Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comparison of the conditional bias and Kokic and Bell methods for Poisson and stratified sampling

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Thomas Deroyon and Cyril Favre-Martinoz

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 309-338
Statistics Canada, Catalogue No. 12-001-X

Comparison of the conditional bias and Kokic and Bell methods for Poisson and stratified sampling

Thomas Deroyon and Cyril Favre-Martinoz[1]

Abstract

In business surveys, it is common to collect economic variables with highly skewed distribution. In this context, winsorization is frequently used to address the problem of influential values. In stratified simple random sampling, there are two methods for selecting the thresholds involved in winsorization. This article comprises two parts. The first reviews the notations and the concept of a winsorization estimator. The second part details the two methods and extends them to the case of Poisson sampling, and then compares them on simulated data sets and on the labour cost and structure of earnings survey carried out by INSEE.

Key Words: Robust estimation; Winsorized estimator; Influential values; Conditional bias.

1 Introduction

In survey statistics, a population unit is influential if the estimators produced on a sample drawn from that population change significantly depending on whether or not that unit is sampled. The concept of an influential unit therefore depends on several factors, which determine what Beaumont, Haziza and Ruiz-Gazen (2013) called a configuration:

- a sampling design for a population;
- one or more variables of interest and a parameter of interest on the distribution of this variable;
- an estimator calculated on the sample for this parameter of interest.

A unit may be influential in one configuration and not in another. For example, it can have a significant effect on the estimator of the total of a variable in a particular domain, but have only a minor influence on the estimator of the total of that variable in the total population.

Chambers (1986) distinguishes two types of influential units: non-representative atypical values are units that have provided erroneous information or are found in these exceptional situations. The information collected on these units cannot be extrapolated to the rest of the population; these units are usually identified during collection or during control of the data collected and processed *via* specific procedures (for answers considered to be erroneous, the information collected is, for example, replaced by a missing and imputed value. It can also be corrected by recontacting the unit in question. For units that are in an exceptional situation and for which we are sure the case is unique, it is common to put their weight to 1).

Representative influential units provided correct answers and are not a priori unique in the population. They are common in surveys of businesses, a population for which many variables have a very skewed distribution. In particular, the variables reflecting volumes or amounts (turnover, value added, payroll,

[1] Thomas Deroyon, INSEE, Paris, France; Cyril Favre-Martinoz, INSEE, Saint-Denis de la Réunion, France. E-mail: cyril.favre-martinoz@insee.fr.

investment, energy consumption, research and development expenditure and anti-pollution expenditure to name some of the key variables of INSEE business surveys) are characterized by a high concentration of low values, corresponding to many small businesses, and some very high values associated with large or very large businesses.

To limit the effects of this wide dispersion of the variables of interest in the population of businesses, the classical sampling design applied to them is a stratified design, in which size, measured in number of employees, is used as a stratification variable. In most cases, this makes it possible to assign businesses inclusion probabilities correlated with the amounts they reported in the survey. In these designs, large businesses are surveyed exhaustively, as are businesses that, according to the auxiliary information available in the sampling frames, are likely to report very large amounts in the survey, regardless of their size.

In practice, however, it is impossible to be entirely protected against influential observations at the sampling design stage. Indeed, the information in the sampling frames may be affected by measurement errors. For example, the number of employees in the sampling frames is a variable derived from returns to social security organizations that requires a significant amount of controls and adjustments and takes two years to reach a definitive value for a given year. It is thus possible, when drawing a sample, to use the last known definitive value, but which relates to a business's previous situation, or to use the nearest preliminary value, which will be affected by more measurement errors. In both cases, the variable used for stratification may not correspond to the actual situation of the business at the time of the survey, creating businesses sampled in the wrong stratum (called "strata jumpers"), whose sampling weight is much too high compared with their survey responses.

The auxiliary variables available to define the sampling designs may also be only weakly correlated with the survey themes. It is therefore difficult to identify businesses that are innovative or involved in research and development activities based solely on their industry, size, region of establishment, duration of existence or legal category. The same goes for the amounts invested in sustainable development (measured in France by the Antipol Survey conducted by INSEE: https://www.insee.fr/fr/metadonnees/source/s1232).

Surveys can also collect several weakly correlated variables of interest. The sampling design, which aims to achieve the highest possible precision for the survey's core variables of interest, may not be appropriate for other, less significant variables, e.g., the portion of turnover generated by online sales. In particular, some businesses that report atypical values for secondary variables of interest in the survey may not have been identified and placed in a comprehensive stratum.

Finally, many business surveys are conducted at regular intervals, most often every year, and aim to estimate both the annual levels of the main variables of interest and their evolution. To meet these two objectives, the sample surveyed in the non-exhaustive strata is not renewed in full each year, but a portion is retained. For example, the sample of business surveys on Information and Communication Technologies (ICT-E) is renewed by half each year; businesses sampled in a given year are surveyed two years in a row (see Demoly, Fizzala and Gros, 2014). In this case, the businesses retain the sampling weight with which they were initially sampled, which may no longer match their characteristics at the time of the survey, resulting in the appearance of “stratum jumpers” and potentially influential units.

Classical estimators in the presence of survey data (for example, the expansion estimator or the estimator adjusted for total non-response) have (virtually) no bias but can be very unstable in the presence of influential values. Robust estimation methods must then be implemented to limit their impact. The principle of these methods is to modify the estimation weights or the declared values by the influential units in order to make the estimators more stable, at the risk of biasing them. More precisely, the estimators to which these methods lead must have a mean square error significantly lower than that of classical expansion estimators in the presence of influential data, without losing too much efficiency in the absence of atypical values in the sample. The processing of influential values therefore lies in a compromise between bias and variance.

The most common method in practice for dealing with the problem of influential values is winsorization, which applies to estimating totals of variables of interest. For a given variable of interest, this consists of partitioning the sample and associating each part of the sample with a threshold; for example, in the case of a sample selected by stratified simple random sampling, the sample is cut according to the drawing strata, and a different threshold is associated with each stratum. Units in the sample for which the values of the variable are greater than the threshold associated with their part of the sample have their response or their weight decreased, while the responses and weights of the other units are not modified. There are two forms of winsorization in the literature, which differ depending on how the variable or weight is modified when the variable of interest exceeds the threshold. In standard winsorization, also known as Type I winsorization, values that exceed the threshold are truncated at the threshold. In this article, we will use the form proposed by Dalén (1987) and Tambay (1988), also called Type II winsorization, because it ensures that winsorized weights greater than 1 are obtained. This method will be briefly reviewed in Section 2.

In the application of winsorization, the choice of thresholds is crucial; a bad choice can lead to winsorized estimators with a higher mean square error than the classical estimators via the introduction of a very high bias that is difficult to correct later. The choice of these thresholds has been the subject of numerous studies, including by Kokic and Bell (1994), Rivest and Hurtubise (1995) and Favre-Martinoz, Haziza and Beaumont (2015). In the case of a simple random stratified design without replacement, Kokic and Bell (1994) determined the theoretical formulas and algorithms for calculating the thresholds that realizations in the winsorized estimator with the lowest mean square error possible, under the hypothesis that the realizations of the variable of interest are identically distributed in each stratum, the mean square error being calculated under the sampling design and the law of the variable of interest. In the case of repeated surveys, they suggest using historical data collected in previous editions of the surveys to calculate these thresholds. Clark (1995) generalized the results of Kokic and Bell (1994) in the case of a ratio estimator by calculating the mean square error with respect to the model only.

Other methods have been proposed for identifying and processing influential units in survey statistics. One of these, introduced by Beaumont et al. (2013), is based on the concept of conditional bias, a measure of influence proposed by Moreno-Rebollo, Muñoz-Reyez and Muñoz-Pichardo (1999) and

Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero and Muñoz-Pichardo (2002). Unlike the winsorization methods mentioned above, which are only suitable for certain sampling designs and require fairly rich information outside the sample, the method proposed by Beaumont et al. (2013) can be applied a priori to any sampling design and uses only the survey responses. However, it does not necessarily lead to the processed estimator of influential units with the smallest mean square error, but to the estimator on which the influence of the most influential unit is the lowest in absolute value. Favre-Martinoz et al. (2015) and Favre-Martinoz, Haziza and Beaumont (2016) proposed adaptations of the conditional bias method for calculating winsorization thresholds and factoring in an additional sampling phase and estimation in domains.

The purpose of this paper is to compare the efficiency of the winsorization and conditional bias methods to treat influential values. In Section 2, we review the winsorization method and the calculation of winsorization thresholds proposed by Kokic and Bell in stratified simple random sampling. We also propose an extension of the Kokic and Bell method for a Poisson sampling design. After briefly reviewing the principles of robust estimation based on conditional bias in Section 3, we present in Section 4 simulations to compare the extension of the Kokic and Bell method with the conditional bias methods in the Poisson case. Finally, an example of the practical application of the Kokic and Bell method and its extension to the Poisson case is presented in Section 4, which compares them with a method based on conditional biases in the context of the labour cost and structure of earnings survey carried out by INSEE.

2 The processing of influential units by winsorization following the approach of Kokic and Bell

In this section, we present the method initially proposed by Kokic and Bell (1994), which applies to samples selected through stratified simple random sampling, and an extension of this method to the case of samples selected through Poisson sampling.

