Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

A comparison of Kokic and Bell and conditional bias methods for outlier treatment

Prepared by Thomas Deroyon & Cyril Favre-Martinoz, Insee, France

INTRODUCTION

1. Following [Chambers, 1986]'s pioneering works, outliers in finite population sampling are usually defined as units belonging to one of two different groups:

- REPRESENTATIVE OUTLIERS are sampled units giving correct but atypic answers, without being absolutely unique.
- NONREPRESENTATIVE OUTLIERS are sampled units whose responses cannot be extrapolated to the rest of the population.

2. Nonrepresentative outliers are in most cases units giving wrong answers, introducing bias in estimates. They are identified at the data editing step where wrong values are replaced by imputations.

3. Representative outliers are units such that the estimates' values change a lot whether they belong to the sample or not. As pointed out by [Beaumont et al., 2013], representative outliers are linked to a configuration, consisting of four elements: a variable of interest, a parameter on this variable's distribution in the population, a sampling design and an estimate of the parameter. A unit may be an outlier for a parameter, such as a total on a specific domain, but not for another one.

4. Representative outliers do not introduce any bias in estimates, as data collected on them are perfectly correct, but they generate an increase in estimation variance. They are frequent in situations where sampling weights are highly scattered and lightly correlated with the variables of interest. But even if that correlation is high, a limited number of units departing from that relationship may cause an excessive increase in variance. In stratified random sampling for instance, precision depends on the responses' empirical dispersion in each stratum. Therefore, one observation whose value is very far from the stratum mean is sufficient to significantly alter the estimates' quality.

5. This type of outlier is frequent in business surveys, where variables' distributions are highly skewed and sampling weights therefore typically highly scattered. The better way to treat them is to avoid them, at the sampling design step, by carefully defining relevant homogeneous strata regarding the main variables of interest. This is however not always sufficient, as a survey may have multiple and lowly correlated variables of interest, as auxiliary information available in the sampling frame and used to design the strata may be of imperfect quality and as samples renewal strategies may generate strata jumpers, units sampled in the past with features that do no longer match their situation at the time of the survey. Treatments are therefore always necessary at the estimation step. These treatments

should be able to lower the variance of estimates affected by outliers without decreasing the precision of estimates not concerned with that problem.

6. Different methods exist in the statistical litterature and the goal of this paper is to compare the results of two of them and their output when they are applied to a survey with a complicated sampling design: historical winsorization techniques tailored to stratified random sampling and new techniques based on a more general measure of outlierness, the conditional bias. We compared them in a simulation study on real data designed to select the best method to deal with representative outliers in the French Wage Structure and Labor Cost Survey. Results show both techniques are able to identify the main outliers and lower their influence enough to significantly increase the estimates' precision.

7. We start by describing both methods; then we describe shortly the Wage Structure and Labor Cost Survey and especially its sampling design and the estimation configurations with which outlier treatments have to deal. To fit onto these configurations, we had to adapt both methods and make simplifying hypotheses we present then. We end by presenting the simulations and their results.

I. Outlier treatment methods

1. In the following section, we will assume we want to estimate Y, the total of variable y in the population U, and will use the usual expansion estimate $\hat{Y} = \sum_{i \in S} w_i \, y_i$, with w_i unbiased estimation weights, such as sampling weights in case of one-stage sampling design.

A. Kokic and Bell winsorization

1. Winsorization techniques consist basically in associating a threshold to each sampled unit. If the value of variable y is higher than the threshold, then it is modified, either truncated at the threshold or at least significantly lower than its original value. Kokic and Bell winsorization (see Kokic and Bell [2016]) applies to a specific framework:

- the sampling design has to be a stratified random sampling;
- the variable y has to be positive;
- observations of variable y in each stratum coming from another source than the survey sample, for instance a previous edition of the survey, are available.

