

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (viii): Selective and macro editing

**USEFULNESS OF ESTIMATION ON TARGET VARIABLES DURING THE DATA
COLLECTION PROCESS AND PRIORITIZING FOLLOW-UPS**

Supporting Paper

Prepared by Benoît Buisson, INSEE, FRANCE

I. INTRODUCTION

1. In ad hoc business surveys, there is a growing perception of the need to calculate estimators for target variables during the survey itself. This estimation requires adjusting certain variables for total and partial non-response while the survey is in progress. The calculation of target variables during the survey feeds a dashboard, which is examined by the survey designers. The dashboard serves two main purposes during data collection: to spot outlier data and to take into fuller account the effects of non-response adjustment. We can thus not only prioritize checks, but also define methods for prioritizing follow-ups of non-responding units. Part II of this document discusses the method for calculating estimators during the survey and their use. Part III describes the implementation of a method for prioritizing follow-ups.

2. We seek to show the value of determining target-variable estimators during a survey, which requires **adjusting for total and partial non-response during the survey itself**, rather than at the end of the survey as is done in some cases. We also describe the implementation **of a method for prioritizing follow-ups for total non-response**, a method defined and applied by the Australian Bureau of Statistics.

II. ESTIMATORS AND CALCULATING PRECISION DURING A SURVEY

A. Why calculate estimators during a survey?

3. During the execution of a business survey by mail—and, optionally, by online data collection—it is critically important **to obtain maximum information** from businesses at this specific point in the statistical production process. Traditionally, once the data have been collected, i.e., when survey clerks have moved on to another operation, it is very unusual for the statistical agency to chase responses—in particular during the determination of the final aggregate. In the collection phase, it is therefore essential to obtain maximum information while targeting contacts to **concentrate resources** where they will be most useful.

4. Generally speaking, a vital step when targeting contacts is to **calculate estimators for certain target variables** early enough in the collection process, notably to assess the influence of individual data on these aggregated estimators. To determine these estimators during collection, we must try “at all times” to act as if we had to disseminate “final” estimators and so to assume that the collection was ending at that moment. This requires us to **adjust for partial non-response**, at least for the variables chosen as target variables, to **adjust for total non-response**, and, when appropriate, to reweight

variables. Naturally, the further along we are in the collection process, the more reliable the estimates become.

5. In sum, calculating estimators during the survey turns out to offer multiple benefits, including:
 - Targeting checks on responding businesses that most strongly influence the estimate of the aggregate of target variable(s)
 - Spotting outlier or “aberrant” data
 - Providing a better measure of the influence of imputed data on the estimate of the aggregate
 - Calculating precision indicators for the estimators, which may allow us to adopt a “stop-collection criterion” approach.

B. How does one calculate estimators during a survey?

6. The estimators will be calculated from data that, by definition, have been partly validated. As a rule, keying checks and automatic checks have already been performed, but manual checks that may require follow-ups are in progress.

7. **Non-response adjustment** is essential at this stage, as it will substantially influence estimator values. What is the most reliable way to adjust for **partial non-response** early in the survey? For recurrent surveys, **we can simply use last year’s adjustment model** (determined on the previous complete sample). For example, in the 2006-2007 INSEE Survey on Information and Communication Technologies (hereafter: ICT Survey), we adjusted the percentage of online sales by random imputation, according to the sector and size of the business. The **situation is more complex for non-recurrent surveys** or the first generation of a recurrent survey. For the Waste Survey, for example, we proposed another approach. The variable of interest was the tonnage of non-hazardous waste produced by private-sector businesses. The analysis of respondents revealed linear relations (by group of economic activities) between tonnage and business size, size being a variable included in the sampling frame. These linear relations, which obviously became more specific as the collection moved ahead, were applied to non-respondents.