2.1 Case of stratified simple random sampling

Consider a finite is a population U of size N and a variable of interest X observed on a sample S of fixed size n and for which we are looking to estimate the total $T(X) = \sum_{i \in U} X_i$ on the population. The approach of Kokic and Bell (1994) is based on the following hypotheses:

- X is a positive or nil variable;
- S is selected according to a stratified simple random sampling design $P,$ following strata $U_h,\ h = 1, \ldots, H.$ In each stratum of size $N_h,$ a sample S_h of size n_h is selected according to a simple random design without replacement. The expectation with respect to the sampling design will be denoted E_p afterwards;
- in each stratum $U_h,$ the values of X in the population are derived from random variables X_{hi} that are independent and identically distributed according to a law $\mathcal{L}_h$ (or of the same model m)

with expectation μ_h. The expectation and the variance with respect to this model will be denoted E_m and V_m respectively hereafter;

- we have, for each stratum U_h, N_h realizations $\breve{X}_{hi}$ of the variable X derived from the same law $\mathcal{L}_h$ but independent of the sample S_h.

In this context, Kokic and Bell (1994) propose applying a Type II winsorization; they associate with each stratum U_h a threshold K_h independent of the sample S and define the winsorized variable $\tilde{X}$, for $i \in S$, by:

$$\tilde{X}_{hi} = \begin{cases} X_{hi} & \text{if } X_{hi} < K_h \\ \dfrac{n_h}{N_h} X_{hi} + \left(1 - \dfrac{n_h}{N_h}\right) K_h & \text{if } X_{hi} \geq K_h. \end{cases}$$

The winsorized estimator of the total X is then the expansion estimator of the total of the winsorized variable $\tilde{X}$: $\hat{T}\left(\tilde{X}\right) = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_h} \tilde{X}_{hi}$.

The thresholds K_h are determined so as to obtain the estimator $\hat{T}\left(\tilde{X}\right)$ with the lowest mean square error with respect to both the sampling design and the law of X in each stratum, i.e.,

$$(K_h^*)_{h=1,\ldots,H} \in \operatorname{Argmin}_{(K_h)_{h=1,\ldots,H}} E_m E_P \left\{\left[\hat{T}\left(\tilde{X}\right) - T(X)\right]^2\right\}.$$

The optimal thresholds must therefore protect the winsorized estimator on average over all possible samples in the population, and on average on the law of the variable of interest, i.e., on average over all the possible populations considering the law of X.

Kokic and Bell (1994) place themselves in an asymptotic framework by considering a set of populations, sampling designs and samples indexed by $\nu \in \mathbb{N}$ such as:

- $\forall \nu \in \mathbb{N}, \forall h = 1, \ldots, H, n_{h_\nu} > 1$;
- $N_\nu, n_\nu \underset{\nu \to +\infty}{\longrightarrow} +\infty$;
- $\exists \epsilon \in \left]0, 1/2\right[, \forall \nu \in \mathbb{N}, \forall h = 1, \ldots, H, \epsilon < \frac{n_{h_\nu}}{N_{h_\nu}} < 1 - \epsilon$;
- the number of strata H is fixed.

They also propose denoting $J_{hi} = \mathbb{I}(X_{hi} \geq K_h)$ the winsorization indicator. To reduce the notations, we will omit in the rest of the article the indicator ν as well as the indicator i in the expression of the expectations and variances E_m and V_m of the random variables and $X_{hi}\, J_{hi}$ under the law of X in the stratum h. Insofar as these variables are assumed to be independent and identically distributed in each stratum, $E_m(X_{hi})$ for example, is indeed the same, regardless of the observation considered.

In this context, Kokic and Bell (1994) show that, at the optimum and asymptotically, all the thresholds are linked to one another by the relation:

$$\left(\frac{N_h}{n_h} - 1\right)(K_h - \mu_h) \sim -B \qquad (2.1)$$

with $B = \sum_{h=1}^{H} N_h \left(1 - \frac{n_h}{N_h}\right)\left[K_h E_m\left(J_h\right) - E_m\left(X_h J_h\right)\right]$ the bias of the winsorized estimator. The notation $\sim$ corresponds to an asymptotic equivalence when n_ν tends toward infinity (which is equivalent to saying when ν tends toward infinity).

If we denote $X_{hi}^* = \left(\frac{N_h}{n_h} - 1\right)\left(X_{hi} - \mu_h\right)$ and $L = -B,$ then we can notice that at the optimum given (2.1), $J_{hi} = J_{hi}^* = \mathbb{I}\left(X_{hi}^* \geq L\right)$ and the bias B is the opposite of the zero-point of the function F defined by:

$$F(L) = L\left\{1 + \sum_{h=1}^{H} n_h E_m\left(J_h^*\right)\right\} - \sum_{h=1}^{H} n_h E_m\left(X_h^* J_h^*\right). \tag{2.2}$$

Determining the zero-point of the function F requires estimates of $\mu_h,\ E_m\left(J_h^*\right)$ and $E_m\left(X_h^* J_h^*\right).$ To do this, Kokic and Bell (1994) rely on observations of the variable X in each stratum. These observations must come from a source independent of the sample, since the demonstration of formulas (2.1) and (2.2) is based on the fact that the thresholds K_h are assumed to be independent of the sample S.

If we assume that for each stratum h we have p_h realizations $\breve{X}_{hi}$ of $X,$ then we can estimate F by:

$$\begin{aligned}\hat{F}(L) &= L\left\{1 + \sum_{h=1}^{H} n_h \frac{\sum_{i=1}^{p_h} \mathbb{I}\left(\breve{X}_{hi}^* \geq L\right)}{p_h}\right\} \\ &\quad - \sum_{h=1}^{H} n_h \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^* \mathbb{I}\left(\breve{X}_{hi}^* \geq L\right)}{p_h}\end{aligned} \tag{2.3}$$

with

$$\breve{X}_{hi}^* = \left(\frac{N_h}{n_h} - 1\right)\left(\breve{X}_{hi} - \frac{\sum_{j=1}^{p_h} \breve{X}_{hj}}{p_h}\right)$$

and estimate the optimal bias B as the opposite of the zero-point of $\hat{F}$.

Now, $\hat{F}$ is an increasing function and is linear by sections, which admits only one zero-point. This can be estimated simply by denoting $\breve{X}_{(i)}^*$ the values of $\breve{X}_{hi}^*$ sorted in ascending order and by calculating $\hat{F}\left(\breve{X}_{(1)}^*\right),\ \hat{F}\left(\breve{X}_{(2)}^*\right), \ldots$ until $\hat{F}$ sign changes.

Indeed, $\hat{F}\left(\breve{X}_{(1)}^*\right) = \breve{X}_{(1)}^* + \sum_{h=1}^{H} \frac{\sum_{i=1}^{p_h}\left(\breve{X}_{(1)}^* - \breve{X}_{hi}^*\right)}{p_h}$ is negative because $\breve{X}_{(1)}^*$ is by definition lower than all the others $\breve{X}_{hi}^*$ and because $\breve{X}_{(1)}^*$ is negative, since $\frac{\sum_{j=1}^{p_h} \breve{X}_{hj}^*}{p_h} = 0.$ However, $\hat{F}\left(\breve{X}_{(p)}^*\right) = \breve{X}_{(p)}^* \geq 0,$ for similar reasons by denoting $p = \sum_{h=1}^{H} p_h$.

By denoting j the indicator such as $\hat{F}\left(\breve{X}_{(j)}^*\right) \leq 0$ and $\hat{F}\left(\breve{X}_{(j+1)}^*\right) \geq 0,$ B can be estimated by linear interpolation, i.e., by

$$\hat{B} = -\frac{\breve{X}_{(j)}^* \hat{F}\left(\breve{X}_{(j)}^*\right) - \breve{X}_{(j+1)}^* \hat{F}\left(\breve{X}_{(j+1)}^*\right)}{\hat{F}\left(\breve{X}_{(j)}^*\right) - \hat{F}\left(\breve{X}_{(j+1)}^*\right)}. \tag{2.4}$$

2.2 Extension to the case of the Poisson sampling design

We now place ourselves in the situation in which the sampling design P by which S is selected is a Poisson sampling design, in which each unit i of the population can belong to the sample with a probability $\pi_i > 0.$ We are always interested in estimating the total in the population $T(X) = \sum_{i \in U} X_i$ of a variable $X.$ The extension of the Kokic and Bell method to this sampling design assumes:

- that X is a positive or nil variable;
- that it is possible to partition the population and the sample into subpopulations U_h and S_h in which all the values $d_{hi} X_{hi}$ are independent realizations from the same model verifying:

$$\forall h = 1, \ldots, H, \forall i \in U_h, d_{hi} \; X_{hi} = \mu_h + \epsilon_{hi}, \tag{2.5}$$

with

$$\begin{cases} E_m(\epsilon_{hi}) & = 0 \\ V_m(\epsilon_{hi}) & = \sigma_h^2 < +\infty \end{cases}$$

where E_m and V_m designates the expectation and variance with respect to the model (2.5).

In this context, we propose, as in the original method applied to stratified simple random sampling, associating a threshold $K_h, h = 1, \ldots, H$ with each part $S_h, h = 1, \ldots, H$ and defining:

- the winsorized variable $\tilde{X}$ by

$$\tilde{X}_{hi} = \begin{cases} X_{hi} & \text{if } \; d_{hi} X_{hi} \leq K_h \\ \dfrac{X_{hi}}{d_{hi}} + \left(1 - \dfrac{1}{d_{hi}}\right) \dfrac{K_h}{d_{hi}} & \text{if } \; d_{hi} X_{hi} > K_h, \end{cases} \tag{2.6}$$

where $d_{hi} = \frac{1}{\pi_i}$ is the weight of the unit i in part $h.$

- the winsorized estimator of the total X as the usual expansion estimator of the total $\tilde{X}$:

$$\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \sum_{i \in S_h} d_{hi} \; \tilde{X}_{hi}. \tag{2.7}$$

In the article by Kokic and Bell (1994), the subpopulations with which the thresholds are associated are the drawing strata, which respect two properties: the draws are independent between strata, and the authors postulate an identical population model for all observations in the same stratum. In the case of Poisson sampling, the drawings are by nature independent between units.