2. A threshold K_h is associated to each stratum $h = 1..H$. With n_h and N_h the sample and population sizes in stratum h, the winsorized variable y^w is defined as:

$$y_i^w = \begin{cases} y_i & \text{if} \quad y_i \leq K_h \\ \frac{n_h}{N_h} \, y_i \, + \, (1 - \frac{n_h}{N_h}) \, K_h & \text{if} \quad y_i > K_h \end{cases} \tag{1}$$

3. The winsorized estimate $\hat{Y}^w$ is then equal to the expansion estimate of the winsorized variable total in the population. As an estimate of Y, $\hat{Y}^w$ is biased, but has a lower variance, as values of Y^w are less scattered in each stratum than values of Y. [Kokic and Bell, 2016] assume y observations in each stratum come from the same distribution and search for the optimum set of thresholds, that is the

thresholds such that $\hat{Y}^w$ has the lowest mean squared error, taking into account the sampling design and the distribution of y. Thresholds obtained with Kokic and Bell methods are therefore designed to protect the estimates against outliers on average on all possible populations and all possible samples taken from these populations.

4. [Kokic and Bell, 2016] show that, at the optimum, all thresholds are linked to a single parameter, the optimum winsorized estimate's bias B, which can be easily computed, thanks to observations of y values in each stratum independent from the sample. The Kokic and Bell winsorization methods has been used with profit since 2007 to treat outliers in the French Structural Business Statistics Surveys (see Deroyon [2015] for more details) and as an help to identify remaining outliers at the output editing step in the same survey (see Gros and Deroyon [2015]).

B. Conditional Bias methods

B.1. *Definition.*

1. Conditional bias was first introduced by Moreno-Rebollo et al. [1999] and Moreno-Rebollo et al. [2002] but Beaumont et al. [2013] first suggested to use it to design outlier-robust estimates in sampling from finite population. Conditional bias of sampled unit i for parameter θ's estimate $\hat{\theta}$ is defined as:

$$\mathbb{B}_{1i}(\hat{\theta}) = \mathbb{E}(\hat{\theta}/I_i = 1) - \theta$$

Where I_i is equal to 1 if unit i is sampled and 0 otherwise and expectation $\mathbb{E}$ is over sampling design. Conditional bias is therefore equal to the difference between the estimate's expectation on all samples containing unit i and the parameter value. [Beaumont et al., 2013] show it to be a direct measure of a sampled unit's outlierness as, in most one-stage sampling designs and if $\theta = Y$ and $\hat{\theta} = \hat{Y}$:

$$\mathbb{B}_{1i}(\hat{Y}) = \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_j \tag{2}$$

$$\mathbb{V}(\hat{Y}) = \sum_{i \in U} \mathbb{B}_{1i}(\hat{Y})\, y_i \tag{3}$$

$$\hat{Y} - Y \approx \sum_{i \in S} \mathbb{B}_{1i}(\hat{Y}) + \sum_{i \in U - S} \mathbb{B}_{0i}(\hat{Y}) \tag{4}$$

Where $\mathbb{B}_{0i}(\hat{Y}) = \mathbb{E}(\hat{\theta}/I_i = 0) - \theta$ is the difference between the estimate's expectation on all samples not containing unit i and the parameter value. Formula (4) directly show the estimation error is the sum of sampled and non-sampled units' conditional biases on the population so that a unit with high conditional bias increases errors. Formula (3) show that this increase of error is linked to an increase in variance.

2. As formula (2) shows, conditional bias depends on all observations from the population and cannot be computed with the informations available in the sample. It may however be estimated without any bias with the sample, as estimate

$$\hat{\mathbb{B}}_{1i}(\hat{Y}) = \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_j} y_j$$

is design-unbiased. In case of Poisson sampling, formula (2) simplifies to $\mathbb{B}_{1i}(\hat{Y}) = (w_i - 1)\, y_i$ so that conditional bias only depends on unit i.