8. It is also essential to **adjust for total non-response**. In ad hoc business surveys, total non-response has generally been adjusted by **reweighting** after identification of **homogeneous response groups (HRGs)**. That is how we calculated e-commerce estimators in the 2007-2008 ICT Survey—using the HRGs identified in the previous year. Although not exactly identical from year to year, the HRGs identified are reasonably similar from one survey generation to the next. The process seems more reliable than looking for HRGs during the survey, as may be necessary in non-recurrent surveys. To estimate target variables during the survey, we effectively treat non-returns as non-responses. Only after the collection has ended, using the “final” non-returns, do we conduct searches in external sources to determine deceased or out-of-scope businesses. This may be regarded as a limitation on the method, but conducting checks based on external sources for a large number of firms has been ruled too costly.

C. An example from the ICT Survey

9. The table below reproduces **the dashboard** used to track collection progress:

Amounts in €billion

	Dec. 7	Dec. 20	Jan 4	Jan. 17	Feb. 1	Feb. 15	Feb. 29	Mar. 14	Mar. 28	file clearing	imputed file
Online sales	49.9	51.2	57.6	51.0	52.7	50.3	50.1	52.8	53.7	57.6	59.8
EDI sales	189.3	275.2	258.3	244.4	234.0	247.7	235.9	238.0	239.9	256.8	243.8
<i>Total e-sales</i>	<i>239.2</i>	<i>326.4</i>	<i>315.9</i>	<i>295.5</i>	<i>286.8</i>	<i>298.0</i>	<i>286.0</i>	<i>290.9</i>	<i>293.6</i>	<i>314.4</i>	<i>303.6</i>
Online purchases	45.1	46.0	65.1	39.2	38.6	40.6	43.5	43.4	48.0	51.9	57.6
EDI purchases	101.3	121.9	111.0	185.4	134.0	141.2	145.7	145.9	149.6	152.3	155.1
<i>Total e-purchases</i>	<i>146.5</i>	<i>167.9</i>	<i>176.1</i>	<i>224.7</i>	<i>172.6</i>	<i>181.8</i>	<i>189.2</i>	<i>189.3</i>	<i>197.7</i>	<i>204.2</i>	<i>212.7</i>
Variation coefficient,											
%											
Online sales	21.6	20.1	16.4	15.7	13.9	12.2	10.0	9.3	9.1	8.1	
EDI sales	10.5	25.4	21.4	14.8	11.1	11.8	4.4	4.3	4.2	4.0	
<i>Total e-sales</i>	<i>10.9</i>	<i>22.1</i>	<i>18.2</i>	<i>12.8</i>	<i>9.6</i>	<i>10.2</i>	<i>4.2</i>	<i>4.0</i>	<i>3.9</i>	<i>3.6</i>	
Online purchases	24.9	21.3	34.1	11.0	8.5	5.9	9.9	9.8	9.7	6.1	
EDI purchases	33.0	20.0	16.2	32.5	11.7	9.7	6.4	6.3	6.0	5.9	
<i>Total e-purchases</i>	<i>29.3</i>	<i>18.2</i>	<i>17.1</i>	<i>27.2</i>	<i>10.2</i>	<i>7.8</i>	<i>5.9</i>	<i>5.9</i>	<i>5.6</i>	<i>4.8</i>	
response rate	36%	46%	57%	65%	76%	81%	83%	84%	84%	85%	

10. The calculation of variance therefore takes into account non-response (total non-response only), but under the simplifying assumption that the HRGs consist of the sampling strata. By using such a dashboard, we can not only track collection but also **spot outlier values rather quickly**: these are values that require prompt adjustment. Two examples are early-January online purchases and mid-January EDI purchases. The EDI-purchase estimator accordingly jumps from €11 billion on January 4 to €85 billion on January 17, with the variation coefficient rising from 16 to 32 in the same interval. Concurrently with this dashboard, we calculated the “**contribution**” of each “responding” business (i.e., responding to the questionnaire) to each e-commerce aggregate. The computation is very rudimentary: we multiply the figure for the business by its weight (after adjustment for total non-response), then divide the result by the aggregate concerned. We therefore answer the question: what would have happened if the business in question had provided a zero value for the target variable involved? This approach is not equivalent to determining what would have happened if the business had not answered the questionnaire, for in that case we would have adjusted by reweighting. Calculating contributions had two purposes. The first was to identify raw figures reported by highly influential businesses and to alert the survey administrators, so as to obtain confirmation of the figures. Previously, we had calculated contributions at the tail end of the process (after imputation/reweighting), when chasing responses by telephone had become a very delicate task. The second purpose was to identify the influence of adjustment for partial non-response on the aggregate concerned. If a figure adjusted by random imputation makes a large contribution, we phone back the firm to obtain the answer to that particular question. The calculation also serves as a warning signal for us so as to ensure the most reliable imputation for this type of firm, should its response be still missing at the end of the survey process.