The strong hypothesis underlying model (2.5) is that values X_{hi} multiplied by weights d_{hi} are assumed to have constant expectation in each stratum. This means that the inclusion probabilities within each stratum are defined proportionally to the variable of interest $X.$ In practice, these inclusion probabilities are often defined proportionally to a known auxiliary variable that is strongly correlated with X, which makes it possible to be close to the hypothesis underlying model (2.5).

Note also that model (2.5) is the one under which the Horvitz-Thompson estimator is optimal in the sense of minimizing the mean square error with respect to the model.

In the following, the random variables $d_{hi} X_{hi}$ being assumed to be independent and identically distributed within each stratum, we will denote $Z_{hi} = d_{hi} X_{hi}$.

We also place ourselves in the same asymptotic framework as Kokic and Bell (1994) by adapting the hypothesis on the inclusion probabilities:

$$\forall h = 1, \ldots, H, \exists \, (\lambda_{1h}, \lambda_{2h}) \in]\, 0, 1 \,[^2, \text{ such that } \forall i \in U_h, \min(\pi_i) > \lambda_{1h} \text{ and } \max(\pi_i) < \lambda_{2h}. \quad (2.8)$$

As in the approach presented in the previous section, the thresholds K_h are determined so as to minimize the mean square error of the winsorized estimator $\hat{T}(\tilde{X})$ with respect to both the model of the variable X and the sampling design P, i.e., on average across all possible populations, given the super-population model applied to X and on average for all samples drawn from these populations, given the sampling design P:

$$(K_h^*)_{h=1, \ldots, H} \in \operatorname{Argmin}_{(K_h)_{h=1, \ldots, H}} E_m E_P \left\{ \left[\hat{T}(\tilde{X}) - T(X) \right]^2 \right\}.$$

It is possible to show (see Appendix A) that at the optimum and asymptotically, denoting as previously $J_{hi} = \mathbb{I}(Z_{hi} > K_h)$ and omitting the indicator i in the expression of expectations and variances under model (2.5) of the variables Z_{hi} and J_{hi}:

$$\forall h = 1, \ldots, H, \, K_h \sim -\frac{A_h}{C_h + D_h} B \quad (2.9)$$

with

$$\begin{cases} A_h = \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}} \right) \\ C_h = \sum_{i \in U_h} \left(\frac{1}{d_{hi}} \right)^2 \left(1 - \frac{1}{d_{hi}} \right)^2 \\ D_h = \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}} \right)^3 \end{cases}$$

and

$$B = \sum_{h=1}^{H} A_h \left[K_h E_m (J_h) - E_m (J_h Z_h) \right]. \quad (2.10)$$

B is the bias of the optimal winsorized estimator $\hat{T}(\tilde{X})$ at the optimum the threshold K_h is therefore equal to a near positive term, in contrast to the bias multiplied by the term $\frac{A_h}{C_h + D_h}$.

If we denote $L = -B$ and $X_{hi}^* = \frac{C_h + D_h}{A_h} Z_{hi}$, then asymptotically $J_{hi} = J_{hi}^* = \mathbb{I}(X_{hi}^* > L)$ using relation (2.9).

By injecting equivalence relation (2.9) into formula (2.10) defining $B,$ we obtain only optimally and asymptotically, B is the opposite of the zero-point of the function F defined by:

$$F(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^*)\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^* X_h^*). \tag{2.11}$$

As in the previous section, we assume finally that we have, for each subpopulation $h,$ of p_h realizations $\breve{X}_{hi}$ drawn from the law of X and independent of the sample $S.$ With these observations, we can estimate F by:

$$\hat{F}(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \mathbb{I}\left(\breve{X}_{hi}^* > L\right)}{p_h}\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^* \mathbb{I}\left(\breve{X}_{hi}^* > L\right)}{p_h} \tag{2.12}$$

and estimate B by the opposite of the zero-point of $\hat{F}.$

We will denote $\breve{X}_{(j)}^*$ the values of the $\breve{X}_{hi}^*$ placed in ascending order. Then, between two successive values $\breve{X}_{(j)}^*$ and $\breve{X}_{(j+1)}^*,$ the indicators $\mathbb{I}\left(\breve{X}_{hi}^* > L\right),$ as functions of $L,$ remain constant and with a positive slope. $\hat{F}$ is therefore a linear and increasing function of $L.$

In addition, $\hat{F}(0) = -\sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^*}{p_h} \leq 0$ and, when L exceeds $\breve{X}_{(p)}^*,$ with $p = \sum_{h=1}^{H} p_h,$ $\hat{F}(L) = L \geq 0.$ To determine the zero-point of $\hat{F},$ it is necessary to operate using a method similar to that proposed by Kokic and Bell (1994) in the case of stratified simple random sampling:

- calculate $\hat{F}(0), \hat{F}\left(\breve{X}_{(1)}^*\right), \hat{F}\left(\breve{X}_{(2)}^*\right), \ldots, \hat{F}\left(\breve{X}_{(p)}^*\right);$
- identify the value j such as $\hat{F}\left(\breve{X}_{(j)}^*\right) \leq 0$ and $\hat{F}\left(\breve{X}_{(j+1)}^*\right) \geq 0,$ assuming that $\breve{X}_{(0)}^* = 0;$
- B is then estimated by interpolation, as in the previous section:

$$\hat{B} = -\frac{\breve{X}_{(j)}^* \hat{F}\left(\breve{X}_{(j)}^*\right) - \breve{X}_{(j+1)}^* \hat{F}\left(\breve{X}_{(j+1)}^*\right)}{\hat{F}\left(\breve{X}_{(j)}^*\right) - \hat{F}\left(\breve{X}_{(j+1)}^*\right)}.$$

3 Review of methods based on conditional bias

3.1 Definition

The conditional bias of an estimator $\hat{\theta}$ for the parameter $\theta,$ for a unit $i \in U$ was defined in the framework of Sampling Theory by Moreno-Rebollo et al. (1999) as follows:

$$B_{1i}^{\hat{\theta}} = E_P\left(\hat{\theta} - \theta \,\middle|\, I_i = 1\right), \tag{3.1}$$

$$B_{0i}^{\hat{\theta}} = E_P\left(\hat{\theta} - \theta \,\middle|\, I_i = 0\right). \tag{3.2}$$

The conditional bias of a sampled unit is equal to the average of the difference between $\hat{\theta}$ and θ on the set of samples containing that unit. Similarly, the conditional bias of an unsampled unit is equal to the average of the sampling error for all samples not containing that unit.

In the case of a one-phase sampling design, the conditional bias of the Horvitz-Thompson estimator $\hat{T}(X) = \sum_{i \in S} \frac{x_i}{\pi_i}$ associated with a sampled unit i is defined by

$$B_{1i}^{\hat{T}(X)} = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) x_j \tag{3.3}$$

where π_{ij} designates the joint inclusion probability of units i and j in the sample. Conditional bias (3.3) is, in general, unknown since the values of the variable of interest are only observed for the units in the sample. In practice, it is possible to estimate it without bias, or in a robust way, from the sample. We consider the conditionally unbiased estimator (see, for example, Beaumont et al., 2013):

$$\hat{B}_{1i}^{\hat{T}(X)} = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) x_j. \tag{3.4}$$

This estimator is conditionally unbiased in the sense that $E_P \left(\hat{B}_{1i}^{\hat{T}(X)} \middle| I_i = 1 \right) = B_{1i}^{\hat{T}(X)}$ only if π_{ij} are strictly positive. Moreover, conditional bias (3.3) and its estimator (3.4) depend on the inclusion probabilities π_i and the joint inclusion probabilities π_{ij}. In other words, conditional bias is a measure that takes the sampling design into account.

For a Poisson design, the conditional bias of the sampled unit i is given by

$$B_i^{\hat{T}(X)} (I_i = 1) = (d_i - 1) x_i. \tag{3.5}$$

Unlike the case of other sampling designs, such as simple random sampling without replacement, conditional bias (3.5) is known directly for all sample units and does not require estimation from the sample because it does not depend on any parameter of the finite population.

Conditional bias, as demonstrated by Beaumont et al. (2013), is a direct measure of the influence of each unit on the estimation error, the second relation being verified for maximum entropy sampling designs:

$$V \left[\hat{T}(X) \right] = \sum_{i \in U} B_{1i}^{\hat{T}(X)} y_i \tag{3.6}$$

$$\hat{T}(X) - T(X) \approx \sum_{i \in S} B_{1i}^{\hat{T}(X)} + \sum_{i \in U - S} B_{0i}^{\hat{T}(X)}. \tag{3.7}$$

3.2 A robust estimator based on conditional bias

As shown by formulas (3.6) and (3.7), the conditional bias (CB) measures the effect of each unit on the estimation error and the estimation variance. A robust estimator should be defined in such a way that observations of the sample have only controlled and limited values of their conditional bias. Based on this idea, Beaumont et al. (2013) suggested using an estimator of the form:

$$\begin{aligned}\hat{T}^{\text{CB}}(X)(c) &= \hat{T}(X) + \sum_{i\in S}\Psi_c\left[\hat{B}_{1i}^{\hat{T}(X)}\right] - \sum_{i\in S}\hat{B}_{1i}^{\hat{T}(X)} \\ &= \hat{T}(X) - \sum_{i\in S}\left[\hat{B}_{1i}^{\hat{T}(X)} - \Psi_c\left(\hat{B}_{1i}^{\hat{T}(X)}\right)\right]\end{aligned}$$

with Ψ_c the Huber function defined by

$$\Psi_c(t) = \begin{cases} c & \text{if } \ t \geq c \\ t & \text{if } \ -c < t < c \\ -c & \text{if } \ -c \leq t \end{cases}$$

and $\hat{B}_{1i}^{\hat{T}(X)}$ the estimator defined in (3.4).

