smoothing method to deal with stratum jumpers in stratified business surveys. The method is simple to implement as the smoothed weights are simply obtained by taking the average of the design weights within appropriate collection strata. It has also been shown to be quite promising in our empirical investigation. If weight smoothing does not yield a sufficient increase in efficiency then nothing precludes, in principle, to combine weight smoothing with an outlier-robust method, such as winsorization or M-estimation. This remains to be investigated.

It is important to point out that finding an appropriate model is a key aspect of our method. To obtain some robustness against model failures, we partitioned the sample into collection strata in a more or less ad hoc way. Research into the issue of determining adequate collection strata boundaries could be useful. An alternative to determine collection strata, which remains to be investigated, would be to use nonparametric methods of estimating the smoothed weights.

ACKNOWLEDGEMENTS

We wish to thank Harold Mantel and Stéphane Tremblay from Statistics Canada for their useful comments and suggestions. We also wish to thank Cynthia Bocci from Statistics Canada for helping us out with the figures.

REFERENCES

Beaumont, J.-F. (2008). "A new approach to weighting and inference in sample surveys". Manuscript under review in *Biometrika*.

- Deville, J.-C., and Särndal, C.-E. (1992). "Calibration estimators in survey sampling". *Journal of the American Statistical Association*, **87**, 376-382.
- Gwet, J.-P., and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, **87**, 1174-1182.
- Liu, B., Ferraro, D., Wilson, E., and Brick, J.M. (2004). "Trimming extreme weights in household surveys". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia.
- Pfeffermann, D., and Sverchkov, M. (1999). "Parametric and semi-parametric estimation of regression models fitted to survey data". *Sankhya*, Series B, **61**, 166-186.
- Potter, F. (1990). "A study of procedures to identify and trim extreme sampling weights". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia, 225-230.
- Rao, J.N.K., and Wu, C.F.J. (1988). "Resampling inference with complex survey data". *Journal of the American Statistical Association*, **83**, 231-241.
- Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). "Some recent work on resampling methods for complex surveys". *Survey Methodology*, **18**, 209-217.
- Rivest, L.-P. (1999). "Stratum jumpers: Can we avoid them?". *Proceedings of the Survey Research Methods Section*, American Statistical Association, 64-72.