11. In sum, calculating estimators during the survey offers many advantages, particularly the targeting of follow-ups for partial non-response. In III.B, we describe a method for prioritizing follow-ups for total non-response, based indirectly on the calculation of target-variable estimators. Beforehand, in §III.A, we recall some approaches already used to prioritize follow-ups for total non-response.

III. PRIORITIZING FOLLOW-UPS FOR TOTAL NON-RESPONSE

A. Different approaches for prioritizing follow-ups

12. Prioritizing follow-ups (contacts of non-respondents) has become a **major aspect** of survey management, particularly for business surveys. Ever tighter management schedules, coupled with constraints on human resources to perform checks and follow-ups, add up to a strong case for contacting only those entities that are deemed to have “priority” status. Later on, we shall discuss prioritization of telephone response-chasing, for all non-responding firms are contacted via automatic postal-mail procedures whose “density” varies according to whether the survey is compulsory or not.

13. A common approach consists in **prioritizing follow-ups on the basis of auxiliary variables**, generally derived from the sampling frame. The decision may or may not be taken to weight these auxiliary-variable values by the entities’ initial weights. For business surveys, for example, we can use sales or number of employees multiplied by sampling weight. The order of priority is therefore static, being defined once and for all at the start of the survey, and is updated only with the list of respondents. Ideally, of course, the chosen auxiliary variable should be very closely related with the variable of interest in the questionnaire; this is not always the case, for example with e-commerce. The reason is that some medium-sized entities (such as group subsidiaries) may carry out 100% of their sales or purchases electronically—unlike very large enterprises. Whatever the link between auxiliary variable and variable of interest, **this approach takes neither collection progress nor adjustment for total non-response into account**. For example, a business belonging to a HRG with a low response rate may prove to be more important to follow up than a firm that is “larger” as measured by the initial weighted variable. The approach is perfectible, notably as it makes no allowance for collection progress and results.

14. A far more interesting approach, from this standpoint, is the “**expected variance loss**” approach developed by Philippe Brion. In what follows, we summarize and comment on his 2007 working paper listed in our bibliography. Brion sets out to target follow-ups on the strata for which we expect a **sharp reduction in variance** between the current collection-in-progress situation and the ideal situation with zero non-response. For stratified simple random sampling—a method used for many business surveys—the variance of the estimated total for a variable Y is:

$$V(\hat{T}(Y)) = \sum_{h=1}^k N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_h^2$$

where h is the stratum index; N_h is the number of firms in stratum h ; n_h is the number of firms surveyed in stratum h ; S_h^2 is the dispersion of variable Y in stratum h .

If we have collected only r_h questionnaires instead of the expected n_h in stratum h , we can proceed with reweighting. Brion shows that if we reweight uniformly in each sampling stratum, and assuming that non-respondents in each stratum exhibit no specific characteristics, the variance of the estimated total, for the share relating to stratum h , becomes $V_h(\hat{T}(Y)) = N_h^2 \left(1 - \frac{r_h}{N_h}\right) \frac{1}{r_h} S_h^2$.