B.2. *A design-robust estimate based on conditional bias.*

1. As formulas (3) and (4) prove, conditional bias directly measures the effect of each unit to the estimation error and to the estimation variance. A design-robust estimate should therefore be defined in such a way that all units have controlled and limited values of conditional bias. Elaborating on this basic principle, Beaumont et al. [2013] suggest to use an estimate of the type:

$$\begin{aligned}\hat{Y}^{CB}(c) &= \hat{Y} + \sum_{i\in S}\Psi_c[\hat{\mathbb{B}}_{1i}(\hat{Y})] - \sum_{i\in S}\hat{\mathbb{B}}_{1i}(\hat{Y}) \\ &= \hat{Y} - \sum_{i\in S}[\hat{\mathbb{B}}_{1i}(\hat{Y}) - \Psi_c(\hat{\mathbb{B}}_{1i}(\hat{Y}))]\end{aligned}$$

$$\text{with } \Psi_c \text{ Huber fonction defined by } \Psi_c(t) = \begin{cases} c & \text{if} \quad t \geq c \\ t & \text{if} \quad -c < t < c \\ -c & \text{if} \quad -c \leq t \end{cases}$$

Where Huber function is used to limit the influence of the most influential units by lowering their conditional bias and with $\hat{\mathbb{B}}_{1i}(\hat{Y})$ the design-unbiased estimate of conditional bias suggested by Beaumont et al. [2013]. Parameter c has to be tuned, according to a given objective. It can be set to lower $\hat{Y}^{CB}$ mean squared error but analytic formulas in that case can be obtained only on a limited set of configurations.

2. Beaumont et al. [2013] suggest to choose $c^* \in argmin_c\, argmax_i\, |\hat{\mathbb{B}}_{1i}(\hat{Y}^{BHR}(c))|$, that is the value of the tuning constant with which the highest value of the obtained estimate's conditional bias is the lowest. They show that, in that case, tuning constant c^{BHR} is such that:

$$\hat{Y}^{CB}(c^{BHR}) = \hat{Y}^{BHR} = \hat{Y} - \frac{min_i\hat{\mathbb{B}}_{1i}(\hat{Y}) + max_i\hat{\mathbb{B}}_{1i}(\hat{Y})}{2} \tag{5}$$

3. Beaumont, Haziza and Ruiz-Gazen design-robust estimate is therefore very easy to implement. Compared to Kokic and Bell method, it is more general and does not need any auxiliary information to be applied. However, it is not intended to lower the estimate's mean squared error and is only design-robust, as opposed to Kokic and Bell winsorized estimate, that also takes into account the variable of interest distribution.
The method has been extended to account for more elements of the sampling design or more general situations: Favre-Martinoz et al. [2016] extend the method to account for unit non-response treated with reweighting methods; Favre-Martinoz et al. [2015] suggest a way to obtain estimates' coherence when HBR method is used for the estimation of a variable of interest total in nested domains of the population.

II. Application to the Wage Structure and Labor Cost Survey

A. Introduction to the survey

A.1. *Sampling design.*

1. The Wage Structure and Labor Cost Survey (*Enquête sur le coût de la main d'œuvre et la structure des salaires, Ecmoss*) is designed to answer to European regulations 530/1999. The sampling design is a two-stage sampling, with stratified random sampling at each stage:

- FIRST STAGE: a sample of local units in the European regulation scope is selected in the business register, strata being defined by crossings of industries, location and sizes of local units

and legal units to which local units belong.

- SECOND STAGE: the first stage sample of local units is matched with social security files containing the list of each local unit's employees. A sample of employees is then selected in each sampled local unit, stratified according to occupation (manager vs non executive staff).

2. Each sampled local unit receives a questionnaire asking detailed information about its total wages and their decompositions. Questions about detailed elements of the sampled employees' wages are also asked. At each stage, unit nonresponse may occur: some local units fail to respond at all and some fail to respond on some of their employees. At each stage, nonresponse is treated as an additional poissonian sampling phase, with response probabilities estimated through homogeneous response groups.