The Huber function is used to limit the influence of the most influential units by truncating their conditional bias. Parameter c can be chosen according to various optimization criteria for the robust estimator. For example, c can be chosen to obtain the estimate having, under the sample design, the smallest mean square error. However, it is relatively complex or sometimes impossible to obtain an analytical expression of c for a given sample design.

Beaumont et al. (2013) suggest choosing $c^* \in \text{argmin}_c\ \text{argmax}_i \left|\hat{B}_{1i}^{\hat{T}^{\text{CB}}(X)}(c)\right|$, i.e., the value of the constant c for which the largest absolute value of the estimated conditional bias for the sample observations on the robust estimator is the lowest. In this case, the robust estimator is equal to:

$$\hat{T}^{\text{CB}}(X)(c^*) = \hat{T}^{\text{BHR}}(X) = \hat{T}(X) - \frac{\min_i \hat{B}_{1i}^{\hat{T}(X)} + \max_i \hat{B}_{1i}^{\hat{T}(X)}}{2}. \tag{3.8}$$

The Beaumont, Haziza and Ruiz-Gazen estimator is thus simple to implement. Compared to the Kokic and Bell method, it is more general because it is valid for all sampling designs and does not require any information outside the sample to be determined. In addition, it does not rely on any hypotheses about the variable of interest. The resulting estimator is robust under the sample design, while the Kokic and Bell estimator considers the sampling design and the distribution of the variable of interest. However, it is not designed to have the smallest mean square error, but to obtain an estimator on which the influence of each unit is limited, by minimizing the influence of the most influential unit.

The method has been extended to integrate more elements of the sample design and to adapt to certain situations. Favre-Martinoz et al. (2016) extended the method for a two-phase sampling design, which makes it possible to take non-response into account when it is assimilated to a second phase of Poisson drawing; Favre-Martinoz et al. (2015) proposed a method for ensuring the consistency of the robust estimators obtained when the parameters of interest are the totals of a variable in different domains included in one another.

4 Comparison of winsorization and conditional bias

In the previous section, we presented two types of methods for processing influential units applied to survey data:

- the Kokic and Bell winsorization, which aims to determine the winsorization thresholds that minimize the mean square error of the winsorized estimator under the sample design and the law of the variable of interest, which was initially conceived for a stratified simple random sampling design, but which we have extended to the case of Poisson sampling. The Kokic and Bell method, like its extension, is thus valid under hypotheses made about the law of the winsorized variable;
- the conditional bias method proposed by Beaumont, Haziza and Ruiz-Gazen, which potentially applies to all sampling designs and does not rely on any hypothesis on the law of the variable of interest; it aims to obtain the estimator for which the most influential unit has the least influence possible.

To compare the efficiency of these two methods, we performed two exercises:

1. simulations applied to the Poisson sampling;
2. a comparison on real data, applied to the data from the French labour cost and structure of earnings survey (ECMOSS).

4.1 Simulations in the case of a Poisson sampling

We performed a simulations study to examine the properties of the two robust estimators proposed in the context of a Poisson drawing. We carried out four scenarios to compare the efficiency of the two estimators, but also to study, in the case of the Kokic and Bell estimator, the model's robustness to a bad specification, i.e., to a modification between the learning model and the model that generated the sample data.

The simulation proceeds as follows:

- We consider $L = 1{,}000$ realizations of a certain model, which makes it possible to generate our learning set of $N = 5{,}000$ units;
- For each of these realizations, we calculate the optimal threshold K_l according to the method proposed in Section (2.2);
- Then we create $M = 10{,}000$ test sampling frames generated according to a (different) model on which we select a sample of expected size $n = 500$ following a Poisson drawing and calculate the robust estimator $\hat{\theta}_{(m)}$ with the threshold K_l calculated. As a comparison, we also calculate the robust estimator resulting from the method based on the conditional bias.

The inclusion probabilities, as well as the values of the variable, X were generated according to the following model:

$$U_i \sim \mathcal{L}\text{og } -\mathcal{N}\,(1; 1.1),$$

$$\pi_i = n \times \frac{U_i}{\sum_{i=1}^{N} U_i},$$

$$X_i = 2{,}000 \times \pi_i + \pi_i \epsilon_i + \delta_i V_i,$$

$$\epsilon_i \sim \mathcal{N}\,(0; 100),\, V_i \sim \mathcal{L}\text{og } -\mathcal{N}\,(\log(500); 1.2),\, \delta_i \sim \mathcal{B}(\omega),$$

where ω is the Bernoulli parameter, reflecting the proportion of influential values whose values are given in Table 4.1. The notation $\mathcal{L}\text{og } - \mathcal{N}$ denotes a log-normal distribution.

Table 4.1
Values of parameter ω used to generate populations

	Values of parameter ω	
Scenario	**Learning model**	**Test model**
1	0	0
2	0.01	0.01
3	0.01	0.1
4	0.1	0.01

Scenario 1 corresponds to the population model for which the extension of the Kokic and Bell method was developed in the Poisson case with $H = 1$, but in which no or very few units are influential (the value of the parameter ω being fixed at 0). Scenario 2 corresponds to a situation in which this model applies, but in which a small proportion (1%) of units are influential. The model is, in scenarios 1 and 2, identical in the population used to calculate the threshold and the sample to which the threshold is applied.

In scenarios 3 and 4, the basic model is the same between the learning population and the sample, but the number of influential units varies between the two. In scenario 3, the learning population contains 10 times fewer influential units than the sample. Scenario 4 corresponds to the opposite scenario.

As a measure of the bias of an estimator $\hat{\theta}$ of a total T, we calculated the relative Monte Carlo bias (as in percentage)

$$\text{BR}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - T\right)}{T} \times 100,$$

where $\hat{\theta}_{(m)}$ is the estimator $\hat{\theta}$ in the sample $m, m = 1, \ldots, M$.

We also calculated the relative efficiency of the robust estimators relative (RE) to the dilation estimator, $\hat{t}$:

$$\text{RE}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - T\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - T\right)^2} \times 100.$$

Tables 4.2 and 4.3 represent the descriptive statistics associated with the $L = 1{,}000$ Monte Carlo values calculated according to the learning population considered.

Table 4.2
Descriptive statistics for scenarios 1 and 2 of the 1,000 simulations for $n = 500$

Statistic	Scenario							
	1				2			
Description	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
Min.	-0.2	100	-0.43	100	-9.0	1	-4.3	26
$Q1$	-0.1	100	-0.32	100	-2.9	35	-1.9	51
Median	0.0	100	-0.27	100	-1.8	50	-1.5	62
Mean	0.0	100	-0.27	100	-2.0	50	-1.6	62
$Q3$	0.0	100	-0.23	100	-1.0	64	-1.3	73
Max.	0.0	100	-0.14	100	-0.1	109	-0.6	91

Scenario 1 corresponds to a situation in which no or very few influential units are present in the population: the performance of the robust estimators is therefore identical to that of the usual Horvitz-Thompson estimator, with a relative bias very close to 0. Scenario 2 corresponds to the situation for which the extension of the Kokic and Bell method to the Poisson case was developed, with the introduction of influential units. The two robust estimators are more effective than the usual estimator, but the performance of the Kokic and Bell estimator in terms of the gain in mean square error is greater, with a median relative efficiency over the 1,000 simulations of 50%, compared to 62% for the conditional bias method. This result is expected given that the threshold of the Kokic and Bell method is explicitly determined to obtain the estimator with the smallest mean square error.

Table 4.3
Descriptive statistics for scenarios 3 and 4 on the 1,000 simulations for $n = 500$

Statistic	Scenario							
	3				4			
Description	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
Min.	-32.2	2	-7.8	27	-4.5	1	-4.3	26
$Q1$	-18.9	50	-5.1	59	-1.8	48	-1.9	51
Median	-13.9	82	-4.6	66	-1.5	70	-1.5	62
Mean	-14.2	89	-4.7	65	-1.5	68	-1.6	62
$Q3$	-9.3	138	-4.2	72	-1.2	91	-1.3	73
Max.	-0.01	537	-2.7	89	-0.6	100	-0.6	91

The performances of the two methods in scenario 3 are more contrasted. While over the set of simulations, the conditional bias method succeeds in reducing the mean square error of the estimators, with a minimum mean square error gain of 27%, the Kokic and Bell method deteriorates precision in more than a quarter of cases. The population on which the threshold was calculated contains, in this scenario, too few influential units compared to the sample for the calculated threshold to be effective.

In scenario 4, where the learning population contains more influential units than the sample, the performances of the two methods are of the same order of magnitude.

Therefore, these simulations show:

- that in the absence of influential units, the two robust estimation methods do not lead to a loss of estimation efficiency;
- that when applied in its hypotheses, the Kokic and Bell method leads to more accurate estimators than the conditional bias method;
- that the Kokic and Bell method is, however, sensitive to the data used to calculate thresholds; if these data are not generated according to the same model as the data to which the thresholds are applied, the method may lead to a loss of precision;
- that the conditional bias method always allows a gain in precision on these simulations, even if this gain is not optimal.

4.2 Application to the Survey on labour costs and wage structure

4.2.1 Presentation of the survey

The Survey on labour cost and structure of earnings (ECMOSS) is conducted by INSEE every year and harmonized at the European level. It is used to respond to European regulations on the production of statistics on both the cost of labour and structure of earnings which contribute to comparisons between European countries in terms of work time and costs.

