Working Paper 18

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (i): Selective and macro editing

Output editing based on winsorization in the French SBS multisource system Esane

Prepared by Thomas Deroyon and Emmanuel Gros, Insee, France

I. Introduction

1. Outliers are sampled units whose responses differ from the responses of units having the same sampling weights, for example belonging to the same stratum in stratified sampling. Outliers, which can correspond to true but atypical values or to wrong answers, cause greater variance for estimators based on survey data.

2. Winsorization is a robust estimation technique aiming at identifying a certain type of outliers and limiting their effect on estimators' variance. In case of stratified sampling, thresholds are defined for each sampling stratum: the responses higher than the thresholds are shrinked. The estimators obtained with this method are biased, but may be more precise. Kokic and Bell [1] suggested a method to compute threshold guaranteeing a gain in precision for estimators.

3. This method is used since 2009 to identify and treat outliers in data used in the French Structural Business Statistics production system, Esane. As statistics produced in the Esane device are subject to many consistency constraints – especially consistency constraints between estimations relating to variables linked by accounting relationships – that the estimation method has to respect, the winsorization in Esane is based only on fiscal turnover: the adjustment winsorization causes on this variable is passed on all other fiscal variables. This method enables to identify each year a limited number of outliers on turnover and to treat their responses, with a significant impact on SBS aggregates precision.

4. However, such a method based on only one core variable does not allow to identify units with average turnover but with unusual values on other fiscal variables, especially on those poorly correlated to turnover. In order to make up for this drawback, we present here an application of the Kokic and Bell winsorization on other variables than the turnover, as an element of the output-editing process.

5. The first part of this paper details the Esane device and the estimation procedure used to produce sector-based estimates. The second part present the Kokic and Bell winsorization procedure. The third part presents the way winsorization is used in Esane as a part of the post-data collection treatment process and the way we plan to use it as part of the output editing process.

II. Main outlines of the new system

A. Structure of the data

6. The system Esane (in French "Élaboration des Statistiques Annuelles d'Entreprises") relies on a combined use of different administrative sources and a statistical survey (figure 1). Two administrative sources are used:

- Annual income returns of enterprises to tax authorities, containing accounting variables;
- Annual social security returns, containing information about employment and wages.

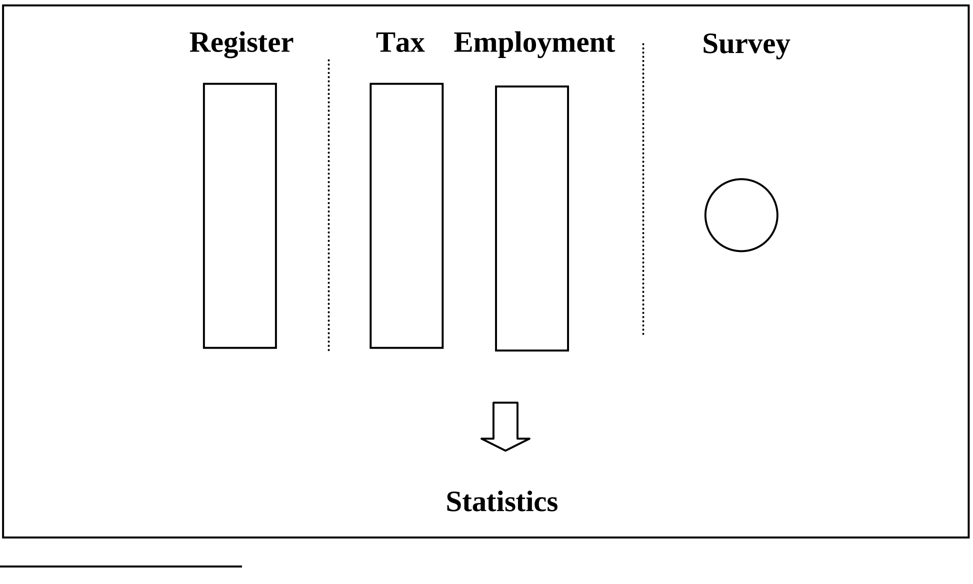
All these data are theoretically complete[1] and the record linkage is made easy with the id-number of the French business register SIRENE. The statistical unit that is used is the legal unit[2] as defined in this register.