A.2. *Parameters of interest.*

1. The main survey output is the estimation of hourly labor costs, averaged on domains of the population: industries at the Nace sections level, slices of legal units's size and locations at the Nuts2 level. Estimates used are based on ratios of total earnings and total number of hours worked:

$$\hat{R}(D) = \frac{\sum_{i \in S \cap D} w_i \, e_i}{\sum_{i \in S \cap D} w_i \, h_i} \tag{6}$$

With S the sample of employees, D a domain of interest, e_i yearly earnings of employee i, h_i his yearly number of worked hours and w_i the employees' estimation weight obtained by multiplying inclusion probabilities of each stage and phase of sample selection.

B. **How to adapt winsorization and conditional bias methods to the Ecmoss**

1. Estimate (6) is not the expansion estimate of a total: it does not match the configuration for which the Kokic and Bell and Beaumont, Haziza and Ruiz-Gazen estimates are designed. However, the problem can be easily adapted to fit the methods' framework.

2. Indeed, an unbiased estimate of $\sum_{i \in S} w_i \, \hat{L}_i(D)$'s variance, with $\hat{L}_i(D) = \frac{e_i - \hat{R}(D) \, h_i}{\sum_{i \in S \cap D} w_i \, h_i} \mathbb{I}(i \in D)$ the linearized variable's estimate of $\hat{R}(D)$, is also an (asymptotically) unbiased estimate of $\mathbb{V}(\hat{R}(D))$. As such, a robust to outliers estimate of linearized variable $\hat{L}_i(D)$'s total is very likely to be a robust to outliers estimate of $\hat{R}(D)$. Each method, applied to the estimated linearized variable, generates an outlier-robust version of it, noted $\hat{L}_i^w[\hat{R}(D)]$. The effects of outlier treatments are then transferred to all other variables through the weights, by computing a winsorized estimation weight:

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

We therefore apply Kokic and Bell method and Beaumont, Haziza and Ruiz-Gazen estimate to the estimation of $\hat{L}_i(D)$'s total. Each method however needs specific hypotheses to be applied to Ecmoss's sampling design.

B.1. *Kokic and Bell Winsorization.*

1. The method's framework does not match Ecmoss's configuration at all. First, Ecmoss' sample is not selected with a stratified random sampling design; Morevover, the variable to winsorize, the estimated linearized variable $\hat{L}_i(\hat{R}(D))$, cannot be always positive. We amend the method to adapt it to these discrepancies:

   (1) we apply the method as if employees were selected directly through a stratified random sampling design in strata defined by local units' industry, number of employees and aggregated location (with three modalities: Paris and its neighbourhood, rest of mainland France and ultramarine France); we therefore do not take into account the estimation weights' dispersion in each stratum, which is far from negligible: the risk is that the method misses some important outliers.

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

2. More precisely, we define for each sampled observation:

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

on which we compute Kokic and Bell's threshold and apply winsorization. Then we compute the winsorized estimation weight as:

$$w_i^w \;=\; w_i\,\frac{\hat{T}_i^w[\hat{R}(D)]}{\hat{T}_i[\hat{R}(D)]}$$

We adapt Kokic and Bell method in such a way that only outliers with high estimated linearized variable are identified and treated, that is employees whose hourly wage is superior to the average hourly wage in domain D. Units with lower values cannot be identified this way, but they also cause less problems in estimation, as hourly wages distribution is skewed to the right.

3. A final adaptation is needed to match the method's requirements and Ecmoss's configuration. Kokic and Bell method may indeed be used provided observations of the winsorized variable in each pseudo-strata are available. Preceeding editions of the survey may be used; however, tests to asses the method's usefulness on SBS surveys showed the risk in using past samples to compute the thresholds is loss of accuracy due to the limited number of available observations, entailing the winsorization of too many units. We therefore choose to use auxiliary information available in social security data on the wages paid yearly to each employee and their numbers of worked hours. These informations are not those measured in the survey, but the variables needed by social security administration to determine each local unit's social taxes. They are however well correlated with the survey's variable and may result in good accuracy of outlier-robust estimates.