If $x_h = \frac{r_h}{n_h}$ is the rate of questionnaires used in stratum h , the loss of variance is therefore

$\Delta V_h = \frac{N_h^2 S_h^2}{n_h} \left(\frac{1}{x_h} - 1\right)$ if we reweight. For each variable of interest, we can thus calculate the quantity

$\sum_h \Delta V_h / V(\hat{T}(Y))$, where $V(\hat{T}(Y))$ is the expected variance, with the n_h surveyed questionnaires in each stratum. **The quantity calculated is the “expected variance loss.”** As the survey proceeds, the loss diminishes, reaching zero if we could retrieve and process all questionnaires. To prioritize follow-ups, we can look at the share of expected variance loss due to each stratum via the different variables ΔV_h . This gives us indications about the **priorities** to assign in processing each stratum: we must begin by focusing

on the strata for which ΔV_h is greatest. We can therefore obtain a **stratum priority ranking**, by sampling stratum. This approach is effectively **dynamic**. It complements and goes further than the often-used approach that consists in **prioritizing follow-up of strata with low response rates**, notably as it takes into account the dispersion of the variable of interest in the stratum. We need to estimate this dispersion, which can be a tricky procedure especially at the start of collection. For a recurrent survey, it may therefore be helpful to use the year-earlier S_h^2 value, at least when survey processing begins. For a non-recurrent survey, S_h^2 will be estimated from questionnaires already received and processed. However, this estimation can be fairly unstable. Philippe Brion points out that the approach proves most relevant for small and medium-sized enterprises. Dealing with very large missing entities calls for a different approach involving a visit by an interviewer and the use of external sources.

B. Method used for ICT and Waste Survey

15. As shown above, the “expected variance loss” method focuses on the variance of the final estimator. The method proposed here is based on the estimator itself rather than its variance. We take up Richard McKenzie’s approach (see bibliography) and apply it to the reweighting procedure, with some simplifying assumptions.

16. Let Y be the quantitative variable of interest. Absent non-response, we obtain Y_1 as estimator of Y . In this case, we assume that all sampled firms have responded and that their responses are deemed “valid.” With non-response, before targeted telephone follow-ups, we obtain Y_2 as estimator of Y . The estimator incorporates the non-response adjustment, which we assume to have been performed by reweighting; the strata used for the adjustment are not defined *a priori*. For a recurrent survey, the strata may consist of the homogeneous response groups identified in the previous wave. For a non-recurrent survey, the strata may consist of the sampling strata, although we need to update them by analyzing the firms’ response behavior during the survey. With non-response and after a targeted telephone follow-up of a particular firm (a follow-up that, we assume, will succeed with a probability of 1), we obtain Y_2^R as estimator of Y . We also assume that units not followed up have not responded. We can consider that we choose to **follow up the firm that minimizes the distance** $F_R = (Y_2^R - Y_1)^2$. Let $F_2 = (Y_2 - Y_1)^2$; we further assume that the absolute value of the gap between Y_2 and Y_1 exceeds the absolute value of the gap between Y_2 and Y_2^R . This hypothesis, which does not seem very restrictive, may be problematic at the very start and the very end of the collection. However, those are not the phases when “prioritizing follow-ups” is crucial. Under the hypothesis, minimizing the distance F_R is equivalent to maximizing the value $(Y_2 - Y_2^R)^2$ under the constraint that $F_2 - F_R > 0$. We must therefore move away from Y_2 as far as possible while moving closer to Y_1 (a graphic analysis offers an immediate illustration). **We therefore wish to maximize the absolute value** $(Y_2 - Y_2^R)$ under constraint. We have:

$$Y_2 = \sum_{rep} w_i^* Y_i = \sum_{rep} \sum_{strate} w_i^* Y_i \text{ and } Y_2^R = \sum_{rep+1strate} w_i^{**} Y_i.$$

Here the strata consist of the strata used for non-response adjustment, i.e., the homogeneous response groups (HRGs). To calculate the difference $(Y_2 - Y_2^R)$, we must bear in mind that it is zero for all strata that do not contain the unit to be followed up. The reason is that the respondents—and the weights—in these strata are the same before and after follow-up.