7. However, the use of administrative data alone is not sufficient to produce the required statistics, on account of the lack of some variables. Especially, one essential piece of information, which is the basis for all sector-based statistics, is not available in the administrative sources: the breakdown of enterprise turnover. This information has two main uses. First, the national accounts need information about the homogeneous branches turnover, which is obtained through this table. Secondly, the breakdown of the turnover is used to compute the value of the principal activity code (in French, "APE" code), referring to the French classification of activities NAF (which is derived from the European NACE). This value of the APE code is obtained through an algorithm, based on the relative share of each component of the turnover. This value is used to produce the aggregates for the economic sectors, and may differ from the value available in the register, which may have not been updated recently.

8. In order to make up for the incompleteness of administrative sources, a statistical survey, called ESA (in French "Enquête Sectorielle Annuelle", i.e. Insee Annual Sectoral Survey on Businesses, see [2] for further details), is conducted on a sample of enterprises. This survey is the French most important survey of businesses. It is carried out in six of the main production sectors[3]. It comprises two parts for each sector:

- A core section, which includes in particular enterprise turnover and its breakdown by different activities at a very detailed level, questions on employment and legal restructuring;
- A sectoral part of the questionnaire relating to characteristics that are specific to a given sector: e.g. sales area for enterprises in the trade sector, spending of fuel for enterprises in the transport sectors, etc.

Figure 1: The different components of the new system of structural business statistics



[1] In practice, we however have to deal with some missing data.
[2] Except for specific units that will be defined for some large groups, for which profiling techniques are used.
[3] Industry excluding food-processing industry, food-processing industry, transport, construction, trade and services excluding banks and insurance companies.

B. Statistical estimates dedicated to the device

9. Thus, we may consider that we have an incomplete rectangular data base: a nearly complete data base for the administrative data, and a part available only from the survey data. Now some of these variables – mainly the principal activity code and the breakdown of turnover – are cornerstone information. In order to take into account at best this information available only on the survey while using all information available in the administrative sources, we use both calibration techniques and specific estimators.

10. First, having the administrative data available allows us to use calibration techniques ([3]) as part of the surveys post-data collection treatment process. Calibration is applied after surveys have been treated for unit non-response and outliers have been identified and treated via winsorization, as we will see in part IV. Calibration leads to modifying the weights according to some calibration equations. More precisely, the equations used here are:

$$\begin{cases} \sum_{i \in R} w_i T^{tax}(i)\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U} T^{tax}(i)\, \mathbb{1}_{APE\_rep=X}(i) \\ \sum_{i \in R} w_i\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U} \mathbb{1}_{APE\_rep=X}(i) \end{cases}$$

where APE_rep is the value of the APE code within the register, and $T^{tax}(i)$ is the value of the turnover of the enterprise i in administrative data. The calibration on the turnover (first equation) uses a "3-digit" level for the sectoral classification, whereas the calibration on the number of enterprises (second equation) uses a "2-digit" level, in order to limit the range of changes of the weights.

11. Moreover, for sector-based estimates, the existence of two APE codes – the one of the register (APE_rep) and the one coming from the survey (APE_enq) – leads us to consider the following difference estimator for the total of Y:

$$\sum_{i \in R} w_i Y_i\, \mathbb{1}_{APE\_enq=X}(i) + \sum_{i \in U} Y_i\, \mathbb{1}_{APE\_rep=X}(i) - \sum_{i \in R} w_i Y_i\, \mathbb{1}_{APE\_rep=X}(i) \qquad \text{(A)}$$

This kind of estimators can be computed for any variable Y available for all units, such as tax and employment variables coming from administrative data source. As the variables $Y_i\, \mathbb{1}_{APE\_enq=X}(i)$ and $Y_i\, \mathbb{1}_{APE\_rep=X}(i)$ are usually well correlated and even often almost identical, this difference estimator is particularly appropriate to the Esane device, and usually improves[4] the quality of sector-based estimates[5].

III. Winsorization: principles and Kokic & Bell method

A. Theoretical framework and principles

12. The winsorization following the method of Kokic and Bell (see [1]) applies to stratified simple random sampling without replacement design. The population U, of size N, is divided into H strata U_h of size N_h. In each stratum, a sample s_h of size n_h is selected by simple random sampling without replacement, the drawings of the s_h being mutually independent.

[4] Compared with the calibrated estimator.

[5] Nevertheless, we had to face in practice to problems of wrongly negative values with these difference estimators. This kind of problems arises occasionally for estimations relating to variables concerning very few units, and more frequently for estimations at detailed level. Consequently, it has been decided to use difference estimators for estimations at the "3-digit" level of the French nomenclature (and upper), and to use different kind of estimators for estimations relating to more detailed level (see [4] and [5] for further details).