B.2. *Beaumont, Haziza and Ruiz-Gazen estimate.*

1. The conditional bias estimate needs less modifications to fit Ecmoss' configuration: for example, it can be applied directly to the variables of interest measured in the survey. Computation of the conditional bias taking into account the whole sampling design is however complex, so that we choose to model it as a one-stage poissonian sampling design, with selection probabilities equal to $1/w_i$. The conditional bias used in outlier's identification is therefore equal to:

$$\mathbb{B}_{1i}\{\hat{L}_i[\hat{R}(D)]\} \;=\; (w_i - 1)\,\hat{L}_i[\hat{R}(D)]$$

2. With formula (5), the Beaumont, Haziza and Ruiz-Gazen estimate only modifies conditional bias for a limited number of sampled observations: the ones with the lowest and highest conditional biases, whose outlier-robust estimation weight is equal to:

$$w_i^{BHR} = \begin{cases} \frac{(2\,|A_{min}|-1)\,w_i+1}{2|A_{min}|} & \text{if } \mathbb{B}_{1i}\{\hat{L}_i[\hat{R}(D)]\} \in A_{min} \\ \frac{(2\,|A_{max}|-1)\,w_i+1}{2|A_{max}|} & \text{if } \mathbb{B}_{1i}\{\hat{L}_i[\hat{R}(D)]\} \in A_{max} \\ w_i & \text{otherwise} \end{cases}$$

with $A_{min} = argmin_{j \in S}\, \mathbb{B}_{1j}\{\hat{L}_j[\hat{R}(D)]\}$, $A_{max} = argmax_{j \in S}\, \mathbb{B}_{1j}\{\hat{L}_j[\hat{R}(D)]\}$ and $|A|$ is set A's cardinal.

3. Compared to Kokic and Bell estimate, conditional bias outlier-robust estimate does not concentrate on the variable's distribution right tail but has an effect on both tails. It also concentrates *a priori* on a limited number of observations as only the observations with the highest and lowest conditional biases are affected.

B.3. *Outlier-robust estimates for estimation on multiple domains.*

1. As described in section A.2, domains of interest are numerous. It causes two types of difficulties for outlier-robust estimates:

- OUTLIER-ROBUST ESTIMATES FOR MORE THAN ONE SET OF DOMAINS

2. European regulations demands dissemination of results on non-nested sets of domains, such as industry sections and numbers of employees' brackets. Sampled observations thus belong to more than one dissemination domain.

3. The best way to treat outliers is to apply methods for each domain of interest, which entails a sampled observation may be associated with a different outlier-robust estimation weight for each dissemination subpopulation of interest. That solution is however hardly conceivable for users, who will possibly have to deal with a different estimation weight for each configuration (domain, variable of interest or parameter).

4. Another solution is to apply both methods for estimations on domain's crossings. The risk, as outlier identification is carried out on small size subpopulations, is to identify too many outliers, as far as estimation on the actual dissemination domains is concerned. Estimates will be optimal on the outlier-identification domains, but may be too biased for dissemination domains.

- OUTLIER-ROBUST ESTIMATES FOR ALL THE DOMAINS OF A SET

5. For a given set of domains (for example industry sections), a unit may be identified and treated as an outlier for the estimation in more than one domain, resulting in more than one outlier-robust estimation weight. It is especially the case if the selection of an observation belonging to a domain is not independant from the selection of units belonging to other domains, as with stratified random sampling when domains intersect strata.

6. This cannot happen with the way we implement Beaumont, Haziza and Ruiz-Gazen estimate, describing sampling design as a single stage poissonian one. It may however be the case with Kokic and Bell method, as some dissemination domains cannot be formed as groupings of strata. The situation is consequently the same as the one exposed with different sets of domains: the only way to maintain a single estimation weight for each sampled observation is to apply methods on outlier identification domains defined as groupings of strata.

7. In the simulations we present in next section, both strategies are implemented, to test whether the use of specific outlier identification domains differing from dissemination domains does not entail too big a decrease of precision's gains.