ECMOSS is a survey of local business units (or establishments). It covers all sectors–both market and non-market–with the exception of agriculture, state administrations and certain activities (extraterritorial activities, embassies, consulates, activities of individuals acting as employers) and businesses with 10 or more employees. It covers establishments located in the metropolitan territory and in the overseas departments. Each sampled business answers two questionnaires: In the first, it must provide a certain amount of aggregated information on its workforce, payroll and a breakdown into its main elements (basic wages, bonuses, social contributions paid by the employer and by employees, etc.) and on the number of work hours of its employees; in the second, it details these elements for a randomly selected sample of its employees.

Given this survey method, the ECMOSS sample design has two stages:

- First stage: A sample of approximately 17,000 establishments is selected according to a stratified sampling design by sector, size of business, size of the establishment and geographical location;

- Second stage: In each establishment, a sample of employees is selected from the lists of employees reported by the establishment to social security organizations. The sampling design is drawn independently in each establishment and stratified by social category of the employees, distinguishing between managers and non-managers. The number of employees surveyed in each establishment varies according to its size, but is limited to 24 to prevent the survey from placing too much burden on businesses. In the end, around 150,000 employees are surveyed each year.

Each year, a certain number of establishments do not respond to the survey, and responding establishments do not systematically provide information for all their employees. Therefore, there is total non-response at each stage, which is handled by reweighting according to the homogeneous response group method. Next, the final sample of respondent employees, on which most operations are performed, is calibrated on the population of employees from the files of social security organizations.

Last, the sample of employees is obtained through a complex sample design, comprising two drawing stages (establishments and employees), with two drawing phases at each stage.

Given the very great variability of the establishments and their wage policy (both in terms of differences in the average levels of wages between establishments and differences in the dispersion of wages in the establishments), the sampling weights of the sampled employees are widely dispersed, and the accuracy of the estimators is sensitive to the influential values of the sample: for example, a very high level executive in a large business, or the athletes employed by a high-level sports club.

4.2.2 Parameter of interest

The main parameter of interest in the survey is the average hourly wage, calculated in different dissemination domains: sectors, sectors crossed with the employment size ranges of the businesses, and sectors crossed with the region in which the establishment is located. The estimators used later in our simulations are obtained by calculating the ratio of estimators by expansion of total remuneration over the total number of hours:

$$\hat{R}(D) = \frac{\sum_{i \in S \cap D} w_i e_i}{\sum_{i \in S \cap D} w_i h_i} \tag{4.1}$$

with S the sample of employees, D the domain of interest, e_i the annual remuneration of the employee i, h_i their annual hourly work volume and w_i the employee's estimation weight obtained by multiplying the selection probabilities and the response probabilities associated with each stage and phase of the sample design. Estimator (4.1) does not correspond to the estimator used in practice because it involves the initial weights corrected for non-response, while the estimator used in practice uses the calibrated weights. In the context of this article, the calibration phase was not taken into account, but it could have been using the classical residual technique and an additional degree of complexity which we deemed unnecessary to compare the two robust estimation methods.

4.2.3 How to adapt the processing methods for influential units to the ECMOSS sampling design

Estimator (4.1) is not the expansion estimator of a total, for which the previously described methods were designed. The problem can, however, be adapted to the framework of these two methods.

Indeed, an unbiased estimator of the variance of $\sum_{i\in S} w_i \hat{L}_i \left[\hat{R}(D)\right],$ with $\hat{L}_i \left[\hat{R}(D)\right] = \frac{e_i - \hat{R}(D) h_i}{\sum_{i\in S\cap D} w_i\, h_i} \mathbb{I}\,(i \in D)$ the estimated linearized variable of $\hat{R}(D),$ is also an asymptotically unbiased estimator of $V(\hat{R}(D)).$ Thus, a robust estimator of the total of the linearized variable $\hat{L}_i \left[\hat{R}(D)\right]$ will also be a robust estimator for the influential units of $\hat{R}(D).$ Each method, applied to the estimated linearized variable, generates a winsorized value of this variable, denoted $\hat{L}_i^w \left[\hat{R}(D)\right].$ The effects of the processing of the influential units are then transferred to all other variables of interest of the survey through the estimation weight, by calculating a winsorized estimation weight:

$$w_i^w = w_i \frac{\hat{L}_i^w \left[\hat{R}\,(D)\right]}{\hat{L}_i \left[\hat{R}\,(D)\right]}.$$

We thus test the two methods of Kokic and Bell and Beaumont, Haziza and Ruiz-Gazen to estimate the total of $\hat{L}_i \left[\hat{R}\,(D)\right].$ However, each of the two methods requires adaptations to be applied to the sampling design and variables of interest of ECMOSS.

4.2.4 Adaptation of winsorization according to the Kokic and Bell method and its extension

The survey and the parameter of interest of the survey, even after linearization, do not fit with the framework of the Kokic and Bell method, whether it is the original method, or the extension presented previously. First, the ECMOSS sample is not selected using a stratified simple random survey or a Poisson sampling. Moreover, the variable to winsorize, the estimated linearized variable $\hat{L}_i \left[\hat{R}\,(D)\right],$ is not always positive. To apply the Kokic and Bell method to the ECMOSS case, we have made the following adaptations.

1. We apply the processing of the influential units as though the employees were directly selected by stratified simple random sampling (Poisson sampling for the extension) in strata defined by the sector, the number of employees of the business and the location of the employing establishment, by grouping certain modalities of this last variable to avoid generating pseudo-strata containing too few observations (we distinguish Île de France, the overseas departments and the rest of the country) and by the social category of the employee (distinguishing managers and non-managers). As the classical method acts as though the sample in each pseudo-stratum was selected by simple random sampling and thus all employees of the same pseudo-stratum have the same sampling weight, we do not consider the dispersion of the estimation weights in the pseudo-strata from the actual sampling design of the survey, and thus risk missing influential

units. In the case of the extension of the method, this dispersion of the weights is properly taken into account.

2. In each of these pseudo-strata, winsorization is not applied directly to the estimated linearized variable, but to a translated version of it.

More precisely, we define for each sampled employee:

$$\hat{T}_i\left[\hat{R}(D)\right] = \hat{L}_i\left[\hat{R}(D)\right] + \min_{j \in S}\hat{L}_j\left[\hat{R}(D)\right]$$

for which we calculate winsorization thresholds in the pseudo-strata according to the method initially proposed by Kokic and Bell and for its extension. We then deduce two sets of estimation weights used to estimate the average hourly wage in each domain of interest of the form:

$$w_i^w = w_i\frac{\hat{T}_i^w\left[\hat{R}(D)\right]}{\hat{T}_i\left[\hat{R}(D)\right]}.$$

We can thus only identify and process influential units with high values of the estimated linearized variable, i.e., employees whose hourly wage is higher than the average hourly wage in the domain of interest D. Units with low hourly wages cannot be identified by this method, but pose less problems for the accuracy of estimates, since the distribution of hourly wages is particularly skewed, with a very long tail on the right.

A final adaptation is necessary to adapt the method to the case of ECMOSS. This can only be used if observations of the variable of interest in each pseudo-stratum are available. Previous editions of the survey can be used. However, the tests performed to evaluate the efficiency of the Kokic and Bell method applied to the Annual Sectoral Surveys (see Deroyon, 2015) have shown that the use of responses to previous editions of the survey to calculate winsorization thresholds does not lead to the largest gains in accuracy. This is because the small number of observations available per stratum to calculate these thresholds are determined with too little precision, so that too many units can be winsorized, or conversely, influential units escape winsorization. We have chosen to use the auxiliary information available in the social security files on total remuneration paid annually to employees and their number of hours worked. These data are not those measured in the survey (in particular, the wages declared in the social security files form the tax base on which are calculated social contributions and tax contributions on wages, and not labour income paid to employees), but are strongly correlated with them.

4.2.5 Adaptation of Beaumont, Haziza and Ruiz-Gazen estimator

Because of its generality, the conditional bias method requires fewer adaptations to be applied to the ECMOSS. It can thus be applied directly to the variables of interest of the survey without the need to mobilize external data. However, calculating conditional biases while considering the whole sampling design is complex; therefore, for our simulations, we chose to apply the conditional bias methods as though the employees had been selected directly by a Poisson sampling, with the selection probabilities $1/w_i$, where w_i designates the estimation weight after correction for non-response of the employee i. The conditional bias used to identify influential units is therefore equal to:

$$B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} = (w_i - 1)\,\hat{L}_i\left[\hat{R}(D)\right].$$

With formula (3.8), the Beaumont, Haziza and Ruiz-Gazen estimator processes only a limited number of units, i.e., the observations with the lowest and highest conditional biases, for which all corresponding indicators define the sets $A_{\min}$ and $A_{\max}$:

$$A_{\min} = \operatorname{argmin}_{j\in S} B_{1j}\left\{\hat{L}_j\left[\hat{R}(D)\right]\right\}$$

$$A_{\max} = \operatorname{argmax}_{j\in S} B_{1j}\left\{\hat{L}_j\left[\hat{R}(D)\right]\right\}.$$

Thus, the processed estimation weight of the influential units is equal to:

$$w_i^{\mathrm{BHR}} = \begin{cases} \dfrac{(2\,|\,A_{\min}\,|-1)\,w_i + 1}{2\,|\,A_{\min}\,|} & \text{if } \; B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} \in A_{\min} \\[2ex] \dfrac{(2\,|\,A_{\max}\,|-1)\,w_i + 1}{2\,|\,A_{\max}\,|} & \text{if } \; B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} \in A_{\max} \\[2ex] w_i & \text{otherwise.} \end{cases}$$

where $|\,A_{\min}\,|$ and $|\,A_{\max}\,|$ respectively designate the cardinal of $A_{\min}$ and $A_{\max}$.