In this framework, the Horvitz-Thompson estimator for the total of any variable of interest Y is:

$$\hat{Y} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right) \sum_{i \in s_h} y_{hi}$$

with y_{hi} the value of variable Y for unit i in stratum h.

This estimator is unbiased, and its variance is equal to $V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$, with $S_h^2 = \frac{\sum_{i \in U_h} (Y_i - \overline{Y}_h)^2}{N_h - 1}$ and $\overline{Y}_h = \frac{\sum_{i \in U_h} Y_i}{N_h}$.

If a stratum contains some outliers, it means that a few units in the stratum present unusually large values of variable X compared to the mean of Y in the stratum, which leads to large value of S_h^2 and directly deteriorate the accuracy of $\hat{Y}$.

13. Winsorization is a method allowing to deal with such influential values. It involves decreasing the value of outliers in order to reduce their impact on estimates accuracy. More precisely, winsorization consists in defining a threshold K_h for each stratum h and then decreasing the value of units in stratum h that are above this threshold K_h, taking their sampling weight into account. The values of the variable of interest after winsorization are defined by:

$$y_{hi}^{w} = \begin{cases} y_{hi} & \text{if} \quad y_{hi} < K_h \\ \frac{n_h}{N_h} y_{hi} + \left(1 - \frac{n_h}{N_h}\right) K_h & \text{if} \quad y_{hi} \geq K_h \end{cases}$$

The winsorized estimator is then defined as the Horvitz-Thompson estimator of the winsorized variable Y^w:

$$\hat{Y}^{w} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right) \sum_{i \in s_h} y_{hi}^{w}$$

Contrary to the Horvitz-Thompson estimator, the winsorized estimator is biased. On the other hand, its variance depends on the winsorized variable Y^w dispersion in each stratum h, which is by construction smaller than the one of the original variable Y. So, winsorization consists in a bias-variance trade-off: it will be efficient if the introduced bias is more than offset by the lower variance of the winsorized estimator, that is if the mean squared error of the winsorized estimator is lower than the one of the Horvitz-Thompson estimator.

14. The choice of the thresholds K_h is crucial, as they determine the quality of the winsorization procedure: a poor choice may lead to winsorized estimators that have a larger error than classical estimators. The problem of choosing these thresholds has been studied by Kokic and Bell in 1994, which proposed a method to determine, under some assumptions, optimal thresholds – that is thresholds that minimizes the estimated mean square error of the winsorized estimators.

B. The Kokic & Bell thresholds

15. Winsorizing presupposes identifying outliers, and thus making assumptions about the distribution of the variable of interest allowing to distinguish outliers from “normal” values. So, Kokic and Bell set up a common mean model (denoted ξ) in each stratum, that is assume that, within each stratum h, the values of the variable of interest is a sequence of independent and identically distributed random variables with

mean μ_h and standard deviation σ_h. They also suppose the interest variable Y is always positive, and the thresholds are determined independently from the sample and the sampling design.

16. Thresholds are then calculated by minimizing the winsorized estimator's mean square error with respect to both the model ξ and the sampling design. Under the previous assumptions, Kokic and Bell established the following result: for the optimum set of thresholds which minimizes the mean square error of the winsorized estimator, the negative bias of this estimator is asymptotically equivalent to $-(N_h/n_h-1)(K_h-\mu_h)$, whatever the stratum. As this bias is an easily computable function of the thresholds K_h, the determination of the optimal thresholds amounts to solve the following system of H equations $B((K_h)_{h=1,..,H})+(N_h/n_h-1)(K_h-\mu_h)=0$. Noting $L=-B$, the opposite of the bias, we can then easily show that the resolution of such a system amounts to deduct the threshold of the identity $K_h=(N_h/n_h-1)^{-1}L+\mu_h$, whith L solution of the equation F(L)=0 with

$$F(L)=L\left(1+\sum_{h=1}^{H}n_h E_\xi(I_h^{\otimes})\right)-\sum_{h=1}^{H}n_h E_\xi(y_{hi}^{\otimes}I_h^{\otimes})$$

$$y_{hi}^{\otimes}=(N_h/n_h-1)(y_{hi}-\mu_h)\ \text{ and }\ I_h^{\otimes}=\begin{cases}1 \text{ if } y_{hi}^{\otimes}\geq L\\ 0 \text{ if } y_{hi}^{\otimes}<L\end{cases}$$