III. Simulations

A. Simulation design

1. Simulations are carried on social security data available in the whole population and on which we can therefore assess estimates' bias. Simulation design is the following:

- Ecmoss' sampling design (including selection of responding local units and employees) is repeated 5 000 times to produce 5 000 samples of employees S_m, $m = 1..5000$;
- we compute the usual expansion estimates of average hourly wages $\hat{R}_m(D)$ by domain on each sample ;
- Kokic and Bell winsorization and conditional bias methods are applied to each sample with different possible specifications:
  - Kokic and Bell winsorization is applied only as if domains of interest were outlier identification domains defined by crossings of dissemination domains.
  - Conditional bias method is applied for each possible dissemination domain separately (industries, location, number of employees) on the one hand, and as if outlier identification domains were the real dissemination domains on the other. For each dissemination domain, we are thus able to compare the results of Beaumont, Haziza and Ruiz-Gazen method applied in its optimal design for that domain (giving estimate $\hat{R}_m^{BHR*}(D)$ of average hourly wages) to conditional bias and Kokic and Bell methods (giving estimates $\hat{R}_m^{BHR}(D)$ and $\hat{R}_m^{KB}(D)$) applied in a sub-optimal but more convenient way.

2. For each outlier-robust estimate and each domain, we compute average and relative bias and average mean squared error as:

$$
\begin{aligned}
AB[\hat{R}^{KB}(D)] &= \frac{\sum_{m=1}^{5000}[\hat{R}_m^{KB}(D) - R(D)]}{5000} \\
AMSE[\hat{R}^{KB}(D)] &= \frac{\sum_{m=1}^{5000}[\hat{R}_m^{KB}(D) - R(D)]^2}{5000} \\
RB[\hat{R}^{KB}(D)] &= 100\,\frac{AB[\hat{R}^{KB}(D)]}{R(D)} \\
RMSE[\hat{R}^{KB}(D)] &= 100\,\frac{AMSE[\hat{R}^{KB}(D)]}{AMSE[\hat{R}(D)]}
\end{aligned}
$$

for instance with Kokic and Bell estimate, $R(D)$ the average hourly wage observed in the population in domain D and $\hat{R}(D)$ the usual expansion estimate of that parameter. Relative bias compares the outlier-robust estimates' bias to the real value of the parameter. Relative mean squared error measures the gain or loss in precision brought by outliers' treatment methods.