Compared to the Kokic and Bell method, the robust estimator based on conditional biases does not focus on influential units located in the right-hand part of the distribution of the estimated linearized variable, but identifies the influential units with very low and very high values for this variable. It also focuses on an a priori limited number of units, since only observations with the minimum and maximum conditional bias are modified.

4.2.6 Robust estimation on several domains of interest

As previously described, the domains of interest for the dissemination of the ECMOSS results are numerous. For the sake of simplicity of dissemination and to comply with the requirements of European regulations, each employee in the individual sample must have only one estimation weight, so adaptations are necessary:

- Robust estimators for several sets of domains of interest

  European regulations require the dissemination of results in sets of domains that intersect and are not included in one another, such as intersections of sectors and ranges of numbers of employees and crossings of sectors and regions. Sampled units may belong to more than one dissemination domain.

Ideally, the processing of influential units should be done in each domain of interest separately, so that a single observation may be associated with a different estimation weight for each dissemination domain to which it belongs. However, this solution is not possible for the reasons mentioned above.

Another solution is to apply both of the methods to the crossings of all the dissemination domains. The risk is then in applying the processes for estimators calculated on very small populations, for which many units are influential. Thus, for the estimation on real dissemination domains, too many units would be winsorized. The resulting estimators will be less precise than the robust estimators adapted to each domain, but potentially also less precise than the unprocessed estimators of the influential units because they are too biased.

- Robust estimators for all modalities of a domain of interest

  For a given set of domains (e.g., industry sectors), an observation can be identified as influential and processed for estimation on more than one dissemination domain, and thus have more than one final estimate weight. This is the case if the selection of an observation belonging to a dissemination domain has an influence on the selection of other units belonging to other dissemination domains (e.g., in the case of a stratified sampling, if the dissemination domains intersect the drawing strata).

  This situation is impossible for the Beaumont, Haziza and Ruiz-Gazen estimator, for which we assume the Poisson sampling design. However, this can happen for the Kokic and Bell method and its extensions as we implement them, because some dissemination domains do not consist of groupings of the pseudo-strata that we have formed. The situation is then the same as that exposed in the case of several sets of dissemination domains: the only way to maintain a unique estimation weight for each sampled unit is to apply the methods to pseudo-dissemination domains close to the real dissemination domains but made up of groupings of winsorization pseudo-strata. These pseudo-domains are in fact formed by intersections of sectors, a range of the number of employees of the businesses and the geographic location of the establishments (distinguishing only the three modalities specified above).

To evaluate the performance in terms of precision gains or losses of the methods defined above, we carried out a set of simulations based on the ECMOSS sampling design and data on wages and hours worked from the social security files, available for all employees and for which we are therefore able to compare the average hourly wages observed in the population with their various estimators. In these simulations, we compared the efficiency of the methods applied directly to each dissemination domain, which lead to the optimal results, and to the pseudo-dissemination domains defined above.

4.2.7 Simulations

The simulations are conducted in the social security files, from which the sample of employees is selected and which are available for all French employees. They are implemented as follows:

- the ECMOSS sampling design (including the selection of responding establishments and employees) is applied 5,000 times to produce 5,000 samples of employees, denoted S_m, $m = 1, \ldots, 5{,}000$;
- for each sample and each dissemination domain, we calculate the usual expansion estimator of $\hat{R}_m(D)$;
- the Kokic and Bell winsorization and conditional bias are applied to each sample according to different specifications:
    - the Kokic and Bell winsorization, classical or adapted to Poisson sampling, is applied only as though the real dissemination domains were the pseudo-dissemination domains defined above.
    - The Beaumont, Haziza and Ruiz-Gazen estimator is applied in each activity sector taken separately on the one hand, and on the other hand as though the pseudo-dissemination domains defined above were the real dissemination domains. For each dissemination domain, we can thus compare the performances of the conditional bias estimator applied in its optimal specification for this domain (producing an estimator $\hat{R}_m^{\text{BHR}*}(D)$ of the average hourly wages in the domain) to the conditional bias method and the Kokic and Bell method (producing estimators $\hat{R}_m^{\text{BHR}}(D)$, $\hat{R}_m^{\text{KB}}(D)$ and $\hat{R}_m^{\text{KB}_{\text{poiss}}}(D)$ for the extension) applied according to specifications that are sub-optimal for this domain but simpler to implement.

For each robust estimator and each domain, we calculate the mean relative bias (RB) and the relative mean square error (RMSE) for all simulations by:

$$\text{AB}\left[\hat{R}^{\text{KB}}(D)\right] = \frac{\sum_{m=1}^{5{,}000}\left[\hat{R}_m^{\text{KB}}(D) - R(D)\right]}{5{,}000}$$

$$\text{AMSE}\left[\hat{R}^{\text{KB}}(D)\right] = \frac{\sum_{m=1}^{5{,}000}\left[\hat{R}_m^{\text{KB}}(D) - R(D)\right]^2}{5{,}000}$$

$$\text{RB}\left[\hat{R}^{\text{KB}}(D)\right] = 100\,\frac{\text{AB}\left[\hat{R}^{\text{KB}}(D)\right]}{R(D)}$$

$$\text{RMSE}\left[\hat{R}^{\text{KB}}(D)\right] = 100\,\frac{\text{AMSE}\left[\hat{R}^{\text{KB}}(D)\right]}{\text{AMSE}\left[\hat{R}(D)\right]}$$

where, for example, for the classical Kokic and Bell method, $R(D)$ designates the average hourly wage observed in the social security files in the domain D and $\hat{R}(D)$ designates the usual expansion estimator of this parameter. Relative bias compares the bias of the robust estimator to the real value of the parameter. The relative mean square error measures the gain or loss of precision provided by the robust estimators relative to the usual estimator.

4.2.8 Simulation results

Among the different estimators tested in our simulations, the estimator obtained by applying the adaptation of the Kokic and Bell method to Poisson sampling is distinguished by extremely poor performances, summarized in Table 4.4. Application of the Kokic and Bell method extended to Poisson sampling for the ECMOSS results in a significant or even dramatic deterioration in the precision of the estimates.

Table 4.4
Statistics on the mean square error (MSE) ratio of the robust Kokic and Bell estimators applied to the Poisson sampling in the different domains of interest

Statistic	$\text{RMSE}\left(\hat{R}_m^{\text{KB}_{\text{poiss}}}(D)\right)$		
	Domain		
	NACE*Workforce	**NACE**	**NACE*NUTS**
Min.	18	128	33
Mean	490	1,858	324
Max.	4,437	8,606	2,466

Figures 4.1, 4.2 and 4.3 focus on presenting the results of the conditional bias and classical Kokic and Bell methods, applied under the hypothesis of a stratified simple random sampling.

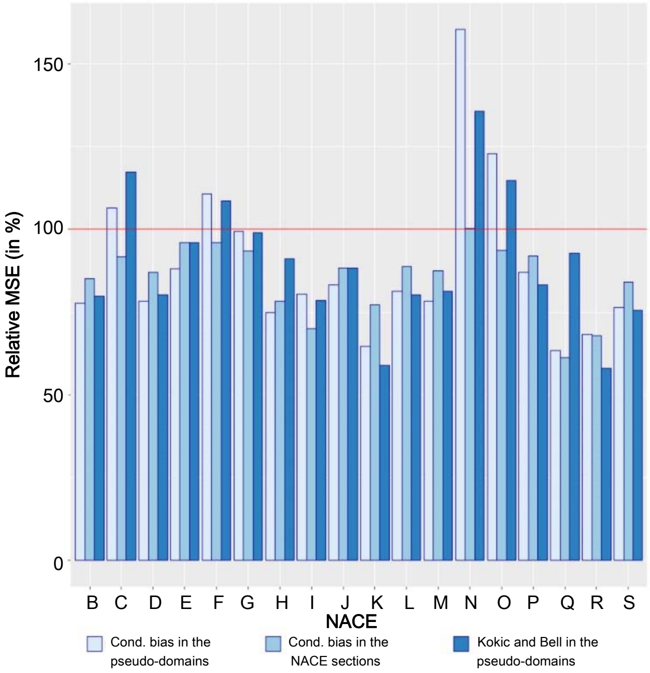

Figure 4.1 Relative mean square errors for the estimators of average hourly wage by sector.

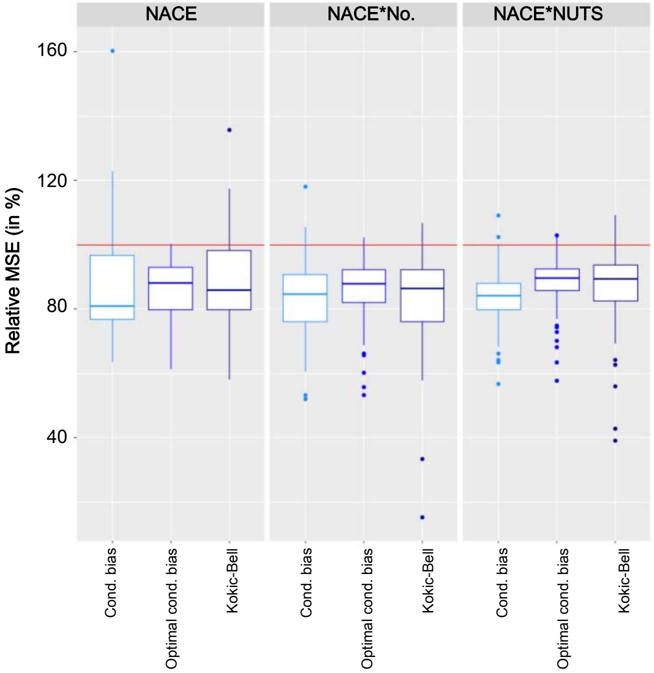

Figure 4.2 Distribution of relative mean square errors in each domain.