17. To compute the zeroes of the function F, we need to estimate μ_h and the values of $E_\xi(I_h^{\otimes})$ and $E_\xi(y_{hi}^{\otimes}I_h^{\otimes})$ depending on L. If we have a set of realizations $(y_{hi})_{1\leq i\leq m_h}$ of the variable Y, μ_h, $E_\xi(I_h^{\otimes})$ and $E_\xi(y_{hi}^{\otimes}I_h^{\otimes})$ will be estimated by:

$$\hat{\mu}_h=\bar{y}_h=\frac{1}{m_h}\sum_{i=1}^{m_h}y_{hi},\ \hat{I}_h^{\otimes}=\begin{cases}1 \text{ if } \hat{y}_{hi}^{\otimes}\geq L\\ 0 \text{ if } \hat{y}_{hi}^{\otimes}<L\end{cases}$$

$$\hat{E}_\xi(I_h^{\otimes})=\frac{1}{m_h}\sum_i \hat{I}_h^{\otimes}\ \text{ and }\ \hat{E}_\xi(y_{hi}^{\otimes}I_h^{\otimes})=\frac{1}{m_h}\sum_i \hat{y}_{hi}^{\otimes}\hat{I}_h^{\otimes}$$

Kokic and Bell propose calculating thresholds K_h in a way that they are independent from the sample to which they are applied – so they can be used for any values of the winsorized variable in the sample. Consequently, the current sample data can not be used as the set of realizations $(y_{hi})_{1\leq i\leq m_h}$ of the variable Y needed to compute the thresholds. are then calculated by minimizing the winsorized estimator's mean square error with respect to both the model and the sampling design. The authors suggest, for repeated surveys, using historical data collected in previous iterations.

IV. Winsorization of Esane surveys as part of the output editing process – the principles

18. Esane estimators quality is likely to be affected by outliers, as often in business surveys:

- The main variables of interest are fiscal variables, that is numeric variables with highly skewed distributions, where the values for a limited part of the population are very different from that of the rest of the population.

- In Esane, fiscal variables are available[6] for all firms in the population of interest, but up-to-date sectoral classification is only available for units in the sample, as explained in part I. So, Esane difference estimates of fiscal variables sectoral totals, such as the total value added of real estate activities, depend on the sampled units fiscal variables. If the sample contains outliers, the precision of Esane estimates may be deteriorated.

[6] Non-response in the fiscal and social security sources is treated through imputation: fiscal variables for active businesses that did not send back their annual income declarations for example are mainly imputed based on their preceding years declarations. Therefore, fiscal variables are available for all firms in the population of interest, whether real or imputed.

- Everything is done at the sampling design stage to protect the estimators against outliers: ESA are sampled with a one-stage stratified sampling, strata being mainly defined by crossings of sector at the 5-digit level of NAF and groups of number of employees. All firms with more than 20 employees are placed in take-all strata, as are businesses with less than 20 employees whose turnover is greater than a threshold depending on the sector. This is intended to create homogeneous strata with regard to the fiscal variables of interest.

- However, the information on each firms' number of employees available at the time the sampling frame for Esane surveys is built is still provisional. Especially, some firms whose information have not been gathered yet by social security administrations are treated as if they have no employee, and are therefore placed in the same stratum as very small businesses.

- Moreover, as Esane system is designed to accurately estimate the levels as well as the growth rates of sectoral fiscal aggregates, only half of the ESA sample is renewed every year, whereas the other half is maintained in the survey. The maintained part was therefore selected in strata corresponding to features that may not match any longer the firms' actual characteristics. This renewal strategy may then generate a special type of outliers, known as strata jumpers.

19. For all these reasons, outliers identification and treatment have been since the beginning a major interest in Esane system. Outliers may correspond to two situations (see [6] for more details):

- Non-representative outliers, mainly observations with wrong or absurd answers.

- Representative outliers, mainly observations whose answers strongly differ from those of the firms in the same stratum.

At the data editing step, micro and selective editing procedures were designed to identify wrong or spurious data in the administrative sources as well as in the survey. The presumed errors with major effect on aggregates are submitted to detailed checks, other errors are treated with automatic corrections (see [7] for more details). At the end of this stage, the main errors are supposed to have been identified and treated.