B. **Simulation results**

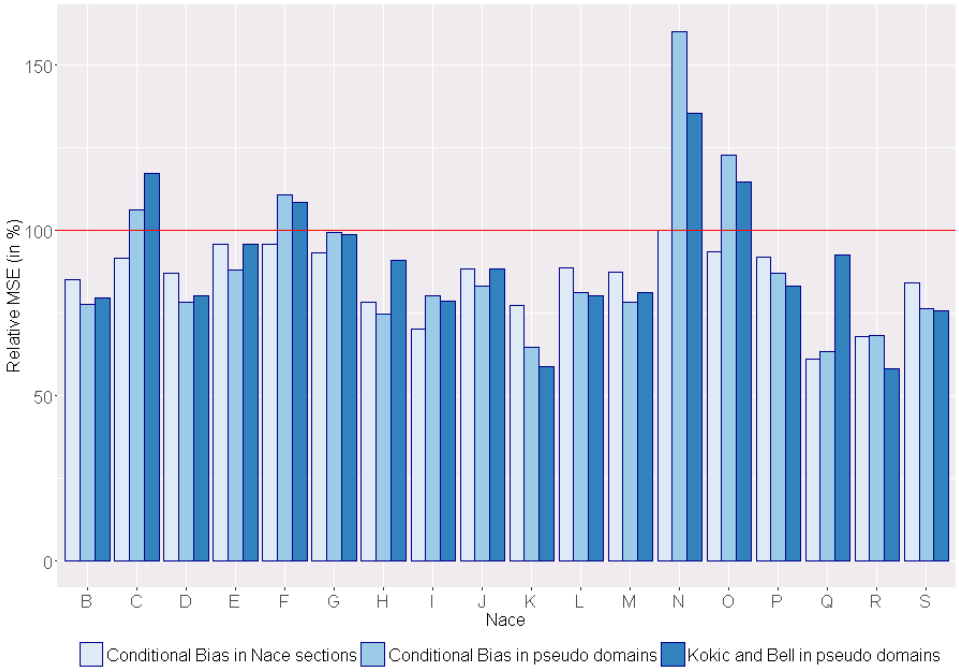

FIGURE 1. Relative MSE of average hourly wages'estimates by Nace sections

1. Figure 1 shows mean squared error of standard and outlier-robust estimates in each Nace section and figure 3 shows the distribution of mean squared errors in each domain [1] For almost all domains of

[1] among all nace sections, all crossings of nace sections and number of employees' slices and all crossings of nace sections and nuts.

interest, outlier-robust estimates have lower MSE than the standard expansion estimate. The domains for which outlier treatments increase estimation errors are those on which the estimation variance is the lowest. The different outlier treatments are therefore able to lower estimates variance when needed without causing too much decrease in precision when estimates are not affected by outliers.

2. Biases in Nace sections estimates are low (figure 2), except for some domains where the sample size is small [2]. The same result is obtained on other domains of interest.

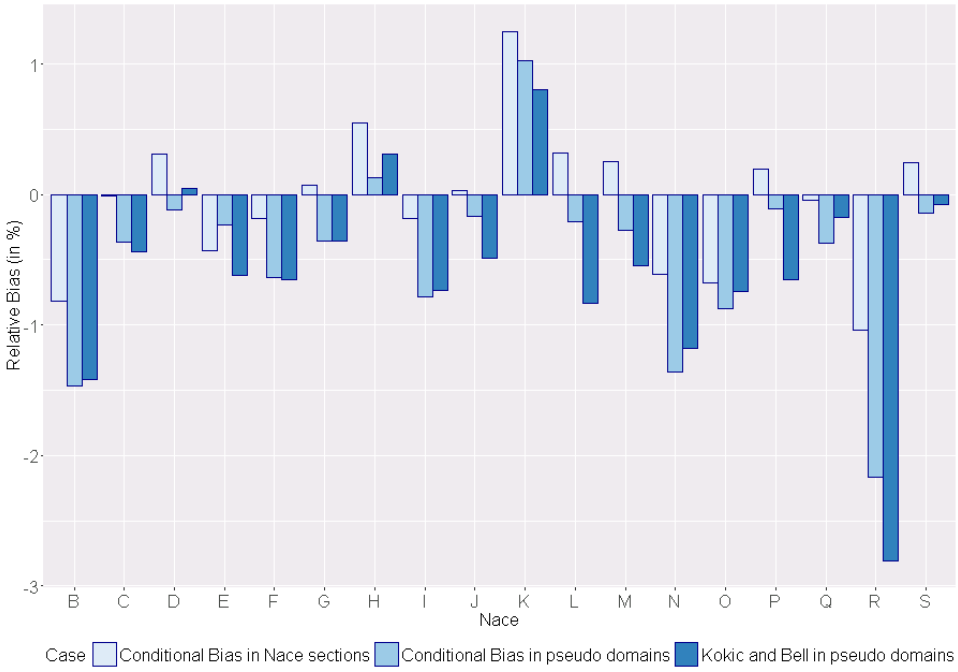

FIGURE 2. Relative Bias of average hourly wages'estimates by Nace sections

3. Application of conditional bias on each domain gives better results than conditional bias or Kokic and Bell method applied on outlier-identification domaines (named peudo-domains in the figures) in Nace sections, but not always for other domains of interest. Nace sections are much more aggregated than outlier-identification domains, so the estimation bias caused by outliers treatment is higher in that case. In the other domains, identifying outliers at a more detailed level enables to treat more units and so to lower more variance, without introducing too much bias. The differences between the real sampling design and the way it is described to apply the treatments may explain the reasons why application of conditional bias on each estimation domain does not always entail better precision.