Figure 4.1 shows the relative mean square errors of the robust average hourly wage estimators in each section of the Statistical classification of economic activities in the European Community (NACE, a grouping of business sectors into 21 categories, of which 18 are in the ECMOSS field) and Figure 4.2 shows the distribution of relative mean square errors in each domain (among all sections, section crossings, and number of business employees, or crossings of sector and location of the establishment).

For almost all domains of interest, the robust estimators considered provide gains in precision over the usual expansion estimator. The domains in which the robust estimators have a higher error than the usual estimator are also those where the estimation variance is the lowest originally. The processes for influential units considered in these figures (conditional bias and classical Kokic and Bell method) are thus able to reduce estimation errors when necessary without causing too much loss of precision when the estimators are not affected by influential units.

The biases of the average hourly wage estimators in the sectors are low (see Figure 4.3), except in some domains where the sample size is small (A: Agriculture, forestry and fishing; K: Financial and insurance activities; R: Arts, entertainment and recreation). The same results are also observed for the other domains.

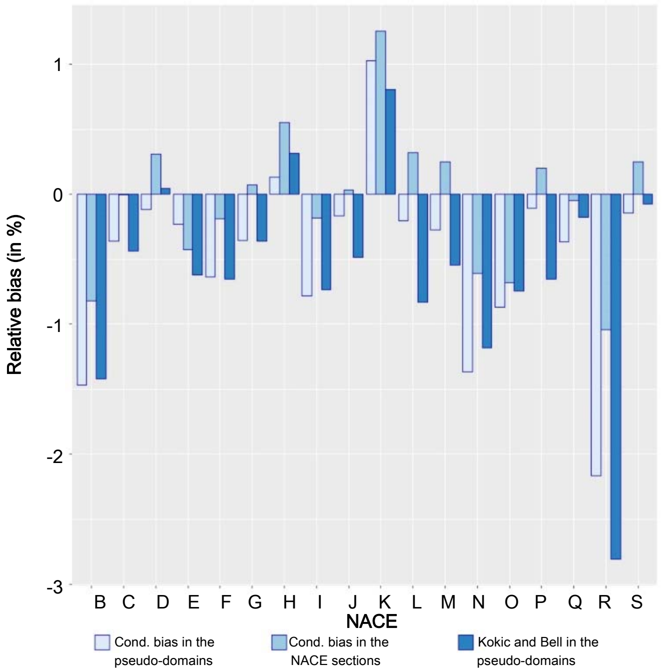

Figure 4.3 Relative biases for estimators of average hourly wage by sector.

The application of conditional bias methods adapted to each domain gives the best results for the estimation in the NACE sections, but not necessarily in the other dissemination domains. The NACE sections are much more aggregated than the pseudo-domains used for the identification of influential units, so the bias introduced by the processing of influential units is more significant in the cases where the application is made on pseudo-domains, compared to the optimal version applied directly to the NACE sections. In the other domains, the identification of influential units at a finer level than the real dissemination domain makes it possible to identify more influential units and thus substantially reduce the estimation variance, without introducing too much additional bias, when the domain used to identify the influential units and the real dissemination domains are close. Differences in how the sampling design is described to apply each of the two methods and the actual sampling design may explain why the use of the Beaumont, Haziza and Ruiz-Gazen robust estimator in each dissemination domain does not necessarily translate into greater precision gains.

The differences between the results obtained with the conditional bias and Kokic and Bell methods under the hypothesis of the stratified simple random sampling design are, however, small. Note however that, for the implementation of these simulations, we use the population data as observations of the additional interest variables not from the sample to calculate the winsorization thresholds in the Kokic and Bell method. Since we also evaluate the performance of the different estimators based on these data, the Kokic and Bell method is favoured a priori.

The extension of the Kokic and Bell method to Poisson sampling results in a significant deterioration in the precision of the estimators.

The discrepancies between the performances of the two implementations of the Kokic and Bell method are thus very high. However, these implementations are both based on two hypotheses:

- a hypothesis on the sampling design used to select the sample;
- a hypothesis on the distribution of the variable of interest in subpopulations U_h.

In both applications of the Kokic and Bell method, the first hypothesis is not respected. The violation of this hypothesis is, however, a priori more significant when we apply the Kokic and Bell method as though the sample had been selected by a stratified simple random sampling in pseudo-strata constructed ad-hoc, because in so doing we assume that the selection probabilities are identical in these pseudo-strata, which is not at all verified. The Kokic and Bell method applied as though the employees had been selected by Poisson sampling, for its part, considers real simple inclusion probabilities, but neglects the links between the indicators of belonging to the sample of different employees.

However, the population model postulated for the Kokic and Bell method extended to the Poisson case is not valid, since the simple inclusion probabilities are not proportional to the variable of interest considered. It is more complex to assess the validity of the population model used for the classical Kokic and Bell method; up to a point, it is still possible to consider that the results of the variable of interest in a pseudo-stratum are derived from the same law whose expectation and variance can be estimated by the mean and the empirical variance of the results of the variable of interest in the stratum.

Also, the performance differences of the two implementations of the Kokic and Bell method are complex to analyze. A first possible explanation is that the performances of the method are more sensitive to violations of the hypothesis on the law of the observations than to those on the form of the sampling design. This finding was shared by Fizzala (2017) in the case of an application of winsorization in the context of corporate profiling. In our ECMOSS simulations, we observe that the classical Kokic and Bell method, based on the hypothesis of stratified simple random samplings, gives very valid results despite the fact that this hypothesis is only partially respected. Future extensions of this work could consist of validating this explanation on the basis of simulations. Another explanation for these differences in performance may lie in the relationship between the two hypotheses in the case of the extension of Kokic and Bell to Poisson sampling. Indeed, while in the case of the classical Kokic and Bell method, the hypotheses on the sampling design and the law of the variable of interest in each stratum are unrelated, in the case of the Poisson sampling, the population model involves selection probabilities and therefore implies additional constraints on the sampling design. Therefore, the fact that the selection probabilities are not proportional to the variable of interest implies that, for the extension of the Kokic and Bell to Poisson sampling, the hypotheses on the sampling design and the population are simultaneously violated, which could explain this explosion of errors of the estimator.

However, the conditional bias and classical Kokic and Bell methods, whatever the configuration, seem to be able to identify influential units for the estimation of the parameters affected, and thus guarantee significant gains in precision even when they are applied in a setting that is remote from their original hypotheses and the actual sampling design of the survey.

Appendix

A Demonstrations of the formulas for the extension of the Kokic and Bell method in the case of a Poisson sampling

A.1 Calculation of the mean square error of the winsorized estimator

First, we will calculate

$$\begin{aligned} E_P\left\{\left[\hat{T}(\tilde{X}) - T(X)\right]^2\right\} &= E_P\left\{\left[\hat{T}(\tilde{X}) - T(\tilde{X})\right]^2 + \left[T(\tilde{X}) - T(X)\right]^2 \right.\\ &\qquad \left. + 2\left[\hat{T}(\tilde{X}) - T(\tilde{X})\right]\left[T(\tilde{X}) - T(X)\right]\right\} \\ &= E_P\left\{\left[\hat{T}(\tilde{X}) - T(\tilde{X})\right]^2\right\} + \left[T(\tilde{X}) - T(X)\right]^2 \end{aligned}$$

with $T(\tilde{X}) = E_P\left[\hat{T}(\tilde{X})\right] = \sum_{h=1}^{H}\sum_{i\in U_h}\tilde{X}_{hi}$.

Furthermore,

$$E_P\left\{\left[\hat{T}(\tilde{X}) - T(\tilde{X})\right]^2\right\} = \sum_{h=1}^{H}\sum_{i\in U_h} d_{hi}\left(1 - \frac{1}{d_{hi}}\right)\tilde{X}_{hi}^2$$

finally:

$$E_P\left\{\left[\hat{T}(\tilde{X}) - T(X)\right]^2\right\} = \sum_{h=1}^{H}\sum_{i\in U_h} d_{hi}\left(1 - \frac{1}{d_{hi}}\right)\tilde{X}_{hi}^2 + \left[\sum_{h=1}^{H}\sum_{i\in U_h}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^2. \tag{A.1}$$

Assuming in each stratum that:

- $E_m(d_{hi}X_{hi}) = \mu_h$;
- $\mathrm{Var}_m(d_{hi}X_{hi}) = \sigma_h^2 < +\infty$;
- and that $d_{hi}X_{hi}$ are independent and of density $g_h(x) > 0$.

and noting that:

$$\begin{aligned} \tilde{X}_{hi} &= X_{hi}(1 - J_{hi}) + J_{hi}\left[\frac{X_{hi}}{d_{hi}} + \left(1 - \frac{1}{d_{hi}}\right)\frac{K_h}{d_{hi}}\right] \\ &= \frac{1}{d_{hi}}\left[d_{hi}X_{hi} + J_{hi}\left(1 - \frac{1}{d_{hi}}\right)(K_h - d_{hi}X_{hi})\right] \end{aligned}$$

and so that:

$$\tilde{X}_{hi} - X_{hi} = \frac{1}{d_{hi}}\left(1 - \frac{1}{d_{hi}}\right)J_{hi}(K_h - d_{hi}X_{hi}), \tag{A.2}$$

we obtain:

$$\begin{aligned}\tilde{X}_{hi}^2 &= \frac{1}{d_{hi}^2}\left[d_{hi}^2 X_{hi}^2 + J_{hi}\left(1-\frac{1}{d_{hi}}\right)^2 \left(K_h^2 + d_{hi}^2 X_{hi}^2 - 2 d_{hi} X_{hi} K_h\right)\right.\\ &\qquad \left. + 2\left(1-\frac{1}{d_{hi}}\right)\left(d_{hi} X_{hi} J_{hi} K_h - J_{hi} d_{hi}^2 X_{hi}^2\right)\right],\end{aligned} \tag{A.3}$$

and that:

$$\begin{aligned}E_m\left\{\left[\sum_{h=1}^{H}\sum_{i\in U_h}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^2\right\} &= \sum_{h=1}^{H}\sum_{i\in U_h} V_m\left(\tilde{X}_{hi} - X_{hi}\right)\\ &\quad + \left[\sum_{h=1}^{H}\sum_{i\in U_h} E_m\left(\tilde{X}_{hi} - X_{hi}\right)\right]^2\\ &= \sum_{h=1}^{H}\sum_{i\in U_h}\left\{E_m\left[\left(\tilde{X}_{hi} - X_{hi}\right)^2\right] - \left[E_m\left(\tilde{X}_{hi} - X_{hi}\right)\right]^2\right\}\\ &\quad + \left[\sum_{h=1}^{H}\sum_{i\in U_h} E_m\left(\tilde{X}_{hi} - X_{hi}\right)\right]^2.\end{aligned} \tag{A.4}$$