20. However, data editing may fail to detect some errors: some non-representative outliers may still hamper the estimates quality. Moreover, data editing is not meant to treat representative outliers. A specific outliers identification and treatment step via winsorization was therefore added in the Esane surveys post-data collection treatment process. Esane surveys are first treated for unit non-response, with imputation techniques in take-all strata and reweighting through homogeneous response groups (HRG) in take-some strata. Then, winsorization thresholds are calculated in each take-some strata with the Kokic and Bell method and applied to detect and treat outliers among respondents. Finally, weights are adjusted via calibration to perfectly match estimates of the sectoral number of firms and their aggregate turnover.

21. The Kokic and Bell method was chosen because:

- Esane surveys sampling design is very close to the framework to which Kokic and Bell method applies, that is one-stage stratified sampling. We however have to make simplifications to use Kokic and Bell formulas. First, we do not take into account the non-response treatment through HRG. Although HRG differ from sampling strata, we indeed act as if respondents where directly selected through a simple random sampling in each stratum. Secondly, we take no account of the renewal strategy, owing to which all units in a stratum do not have the same sampling weight, depending on the year they entered the sample.

- As fiscal variables are available for all units in the population, we have realizations of the variable independent from the sample to estimate the winsorization thresholds for each fiscal variable we want to winsorize.

22. SBS aggregates are linked by numerous relationships, mainly accounting principles or definitions. If winsorization thresholds were defined and applied for each variable independently, these

relationships would no longer hold for the winsorized units in most cases. That is the reason why the choice was made to base the winsorization on a core variable, the firm's turnover in the fiscal source. More precisely, each year, in each sector at the three-digit level of the NACE, winsorization thresholds are computed thanks to the Kokic and Bell method to detect and treat outliers, that is influential observations with respect to the estimation of the total sectoral turnover. These thresholds are applied to all take-some strata in the sector. The correction on turnover this winsorization entails is then applied to all other fiscal variables and all survey variables.

23. This method is expected to identify and treat well outliers for variables strongly correlated to turnover. But outliers may remain after this step for other variables. Indeed, the subject-matter experts implementing the output editing procedure before SBS aggregates dissemination report constant problems with some aggregates, such as total sales or purchases in real estate activities. Output editing procedures are based on a top-down analysis of growth contributions. In this procedure, three partitions of the economy in sectors are considered: a partition in seventeen sectors (called A17), a partition in 38 sectors (called A38) and the partition obtained with the first 3 digits of the NACE. For the main variables of interest, the aggregates in the A17 sectors are first analysed and spurious growth rates identified. For each of these sectors, the A38 sectoral aggregates whose contributions to the spurious A17 sectoral aggregate growth rates are the highest are identified. Then, for each of these A38 sectors, the 3-digit sectors whose contribution to the A38 sectoral aggregates growth rates are the highest are identified. Finally, in each of the identified 3-digits sectors, units whose contributions to the aggregate growth rate are the highest are identified and checked. Once the A17 sector aggregates are validated, the output editing procedure is applied to the A38 sectoral aggregates, then to the 3-digits sectoral aggregates. Output editing procedures are also applied to the aggregates levels. This process cannot be fully automated and is time consuming. Winsorization may however be used as a tool to fasten this process.

24. Even if winsorization for variables other than the turnover cannot be fully implemented due to the constraints accounting principles entail, it can still be used to identify potential outliers, that is units that *would be* winsorized, if winsorization for these variables was applied. The effect winsorization would have on the units answers and therefore on the aggregates can also be used as measures of outlierness. These potential outliers, especially those with a high measure of outlierness, are likely to have a strong impact on the aggregates and their growth rates, that is to belong to the list of units identified by the output editing process. Winsorization may therefore be treated as a "shortcut" automatically pointing to a sub-list of problematic units, that is units that will have to be checked in the output editing process.

25. To be helpful to subject-matter experts in the output editing process, the measure of outlierness defined by the effect of winsorization on aggregates has to be computed on the actual disseminated aggregates. However, winsorization with the Kokic and Bell method is meant to identify and treat outliers with regard to the classical Horvitz-Thompson estimator of the total of a variable of interest. In Esane, the variables of interest are fiscal or employment variables multiplied by sectoral dummies. However, as explained in part I, Esane aggregates of sectoral totals are not Horvitz-Thompson estimators, but difference estimators that enable us to benefit from the fact fiscal variables are known for all units in the population and whose formula is equation (A) page 3. Let's define Y the variable of interest in the output editing process, *i* a potential outlier identified through winsorization, w_i its calibrated weight, Y_i the value of Y for i and Y_i^W the value that would be obtained if winsorization of Y was applied. Then, the effect of unit *i* winsorization on sectoral totals of Y is given by:

$$\Delta_i^W(Y) = w_i(Y_i^W - Y_i)(\mathbb{1}_{APE\_enq}(i) - \mathbb{1}_{APE\_rep}(i))$$

Two situations can occur:

- The APE in the register is identical to the APE updated in the survey ⇒ winsorization of unit *i* has no effect on any sectoral total (and growth rate) of the Y variable.