Consider stratum s , which contains the unit to be followed up. We have:

$$(Y_2 - Y_2^R) = \sum_{rep} w_i^* Y_i - \sum_{rep+1} w_i^{**} Y_i = -w_{rel}^{**} Y^{rel} + \sum_{rep} Y_i (w_i^* - w_i^{**}); \text{ with}$$

$$w_i^* = w_i * \frac{n_s}{r_s}$$

$$w_i^{**} = w_i * \frac{n_s}{r_s + 1}.$$

r_s denotes the number of respondents in the stratum (HRG) of the unit to be followed up; n_s the number of respondents units sampled in this same stratum; w_{rel} the initial weight of the unit to be followed up. We thus obtain:

$$(Y_2 - Y_2^R) = -w_{rel} \frac{n_s}{r_s + 1} Y^{rel} + \frac{n_s}{r_s(r_s + 1)} \sum_{rep}^s w_i Y_i = \frac{n_s}{r_s + 1} \left[\left(\frac{1}{r_s} \sum_{rep}^s w_i Y_i \right) - w_{rel} Y^{rel} \right] \quad (\mathbf{A})$$

17. It is important to realize that when we target the telephone follow-up, Y^{rel} is unknown. **To prioritize follow-ups, we must therefore estimate Y for non-respondents.** In a recurrent survey, we can estimate Y for units to be followed up from the response to the previous survey (if the unit responded to that survey). In a non-recurrent survey (or for a unit that did not respond in previous waves), we need to use a model approach to obtain an initial estimate of the value of the variable of interest for non-responding units. How should we deal with units that responded to the questionnaire but failed to provide an answer for the variable of interest (i.e., partial non-response for the variable of interest)? In the previous formula, we needed values for the target variable for all respondents (or estimates for partial non-respondents). We therefore propose adjusting for partial non-response at this level, by means of a random imputation mechanism. For recurrent surveys, this may involve re-using the model for partial non-response adjustment chosen in the previous survey. For non-recurrent surveys, we must consequently suggest—then refine—a model for partial non-response adjustment for the target variable during the survey. This may be a delicate task, especially at the start of the process. It is worth emphasizing that this model should diverge as little as possible from the method used to estimate the target variable for total non-respondents. Despite the difference in purpose, the same model can be used—a choice that will save time. A particularly interesting case is the one where **non-response adjustment is performed on the sampling strata**. Here, as the non-response adjustment strata coincide with the sampling strata, the w_i values in formula (A) are equal: they consist of the inverse of the stratum sampling rate. Formula (A) therefore becomes:

$$(Y_2 - Y_2^R) = \frac{n_s w_s}{r_s + 1} \left[\left(\frac{1}{r_s} \sum_{rep}^s Y_i \right) - Y^{rel} \right].$$

In business surveys, we commonly have **several quantitative target variables**. Beyond prioritizing the target variables, it may be useful to develop a single list of priorities combining all target variables. This may be particularly helpful for survey clerks. We can operate on **ranks**. We can define a follow-up priority for each target variable. By summing the ranks, we can define an order of priority for this rank sum.

18. To prioritize follow-ups, we have simply adopted one criterion here: to converge as closely as possible toward the estimator obtained in the absence of non-response, which is generally bias-free. As a result, in the prioritization process, **the priority firms to be followed up are those most distant from the average in their stratum**. We can clearly see that this approach is justified under the chosen criterion, but it raises a problem with respect to the variance estimator. By “seeking outliers” we increase the sample variance, and so probably **overestimate the population variance**. This creates a risk that the population variance, hence the confidence intervals, may be poorly estimated or over-pessimistic.

C. Implementing the method used for ICT and Waste Survey

19. As regards **the operating mode**, we must therefore allow the calculation of term (A) for non-responding units, i.e.:

- Calculate the number of respondents in the non-response adjustment HRGs
- Estimate the target variable for all non-respondents (no doubt the most delicate step)
- Estimate the target variable for partial non-respondents.

In practice, for each target variable, once we have prioritized follow-ups and sent the list to the survey clerks, we must **monitor target-variable estimation during the survey**, giving special attention to major variations. For each unit potentially designated for follow-up, we can determine **the estimator’s variation if the unit were to respond consistently with the expected figure**. Tracking the estimator’s

variation over time—for example, for the first unit to follow up—yields valuable information. We naturally expect the variation to decrease.