In the end, taking the expectation under the model of expression (A.1) and applying simplifications (A.2), (A.3), (A.4), we obtain, after some additional simplifications:

$$\begin{aligned}E_m E_P\left\{\left[\hat{\tilde{T}}\left(\tilde{X}\right) - T\left(X\right)\right]^2\right\} &= \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left\{\mu_h^2 + \sigma_h^2\right.\\ &\quad + \left(1-\frac{1}{d_{hi}}\right)^2\left[K_h^2 E_m\left(J_{hi}\right) + E_m\left(J_{hi} d_{hi}^2 X_{hi}^2\right) - 2K_h E_m\left(J_{hi} d_{hi} X_{hi}\right)\right]\\ &\quad \left. + 2\left(1-\frac{1}{d_{hi}}\right)\left[K_h E_m\left(J_{hi} d_{hi} X_{hi}\right) - E_m\left(J_{hi} d_{hi}^2 X_{hi}^2\right)\right]\right\}\\ &\quad + \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)^2\left(1-\frac{1}{d_{hi}}\right)^2\left\{K_h^2 E_m\left(J_{hi}\right) + E_m\left(J_{hi} d_{hi}^2 X_{hi}^2\right)\right.\\ &\quad \left. - 2K_h E_m\left(J_{hi} d_{hi} X_{hi}\right) + \left[K_h E_m\left(J_{hi}\right) - E_m\left(J_{hi} d_{hi} X_{hi}\right)\right]^2\right\}\\ &\quad + \left\{\sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left[K_h E_m\left(J_{hi}\right) - E_m\left(J_{hi} d_{hi} X_{hi}\right)\right]\right\}^2.\end{aligned}$$

Given that the $d_{hi} X_{hi}$ are assumed to be independent and follow the same law within the strata, it is sufficient to consider a random variable Z_h that has the same law as one of the $d_{hi} X_{hi}$, , i.e., verifying:

- $E_m\left(Z_h\right) = \mu_h$;
- $\text{Var}_m\left(Z_h\right) = \sigma_h^2 < +\infty$;

- and that Z_h are independent and of density $g_h(x) > 0$.

Thus, we can also consider that a random variable $J_h = \mathbb{I}_{Z_h > K_h}$ to calculate the expectation with respect to the model of the winsorized indicator. The previous expression is rewritten:

$$\begin{aligned} E_m E_P\left[\left(\hat{\tilde{T}}(\tilde{X}) - T(X)\right)^2\right] = & \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\Bigg\{\mu_h^2 + \sigma_h^2 \\ & + \left(1-\frac{1}{d_{hi}}\right)^2 [K_h^2 E_m(J_h) + E_m(J_h Z_h^2) - 2K_h E_m(J_h Z_h)] \\ & + 2\left(1-\frac{1}{d_{hi}}\right)[K_h E_m(J_h Z_h) - E_m(J_h Z_h^2)]\Bigg\} \\ & + \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)^2\left(1-\frac{1}{d_{hi}}\right)^2\Big\{K_h^2 E_m(J_h) + E_m(J_h Z_h^2) - 2K_h E_m(J_h Z_h) \\ & - [K_h E_m(J_h) - E_m(J_h Z_h)]^2\Big\} \\ & + \left\{\sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)[K_h E_m(J_h) - E_m(J_h Z_h)]\right\}^2. \end{aligned}$$

A.2 Search for thresholds to minimize the MSE

To determine the value of the thresholds K_h leading to the optimum of $E_m E_P\left\{\left[\hat{T}(\tilde{X}) - T(X)\right]^2\right\}$, we use the same property as Kokic and Bell in their demonstration, i.e., that:

$$E_m\left(Z_h^p J_h\right) = \int_{K_h}^{+\infty} t_h^p\, g_h(t)\, dt,$$

and so that

$$\frac{\partial}{\partial K_h} E_m\left(Z_h^p J_h\right) = -K_h^p g_h(K_h).$$

By deriving relative to K_h, and after simplification, we obtain that:

$$\begin{aligned} \frac{\partial}{\partial K_h} E_m E_P\left\{\left[\hat{\tilde{T}}(\tilde{X}) - T(X)\right]^2\right\} = & \ 2B \times A_h E_m(J_h) \\ & + 2C_h\{[K_h E_m(J_h) - E_m(J_h Z_h)][1 - E_m(J_h)]\} \\ & + 2D_h[K_h E_m(J_h) - E_m(J_h Z_h)] + 2F_h E_m(J_h Z_h) \end{aligned} \qquad \text{(A.5)}$$

where

- $B = \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)[K_h E_m(J_h) - E_m(J_h Z_h)]$,
- $A_h = \sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)$,
- $C_h = \sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)^2\left(1-\frac{1}{d_{hi}}\right)^2$,

- $D_h = \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)^3,$
- $F_h = \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)^2.$

Equation (A.5) is reduced to:

$$\frac{\partial}{\partial K_h} E_m E_P \left[\left(\hat{\tilde{T}}\left(\tilde{X}\right) - T\left(X\right)\right)^2\right] = 0$$

$$\Leftrightarrow$$

$$A_h \times B \times E_m\left(J_h\right) + \left(C_h + D_h\right) K_h E_m\left(J_h\right) - C_h E_m\left(J_h\right)\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right] + \left(F_h - C_h - D_h\right) E_m\left(J_h Z_h\right) = 0.$$

Finally, by noting that $\left(F_h - C_h - D_h\right) = 0$ and assuming that $E_m\left(J_h\right) > 0,$ we obtain that the threshold K_h minimizing the MSE verifies the equation:

$$A_h \times B + \left(C_h + D_h\right) K_h - C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right] = 0$$

which is reduced further to

$$B + \frac{\left(C_h + D_h\right)}{A_h} K_h = \frac{C_h}{A_h}\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right].$$

It remains to be shown that $\frac{C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right]}{A_h B}$ tends toward zero when $n \rightarrow \infty.$ However,

$$\frac{C_h \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h B\right|} = \frac{C_h \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h \right| \sum_{l=1}^{H} \left| A_l \right| \left| K_l E_m\left(J_l\right) - E_m\left(J_l Z_l\right)\right|}$$

and according to hypothesis (2.8) relating to inclusion probabilities, we have that, $\forall h = 1, \ldots, H, \forall i \in U_h$ $d_{hi} > 1.$ Which implies $A_h > 0,$ and thus:

$$\begin{aligned}\frac{\left| C_h \right| \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h B\right|} &\leq \frac{C_h}{A_h^2} \\ &\leq \frac{\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)^2 \left(1 - \frac{1}{d_{hi}}\right)^2}{\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]^2} \\ &\leq \frac{1}{\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]}.\end{aligned}$$

However, it is possible to demonstrate from hypothesis (2.8) that $\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]^{-1} = O\left(\frac{1}{N_h}\right).$ Thus: $\frac{C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right]}{A_h B}$ tends toward zero when $n \rightarrow \infty.$

K_h is thus equivalent in each stratum to $-\frac{A_h}{\left(C_h + D_h\right)} B,$ when the size of the population and the sample tend toward infinity.

References

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

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

Clark, R.G. (1995). Winsorization methods in sample surveys. Master's thesis, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R & D Report, Statistics Sweden.

Demoly, E., Fizzala, A. and Gros, E. (2014). Méthodes et pratiques des enquêtes entreprises à l'Insee. *Journal de la Société Française de Statistique*, 155-4.

Deroyon, T. (2015). Traitement des observations atypiques d'une enquête par winsorisation : application aux Enquêtes Sectorielles Annuelles. *Actes des Journées de Méthodologie Statistique*.

Fizzala, A. (2017). *Adaptations of Winsorization Caused by Profiling - An Example Based on the French SBS Survey*. European Establishment Survey Workshop, Southampton.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Survey Methodology*, 41, 1, 57-77. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14199-eng.pdf.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scandinavian Journal of Statistics*, 43, 1019-1034.

Kokic, P.N., and Bell, P.A. (1994). Optimal winsorizing cut-offs for a stratified finite population estimation. *Journal of Official Statistics*, 10-4, 419-435.

Moreno-Rebollo, J.-L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika*, 86, 923-968.

Moreno-Rebollo, J.-L., Muñoz-Reyez, A.M., Jimenez-Gamero, J.-L. and Muñoz-Pichardo, J.M. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika*, 55, 209-214.

Rivest, L.-P., and Hurtubise, D. (1995). On searls' winsorized mean for skewed populations. *Survey Methodology*, 21, 2, 107-116. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995002/article/14399-eng.pdf.

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.