- Register and survey APE differ ⇒ unit *i* winsorization affects two sectoral totals of Y: the one of the register APE, with an effect equal to $-w_i(Y_i^W - Y_i)$, and the one of the survey APE, with an effect equal to $w_i(Y_i^W - Y_i)$. If Y is a positive variable, and as $Y_i^W < Y_i$, the effect on the register APE sectoral aggregate is positive, whereas the effect on the survey APE sectoral aggregate is negative.

The lists sent to subject matter experts focus on winsorized units having a non-zero effect on difference estimates: units whose survey and register APE are equal are not mentioned. For each unit in the list, the effects on sectoral difference estimates at the A17, A38 and 3-digit level of the NACE are given for register and survey APE sectors, as measures of outlierness.

25. Winsorization as a tool in the output editing process will be used starting in 2015, for the 2014 annual estimation campaign, according to the following principles:

- Two groups of variables of interests are defined with subject matter experts. So far, the first group gathers 21 variables belonging to the profits and loss accounts (total of sales, total of purchases, salaries, total of exportations) and employment variables. The other group so far gathers 13 variables about investments and assets.

- Winsorization thresholds are calculated for each of these variables according to the same method as the one applied to turnover. For each of the variables groups, a list of units is produced and sent to the subject-matter experts. Each list contains the id-numbers of the units that would be winsorized, with the sectors on whose aggregates winsorization would have an effect, and the effect winsorization would have.

- For each variable of interest and sector with suspect growth rate or level, these lists will directly point potential outliers, with a measure or their outlierness that can be used as a score to define the order in which these units will be checked, or to identify the units it is worth checking.

26. In the next part, we present the results we obtained by applying this method to the 2013 aggregates. As the method was not implemented during the actual estimation campaign, it could not be used as an help in the output editing procedure. These results however give indications about the sectoral aggregates for which strong potential outliers are identified, and the numbers of these candidates for editing.

V. Winsorization of Esane surveys as part of the output editing process – first results

27. The winsorization procedure applied to the 2013 sectoral aggregates would have identified 1 161[7] units whose winsorization would have changed the A17 sectoral aggregates for at least one variable of the first group, and 473 units for the variables of the second group. Tables 1 and 2 sum up the effect winsorization would have on sectoral totals at the A17 level for four variables of interest: sales and purchases (with changes in inventories) in table 1, and value-added and salaries in table 2.

28. Sectoral aggregates with strange evolutions or levels often concern variables secondary to the activity of the sector. For instance, totals of sales are often problematic in sectors such as real estate activities, services or transportation, in which the major part of the turnover is represented by production activities, whereas they show credible trends in wholesale and retail trade. In trade, all firms indeed contribute to the sales aggregate. It is thus unlikely that a firm can be influential in an aggregate to which all firms make substantial contributions.

29. Table 1 tends to confirm this intuition. Winsorization on sales on the one hand and purchases and changes in inventories on the other would show very strong outliers in real estate activities, construction, mining and quarrying, agriculture and services. For instance, the winsorization of 3 units would increase the estimated total of sales in real estate activities by 17 %. In construction, the winsorization of 10 units

[7] Among approximately 45 000 responding units in the take-some strata.