20. We first applied the method to **the Waste Survey in early 2007**. The survey’s main aim was to measure the tonnage of non-hazardous waste generated by business establishments. This was the first time that the survey was conducted in France—which raised **two difficulties** in regard to the application of the method. First, we needed to estimate total non-hazardous waste for non-respondents. For this purpose, we prepared a model for respondents, correlating the number of local-unit employees and the waste tonnage emitted, by type of activity. Second, we had to **“assume” an *a priori* non-response adjustment model at the start of the survey**. We therefore posited that the homogeneous response groups (HRGs) were identical to the sampling strata. For lack of time, we maintained this hypothesis throughout the survey application. After the end of the survey management phase, during non-response adjustment, we realized that we needed to take the geographic criterion into account in defining the HRGs—especially owing to the low response rate in large cities, in particular among retailers. The differences between the HRGs adopted after non-response adjustment and the HRGs defined *a priori* could **undermine the method**. In practice, we sent priority lists to each survey clerk every week during the survey management phase. The operation was well received by the clerks. **However, they were “disturbed” by the method’s “dynamic” aspect**. The prioritization of local units that remained non-respondents could change over time, a situation that was not always easy for clerks to interpret.

21. We applied the method a second time in **the 2009 ICT Survey**. It is arguably a recurrent survey, for fully two-thirds of the questions are asked annually. This is an undeniable advantage over the Waste Survey. Survey designers chose online sales as the variable of interest. As we conducted a survey on the same topic in the previous year, several points are worth emphasizing. For the firms present in both samples, we take last year’s sales estimates as the estimated values for this year. They may consist of last year’s raw values or values obtained after non-response adjustment. The estimates are of use when the unit concerned is potentially designated for follow-up (total non-response) or when it has not supplied answers for the target variables (partial non-response). To **adjust for partial non-response on e-sales** by other firms, we chose the partial non-response adjustment model used for the same variable in the previous survey. To **estimate e-sales for total non-respondents** not questioned in the previous wave, we decided to apply the same partial non-response adjustment model defined in the previous survey. In this prioritization of follow-ups, we frequently **analyzed sales estimators**. We also analyzed the impact of the first unit to be followed up on the estimators. Accordingly, we sent follow-up priority lists to survey clerks every two weeks. In practice, the operation went very well. It allowed us to improve the quality of the online-sales estimator by targeting checks and follow-ups and by engaging in greater dialogue with firms on the subject during collection. However, there is a need **to provide clerks with support for this type of operation**.

IV. CONCLUSION

22. The main objective of this contribution was to show **the importance of preparing a dashboard to monitor target variables, with a precision indicator, during collection**. Although the basic idea is rather simple, the approach may not be applied to all surveys. As we have seen, the implementation of the dashboard and its statistical outputs allows a better identification of outlier values, a targeting of checks, and a prioritization of follow-ups for total non-response. As regards **prioritization of follow-ups for total non-response, further progress and testing are needed**.

Bibliography

Berger, Yves (2009), *Developing a scoring system for prioritising response-chasing*, PowerPoint presentation for Neuchâtel congress, June 2009
http://www2.unine.ch/webdav/site/colloque_deville/shared/documents/Berger.pdf

Brion, Philippe (2007), *Guidelines for balance between accuracy and delays in the statistical surveys*, INSEE working paper E2007/018
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/GUIDELINES_FOR_BALANCE_BETWEEN_ACCURACY_AND_DELAYS.pdf

Daoust, Pierre (2006), *Prioritizing follow-up of non-respondents using scores for the Canadian Quarterly survey of financial statistics for enterprises*, UN/ECE Work Session on Statistical Data Editing, Bonn, 2006
<http://www.unecce.org/stats/documents/ece/ces/ge.44/2006/wp.20.e.pdf>

James, Gareth (2007), *A strategy for prioritising non-response follow-up to reduce costs without reducing output quality*, Third international conference on Establishment surveys, Montreal, 2007
<http://www.amstat.org/meetings/ices/2007/presentations/Session72/James.ppt>

McKenzie, Richard (2000), *A framework for priority contact of non respondents*, Australian Bureau of Statistics paper. Available on OECD website:
<http://www.oecd.org/dataoecd/60/16/30890652.pdf>