4. Differences between conditional bias and Kokic and Bell method outputs are low, even if conditional bias applied on outlier-identification domains seem to perform a little better, as it is able to take into account the estimation weights dispersion whereas Kokic and Bell method is compelled to treat them as identical in each pseudo-strata.

[2] A: agriculture, forestry and fishing; K: financial and insurance activities; R: arts, entertainment and recreation.

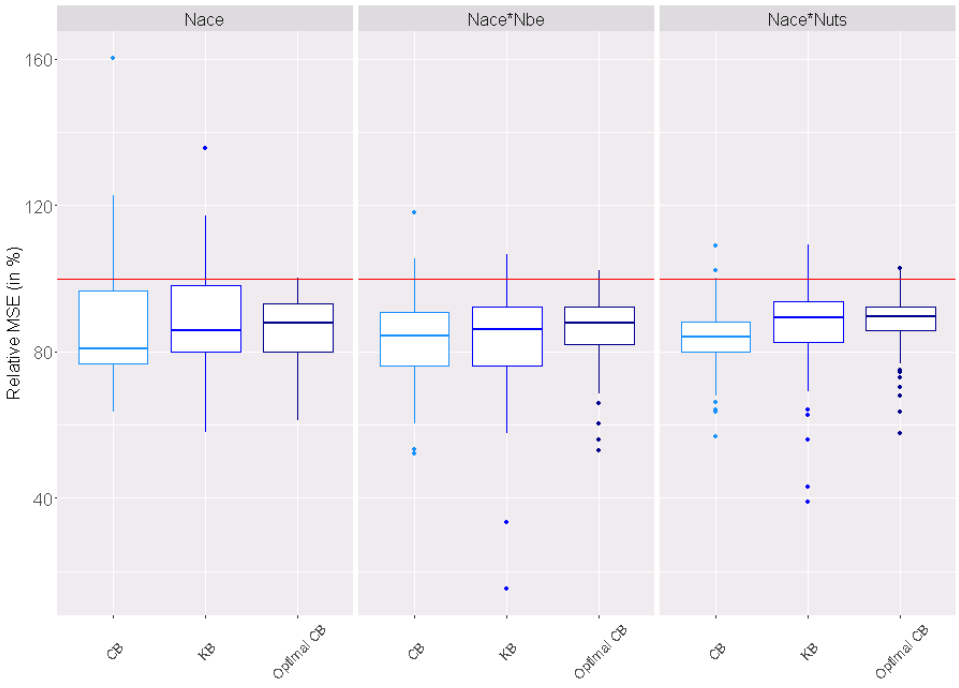

FIGURE 3. MSE's Distribution of average hourly wages' estimates by domains

IV. Conclusion

1. Outlier treatments method applied on Ecmoss's data are able to lower estimates' mean squared error, even if the hypotheses needed to match their framework with the survey's sampling design are strong. Both methods, conditional bias and Kokic and Bell winsorization, give similar results. An outlier treatment will be applied each year to Ecmoss's data and will be followed by a calibration on margins computed on the business register and on social data to lower the introduced estimation bias, which can be non negligible for certain small aggregated domains.

References

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

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

T. Deroyon. Traitement des observations atypiques d'une enquête par winsorisation : application aux enquêtes sectorielles annuelles. In *Actes des Journées de Méthodologie Statistique*. Insee, 2015.

C. Favre-Martinoz, D. Haziza, and J.F. Beaumont. A method for determining the cut-off points for winsorized estimators with application to domain estimation. *Survey Methodology*, 41:51–77, 2015.

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

E. Gros and T. Deroyon. Output editing based on winsorization in the french sbs multisource system esane. In *Work Session on Statistical Data Editing, Budapest*. Unece, 2015.

P.N. Kokic and P.A. Bell. Optimal winsorizing cut-offs for a stratified finite population estimation. *Scandinavian Journal of Statistics*, 43:1019–1034, 2016.

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

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