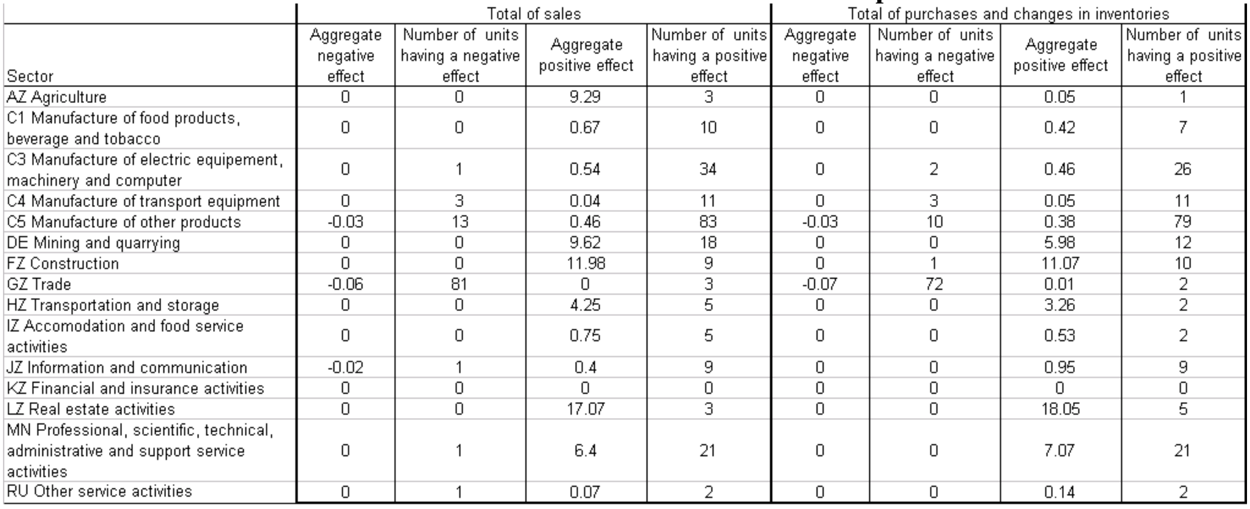
would increase the total of purchases and changes in inventories by 11 %. In fact, 3 units among the 10 concentrate the major part (80 %) of this effect. The major effects on aggregates are positive: they are caused by firms that were classified in services, construction or real estate activities in the register but that belong to trade according to the survey. They have a great effect on their register sector aggregate, but a very limited negative effect on the trade aggregate.

30. Winsorization has however a small effect on total value-added or salaries and may work poorly to identify outliers on these variables. Two reasons may explain this phenomenon:

- First, as with sales in trade, a major share of firms in each sector pay salaries and have a value-added. It is thus more difficult to be an outlier for an aggregate to which almost all firms in the sample contribute.

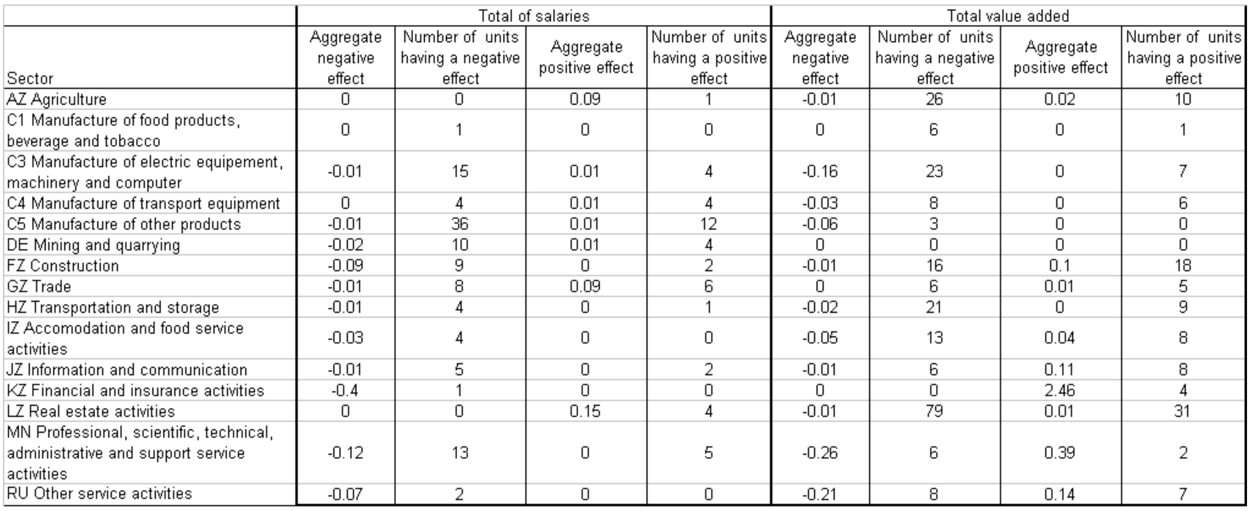
- Secondly, value-added and salaries are usually highly correlated to turnover. The variables' values on which the winsorization as an output editing technique is applied are the values contributing to the disseminated aggregates, that is the values after the winsorization on the turnover. It is therefore possible that the major outliers for value-added or salaries have already been identified and treated as the turnover winsorization step.

Table 1: Effect of winsorization on total sales and purchases



	Total of sales				Total of purchases and changes in inventories			
Sector	Aggregate negative effect	Number of units having a negative effect	Aggregate positive effect	Number of units having a positive effect	Aggregate negative effect	Number of units having a negative effect	Aggregate positive effect	Number of units having a positive effect
AZ Agriculture	0	0	9.29	3	0	0	0.05	1
C1 Manufacture of food products, beverage and tobacco	0	0	0.67	10	0	0	0.42	7
C3 Manufacture of electric equipement, machinery and computer	0	1	0.54	34	0	2	0.46	26
C4 Manufacture of transport equipment	0	3	0.04	11	0	3	0.05	11
C5 Manufacture of other products	-0.03	13	0.46	83	-0.03	10	0.38	79
DE Mining and quarrying	0	0	9.62	18	0	0	5.98	12
FZ Construction	0	0	11.98	9	0	1	11.07	10
GZ Trade	-0.06	81	0	3	-0.07	72	0.01	2
HZ Transportation and storage	0	0	4.25	5	0	0	3.26	2
IZ Accomodation and food service activities	0	0	0.75	5	0	0	0.53	2
JZ Information and communication	-0.02	1	0.4	9	0	0	0.95	9
KZ Financial and insurance activities	0	0	0	0	0	0	0	0
LZ Real estate activities	0	0	17.07	3	0	0	18.05	5
MN Professional, scientific, technical, administrative and support service activities	0	1	6.4	21	0	0	7.07	21
RU Other service activities	0	1	0.07	2	0	0	0.14	2

Table 2: Effect of winsorization on total value added and salaries



	Total of salaries				Total value added			
Sector	Aggregate negative effect	Number of units having a negative effect	Aggregate positive effect	Number of units having a positive effect	Aggregate negative effect	Number of units having a negative effect	Aggregate positive effect	Number of units having a positive effect
AZ Agriculture	0	0	0.09	1	-0.01	26	0.02	10
C1 Manufacture of food products, beverage and tobacco	0	1	0	0	0	6	0	1
C3 Manufacture of electric equipement, machinery and computer	-0.01	15	0.01	4	-0.16	23	0	7
C4 Manufacture of transport equipment	0	4	0.01	4	-0.03	8	0	6
C5 Manufacture of other products	-0.01	36	0.01	12	-0.06	3	0	0
DE Mining and quarrying	-0.02	10	0.01	4	0	0	0	0
FZ Construction	-0.09	9	0	2	-0.01	16	0.1	18
GZ Trade	-0.01	8	0.09	6	0	6	0.01	5
HZ Transportation and storage	-0.01	4	0	1	-0.02	21	0	9
IZ Accomodation and food service activities	-0.03	4	0	0	-0.05	13	0.04	8
JZ Information and communication	-0.01	5	0	2	-0.01	6	0.11	8
KZ Financial and insurance activities	-0.4	1	0	0	0	0	2.46	4
LZ Real estate activities	0	0	0.15	4	-0.01	79	0.01	31
MN Professional, scientific, technical, administrative and support service activities	-0.12	13	0	5	-0.26	6	0.39	2
RU Other service activities	-0.07	2	0	0	-0.21	8	0.14	7

31. The real test for the use of winsorization as part of the output editing process will however take place in late autumn 2015, when SBS experts will validate a first set of SBS aggregates used by the French national accounts division. In this process, they will have for the first time the list of potential outliers identified by winsorization as a guide to help them in their output editing process.

VI. References

[1] P.N. Kokic, P.A.Bell, Optimal winsorizing cut-offs for a stratified finite population estimator, Journal of Official Statistics, 1994.

[2] O. Haag, Reengineering French structural business statistics : redesign of the annual survey, paper presented at the Q2010 conference, Helsinki, 2010.

[3] J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, Journal of the American Statistical Association, 87, pp. 376-382, 1992.

[4] Ph. Brion, A first assessment of the French system of production of structural business statistics combining administrative data and survey data, Proceedings of the Fourth International Conference on Establishment Surveys, June 11–14, 2012, Montréal, Canada, 2012.

[5] E. Gros, Esane ou les malheurs de l’estimation composite : comment gérer les valeurs négatives d’estimateurs par différence, Paper presented at the « Journées de méthodologie statistique », Insee, 2012.

[6] R. Chambers, Outlier robust finite population estimation, Journal of the American Statistical Association, 1986.

[7] E. Gros, Assessment and improvement of the selective editing process in Esane (French SBS), Work Session on Data Editing, Oslo, 2012.